UNIVERSITAT DE BARCELONA

Biodiversity assessment of marine benthic communities with COI metabarcoding:
methods and applications

Adrià Antich González

Biodiversity assessment of marine benthic communities with COI metabarcoding: methods and applications

Adrià Antich González

Doctoral Thesis

2022

UNIVERSITAT DE BARCELONA

Programa de Doctorat Biodiversitat

Biodiversity assessment of marine benthic communities with COI metabarcoding:

methods and applications

Adrià Antich González

Antich, A., 2022. Biodiversity assessment of marine benthic communities with COI metabarcoding: methods and applications. PhD Thesis. Universitat de Barcelona. 370pp

Tesis Doctoral

UNIVERSITAT DE BARCELONA

CEAB exploring aquatic ecosystems

Facultat de Biologia, Universitat de Barcelona

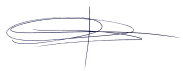
Programa de Doctorat de Biodiversitat

Biodiversity assessment of marine benthic communities with COI metabarcoding:

methods and applications

Avaluació de la biodiversitat de les comunitats bentòniques marines amb metabarcoding de COI: metodologia i aplicacions

Memòria presentada per Adrià Antich González per obtar al Grau de Doctor per la Universitat de Barcelona



Adrià Antich González

Departament d'Ecologia Marina, Centre d'Estudis Avançats de Blanes (CEAB)

Consejo Superior de Investigaciones Científicas (CSIC)

Juliol 2022

Director de tesi
Dr. Xavier Turon Barrera
Centre d'Estudis Avançats de Blanes (CEAB)

Directora i Tutora de tesi
Dra. Creu Palacín Cabañas
Universitat de Barcelona (UB)

Director de tesi
Dr. Owen S. Wangensteen Fuentes
The Arctic University of Norway (UiT)

A mi querida Antonia

Agraïments

Aquesta Tesi és el final d'una etapa, els fruits d'una feina que em sento meva i que, a la vegada, també és gratament compartida. Quan passeu aquestes primeres pàgines, podreu llegir el que he estat fent a Blanes i al mar aquests anys. En els següents capítols, espero que els resultats que hem pogut generar es transmetin i que serveixin per fer créixer el coneixement de la nostra societat. Però abans, em reservo aquestes línies una mica més personals.

Seria obviar l'elefant en l'habitació si no confessés que la pandèmia ha tingut un gran impacte en la realització d'aquesta Tesi. Uns anys que, per molta gent, han sigut molt durs en l'àmbit personal i laboral. Pel camí he perdut l'oportunitat de fer les estances que m'hagués agradat, col·laboracions que no han sorgit i congressos que no s'han celebrat. Res comparat amb les relacions personals que hem hagut de posposar, festes anul·lades, vacances no planejades i, sobretot, el dolor de dir adéu a qui ens ha deixat. Tinc la sort que heu sigut molts els que m'heu ajudat, sense voler-ho, a superar psicològicament aquest període. Sense vosaltres, que m'heu aplanat el camí, aquest hagués sigut una paret difícil d'escalar. De totes les persones amb les quals interactuem al llarg de la nostra vida, algunes deixen llavors que amb el temps creixen per formar les arrels de qui som. Aquesta Tesi ha sigut possible gràcies a totes aquelles persones que han plantat grans i petites llavors que, en major o menor mesura, han ajudat que l'arbre donés aquest fruit.

Primer, gràcies a tots els Ceabins i Ceabines. Quan arribes a un lloc nou, on la gent ja es coneix, on ja hi ha amistats formades i on el primer dia et presenten desenes de noms que et costa recordar (sóc d'aquestes persones fatals pels noms), no esperes ser tan ben acollit per tothom en tan poc temps. M'heu fet sentir un més. La pandèmia ha canviat les nostres dinàmiques. El teletreball i les mascaretes podrien haver sigut les barreres que trenquessin les sanes relacions personals que es respiren al centre, però

per sort no ha sigut així. Les ganes de poder veure-us per fer el cafè i dinar amb vosaltres és una sort que pocs llocs de feina et fan sentir. Espero tenir la sort de poder compartir moltes més festes en el futur. Al cap de pocs mesos d'entrar ja em vau convidar a un fieston padre on vaig ballar, riure i gaudir amb vosaltres. La festa de la defensa de la Marta (un 17 de gener que se'm quedarà al cap, ja que va ser l'última que vaig tenir abans de la pandèmia) va ser increïble. A més fa poc hem tingut l'oportunitat de repetir experiència i hem demostrat que aquests anys no ens han fet baixar el ritme. Ja tinc ganes de les pròximes!!!!!! Alejandro, Buñuel, Cèlia, Cris, Eli, Ibor, Itziar, Jana, Joan, Mar, Mario, les Martes, Mateu i Elena, Laura, Xavi, Vicente, Pol, Victor... la llista ocuparia pàgines. Gràcies a la Teresa, la Candela, al Kike i a la Emma amb qui vaig poder gaudir d'una de les millors campanyes de la meva vida. Jorge m'encanta l'energia que transmets. Roger, quins riures després del capítol *fast and furious*. Marta Turon, et seré eternament agraït per acollir-me allà al NORD amb majúscules. I... parlant del NORD, voldria agrair, en aquest petit parèntesi, a les persones que vaig conèixer i amb les quals vaig conviure per aquelles terres. Luke, Sandra, Laia, Ciulia, Federica, Andrea, Greta, Paul and the others. Being there was hard for me, but you have made me survive the true winter. All of you have made my days there lighter during the northern night. You have been the warmness in the coldness. Moltes gràcies a tots.

Durant aquest temps també he tingut l'oportunitat de co-tutoritzar dos treballs de final de grau. Hi ha, però, un d'ells que m'ha marcat especialment. Si en el futur tinc la sort de poder dedicar-me ni que sigui parcialment a la docència no oblidaré a la persona que durant el primer any d'aquesta història ha fet créixer aquest desig. Mireia, ha sigut un plaer ajudar-te en tot el que vaig poder i no saps el feliç que em va fer les paraules d'agraïment que em vas donar (potser eren mentida, però no cal que m'ho confessis XD). Et desitjo i auguro un total èxit en el futur en tot el que et proposis.

M'enduc molts bons records i experiències que gràcies a aquesta feina he pogut anar acumulant i que ja es quedaran per sempre en mi. La vida, així i tot, ha sigut igual o més emocionant fora de la bombolla de la Tesi i això es mereix, o més aviat requereix, un reconeixement, ja que, sense les persones

que m'heu envoltat a casa aquests anys, això hagués sigut impossible.

Aquests anys he tingut la sort d'omplir un gran buit que m'ha ajudat molt mentalment durant aquests anys, el teatre. Gràcies a la Laia per fer reviure aquestes ganes que feia temps que tenia amagades. Tant de bo hagués conegut escena 25 abans, però gràcies a tu vaig poder esprémer l'últim any al màxim i em va donar ales per poder fer el pas a buscar un racó on poder aprendre i experimentar què és posar-se en la pell d'un personatge. Gràcies a la Bibiana per fer-me gaudir, fer-me créixer i fer-me plorar. Has sigut la millor professora que podríem haver tingut. Sempre t'hem tingut com una més del grup, com una amiga que ens ha guiat en aquest camí. Et desitjo moltíssima sort en els teus futurs projectes i vull que sàpigues que l'escola s'ha quedat en els nostres cors per sempre. A totes les meves companyes d'escenari, Marina, Anna, Zoraida i Albert, a les quals he tingut l'oportunitat de conèixer aquest últim any, i Annes i Marina, amb qui he pogut gestionar emocionalment un confinament que hagués sigut molt pitjor sense vosaltres. Juntament amb la Diana i un cos de ball fenomenal hem creat una família que desitjo que es torni a reunir adalt de l'escenari aviat. Ha sigut una de les experiències més flipants de la meva vida haver fet de Billy al vostre costat. Els dimarts han sigut molt importants per la meva gestió emocional de tot plegat aquests tres anys. És curiós que aquesta Tesi hagi durat el mateix que el meu període en l'escola. Però si tinc clar que després del doctorat hi ha camí, també ho tinc clar pel teatre. Fura!!!

Hay un grupo de locos y locas, una secta unida por la sangre a quien quiero dar las gracias. A toda mi familia la cual, sin poder escogerla, siento como el premio de lotería más grande que jamás tendré. Comidas familiares, campings, Romanyà y Feitús... me habéis demostrado que mis padres me criaron en un árbol de raíces profundas, tronco firme y copa enorme. A "la Tribu", a mis primos y tias a los que tengo como familia y amigos. Sois un pilar indispensable en mi vida. A ti Carmen, que eres como mi segunda madre, supongo que está en los genes. Rubén, Iván y Carlos (Arnau tiene su apartado más abajo), mi infancia en Feitús lo ha convertido en mi paraíso en la tierra. Ana, Chus y Nil sois, para mí, las eternas juventudes, los años me han acercado a vosotras y también me habéis guiado en mi adolescencia.

Mauricio y Juan Antonio, solo con imaginar las batallitas que contáis ya me empiezo a reír. Martí, Magí i Oriol, sou i sereu els peques per nosaltres. Ens gasteu tota l'energia però ens ompliu el cor. Meritxell, has criat a dues criatures que tot i les baralles que puguin tenir, es nota l'estima que es tenen per ells i per la seva família. Manel, tinc gravat al cap quan em vas portar en moto quan era només un marrec. I tots els meus avis i abuelos a qui he tingut la sort de conèixer i estimar. Amb tots vosaltres m'uneix la sang, però també heu portat a gent amb la qual he crescut a la família i a qui no vull oblidar perquè per mi són i seran sempre uns més. Juanjo, Xavi, Montse, Mei, sois mis tíos y lo seréis siempre, Anabel, Marina, lo siento por vosotras, que la entrada a esta secta haya sido por vía de dos energúmenos como lo son mis primos. Sin embargo, de toda mi familia, esta tesis no solamente se la puedo agradecer, sino que dedico a mi Abuela, eres la que un día se fue para estar a mi lado para siempre. Ojalá me hubieras visto hacer realidad este sueño. No tengo palabras para describir quanto te he querido y cuanto siento ausencia cuando nos reunimos todos, te llevo en el corazón por siempre. Ojalá poder abrazarte por última vez. Esto es también tuyo.

Ara sí que ja et toca a tu, Arnau. Crec que tenim una sort molt gran de tenir-nos com a germans i amics. Espero que tinguis en mi un suport per tot el que vulguis en el futur igual que el que tu m'has fet sentir. La meva infància i adolescència ha sigut al teu costat, Romanyà, Feitús i tots els viatges que hem fet han sigut al costat d'un gran amic. Afegeixo aquí a la meva cunyi preferida, Jessica. Els dos em teniu pel que necessiteu, però us quedeu amb la Duna i el Roc vosaltres, entesos?

Als meus Pares que no només m'han donat la vida, sinó les ganes de voler-la estudiar. Tu Papa que m'has ensenyat, en totes les discussions que hem tingut, que tenir la capacitat d'enraonar no és només una característica que ens fa humans, sinó que, a més, és molt divertit. Espero que gaudeixis estar en aquesta Tesi com si fos la primera en la qual hi ets. A tu Mama, que si sóc biòleg és culpa teva. M'has donat la curiositat per la natura des de ben petit. Un dels grans tresors que conservo és l'herbari que em vas ajudar a fer i que va ser segurament l'inici del meu amor per la Biologia.

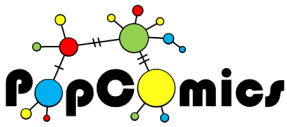
Sou uns referents en la meva vida. Us estimo molt i gràcies per estar sempre que ho he necessitat!

Per arribar on he arribat el pas per la universitat, clar està, era indispensable. Fer la carrera dels meus somnis no hagués sigut ni la meitat de plaent si no hagués estat acompanyat d'unes persones amb les quals he passat els millors anys de la meva vida. A l'Estela, la millor ex que pugui tenir amb qui he compartit moments especials com a parella i amics; al Pau, el meu co-entrenador; al Giralt, hippie amb qui he passat els grans moments de la uni; a l'Uri, amb qui compto per qualsevol festival i podcast futur; a la Júlia, a qui només demano que entre viatge i viatge em pugui dedicar una estoneta per una birreta; al Toni, de qui els "piropos" em fan que em creixi l'ego (;p); a les Paules, una de les quals estic feliç de tornar a tenir a la meva vida i que sempre serà de bio tot i que s'equivoqués en canviar-se a biomed; i la resta, si bé avui en dia ja gairebé només ens veiem un parell de cop a l'any, espero que les masies amb vosaltres segueixin per sempre, caps de setmana on reviure els dies en que conviviem el dia a dia. Tot i això hi ha tres noms entre vosaltres especials. Javi, eres una persona realmente especial, a secas y para mi jajaja, te tengo como uno de mis grandes amigos y me alegro mucho de haberte conocido y tenerte como entrenador. Jesús, me fui contigo a Canarias y me alegro de que hayas venido conmigo a Blanes. Espero que nuestra amistad dure toda la vida y que en un fututo seamos de los investigadores cuñaos que no paran de decir que se conocieron en la carrera. I tu Laura, amb el perdó de la resta, amb la qual he tingut, des que vas sortir del metro a fer-te amiga dels d'història, com la peça clau de la meva vida. Molta gent m'assentarà arrels, però tu ets l'aigua que m'ha fet créixer. A tots vosaltres us tinc com la família escollida.

I si al començament he donat les gràcies a la meva família artística, a la família de sang i a la família escollida, només em falta agrair enormement a la meva família acadèmica. Creu, em vas acollir al grup a finals de 2015 i tots aquests anys t'he tingut en una estima enorme. Tutora, directora però sobretot mentora. Som molts els que agraïm poder continuar tenint-te al nostre costat. Em quedarà sempre l'espina de no haver aconseguit ordenar el teu despatx. Owen, has hecho crecer en mí la curiosidad por

la bioinformática y la programación, lo cual no tengo duda que será parte de mi futuro. Te agradezco mucho haberme acogido en Trømso y creo que tener la oportunidad de hablar contigo es realmente productivo, tanto en lo puramente laboral como en alimentar la curiosidad por el mar. I Xavier, crec que ets el millor director de tesi que mai ningú podria tenir. Has sigut tant inspiració com referent per mi durant tot aquest temps i crec que sense tu això no hagués estat possible. Moltes gràcies per confiar en mi, gràcies per tenir en compte la meva opinió i gràcies per tota l'ajuda que m'has donat. Ha sigut un plaer mostrejar i anar de congrés amb tu. M'has donat l'oportunitat de desenvolupar vies que no estaven previstes des d'un inici i que m'han permès créixer en el que més m'ha interessat. Per desgràcia la teva manera de dirigir a la gent que treballem amb tu no és la norma. No són pocs els casos en què una mala relació doctorand-director suposa un punt crític en la realització de la tesi, però no ha sigut el cas en aquesta. M'has fet sentir un més i realment ets un exemple a seguir i que espero seguir.

Finalment, donar també les gràcies a la Dolors Vinyoles a qui demano disculpes per tots els maldecaps que li hagi pogut ocasionar, però que ens ha solucionat tots els dubtes i problemes que han sorgit durant els últims mesos d'aquesta Tesi.



This Thesis has been supported by a grant (PRE2018-085664) associated to the PopCOmics project (CTM2017-88080-C2-1-R) funded by the Spanish Ministry of Science and Innovation (MCIN/AEI/ 10.13039/501100011033) and by "ESF Investing in your future". The research performed has been funded by projects PopCOmics and MARGECH (PID2020-118550RB, MCIN/AEI/10.13039/501100011033), and by project BIGPARK (OAPN, 2462/2017) of the Spanish Ministry of Ecological Transtion and Demographic Challenge.

Abstract

Ecosystem biomonitoring is crucial for proper management of natural communities during the Anthropocene era. With the advent of new sequencing technologies, DNA metabarcoding has been proposed as a game-changing tool for biomonitoring. In this Thesis we plead for the use of metabarcoding of a highly variable marker to infer not only the interspecies but also the intraspecies variability to assess both biogeographic, at the species level, and metaphylogeographic patterns, at the haplotype level. We focused on highly complex hard-substratum benthic littoral communities. The term "Metaphylogeography", coined in this Thesis, refers to the study of phylogeographic patterns of many species at the same time using metabarcoding data. However, as of the start of this Thesis, only a few studies had tested the metabarcoding method to directly characterize the whole eukaryotic community in highly diverse benthic ecosystems. This required to set up and calibrate methods for these communities as a prior step.

We first evaluated both the sampling methods and the bioinformatic pipelines. We assessed the viability of detecting the environmental DNA released from the benthic community into the adjacent water layer using metabarcoding of COI with highly degenerated primers targeting the whole eukaryotic community. We sampled water from 0 to 20m from shallow rocky benthic communities and compared the DNA signal with the results obtained from metabarcoding directly the benthic communities by traditional quadrat sampling. We also designed a pipeline combining clustering and denoising methods to treat metabarcoding data of COI. We considered the entropy of each codon position of this coding fragment both to improve the detection of spurious sequences and to calibrate the best performing parameters of the software used. In addition, we created our own denoising program, DnoisE, to incorporate information on the codon position. This new code and parameter calibration were required as the commonly used bioinformatic pipelines had been designed and tested mostly for less variable ribosomal

fragments and, particularly, in prokaryotes.

Results showed that the DNA signal from the benthos decreased with the distance but was too weak for a correct assessment of benthic biodiversity. The proportion of eukaryotic DNA sequenced was also very low in water samples due to the amplification of prokaryotic DNA. We thus concluded that the benthos must be sampled directly to properly assess its biodiversity composition. The new bioinformatic developments allowed us to propose new methods for processing metabarcoding reads, combining clustering and denoising steps, and to set optimal values for the parameters used at each step. These contributions effectively expanded the field to the novel analysis of inter- and intraspecies genetic variability with metabarcoding data.

Finally, we applied this methodology to 12 localities of the Western Iberian Coast along two well studied fronts, the Almeria-Oran Front (AOF) and the Ibiza Channel (IC). We analysed the species and haplotypes using the COI barcode. From a biogeographical perspective, the AOF had a strong effect in separating regions, while IC effect was less marked, but still half of the MOTUs were found in only one side of this divide. For the metaphylogeographic analysis, only 10% of the MOTUs could be used. However, they showed a good separation between populations of the three regions with a strong effect of the AOF break. The IC, on the other hand, seemed to be more a transitional zone than a fixed break.

This Thesis laid the ground for the efficient use of metabarcoding in the biomonitoring of benthic reef habitats, allowing community composition, β-diversity, and biogeographic patterns to be analysed in a fast, repeatable, and cost-efficient way. We also developed the metaphylogeography approach as a new tool to assess population genetic structure at the community-wide level.

Contents

Chapters

Chapter 1

General Introduction

1.1 Marine benthic environments

Among all known planets in the universe, what makes Earth special and unique is life. Life is shown in every corner of our planet in multiple ways. Such variety of living forms is what we know as Biodiversity. Across all ecosystems in the biosphere, shallow rocky benthic communities are considered to be one of the most diverse habitats on our planet (Agardy et al., 2005; Reaka-Kudla, 1997). These ecosystems composed by multiple three-dimensional structures create a vast variety of niches for organisms to thrive in.

Humanity has been connected to the sea for millennia. From food to other materials, the sea has been an important source of different goods. Coastal waters in particular have been the most exploited due to their proximity. After the industrialization, deep waters and their benthos have also been exploited for the extraction of minerals, global transport through ocean has increased, and the exploitation of the shallower waters has intensified. Fishing, urbanization and anchoring are some of the direct impacts on these environments but also other indirect stressors are present such as pollution, industry, alien species and, of course, global climate change. Such human stressors are present in almost every corner of the planet and have brought geologists to define the actual era as the Anthropocene, a period characterized by the human impact in the planet (Crutzen and Stoermer, 2000).

It is clear that all environments are affected by the Anthropocene, however, the impacts are not equal both in type and magnitude. To properly evaluate how ecosystems change as a response to such stressors it is

mandatory to analyse the state of them before, during and after the impact, but such data are not always available. Naturalists and scientists have focused mainly on terrestrial ecosystems in detriment of the sea for the last centuries (Costello and Chaudhary, 2017). The higher accessibility engages the interest of the society in ecosystems such as rain forests leading to a higher comprehension of how they change, from an almost non-disturbed to a totally impacted state. The "out of sight, out of mind" principle makes society less aware of problems and impacts on the marine realm. In addition, marine benthic communities are difficult to survey (Duarte, 2000; Gaither et al., 2022). Scuba diving, echo sounding, dredging or satellites are some of the methods used to study marine communities. Such methods are relatively new and require specific equipment and training which are usually expensive. Hence data on the state of marine environments previous to any impact usually do not exist and scientists have to infer the pristine state of the systems analysing less disturbed areas for comparison to those impacted. But even when sampling is performed, the high complexity of marine benthic communities makes the biodiversity characterisation challenging (Wangensteen and Turon, 2017). Such arduous challenge stems from a vast variability of living forms of different complexity, from unicellular to pluricellular organisms, from solitary to modular organisation, and the associations of symbiosis, parasitism or competition for the same niche, that have rendered these communities largely inaccessible to exhaustive biodiversity assessments.

Yet biomonitoring marine communities is mandatory to understand how and why they change (Baird and Hajibabaei, 2012). Local experiments are often easier to perform logistically but marine habitats can differ markedly even in nearby regions. Although practically all sea waters are connected, physical factors such as salinity, temperature and light originate different regions separated by invisible barriers (Toonen et al., 2016). Detection of these discontinuities allows to delimit regions and to study possible genetic fluxes. Fast and robust methodology is essential for efficient biomonitoring (Baird and Hajibabaei, 2012) and focusing the efforts on these interesting areas may offer a more holistic view of how different marine biomes interact while they evolve.

1.2 Biodiversity and how we survey it

Living organisms interact with themselves and with the surrounding environment. Thus, when studying the processes that occur in any nook of the earth surface, biotic and abiotic processes have to be taken into account. Biodiversity assessment is at the basis of any biological and ecological research, of basic or applied focus. To describe biotic processes or ecological interactions the characterisation of the community is mandatory (Bakker et al., 2019). Traditionally, inventories have been performed using the morphological characters that distinguish species to identify and classify them. Taxonomists have been and still are the specialists that perform this work. However, to reach the knowledge required to be a specialist in any taxon requires a lot of time and training, and to characterise the whole community composition of a sample, many taxonomists are necessary. Hebert et al. (2003) estimated that approximately 15,000 taxonomists are required in perpetuity to identify the whole canopy of life. This contrasts with the marked shortage of taxonomists in recent years, known as the "taxonomic impediment" (Dupérré, 2020; Engel et al., 2021; Giangrande, 2003). Thus, at the time when biodiversity assessment is more urgent than ever due to human impacts and loss of species, the task is overwhelmingly out of reach for the dwindling taxonomic workforce.

Moreover, morphological traits are often hard to assess even for taxonomists due to plasticity, collection and preservation artefacts, or the existence of different life stages. However, as morphological traits are determined by its codification in the DNA, it is expected that this molecule can also be used for species identification. Over the last years scientists have thus used DNA as a complementary tool to assess biodiversity (Gaither et al., 2022).

1.2.1 From DNA Barcoding to Metabarcoding

The identification of species using DNA barcoding was first proposed by Hebert et al. (2003). This technique is based on extracting DNA to amplify and sequence certain genes or gene fragments, known as barcodes (a catchy word that evokes the barcodes commonly used to identify items in any store).

When this is performed on identified specimens, a database of sequences is built. Once the database has enough coverage, this information can be used for reverse taxonomy, by comparing the sequences obtained from unidentified individuals with them to assign a proper name to the sample. As pointed out before, traditional taxonomy is a harsh task that requires thousands of experts to cover just a low percentage of the whole biodiversity. However, if this expert knowledge can be encapsulated in accurate databases of barcodes, then assessments can be based on the analysis of DNA, which can be performed bioinformatically without specific knowledge of the morphological traits that distinguish species. In the last decades, scientists have complemented biodiversity analyses with the generation of DNA barcodes, often through huge international initiatives such as the international Barcode of Life (iBoL) consortium. Notwithstanding, not all genes are susceptible to be used as barcodes. They have to fulfil some conditions:

The barcode gap. This is the lack of overlap of the intra- and interspecies variability of the candidate fragment (Meyer and Paulay, 2005). Such gap allows to cluster close sequences to operational taxonomic units (OTUs, or Molecular Taxonomic Units, henceforth MOTUs) differentiated from other clusters. If the barcode gap is good enough, such MOTUs can be used as a proxy of species.

Primer design. Barcodes are variable fragments of DNA, however, they must be flanked by conserved regions to be used as binding points for the primers that initiate the amplification reaction. To amplify the barcode region using the Polimerase Chain Reaction (PCR), primer design is crucial. Primers targeting highly conserved positions and including variable bases where necessary to amplify wide taxonomic groups while other primers can be designed to target particular groups. Scientists distinguish between universal and specific primers depending on their use for higher or lower taxonomic ranks.

Barcode length. Since not all nucleotide positions in the genome are equally affected by natural selection, some positions are more conserved

Table 1.1: Taken from Taberlet et al. (2012a). High-quality reference databases overview.

Database name	Organisms covered	DNa region targeted	Website	Reference
SILVA	Bacteria, Archaea, Eukaryota	16S/23S rRNA genes 18S/28S rRNA genes	www.arb-silva.de	Quast et al. (2013)
RDP	Bacteria, Archaea, Fungi	16S rRNA genes 28S rRNA gene, ITS region	rdp.cme.msu.edu	Cole et al. (2014)
Greengenes	Bacteria, Archaea	16S rRNA genes	greengenes.lbl.gov	DeSantis et al. (2006)
EzTaxon	Bacteria, Archaea	16S rRNA genes	www.ezbiocloud.net	Chun et al. (2007)
PR2	Protists	18S rRNA gene	ssu-rrna.org/pr2	Guillou et al. (2012)
PhytoREF	Photosynthetic eukaryotes	Plastidial 16S rRNA gene	phytoref.org	Decelle et al. (2015)
BOLD	Animals, plants, Fungi	Mitochondrial COI Chloroplastic rbcL, matK ITS region	www.boldsystems.org	Ratnasingham and Hebert (2007)
UNITE	Fungi	ITS region	unite.ut.ee	Kõljalg et al. (2013)
MaarjAM	Arbuscular mycorrhizal Fungi	Multiple	maarjam.botany.ut.ee	Öpik et al. (2010)
AFTOL	Fungi	Multiple	www.aftol.org	Celio et al. (2017)

among species and others are more variable. For example, in coding genes, the third position in the codon is more variable than the first or the second because a change in the third position usually does not change the aminoacid that is coded for. However, a change in a second position will probably change the aminoacid and thus the protein composition, being more prone to being affected by natural selection. Likewise, in ribosomal genes, some nucleotide positions determine the tertiary structure and are less variable. Thus, the length of the barcode should include a sufficient number of variable positions to allow reliable species assignment.

Proper barcode for proper group. Barcoding is not a "one gene fits all" field, and depending on the target group different barcodes have been used. Certain genes can only be found in some phylogenetic branches and can be reliable for use as barcodes in them, such as for instance the RuBisCo and other chloroplast genes in plants. Other genes can be found in all organisms but have more resolution in some groups than in others, which determines their suitability as barcodes. For instance, ITS has been broadly used in fungi, 12S in vertebrates and 16S in bacteria thanks to their resolution in these groups. The mitochondrial gene COI has been used as a general barcode for eukaryotes in general although for some groups there are issues with the design of adequate primers. Finally, the coverage of the targeted taxa in the reference Databases will be also crucial to decide which barcode to use (tab. 1.1).

During the first decade of the present millennium, DNA barcoding gained

interest rapidly due to its potential to identify any species using a robust and easy to perform technique. However, there was still an impediment for its use in ecological studies where many individuals should be sequenced at the same time. The advent of the new sequencing technologies, known as High-Throughput Sequencing (HTS) or Next-Generation Sequencing (NGS), allowed sequencing a heterogeneous DNA template (unlike Sanger sequencing) and thus the obtention of sequences of pooled amplicons. This fact, and the reduction in cost per sequence, brought about what is known as DNA metabarcoding. If DNA barcoding is based on sequencing the targeted fragment of a single individual at a time, the new sequencing technologies allow to do it with several hundreds, thousands, or millions of individuals simultaneously, thus the name Meta-barcoding. Metabarcoding was originated and is commonly used by microbiologists, and indeed the method is well established and standardised with accepted protocols for the analysis of prokaryote communities. However, in the field of ecology, where eukaryotic organisms are the common target, factors associated to the environment or the taxonomic group make it difficult to standardise the method as in microbiology.

1.2.2 Environmental DNA

In opposition to prokaryotic forms, eukaryotic communities differ strongly in organization, both in terms of complexity (from uni- to pluricellular organisms), individual organization (i.e., single individuals, clonal forms, or symbiotic associations among others), and size, from μm in some unicellular organisms to meters in large mammals as blue whales. The targeted community will determine the process in metabarcoding studies for eukaryotic communities. Capturing the whole organisms or tissue samples is sometimes arduous. However, DNA can be found in multiple forms in the environment and not only inside living organisms. The environmental DNA (eDNA) has been described as the DNA that can be obtained from environmental samples (i.e. soil, water, faeces, etc.) without isolating any organism, as opposed to community DNA (comDNA) that is obtained from bulk samples of previously isolated organisms (Deiner et al., 2017; Taberlet et al., 2012b). The eDNA is originated from living organisms but can be found in the envi-

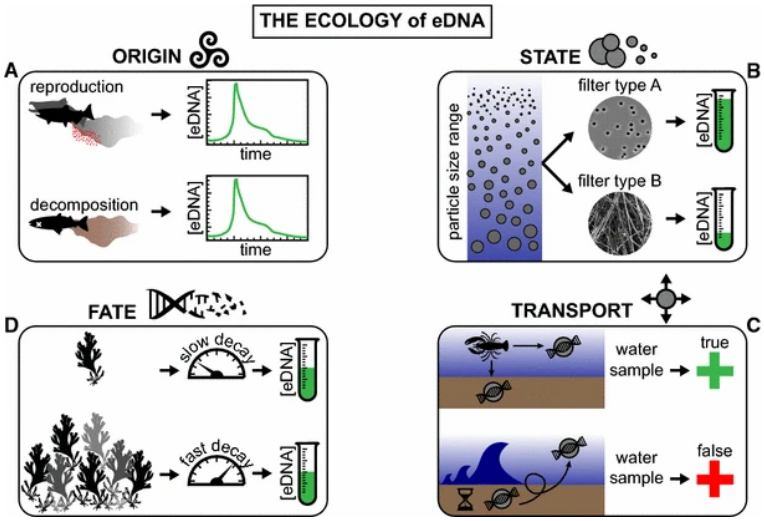


Figure 1.1: Taken from Barnes and Turner (2016). eDNA ecology affects population inferences. **a** eDNA from reproduction and decomposition could produce similar temporal patterns despite different origins. **b** Different filter types could yield different eDNA concentrations that reflect particle size classes rather than population size differences. **c** Resuspension of old sedimentary eDNA could produce false inferences of presence after organisms are gone. **d** Different environmentally-mediated eDNA decay rates could confound inferences about population size or biomass from eDNA concentration influence.

ronment as mucus, exudates, or dead cells that have been separated from their individual. This DNA can be encapsulated inside cells or organelles but can also be found in free form, either adsorbed to other molecules or not (Nagler et al., 2022).

One important consideration that scientists point out when using eDNA is what has been named as the ecology of the eDNA (Barnes and Turner, 2016). Degradation and transport of DNA in an environment is highly dependent on its nature (fig. 1.1). Thus, if the molecule is found inside the living organism, transport and degradation will be affected mainly by the behaviour of the individual with little degradation of this DNA (Creedy et al., 2022). If DNA is encapsulated inside a cell that is shed to the environment, abiotic factors will highly affect the transport but degradation will not be important. Finally, when the DNA is in free form (extracellular), environmental factors will be determinant of its fate and

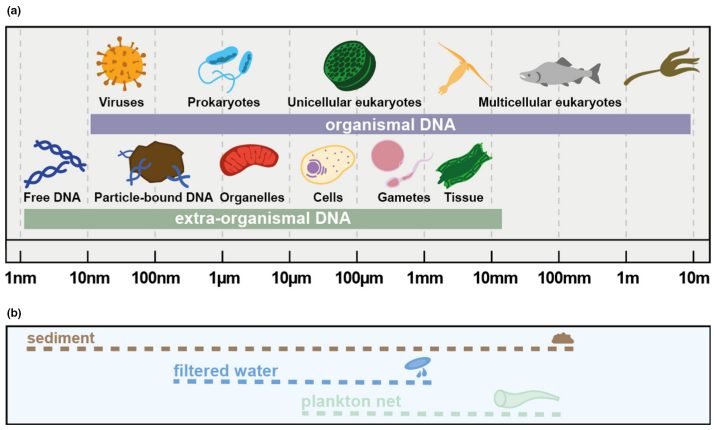


Figure 1.2: Taken from Rodriguez-Ezpeleta et al. (2021). **a** Types of environmental DNA (organismal and extra- organismal, including extracellular) with possible sources and approximate size ranges. **b** Illustrative examples of sampling methods with intended captured particle size ranges

degradation (Nagler et al., 2022). It is important, however, to recognize that every sample from the environment has a mixture of extra- and intra-organismal DNA (Rodriguez-Ezpeleta et al., 2021, fig. 1.2) and the target group, sampling method and sample processing (f.i., filtering or separating the organisms manually) determine the relative proportion of one or another in the sample. Separating individuals from the environmental sample will also reduce the proportion of nontarget species and will enlarge the scope for complementing, refining and/or validating the community composition data obtained with metabarcoding against traditional morphology based data (Creedy et al., 2022). The primer design for very different organisms and different DNA degradation rates depending on the environment (which leads to shorter DNA fragments available) is another factor that makes the metabarcoding of eukaryotic eDNA hard to be standardized.

1.2.3 COI Metabarcoding

The need of standardised DNA metabarcoding methods in the field of ecology has been a major concern in the scientific community applying this

technique to diversity assessment. Several barcodes have been proposed but those found in the mitochondria have gained popularity. The lack of introns and the haploidy make the mitochondrial genes more amenable to be used as DNA barcodes. Among all these, Cytocrome Oxidase I (COI) is the most commonly used and has been proposed as the main barcode for metazoan community assessment using metabarcoding (Andújar et al., 2018; Creedy et al., 2022; Porter and Hajibabaei, 2020). Other markers have been used to analyse specific groups of taxa, such as 12S for vertebrates, ITS for fungi, or chloroplast genes for plants. However, the advantages of COI as a community barcode for eukaryotes and especially for metazoans have been pointed out by many authors (Andújar et al., 2018; Creedy et al., 2022; Hebert et al., 2003; Porter and Hajibabaei, 2020).

The region of COI most commonly used for barcoding is the so-called Folmer fragment (Folmer et al., 1994). This fragment of ca. 658 bp is amplified using very robust universal primers. However, for some groups as nematodes, gastropods and echinoderms the efficiency has been reported to be poor (as reviewed in Leray et al., 2013). For metabarcoding, however, a shorter region (approximately the 3' half of the Folmer fragment) has been proposed. This region of ca. 313 bp, known as the Leray fragmentLeray et al. (2013) is amplified using a new forward primer internal to the Folmer barcoding fragment. Wangensteen et al. (2018b) suggested an improvement of this primer to increase its coverage. Vamos et al. (2017) also designed primers for even smaller fragments to amplify highly degraded DNA. Elbrecht and Leese (2017) made a comparison between different primers amplifying different regions of the Folmer region for 15 freshwater invertebrate groups. Indels are rare in COI as they often result in shifts in the reading frame (Hebert et al., 2003) and this favours the alignment of the sequences which will be necessary for some steps of the data analysis, for instance to compare sequences to databases to assign them to a certain taxon. But COI has also characteristics that can be used during the filtering steps of the bioinformatic process. As pointed out before, since COI is a mitochondrial gene without introns, the Folmer fragment, or subsets thereof, are coding regions. We can use this information to detect amplification or sequencing artefacts. For instance, as no codon stops are expected in the fragment of interest, if any

is found the sequence can be labelled as erroneous. In addition, for certain groups as Metazoans, some aminoacids are conserved in the Folmer region and such information can also be used to detect erroneous sequences. On the other hand, nuclear pseudogene copies of COI known as NUMTs are of high concern so they can inflate diversity (Andújar et al., 2021; Schultz and Hebert, 2022; Song et al., 2008).

However, one of the most important pros for COI is the huge database of identified sequences available. To obtain the taxonomic information of each sequence, the metabarcoding pipeline uses algorithms to match the query sequence against a database to assign it to a given taxon (at species or higher taxonomic ranks). Particularly, the Barcode of life Database (Ratnasingham and Hebert, 2007) so far contains over 14.4M COI sequences of >500 bp in length representing over 334.4K species.

Moreover, COI sequences have been extensively used in studies of population genetics and phylogeography of terrestrial, freshwater, and marine organisms (Avise, 2009; Emerson et al., 2011). Hebert et al. (2003) pointed to the phylogenetic signal of COI as a good advantage to consider it as a barcode candidate. COI metabarcoding has, therefore, the potential of creating databases containing an untapped reservoir of intraspecies variation that can allow characterizing intra- and interpopulation genetic features of many species simultaneously and Andújar et al. (2018).

1.3 Need for standardisation of methodology

1.3.1 From field to laboratory

Although hard bottom communities are among the most diverse habitats in the world, the use of metabarcoding in such ecosystems has been poorly developed. Before the beginning of this Thesis only Wangensteen et al. (2018b) had applied metabarcoding to samples of rocky benthic communities (1.3) using the same collection methods used to infer biodiversity of such environments with traditional methods (namely quadrat sampling). Metabarcoding has the potential to solve the problem of assessing these complex communities even though the invasiveness of the sampling method

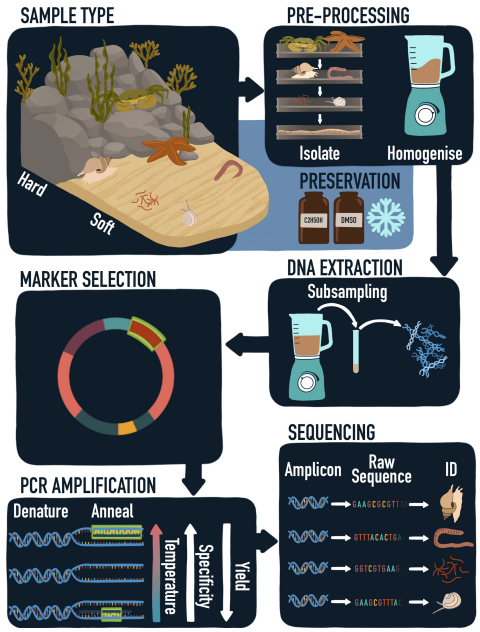


Figure 1.3: Redrawn from van der Loos and Nijland (2021). Scheme of the use of the metabarcoding approach to marine benthic communities.

is still a problematic issue. An obvious choice would be to sample the water in the vicinity of the benthos to recover benthic DNA for metabarcoding applications. Such indirect sampling methods using environmental DNA (eDNA) have proved to be appropriate to detect species without capturing directly their individuals.

DNA can be spread to the environment in multiple ways, such as faeces, skin cells or mucus, among others. Sampling soil or water samples from terrestrial environments, many experiments have demonstrated that this DNA can be properly amplified and thus analysed using molecular methods such as qPCR (for quantification of single species) or metabarcoding (multiple species assessment). In the marine realm several studies have analysed sediment samples (even from the deep sea) for characterization of soft-bottom assemblages. The use of water metabarcoding, however, has largely been restricted to analysing fishery stocks or phytoplankton monitoring for bloom assessment using water samples.

The potential usefulness of water metabarcoding to detect the eDNA

originated from benthic organisms has not been analysed as no study had targeted the water immediately adjacent to the marine benthos. Only in freshwater ecosystems Hajibabaei et al. (2019) and Gleason et al. (2021) performed experiments concluding that eDNA found in water samples retrieved a poor signal from the benthos. Moreover, the size of the extra-organismal DNA forms correspond to the size spectrum of many non-eukaryotic organisms (fig. 1.2) whose DNA can also be amplified by universal primers, reducing the proportion of target DNA in the amplification process by PCR. Testing and analysing the use of eDNA from water samples adjacent to benthic communities in marine environments is therefore crucial to determine the best techniques for assessing the biodiversity of such communities minimising the impact.

1.3.2 Bioinformatic pipelines

Bioinformatic treatment of sequence data is mandatory to turn uninterpretable raw data, such as the flood of sequences that the HTS platforms return, into understandable and useful information. Besides, during the laboratory and sequencing processes, many sources of errors add bias to sequence data. As of the beginning of this Thesis most of the bioinformatic pipelines (pipeline refers to the sequential run of different software in an informatic process) and the software designed to process all these data were designed for ribosomal markers such as 16S or 18S, adapted from the pioneering approaches developed by microbiologists (Creedy et al., 2022).

The main steps of the metabarcoding bioinformatic pipeline process are (fig. 1.4):

1) De-replicating and de-multiplexing sequences. The sequencing process reads each string of the double chain of the DNA separately. Hence the forward and the reverse reads are provided by the sequencing services in two different files. For each read both strings have to be merged using the overlapping segments to generate paired reads, also called amplicons. Since each amplicon can appear many times, identical reads are merged together in so-called unique sequences, keeping a count of the number of reads that are merged (de-replicating). During this procedure some quality

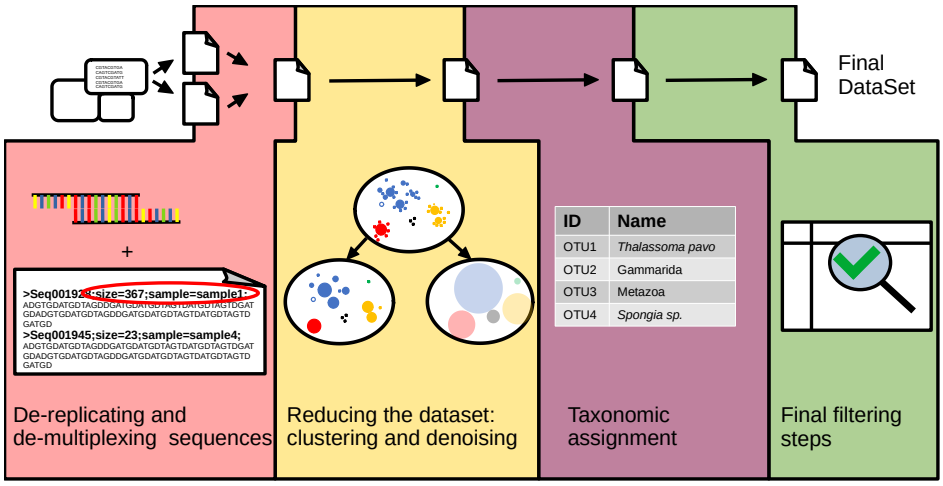


Figure 1.4: Metabarcoding pipeline divided in 4 steps. 1) De-replicating and de-multiplexing of the raw sequences obtained. In this step some quality filters are also applied. 2) Clustering and Denoising delimiting the OTUs or the ESVs that will be used as taxonomic units for further analysis. 3) Taxonomic assignment of the units delimited in the previous step. 4) Final filtering steps such as removal of contamination or minimum abundance filtering.

filtering steps can be performed (for instance, using base call qualities, or using total amplicon length). Finally, different tagging procedures can be used to uniquely label the amplicons generated from each sample, allowing the partition of the total reads of each unique sequence into the different samples (de-multiplexing).

2) Reducing the dataset: clustering and denoising. The different amplicons obtained include real sequences and also erroneous sequences generated during the different processes of amplification and sequencing. In addition, as the COI marker is highly variable, each species is represented in the amplicon pool by different sequences. The clustering algorithms try to join similar sequences that are assumed to belong to the same biological entity. There are several strategies for clustering, based on fixed or variable distance thresholds, or on phylogenetic information. The result of the clustering process is the generation of a reduced dataset of so-called Operational Taxonomic Units (OTUs) or Molecular Operational

Taxonomic Units (MOTUs) of which only the most abundant sequence is retained as the representative of the cluster (thus clustering often results in loss of sequence information). The goal of the clustering process is that the MOTUs reflect as closely as possible the species-level diversity of the samples (thus ideally each MOTU corresponds to a species). During clustering, erroneous sequences are most often joined to the correct MOTU (as they have little divergence with the correct sequences from which they originate), but note that clustering is not per se a denoising procedure. Denoising, on the other hand, include methods that try to detect erroneous sequences, based normally in abundance structures and identity levels (as an erroneous sequence derives from a similar, more abundant sequence). Sequences labelled as erroneous are eliminated and their reads added to the correct sequences from which they supposedly originated. This results also in a reduced dataset of correct sequences, but only erroneous sequences are lost in this case. Note, however, that denoising procedures make no inference as to the underlying structure of biological units (species) present.

3) Taxonomic assignment. Specific software can apply diverse algorithms to compare sequences with databases (tab. 1.1) to assign taxonomic information to each unit (be it a MOTU or a correct sequence, see below). These algorithms are usually based on similarity values, phylogenetic information, or both.

4) Final filtering steps. These steps usually include corrections for the sequences that appear in the technical controls (blanks and negatives), as well as removing potential tag jumping artefacts (i.e., sequences assigned to the wrong sample) and performing minimal abundance filtering if deemed necessary. Additional algorithms can detect nuclear copies of the COI gene to eliminate them (i.e., metaMATE, Andújar et al., 2022, for NUMTs, Lopez et al., 1994)

The second step has been particularly controversial in recent times. While clustering and MOTU generation has been the usual method to infer biological units and eliminate erroneous sequences simultaneously, new software development has brought about powerful and robust programs

(DADA2, Callahan et al., 2016; UNOISE Edgar, 2016; deblur, Amir et al., 2017) to denoise sequence datasets and to retain only the true sequences (called ESV, exact sequence variants, Porter and Hajibabaei, 2018b; ASV, Amplicon Sequence Variants, Callahan et al., 2017; or ZOTU, Zero-radius Operational Taxonomic Units, Edgar, 2016). Eliminating erroneous sequences keeps all usable sequence information and facilitates inter-study comparisons, while MOTUs are not directly comparable across studies. This has led to the suggestion that exact sequence variants should replace MOTUs in metabarcoding studies (Callahan 2017). We contend, however, that this can be adequate for ribosomal genes for which the method was designed, but that for variable mitochondrial genes such as COI it is advisable to keep the correct sequences (as a proxy for haplotype diversity) but also the MOTU structure (as a proxy for species diversity) where they belong. There has been some confusion between these two different (philosophically and computationally) strategies, but we suggest to make the most of the two approaches, as they are not incompatible but complementary. Combining them we can obtain both the inter and intraspecies variability for ecological analysis.

1.4 Marine biogeography and phylogeography

Management and conservation of marine habitats in the Anthropocene is one of the main concerns of our present societies. However, the resources allocated to that purpose are scarce and thus the most efficient methods for biomonitoring must be used. To make the correct decisions, managers require to know as much as possible the communities that are meant to be protected. New methods to determine the temporal and spatial distribution of biodiversity in a fast and comprehensive way are needed in the face of fast ongoing change resulting from human activities.

The distribution in time and space of living organisms is studied by biogeography. Marine environments have an apparent continuity, lacking obvious hard barriers, but communities differ clearly across regions of the globe. Marine barriers are those physical discontinuities that separate regions and, given enough time, give rise to well differentiated biota (Avise,

2009). Costello et al. (2017) provided the most comprehensive analyses of marine biogeographic patterns by compiling data from 65,000 marine species available in public databases across all higher taxa.

However, these barriers not only separate communities composed by different species but can also hinder gene flow among different populations of the same species. The importance of studying how populations are distributed in space and their genetic connectivity is crucial to understand both the past and present but also the future of the communities(Beng and Corlett, 2020). Population genetics has been used in the last decades to study colonization events, invasions, endemism, and to explain the present state of populations of different species. Moreover, with this information we can also infer the susceptibility of populations in front of environmental changes. Low genetic diversity in a population can be interpreted as a recent colonization event and thus the haplotypes of colonizers are dominant in the area. But it can also be explained by high specialization of the species to a particular niche. In both cases a low genetic variability, if coupled with low connectivity with other populations, can result in a population collapse should the environment change. On the other hand, a genetically diverse population has a higher probability to survive a change in the environment due to a higher probability of presence of genetic variants that offer resistance and adaptability. Likewise, an intense gene flow between populations can allow fast population recovery after local events (at the cost of less scope for specialization).

Most population genetic studies in the marine or terrestrial realms have been performed on single species using one or several markers. Few instances encompassing up to tens of species (e.g., Ayre et al., 2009; Cahill et al., 2017; Haye et al., 2014; Kelly and Palumbi, 2010) or reviewing available information from multiple groups (e.g., Hardy et al., 2011; Pascual et al., 2017; Patarnello et al., 2007; Teske et al., 2011) are available in the marine ecosystems.

The metabarcoding approach can fully capture the biodiversity of the complex reef benthic communities. This by itself is an enormous achievement,

but the potential of metabarcoding goes beyond simple biodiversity assessment. As said above, information about intraspecies variability can be mined from metabarcoding datasets, and with this information phylogeographic patterns of many MOTUs at a time (here called metaphylogeography) can be uncovered for the same spatial sampling scheme. Thus, metabarcoding can provide reliable data for biodiversity assessment, biogeography, and phylogeography simultaneously. It therefore has the potential to become an invaluable method for biomonitoring benthic communities, which in turn is a necessary step for all basic research, conservation, and management of these hyper-diverse key communities.

Chapter 2

Thesis aims and objectives

This Thesis is presented as a compendium of five articles, of which four have been already peer-reviewed and published in high impact journals. This Thesis tries to advance the field of biogeography and metaphylogeography for COI metabarcoding in marine shallow benthic communities. Given the lack of methods adapted for this type of community and the fact that available pipelines were mostly developed for ribosomal markers, the Thesis has a strong methodological component, consisting of several chapters where methods are set up and a final chapter where they are applied to a selected model. The main objectives of this Thesis are:

1) To explore different sampling methods to assess the marine benthic diversity. Is it possible to retrieve the benthic biodiversity using water samples from the benthic boundary layer or traditional sampling methods scraping the rocky surfaces are required? To respond to this objective, in chapter 4 we performed an experiment comparing the biodiversity results obtained from rocky benthic samples of two communities, photophilous and sciaphilous, and water samples collected at a distance of 0 to 20 m from the benthic communities. We therefore assessed the dynamics and detectability of benthic DNA in the adjacent water column.

2) Develop the pipelines and software required to obtain both inter and intraspecific variability from marine benthic commmunities of eukaryotes using COI. The chapters 5 and 6 respond to this purpose proposing an algorithm that combines both clustering and denoising methods to obtain the ESV and the MOTUs into which they are clustered to. These two chapters also calibrate the parameters of the software used for the COI barcode and chapter 7 presents a new program designed specially to denoise sequence data of coding genes using the entropy variability between

the nucleotide positions of the codon.

3) To apply the methodology provided by the chapters 4, 5, 6 and 7 to analyse the biogeographic and metaphylogeographic patterns from benthic communities across two well described fronts in the eastern Iberian littoral. In the chapter 8 12 localities of photophilous rocky benthic communities, from the southernmost Iberian point to the northeast coast, were sampled using traditional methods. Samples were analysed with metabarcoding techniques developed in the previous chapters and the inter and intraspecies variability were obtained. New analysis methods were used to retrieve the patterns of community and population distributions across the Almeria-Oran Front (AOF) and the Ibiza Channel (IC) barrier.

Finally the chapter 9 presents a global discussion of the results obtained in this Thesis.

Chapter 3

Directors Report

Dr. Xavier Turon Barrera, Dr. Owen S. Wangensteen Fuentes and Dr. Creu Palacín Cabañas, supervisors of the Doctoral Thesis written by the candidate Adrià Antich González and entitled "Biodiversity assessment of marine benthic communities with COI metabarcoding: methods and applications"

STATE

That the research work carried out by Adrià Antich González as part of his pre-doctoral training and included in his Doctoral Thesis has resulted in five chapters, four of which have already been published in international journals and one that is currently under review. To prepare this report, we have relied on the Web of Science database (Clarivate) to obtain the impact factors, quartiles and number of citations of the articles.

That none of the articles included in this Doctoral Thesis has been or will be used in any other Doctoral Thesis.

For each of the articles, the participation of Adrià Antich González is detailed as follows:

1 Turon, X., **Antich, A.**, Palacín, C., Præbel, K., Wangensteen, O. S. 2020. From metabarcoding to metaphylogeography: separating the wheat from the chaff. Ecological Applications, 30(2):e02036. doi: 10.1002/eap.2036. **IF 2020: 4.657. First Quartile in "Ecology"**

(31/166) and Second Quartile in "Environmental Sciences" (80/274). 39 citations

In this work, the doctoral student signs as second author, being the first one of the thesis supervisors. In this article, we worked with sequences that were already available to the group. When the PhD student joined the group, he wrote the code necessary for the article, analyzed the data, produced the figures and contributed to the drafting of the first manuscript.

2 **Antich, A.**, Palacín, C., Cebrian, E., Golo, R., Wangensteen, O. S., Turon, X. 2020. Marine biomonitoring with eDNA: can metabarcoding of water samples cut it as a tool for surveying benthic communities? Molecular ecology, pages 1–14. doi: 10.1111/mec.15641. **IF 2020: 6.185. Primer Decil a "Ecology" (16/166), First Quartile a "Evolutionary Biology" (7/50) and First Quartile in "Biochemistry & Molecular Biology" (62/295). 18 citations**

In this article, the doctoral student signs in first position as relevant author. He took care of all laboratory work, sequence analysis, statistical analyses, figure preparation and drafting of the first manuscript.

3 **Antich, A.**, Palacín, C., Wangensteen, O. S., Turon, X. 2021. To denoise or to cluster, that is not the question: optimizing pipelines for COI metabarcoding and metaphylogeography. BMC Bioinformatics, 22(1):177. doi: 10.1186/s12859-021-04115-6. **IF 2021: 3.307. Second Quartile in "Mathematical and Computational Biology" (20/57), Second Quartile in "Biochemical Research Methods" (37/79) and Second Quartile in "Biotechnology & Applied Microbiology" (86/158). 18 citations**

In this article, the doctoral student signs in first position as relevant author. He took care of all laboratory work, sequence analysis, statistical

analyses, figure preparation and drafting of the first manuscript.

4 **Antich, A.**, Palacín, C., Turon, X., Wangensteen, O. S. 2022. DnoisE: Distance denoising by entropy. An open-source parallelizable alternative for denoising sequence datasets. PeerJ, 10:e12758. doi: 10.7717/peerj.12758. **IF 2021: 3.061. Second Quartile in "Multidisciplinary Sciences" (33/73).**

In this article, the doctoral student signs in first position as relevant author. He wrote the code of the program, carried out its tuning and optimization using databases, and took care of the preparation of figures and writing of the first manuscript.

5 **Antich, A.**, Palacín, C., Zarcero J., Turon, X., Wangensteen, O. S. (submitted). Metabarcoding reveals high-resolution biogeographic and metaphylogeographic patterns through marine barriers. Submitted to Journal of Biogeography. **IF 2022: 4.810. First Quartile in "Geography, Physical" (9/48) and First Quartile in "Ecology" (41/173)**

In this article, the doctoral student signs in first position as relevant author. He participated in sample collection, took care of all laboratory work, sequence analysis, statistical analyses, figure preparation and writing of the first manuscript.

Thesis Supervisor
Dr. Xavier Turon Barrera
Center of Advanced Studies of Blanes (CEAB)

Thesis Supervisor and Tutor
Dr. Creu Palacín Cabañas
University of Barcelona (UB)

Thesis Supervisor
Dr. Owen S. Wangensteen Fuentes
The Arctic University of Norway (UiT)

Chapter 4

Marine biomonitoring with eDNA: Can metabarcoding of water samples cut it as a tool for surveying benthic communities?

4.1 Abstract

In the marine realm, biomonitoring using environmental DNA (eDNA) of benthic communities requires destructive direct sampling or the setting-up of settlement structures. Comparatively much less effort is required to sample the water column, which can be accessed remotely. In this study we assess the feasibility of obtaining information from the eukaryotic benthic communities by sampling the adjacent water layer. We studied two different rocky-substrate benthic communities with a technique based on quadrat sampling. We also took replicate water samples at four distances (0, 0.5, 1.5, and 20 m) from the benthic habitat. Using broad range primers to amplify a ca. 313 bp fragment of the cytochrome oxidase subunit I gene, we obtained a total of 3,543 molecular operational taxonomic units (MOTUs). The structure obtained in the two environments was markedly different, with Metazoa, Archaeplastida and Stramenopiles being the most diverse groups in benthic samples, and Hacrobia, Metazoa and Alveolata in the water. Only 265 MOTUs (7.5%) were shared between benthos and water samples and, of these, 180 (5.1%) were identified as benthic taxa that left their DNA in the water. Most of them were found immediately adjacent to the benthos, and their number decreased as we moved apart from the benthic habitat. It was concluded that water eDNA, even in the close vicinity of the benthos, was a poor proxy for the analysis of benthic structure, and

that direct sampling methods are required for monitoring these complex communities via metabarcoding.

4.2 Introduction

Metabarcoding is by now a well-established technique for assessing biodiversity in a variety of terrestrial, freshwater, and marine environments (reviewed in Bohmann et al., 2014; Creer et al., 2016; Cristescu, 2014; Deiner et al., 2017; Taberlet et al., 2012b). The wealth of published papers dealing with technical issues and generating new data with this method testifies to the widening scope of applications of metabarcoding. One such application, where metabarcoding is becoming a game-changer, is in the field of biomonitoring (Aylagas et al., 2018; Hajibabaei et al., 2016; Kelly et al., 2014a; Porter and Hajibabaei, 2018c). Not in vain the use of DNA-based approaches for monitoring applications has been christened Biomonitoring 2.0 (Baird and Hajibabaei, 2012; Leese et al., 2018).

In the marine realm, all current policies, such as the European Union Marine Strategy Framework Directive, mandate comprehensive, community-wide approaches to monitoring (Danovaro et al., 2016; Goodwin et al., 2017; Hering et al., 2018; Leese et al., 2018). Metabarcoding provides a cost-effective, ecosystem-wide method for the assessment of biodiversity, which lies at the basis of all monitoring efforts (Aylagas et al., 2018; Krehenwinkel et al., 2019; Leray and Knowlton, 2016; Shaw et al., 2017). An ever widening range of ecological and socioeconomic issues, such as invasive species management (Darling et al., 2017; Holman et al., 2019), marine protected areas design (Bani et al., 2020), pathogen monitoring (Peters et al., 2018), fisheries management (Zou et al., 2020), or deep-sea mining (Cowart et al., 2020), among others, require powerful and fast biomonitoring tools. Metabarcoding provides these tools at a pace, cost, and depth that are not achievable using conventional, morphology-based surveys (Porter and Hajibabaei, 2018c). Alpha- and beta-diversity estimates, as well as biotic indices, can be reliably obtained using metabarcoding (Aylagas et al., 2018; Bani et al., 2020; Hering et al., 2018; Pawlowski et al., 2018). The amount of data typically generated in metabarcoding data sets also allows

bioassessments based on taxonomy-free and machine learning techniques (Cordier and Pawlowski, 2018; Gerhard and Gunsch, 2019), or the analysis of diversity at the within-species level (Turon et al., 2020).

Of course, gaps and problems are also recognized in this burgeoning field (e.g., Alberdi et al., 2018; Kelly et al., 2019; McGee et al., 2019), among which the need to obtain better reference databases (Sinniger et al., 2016; Wangensteen et al., 2018b; Weigand et al., 2019) and the need to standardize field and laboratory procedures (McGee et al., 2019; Weigand et al., 2019). Among the latter, the type of substrate sampled is of paramount importance (Koziol et al., 2019). In the sea, most studies to date have sampled either the sediment (e.g., Atienza et al., 2020; Brannock et al., 2016; Fonseca et al., 2014; Guardiola et al., 2016), or the water column (e.g., Brannock et al., 2018; Fraija-Fernández et al., 2020; Sigsgaard et al., 2019; Stefanni et al., 2018). Less effort has been devoted to the study of hard-substrate natural benthic communities. These have been analysed either using indirect methods based on deploying artificial substrates (Cahill et al., 2018; Leray and Knowlton, 2015; Pearman et al., 2019; Ransome et al., 2017), or by directly taking samples by scraping off standardized surfaces (Shum et al., 2019; Wangensteen et al., 2018a,b), or using suction devices (Cowart et al., 2020; de Jode et al., 2019).

Either deploying settlement surfaces (that need to be recovered) or using direct collection methods, the sampling of benthic hard-bottom habitats requires direct access to the environment and involves more effort than sampling substrates such as water or sediment, which can be accessed remotely. In addition, direct methods are destructive, which is an inconvenience for the sustained sampling necessary for biomonitoring. It is, therefore, highly convenient to develop alternative methods for assessing benthic biodiversity, and an obvious choice would be to sample the water in the vicinity of the benthos to recover benthic DNA for metabarcoding applications. While water environmental DNA (eDNA) has been used for the study of protists, fito- and zooplankton or fish assemblages (e.g., Djurhuus et al., 2018; Massana et al., 2015; Shu et al., 2020), its potential utility to analyse benthic communities is much less understood. Some authors (Koziol et al., 2019;

Rey et al., 2020) compared eDNA from water, sediment and settlement plates in port environments, finding clearly distinct community profiles. Leduc et al. (2019) similarly found significant differences in community composition between eDNA from water samples and standard invertebrate collection methods in Arctic harbours. West et al. (2020) used surface water samples to assess coral reef community variation, but did not perform a comparison with the actual benthic communities. Alexander et al. (2020) used eDNA from surface waters to target scleractinian diversity, and found the method promising, albeit with notable differences with results from visual censuses. Stat et al. (2017) compared two different methods to study the eDNA from tropical marine reefs using shallow water and found eDNA metabarcoding more promising than the shotgun approach for assessing eukaryotic diversity.

The usefulness of DNA obtained from water samples as a proxy for benthic communities will depend on the many factors that affect DNA release, transport, and degradation (Barnes and Turner, 2016; Collins et al., 2018; Salter, 2018; Stewart, 2019). While some studies have assessed the spatial distribution of eDNA in coastal habitats, they have been done at scales too large to link water samples with particular benthic habitats. Bakker et al. (2019) analysed water eDNA from coastal shelf habitats spanning the Caribbean Sea. O'Donnell et al. (2017) found fine scale patterns in the distribution of water eDNA, but they used transects perpendicular to the shore spanning a few kilometres. Jeunen et al. (2019) analysed the vertical stratification of eDNA at the scale of metres, but did not focus on any relationship with benthic communities. Jacobs-Palmer et al. (2020) analysed eDNA from water taken in the vicinity (from 1 to 15 m) of the edges of *Zostera marina* patches, and could detect an inhibitory effect of the seagrass community on the dinoflagellate abundances in the plankton. To our knowledge, however, no study has assessed marine eDNA dynamics at the benthic boundary layer, which is the water immediately adjacent (from centimetres to metres) to the benthos, where steep gradients in abiotic and biotic parameters occur (Boudreau and Jorgensen, 2001). Only Hajibabaei et al. (2019) and (Gleason et al., 2021) have compared, in freshwater environments, the results from DNA obtained from matched

water and benthic samples, and found water eDNA to be a poor surrogate for benthic community composition.

In this work, and using two hard-bottom communities on vertical walls in the NW Mediterranean, we compared the information obtained from analysing the DNA obtained from benthic (using direct methods as in Wangensteen et al., 2018b) and water samples collected at increasing distances (from centimetres to metres) from these communities. We used metabarcoding of the COI gene with broad range primers as our focus was on recovering the taxonomically diverse eukaryotic communities present. Our goals were to assess the eDNA dynamics in the boundary layer of the benthos and to determine the feasibility of analysing benthic diversity by collecting water samples.

4.3 Material and Methods

4.3.1 Sample collection

In the present study samples were taken from two different hard-bottom communities, a shallower (photophilous) and a deeper (sciaphilous) communities found in the same vertical wall facing SSE, in the National Park of Cabrera Archipelago in the Balearic Islands (Western Mediterranean, 39º07'30.32"N, 2º57'37.14"E, Fig. A.1.1). The photophilous community at 10 m depth was dominated by the seaweeds *Padina pavonica* and *Dictyopteris membranacea*. In the sciaphilous community at 30 m depth, the seaweed *Halimeda tuna*, sponges and other invertebrates were the dominant biota. For more detailed information of these communities see Wangensteen et al. (2018b).

Two different sampling methods were used in the present study. Benthic samples (three replicates per community) were obtained by scraping to bare rock quadrats of 25 × 25 cm with hammer and chisel. All the material was collected underwater in plastic bags. Two divers performed the sampling, with one keeping the sample bag open just over the zone being scraped to avoid escape of small motile fauna. Water samples (four replicates at each point) were obtained with 1.5 L bottles at different distances from the benthos (0, 0.5 and 1.5 m) for each community. The sample labelled 0 m

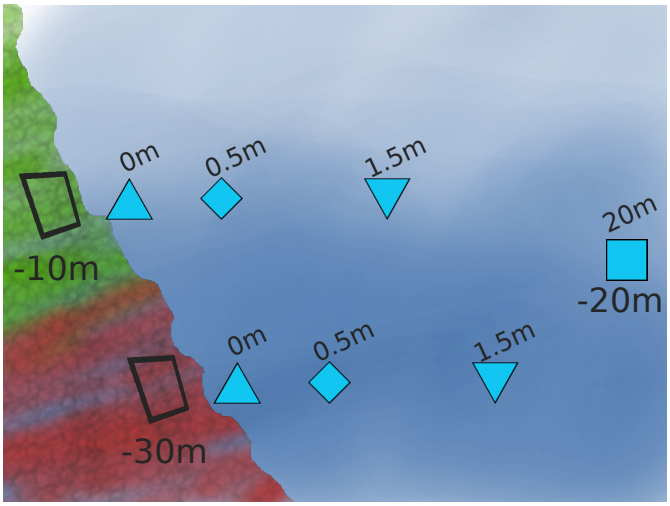


Figure 4.1: Schema of the sampling design. We sampled two hard bottom communities (green: photophilous; red: sciaphilous) at -10 and -30 m of depth, respectively, by sampling quadrats of 25 × 25 cm (three replicates each). Water samples (1.5 L) were collected at different distances from each community (0, 0.5 and 1.5 m, four replicates each). Pelagic samples were taken at intermediate (-20 m) depth and at 20 m from the wall (three replicates).

was obtained in the water layer just adjacent (ca. 5 cm) to the benthos. As an external pelagic control, water samples (three replicates) of 1.5 L were obtained at 20 m from the benthos and at an intermediate depth (-20 m). The sampling design is sketched in Figure 4.1. Hereafter we will use the names photophilous and sciaphilous samples to designate both the benthic and the water samples $\leq$1.5 m from the wall at each of the two depth levels sampled, and the name pelagic samples to designate the water samples collected 20 m apart from the rocky wall at -20 m. New, unopened mineral water plastic bottles were used for water collection, one per sample. They were first filled with sterilized water and, once in the collection point, they were held upside-down and water was displaced using air bubbled from a spare SCUBA regulator. The bottles were then righted and water from the exact point of collection was allowed to fill them.

4.3.2 Sample processing

Water samples were processed on site immediately after collection. The whole collected volume (1.5 L, comparable to other studies, e.g., Collins et al., 2019; Sales et al., 2019) was prefiltered with a 200 μm mesh to eliminate coarse particles and then filtered through 0.22 μm Sterivex millipore filters (Merck) using sterile, disposable syringes (a new syringe per sample). The filter cartridges were then stored at -20^{o}C in sterile plastic bags. Benthic samples were fixed with ethanol immediately after collection and kept at -20^{o}C until processed in the laboratory. Following Wangensteen and Turon (2017), Wangensteen et al. (2018b) and Wangensteen et al. (2018a), benthic samples were separated in the laboratory in three different size fractions (A: >10 mm; B: 1–0 mm; C: 63 μm–1 mm) using a stainless steel mesh sieve column (Cisa S.L., www.cisa.net). Each fraction was homogenized with a blender and stored in ethanol at -20^{o}C until DNA extraction. All equipment was carefully bleached between samples.

Our sample data set thus consisted of 18 benthic samples (two communities $\times$ three replicates $\times$ three fractions) and 27 water samples (two communities $\times$ three distances $\times$ four replicates + three pelagic samples).

4.3.3 DNA straction

All procedures were made in a laminar flow cabinet sterilised with UV light between samples. DNA from benthic samples was extracted using 10 g of homogenized material and the DNeasy PowerMax Soil Kit (Qiagen). The Sterivex filter cartridges were opened with sterile pincers in the cabinet and DNA from the filters was then extracted using the DNeasy PowerWater kit (Qiagen). A Qubit fluorometer (ThermoFisher) was used to check the concentration of DNA (higher than 5 ng/μl in all cases).

4.3.4 PCR amplification and library preparation

A fragment of ca. 313 bp of the Cytochrome Oxidase 1 (COI) gene was amplified with a set of universal primers targeting eukaryotes. We used the Leray-XT primer set (Wangensteen

et al., 2018a,b): forward jgHCO2198 (Geller et al., 2013): 5'-TAIACYTCIGGRTGICCRAARAAYCA-3', reverse mlCOIintF-XT (Wangensteen et al., 2018b): 5'-GGWACWRGWTGRACWITITAYCCYCC-3'. All primers had an 8-base specific tag attached. The tags had a minimum difference of 3 bases from each other, and were designed with the program Oligotaq (Boyer et al., 2016). Forward and reverse primers used for amplification of each sample had the same tag. A variable number of degenerate (N) bases (from two to four) were also attached to the forward and reverse primers to improve sequence diversity for illumina processing.

Three PCR replicates were performed for each DNA extraction. PCR conditions for COI amplification followed Wangensteen et al. (2018b). DNA was then purified and concentrated using MinElute PCR Purification Kit (Qiagen) and an electrophoresis gel was performed to check amplification success.

Amplification controls were added as follows: two PCR blanks were run by amplifying the PCR mixture without any DNA template. Negative controls were made for the benthic samples by processing triplicate sand samples that were charred in a furnace (400ºC for 24 hr) and then sieved and processed as above. For the water samples we filtered in situ sterilized ultrapure water with three Sterivex filters that were then treated in the same manner as the seawater filters. Amplification products were pooled to build two Illumina libraries using Nextflex PCR-free library preparation kit (Perkin-Elmer). Both libraries were sequenced together in an Illumina MiSeq V3 run using 2 × 250 bp paired-end sequencing.

4.3.5 Bioinformatic analyses

The bioinformatic analyses followed the same pipeline of Atienza et al. (2020) with slight modifications. Most steps used the OBITools package (Boyer et al., 2016). Illuminapairedend was used to align paired-end reads and keep only those with >40 alignment quality score. Reads were demultiplexed using ngsfilter. Those with mismatched primer tags at any end were discarded. Obigrep and obiuniq were used to perform a length filter (retaining only those between 310–317 bp) and dereplicate sequences.

Uchime-denovo algorithm from VSEARCH v2.7.1 was used to remove chimeric amplicons. The resulting read data set in fasta format, with the abundances in each sample, was uploaded to the Dryad repository (https://doi-org.sire.ub.edu/10.5061/dryad.vt4b8gtq2).

Sequences were then clustered into molecular operational taxonomic units (MOTUs) with SWARM v2.1.7 using $d = 13$ (Bakker et al., 2019; Siegenthaler et al., 2019a). Singletons (MOTUs with just one read) were removed after this step to minimize data loss (Atienza et al., 2020). Taxonomic assignment was performed using ecotag and a custom database containing sequences from the EMBL nucleotide database and sequences obtained from the Barcode of Life Database (BOLD), using a custom script to select the appropriate fragment (see details and a summary of the taxonomic groups represented in Wangensteen et al., 2018b). This database contains 188,960 reference sequences covering most eukaryotic groups and is available from https://github.com/metabarpark/Reference-databases. Assignment of metazoan sequences was further improved by querying the BOLD database. Sequences with a species name assigned and with an identity match >95% in BOLD were kept, whereas matches below this threshold, even if assigned to species level by ecotag, were downgraded to genus level.

The final refining steps consisted of deleting any MOTU for which reads in blank or negative controls represented more than 10% of total reads for that MOTU in all samples. A minimum relative abundance filter was also applied, removing, for a given PCR replicate, the MOTUs that represented less than 0.005% of total reads of that replicate. We also removed MOTUs that had a combined total of <5 reads after the previous steps. Finally, all MOTUs that were not assigned to marine eukaryotes (i.e., MOTUs assigned to nonmarine organisms, prokaryotes, or to the root of the Tree of Life) were eliminated. We then pooled the three PCRs of each sample. We used the higher classification of eukaryotes proposed by Guillou et al. (2012) at the super-group level, with one exception: Opisthokontha was split into Metazoa and Fungi.

4.3.6 Data analyses

Analyses were performed with the R package vegan (Oksanen et al., 2019). Rarefaction curves of the number of MOTUs obtained at an increasing number of reads were obtained with function rarecurve, separately for benthos and water samples. Likewise, MOTU accumulation curves with increasing numbers of samples were obtained for benthos and water with specaccum. MOTU richness values were compared with standard ANOVAs (factors community and sample type: benthos or water). Between-sample distances were computed using the Jaccard index based on presence/absence data of each MOTU per sample. These distances were then used to obtain ordinations of the samples in nonmetric multidimensional scaling (nmMDS) representations using function metaMDS with 500 random starts. Permutational analyses of variance were performed on Jaccard distances with function adonis to test differences between relevant factors: a one-way analysis was performed between benthos and water (all samples combined), a three-way analysis was done for the benthos with community and fraction as main factors and sample as a blocking factor nested in community. For the water, a two-way analysis was performed with community and distance to the wall (pelagic samples excluded as they were taken at an intermediate depth). Main factors were also tested for differences in multivariate dispersion (permdisp analysis using function betadisper) to check whether significant outcomes were a result of different multivariate heterogeneity (spread) or different centroid location of the groups. A Venn diagram was prepared with the VennDiagram package (Chen, 2018) to represent the degree of MOTU overlap between benthos and water. Upset diagrams were used to plot shared MOTUs at increasing distances of the benthic communities using package UpSetR (Conway et al., 2017).

4.4 Results

We obtained a total of 7,391,160 reads in total for the benthic samples (18 samples) and 13,652,493 reads for the water samples (27 samples). The controls had a negligible number of reads (85.29 $\pm$ 19.80, mean $\pm$ SE). After quality filtering, demultiplexing, dereplicating and chimera elimination we

had a total 3,868,827 unique COI sequences. These were clustered into 15,954 nonsingleton MOTUs. The final refining steps and, particularly, the elimination of MOTUs not assignable to marine eukaryotes using our reference database greatly reduced the data set to a final list of 3,543 MOTUs. The impact of removing noneukaryotic MOTUs was much greater in the water samples: only 14.35% of initial reads were retained at this step, while 99.36% were kept in the benthic samples. In the final data set, benthic samples had 2,396 MOTUs, while water samples had 1,412 MOTUs. The final average number of eukaryotic reads in benthic samples was 233.957 ± 25.40 (mean ± SE) and in water samples was much lower, 34.708 ± 2.50, as a result of the elimination of noneukaryotic MOTUs. Table A.2.1 presents the final MOTU table with the taxonomic assignment and number of reads per sample. Rarefaction curves (Fig. A.1.2) showed that a plateau is reached in the number of MOTUs with the sequencing depth obtained in most samples from benthos and water (exceptions corresponded to some of the finer fractions in benthic samples). Likewise, MOTU accumulation curves (Fig. A.1.3) tended to saturate in water samples but not in benthic samples, so addition of more samples would probably increase the total number of MOTUs recovered from this habitat. In spite of the different number of total reads, we compared MOTU richness without rarefaction as in most samples the richness values plateaued at the sequencing depth obtained. Somewhat higher values were found in benthos (637.78 ± 59.00 and 420.34 ± 47.96 MOTUs in the photophilous and sciaphilous communities, respectively) compared to those in water at 0–1.5 m of distance (541.58 ± 29.40 and 389.92 ± 20.58 MOTUs, respectively). A two-way ANOVA showed that the number of MOTUs was not significantly different between benthos and water samples, but it was significantly higher in the photophilous than in the sciaphilous community (community effect, $p < .001$; sample type effect, $p = .110$; interaction, $p = .401$). The pelagic samples had 474.33 ± 28.50 MOTUs.

Taxonomic assignment revealed a total of seven super-groups in the samples, of which the most diverse was Metazoa (996 MOTUs, 45.47% of reads, all samples combined) followed by Archaeplastida (351 MOTUs, 16.47% of reads, mostly belonging to Rhodophyta), and Stramenopiles

(287 MOTUs, 3.25% of reads). A total of 1,565 eukaryotic MOTUs could not be assigned to a given super-group. They represent 32.25% of total reads, but the share of unassigned reads was highly uneven: 21.94% of reads in benthic samples, and 78.58% in water samples. Within metazoans we identified 15 phyla, of which the most diverse were Arthropoda (211 MOTUs, 2.17% of total reads, all samples combined), followed by Annelida (116 MOTUs, 1.71% of reads), Cnidaria (74 MOTUs, 11.65% of reads), Porifera (59 MOTUs, 6.35% of reads) and Mollusca (50 MOTUs, 1.20% of reads). Among metazoans, 382 MOTUs could not be assigned at phylum or lower levels. In addition, 165 MOTUs could be assigned at the species level by ecotag with more than 0.95 identity with the best match in the reference database.

The relative number of MOTUs as per super-group and metazoan phylum obtained in the benthos and water samples is shown in Figure 4.2. The general patterns recovered were notably different in the two habitats surveyed. Metazoa were markedly dominant in the benthos in terms of number of MOTUs, followed by Archaeplastida (mostly Rhodophyta). On the other hand, Hacrobia (mostly Haptophyta) had the highest diversity in water samples, where other important planktonic groups such as the Alveolata had a much higher representation than in the benthos. Nevertheless, Metazoa was the second most MOTU-rich group in the water. As for metazoan phyla, the distribution was more similar: Arthropoda was the most diverse group in both habitats, and Annelida, Cnidaria, Mollusca and Porifera (albeit in different order) came next. However, the picture is different considering the relative number of reads: Cnidaria were dominant in the benthos (26.05% of metazoan reads), where the abundance of Arthropoda was much lower (3.88%). Conversely, in the water Arthropoda was the most abundant by far in proportion of metazoan reads (46.70%).

The number of MOTUs of the main metazoan phyla, Arthropoda, Annelida, Cnidaria, and Mollusca was further assessed at lower taxonomic levels (Order) in Table A.2.2. In arthropods, Amphipoda, Decapoda, Isopoda and Harpacticoida were highly diverse in the benthos but practically absent from water samples, which were dominated by planktonic groups such as

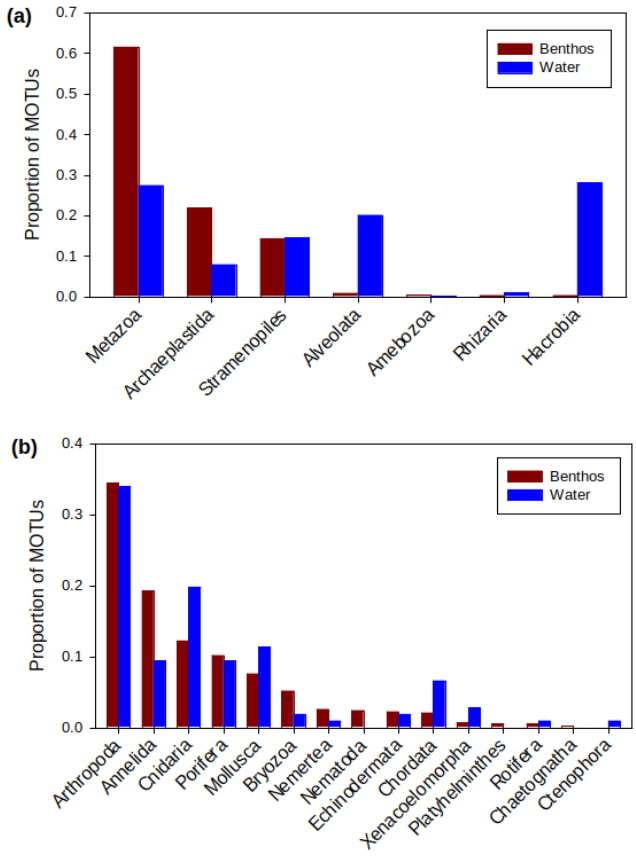


Figure 4.2: Barplot of relative MOTU richness of the super-groups (a) and metazoan phyla (b) detected in benthic and water samples.

Calanoida and Cyclopoida. In annelids, Sabellida and Sipuncula were the most diverse groups in the plankton, while the dominant group in benthos (Phyllodocida) was practically absent in water samples (only four MOTUs in total). Among Cnidaria, only hydrozoans (Trachymedusae, Siphonophora, and Leptothecata) are diverse in the plankton samples, with a negligible representation of anthozoan orders which, together with Leptothecata, dominate in the benthic samples. Among Mollusca, highly diverse groups in the benthos such as Mytiloida, or gastropoda in general (with the exception of the pelagic Pteropoda) were absent or poorly represented in water samples. This perusal indicates that we did not capture in our samples planktonic

stages of many benthic groups, and that the rates of DNA shedding from benthos to the water are in general low.

The sample ordination using the Jaccard index is shown in Figure 4.3a. A clear separation of benthic and water samples is evident, which is in agreement with one-way results comparing benthos and water, all samples pooled (PERMANOVA $p < .001$, and permdisp $p < .001$). In the benthos, the shallower and deeper communities formed clearly separated clusters. A PERMANOVA analysis on benthic samples alone showed a significant effect of community ($p < .001$) and of the nested factor sample (within community); while fraction or the interaction between community and fraction were not significant (Table 4.1). The permdisp test showed that there was also a different dispersion of data in the two communities ($p < .001$), which is also visible in the nmMDS. A second nmMDS was performed only with the water samples (Fig. 4.3b), where a separation by communities can also be seen, albeit with some overlap. A PERMANOVA of water samples using community and distance to the wall as factors (pelagic samples were excluded in this analysis) showed a significant interaction term ($p = .027$, Table 4.2), indicating different effects of the community with increasing distances. A comparison of the factor community at fixed distances showed that differences between photophilous and sciaphilous samples were significant at all distances (0, 0.5, and 1.5 m, all $p < .031$), and this was not due to differences in heterogeneity (all permdisp tests not significant). Likewise, a comparison of the factor distance at each depth level showed that distance to the rocky wall did not have a significant effect on the overall water assemblage composition ($p = .063$ and .056 for the photophilous and sciaphilous communities, respectively).

Of the total 3,543 MOTUs, only 265 were shared between benthos and water (Fig. 4.4, Tables A.2.3 and A.2.4), which represented 11.06% of the MOTUs found in benthos. However, these 265 MOTUs accounted for 70.40% of the reads of the benthos, indicating that they correspond to abundant taxa. These same MOTUs accounted for 56.37% of the reads in the water samples. The MOTUs shared between benthos and water could be assigned to two main groups, those whose relative read abundance in the benthos

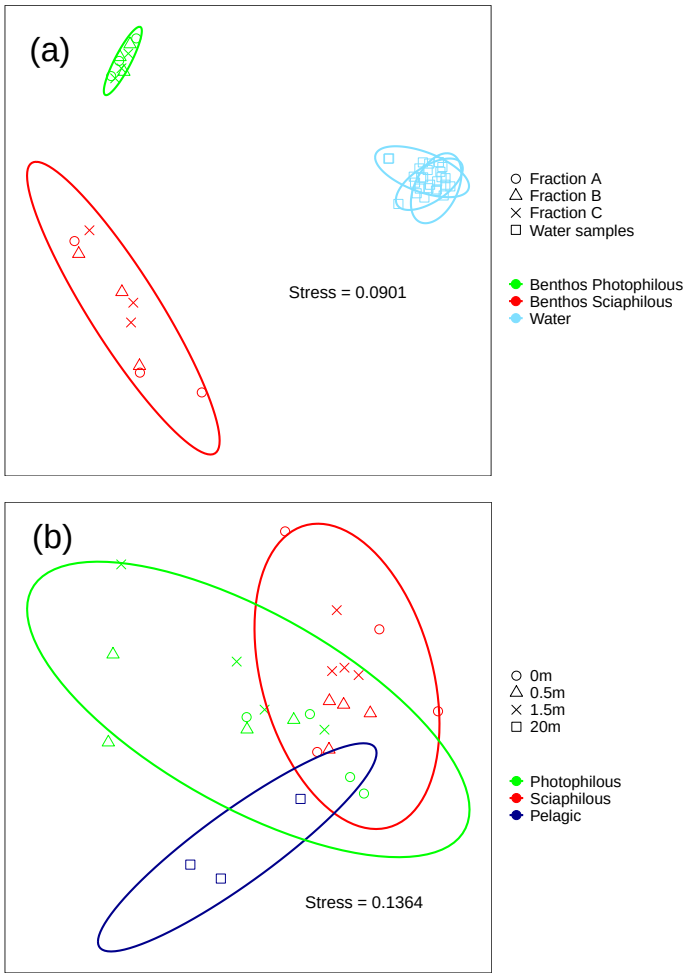


Figure 4.3: Nonmetric multidimentional scaling representation of all samples (a) and only water samples (b) using the Jaccard distance. Benthic samples (a) were separated in three different size fractions: A (>10 mm); B (between 10 mm and 1 mm); and C (between 1 mm and 63 μm). Communities are coded by colours and fractions (benthos) and distances (water) by symbols.

Table 4.1: Results of the PERMANOVA analysis performed on Jaccard distances among the samples collected in two benthic communities (photophilous and sciaphilous) and separated into three size classes (fractions). Sample was added as a nested factor within community. Columns are: degrees of freedom (DF), sum of squares (SS), F-statistic of the model, with its associated probability (p-value), and probability of the permdisp test of multivariate homogeneity of group dispersions (Permdisp). Significant values marked with asterisk.

Factor	**df**	**SS**	**F-statistic**	**p-value**	**Permdisp**
Community	1	1.581	5.442	0.001*	0.001*
Fraction	2	0.731	1.258	0.140	0.869
Community*fraction	2	0.653	1.124	0.267	
Sample (community)	2	1.158	1.993	0.002*	
Residuals	10	2.905			

Table 4.2: Results of the PERMANOVA analysis performed on Jaccard distances among the water samples collected in two communities (photophilous and sciaphilous) and at three distances from the benthos (Distance factor: 0, 0.5 and 1.5 m). Columns are: degrees of freedom (DF), sum of squares (SS), F-statistic of the model, with its associated probability (p-value), and probability of the permdisp test of multivariate homogeneity of group dispersions (Permdisp). Significant values marked with asterisk. Significant values marked with asterisk.

Factor	**df**	**SS**	**F-statistic**	**p-value**	**Permdisp**
Community	1	0.265	4.127	0.001*	0.216
Distance	2	0.166	1.293	0.129	0.940
Community*distance	2	0.216	1.682	0.027*	
Residuals	18	1.157			

was higher than in the water and those displaying the opposite pattern. We assume that the first group corresponds mainly to benthic MOTUs that left their DNA signature in the water (hereafter "shared benthic MOTUs" or SBM), while the second group probably corresponds to planktonic MOTUs (hereafter "shared pelagic MOTUs" or SPM). Only one MOTU could not be assigned to any of these categories as it had the same number of reads in both environments.

The first group (SBM) comprised 180 MOTUs (Table A.2.3), which represented 7.51% and 70.33% of MOTUs and reads in the benthos, respectively, while they constituted 12.75% and 1.99% of the MOTUs and reads in the water. Of these MOTUs, almost half (84, 46.67%) belonged to

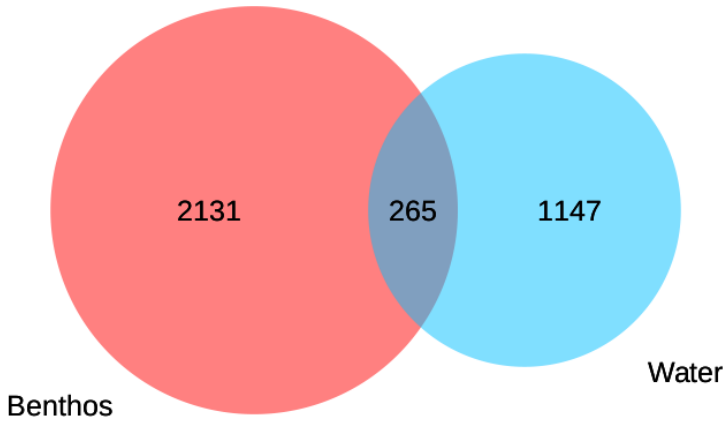


Figure 4.4: Venn diagram showing the overall MOTU overlap between the two types of community considered.

metazoan groups, but only seven of them were arthropods (the dominant metazoan group in the plankton); the second most important group were the red algae (a mostly benthic group), with 25 (13.89%) MOTUs. Of the dominant planktonic groups, only 11 (6.11%) SBM were diatoms and two were dinoflagellates. The taxonomic assignments were, therefore, mostly coherent with the idea that this subset of MOTUs belong mainly to benthic groups (Table A.2.3). A total of 45 SBM MOTUs (25%) could not be assigned to any super-group.

The 84 shared pelagic MOTUs (SPM, Table A.2.4) made up 3.51% of MOTUs but only 0.07% of reads in the benthos. On the other hand, while they comprised 5.95% of pelagic MOTUs they accounted for 54.44% of pelagic reads. Their taxonomic assignments showed that 22 (26.19%) MOTUs were metazoans, of which a majority (17) were arthropods. On the other hand, 18 (21.43%) MOTUs belonged to typical planktonic protists (diatoms, dinoflagellates, Hacrobia, Rhizaria) (Table A.2.4). Finally, 42 (50%) SPM could not be assigned to any super-group. The higher number of unassigned MOTUs and the taxonomic composition suggest a dominance of nonbenthic groups in the SPM subset.

When the distribution of the 180 shared benthic MOTUs was examined, they clearly decreased with distance to the wall (Fig. 4.5), with 135, 74, 24,

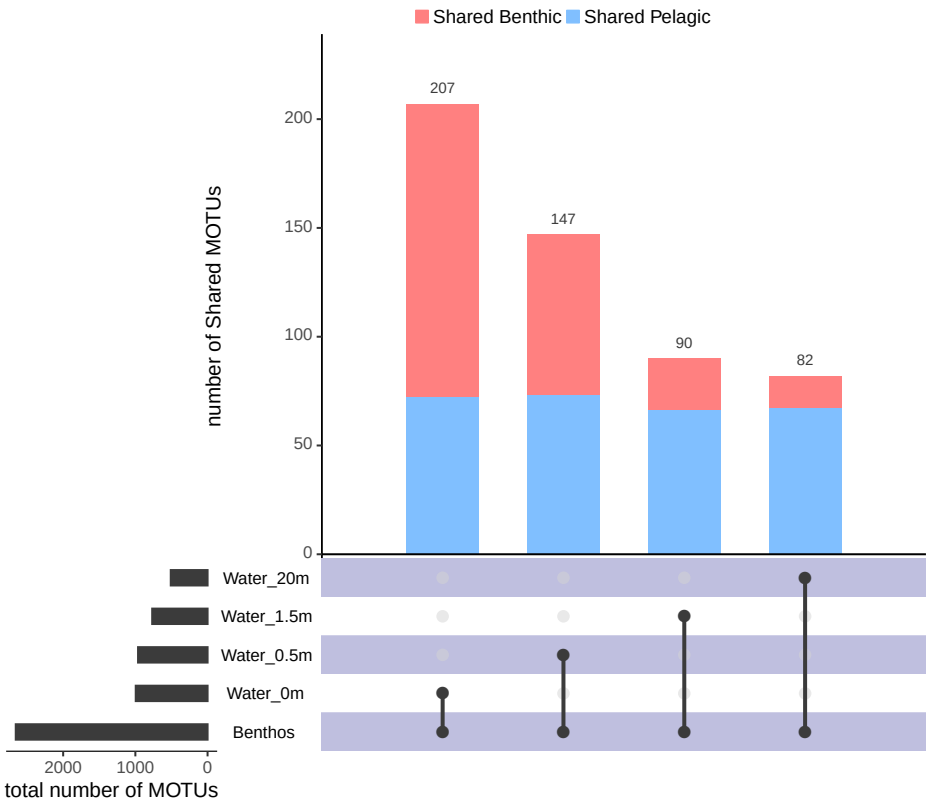


Figure 4.5: Upset plot with the number of shared MOTUs between the benthos and the water samples and the total number of MOTUs detected. Shared benthic MOTUs (SBM) are represented in pink and shared pelagic MOTUs (SPM) in light blue.

and 15 MOTUs shared between benthos and water samples at 0, 0.5, 1.5 and 20 m, respectively. Their abundance in relative read numbers also decreased (from 0.056 to 0.002, Table A.2.3), which supports the idea of their benthic origin. This same general pattern was found when both communities studied were analysed separately (Figs. A.1.4 and A.1.5).

By contrast, the comparison of shared pelagic MOTUs did not show any clear trend with distance to the wall (Fig. 4.5): 72, 73, 66, and 67 at 0, 0.5, 1.5, and 20 m, respectively. Neither was a trend found in relative read abundances per sample (between 0.570 and 0.526 irrespective of distance, Table A.2.4). Again, this same general pattern was found in both communities separately (Figs. A.1.4 and A.1.5).

4.5 Discussion

Metabarcoding of benthos and water samples, using a broad range eukaryotic marker (COI), retrieved clearly different communities. The patterns of MOTU richness and abundance of reads from the different environments were distinct, showing a dominance of taxa with important planktonic components (such as dinoflagellates, diatoms, and haptophytes) in the water samples, while metazoans and rhodophytes were the most diverse and abundant in the benthos. Metazoans, notwithstanding, were also well represented in water samples, with a dominance of arthropods (mostly calanoids and cyclopoids) in both number of MOTUs and reads. The rarefaction and MOTU accumulation curves showed that we captured adequately the richness present in the samples with our sequencing depth, and that the total eukaryotic diversity in the benthos was higher than that in the water. More replicates of benthic samples would be necessary to recover the overall MOTU richness of this habitat.

However, we acknowledge that the sampling methods used were different for benthos and plankton. We have used techniques currently applied to sample these environments. In complex communities such as the benthos, with organism sizes spanning several orders of magnitude, size-fractionation is necessary to recover the biodiversity present (Elbrecht et al., 2017a; Wangensteen and Turon, 2017; Wangensteen et al., 2018b). In addition, the mesh size used for the smallest sieve was 63 μm, meaning that most prokaryotes and a significant part of the smallest microeukaryotes were washed out, along with cell debris and extracellular DNA. In the filters, on the other hand, we retained everything down to colloidal level, thus the prokaryotic community, for instance, was captured in our samples. This explains the amount of reads that had to be discarded in the water samples as not assignable to eukaryotes and, within eukaryotes, the high number of reads that could not be assigned to any supergroup (the smallest eukaryotes being the less represented in the reference database for COI). Our point was not to test both techniques or to compare their particularities, but rather to check if the information retrieved from currently established methods for the analysis of water DNA is comparable to that from current analytical

techniques for benthos.

While the DNA obtained from the filters would be labelled as eDNA, the sampling from the benthos would be qualified as community or bulk DNA by many. eDNA is defined as the DNA obtained from an environmental matrix such as water or sediment without isolating the organisms (Barnes and Turner, 2016; Creer et al., 2016; Stewart, 2019; Taberlet et al., 2012a); and is usually opposed to bulk or community DNA, referring to DNA obtained from organisms previously isolated from the environment (Andújar et al., 2018; Creer et al., 2016; Deiner et al., 2017). In a more restricted sense (e.g., Andújar et al., 2018; Cristescu and Hebert, 2018; Thomsen et al., 2012; Tsuji et al., 2019), the term eDNA is used as equivalent to trace DNA released from organisms (in the form of mucus, faeces, cells, hairs, etc), so when studying eDNA the organisms themselves are not in the sample. We consider, however, that eDNA should be used as a general term, to designate any DNA extracted from an environmental sample. It is commonly made up of a mix of intraorganismal (in the form of small organisms relative to the sample size) and extra-organismal or trace eDNA shed from large organisms (Creer et al., 2016; Pawlowski et al., 2018; Porter and Hajibabaei, 2018c; Salter, 2018; Taberlet et al., 2012a). The relative amount of both components is highly variable, though, and it depends on the sampling method and the target group, and hence the primers used. In our case, we used a broadly universal primer set for eukaryotes, capable of amplifying both intraorganismal and trace DNA from most eukaryotic taxa. So the benthic samples are more enriched in intraorganismal DNA (since most trace DNA was removed by sieving), while the water samples contain a mix of a high amount of intraorganismal DNA from planktonic microeukaryotes and a smaller fraction of extra-organismal DNA from larger organisms.

The ordination and PERMANOVA results confirmed the marked differentiation between the samples from both environments. An assessment at the Order level in the main metazoan phyla confirmed that the composition of the two environments is highly different. Moreover, the differences between the two depths sampled, which corresponded to two different communities (photophilous and sciaphilous) on precisely the same wall, were pronounced

in the benthic samples, but were also significant in the water samples taken between 0 and 1.5 m of the rocky wall. Thus, the method is sensitive enough to detect ecological differences not just in the sessile communities, but also in the more dynamic planktonic habitat. This is in agreement with other studies that have also shown that the eDNA in seawater samples can detect differences in composition of several groups at relatively small scales (from metres to tens of metres, Jacobs-Palmer et al., 2020; Jeunen et al., 2019; Port et al., 2016).

A total of 3,543 eukaryotic MOTUs were detected in the whole data set. In spite of the lower number of eukaryotic reads retrieved from the water (15% of those retrieved from the benthos), the number of eukaryotic MOTUs in the water was ca. 60% of those in the benthos (1,412 as compared to 2,396). Only 265 MOTUs were found to be shared between the benthos and the water samples. This represents only ca. 11% and 19% of the MOTUs in the two environments, respectively. In addition, a closer scrutiny allowed us to separate those shared MOTUs into those of possibly benthic origin (shared benthic MOTUs, SBM) and those of probably planktonic origin (shared pelagic MOTUs, SPM).

The 180 SBM comprised ca. 7.5% of the benthic MOTUs but represented ca. 70% or benthic reads (while only ca. 2% of water-derived reads), indicating that abundant benthic MOTUs are the ones more prone to leave their signature in the surrounding water. The 84 SPM accounted to ca. 6% of pelagic MOTUs but ca. 54% or eukaryotic pelagic reads (and only 0.07% of reads in the benthos), again indicating that the most abundant MOTUs are the ones that can be detected also in the other habitat.

The fine-scale distribution of the 180 SBM showed a clear trend: more MOTUs were shared in the immediate vicinity of the benthos (135 with water at 0 m), and the number decreased with distance down to only 15 MOTUs shared with the water at 20 m. The shared MOTUs also represented a decreasing percent of reads in the water samples as we moved away from the rocky wall. On the other hand, there was no clear pattern of abundance changes with distance in the richness or amount of reads shared between

benthos and water for the 84 PSM.

We found therefore evidence for DNA originating from the benthic communities being present in the adjacent water layer and, conversely, DNA of presumably pelagic origin could be detected in the benthos. The interest of this article was in detecting the presence of benthic DNA in the water column, of which only a modest amount could be retrieved. The form of this benthic DNA in the water cannot be assessed with our sampling design, but it probably includes naturally released meroplanktonic components, such as gametes (Tsuji and Shibata, 2020) or larvae, and degradation products in the form of fragments, mucus, cell aggregates, exudates, or extracellular DNA.

Our results clearly indicated that DNA from water samples is a poor surrogate for the analysis of benthic communities, as found previously in freshwater environments (Gleason et al., 2021; Hajibabaei et al., 2019). Even in the water within a few centimetres from the benthos, only a modest portion (135) of the benthic MOTUs could be detected. In addition, we found that considering the relative number of reads of the shared MOTUs provided useful insights about the origin of the MOTUs and their dynamics as we move farther from the rocky wall. The lack of accordance between benthos and water is in agreement with previous comparisons of different substrates for eDNA made in port environments (e.g., Koziol et al., 2019; Rey et al., 2020) which found different community profiles in water and in sediments or settlement plates. We must keep in mind that we have used universal primers as we targeted the whole eukaryotic communities. With more specific targets, the results could be different. For instance, using vertebrate-specific primers to detect fish in the water has proved to be a sensitive method (e.g., Bakker et al., 2017; Sales et al., 2019; Salter et al., 2019; Sigsgaard et al., 2019; Thomsen et al., 2016), even at the intraspecific level (Sigsgaard et al., 2019), since it is possible to amplify selectively the DNA of the target group. Likewise, species-specific primers have been successfully used to detect particular marine benthic species in the water column, usually as a means of monitoring invasive species (e.g., Pochon et al., 2013; Simpson et al., 2017; von Ammon et al., 2019).

It seems reasonable to expect that DNA shedding rates from a highly diverse community such as sublittoral rocky bottom assemblages would be unbalanced between groups, and that this unevenness would hinder our ability to extract reliable monitoring information from seawater eDNA. This expectation is borne out by our results. Thus, albeit for group-specific or species-specific studies useful information from benthic groups may be gleaned from water DNA, the method is presently unsuitable for the community-wide diversity assessment required for many biomonitoring applications. New technologies affording much higher sequencing depth or metagenomic approaches (Singer et al., 2019, 2020) might improve our ability to extract information from water samples. But for the time being we must continue to rely on methods that can sample directly the benthos for reliable biodiversity assessment of these complex assemblages.

Chapter 5

From metabarcoding to metaphylogeography: separating the wheat from the chaff

5.1 Abstract

Metabarcoding is by now a well-established method for biodiversity assessment in terrestrial, freshwater, and marine environments. Metabarcoding data sets are usually used for α- and β-diversity estimates, that is, interspecies (or inter-MOTU [molecular operational taxonomic unit]) patterns. However, the use of hypervariable metabarcoding markers may provide an enormous amount of intraspecies (intra-MOTU) information—mostly untapped so far. The use of cytochrome oxidase (COI) amplicons is gaining momentum in metabarcoding studies targeting eukaryote richness. COI has been for a long time the marker of choice in population genetics and phylogeographic studies. Therefore, COI metabarcoding data sets may be used to study intraspecies patterns and phylogeographic features for hundreds of species simultaneously, opening a new field that we suggest to name metaphylogeography. The main challenge for the implementation of this approach is the separation of erroneous sequences from true intra-MOTU variation. Here, we develop a cleaning protocol based on changes in entropy of the different codon positions of the COI sequence, together with co-occurrence patterns of sequences. Using a data set of community DNA from several benthic littoral communities in the Mediterranean and Atlantic seas, we first tested by simulation on a subset of sequences a two-step cleaning approach consisting of a denoising step followed by a minimal abundance filtering. The procedure was then applied to the whole data set. We obtained a total of 563 MOTUs that were usable for phylogeographic inference. We used

semiquantitative rank data instead of read abundances to perform AMOVAs and haplotype networks. Genetic variability was mainly concentrated within samples, but with an important between seas component as well. There were intergroup differences in the amount of variability between and within communities in each sea. For two species, the results could be compared with traditional Sanger sequence data available for the same zones, giving similar patterns. Our study shows that metabarcoding data can be used to infer intra- and interpopulation genetic variability of many species at a time, providing a new method with great potential for basic biogeography, connectivity and dispersal studies, and for the more applied fields of conservation genetics, invasion genetics, and design of protected areas.

5.2 Introduction

Metabarcoding, whereby information on species present in a variety of communities can be obtained from so-called environmental DNA (eDNA), or from bulk or community DNA (Creer et al., 2016; Macher et al., 2018), is by now established as a robust method for biodiversity assessment (Adamowicz et al., 2019; Baird and Hajibabaei, 2012; Deiner et al., 2017; Taberlet et al., 2018).

Metabarcoding provides a fast and accurate method for measuring biodiversity, allowing identification of many more taxa (Molecular Operational Taxonomic Units or MOTUs) than morphological methods (Cowart et al., 2015; Dafforn et al., 2014; Elbrecht et al., 2017b), as small and cryptic organisms, early life stages, and fragments or trace DNA left in the environment can be targeted. Further, metabarcoding is largely independent of taxonomic expertise, which is dwindling worldwide (Wheeler et al., 2004), albeit it is highly dependent on the completeness of reference databases to reliably assign taxonomic names to MOTUs (Briski et al., 2016; Cowart et al., 2015). Taxonomic expertise, of course, will always be necessary to construct and expand accurate reference databases. Biodiversity assessment, detection of invasive or endangered species, paleoecological reconstruction, or diet analyses are among the main applications of metabarcoding to date (e.g., Ficetola et al., 2018; Hajibabaei et al., 2016; Ji et al., 2013; Kelly et al.,

2014b; Pochon et al., 2013). All of them are highly relevant for basic biodiversity research and for establishing management policies. There is, however, more information in metabarcoding data sets than just α- and β-diversity related issues. Further exploitation requires a shift from interspecies genetic patterns, that constitute most of the metabarcoding applications so far, to intraspecies genetic patterns (reviewed by Adams et al., 2019), making use of the within-MOTU genetic variability uncovered by metabarcoding.

Being heirs to studies in prokaryotes, eukaryotic metabarcoding initially relied heavily on ribosomal RNA sequences for MOTU delimitation (mostly nuclear 18S rDNA sequences). These sequences lack variability for within-MOTU studies in many groups, particularly metazoans (Leray and Knowlton, 2016; Tang et al., 2012; Wangensteen et al., 2018b). However, in recent years, intense efforts have been devoted to optimize the use of mitochondrial COI sequences in metabarcoding (Andújar et al., 2018). Their use was hindered by the lack of universal primers (Deagle et al., 2014), but new sets of COI primers for general purposes or for specific groups (Elbrecht and Leese, 2017; Günther et al., 2018; Leray et al., 2013; Vamos et al., 2017) are overcoming this problem and COI sequences are now being increasingly used in general biodiversity studies (e.g., Aylagas et al., 2016; Leray and Knowlton, 2015; Macher et al., 2018; Porter and Hajibabaei, 2018a), where they typically uncover a much higher degree of α-diversity than 18S rDNA (Stefanni et al., 2018; Wangensteen et al., 2018a,b). Furthermore, the use of COI opens the door to taxonomic assignment using the extensive database of the Barcode of Life Datasystems (BOLD), which is continuously increasing in depth and coverage (Porter and Hajibabaei, 2018b; Ratnasingham and Hebert, 2007).

COI sequences have been extensively used in studies of population genetics and phylogeography of terrestrial, freshwater, and marine organisms (Avise, 2009; Emerson et al., 2011). The shift to COI-based metabarcoding (Andújar et al., 2018), therefore, implies the generation of databases containing an untapped reservoir of intraspecies variation that can allow characterizing intra- and interpopulation genetic features of many species simultaneously. This could constitute a gigantic leap from the current single-species studies, effectively opening a new field in population genetics

for which we suggest the name of metaphylogeography.

The possibility of using metabarcoding for population genetics was hinted at by Bohmann et al. (2014) and Adams et al. (2019), but has been hardly developed. Current instances are in general preliminary, proof of concept, applications, always referred to particular taxa, not to whole community assessments. For instance, within- and between-population genetic structure using bulk DNA has been assessed for ichthyosporean parasites of the cladoceran *Daphnia* (González-Tortuero et al., 2015), for a *Xyleborus* beetle collected at two locations with differing management practices (Pedro et al., 2017) or for coral reef fishes of the genus *Lethrinus* (Stat et al., 2017). In the marine realm, eDNA from water has been used to obtain haplotype and ecotype information for species that are hard to sample, such as whale sharks (Sigsgaard et al., 2016), harbour porpoises (Parsons et al., 2018), or killer whales (Baker et al., 2018). In invasion biology, eDNA was proven useful to assess native vs. nonnative strains of common carp in Japan (Uchii et al., 2016).

An integrated phylogeography encompassing a range of species would be a powerful tool to investigate landscape-level processes (either natural or anthropogenic), over and above the signal given by each species. Studies that combine population genetics data on multiple species by traditional methods are costly and usually involve just a handful of species (e.g., Haye et al., 2014). The alternative is to use meta-analyses to collate the information scattered in different works (e.g., Pascual et al., 2017; Zink, 2002), or to use the information contained in georeferenced genetic databases (Gratton et al., 2017). However, the pace at which climate change affect our ecosystems and the projected increased exploration of our resources in the coming decades urge for increased knowledge of population structure and phylogeography at the global biome level. The potential of metaphylogeography ranges from basic questions about biogeography, connectivity, and dispersal patterns to more applied fields of conservation genetics, invasion genetics, and protected areas design. Nowadays, the consideration of multispecies genetic conservation objectives is seen as crucial to preserve community-wide genetic and evolutionary patterns (Nielsen et al.,

2017; Vellend et al., 2014).

The main problem for the application of eDNA or community DNA to analyze intraspecies patterns lies in the fact that this technique generates a high number of reads containing sequencing errors, which can occur at different steps in the procedure. Reads obtained by amplification and sequencing can be thought of as a "cloud" of erroneous sequences surrounding the correct one (Edgar and Flyvbjerg, 2015). Sequencing errors will typically occur as low-abundance reads with one or few base changes, while errors during amplification (PCR point errors, chimeras) have the potential of generating "daughter clouds" as they can reach higher read abundances (Edgar and Flyvbjerg, 2015). As erroneous sequences in general diverge very little from the true sequences, they are often incorporated into the right MOTU during the clustering step, thus reducing potential impacts on the results of "standard" metabarcoding approaches. However, they can severely bias intraspecies genetic patterns by artificially inflating the true haplotype diversity. Thus, separating the "wheat" (true sequences) from the "chaff" (false sequences) is the main challenge for the application of metabarcoding data to metaphylogeography.

To our knowledge, the problem of the correct assessment of intraspecific genetic diversity from community DNA in complex samples has been explicitly addressed only in a recent work by Elbrecht et al. (2018b). Using a single-species mock sample with known Sangersequenced haplotypes, they assayed a combination of denoising procedures to reduce the number of spurious haplotypes obtained using a metabarcoding pipeline. They then applied the best performing strategy to natural samples of freshwater invertebrates, deriving population genetic patterns for some of the species present.

We sought here to develop a practical strategy to make metabarcoding data sets amenable to phylogeographic studies. There are an ever-increasing number of such data sets publicly available in repositories. Mining COImetabarcoding data has been suggested for species discovery (Porter and Hajibabaei, 2018b), and these databases can be a resource for phylo-

geography as well. These data comprise different information, from raw sequences to filtered and paired sequences to simply MOTU tables. In many cases, no ground truth data or mock community analyses exist for them. We therefore need a strategy for cleaning noisy databases in the absence of ground truth information. We contend that the properties of coding sequences such as COI can provide such a strategy. Indeed, coding DNA sequences naturally have a high amount of variation concentrated in the third position of the codons, while errors at any step of the metabarcoding pipeline would be randomly distributed across codon positions. Examination of the change of diversity values (measured here as the entropy of each position; Schmitt and Herzel, 1997) as we eliminate noisy sequences can therefore guide the choice of the best cleaning parameters in the presence of an unknown amount of noisy data. Entropy values have been used previously to guide sequence trimming (Porter and Zhang, 2017) and OTU clustering (Eren et al., 2015), but never before in the context of distinguishing true variation from erroneous sequences.

A parallel inspection of the distribution of sequences across samples is also necessary. Error-containing sequences will typically co-occur in the same sample with the correct sequence, albeit with less abundance, and co-occurrence patterns can be incorporated to detect these sequences in cleaning steps. At the same time, while error sequences are likely to appear randomly in the samples, true sequences should feature a given ecological distribution, meaning that a sequence appearing in all replicates of a community, for instance, is unlikely to be an error. Distribution patterns of sequences have been suggested to guide MOTU calling or MOTU curating procedures (Frøslev et al., 2017; Olesen et al., 2017), but have not been applied, to our knowledge, for within-MOTU sequence curation.

Combining patterns of variation in entropy and sequence distribution patterns can lead to meaningful ways to reduce noisy data sets to operational data sets. This approach can be used to generate customized procedures for each different study system that take into consideration its particulars (replication level, pre-filtering applied, clustering procedure). It only requires that, for a given study, the information about which sequences have been

pooled in each MOTU in the clustering step, with their sample distribution, is provided.

We want to point out that the "metaphylogeography" concept is not equivalent to "conventional phylogeography of many species," and we therefore need to adapt some definitions. In particular, relative frequencies of reads of the different haplotypes are available instead of the relative frequencies of individuals bearing these. These are unlikely to be equivalent. The high difference in number of reads that can be obtained in metabarcoding can easily reach orders of magnitude and is hardly representative of conventional frequencies based on the number of individuals bearing a particular haplotype. Further, the quantitative value of metabarcoding data is debatable (Elbrecht and Leese, 2015; Piñol et al., 2019; Wares and Pappalardo, 2016). Once we have a curated data set, we suggest performing phylogeographic inference using a semiquantitative abundance ranking applied within each MOTU as a compromise between a strictly quantitative interpretation of the data, on one hand, and losing all the information contained in the number of reads on the other. For comparative inference, the traditional analytical framework including haplotype networks, AMOVA, and the like, is perfectly valid if one keeps in mind these differences in the interpretation of results.

In the present study, we developed cleaning strategies to make community data derived from COI amplicon sequencing amenable to the analysis of intraspecific variation. As a case study, we used a COI-based metabarcoding survey of biodiversity of sublittoral marine benthic communities. We then extracted phylogeographic trends from the MOTUs obtained with the best pruning parameters selected. We finally compared results with those of traditional phylogeographic studies for two species for which information exists for the same (or nearby) sampling areas. Our general goal was to show the feasibility of the metaphylogeographic approach using a "standard" metabarcarcoding data set obtained from natural samples.

Table 5.1: Sample characteristics, with indication of locality, type of community, dominant species, depth, coordinates, and number of replicate samples collected in each study year.

			No. samples	
National park, community, and dominant species	**Depth (m)**	**Coordinates**	**2014**	**2015**
Cabrera Archipelago				
Photophilic algae				
Lophocladia lallemandii	7–10	39.1250° N, 2.9603° E	3	3
Padina pavonica	7–10	39.1250° N, 2.9603° E	3	3
Sciaphilic algae				
Sponges and invertebrates	30	39.1250° N, 2.9603° E	3	3
Caulerpa cylindracea	30	39.1250° N, 2.9603° E	-	3
Detritic bottoms				
Coralline algae	50	39.1249° N, 2.9604° E	3	3
Atlantic Islands				
Photophilic algae				
Cystoseira nodicaulis	3-5	42.2259° N, 8.8969° W	3	3
Cystoseira tamariscifolia	3-5	42.2260° N, 8.8970° W	3	-
Asparagopsis armata	4-6	42,2146° N, 8.8973° W	-	3
Sciaphilic algae				
Saccorhiza polyschides	16	42.1917° N, 8.8885° W	3	3
Detritic bottoms				
Coralline algae	20	42.2123° N, 8.8972° W	3	3

5.3 Material and Methods

5.3.1 Data set

The data set consisted of COI-based biodiversity data obtained from benthic marine communities in two Spanish National Parks, one in the Atlantic and one in the Mediterranean (Fig. B.0.1). The data set has different replication levels: over time (two years), within communities (sample replicates), and within samples (size fractions). Sample collection and processing followed Wangensteen and Turon (2017) and Wangensteen et al. (2018b). In short, several communities were sampled in 2014 and 2015 by completely scraping off standardized 25 × 25 cm quadrats in hard bottom substrates or by sampling with PVC corers, 24 cm in diameter, in detritic communities. Three replicate samples were collected per community, and each sample was then separated through sieving into three size fractions (>10 mm, 1–10 mm, 63 μm–1 mm, roughly corresponding to mega-, macro-, and meiobenthos; Rex and Etter, 2010). A total of 51 samples separated in 153 fractions were included in the present study (Table 5.1).

The sampling performed in 2014 included four communities in the Mediterranean Park (Cabrera Archipelago, Balearic Islands) and four in

the Atlantic Park (Atlantic Islands of Galicia). These communities were, in each Park, two well-lit communities, one deeper, invertebrate-dominated, community, and a detritic bottom with coralline algae (Table 5.1). In 2015, the sampling was repeated on the same localities and communities, except for a new community sampled in Cabrera (*Caulerpa cylindracea* community) and the change of one of the two well-lit communities in the Atlantic (*Asparagopsis armata* community instead of *Cystoseira tamariscifolia* community, Table 5.1). Wangensteen et al. (2018b) reported α- and β-diversity results of the sampling performed in 2014, while some of the communities sampled in 2015 were used in a study of the effect of invasive seaweeds (Wangensteen et al., 2018a).

Samples were extracted and sequenced using the Leray-XT primer set, a modification of the Leray et al. (2013) primers for a 313 base pair (bp) fragment of COI, with the adequate blanks and negatives, following procedures detailed in Wangensteen et al. (2018b). Separate libraries were built with samples from 2014 and 2015 and sequenced in two runs on an Illumina MiSeq platform (2 $\times$ 300 bp paired-end) at Fasteris SA (Plan-les-Ouates, Switzerland).

For the present study, we pooled the reads of the two years and analyzed the joint data set with a pipeline based mostly on the OBITools suite (Boyer et al., 2016). The length of the raw reads was trimmed to a median Phred quality score higher than 30, after which pairedreads were assembled using illuminapairedend. The reads with paired-end alignment quality scores higher than 40 were demultiplexed using ngsfilter, which also removed the primer sequences. For this study, we applied a strict length filter keeping only sequences of the expected length (313 bp). Identical sequences were then dereplicated (using obiuniq) and chimeric sequences were detected and removed using the uchime_denovo algorithm implemented in vsearch v1.10.1 (Rognes et al., 2016). At this step, we discarded sequences with just one read in all the data set, as is common practice in metabarcoding studies. We clustered sequences into MOTUs using the SWARM2 method (Mahé et al., 2015), with a d-parameter of 13. This parameter was set for the COI fragment used here after comparing the number of MOTUs obtained

at different values and checking that this number remained constant for values of d in the range of 9–13. The value of $d = 13$ has been previously used in other studies involving the same COI fragment (Kemp et al., 2019; Macías-Hernández et al., 2018; Siegenthaler et al., 2019a).

The taxonomic assignment of the MOTU was performed using ecotag (Boyer et al., 2016), which uses a local reference database and a phylogenetic tree-based approach (using the NCBI taxonomy) for assigning sequences without a perfect match. Ecotag searches the best hit in the reference database and builds the set of sequences in the database that are at least as similar to the best hit as the query sequence is. Then, the MOTU is assigned to the most recent common ancestor to all these sequences in the NCBI taxonomy tree. With this procedure, the assigned taxonomic rank varies depending on the similarity of the query sequences and the density of the reference database. We developed a mixed reference database by joining sequences obtained from two sources: in silico ecoPCR against the release 117 of the EMBL nucleotide database and a second set of sequences obtained from the Barcode of Life Datasystems (Ratnasingham and Hebert, 2007) using a custom R script to select the Leray fragment. Details of this newly generated database (db_COI_MBPK) are given in Wangensteen et al. (2018b). It includes 188,929 reference sequences and is available online (http://github.com/metabarpark/Reference-databases).

Following the pipeline, we generated an MOTU list and assigned a taxonomical rank to each MOTU. Noneukaryotic MOTUs were removed. Occasionally, two or more MOTUs received the same species-level assignment, in which case, only the most abundant MOTU was retained and the reads of the others were added to it (this happened in 349 species). We also pooled the sequences of the three fractions of each sample for downstream analyses. For the goal of this study, not all MOTUs carried the phylogeographic information sought (i.e., genetic variation within and between communities and seas). We therefore performed a previous selection in which we included MOTUs that had at least two different sequences (i.e., displayed intra-MOTU structure). We also required that the MOTU appeared in the two Parks with 20 or more reads in each one, and appeared at least once in each

of the two study years. We acknowledge that this selection is arbitrary, but these limits were set to ensure that the MOTUs were minimally abundant and widely distributed for reliable phylogeographic inference. Note that this MOTU selection does not imply that discarded MOTUs are artefacts, but simply that they are not useful for population genetics inference (e.g., one MOTU appearing only in a given community, even if abundant).

Using the list of retained MOTUs, the original sequence file, and the information of which sequence belongs to each MOTU (contained in the output of the clustering program used to generate MOTUs), we obtained separate MOTU files containing, for each MOTU, all sequences included with their abundances in the different samples. We then aligned sequences within each MOTU with the msa R package (Bodenhofer et al., 2015), and misaligned sequences, likely due to slippage of degenerate primers (Elbrecht et al., 2018a), were detected and eliminated.

5.3.2 Simulation analysis

All data manipulation and analyses were conducted using R software (R Core Team, 2020). To avoid confusion between different terms, sometimes used interchangeably, we will use the name denoising to refer to any procedure that tries to infer which sequences contain errors and merges their reads with those of the correct "mother" sequence. We will call filtering any method that actually deletes sequences from the data set, based on abundance thresholds or otherwise. Clustering will refer to any procedure for combining sequences, without regard to whether they are correct or not, into meaningful MOTUs.

We ran a simulation study to infer the best cleaning strategy and the best parameters for our data. The rationale was to start with a known data set, introduce sequencing errors, and clean it again to recover the original data set. We used a custom R script for this simulation. Following (Wang et al., 2012), we considered that the 1,000 sequences with highest frequency (in read number) in our data set were error free, and used them for parameter estimation on a data set representative of our actual sequences. For this simulation, we did not keep the ecological information and used just the total number of reads of each of these 1,000 top sequences.

We simulated that these allegedly correct amplicons were sequenced with error rates between 0.001 and 0.01 per base, bracketing values published for HTS sequencers and, in particular, for the MiSeq platform (Pfeiffer et al., 2018; Schirmer et al., 2016). For simplicity, we assumed a constant error rate for all bases in a sequence, albeit we acknowledge that this is a simplification as sequence features such as homopolymer regions make some positions more prone to errors (Taberlet et al., 2018).

For the highest error rate (0.01), we then denoised the resulting sequences using a procedure adapted from the algorithm of Edgar (2016). We merged the reads of presumably incorrect daughter sequences with those of the correct mother sequences if the number of sequence differences (d) is small and the abundance of the incorrect sequence with respect to the correct one (abundance ratio) is low. The higher the number of differences, the lower the ratio should be for the sequences to be merged. This was formalized by the expression Edgar (2016)

$$\beta(d) = 1/2^{\alpha d+1}$$

where $\beta(d)$ is the maximum abundance ratio allowed between two sequences separated by d changes so that the less abundant was merged with the more abundant. The α parameter is user-settable to seek a compromise between accepting as correct erroneous sequences (high α values) or merging true sequences (low α values). The denoising was done for values of α from 10 to 1.

We analyzed changes in diversity of the different codon positions as we introduced increasing levels of noise (erroneous reads) and as we denoised the data set with increased stringency (lower α values). As a measure of diversity, we used the Shannon entropy value computed with the R package entropy (Hausser and Strimmer, 2009). We expected that random error will increase more the entropy of the less variable position (second position of the codons) and less the entropy of the third, more variable, position. Thus, the entropy ratio (hereafter E_r)

$$E_r = entropy\ position2/entropy\ position3$$

was expected to increase as simulated error rates increased and to decrease when denoising. After each round of denoising we noted the number of original sequences remaining, the number of noisy sequences remaining, and the entropy ratio of the sequences. We expected that at some value of α the E_r will reach the original value and remain more or less constant afterwards. As at this point many erroneous sequences remained in the data set (see *Results*), we completed the simulation with a filtering procedure in which low frequency sequences were eliminated.

We assayed a range of minimal number of reads to keep a sequence and looked at the number of original and noisy sequences remaining, as well as their entropy ratio. As before, we expected the E_r to decrease markedly and stabilize after some threshold is reached. The best α parameter and the best minimal number of reads should allow us to recover most of the original sequences with as few erroneous sequences as possible.

5.3.3 Data set cleaning

The cleaning procedure followed the findings of the simulation and was therefore based on two steps: denoising (without loss of reads) and filtering by minimal abundance (with loss of reads). We applied denoising within defined MOTUs, under the assumption that most erroneous sequences would have been included in the same MOTU as the correct sequence, and thus sequence distances and abundances, a key part of the denoising algorithm, are more meaningful if compared within MOTUs. Once denoising was complete and, thus, all "salvageable" sequences had been merged with the correct sequence, the second step consisted of an abundance filtering, in which low-abundance sequences, likely erroneous, "surviving" the denoising step were eliminated.

During the previous steps, co-occurrence patterns were used to avoid merging or eliminating sequences whose sample distribution and co-occurrence patterns suggested they were not artifacts (for instance, sequences that do not co-occur with similar sequences will not be merged with them, and sequences found in all replicates of a community will not be filtered out). The use of distribution data can reduce the risk of eliminating

true sequences, particularly when they are present at low abundances (e.g., reflecting a low biomass of the organism).

To allow a daughter sequence presumed to be a sequencing error to be merged with a more abundant mother sequence, we required that the former co-occurs with the latter. This is formalized by a co-occurrence (C_{occ}) ratio in the form

$$C_{occ} = daughter/(daughter + mother)$$

were daughter is the number of samples with only the daughter sequence and daughter + mother is the number of samples with the daughter and the mother sequence. The higher the ratio, the less we will merge sequences, as we require a higher co-occurrence with the mother sequence.

We set this parameter to a value of 1 (i.e., whenever a daughter sequence was present, the mother sequence was present in the same sample). Any "daughter" sequence with co-occurrence ratio <1 was considered a genuine sequence and was not merged. This is a conservative value that seeks to avoid merging potentially good sequences. It was set considering that we enforce the presence at the sample level, and not at the fraction level, which means that the sequence needs to be present in just one of the three fractions (10 mm, 1 mm, 63 μm) of the sample. In preliminary assays, changing C_{occ} influenced the number of sequences retained, but represented little change in the entropy ratios obtained. In addition, in the filtering step sequences appearing in all replicates of a given community were considered correct and not filtered out, even if present at low abundance.

Taking these distribution patterns into consideration we applied the denoising and filtering steps. A diagrammatic representation of the pipeline used is presented in Figure 5.1. Denoising was performed at α values between 10 and 1, and for the best-performing α, filtering was done for increasing minimal numbers of reads from 2 to 100. After each round of sequence denoising or filtering, the MOTUs were examined and retained only if they still met the requirements of having at least two sequences, appearing in the two Parks with 20 or more reads in each one, and appearing at least once in the two study years. The changes in E_r of the retained MOTUs were

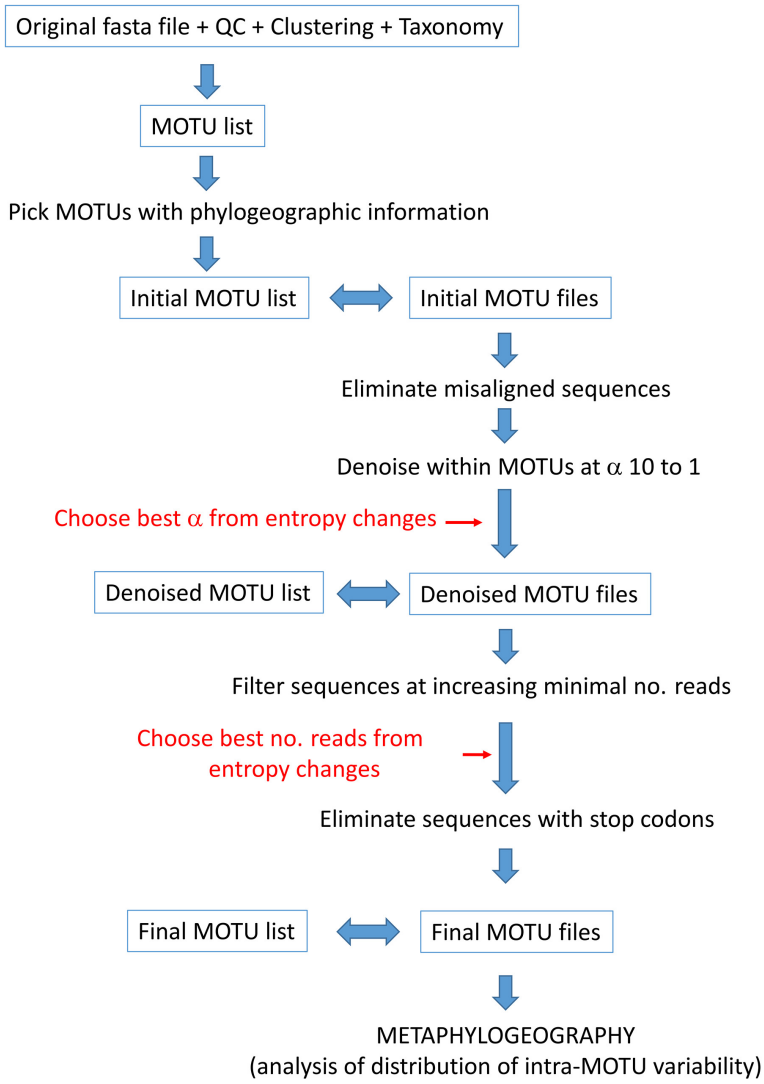


Figure 5.1: Schematic representation of the pipeline followed in this study. See *Methods* for details. The red arrows and text indicate the two steps in the pipeline where parameter selection should be carried out based on entropy values. MOTU, molecular operational taxonomic unit.

examined over the range of α and minimal abundance values. In both cases, the entropy ratio should decrease and, following the simulation results, the points where it became stabilized (we chose as a threshold the point at which the slope fell below 0.005) were used as optimal parameter cutoffs.

Finally, even if sequences retained were mostly correct, they can still

include a number of nontarget variants due to heteroplasmy or numts (Elbrecht et al., 2018b). However, numts tend to accumulate mutations resulting in stop codons (Song et al., 2008). They can also present amino acid substitutions that result in a non-functional protein: Pentinsaari et al. (2016) found 23 amino acids completely conserved across the COI barcode region in Metazoa, corresponding mostly to the helices of the protein that penetrate the mitochondrial membrane. Five of these amino acid positions occur in the fragment sequenced here. Some numts can therefore be detected by inspecting the sequences retained, as has been done in previous metabarcoding studies (Leray et al., 2013). As the data set included many different eukaryotic groups with different genetic codes, we adopted a conservative approach. For each MOTU, we tried the 20 genetic code variants stored in the Biostrings R package (Pagès et al., 2008) and used the *translate* function to obtain the corresponding amino acid sequence. We then chose, for each MOTU, the genetic code giving the lower number of stop codons (often several code variants resulted in no stop codons). In addition, we verified (for the metazoan MOTUs) that the five conserved positions described above did not have any amino acid substitution. The MOTUs denoised with the optimal α-value (first step), once filtered with the optimal abundance cutoff (second step) were checked for the presence of stop codons and amino acid changes, and the sequences presenting them were removed from the data set. The remaining MOTUs and sequences constituted the curated data set for further analyses (Fig. 5.1).

5.3.4 Metaphylogeographic analyses

We performed network analyses with function HaploNet of the R package pegas (Paradis, 2010). We used function amova of the R package ade4 (Dray and Dufour, 2007) to compute analyses of molecular variance (AMOVA) in order to ascertain the percent variation associated with the hierarchical organization of the samples. For AMOVA, we used the proportion of the different sequences present (option distances = NULL). Preliminary assays considering also sequence distances (not just sequence frequencies) gave highly similar results and were computationally slower.

In these analyses, we needed to capture the quantitative information regarding frequencies of the different sequence variants. As mentioned above, using number of reads as a proxy for individual-based abundances can be misleading. We adopted a semiquantitative index based on Wangensteen et al. (2018a) applied within each MOTU. To obtain this semiquantitative ranking, we ordered the sequences of each sample in each MOTU by increasing number of reads and ranked them from 0 to 4, indicating that the sequence is either absent in that sample (rank 0) or falls in the following percentiles of the distribution of ordered sequences: rank 1, $\leq$50%; rank 2, $>50 \leq 75\%$; rank 3, $>75 \leq 90\%$; rank 4, >90%. These semiquantitative ranks were used as proxies for haplotype abundances in the analyses.

5.3.5 Comparison with previous studies

After examination of the curated MOTU data set, we found only two species for which conventional phylogeographic analyses had been performed using COI information in the same geographic area: the sea urchin *Paracentrotus lividu* and the brittle star *Ophiothrix fragilis*.

For *Paracentrotus lividus*, we collated haplotype information from studies spanning the Atlanto-Mediterranean transition (Duran et al., 2004), trimmed the sequences to the same fragment amplified in our study, and compared the haplotypes with the ones encountered in our metabarcoding data set. Duran et al. (2004) included two populations close to our localities: Eivissa Island in the Balearic Archipelago, and Ferrol in the Galician coast. Networks were generated with the haplotypes found in these localities and compared with our results.

For *Ophiothrix fragilis*, our MOTU corresponded to Lineage II of Pérez-Portela et al. (2013). This brittle star is in fact a complex of species, and Lineage II is likely a cryptic species (Taboada and Pérez-Portela, 2016), but it remains unnamed so far. As before, we extracted haplotype information from all localities in Pérez-Portela et al. (2013), spanning the Atlanto-Mediterranean area, and compared with our results. We also obtained haplotype networks for the two closest populations studied in that work: Alcudia in the Balearic Archipelago and Ferrol in the Galician coast.

5.4 Results

5.4.1 The data set

The original data set, once quality and length filtered, contained 25,772,264 sequences of 8,900,080 unique sequences. Without singletons, the numbers were reduced to 17,808,524 reads and 936,340 unique sequences. Following the pipeline, we obtained a MOTU list of 26,561 eukaryote MOTUs. Of these, 13,410 MOTUs were present only in the Mediterranean site, 8,247 only in the Atlantic locality, and 4,904 were shared by both basins. Of the latter, only 722 MOTUs (with a total of 362,177 unique sequences and 9,430,236 reads) fulfilled the conditions that we set for the metaphylogeographic analyses (see *Methods*) of having at least two sequences, being present in the two Parks with at least 20 reads in each one, and having appeared in the two years of study. After checking the alignment, only 158 sequences, comprising 689 reads, appeared as misaligned, mostly as a result of 1 bp slippage, and were removed. The singleton-free fasta sequence file (paired, demultiplexed, and quality-filtered), the original MOTU list, and the output of the SWARM analyses have been uploaded as a Mendeley data set (see *Data Availability*). The 722 MOTUs selected for the study are listed in Data B.1.1, together with their taxonomic assignment and abundance (number of reads) per sample. The actual sequences of each MOTU, with their abundances per sample, are available at the Mendeley data set.

5.4.2 Simulation study

In our case, the top 1,000 sequences in the 722 MOTUs data set contained 5,948,135 reads. The entropy values of the codon positions of these sequences were: first position, 0.4298 ± 0.037 bits (mean ± SE); second position, 0.1833 ± 0.028 bits; third position, 0.9256 ± 0.023 bits. The simulation of increasing sequencing error rates clearly increased the entropy of the three positions (Fig. 5.2A), but more so for the less variable second position, which increased its value 30% at the highest error rate. On the other hand, the third position increased entropy only about 1.8%. As a result, the entropy ratio (E_r , entropy2/entropy3) increased linearly with error rate, from 0.198 to 0.252 (Fig. 5.2B).

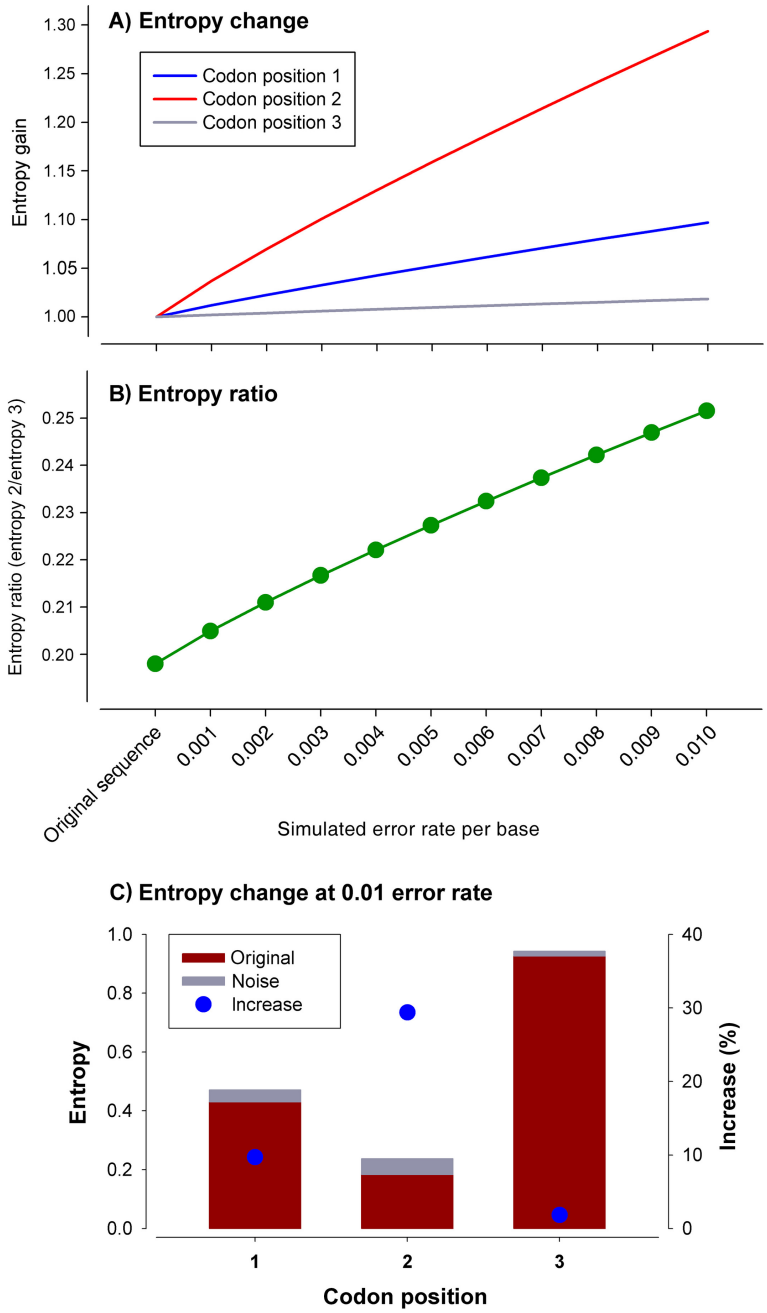


Figure 5.2: Simulation analysis. (A) Relative increase (initial value = 1) of the entropy values of each position at increased error rates. Bar plot shows the original and added entropy of each position at the highest (0.01) error rate. (B) Change in the entropy ratio. (C) Bar plot showing the original and added entropy of each position at the highest (0.01) error rate.

We then used the "noisy-most" data set, the one simulated at the highest (0.01%) error rate. It had the same original number of reads, but 5,141,683 erroneous sequences (besides the 1,000 correct ones) were generated. For coherence with the global data set used, singletons were removed, leaving 144,791 sequences. This data set was then denoised at a values between 10 (least stringent) and 1 (most stringent). The E_r decreased drastically at the initial steps, concomitantly with a decrease in the number of erroneous sequences (Fig. 3A). The E_r value of the simulated data set reached the original value at α between 6 and 5. Taking the more conservative $\alpha = 5$, which is also the point where the entropy curve levelled off (slope < 0.005), we found that the data set contained 895 of the original sequences and 17,799 erroneous sequences. In other words, while 10% of the original sequences have been incorrectly merged, there remained still a high number of errors in the data set. Using only the denoise procedure, we got completely rid of erroneous sequences only at $\alpha = 1$. But at this value only 66% of the correct sequences were retained.

We therefore applied a round of filtering by minimal number of reads to the data set denoised at $\alpha = 5$. Again, the E_r decreased sharply at increasing thresholds of minimal reads, following the elimination of erroneous sequences (Fig. 5.3B), and stabilized clearly at seven reads (Fig. 5.3B). The combination of denoising ($\alpha = 5$) and filtering (minimal abundance = 7) allowed us to recover 924 sequences, of which 895 (97%) were among the 1,000 original sequences and only 3% were erroneous sequences. The frequency distribution of the number of reads in both the original (1,000) and the recovered (924) sequences was almost identical (not shown). Importantly, the shape of the E_r curve, specifically the stabilization points, proved informative to select the cut-points for the two variables.

5.4.3 Data set cleaning

As a first step, we tried to identify PCR errors during amplification, as they can result in abundant sequences and be more difficult to spot. We assumed that PCR errors will affect one nucleotide at most, will occur in few samples, where they will coexist with the original sequence, and will

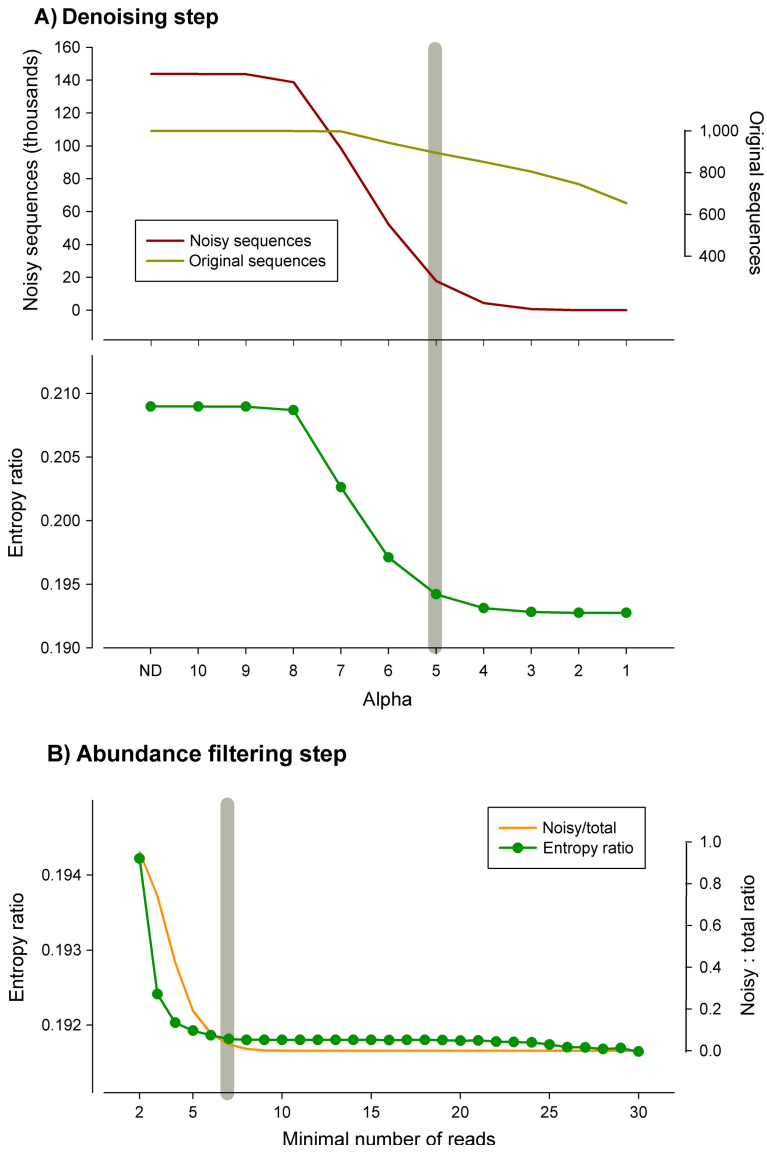


Figure 5.3: Simulation analysis. (A) Variation in the number of original and erroneous ("noisy") sequences and entropy ratio at decreasing values of the alpha parameter of the denoising algorithm (ND, no denoising). (B) Change in the entropy ratio and in proportion of noisy vs. original sequences after filtering the data set by minimal abundance. The gray bars indicate the selected values of alpha (5) and minimal number of reads (7).

be abundant. Therefore, we looked within the 722 MOTUs for sequences differing by one nucleotide from a more abundant one, co-occurring always with it, being present in at most three samples (out of 51 samples), and having an abundance of >200 reads (set as a threshold to identify relatively abundant sequences). Only 14 such sequences were identified and merged with the more abundant ones.

After applying the denoising step for α values from 10 to 1 and a co-occurrence index of 1 to the whole data set of 722 MOTUs, we examined the change in number of retained MOTUs and entropy ratio (Fig. 5.4A). The number of MOTUs remained constant but started decreasing at $\alpha = 6$. As expected, the E_r decreased fast at first and more slowly at lower α-values (i.e., with higher merging power) (Fig. 5.4A). The curve leveled off (slope below 0.005) at $\alpha = 5$, with only a slight loss of MOTUs (six out of 722). We thus retained $\alpha = 5$ as the optimal denoising parameter.

The MOTU list corresponding to the denoised data set had 716 MOTUs, with 49,995 sequences (86% of the original sequences had been merged) and 9,426,339 reads (Data B.1.1). The corresponding MOTU files (available at the Mendeley data set; see *Data Availability*) were submitted to an abundance filter, with a threshold from 2 to 100 reads. There was a decrease the number of MOTUs retained at increasing minimal numbers of reads, particularly in the interval 2–50 (Fig. 5.4B). The entropy ratio fell markedly and became stabilized at a value of 20 reads, after which it remained more or less constant (Fig. 5.4B). Thus, 20 reads was used as a minimal abundance threshold.

The sequences of the resulting MOTU files were translated and checked. Only eight sequences had stop codons, while a further 52 metazoan sequences had amino acid changes in the five positions invariable in Metazoa. These 60 sequences were eliminated, and the final MOTU list thus consisted of 563 MOTUs, with 7,146 sequences and 8,910,913 reads (Data B.1.2). The final MOTU files were uploaded to the Mendeley data set (see *Data Availability*).

As for the taxonomy assigned, the most diverse groups of Eukarya in the final data set were Rhodophyta (91 MOTUs), Stramenopiles (90 MO-

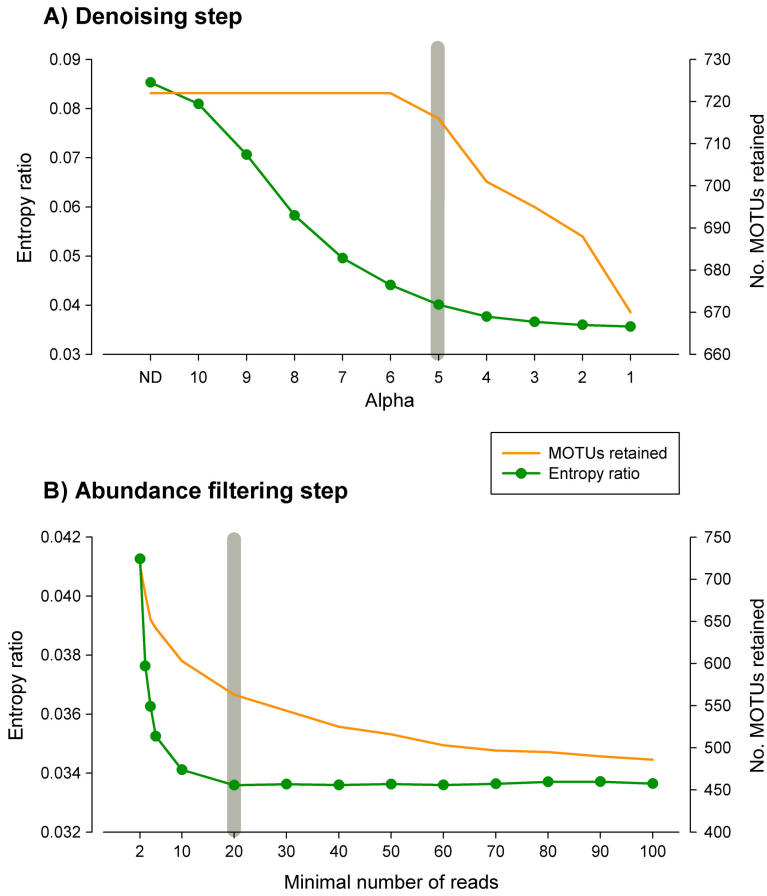


Figure 5.4: Final analyses of the littoral communities data set. (A) Variation in the number of sequences and number of MOTUs remaining at decreasing values of the alpha parameter (ND, no denoising) of the denoising algorithm. (B) Change in the entropy ratio and (C) change in residual (within-sample) variance of the amova model. The gray bars indicate the selected alpha value (5) and abundance threshold (20).

TUs, mostly diatoms and brown algae), and Metazoa (273 MOTUs) (Data B.1.2). A total of 99 eukaryotic MOTUs remained unassigned taxonomically (identified as Eukarya). Among metazoans, 112 MOTUs were assigned a species-level taxon, while 225 MOTUs were assigned at least at the phylum level and 48 MOTUs remained unassigned (Data B.1.2). The phyla of metazoans identified in the final MOTU list were Annelida (34 MOTUs), Arthropoda (56 MOTUs), Bryozoa (17 MOTUs), Chordata (eight MOTUs), Echinodermata (seven MOTUs), Mollusca (22 MOTUs), Nemertea (six MOTUs), Porifera (30 MOTUs), and Xenacoelomorpha (one MOTU).

Further analyses concentrated in the major groups detected, which accounted for 437 of the 464 MOTUs that could be assigned: red algae (Rhodophyta), diatoms (Bacillariophyta), brown algae (Phaeophyceae), and metazoans (Metazoa). In the latter, phylum-level analyses were performed.

5.4.4 Phylogeography

Network graphs of the MOTUs (Data B.1.3) showed different patterns, albeit in most cases one or a few haplotypes appeared as the most abundant, linked to a varying number of low abundance haplotypes. Some selected instances are presented in Figure 5.5, showing also the change in network shape along the process of cleaning. It can be seen that the major pruning effect was due to the initial denoising step.

AMOVAs were used to partition the genetic variance hierarchically into components due to the differences between seas, between communities within seas, between samples (replicates) within communities, and within samples. The average values of these variance components for the major groups detected, and for metazoan phyla separately, displayed a clear overall trend: genetic variance was concentrated within samples (60–75%) in all major groups (Fig. 5.6A). The other components of variance followed a decreasing trend, with a remarkable variance associated to differentiation between the two seas (14–25% of variance), and smaller variance between communities within each sea, and even lower between replicate samples of a given community. The latter component was almost negligible (<1.2%) in the nonmetazoan groups considered, but reached 5.4% in metazoans. The

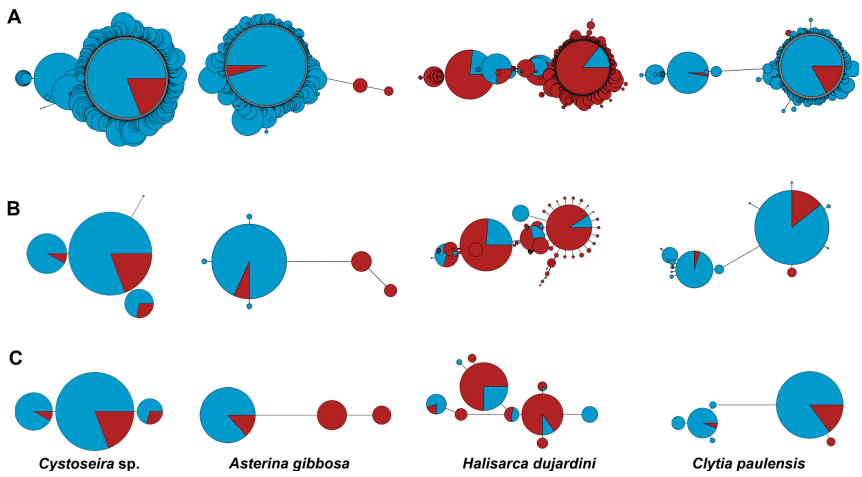


Figure 5.5: Selected instances of networks obtained at different stages of the pipeline: (A) without filters; (B) after denoising at alpha = 5; (C) after denoising at alpha = 5 plus minimal abundance filtering (threshold 20 reads). Circles represent haplotypes, and their diameters are proportional to their abundance (in semiquantitative ranks) in the samples. Blue color represent abundance in Mediterranean samples, red color in Atlantic samples. Length of links is proportional to the number of mutational steps between haplotypes. Note that circles in panels A, B, and C are not drawn to the same scale. The names correspond to the taxonomical identification of the MOTUs with ecotag (OBITools package). The MOTU ids (as per Data B.1.1) are, from left to right, 143, 1740, 2500, and 25366.

different components were compared across groups with ANOVA (followed by Student-NewmannKeuls post hoc tests if significant). The between sample component was significantly higher (all $P < 0.001$) in metazoans than in the other groups. For the other components, the values were in general comparable, the only significant differences being a higher between seas differentiation in diatoms than in metazoans, and a higher within sample variance in red algae than in diatoms.

Metazoans therefore showed a higher heterogeneity between replicate samples of a given community than the other groups. When examined across phyla (Fig. 5.6B), albeit the overall trend was in general maintained, a dominant within sample component and a variance between seas > between communities > between samples, there were exceptions. In particular, molluscs had a high between sample variability, and other groups presented important small-scale (between communities and/or between samples) vari-

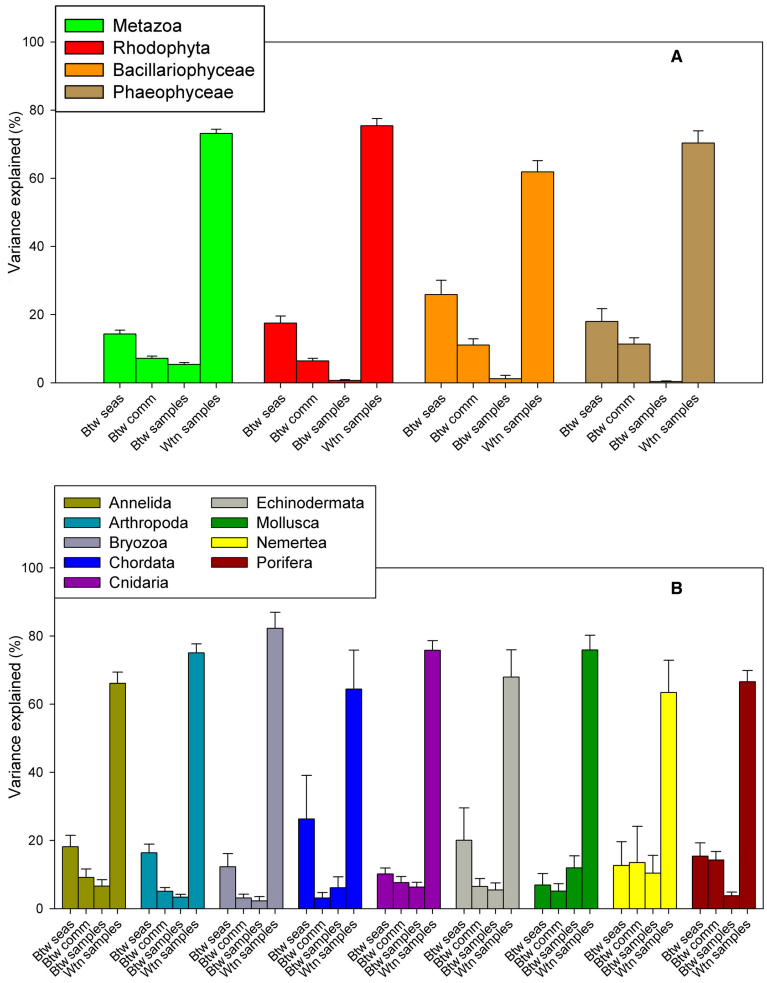


Figure 5.6: Summary of the mean percentage of variance explained by the hierarchical structure of the AMOVA: (A) as per eukaryote groups; (B) per metazoan phyla. Error bars are standard errors. Btw seas, between seas; btw comm, between communities within seas; btw samples, between samples within communities; wtn samples, within samples.

ability as compared to the between seas differentiation (Cnidaria, Nemertea, Porifera). ANOVA showed few significant differences between phyla, the only significant comparisons involving the between samples component in molluscs, which was significantly higher than in bryozoans or sponges.

As for the comparison with previous studies, MOTU 697 was identified as the sea urchin *Paracentrotus lividus* with 100% sequence identity. This MOTU had 15 sequences. This species has an Atlanto-Mediterranean distribution and Duran et al. (2004) analyzed populations spanning the western Mediterranean and northeast Atlantic with COI. In that work, 65 different haplotypes (of a longer fragment of COI) were detected. Once trimmed to our sequence length and collapsed, there were 32 remaining haplotypes. Nine out of the 15 sequences detected in our study had already been found by Duran and co-workers, while the remaining six were new.

We then selected the haplotypes found in the previous work in the two localities closest to our sampling points (Eivissa in Balearic Islands and Ferrol in Galicia). There were 11 haplotypes (four of which were also present in our MOTU). We performed a network with the 2004 information and compared it with the one obtained for MOTU 697 with our semiquantitative abundance rank (Fig. 5.7A, B). The two networks had a similar shape, with a highest abundance of haplotype 2 (named after the order of abundance of sequences obtained for this MOTU), followed by haplotypes 1, 3, and 6. For the shared haplotypes, the between seas distribution was the same in the two studies (1, 2, and 3 shared between seas, six present only in the Atlantic). An AMOVA with a randomization test ($n = 1{,}000$) of our MOTU 697 revealed a significant differentiation between seas and between and within samples ($P < 0.001$) but not between communities ($P = 0.812$).

The MOTU 15396, comprising 37 sequences, was identified (100% identity) with *Ophiothrix* sp. in Pérez-Portela et al. (2013). In that work, the authors studied a controversial species complex of the genus *Ophiothrix* in the European waters using 16S and COI. Our sequences corresponded to the Lineage II of *Ophiothrix fragilis* in that work, that spanned from Britanny to Turkey. Pérez-Portela et al. (2013) reported 125 haplotypes of Lineage II

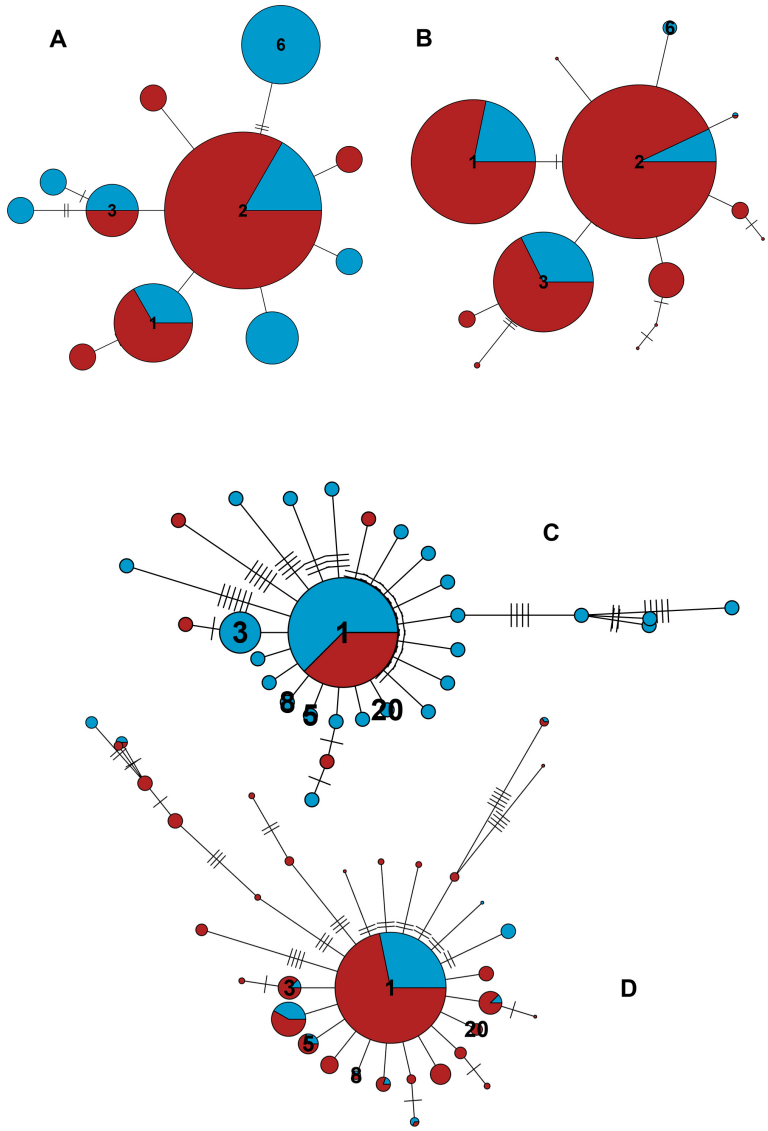


Figure 5.7: (A) Network constructed with the 11 haplotypes of the sea urchin *Paracentrotus lividus* found by Duran et al. (2004) in localities close to our sampling points and (B) network constructed with the 13 haplotypes comprising the MOTU corresponding to this species (id 697). Haplotypes common to both studies are numbered. (C) Network with the 29 haplotypes of the brittle star *Ophiothrix fragilis* identified by Pérez-Portela et al. (2013) in localities close to our sampling points. (D) Network of the 34 haplotypes found in the present study in the MOTU corresponding to this species (id 15396). Haplotypes common to both studies are numbered. The short slashes in the links between haplotypes represent mutational steps. Colors as in Fig. 5.5.

that, once trimmed to our 313 bp length, resulted in 90 different haplotypes. When merged with our data set, nine out of 37 sequences in MOTU 15396 had already been found in the previous study, while another 28 were new.

As before, we selected in Pérez-Portela et al. (2013) the two localities closest to our sampling points (Alcudia in Balearic Islands, and Ferrol in Galicia). There were 29 haplotypes in these localities, of which five were shared with our study. The corresponding networks (Fig. 5.7C, D) showed a star-shaped structure with a dominant haplotype 1 found in the two studies, with many low abundance sequences separated by one or a few mutations from the central haplotype and some longer branches. It is noteworthy that, in this case, the shared haplotypes do not have always the same inter-basin distribution, thus, haplotype 1 was present in both oceans, but haplotypes 3, 8, and 5 present only in the Mediterranean site in the previous work, appeared now in the two seas (it should be noted that haplotype 3 did appear in other Atlantic sites in Pérez-Portela et al., 2013). Finally, haplotype 20 was present only in the Mediterranean site in Pérez-Portela et al. (2013) and only in the Atlantic locality in information contained in the data sets, we think that it is more advisable to define meaningful MOTUs and perform denoising procedures within them, in order to obtain a "clean" data set and be able to use the intraMOTU sequence variability to make phylogeographic and population genetics inference. Clearly, our procedure is applicable only to coding sequences, which excludes much work done on protists based on ribosomal DNA. However, the growing number of metabarcoding studies using COI sequence data, together with the present work. An AMOVA with a randomization test (n = 1,000) of our MOTU 15396 showed a significant component of variation related to between and within samples genetic variability ($P < 0.001$), but not between seas ($P = 0.729$) or between communities within seas ($P = 0.212$).

5.5 Discussion

In this study, we have developed a method to apply metabarcoding data sets to the study of intraspecies patterns of many species at a time using a highly variable coding fragment (COI). An initial denoising step, aimed at merging

erroneous sequences with the correct ones, was followed by an abundance filtering step aimed at removing the remaining erroneous sequences. We used information from the variability of the different codon positions, following a simulation study, to select the best parameter values in the denoising and filtering steps. In addition, sample distribution information was used in the different steps to minimize loss of low abundance true sequences.

All cleaning procedures are a compromise between eliminating spurious sequences and losing true signal. In the benchmarking approach of Elbrecht et al. (2018b), 943 erroneous haplotypes appeared in a sample known to have only 15 before any processing. After a denoising process, 15 haplotypes remained but, of these, 6 (40%) were still sequences not present in the original sample, while 6 of the 15 original variants were discarded during the process. Clearly, separating wheat from chaff is a challenging problem.

In this study, we suggest an operational approach based on the stabilization of the entropy ratio to guide the cleaning procedures. Both the simulation approach and the analysis of the real data set pointed to an avalue of 5 in the denoising step, which was also the optimal value selected in Elbrecht et al. (2018b). Whether this value can be taken as a general rule of thumb or not will require analyses of more data sets. For the filtering step, our method indicated 20 reads as the optimal threshold. This is a parameter that will likely vary between studies and should be optimized for each particular data set.

Some authors proposed that denoising should be performed before clustering to identify genuine sequence variants, using different procedures, such as the UNOISE2 algorithm that we have adapted here (Edgar, 2016), the MED (minimum entropy decomposition; Eren et al., 2015) procedure, or the DADA2 algorithm (divisive amplicon denoising algorithm; Callahan et al., 2016). It has also been suggested that sequence variants should replace MOTUs to capture relevant biological variation (Callahan et al., 2017; Edgar, 2016). This suggestion may be adequate in prokaryotes, where strains of the same species can have different characteristics (e.g., pathogenicity). However, for eukaryotes, and particularly metazoans, given the high

amount of intraspecies information contained in the data sets, we think that it is more advisable to define meaningful MOTUs and perform denoising procedures within them, in order to obtain a "clean" data set and be able to use the intraMOTU sequence variability to make phylogeographic and population genetics inference. Clearly, our procedure is applicable only to coding sequences, which excludes much work done on protists based on ribosomal DNA. However, the growing number of metabarcoding studies using COI sequence data, together with the steady development of the BOLD database, makes us confident that many metabarcoding data sets of enormous potential for metaphylogeographic inference will become available in the near future.

We found a couple of instances of previous studies that have analysed COI structure in species recovered in our MOTU data set and in nearby localities. For *Paracentrotus lividus*, there were phylogeographic studies of the Atlanto-Mediterranean area using COI (Duran et al., 2004), 16S (Calderón et al., 2008), and the nuclear ANT intron (Calderón et al., 2008). In all cases, a low, but significant, signal corresponding to the separation between Atlantic and Mediterranean was found. Our COI results were in agreement with those of Duran et al. (2004) for the localities that could be compared. We detected a somewhat higher number of haplotypes (11 in the previous work, 15 in our study) and the most common haplotypes were shared. The shape of the network was also similar. We want to emphasize that, as far as we could detect, not a single sea urchin of this species was present in our samples, so we obtained a similar level of haplotype diversity with community DNA than in a study specifically devoted to collect sea urchin specimens. For *Ophiothrix fragilis*, we also found a higher haplotype diversity (37 haplotypes) than in comparable localities in the work of Pérez-Portela et al. (2013, 29 haplotypes). We identified five haplotypes that were shared in the two studies, including the commonest one in both data sets, and the networks again had similar structure. Of note here is that we could expand the distribution range of some of the haplotypes. Our AMOVA results for these two instances were equivalent to previous results for the only component that was analyzed in both studies (the between-seas differentiation). Thus, Duran et al. (2004) found a significant ($P < 0.05$)

between-basin differentiation in *Paracentrotus lividus*, while Pérez-Portela et al. (2013) did not find any significant genetic variability between Atlantic and Mediterranean for Lineage II of *Ophiothrix fragilis* ($P = 0.790$). This is consistent with our metabarcoding-derived AMOVAs ($P < 0.001$ and $P = 0.729$, respectively). The two species are of remarkable ecological importance, *Paracentrotus lividus* is an engineer species able to modify the littoral landscape through its browsing activity (Palacín et al., 1998; Wangensteen et al., 2011), and is also a commercially exploited species (Barnes and Crook, 2001). The different lineages of *Ophiothrix fragilis* are highly abundant components of the littoral communities and can form dense beds, with an important role in clearing particulate matter with their filtering activities (Davoult, 1989; Davoult and Gounin, 1995). For both species, therefore, an accurate assessment of the genetic relationships across the different basins is of utmost importance for conservation and management purposes.

We have used an already collected data set, which can mimic the situation that many a posteriori studies can encounter. However, future metabarcoding studies can be planned taking into consideration the potential application for intraspecies analyses as well. For instance, PCR replicates for each sample can be of tremendous advantage to eliminate noise in the first steps. Increasing ecological replication can also be of great value for metaphylogeographic studies. We strongly advocate that published metabarcoding studies include in their data sets the information about which sequences are grouped into each MOTU with their sample distribution. This information is not commonly provided, and is necessary to make these studies amenable for intraspecies and metaphylogeographic analyses.

Metabarcoding now occupies a well-deserved prominent place among the methods for assessing communitylevel diversity (Adamowicz et al., 2019; Kelly et al., 2014b). We have shown that it can be also an important source for species-level genetic diversity information for a wide assemblage of taxonomic groups. The mining of metabarcoding data for intraspecies information opens up a vast field with both basic and applied implications

(Adams et al., 2019). Among the latter, the possibility of effectively basing conservation efforts on multispecies genetic metrics to preserve community-level evolutionary patterns (Nielsen et al., 2017). It will also open the phylogeography field, nowadays restricted almost exclusively to macroorganisms, to the myriad of meio- and micro-eukaryotes that make up most of the diversity present in natural communities.

Another related field is the assessment of connectivity between populations. This is important for endangered species, invasive species, protected areas design, and management in general. For instance, in the marine environment, differences in larval dispersal have often been suggested as responsible for determining population genetic structure, but other factors, such as variation in divergence times and changes in effective population sizes, must be taken into account (Hart and Marko, 2010). A powerful test for these contrasting assumptions is to compare phylogeographic patterns among species that concur or differ in larval type. Metaphylogeography can provide such comparative data. For instance, in our study we have found that metazoans in general have more between-replicate variability than other groups, and within metazoans the between community and between-replicate components of genetic variation can be significantly different between phyla.

In conclusion, our study shows the feasibility of mining metabarcoding data sets for the analysis of intraspecies genetic diversity using objective parameters for denoising and filtering spurious sequences. We cannot at present advice a set pipeline to do this, as procedures should be customized for the particulars (e.g., replication level, number of habitats, number of localities) of each study data set. With this article, we hope to stir further discussion and developments in this field. The metaphylogeography application should be borne in mind to guide the planning and reporting of metabarcoding studies to ease the recovery of this, so far unexplored, vast amount of information.

Chapter 6

To denoise or to cluster, that is not the question: optimizing pipelines for COI metabarcoding and metaphylogeography

6.1 Abstract

Background. The recent blooming of metabarcoding applications to biodiversity studies comes with some relevant methodological debates. One such issue concerns the treatment of reads by denoising or by clustering methods, which have been wrongly presented as alternatives. It has also been suggested that denoised sequence variants should replace clusters as the basic unit of metabarcoding analyses, missing the fact that sequence clusters are a proxy for species-level entities, the basic unit in biodiversity studies. We argue here that methods developed and tested for ribosomal markers have been uncritically applied to highly variable markers such as cytochrome oxidase I (COI) without conceptual or operational (e.g., parameter setting) adjustment. COI has a naturally high intraspecies variability that should be assessed and reported, as it is a source of highly valuable information. We contend that denoising and clustering are not alternatives. Rather, they are complementary and both should be used together in COI metabarcoding pipelines.

Results: Using a COI dataset from benthic marine communities, we compared two denoising procedures (based on the UNOISE3 and the DADA2 algorithms), set suitable parameters for denoising and clustering, and applied these steps in different orders. Our results indicated that the UNOISE3

algorithm preserved a higher intra-cluster variability. We introduce the program DnoisE to implement the UNOISE3 algorithm taking into account the natural variability (measured as entropy) of each codon position in protein-coding genes. This correction increased the number of sequences retained by 88%. The order of the steps (denoising and clustering) had little influence on the final outcome.

Conclusions: We highlight the need for combining denoising and clustering, with adequate choice of stringency parameters, in COI metabarcoding. We present a program that uses the coding properties of this marker to improve the denoising step. We recommend researchers to report their results in terms of both denoised sequences (a proxy for haplotypes) and clusters formed (a proxy for species), and to avoid collapsing the sequences of the latter into a single representative. This will allow studies at the cluster (ideally equating species-level diversity) and at the intra-cluster level, and will ease additivity and comparability between studies.

6.2 Background

The field of eukaryotic metabarcoding is witnessing an exponential growth, both in the number of communities and substrates studied and the applications reported (reviewed in Aylagas et al., 2018; Bani et al., 2020; Compson et al., 2020; Deiner et al., 2017). In parallel, technical and conceptual issues are being discussed (e.g., Mathieu et al., 2020; Rodriguez-Ezpeleta et al., 2021) and new methods and pipelines generated. In some cases, however, new practices are established after a paper reporting a technique is published and followed uncritically, sometimes pushing its application outside the context in which it was first developed.

A recently debated matter concerns the treatment of reads by denoising procedures or by clustering techniques (Porter and Hajibabaei, 2020). Both methods are often presented as alternative approaches to the same process (e.g., Forster et al., 2019; Giebner et al., 2020; Macheriotou et al., 2019; O'Rourke et al., 2020; Porter and Hajibabaei, 2020). However, both are philosophically and analytically different (Turon et al., 2020). While

denoising strives to detect erroneous sequences and to merge them with the correct "mother" sequence, clustering tries to combine a set of sequences (without regard to whether they contain or not errors) into meaningful biological entities, ideally approaching the species level, called OTUs or MOTUs (for Molecular Operational Taxonomic Units). Usually only one representative sequence from each MOTU is kept (but note that this is only common practice, not a necessary characteristic of the method). Thus, while both procedures result in a reduced dataset and in error correction (by merging reads of erroneous sequences with the correct one or by combining them with the other reads in the MOTU), they are not equivalent. More importantly, they are not incompatible at all and can (and should) be used together.

A recent paper (Callahan et al., 2017) proposes that denoised sequences should replace MOTUs as the unit of metabarcoding analyses. We contend that it may be so for ribosomal DNA datasets such as the one used in that paper, but this notion has gained momentum also in other fields of metabarcoding for which it is not adequate. In particular, when it comes to highly variable markers such as COI. This proposal misses the fact that sequence clusters are a proxy for species-level entities, the basic unit in eukaryotic biodiversity studies. The 3' half (also called Leray fragment) of the standard barcode fragment of COI (Folmer fragment) is becoming a popular choice for metabarcoding studies addressed at metazoans or at eukaryotic communities at large (Andújar et al., 2018), reaching now 28% of all metabarcoding studies (van der Loos and Nijland, 2021). Metabarcoding stems from studies of microbes where 16S rRNA is the gene of choice, and the concept was then applied to analyses of the 18S rRNA gene of eukaryotes. With the recent rise of COI applications in metabarcoding, programs and techniques developed for rDNA are sometimes applied to COI without reanalysis and with no parameter adjusting given the highly contrasting levels of variation of these markers.

The idea that denoising should be used instead of clustering has been followed by some (e.g., Holman et al., 2021; Pearman et al., 2020; Steyaert et al., 2020; Tapolczai et al., 2019; Zamora-Terol et al., 2020), while other

authors have combined the two approaches (e.g., Brandt et al., 2021; Laroche et al., 2020; Nguyen et al., 2020). Indeed, denoising has the advantages of reducing the dataset and to ease pooling or comparing studies, which is necessary in long term biomonitoring applications. However, with COI there is a wealth of intraspecific information that is missed if only denoising is applied (Zizka et al., 2020). COI has been a prime marker of phylogeographic studies to date (Avise, 2009; Emerson et al., 2011), and these studies can be extended to metabarcoding datasets by mining the distribution of haplotypes within MOTUs (metaphylogeography, Turon et al., 2020). The latter authors suggested to perform clustering first, and that denoising should be done within MOTUs to provide the right context of sequence variation and abundance skew. They also advised to perform a final abundance filtering step. In other studies, denoising is performed first, followed by clustering and refining steps (e.g., Laroche et al., 2020; Nguyen et al., 2020).

There are several methods for denoising (reviewed in Peng and Dorman, 2020) and for clustering (reviewed in Kopylova et al., 2016). We will use two of the most popular denoising techniques, based on the DADA2 algorithm (Divisive Amplicon Denoising Algorithm, Callahan et al., 2016) and the UNOISE3 algorithm (Edgar, 2016). The results of the former are called Amplicon Sequence Variants (ASVs) and those of the latter ZOTUs (zero-radius OTUs). In practice, the terminology is mixed and ASV, ZOTU, ESV (Exact Sequence Variant), sOTU (sub-OTU) or ISU (Individual Sequence Variant), among others, are used more or less interchangeably. For simplicity, as all of them are equivalent, we will use henceforth the term ESV. Clustering, on the other hand, can be performed using similarity thresholds (e.g., Edgar, 2013; Rognes et al., 2016), Bayesian Methods (CROP, Hao et al., 2011), or methods based on single-linkage-clustering (SWARM, Mahé et al., 2015), among others. We will focus on denovo clustering methods (i.e., independent of a reference database), while denoising is always denovo by its very nature (Callahan et al., 2017). We will here use SWARM as our choice of clustering program due to its good performance compared to other methods (Kopylova et al., 2016). It is noteworthy that all these programs were originally developed and tested on ribosomal DNA datasets. When applied to other markers, often no indication of parameter setting is given (i.e., omega_A for

DADA2, α for UNOISE3, d for SWARM), suggesting that default parameter values are used uncritically.

In this article, we aim to use a COI metabarcoding dataset of benthic littoral communities to (1) set the optimal parameters of the denoising and clustering programs for COI markers, (2) compare results of the DADA2 algorithm with the UNOISE3 algorithm, (3) compare the results of performing only denoising, only clustering, or combining denoising with clustering in different orders, and (4), suggest and test improvements in the preferred denoising algorithm to take into account the fact that COI is a coding gene. We implement these modifications in the new program DnoisE. Our aims are to provide guidelines for using these key bioinformatic steps in COI metabarcoding and metaphylogeography. The conceptual framework of our approach is sketched in Figure 6.1.

6.3 Methods

6.3.1 The dataset

We used as a case study an unpublished dataset of COI sequences obtained from benthic communities in 12 locations of the Iberian Mediterranean. Information on the sampling and sample processing is given in Appendix C. Sequences were obtained in a full run of an Illumina MiSeq (2 $\times$ 250 bp paired-end reads).

6.3.2 Bioinformatic analyses

The initial steps of the bioinformatic pipeline followed Turon et al. (2020) and were based on the OBItools package (Boyer et al., 2016). Reads were paired and quality filtered, demultiplexed, and dereplicated. A strict length filter of 313 bp was used. We also eliminated sequences with only one read. Chimera detection was performed on the whole dereplicated dataset with uchime3_denovo as embedded in unoise3 (USEARCH 32-bit free version, Edgar, 2010). We used minsize = 2 to include all sequences. Those identified as chimeras were recovered from the –tabbedout file and eliminated from the dataset. Sequences with small offsets (misaligned), identified as shifted

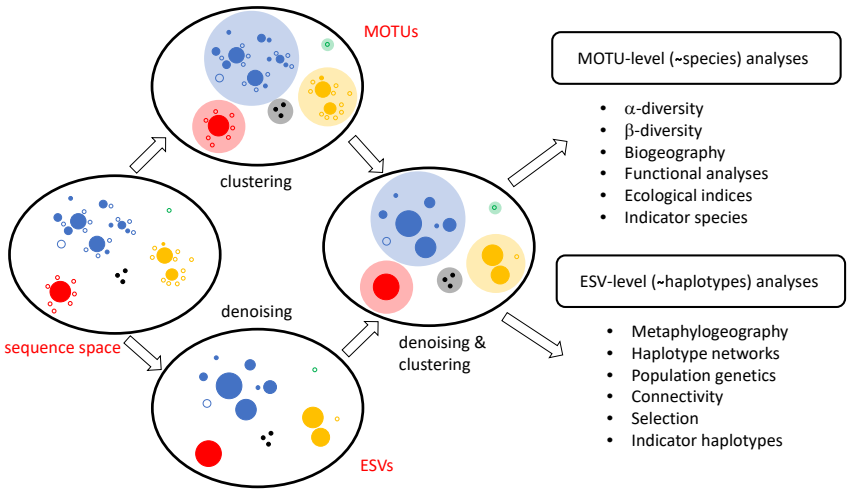


Figure 6.1: Conceptual overview of the denoising and clustering processes. The oval on the left sketches a fragment of the sequence space with four biological species plus an artefact divergent sequence (denoted by colours). Correct sequences are indicated by filled circles and artefacts by empty circles, with indication of abundance (circle size). Denoising results in the detection of putatively correct sequences to which the reads of putatively incorrect sequences are merged (leading to a reduced dataset). The outcome of denoising should ideally approach the true haplotype composition of the samples. Clustering generates MOTUs without regard as to whether the grouped sequences are erroneous or not. This is usually accompanied by read pooling and keeping only one representative sequence per MOTU (leading to a reduced dataset). The outcome of clustering should ideally approach the species composition of the samples. Combining both processes results in a dataset that is reduced in size, comparable across studies, and amenable to analyses at the MOTU (species) and ESV (haplotype) levels. Note that errors likely persist in the final dataset both as artefact MOTUs and artefact ESVs within MOTUs, and carefully designed filters should be used to minimize them (abundance filtering, chimera filtering, numts removal).

in the output, were likewise deleted. The working dataset thus comprised well-aligned, chimera-free, unique sequences which had appeared with at least two reads in the samples.

Note that for this technical study we didn't consider the sample distribution of the reads. A complete biogeographic study of the samples is ongoing and will be published elsewhere. For the present analysis, for each unique sequence only the actual DNA sequence and the total number of reads were retained.

6.3.3 The denoisers: UNOISE3 and DADA2

Comparing denoising algorithms is challenging because each method comes with a different software suite with embedded features and recommendations (Peng and Dorman, 2020). For instance, uchime3_denovo is embedded in the unoise3 command as implemented in USEARCH, while a chimera removal procedure (removeBimeraDenovo) is an optional feature in the DADA2 pipeline. Furthermore, while UNOISE3 uses paired reads, DADA2 recommends denoising forward and reverse reads separately, and then performing a merging step. We have tried to isolate the algorithms from their pipelines for comparability. This was done by generating a Python script (Antich, 2020) that implements the algorithm described in Edgar (2016) and by using DADA2 from its R package v. 1.14.1 and not as embedded into the qiime2 pipeline (Bolyen et al., 2019).

For UNOISE3, our program (henceforth DnoisE) was compared on the working dataset described above with command unoise3 in USEARCH with minsize = 2, alpha = 5 and without the otutab step. That is, we recovered the ESV composition and abundance with an R script directly from the output of unoise3 (using the output files –tabbedout and –ampout), without a posterior re-assignment of sequences to ESVs via otutab. This step was not necessary as all sequences were included in the ESV calculations. The results of DnoisE and unoise3 were $> 99.99\%$ identical in ESVs recovered and reads assigned to them, so we continued to use our script for performing the comparisons and for further improvements of the algorithm (see below).

The recommended approach for DADA2 is to denoise separately the forward and reverse reads of each sequence. This complicates the technical comparison, as all initial filtering steps cannot be equally performed (e.g., we won't know if there is just one read of a particular sequence, or if the merged pair will be discarded for low quality of the assembly or for unsuitable final length) and thus we cannot have two identical starting datasets. More importantly, we cannot use this procedure when we test the effects of denoising at later steps (i.e., after clustering), so we would be unable to compare the denoisers at this level. Thus, for our comparative analysis we need to use DADA2 on paired reads. According to Callahan et al.

(2016), this can result in a loss of accuracy, but this point has never been tested to our knowledge. We addressed this issue by comparing denoising before and after pairing on half of the reads in the final dataset. After this analysis, we decided to continue our comparison of DADA2 and UNOISE3 on paired reads.

Additionally, denoising before pairing is not optimal if a PCR-free library preparation protocol is used, as in our case, because half of the reads are in one direction and the other half are in the opposite direction (hence the use of half of the reads in the above comparison). Forward and reverse reads can of course be recombined to generate new files with all reads in the same direction, but the quality of the reads with original forward and reverse orentation is different. Alternatively, two rounds of DADA2 (one per orientation) must be performed and combined at later steps.

To run DADA2 on paired reads, we entered them in the program as if they were the forward reads and did not use a merging step after denoising. In all DADA2 runs we did not perform the recommended chimera removal procedure as the input sequences were already chimera-free according to uchime3_denovo. Note that, when denoising was done after clustering, we used error rates calculated for the whole dataset, and not for each MOTU separately (most of them do not have enough number of sequences for a reliable estimation of error rates).

UNOISE3 relies heavily on the stringency parameter α, which weights the distance j given their abundance values and the inferred error rates. If the observed value is higher than omega_A, then sequence i is considered an error of sequence j. Omega_A is by default set to a very low value (10^{-40}), but no study has analysed the impact of changing this parameter for COI datasets. To our knowledge, only Tsuji and Shibata (2020), based on a comparison of 3 values, concluded that the default value of omega_A was adequate for a marker based on the control region of the mitochondrial DNA.

6.3.4 The clustering algorithm

Our preferred clustering method is SWARM v3 (Mahé et al., 2014), as it is not based on a fixed distance threshold and is independent of input order. It is a very fast procedure that relies on a single-linkage method with a clustering distance (d), followed by a topological refining of the clusters using abundance structures to divide MOTUs. As we were interested in keeping all sequences within MOTUs, and not just a representative sequence, we mined the SWARM output with an R script to generate MOTU files, each with its sequence composition and abundance.

The crucial parameter in this approach is d, the clustering distance threshold for the initial phase. The default value is 1 (that is, amplicons separated by more than one difference will not be clustered together), and this value has been tested in ribosomal DNA. However, Mahé et al. (2014) pointed out that higher d values can be necessary for fast evolving markers (such as COI) and advised to analyse a range of d to identify the best fitting parameter (i.e., avoiding over- or under-clustering) for a particular dataset or scientific question. A d value of 13 (thus, allowing 13 differences over ca. 313 bp to make a connection) has been recently used for the Leray fragment of COI (e.g., Antich et al., 2020; Atienza et al., 2020; Bakker et al., 2019; Garcés-Pastor et al., 2019; Siegenthaler et al., 2019b), but a formal study of its adequacy has not been published yet.

6.3.5 Setting the right parameters

With our dataset, we assessed the best-fitting parameters for UNOISE3, DADA2 and SWARM as applied to COI data. For the first two, we used changes in diversity values per codon position (measured as entropy, Schmitt and Herzel, 1997), as calculated with the R package *entropy* (Hausser and Strimmer, 2009). Coding sequences have properties that can be used in denoising procedures (Tsuji and Shibata, 2020; Turon et al., 2020). They have naturally a high amount of variation concentrated in the third position of the codons, while errors at any step of the metabarcoding pipeline would be randomly distributed across codon positions. Thus, examining the change in entropy values according to codon position can guide the choice of the

best cleaning parameters. Turon et al. (2020) suggested to use the entropy ratio (E_r) between position 2 of the codons (least variable) and position 3 (most variable). In a simulation study these authors showed that E_r decreased as more stringent denoising was applied until reaching a plateau, which was taken as the indication that the right parameter value had been reached.

Using the E_r to set cut-points, we re-assessed the adequate value of α in UNOISE3 testing the interval of $\alpha = 1$ to 10. With the same procedure, we tested DADA2 for values of omega_A between $10^{-0.05}$ (ca. 0.9) and $10^{-0.05}$. For SWARM, we compared the output of SWARM with a range of values of d from 1 to 30 applied to our dataset (prior to denoising). We monitored the number of MOTUs generated and the mean intra- and inter-MOTU distances to find the best-performing value of d for our fragment.

6.3.6 The impact of the steps and their order

With the selected optimal parameters for each method, we combined the two denoising procedures and the clustering step in different orders. We therefore combined denoising (Du for UNOISE3 algorithm implemented in DnoisE, Da for DADA2) and clustering with SWARM (S) and generated and compared datasets of ESVs and MOTUs as follows (for instance, Da_S means that the dataset was first denoised with DADA2, then clustered with SWARM):

ESVs: Du, Da

MOTUs: Du_S, Da_S, S_Du, S_Da

For comparison of datasets, we used Venn diagrams and an average match index of the form

$$Match\ Index(A, B) = (N_{match\_A}/N_A + N_{match\_B}/N_B)/2$$

where N_{match_A} is the number of a particular attribute in dataset A that is shared with dataset B, and N_A is the total number of that attribute in dataset A. The same for N_{match_B} and N_B. The matches can be the number of ESVs shared, the number of MOTUs shared, the number of ESVs in the

shared MOTUs, or the number of reads in the shared ESVs or MOTUs, depending on the comparison.

6.3.7 Improving the denoising algorithm

The preferred denoising algorithm (UNOISE3, see Results) has been further modified in two ways. Let i be a potential error sequence derived from sequence j. The UNOISE3 procedure is based on two parameters: the number of sequence differences between i and j (d, as measured by the Levenshtein distance) and the abundance skew (β, abundance i/abundance j) between them. These parameters are related by the simple formula (Edgar, 2016):

$$\beta(d) = 1/2^{\alpha d+1}$$

where $\beta(d)$ is the threshold abundance skew allowed between two sequences separated by distance d so that below it the less abundant would be merged with the more abundant, and α is the stringency parameter. Thus, presumably incorrect "daughter" sequences are merged with the correct "mother" sequences if the number of sequence differences (d) is small and the abundance of the incorrect sequence with respect to the correct one (abundance skew) is low. The higher the number of differences, the lower the skew should be for the sequences to be merged.

For COI, however, the fact that it is a coding gene is a fundamental difference with respect to ribosomal genes. In a coding fragment, the amount of variability is substantially different among codon positions. This is not considered in the UNOISE3 formulation (nor in DADA2 or other denoising programs that we knew of, for that matter). We suggest to incorporate this information in DnoisE by differentially weighting the d values according to whether the change occurs in the first, second, or third codon position. Note that our sequences are all aligned and without indels, which makes this weighting scheme straightforward. The differences in variability can be quantified as differences in entropy values (Schmitt and Herzel, 1997); position 3 of the codons has the highest entropy, followed by position 1 and position 2. In other words, two sequences separated by n differences in third positions are more likely to be naturally-occurring sequences than

if the n differences happen to occur in second positions, because position 3 is naturally more variable. To weight the value of d, we first record the number of differences in each of the three codon positions ($d(1)$ to $d(3)$), we then correct the d value using the formula

$$d_{corr} = \sum_{i=1}^{3} d(i) \times entropy(i) \times 3/(entropy(1) + entropy(2) + entropy(3))$$

where i is the position in the codon, and d_{corr} is the corrected distance that will be used in the UNOISE3 formula instead of d.

With this formula, two sequences separated by just one difference in each codon position will continue to have a d of 3, but a change in a high entropy position (3) will translate in a higher d than the same change in a low entropy position (2), thus the program will tend to keep the former and to merge the later. The entropy of the three positions of the codons for the weighting was obtained from the original dataset prior to any denoising, thus entropy(1) = 0.473, entropy(2) = 0.227, and entropy(3) = 1.021. Note that d(i) is based on the number of differences occurring at each codon position. The Levenshtein distance used in the non-corrected d measures is not adequate for this purpose, as it cannot keep track of codon positions. However, for sequences of equal length, aligned, and without indels, as in our case, the number of differences is in practice equivalent to the Levenshtein distance.

The present algorithm of UNOISE3 gives precedence to the abundance skew over the number of differences (d) because sequences are considered in order of decreasing abundance. Thus, a very abundant sequence will form a centroid that can "capture" a rare one even if d is relatively high. Other, somewhat less abundant, sequences can be more similar (less d) to the rare sequence and can fulfil the conditions to capture it, but this will never happen as the rare sequence will be incorporated to the first centroid and will become unavailable for further comparisons. In our modification, DnoisE does not automatically join sequences to the first centroid that fits the condition. Rather, for each sequence the potential "mothers" are stored (with their abundance skew and d) and the sequences are left in the dataset. After the round of comparisons is completed, for each daughter sequence

we can choose, among the potential mothers, the one whose abundance skew is lower (precedence to abundance skew, corresponding to the usual UNOISE3 procedure), the mother with the lowest distance (precedence to d), or the one for which the ratio (abundance skew/max abundance skew for the observed d, $\beta(d)$) is lower, thus combining the two criteria.

We compared in our dataset the results of the different formulations of DnoisE: precedence to abundance skew, precedence to distance, combined precedence, and correcting distances according to codon position of the differences. A beta version of DnoisE is available from github (Antich, 2020).

6.3.8 Benchmarking

Ground truthing is a difficult task in metabarcoding studies. Constructing mock communities is the most common method. However, mock communities, even the largest ones, are orders of magnitude simpler than complex biological communities. Thus, some technical aspects cannot be tested accurately. For instance, metabarcoding results of mock communities in general lack true sequences at very rare abundances (the most problematic ones). For complex communities, we need to rely on metrics that can evaluate the fit of denoising and filtering procedures. The coding properties of COI can help design useful parameters, such as the entropy ratio mentioned above. Another possible metric stems from the evaluation of the prevalence of incorrect ESVs (defined by having indels or stop codons) across denoising and filtering procedures (Andújar et al., 2021).

In this work, we have performed two benchmarking procedures that rely on taxonomic assignment of the MOTUs. This assignment was done using the ecotag procedure in OBItools against the db-COI_MBPK database (Wangensteen et al., 2018a), containing 188,929 eukaryote COI reference sequences (available at Wangensteen, 2020). Ecotag assigns a sequence to the common ancestor of the candidate sequences selected in the database, using the NCBI taxonomy tree. This results in differing taxonomic rank of the assignments depending on the density of the reference database for a given taxonomic group.

First, we checked the performance of the entropy correction of DnoisE by examining the percent of incorrect to total ESVs. To this end, we retained only the MOTUs assigned to metazoans and, following Turon et al. (2020), examined the presence of stop codons and changes in the 5 aminoacids present in the fragment amplified that are conserved among metazoans (Pentinsaari et al., 2016). To be on the conservative side, for a given MOTUs we evaluated the different genetic codes and selected the ones that produced the smaller number of stop codons. The five aminoacids were then checked using these codes and the minimal number of "wrong" aminoacids was recorded. The R package Biostrings (Pagès et al., 2008) was used for the translations. The ESVs featuring stop codons and/or aminoacid changes in the five conserved positions were labelled as erroneous. The rationale is that a suitable denoising procedure would reduce the ratio of error vs total ESVs.

Second, we performed a taxonomic benchmarking. As MOTUs should ideally reflect species-level entities, we selected those sequences assigned at the species level as a benchmark for the MOTU datasets. We also enforced a 97% minimal best identity with the reference sequence. We traced these sequences in the output files of our procedures and classified the MOTUs containing them into three categories (following the terminology in Forster et al., 2019): closed MOTUs, when they contain all sequences assigned to a species and only those; open MOTUs, when they contain some, but not all, sequences assigned to one species and none from other species, and hybrid MOTUs. The latter included MOTUs with sequences assigned to more than one species, or MOTUs with a combination of sequences assigned to one species and sequences not assigned (i.e., they don't have species-level assignment, or they do with less than 97% similarity).

This analysis was intended as a tool for comparative purposes, to benchmark the ability of the different MOTU sets generated to recover species-level entities. In other words, which procedure retains more ESVs with species-level assignment and places them in closed (as opposed to open or hybrid) MOTUs.

6.4 Results

6.4.1 The dataset

After pairing, quality filters, and retaining only 313 bp-long reads, we had a dataset of 16,325,751 reads that were dereplicated into 3,507,560 unique sequences. After deleting singletons (sequences with one read), we kept 423,164 sequences (totalling 10,305,911 reads). Of these sequences, 92,630 were identified as chimeras and 152 as misaligned sequences and eliminated. Our final dataset for the study, therefore, comprised 330,382 sequences and 9,718,827 reads (the original and the refined datasets were deposited in Mendeley Data, Antich et al., 2021b).

For testing the performance of DADA2 on unpaired and paired reads on a coherent dataset, we selected the reads that were in the forward direction, that is, the forward primer was in the forward read (R1). As expected, they comprised ca. half of the reads (4,892,084). For these reads we compared the output of applying DADA2 before and after pairing, as detailed in Appendix C. The results were similar, with most reads placed in the same ESVs in both datasets, albeit 21% more low-abundance ESVs were retained using the paired reads. Henceforth we will use DADA2 on paired sequences, as this was necessary to perform our comparisons.

6.4.2 Setting the right parameters

We used the change in entropy ratio (E_r) of the retained sequences of the global dataset (330,382 sequences and 9,718,827 reads) for selecting the best performing α–value in UNOISE3 and the best omega_A in DADA2 across a range of values. We also assessed the number of ESVs resulting from the procedures.

For UNOISE3 as implemented in our DnoisE script, the E_r diminished sharply for α–values of 10 to 7, and more smoothly afterwards (Fig. 6.2a). The number of ESVs detected likewise decreased sharply with lower α–values, but tended to level off at $\alpha = 5$ (Fig. 6.2a). The value of 5 seems a good compromise between minimizing the E_r and keeping the maximum number

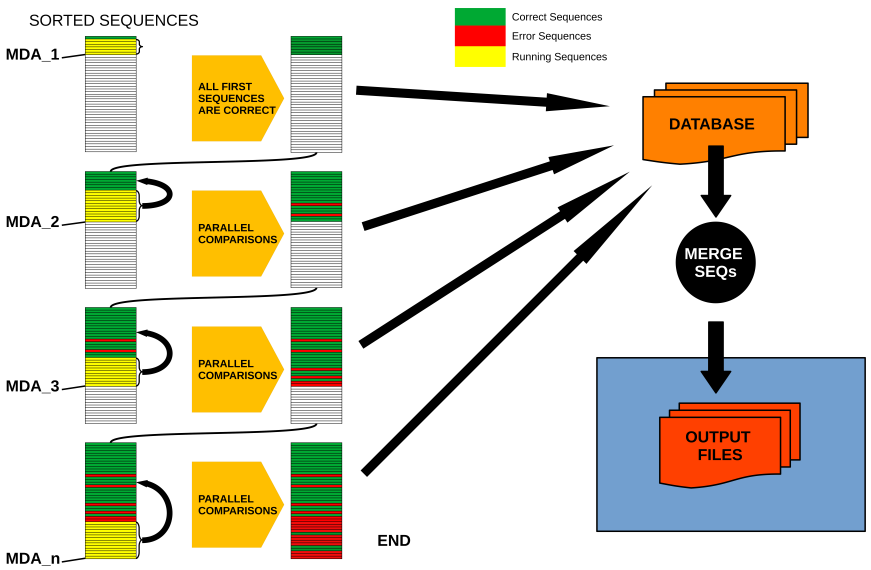


Figure 6.2: Values of the Entropy ratio (E_r) of the set of ESVs obtained with the UNOISE3 algorithm at decreasing values of α (**a**), and of those obtained with the DADA2 algorithm at decreasing values of omega_A (**b**). Arrows point at the selected value for each parameter. Horizontal blue line in (**b**) represents the E_r value reached in (**a**) at $\alpha = 5$, horizontal red line marks the number of ESVs detected in (**a**) at $\alpha = 5$.

of putatively correct sequences.

For the DADA2 algorithm we tested a wide range of omega_A from $10^{-0,05}$ to 10^{-90} (we set parameter omega_C to 0 in all tests, so all erroneous sequences were corrected). The results showed that, even at the highest value ($10^{-0,05}$, or ca. 0.9 p-value, thus accepting as new partitions a high number of sequences), there was a substantial drop in number of sequences (ca. 75% reduction) and in E_r with respect to the original dataset (Fig. 6.2b). Both variables remained relatively flat with a slight decrease between omega_A 10^{-2} and 10^{-15}, becoming stable again afterwards (Fig. 6.2b).

The number of ESVs retained was considerably lower than for UNOISE3. In fact, the number obtained at $\alpha = 5$ by the latter (60,198 ESVs) was approximately reached at omega_A $= 10^{-5}$ (58,191 ESVs). On the other hand, the entropy value obtained at $\alpha = 5$ in UNOISE3 (0.2182) was not reached until omega_A $= 10^{-60}$. As a compromise, we will use in this study

the default value of the dada function (10^{-40}), while acknowledging that the behaviour of DADA2 with changes in omega_A for the parameters analysed was unexpected and deserves further research.

For the clustering algorithm SWARM v.2, we monitored the outcome of changing the d parameter between 1 and 30. For each value, we tracked the number of clusters formed (separately for all MOTUs and for those with 2 or more sequences), as well as the mean intra-MOTU and the mean inter-MOTU genetic distances (considering only the most abundant sequence per MOTU for the lattE_r). The goal was to find the value that maximizes the intra-MOTU variability while keeping a sharp difference between both values (equivalent to the barcode gap).

The total number of MOTUs decreased sharply from 38,560 ($d = 1$) to around 19,000 with a plateau from $d = 9$ to $d = 13$, and then decreased again (Fig. 6.3a). If we only consider the MOTUs with 2 or more sequences, the overall pattern is similar, albeit the curve is much less steep. The numbers decreased from 8,684 for $d = 1$ to 6,755 at $d = 12$ and 13, and decreasing again at higher values (Fig. 6.3a).

Inter-MOTU distances had a similar distribution with all values of the parameter d, albeit with a small shoulder at distances of 10-20 differences with $d = 1$ (selected examples in Fig. 6.3b). Intra-MOTU distances, on the other hand, became more spread with higher values of d as expected. Values from 9 to 13 showed a similar distribution of number of differences, but for d values higher than 14, intra-MOTU distances started to overlap with the inter-MOTU distribution (Fig. 6.3b). The value of $d = 13$ seems, therefore, to be the best choice to avoid losing too much MOTU variability (both in terms of number of MOTUs and intra-MOTU variation), and at the same time keeping intra- and inter-MOTU distances well separated. The mean intra-MOTU distance in our dataset at $d = 13$ was 9.10 (equivalent to 97.09% identity), and the mean inter-MOTU distance was 108.78 (65.25% identity).

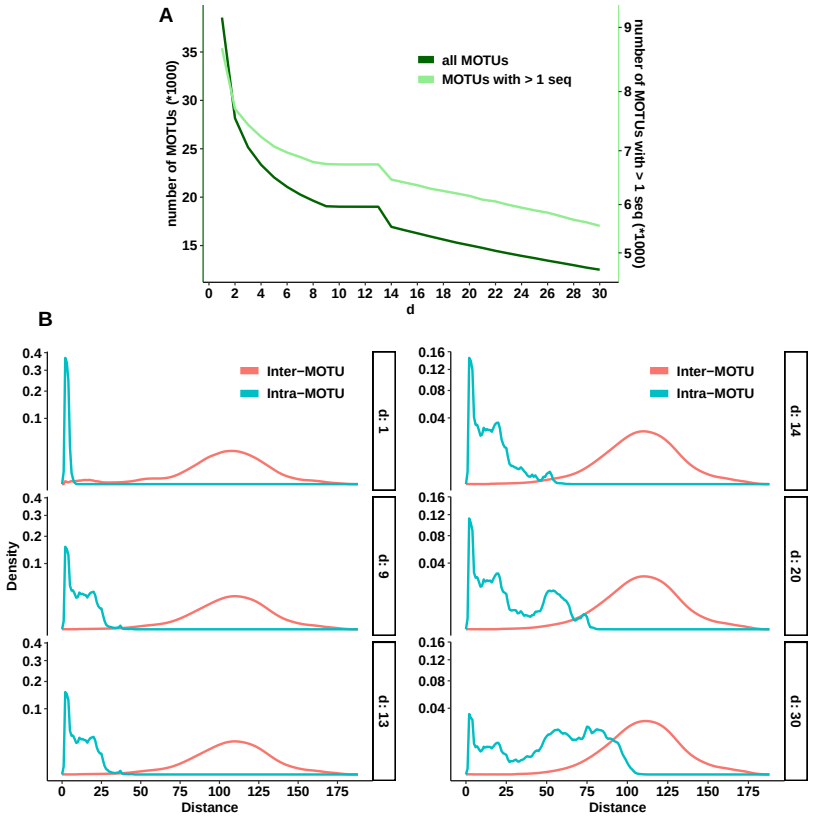


Figure 6.3: **a** Number of MOTUs obtained at different values of d using SWARM. Total number of MOTUs (dark green) and of MOTUs with two or more sequences (light green) are represented (note different Y-axes). **b** Density plots (note quadratic scale) showing the distribution of number of differences between different clusters (inter-MOTU, red) and sequences within clusters (intra-MOTU, blue) obtained by SWARM for selected values of the parameter d (1, 9, 13, 14, 20 and 30).

6.4.3 The impact of the steps and their order

Table 6.1 shows the main characteristics of the original and the generated datasets, as well as the datasets obtained by modifying the UNOISE3 algorithm (see below). All datasets are available from Mendeley Data (Antich et al., 2021b).

We first compared the outcomes of denoising the original reads with UNOISE3 and DADA2 (Du vs Da), with the stringency parameters set as above. The error rates of the different substitution types as a function of

Table 6.1: Main characteristics of the original and the generated datasets.

	n. ESVs(*)	n. MOTUs	Single-ESV MOTUs	ESVs/MOTU(*)	reads/MOTU
Original	330,382	–	–	–	–
Du(**)	60,198	–	–	–	–
Da	32,798	–	–	–	–
Du_e(***)	113,133	–	–	–	–
S	330,382	19,012	12,257	17.378	511.194
Du_S	60,198	19,058	12,471	3.159	509.961
S_Du	75,069	19,012	12,433	3.949	511.194
Da_S	32,798	19,167	15,565	1.711	507.060
S_Da	35,376	19,012	15,198	1.861	511.194
Du_d_S	60,198	19,058	12,471	3.159	509.960
Du_c_S	60,198	19,058	12,471	3.159	509.960
Du_e_S	113,133	19,016	12,365	5.949	511.087
Du_e_d_S	113,133	19,016	12,365	5.949	511.087
Du_e_c_S	113,133	19,016	12,365	5.949	511.087

All datasets had 9,718,827 reads. 1-ESV MOTUs refer to the number of MOTUs with just one ESV. Codes of the datasets: Du, denoised with UNOISE3 algorithm (unless otherwise stated, it refers to the original formulation giving precedence to abundance ratio); Da, denoised with DADA2 algorithm; S, clustered with SWARM algorithm; Du_S, denoised (UNOISE3) and clustered; S_Du, clustered and denoised (UNOISE3); Da_S, denoised (DADA2) and clustered; S_Da, clustered and denoised (DADA2); Du_d_S, denoised (UNOISE3) with precedence to distance and clustered; Du_c_S, denoised (UNOISE3) with combined precedence and clustered; Du_e _S, denoised (UNOISE3) with correction taking into account the entropy of the codon positions and clustered; Du_e_d_S, denoised (UNOISE3) with correction plus precedence to distance and clustered; Du_e_c_S, denoised (UNOISE3) with correction plus combined precedence and clustered
*For the original and S datasets the number of sequences instead of ESVs is used
**The same values apply to Du_d (distance precedence) and Du_c (combined precedence)
***The same values apply to Du_e_d (distance precedence) and Du_e_c (combined precedence)

quality scores were highly correlated in the DADA2 learnErrors procedure. The lowest Pearson correlation was obtained between the substitutions T to C and A to G (r = 0.810), and all correlations (66 pairs of substitution types) were significant after a False Discovery Rate correction (Benjamini and Hochberg, 1995).

The main difference found is that the Du dataset retained almost double number of ESVs than the Da dataset: 60,198 vs 32,798. Of these, 31,696 were identical in the two datasets (Fig. 6.4), representing a match index of 0.746. Of the shared ESVs, 20,691 (65.28%) had exactly the same number of reads, suggesting that the same reads have been merged in these ESVs.

On the other hand, the shared ESVs concentrated most of the reads (Fig. 6.4): the match index for the reads was 0.986. This is coherent with the fact that most of the non-shared ESVs of the Du dataset had a low number of reads (mean = 3.66). Thus, the two denoising algorithms with

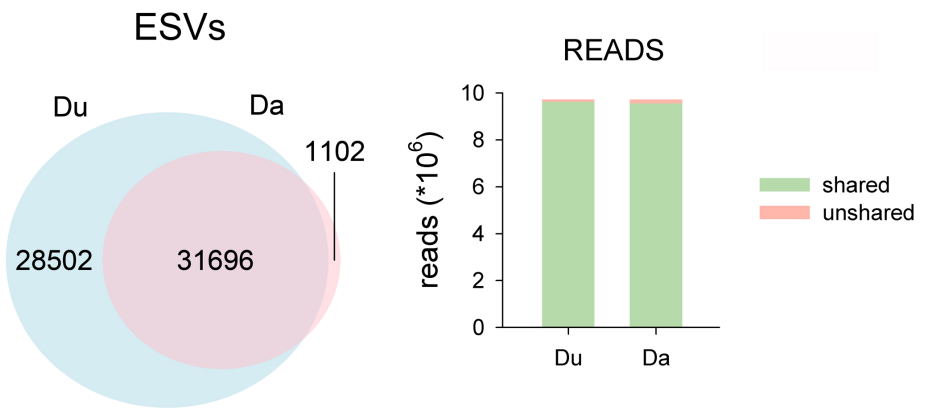


Figure 6.4: Venn Diagram showing the number of ESVs shared between the two denoising procedures (Du vs Da). Bar chart shows the number of reads in the shared and unshared ESVs.

the chosen parameter values provided similar results as for the abundant ESVs, but UNOISE3 retained a high number of low abundance ESVs as true sequences.

We then evaluated the output of combining denoising and clustering, using either of them as a first step. Thus, we compared the datasets Du_S, S_Du, Da_S, and S_Da. The results showed that the final number of MOTUs obtained was similar (ca. 19,000) irrespective of the denoising method and the order used (Table 6.1). Moreover, the shared MOTUs (flagged as MOTUs that have the same representative sequence) were the overwhelming majority (Fig. 6.5), with MOTU match indices over 0.96 in all comparisons.

As for the number of ESVs, clustering first results in a higher number of retained sequence variants than clustering last, ca. 25% more for Du and ca. 8% for Da. In all comparisons, the majority of ESVs were to be found in the shared MOTUs, and the same applies to the number of reads (Fig. 6.6, match indices for the ESVs, all > 0.95, match indices for the reads, all > 0.99). Ca. 2/3 of the MOTUs comprised a single ESV when using Du, and this number increased notably with Da (ca. 80% of MOTUs, Table 6.1). In both cases, clustering first resulted in a slight decrease of the number of single-ESV MOTUs.

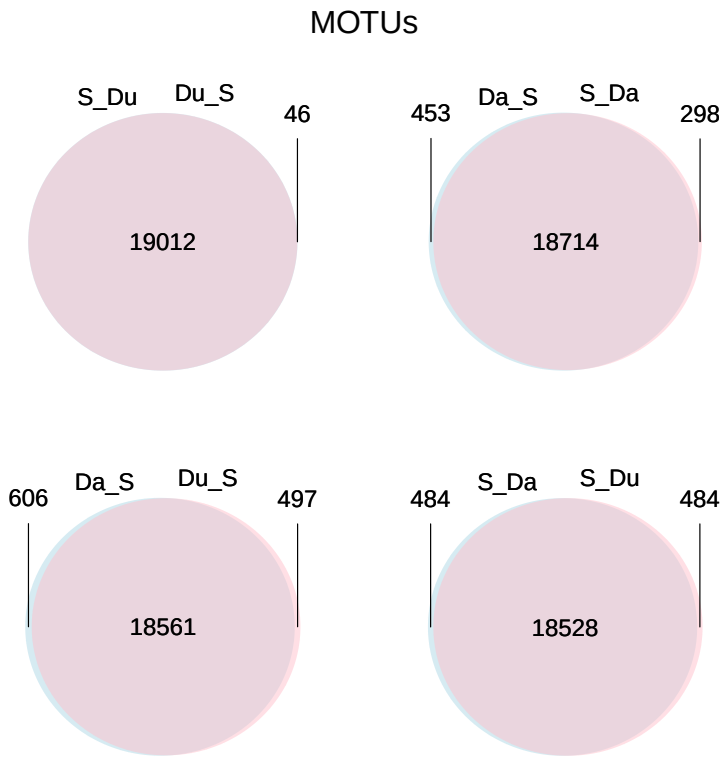


Figure 6.5: Venn diagrams showing the number of MOTUs shared between the two denoising procedures and a clustering step performed in different orders.

6.4.4 Improving the denoising algorithm

We tried different options of our DnoisE algorithm. The use of the Levenshtein distance without any correction and with priority to abundance skew corresponds to the original UNOISE3 algorithm (i.e., the Du dataset used previously). We also tried priority to distance and a combination of skew and abundance to choose among the potential "mother" ESVs to which a given "daughter" sequence will be joined. The same three options were applied when correcting distances according to the entropy of each codon position. In this case we used a pairwise distance accounting for the codon position where a substitution was found. We further applied a clustering step (SWARM) to the DnoisE results to generate MOTU sets (Du_S, Du_d_S, Du_c_S, Du_e_S, Du_e_d_S, Du_e_c_S, see Table

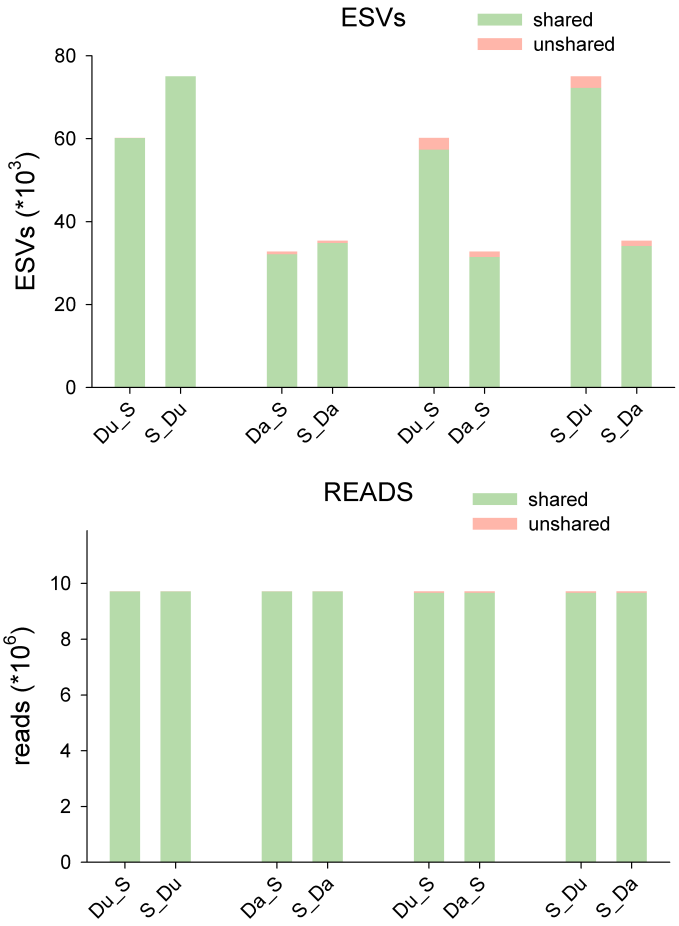


Figure 6.6: Bar charts of the number of ESVs and the number of reads found in the shared and unshared MOTUs in the same comparisons as in Fig. 6.5.

6.1 for explanation of codes) for comparison with those obtained previously.

The three ways to join sequences have necessarily the same ESVs, only the sequences that are joined under each centroid can vary and, thus, the abundance of each ESV and how these are clustered in MOTUs. However, this had a very small effect in our case. For the three datasets generated without distance correction, most MOTUs were shared, and the shared MOTUs comprised most ESVs. In turn most ESVs have the same number of reads, suggesting that the same sequences have been grouped in each ESV. All match indices were ca. 0.99. The same was found for the three

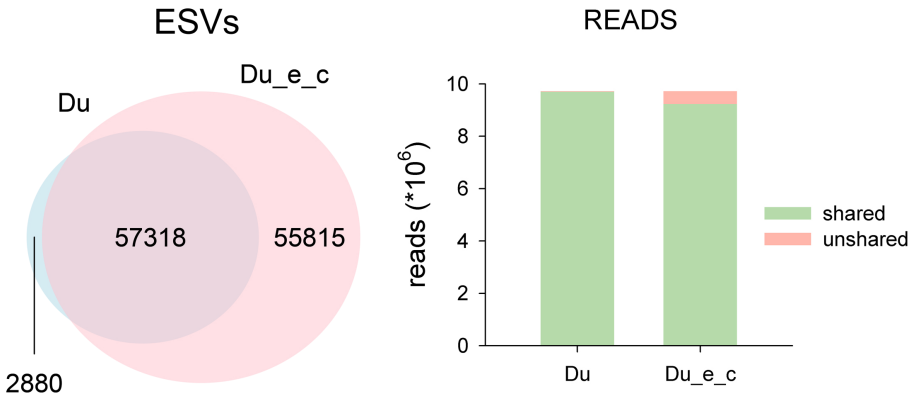


Figure 6.7: Venn Diagram showing the number of ESVs shared between two denoised datasets (Du vs Du_e_c). Bar chart shows the number of reads in the shared and unshared ESVs.

entropy-corrected datasets.

On the other hand, if we consider the entropy of codon positions the results change notably in terms of ESV recovered. The corrected datasets have 113,133 ESVs (against 60,198 of the uncorrected datasets). So, when considering the entropy in distance calculations the number of retained ESVs increased by 88%. This is the result of accepting sequences that have variation in third codon positions as legitimate. When comparing the entropy-corrected and uncorrected datasets 57,318 ESVs were found in common (ESV match index of 0.729). These ESVs comprise a majority of reads, though (read match indices of ca. 0.97 in all possible comparisons). Figure 6.7 illustrates one of these comparisons (Du vs Du_e_c).

When clustering the ESVs obtained with the different methods, the final number of MOTUs obtained was similar to those generated in the previous sections (ca. 19,000 in all cases, Table 6.1). This indicates that the entropy corrected datasets provided more intra-MOTU variability, but no appreciable increase in the number of MOTUs. As an example, the mean number of ESVs per MOTU was 3.159 for the Du_S dataset, and 5.949 for the Du_e_c_S dataset. The number of single-ESV MOTUs decreased slightly (12,471 for Du_S, 12,365 for Du_e_c_S). Taking this comparison as an example, most MOTUs (as indicated by identity in the representative

sequence) were shared between datasets. In addition, most of the ESVs and most of the reads were found in the shared MOTUs (match indices for MOTUs, ESVs and reads > 0.99).

6.4.5 Benchmarking

We computed the percent of erroneous ESVs (either because they have stop codons or changes in the five conserved aminoacids) in the MOTUs assigned to metazoans for the datasets obtained with and without entropy correction. The original dataset clustered without any denoising (dataset S) had 9,702 erroneous ESVs (or 4.65% of the total number of ESVs). The denoised dataset Du_S had 559 erroneous ESVs (1.58%), while the dataset denoised considering the variability of the codon positions (Du_e_c_S) had 500 erroneous ESVs (0.70%). Thus, albeit the uncorrected UNOISE3 procedure reduced the proportion of errors to one third, when a correction for codon position is applied the absolute number of errors is reduced, out of almost double total number of ESVs, thus the relative number is cut by more than one half.

The results of the taxonomic benchmarking are given in detail in Appendix C, while the obtained species-level dataset is available as Data C.4.1. In short, all datasets recovered a majority of closed MOTUs, meaning that ESVs assigned to a given species were placed in the same MOTU. The proportion of hybrid MOTUs was lower for the more stringent DADA2 datasets. On the contrary, the proportion of species recovered and the proportion of ESVs with species-level assignment was lowest for the DADA2 datasets and highest for the entropy-corrected UNOISE3 datasets.

6.5 Discussion

After adjusting the different parameters of the algorithms based on ad hoc criteria for COI amplicons, between ca. 33,000 and ca. 113,000 ESVs were obtained depending on the denoising procedure used. Irrespective of the method, however, they clustered into ca. 19,000 MOTUs. This implies that there was a noticeable intra-MOTU variability even for the

most stringent denoising method. The application of SWARM directly to the original dataset (without any denoising) generated likewise ca. 19,000 MOTUs. This suggests that the SWARM algorithm is robust in recovering alpha-diversity even in the presence of noisy sequences. Thus, denoising and clustering clearly accomplish different functions and, in our view, both are complementary and should be used in combination. The fact that some studies detect more MOTUs than ESVs when analysing datasets using clustering and denoising algorithms separately (e.g., Macheriotou et al., 2019; Nearing et al., 2018) reflects a logical flaw: MOTUs seek to recover meaningful species-level entities, ESVs seek to recover correct sequences. There should be more sequences than species, otherwise something is wrong with the respective procedures. It has even been suggested that ESVs or MOTUs represent a first level of sequence grouping and that a second round using network analysis is convenient (Forster et al., 2019). We contend that, with the right parameter settings, this is unnecessary for eukaryotic COI datasets.

We do not endorse the view of Callahan et al. (2017) that ESVs should replace MOTUs as the standard unit analysis of amplicon-sequencing datasets. Using information at the strain level may be useful in the case of prokaryotes, and in low-variability eukaryote markers such as ribosomal 18S rDNA there may be correspondence between species and unique sequences (indeed, in many cases different species share sequences). But even in more variable nuclear markers such as ITS, a clustering step is necessary (Estensmo et al., 2021). In eukaryotes the unit of diversity analyses is the species. MOTUs and not ESVs target species-level diversity and, in our view, should be used as the standard unit of analyses for most ecological and monitoring applications. Most importantly, that ESVs are organized into MOTUs is highly relevant information added at no cost. We do not agree that clustering ESVs into MOTUs eliminates biological information (Callahan et al., 2016). This only happens if only one representative sequence per MOTU is kept. We strongly advocate here for keeping track of the different sequences clustered in every MOTU and reporting them in metabarcoding studies. In this way analyses can be performed at the MOTU level or at the ESV level, depending on the question addressed.

Denoising has been suggested as a way to overcome problems of MOTU construction and to provide consistent biological entities (the correct sequences) that can be compared across studies (Callahan et al., 2017). We fully agree with the last idea: ESVs are interchangeable units that allow comparisons between datasets and can avoid generating too big datasets when combining reads of, say, temporally repeated biomonitoring studies. But clustering ESVs into MOTUs comes as a bonus, provided the grouped sequences are kept and not collapsed under a representative sequence, thus being available for future reanalyses.

The denoising and clustering methods here tested have been developed for ribosomal markers and uncritically applied to COI data in the past, with default parameter values often taken at face value (in fact, parameters are rarely mentioned in methods sections). We confirm that the UNOISE3 parameter $\alpha = 5$ is adequate for COI data, in agreement with previous research using three independent approaches (Elbrecht et al., 2018b; Shum and Palumbi, 2021; Turon et al., 2020). We also tested and confirmed the suitability of a d value of 13 for SWARM that has been used in previous works with COI datasets (e.g., Antich et al., 2020; Atienza et al., 2020; Bakker et al., 2019; Garcés-Pastor et al., 2019; Siegenthaler et al., 2019b). As Mahé et al. (2014) noted, higher d values can be necessary for fast evolving markers. They advised to track MOTU coalescing events as d increases to find the value best-fitting the sequence marker chosen. We have followed this approach, together with the course of the intra- and inter-MOTU distances, to select the d-value for the COI marker. In our view, fixed-threshold clustering procedures should be avoided, as even for a given marker the intra- and interspecies distances can vary according to the group of organisms considered. With SWARM, even if the initial clusters were made at $d = 13$ (for a fragment of 313 this means an initial threshold of 4.15% for connecting sequences), after the refining procedure the mean intra-MOTU distances obtained was 2.91%, which is in line with values suggested using the whole barcoding region of COI (Ratnasingham and Hebert, 2013). Furthermore, in our taxonomic benchmarking, we found a high proportion of closed MOTUs, irrespective of the denoising method used, indicating that the SWARM procedure adequately and robustly grouped

the sequences with known species-level assignments.

Our preferred algorithm for denoising is UNOISE3. It is a one-pass algorithm based on a simple formula with few parameters, it is computationally fast and can be applied at different steps of the pipelines. It keeps almost double ESVs than DADA2 and, combined with a clustering step, results in less single-sequence MOTUs and a higher number of ESVs per MOTU, thus capturing a higher intra-MOTU diversity. It also produced 60% more closed group MOTUs than DADA2 in our taxonomic benchmarking. Edgar (2016), by comparing both algorithms in mock and in vivo datasets, also found that UNOISE had comparable or better accuracy than DADA2. Similarly, Tsuji and Shibata (2020) found that UNOISE3 retained less false haplotypes than DADA2 in samples from tank water containing fish DNA. We also found that the entropy values of the sequences changed as expected when denoising becomes more stringent with UNOISE3, indicating that the algorithm performs well with coding sequences. We also suggest ways of improving this algorithm (see below).

DADA2, on the other hand, is being increasingly used in metabarcoding studies but its suitability for a coding gene such as COI remains to be demonstrated. We had to use paired reads (against recommendation) to be able to make meaningful comparisons, but our results indicate that with unpaired sequences the number of ESVs retained would have been even lower. The DADA2 algorithm, when tested with increasingly stringent parameters, did not progressively reduce the entropy ratio values that should reflect an adequate denoising of coding sequences. Further, the high correlation of error rates between all possible substitution types suggests that the algorithm may be over-parameterized, at least for COI, which comes at a computational cost. Comparisons based on known communities (as in Tsuji and Shibata, 2020) and using COI are needed to definitely settle the appropriateness of the two algorithms for metabarcoding with this marker.

In addition, PCR-free methods now popular in library preparation procedures complicate the use of DADA2 as there is no consistent direction (forward or reverse) of the reads. We acknowledge that our paired sequences

still included a mixture of reads that were originally in one or another direction and, thus, with different error rates. However, the non-overlapped part is only the initial ca. 100 bp, and these are in general good quality positions in both the forward and reverse reads.

Another choice to make is to decide what should come first, denoising or clustering. Both options have been adopted in previous studies (note that clustering first is not possible with DADA2 unless paired sequences are used). Turon et al. (2020) advocated that denoising should be made within MOTUs, as they provide the natural "sequence environment" where errors occur and where they should be targeted by the cleaning procedure. We found that clustering first retained more ESVs, because sequences that would otherwise be merged with another from outside its MOTU were preserved. It also resulted in less single-ESV MOTUs, retaining more intra-MOTU variability. It can also be mentioned that denoising the original sequences took approximately 10 times more computing time than denoising within clusters, which can be an issue depending on the dataset and the available computer facilities. We acknowledge, however, that most MOTUs are shared and most ESVs and reads are in the shared MOTUs when comparing the two possible orderings, irrespective of denoising algorithm. The final decision may come more from the nature and goals of each study. For instance, a punctual research may go for clustering first and denoising within clusters to maximize the intra-MOTU variability obtained. A long-term research that implies multiple samplings over time that need to be combined together may use denoising first and then perform the clustering procedure at each reporting period with the ESVs obtained in the datasets collected so far pooled.

There are other important steps at which errors can be reduced and that require key choices, but they are outside the scope of this work as we addressed only clustering and denoising steps. In particular, nuclear insertions (numts) may be difficult to distinguish from true mitochondrial sequences (Andújar et al., 2021; Porter and Hajibabaei, 2021). Singletons (sequences with only one read) are also a problem for all denoising algorithms (as it is difficult to discern rare sequences from errors). Singletons are often

eliminated right at the initial steps, as we did in this work. Likewise, a filtering step, in which ESVs with less than a certain amount of reads are eliminated, is deemed necessary to obtain biologically reliable datasets. A 5% relative abundance cut-off value was suggested by Elbrecht et al. (2018b), while Turon et al. (2020) proposed an absolute threshold of 20 reads. However, the procedure and the adequate threshold are best adjusted according to the marker and the study system, so, albeit we acknowledge that a filtering step is necessary, this has not been addressed in this paper.

We recommend that the different denoising algorithms be programmed as stand-alone steps (not combined, for instance, with chimera filtering) so anyone interested could combine the denoising step with the preferred choices for other steps. We also favour open source programs that could be customized if needed. For UNOISE3 algorithm we suggest that a combination between distance and skew ratio be considered to assign a read to the most likely centroid. This had little effect in our case, but can be significant in other datasets. For DADA2 algorithm, we advise to weight the gain of considering the two reads separately vs using paired sequences. The advantages of the latter involve a higher flexibility of the algorithm as it does not need to be performed right at the beginning of the pipeline. For both algorithms, we think it is important to consider the natural variation of the three positions of the codons of a coding sequence such as COI, which can allow a more meaningful computation of distances between sequences and error rates. This of course applies to other denoising algorithms not tested in the present study (e.g., AmpliCI Peng and Dorman, 2020, deblur Amir et al., 2017). Our DnoiSE program, based on the UNOISE3 algorithm, includes the option of incorporating codon information in the denoising procedure. With this option, we found ca 50,000 more ESVs than with the standard approach. Importantly, this fact did not increase the proportion of erroneous sequences, as determined using aminoacid substitution patterns in metazoan MOTUs. Rather, this proportion was cut by one-half, and erroneous sequences were less even in absolute numbers. In our taxonomic benchmarking, a higher proportion of ESVs with species-level matches in the reference database were detected with the codon position-corrected method. We used a dataset of fixed sequence length and eliminated misaligned se-

quences. The correction for codon position would be more complicated in the presence of indels and dubious alignments. We also acknowledge the lack of a mock community to ground truth our method, but we contend that mock communities are hardly representative of highly complex communities such as those here analysed. We hope our approach will be explored further and adequately benchmarked in future studies on different communities.

6.6 Conclusions

COI has a naturally high intraspecies variability that should be assessed and reported in metabarcoding studies, as it is a source of highly valuable information. Denoising and clustering of sequences are not alternatives. Rather, they are complementary and both should be used together to make the most of the inter- and intraspecies information contained in COI metabarcoding datasets. We emphasize the need to carefully choose the stringency parameters of the different steps according to the variability of this marker.

Our results indicated that the UNOISE3 algorithm preserved a higher intra-cluster variability than DADA2. We introduce the program DnoisE to implement the UNOISE3 algorithm considering the natural variability (measured as entropy) of each codon position in protein-coding genes. This correction increased the number of sequences retained by 88%. The order of the steps (denoising and clustering) had little influence on the final outcome.

We provide recommendations for the preferred algorithms of denoising and clustering, as well as step order, but these may be tuned according to the goals of each study, feasibility of preliminary tests, and ground-truthing options, if any. Other important steps of metabarcoding pipelines, such as abundance filtering, have not been addressed in this study and should be adjusted according to the marker and the study system.

We advise to report the results in terms of both MOTUs and ESVs included in each MOTU, rather than reporting only MOTU tables with collapsed information and just a representative sequence. We also advise

that the coding properties of COI should be used both to set the right parameters of the programs and to guide error estimation in denoising procedures. We wanted to spark further studies on the topic, and our procedures should be tested and validated or refined in different types of community.

There is a huge amount of intra- and inter-MOTU information in metabarcoding datasets that can be exploited for basic (e.g., biodiversity assessment, connectivity estimates, metaphylogeography) and applied (e.g., management) issues in biomonitoring programs, provided the results are reported adequately.

Chapter 7

DnoisE: distance denoising by entropy. An open-source parallelizable alternative for denoising sequence datasets

7.1 Abstract

DNA metabarcoding is broadly used in biodiversity studies encompassing a wide range of organisms. Erroneous amplicons, generated during amplification and sequencing procedures, constitute one of the major sources of concern for the interpretation of metabarcoding results. Several denoising programs have been implemented to detect and eliminate these errors. However, almost all denoising software currently available has been designed to process non-coding ribosomal sequences, most notably prokaryotic 16S rDNA. The growing number of metabarcoding studies using coding markers such as COI or RuBisCO demands a re-assessment and calibration of denoising algorithms. Here we present DnoisE, the first denoising program designed to detect erroneous reads and merge them with the correct ones using information from the natural variability (entropy) associated to each codon position in coding barcodes. We have developed an open-source software using a modified version of the UNOISE algorithm. DnoisE implements different merging procedures as options, and can incorporate codon entropy information either retrieved from the data or supplied by the user. In addition, the algorithm of DnoisE is parallelizable, greatly reducing runtimes on computer clusters. Our program also allows different input file formats, so it can be readily incorporated into existing metabarcoding pipelines.

7.2 Background

Biodiversity studies have experienced a revolution in the last decade with the application of high throughput sequencing (HTS) techniques. In particular, the use of metabarcoding in ecological studies has increased notably in recent years. For both prokaryotic and eukaryotic organisms, a large number of applications have been developed, ranging from biodiversity assessment (Wangensteen et al., 2018b), detection of particular species (Kelly et al., 2014a), analysis of impacts (Pawlowski et al., 2018), and diet studies (Clarke et al., 2020; Sousa et al., 2019), among others. Also, different sample types have been used: terrestrial soil, freshwater, marine water, benthic samples, arthropod traps, or animal faeces (Creer et al., 2016; Deiner et al., 2017). Many of these studies have direct implications on management and conservation of ecosystems and are thus providing direct benefits to society. They have also brought to light a bewildering diversity of organisms in habitats difficult to study with traditional techniques.

Metabarcoding studies have greatly contributed to so-called big community data (Pichler and Hartig, 2021) by generating an enormous amount of sequence data that, in most cases, is available online. Handling these datasets is memory intensive and filtering steps are required to analyze such information. Clustering and denoising are the two main strategies to compress data into Molecular Operational Taxonomic Units (MOTUs, aka OTUs) or Exact Sequence Variants (ESVs; also ASVs, Amplicon Sequence Variants, or ZOTUs, zero ratio OTUs) to extract biodiversity composition (Antich et al., 2021a). Both methods rely on minimizing sequencing and PCR errors either by clustering sequences into purportedly meaningful biological entities (MOTUs) or by merging erroneous sequences with the correct ones from which they possibly originated, and keeping just correct amplicons (ESVs). Hence, both methods differ philosophically and analytically. Furthermore, they are not incompatible and can be jointly applied. Software development is crucial to create tools capable of performing these tasks in a fast and efficient way. The type of samples, the marker, and the target organisms are also instrumental in choosing the adequate bioinformatic pipelines to provide interpretable results.

Recent studies have explored the joint application of both methods to filter metabarcoding data (Antich et al., 2021a; Brandt et al., 2021; Elbrecht et al., 2018b; Turon et al., 2020). Importantly, the combination of clustering and denoising opens the door to the analysis of intraspecies (intra-MOTU) variability (Antich et al., 2021a). Turon et al. (2020) proposed the term metaphylogeography for the study of population genetics using metabarcoding data, and Zizka et al. (2020) found different haplotype composition between perturbed and unperturbed rivers, both studies using a combination of clustering and denoising steps.

The software presented here focuses on the denoising step. There are currently several software programs developed to denoise sequencing and PCR errors, such as DADA2 (Callahan et al., 2016), AmpliCL (Peng and Dorman, 2020), Deblur (Amir et al., 2017), or UNOISE (Edgar, 2016). These programs have been widely used in metabarcoding studies to generate ESVs, using sequence quality information for the first two and simple analytical methods for the latter two. All were originally tested for ribosomal DNA (non-coding) and thus some adjustment is necessary for application to other markers (Antich et al., 2021a).

Here we present DnoisE, a parallelizable Python3 software for denoising sequences using a modification of the UNOISE algorithm and tested for metabarcoding of eukaryote communities using mitochondrial markers (COI, Cytochrome Oxidase subunit I). We introduce a novel correction procedure for coding sequences using changes in diversity values per codon position. In coding genes, the natural entropy of the different positions is markedly different, with the third position being always the most variable. We therefore contend that differences in each position should have different weights when deciding whether a change in a given position is legitimate or is attributable to random PCR or

sequencing errors. DnoisE is also applicable to other markers due to the settable options and offers a fast and open source alternative to non-parallelizable closed source programs. Scripts for installation and example files to run DnoisE are provided in the GitHub repository:

https://github.com/adriantich/DnoisE.

7.3 Workflow

7.3.1 Structure of input files

DnoisE is designed to run with HTS datasets (after paired-end merging and de-replicating sequences) to obtain ESVs, or after clustering with SWARM (Mahé et al., 2015) to obtain haplotypes within MOTUs. Due to variability in format files, we have designed an algorithm that can read both fasta and csv files. In the present version, however, sample information (if present) is kept only for csv input.

7.3.2 Combining the UNOISE algorithm and the entropy correction

Sequences are stored as a data frame, with each row corresponding to a sequence record and the columns to the abundances (either total or per sample). The original Edgar's 2016 function used by UNOISE to determine whether two sequences should be merged is:

$$\beta(d) = 1/2^{\alpha d+1}$$

where $\beta(d)$is the threshold abundance ratio of a less abundant sequence with respect to a more abundant one (from which it differs by distance d) below which they are merged. The distance d is the Levenshtein genetic distance measured in DnoisE with the Levenshtein module (https://maxbachmann.github.io/Levenshtein/) and α is the stringency parameter (the higher α, the lower the abundance skew required for merging two sequences).

The UNOISE algorithm sorts sequences by decreasing abundance and each one is compared with the less abundant ones. At each comparison, the distance between sequences (d) is computed and, if the abundance ratio between the less abundant and the more abundant sequence is lower than $\beta(d)$, the former is assumed to be an error. In UNOISE terminology, the sequences form clusters, of which the correct one is the centroid and the

remaining members are inferred to derive from the centroid template but contain errors. In his original paper, Edgar (2016) suggests constructing a table of centroids excluding low abundance reads, and then constructing a ZOTU table by mapping all reads (before the abundance filtering) to the centroids table using the same merging criterion but without creating new centroids. So, the original formulation of this algorithm gives priority to the abundance ratio over the genetic distance. The first, very abundant, sequences will "capture" rare sequences even if d is relatively high. Other, less abundant sequences may be closer (lower d) and still fulfill Edgar's formula for merging the rare sequence, but this will never happen as the rare sequence will be joined with the very abundant one and will not be available for further comparisons. However, in the standard procedure of this algorithm implemented as UNOISE3 in the USEARCH pipeline (Edgar, 2010; https://drive5.com/usearch/), the reads are mapped to the centroid table using a similarity criterion (identity threshold in the otutab command), so in practice a distance criterion is used during the mapping.

DnoisE is a one pass algorithm, with no posterior mapping of reads to centroids (which is indeed repetitive, as reads have already been evaluated against the centroids when constructing the centroid table) and with a choice of merging criteria. If deemed necessary, low abundance reads can be eliminated previously or, alternatively, ESVs with one or a few reads can be discarded after denoising. Chimeric amplicons can likewise be eliminated before or after denoising. DnoisE follows previously used terminology (Antich et al., 2021a; Turon et al., 2020) in which the correct sequences (centroids in UNOISE terms) are called "mother" sequences and the erroneous sequences derived from them are labelled "daughter" sequences. DnoisE provides different options for merging the sequences. Let PMS (potential "mother" sequence) and PDS (potential "daughter" sequence) denote the more abundant and the less abundant sequences that are being compared, respectively, and let d be the genetic distance between them. When the abundance ratio PDS/PMS is lower than $\beta(d)$, the PDS is tagged as an error sequence but is not merged with the PMS. Instead, a round with all comparisons is performed and, for a given PDS, all PMS fulfilling the UNOISE criterion for merging are stored. After this round is completed,

the merging is performed following one of three possible criteria: (1) Ratio criterion, joining a PDS to its more abundant PMS (lowest abundance ratio, corresponding to the original UNOISE formulation); (2) Distance criterion, joining a sequence to the closest (least d value) possible "mother"; and the (3) Ratio-Distance criterion, whereby a PDS is merged with the PMS for which the quotient $\beta/\beta(d)$ (i.e., between the abundance ratio PDS/PMS and the maximal abundance ratio allowed for the observed d), is lowest, thus combining the two previous criteria. For each criterion, the best PMS and the corresponding values (ratio, d and ratio skew values) are stored. The user then has the choice to select one or another for merging sequences. As an option, if the user wants to apply only the Ratio criterion, each PDS is assigned to the first (i.e., the most abundant) PMS that fulfills the merging inequality and becomes unavailable for further comparisons, thus decreasing computing time. Figure 7.1 shows a conceptual scheme of this workflow process.

In addition, for coding markers such as COI, the codon position provides crucial additional information that must be taken into account. In nature, the third codon position is the most variable, followed by the first and the second position. This variation can be measured as entropy (Schmitt and Herzel, 1997) of the different positions. A change in third position is more likely to be a natural change (and not an error) than the same change in a second position, much less variable naturally. To our knowledge, no denoising algorithm incorporates this important information. We propose to use the entropy values of each codon position to correct the distance d in Edgar's formula as follows:

$$d_{corr} = \sum_{i=1}^{3} d(i) \times entropy(i) \times 3/(entropy(1) + entropy(2) + entropy(3))$$

where i is the codon position and d is the number of differences in each position. The d_{corr} value is then used instead of d in the formula. This correction results in a higher d_{corr} when a change occurs in a third position than in the first or second position, thus a sequence with changes in third positions will be less likely to be merged. In practice, as many changes occur naturally in third positions, this correction will lead to a higher number of

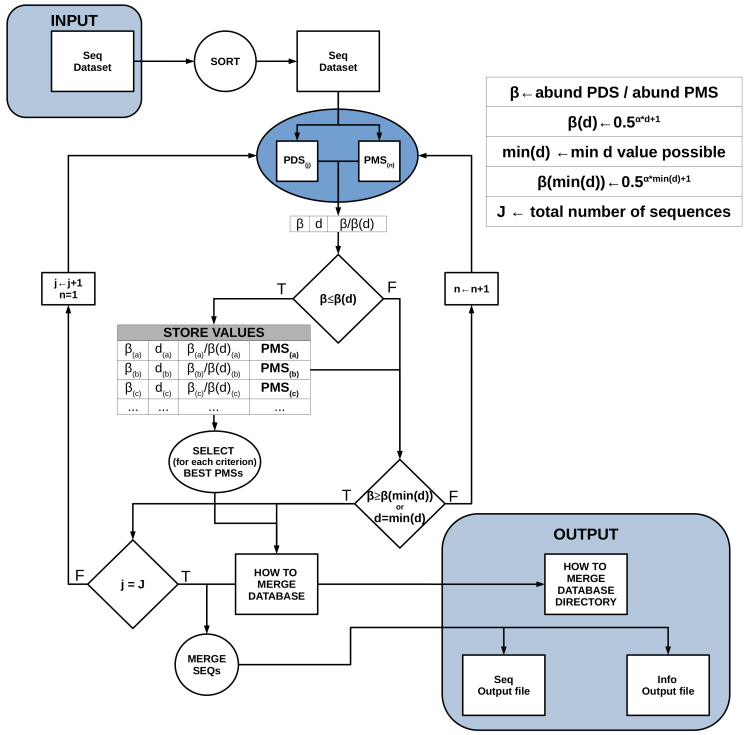


Figure 7.1: Scheme of the workflow of DnoisE. Starting from an abundance-sorted sequence dataset, subsets of possible daughter sequences (PDS) and possible mother sequences (PMS) are selected as detailed in Fig. 7.2. For each subset, all PDS are compared with all compatible PMS (in terms of MDA and MMA). If the merging inequality is met, the values of the main parameters are stored. After all subsets have been evaluated, for each merging criterion the best PMS for each PDS is chosen and a sequence file is generated, together with a file with information on the merging process.

ESVs retained that would otherwise be considered errors. Careful choice of entropy values is crucial, and it is recommended that they are adjusted for each marker and particular study. The values of entropy for each position can be obtained from the data (computed directly by the program) or added manually by the user.

Note that, when applying this correction, the Levenshtein distance is not used as it cannot consider codon positions. Instead, the number of differences is used. In practice, in aligned sequences with no indels both distances are equivalent. In addition, with the entropy correction, lengths

should be equal when comparing two sequences. The dataset is thus analysed separately by sequence length sets. These sets must differ from the modal length (the modal sequence length can also be set using the -m parameter) of the complete dataset by n number of codons (groups of three nucleotides), as in general indels in coding sequences are additions or deletions of whole codons. A sequence differing from these accepted lengths is considered erroneous and removed. Sequences of the same length must be aligned for the algorithm to run properly.

7.3.3 Parallel processing

Parallel processing is a useful tool to increase speed when multicore computers are available. DnoisE implements parallel processing in the algorithm so the required time to run huge datasets decreases drastically as more cores are used. Parallel processing was applied using the multiprocessing module of Python3 (McKerns et al., 2011). A computational bottleneck of denoising procedures is their sequential nature, which is hardly parallelizable, and more so in the case of DnoisE that computes all comparisons before merging. In particular, a sequence that has been tagged as "daughter" (error) cannot be a "mother" of a less abundant sequence. Therefore, to compare a PDS to all its PMS requires that those more abundant sequences have been identified as correct before.

We incorporate two concepts, based on the highest skew ratio required for a sequence to be merged with a more abundant one. This is of course $\beta(min(d))$, where *min(d)* is one if entropy correction is not performed, and it equals the d cor r corresponding to a single change in the position with less entropy (position 2) if entropy is considered. From this maximal abundance ratio we can obtain, for a given potential "mother", the maximal "daughter" abundance (MDA, any sequence more abundant than that cannot be a "daughter" of the former). Conversely, for a given "daughter" sequence we can obtain the minimum "mother" abundance (MMS, any sequence less abundant than that cannot be the "mother" of the former). The formulae are:

$$MDA = abundancePMS/\beta(min(d))$$
$$MMA = \beta(min(d))/abundancePDS$$

$$\beta(min(d)) = 0.5^{\alpha \cdot 1 + 1}$$

$$OR$$

$$\beta(min(d)) = 0.5^{\alpha \cdot min(entropy(i) \cdot 3/(entropy(1) + entropy(2) + entropy(3))) + 1}$$

The use of MDA and MMA simplifies the workload of the program as it greatly reduces the number of comparisons (a PMS will not be evaluated against sequences more abundant than the MDA, and a PDS will not be compared with sequences with less abundance than the MMA). Likewise, it allows for a parallel processing of sequences using the MDA as follows:

1- Sequences are ordered by decreasing abundance.

2- The first sequence is automatically tagged as a correct sequence.

3- MDA is calculated for this sequence (MDA_1).

4- All sequences with abundances between the first sequence and the MDA are, by definition, tagged also as correct sequences.

5- For the last sequence tagged as correct, the MDA is calculated (MDA_2).

6- Every sequence with abundance between the last correct sequence and MDA_2 is evaluated in parallel against all correct sequences that are more abundant than its MMA. Those for which no valid "mother" is found are tagged as correct, the rest are "daughter" (error) sequences.

7- Repeat steps 5 and 6 (i.e., calculating MDA_3 to n) until all sequences have been evaluated.

Figure 7.2 provides a conceptual scheme of this procedure. Note that, for each block of sequences that is evaluated in parallel, no comparisons need to be performed between them as they will never fulfill the merging inequality. After this process is completed, all sequences not labelled as "daughter" are kept as ESVs, and all "daughters" are merged to them according to the merging criterion chosen.

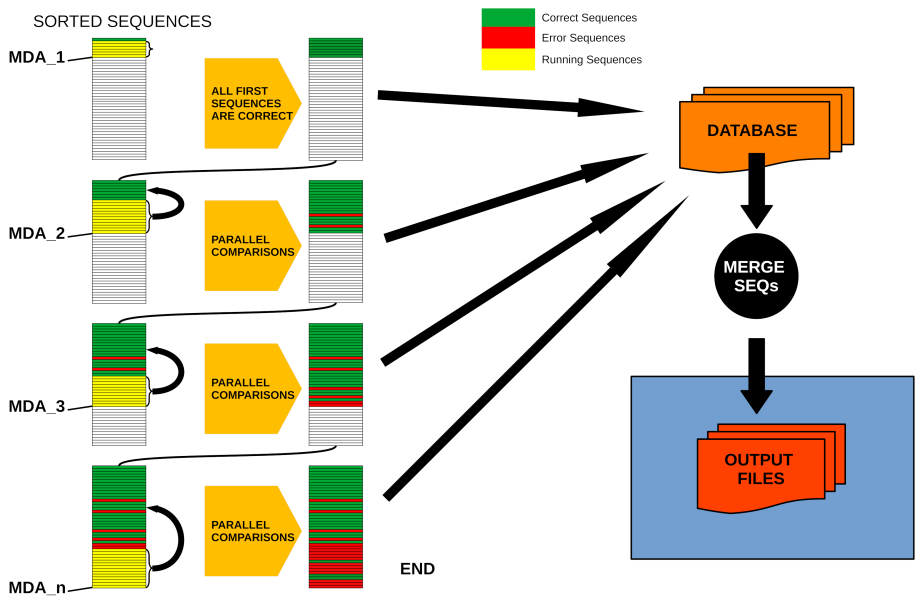


Figure 7.2: Schematic workflow of parallel processing of DnoisE. When running in parallel, comparisons between sequences are computed in sets of sequences defined by their abundances. Using the Maximum Daughter Abundance (MDA) value, computed from the last correct sequence of the previous step, we can define sets of sequences that are compared in parallel with the previously tagged correct sequences.

7.4 DnoisE Performance

A previous version of DnoisE was tested in Antich et al. (2021a) on a COI metabarcoding dataset of marine benthic communities. The version used in Antich et al. (2021a) implemented the same basic algorithm but was not curated for general use. For the present version, we have corrected bugs, made the program user-friendly, and added more settable options and features. The dataset consisted of 330,382 chimera-filtered COI sequences of 313 bp (all sequences had more than one read). They came from benthic marine communities in 12 locations of the Iberian Mediterranean coast (see Antich et al., 2021a for details), and are available as a Mendeley Dataset (https://data.mendeley.com/datasets/84zypvmn2b/). DnoisE was used in Antich et al. (2021a) in combination with the clustering algorithm SWARM, and was compared with the results of DADA2 denoising algorithm. Antich et al. (2021a) also compared DnoisE with and without entropy correction,

and obtained twice the number of ESVs with correction, while he proportion of erroneous sequences (defined as those having stop codons or substitutions in conserved positions) decreased to one half as compared with not correcting for codon position variation, as discussed in Antich et al. (2021a).

7.4.1 Comparison with UNOISE3

We benchmarked the current version of DnoisE (with alpha = 5) against the current implementation of the UNOISE algorithm: UNOISE3 (USEARCH 32-bit, free version, with alpha = 5 and minsize = 2) on this same dataset. To be able to make a direct comparison, for UNOISE3 we didn't perform an otutab step, rather, we recovered the ESVs and their abundance directly from the output files generated with -tabbedout and -ampout. As chimeric sequences were already removed from the dataset, and for the sake of comparability, we didn't exclude the few sequences flagged as such by the chimera filtering procedure embedded in UNOISE3. The number of ESVs obtained was almost the same: 60,198 and 60,205, respectively, if no entropy correction was performed. In addition, 60,196 ESVs were shared (comprising $> 99.999\%$ of the total reads) among the two programs, confirming that DnoisE (without correction) and UNOISE3 were practically equivalent. For further analyses of the effect of entropy correction we will therefore compare DnoisE with and without this correction.

7.4.2 Running performance

We compared the run speed of DnoisE with and without entropy correction for the same dataset of sequences. We used different numbers of cores, from 1 to 59, for parallelization. We applied the entropy correction values from Antich et al. (2021a).

Running DnoisE with just one core (without entropy correction) took about 29 h, decreasing sharply when using parallel processing with just a few cores. DnoisE took 4.5 h with 6 cores and 2.78 h with 10 cores. As a reference, the execution time of UNOISE3 (32-bit version, not parallelizable) without the otutab step was ca. 7 h, albeit this execution time is not directly comparable as UNOISE3 has a chimera filtering step embedded.

Using entropy correction, run times increased (Fig. 7.3) as there is a higher number of comparisons needed because the MMA values are generally lower. This slows the process as any given PSD has more PMS to compare with. With entropy correction, DnoisE retrieved ca. twice the number of ESVs, further increasing run time. For the Ratio-Distance merging criterion, when entropy correction was performed, 16 cores were required for DnoisE to run at a similar time speed than 6 cores with no entropy correction (Fig. 7.3). Above 10 cores (without correction) or 20 cores (with correction), run times reached a plateau and did not further improve, while memory usage continued to increase steadily. A trade-off between both parameters should be sought depending on the cluster architecture and the dataset being run.

7.4.3 Merging performance

Due to the practical impossibility of building a mock community of the complexity required with known COI haplotypes for multiple species, in order to compare the merging performance of the original formula of UNOISE with the entropy correction available in DnoisE, we performed a simulation following the procedure described in Turon et al. (2020), and using the same dataset of 1,000 "good" sequences from marine samples used in that study. The rationale was to start with a dataset of good sequences with realistic read abundance distribution, simulate sequencing errors at a given error rate (henceforth "error" amplicons), and then denoise the resulting dataset to recover the original one. In addition, in the present study we kept track of which original sequences produced each error amplicon and used this information to check if error sequences are merged or not with their "true" mother. We applied a random error rate per base of 0.005, which is intermediate among reported values for Illumina platforms (Pfeiffer et al., 2018; Schirmer et al., 2016). After the simulation, we removed all sequences with only one read. This resulted in a dataset with the 1,000 original sequences and 265,297 error sequences.

We used the DnoisE software with and without entropy correction (the latter equivalent to the UNOISE3 results, see above) to denoise the simulated dataset. The entropy values were automatically computed from the data by

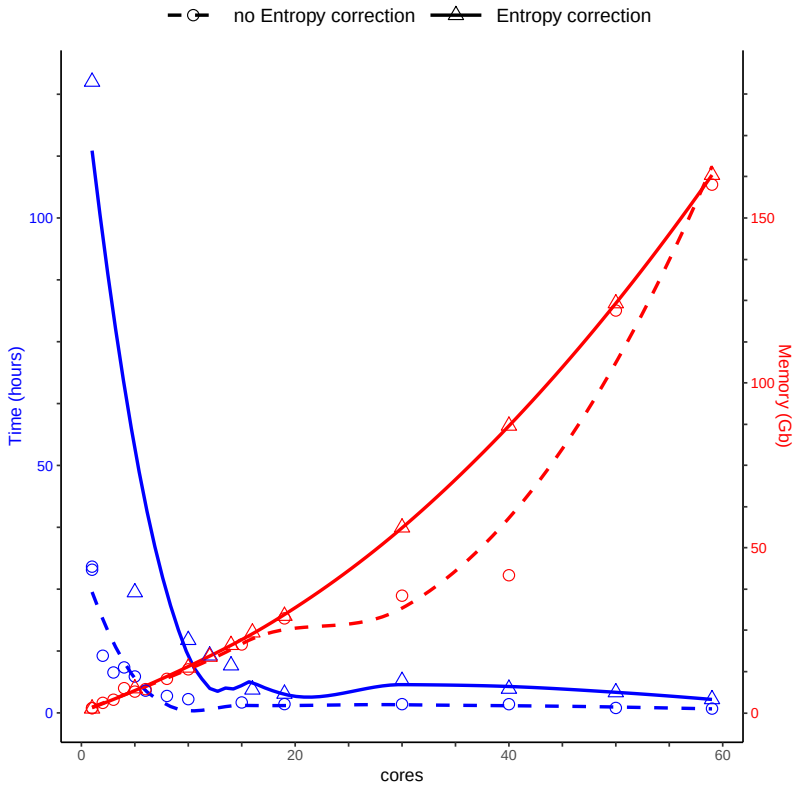


Figure 7.3: Time (blue) and memory (red) used by DnoisE to denoise and merge sequences with the Ratio-Distance criterion using different cores on a computer cluster. Denoising using entropy correction (triangles and dashed line) is compared against no correction (circles and dashed line). Lines are computed using the geom_smooth() function of the ggplot2 package with method = 'loess'.

the program and we tested alpha values from 10 to 1 (from lowest to highest stringency level). The results showed a decreasing number of total remaining sequences with more stringent (lower) alpha values (Fig. 7.4). There was also a drop in the number of good sequences remaining as alpha diminished. Except for the less stringent alpha values, however, data denoised with entropy correction kept a higher number of true sequences. With entropy correction, they remained almost constant for alpha values of 5 or higher, and decreased at lower values. Without entropy correction, the number of true sequences started to decrease at alpha values below 8. On the other hand, the entropy correction procedure also retrieved a higher number of false positives (i.e., error sequences) at intermediate alpha values, but the

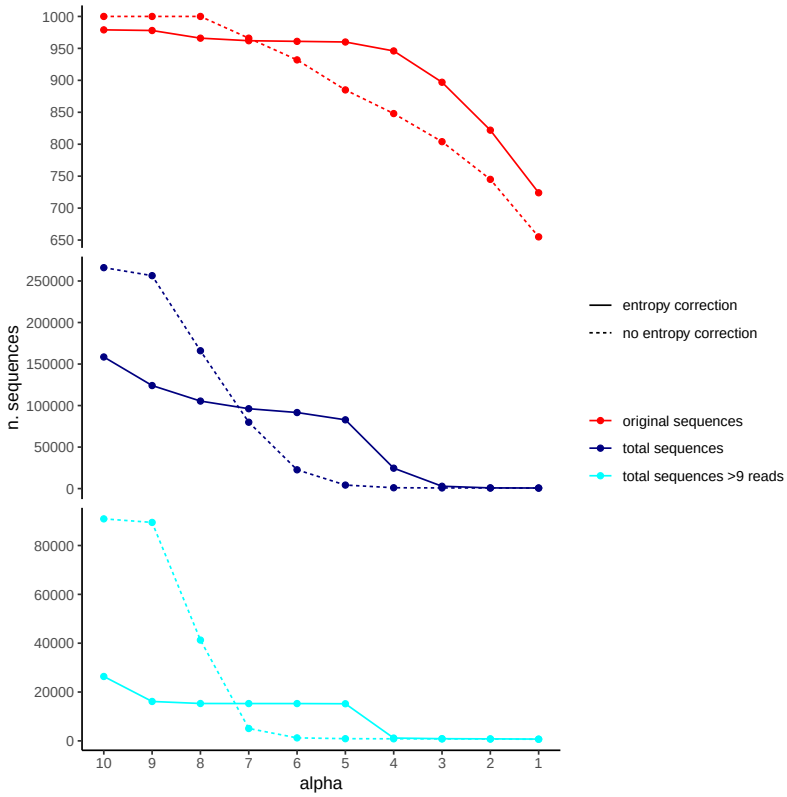


Figure 7.4: Number of original (correct) sequences (red), total sequences (dark blue) and total sequences filtered by read abundance (light blue) retrieved by DnoisE with entropy correction (solid line) and without entropy correction (equivalent to UNOISE). Values with abundance filtering were computed using a minimum abundance of 10 reads (–min_abund 10).

vast majority of them could be removed by applying a minimum abundance filter of 10 reads (–min_abund 10).

We also computed the match ratio, which is the ratio of sequences that merged with their "true" mothers divided by the number of merged sequences (Fig. 7.5). For alpha values of 6 or higher, the match ratio was close to 1 irrespective of the use of entropy correction or not, albeit it was slightly better without correction. At lower values of alpha, the match ratio decreased markedly for the Ratio merging criterion, and more so without correction, reaching values of ca. 75% at alpha = 1. There were also marked differences in the three joining criteria (compared only for the runs with

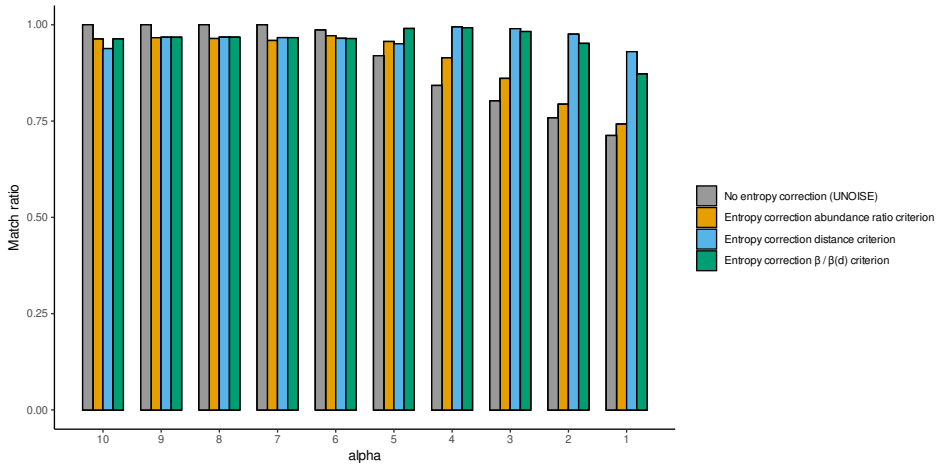


Figure 7.5: Match ratio (error sequences merged to their "true" mothers/total number of merged sequences) of DnoisE without entropy correction and abundance ratio joining criterion (equivalent to UNOISE) grey bars) and DnoisE with entropy correction. For DnoisE with entropy correction the three merging criteria were compared, abundance ratio criterion (orange bars), the genetic distance criterion (blue bars) and the criterion based on the cocient between the abundance ratio and the $\beta(d)$ (green bars)

entropy correction). While the abundance Ratio criterion resulted in a strong decrease of the match ratio, using the Distance or the Ratio-Distance joining criteria, the match ratios remained close to 1 until values of alpha 3 and decreased slightly at alpha 2 and 1. Note that the different joining criteria do not affect the number of ESVs produced, but the number of sequences merged with each ESV and, thus, their relative abundances. By keeping track of which original sequence produced each error sequence, we could compare how the relative performance of the different methods changed with alpha values.

While this simulated dataset may not be a perfect representative of true metabarcoding datasets, it nevertheless highlights the importance of choosing the correct parameters of both alpha and minimum abundance filtering values as well as the need of choosing the proper joining criterion, especially at more stringent denoising levels (lower values of alpha). Note also that the results can vary depending on the error rate (we acknowledge that applying an uniform error rate of 0.005 is a simplification). Alpha values

of 5 have been proposed for datasets of this COI fragment (Elbrecht et al., 2018b; Shum and Palumbi, 2021; Turon et al., 2020) using several lines of evidence, but none of these studies included entropy correction. In addition, a minimal abundance filtering step is deemed necessary (Elbrecht et al., 2018b; Turon et al., 2020) but an adequate threshold should be determined in each case. With our dataset and the explored error rate, values of 4 for alpha and 10 for minimal abundance seem a good compromise between keeping ca. 95% good sequences and accepting only a few error sequences. Our results emphasize the importance of calibrating the parameters for each type of data using any available evidence, including mock community data when available. The flexibility of DnoisE can greatly facilitate this exercise in future studies.

7.5 Conclusions

DnoisE is a novel denoising program that can be incorporated into any metabarcoding pipeline. It is a stand-alone program that addresses exclusively the denoising step, so that users can apply their favourite programs at all other steps (e.g., chimera filtering, clustering...). Moreover, DnoisE is open-source code. Other programs used in metabarcoding pipelines also have open codes, such as DADA2 (Callahan et al., 2016), OBITOOLS (Boyer et al., 2016), SWARM (Mahé et al., 2015), or VSEARCH (Rognes et al., 2016). We strongly adhere to the open software concept for continuous and collaborative development of computing science and, in particular, in the metabarcoding field.

DnoisE is based on the UNOISE algorithm developed by Edgar (2016), but with three main improvements: first, it allows to select among different criteria for joining sequences to optimize the match ratio; second, it incorporates the option to perform an entropy correction for coding genes, thus keeping more true sequences with high natural variability in third nucleotide positions in the codon; third, it is parallelizable to take advantage of the cluster architecture of modern computers.

Our correction by entropy opens a new field of analysis of coding genes,

considering the different natural variability between codon positions. The flexibility of DnoisE with its settable options make this program a good tool for optimizing parameters in metabarcoding pipelines and for running the denoising step at any desired point of the pipeline (before or after clustering sequences into MOTUs).

In the next few years, processors are expected to reach the minimum size permitted by quantum laws. Parallel processing is needed to optimize future computer performance (Gebali, 2011; Zomaya, 2005). DnoisE offers a new parallel processing algorithm based on the MDA (maximum "daughter" abundance) to run analyses in parallel by groups of sequences that do not need to be compared between them. Parallel processing allows users to run huge datasets in a fast way using multithread computers. In our example, when running with 10 cores, DnoisE took about 2.78 h to compute a large dataset. On the other hand, memory management can be critical when running a high number of cores and large datasets and should be considered when setting the running parameters. DnoisE is written in Python3, one of the most popular languages, so it is a good option for users who want to modify or customize the code. We indeed encourage new developments of this software.

We consider that DnoisE is a good option to denoise metabarcoding sequence datasets from all kinds of markers, but especially for coding genes, given the entropy differences of codon positions. More details, sample files and complete instructions are available at GitHub (https://github.com/adriantich/DnoisE).

Chapter 8

Metabarcoding reveals high-resolution biogeographic and metaphylogeographic patterns through marine barriers

8.1 Abstract

Aim The marine environment features oceanographic barriers that affect both the distribution and the connectivity of the marine biota. Biogeography can be extended by phylogeography, which analyses the distribution of genetic diversity within species. Metabarcoding can represent a leap forward in our ability to assess biogeographic and phylogeographic patterns, as it allows us to study many species at a time, including the often neglected small meio- and micro-organisms.

Location We tested the utility of the metabarcoding approach in one key biogeographic area, the Atlanto-Mediterranean transition along the E Iberian coast. This transition is marked by two barriers, the Almeria-Oran Front (AOF) and the Ibiza Channel (IC).

Time period Present

Major taxa studied Eukaryotes

Methods We sampled shallow hard-bottom communities at 12 sites over the littoral and performed community DNA metabarcoding using the cytochrome oxidase I (COI) marker. The resulting dataset was analysed at several levels: beta diversity of MOTUs (Molecular Operational Taxonomic

Units, surrogate for species) and ESVs (Exact Sequence Variants, surrogate for haplotypes), and genetic differentiation within MOTUs (metaphylogeography).

Results In a context of high differentiation and isolation by distance, we nevertheless found a strong effect of the AOF at all levels, which marks the main boundary between the Atlantic and Mediterranean waters. The IC had a comparatively minor role. With the MOTU dataset we obtained more clear cut patterns than with ESVs, and we discourage the use of the latter as the unit of biogeographic analyses. On the other hand, the metaphylogeographic approach provided the highest resolution in terms of differentiating localities and identifying geographic barriers.

Main conclusions Metabarcoding coupled with metaphylogeography provides a new tool to integrate the simultaneous analysis of beta diversity and genetic differentiation, unlocking a vast amount of information on the geographic distribution of biodiversity for basic and applied research.

8.2 Introduction

The marine environment, despite its apparent continuity, has physical and oceanographic barriers that determine the distribution of the different biota. The study of marine biogeography is a well-established field, and different regions and provinces have been proposed over the years, from Ekman's seminal review (Ekman, 1953) to more recent accounts (e.g., Briggs, 1995; Longhurst, 1998; Spalding et al., 2007; Toonen et al., 2016). These regions are usually defined by species turnover or changes in species abundances (beta-diversity) concomitant with geographic and oceanographic features. In addition, the advent of genetic techniques added a new component to the study of marine biogeography, thus giving rise to the field of phylogeography (Avise, 2009; Avise et al., 1987) which sought to assess how the present-day distribution of genetic diversity within species was reached (Riddle et al., 2008; Vellend et al., 2014). Barriers may be reflected, not just in species change, but also in genetic divergence within species due to restricted connectivity coupled with drift/selection. Delimiting homoge-

neous biogeographic regions has relevance for management, marine reserves' delimitation, evolutionary approaches, and socio-economic issues (Costello et al., 2017; Thiel et al., 2007).

Biogeographic breaks have been commonly studied on particular taxa, while studies with broad taxonomic coverage are rarer. Costello et al. (2017) provided the most comprehensive analyses of marine realms by compiling data from 65,000 marine species from public databases. Likewise, phylogeographic studies have usually addressed one species at a time, with few instances encompassing up to tens of species (e.g., Ayre et al., 2009; Cahill et al., 2017; Haye et al., 2014; Kelly and Palumbi, 2010) or reviewing available information from multiple groups (e.g., Hardy et al., 2011; Pascual et al., 2017; Patarnello et al., 2007; Teske et al., 2011). Most often, however, biogeographic and phylogeographic studies concern macro-organismal components of biodiversity, while the small meio- and micro-eukaryotes have been comparatively neglected, in spite of their importance and evidence of genetic breaks in them (e.g., Derycke et al., 2008; Tulchinsky et al., 2012). It is crucial to analyse patterns across macro- and micro-organisms to determine underpinning processes (Shade et al., 2018).

The rise of metabarcoding techniques during the last decade provided a new tool for assessing marine diversity in an integrative way, encompassing thousands of organisms (so-called MOTUs, or Molecular Operational Taxonomic Units), and including from micro- to macro-organisms to efficiently detect biodiversity patterns and processes. Metabarcoding has become an invaluable tool for biomonitoring, impact assessment, and detection of introduced species, among others (reviewed in Bowers et al., 2021; Cordier and Pawlowski, 2018; Deiner et al., 2017; Miya, 2022; Pawlowski et al., 2022; Rodríguez-Ezpeleta et al., 2021). Likewise, metabarcoding datasets using highly variable markers can be mined for intraspecies genetic diversity (Adams et al., 2019; Andújar et al., 2022; Elbrecht et al., 2018b; Sigsgaard et al., 2020) thereby opening the field for multispecies phylogeography (metaphylogeography, Turon et al., 2020). For metaphylogeographic analysis, stringent denoising of sequences to eliminate errors is necessary, generating Exact Sequence Variants (ESVs, e.g., Andújar et al., 2021; Antich et al.,

2022; Callahan et al., 2016; Edgar, 2016).

Coastal communities are among the most diverse marine habitats (Agardy et al., 2005; Reaka-Kudla, 1997) and benthic species are likely to be more affected by oceanographic discontinuities than pelagic species (Costello et al., 2017). The study of the benthos, therefore, is a powerful tool to assess biogeographic breaks. As in other environments, metabarcoding has boosted our ability to assess biodiversity in benthic communities, where most studies have been performed on soft-bottoms (e.g., Atienza et al., 2020; Brannock et al., 2018; Fonseca et al., 2017; Guardiola et al., 2016), with comparatively less work on rocky substrates, which are analysed either deploying artificial settlement units Atienza et al., 2020; Brannock et al., 2018; Fonseca et al., 2017; Guardiola et al., 2016 or by collecting samples directly from the natural communities (Shum et al., 2019; Wangensteen et al., 2018b).

Metabarcoding has been commonly used for community analysis, but it has seldom been applied to the formal assessment of biogeographic breaks in coastal areas (Gaither et al., 2022). Some instances focused on particular groups of organisms (e.g., Pagenkopp Lohan et al., 2017, protists; Santoferrara et al., 2018, ciliates; Closek et al., 2019, Czachur et al., 2021, vertebrates; Pitz et al., 2020, zooplankton), while other studies encompassed several groups (Cahill et al., 2018; DiBattista et al., 2022) or even across-kingdom comparisons(Holman et al., 2021). In all cases so far, however, these contributions were based on alpha- and beta-diversity changes. However, metabarcoding has the potential to uncover not only turnover rates and abundance changes of biotic components, but also to detect phylogeographic patterns of many species simultaneously as related to biogeographic breaks. Although it has been suggested that ESVs should be the unit of study instead of MOTUs for ribosomal markers (Callahan et al., 2017), for markers with a high intraspecies variability such as the cytochrome oxidase I (COI) gene, this can lead to an overestimation of the alpha and beta biodiversity and to interpret as biogeographic breaks what in fact are phylogeographic discontinuities. The combined use of MOTUs (as surrogate of species) and ESVs (as surrogate of haplotypes) allows to extract both

biogeographic and phylogeographic patterns (Antich et al., 2021a; Brandt et al., 2021; Turon et al., 2020), thus widening the scope of biogeographic studies for the assessment of marine discontinuities.

The Mediterranean is a well-known sea from the point of view of oceanographic features and biogeographic regions (Bianchi, 2007; Bianchi and Morri, 2000). The Atlanto-Mediterranean transition is one of the most important biogeographic boundaries worldwide. Albeit the geographic border lies in the Gibraltar Strait, the main barrier is considered to be eastwards in the nearby Almería-Oran Front (AOF) (Folkard et al., 1994; Tintore et al., 1988), a density front where the inflowing Atlantic water is deflected southeastward. The AOF poses an effective limitation for the dispersion of marine organisms, and it is the effective genetic barrier between both seas for diverse groups of organisms (e.g., Carreras et al., 2020; Naciri et al., 1999; Patarnello et al., 2007. he westernmost Mediterranean Sea features a sharp transition from Atlantic to Mediterranean waters, both along the N African coast and along the Iberian Peninsula. In this work we apply metabarcoding to characterise the biotic component of hard bottom benthic communities along the Iberian Mediterranean coast. We seek to analyse previously defined biogeographic breaks in these important transitional waters using a multilevel approach encompassing beta diversity analysis using both MOTUs (species level) and ESVs (haplotype level), and phylogeographic structures within MOTUs. Our aim is to test the potential of the metabarcoding approach to capture biogeographic and metaphylogeographic patterns across established oceanographic breaks.

8.3 Material and Methods

8.3.1 Sampling sites

We collected samples from 12 localities along the Mediterranean coast of the Iberian Peninsula. From South to North: Tarifa (TAR), Costa del Sol (SOL), La Herradura (LHE), Granada coast (GRA), Carboneras (CAR), Azohia (AZO), Cape Palos (PAL), Villajoyosa (JOY), Cullera (CLL), Calafat (CAL), Tossa de Mar (TOS) and Roses (ROS). These localities

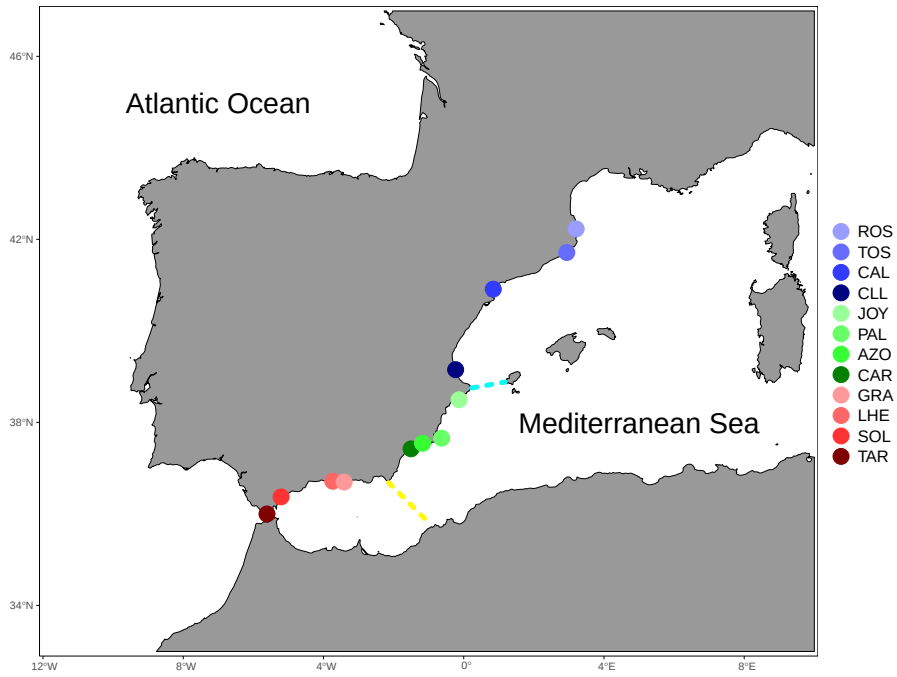


Figure 8.1: Map of the Iberian Mediterranean coast with the sampling localities and the two fronts studied: Ibiza Channel (IC, light blue) and Almeria Oran Front (AOF, yellow).

encompass two well-known oceanographic discontinuities: the Almeria-Oran front (AOF), between GRA and CAR (Folkard et al., 1994; L'Helguen et al., 2002; Tintore et al., 1988), commonly considered the true boundary between the Atlantic and the Mediterranean, and the Ibiza Channel (IC), between JOY and CLL (Bouffard et al., 2010; Pinot et al., 2002).

Accordingly, we grouped locations into three regions separated by these potential barriers: southern (TAR, SOL, LHE and GRA), central (CAR, AZO, PAL and JOY) and northern (CLL, CAL, TOS and ROS) regions (Fig. 8.1 and Tab. D.1).

8.3.2 Sample collection and laboratory procedures

We targeted the eukaryote component of the photophilous community found between 4 and 8 m depth in subvertical rocky walls. These communities are dominated by seaweeds with a highly diverse understorey of macro- and

meio-organisms. Sampling and laboratory processing were performed as described in Wangensteen et al. (2018b). In short, three sample replicates per locality were collected by scraping to bare rocky quadrats of 25 × 25 cm using a hammer and chisel. The material was collected in zip bags underwater, fixed with 95% ethanol within the hour, and stored at -20ºC. Sample processing included a size fractionation step in two sizes, large (L, >1mm) and small (S, between 1mm and 63μm) using stainless steel sieves. The two fractions were then homogenised separately with a blender, and 10 g of each were used for DNA extraction with the DNeasy PowerMax Soil Kit (Qiagen). Our initial dataset had thus a total of 72 samples (2 fractions × 3 replicates × 12 localities). All laboratory hardware was rinsed and bleached between samples. Negative controls were prepared by processing charred sand (Wangensteen and Turon, 2017) instead of actual samples.

A fragment of the COI mitochondrial gene (Leray fragment) was amplified with the degenerated primer set Leray-XT from (Wangensteen et al., 2018b) with PCR conditions as indicated in that work. Amplification blanks were obtained using the PCR mix without addition of DNA template. Primers were tagged (as in Wangensteen2018) to allow sample demultiplexing after sequencing. Library preparation was done with the BIOO NEXTFLEX PCR-Free DNA-Seq Kit (Perkin-Elmer) and sequencing was performed in an Illumina MiSeq V3 run with 2 × 250 bp paired-ends.

8.3.3 Bioinformatics pipeline

We processed the sequencing reads following a pipeline based on the OBITools package (Boyer et al., 2016). Illuminapairedend was used to align paired-end reads keeping only those with >40 quality score. Reads were demultiplexed using ngsfilter. Those with mismatched primer tags at any end were discarded. Obigrep and obiuniq were used to perform a length filter (retaining reads 299-320 bp long) and dereplicate sequences. Uchime-denovo algorithm from VSEARCH (Rognes et al., 2016) was used to remove chimeric amplicons.

The downstream processing included clustering sequences into MOTUs with SWARM (with $d = 13$ following Antich et al., 2021a). We

removed all MOTUs with less than 5 reads and used ecotag for taxonomic assignment against a local reference database, which is available at https://github.com/uit-metabarcoding/DUFA/ and contains 185,015 COI sequences. We then ran LULU (Frøslev et al., 2017) to remove potentially erroneous MOTUs and manually filtered the MOTU dataset to retain only the marine eukaryotes.

We then generated a sequence table for each MOTU using the output information of SWARM that contains a list of all sequences clustered in each MOTU. We denoised the sequences within each MOTU using DnoisE (Antich et al., 2022) to generate a table of exact sequence variants (ESV, Antich et al., 2021a) for each retained MOTU. DnoisE takes into account the natural variability (measured as entropy values) of each codon position for coding genes (such as COI) to improve the denoising algorithm. The entropy values (0.4812, 0.2407, 1.0285 for the first, second, and third codon position, respectively) were obtained from the whole dataset before clustering using DnoisE. The stringency parameter (alpha) was set to 4 following Antich et al. (2022). Final filtering steps were as follows: i) we removed any ESV for which the abundance in the blanks or negative controls was higher than 10% of its total read abundance; ii) in each sample, we applied a minimum relative abundance threshold, setting to zero the reads of any ESV with abundance below 0.005% of the total reads of this sample (this was done to eliminate tag-switching between samples); iii) we eliminated all remaining ESVs with <5 total reads; iv) from the whole ESV table we removed sequences with lengths deemed as incorrect: as for most species the length of the fragment used is 313, a correct sequence is expected to have $313 \pm 3 \cdot n$, being n the number of codons added or removed in indels; v) we finally removed sequences with stop codons and (for Metazoans) sequences with changes in conserved amino acids, since they probably arise from NUMTs, as described in Turon et al. (2020). The relative read abundances of each ESV in the two fractions of each sample were averaged for downstream analyses.

After these filtering steps, we obtained a dataset of MOTUs with taxonomic information and a dataset of ESVs (including all ESVs of all MOTUs). This allowed us to perform analyses at both levels: MOTUs (as surro-

gate of species) and ESVs (as surrogate of haplotypes) using relative read abundances as the analysed variable. Rarefaction curves and sample accumulation plots for both datasets were done using rarecurve and specaccum functions from the R package vegan (Oksanen et al., 2019).

8.3.4 Metaphylogeography dataset

The ESVs obtained in the previous analysis can be used to construct haplotype tables for phylogeographic inference for each MOTU (Antich et al., 2021a). To be able to capture potential patterns we selected only MOTUs that were present in at least two localities of two adjacent regions and with at least two ESVs each. As a proxy for haplotype abundances, ESV read abundances were converted to semi-quantitative values (following Turon et al., 2020): for each MOTU all ESVs were sorted in each sample in order of increasing abundance and ranked from 0 to 4 following percentiles of the ordered distribution: rank 0 for sequences with 0 reads; rank 1, for sequences that fell below the 51 percentile of the distribution; rank 2, sequences in percentiles $>50 \leq 75$; rank 3, sequences in percentiles $>75 \leq 90$; rank 4, sequences in the top >90 percentiles. The fractions of the same sample were ranked separately and then averaged to obtain the final semiquantitative abundance of each sequence in each sample.

8.3.5 Analyses

To assess community composition, MOTUs and ESVs were grouped into taxonomic super-groups (as in Guardiola et al., 2015) and, for metazoans, into phyla.

For biogeographic inference, Bray-Curtis (BC, with four-root transformation of relative read abundance per sample) and Jaccard (with presence-absence data) dissimilarities between samples were calculated using either the MOTU and the ESV dataset. These dissimilarities were used to ordinate samples in non-metric multidimensional scaling (nmMDS) using the metaMDS function from vegan package.

For the analysis of metaphylogeographic patterns, we computed a genetic

differentiation matrix using the D estimator (Jost, 2008) with the function pairwise_D() from the mmod R package (Winter et al., 2017). D values ranged from 0 to 1 (maximal dissimilarity). D values were obtained for each MOTU selected for phylogeographic analysis (see above) by performing pairwise comparisons of all samples in which the MOTU was present. Finally, for each pair of samples, the average D values across all shared MOTUs was computed and used to construct a genetic dissimilarity matrix. This matrix was used to create a nmMDS, and a network analysis with EDENetworks (Kivelä et al., 2015). For the latter, we used the mean D values across all-shared MOTUs among localities as the dissimilarity matrix and then transformed them into a network. The program automatically computes the percolation threshold (at which the all-including network breaks down into its main components), and we plotted the network just below this threshold. Finally, we plotted haplotype networks for all selected MOTUs, using the function haplonet of the R package pegas (Paradis, 2010).

Mantel tests were performed with the three dissimilarity measures (BC for MOTUs and ESVs, D for genetic differentiation) and the logarithm of the shortest distances by sea among localities with the mantel function of the vegan package. These analyses were repeated separately for the localities within each of the three regions defined. As localities separated by fronts tended to be also more distant geographically, to disentangle the effects of geographic distance from those of the fronts, the different dissimilarities between adjacent localities were calculated to assess whether there is a peak in dissimilarity associated with the transition between fronts.

We also assessed the pattern of distributional breaks of MOTUs and ESVs in adjacent localities using a randomization approach (partly based on Arranz et al., 2021). For each pair of adjacent communities, the number of breaks (defined as the number of MOTUs or ESVs present at only one of the two localities) was assessed and compared with the number found when the matrix of presence-absence was randomised across samples independently for each MOTU or ESV, thus effectively removing any geographic structure. This process was repeated 10,000 times and the distribution of breaks was determined and compared with the observed value. The variable used was

number of breaks / number of MOTUs (ESVs) present at the two localities being compared, and significance was assessed when the observed value fell outside the generated distribution or at its extremes (using a two-tailed test with Bonferroni correction).

To further separate the effect of differentiation among localities and of potential breaks, we performed permutational analysis of variance (PERMANOVA) on the three dissimilarity matrices. We compared adjacent regions (South-Center and Center-North) using region and locality (nested within region) as factors. In this way the effect of the two discontinuities could be assessed once the contribution of differences between localities was factored out. The PERMANOVA module incorporated in the Primer v6 statistical package (Anderson et al., 2008) was used. Tests of multivariate dispersions (permdisp) were run when the main factors were significant to determine whether this outcome was a result of different multivariate means or different heterogeneity (spread) of the groups. A second PERMANOVA, followed by permutational pair-wise tests, was run with just the locality factor (12 levels) on the three dissimilarity matrices to assess the degree of differentiation between localities.

8.4 Results

We obtained 16,096,788 reads comprising 4,149,955 unique COI sequences after demultiplexing, quality filtering and chimera removal. The original raw sequences have been deposited in the NCBI SRA archive (accession numbers pending)

All sequences were clustered with SWARM followed by LULU, resulting in 257,719 MOTUs of which only 17,944 had 5 or more reads. We filtered taxonomically all MOTUs to retain only those assigned to marine eukaryotes, 8,696 MOTUs in total. We then obtained the ESVs using DnoisE within MOTUs. After all filtering steps we retained 18,026 ESVs, 3,392 MOTUS and 9,423,471 reads. The list of MOTUs and ESVs is provided as Data D.1.1, and the taxonomic assignment of MOTUs in Data D.1.2. As per sample, we had 588 ± 20 (mean±SE) ESVs, 263 ± 10 MOTUs and $130{,}882 \pm 6{,}138$ reads.

At the locality level (combining samples), there is a significant correlation between the number of MOTUs and the number of ESVs (Pearson's r = 0.641, p = 0.025). From the retained MOTUs only 339 (indicated in Data D.1.2) fulfilled the conditions to be used for metaphylogeographic analyses. They had a median of 11 haplotypes (ESVs) each, with 3 and 60 as 10% and 90% percentiles, respectively.

The rarefaction curves showed that all samples reached an asymptote (Fig. D.0.1). However, for the species accumulation curve no clear plateau was reached as more samples were added (Fig. D.0.2).

8.4.1 Community composition

Metazoans were the dominant group in all localities both in number of MOTUs and ESVs (Fig. D.0.3). They were also the most abundant in relative number of reads, except in JOY, where Rhodophyta were dominant (Fig. 8.2). The latter group was the second most abundant in relative read abundance in all other localities except in TAR where Stramenopiles was the second group (Fig. 8.2a). For metazoans (Fig. 8.2b), a similar distribution in the number of reads across samples was found, with Porifera, Annelida, Arthropoda and Mollusca being the most abundant groups. MOTU composition across localities was homogeneous at the phylum level, but the composition in terms of ESV was more variable (Fig. D.0.3). The abundance of unidentified metazoans was higher in ROS and AZO, (over 30% of metazoan reads unassigned).

8.4.2 Biogeography

We computed non-metric Multidimensional Scaling using the Bray-Curtis (BC) (Fig. 8.3) and Jaccard dissimilarities (Fig. D.0.4) to obtain a reduced space representation of the samples for MOTUs and ESVs. BC and Jaccard dissimilarities distances provided highly congruent results. In general, the different localities appeared well separated, with no overlap of the inertia ellipses in the nmMDS plots of MOTUs for both BC and Jaccard indices, while some overlap was found for ESV data between TOS, CAL and AZO. A geographic distribution was apparent, with a differentiation of the southern

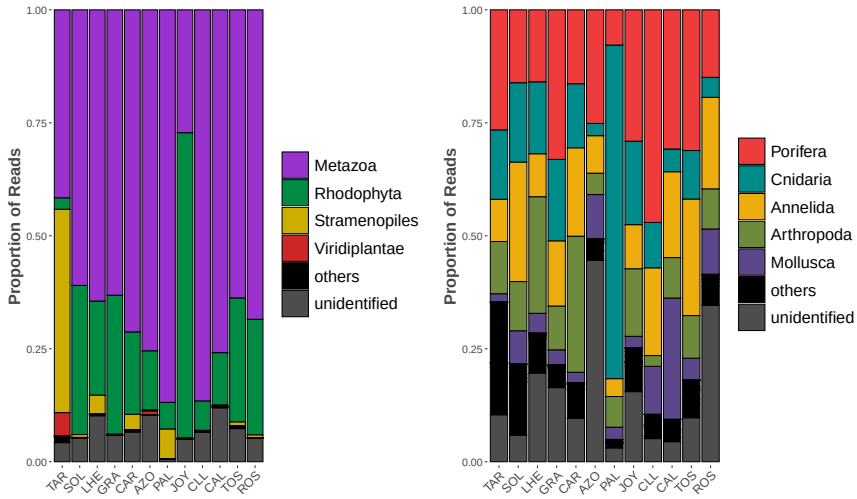


Figure 8.2: Supergroup (a) and metazoan phyla (b) composition in relative read abundance for each locality.

region from the other two along the first axis. The central and northern region did not form clearly separated clusters for MOTUs, and even less so for ESV data.

PERMANOVA analyses (Table 8.1) of Bray-Curtis dissimilarities showed for MOTUs a significant effect of the differentiation between southern and central regions, and not between central and northern regions. For the ESVs, no significant differentiation associated to regions was detected. In all cases, the nested locality factor explained most of the variation and was highly significant ($p<0.001$). No dispersion differences were detected for levels of significant factors (permdist tests).

PERMANOVAs for the locality factor alone were highly significant for both MOTUs and ESVs ($p<0.001$, while permdisp tests were not significant), and pairwise tests revealed that all pairs of localities were significantly differentiated in the MOTU dataset (with the exception of the two northernmost localities, TOS and ROS), while for the ESV dataset the following pairs of populations were not significantly different: TAR-SOL, SOL-LHE, SOL-GRA, GRA-AZO, AZO-JOY, AZO-CLL, TOS-ROS.

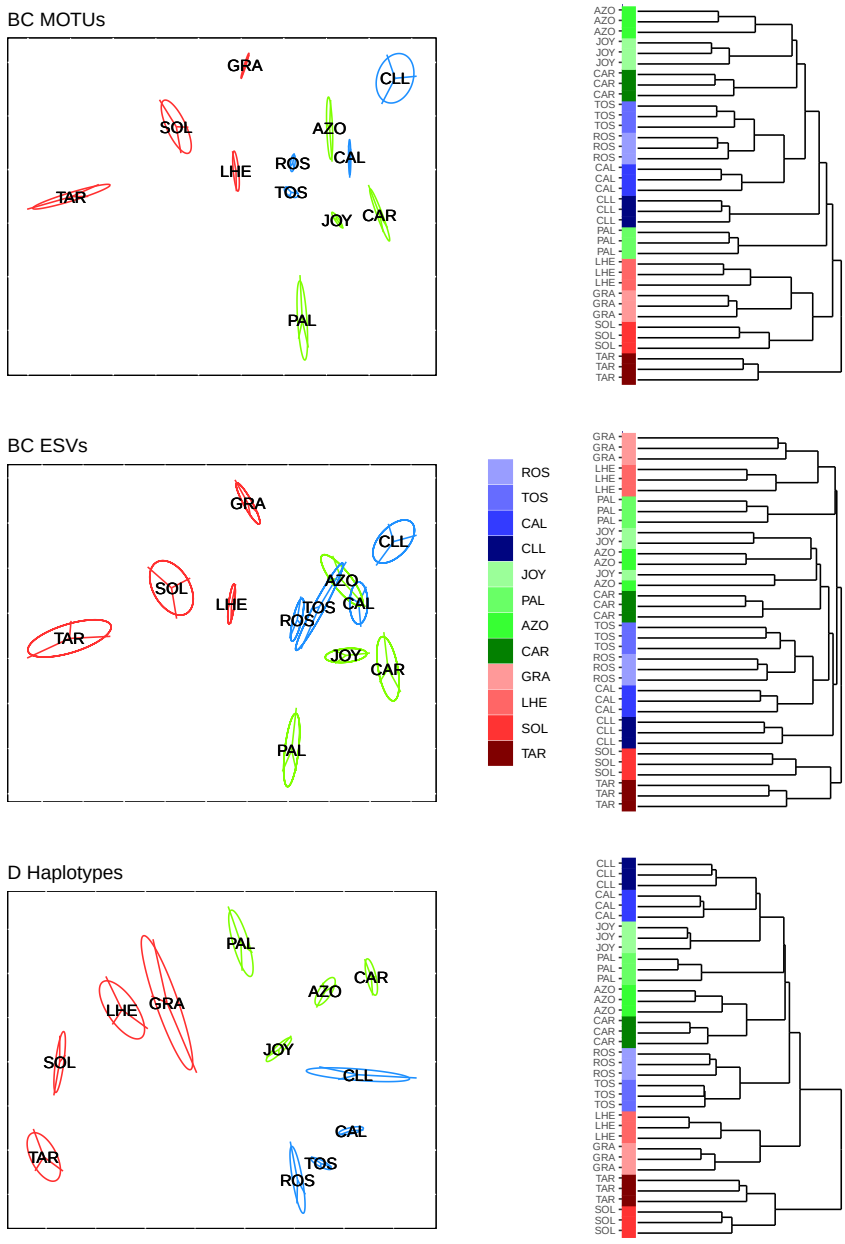


Figure 8.3: Non-metric Multidimensional Scalings (left) and clusters (right) of samples using Bray-Curtis dissimilarities for MOTUs and ESVs and mean D dissimilarities for haplotypes within MOTUs. Samples grouped by locality. Factor region is represented by colours (northern, blues; central, greens; southern, reds).

Table 8.1: PERMANOVA results obtained from the three distance matrices (MOTUs and ESVs based on Bray-Curtis distances and Jost's D distances) using adjacent Regions (southern vs central and central vs northern) and Locality (nested in Region) as factors. The p-values of the permutational analysis of multivariate dispersions (permdisp) are also given for significant factors.

MOTUs	**df**	**SS**	**pseudo-F**	**p-value**	**permdisp p**
Region (S vs C)	1	1.361	2.190	0.036	0.191
Locality(Region)	6	3.728	5.543	<0.001	0.126
Residuals	16	1.793			
Region (C vs N)	1	0.911	1.586	0.129	
Locality(Region)	6	3.448	6.360	<0.001	0.645
Residuals	16	1.446			
ESVs	**df**	**SS**	**pseudo-F**	**p-value**	**permdisp p**
Region (S vs C)	1	1.255	1.647	0.088	
Locality(Region)	6	4.572	3.596	<0.001	0.352
Residuals	16	3.390			
Region (C vs N)	1	1.080	1.493	0.116	
Locality(Region)	6	4.340	3.674	<0.001	0.786
Residuals	16	3.150			
D	**df**	**SS**	**pseudo-F**	**p-value**	**permdisp p**
Region (S vs C)	1	0.818	3.071	0.011	0.137
Locality(Region)	6	1.599	5.978	<0.001	0.002
Residuals	16	0.713			
Region (C vs N)	1	0.584	2.386	0.033	0.536
Locality(Region)	6	1.469	6.911	<0.001	0.046
Residuals	16	0.567			

The values of BC and Jaccard dissimilarities were higher for the analysis of ESVs (means of 0.893 and 0.942, respectively) compared to MOTUs (means of 0.757 and 0.841), which was expectable as localities should share less ESVs than MOTUs. For both MOTUs and ESVs, using the Jaccard distance the low number of shared taxonomic units between samples of different localities was evident, especially for TAR (which had the highest BC and Jaccard dissimilarities with other localities (Fig. D.0.4). TOS and ROS had the lowest values of both dissimilarities (Fig. 8.3 and Fig. D.0.4). Overall, BC values for MOTUs spanned a wider range of values (inter-locality comparisons, from 0.472 to 0.946) than for ESVs (from 0.665 to 0.988).

When comparing dissimilarity values from adjacent localities (Fig. 8.4 and Fig. D.0.5), the transition associated with the Almeria-Oran Front (GRA-CAR) had the highest mean values both for MOTUs and ESVs, followed by the comparisons between TAR and SOL and PAL and AZO. The Ibiza Channel (JOY-CLL) came next, with relatively high values but lower than some intra-region comparisons. The same results were obtained using the Jaccard distance (not shown).

The study of distributional breaks through permutation (Fig. 8.5) showed that most transitions had significantly more breaks than expected if breaks were randomly distributed, again revealing a strong structure between populations. However, for MOTUs the two highest values were found in the AOF transition (GRA-CAR, 26% more breaks than expected) and the IC transition (JOY-CLL, 22%), while there were significantly less breaks than expected between the two northernmost localities (TOS-ROS). For ESVs, the values were in general lower, with again the highest deviation (11%) from random expectation in the AOF transition, while the IC did not show any increase in associated breaks.

8.4.3 Metaphylogeography

A total of 339 MOTUs were selected for the metaphylogeographic analysis. Of these, 160 were found in at least 2 localities of each region, of which 12 were found in all localities. Of the 339 MOTUs, 85 were tagged with a

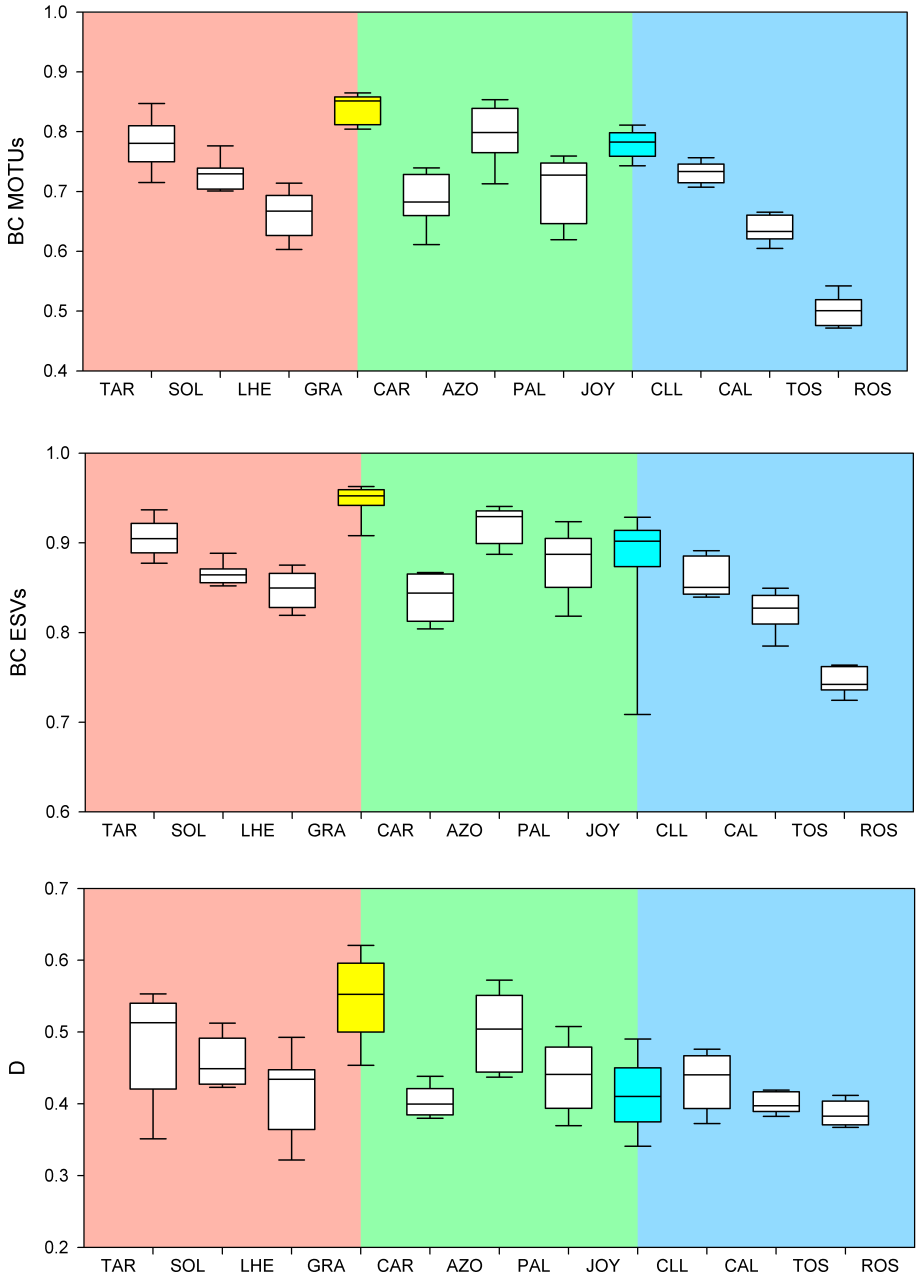


Figure 8.4: Bray-Curtis dissimilarities of MOTUs (top) and ESVs (central) and average D dissimilarities of haplotypes (bottom) from adjacent localities. Fronts are represented in yellow for the AOF and light blue for IC. Background colour corresponds to regions (red: northern, green: central, blue: southern).

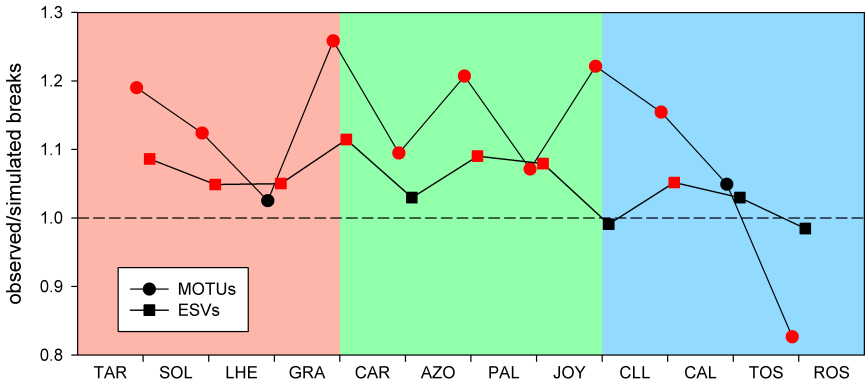


Figure 8.5: Number of observed vs simulated (through randomization) breaks between each pair of adjacent localities for MOTUs and ESVs. In red are transitions with significantly more (or less) breaks than expected. Background colour corresponds to regions (red: northern, green: central, blue: southern).

species name, 8 had genus and 13 family assignments, and the remaining 233 MOTUs were assigned at order or higher taxonomic rank. 240 of the MOTUs were Metazoa, 53 Rhodophyta, 10 Stramenopiles, 3 Viridiplantae, 2 Alveolata and the remaining 31 were unassigned eukaryotes. The best represented metazoan phylum was Annelida (53 MOTUs), followed by Arthropoda (51 MOTUs), Cnidaria (28 MOTUs), Porifera (20 MOTUs) and Mollusca (17 MOTUs). Haplotype networks for these 339 MOTUs are presented in Data D.1.3. They show a variety of patterns, but are predominantly star-shaped with a few dominant haplotypes. The latter are in general shared among regions, albeit with different proportions.

We computed a dissimilarity matrix of 36x36 (3 samples x 12 localities) with the average D values for each pair of samples computed from the shared MOTUs. These values were used to map samples in a nmMDS (Fig. 8.3) that showed a sharp separation between localities, with no overlap of the inertia ellipses. The first axis separated the southern region from the other two, which in turn formed distinct clusters along the second dimension. PERMANOVA analyses showed a significant effect of the Region factor, in both the differentiation between southern and central, and between central and northern regions ($p = 0.011$ and $p = 0.033$, respectively). The nested

locality factor again explained most of the variance and was highly significant ($p<0.001$) with significant differences also in dispersion levels (permdisp tests, Table 8.1).The analysis of the factor locality as the main factor showed a highly significant effect ($p<0.001$) but also a significant difference in dispersion values (permdisp $p = 0.003$). All pairwise comparisons were significant except between the two southernmost localities TAR and SOL.

The analysis of D dissimilarities from adjacent localities (Fig. 8.4) showed that GRA and CAR (corresponding to the AOF) had the highest average differentiation, followed by TAR and SOL and PAL and AZO. The lowest differentiation between adjacent localities was found in the northern region (TOS and ROS followed by CAL and TOS). No clear differentiation was detected associated with the IC break.

The network analysis using EDENetworks detected the percolation threshold at a D value of 0.51. The network obtained just below this threshold ($D = 0.50$, Fig. 8.6) showed a separation between the southern region and the central and northern regions corresponding to the AOF. In turn, the central and northern regions were connected by a few weak links involving mostly the northernmost central region locality (JOY). Only the link between the two localities at both sides of the break, JOY and CLL, was relatively strong, which is consistent with the pattern shown in Fig. 8.4. The northern region showed strong internal links, particularly between CAL, TOS, and ROS. The node with the highest betweenness centrality (indicating its importance in connecting other nodes, Kivelä et al., 2015) is JOY, which also has the highest number of links.

The Mantel tests (Fig. D.0.6) showed that, for the three variables considered (BC dissimilarity between localities calculated with the MOTU and ESV, and genetic distance D) there was a highly significant correlation with geographic distance (all Mantel $r>0.803$, $p<0.001$). The same result was obtained within regions (all Mantel $r>0.748$, $p<0.001$), indicating a clear signal of isolation by distance.

Finally, we computed the relationship between the Bray-Curtis dissimilarities between localities for the MOTU and ESV datasets and also plotted

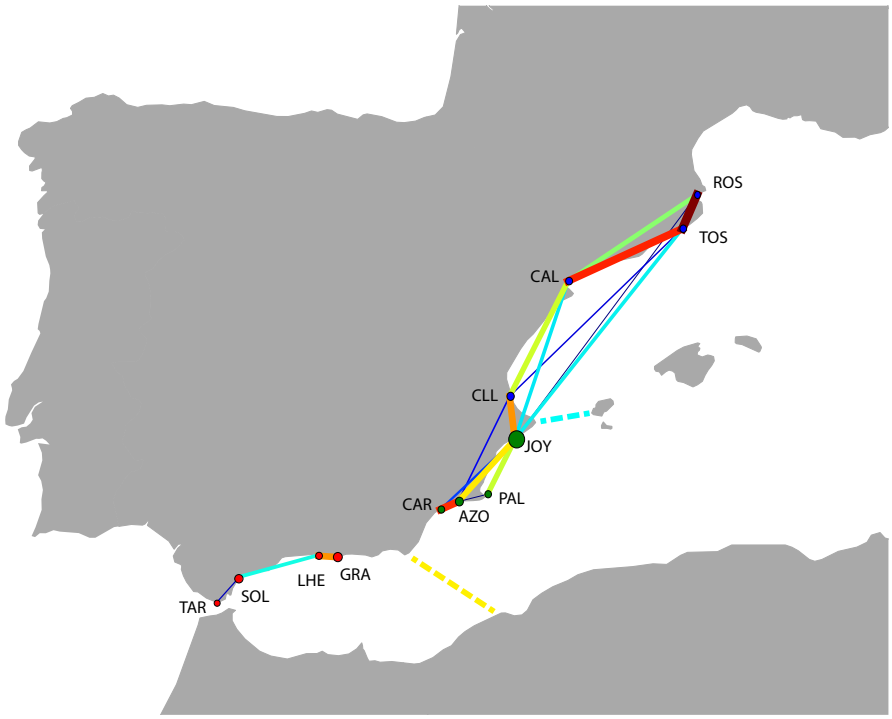


Figure 8.6: Network analysis using EDENetworks with D values. Wider and warmer colours represent stronger connections and thinner and colder colours represent weaker connections. The size of the locality symbols is proportional to the betweenness centrality of the nodes. The two breaks are represented in dashed lines; Almeria-Oral Front (AOF) in yellow and Ibiza Channel (IC) in light blue.

the D value of the MOTUs selected for the metaphylogeographic analysis (Fig. 8.7). The results showed a general good correlation of the three measures. Overall, pairwise differentiation values are higher for ESVs than for MOTUs, and the difference is reduced for highly differentiated localities (i.e., with values close to 1 for both datasets).

8.5 Discussion

Metabarcoding of highly diverse shallow benthic communities, using a broadly used mitochondrial marker (COI), retrieved both biological and genetic diversity from the Atlanto-Mediterranean transition along the eastern Iberian coast, marked by two known discontinuities. The present study is the first to explore the effects of barriers to gene flow in the marine realm

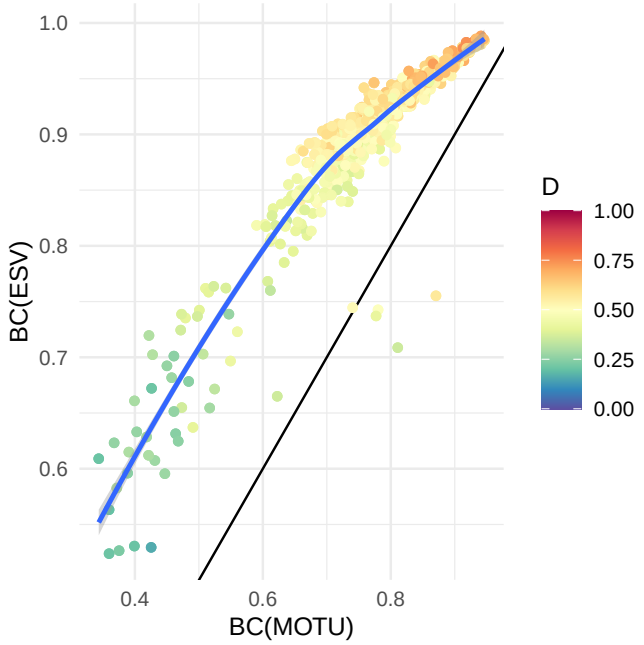


Figure 8.7: Relation between Bray-Curtis dissimilarities of MOTUs (x-axis) and ESV data (y-axis). D values computed for the selected MOTUs are represented by colours. Each point represents a pairwise comparison between samples and the blue line is the trend line obtained by the 'loess' method of ggplot2.

simultaneously with biogeographic patterns using metabarcoding data and encompassing different groups of eukaryotes. Both the biogeographic and the phylogeographic perspectives showed similar patterns of community differentiation but with different resolution. The different approaches reveal important information at several levels of biological organisation.

8.5.1 Biogeographic (MOTUs and ESVs diversity) patterns

Along the 1,200 km of the Iberian coast we retrieved a high diversity of taxa in all localities. Rarefaction curves showed an adequate number of reads per sample but more replicated samples seem necessary to capture all diversity present in such complex assemblages. Communities were dominated by metazoans in both number of MOTUs and relative read abundance, with Porifera, Cnidaria, Annelida and Arthropoda being the most abundant phyla. Across all samples, taxa composition was similar in terms of relative number

of MOTUs, but TAR showed a higher relative abundance of Stramenopiles than other localities. It had a community dominated by brown algae, unlike the other sites. PAL was also different in the read abundance of metazoan groups compared to other localities, with a higher abundance of Cnidaria. The benthic community in PAL visually differed from other locations, with low abundance of algae and high abundance of Anthozoa and Hydrozoa. Overall, about 25% of metazoan MOTUs did not match with any phyla, emphasising the importance of completing current reference databases (Mugnai et al., 2021; Wangensteen et al., 2018b). However, even if low-level taxonomic assignments are lacking, unassigned MOTUs can still be used to calculate diversity metrics.

In the present study three regions were considered, separated by two previously described fronts. We ordered samples of these areas in nmMDS plots using both MOTUs and ESVs. Localities from the southern region were well separated from those of central and northern regions. In addition, the localities of GRA and CAR, separated by the AOF, showed the highest values of BC dissimilarity of all comparisons of adjacent localities. This emphasises the importance of this barrier. The AOF is a geostrophic front that separates Atlantic waters entering through the Gibraltar Strait from Mediterranean waters, thus marking the main boundary in the Atlanto-Mediterranean transition (Folkard et al., 1994; L'Helguen et al., 2002; Tintore et al., 1988). However, although its role in the genetic structure of many species has been investigated (see, f.i., El Ayari et al., 2019; Pascual et al., 2017; Patarnello et al., 2007), there is to date no comprehensive analysis of its effect in species beta diversity. The role of this front is a clear-cut feature of our analyses. In fact, of the MOTUs present in the southern and central regions (separated by the AOF), 57.1% were present at only one side of this divide. In terms of ESVs, this figure was 81.8%. The IC was not as strong a break in terms of regional differentiation. The nMDS analyses showed an imperfect separation of the northern and central regions, and the BC dissimilarity of JOY and CLL (separated by the IC) was relatively high but smaller than values obtained from other comparisons of adjacent localities. Overall, dissimilarity values show that AOF is a strong biogeographic barrier but the IC is not and, thus, the central and northern regions are not well

differentiated in biotic composition.

We also note that TAR appeared separated from all other localities in our nmMDS analyses of both MOTUs and ESVs. As mentioned earlier, this community was the only one dominated by brown algae, but being located just on the Gibraltar Strait, this result could also indicate an effect of this geographic boundary. Unfortunately, our sampling scheme was not designed to test this effect, which should be considered in future studies. PAL was also somewhat offset from the other central region localities, which is attributable to a different community with less algae and higher invertebrate dominance.

While both descriptors (MOTUs and ESVs) provided basically the general pattern, there were nevertheless differences. The PERMANOVA analyses showed a significant differentiation between the southern and central localities with the MOTU dataset, which was not found with ESVs. Likewise, almost all pairwise comparisons between localities (PERMANOVA) revealed significant differences with MOTUs (except TOS-ROS), while seven comparisons were not significant with ESVs. This is also reflected in the overlap of some localities in the nmMDS analyses. When considering the pattern of distributional breaks between adjacent localities, in general more breaks than expected in a random simulation were found, supporting the idea that localities were well differentiated. However, values were higher for MOTUs than for ESVs, and for the former the AOF and IC transitions had the highest values, while IC was not significant for ESVs. The narrower range of dissimilarities obtained with ESVs (0.524 to 0.988) than with MOTUs (0.343 to 0.946) may have hampered resolution when using the ESV dataset.

The analysis of MOTU-level turnover is the metabarcording equivalent to the standard biogeographic species-level analysis. On the other hand, using ESVs instead is equivalent to analyse haplotype turnover, whose interpretation is unclear. In particular, if we miss the MOTU information we would be giving the same weight to a distributional change in a haplotype from a species with high genetic variability (many ESVs) than from a species with just a few haplotypes. Diversity calculations based on ESVs are thus driven and biased by species with high intraspecies variability. We

would be lumping together biogeographic (interspecific) and phylogeographic (intraspecific) information. On the contrary, if ESV information is grouped into meaningful biological units (MOTUs), then the well-developed and tested tools for analysis of geographic genetic differentiation can be applied in a metaphylogeographic context. We used in this work only basic analyses (population differentiation and network analyses), but the whole panoply of phylogeographic analytical tools (Knowles, 2004) can be applied depending on the question of interest.

Recent articles have discussed the relative merits of using MOTUs and/or ESVs for metabarcoding studies (Antich et al., 2021a; Brandt et al., 2021). These works emphasise that using ESVs as a standard unit of analysis, as suggested previously (Callahan et al., 2017), may be valid for ribosomal markers but not when studying eukaryotes and highly variable markers such as COI. In our view, in these cases diversity patterns must be studied using MOTUs as a proxy for species, while ESVs must be used within MOTUs as a proxy for haplotypes and intraspecies variability (metaphylogeographic approach), allowing a hierarchical analysis of the distribution of diversity.

8.5.2 Phylogeographic perspective

Geophysical barriers play a crucial role in population fragmentation even in apparently continuous marine environments. Phylogeography analyses the geographic distribution of genetic lineages, linking geography and genealogy, and has been developed since the eighties of the last century (Avise, 2009; Avise et al., 1987). Phylogeographic studies rely on species that are easy to sample, being therefore restricted in general to macro-organisms, commercially interesting species, or flagship iconic species. Small organisms are only rarely studied due to the difficulty of sampling individuals. There is therefore a lack of information on whether phylogeographic patterns of marine macro-organisms are coherent with those of meio- and micro-organisms.

From all the MOTUs that we found, only 10% could be used for genetic dissimilarity analysis. This is caused by a high number of low abundance MOTUs found in few samples compared to those selected for the analyses, which had a broader distribution. Our results show high values of genetic

dissimilarity when comparing samples from different localities, with almost all comparisons being significant in PERMANOVA analyses. However, dissimilarities between localities of the same region were smaller (0.470 ± 0.004) than those between regions (0.560 ± 0.002) meaning that gene flow is higher within than between regions. The three regions appeared well separated in nmMDS ordinations, and PERMANOVA analyses showed significant differences between southern and central and between central and northern regions. Furthermore, a network analysis reflected disconnected networks in the southern and the central plus northern regions. Among the latter, the links between regions were feeble with the exception of the edge between JOY and CLL. The JOY locality, on the other hand, had links with all other localities in the central and northern regions and the highest betweenness centrality in the whole network, thus constituting a hotspot for genetic connectivity in the area. If we perform the network analysis without JOY, the central and northern regions appear disconnected (results not shown). Rather than a clearcut divide, the network analysis indicated that the IC is placed in a transition zone connected to both sides.

Phylogeographic structure and species beta diversity are two complementary dimensions of integrative biogeography in a broad sense (Riddle et al., 2008). However, the former is much harder information to acquire. Phylogeographic marine breaks have been usually studied on a single species basis, sampling populations and analysing a set of genetic markers, depending on the study. Multispecies studies are rare and include only a handful of species (e.g., Haye et al., 2014; Kelly and Palumbi, 2010). Alternatively, meta-analyses of published data can be used to make inferences (Arranz et al., 2021, 2022; Dawson, 2014; Pascual et al., 2017). Metaphylogeography is a new way to study population genetic differentiation for the whole community using metabarcoding data. This new tool has the potential to detect subtle patterns and study genetic connectivity with a relatively low sampling effort and targeting a huge amount of taxa of any size. As pointed out by Zizka et al. (2020), the study of haplotypic diversity can provide crucial information on the state of the ecosystem and predict which populations are more sensitive to environmental changes. Moreover, the study of barriers affecting gene flow is mandatory to manage not only biodiversity but also

genetic diversity (Sandström et al., 2019), and metaphylogeography can become a key tool to achieve integrated management programs.

8.6 Conclusions

The simultaneous study of biogeographic and phylogeographic patterns captured important information at different levels of biological organisation. There was an overall pattern of high structure between localities and a significant relationship with geographic distance. Superimposed to this pattern, the Almeria-Orant Front (AOF) had a strong structuring effect in most analyses, confirming expectations. On the other hand, the Ibiza Channel (IC) barrier had a minor effect, detected only with the genetic differentiation analyses (metaphylogeography) and the distribution of breaks using MOTU data, but not with ESV data. The distribution of species can be determined by a broad range of biotic and abiotic factors, leading to differences in community composition. However, both isolation and local adaptation can have an effect not only on the species distribution but also determine shifts in haplotype frequencies within species. We do not favour the study of ESVs alone without the species (MOTU) context. The haplotypes are not independent units, they are distributed, adapt, and evolve encapsulated in biological units (species), which have biological, historical, and demographic traits that determine their haplotype richness and distribution. We therefore suggest using MOTUs as the unit for species turnover analysis and ESVs within MOTUs for phylogeographic analysis when using metabarcoding data.

Metabarcoding coupled with metaphylogeography provide a new tool to integrate the simultaneous analysis of species turnover and genetic differentiation, unlocking a vast amount of information on the geographic distribution of biodiversity for basic and applied research.

Chapter 9

General Discussion

Climate change and human impacts are affecting all environments, and particularly so marine ecosystems (Halpern et al., 2008). Biomonitoring and managing have become crucial for our society to understand how biological communities change and to preserve the biodiversity. "Healthy" environments, those less disturbed by human activities, are a sustainable source of indispensable goods for our civilization (Borja et al., 2020; Duarte, 2000). Rapid and efficient methodologies are demanded for a correct monitoring of these ecosystems (Borja and Elliott, 2018). Since the beginning of the millennium, molecular methods have gained popularity to analyse the biotic composition of the ecosystems, as a response to the time-consuming traditional methods, that require also a taxonomic expertise that is dwindling worldwide (Orr et al., 2020).

Metabarcoding, where a particular gene (or a combination of a few markers) is amplified from samples obtained by diverse methods, has been developed as a technique for inferring in a fast and repeatable way the biodiversity of natural communities (Beng and Corlett, 2020; Gaither et al., 2022). Our main goal in this Thesis has been to develop metabarcoding protocols for the analysis of complex, hard-substratum benthic communities, arguably among the most diverse on Earth (Helmuth et al., 2006). At sea, metabarcoding was first applied to the analysis of sediments or pelagic components (e.g., de Vargas et al., 2015; Guardiola et al., 2016). Hard-bottom communities had been studied indirectly, by deploying standardized settlement surfaces (Pearman et al., 2018). At the time the Thesis started, only Wangensteen and Turon (2017); Wangensteen et al. (2018b) had devised methods for the direct sampling and metabarcoding of natural hard-bottom communities and had designed improved universal primers. We wanted to capitalise on these previous studies to design a complete protocol for this

type of community based on the analysis of mitochondrial (COI) sequence data targeting the whole eukaryotic assemblage present, thus including from micro- to meio- and macro-organisms.

We were first interested in determining the best sampling method. In the last 10 years, the concept of "conservation in a cup of water" (Lodge et al., 2012, and the accompanying special volume of Molecular Ecology) has emphasized the idea that the so-called environmental DNA (eDNA) could serve as a multi-purpose tool for the analysis of all kinds of communities (Deiner et al., 2016, 2017). The term "environmental DNA" has been subject to recent debate (Nagler et al., 2022; Pawlowski et al., 2020; Rodriguez-Ezpeleta et al., 2021), and it is now acknowledged that most eDNA samples include both intra-organismal and extra-organismal (trace) DNA in variable proportion according to sampling method, primers, and target group. Most commonly, however, the term eDNA is used when the environment (water, sediment, soil) is sampled without isolating the organisms. When the biotic community is isolated from the environment before the analysis, the terms community DNA or bulk DNA are commonly used. A paramount application of eDNA in aquatic ecosystems consists of the filtration of water with sub-micron pore sizes to capture DNA from microorganisms, exudates (mucus, hairs, scales, faeces...) of macro-organisms, and free DNA. From this pool of DNA information about pelagic and, potentially, benthic communities can be obtained. However, the dynamics of eDNA in the benthic boundary layer were unexplored, and thus the possibility of assessing the biodiversity of benthic communities using DNA from the adjacent water had not been tested. We addressed these issues in Chapter 4 based on samples collected during this Thesis in the Cabrera Archipelago.

Moreover, while metabarcoding data had been used to obtain presence/absence information (or, at most, relative frequency), during this thesis we have evaluated the possibility to infer also intraspecies variability. We coined the term "metaphylogeography" to refer to the analysis of patterns of distribution of genetic variability for many taxa simultaneously, exploring the so far untapped reservoir of intraspecies information available in metabarcoding datasets. The use of this information required the

curation of sequences to distinguish sequencing errors from true sequences, for which we tested a combination of denoising and filtering procedures. Since all marine ecosystems are somehow connected, understanding how both species and genetic diversity co-distribute in different areas allows to understand community changes, species shifts, and genetic flow across all kinds of gradients and breaks. We developed the metaphylogeographic approach in Chapter 5 using a dataset previously published from Spanish National Parks with maritime domains.

In addition, since the technique of metabarcoding has its origins in microbiology and, particularly, the study of prokaryotes, the commonly used bioinformatic pipelines have been originally developed for the analysis of ribosomal genes and in a context of genetic variability different from what is found in eukaryotes. These programs have too often been uncritically adopted for eukaryotic metabarcoding, without proper parameter calibration (Elbrecht et al., 2018b). The problem is more acute when they are applied to highly variable genes such as COI, where the particular features of coding regions have been totally overlooked (Creedy et al., 2022). In Chapter 6, using samples collected in littoral communities of the Iberian Mediterranean, we have calibrated a bioinformatic pipeline tailored to eukaryotic marine communities assessed using COI metabarcoding. We have contributed to recent issues about the use of clustering (to delineate MOTUs) or denoising (to generate ESVs) to obtain the units for analysis (Brandt et al., 2021; Callahan et al., 2017), and showed that both are not just complementary, but also necessary for COI metabarcoding if both the inter- and intraspecies variability is targeted. Furthermore, we contend that, for coding genes, the different level of variation in the three codon positions provides an invaluable information that should be incorporated into the denoising programs. This led us to develop a new denoising method in the software DnoisE (Chapter 7), which has been made publicly available.

The previous steps allowed us to determine the best methods for metabarcoding the complex eukaryotic communities thriving on hard substratum benthic habitats. We sought to apply them in Chapter 8 (using the same samples as in Chapter 6) to determine the biogeographic and metaphylo-

geographic structure of littoral communities of the Atlanto-Mediterranean transition following the Iberian littoral from the Gibraltar Strait to the Cap de Creus.

9.1 Testing methodology

9.1.1 Sampling

Biomonitoring of marine environments using metabarcoding approaches has proved to be comparable, or even better in terms of time and information generated, to traditional methods (Aylagas et al., 2018). In the context of the Anthropocene era, biomonitoring is mandatory to assess the status of the ecosystems and to apply managing strategies properly (Borja and Elliott, 2018). However, traditional sampling methods to capture the whole diversity of living organisms are cumbersome and usually cause an impact on the ecosystem. Thus, the use of less invasive methods is advisable, especially to monitor the most fragile environments (Adams et al., 2019; Gilbey et al., 2021). The detection of the environmental DNA in aquatic ecosystems has proved to be useful to detect single species without capturing the whole organism (Ficetola et al., 2008). Yet the use of metabarcoding to characterize the community composition using the eDNA in the water has been restricted almost to freshwater environments and those studies in marine environments usually focused on the vertebrate community, especially fishes, using specific primers targeting these groups. The potential of using noninvasive water eDNA metabarcoding to assess biodiversity in marine shallow benthic communities is thus promising but had not been evaluated.

In the chapter 4 of this Thesis we performed an experiment comparing benthic samples obtained from scraping the rocky substrate and water samples from 0m to 20m from the benthic community to analyse the eDNA signal of the benthic community in the water. Our goal was to compare two commonly used methods in these two habitats, rather than trying to adapt them.

We acknowledge that the sampling methods are not the same for both sample types. The different fractioning of the DNA of the samples can reveal

different levels of information and have to be taken into account (Nagler et al., 2022). For the benthic samples the smaller sieve of 63μm favor the removal of most of prokaryotes and most of the retained eukaryotic DNA is encapsuled within cells. This is not the case in water samples where we used a 200 μm pre-filter and a 0.22 μm filter that retained most of the prokaryotic DNA and for which part of the eukaryotic DNA is expected to be extra-organismal. In addition the primers used, although they are designed for eukaryotes, amplify also prokaryotic DNA. Primer biases are among the major concerns in the use of PCR for metabarcoding studies (Beng and Corlett, 2020). In this context, results showed that only 14.35% of the total reads obtained from water samples were assigned to Eukaryotes while in benthic samples this percentage was 99.35% of the total reads.

Although the total MOTUs found was substantial, ca. 3,500, the proportion of those shared between benthic and water samples was low (11% and 19% of the MOTUs detected in benthos and in water, respectively). Conversely, in terms of relative read abundance, these MOTUs represented the majority of reads (70.40% in benthos and 56.37% in water). We also differentiated between two main groups of those MOTUs shared, the shared benthic MOTUs (SBM), which are more abundant in benthic than water samples, and the shared pelagic MOTUs (SPM), those for which the distribution of reads is the opposite. We acknowledge that this was a crude approach, but nevertheless allowed us a rough separation of the shared MOTUs of potential benthic (SBM) and potential pelagic (SPM) origin. The picture for each group was markedly contrasting, SBM represented a majority of the reads (ca. 70%) in the benthic samples but a very low percentage in the water (ca. 2%), and SPM constituted the 54.44% of the pelagic reads and 0.07% of the benthic reads. The results showed that the number of MOTUs shared between the benthos and the surrounding waters (rather low to start with) decreased with distance, especially those MOTUs of potential benthic origin (SBM). On the contrary, the number of SPM remained stable. This pattern indicated a decreasing detection of the DNA originating from benthic organisms even at very short distances (the first 50 cm) from the benthos, while for DNA of presumably pelagic origin the pattern was more homogeneous.

We therefore recommend the use of the direct sampling method by scraping the rocky benthos to capture its biodiversity composition if COI with universal primers are used. Unspecific amplification and the great proportion of non-eukaryotic DNA in the water filters resulted in a low percent of assignable reads in the water samples. Even considering the eukaryotic MOTUs detected, a very low percent of those found in benthos were detected in water and, from these, the numbers with potential benthic origin decreased sharply at very short distances from the rocky wall. It remains to be tested if increasing sequencing depth (as is achievable with the new sequencing platforms such as NovaSeq) could allow a more effective detection of the (presumably very rare) benthic DNA in the water. Alternatively, the use of water eDNA to detect benthic biodiversity can be feasible if specific primers (instead of universal ones) are used (Miya et al., 2020; Stoeckle et al., 2018), as this will reduce the unspecific amplifications, but of course in these cases only the target group will be detected, not the whole community that was our goal.

9.1.2 Bioinformatic pipeline

The raw metabarcoding datasets contain many erroneous sequences derived from errors originated during PCR amplification and subsequent sequencing. Some errors consist of chimeric sequences that can be spotted using several comparison algorithms (our preferred one was uchime_denovo from Rognes et al., 2016 based on the UCHIME algorithm of Edgar et al., 2011). Other errors are the result of tag-jumping, whereby a sequence is assigned to the wrong sample. Again, some filters can be set to detect and eliminate these sequences. Nuclear pseudogenes are another source of noise in metabarcoding datasets of mitochondrial sequences; they are easily detectable if they contain stop codons, but hardly so otherwise (Andújar et al., 2021; Schultz and Hebert, 2022). Finally, many errors persist as spurious sequences, and thus metabarcoding datasets require a reduction and cleaning. Two main procedures are used to reduce the size and eliminate erroneous sequences from the datasets. One of them is clustering, whereby similar sequences are joined together in Molecular Operational Taxonomic Units (MOTUs), and the most abundant sequence is taken as the representative of the MOTU. These units

should ideally approach the species-level composition of the sample. After clustering, the dataset is reduced to MOTUs, and the erroneous sequences are assumed to have been incorporated into the correct MOTU, so clustering eliminates erroneous sequences as a by-product (even if it is not its main purpose). Note, however, that if only the representative sequence is kept, all intra-MOTU information is lost after clustering (Hajibabaei et al., 2019). There are diverse algorithms for this step, ranging from the use of simple similarity thresholds, to phylogenetic approaches.

The second method is denoising, that tries to detect erroneous sequences and to add their reads to the correct ones from which they derive. The methods to detect spurious sequences depend on three principles: 1) Erroneus sequences (daughter sequences) are produced when copying or sequencing true sequences (mother sequences) and thus ideally for each erroneous daughter there is a correct mother, 2) Mothers and daughters are similar, and 3) the mother is more abundant than the daughter, as sequencing error rates are generally low. There are different ways to compute the likelihood of each sequence to be an error, and hence different denoising algorithms exist. The results are a set of correct sequences, here called ESVs (Exact Sequence Variants).

Clustering had been the method of choice in metabarcoding until recently, when denoising programs have become popular (Callahan et al., 2016; Edgar, 2016; Peng and Dorman, 2020). This has led to the proposal that ESVs, and not MOTUs, should be the unit of analysis for metabarcoding (Callahan et al., 2017). In recent years, clustering and denoising methods have been widely used separately (Creedy et al., 2022). These two strategies are philosophically and operationally different, and we endorse the idea that both are complementary and should be used together in COI metabarcoding. However, most of the published pipelines and the software available have been designed and tested for prokaryotes. Moreover, the use of coding barcodes such as COI has also gained popularity (Andújar et al., 2018; Duarte et al., 2021a), but most available software has been developed for ribosomal markers. All this required calibration and software modifications.

The high variability of COI both between and within species renders COI metabarcoding a potential tool to retrieve the species diversity but also intraspecies variability to perform population genetics analyses. In chapter 5 we introduced the concept of "metaphylogeography" as a method to study both inter and intra-specific patterns using metabarcoding data. The underlying idea is that if metabarcoding techniques can reliably detect the true sequences of highly variable markers, then these sequences can be used as a proxy of the haplotypes present in a given population and thus perform phylogeographic analyses with metabarcoding data. Chapter 5 is a proof-of-concept of the metaphylogeography approach. We used a combination of clustering with the algorithm SWARM (Mahé et al., 2021), which starts with a distance threshold (parameter d) and then uses the clusters' structure to refine them. Crucially, we did keep not only the representative sequence of each MOTU, but retrieved all sequences grouped in each cluster from the output of the program. We then denoised them using the UNOISE algorithm (Edgar, 2016), which is based on a simple, non-iterative formula that relates the abundance ratios with the number of changes to determine if a sequence is an error of another more abundant sequence. Finally, a filtering step was performed to eliminate low-abundance sequences that are assumed to be erroneous. The main problem in these steps is parameter calibration for real metabarcoding datasets with high complexity. Mock communities are too simple to be a realistic approximation to this type of datasets and we needed to rely on other features of the sequences. We used the coding properties of COI, namely the fact that the variability (entropy) of third codon positions is higher than in the first or second positions. We suggested the use of an entropy ratio (entropy of position 2/entropy of position 3), that was shown in simulations to increase as errors accumulate and to decrease and finally stabilize as errors were eliminated. Using the entropy ratio, we optimized the stringency parameter (alpha) for UNOISE and the minimal number of reads for filtering. We could perform metaphylogeographic analyses for 563 MOTUs and generated haplotype networks and analyses of molecular variance coherent with data published in the literature. Our results demonstrated the feasibility of mining metabarcoding datasets for the analysis of intraspecies genetic diversity using objective parameters for

denoising and filtering spurious sequences.

In Chapter 6 we analysed and compared pipelines performing the denoising and clustering steps in different orders and with different programs. For clustering, we empirically assessed the best value for parameter d of SWARM (Mahé et al., 2021) by tracking the number of clusters formed and the mean intra- and inter-MOTU distances. We found that for the fragment of COI analysed, d=13 was the best compromise to maximize the intra-MOTU variability while keeping a sharp difference between both values (equivalent to the barcode gap (Meyer and Paulay, 2005)).

For the denoising procedures, we compared the DADA2 (Callahan et al., 2016) and the UNOISE (Edgar, 2016) algorithms. We introduced a correction in the UNOISE formula to take into account the different variability (entropy) in each codon position. This correction parameterizes the notion that a change in a third position is more likely to be natural variability than a change in a first or second position. Such a correction was not possible for the DADA2 algorithm. We determined that performing DADA2 on paired reads gave much the same results as on unpaired reads (the default procedure) and that the program was over-parameterized. Likewise, the change in entropy ratio of the denoised sequences when modifying parameters in DADA2 did not follow a clear pattern. We therefore chose to work with the corrected UNOISE algorithm, for which the best performing alpha parameter, as assessed by changes in the entropy ratio, was determined to be 5.

We concluded that using first clustering and then denoising or vice-versa had no major impact in the final results. However, from the nature of this thesis and under a metaphylogeographic perspective, we have followed the strategy of clustering before denoising. This strategy, followed by adequate filtering steps, will retain the reads of merged sequences into the same MOTUs. The idea behind it is to avoid that a high abundance sequence "absorbs" correct sequences from other low abundance MOTUs because of the high abundance skew. The computing times were also lower when performing denoising within MOTUs rather than on the whole dataset. Finally, we recommend researchers to report their results in terms of both

denoised sequences (a proxy for haplotypes) and clusters formed (a proxy for species), and to avoid collapsing the sequences of the latter into a single representative. This will allow studies at the cluster (ideally equating species-level diversity) and at the intra-cluster level, and will ease additivity and comparability between studies.

In Chapter 7 we formalized the new denoising algorithm, called DnoisE, and made it publicly available in Github. This program uses the initial entropy values of each codon position to compute a correction factor to differentiate between those changes more likely to be spurious than natural mutations in a certain codon position. We based our algorithm on the published formula of the software UNOISE (Edgar, 2016) and added a correction parameter to it. During the writing of this software we also added features such as parallelizable options, different merging criteria, or different formats of both the input and output files. DnoisE is then therefore the first software designed and tested to denoise COI sequences obtained using metabarcoding and the first one to implement a codon entropy-ratio correction to denoise coding DNA regions.

9.2 Application to Biogeography and Metaphylogeography

As pointed out before, the standardization of the sampling and bioinformatic methods is mandatory to obtain meaningful data to be used for biodiversity analysis. Moreover, the bioinformatic pipeline developed in this thesis has the potential to retrieve both intra and interspecies variability when using COI metabarcoding to assess marine benthic communities. By keeping both the MOTUs and the ESVs clustered into them we have proxies of the species and their haplotypes that allow us to biomonitor communities and populations at the inter- and intraspecies levels.

In the chapter 8 we have tested the ability of metabarcoding to retrieve both biogeographic and metaphylogeographic patterns in one key area, the Atlanto-Mediterranean transition along the eastern Iberian coast. This area features a sharp shift from Atlantic to Mediterranean waters, and

is marked by two well-known barriers, the Almeria-Oran Front (AOF) (Folkard et al., 1994; L'Helguen et al., 2002; Tintore et al., 1988) and the Ibiza Channel (IC) (Bouffard et al., 2010; Pinot et al., 2002). The AOF in particular is a geostrophic front that separates Atlantic waters entering through the Gibraltar Strait from Mediterranean waters, thus marking the quasi-permanent true hydrological boundary between Atlantic and Mediterranean (Schunter et al., 2011; Tintore et al., 1988), while the IC is a secondary break that delimit the south of the Balearic Sea (Bouffard et al., 2010). Our sampling therefore encompassed a southern, a central, and a northern region separated by these breaks. The Atlanto-Mediterranean transition has been studied from the point of view of population genetics in a handful of species (reviewed in Pascual et al., 2017; Patarnello et al., 2007), but this Thesis represents the first study to explore the effect of these discontinuities using metabarcoding data with many different taxa at the same time in marine benthic communities.

From the analysis at MOTU level (biogeographic approach), a good separation between the southern localities and the others, corresponding to the AOF, was apparent in nMDS results and permanova analyses. The same discontinuity had a higher dissimilarity between communities when comparing adjacent localities. The IC had also an effect separating in the second axis of the nMDS the localities of the central and northern region. However, this effect is less marked than that of the AOF as seen in the comparison of adjacent localities and was not significant in permanova. The proportion of MOTUs of the different groups was similar across localities both for the main eukaryotic groups and the metazoan phyla. On the other hand, the proportion of reads had clear differences in some localities. The proportion of metazoan reads was lower in TAR and JOY, where the proportions of Stramenopiles (including Phaeophyta) and Rhodophyta, respectively, were especially high in these localities.

We also performed the analysis using ESVs instead of MOTUs, and the patterns found were less clear-cut. In particular, the AOF didn't appear as a significant divide anymore (permanova test). Brandt et al. (2021) discussed the relative merits of using MOTUs and/or ESVs for

metabarcoding studies. We endorse their vision that using ESVs as a standard unit of analysis may be valid for ribosomal markers (Callahan et al., 2017), yet, when studying eukaryotes and highly variable markers such as COI, MOTUs must be used as a proxy for species, while ESVs must be used only within meaningful biological units (MOTUs) as a proxy for haplotypes and intraspecies variability (metaphylogeographic approach), allowing a hierarchical analysis of the distribution of diversity. Using ESVs without their MOTU context is equivalent to lumping together biogeographic (interspecific) and phylogeographic (intraspecific) information. We therefore discourage the use of ESVs as the unit for biogeographic analyses.

For our metaphylogeographic analyses we retained 339 MOTUs that were shared between a minimum of two regions and were present at least in two localities within each region. For these MOTUs, we used the abundance of their ESVs (in a semiquantitative scale) to compute values of genetic dissimilarity with the D estimator (Jost, 2008). The analysis of mean D values of all shared MOTUs between localities showed a good separation between the three regions in the nMDS, with significant differences associated to the AOF and the IC divides.

Using these data we also computed a network analysis to assess the strength of the genetic relationships between localities. For the southern region, links were established only between adjacent localities, with the connection between TAR and SOL being the weakest. These localities appeared disconnected in the network from those of other areas, indicating that the AOF is a strong divide between the southernmost localities and the central and northern regions. On the contrary JOY, the closest locality to the IC, had a good connection with other localities of both the northern and central regions. This seemed to suggest that, while the AOF has a strong effect separating genetically the populations at both sides, the IC is more a transition zone than a true break. Moreover, it has been recently reported that these fronts can have some fluctuations that can affect the populations arround them (Ojeda et al., 2022) which could explain the connectivity of JOY if the IC is not fixed over time. Future studies should focus on this area and the Balearic Islands to better assess the role of the Ibiza Channel

in the connectivity of populations. We also envisage for the future a more detailed analysis of the area surrounding the Strait of Gibraltar.

We acknowledge that the analyses that we have performed used only a small representation of the range of tools available for phylogeographic studies, so we only touched the surface of the potential amount of information that can be gleaned with this approach. From our results, obtained from 1,200km of coast, future studies can be designed to focus resources on particular areas of interest. The analysis of the biogeographic and phylogeographic breaks is crucial to understand how different areas are connected, how communities change, and the connectivity of populations between them. Moreover, transitional zones are areas of high interest for management and can be reservoirs of biodiversity, thus the importance of continued monitoring of these communities. Metaphylogeographic results can be mined to detect which species have interesting patterns of distribution and connectivity. Selected populations of these species can then be targeted using traditional phylogeographic approaches or the new population genomics methods (Reitzel et al., 2013) to fully understand the driving mechanisms behind the distribution of their genetic variability.

9.3 Looking through the Crystal Ball

This Thesis lays the groundwork for the study of the eukaryotic communities inhabiting marine reef habitats. Metabarcoding is a fast-developing field, but methods for this type of community were still to be developed at the start of this Thesis, from sampling procedures to bioinformatic treatment of sequences obtained from the highly variable marker selected (the Leray fragment of COI, Leray et al., 2013). Our methods allowed us to uncover a bewildering biodiversity in these complex communities, spanning typically thousands of MOTUs in a single locality. We also took advantage of the coding properties of COI, so far unexplored in studies applying metabarcoding methods, to improve the bioinformatic procedures by adjusting parameters in commonly used programs and by developing new denoising methods.

An efficient denoising procedure is the pre-requisite for the study of

intraspecies diversity with metabarcoding data, a new field that we have christened "metaphylogeography". The combination of the traditional metabarcoding approach with metaphylogeography opens the door to fast and efficient assessment of the biogeographic and phylogeographic patterns of the communities and populations in the marine ecosystems. It also allows to focus resources on areas of interest if used as exploratory methodology in poorly-known areas. The generation of a huge amount of information with relatively low effort provides a holistic perspective of how ecosystems are connected and how they will interact in case of disturbance. For managers this information will be crucial in an scenario of fast global change.

It is difficult to forecast the future of the metabarcoding approach as applied to marine and other communities. Metabarcoding is becoming more and more accepted as the new gold standard for biomonitoring, but its weaknesses must be also acknowledged (Duarte et al., 2021b; Elbrecht et al., 2017b; Zaiko et al., 2015). We have shown that the current trend to use water eDNA to assess biodiversity in aquatic ecosystems has its limitations when applied to benthic communities. We also contend that the number of MOTUs assigned at high taxonomic levels or simply unassigned reflects the lack of completeness of the reference databases, particularly when it comes to meio- and microeukaryotes. Clearly this is an aspect of eukaryote metabarcoding that needs improvement, and no doubt the current barcoding projects worldwide will provide the much needed updated databases. It is also mandatory to devise new software to automatically generate the reference databases to keep pace with the growing number of sequences being included in public repositories. Moreover, the amount of information retrieved from new and future sequencing technologies demands efficient and powerful algorithms to process millions of sequences in a reasonable time.

Yet, metabarcoding approaches have been proved to be robust and its applicability on a daily basis will increase in popularity thus managers and policymakers must be trained in these technologies but it is everyone's business that knowledge is transferred to society. The development of fast and easy to perform protocols to detect sequences in the environment as possible

bioindicators of impacts, alien species or harmful blooms will introduce eDNA-based technologies into management. In this area, software development applying artificial neural networks and machine learning (Cordier et al., 2018; Pawlowski et al., 2018; Pichler and Hartig, 2021) and hardware specially designed to detect this DNA patterns (Nakao et al., 2022) are promising and will enhance the use of eDNA approaches in biomonitoring and managing.

Metabarcoding (amplicon sequencing) is a well-established method in the study of prokaryotes, but it is being progressively complemented, if not replaced, by metagenomics, whereby the whole DNA present in the samples is analysed via shotgun sequencing, rather than using amplicons of single markers (Thomas et al., 2012)). The high number of prokaryote genomes available and their relative simplicity makes this approach feasible. Likewise, prokaryote metatranscriptomics analyse the transcripts present in the environment. Metagenomics and metatranscriptomics enable a more functional profiling, as compared to a descriptive taxonomic approach, of the communities (Semenov, 2021). It is likely that the study of eukaryotic communities will follow the same trend. For the time being, however, the complexity of eukaryotic genomes and the limited availability of published genomes (particularly in small organisms) hinders the switch from metabarcoding to metagenomics (Singer et al., 2020; Stat et al., 2017). Undoubtedly, as sequencing platforms increase their output, and in the wake of global efforts (such as the Earth Biogenome Project; Lewin et al., 2018) to escalate the number of genomes available, the time for eukaryotic metagenomics will eventually come. In the meantime, the application of mito-metagenomics may bridge the gap between amplicon sequencing and metagenomics (Andújar et al., 2015; Bista et al., 2018; Crampton-Platt et al., 2016). For the time being, however, there is still a lot of information to be gleaned from metabarcoding analysis of complex eukaryotic communities such as the benthic assemblages on which this Thesis is focused.

Chapter 10

Conclusions

Taking into account the objectives of this thesis, the main conclusions reached are:

1.1) Direct sampling is necessary to properly assess the biodiversity of eukaryotic marine benthic communities using metabarcoding approaches. With universal primers targeting the highly variable COI marker, the environmental DNA captured from the water at the boundary layer failed to retrieve the benthic diversity and thus the traditional method of 'quadrat sampling' is recommended.

2.1) We have developed a pipeline designed and tested for a highly variable metabarcode, the Leray fragment of COI, to obtain both the inter- and intra-MOTU variability. This pipeline combines the two common strategies to process sequences in metabarcoding studies, clustering and denoising.

2.2) We introduced the field of Metaphylogeography as the study of phylogeographic patterns of hundreds of species simultaneously using metabarcoding data.

2.3) We have calibrated the parameters of our preferred clustering program, SWARM, for COI metabarcoding of marine benthic communities.

2.4) We have designed, programmed, calibrated and tested the DnoisE program. This parallelizable open-source software is a new formulation of the UNOISE algorithm and includes a correction factor based on the different entropy values of each position in the codon to better assess the probability that a change in a given codon position is a natural variation or a sequencing artifact. DnoisE has been designed and tested for metabarcoding

data of coding genes but can also be used (disabling entropy correction) with noncoding markers.

3.1) We have successfully applied the metabarcoding methods developed to assess the biogeographic and metaphylogeographic patterns of the Eastern Iberian Coast to analyse the Atlanto-Mediterranean transition along two well studied barriers, the Almeria-Oran Front (AOF) and the Ibiza Channel (IC).

3.2) In all the analyses, the AOF had a strong effect separating regions, confirming previous reports. For the IC, a clear effect was detected only with the metaphylogeographic approach. Network analysis confirmed the important role of the AOF and showed that the localities close to the IC are highly connected with central and northern regions, indicating a potential transition area near the IC, rather than a well-delimited break.

3.3) We favour the use of MOTUs as a proxy of species and ESVs within MOTUs as a proxy of haplotypes for metabarcoding with highly variable markers.

3.4) Metabarcoding offers the opportunity to fast and efficiently assess biogeographic and metaphylogeographic patterns and has the potential to become a cornerstone in biodiversity assessment of complex littoral benthic communities.

Bibliography

Adamowicz, S. J., Boatwright, J. S., Chain, F. J. J., Fisher, B. L., Hogg, I. D., Leese, F., Lijtmaer, D. A., Mwale, M., Naaum, A. M., Pochon, X., Steinke, D., Wilson, J.-J., Wood, S., Xu, J., Xu, S., Zhou, X., and van der Bank, M. 2019. Trends in DNA barcoding and metabarcoding. *Genome*, **62**(**3**):v–viii. doi: 10.1139/gen-2019-0054.

Adams, C. I., Knapp, M., Gemmell, N. J., Jeunen, G.-J., Bunce, M., Lamare, M. D., Taylor, H. R., Adams, C. I., Knapp, M., Gemmell, N. J., Jeunen, G.-J., Bunce, M., Lamare, M. D., and Taylor, H. R. 2019. Beyond Biodiversity: Can Environmental DNA (eDNA) Cut It as a Population Genetics Tool? *Genes*, **10**(**3**):192. doi: 10.3390/genes10030192.

Agardy, T., Alder, J., Dayton, P., Curran, S. R., Kitchingman, A., Wilson, M., Catenazzi, A., Restrepo, J., Birkeland, C., Blaber, S., Saifullah, S., Branch, G., Boersma, D., Nixon, S., Dugan, P., Davidson, N., and Vörösmarty, C. 2005. Coastal systems. In Reid, W. V., editor, *Millennium ecosystem assessment: ecosystems and human well-being*, chapter 19, pages 513–549. Island Press, Washington D.C.

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., and Bohmann, K. 2018. Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution*, **9**(**1**):134–147. doi: 10.1111/2041-210X.12849.

Alexander, J. B., Bunce, M., White, N., Wilkinson, S. P., Adam, A. A., Berry, T. E., Stat, M., Thomas, L., Newman, S. J., Dugal, L., and Richards, Z. T. 2020. Development of a multi-assay approach for monitoring coral diversity using eDNA metabarcoding. *Coral Reefs*, **39**(**1**):159–171. doi: 10.1007/s00338-019-01875-9.

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., Kightley, E. P., Thompson, L. R., Hyde, E. R., Gonzalez, A., and Knight, R. 2017. Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems*, **2**(**2**): e00191–16. doi: 10.1128/msystems.00191-16.

Anderson, M. J., Gorley, R. N., and Clarke, K. R. 2008. *PERMANOVA+ for PRIMER: guide to software and statistical methods*. Plymouth: Primer-E Ltd.

Andújar, C., Arribas, P., Ruzicka, F., Crampton-Platt, A., Timmermans, M. J., and Vogler, A. P. 2015. Phylogenetic community ecology of soil biodiversity using mitochondrial metagenomics. *Molecular Ecology*, **24**(**14**):3603–3617. doi: 10.1111/mec.13195.

Andújar, C., Arribas, P., Yu, D. W., Vogler, A. P., and Emerson, B. C. 2018. Why the COI barcode should be the community DNA metabarcode for the metazoa. *Molecular Ecology*, **27**(**20**):3968–3975. doi: 10.1111/mec.14844.

Andújar, C., Creedy, T. J., Arribas, P., López, H., Salces-Castellano, A., Pérez-Delgado, A. J., Vogler, A. P., and Emerson, B. C. 2021. Validated removal of nuclear pseudogenes and sequencing artefacts from mitochondrial metabarcode data. *Molecular Ecology Resources*, **21**(**6**):1772–1787. doi: 10.1111/1755-0998.13337.

Andújar, C., Arribas, P., López, H., Arjona, Y., Pérez-Delgado, A., Oromí, P., Vogler, A. P., and Emerson, B. C. 2022. Community assembly and metaphylogeography of soil biodiversity: insights from haplotype-level community DNA metabarcoding within an oceanic island. *Molecular Ecology*. doi: 10.1111/mec.16560.

Antich, A., 2020. DnoisE, Distance denoise by Entropy. GitHub repository. https://github.com/adriantich/DnoisE.

Antich, A., Palacín, C., Cebrian, E., Golo, R., Wangensteen, O. S., and Turon, X. 2020. Marine biomonitoring with eDNA: can metabarcoding of water samples cut it as a tool for surveying benthic communities? *Molecular ecology*, pages 1–14. doi: 10.1111/mec.15641.

Antich, A., Palacín, C., Wangensteen, O. S., and Turon, X. 2021a. To denoise or to cluster, that is not the question: optimizing pipelines for COI metabarcoding and metaphylogeography. *BMC Bioinformatics*, **22** (**1**):177. doi: 10.1186/s12859-021-04115-6.

Antich, A., Palacin, C., Wangensteen, O. S., and Turon, X. 2021b. Dataset for "To denoise or to cluster? That is not the question. Optimizing pipelines for COI metabarcoding and metaphylogeography". *Mendeley*

Data.

Antich, A., Palacín, C., Turon, X., and Wangensteen, O. S. 2022. DnoisE: Distance denoising by entropy. An open-source parallelizable alternative for denoising sequence datasets. *PeerJ*, **10**:e12758. doi: 10.7717/peerj.12758.

Arranz, V., Fewster, R. M., and Lavery, S. D. 2021. Geographic concordance of genetic barriers in New Zealand coastal marine species. *Aquatic Conservation: Marine and Freshwater Ecosystems*, **31**(**12**):3607–3625. doi: 10.1002/aqc.3735.

Arranz, V., Fewster, R. M., and Lavery, S. D. 2022. Genogeographic clustering to identify cross-species concordance of spatial genetic patterns. *Diversity and Distributions*, **00**:1–13. doi: 10.1111/ddi.13474.

Atienza, S., Guardiola, M., Præbel, K., Antich, A., Turon, X., and Wangensteen, O. S. 2020. DNA Metabarcoding of Deep-Sea Sediment Communities Using COI: Community Assessment, Spatio-Temporal Patterns and Comparison with 18S rDNA. *Diversity*, **12**(**4**):123. doi: 10.3390/d12040123.

Avise, J. C. 2009. Phylogeography: retrospect and prospect. *Journal of Biogeography*, **36**(**1**):3–15. doi: 10.1111/j.1365-2699.2008.02032.x.

Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E., Reeb, C. A., and Saunders, N. C. 1987. INTRASPECIFIC PHYLOGEOGRAPHY: The Mitochondrial DNA Bridge Between Population Genetics and Systematics. *Annual Review of Ecology and Systematics*, **18**(**1**):489–522. doi: 10.1146/annurev.es.18.110187.002421.

Aylagas, E., Borja, Á., Irigoien, X., and Rodríguez-Ezpeleta, N. 2016. Benchmarking DNA Metabarcoding for Biodiversity-Based Monitoring and Assessment. *Frontiers in Marine Science*, **3**(**96**). doi: 10.3389/fmars.2016.00096.

Aylagas, E., Borja, Á., Muxika, I., and Rodríguez-Ezpeleta, N. 2018. Adapting metabarcoding-based benthic biomonitoring into routine marine ecological status assessment networks. *Ecological Indicators*, **95**:194–202. doi: 10.1016/j.ecolind.2018.07.044.

Ayre, D. J., Minchinton, T. E., and Perrin, C. 2009. Does life history predict past and current connectivity for rocky intertidal invertebrates

across a marine biogeographic barrier? *Molecular Ecology*, **18**(**9**):1887–1903. doi: 10.1111/j.1365-294X.2009.04127.x.

Baird, D. J. and Hajibabaei, M. 2012. Biomonitoring 2.0: A new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology*, **21**:2039–2044. doi: 10.1111/j.1365-294X.2012.05519.x.

Baker, C. S., Steel, D., Nieukirk, S., and Klinck, H. 2018. Environmental DNA (eDNA) From the Wake of the Whales: Droplet Digital PCR for Detection and Species Identification. *Frontiers in Marine Science*, **5**:133. doi: 10.3389/fmars.2018.00133.

Bakker, J., Wangensteen, O. S., Chapman, D. D., Boussarie, G., Buddo, D., Guttridge, T. L., Hertler, H., Mouillot, D., Vigliola, L., and Mariani, S. 2017. Environmental DNA reveals tropical shark diversity in contrasting levels of anthropogenic impact. *Scientific Reports*, **7**(**1**):16886. doi: 10.1038/s41598-017-17150-2.

Bakker, J., Wangensteen, O. S., Baillie, C., Buddo, D., Chapman, D. D., Gallagher, A. J., Guttridge, T. L., Hertler, H., and Mariani, S. 2019. Biodiversity assessment of tropical shelf eukaryotic communities via pelagic eDNA metabarcoding. *Ecology and Evolution*, page ece3.5871. doi: 10.1002/ece3.5871.

Bani, A., De Brauwer, M., Creer, S., Dumbrell, A. J., Limmon, G., Jompa, J., von der Heyden, S., and Beger, M. 2020. Informing marine spatial planning decisions with environmental DNA. In *Advances in Ecological Research*, volume 62, pages 375–407. Academic Press Inc. ISBN 9780128211342. doi: 10.1016/bs.aecr.2020.01.011.

Barnes, D. K. and Crook, A. C. 2001. Implications of temporal and spatial variability in Paracentrotus lividus populations to the associated commercial coastal fishery. *Hydrobiologia*, **465**(**1**):95–102. doi: 10.1023/A:1014568103867.

Barnes, M. A. and Turner, C. R. 2016. The ecology of environmental DNA and implications for conservation genetics. *Conservation Genetics*, **17**(**1**):1–17. doi: 10.1007/s10592-015-0775-4.

Beng, K. C. and Corlett, R. T. 2020. Applications of environmental DNA (eDNA) in ecology and conservation: opportunities, challenges

and prospects. *Biodiversity and Conservation*, **29**(**7**):2089–2121. doi: 10.1007/s10531-020-01980-0.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**(**1**):289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Bianchi, C. N. 2007. Biodiversity issues for the forthcoming tropical Mediterranean Sea. *Hydrobiologia 2007 580:1*, **580**(**1**):7–21. doi: 10.1007/s10750-006-0469-5.

Bianchi, C. N. and Morri, C. 2000. Marine Biodiversity of the Mediterranean Sea: Situation, Problems and Prospects for Future Research. *Marine Pollution Bulletin*, **40**(**5**):367–376. doi: 10.1016/S0025-326X(00)00027-8.

Bista, I., Carvalho, G. R., Tang, M., Walsh, K., Zhou, X., Hajibabaei, M., Shokralla, S., Seymour, M., Bradley, D., Liu, S., Christmas, M., and Creer, S. 2018. Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. *Molecular Ecology Resources*, **18**(**5**):1020–1034. doi: 10.1111/1755-0998.12888.

Bodenhofer, U., Bonatesta, E., Horejš-Kainrath, C., and Hochreiter, S. 2015. msa: an R package for multiple sequence alignment. *Bioinformatics*, **31**(**24**):3997–3999. doi: 10.1093/bioinformatics/btv494.

Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., Yu, D. W., and de Bruyn, M. 2014. Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology & Evolution*, **29**(**6**):358–367. doi: 10.1016/j.tree.2014.04.003.

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., Cope, E. K., Da Silva, R., Diener, C., Dorrestein, P. C., Douglas, G. M., Durall, D. M., Duvallet, C., Edwardson, C. F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J. M., Gibbons, S. M., Gibson, D. L., Gonzalez,

A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G. A., Janssen, S., Jarmusch, A. K., Jiang, L., Kaehler, B. D., Kang, K. B., Keefe, C. R., Keim, P., Kelley, S. T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M. G., Lee, J., Ley, R., Liu, Y. X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B. D., McDonald, D., McIver, L. J., Melnik, A. V., Metcalf, J. L., Morgan, S. C., Morton, J. T., Naimey, A. T., Navas-Molina, J. A., Nothias, L. F., Orchanian, S. B., Pearson, T., Peoples, S. L., Petras, D., Preuss, M. L., Pruesse, E., Rasmussen, L. B., Rivers, A., Robeson, M. S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S. J., Spear, J. R., Swafford, A. D., Thompson, L. R., Torres, P. J., Trinh, P., Tripathi, A., Turnbaugh, P. J., Ul-Hasan, S., van der Hooft, J. J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W. A., Wan, Y., Wang, M., Warren, J., Weber, K. C., Williamson, C. H., Willis, A. D., Xu, Z. Z., Zaneveld, J. R., Zhang, Y., Zhu, Q., Knight, R., and Caporaso, J. G. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, **37**(**8**):852–857. doi: 10.1038/s41587-019-0209-9.

Borja, Á. and Elliott, M., 2018. There is no Planet B: A healthy Earth requires greater parity between space and marine research. doi: 10.1016/j.marpolbul.2018.03.015.

Borja, Á., White, M. P., Berdalet, E., Bock, N., Eatock, C., Kristensen, P., Leonard, A., Lloret, J., Pahl, S., Parga, M., Prieto, J. V., Wuijts, S., and Fleming, L. E., 2020. Moving Toward an Agenda on Ocean Health and Human Health in Europe. doi: 10.3389/fmars.2020.00037.

Boudreau, B. P. and Jorgensen, B. B., editors. 2001. *The Benthic Boundary Layer: Transport Processes and Biogeochemistry*. Oxford University Press, New York, NY. ISBN 0-19-5118811-2.

Bouffard, J., Pascual, A., Ruiz, S., Faugère, Y., and Tintoré, J. 2010. Coastal and mesoscale dynamics characterization using altimetry and gliders: A case study in the Balearic Sea. *Journal of Geophysical*

Research: Oceans, **115**(**C10**):10029. doi: 10.1029/2009JC006087.

Bowers, H. A., Pochon, X., von Ammon, U., Gemmell, N., Stanton, J. A. L., Jeunen, G. J., Sherman, C. D., and Zaiko, A. 2021. Towards the optimization of edna/erna sampling technologies for marine biosecurity surveillance. *Water*, **13**:1113. doi: 10.3390/w13081113.

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., and Coissac, E. 2016. Obitools: a unix-inspired software package for DNA metabarcoding. *Molecular ecology resources*, **16**:176–182. doi: 10.1111/1755-0998.12428.

Brandt, M. I., Trouche, B., Quintric, L., Günther, B., Wincker, P., Poulain, J., and Arnaud-Haond, S. 2021. Bioinformatic pipelines combining denoising and clustering tools allow for more comprehensive prokaryotic and eukaryotic metabarcoding. *Molecular Ecology Resources*, **21**(**6**):1904–1921. doi: 10.1111/1755-0998.13398.

Brannock, P. M., Ortmann, A. C., Moss, A. G., and Halanych, K. M. 2016. Metabarcoding reveals environmental factors influencing spatio-temporal variation in pelagic micro-eukaryotes. *Molecular Ecology*, **25**(**15**):3593–3604. doi: 10.1111/mec.13709.

Brannock, P. M., Learman, D., Mahon, A., Santos, S., and Halanych, K. 2018. Meiobenthic community composition and biodiversity along a 5500 km transect of Western Antarctica: a metabarcoding analysis. *Marine Ecology Progress Series*, **603**:47–60. doi: 10.3354/meps12717.

Briggs, J. C. 1995. Global biogeography. *Developments in Palaeontology and Stratigraphy*, **14**:i–xvii, 1–452.

Briski, E., Ghabooli, S., Bailey, S. A., and MacIsaac, H. J. 2016. Are genetic databases sufficiently populated to detect non-indigenous species? *Biological Invasions*, **18**(**7**):1911–1922. doi: 10.1007/S10530-016-1134-1.

Cahill, A. E., De Jode, A., Dubois, S., Bouzaza, Z., Aurelle, D., Boissin, E., Chabrol, O., David, R., Egea, E., Ledoux, J. B., Mérigot, B., Weber, A. A. T., and Chenuil, A. 2017. A multispecies approach reveals hot spots and cold spots of diversity and connectivity in invertebrate species with contrasting dispersal modes. *Molecular Ecology*, **26**(**23**):6563–6577. doi: 10.1111/mec.14389.

Cahill, A. E., Pearman, J. K., Borja, Á., Carugati, L., Carvalho,

S., Danovaro, R., Dashfield, S., David, R., Féral, J. P., Olenin, S., Šiaulys, A., Somerfield, P. J., Trayanova, A., Uyarra, M. C., and Chenuil, A. 2018. A comparative analysis of metabarcoding and morphology-based identification of benthic communities across different regional seas. *Ecology and Evolution*, **8**(**17**):8908–8920. doi: 10.1002/ece3.4283.

Calderón, I., Giribet, G., and Turon, X. 2008. Two markers and one history: Phylogeography of the edible common sea urchin Paracentrotus lividus in the Lusitanian region. *Marine Biology*, **154**(**1**):137–151. doi: 10.1007/S00227-008-0908-0.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, **13**(**7**):581–583. doi: 10.1038/nmeth.3869.

Callahan, B. J., McMurdie, P. J., and Holmes, S. P. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME Journal*, **11**(**12**):2639–2643. doi: 10.1038/ismej.2017.119.

Carreras, C., García-Cisneros, A., Wangensteen, O. S., Ordóñez, V., Palacín, C., Pascual, M., and Turon, X. 2020. East is East and West is West: Population genomics and hierarchical analyses reveal genetic structure and adaptation footprints in the keystone species Paracentrotus lividus (Echinoidea). *Diversity and Distributions*, **26**(**3**):382–398. doi: 10.1111/ddi.13016.

Celio, G., Padamsee, M., Dentinger, B., Bauer, R., and McLaughlin, D. 2017. Assembling the Fungal Tree of Life: constructing the Structural and Biochemical Database. *Mycologia*, **98**(**6**):850–859. doi: 10.1080/15572536.2006.11832615.

Chen, H., **2018**. VennDiagram: Generate High-Resolution Venn and Euler Plots. R package version 1.6.20. https://CRAN.R-project.org/package=VennDiagram.

Chun, J., Lee, J. H., Jung, Y., Kim, M., Kim, S., Kim, B. K., and Lim, Y. W. 2007. EzTaxon: A web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences. *International*

Journal of Systematic and Evolutionary Microbiology, **57**(**10**):2259–2261. doi: 10.1099/ijs.0.64915-0.

Clarke, L. J., Trebilco, R., Walters, A., Polanowski, A. M., and Deagle, B. E. 2020. DNA-based diet analysis of mesopelagic fish from the southern Kerguelen Axis. *Deep Sea Research Part II: Topical Studies in Oceanography*, **174**:104494. doi: 10.1016/j.dsr2.2018.09.001.

Closek, C. J., Santora, J. A., Starks, H. A., Schroeder, I. D., Andruszkiewicz, E. A., Sakuma, K. M., Bograd, S. J., Hazen, E. L., Field, J. C., and Boehm, A. B. 2019. Marine vertebrate biodiversity and distribution within the central California Current using environmental DNA (eDNA) metabarcoding and ecosystem surveys. *Frontiers in Marine Science*, **6**:732. doi: 10.3389/fmars.2019.00732.

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., Brown, C. T., Porras-Alfaro, A., Kuske, C. R., and Tiedje, J. M. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, **42**(**D1**): D633–D642. doi: 10.1093/nar/gkt1244.

Collins, R. A., Wangensteen, O. S., O'Gorman, E. J., Mariani, S., Sims, D. W., and Genner, M. J. 2018. Persistence of environmental DNA in marine systems. *Communications Biology*, **1**(**1**):185. doi: 10.1038/s42003-018-0192-6.

Collins, R. A., Bakker, J., Wangensteen, O. S., Soto, A. Z., Corrigan, L., Sims, D. W., Genner, M. J., and Mariani, S. 2019. Non-specific amplification compromises environmental DNA metabarcoding with COI. *Methods in Ecology and Evolution*, **10**(**11**). doi: 10.1111/2041-210X.13276.

Compson, Z. G., McClenaghan, B., Singer, G. A., Fahner, N. A., and Hajibabaei, M. 2020. Metabarcoding From Microbes to Mammals: Comprehensive Bioassessment on a Global Scale. *Frontiers in Ecology and Evolution*, **8**:379. doi: 10.3389/fevo.2020.581835.

Conway, J., Lex, A., and Gehlenborg, N. 2017. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*, **33**(**18**):2938–2940. doi: 10.1093/bioinformatics/btx364.

Cordier, T. and Pawlowski, J. 2018. BBI: an R package for the com-

putation of Benthic Biotic Indices from composition data. *Metabarcoding and Metagenomics*, **2**:e25649. doi: 10.3897/mbmg.2.25649.

Cordier, T., Forster, D., Dufresne, Y., Martins, C. I. M., Stoeck, T., and Pawlowski, J. 2018. Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Molecular Ecology Resources*, **18**(**6**):1381–1391. doi: 10.1111/1755-0998.12926.

Costello, M. J. and Chaudhary, C. 2017. Marine Biodiversity, Biogeography, Deep-Sea Gradients, and Conservation. *Current Biology*, **27** (**11**):R511–R527. doi: 10.1016/j.cub.2017.04.060.

Costello, M. J., Tsai, P., Wong, P. S., Cheung, A. K. L., Basher, Z., and Chaudhary, C. 2017. Marine biogeographic realms and species endemicity. *Nature Communications 2017 8:1*, **8**(**1**):1–10. doi: 10.1038/s41467-017-01121-2.

Cowart, D. A., Pinheiro, M., Mouchel, O., Maguer, M., Grall, J., Miné, J., and Arnaud-Haond, S. 2015. Metabarcoding is powerful yet still blind: A comparative analysis of morphological and molecular surveys of seagrass communities. *PLoS ONE*, **10**(**2**). doi: 10.1371/journal.pone.0117562.

Cowart, D. A., Matabos, M., Brandt, M. I., Marticorena, J., and Sarrazin, J. 2020. Exploring Environmental DNA (eDNA) to Assess Biodiversity of Hard Substratum Faunal Communities on the Lucky Strike Vent Field (Mid-Atlantic Ridge) and Investigate Recolonization Dynamics After an Induced Disturbance. *Frontiers in Marine Science*, **6**:783. doi: 10.3389/fmars.2019.00783.

Crampton-Platt, A., Yu, D. W., Zhou, X., and Vogler, A. P. 2016. Mitochondrial metagenomics: Letting the genes out of the bottle. *GigaScience*, **5**(**1**). doi: 10.1186/s13742-016-0120-y.

Creedy, T. J., Andújar, C., Meramveliotakis, E., Noguerales, V., Overcast, I., Papadopoulou, A., Morlon, H., Vogler, A. P., Emerson, B. C., Arribas, P., Papadopoulou, A., Morlon, H., Vogler, A. P., Emerson, B. C., and Arribas, P. 2022. Coming of age for COI metabarcoding of whole organism community DNA: Towards bioinformatic harmonisation. *Molecular Ecology Resources*, **22**(**3**):847–861.

doi: 10.1111/1755-0998.13502.

Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Thomas, W. K., Potter, C., and Bik, H. M. 2016. The ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology and Evolution*, **7**(**9**):1008–1018. doi: 10.1111/2041-210X.12574.

Cristescu, M. E., 2014. From barcoding single individuals to metabarcoding biological communities: Towards an integrative approach to the study of global biodiversity. doi: 10.1016/j.tree.2014.08.001.

Cristescu, M. E. and Hebert, P. D. 2018. Uses and Misuses of Environmental DNA in Biodiversity Science and Conservation. *Annual Review of Ecology, Evolution, and Systematics*, **49**(**1**):209–230. doi: 10.1146/annurev-ecolsys-110617-062306.

Crutzen, P. J. and Stoermer, E. F. 2000. The "Anthropocene". *IGBP Newsletter*, **41**:17–18.

Czachur, M. V., Seymour, M., Creer, S., and von der Heyden, S. 2021. Novel insights into marine fish biodiversity across a pronounced environmental gradient using replicated environmental DNA analyses. *Environmental DNA*, **00**:1–10. doi: 10.1002/edn3.238.

Dafforn, K. A., Baird, D. J., Chariton, A. A., Sun, M. Y., Brown, M. V., Simpson, S. L., Kelaher, B. P., and Johnston, E. L. 2014. Faster, Higher and Stronger? The Pros and Cons of Molecular Faunal Data for Assessing Ecosystem Condition. *Advances in Ecological Research*. doi: 10.1016/B978-0-08-099970-8.00003-8.

Danovaro, R., Carugati, L., Berzano, M., Cahill, A. E., Carvalho, S., Chenuil, A., Corinaldesi, C., Cristina, S., David, R., Dell'Anno, A., Dzhembekova, N., Garcés, E., Gasol, J. M., Goela, P., Féral, J.-P., Ferrera, I., Forster, R. M., Kurekin, A. A., Rastelli, E., Marinova, V., Miller, P. I., Moncheva, S., Newton, A., Pearman, J. K., Pitois, S. G., Reñé, A., Rodríguez-Ezpeleta, N., Saggiomo, V., Simis, S. G. H., Stefanova, K., Wilson, C., Lo Martire, M., Greco, S., Cochrane, S. K. J., Mangoni, O., and Borja, Á. 2016. Implementing and Innovating Marine Monitoring Approaches for Assessing Marine Environmental Status. *Frontiers in Marine Science*, **3**:213. doi: 10.3389/fmars.2016.00213.

Darling, J. A., Galil, B. S., Carvalho, G. R., Rius, M., Viard, F., and Piraino, S. 2017. Recommendations for developing and applying genetic tools to assess and manage biological invasions in marine ecosystems. *Marine Policy*, **85**:54–64. doi: 10.1016/j.marpol.2017.08.014.

Davoult, D. 1989. Structure démografique et production de la population d'Ophiothrix fragilis (Abildgaard) du Détroit du Pas-de-Calais, France. *Vie Marine*, **10**:116–127.

Davoult, D. and Gounin, F. 1995. Suspension-feeding activity of a dense Ophiothrix fragilis (Abildgaard) population at the water-sediment interface: Time coupling of food availability and feeding behaviour of the species. *Estuarine, Coastal and Shelf Science*, **41**(**5**):567–577. doi: 10.1016/0272-7714(95)90027-6.

Dawson, M. N. 2014. Natural experiments and meta-analyses in comparative phylogeography. *Journal of Biogeography*, **41**(**1**):52–65. doi: 10.1111/jbi.12190.

de Jode, A., David, R., Dubar, J., Rostan, J., Guillemain, D., Sartoretto, S., Feral, J.-P., and Chenuil, A. 2019. Community ecology of coralligenous assemblages using a metabarcoding approach. In *3rd Mediterranean Symposium on the conservation of Coralligenous & other Calcareous Bio-Concretions*, pages 41–45, Antalya, Turkey. Specially Protected Areas Regional Activity Centre (SPA/RAC). ISBN 9789938957457.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.-M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horák, A., Jaillon, O., Lima-Mendez, G., Lukeš, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Tara Oceans Coordinators, T. O., Acinas, S. G., Bork, P., Bowler, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Raes, J., Sieracki, M. E., Speich, S., Stemmann, L., Sunagawa, S., Weissenbach, J., Wincker, P., and Karsenti, E. 2015. Ocean plankton. Eukaryotic

plankton diversity in the sunlit ocean. *Science (New York, N.Y.)*, **348** (**6237**):1261605. doi: 10.1126/science.1261605.

Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., and Taberlet, P. 2014. DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology Letters*, **10**(**9**). doi: 10.1098/rsbl.2014.0562.

Decelle, J., Romac, S., Stern, R. F., Bendif, E. M., Zingone, A., Audic, S., Guiry, M. D., Guillou, L., Tessier, D., Le Gall, F., Gourvil, P., Dos Santos, A. L., Probert, I., Vaulot, D., de Vargas, C., and Christen, R. 2015. PhytoREF: a reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated taxonomy. *Molecular Ecology Resources*, **15**(**6**):1435–1445. doi: 10.1111/1755-0998.12401.

Deiner, K., Fronhofer, E. A., Mächler, E., Walser, J.-C., and Altermatt, F. 2016. Environmental DNA reveals that rivers are conveyer belts of biodiversity information. *Nature Communications*. doi: 10.1038/ncomms12544.

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., de Vere, N., Pfrender, M. E., and Bernatchez, L. 2017. Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, **26**(**21**):5872–5895. doi: 10.1111/mec.14350.

Derycke, S., Remerie, T., Backeljau, T., Vierstraete, A., Vanfleteren, J., Vincx, M., and Moens, T. 2008. Phylogeography of the Rhabditis (Pellioditis) marina species complex: evidence for long-distance dispersal, and for range expansions and restricted gene flow in the northeast Atlantic. *Molecular Ecology*, **17**(**14**):3306–3322. doi: 10.1111/j.1365-294X.2008.03846.x.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, **72**(**7**):5069–5072. doi: 10.1128/AEM.03006-05.

DiBattista, J. D., Berumen, M. L., Priest, M. A., De Brauwer, M., Coker, D. J., Sinclair-Taylor, T. H., Hay, A., Bruss, G., Mansour, S., Bunce, M., Goatley, C. H. R., Power, M., and Marshell, A. 2022. Environmental DNA reveals a multi-taxa biogeographic break across the Arabian Sea and Sea of Oman. *Environmental DNA*, **4**(**1**): 206–221. doi: 10.1002/edn3.252.

Djurhuus, A., Pitz, K., Sawaya, N. A., Rojas-Márquez, J., Michaud, B., Montes, E., Muller-Karger, F., and Breitbart, M. 2018. Evaluation of marine zooplankton community structure through environmental DNA metabarcoding. *Limnology and Oceanography: Methods*, **16**(**4**):209–221. doi: 10.1002/lom3.10237.

Dray, S. and Dufour, A. B. 2007. The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, **22** (**4**):1–20. doi: 10.18637/jss.v022.i04.

Duarte, C. M. 2000. Marine biodiversity and ecosystem services: an elusive link. *Journal of Experimental Marine Biology and Ecology*, **250** (**1-2**):117–131. doi: 10.1016/S0022-0981(00)00194-5.

Duarte, S., Leite, B. R., Feio, M. J., Costa, F. O., and Filipe, A. F. 2021a. Integration of DNA-Based Approaches in Aquatic Ecological Assessment Using Benthic Macroinvertebrates. *Water 2021, Vol. 13, Page 331*, **13**(**3**):331. doi: 10.3390/w13030331.

Duarte, S., Vieira, P. E., Lavrador, A. S., and Costa, F. O. 2021b. Status and prospects of marine NIS detection and monitoring through (e)DNA metabarcoding. *Science of The Total Environment*, **751**:141729. doi: 10.1016/j.scitotenv.2020.141729.

Dupérré, N. 2020. Old and new challenges in taxonomy: what are taxonomists up against? *Megataxa*, **1**(**1**):59–62–59–62. doi: 10.11646/megataxa.1.1.12.

Duran, S., Palacín, C., Becerro, M. A., Turon, X., and Giribet, G. 2004. Genetic diversity and population structure of the commercially harvested sea urchin Paracentrotus lividus (Echinodermata, Echinoidea). *Molecular Ecology*, **13**(**11**):3317–3328. doi: 10.1111/j.1365-294X.2004.02338.x.

Edgar, R. C. 2010. Search and clustering orders of magnitude faster than

BLAST. *Bioinformatics*, **26**(**19**):2460–2461. doi: 10.1093/bioinformatics/btq461.

Edgar, R. C. 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods 2013 10:10*, **10**(**10**):996–998. doi: 10.1038/nmeth.2604.

Edgar, R. C. 2016. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv*, page 081257. doi: 10.1101/081257.

Edgar, R. C. and Flyvbjerg, H. 2015. Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics*, **31**(**21**):3476–3482. doi: 10.1093/bioinformatics/btv401.

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R. 2011. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, **27**(**16**):2194–2200. doi: 10.1093/bioinformatics/btr381.

Ekman, S. 1953. *Zoogeography of the Sea*. Sidgwick & Jackson, London, 8 edition. doi: 10.1080/00222935308654417.

El Ayari, T., Trigui El Menif, N., Hamer, B., Cahill, A. E., and Bierne, N. 2019. The hidden side of a major marine biogeographic boundary: a wide mosaic hybrid zone at the Atlantic–Mediterranean divide reveals the complex interaction between natural and genetic barriers in mussels. *Heredity 2019 122:6*, **122**(**6**):770–784. doi: 10.1038/s41437-018-0174-y.

Elbrecht, V. and Leese, F. 2015. Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass-sequence relationships with an innovative metabarcoding protocol. *PLoS ONE*, **10**(**7**). doi: 10.1371/journal.pone.0130324.

Elbrecht, V. and Leese, F. 2017. Validation and Development of COI Metabarcoding Primers for Freshwater Macroinvertebrate Bioassessment. *Frontiers in Environmental Science*, **5**:11. doi: 10.3389/fenvs.2017.00011.

Elbrecht, V., Peinert, B., and Leese, F. 2017a. Sorting things out: Assessing effects of unequal specimen biomass on DNA metabarcoding. *Ecology and Evolution*, **7**(**17**):6918–6926. doi: 10.1002/ece3.3192.

Elbrecht, V., Vamos, E. E., Meissner, K., Aroviita, J., and Leese, F. 2017b. Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. *Methods in*

Ecology and Evolution, **8**(**10**):1265–1275. doi: 10.1111/2041-210X.12789.

Elbrecht, V., Hebert, P. D., and Steinke, D. 2018a. Slippage of degenerate primers can cause variation in amplicon length. *Scientific Reports*, **8**(**1**):1–5. doi: 10.1038/s41598-018-29364-z.

Elbrecht, V., Vamos, E. E., Steinke, D., and Leese, F. 2018b. Estimating intraspecific genetic diversity from community DNA metabarcoding data. *PeerJ*, **2018**(**4**):e4644. doi: 10.7717/peerj.4644.

Emerson, B. C., Cicconardi, F., Fanciulli, P. P., and Shaw, P. J. 2011. Phylogeny, phylogeography, phylobetadiversity and the molecular analysis of biological communities. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **366**(**1576**):2391–2402. doi: 10.1098/rstb.2011.0057.

Engel, M. S., Ceríaco, L. M., Daniel, G. M., Dellapé, P. M., Löbl, I., Marinov, M., Reis, R. E., Young, M. T., Dubois, A., Agarwal, I., Lehmann, P. A., Alvarado, M., Alvarez, N., Andreone, F., Araujo-Vieira, K., Ascher, J. S., Baêta, D., Baldo, D., Bandeira, S. A., Barden, P., Barrasso, D. A., Bendifallah, L., Bockmann, F. A., Böhme, W., Borkent, A., Brandão, C. R., Busack, S. D., Bybee, S. M., Channing, A., Chatzimanolis, S., Christenhusz, M. J., Crisci, J. V., D'Elía, G., Da Costa, L. M., Davis, S. R., De Lucena, C. A. S., Deuve, T., Fernandes Elizalde, S., Faivovich, J., Farooq, H., Ferguson, A. W., Gippoliti, S., Gonçalves, F. M., Gonzalez, V. H., Greenbaum, E., Hinojosa-Díaz, I. A., Ineich, I., Jiang, J., Kahono, S., Kury, A. B., Lucinda, P. H., Lynch, J. D., Malécot, V., Marques, M. P., Marris, J. W., McKellar, R. C., Mendes, L. F., Nihei, S. S., Nishikawa, K., Ohler, A., Orrico, V. G., Ota, H., Paiva, J., Parrinha, D., Pauwels, O. S., Pereyra, M. O., Pestana, L. B., Pinheiro, P. D., Prendini, L., Prokop, J., Rasmussen, C., Rödel, M. O., Rodrigues, M. T., Rodríguez, S. M., Salatnaya, H., Sampaio, Í., Sánchez-García, A., Shebl, M. A., Santos, B. S., Solórzano-Kraemer, M. M., Sousa, A. C., Stoev, P., Teta, P., Trape, J. F., Dos Santos, C. V. D., Vasudevan, K., Vink, C. J., Vogel, G., Wagner, P., Wappler, T., Ware, J. L., Wedmann, S., and Zacharie, C. K.

2021. The taxonomic impediment: a shortage of taxonomists, not the lack of technical approaches. *Zoological Journal of the Linnean Society*, **193**(**2**):381–387. doi: 10.1093/zoolinnean/zlab072.

Eren, A. M., Morrison, H. G., Lescault, P. J., Reveillaud, J., Vineis, J. H., and Sogin, M. L. 2015. Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME Journal*, **9**(**4**):968–979. doi: 10.1038/ismej.2014.195.

Estensmo, E. L. F., Maurice, S., Morgado, L., Martin-Sanchez, P. M., Skrede, I., and Kauserud, H. 2021. The influence of intraspecific sequence variation during DNA metabarcoding: A case study of eleven fungal species. *Molecular Ecology Resources*, **21**(**4**):1141–1148. doi: 10.1111/1755-0998.13329.

Ficetola, G. F., Miaud, C., Pompanon, F., and Taberlet, P. 2008. Species detection using environmental DNA from water samples. *Biology Letters*, **4**(**4**):423–425. doi: 10.1098/rsbl.2008.0118.

Ficetola, G. F., Poulenard, J., Sabatier, P., Messager, E., Gielly, L., Leloup, A., Etienne, D., Bakke, J., Malet, E., Fanget, B., Støren, E., Reyss, J. L., Taberlet, P., and Arnaud, F. 2018. DNA from lake sediments reveals long-term ecosystem changes after a biological invasion. *Science Advances*, **4**(**5**). doi: 10.1126/sciadv.aar4292.

Folkard, A. M., Davies, P. A., and Prieur, L. 1994. The surface temperature field and dynamical structure of the Almeria-Oran front from simultaneous shipboard and satellite data. *Journal of Marine Systems*, **5** (**3-5**):205–222. doi: 10.1016/0924-7963(94)90047-7.

Folmer, O., Black, M., Hoeh, W., Lutz, R., and Vrijenkoek, R. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, **3**(**5**):294–299.

Fonseca, V. G., Sinniger, F., Gaspar, J. M., Quince, C., Creer, S., Power, D. M., Peck, L. S., and Clark, M. S. 2017. Revealing higher than expected meiofaunal diversity in Antarctic sediments: a metabarcoding approach. *Scientific Reports*, **7**(**1**):1–11. doi: 10.1038/s41598-017-06687-x.

Fonseca, V. G., Carvalho, G. R., Nichols, B., Quince, C., Johnson, H. F., Neill, S. P., Lambshead, J. D., Thomas, W. K., Power, D. M., and Creer, S. 2014. Metagenetic analysis of patterns of distribution and diversity of marine meiobenthic eukaryotes. *Global Ecology and Biogeography*, **23**(**11**):1293–1302. doi: 10.1111/geb.12223.

Forster, D., Lentendu, G., Filker, S., Dubois, E., Wilding, T. A., and Stoeck, T. 2019. Improving eDNA-based protist diversity assessments using networks of amplicon sequence variants. *Environmental Microbiology*, **21**(**11**):4109–4124. doi: 10.1111/1462-2920.14764.

Fraija-Fernández, N., Bouquieaux, M.-C., Rey, A., Mendibil, I., Cotano, U., Irigoien, X., Santos, M., and Rodríguez-Ezpeleta, N. 2020. Marine water environmental DNA metabarcoding provides a comprehensive fish diversity assessment and reveals spatial patterns in a large oceanic area. *Ecology and Evolution*, **10**(**14**):7560–7584. doi: 10.1002/ece3.6482.

Frøslev, T. G., Kjøller, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., and Hansen, A. J. 2017. Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nature Communications*, **8**(**1**):1–11. doi: 10.1038/s41467-017-01312-x.

Gaither, M. R., DiBattista, J. D., Leray, M., and Heyden, S. 2022. Metabarcoding the marine environment: from single species to biogeographic patterns. *Environmental DNA*, **4**(**1**):3–8. doi: 10.1002/edn3.270.

Garcés-Pastor, S., Wangensteen, O. S., Pérez-Haase, A., Pèlachs, A., Pérez-Obiol, R., Cañellas-Boltà, N., Mariani, S., and Vegas-Vilarrúbia, T. 2019. DNA metabarcoding reveals modern and past eukaryotic communities in a high-mountain peat bog system. *Journal of Paleolimnology*, **62**(**4**):425–441. doi: 10.1007/s10933-019-00097-x.

Gebali, F. 2011. *Algorithms and Parallel Computing*. John Wiley and Sons. ISBN 9780470902103. doi: 10.1002/9780470932025.

Geller, J., Meyer, C., Parker, M., and Hawk, H. 2013. Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. *Molecular Ecology*

Resources, **13**(**5**):851–861. doi: 10.1111/1755-0998.12138.

Gerhard, W. A. and Gunsch, C. K. 2019. Metabarcoding and machine learning analysis of environmental DNA in ballast water arriving to hub ports. *Environment International*, **124**:312–319. doi: 10.1016/j.envint.2018.12.038.

Giangrande, A. 2003. Biodiversity, conservation, and the 'Taxonomic impediment'. *Aquatic Conservation: Marine and Freshwater Ecosystems*, **13**(**5**):451–459. doi: 10.1002/aqc.584.

Giebner, H., Langen, K., Bourlat, S. J., Kukowka, S., Mayer, C., Astrin, J. J., Misof, B., and Fonseca, V. G. 2020. Comparing diversity levels in environmental samples: DNA sequence capture and metabarcoding approaches using 18S and COI genes. *Molecular Ecology Resources*, **20**(**5**):1333–1345. doi: 10.1111/1755-0998.13201.

Gilbey, J., Carvalho, G. R., Castilho, R., Coscia, I., Coulson, M. W., Dahle, G., Derycke, S., Francisco, S. M., Helyar, S. J., Johansen, T., Junge, C., Layton, K. K., Martinsohn, J., Matejusova, I., Robalo, J. I., Rodríguez-Ezpeleta, N., Silva, G., Strammer, I., Vasemägi, A., and Volckaert, F. A. 2021. Life in a drop: Sampling environmental DNA for marine fishery management and ecosystem monitoring. *Marine Policy*, **124**:104331. doi: 10.1016/j.marpol.2020.104331.

Gleason, J. E., Elbrecht, V., Braukmann, T. W. A., Hanner, R. H., and Cottenie, K. 2021. Assessment of stream macroinvertebrate communities with eDNA is not congruent with tissue-based metabarcoding. *Molecular Ecology*, **30**(**13**):3239–3251. doi: 10.1111/mec.15597.

González-Tortuero, E., Rusek, J., Petrusek, A., Gießler, S., Lyras, D., Grath, S., Castro-Monzón, F., and Wolinska, J. 2015. The Quantification of Representative Sequences pipeline for amplicon sequencing: case study on within-population ITS1 sequence variation in a microparasite infecting Daphnia. *Molecular Ecology Resources*, **15**(**6**):1385–1395. doi: 10.1111/1755-0998.12396.

Goodwin, K. D., Thompson, L. R., Duarte, B., Kahlke, T., Thompson, A. R., Marques, J. C., and Caçador, I., **2017**. DNA sequencing as a tool to monitor marine ecological status. doi:

10.3389/fmars.2017.00107.

Gratton, P., Marta, S., Bocksberger, G., Winter, M., Trucchi, E., and Kühl, H. 2017. A world of sequences: can we use georeferenced nucleotide databases for a robust automated phylogeography? *Journal of Biogeography*, **44**(**2**):475–486. doi: 10.1111/jbi.12786.

Guardiola, M., Uriz, M. J., Taberlet, P., Coissac, E., Wangensteen, O. S., and Turon, X. 2015. Deep-Sea, Deep-Sequencing: Metabarcoding Extracellular DNA from Sediments of Marine Canyons. *PloS one*, **10** (**10**):e0139633. doi: 10.1371/journal.pone.0153836.

Guardiola, M., Wangensteen, O. S., Taberlet, P., Coissac, E., Uriz, M. J., and Turon, X. 2016. Spatio-temporal monitoring of deep-sea communities using metabarcoding of sediment DNA and RNA. *PeerJ*, **4**: e2807. doi: 10.7717/peerj.2807.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., Decelle, J., del Campo, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Holzmann, M., Kooistra, W. H., Lara, E., Le Bescot, N., Logares, R., Mahé, F., Massana, R., Montresor, M., Morard, R., Not, F., Pawlowski, J., Probert, I., Sauvadet, A.-L., Siano, R., Stoeck, T., Vaulot, D., Zimmermann, P., and Christen, R. 2012. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, **41**(**D1**):D597–D604. doi: 10.1093/nar/gks1160.

Günther, B., Knebelsberger, T., Neumann, H., Laakmann, S., and Martínez Arbizu, P. 2018. Metabarcoding of marine environmental DNA based on mitochondrial and nuclear genes. *Scientific Reports*, **8**(**1**): 14822. doi: 10.1038/s41598-018-32917-x.

Hajibabaei, M., Baird, D. J., Fahner, N. A., Beiko, R., and Golding, G. B. 2016. A new way to contemplate Darwin's tangled bank: how DNA barcodes are reconnecting biodiversity science and biomonitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **371**(**1702**):20150330. doi: 10.1098/rstb.2015.0330.

Hajibabaei, M., Porter, T. M., Robinson, C. V., Baird, D. J., Shokralla, S., and Wright, M. T. 2019. Watered-down biodiver-

sity? A comparison of metabarcoding results from DNA extracted from matched water and bulk tissue biomonitoring samples. *PLoS ONE*, **14** (**12**):e0225409. doi: 10.1371/journal.pone.0225409.

Halpern, B. S., Walbridge, S., Selkoe, K. A., Kappel, C. V., Micheli, F., D'Agrosa, C., Bruno, J. F., Casey, K. S., Ebert, C., Fox, H. E., Fujita, R., Heinemann, D., Lenihan, H. S., Madin, E. M., Perry, M. T., Selig, E. R., Spalding, M., Steneck, R., and Watson, R. 2008. A global map of human impact on marine ecosystems. *Science*, **319**(**5865**):948–952. doi: 10.1126/science.1149345.

Hao, X., Jiang, R., and Chen, T. 2011. Clustering 16S rRNA for OTU prediction: A method of unsupervised Bayesian clustering. *Bioinformatics*, **27**(**5**):611–618. doi: 10.1093/bioinformatics/btq725.

Hardy, S. M., Carr, C. M., Hardman, M., Steinke, D., Corstorphine, E., and Mah, C. 2011. Biodiversity and phylogeography of Arctic marine fauna: insights from molecular tools. *Marine Biodiversity*, **41**(**1**):195–210. doi: 10.1007/s12526-010-0056-x.

Hart, M. W. and Marko, P. B. 2010. It's About Time: Divergence, Demography, and the Evolution of Developmental Modes in Marine Invertebrates. *Integrative and Comparative Biology*, **50**(**4**):643–661. doi: 10.1093/icb/icq068.

Hausser, J. and Strimmer, K. 2009. Entropy Inference and the James-Stein Estimator, with Application to Nonlinear Gene Association Networks Korbinian Strimmer. *Journal of Machine Learning Research*, **10**:1469–1484.

Haye, P. A., Segovia, N. I., Muñoz-Herrera, N. C., Gálvez, F. E., Martínez, A., Meynard, A., Pardo-Gandarillas, M. C., Poulin, E., and Faugeron, S. 2014. Phylogeographic Structure in Benthic Marine Invertebrates of the Southeast Pacific Coast of Chile with Differing Dispersal Potential. *PLOS ONE*, **9**(**2**):e88613. doi: 10.1371/journal.pone.0088613.

Hebert, P. D., Cywinska, A., Ball, S. L., and DeWaard, J. R. 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, **270**:313–321. doi: 10.1098/rspb.2002.2218.

Helmuth, B., Mieszkowska, N., Moore, P., and Hawkins, S. J.

2006. Living on the Edge of Two Changing Worlds: Forecasting the Responses of Rocky Intertidal Ecosystems to Climate Change. *Annual Review of Ecology, Evolution, and Systematics*, **37**:373–404. doi: 10.1146/annurev.ecolsys.37.091305.110149.

Hering, D., Borja, Á., Jones, J., Pont, D., Boets, P., Bouchez, A., Bruce, K., Drakare, S., Hänfling, B., Kahlert, M., Leese, F., Meissner, K., Mergen, P., Reyjol, Y., Segurado, P., Vogler, A., and Kelly, M. 2018. Implementation options for DNA-based identification into ecological status assessment under the European Water Framework Directive. *Water Research*, **138**:192–205. doi: 10.1016/j.watres.2018.03.003.

Holman, L. E., de Bruyn, M., Creer, S., Carvalho, G. R., Robidart, J., and Rius, M. 2019. Detection of introduced and resident marine species using environmental DNA metabarcoding of sediment and water. *Scientific Reports*, **9**(**1**):11559. doi: 10.1038/s41598-019-47899-7.

Holman, L. E., de Bruyn, M., Creer, S., Carvalho, G., Robidart, J., and Rius, M. 2021. Animals, protists and bacteria share marine biogeographic patterns. *Nature Ecology and Evolution*, **5**(**6**):738–746. doi: 10.1038/s41559-021-01439-7.

Jacobs-Palmer, E., Gallego, R., Ramón-Laca, A., Kunselman, E., Cribari, K., Horwith, M., and Kelly, R. P. 2020. A halo of reduced dinoflagellate abundances in and around eelgrass beds. *PeerJ*, **8**:e8869. doi: 10.7717/peerj.8869.

Jeunen, G., Lamare, M. D., Knapp, M., Spencer, H. G., Taylor, H. R., Stat, M., Bunce, M., and Gemmell, N. J. 2019. Water stratification in the marine biome restricts vertical environmental DNA (eDNA) signal dispersal. *Environmental DNA*, page edn3.49. doi: 10.1002/edn3.49.

Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., Kitching, R., Dolman, P. M., Woodcock, P., Edwards, F. A., Larsen, T. H., Hsu, W. W., Benedick, S., Hamer, K. C., Wilcove, D. S., Bruce, C., Wang, X., Levi, T., Lott, M., Emerson, B. C., and Yu, D. W. 2013. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*,

16:1245–1257. doi: 10.1111/ele.12162.

Jost, L. 2008. GST and its relatives do not measure differentiation. *Molecular Ecology*, **17**(**18**):4015–4026. doi: 10.1111/j.1365-294X.2008.03887.x.

Kelly, R. P. and Palumbi, S. R. 2010. Genetic Structure Among 50 Species of the Northeastern Pacific Rocky Intertidal Community. *PLOS ONE*, **5**(**1**):e8594. doi: 10.1371/journal.pone.0008594.

Kelly, R. P., Port, J. A., Yamahara, K. M., and Crowder, L. B. 2014a. Using Environmental DNA to Census Marine Fishes in a Large Mesocosm. *PLoS ONE*, **9**(**1**):e86175. doi: 10.1371/journal.pone.0086175.

Kelly, R. P., Port, J. A., Yamahara, K. M., Martone, R. G., Lowell, N., Thomsen, P. F., Mach, M. E., Bennett, M., Prahler, E., Caldwell, M. R., and Crowder, L. B. 2014b. Harnessing DNA to improve environmental management. *Science*, **344**(**6191**):1455–1456. doi: 10.1126/science.1251156.

Kelly, R. P., Shelton, A. O., and Gallego, R. 2019. Understanding PCR Processes to Draw Meaningful Conclusions from Environmental DNA Studies. *Scientific Reports*, **9**(**1**):1–14. doi: 10.1038/s41598-019-48546-x.

Kemp, J., López-Baucells, A., Rocha, R., Wangensteen, O. S., Andriatafika, Z., Nair, A., and Cabeza, M. 2019. Bats as potential suppressors of multiple agricultural pests: A case study from Madagascar. *Agriculture, Ecosystems & Environment*, **269**:88–96. doi: 10.1016/j.agee.2018.09.027.

Kivelä, M., Arnaud-Haond, S., and Saramäki, J. 2015. EDENetworks: A user-friendly software to build and analyse networks in biogeography, ecology and population genetics. *Molecular Ecology Resources*, **15**(**1**):117–122. doi: 10.1111/1755-0998.12290.

Knowles, L. L. 2004. The burgeoning field of statistical phylogeography. *Journal of Evolutionary Biology*, **17**(**1**):1–10. doi: 10.1046/j.1420-9101.2003.00644.x.

Kõljalg, U., Nilsson, R. H., Abarenkov, K., Tedersoo, L., Taylor, A. F., Bahram, M., Bates, S. T., Bruns, T. D., Bengtsson-Palme, J., Callaghan, T. M., Douglas, B., Drenkhan, T., Eberhardt, U., Dueñas, M., Grebenc, T., Griffith, G. W., Hartmann, M., Kirk,

P. M., Kohout, P., Larsson, E., Lindahl, B. D., Lücking, R., Martín, M. P., Matheny, P. B., Nguyen, N. H., Niskanen, T., Oja, J., Peay, K. G., Peintner, U., Peterson, M., Põldmaa, K., Saag, L., Saar, I., Schüßler, A., Scott, J. A., Senés, C., Smith, M. E., Suija, A., Taylor, D. L., Telleria, M. T., Weiss, M., and Larsson, K. H. 2013. Towards a unified paradigm for sequence-based identification of fungi. *Molecular Ecology*, **22**(**21**):5271–5277. doi: 10.1111/mec.12481.

Kopylova, E., Navas-Molina, J. A., Mercier, C., Xu, Z. Z., Mahé, F., He, Y., Zhou, H.-W., Rognes, T., Caporaso, J. G., and Knight, R. 2016. Open-Source Sequence Clustering Methods Improve the State Of the Art. *mSystems*, **1**(**1**):e00003–15. doi: 10.1128/mSystems.00003-15.

Koziol, A., Stat, M., Simpson, T., Jarman, S., DiBattista, J. D., Harvey, E. S., Marnane, M., McDonald, J., and Bunce, M. 2019. Environmental DNA metabarcoding studies are critically affected by substrate selection. *Molecular Ecology Resources*, **19**(**2**):366–376. doi: 10.1111/1755-0998.12971.

Krehenwinkel, H., Pomerantz, A., and Prost, S. 2019. Genetic Biomonitoring and Biodiversity Assessment Using Portable Sequencing Technologies: Current Uses and Future Directions. *Genes*, **10**(**11**):858. doi: 10.3390/genes10110858.

Laroche, O., Kersten, O., Smith, C. R., and Goetze, E. 2020. Environmental DNA surveys detect distinct metazoan communities across abyssal plains and seamounts in the western Clarion Clipperton Zone. *Molecular Ecology*, **29**(**23**):4588–4604. doi: 10.1111/mec.15484.

Leduc, N., Lacoursière-Roussel, A., Howland, K. L., Archambault, P., Sevellec, M., Normandeau, E., Dispas, A., Winkler, G., McKindsey, C. W., Simard, N., and Bernatchez, L. 2019. Comparing eDNA metabarcoding and species collection for documenting Arctic metazoan biodiversity. *Environmental DNA*, **1**(**4**):342–358. doi: 10.1002/edn3.35.

Leese, F., Bouchez, A., Abarenkov, K., Altermatt, F., Borja, Á., Bruce, K., Ekrem, T., Čiampor, F., Čiamporová-Zaťovičová, Z.,

Costa, F. O., Duarte, S., Elbrecht, V., Fontaneto, D., Franc, A., Geiger, M. F., Hering, D., Kahlert, M., Kalamujić Stroil, B., Kelly, M., Keskin, E., Liska, I., Mergen, P., Meissner, K., Pawlowski, J., Penev, L., Reyjol, Y., Rotter, A., Steinke, D., van der Wal, B., Vitecek, S., Zimmermann, J., and Weigand, A. M. 2018. Why We Need Sustainable Networks Bridging Countries, Disciplines, Cultures and Generations for Aquatic Biomonitoring 2.0: A Perspective Derived From the DNAqua-Net COST Action. In *Advances in Ecological Research*, volume 58, pages 63–99. Academic Press Inc. doi: 10.1016/bs.aecr.2018.01.001.

Leray, M. and Knowlton, N. 2015. DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences of the United States of America*, **112**(**7**):2076–81. doi: 10.1073/pnas.1424997112.

Leray, M. and Knowlton, N., 2016. Censusing marine eukaryotic diversity in the twenty-first century. doi: 10.1098/rstb.2015.0331.

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., Boehm, J. T., and Machida, R. J. 2013. A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, **10**(**1**):34. doi: 10.1186/1742-9994-10-34.

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E., Patrinos, A., Richards, S., Castilla-Rubio, J. C., van Sluys, M. A., Soltis, P. S., Xu, X., Yang, H., and Zhang, G. 2018. Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences of the United States of America*, **115** (**17**):4325–4333. doi: 10.1073/pnas.1720115115.

L'Helguen, S., Le Corre, P., Madec, C., and Morin, P. 2002. New and regenerated production in the Almeria-Oran front area, eastern Alboran Sea. *Deep-Sea Research Part I*, **49**(**1**):83–99. doi: 10.1016/

S0967-0637(01)00044-9.

Lodge, D. M., Turner, C. R., Jerde, C. L., Barnes, M. A., Chadderton, L., Egan, S. P., Feder, J. L., Mahon, A. R., and Pfrender, M. E. 2012. Conservation in a cup of water: estimating biodiversity and population abundance from environmental DNA. *Molecular Ecology*, **21**(**11**):2555–2558. doi: 10.1111/j.1365-294X.2012.05600.x.

Longhurst, A. 1998. *Ecological geography of the sea*. Longhurst, Alan. ISBN 0-12-455558-6.

Lopez, J. V., Yuhki, N., Masuda, R., Modi, W., and O'Brien, S. J. 1994. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *Journal of Molecular Evolution*, **39**(**2**):174–190. doi: 10.1007/BF00163806.

Macher, J.-N., Vivancos, A., Piggott, J. J., Centeno, F. C., Matthaei, C. D., and Leese, F. 2018. Comparison of environmental DNA and bulk-sample metabarcoding using highly degenerate cytochrome c oxidase I primers. *Molecular Ecology Resources*, **18**(**6**): 1456–1468. doi: 10.1111/1755-0998.12940.

Macheriotou, L., Guilini, K., Bezerra, T. N., Tytgat, B., Nguyen, D. T., Phuong Nguyen, T. X., Noppe, F., Armenteros, M., Boufahja, F., Rigaux, A., Vanreusel, A., and Derycke, S. 2019. Metabarcoding free-living marine nematodes using curated 18S and CO1 reference sequence databases for species-level taxonomic assignments. *Ecology and Evolution*, **9**(**3**):1211–1226. doi: 10.1002/ece3.4814.

Macías-Hernández, N., Athey, K., Tonzo, V., Wangensteen, O. S., Arnedo, M., and Harwood, J. D. 2018. Molecular gut content analysis of different spider body parts. *PLOS ONE*, **13**(**5**):e0196589. doi: 10.1371/journal.pone.0196589.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. 2014. Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*, **2**:e593. doi: 10.7717/peerj.593.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. 2015. Swarmv2: Highly-scalable and high-resolution amplicon clustering. *PeerJ*, **2015**(**12**):e1420. doi: 10.7717/peerj.1420.

Mahé, F., Czech, L., Stamatakis, A., Quince, C., De Vargas, C.,

Dunthorn, M., and Rognes, T. 2021. Swarm v3: towards tera-scale amplicon clustering. *Bioinformatics*, **38**(**1**):267–269. doi: 10.1093/bioinformatics/btab493.

Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., Chambouvet, A., Christen, R., Claverie, J. M., Decelle, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Forn, I., Forster, D., Guillou, L., Jaillon, O., Kooistra, W. H., Logares, R., Mahé, F., Not, F., Ogata, H., Pawlowski, J., Pernice, M. C., Probert, I., Romac, S., Richards, T., Santini, S., Shalchian-Tabrizi, K., Siano, R., Simon, N., Stoeck, T., Vaulot, D., Zingone, A., and de Vargas, C. 2015. Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. *Environmental Microbiology*, **17**(**10**):4035–4049. doi: 10.1111/1462-2920.12955.

Mathieu, C., Hermans, S. M., Lear, G., Buckley, T. R., Lee, K. C., and Buckley, H. L. 2020. A Systematic Review of Sources of Variability and Uncertainty in eDNA Data for Environmental Monitoring. *Frontiers in Ecology and Evolution*, **8**:1–14. doi: 10.3389/fevo.2020.00135.

McGee, K. M., Robinson, C. V., and Hajibabaei, M. 2019. Gaps in DNA-Based Biomonitoring Across the Globe. *Frontiers in Ecology and Evolution*, **7**:337. doi: 10.3389/fevo.2019.00337.

McKerns, M., Strand, L., Sullivan, T., Fang, A., and Aivazis, M. 2011. Building a Framework for Predictive Science. In *Proceedings of the 10th Python in Science Conference*, pages 76–86. SciPy. doi: 10.25080/Majora-ebaa42b7-00d.

Meyer, C. P. and Paulay, G. 2005. DNA barcoding: Error rates based on comprehensive sampling. *PLoS Biology*, **3**(**12**):2229–2238. doi: 10.1371/journal.pbio.0030422.

Miya, M. 2022. Environmental DNA Metabarcoding: A Novel Method for Biodiversity Monitoring of Marine Fish Communities. *Annual Review of Marine Science*, **14**(**1**):161–185. doi: 10.1146/annurev-marine-041421-082251.

Miya, M., Gotoh, R. O., and Sado, T., 2020. MiFish metabarcoding: a high-throughput approach for simultaneous detection of multiple fish species from environmental DNA and other samples. doi:

10.1007/s12562-020-01461-x.

Mugnai, F., Meglécz, E., Costantini, F., Abbiati, M., Bavestrello, G., Bertasi, F., Bo, M., Capa, M., Chenuil, A., Colangelo, M. A., De Clerck, O., Gutiérrez, J. M., Lattanzi, L., Leduc, M., Martin, D., Matterson, K. O., Mikac, B., Plaisance, L., Ponti, M., Riesgo, A., Rossi, V., Turicchia, E., Waeschenbach, A., and Wangensteen, O. S. 2021. Are well-studied marine biodiversity hotspots still blackspots for animal barcoding? *Global Ecology and Conservation*, **32**:e01909. doi: 10.1016/j.gecco.2021.e01909.

Naciri, M., Lemaire, C., Borsa, P., and Bonhomme, F. 1999. Genetic study of the Atlantic/Mediterranean transition in sea bass (Dicentrarchus labrax). *Journal of Heredity*, **90**(**6**):591–596. doi: 10.1093/jhered/90.6.591.

Nagler, M., Podmirseg, S. M., Ascher-Jenull, J., Sint, D., and Traugott, M. 2022. Why eDNA fractions need consideration in biomonitoring. *Molecular Ecology Resources*. doi: 10.1111/1755-0998.13658.

Nakao, R., Miyata, R., Nakamura, ·. N., Muramatsu, M., Okamura, ·. H., Imamura, F., and Akamatsu, Y. 2022. Development of environmental DNA chip for monitoring the invasive alien fishes in dam reservoirs. *Landscape and Ecological Engineering 2022*, **1**:1–9. doi: 10.1007/S11355-022-00513-X.

Nearing, J. T., Douglas, G. M., Comeau, A. M., and Langille, M. G. 2018. Denoising the Denoisers: An independent evaluation of microbiome sequence error- correction approaches. *PeerJ*, **2018**(**8**):e5364. doi: 10.7717/peerj.5364.

Nguyen, B. N., Shen, E. W., Seemann, J., Correa, A. M., O'Donnell, J. L., Altieri, A. H., Knowlton, N., Crandall, K. A., Egan, S. P., McMillan, W. O., and Leray, M. 2020. Environmental DNA survey captures patterns of fish and invertebrate diversity across a tropical seascape. *Scientific Reports*, **10**(**1**):1–14. doi: 10.1038/s41598-020-63565-9.

Nielsen, E. S., Beger, M., Henriques, R., Selkoe, K. A., and von der Heyden, S. 2017. Multispecies genetic objectives in spatial conservation planning. *Conservation Biology*, **31**(**4**):872–882. doi: 10.1111/cobi.12875.

O'Donnell, J. L., Kelly, R. P., Shelton, A. O., Samhouri, J. F.,

Lowell, N. C., Williams, G. D., O'Donnell, J. L., Kelly, R. P., Shelton, A. O., Samhouri, J. F., Lowell, N. C., and Williams, G. D. 2017. Spatial distribution of environmental DNA in a nearshore marine habitat. *PeerJ*, **5**(**2**):e3044. doi: 10.7717/peerj.3044.

Ojeda, V., Serra, B., Lagares, C., Rojo-Francàs, E., Sellés, M., Marco-Herrero, E., García, E., Farré, M., Arenas, C., Abelló, P., and Mestres, F. 2022. Interannual fluctuations in connectivity among crab populations (Liocarcinus depurator) along the Atlantic-Mediterranean transition. *Scientific Reports*, **12**(**1**):1–14. doi: 10.1038/s41598-022-13941-4.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., and Wagner, H., 2019. vegan: Community Ecology Package. R package version 2.5-7. https://CRAN.R-project.org/package=vegan.

Olesen, S. W., Duvallet, C., and Alm, E. J. 2017. dbOTU3: A new implementation of distribution-based OTU calling. *PLOS ONE*, **12**(**5**): e0176335. doi: 10.1371/journal.pone.0176335.

Öpik, M., Vanatoa, A., Vanatoa, E., Moora, M., Davison, J., Kalwij, J. M., Reier, Ü., and Zobel, M. 2010. The online database MaarjAM reveals global and ecosystemic distribution patterns in arbuscular mycorrhizal fungi (Glomeromycota). *New Phytologist*, **188**(**1**):223–241. doi: 10.1111/j.1469-8137.2010.03334.x.

O'Rourke, D. R., Bokulich, N. A., Jusino, M. A., MacManes, M. D., and Foster, J. T. 2020. A total crapshoot? Evaluating bioinformatic decisions in animal diet metabarcoding analyses. *Ecology and Evolution*, **10**(**18**):9721–9739. doi: 10.1002/ece3.6594.

Orr, M. C. C., Ascher, J. S., Bai, M., Chesters, D., and Zhu, C.-D. 2020. Three questions: How can taxonomists survive and thrive worldwide? *Megataxa*, **1**(**1**):19–27. doi: 10.11646/megataxa.1.1.4.

Pagenkopp Lohan, K. M., Fleischer, R. C., Torchin, M. E., and Ruiz, G. M. 2017. Protistan Biogeography: A Snapshot Across a Major Shipping Corridor Spanning Two Oceans. *Protist*, **168**(**2**):183–196. doi: 10.1016/j.protis.2016.12.003.

Pagès, H., Aboyoun, P., Gentleman, R., and DebRoy, S., **2008**. Biostrings: Efficient manipulation of biological strings. R package version 2.58.0. https://bioconductor.org/packages/Biostrings.

Palacín, C., Giribet, G., Carner, S., Dantart, L., and Turon, X. 1998. Low densities of sea urchins influence the structure of algal assemblages in the western Mediterranean. *Journal of Sea Research*, **39** (**3-4**):281–290. doi: 10.1016/S1385-1101(97)00061-0.

Paradis, E. 2010. pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics*, **26**:419–420. doi: 10.1093/bioinformatics/btp696.

Parsons, K. M., Everett, M., Dahlheim, M., and Park, L. 2018. Water, water everywhere: environmental DNA can unlock population structure in elusive marine species. *Royal Society Open Science*, **5**(**8**): 180537. doi: 10.1098/rsos.180537.

Pascual, M., Rives, B., Schunter, C., and Macpherson, E. 2017. Impact of life history traits on gene flow: A multispecies systematic review across oceanographic barriers in the Mediterranean Sea. *PLOS ONE*, **12** (**5**):e0176419. doi: 10.1371/journal.pone.0176419.

Patarnello, T., Volckaert, F. A., and Castilho, R. 2007. Pillars of Hercules: is the Atlantic–Mediterranean transition a phylogeographical break? *Molecular Ecology*, **16**(**21**):4426–4444. doi: 10.1111/j.1365-294X.2007.03477.x.

Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., Borja, Á., Bouchez, A., Cordier, T., Domaizon, I., Feio, M. J., Filipe, A. F., Fornaroli, R., Graf, W., Herder, J., van der Hoorn, B., Iwan Jones, J., Sagova-Mareckova, M., Moritz, C., Barquín, J., Piggott, J. J., Pinna, M., Rimet, F., Rinkevich, B., Sousa-Santos, C., Specchia, V., Trobajo, R., Vasselon, V., Vitecek, S., Zimmerman, J., Weigand, A., Leese, F., and Kahlert, M. 2018. The future of biotic indices in the ecogenomic era: Integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. *Science of the Total Environment*, **637-638**:1295–1310. doi: 10.1016/j.scitotenv.2018.05.002.

Pawlowski, J., Apothéloz-Perret-Gentil, L., and Altermatt, F.

2020. Environmental DNA: What's behind the term? Clarifying the terminology and recommendations for its future use in biomonitoring. *Molecular Ecology*, **29**(**22**):4258–4264. doi: 10.1111/mec.15643.

Pawlowski, J., Bruce, K., Panksep, K., Aguirre, F., Amalfitano, S., Apothéloz-Perret-Gentil, L., Baussant, T., Bouchez, A., Carugati, L., Cermakova, K., Cordier, T., Corinaldesi, C., Costa, F. O., Danovaro, R., Dell'Anno, A., Duarte, S., Eisendle, U., Ferrari, B., Frontalini, F., Frühe, L., Haegerbaeumer, A., Kisand, V., Krolicka, A., Lanzén, A., Leese, F., Lejzerowicz, F., Lyautey, E., Maček, I., Sagova-Marečková, M., Pearman, J. K., Pochon, X., Stoeck, T., Vivien, R., Weigand, A. M., and Fazi, S. 2022. Environmental DNA metabarcoding for benthic monitoring: A review of sediment sampling and DNA extraction methods. *Science of The Total Environment*, **818**:151783. doi: 10.1016/j.scitotenv.2021.151783.

Pearman, J. K., Leray, M., Villalobos, R., Machida, R. J., Berumen, M. L., Knowlton, N., and Carvalho, S. 2018. Cross-shelf investigation of coral reef cryptic benthic organisms reveals diversity patterns of the hidden majority. *Scientific Reports*, **8**(**1**):1–17. doi: 10.1038/s41598-018-26332-5.

Pearman, J. K., Aylagas, E., Voolstra, C. R., Anlauf, H., Villalobos, R., and Carvalho, S. 2019. Disentangling the complex microbial community of coral reefs using standardized Autonomous Reef Monitoring Structures (ARMS). *Molecular Ecology*, **28**(**15**):3496–3507. doi: 10.1111/mec.15167.

Pearman, J. K., Chust, G., Aylagas, E., Villarino, E., Watson, J. R., Chenuil, A., Borja, A., Cahill, A. E., Carugati, L., Danovaro, R., David, R., Irigoien, X., Mendibil, I., Moncheva, S., Rodríguez-Ezpeleta, N., Uyarra, M. C., and Carvalho, S. 2020. Pan-regional marine benthic cryptobiome biodiversity patterns revealed by metabarcoding Autonomous Reef Monitoring Structures. *Molecular Ecology*, **29**(**24**):4882–4897. doi: 10.1111/mec.15692.

Pedro, P. M., Piper, R., Bazilli Neto, P., Cullen, L., Dropa, M., Lorencao, R., Matté, M. H., Rech, T. C., Rufato, M. O., Silva, M., and Turati, D. T. 2017. Metabarcoding Analyses Enable

Differentiation of Both Interspecific Assemblages and Intraspecific Divergence in Habitats With Differing Management Practices. *Environmental Entomology*, **46**(**6**):1381–1389. doi: 10.1093/ee/nvx166.

Peng, X. and Dorman, K. S. 2020. AmpliCI: a high-resolution model-based approach for denoising Illumina amplicon data. *Bioinformatics*, **36** (**21**):5151–5158. doi: 10.1093/bioinformatics/btaa648.

Pentinsaari, M., Salmela, H., Mutanen, M., and Roslin, T. 2016. Molecular evolution of a widely-adopted taxonomic marker (COI) across the animal tree of life. *Scientific reports*, **6**. doi: 10.1038/srep35275.

Pérez-Portela, R., Almada, V., and Turon, X. 2013. Cryptic speciation and genetic structure of widely distributed brittle stars (Ophiuroidea) in Europe. *Zoologica Scripta*, **42**(**2**):151–169. doi: 10.1111/j.1463-6409.2012.00573.x.

Peters, L., Spatharis, S., Dario, M. A., Dwyer, T., Roca, I. J. T., Kintner, A., Kanstad-Hanssen, Ø., Llewellyn, M. S., and Praebel, K. 2018. Environmental DNA: A New Low-Cost Monitoring Tool for Pathogens in Salmonid Aquaculture. *Frontiers in Microbiology*, **9**:3009. doi: 10.3389/fmicb.2018.03009.

Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J. L., and Mayer, G. 2018. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Scientific Reports*, **8**(**1**):1–14. doi: 10.1038/s41598-018-29325-6.

Pichler, M. and Hartig, F. 2021. A new joint species distribution model for faster and more accurate inference of species associations from big community data. *Methods in Ecology and Evolution*, **12**(**11**):2159–2173. doi: 10.1111/2041-210X.13687.

Piñol, J., Senar, M. A., and Symondson, W. O. C. 2019. The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative. *Molecular Ecology*, **28**(**2**):407–419. doi: 10.1111/mec.14776.

Pinot, J. M., López-Jurado, J. L., and Riera, M. 2002. The CANALES experiment (1996-1998). Interannual, seasonal, and mesoscale variability of the circulation in the Balearic Channels. *Progress in Oceanography*, **55**(**3-4**):335–370. doi: 10.1016/S0079-6611(02)00139-8.

Pitz, K. J., Guo, J., Johnson, S. B., Campbell, T. L., Zhang, H., Vrijenhoek, R. C., Chavez, F. P., and Geller, J. 2020. Zooplankton biogeographic boundaries in the California Current System as determined from metabarcoding. *PLOS ONE*, **15**(**6**):e0235159. doi: 10.1371/journal.pone.0235159.

Pochon, X., Bott, N. J., Smith, K. F., and Wood, S. A. 2013. Evaluating Detection Limits of Next-Generation Sequencing for the Surveillance and Monitoring of International Marine Pests. *PLoS ONE*, **8**(**9**):e73935. doi: 10.1371/journal.pone.0073935.

Port, J. A., O'Donnell, J. L., Romero-Maraccini, O. C., Leary, P. R., Litvin, S. Y., Nickols, K. J., Yamahara, K. M., and Kelly, R. P. 2016. Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA. *Molecular Ecology*, **25**(**2**):527–541. doi: 10.1111/mec.13481.

Porter, J. and Zhang, L. 2017. InfoTrim: A DNA Read Quality Trimmer Using Entropy. *bioRxiv*, page 201442. doi: 10.1101/201442.

Porter, T. M. and Hajibabaei, M. 2018a. Automated high throughput animal CO1 metabarcode classification. *Scientific Reports*, **8**(**1**). doi: 10.1038/s41598-018-22505-4.

Porter, T. M. and Hajibabaei, M. 2018b. Over 2.5 million COI sequences in GenBank and growing. *PLOS ONE*, **13**(**9**):e0200177. doi: 10.1371/journal.pone.0200177.

Porter, T. M. and Hajibabaei, M. 2018c. Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Molecular Ecology*, **27**(**2**):313–338. doi: 10.1111/mec.14478.

Porter, T. M. and Hajibabaei, M. 2020. Putting COI Metabarcoding in Context: The Utility of Exact Sequence Variants (ESVs) in Biodiversity Analysis. *Frontiers in Ecology and Evolution*, **8**:248. doi: 10.3389/fevo.2020.00248.

Porter, T. M. and Hajibabaei, M. 2021. Profile hidden Markov model sequence analysis can help remove putative pseudogenes from DNA barcoding and metabarcoding datasets. *BMC Bioinformatics*, **22**(**1**). doi: 10.1186/s12859-021-04180-x.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza,

P., Peplies, J., and Glöckner, F. O. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, **41**(**D1**):D590–D596. doi: 10.1093/nar/gks1219.

R Core Team, **2020**. R: A Language and Environment for Statistical Computing.

Ransome, E., Geller, J. B., Timmers, M., Leray, M., Mahardini, A., Sembiring, A., Collins, A. G., and Meyer, C. P. 2017. The importance of standardization for biodiversity comparisons: A case study using autonomous reef monitoring structures (ARMS) and metabarcoding to measure cryptic diversity on Mo'orea coral reefs, French Polynesia. *PLOS ONE*, **12**(**4**):e0175066. doi: 10.1371/journal.pone.0175066.

Ratnasingham, S. and Hebert, P. D. 2007. BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular ecology notes*, **7**(**3**):355–364. doi: 10.1111/j.1471-8286.2007.01678.x.

Ratnasingham, S. and Hebert, P. D. 2013. A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. *PLoS ONE*, **8**(**7**):e66213. doi: 10.1371/journal.pone.0066213.

Reaka-Kudla, M. L. 1997. The global biodiversity of coral reefs: a comparison with rainforests. Biodiversity II: understanding and protecting our natural resources. In *Biodiversity II: Understanding and Protecting Our Biological Resources*, pages 83–108. Joseph Henry/National Academy Press, Washington, DC. ISBN 0309052270.

Reitzel, A. M., Herrera, S., Layden, M. J., Martindale, M. Q., and Shank, T. M. 2013. Going where traditional markers have not gone before: utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics. *Molecular Ecology*, **22**(**11**): 2953–2970. doi: 10.1111/mec.12228.

Rex, M. A. and Etter, R. J. 2010. *Deep-sea biodiversity. Pattern and scale.* Harvard University Press, Cambridge, Mas- sachusetts, USA. ISBN 978-0-674-03607-9.

Rey, A., Basurko, O. C., and Rodriguez-Ezpeleta, N. 2020. Considerations for metabarcoding-based port biological baseline surveys aimed at marine nonindigenous species monitoring and risk assessments. *Ecology and Evolution*, **10**(**5**):2452–2465. doi: 10.1002/ece3.6071.

Riddle, B. R., Dawson, M. N., Hadly, E. A., Hafner, D. J., Hickerson, M. J., Mantooth, S. J., and Yoder, A. D. 2008. The role of molecular genetics in sculpting the future of integrative biogeography. *Progress in Physical Geography: Earth and Environment*, **32**(**2**):173–202. doi: 10.1177/0309133308093822.

Rodriguez-Ezpeleta, N., Morissette, O., Bean, C. W., Manu, S., Banerjee, P., Lacoursière-Roussel, A., Beng, K. C., Alter, S. E., Roger, F., Holman, L. E., Stewart, K. A., Monaghan, M. T., Mauvisseau, Q., Mirimin, L., Wangensteen, O. S., Antognazza, C. M., Helyar, S. J., de Boer, H., Monchamp, M. E., Nijland, R., Abbott, C. L., Doi, H., Barnes, M. A., Leray, M., Hablützel, P. I., and Deiner, K. 2021. Trade-offs between reducing complex terminology and producing accurate interpretations from environmental DNA: Comment on "Environmental DNA: What's behind the term?" by Pawlowski et al., (2020). *Molecular Ecology*, **30**(**19**):4601–4605. doi: 10.1111/mec.15942.

Rodríguez-Ezpeleta, N., Zinger, L., Kinziger, A., Bik, H. M., Bonin, A., Coissac, E., Emerson, B. C., Lopes, C. M., Pelletier, T. A., Taberlet, P., and Narum, S. 2021. Biodiversity monitoring using environmental DNA. *Molecular Ecology Resources*, **21**(**5**):1405–1409. doi: 10.1111/1755-0998.13399.

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, **4**:e2584. doi: 10.7717/peerj.2584.

Sales, N. G., Wangensteen, O. S., Carvalho, D. C., and Mariani, S. 2019. Influence of preservation methods, sample medium and sampling time on eDNA recovery in a neotropical river. *Environmental DNA*, **1**(**2**): 119–130. doi: 10.1002/edn3.14.

Salter, I. 2018. Seasonal variability in the persistence of dissolved environmental DNA (eDNA) in a marine system: The role of microbial nutrient limitation. *PLOS ONE*, **13**(**2**):e0192409. doi: 10.1371/journal.pone.0192409.

Salter, I., Joensen, M., Kristiansen, R., Steingrund, P., and Vestergaard, P. 2019. Environmental DNA concentrations are correlated with

regional biomass of Atlantic cod in oceanic waters. *Communications Biology*, **2**(**1**):1–9. doi: 10.1038/s42003-019-0696-8.

Sandström, A., Lundmark, C., Andersson, K., Johannesson, K., and Laikre, L. 2019. Understanding and bridging the conservation-genetics gap in marine conservation. *Conservation Biology*, **33**(**3**):725–728. doi: 10.1111/cobi.13272.

Santoferrara, L. F., Rubin, E., and McManus, G. B. 2018. Global and local DNA (meta)barcoding reveal new biogeography patterns in tintinnid ciliates. *Journal of Plankton Research*, **40**(**3**):209–221. doi: 10.1093/plankt/fby011.

Schirmer, M., D'Amore, R., Ijaz, U. Z., Hall, N., and Quince, C. 2016. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics*, **17**(**1**):125. doi: 10.1186/s12859-016-0976-y.

Schmitt, A. O. and Herzel, H. 1997. Estimating the entropy of DNA sequences. *Journal of Theoretical Biology*, **188**(**3**):369–377. doi: 10.1006/jtbi.1997.0493.

Schultz, J. A. and Hebert, P. D. N. 2022. Do pseudogenes pose a problem for metabarcoding marine animal communities? *Molecular Ecology Resources*. doi: 10.1111/1755-0998.13667.

Schunter, C., Carreras-Carbonell, J., MacPherson, E., TintorÉ, J., Vidal-Vijande, E., Pascual, A., Guidetti, P., and Pascual, M. 2011. Matching genetics with oceanography: directional gene flow in a Mediterranean fish species. *Molecular Ecology*, **20**(**24**):5167–5181. doi: 10.1111/j.1365-294X.2011.05355.x.

Semenov, M. V. 2021. Metabarcoding and Metagenomics in Soil Ecology Research: Achievements, Challenges, and Prospects. *Biology Bulletin Reviews 2021 11:1*, **11**(**1**):40–53. doi: 10.1134/S2079086421010084.

Shade, A., Dunn, R. R., Blowes, S. A., Keil, P., Bohannan, B. J., Herrmann, M., Küsel, K., Lennon, J. T., Sanders, N. J., Storch, D., and Chase, J. 2018. Macroecology to Unite All Life, Large and Small. *Trends in Ecology & Evolution*, **33**(**10**):731–744. doi: 10.1016/j.tree.2018.08.005.

Shaw, J. L. A., Weyrich, L., and Cooper, A. 2017. Using environ-

mental (e)DNA sequencing for aquatic biodiversity surveys: a beginner's guide. *Marine and Freshwater Research*, **68**(**1**):20. doi: 10.1071/MF15361.

Shu, L., Ludwig, A., and Peng, Z. 2020. Standards for Methods Utilizing Environmental DNA for Detection of Fish Species. *Genes*, **11** (**3**):296. doi: 10.3390/genes11030296.

Shum, P. and Palumbi, S. R. 2021. Testing small-scale ecological gradients and intraspecific differentiation for hundreds of kelp forest species using haplotypes from metabarcoding. *Molecular Ecology*, **30**(**13**): 3355–3373. doi: 10.1111/mec.15851.

Shum, P., Barney, B. T., O'Leary, J. K., and Palumbi, S. R. 2019. Cobble community DNA as a tool to monitor patterns of biodiversity within kelp forest ecosystems. *Molecular Ecology Resources*, **19**(**6**):1470–1485. doi: 10.1111/1755-0998.13067.

Siegenthaler, A., Wangensteen, O. S., Benvenuto, C., Campos, J., and Mariani, S. 2019a. DNA metabarcoding unveils multiscale trophic variation in a widespread coastal opportunist. *Molecular Ecology*, **28**(**2**): 232–249. doi: 10.1111/mec.14886.

Siegenthaler, A., Wangensteen, O. S., Soto, A. Z., Benvenuto, C., Corrigan, L., and Mariani, S. 2019b. Metabarcoding of shrimp stomach content: Harnessing a natural sampler for fish biodiversity monitoring. *Molecular Ecology Resources*, **19**(**1**):206–220. doi: 10.1111/1755-0998.12956.

Sigsgaard, E. E., Nielsen, I. B., Bach, S. S., Lorenzen, E. D., Robinson, D. P., Knudsen, S. W., Pedersen, M. W., Jaidah, M. A., Orlando, L., Willerslev, E., Møller, P. R., and Thomsen, P. F. 2016. Population characteristics of a large whale shark aggregation inferred from seawater environmental DNA. *Nature Ecology & Evolution 2016 1:1*, **1**(**1**):1–5. doi: 10.1038/s41559-016-0004.

Sigsgaard, E. E., Torquato, F., Frøslev, T. G., Moore, A. B. M., Sørensen, J. M., Range, P., Ben-Hamadou, R., Bach, S. S., Møller, P. R., and Thomsen, P. F. 2019. Using vertebrate environmental DNA from seawater in biomonitoring of marine habitats. *Conservation Biology*. doi: 10.1111/cobi.13437.

Sigsgaard, E. E., Jensen, M. R., Winkelmann, I. E., Møller, P. R.,

Hansen, M. M., and Thomsen, P. F. 2020. Population-level inferences from environmental DNA—Current status and future perspectives. *Evolutionary Applications*, **13**(**2**):245–262. doi: 10.1111/eva.12882.

Simpson, T. J., Smale, D. A., McDonald, J. I., and Wernberg, T. 2017. Large scale variability in the structure of sessile invertebrate assemblages in artificial habitats reveals the importance of local-scale processes. *Journal of Experimental Marine Biology and Ecology*, **494**: 10–19. doi: 10.1016/j.jembe.2017.05.003.

Singer, G. A. C., Fahner, N. A., Barnes, J. G., McCarthy, A., and Hajibabaei, M. 2019. Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: a case study of eDNA metabarcoding seawater. *Scientific Reports*, **9**(**1**):1–12. doi: 10.1038/s41598-019-42455-9.

Singer, G. A. C., Shekarriz, S., McCarthy, A., Fahner, N., and Hajibabaei, M. 2020. The utility of a metagenomics approach for marine biomonitoring. *bioRxiv*. doi: 10.1101/2020.03.16.993667.

Sinniger, F., Pawlowski, J., Harii, S., Gooday, A. J., Yamamoto, H., Chevaldonné, P., Cedhagen, T., Carvalho, G. R., and Creer, S. 2016. Worldwide analysis of sedimentary DNA reveals major gaps in taxonomic knowledge of deep-sea benthos. *Frontiers in Marine Science*, **3**:92. doi: 10.3389/fmars.2016.00092.

Song, H., Buhay, J. E., Whiting, M. F., and Crandall, K. A. 2008. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the National Academy of Sciences of the United States of America*, **105** (**36**):13486–13491. doi: 10.1073/pnas.0803076105.

Sousa, L. L., Silva, S. M., and Xavier, R. 2019. DNA metabarcoding in diet studies: Unveiling ecological aspects in aquatic and terrestrial ecosystems. *Environmental DNA*, page edn3.27. doi: 10.1002/edn3.27.

Spalding, M. D., Fox, H. E., Allen, G. R., Davidson, N., Ferdaña, Z. A., Finlayson, M., Halpern, B. S., Jorge, M. A., Lombana, A., Lourie, S. A., Martin, K. D., Mcmanus, E., Molnar, J., Recchia, C. A., and Robertson, J. 2007. Marine Ecoregions of the World: A Bioregionalization of Coastal and Shelf Areas. *BioScience*, **57**(**7**):573–583. doi: 10.1641/B570707.

Stat, M., Huggett, M. J., Bernasconi, R., DiBattista, J. D., Berry, T. E., Newman, S. J., Harvey, E. S., and Bunce, M. 2017. Ecosystem biomonitoring with eDNA: metabarcoding across the tree of life in a tropical marine environment. *Scientific Reports*, **7**(**1**):12240. doi: 10.1038/s41598-017-12501-5.

Stefanni, S., Stanković, D., Borme, D., de Olazabal, A., Juretić, T., Pallavicini, A., and Tirelli, V. 2018. Multi-marker metabarcoding approach to study mesozooplankton at basin scale. *Scientific Reports*, **8** (**1**):1–13. doi: 10.1038/s41598-018-30157-7.

Stewart, K. A. 2019. Understanding the effects of biotic and abiotic factors on sources of aquatic environmental DNA. *Biodiversity and Conservation*, **28**(**5**):983–1001. doi: 10.1007/s10531-019-01709-8.

Steyaert, M., Priestley, V., Osborne, O., Herraiz, A., Arnold, R., and Savolainen, V. 2020. Advances in metabarcoding techniques bring us closer to reliable monitoring of the marine benthos. *Journal of Applied Ecology*, **57**(**11**):2234–2245. doi: 10.1111/1365-2664.13729.

Stoeckle, M. Y., Das Mishu, M., and Charlop-Powers, Z. 2018. GoFish: A versatile nested PCR strategy for environmental DNA assays for marine vertebrates. *PLOS ONE*, **13**(**12**):e0198717. doi: 10.1371/journal.pone.0198717.

Taberlet, P., Coissac, E., Hajibabaei, M., and Rieseberg, L. H. 2012a. Environmental DNA. *Molecular Ecology*, **21**(**8**):1789–1793. doi: 10.1111/j.1365-294X.2012.05542.x.

Taberlet, P., Coissac, E., Pompanon, F., Bronchmann, C., and Willerslev, E. 2012b. Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, **21**:2045–2050. doi: 10.1111/j.1365-294X.2012.05470.x.

Taberlet, P., Bonin, A., Zinger, L., and Coissac, E. 2018. *Environmental DNA: For Biodiversity Research and Monitoring*. Oxford Scholarship Online. ISBN 9780198767220. doi: 10.1093/oso/9780198767220.001.0001.

Taboada, S. and Pérez-Portela, R. 2016. Contrasted phylogeographic patterns on mitochondrial DNA of shallow and deep brittle stars across the Atlantic-Mediterranean area. *Scientific Reports*, **6**(**1**):1–13. doi:

10.1038/srep32425.

Tang, C. Q., Leasi, F., Obertegger, U., Kieneke, A., Barraclough, T. G., and Fontaneto, D. 2012. The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(**40**):16208–16212. doi: 10.1073/pnas.1209160109.

Tapolczai, K., Keck, F., Bouchez, A., Rimet, F., Kahlert, M., and Vasselon, V. 2019. Diatom DNA Metabarcoding for Biomonitoring: Strategies to Avoid Major Taxonomical and Bioinformatical Biases Limiting Molecular Indices Capacities. *Frontiers in Ecology and Evolution*, **7**: 409. doi: 10.3389/fevo.2019.00409.

Teske, P. R., Von der Heyden, S., McQuaid, C. D., and Barker, N. P. 2011. A review of marine phylogeography in southern Africa. *South African Journal of Science*, **107**(**5/6**). doi: 10.4102/sajs.v107i5/6.514.

Thiel, M., Macaya, E. C., Acuña, E., Arntz, W. E., Bastias, H., Brokordt, K., Camus, P. A., Castilla, J. C., Castro, L. R., Cortés, M., Dumont, C. P., Escribano, R., Fernandez, M., Gajardo, J. A., Gaymer, C. F., Gomez, I., González, A. E., González, H. E., Haye, P. A., Illanes, J.-E., Iriarte, J. L., Lancellotti, D. A., Luna-Jorquera, G., Luxoro, C., Manriquez, P. H., Marín, V., Muñoz, P., Navarrete, S. A., Perez, E., Poulin, E., Sellanes, J., Sepúlveda, H. H., Stotz, W., Tala, F., Thomas, A., Vargas, C. A., Vasquez, J. A., and Vega, J. A. 2007. The Humboldt Current System of Northern and Central Chile. In Gibson, R. N., Atkinson, R. J. A., and Gordon, J. D. M., editors, *Oceanography and Marine Biology : An annual Review*, volume 45, pages 195–344. Taylor & Francis. doi: 10.1201/9781420050943.ch6.

Thomas, T., Gilbert, J., and Meyer, F. 2012. Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation*, **2**(**1**):1–12. doi: 10.1186/2042-5783-2-3.

Thomsen, P. F., Kielgast, J., Iversen, L. L., Wiuf, C., Rasmussen, M., Gilbert, M. T. P., Orlando, L., and Willerslev, E. 2012. Monitoring endangered freshwater biodiversity using environmental DNA.

Molecular Ecology, **21**(**11**):2565–2573. doi: 10.1111/j.1365-294X.2011.05418.x.

Thomsen, P. F., Møller, P. R., Sigsgaard, E. E., Knudsen, S. W., Jørgensen, O. A., and Willerslev, E. 2016. Environmental DNA from Seawater Samples Correlate with Trawl Catches of Subarctic, Deepwater Fishes. *PLOS ONE*, **11**(**11**):e0165252. doi: 10.1371/journal.pone.0165252.

Tintore, J., Violette, P. E. L., Blade, I., and Cruzado, A. 1988. A Study of an Intense Density Front in the Eastern Alboran Sea: The Almeria–Oran Front. *Journal of Physical Oceanography*, **18**(**10**). doi: 10.1175/1520-0485(1988)018<1384:ASOAID>2.0.CO;2.

Toonen, R. J., Bowen, B. W., Iacchei, M., and Briggs, J. C. 2016. Biogeography, Marine. In Kliman, R. M., editor, *Encyclopedia of Evolutionary Biology*, pages 166–178. Elsevier Inc., Oxford. ISBN 9780128004265. doi: 10.1016/B978-0-12-800049-6.00120-7.

Tsuji, S. and Shibata, N. 2020. Identifying spawning events in fish by observing a spike in environmental DNA concentration after spawning. *Environmental DNA*, page edn3.153. doi: 10.1002/edn3.153.

Tsuji, S., Miya, M., Ushio, M., Sato, H., Minamoto, T., and Yamanaka, H. 2019. Evaluating intraspecific genetic diversity using environmental DNA and denoising approach: A case study using tank water. *Environmental DNA*, page edn3.44. doi: 10.1002/edn3.44.

Tulchinsky, A. Y., Norenburg, J. L., and Turbeville, J. M. 2012. Phylogeography of the marine interstitial nemertean Ototyphlonemertes parmula (Nemertea, Hoplonemertea) reveals cryptic diversity and high dispersal potential. *Marine Biology*, **159**(**3**):661–674. doi: 10.1007/s00227-011-1844-y.

Turon, X., Antich, A., Palacín, C., Præbel, K., and Wangensteen, O. S. 2020. From metabarcoding to metaphylogeography: separating the wheat from the chaff. *Ecological Applications*, **30**(**2**):e02036. doi: 10.1002/eap.2036.

Uchii, K., Doi, H., and Minamoto, T. 2016. A novel environmental DNA approach to quantify the cryptic invasion of non-native genotypes. *Molecular Ecology Resources*, **16**(**2**):415–422. doi: 10.1111/1755-0998.12460.

Vamos, E., Elbrecht, V., and Leese, F. 2017. Short COI markers for freshwater macroinvertebrate metabarcoding. *Metabarcoding and Metagenomics*, **1**:e14625. doi: 10.3897/mbmg.1.14625.

van der Loos, L. M. and Nijland, R. 2021. Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. *Molecular Ecology*, **30**(**13**):3270–3288. doi: 10.1111/mec.15592.

Vellend, M., Lajoie, G., Bourret, A., Múrria, C., Kembel, S. W., and Garant, D. 2014. Drawing ecological inferences from coincident patterns of population- and community-level biodiversity. *Molecular Ecology*, **23**(**12**):2890–2901. doi: 10.1111/mec.12756.

von Ammon, U., Wood, S. A., Laroche, O., Zaiko, A., Lavery, S. D., Inglis, G. J., and Pochon, X. 2019. Linking Environmental DNA and RNA for Improved Detection of the Marine Invasive Fanworm Sabella spallanzanii. *Frontiers in Marine Science*, **6**. doi: 10.3389/fmars.2019.00621.

Wang, X. V., Blades, N., Ding, J., Sultana, R., and Parmigiani, G. 2012. Estimation of sequencing error rates in short reads. *BMC Bioinformatics*, **13**(**1**):1–12. doi: 10.1186/1471-2105-13-185.

Wangensteen, O. S., 2020. Reference-databases Metabarpark. GitHub repository. http://github.com/metabarpark/Reference-databases.

Wangensteen, O. S. and Turon, X. 2017. Metabarcoding Techniques for Assessing Biodiversity of Marine Animal Forests. In *Marine Animal Forests*, pages 445–473. Springer International Publishing, Cham. doi: 10.1007/978-3-319-21012-4_53.

Wangensteen, O. S., Turon, X., García-Cisneros, A., Recasens, M., Romero, J., and Palacín, C. 2011. A wolf in sheep's clothing: carnivory in dominant sea urchins in the Mediterranean. *Marine Ecology Progress Series*, **441**:117–128. doi: 10.3354/meps09359.

Wangensteen, O. S., Cebrian, E., Palacín, C., and Turon, X. 2018a. Under the canopy: Community-wide effects of invasive algae in Marine Protected Areas revealed by metabarcoding. *Marine Pollution Bulletin*, **127**:54–66. doi: 10.1016/j.marpolbul.2017.11.033.

Wangensteen, O. S., Palacín, C., Guardiola, M., and Turon, X. 2018b. DNA metabarcoding of littoral hard-bottom communities: high

diversity and database gaps revealed by two molecular markers. *PeerJ*, **6**: e4705. doi: 10.7717/peerj.4705.

Wares, J. P. and Pappalardo, P. 2016. Can Theory Improve the Scope of Quantitative Metazoan Metabarcoding? *Diversity*, **8**(**1**):1. doi: 10.3390/d8010001.

Weigand, A., Bouchez, A., Boets, P., Bruce, K., Ciampor, F., Ekrem, T., Fontaneto, D., Franc, A., Hering, D., Kahlert, M., Keskin, E., Mergen, P., Pawlowski, J., Kueckmann, S., and Leese, F. 2019. Taming the Wild West of Molecular Tools Application in Aquatic Research and Biomonitoring. *Biodiversity Information Science and Standards*, **3**:e37215. doi: 10.3897/biss.3.37215.

West, K. M., Stat, M., Harvey, E. S., Skepper, C. L., DiBattista, J. D., Richards, Z. T., Travers, M. J., Newman, S. J., and Bunce, M. 2020. eDNA metabarcoding survey reveals fine-scale coral reef community variation across a remote, tropical island ecosystem. *Molecular Ecology*, **29**(**6**):1069–1086. doi: 10.1111/mec.15382.

Wheeler, Q. D., Raven, P. H., and Wilson, E. O. 2004. Taxonomy: Impediment or Expedient? *Science*, **303**(**5656**):285. doi: 10.1126/science.303.5656.285.

Winter, D., Green, P., Kamvar, Z., and Gosselin, T., 2017. Modern Measures of Population Differentiation. R package version 1.3.3. https://CRAN.R-project.org/package=mmod.

Zaiko, A., Martinez, J. L., Schmidt-Petersen, J., Ribicic, D., Samuiloviene, A., and Garcia-Vazquez, E. 2015. Metabarcoding approach for the ballast water surveillance – An advantageous solution or an awkward challenge? *Marine Pollution Bulletin*, **92**(**1-2**):25–34. doi: 10.1016/j.marpolbul.2015.01.008.

Zamora-Terol, S., Novotny, A., and Winder, M. 2020. Reconstructing marine plankton food web interactions using DNA metabarcoding. *Molecular Ecology*, **29**(**17**):3380–3395. doi: 10.1111/mec.15555.

Zink, R. M. 2002. Methods in Comparative Phylogeography, and Their Application to Studying Evolution in the North American Aridlands. *Integrative and Comparative Biology*, **42**(**5**):953–959. doi: 10.1093/icb/42.5.953.

Zizka, V. M. A., Weiss, M., and Leese, F. 2020. Can metabarcoding resolve intraspecific genetic diversity changes to environmental stressors? A test case using river macrozoobenthos. *Metabarcoding and Metagenomics*, **4**:23–34. doi: 10.3897/mbmg.4.51925.

Zomaya, A. Y., editor. 2005. *Parallel Computing for Bioinformatics and Computational Biology*. John Wiley & Sons, Inc., Hoboken, NJ, USA. ISBN 9780471756507. doi: 10.1002/0471756504.

Zou, K., Chen, J., Ruan, H., Li, Z., Guo, W., Li, M., and Liu, L. 2020. eDNA metabarcoding as a promising conservation tool for monitoring fish diversity in a coastal wetland of the Pearl River Estuary compared to bottom trawling. *Science of The Total Environment*, **702**: 134704. doi: 10.1016/j.scitotenv.2019.134704.

Appendix A

Chapter 4 Supporting Information

A.1 Supplementary Figures

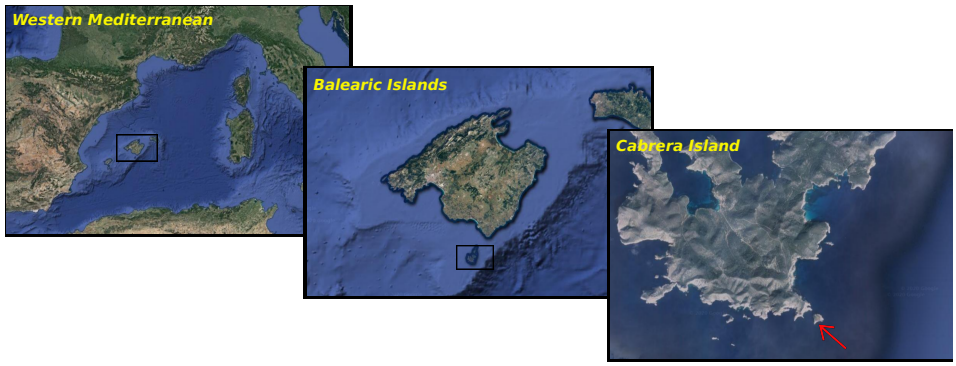


Figure A.1.1: Map of study area in the Western Mediterranean. The sampling zone is indicated with an asterisk. Maps from Google Earth public domain.

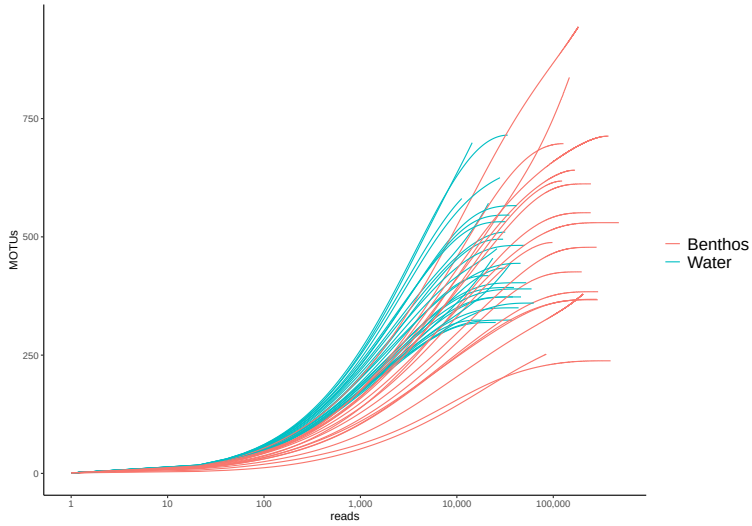


Figure A.1.2: Rarefaction curves of the number of MOTUs at increasing numbers of reads. Note logarithmic scale in the X axis.

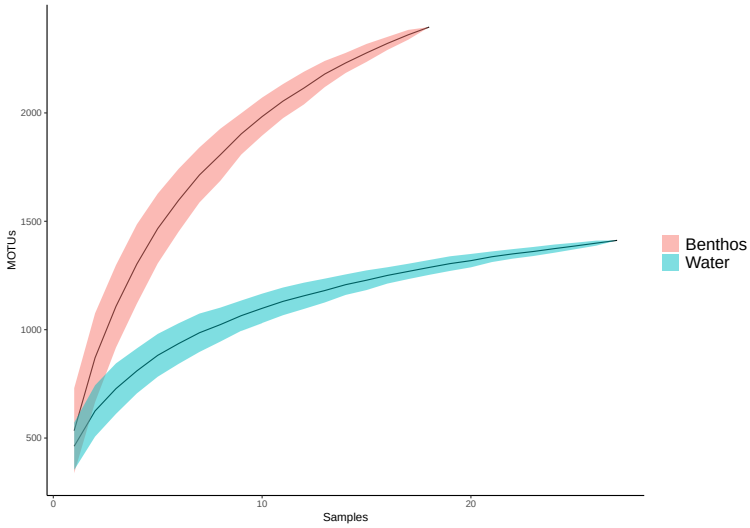


Figure A.1.3: MOTU accumulation curves of the number of reads detected as samples are pooled.

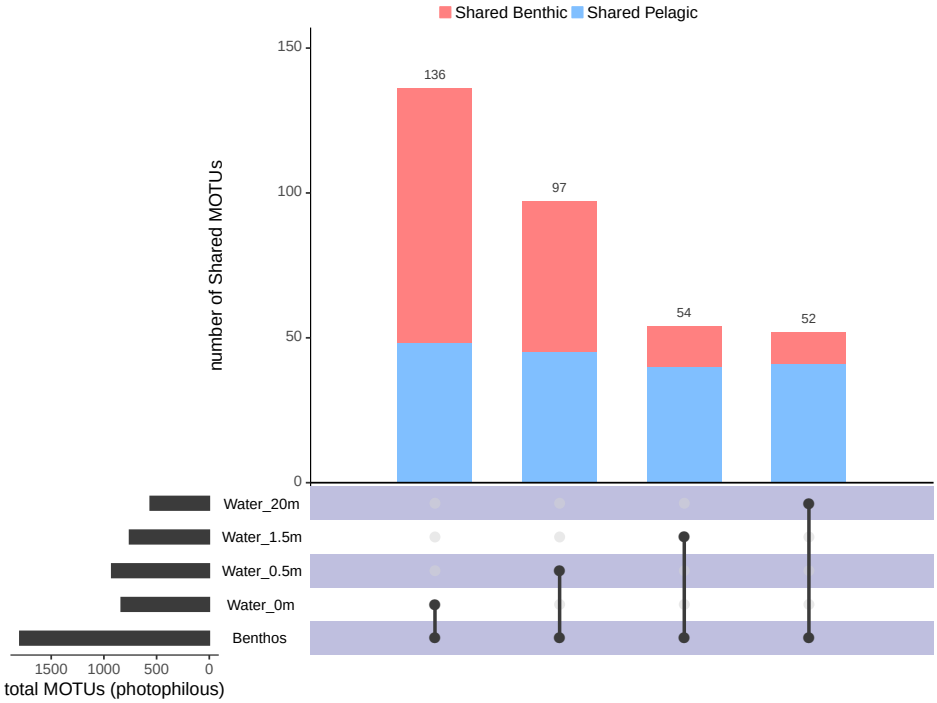


Figure A.1.4: Upset plot with the number of shared MOTUs between the benthos and the water samples and the total number of MOTUs detected in the Photophilous community (plus the pelagic samples). Shared benthic MOTUs (SBM) are represented in pink and shared pelagic MOTUs (SPM) in light blue.

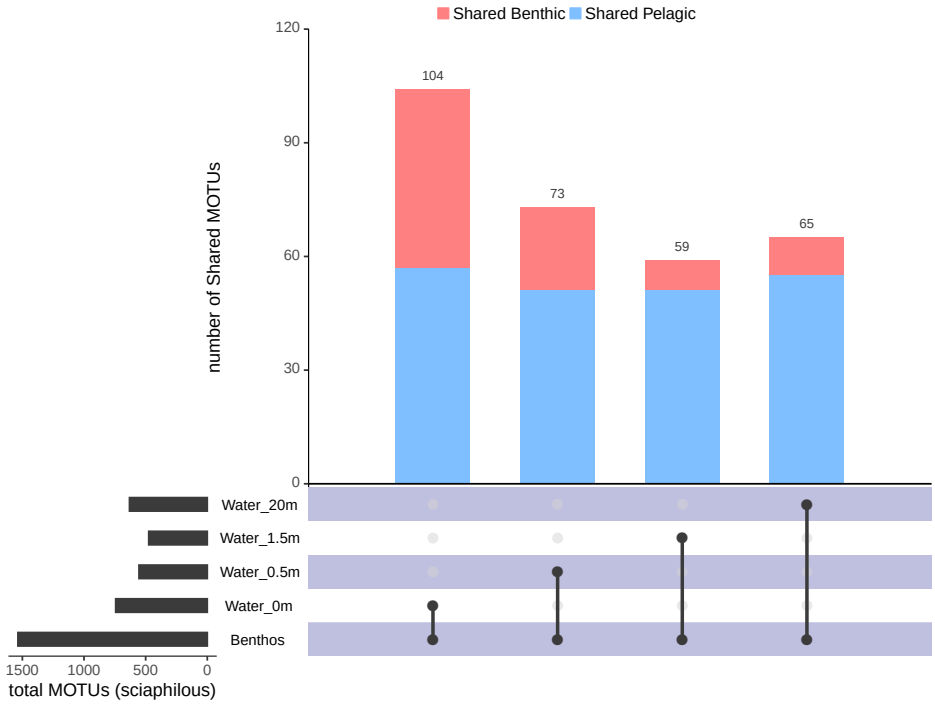


Figure A.1.5: Upset plot with the number of shared MOTUs between the benthos and the water samples and the total number of MOTUs detected in the Sciaphilous community (plus the pelagic samples). Shared benthic MOTUs (SBM) are represented in pink and shared pelagic MOTUs (SPM) in light blue.

A.2 Supplementary Data

QR A.2.1: Supplementary Data stored in the GitHub repository, see the following link https://github.com/adriantich/Thesis/tree/main/Chapter_2 to download supplementary data of chapter 4. Scan the QR code to download a zip file containing all chapter's supplementary data.

Supplementary Data A.2.1: List of the 3,543 MOTUs detected. Cell entries are number of reads in each sample. Download data using the QR A.2.1. Table S1 from file **SD2_1.xlsx**

Supplementary Data A.2.2: Number of MOTUs of the most important metazoan Phyla as per Class and Order in the different samples. Download data using the QR A.2.1. Table S2 from file **SD2_1.xlsx**

Supplementary Data A.2.3: List of the 180 shared benthic MOTUs (SBM). Cell entries are number of reads in each sample. Read abundance relative to the total of the samples of these MOTUs is also indicated. Download data using the QR A.2.1. Table S3 from file **SD2_1.xlsx**

Supplementary Data A.2.4: List of the 84 shared pelagic MOTUs (SPM). Cell entries are number of reads in each sample. Read abundance relative to the total of the samples of these MOTUs is also indicated. Download data using the QR A.2.1. Table S4 from file **SD2_1.xlsx**

Appendix B

Chapter 5 Supporting Information

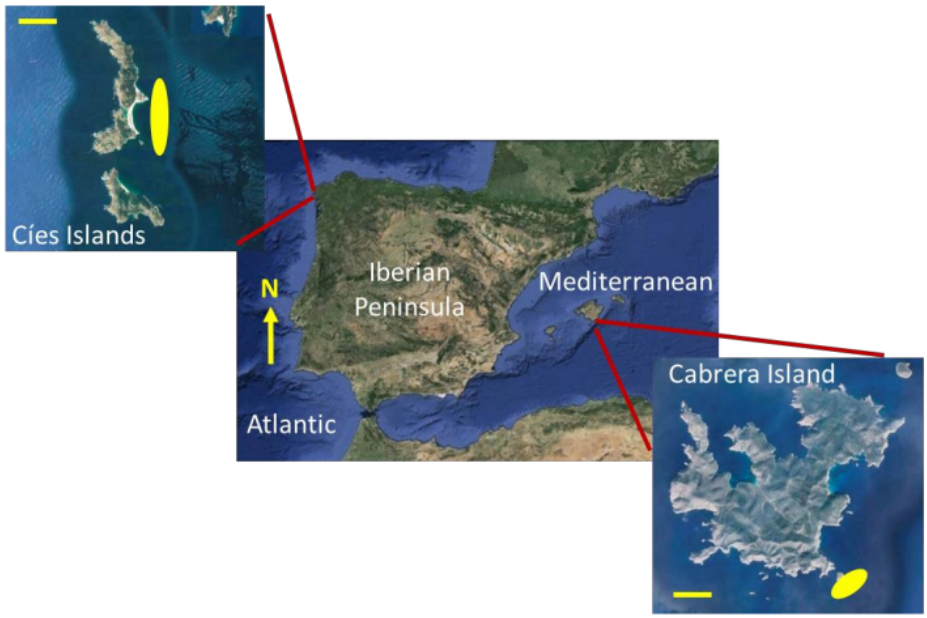


Figure B.0.1: Map of the Iberian Peninsula showing the two sampling sites in the Cabrera Archipelago (Mediterranean) and Cies Islands (Atlantic). The sampling zone in each site is indicated in yellow. Scale bars measure 1 Km.

B.1 Supplementary Data

QR B.1.1: Supplementary Data stored in the GitHub repository, see the following link https://github.com/adriantich/Thesis/tree/main/Chapter_3 to download supplementary data of chapter 5. Scan the QR code to download a zip file containing all chapter's supplementary data.

Supplementary Data B.1.1: Table of the initial 722 MOTUs used in the study. Download data using the QR B.1.1. Files **SD3_1.csv** (table) and **MetadataSD3_1.pdf** (metadata).

Supplementary Data B.1.2: Table of the final 563 MOTUs used in the study after the denoising and filtering steps. Download data using the QR B.1.1. Files **SD3_2.csv** (table) and **MetadataSD3_2.pdf** (metadata).

Supplementary Data B.1.3: Network analyses of the MOTUs retained after denoising and filtering. The size of the pies is proportional to the semiquantitative rank abundances used. Blue color represent abundance in Mediterranean samples, red color in Atlantic samples. The code of the MOTU and the main group where they belong are indicated. For details on each MOTU, refer to Data B.1.2. Only graphs for MOTUs with >2 and <230 sequences are represented. Download data using the QR B.1.1. File **SD3_3.pdf**.

Appendix C

Chapter 6 Supporting Information

C.1 The dataset

We used as a case study an unpublished dataset of COI sequences obtained from benthic communities in 12 locations of the Iberian Mediterranean. These locations are shown in Figure C.1.1. The seaweed-dominated shallow community inhabiting vertical rocky surfaces between -4 and -8 m was sampled by completely scraping off with hammer and chisel standardized surfaces of 25 × 25 cm. Three replicate samples were taken per location, and all samplings were performed in autumn of 2017. Sample processing was based on (Wangensteen et al., 2018b) and included a size fractionation step. Extraction and amplification were also performed as in that work using a modified version of the Leray et al. (2013) primer set (called Leray-XT in Wangensteen et al., 2018b), adding also unique 8-bp sample tags at both ends. HTS library preparation was performed using the NextFlex PCR-free DNA-Seq kit (Perkin-Elmer), based on ligation of the Illumina adapters at both ends of the amplicons. See the ms for the implications of PCR-free library construction methods in the application of one of the denoising algorithms (DADA2). We used a full run of a V3 Illumina MiSeq kit with 2*250 bp paired-end sequencing.

C.2 Comparison of DADA2 on unpaired and paired reads

For testing the performance of DADA2 on unpaired and paired reads on a coherent dataset, we selected the reads that were in the forward direction, that is, the forward primer was in the forward read (R1). To do this, we selected the forward-oriented paired reads before de-replicating (as indicated

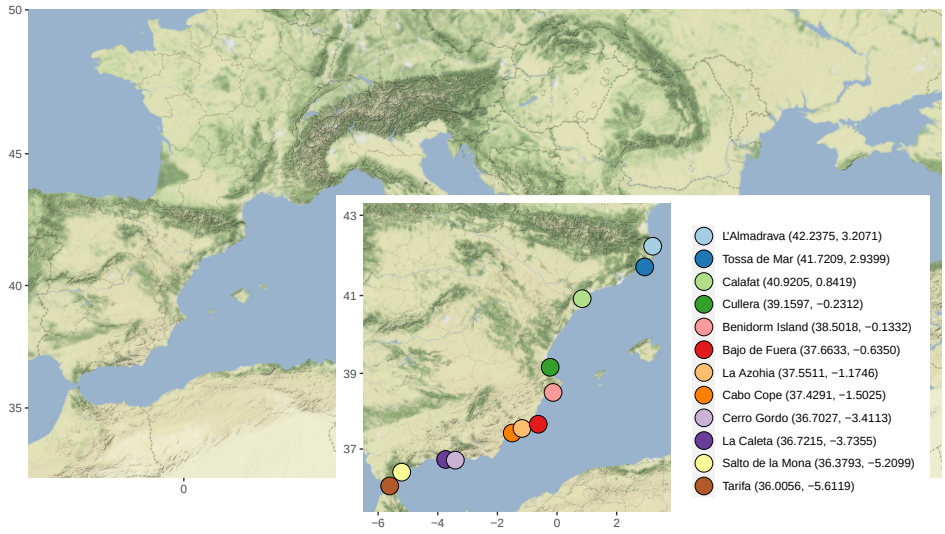


Figure C.1.1: Map of the sampling localities in the Iberian Peninsula, with indication of their coordinates.

by the tag "direction=forward" added by the merging procedure), but kept only those corresponding to sequences that will pass all the filters and will therefore make part of the final 9,718,827 reads.

As expected, the forward directed reads comprised ca. half of the total (4,892,084). This is due to the ligation-based library preparation protocol. We retrospectively picked the corresponding R1 and R2 reads from the sequencer output before pairing and eliminated the tags and primers. The last 20 bases from each read were trimmed. Thus, we had exactly the same 4,892,084 reads, paired and unpaired, for testing DADA2. This dataset, and the resulting ESV tables, are available from Mendeley Data (https://data.mendeley.com/datasets/84zypvmn2b/).

We first applied DADA2 to the unpaired R1 and R2 reads using the R package dada2 v. 1.14, with error rates estimated from the data with learnErrors. The dada command was applied to R1 and R2 reads with the default value for omega_A (10^{-40}) and setting omega_C to 0 (so all sequences with errors were corrected) and DETECT_SINGLETONS to True (to use all reads). The resulting reads were merged with mergePairs.

As a final output, we obtained 20,322 ESVs including the 4,892,084 reads.

The same procedure was repeated with the 4,892,084 paired reads. We input the sequences as if they were forward reads, no reverse reads were input and no merging step was performed. The quality profiles showed the expected jump towards higher quality in the overlapped fragment (ca. 106 bp). The mean quality score of all positions is 51.29, of the non-overlapping positions is 39.66, and of the overlapping bases is 73.87. The error rates were computed from the data and dada was applied as before. We obtained 24,573 ESVs, also totalling 4,892,084 reads. Therefore, using paired reads we obtained a number of ESVs 21% higher than with the unpaired reads. When comparing the outputs, we noted that 18,194 ESVs were identical (Fig 1). The match index of the ESVs was 0.818. In addition, the shared ESVs comprised most of the reads of the two datasets (98.81% of the reads of the unpaired dataset and 98.65% of the reads from the paired dataset). The match index of the reads was 0.987. The ESVs in the paired output not shared with the unpaired dataset had a low number of reads in general (average 10.39 reads).

We also noted that the estimated error rates for each substitution type (12 types) and quality score were highly correlated between the R1 and R2 reads ($r = 0.870$, $p < 0.0001$). In addition, the error rates as a function of quality score were also highly correlated between the 12 substitution types in each dataset. The lowest Pearson correlation coefficient for the estimated error rates of the R1 reads was 0.653 (between G to C and G to T changes), for R2 reads it was 0.741 (between G to T and C to G), and for thepaired sequences it was 0.894 (between G to A and C to G). All correlations proved highly significant after a False Discovery Rate correction (Benjamini and Hochberg, 1995).

Thus, our results using merged reads instead of using the forward and reverse sequences separately resulted in most reads being placed in the same ESVs, but more (21%) ESVs were kept when using merged reads. This result stems from the fact that a higher confidence in the bases of the long (ca. one third) overlapped region in turn results in accepting as correct

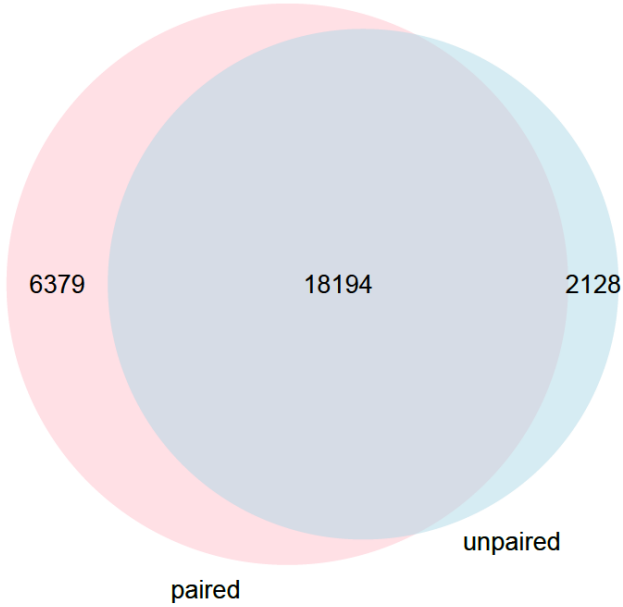


Figure C.2.2: Venn Diagram of the number of ESVs after applying DADA2 before (unpaired) or after (paired) merging the two reads for each sequence in the dataset analysed.

sequences that would otherwise be labelled as erroneous. We could therefore retain low abundance ESVs that would have been merged in the unpaired dataset analysis. Indeed, the ability to tell low-abundance, but legitimate, sequences from errors is the goal of all denoising procedures. Using paired reads also improves the applicability of the DADA2 algorithm at any step in the bioinformatic processing (not at the very beginning), thus making it a more flexible tool. As this is a requirement to perform our comparative analyses, we will use DADA2 on merged sequences, while keeping in mind that we lose stringency (retain more ESVs) by doing so.

C.3 Taxonomic benchmarking

We combined all unique ESVs retained after the denoising algorithms (those retained by the different versions of DnoisE and those retained by DADA2, for a total of 116,218 ESVs) and assigned them taxonomically with ecotag. We found that 25,197 sequences had a species-level assignment, comprising 690 species, of which 187 were represented by a single sequence. We further refined this dataset by accepting only sequences whose best hit in the reference database was $\geq$0.97, which is in accordance with the mean intra-MOTU distance we found in our dataset with SWARM. This pruned dataset (henceforth species-level dataset, available as Additional file 4), consisted of 14,487 assigned sequences belonging to 422 species, with 130 having only one sequence. Without the inclusion of the entropy-corrected ESV dataset (which kept more ESVs than the other methods) we had 5,147 sequences assigned at species level, belonging to 417 species. Thus, the inclusion of this dataset almost tripled the number of sequences with species-level hit but these represented only five extra species with respect to the other datasets, indicating that the gain in ESVs in the entropy-corrected procedure mainly increases within-MOTU variability.

We checked how many of the sequences in the species-level dataset were recovered with the different denoising and clustering methods. We also assessed whether these sequences were grouped in closed MOTUs (meaning all sequences in the MOTU belonged to the same species and no other sequences of this species were found in other MOTUs), open MOTUs (i.e., all sequences belonged to the same species, but not all sequences assigned to the species were included) and hybrid MOTUs. The latter included MOTUs with sequences assigned to more than one species, or MOTUs with a combination of sequences assigned to one species and sequences not in the species-list dataset (i.e., they don't have species-level assignment, or they do with less than 97% similarity). Closed MOTUs were further subdivided among those with only one sequence (closed singleton) and those with several sequences (closed group).

The proportion of the 422 species that were recovered by the different

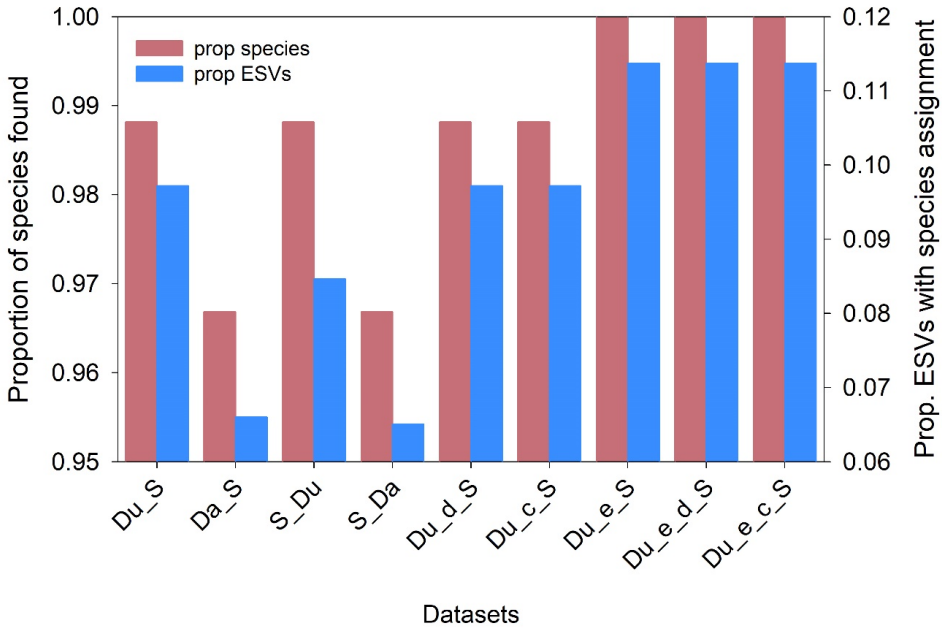


Figure C.3.3: Proportion of the species in the species-level dataset recovered and proportion of ESVs with species-level assignment found in the different datasets.

methods was in all cases high (above 96%), but the datasets denoised with DADA2 featured the lowest proportions (Figure 1). The entropy-corrected datasets, on the other hand, recovered all species. Differences were also apparent in the proportion of ESVs with species-level assignment that were found in the different datasets. In general, DADA2- based datasets had a lower proportion of ESVs with species assignment indicating that more sequences assignable to species have been merged during denoising. Clustering first reduced appreciably this proportion in the UNOISE3-based datasets (Du_S vs. S_Du, ca. 13% reduction), while it didn't vary in the comparison Da_S vs S_Da. Entropy- corrected datasets had not only a higher number of ESVs, but a higher proportion of them (>11%) with species-level assignment (Figure C.3.3).

As shown in Figure C.3.4, when we checked the different datasets generated, irrespective of the method the majority (62-75%) of MOTUs that had sequences assigned to the species rank were closed, 7-10% were open,

and 18-28% were hybrid MOTUs. This indicates that, in all cases, the denoising plus clustering methods performed reasonably well in recovering species that were identified as such in taxonomic assignment of the ESVs. The UNOISE3 algorithm, however, recovered ca. 60% more closed group MOTUs that DADA2, and the opposite occurred for closed singleton MOTUs, for a similar total. This is the result of the higher number of sequences retained by UNOISE3, that translated into a higher ability to recover MOTUs with internal diversity. The proportion of hybrid MOTUs was lower in the DADA2 denoised datasets, which were the most stringent in terms of ESVs retained, while the datasets with entropy correction, the ones with higher number of ESVs, featured a slightly higher proportion of hybrid MOTUs than those not-corrected. We verified manually these hybrid MOTUs, and in most cases they were due to the inclusion of some sequences not in the species-level dataset (for instance, sequences with less than 97% similarity with their matches in the reference database), rather than to the lumping of sequences assigned to different species. To check this point, we repeated the analysis without enforcing the 97% similarity, and the proportion of hybrid MOTUs decreased by half. This indicates that in many cases the hybrid MOTUs found in the 97% restricted-similarity analysis comprised sequences assigned by ecotag to the same species, but some of them with similarity levels lower than 97%. Overall, then, the taxonomic benchmarking showed a good correlation between MOTUs and species assignments performed with ecotag.

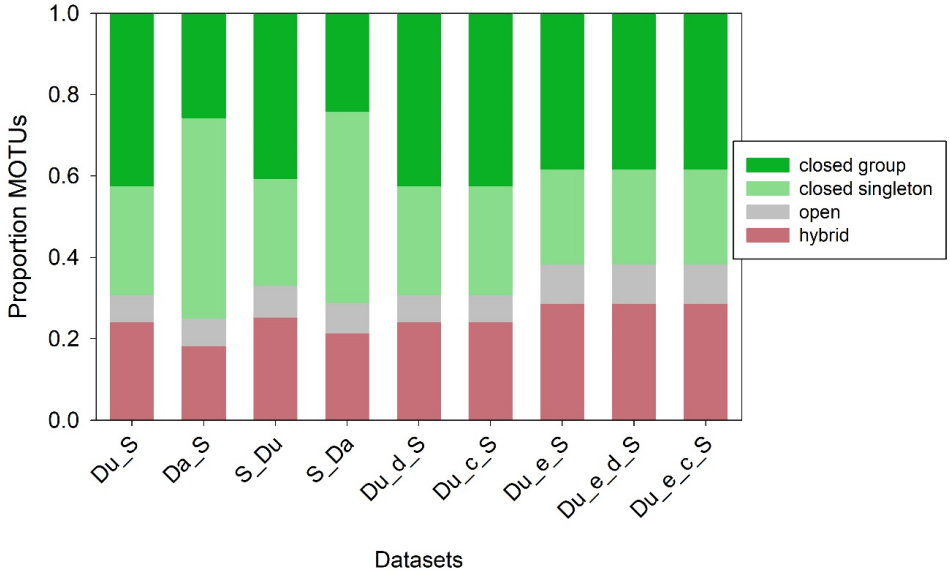


Figure C.3.4: Proportion of closed, open, and hybrid MOTUs found in the different datasets.

C.4 Supplementary Data

QR C.4.1: Supplementary Data stored in the GitHub repository, see the following link https://github.com/adriantich/Thesis/tree/main/Chapter_4 to download supplementary data of chapter 6. Scan the QR code to download a zip file containing all chapter's supplementary data.

Supplementary Data C.4.1: Table with the ESVs identified at the species level with >97% similarity. The taxonomy assigned is indicated, as well as the best-match in the reference database, the taxid, and the sequence. Download data using the QR C.4.1. File **SD4_1.xlsx**.

Appendix D

Chapter 8 Supporting Information

Table D.1: Names and codes of localities and fronts. Coordinates and regions of sampling sites and localities adjacent to the fronts are given.

Locality	Code	Latitude	Longitude	Region
ROSES	ROS	42.237	3.207	Northern
TOSSA	TOS	41.721	2.940	Northern
CALAFAT	CAL	40.920	0.842	Northern
CULLERA	CLL	39.160	-0.231	Northern
VILLAJOYOSA	JOY	38.502	-0.133	Central
PALOS	PAL	37.663	-0.631	Central
AZOHIA	AZO	37.551	-1.175	Central
CARBONERAS	CAR	37.429	-1.502	Central
COSTA DE GRANADA	GRA	36.703	-3.411	Southern
LA HERRADURA	LHE	36.722	-3.736	Southern
COSTA DEL SOL	SOL	36.379	-5.210	Southern
TARIFA	TAR	36.006	-5.612	Southern
Front	Code	Locality above	Locality below	
Ibiza Channel	ICF	CLL	JOY	
Almeria-Oran	AOF	CAR	GRA	

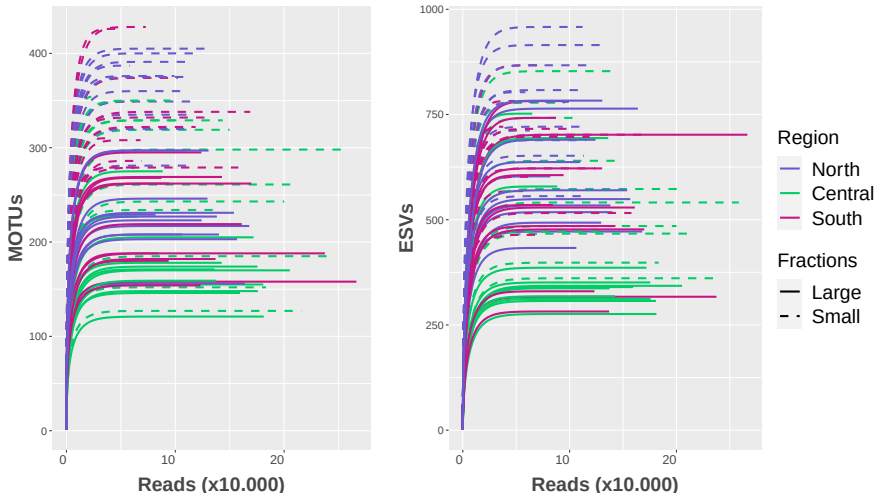


Figure D.0.1: Rarefaction curves for all samples for MOTU (left) and ESV (right) data. Colours represent the region and linetypes the fraction size.

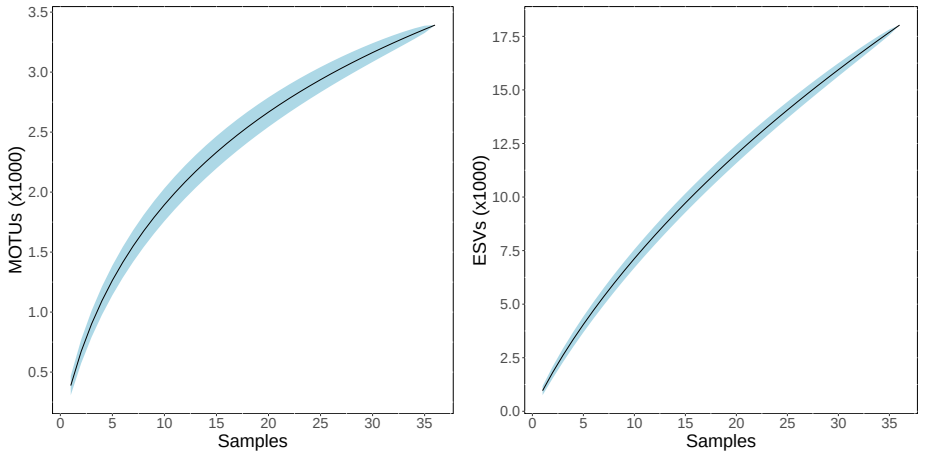


Figure D.0.2: MOTU (left) and ESV (right) accumulation curves of all sample sites.

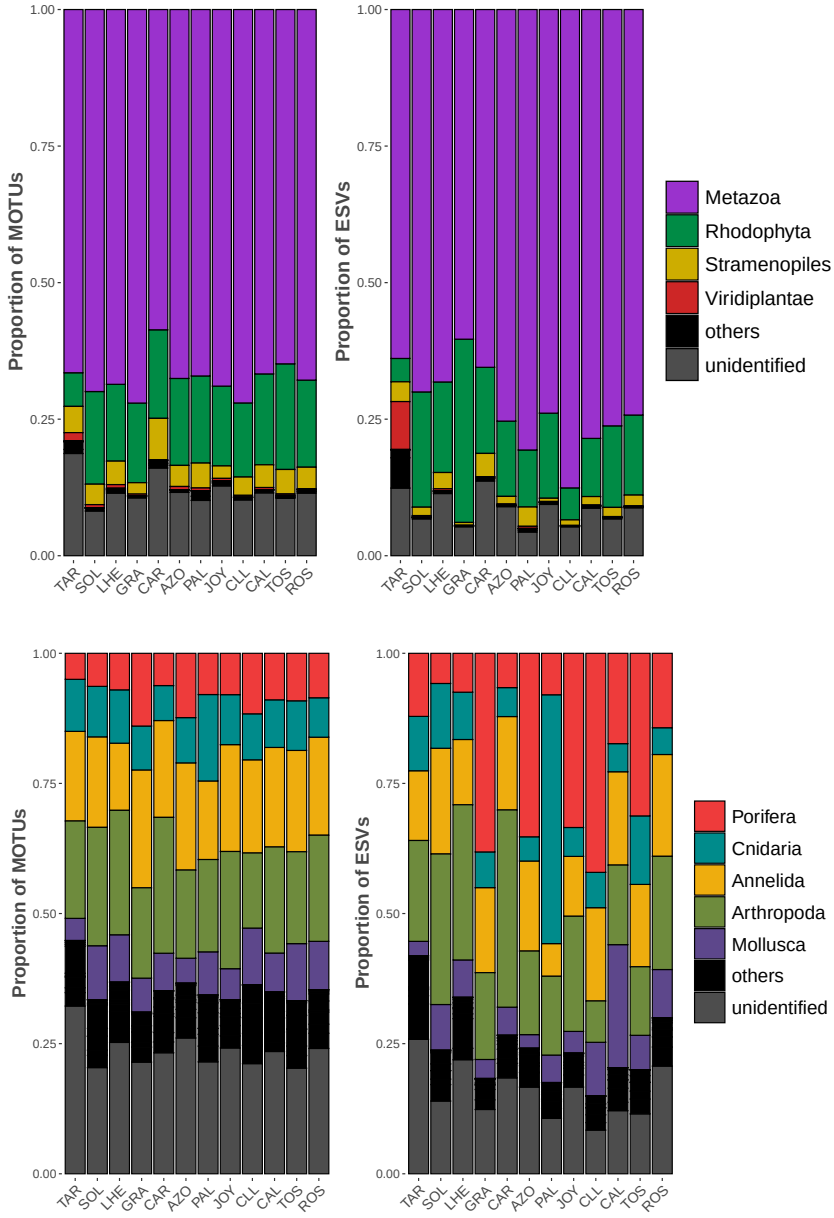


Figure D.0.3: Supergroup (top) and metazoan phyla (bottom) composition in relative MOTU number (left) and ESV number (right) for each locality.

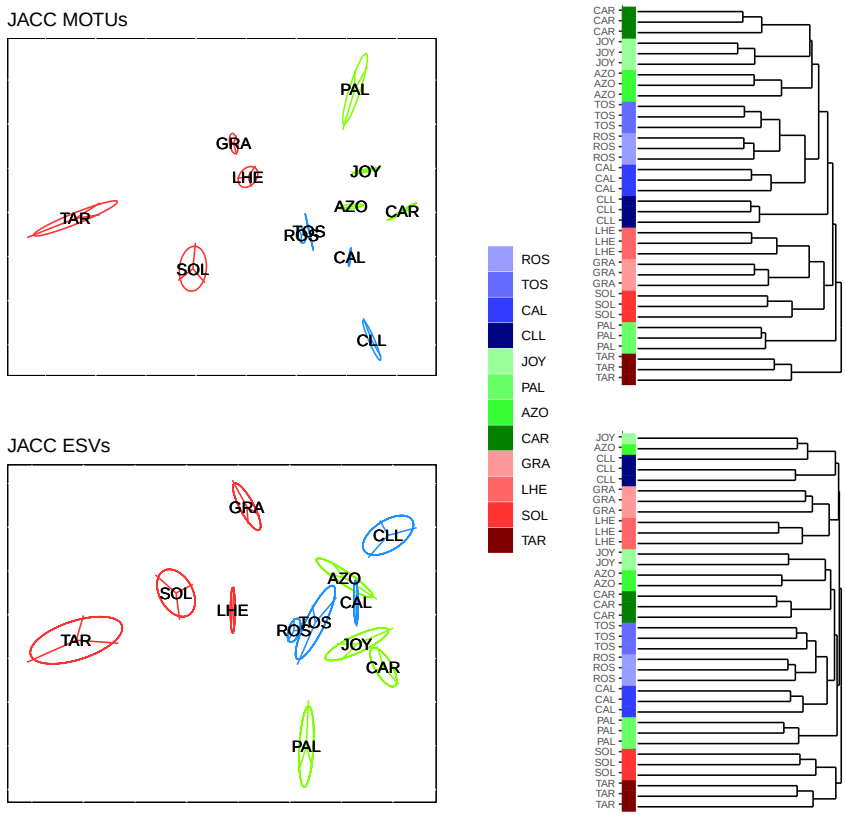


Figure D.0.4: Non-metric Multidimensional Scaling (left) and clusters (right) of samples using Jaccard dissimilarities for MOTUs and ESVs. Samples grouped by locality. Region factor is represented by colours (Northern, blues; Central, greens; Southern, reds).

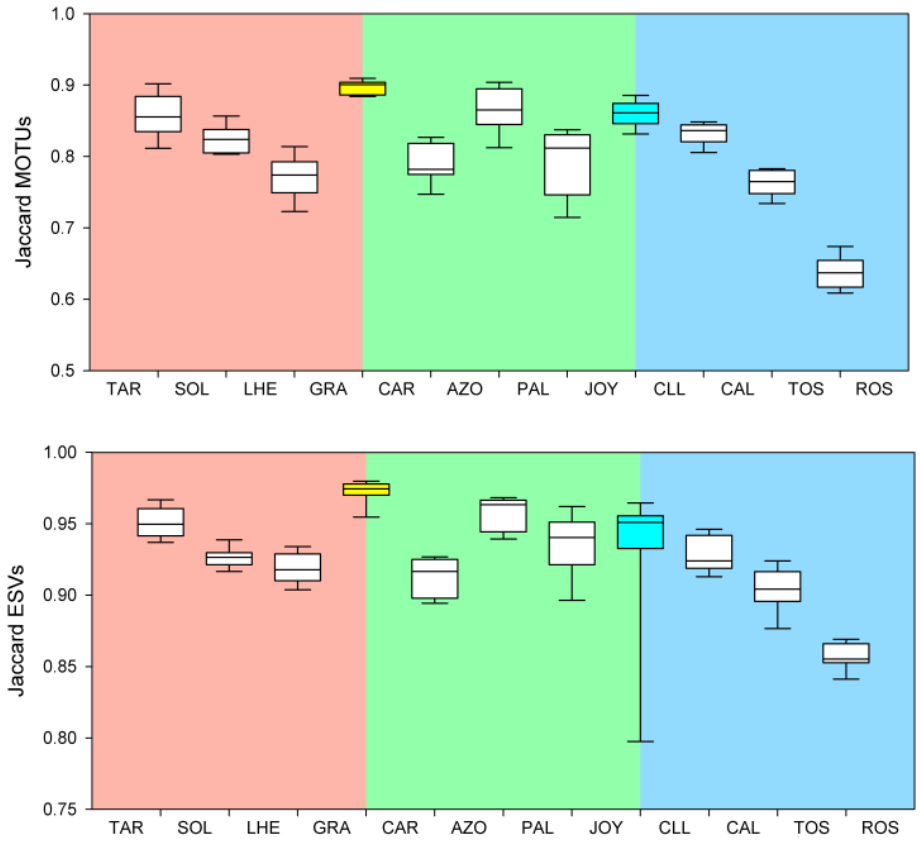


Figure D.0.5: Jaccard dissimilarities of MOTUs (top) and ESVs (bottom). Fronts are represented in yellow for the AOF and light blue for IC. Background colour corresponds to regions (red: Southern, green: Central, blue: Northern).

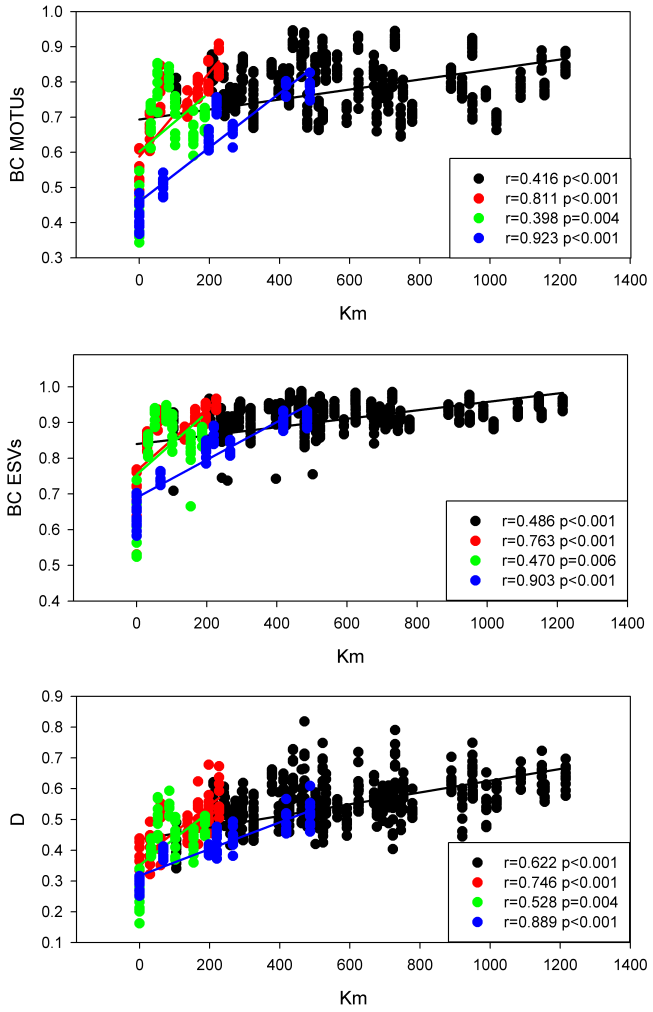


Figure D.0.6: Mantel test plotting the geographic distance against the BC dissimilarities of MOTUs and ESV and D dissimilarities of haplotypes. Each point represents a pairwise comparison between samples: red for comparisons within the southern region, green within the central region, and blue within the northern region. Black dots represent inter-region comparisons, but note that all points are used in the computation of the overall Mantel test (represented by the black symbol in legend). The Mantel r statistics and p-values are indicated. The regression lines plotted correspond to the complete set of comparisons (black line) and to each of the southern, central, and northern regions separately (red, green, and blue lines, respectively).

D.1 Supplementary Data

QR D.1.1: Supplementary Data stored in the GitHub repository, see the following link https://github.com/adriantich/Thesis/tree/main/Chapter_6 to download supplementary data of chapter 8. Scan the QR code to download a zip file containing all chapter's supplementary data.

Supplementary Data D.1.1: Table of the 18026 ESV from 3392 MOTUs with their read abundance in each sample used for the analyses. Download data using the QR D.1.1. File **SD6_1.csv**.

Supplementary Data D.1.2: Table of the 3392 MOTUs with their taxonomic information and representative sequence. Download data using the QR D.1.1. File **SD6_2.csv**.

Supplementary Data D.1.3: Network analyses of the MOTUs used for Metaphylogeographycal analysis. The size of the pies is proportional to the semiquantitative rank abundances used. Region factor is represented by colours (Northern, blues; Central, greens; Southern, reds). The code of the MOTU, the main group where they belong and best matched identity are indicated. Download data using the QR D.1.1. File **SD6_3.csv**.

Appendix E

Published Chapters

Received: 16 April 2020 | Revised: 6 August 2020 | Accepted: 7 September 2020

DOI: 10.1111/mec.15641

SPECIAL ISSUE

MOLECULAR ECOLOGY WILEY

Marine biomonitoring with eDNA: Can metabarcoding of water samples cut it as a tool for surveying benthic communities?

Adrià Antich[1] | Cruz Palacín[2] | Emma Cebrian[3] | Raül Golo[3] | Owen S. Wangensteen[4] | Xavier Turon[1]

[1]Department of Marine Ecology, Center for Advanced Studies of Blanes (CEAB-CSIC), Girona, Spain

[2]Department of Evolutionary Biology, Ecology and Environmental Sciences, University of Barcelona, and Research Institute of Biodiversity (IRBIO), Barcelona, Spain

[3]Institute of Aquatic Ecology, University of Girona, Girona, Spain

[4]Norwegian College of Fishery Science, UiT The Arctic University of Norway, Tromsø, Norway

Correspondence
Xavier Turon, Department of Marine Ecology, Center for Advanced Studies of Blanes (CEAB-CSIC), Blanes, Girona, Catalonia, Spain.
Email: xturon@ceab.csic.es

Owen S. Wangensteen, Norwegian College of Fishery Science, UiT The Arctic University of Norway, Tromsø, Norway.
Email: owen.wangensteen@uit.no

Funding information
Autonomous Organism of National Parks, Grant/Award Number: 2017-2462; Ministerio de Economía y Competitividad, Grant/Award Number: CGL2016-76341-R and CTM2017-88080

Abstract

In the marine realm, biomonitoring using environmental DNA (eDNA) of benthic communities requires destructive direct sampling or the setting-up of settlement structures. Comparatively much less effort is required to sample the water column, which can be accessed remotely. In this study we assess the feasibility of obtaining information from the eukaryotic benthic communities by sampling the adjacent water layer. We studied two different rocky-substrate benthic communities with a technique based on quadrat sampling. We also took replicate water samples at four distances (0, 0.5, 1.5, and 20 m) from the benthic habitat. Using broad range primers to amplify a ca. 313 bp fragment of the cytochrome oxidase subunit I gene, we obtained a total of 3,543 molecular operational taxonomic units (MOTUs). The structure obtained in the two environments was markedly different, with Metazoa, Archaeplastida and Stramenopiles being the most diverse groups in benthic samples, and Hacrobia, Metazoa and Alveolata in the water. Only 265 MOTUs (7.5%) were shared between benthos and water samples and, of these, 180 (5.1%) were identified as benthic taxa that left their DNA in the water. Most of them were found immediately adjacent to the benthos, and their number decreased as we moved apart from the benthic habitat. It was concluded that water eDNA, even in the close vicinity of the benthos, was a poor proxy for the analysis of benthic structure, and that direct sampling methods are required for monitoring these complex communities via metabarcoding.

KEYWORDS
benthos, biomonitoring, eDNA, marine, metabarcoding, water

1 | INTRODUCTION

Metabarcoding is by now a well-established technique for assessing biodiversity in a variety of terrestrial, freshwater, and marine environments (reviewed in Bohmann et al., 2014; Creer et al., 2016; Cristescu, 2014; Deiner et al., 2017; Taberlet, Coissac, Pompanon, et al., 2012). The wealth of published papers dealing with technical issues and generating new data with this method testifies to the widening scope of applications of metabarcoding. One such application, where metabarcoding is becoming a game-changer, is in the field of biomonitoring (Aylagas et al., 2018; Hajibabaei et al., 2016; Kelly et al., 2014; Porter & Hajibabaei, 2018). Not in vain the use of DNA-based approaches for monitoring applications has been christened Biomonitoring 2.0 (Baird & Hajibabaei, 2012; Leese et al., 2018).

In the marine realm, all current policies, such as the European Union Marine Strategy Framework Directive, mandate comprehensive, community-wide approaches to monitoring (Danovaro et al., 2016; Goodwin et al., 2017; Hering et al., 2018; Leese et al., 2018). Metabarcoding provides a cost-effective, ecosystem-wide method for the assessment of biodiversity, which lies at the basis of all monitoring efforts (Aylagas et al., 2018; Krehenwinkel

et al., 2019; Leray & Knowlton, 2016; Shaw et al., 2017). An ever widening range of ecological and socioeconomic issues, such as invasive species management (Darling et al., 2017; Holman et al., 2019), marine protected areas design (Bani et al., 2020), pathogen monitoring (Peters et al., 2018), fisheries management (Zou et al., 2020), or deep-sea mining (Cowart et al., 2020), among others, require powerful and fast biomonitoring tools. Metabarcoding provides these tools at a pace, cost, and depth that are not achievable using conventional, morphology-based surveys (Porter & Hajibabaei, 2018). Alpha- and beta-diversity estimates, as well as biotic indices, can be reliably obtained using metabarcoding (Aylagas et al., 2018; Bani et al., 2020; Hering et al., 2018; Pawlowski et al., 2018). The amount of data typically generated in metabarcoding data sets also allows bioassessments based on taxonomy-free and machine learning techniques (Cordier & Pawlowski, 2018; Gerhard & Gunsch, 2019), or the analysis of diversity at the within-species level (Turon et al., 2020).

Of course, gaps and problems are also recognized in this burgeoning field (e.g., Alberdi et al., 2018; Kelly et al., 2019; McGee et al., 2019), among which the need to obtain better reference databases (Sinniger et al., 2016; Wangensteen, Palacín, et al., 2018; Weigand et al., 2019) and the need to standardize field and laboratory procedures (McGee et al., 2019; Weigand et al., 2019). Among the latter, the type of substrate sampled is of paramount importance (Koziol et al., 2019). In the sea, most studies to date have sampled either the sediment (e.g., Atienza et al., 2020; Brannock et al., 2016; Fonseca et al., 2014; Guardiola et al., 2016), or the water column (e.g., Brannock et al., 2018; Fraija-Fernández et al., 2019; Sigsgaard et al., 2019; Stefanni et al., 2018). Less effort has been devoted to the study of hard-substrate natural benthic communities. These have been analysed either using indirect methods based on deploying artificial substrates (Cahill et al., 2018; Leray & Knowlton, 2015; Pearman et al., 2019; Ransome et al., 2017), or by directly taking samples by scraping off standardized surfaces (Shum et al., 2019; Wangensteen, Cebrian, et al., 2018; Wangensteen, Palacín, et al., 2018), or using suction devices (Cowart et al., 2020; De Jode et al., 2019).

Either deploying settlement surfaces (that need to be recovered) or using direct collection methods, the sampling of benthic hard-bottom habitats requires direct access to the environment and involves more effort than sampling substrates such as water or sediment, which can be accessed remotely. In addition, direct methods are destructive, which is an inconvenience for the sustained sampling necessary for biomonitoring. It is, therefore, highly convenient to develop alternative methods for assessing benthic biodiversity, and an obvious choice would be to sample the water in the vicinity of the benthos to recover benthic DNA for metabarcoding applications. While water environmental DNA (eDNA) has been used for the study of protists, fito- and zooplankton or fish assemblages (e.g., Djurhuus et al., 2018; Massana et al., 2015; Shu et al., 2020), its potential utility to analyse benthic communities is much less understood. Some authors (Koziol et al., 2019; Rey et al., 2020) compared eDNA from water, sediment and settlement plates in port environments, finding clearly distinct community profiles. Leduc et al. (2019) similarly found significant differences in community composition between eDNA from water samples and standard invertebrate collection methods in Arctic harbours. West et al. (2020) used surface water samples to assess coral reef community variation, but did not perform a comparison with the actual benthic communities. Alexander et al. (2020) used eDNA from surface waters to target scleractinian diversity, and found the method promising, albeit with notable differences with results from visual censuses. Stat et al. (2017) compared two different methods to study the eDNA from tropical marine reefs using shallow water and found eDNA metabarcoding more promising than the shotgun approach for assessing eukaryotic diversity.

The usefulness of DNA obtained from water samples as a proxy for benthic communities will depend on the many factors that affect DNA release, transport, and degradation (Barnes & Turner, 2016; Collins et al., 2018; Salter, 2018; Stewart, 2019). While some studies have assessed the spatial distribution of eDNA in coastal habitats, they have been done at scales too large to link water samples with particular benthic habitats. Bakker et al. (2019) analysed water eDNA from coastal shelf habitats spanning the Caribbean Sea. O'Donnell et al. (2017) found fine scale patterns in the distribution of water eDNA, but they used transects perpendicular to the shore spanning a few kilometres. Jeunen et al. (2019) analysed the vertical stratification of eDNA at the scale of metres, but did not focus on any relationship with benthic communities. Jacobs-Palmer et al. (2020) analysed eDNA from water taken in the vicinity (from 1 to 15 m) of the edges of *Zostera marina* patches, and could detect an inhibitory effect of the seagrass community on the dinoflagellate abundances in the plankton. To our knowledge, however, no study has assessed marine eDNA dynamics at the benthic boundary layer, which is the water immediately adjacent (from centimetres to metres) to the benthos, where steep gradients in abiotic and biotic parameters occur (Boudreau & Jorgensen, 2001). Only Hajibabaei et al. (2019) and Gleason et al. (2020) have compared, in freshwater environments, the results from DNA obtained from matched water and benthic samples, and found water eDNA to be a poor surrogate for benthic community composition.

In this work, and using two hard-bottom communities on vertical walls in the NW Mediterranean, we compared the information obtained from analysing the DNA obtained from benthic (using direct methods as in Wangensteen, Palacín, et al., 2018) and water samples collected at increasing distances (from centimetres to metres) from these communities. We used metabarcoding of the COI gene with broad range primers as our focus was on recovering the taxonomically diverse eukaryotic communities present. Our goals were to assess the eDNA dynamics in the boundary layer of the benthos and to determine the feasibility of analysing benthic diversity by collecting water samples.

2 | MATERIALS AND METHODS

2.1 | Sample collection

In the present study samples were taken from two different hard-bottom communities, a shallower (photophilous) and a deeper

(sciaphilous) communities found in the same vertical wall facing SSE, in the National Park of Cabrera Archipelago in the Balearic Islands (Western Mediterranean, 39°07′30.32″N, 2°57′37.14″E, Figure S1). The photophilous community at 10 m depth was dominated by the seaweeds *Padina pavonica* and *Dictyopteris membranacea*. In the sciaphilous community at 30 m depth, the seaweed *Halimeda tuna*, sponges and other invertebrates were the dominant biota. For more detailed information of these communities see Wangensteen, Palacín, et al. (2018).

Two different sampling methods were used in the present study. Benthic samples (three replicates per community) were obtained by scraping to bare rock quadrats of 25 × 25 cm with hammer and chisel. All the material was collected underwater in plastic bags. Two divers performed the sampling, with one keeping the sample bag open just over the zone being scraped to avoid escape of small motile fauna. Water samples (four replicates at each point) were obtained with 1.5 L bottles at different distances from the benthos (0, 0.5 and 1.5 m) for each community. The sample labelled 0 m was obtained in the water layer just adjacent (ca. 5 cm) to the benthos. As an external pelagic control, water samples (three replicates) of 1.5 L were obtained at 20 m from the benthos and at an intermediate depth (−20 m). The sampling design is sketched in Figure 1. Hereafter we will use the names photophilous and sciaphilous samples to designate both the benthic and the water samples ≤1.5 m from the wall at each of the two depth levels sampled, and the name pelagic samples to designate the water samples collected 20 m apart from the rocky wall at −20 m. New, unopened mineral water plastic bottles were used for water collection, one per sample. They were first filled with sterilized water and, once in the collection point, they were held upside-down and water was displaced using air bubbled from a spare SCUBA regulator. The bottles were then righted and water from the exact point of collection was allowed to fill them.

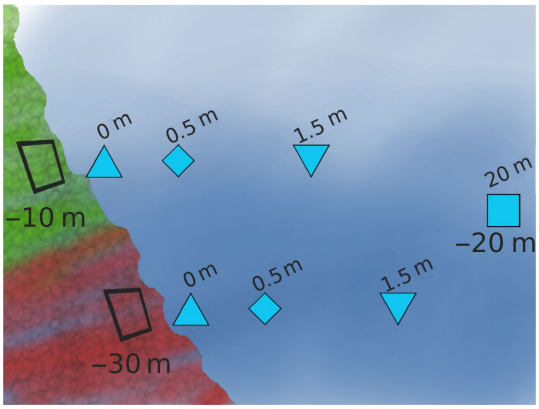


FIGURE 1 Schema of the sampling design. We sampled two hard bottom communities (green: photophilous; red: sciaphilous) at −10 and −30 m of depth, respectively, by sampling quadrats of 25 × 25 cm (three replicates each). Water samples (1.5 L) were collected at different distances from each community (0, 0.5 and 1.5 m, four replicates each). Pelagic samples were taken at intermediate (−20 m) depth and at 20 m from the wall (three replicates)

2.2 | Sample processing

Water samples were processed on site immediately after collection. The whole collected volume (1.5 L, comparable to other studies, e.g., Collins et al., 2019; Sales et al., 2019) was prefiltered with a 200 μm mesh to eliminate coarse particles and then filtered through 0.22 μm Sterivex millipore filters (Merck) using sterile, disposable syringes (a new syringe per sample). The filter cartridges were then stored at −20°C in sterile plastic bags. Benthic samples were fixed with ethanol immediately after collection and kept at −20°C until processed in the laboratory. Following Wangensteen and Turon (2017), Wangensteen, Palacín, et al. (2018) and Wangensteen, Cebrian, et al. (2018), benthic samples were separated in the laboratory in three different size fractions (A: >10 mm; B: 1–0 mm; C: 63 μm–1 mm) using a stainless steel mesh sieve column (Cisa S.L., www.cisa.net). Each fraction was homogenized with a blender and stored in ethanol at −20°C until DNA extraction. All equipment was carefully bleached between samples.

Our sample data set thus consisted of 18 benthic samples (two communities × three replicates × three fractions) and 27 water samples (two communities × three distances × four replicates + three pelagic samples).

2.3 | DNA extraction

All procedures were made in a laminar flow cabinet sterilised with UV light between samples. DNA from benthic samples was extracted using 10 g of homogenized material and the DNeasy PowerMax Soil Kit (Qiagen). The Sterivex filter cartridges were opened with sterile pincers in the cabinet and DNA from the filters was then extracted using the DNeasy PowerWater kit (Qiagen). A Qubit fluorometer (ThermoFisher) was used to check the concentration of DNA (higher than 5 ng/μl in all cases).

2.4 | PCR amplification and library preparation

A fragment of ca. 313 bp of the Cytochrome Oxidase 1 (COI) gene was amplified with a set of universal primers targeting eukaryotes. We used the Leray-XT primer set (Wangensteen, Cebrian, et al., 2018; Wangensteen, Palacín, et al., 2018): forward jgHCO2198 (Geller et al., 2013): 5′-TAIACYTCIGGRTGICCRAARAAYCA-3′, reverse mlCOIintF-XT (Wangensteen, Palacín, et al., 2018): 5′-GGWACWRGWTGRACWITITAYCCYCC-3′. All primers had an 8-base specific tag attached. The tags had a minimum difference of 3 bases from each other, and were designed with the program Oligotaq (Boyer et al., 2016). Forward and reverse primers used for amplification of each sample had the same tag. A variable number

of degenerate (N) bases (from two to four) were also attached to the forward and reverse primers to improve sequence diversity for illumina processing.

Three PCR replicates were performed for each DNA extraction. PCR conditions for COI amplification followed (Wangensteen, Palacín, et al., 2018). DNA was then purified and concentrated using MinElute PCR Purification Kit (Qiagen) and an electrophoresis gel was performed to check amplification success.

Amplification controls were added as follows: two PCR blanks were run by amplifying the PCR mixture without any DNA template. Negative controls were made for the benthic samples by processing triplicate sand samples that were charred in a furnace (400°C for 24 hr) and then sieved and processed as above. For the water samples we filtered in situ sterilized ultrapure water with three Sterivex filters that were then treated in the same manner as the seawater filters. Amplification products were pooled to build two Illumina libraries using Nextflex PCR-free library preparation kit (Perkin-Elmer). Both libraries were sequenced together in an Illumina MiSeq V3 run using 2 × 250 bp paired-end sequencing.

2.5 | Bioinformatic analyses

The bioinformatic analyses followed the same pipeline of Atienza et al. (2020) with slight modifications. Most steps used the OBITools package (Boyer et al., 2016). Illuminapairedend was used to align paired-end reads and keep only those with >40 alignment quality score. Reads were demultiplexed using ngsfilter. Those with mismatched primer tags at any end were discarded. Obigrep and obiuniq were used to perform a length filter (retaining only those between 310–317 bp) and dereplicate sequences. Uchime-denovo algorithm from VSEARCH v2.7.1 was used to remove chimeric amplicons. The resulting read data set in fasta format, with the abundances in each sample, was uploaded to the Dryad repository (https://doi.org/10.5061/dryad.vt4b8gtq2).

Sequences were then clustered into molecular operational taxonomic units (MOTUs) with SWARM v2.1.7 using d = 13 (Bakker et al., 2019; Siegenthaler et al., 2019). Singletons (MOTUs with just one read) were removed after this step to minimize data loss (Atienza et al., 2020). Taxonomic assignment was performed using ecotag and a custom database containing sequences from the EMBL nucleotide database and sequences obtained from the Barcode of Life Database (BOLD), using a custom script to select the appropriate fragment (see details and a summary of the taxonomic groups represented in Wangensteen, Palacín, et al., 2018). This database contains 188,960 reference sequences covering most eukaryotic groups and is available from https://github.com/metabarpark/Reference-databases. Assignment of metazoan sequences was further improved by querying the BOLD database. Sequences with a species name assigned and with an identity match >95% in BOLD were kept, whereas matches below this threshold, even if assigned to species level by ecotag, were downgraded to genus level.

The final refining steps consisted of deleting any MOTU for which reads in blank or negative controls represented more than 10% of total reads for that MOTU in all samples. A minimum relative abundance filter was also applied, removing, for a given PCR replicate, the MOTUs that represented less than 0.005% of total reads of that replicate. We also removed MOTUs that had a combined total of <5 reads after the previous steps. Finally, all MOTUs that were not assigned to marine eukaryotes (i.e., MOTUs assigned to nonmarine organisms, prokaryotes, or to the root of the Tree of Life) were eliminated. We then pooled the three PCRs of each sample. We used the higher classification of eukaryotes proposed by Guillou et al. (2013) at the super-group level, with one exception: Opisthokontha was split into Metazoa and Fungi.

2.6 | Data analyses

Analyses were performed with the R package vegan (Oksanen et al., 2019). Rarefaction curves of the number of MOTUs obtained at an increasing number of reads were obtained with function rarecurve, separately for benthos and water samples. Likewise, MOTU accumulation curves with increasing numbers of samples were obtained for benthos and water with specaccum. MOTU richness values were compared with standard ANOVAs (factors community and sample type: benthos or water). Between-sample distances were computed using the Jaccard index based on presence/absence data of each MOTU per sample. These distances were then used to obtain ordinations of the samples in nonmetric multidimensional scaling (nmMDS) representations using function metaMDS with 500 random starts. Permutational analyses of variance were performed on Jaccard distances with function adonis to test differences between relevant factors: a one-way analysis was performed between benthos and water (all samples combined), a three-way analysis was done for the benthos with community and fraction as main factors and sample as a blocking factor nested in community. For the water, a two-way analysis was performed with community and distance to the wall (pelagic samples excluded as they were taken at an intermediate depth). Main factors were also tested for differences in multivariate dispersion (permdisp analysis using function betadisper) to check whether significant outcomes were a result of different multivariate heterogeneity (spread) or different centroid location of the groups. A Venn diagram was prepared with the VennDiagram package (Chen, 2018) to represent the degree of MOTU overlap between benthos and water. Upset diagrams were used to plot shared MOTUs at increasing distances of the benthic communities using package UpSetR (Conway et al., 2017).

3 | RESULTS

We obtained a total of 7,391,160 reads in total for the benthic samples (18 samples) and 13,652,493 reads for the water samples (27 samples). The controls had a negligible number of reads

(85.29 ± 19.80, mean ± *SE*). After quality filtering, demultiplexing, dereplicating and chimera elimination we had a total 3,868,827 unique COI sequences. These were clustered into 15,954 nonsingleton MOTUs. The final refining steps and, particularly, the elimination of MOTUs not assignable to marine eukaryotes using our reference database greatly reduced the data set to a final list of 3,543 MOTUs. The impact of removing noneukaryotic MOTUs was much greater in the water samples: only 14.35% of initial reads were retained at this step, while 99.36% were kept in the benthic samples. In the final data set, benthic samples had 2,396 MOTUs, while water samples had 1,412 MOTUs. The final average number of eukaryotic reads in benthic samples was 233.957 ± 25.40 (mean ± *SE*) and in water samples was much lower, 34.708 ± 2.50, as a result of the elimination of noneukaryotic MOTUs. Table S1 presents the final MOTU table with the taxonomic assignment and number of reads per sample. Rarefaction curves (Figure S2) showed that a plateau is reached in the number of MOTUs with the sequencing depth obtained in most samples from benthos and water (exceptions corresponded to some of the finer fractions in benthic samples). Likewise, MOTU accumulation curves (Figure S3) tended to saturate in water samples but not in benthic samples, so addition of more samples would probably increase the total number of MOTUs recovered from this habitat. In spite of the different number of total reads, we compared MOTU richness without rarefaction as in most samples the richness values plateaued at the sequencing depth obtained. Somewhat higher values were found in benthos (637.78 ± 59.00 and 420.34 ± 47.96 MOTUs in the photophilous and sciaphilous communities, respectively) compared to those in water at 0–1.5 m of distance (541.58 ± 29.40 and 389.92 ± 20.58 MOTUs, respectively). A two-way ANOVA showed that the number of MOTUs was not significantly different between benthos and water samples, but it was significantly higher in the photophilous than in the sciaphilous community (community effect, $p < .001$; sample type effect, $p = .110$; interaction, $p = .401$). The pelagic samples had 474.33 ± 28.50 MOTUs.

Taxonomic assignment revealed a total of seven super-groups in the samples, of which the most diverse was Metazoa (996 MOTUs, 45.47% of reads, all samples combined) followed by Archaeplastida (351 MOTUs, 16.47% of reads, mostly belonging to Rhodophyta), and Stramenopiles (287 MOTUs, 3.25% of reads). A total of 1,565 eukaryotic MOTUs could not be assigned to a given super-group. They represent 32.25% of total reads, but the share of unassigned reads was highly uneven: 21.94% of reads in benthic samples, and 78.58% in water samples. Within metazoans we identified 15 phyla, of which the most diverse were Arthropoda (211 MOTUs, 2.17% of total reads, all samples combined), followed by Annelida (116 MOTUs, 1.71% of reads), Cnidaria (74 MOTUs, 11.65% of reads), Porifera (59 MOTUs, 6.35% of reads) and Mollusca (50 MOTUs, 1.20% of reads). Among metazoans, 382 MOTUs could not be assigned at phylum or lower levels. In addition, 165 MOTUs could be assigned at the species level by ecotag with more than 0.95 identity with the best match in the reference database.

The relative number of MOTUs as per super-group and metazoan phylum obtained in the benthos and water samples is shown in Figure 2. The general patterns recovered were notably different in the two habitats surveyed. Metazoa were markedly dominant in the benthos in terms of number of MOTUs, followed by Archaeplastida (mostly Rhodophyta). On the other hand, Hacrobia (mostly Haptophyta) had the highest diversity in water samples, where other important planktonic groups such as the Alveolata had a much higher representation than in the benthos. Nevertheless, Metazoa was the second most MOTU-rich group in the water. As for metazoan phyla, the distribution was more similar: Arthropoda was the most diverse group in both habitats, and Annelida, Cnidaria, Mollusca and Porifera (albeit in different order) came next. However, the picture is different considering the relative number of reads: Cnidaria were dominant in the benthos (26.05% of metazoan reads), where the abundance of Arthropoda was much lower (3.88%). Conversely, in the water Arthropoda was the most abundant by far in proportion of metazoan reads (46.70%).

The number of MOTUs of the main metazoan phyla, Arthropoda, Annelida, Cnidaria, and Mollusca was further assessed at lower taxonomic levels (Order) in Table S2. In arthropods, Amphipoda, Decapoda, Isopoda and Harpacticoida were highly diverse in the benthos but practically absent from water samples, which were dominated by planktonic groups such as Calanoida and Cyclopoida. In annelids, Sabellida and Sipuncula were the most diverse groups in the plankton, while the dominant group in benthos (Phyllodocida) was practically absent in water samples (only four MOTUs in total). Among Cnidaria, only hydrozoans (Trachymedusae, Siphonophora, and Leptothecata) are diverse in the plankton samples, with a negligible representation of anthozoan orders which, together with Leptothecata, dominate in the benthic samples. Among Mollusca, highly diverse groups in the benthos such as Mytiloida, or gastropoda in general (with the exception of the pelagic Pteropoda) were absent or poorly represented in water samples. This perusal indicates that we did not capture in our samples planktonic stages of many benthic groups, and that the rates of DNA shedding from benthos to the water are in general low.

The sample ordination using the Jaccard index is shown in Figure 3a. A clear separation of benthic and water samples is evident, which is in agreement with one-way results comparing benthos and water, all samples pooled (PERMANOVA $p < .001$, and permdisp $p < .001$). In the benthos, the shallower and deeper communities formed clearly separated clusters. A PERMANOVA analysis on benthic samples alone showed a significant effect of community ($p < .001$) and of the nested factor sample (within community); while fraction or the interaction between community and fraction were not significant (Table 1). The permdisp test showed that there was also a different dispersion of data in the two communities ($p < .001$), which is also visible in the nmMDS. A second nmMDS was performed only with the water samples (Figure 3b), where a separation by communities can also be seen, albeit with some overlap. A PERMANOVA of water samples using community and distance to the wall as factors (pelagic samples were excluded in this analysis) showed a significant interaction term ($p = .027$, Table 2), indicating different effects of the community with increasing distances. A comparison

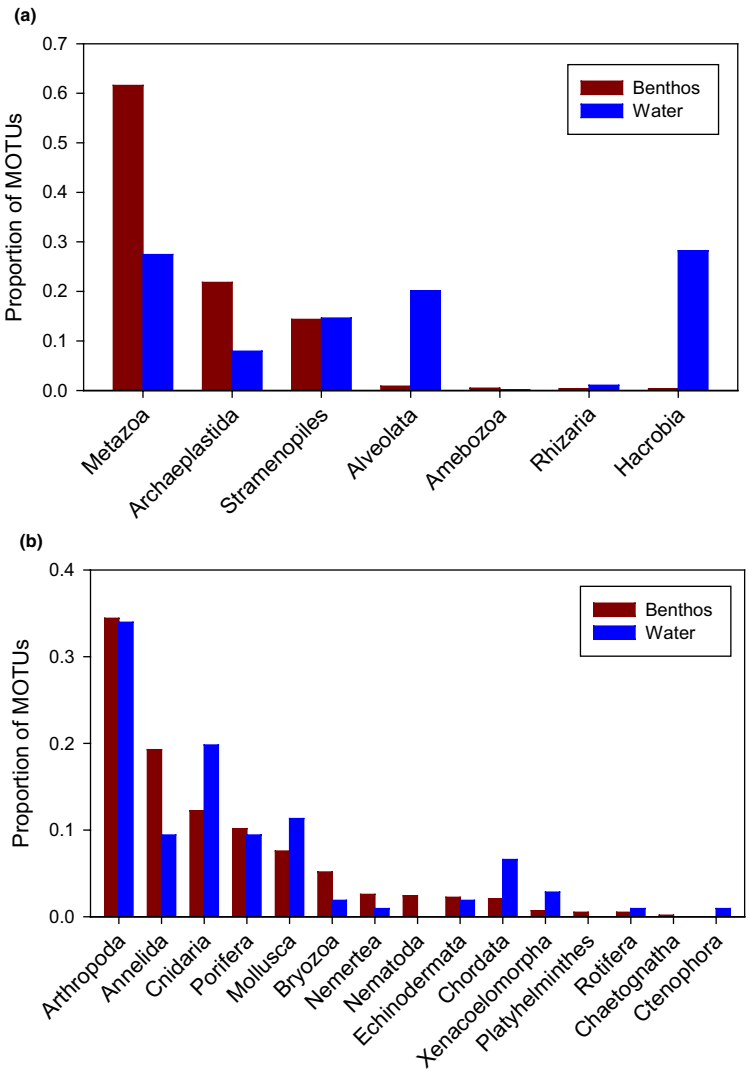


FIGURE 2 Barplot of relative MOTU richness of the super-groups (a) and metazoan phyla (b) detected in benthic and water samples

of the factor community at fixed distances showed that differences between photophilous and sciaphilous samples were significant at all distances (0, 0.5, and 1.5 m, all $p < .031$), and this was not due to differences in heterogeneity (all permdisp tests not significant). Likewise, a comparison of the factor distance at each depth level showed that distance to the rocky wall did not have a significant effect on the overall water assemblage composition ($p = .063$ and .056 for the photophilous and sciaphilous communities, respectively).

Of the total 3,543 MOTUs, only 265 were shared between benthos and water (Figure 4, Tables S3 and S4), which represented 11.06% of the MOTUs found in benthos. However, these 265 MOTUs accounted for 70.40% of the reads of the benthos, indicating that they correspond to abundant taxa. These same MOTUs accounted for 56.37% of the reads in the water samples. The MOTUs shared between benthos and water could be assigned to two main groups, those whose relative read abundance in the benthos was higher than in the water and those displaying the opposite pattern. We assume that the first group corresponds mainly to benthic MOTUs that left their DNA signature in the water (hereafter "shared benthic MOTUs" or SBM), while the second group probably corresponds to planktonic MOTUs (hereafter "shared pelagic MOTUs" or SPM). Only one MOTU could not be assigned to any of these categories as it had the same number of reads in both environments.

The first group (SBM) comprised 180 MOTUs (Table S3), which represented 7.51% and 70.33% of MOTUs and reads in the benthos, respectively, while they constituted 12.75% and 1.99% of the MOTUs and reads in the water. Of these MOTUs, almost half (84, 46.67%) belonged to metazoan groups, but only seven of them were arthropods (the dominant metazoan group in the plankton); the second most important group were the red algae (a mostly

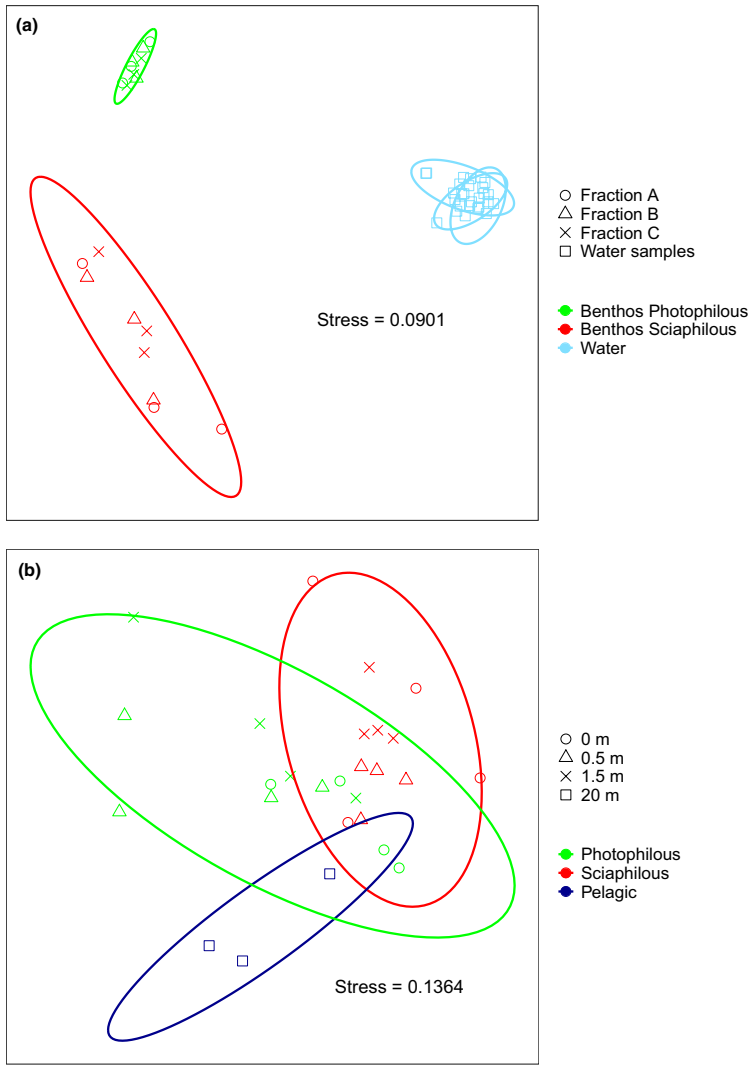


FIGURE 3 Nonmetric multidimentional scaling representation of all samples (a) and only water samples (b) using the Jaccard distance. Benthic samples (a) were separated in three different size fractions: A (>10 mm); B (between 10 mm and 1 mm); and C (between 1 mm and 63 μm). Communities are coded by colours and fractions (benthos) and distances (water) by symbols

TABLE 1 Results of the PERMANOVA analysis performed on Jaccard distances among the samples collected in two benthic communities (photophilous and sciaphilous) and separated into three size classes (fractions). Sample was added as a nested factor within community. Columns are: degrees of freedom (DF), sum of squares (SS), *F*-statistic of the model, with its associated probability (*p*-value), and probability of the permdisp test of multivariate homogeneity of group dispersions (Permdisp). Significant values marked with asterisk

Factor	*df*	SS	*F*-statistic	*p*-value	Permdisp
Community	1	1.581	5.442	0.001*	0.001*
Fraction	2	0.731	1.258	0.140	0.869
Community*fraction	2	0.653	1.124	0.267	
Sample (community)	2	1.158	1.993	0.002*	
Residuals	10	2.905			

Factor	*df*	SS	*F*-statistic	*p*-value	Permdisp
Community	1	0.265	4.127	0.001*	0.216
Distance	2	0.166	1.293	0.129	0.940
Community*distance	2	0.216	1.682	0.027*	
Residuals	18	1.157			

TABLE 2 Results of the PERMANOVA analysis performed on Jaccard distances among the water samples collected in two communities (photophilous and sciaphilous) and at three distances from the benthos (Distance factor: 0, 0.5 and 1.5 m). Columns are: degrees of freedom (DF), sum of squares (SS), *F*-statistic of the model, with its associated probability (*p*-value), and probability of the permdisp test of multivariate homogeneity of group dispersions (Permdisp). Significant values marked with asterisk

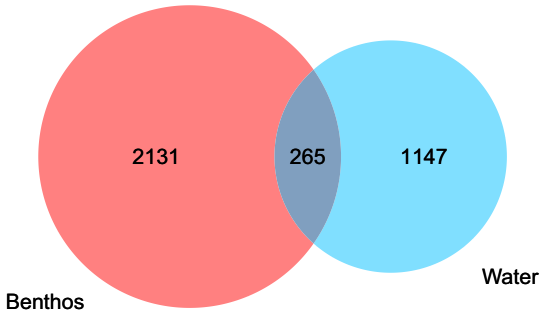


FIGURE 4 Venn diagram showing the overall MOTU overlap between the two types of community considered

benthic group), with 25 (13.89%) MOTUs. Of the dominant planktonic groups, only 11 (6.11%) SBM were diatoms and two were dinoflagellates. The taxonomic assignments were, therefore, mostly coherent with the idea that this subset of MOTUs belong mainly to benthic groups (Table S3). A total of 45 SBM MOTUs (25%) could not be assigned to any super-group.

The 84 shared pelagic MOTUs (SPM, Table S4) made up 3.51% of MOTUs but only 0.07% of reads in the benthos. On the other hand, while they comprised 5.95% of pelagic MOTUs they accounted for 54.44% of pelagic reads. Their taxonomic assignments showed that 22 (26.19%) MOTUs were metazoans, of which a majority (17) were arthropods. On the other hand, 18 (21.43%) MOTUs belonged to typical planktonic protists (diatoms, dinoflagellates, Hacrobia, Rhizaria) (Table S4). Finally, 42 (50%) SPM could not be assigned to any super-group. The higher number of unassigned MOTUs and the taxonomic composition suggest a dominance of nonbenthic groups in the SPM subset.

When the distribution of the 180 shared benthic MOTUs was examined, they clearly decreased with distance to the wall (Figure 5), with 135, 74, 24, and 15 MOTUs shared between benthos and water samples at 0, 0.5, 1.5 and 20 m, respectively. Their abundance in

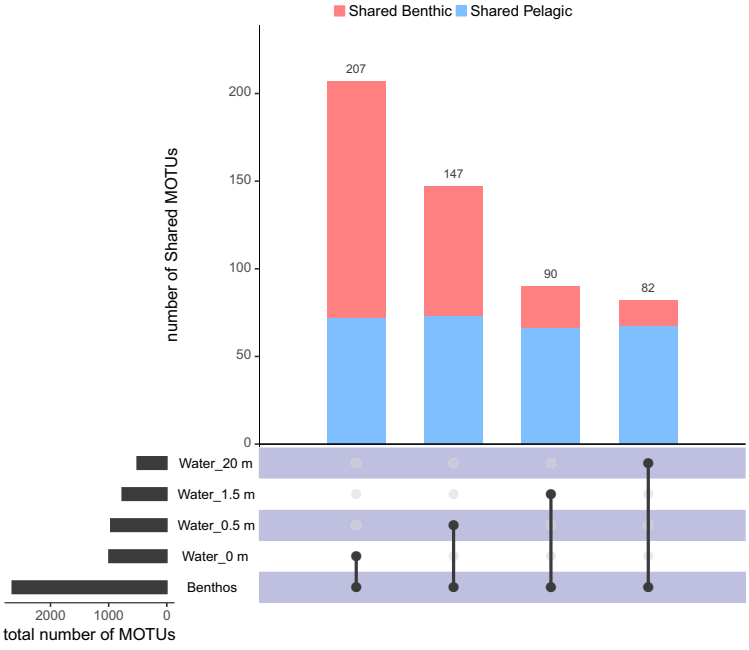


FIGURE 5 Upset plot with the number of shared MOTUs between the benthos and the water samples and the total number of MOTUs detected. Shared benthic MOTUs (SBM) are represented in pink and shared pelagic MOTUs (SPM) in light blue

relative read numbers also decreased (from 0.056 to 0.002, Table S3), which supports the idea of their benthic origin. This same general pattern was found when both communities studied were analysed separately (Figures S4 and S5).

By contrast, the comparison of shared pelagic MOTUs did not show any clear trend with distance to the wall (Figure 5): 72, 73, 66, and 67 at 0, 0.5, 1.5, and 20 m, respectively. Neither was a trend found in relative read abundances per sample (between 0.570 and 0.526 irrespective of distance, Table S4). Again, this same general pattern was found in both communities separately (Figures S4 and S5).

4 | DISCUSSION

Metabarcoding of benthos and water samples, using a broad range eukaryotic marker (COI), retrieved clearly different communities. The patterns of MOTU richness and abundance of reads from the different environments were distinct, showing a dominance of taxa with important planktonic components (such as dinoflagellates, diatoms, and haptophytes) in the water samples, while metazoans and rhodophytes were the most diverse and abundant in the benthos. Metazoans, notwithstanding, were also well represented in water samples, with a dominance of arthropods (mostly calanoids and cyclopoids) in both number of MOTUs and reads. The rarefaction and MOTU accumulation curves showed that we captured adequately the richness present in the samples with our sequencing depth, and that the total eukaryotic diversity in the benthos was higher than that in the water. More replicates of benthic samples would be necessary to recover the overall MOTU richness of this habitat.

However, we acknowledge that the sampling methods used were different for benthos and plankton. We have used techniques currently applied to sample these environments. In complex communities such as the benthos, with organism sizes spanning several orders of magnitude, size-fractionation is necessary to recover the biodiversity present (Elbrecht et al., 2017; Wangensteen, Palacín, et al., 2018; Wangensteen & Turon, 2017). In addition, the mesh size used for the smallest sieve was 63 μm, meaning that most prokaryotes and a significant part of the smallest microeukaryotes were washed out, along with cell debris and extracellular DNA. In the filters, on the other hand, we retained everything down to colloidal level, thus the prokaryotic community, for instance, was captured in our samples. This explains the amount of reads that had to be discarded in the water samples as not assignable to eukaryotes and, within eukaryotes, the high number of reads that could not be assigned to any supergroup (the smallest eukaryotes being the less represented in the reference database for COI). Our point was not to test both techniques or to compare their particularities, but rather to check if the information retrieved from currently established methods for the analysis of water DNA is comparable to that from current analytical techniques for benthos.

While the DNA obtained from the filters would be labelled as eDNA, the sampling from the benthos would be qualified as community or bulk DNA by many. eDNA is defined as the DNA obtained from an environmental matrix such as water or sediment without isolating the organisms (Barnes & Turner, 2016; Creer et al., 2016; Stewart, 2019; Taberlet, Coissac, Hajibabaei, et al., 2012); and is usually opposed to bulk or community DNA, referring to DNA obtained from organisms previously isolated from the environment (Andújar et al., 2018; Creer et al., 2016; Deiner et al., 2017). In a more restricted sense (e.g., Andújar et al., 2018; Cristescu & Hebert, 2018; Thomsen et al., 2012; Tsuji et al., 2019), the term eDNA is used as equivalent to trace DNA released from organisms (in the form of mucus, faeces, cells, hairs, etc), so when studying eDNA the organisms themselves are not in the sample. We consider, however, that eDNA should be used as a general term, to designate any DNA extracted from an environmental sample. It is commonly made up of a mix of intraorganismal (in the form of small organisms relative to the sample size) and extra-organismal or trace eDNA shed from large organisms (Creer et al., 2016; Pawlowski et al., 2018; Porter & Hajibabaei, 2018; Salter, 2018; Taberlet, Coissac, Hajibabaei, et al., 2012). The relative amount of both components is highly variable, though, and it depends on the sampling method and the target group, and hence the primers used. In our case, we used a broadly universal primer set for eukaryotes, capable of amplifying both intraorganismal and trace DNA from most eukaryotic taxa. So the benthic samples are more enriched in intraorganismal DNA (since most trace DNA was removed by sieving), while the water samples contain a mix of a high amount of intraorganismal DNA from planktonic microeukaryotes and a smaller fraction of extra-organismal DNA from larger organisms.

The ordination and PERMANOVA results confirmed the marked differentiation between the samples from both environments. An assessment at the Order level in the main metazoan phyla confirmed that the composition of the two environments is highly different. Moreover, the differences between the two depths sampled, which corresponded to two different communities (photophilous and sciaphilous) on precisely the same wall, were pronounced in the benthic samples, but were also significant in the water samples taken between 0 and 1.5 m of the rocky wall. Thus, the method is sensitive enough to detect ecological differences not just in the sessile communities, but also in the more dynamic planktonic habitat. This is in agreement with other studies that have also shown that the eDNA in seawater samples can detect differences in composition of several groups at relatively small scales (from metres to tens of metres, Jacobs-Palmer et al., 2020; Jeunen et al., 2019; Port et al., 2016).

A total of 3,543 eukaryotic MOTUs were detected in the whole data set. In spite of the lower number of eukaryotic reads retrieved from the water (15% of those retrieved from the benthos), the number of eukaryotic MOTUs in the water was ca. 60% of those in the benthos (1,412 as compared to 2,396). Only 265 MOTUs were found to be shared between the benthos and the water samples. This represents only ca. 11% and 19% of the MOTUs in the two environments, respectively. In addition, a closer scrutiny allowed us to

separate those shared MOTUs into those of possibly benthic origin (shared benthic MOTUs, SBM) and those of probably planktonic origin (shared pelagic MOTUs, SPM).

The 180 SBM comprised ca. 7.5% of the benthic MOTUs but represented ca. 70% or benthic reads (while only ca. 2% of water-derived reads), indicating that abundant benthic MOTUs are the ones more prone to leave their signature in the surrounding water. The 84 SPM accounted to ca. 6% of pelagic MOTUs but ca. 54% or eukaryotic pelagic reads (and only 0.07% of reads in the benthos), again indicating that the most abundant MOTUs are the ones that can be detected also in the other habitat.

The fine-scale distribution of the 180 SBM showed a clear trend: more MOTUs were shared in the immediate vicinity of the benthos (135 with water at 0 m), and the number decreased with distance down to only 15 MOTUs shared with the water at 20 m. The shared MOTUs also represented a decreasing percent of reads in the water samples as we moved away from the rocky wall. On the other hand, there was no clear pattern of abundance changes with distance in the richness or amount of reads shared between benthos and water for the 84 PSM.

We found therefore evidence for DNA originating from the benthic communities being present in the adjacent water layer and, conversely, DNA of presumably pelagic origin could be detected in the benthos. The interest of this article was in detecting the presence of benthic DNA in the water column, of which only a modest amount could be retrieved. The form of this benthic DNA in the water cannot be assessed with our sampling design, but it probably includes naturally released meroplanktonic components, such as gametes (Tsuji & Shibata, 2020) or larvae, and degradation products in the form of fragments, mucus, cell aggregates, exudates, or extracellular DNA.

Our results clearly indicated that DNA from water samples is a poor surrogate for the analysis of benthic communities, as found previously in freshwater environments (Hajibabaei et al., 2019; Gleason et al., 2020). Even in the water within a few centimetres from the benthos, only a modest portion (135) of the benthic MOTUs could be detected. In addition, we found that considering the relative number of reads of the shared MOTUs provided useful insights about the origin of the MOTUs and their dynamics as we move farther from the rocky wall. The lack of accordance between benthos and water is in agreement with previous comparisons of different substrates for eDNA made in port environments (e.g., Koziol et al., 2019; Rey et al., 2020) which found different community profiles in water and in sediments or settlement plates. We must keep in mind that we have used universal primers as we targeted the whole eukaryotic communities. With more specific targets, the results could be different. For instance, using vertebrate-specific primers to detect fish in the water has proved to be a sensitive method (e.g., Bakker et al., 2017; Sales et al., 2019; Salter et al., 2019; Sigsgaard et al., 2019; Thomsen et al., 2016), even at the intraspecific level (Sigsgaard et al., 2020), since it is possible to amplify selectively the DNA of the target group. Likewise, species-specific primers have been successfully used to detect particular marine benthic species in the water column, usually as a means of monitoring invasive species (e.g., Pochon, Bott, Smith, & Wood, 2013; Simpson et al., 2017; von Ammon et al., 2019).

It seems reasonable to expect that DNA shedding rates from a highly diverse community such as sublittoral rocky bottom assemblages would be unbalanced between groups, and that this unevenness would hinder our ability to extract reliable monitoring information from seawater eDNA. This expectation is borne out by our results. Thus, albeit for group-specific or species-specific studies useful information from benthic groups may be gleaned from water DNA, the method is presently unsuitable for the community-wide diversity assessment required for many biomonitoring applications. New technologies affording much higher sequencing depth or metagenomic approaches (Singer et al., 2019; Singer et al., 2020) might improve our ability to extract information from water samples. But for the time being we must continue to rely on methods that can sample directly the benthos for reliable biodiversity assessment of these complex assemblages.

ACKNOWLEDGEMENTS

We thank the authorities of the Cabrera Archipelago National Park for granting the permission to perform the field work. We also thank Daniel San Román for help with the laboratory procedures. This research has been funded by project BIGPARK of the Spanish Autonomous Organism of National Parks (OAPN, project 2017-2462), and by projects PopCOmics (CTM2017-88080, MINECO/AEI/FEDER,UE) and ANIMA (CGL2016-76341-R, MINECO/AEI/FEDER,UE) of the Spanish Government. AA was funded by a predoctoral FPI contract of the Spanish Government. This is a contribution from the Consolidated Research Group "Benthic Biology and Ecology" SGR2017-1120 (Catalan Government).

AUTHOR CONTRIBUTIONS

A.A. performed laboratory and bioinformatics work, prepared tables and figures and drafted the paper; C.P., designed research, analysed data and revised the paper; E.C., performed field work, contributed funding and revised the paper; R.G., performed field work, analysed data and revised the paper; O.S.W., designed research, contributed reagents and analytical tools, analysed data and revised the paper; X.T., designed research, performed field work, contributed funding, analysed data and revised the paper.

DATA AVAILABILITY STATEMENT

The original read data set, with the abundances in each sample, was uploaded to the Dryad Data repository (https://doi.org/10.5061/dryad.vt4b8gtq2).

The final MOTU data set has been uploaded as online Supporting Information.

REFERENCES

Alberdi, A., Aizpurua, O., Thomas, M., Gilbert, P., & Bohmann, K. (2018). Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution*, *9*, 134–147. https://doi.org/10.1111/2041-210X.12849

Alexander, J. B., Bunce, M., White, N., Wilkinson, S. P., Adam, A. A. S., Berry, T., Stat, M., Thomas, L., Newman, S. J., Dugal, L., & Richards, Z. T. (2020). Development of a multi-assay approach for monitoring coral diversity using eDNA metabarcoding. *Coral Reefs*, *39*(1), 159–171. https://doi.org/10.1007/s00338-019-01875-9

Andújar, C., Arribas, P., Yu, D. W., Vogler, A. P., & Emerson, B. C. (2018). Why the COI barcode should be the community DNA metabarcode for the metazoa. *Molecular Ecology*, *27*(20), 3968–3975. https://doi.org/10.1111/mec.14844

Atienza, S., Guardiola, M., Præbel, K., Antich, A., Turon, X., & Wangensteen, O. S. (2020). DNA metabarcoding of deep-sea sediment communities using COI: Community assessment, spatio-temporal patterns and comparison with 18S rDNA. *Diversity*, *12*(4), 123. https://doi.org/10.3390/d12040123

Aylagas, E., Borja, Á., Muxika, I., & Rodríguez-Ezpeleta, N. (2018). Adapting metabarcoding-based benthic biomonitoring into routine marine ecological status assessment networks. *Ecological Indicators*, *95*, 194–202. https://doi.org/10.1016/J.ECOLIND.2018.07.044

Baird, D. J., & Hajibabaei, M. (2012). Biomonitoring 2.0: A new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology*, *21*, 2039–2044. https://doi.org/10.1111/j.1365-294X.2012.05519.x

Bakker, J., Wangensteen, O. S., Baillie, C., Buddo, D., Chapman, D. D., Gallagher, A. J., Guttridge, T. L., Hertler, H., & Mariani, S. (2019). Biodiversity assessment of tropical shelf eukaryotic communities via pelagic eDNA metabarcoding. *Ecology and Evolution*, *9*(24), 14341–14355. https://doi.org/10.1002/ece3.5871

Bakker, J., Wangensteen, O. S., Chapman, D. D., Boussarie, G., Buddo, D., Guttridge, T. L., Hertler, H., Mouillot, D., Vigliola, L., & Mariani, S. (2017). Environmental DNA reveals tropical shark diversity in contrasting levels of anthropogenic impact. *Scientific Reports*, *7*(1), 16886. https://doi.org/10.1038/s41598-017-17150-2

Bani, A., De Brauwer, M., Creer, S., Dumbrell, A. J., Limmon, G., Jompa, J., & Beger, M. (2020). Informing marine spatial planning decisions with environmental DNA. *Advances in Ecological Research*, *62*, 375–407. https://doi.org/10.1016/bs.aecr.2020.01.011

Barnes, M. A., & Turner, C. R. (2016). The ecology of environmental DNA and implications for conservation genetics. *Conservation Genetics*, *17*(1), 1–17. https://doi.org/10.1007/s10592-015-0775-4

Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., Yu, D. W., & de Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology & Evolution*, *29*(6), 358–367. https://doi.org/10.1016/J.TREE.2014.04.003

Boudreau, B. P., & B. B. Jorgensen (Eds.) (2001). *The benthic boundary layer: Transport processes and biogeochemistry*. Oxford University Press.

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., & Coissac, E. (2016). Obitools: A unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, *16*, 176–182. https://doi.org/10.1111/1755-0998.12428

Brannock, P., Learman, D., Mahon, A., Santos, S., & Halanych, K. (2018). Meiobenthic community composition and biodiversity along a 5500 km transect of Western Antarctica: A metabarcoding analysis. *Marine Ecology Progress Series*, *603*, 47–60. https://doi.org/10.3354/meps12717

Brannock, P. M., Ortmann, A. C., Moss, A. G., & Halanych, K. M. (2016). Metabarcoding reveals environmental factors influencing spatio-temporal variation in pelagic micro-eukaryotes. *Molecular Ecology*, *25*(15), 3593–3604. https://doi.org/10.1111/mec.13709

Cahill, A. E., Pearman, J. K., Borja, A., Carugati, L., Carvalho, S., Danovaro, R., Dashfield, S., David, R., Féral, J.-P., Olenin, S., Šiaulys, A., Somerfield, P. J., Trayanova, A., Uyarra, M. C., & Chenuil, A. (2018). A comparative analysis of metabarcoding and morphology-based identification of benthic communities across different regional seas. *Ecology and Evolution*, *8*(17), 8908–8920. https://doi.org/10.1002/ece3.4283

Chen, H. (2018). VennDiagram: Generate high-resolution Venn and Euler plots (1.6.20). Retrieved from https://cran.r-project.org/package=VennDiagram

Collins, R. A., Bakker, J., Wangensteen, O. S., Soto, A. Z., Corrigan, L., Sims, D. W., Genner, M. J., & Mariani, S. (2019). Non-specific amplification compromises environmental DNA metabarcoding with COI. *Methods in Ecology and Evolution*, *10*(11), 1985–2001. https://doi.org/10.1111/2041-210X.13276

Collins, R. A., Wangensteen, O. S., O'Gorman, E. J., Mariani, S., Sims, D. W., & Genner, M. J. (2018). Persistence of environmental DNA in marine systems. *Communications Biology*, *1*(1), 185. https://doi.org/10.1038/s42003-018-0192-6

Conway, J., Lex, A., & Gehlenborg, N. (2017). UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics*, *33*(18), 2938–2940. https://doi.org/10.1093/bioinformatics/btx364

Cordier, T., & Pawlowski, J. (2018). BBI: An R package for the computation of Benthic Biotic Indices from composition data. *Metabarcoding and Metagenomics*, *2*, e25649. https://doi.org/10.3897/mbmg.2.25649

Cowart, D. A., Matabos, M., Brandt, M. I., Marticorena, J., & Sarrazin, J. (2020). Exploring environmental DNA (eDNA) to assess biodiversity of hard substratum faunal communities on the lucky strike vent field (mid-Atlantic ridge) and investigate recolonization dynamics after an induced disturbance. *Frontiers in Marine Science*, *6*, 783. https://doi.org/10.3389/fmars.2019.00783

Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Thomas, W. K., Potter, C., & Bik, H. M. (2016). The ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology and Evolution*, *7*(9), 1008–1018. https://doi.org/10.1111/2041-210X.12574

Cristescu, M. E. (2014). From barcoding single individuals to metabarcoding biological communities: Towards an integrative approach to the study of global biodiversity. *Trends in Ecology and Evolution*, *29*, 566–571. https://doi.org/10.1016/j.tree.2014.08.001

Cristescu, M. E., & Hebert, P. D. N. (2018). Uses and misuses of environmental DNA in biodiversity science and conservation. *Annual Review of Ecology, Evolution, and Systematics*, *49*(1), 209–230. https://doi.org/10.1146/annurev-ecolsys-110617-062306

Danovaro, R., Carugati, L., Berzano, M., Cahill, A. E., Carvalho, S., Chenuil, A., Corinaldesi, C., Cristina, S., David, R., Dell'Anno, A., Dzhembekova, N., Garcés, E., Gasol, J. M., Goela, P., Féral, J.-P., Ferrera, I., Forster, R. M., Kurekin, A. A., Rastelli, E., … Borja, A. (2016). Implementing and innovating marine monitoring approaches for assessing marine environmental status. *Frontiers in Marine Science*, *3*, 213. https://doi.org/10.3389/fmars.2016.00213

Darling, J. A., Galil, B. S., Carvalho, G. R., Rius, M., Viard, F., & Piraino, S. (2017). Recommendations for developing and applying genetic tools to assess and manage biological invasions in marine ecosystems. *Marine Policy*, *85*, 54–64. https://doi.org/10.1016/j.marpol.2017.08.014

De Jode, A., David, R., Dubar, J., Rostan, J., Guillemain, D., Sartoretto, S., &Chenuil, A. (2019). Community ecology of coralligenous assemblages using a metabarcoding approach. 3rd Mediterranean Symposium on the Conservation of Coralligenous & Other Calcareous Bio-Concretions, 41–45. Retrieved from https://hal.archives-ouvertes.fr/hal-02048506

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., & Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, *26*(21), 5872–5895. https://doi.org/10.1111/mec.14350

Djurhuus, A., Pitz, K., Sawaya, N. A., Rojas-Márquez, J., Michaud, B., Montes, E., Muller-Karger, F., & Breitbart, M. (2018). Evaluation of

marine zooplankton community structure through environmental DNA metabarcoding. *Limnology and Oceanography: Methods*, *16*(4), 209–221. https://doi.org/10.1002/lom3.10237

Elbrecht, V., Peinert, B., & Leese, F. (2017). Sorting things out: Assessing effects of unequal specimen biomass on DNA metabarcoding. *Ecology and Evolution*, *7*(17), 6918–6926. https://doi.org/10.1002/ece3.3192

Fonseca, V. G., Carvalho, G. R., Nichols, B., Quince, C., Johnson, H. F., Neill, S. P., Lambshead, J. D., Thomas, W. K., Power, D. M., & Creer, S. (2014). Metagenetic analysis of patterns of distribution and diversity of marine meiobenthic eukaryotes. *Global Ecology and Biogeography*, *23*(11), 1293–1302. https://doi.org/10.1111/geb.12223

Fraija-Fernández, N., Bouquieaux, M.-C., Rey, A., Mendibil, I., Cotano, U., Irigoien, X., & Rodríguez-Ezpeleta, N. (2019). Marine water environmental DNA metabarcoding provides a comprehensive fish diversity assessment and reveals spatial patterns in a large oceanic area. *BioRxiv*, 864710, https://doi.org/10.1101/864710

Geller, J., Meyer, C., Parker, M., & Hawk, H. (2013). Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. *Molecular Ecology Resources*, *13*(5), 851–861.

Gerhard, W. A., & Gunsch, C. K. (2019). Metabarcoding and machine learning analysis of environmental DNA in ballast water arriving to hub ports. *Environment International*, *124*, 312–319. https://doi.org/10.1016/J.ENVINT.2018.12.038

Gleason, J. E., Elbrecht, V., Braukmann, T. W. A., Hanner, R. H., & Cottenie, K. (2020). Assessment of stream macroinvertebrate communities with eDNA is not congruent with tissue based metabarcoding. *Molecular Ecology*, early view. https://doi.org/10.1111/mec.15597

Goodwin, K. D., Thompson, L. R., Duarte, B., Kahlke, T., Thompson, A. R., Marques, J. C., & Caçador, I. (2017). DNA sequencing as a tool to monitor marine ecological status. *Frontiers in Marine Science*, *4*, 107. https://doi.org/10.3389/fmars.2017.00107

Guardiola, M., Wangensteen, O. S., Taberlet, P., Coissac, E., Uriz, M. J., & Turon, X. (2016). Spatio-temporal monitoring of deep-sea communities using metabarcoding of sediment DNA and RNA. *PeerJ*, *4*, e2807. https://doi.org/10.7717/peerj.2807

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., & Christen, R. (2013). The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, *41*(D1), 13–14. https://doi.org/10.1093/nar

Hajibabaei, M., Baird, D. J., Fahner, N. A., Beiko, R., & Golding, G. B. (2016). A new way to contemplate Darwin's tangled bank: How DNA barcodes are reconnecting biodiversity science and biomonitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1702), 20150330. https://doi.org/10.1098/rstb.2015.0330

Hajibabaei, M., Porter, T. M., Robinson, C. V., Baird, D. J., Shokralla, S., & Wright, M. T. G. (2019). Watered-down biodiversity? A comparison of metabarcoding results from DNA extracted from matched water and bulk tissue biomonitoring samples. *PLoS One*, *14*(12), e0225409. https://doi.org/10.1371/journal.pone.0225409

Hering, D., Borja, A., Jones, J. I., Pont, D., Boets, P., Bouchez, A., Bruce, K., Drakare, S., Hänfling, B., Kahlert, M., Leese, F., Meissner, K., Mergen, P., Reyjol, Y., Segurado, P., Vogler, A., & Kelly, M. (2018). Implementation options for DNA-based identification into ecological status assessment under the European Water Framework Directive. *Water Research*, *138*, 192–205. https://doi.org/10.1016/J.WATRES.2018.03.003

Holman, L. E., de Bruyn, M., Creer, S., Carvalho, G., Robidart, J., & Rius, M. (2019). Detection of introduced and resident marine species using environmental DNA metabarcoding of sediment and water. *Scientific Reports*, *9*(1), 11559. https://doi.org/10.1038/s41598-019-47899-7

Jacobs-Palmer, E., Gallego, R., Ramón-Laca, A., Kunselman, E., Cribari, K., Horwith, M., & Kelly, R. P. (2020). A halo of reduced dinoflagellate abundances in and around eelgrass beds. *PeerJ*, *8*, e8869. https://doi.org/10.7717/peerj.8869

Jeunen, G.-J., Lamare, M. D., Knapp, M., Spencer, H. G., Taylor, H. R., Stat, M., Bunce, M., & Gemmell, N. J. (2019). Water stratification in the marine biome restricts vertical environmental DNA (eDNA) signal dispersal. *Environmental DNA*, *2*(1), 99–111. https://doi.org/10.1002/edn3.49

Kelly, R. P., Port, J. A., Yamahara, K. M., & Crowder, L. B. (2014). Using environmental DNA to census marine fishes in a large mesocosm. *PLoS One*, *9*(1), e86175. https://doi.org/10.1371/journal.pone.0086175

Kelly, R. P., Shelton, A. O., & Gallego, R. (2019). Understanding PCR processes to draw meaningful conclusions from environmental DNA studies. *Scientific Reports*, *9*(1), 1–14. https://doi.org/10.1038/s41598-019-48546-x

Koziol, A., Stat, M., Simpson, T., Jarman, S., DiBattista, J. D., Harvey, E. S., Marnane, M., McDonald, J., & Bunce, M. (2019). Environmental DNA metabarcoding studies are critically affected by substrate selection. *Molecular Ecology Resources*, *19*(2), 366–376. https://doi.org/10.1111/1755-0998.12971

Krehenwinkel, H., Pomerantz, A., & Prost, S. (2019). Genetic biomonitoring and biodiversity assessment using portable sequencing technologies: Current uses and future directions. *Genes*, *10*(11), 858. https://doi.org/10.3390/genes10110858

Leduc, N., Lacoursière-Roussel, A., Howland, K. L., Archambault, P., Sevellec, M., Normandeau, E., Dispas, A., Winkler, G., McKindsey, C. W., Simard, N., & Bernatchez, L. (2019). Comparing eDNA metabarcoding and species collection for documenting Arctic metazoan biodiversity. *Environmental DNA*, *1*(4), 342–358. https://doi.org/10.1002/edn3.35

Leese, F., Bouchez, A., Abarenkov, K., Altermatt, F., Borja, Á., Bruce, K., & Weigand, A. M. (2018). Why we need sustainable networks bridging countries, disciplines, cultures and generations for aquatic biomonitoring 2.0: A perspective derived from the DNAqua-Net COST action. *Advances in Ecological Research*, *58*, 63–99. https://doi.org/10.1016/bs.aecr.2018.01.001

Leray, M., & Knowlton, N. (2015). DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(7), 2076–2081. https://doi.org/10.1073/pnas.1424997112

Leray, M., & Knowlton, N. (2016). Censusing marine eukaryotic diversity in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1702), 20150331. https://doi.org/10.1098/rstb.2015.0331

Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., Chambouvet, A., Christen, R., Claverie, J.-M., Decelle, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Forn, I., Forster, D., Guillou, L., Jaillon, O., Kooistra, W. H. C. F., Logares, R., … de Vargas, C. (2015). Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. *Environmental Microbiology*, *17*(10), 4035–4049. https://doi.org/10.1111/1462-2920.12955

McGee, K. M., Robinson, C. V., & Hajibabaei, M. (2019). Gaps in DNA-based biomonitoring across the globe. *Frontiers in Ecology and Evolution*, *7*, 337. https://doi.org/10.3389/fevo.2019.00337

O'Donnell, J. L., Kelly, R. P., Shelton, A. O., Samhouri, J. F., Lowell, N. C., & Williams, G. D. (2017). Spatial distribution of environmental DNA in a nearshore marine habitat. *PeerJ*, *5*, e3044. https://doi.org/10.7717/peerj.3044

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., & Wagner, H. (2019). *vegan: Community ecology package*. Retrieved from https://cran.r-project.org/web/packages/vegan/index.html

Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., Borja, A., Bouchez, A., Cordier, T., Domaizon, I., Feio, M. J., Filipe, A. F., Fornaroli, R., Graf, W., Herder, J., van der Hoorn, B., Iwan Jones, J., Sagova-Mareckova,

M., Moritz, C., … Kahlert, M. (2018). The future of biotic indices in the ecogenomic era: Integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. *Science of the Total Environment*, *637–638*, 1295–1310. https://doi.org/10.1016/J.SCITOTENV.2018.05.002

Pearman, J. K., Aylagas, E., Voolstra, C. R., Anlauf, H., Villalobos, R., & Carvalho, S. (2019). Disentangling the complex microbial community of coral reefs using standardized Autonomous Reef Monitoring Structures (ARMS). *Molecular Ecology*, *28*(15), 3496–3507. https://doi.org/10.1111/mec.15167

Peters, L., Spatharis, S., Dario, M. A., Dwyer, T., Roca, I. J. T., Kintner, A., Kanstad-Hanssen, Ø., Llewellyn, M. S., & Praebel, K. (2018). Environmental DNA: A new low-cost monitoring tool for pathogens in salmonid aquaculture. *Frontiers in Microbiology*, *9*, 3009. https://doi.org/10.3389/fmicb.2018.03009

Pochon, X., Bott, N. J., Smith, K. F., & Wood, S. A. (2013). Evaluating detection limits of next-generation sequencing for the surveillance and monitoring of international marine pests. *PLoS ONE*, *8*(9), e73935. http://dx.doi.org/10.1371/journal.pone.0073935

Port, J. A., O'Donnell, J. L., Romero-Maraccini, O. C., Leary, P. R., Litvin, S. Y., Nickols, K. J., Yamahara, K. M., & Kelly, R. P. (2016). Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA. *Molecular Ecology*, *25*(2), 527–541. https://doi.org/10.1111/mec.13481

Porter, T. M., & Hajibabaei, M. (2018). Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Molecular Ecology*, *27*(2), 313–338. https://doi.org/10.1111/mec.14478

Ransome, E., Geller, J. B., Timmers, M., Leray, M., Mahardini, A., Sembiring, A., Collins, A. G., & Meyer, C. P. (2017). The importance of standardization for biodiversity comparisons: A case study using autonomous reef monitoring structures (ARMS) and metabarcoding to measure cryptic diversity on Mo'orea coral reefs. *French Polynesia. PLOS One*, *12*(4), e0175066. https://doi.org/10.1371/journal.pone.0175066

Rey, A., Basurko, O. C., & Rodriguez-Ezpeleta, N. (2020). Considerations for metabarcoding-based port biological baseline surveys aimed at marine nonindigenous species monitoring and risk assessments. *Ecology and Evolution*, *10*(5), 2452–2465. https://doi.org/10.1002/ece3.6071

Sales, N. G., Wangensteen, O. S., Carvalho, D. C., & Mariani, S. (2019). Influence of preservation methods, sample medium and sampling time on eDNA recovery in a neotropical river. *Environmental DNA*, *1*(2), 119–130. https://doi.org/10.1002/edn3.14

Salter, I. (2018). Seasonal variability in the persistence of dissolved environmental DNA (eDNA) in a marine system: The role of microbial nutrient limitation. *PLoS One*, *13*(2), e0192409. https://doi.org/10.1371/journal.pone.0192409

Salter, I., Joensen, M., Kristiansen, R., Steingrund, P., & Vestergaard, P. (2019). Environmental DNA concentrations are correlated with regional biomass of Atlantic cod in oceanic waters. *Communications Biology*, *2*(1), 1–9. https://doi.org/10.1038/s42003-019-0696-8

Shaw, J. L. A., Weyrich, L., & Cooper, A. (2017). Using environmental (e)DNA sequencing for aquatic biodiversity surveys: A beginner's guide. *Marine and Freshwater Research*, *68*, 20–33. https://doi.org/10.1071/MF15361

Shu, L., Ludwig, A., & Peng, Z. (2020). Standards for methods utilizing environmental DNA for detection of fish species. *Genes*, *11*(3), 296. https://doi.org/10.3390/genes11030296

Shum, P., Barney, B. T., O'Leary, J. K., & Palumbi, S. R. (2019). Cobble community DNA as a tool to monitor patterns of biodiversity within kelp forest ecosystems. *Molecular Ecology Resources*, *19*(6), 1470–1485. https://doi.org/10.1111/1755-0998.13067

Siegenthaler, A., Wangensteen, O. S., Benvenuto, C., Campos, J., & Mariani, S. (2019). DNA metabarcoding unveils multiscale trophic variation in a widespread coastal opportunist. *Molecular Ecology*, *28*(2), 232–249. https://doi.org/10.1111/mec.14886

Sigsgaard, E. E., Jensen, M. R., Winkelmann, I. E., Møller, P. R., Hansen, M. M., & Thomsen, P. F. (2020). Population-level inferences from environmental DNA—Current status and future perspectives. *Evolutionary Applications*, *13*(2), 245–262. https://doi.org/10.1111/eva.12882

Sigsgaard, E. E., Torquato, F., Frøslev, T. G., Moore, A. B. M., Sørensen, J. M., Range, P., Ben-Hamadou, R., Bach, S. S., Møller, P. R., & Thomsen, P. F. (2019). Using vertebrate environmental DNA from seawater in biomonitoring of marine habitats. *Conservation Biology*, *34*(3), 697–710. https://doi.org/10.1111/cobi.13437

Simpson, T. J. S., Smale, D. A., McDonald, J. I., & Wernberg, T. (2017). Large scale variability in the structure of sessile invertebrate assemblages in artificial habitats reveals the importance of local-scale processes. *Journal of Experimental Marine Biology and Ecology*, *494*, 10–19. https://doi.org/10.1016/J.JEMBE.2017.05.003

Singer, G. A. C., Fahner, N. A., Barnes, J. G., McCarthy, A., & Hajibabaei, M. (2019). Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: A case study of eDNA metabarcoding seawater. *Scientific Reports*, *9*(1), 1–12. https://doi.org/10.1038/s41598-019-42455-9

Singer, G. A. C., Shekarriz, S., McCarthy, A., Fahner, N., & Hajibabaei, M. (2020). The utility of a metagenomics approach for marine biomonitoring. *BioRxiv*, https://doi.org/10.1101/2020.03.16.993667

Sinniger, F., Pawlowski, J., Harii, S., Gooday, A. J., Yamamoto, H., Chevaldonné, P., Cedhagen, T., Carvalho, G., & Creer, S. (2016). Worldwide analysis of sedimentary DNA reveals major gaps in taxonomic knowledge of deep-sea benthos. *Frontiers in Marine Science*, *3*, 92. https://doi.org/10.3389/fmars.2016.00092

Stat, M., Huggett, M. J., Bernasconi, R., DiBattista, J. D., Berry, T. E., Newman, S. J., Harvey, E. S., & Bunce, M. (2017). Ecosystem biomonitoring with eDNA: Metabarcoding across the tree of life in a tropical marine environment. *Scientific Reports*, *7*(1), 12240. https://doi.org/10.1038/s41598-017-12501-5

Stefanni, S., Stanković, D., Borme, D., de Olazabal, A., Juretić, T., Pallavicini, A., & Tirelli, V. (2018). Multi-marker metabarcoding approach to study mesozooplankton at basin scale. *Scientific Reports*, *8*(1), 12085. https://doi.org/10.1038/s41598-018-30157-7

Stewart, K. A. (2019). Understanding the effects of biotic and abiotic factors on sources of aquatic environmental DNA. *Biodiversity and Conservation*, *28*(5), 983–1001. https://doi.org/10.1007/s10531-019-01709-8

Taberlet, P., Coissac, E., Hajibabaei, M., & Rieseberg, L. H. (2012). Environmental DNA. *Molecular Ecology*, *21*(8), 1789–1793. https://doi.org/10.1111/j.1365-294X.2012.05542.x

Taberlet, P., Coissac, E., Pompanon, F., Bronchmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*, 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Thomsen, P. F., Kielgast, J., Iversen, L. L., Wiuf, C., Rasmussen, M., Gilbert, M. T. P., Orlando, L., & Willerslev, E. (2012). Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology*, *21*(11), 2565–2573. https://doi.org/10.1111/j.1365-294X.2011.05418.x

Thomsen, P. F., Møller, P. R., Sigsgaard, E. E., Knudsen, S. W., Jørgensen, O. A., & Willerslev, E. (2016). Environmental DNA from seawater samples correlate with trawl catches of subarctic, deepwater fishes. *PLoS One*, *11*(11), e0165252. https://doi.org/10.1371/journal.pone.0165252

Tsuji, S., Miya, M., Ushio, M., Sato, H., Minamoto, T., & Yamanaka, H. (2019). Evaluating intraspecific genetic diversity using environmental DNA and denoising approach: A case study using tank water. *Environmental DNA*, *2*(1), 42–52. https://doi.org/10.1002/edn3.44

Tsuji, S., & Shibata, N. (2020). Identifying spawning activity in aquatic species based on environmental DNA spikes. *BioRxiv*. https://doi.org/10.1101/2020.01.28.924167

Turon, X., Antich, A., Palacín, C., Præbel, K., & Wangensteen, O. S. (2020). From metabarcoding to metaphylogeography: Separating the wheat from the chaff. *Ecological Applications*, *30*(2), e02036. https://doi.org/10.1002/eap.2036

von Ammon, U., Wood, S. A., Laroche, O., Zaiko, A., Lavery, S. D., Inglis, G. J., & Pochon, X. (2019). Linking environmental DNA and RNA for improved detection of the marine invasive fanworm Sabella spallanzanii. *Frontiers in Marine Science*, *6*, 621. https://doi.org/10.3389/fmars.2019.00621

Wangensteen, O. S., Cebrian, E., Palacín, C., & Turon, X. (2018). Under the canopy: Community-wide effects of invasive algae in Marine Protected Areas revealed by metabarcoding. *Marine Pollution Bulletin*, *127*, 54–66. https://doi.org/10.1016/J.MARPOLBUL.2017.11.033

Wangensteen, O. S., Palacín, C., Guardiola, M., & Turon, X. (2018). DNA metabarcoding of littoral hard-bottom communities: High diversity and database gaps revealed by two molecular markers. *PeerJ*, *6*, e4705. https://doi.org/10.7717/peerj.4705

Wangensteen, O. S., & Turon, X. (2017). Metabarcoding techniques for assessing biodiversity of marine animal forests. In S. Rossi, L. Bramanti, A. Gori, & C. Orejas (Eds.), *Marine animal forests* (pp. 445–473). Springer. https://doi.org/10.1007/978-3-319-21012-4_53

Weigand, A., Bouchez, A., Boets, P., Bruce, K., Ciampor, F., Ekrem, T., Fontaneto, D., Franc, A., Hering, D., Kahlert, M., Keskin, E., Mergen, P., Pawlowski, J., Kueckmann, S., & Leese, F. (2019). Taming the wild west of molecular tools application in aquatic research and biomonitoring. *Biodiversity Information Science and Standards*, *3*, e37215. https://doi.org/10.3897/biss.3.37215

West, K. M., Stat, M., Harvey, E. S., Skepper, C. L., DiBattista, J. D., Richards, Z. T., Travers, M. J., Newman, S. J., & Bunce, M. (2020). eDNA metabarcoding survey reveals fine-scale coral reef community variation across a remote, tropical island ecosystem. *Molecular Ecology*, *29*(6), 1069–1086. https://doi.org/10.1111/mec.15382

Zou, K., Chen, J., Ruan, H., Li, Z., Guo, W., Li, M., & Liu, L. (2020). eDNA metabarcoding as a promising conservation tool for monitoring fish diversity in a coastal wetland of the Pearl River Estuary compared to bottom trawling. *Science of the Total Environment*, *702*, 134704. https://doi.org/10.1016/j.scitotenv.2019.134704

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Antich A, Palacín C, Cebrian E, Golo R, Wangensteen OS, Turon X. Marine biomonitoring with eDNA: Can metabarcoding of water samples cut it as a tool for surveying benthic communities?. *Mol Ecol* 2020;00:1–14. https://doi.org/10.1111/mec.15641

Ecological Applications, 30(2), 2020, e02036

From metabarcoding to metaphylogeography: separating the wheat from the chaff

Xavier Turon,[1,4] Adrià Antich,[1] Creu Palacín,[2] Kim Præbel,[3] and Owen Simon Wangensteen[3]

[1]*Department of Marine Ecology, Centre for Advanced Studies of Blanes (CEAB, CSIC), Blanes, Catalonia, Spain*
[2]*Department of Evolutionary Biology, Ecology and Environmental Sciences, and Institute of Biodiversity Research (IRBio), University of Barcelona, Barcelona, Catalonia, Spain*
[3]*Norwegian College of Fishery Science, UiT the Arctic University of Norway, Tromsø, Norway*

Citation: Turon, X., A. Antich, C. Palacín, K. Præbel, and O. S. Wangensteen. 2020. From metabarcoding to metaphylogeography: separating the wheat from the chaff. Ecological Applications 30(2):e02036. 10.1002/eap.2036

Abstract. Metabarcoding is by now a well-established method for biodiversity assessment in terrestrial, freshwater, and marine environments. Metabarcoding data sets are usually used for α- and β-diversity estimates, that is, interspecies (or inter-MOTU [molecular operational taxonomic unit]) patterns. However, the use of hypervariable metabarcoding markers may provide an enormous amount of intraspecies (intra-MOTU) information—mostly untapped so far. The use of cytochrome oxidase (COI) amplicons is gaining momentum in metabarcoding studies targeting eukaryote richness. COI has been for a long time the marker of choice in population genetics and phylogeographic studies. Therefore, COI metabarcoding data sets may be used to study intraspecies patterns and phylogeographic features for hundreds of species simultaneously, opening a new field that we suggest to name metaphylogeography. The main challenge for the implementation of this approach is the separation of erroneous sequences from true intra-MOTU variation. Here, we develop a cleaning protocol based on changes in entropy of the different codon positions of the COI sequence, together with co-occurrence patterns of sequences. Using a data set of community DNA from several benthic littoral communities in the Mediterranean and Atlantic seas, we first tested by simulation on a subset of sequences a two-step cleaning approach consisting of a denoising step followed by a minimal abundance filtering. The procedure was then applied to the whole data set. We obtained a total of 563 MOTUs that were usable for phylogeographic inference. We used semiquantitative rank data instead of read abundances to perform AMOVAs and haplotype networks. Genetic variability was mainly concentrated within samples, but with an important between seas component as well. There were intergroup differences in the amount of variability between and within communities in each sea. For two species, the results could be compared with traditional Sanger sequence data available for the same zones, giving similar patterns. Our study shows that metabarcoding data can be used to infer intra- and interpopulation genetic variability of many species at a time, providing a new method with great potential for basic biogeography, connectivity and dispersal studies, and for the more applied fields of conservation genetics, invasion genetics, and design of protected areas.

Key words: AMOVA; cytochrome oxidase; connectivity; eukaryotes; haplotype networks; Illumina; metabarcoding; phylogeography; sequencing errors.

Introduction

Metabarcoding, whereby information on species present in a variety of communities can be obtained from so-called environmental DNA (eDNA), or from bulk or community DNA (Creer et al. 2016, Macher et al. 2018), is by now established as a robust method for biodiversity assessment (Baird and Hajibabaei 2012, Deiner et al. 2017, Taberlet et al. 2018, Adamowicz et al. 2019). Metabarcoding provides a fast and accurate method for measuring biodiversity, allowing identification of many more taxa (Molecular Operational Taxonomic Units or MOTUs) than morphological methods (Dafforn et al. 2014, Cowart et al. 2015, Elbrecht et al. 2017), as small and cryptic organisms, early life stages, and fragments or trace DNA left in the environment can be targeted. Further, metabarcoding is largely independent of taxonomic expertise, which is dwindling worldwide (Wheeler et al. 2004), albeit it is highly dependent on the completeness of reference databases to reliably assign taxonomic names to MOTUs (Cowart et al. 2015, Briski et al. 2016). Taxonomic expertise, of course, will always be necessary to construct and expand accurate reference

Manuscript received 13 May 2019; revised 31 July 2019; accepted 3 October 2019. Corresponding Editor: Stefano Mariani.
[4]E-mail: xturon@ceab.csic.es

databases. Biodiversity assessment, detection of invasive or endangered species, paleoecological reconstruction, or diet analyses are among the main applications of metabarcoding to date (e.g., Ji et al. 2013, Pochon et al. 2013, Kelly et al. 2014, Hajibabaei et al. 2016, Ficetola et al. 2018). All of them are highly relevant for basic biodiversity research and for establishing management policies. There is, however, more information in metabarcoding data sets than just α- and β-diversity related issues. Further exploitation requires a shift from interspecies genetic patterns, that constitute most of the metabarcoding applications so far, to intraspecies genetic patterns (reviewed by Adams et al. 2019), making use of the within-MOTU genetic variability uncovered by metabarcoding.

Being heirs to studies in prokaryotes, eukaryotic metabarcoding initially relied heavily on ribosomal RNA sequences for MOTU delimitation (mostly nuclear 18S rDNA sequences). These sequences lack variability for within-MOTU studies in many groups, particularly metazoans (Tang et al. 2012, Leray and Knowlton 2016, Wangensteen et al. 2018*a*). However, in recent years, intense efforts have been devoted to optimize the use of mitochondrial COI sequences in metabarcoding (Andújar et al. 2018). Their use was hindered by the lack of universal primers (Deagle et al. 2014), but new sets of COI primers for general purposes or for specific groups (Leray et al. 2013, Elbrecht and Leese 2017, Vamos et al. 2017, Gunther et al. 2018) are overcoming this problem and COI sequences are now being increasingly used in general biodiversity studies (e.g. Leray and Knowlton 2015, Aylagas et al. 2016, Macher et al. 2018, Porter and Hajibabaei 2018*a*), where they typically uncover a much higher degree of α-diversity than 18S rDNA (Stefanni et al. 2018, Wangensteen et al. 2018*a*, *b*). Furthermore, the use of COI opens the door to taxonomic assignment using the extensive database of the Barcode of Life Datasystems (BOLD), which is continuously increasing in depth and coverage (Ratnasingham and Hebert 2007, Porter and Hajibabaei 2018*b*).

COI sequences have been extensively used in studies of population genetics and phylogeography of terrestrial, freshwater, and marine organisms (Avise 2009, Emerson et al. 2011). The shift to COI-based metabarcoding (Andújar et al. 2018), therefore, implies the generation of databases containing an untapped reservoir of intraspecies variation that can allow characterizing intra- and interpopulation genetic features of many species simultaneously. This could constitute a gigantic leap from the current single-species studies, effectively opening a new field in population genetics for which we suggest the name of metaphylogeography.

The possibility of using metabarcoding for population genetics was hinted at by Bohmann et al. (2014) and Adams et al. (2019), but has been hardly developed. Current instances are in general preliminary, proof of concept, applications, always referred to particular taxa, not to whole community assessments. For instance, within- and between-population genetic structure using bulk DNA has been assessed for ichthyosporean parasites of the cladoceran *Daphnia* (González-Tortuero et al. 2015), for a *Xyleborus* beetle collected at two locations with differing management practices (Pedro et al. 2017) or for coral reef fishes of the genus *Lethrinus* (Stat et al. 2017). In the marine realm, eDNA from water has been used to obtain haplotype and ecotype information for species that are hard to sample, such as whale sharks (Sigsgaard et al. 2016), harbour porpoises (Parsons et al. 2018), or killer whales (Baker et al. 2018). In invasion biology, eDNA was proven useful to assess native vs. nonnative strains of common carp in Japan (Uchii et al. 2016).

An integrated phylogeography encompassing a range of species would be a powerful tool to investigate landscape-level processes (either natural or anthropogenic), over and above the signal given by each species. Studies that combine population genetics data on multiple species by traditional methods are costly and usually involve just a handful of species (e.g., Haye et al. 2014). The alternative is to use meta-analyses to collate the information scattered in different works (e.g., Zink 2002, Pascual et al. 2017), or to use the information contained in georeferenced genetic databases (Gratton et al. 2017). However, the pace at which climate change affect our ecosystems and the projected increased exploration of our resources in the coming decades urge for increased knowledge of population structure and phylogeography at the global biome level. The potential of metaphylogeography ranges from basic questions about biogeography, connectivity, and dispersal patterns to more applied fields of conservation genetics, invasion genetics, and protected areas design. Nowadays, the consideration of multispecies genetic conservation objectives is seen as crucial to preserve community-wide genetic and evolutionary patterns (Vellend et al. 2014, Nielsen et al. 2017).

The main problem for the application of eDNA or community DNA to analyze intraspecies patterns lies in the fact that this technique generates a high number of reads containing sequencing errors, which can occur at different steps in the procedure. Reads obtained by amplification and sequencing can be thought of as a "cloud" of erroneous sequences surrounding the correct one (Edgar and Flyvbjerg 2015). Sequencing errors will typically occur as low-abundance reads with one or few base changes, while errors during amplification (PCR point errors, chimeras) have the potential of generating "daughter clouds" as they can reach higher read abundances (Edgar and Flyvbjerg 2015). As erroneous sequences in general diverge very little from the true sequences, they are often incorporated into the right MOTU during the clustering step, thus reducing potential impacts on the results of "standard" metabarcoding approaches. However, they can severely bias intraspecies genetic patterns by artificially inflating the true haplotype diversity. Thus, separating the "wheat" (true

sequences) from the "chaff" (false sequences) is the main challenge for the application of metabarcoding data to metaphylogeography.

To our knowledge, the problem of the correct assessment of intraspecific genetic diversity from community DNA in complex samples has been explicitly addressed only in a recent work by Elbrecht et al. (2018*a*). Using a single-species mock sample with known Sanger-sequenced haplotypes, they assayed a combination of denoising procedures to reduce the number of spurious haplotypes obtained using a metabarcoding pipeline. They then applied the best performing strategy to natural samples of freshwater invertebrates, deriving population genetic patterns for some of the species present.

We sought here to develop a practical strategy to make metabarcoding data sets amenable to phylogeographic studies. There are an ever-increasing number of such data sets publicly available in repositories. Mining COI-metabarcoding data has been suggested for species discovery (Porter and Hajibabaei 2018*b*), and these databases can be a resource for phylogeography as well. These data comprise different information, from raw sequences to filtered and paired sequences to simply MOTU tables. In many cases, no ground truth data or mock community analyses exist for them. We therefore need a strategy for cleaning noisy databases in the absence of ground truth information. We contend that the properties of coding sequences such as COI can provide such a strategy. Indeed, coding DNA sequences naturally have a high amount of variation concentrated in the third position of the codons, while errors at any step of the metabarcoding pipeline would be randomly distributed across codon positions. Examination of the change of diversity values (measured here as the entropy of each position; Schmidt and Herzel 1997) as we eliminate noisy sequences can therefore guide the choice of the best cleaning parameters in the presence of an unknown amount of noisy data. Entropy values have been used previously to guide sequence trimming (Porter and Zhang 2017) and OTU clustering (Eren et al. 2015), but never before in the context of distinguishing true variation from erroneous sequences.

A parallel inspection of the distribution of sequences across samples is also necessary. Error-containing sequences will typically co-occur in the same sample with the correct sequence, albeit with less abundance, and co-occurrence patterns can be incorporated to detect these sequences in cleaning steps. At the same time, while error sequences are likely to appear randomly in the samples, true sequences should feature a given ecological distribution, meaning that a sequence appearing in all replicates of a community, for instance, is unlikely to be an error. Distribution patterns of sequences have been suggested to guide MOTU calling or MOTU curating procedures (Frøslev et al. 2017, Olesen et al. 2017), but have not been applied, to our knowledge, for within-MOTU sequence curation.

Combining patterns of variation in entropy and sequence distribution patterns can lead to meaningful ways to reduce noisy data sets to operational data sets. This approach can be used to generate customized procedures for each different study system that take into consideration its particulars (replication level, pre-filtering applied, clustering procedure). It only requires that, for a given study, the information about which sequences have been pooled in each MOTU in the clustering step, with their sample distribution, is provided.

We want to point out that the "metaphylogeography" concept is not equivalent to "conventional phylogeography of many species," and we therefore need to adapt some definitions. In particular, relative frequencies of reads of the different haplotypes are available instead of the relative frequencies of individuals bearing these. These are unlikely to be equivalent. The high difference in number of reads that can be obtained in metabarcoding can easily reach orders of magnitude and is hardly representative of conventional frequencies based on the number of individuals bearing a particular haplotype. Further, the quantitative value of metabarcoding data is debatable (Elbrecht and Leese 2015, Wares and Pappalardo 2016, Piñol et al. 2019). Once we have a curated data set, we suggest performing phylogeographic inference using a semiquantitative abundance ranking applied within each MOTU as a compromise between a strictly quantitative interpretation of the data, on one hand, and losing all the information contained in the number of reads on the other. For comparative inference, the traditional analytical framework including haplotype networks, AMOVA, and the like, is perfectly valid if one keeps in mind these differences in the interpretation of results.

In the present study, we developed cleaning strategies to make community data derived from COI amplicon sequencing amenable to the analysis of intraspecific variation. As a case study, we used a COI-based metabarcoding survey of biodiversity of sublittoral marine benthic communities. We then extracted phylogeographic trends from the MOTUs obtained with the best pruning parameters selected. We finally compared results with those of traditional phylogeographic studies for two species for which information exists for the same (or nearby) sampling areas. Our general goal was to show the feasibility of the metaphylogeographic approach using a "standard" metabarcarcoding data set obtained from natural samples.

Material and Methods

Data set

The data set consisted of COI-based biodiversity data obtained from benthic marine communities in two Spanish National Parks, one in the Atlantic and one in the Mediterranean (Appendix S1: Fig. S1). The data set has different replication levels: over time (two years), within

communities (sample replicates), and within samples (size fractions). Sample collection and processing followed Wangensteen and Turon (2017) and Wangensteen et al. (2018*a*). In short, several communities were sampled in 2014 and 2015 by completely scraping off standardized 25 × 25 cm quadrats in hard bottom substrates or by sampling with PVC corers, 24 cm in diameter, in detritic communities. Three replicate samples were collected per community, and each sample was then separated through sieving into three size fractions (>10 mm, 1–10 mm, 63 μm–1 mm, roughly corresponding to mega-, macro-, and meiobenthos; Rex and Ettter 2010). A total of 51 samples separated in 153 fractions were included in the present study (Table 1).

The sampling performed in 2014 included four communities in the Mediterranean Park (Cabrera Archipelago, Balearic Islands) and four in the Atlantic Park (Atlantic Islands of Galicia). These communities were, in each Park, two well-lit communities, one deeper, invertebrate-dominated, community, and a detritic bottom with coralline algae (Table 1). In 2015, the sampling was repeated on the same localities and communities, except for a new community sampled in Cabrera (*Caulerpa cylindracea* community) and the change of one of the two well-lit communities in the Atlantic (*Asparagopsis armata* community instead of *Cystoseira tamariscifolia* community, Table 1). Wangensteen et al. (2018*a*) reported α- and β-diversity results of the sampling performed in 2014, while some of the communities sampled in 2015 were used in a study of the effect of invasive seaweeds (Wangensteen et al. 2018*b*).

Samples were extracted and sequenced using the Leray-XT primer set, a modification of the Leray et al. (2013) primers for a 313 base pair (bp) fragment of COI, with the adequate blanks and negatives, following procedures detailed in Wangensteen et al. (2018*a*). Separate libraries were built with samples from 2014 and 2015 and sequenced in two runs on an Illumina MiSeq platform (2 × 300 bp paired-end) at Fasteris SA (Plan-les-Ouates, Switzerland).

For the present study, we pooled the reads of the two years and analyzed the joint data set with a pipeline based mostly on the OBITools suite (Boyer et al. 2016). The length of the raw reads was trimmed to a median Phred quality score higher than 30, after which paired-reads were assembled using illuminapairedend. The reads with paired-end alignment quality scores higher than 40 were demultiplexed using ngsfilter, which also removed the primer sequences. For this study, we applied a strict length filter keeping only sequences of the expected length (313 bp). Identical sequences were then dereplicated (using obiuniq) and chimeric sequences were detected and removed using the uchime_denovo algorithm implemented in vsearch v1.10.1 (Rognes et al. 2016). At this step, we discarded sequences with just one read in all the data set, as is common practice in metabarcoding studies. We clustered sequences into MOTUs using the SWARM2 method (Mahé et al. 2015), with a *d*-parameter of 13. This parameter was set for the COI fragment used here after comparing the number of MOTUs obtained at different values and checking that this number remained constant for values of *d* in the range of 9–13. The value of $d = 13$ has been previously used in other studies involving the same COI fragment (Macías-Hernández et al. 2018, Kemp et al. 2019, Siegenthaler et al. 2019).

Table 1. Sample characteristics, with indication of locality, type of community, dominant species, depth, coordinates, and number of replicate samples collected in each study year.

National park, community, and dominant species	Depth (m)	Coordinates	No. samples 2014	No. samples 2015
Cabrera Archipelago				
Photophilic algae				
Lophocladia lallemandii	7–10	39.1250° N, 2.9603° E	3	3
Padina pavonica	7–10	39.1250° N, 2.9603° E	3	3
Sciaphilic algae				
Sponges and invertebrates	30	39.1250° N, 2.9603° E	3	3
Caulerpa cylindracea	30	39.1250° N, 2.9603° E	–	3
Detritic bottoms				
Coralline algae	50	39.1249° N, 2.9604° E	3	3
Atlantic Islands				
Photophilic algae				
Cystoseira nodicaulis	3–5	42.2259° N, 8.8969° W	3	3
Cystoseira tamariscifolia	3–5	42.2260° N, 8.8970° W	3	–
Asparagopsis armata	4–6	42,2146° N, 8.8973° W	–	3
Sciaphilic algae				
Saccorhiza polyschides	16	42.1917° N, 8.8885° W	3	3
Detritic bottoms				
Coralline algae	20	42.2123° N, 8.8972° W	3	3

The taxonomic assignment of the MOTU was performed using ecotag (Boyer et al. 2016), which uses a local reference database and a phylogenetic tree-based approach (using the NCBI taxonomy) for assigning sequences without a perfect match. Ecotag searches the best hit in the reference database and builds the set of sequences in the database that are at least as similar to the best hit as the query sequence is. Then, the MOTU is assigned to the most recent common ancestor to all these sequences in the NCBI taxonomy tree. With this procedure, the assigned taxonomic rank varies depending on the similarity of the query sequences and the density of the reference database. We developed a mixed reference database by joining sequences obtained from two sources: in silico ecoPCR against the release 117 of the EMBL nucleotide database and a second set of sequences obtained from the Barcode of Life Datasystems (Ratnasingham and Hebert 2007) using a custom R script to select the Leray fragment. Details of this newly generated database (db_COI_MBPK) are given in Wangensteen et al. (2018*a*). It includes 188,929 reference sequences and is *available online*.[5]

Following the pipeline, we generated an MOTU list and assigned a taxonomical rank to each MOTU. Non-eukaryotic MOTUs were removed. Occasionally, two or more MOTUs received the same species-level assignment, in which case, only the most abundant MOTU was retained and the reads of the others were added to it (this happened in 349 species). We also pooled the sequences of the three fractions of each sample for downstream analyses. For the goal of this study, not all MOTUs carried the phylogeographic information sought (i.e., genetic variation within and between communities and seas). We therefore performed a previous selection in which we included MOTUs that had at least two different sequences (i.e., displayed intra-MOTU structure). We also required that the MOTU appeared in the two Parks with 20 or more reads in each one, and appeared at least once in each of the two study years. We acknowledge that this selection is arbitrary, but these limits were set to ensure that the MOTUs were minimally abundant and widely distributed for reliable phylogeographic inference. Note that this MOTU selection does not imply that discarded MOTUs are artefacts, but simply that they are not useful for population genetics inference (e.g., one MOTU appearing only in a given community, even if abundant).

Using the list of retained MOTUs, the original sequence file, and the information of which sequence belongs to each MOTU (contained in the output of the clustering program used to generate MOTUs), we obtained separate MOTU files containing, for each MOTU, all sequences included with their abundances in the different samples. We then aligned sequences within each MOTU with the msa R package (Bodenhofer et al. 2015), and misaligned sequences, likely due to slippage of degenerate primers (Elbrecht et al. 2018*b*), were detected and eliminated.

[5] http://github.com/metabarpark/Reference-databases

Simulation analysis

All data manipulation and analyses were conducted using R software (R Development Core Team 2008). To avoid confusion between different terms, sometimes used interchangeably, we will use the name denoising to refer to any procedure that tries to infer which sequences contain errors and merges their reads with those of the correct "mother" sequence. We will call filtering any method that actually deletes sequences from the data set, based on abundance thresholds or otherwise. Clustering will refer to any procedure for combining sequences, without regard to whether they are correct or not, into meaningful MOTUs.

We ran a simulation study to infer the best cleaning strategy and the best parameters for our data. The rationale was to start with a known data set, introduce sequencing errors, and clean it again to recover the original data set. We used a custom R script for this simulation. Following Wang et al. (2012), we considered that the 1,000 sequences with highest frequency (in read number) in our data set were error free, and used them for parameter estimation on a data set representative of our actual sequences. For this simulation, we did not keep the ecological information and used just the total number of reads of each of these 1,000 top sequences.

We simulated that these allegedly correct amplicons were sequenced with error rates between 0.001 and 0.01 per base, bracketing values published for HTS sequencers and, in particular, for the MiSeq platform (Schirmer et al. 2016, Pfeiffer et al. 2018). For simplicity, we assumed a constant error rate for all bases in a sequence, albeit we acknowledge that this is a simplification as sequence features such as homopolymer regions make some positions more prone to errors (Taberlet et al. 2018).

For the highest error rate (0.01), we then denoised the resulting sequences using a procedure adapted from the algorithm of Edgar (2016). We merged the reads of presumably incorrect daughter sequences with those of the correct mother sequences if the number of sequence differences (d) is small and the abundance of the incorrect sequence with respect to the correct one (abundance ratio) is low. The higher the number of differences, the lower the ratio should be for the sequences to be merged. This was formalized by the expression (Edgar 2016)

$$\beta(d) = 1/2^{\alpha d+1}$$

where $\beta(d)$ is the maximum abundance ratio allowed between two sequences separated by d changes so that the less abundant was merged with the more abundant. The α parameter is user-settable to seek a compromise between accepting as correct erroneous sequences (high

α values) or merging true sequences (low α values). The denoising was done for values of α from 10 to 1.

We analyzed changes in diversity of the different codon positions as we introduced increasing levels of noise (erroneous reads) and as we denoised the data set with increased stringency (lower α values). As a measure of diversity, we used the Shannon entropy value computed with the R package entropy (Hausser and Strimmer 2009). We expected that random error will increase more the entropy of the less variable position (second position of the codons) and less the entropy of the third, more variable, position. Thus, the entropy ratio (hereafter E_r)

$$E_r = \text{entropy position2}/\text{entropy position3}$$

was expected to increase as simulated error rates increased and to decrease when denoising. After each round of denoising we noted the number of original sequences remaining, the number of noisy sequences remaining, and the entropy ratio of the sequences. We expected that at some value of α the E_r will reach the original value and remain more or less constant afterwards. As at this point many erroneous sequences remained in the data set (see *Results*), we completed the simulation with a filtering procedure in which low frequency sequences were eliminated.

We assayed a range of minimal number of reads to keep a sequence and looked at the number of original and noisy sequences remaining, as well as their entropy ratio. As before, we expected the E_r to decrease markedly and stabilize after some threshold is reached. The best α parameter and the best minimal number of reads should allow us to recover most of the original sequences with as few erroneous sequences as possible.

Data set cleaning

The cleaning procedure followed the findings of the simulation and was therefore based on two steps: denoising (without loss of reads) and filtering by minimal abundance (with loss of reads). We applied denoising within defined MOTUs, under the assumption that most erroneous sequences would have been included in the same MOTU as the correct sequence, and thus sequence distances and abundances, a key part of the denoising algorithm, are more meaningful if compared within MOTUs. Once denoising was complete and, thus, all "salvageable" sequences had been merged with the correct sequence, the second step consisted of an abundance filtering, in which low-abundance sequences, likely erroneous, "surviving" the denoising step were eliminated.

During the previous steps, co-occurrence patterns were used to avoid merging or eliminating sequences whose sample distribution and co-occurrence patterns suggested they were not artifacts (for instance, sequences that do not co-occur with similar sequences will not be merged with them, and sequences found in all replicates of a community will not be filtered out). The use of distribution data can reduce the risk of eliminating true sequences, particularly when they are present at low abundances (e.g., reflecting a low biomass of the organism).

To allow a daughter sequence presumed to be a sequencing error to be merged with a more abundant mother sequence, we required that the former co-occurs with the latter. This is formalized by a co-occurrence (C_{occ}) ratio in the form

$$C_{OCC} = \text{daughter}/(\text{daughter} + \text{mother})$$

were daughter is the number of samples with only the daughter sequence and daughter + mother is the number of samples with the daughter and the mother sequence. The higher the ratio, the less we will merge sequences, as we require a higher co-occurrence with the mother sequence.

We set this parameter to a value of 1 (i.e., whenever a daughter sequence was present, the mother sequence was present in the same sample). Any "daughter" sequence with co-occurrence ratio <1 was considered a genuine sequence and was not merged. This is a conservative value that seeks to avoid merging potentially good sequences. It was set considering that we enforce the presence at the sample level, and not at the fraction level, which means that the sequence needs to be present in just one of the three fractions (10 mm, 1 mm, 63 μm) of the sample. In preliminary assays, changing C_{occ} influenced the number of sequences retained, but represented little change in the entropy ratios obtained. In addition, in the filtering step sequences appearing in all replicates of a given community were considered correct and not filtered out, even if present at low abundance.

Taking these distribution patterns into consideration we applied the denoising and filtering steps. A diagrammatic representation of the pipeline used is presented in Fig. 1. Denoising was performed at α values between 10 and 1, and for the best-performing α, filtering was done for increasing minimal numbers of reads from 2 to 100. After each round of sequence denoising or filtering, the MOTUs were examined and retained only if they still met the requirements of having at least two sequences, appearing in the two Parks with 20 or more reads in each one, and appearing at least once in the two study years. The changes in E_r of the retained MOTUs were examined over the range of α and minimal abundance values. In both cases, the entropy ratio should decrease and, following the simulation results, the points where it became stabilized (we chose as a threshold the point at which the slope fell below 0.005) were used as optimal parameter cutoffs.

Finally, even if sequences retained were mostly correct, they can still include a number of nontarget variants due to heteroplasmy or numts (Elbrecht et al. 2018*a*). However, numts tend to accumulate mutations resulting in stop codons (Song et al. 2008). They can also present

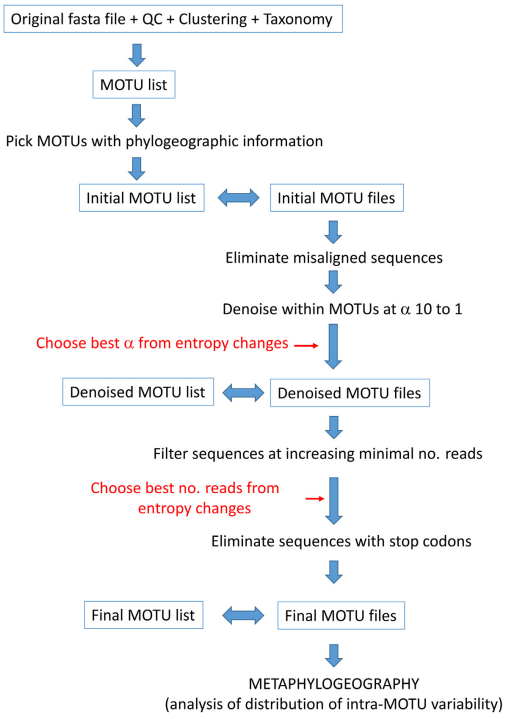


FIG. 1. Schematic representation of the pipeline followed in this study. See *Methods* for details. The red arrows and text indicate the two steps in the pipeline where parameter selection should be carried out based on entropy values. MOTU, molecular operational taxonomic unit.

amino acid substitutions that result in a non-functional protein: Pentinsaari et al. (2016) found 23 amino acids completely conserved across the COI barcode region in Metazoa, corresponding mostly to the helices of the protein that penetrate the mitochondrial membrane. Five of these amino acid positions occur in the fragment sequenced here. Some numts can therefore be detected by inspecting the sequences retained, as has been done in previous metabarcoding studies (Leray et al. 2013). As the data set included many different eukaryotic groups with different genetic codes, we adopted a conservative approach. For each MOTU, we tried the 20 genetic code variants stored in the Biostrings R package (Pagès et al. 2018) and used the *translate* function to obtain the corresponding amino acid sequence. We then chose, for each MOTU, the genetic code giving the lower number of stop codons (often several code variants resulted in no stop codons). In addition, we verified (for the metazoan MOTUs) that the five conserved positions described above did not have any amino acid substitution. The MOTUs denoised with the optimal α-value (first step), once filtered with the optimal abundance cutoff (second step) were checked for the presence of stop codons and amino acid changes, and the sequences presenting them were removed from the data set. The remaining MOTUs and sequences constituted the curated data set for further analyses (Fig. 1).

Metaphylogeographic analyses

We performed network analyses with function HaploNet of the R package pegas (Paradis 2010). We used function amova of the R package ade4 (Dray and Dufour 2007) to compute analyses of molecular variance (AMOVA) in order to ascertain the percent variation associated with the hierarchical organization of the samples. For AMOVA, we used the proportion of the different sequences present (option distances = NULL). Preliminary assays considering also sequence distances (not just sequence frequencies) gave highly similar results and were computationally slower.

In these analyses, we needed to capture the quantitative information regarding frequencies of the different sequence variants. As mentioned above, using number of reads as a proxy for individual-based abundances can be misleading. We adopted a semiquantitative index based on Wangensteen et al. (2018*b*) applied within each MOTU. To obtain this semiquantitative ranking, we ordered the sequences of each sample in each MOTU by increasing number of reads and ranked them from 0 to 4, indicating that the sequence is either absent in that sample (rank 0) or falls in the following percentiles of the distribution of ordered sequences: rank 1, ≤50%; rank 2, >50 ≤ 75%; rank 3, >75 ≤ 90%; rank 4, >90%. These semiquantitative ranks were used as proxies for haplotype abundances in the analyses.

Comparison with previous studies

After examination of the curated MOTU data set, we found only two species for which conventional phylogeographic analyses had been performed using COI information in the same geographic area: the sea urchin *Paracentrotus lividus* and the brittle star *Ophiothrix fragilis*.

For *Paracentrotus lividus*, we collated haplotype information from studies spanning the Atlanto-Mediterranean transition (Duran et al. 2004), trimmed the sequences to the same fragment amplified in our study, and compared the haplotypes with the ones encountered in our metabarcoding data set. Duran et al. (2004) included two populations close to our localities: Eivissa Island in the Balearic Archipelago, and Ferrol in the Galician coast. Networks were generated with the haplotypes found in these localities and compared with our results.

For *Ophiothrix fragilis*, our MOTU corresponded to Lineage II of Pérez-Portela et al. (2013). This brittle star is in fact a complex of species, and Lineage II is likely a cryptic species (Taboada and Pérez-Portela 2016), but it remains unnamed so far. As before, we extracted haplotype information from all localities in Pérez-Portela

et al. (2013), spanning the Atlanto-Mediterranean area, and compared with our results. We also obtained haplotype networks for the two closest populations studied in that work: Alcudia in the Balearic Archipelago and Ferrol in the Galician coast.

Results

The data set

The original data set, once quality and length filtered, contained 25,772,264 sequences of 8,900,080 unique sequences. Without singletons, the numbers were reduced to 17,808,524 reads and 936,340 unique sequences. Following the pipeline, we obtained a MOTU list of 26,561 eukaryote MOTUs. Of these, 13,410 MOTUs were present only in the Mediterranean site, 8,247 only in the Atlantic locality, and 4,904 were shared by both basins. Of the latter, only 722 MOTUs (with a total of 362,177 unique sequences and 9,430,236 reads) fulfilled the conditions that we set for the metaphylogeographic analyses (see *Methods*) of having at least two sequences, being present in the two Parks with at least 20 reads in each one, and having appeared in the two years of study. After checking the alignment, only 158 sequences, comprising 689 reads, appeared as misaligned, mostly as a result of 1 bp slippage, and were removed. The singleton-free fasta sequence file (paired, demultiplexed, and quality-filtered), the original MOTU list, and the output of the SWARM analyses have been uploaded as a Mendeley data set (see *Data Availability*). The 722 MOTUs selected for the study are listed in Data S1, together with their taxonomic assignment and abundance (number of reads) per sample. The actual sequences of each MOTU, with their abundances per sample, are available at the Mendeley data set.

Simulation study

In our case, the top 1,000 sequences in the 722 MOTUs data set contained 5,948,135 reads. The entropy values of the codon positions of these sequences were: first position, 0.4298 ± 0.037 bits (mean $\pm$ SE); second position, 0.1833 ± 0.028 bits; third position, 0.9256 ± 0.023 bits. The simulation of increasing sequencing error rates clearly increased the entropy of the three positions (Fig. 2A), but more so for the less variable second position, which increased its value ~30% at the highest error rate. On the other hand, the third position increased entropy only about 1.8%. As a result, the entropy ratio (E_r, entropy2/entropy3) increased linearly with error rate, from 0.198 to 0.252 (Fig. 2B).

We then used the "noisy-most" data set, the one simulated at the highest (0.01%) error rate. It had the same original number of reads, but 5,141,683 erroneous sequences (besides the 1,000 correct ones) were generated. For coherence with the global data set used, singletons were removed, leaving 144,791 sequences. This data set was then denoised at α values between 10 (least stringent) and 1 (most stringent). The E_r decreased drastically at the initial steps, concomitantly with a decrease in the number of erroneous sequences (Fig. 3A). The E_r value of the simulated data set reached the original value at α between 6 and 5. Taking the more conservative $\alpha = 5$, which is also the point where the entropy curve levelled off (slope < 0.005), we found that the data set contained 895 of the original sequences and 17,799 erroneous sequences. In other words, while ~10% of the original sequences have been incorrectly merged, there remained still a high number of errors in the data set. Using only the denoise procedure, we got completely rid of erroneous sequences only at $\alpha = 1$. But at this value only 66% of the correct sequences were retained.

We therefore applied a round of filtering by minimal number of reads to the data set denoised at $\alpha = 5$. Again, the E_r decreased sharply at increasing thresholds of minimal reads, following the elimination of erroneous sequences (Fig. 3B), and stabilized clearly at seven reads (Fig. 3B). The combination of denoising ($\alpha = 5$) and filtering (minimal abundance = 7) allowed us to recover 924 sequences, of which 895 (97%) were among the 1,000 original sequences and only 3% were erroneous sequences. The frequency distribution of the number of reads in both the original (1,000) and the recovered (924) sequences was almost identical (not shown). Importantly, the shape of the E_r curve, specifically the stabilization points, proved informative to select the cut-points for the two variables.

Data set cleaning

As a first step, we tried to identify PCR errors during amplification, as they can result in abundant sequences and be more difficult to spot. We assumed that PCR errors will affect one nucleotide at most, will occur in few samples, where they will coexist with the original sequence, and will be abundant. Therefore, we looked within the 722 MOTUs for sequences differing by one nucleotide from a more abundant one, co-occurring always with it, being present in at most three samples (out of 51 samples), and having an abundance of >200 reads (set as a threshold to identify relatively abundant sequences). Only 14 such sequences were identified and merged with the more abundant ones.

After applying the denoising step for α values from 10 to 1 and a co-occurrence index of 1 to the whole data set of 722 MOTUs, we examined the change in number of retained MOTUs and entropy ratio (Fig. 4A). The number of MOTUs remained constant but started decreasing at $\alpha = 6$. As expected, the E_r decreased fast at first and more slowly at lower α-values (i.e., with higher merging power) (Fig. 4A). The curve leveled off (slope below 0.005) at $\alpha = 5$, with only a slight loss of MOTUs (six out of 722). We thus retained $\alpha = 5$ as the optimal denoising parameter.

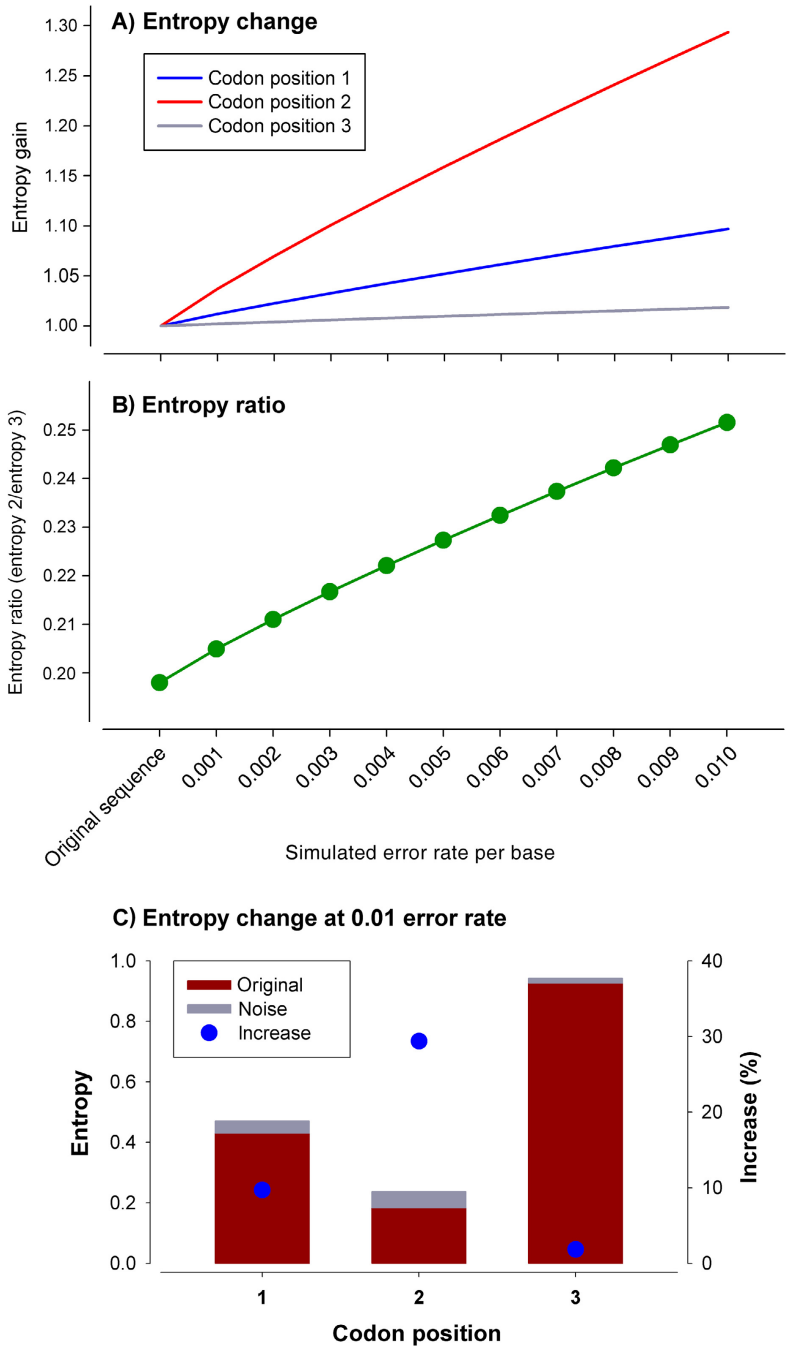


FIG. 2. Simulation analysis. (A) Relative increase (initial value = 1) of the entropy values of each position at increased error rates. Bar plot shows the original and added entropy of each position at the highest (0.01) error rate. (B) Change in the entropy ratio. (C) Bar plot showing the original and added entropy of each position at the highest (0.01) error rate.

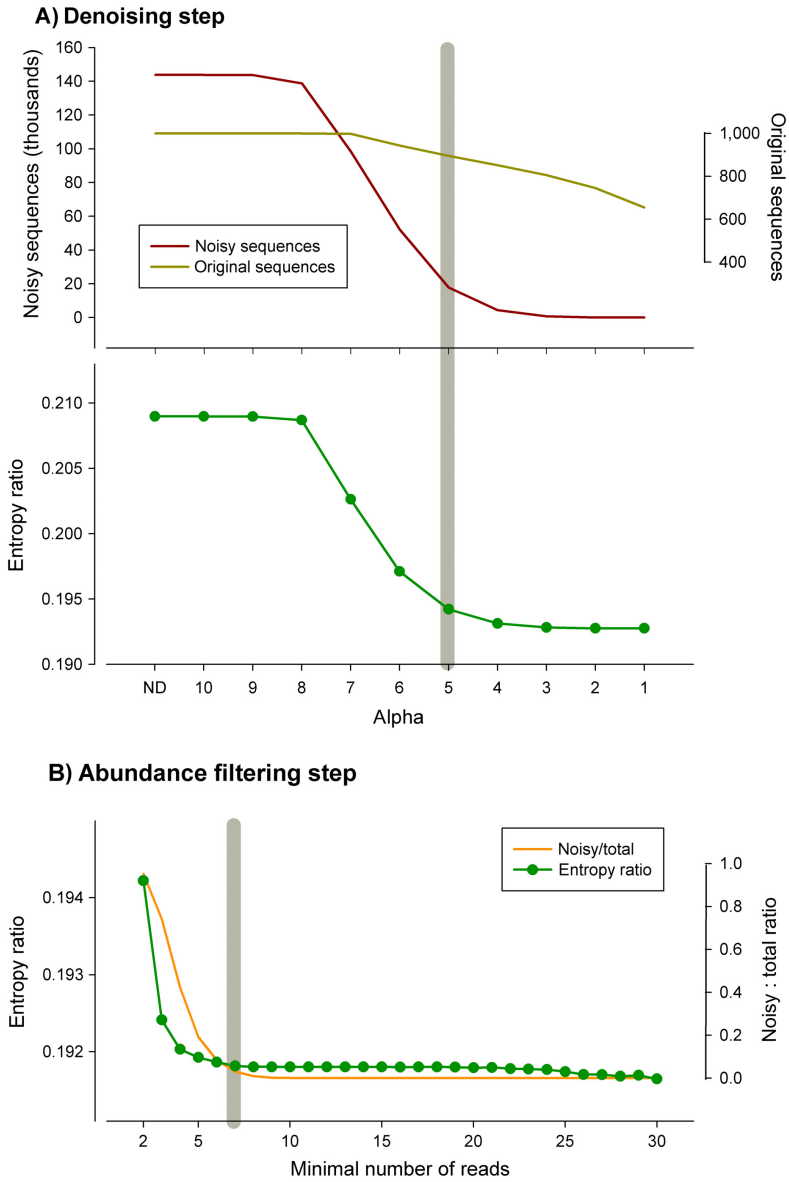


FIG. 3. Simulation analysis. (A) Variation in the number of original and erroneous ("noisy") sequences and entropy ratio at decreasing values of the alpha parameter of the denoising algorithm (ND, no denoising). (B) Change in the entropy ratio and in proportion of noisy vs. original sequences after filtering the data set by minimal abundance. The gray bars indicate the selected values of alpha (5) and minimal number of reads (7).

The MOTU list corresponding to the denoised data set had 716 MOTUs, with 49,995 sequences (86% of the original sequences had been merged) and 9,426,339 reads (Data S1). The corresponding MOTU files (available at the Mendeley data set; see *Data Availability*) were submitted to an abundance filter, with a threshold from 2 to 100 reads. There was a decrease the number of MOTUs retained at increasing minimal numbers of reads, particularly in the interval 2–50 (Fig. 4B). The entropy ratio fell markedly and became stabilized at a

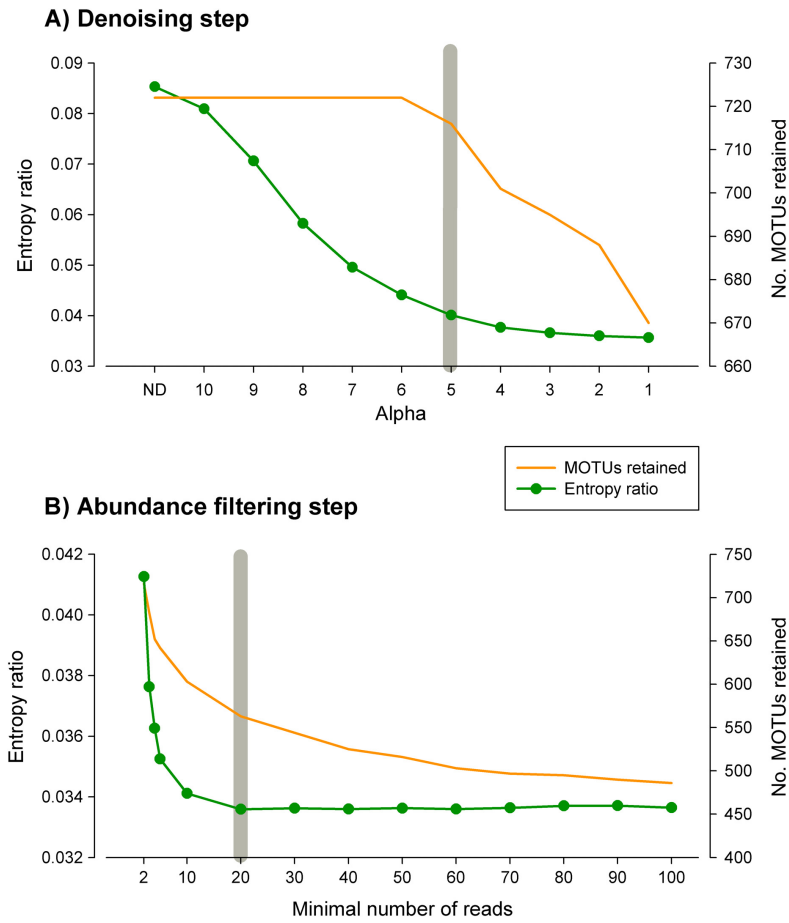


FIG. 4. Final analyses of the littoral communities data set. (A) Variation in the number of sequences and number of MOTUs remaining at decreasing values of the alpha parameter (ND, no denoising) of the denoising algorithm. (B) Change in the entropy ratio and (C) change in residual (within-sample) variance of the amova model. The gray bars indicate the selected alpha value (5) and abundance threshold (20).

value of 20 reads, after which it remained more or less constant (Fig. 4B). Thus, 20 reads was used as a minimal abundance threshold.

The sequences of the resulting MOTU files were translated and checked. Only eight sequences had stop codons, while a further 52 metazoan sequences had amino acid changes in the five positions invariable in Metazoa. These 60 sequences were eliminated, and the final MOTU list thus consisted of 563 MOTUs, with 7,146 sequences and 8,910,913 reads (Data S2). The final MOTU files were uploaded to the Mendeley data set (see *Data Availability*).

As for the taxonomy assigned, the most diverse groups of Eukarya in the final data set were Rhodophyta (91 MOTUs), Stramenopiles (90 MOTUs, mostly diatoms and brown algae), and Metazoa (273 MOTUs) (Data S2). A total of 99 eukaryotic MOTUs remained unassigned taxonomically (identified as Eukarya). Among metazoans, 112 MOTUs were assigned a species-level taxon, while 225 MOTUs were assigned at least at the phylum level and 48 MOTUs remained unassigned (Data S2). The phyla of metazoans identified in the final MOTU list were Annelida (34 MOTUs), Arthropoda (56 MOTUs), Bryozoa (17 MOTUs), Chordata (eight MOTUs), Echinodermata (seven MOTUs), Mollusca (22 MOTUs), Nemertea (six MOTUs), Porifera (30 MOTUs), and Xenacoelomorpha (one MOTU).

Further analyses concentrated in the major groups detected, which accounted for 437 of the 464 MOTUs that could be assigned: red algae (Rhodophyta), diatoms (Bacillariophyta), brown algae (Phaeophyceae), and metazoans (Metazoa). In the latter, phylum-level analyses were performed.

Phylogeography

Network graphs of the MOTUs (Appendix S2) showed different patterns, albeit in most cases one or a few haplotypes appeared as the most abundant, linked to a varying number of low abundance haplotypes. Some selected instances are presented in Fig. 5, showing also the change in network shape along the process of cleaning. It can be seen that the major pruning effect was due to the initial denoising step.

AMOVAs were used to partition the genetic variance hierarchically into components due to the differences between seas, between communities within seas, between samples (replicates) within communities, and within samples. The average values of these variance components for the major groups detected, and for metazoan phyla separately, displayed a clear overall trend: genetic variance was concentrated within samples (60–75%) in all major groups (Fig. 6A). The other components of variance followed a decreasing trend, with a remarkable variance associated to differentiation between the two seas (14–25% of variance), and smaller variance between communities within each sea, and even lower between replicate samples of a given community. The latter component was almost negligible (<1.2%) in the non-metazoan groups considered, but reached 5.4% in metazoans. The different components were compared across groups with ANOVA (followed by Student-Newmann-Keuls post hoc tests if significant). The between sample component was significantly higher (all $P < 0.001$) in metazoans than in the other groups. For the other components, the values were in general comparable, the only significant differences being a higher between seas differentiation in diatoms than in metazoans, and a higher within sample variance in red algae than in diatoms.

Metazoans therefore showed a higher heterogeneity between replicate samples of a given community than the other groups. When examined across phyla (Fig. 6B), albeit the overall trend was in general maintained, a dominant within sample component and a variance between seas > between communities > between samples, there were exceptions. In particular, molluscs had a high between sample variability, and other groups presented important small-scale (between communities and/or between samples) variability as compared to the between seas differentiation (Cnidaria, Nemertea, Porifera). ANOVA showed few significant differences between phyla, the only significant comparisons involving the between samples component in molluscs, which was significantly higher than in bryozoans or sponges.

As for the comparison with previous studies, MOTU 697 was identified as the sea urchin *Paracentrotus lividus* with 100% sequence identity. This MOTU had 15 sequences. This species has an Atlanto-Mediterranean

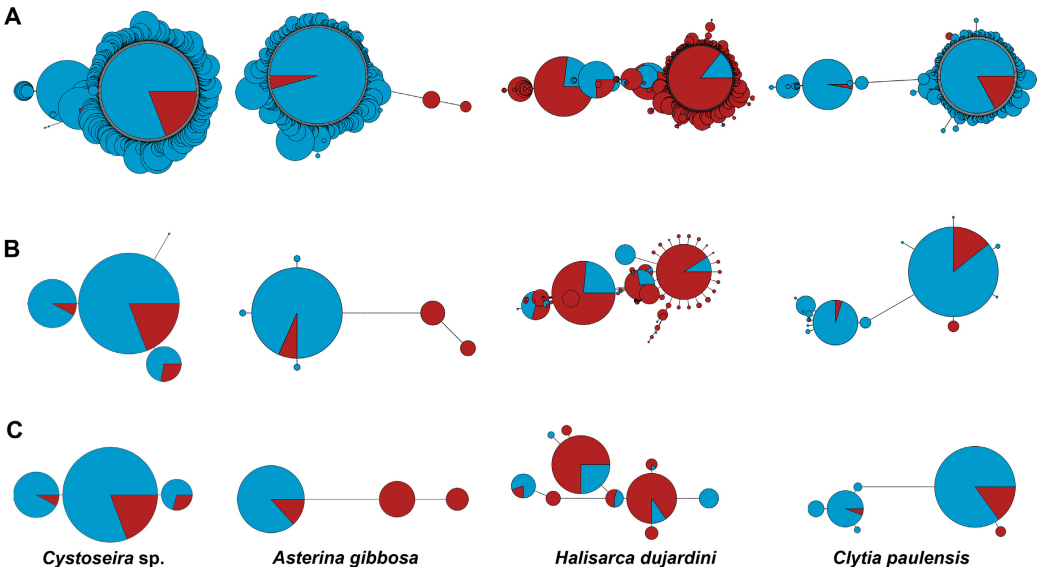


FIG. 5. Selected instances of networks obtained at different stages of the pipeline: (A) without filters; (B) after denoising at alpha = 5; (C) after denoising at alpha = 5 plus minimal abundance filtering (threshold 20 reads). Circles represent haplotypes, and their diameters are proportional to their abundance (in semiquantitative ranks) in the samples. Blue color represent abundance in Mediterranean samples, red color in Atlantic samples. Length of links is proportional to the number of mutational steps between haplotypes. Note that circles in panels A, B, and C are not drawn to the same scale. The names correspond to the taxonomical identification of the MOTUs with ecotag (OBITools package). The MOTU ids (as per Data S1) are, from left to right, 143, 1740, 2500, and 25366.

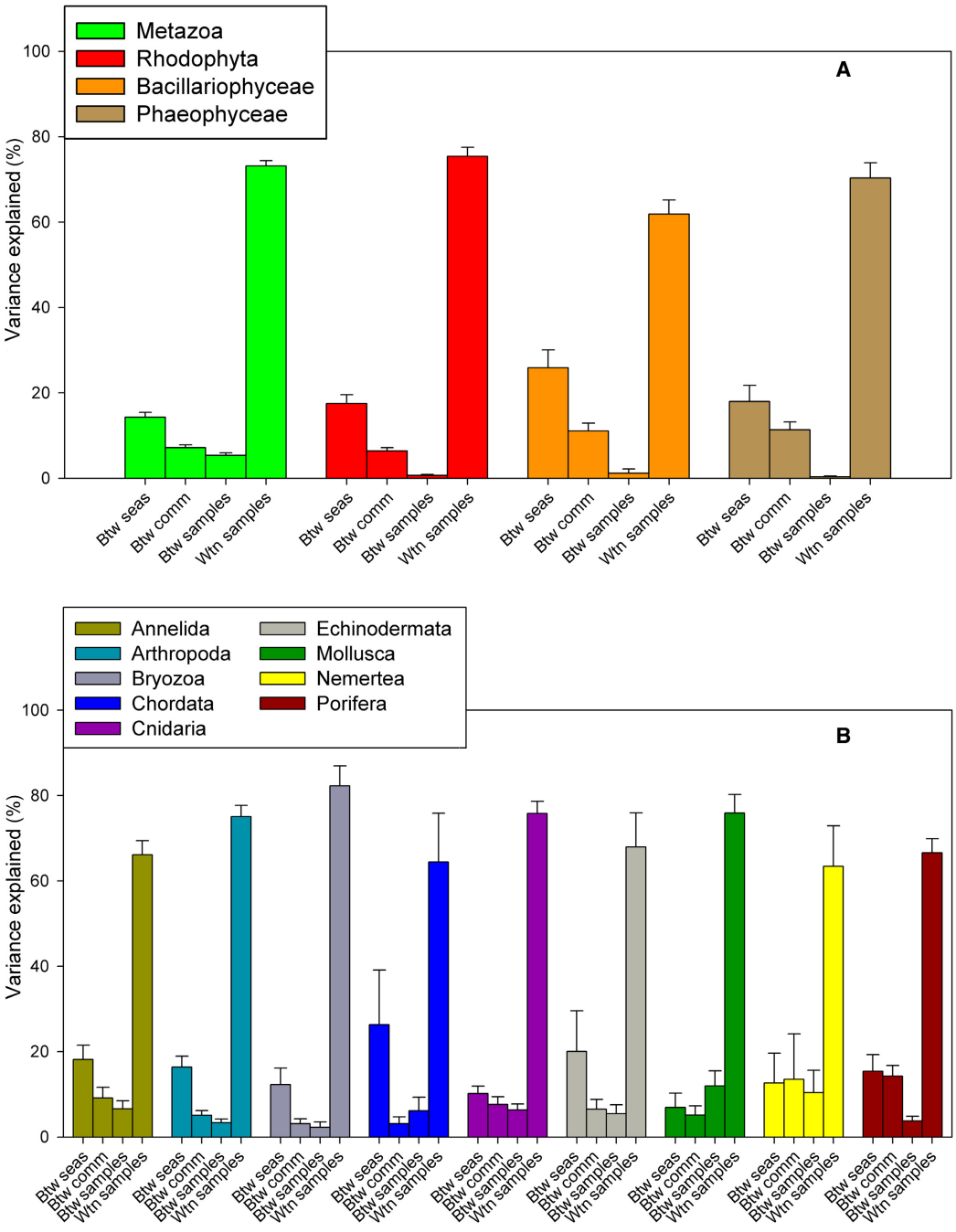


FIG. 6. Summary of the mean percentage of variance explained by the hierarchical structure of the AMOVA: (A) as per eukaryote groups; (B) per metazoan phyla. Error bars are standard errors. Btw seas, between seas; btw comm, between communities within seas; btw samples, between samples within communities; wtn samples, within samples.

distribution and Duran et al. (2004) analyzed populations spanning the western Mediterranean and northeast Atlantic with COI. In that work, 65 different haplotypes (of a longer fragment of COI) were detected. Once trimmed to our sequence length and collapsed, there were 32 remaining haplotypes. Nine out of the 15 sequences detected in our study had already been found by Duran and co-workers, while the remaining six were new.

We then selected the haplotypes found in the previous work in the two localities closest to our sampling points (Eivissa in Balearic Islands and Ferrol in Galicia). There were 11 haplotypes (four of which were also present in our MOTU). We performed a network with the 2004 information and compared it with the one obtained for MOTU 697 with our semiquantitative abundance rank (Fig. 7A, B). The two networks had a similar shape, with a highest abundance of haplotype 2 (named after the order of abundance of sequences obtained for this MOTU), followed by haplotypes 1, 3, and 6. For the shared haplotypes, the between seas distribution was the same in the two studies (1, 2, and 3 shared between seas, six present only in the Atlantic). An AMOVA with a randomization test (n = 1,000) of our MOTU 697 revealed a significant differentiation between seas and between and within samples ($P < 0.001$) but not between communities ($P = 0.812$).

The MOTU 15396, comprising 37 sequences, was identified (100% identity) with *Ophiothrix* sp. in Pérez-Portela et al. (2013). In that work, the authors studied a controversial species complex of the genus *Ophiothrix* in the European waters using 16S and COI. Our sequences corresponded to the Lineage II of *Ophiothrix fragilis* in that work, that spanned from Britanny to Turkey. Pérez-Portela et al. (2013) reported 125 haplotypes of Lineage II that, once trimmed to our 313 bp length, resulted in 90 different haplotypes. When merged with our data set, nine out of 37 sequences in MOTU 15396 had already been found in the previous study, while another 28 were new.

As before, we selected in Pérez-Portela et al. (2013) the two localities closest to our sampling points (Alcudia in Balearic Islands, and Ferrol in Galicia). There were 29 haplotypes in these localities, of which five were shared with our study. The corresponding networks (Fig. 7C, D) showed a star-shaped structure with a dominant haplotype 1 found in the two studies, with many low abundance sequences separated by one or a few mutations from the central haplotype and some longer branches. It is noteworthy that, in this case, the shared haplotypes do not have always the same inter-basin distribution, thus, haplotype 1 was present in both oceans, but haplotypes 3, 8, and 5 present only in the Mediterranean site in the previous work, appeared now in the two seas (it should be noted that haplotype 3 did appear in other Atlantic sites in Pérez-Portela et al. 2013). Finally, haplotype 20 was present only in the Mediterranean site in Pérez-Portela et al. (2013) and only in the Atlantic locality in the present work. An AMOVA with a randomization test (n = 1,000) of our MOTU 15396 showed a significant component of variation related to between and within samples genetic variability ($P < 0.001$), but not between seas ($P = 0.729$) or between communities within seas ($P = 0.212$).

Discussion

In this study, we have developed a method to apply metabarcoding data sets to the study of intraspecies patterns of many species at a time using a highly variable coding fragment (COI). An initial denoising step, aimed at merging erroneous sequences with the correct ones, was followed by an abundance filtering step aimed at removing the remaining erroneous sequences. We used information from the variability of the different codon positions, following a simulation study, to select the best parameter values in the denoising and filtering steps. In addition, sample distribution information was used in the different steps to minimize loss of low abundance true sequences.

All cleaning procedures are a compromise between eliminating spurious sequences and losing true signal. In the benchmarking approach of Elbrecht et al. (2018*a*), 943 erroneous haplotypes appeared in a sample known to have only 15 before any processing. After a denoising process, 15 haplotypes remained but, of these, 6 (40%) were still sequences not present in the original sample, while 6 of the 15 original variants were discarded during the process. Clearly, separating wheat from chaff is a challenging problem.

In this study, we suggest an operational approach based on the stabilization of the entropy ratio to guide the cleaning procedures. Both the simulation approach and the analysis of the real data set pointed to an α-value of 5 in the denoising step, which was also the optimal value selected in Elbrecht et al. (2018*a*). Whether this value can be taken as a general rule of thumb or not will require analyses of more data sets. For the filtering step, our method indicated 20 reads as the optimal threshold. This is a parameter that will likely vary between studies and should be optimized for each particular data set.

Some authors proposed that denoising should be performed before clustering to identify genuine sequence variants, using different procedures, such as the UNOISE2 algorithm that we have adapted here (Edgar 2016), the MED (minimum entropy decomposition; Eren et al. 2015) procedure, or the DADA2 algorithm (divisive amplicon denoising algorithm; Callahan et al. 2016). It has also been suggested that sequence variants should replace MOTUs to capture relevant biological variation (Edgar 2016, Callahan et al. 2017). This suggestion may be adequate in prokaryotes, where strains of the same species can have different characteristics (e.g., pathogenicity). However, for eukaryotes, and particularly metazoans, given the high amount of intraspecies

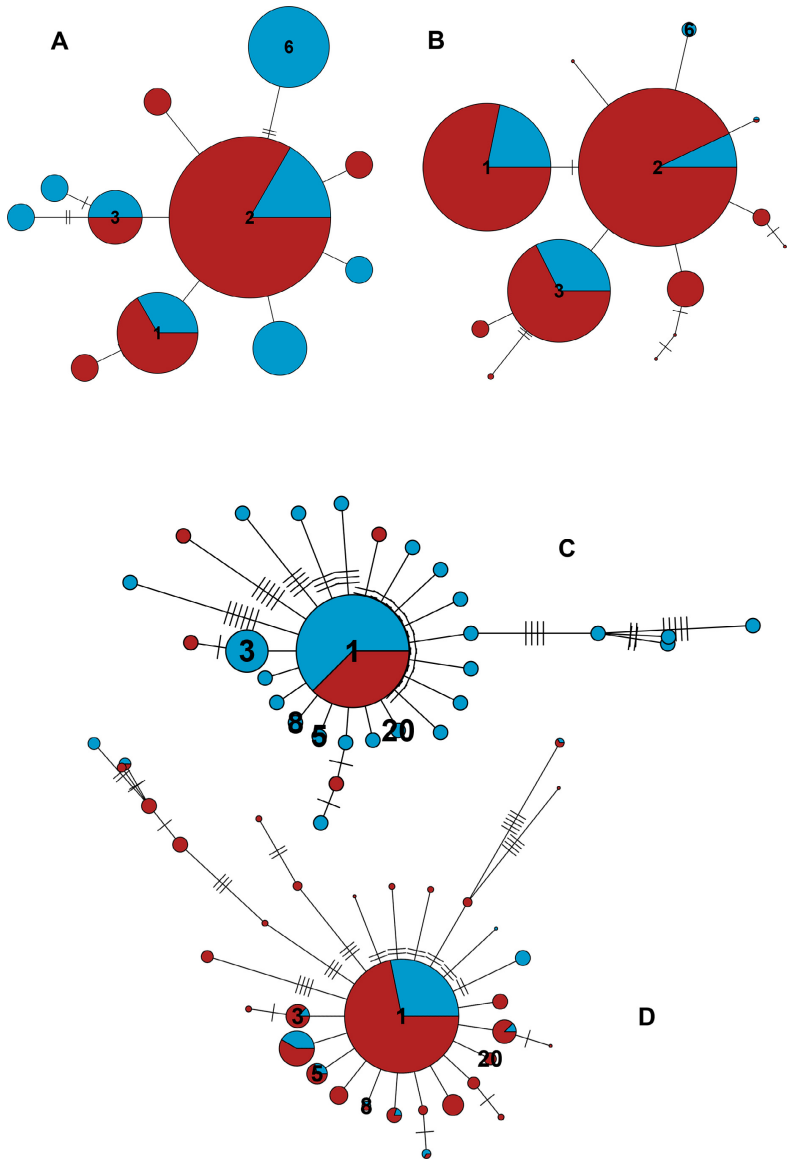


FIG. 7. (A) Network constructed with the 11 haplotypes of the sea urchin *Paracentrotus lividus* found by Duran et al. (2004) in localities close to our sampling points and (B) network constructed with the 13 haplotypes comprising the MOTU corresponding to this species (id 697). Haplotypes common to both studies are numbered. (C) Network with the 29 haplotypes of the brittle star *Ophiothrix fragilis* identified by Pérez-Portela et al. (2013) in localities close to our sampling points. (D) Network of the 34 haplotypes found in the present study in the MOTU corresponding to this species (id 15396). Haplotypes common to both studies are numbered. The short slashes in the links between haplotypes represent mutational steps. Colors as in Fig. 5.

information contained in the data sets, we think that it is more advisable to define meaningful MOTUs and perform denoising procedures within them, in order to obtain a "clean" data set and be able to use the intra-MOTU sequence variability to make phylogeographic and population genetics inference. Clearly, our procedure is applicable only to coding sequences, which excludes much work done on protists based on ribosomal DNA. However, the growing number of metabarcoding studies using COI sequence data, together with

the steady development of the BOLD database, makes us confident that many metabarcoding data sets of enormous potential for metaphylogeographic inference will become available in the near future.

We found a couple of instances of previous studies that have analysed COI structure in species recovered in our MOTU data set and in nearby localities. For *Paracentrotus lividus*, there were phylogeographic studies of the Atlanto-Mediterranean area using COI (Duran et al. 2004), 16S (Calderón et al. 2008), and the nuclear ANT intron (Calderón et al. 2008). In all cases, a low, but significant, signal corresponding to the separation between Atlantic and Mediterranean was found. Our COI results were in agreement with those of Duran et al. (2004) for the localities that could be compared. We detected a somewhat higher number of haplotypes (11 in the previous work, 15 in our study) and the most common haplotypes were shared. The shape of the network was also similar. We want to emphasize that, as far as we could detect, not a single sea urchin of this species was present in our samples, so we obtained a similar level of haplotype diversity with community DNA than in a study specifically devoted to collect sea urchin specimens. For *Ophiothrix fragilis*, we also found a higher haplotype diversity (37 haplotypes) than in comparable localities in the work of Pérez-Portela et al. (2013; 29 haplotypes). We identified five haplotypes that were shared in the two studies, including the commonest one in both data sets, and the networks again had similar structure. Of note here is that we could expand the distribution range of some of the haplotypes. Our AMOVA results for these two instances were equivalent to previous results for the only component that was analyzed in both studies (the between-seas differentiation). Thus, Duran et al. (2004) found a significant ($P < 0.05$) between-basin differentiation in *Paracentrotus lividus*, while Pérez-Portela et al. (2013) did not find any significant genetic variability between Atlantic and Mediterranean for Lineage II of *Ophiothrix fragilis* ($P = 0.790$). This is consistent with our metabarcoding-derived AMOVAs ($P < 0.001$ and $P = 0.729$, respectively). The two species are of remarkable ecological importance, *Paracentrotus lividus* is an engineer species able to modify the littoral landscape through its browsing activity (Palacin et al. 1998, Wangensteen et al. 2011), and is also a commercially exploited species (Barnes and Crook 2001). The different lineages of *Ophiothrix fragilis* are highly abundant components of the littoral communities and can form dense beds, with an important role in clearing particulate matter with their filtering activities (Davoult 1989, Davoult and Gounin 1995). For both species, therefore, an accurate assessment of the genetic relationships across the different basins is of utmost importance for conservation and management purposes.

We have used an already collected data set, which can mimic the situation that many a posteriori studies can encounter. However, future metabarcoding studies can be planned taking into consideration the potential application for intraspecies analyses as well. For instance, PCR replicates for each sample can be of tremendous advantage to eliminate noise in the first steps. Increasing ecological replication can also be of great value for metaphylogeographic studies. We strongly advocate that published metabarcoding studies include in their data sets the information about which sequences are grouped into each MOTU with their sample distribution. This information is not commonly provided, and is necessary to make these studies amenable for intraspecies and metaphylogeographic analyses.

Metabarcoding now occupies a well-deserved prominent place among the methods for assessing community-level diversity (Kelly et al. 2014, Adamowicz et al. 2019). We have shown that it can be also an important source for species-level genetic diversity information for a wide assemblage of taxonomic groups. The mining of metabarcoding data for intraspecies information opens up a vast field with both basic and applied implications (Adams et al. 2019). Among the latter, the possibility of effectively basing conservation efforts on multispecies genetic metrics to preserve community-level evolutionary patterns (Nielsen et al. 2017). It will also open the phylogeography field, nowadays restricted almost exclusively to macroorganisms, to the myriad of meio- and micro-eukaryotes that make up most of the diversity present in natural communities.

Another related field is the assessment of connectivity between populations. This is important for endangered species, invasive species, protected areas design, and management in general. For instance, in the marine environment, differences in larval dispersal have often been suggested as responsible for determining population genetic structure, but other factors, such as variation in divergence times and changes in effective population sizes, must be taken into account (Hart and Marko 2010). A powerful test for these contrasting assumptions is to compare phylogeographic patterns among species that concur or differ in larval type. Metaphylogeography can provide such comparative data. For instance, in our study we have found that metazoans in general have more between-replicate variability than other groups, and within metazoans the between community and between-replicate components of genetic variation can be significantly different between phyla.

In conclusion, our study shows the feasibility of mining metabarcoding data sets for the analysis of intraspecies genetic diversity using objective parameters for denoising and filtering spurious sequences. We cannot at present advice a set pipeline to do this, as procedures should be customized for the particulars (e.g., replication level, number of habitats, number of localities) of each study data set. With this article, we hope to stir further discussion and developments in this field. The metaphylogeography application should be borne in mind to guide the planning and reporting of metabarcoding studies to ease the recovery of this, so far unexplored, vast amount of information.

Acknowledgments

We are indebted to the staff of the Atlantic Islands of Galicia and Cabrera Archipelago National Parks for sampling permits and invaluable logistic help. Thanks also to Xavier Roijals for his skillful bioinformatic assistance in the CEAB computing cluster facilities. This work has been funded by projects Metabarpark (Spanish National Parks Autonomous Agency, OAPN 1036/2013) and PopCOmics (CTM2017-88080, AEI/FEDER, UE) of the Spanish Government.

Literature Cited

Adamowicz, S. J., et al. 2019. Trends in DNA barcoding and metabarcoding. Genome 62:5–8.

Adams, C. I. M., M. Knapp, N. J. Gemmell, G. J. Jeunen, M. Bunce, M. Lamare, and H. R. Taylor. 2019. Beyond diversity: can environmental DNA (eDNA) cur it as a population genetic tool? Genes 10:192.

Andújar, C., P. Arribas, D. W. Yu, A. P. Vogler, and B. C. Emerson. 2018. Why the COI barcode should be the community DNA metabarcode for the Metazoa. Molecular Ecology 27:3968–3975.

Avise, J. C. 2009. Phylogeography: retrospect and prospect. Journal of Biogeography 36:3–15.

Aylagas, E., A. Borja, X. Irigoien, and N. Rodríguez-Ezpeleta. 2016. Benchmarking DNA metabarcoding for biodiversity-based monitoring and assessment. Frontiers in Marine Science 3:1–12.

Baird, D. J., and M. Hajibabaei. 2012. Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. Molecular Ecology 21:2039–2044.

Baker, C. S., D. Steel, S. Nieukirk, and H. Klinck. 2018. Environmental DNA (eDNA) from the wake of the whales: droplet digital PCR for detection and species identification. Frontiers in Marine Science 5:133.

Barnes, D. K. A., and A. C. Crook. 2001. Implications of temporal and spatial variability in Paracentrotus lividus populations to the associated commercial coastal fishery. Hydrobiologia 465:95–102.

Bodenhofer, U., E. Bonatesta, C. Horejs-Kainrath, and S. Hochreiter. 2015. msa: an Rp ackage for multiple sequence alignment. Bioinformatics 31:3997–3999.

Bohmann, K., A. Evans, M. T. P. Gilbert, G. R. Carvalho, S. Creer, M. Knapp, D. W. Yu, and M. de Bruyn. 2014. Environmental DNA for wildlife biology and biodiversity monitoring. Trends in Ecology and Evolution 29:358–367.

Boyer, F., C. Mercier, A. Bonin, Y. Le Bras, P. Taberlet, and E. Coissac. 2016. OBITOOLS: a unix-inspired software package for DNA metabarcoding. Molecular Ecology Resources 16:176–182.

Briski, E., S. Ghabooli, S. A. Bailey, and H. J. MacIsaac. 2016. Are genetic databases sufficiently populated to detect non-indigenous species? Biological Invasions 18:1911–1922.

Calderón, I., G. Giribet, and X. Turon. 2008. Two markers and one history: phylogeography of the edible common sea urchin *Paracentrotus lividus* in the Lusitanian region. Marine Biology 154:137–151.

Callahan, B. J., P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes. 2016. DADA2: High resolution sample inference from Illumina amplicon data. Nature Methods 13:581–583.

Callahan, B. J., P. J. McMurdie, and S. P. Holmes. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME Journal 11:2639–2643.

Cowart, D. A., M. Pinheiro, O. Mouchel, M. Maguer, J. Grall, J. Miné, and S. Arnaud-Haond. 2015. Metabarcoding is powerful yet still blind: a comparative analysis of morphological and molecular surveys of seagrass communities. PLoS ONE 10:e0117562.

Creer, S., K. Deiner, S. Frey, D. Porazinska, P. Taberlet, K. Thomas, C. Potter, and H. Bik. 2016. The ecologist's field guide to sequence-based identification of biodiversity. Methods in Ecology and Evolution 7:1008–1018.

Dafforn, K. A., D. J. Baird, A. A. Chariton, M. Y. Sun, M. V. Brown, S. L. Simpson, B. P. Kelaher, and E. M. Johnston. 2014. Faster, higher and stronger? the pros and cons of molecular faunal data for assessing ecosystem condition. Advances in Ecological Research 51:1–40.

Davoult, D. 1989. Structure démografique et production de la population *d'Ophiothrix fragilis* (Abildgaard) du Détroit du Pas-deCalais, France. Vie Marine 10:116–127.

Davoult, D., and F. Gounin. 1995. Suspension-feeding activity of a dense *Ophiothrix fragilis* (Abildgaard) population at the water–sediment interface: time coupling of food availability and feeding behaviour of the species. Estuarine and Coastal Shelf Science 41:567–577.

Deagle, B. E., S. N. Jarman, E. Coissac, F. Pompanon, and P. Taberlet. 2014. DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. Biology Letters 10:20140562.

Deiner, K., et al. 2017. Environmental DNA metabarcoding: transforming how we survey animal and plant communities. Molecular Ecology 26:5872–5895.

Dray, S., and A. Dufour. 2007. The ade4 package: Implementing the duality diagram for ecologists. Journal of Statistical Software 22:1–20.

Duran, S., C. Palacín, M. A. Becerro, X. Turon, and G. Giribet. 2004. Genetic diversity and population structure of the commercially harvested sea urchin *Paracentrotus lividus* (Echinodermata, Echinoidea). Molecular Ecology 13:3317–3328.

Edgar, R. C. 2016. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv. https://doi.org/10.1101/081257

Edgar, R. C., and H. Flyvbjerg. 2015. Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics 31:3476–3482.

Elbrecht, V., and F. Leese. 2015. Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass-sequence relationships with an innovative metabarcoding protocol. PLoS ONE 10:e0130324.

Elbrecht, V., and F. Leese. 2017. Validation and development of COI metabarcoding primers for freshwater macroinvertebrate bioassessment. Frontiers in Environmental Science 5:11.

Elbrecht, V., E. Vamos, K. Meissner, J. Aroviita, and F. Leese. 2017. Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. Methods in Ecology and Evolution 8:1265–1275.

Elbrecht, V., E. E. Vamos, D. Steinke, and F. Leese. 2018*a*. Estimating intraspecific genetic diversity from community DNA metabarcoding data. PeerJ 6:e4644.

Elbrecht, V., P. D. N. Hebert, and D. Steinke. 2018*b*. Slippage of degenerate primers can cause variation in amplicon length. Scientific Reports 8:10999.

Emerson, B. C., F. Cicconardi, P. P. Fanciulli, and P. J. A. Shaw. 2011. Phylogeny, phylogeography, phylobetadiversity and the molecular analysis of biological communities. Philosophical Transactions of the Royal Society B 366:2391–2402.

Eren, A. M., H. G. Morrison, P. J. Lescault, J. Reveillaud, J. H. Vineis, and M. L. Sogin. 2015. Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of

high-throughput marker gene sequences. ISME Journal 9:968–979.

Ficetola, G. F., et al. 2018. DNA from lake sediments reveals long-term ecosystem changes after a biological invasion. Science Advances 4:eaar4292.

Frøslev, T. G., R. Kjoller, H. H. Bruun, R. Erjnaes, A. K. Brundjerg, C. Pietroni, and A. J. Hansen. 2017. Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. Nature Communications 8:1188.

González-Tortuero, E., J. Rusek, A. Petrusek, S. Giessler, D. Lyras, S. Grath, F. Castro-Monzón, and J. Wolinska. 2015. The quantification of representative sequences pipeline for amplicon sequencing: case study on within-population ITS1 sequence variation in a microparasite infecting *Daphnia*. Molecular Ecology Resources 15:1385–1395.

Gratton, P., S. Marta, G. Bocksberger, M. Winter, E. Trucchi, and H. Kühl. 2017. A world of sequences: can we use georeferenced nucleotide databases for a robust automated phylogeography? Journal of Biogeography 44:475–486.

Gunther, B., T. Knebelsberger, H. Neumann, S. Laakmann, and P. Martínez Arbizu. 2018. MEtabarcoding of marine environmental DNA based on mitochondrial and nuclear genes. Scientific Reports 8:14822.

Hajibabaei, M., D. J. Baird, N. A. Fahnere, R. Beiko, and G. B. Golding. 2016. A new way to contemplate Darwin's tangled bank: how DNA barcodes are reconnecting biodiversity science and biomonitoring. Philosophical Transactions of the Royal Society B 371:20150330.

Hart, M. W., and P. B. Marko. 2010. It's about time: divergence, demography, and the evolution of developmental modes in marine invertebrates. Integrative and Comparative Biology 50:643–661.

Hausser, J., and K. Strimmer. 2009. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. Journal of Machine Learning Research 10:1469–1484.

Haye, P. A., N. I. Segovia, N. C. Muñoz-Herrera, F. E. Gálvez, A. Martínez, A. Meynard, M. C. Pardo-Gandarillas, E. Poulin, and S. Faugeron. 2014. Phylogeographic structure in benthic marine invertebrates of the Southern Pacific Coast of Chile with differing dispersal potential. PLoS ONE 9:e88613.

Ji, Y., et al. 2013. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecology Letters 16:1245–1257.

Kelly, R. P., et al. 2014. Harnessing DNA to improve environmental management. Science 344:1455–1456.

Kemp, J., A. López-Baucells, R. Rocha, O. S. Wangensteen, Z. Andriatafika, A. Nair, and M. Cabeza. 2019. Bats as potential suppressors of multiple agricultural pests: a case study from Madagascar. Agriculture, Ecosystems & Environment 269:88–96.

Leray, M., and N. Knowlton. 2015. DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. Proceedings of the National Academy of Sciences USA 112:2076–2081.

Leray, M., and N. Knowlton. 2016. Censusing marine eukaryotic diversity in the twenty-first century. Philosophical Transactions of the Royal Society B 371:20150331.

Leray, M., J. Y. Yang, C. P. Meyer, S. C. Mills, N. Agudelo, V. Ranwez, J. T. Boehm, and R. J. Machida. 2013. A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. Frontiers in Zoology 10:34.

Macher, J. N., A. Vivancos, M. P. Piggott, F. C. Centeno, C. D. Matthaei, and F. Leese. 2018. Comparison of environmental DNA and bulk-sample metabarcoding using highly degenerate COI primers. Molecular Ecology Resources 18:1456–1468.

Macías-Hernández, N., K. Athey, V. Tonzo, O. S. Wangensteen, M. A. Arnedo, and J. D. Harwood. 2018. Molecular gut content analysis of different spider body parts. PLoS ONE 13: e0196589.

Mahé, F., T. Rognes, C. Quince, C. de Vargas, and M. Dunthorn. 2015. Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ 3:e1420.

Nielsen, E. S., M. Beger, R. Henriques, K. A. Selkoe, and S. von der Heyden. 2017. Multispecies genetic objectives in spatial conservation planning. Conservation Biology 31:872–882.

Olesen, S. W., C. Duvallet, and E. J. Alm. 2017. dbOTU3: A new implementation of distribution-based OTU calling. PLoS ONE 12:e0176335.

Pagès, H., P. Aboyoun, R. Gentleman, and S. DebRoy. 2018. Biostrings: Efficient manipulation of biological strings. R package version 2.50.1. https://bioconductor.org/packages/release/bioc/html/Biostrings.html

Palacin, C., G. Giribet, S. Garner, L. Dantart, and X. Turon. 1998. Low densities of sea urchins influence the structure of algal assemblages in the western Mediterranean. Journal of Sea Research 39:281–290.

Paradis, E. 2010. pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics 26:419–420.

Parsons, K. M., M. Everett, M. Dahlheim, and L. Park. 2018. Water, water everywhere: environmental DNA can unlock population structure in elusive marine species. Royal Society Open Science 5:180537.

Pascual, M., B. Rives, C. Schunter, and E. Macpherson. 2017. Impact of life history traits on gene flow: A multispecies systematic review across oceanographic barriers in the Mediterranean Sea. PLoS ONE 12:e0176419.

Pedro, P. M., et al. 2017. Metabarcoding analyses enable differentiation of both interspecific assemblages and intraspecific divergence in habitats with differing management practices. Environmental Entomology 46:1381–1389.

Pentinsaari, M., H. Salmela, M. Mutanen, and T. Roslin. 2016. Molecular evolution of a widely-adopted taxonomic marker (COI) across the animal tree of life. Scientific Reports 6:35275.

Pérez-Portela, R., V. Almada, and X. Turon. 2013. Cryptic speciation and genetic structure of widely distributed brittle stars (Ophiuroidea) in Europe. Zoologica Scripta 42:151–169.

Pfeiffer, F., C. Gröber, M. Blank, K. Händler, M. Beyer, J. L. Schultze, and G. Mayer. 2018. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Scientific Reports 8:10950.

Piñol, J., M. A. Senar, and O. C. Symondson. 2019. The choice of universal primers and the characteristics of the species mixture determines when DNA metabarcoding can be quantitative. Molecular Ecology 28:407–419.

Pochon, X., N. J. Bott, K. F. Smith, and S. A. Wood. 2013. Evaluating detection limits of next-generation sequencing for the surveillance and monitoring of international marine pest. PLoS ONE 8:e73935.

Porter, T. M., and M. Hajibabaei. 2018*a*. Automated high throughput animal CO1 metabarcode classification. Scientific Reports 8:4226.

Porter, T. M., and M. Hajibabaei. 2018*b*. Over 2.5 million COI sequences in GenBank and growing. PLoS ONE 13: e0200177.

Porter, J., and L. Zhang. 2017. InfoTrim: A DNA read quality trimmer using entropy. bioRxiv. https://doi.org/10.1101/201442

R Development Core Team 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

Ratnasingham, S., and P. D. N. Hebert. 2007. bold: The Barcode of Life data system (http://www.barcodinglife.org). Molecular Ecology Notes 7:355–364.

Rex, M. A., and R. J. Ettter. 2010. Deep-sea biodiversity. Pattern and scale. Harvard University Press, Cambridge, Massachusetts, USA.

Rognes, T., T. Flouri, B. Nichols, C. Quince, and F. Mahé 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e0200177.

Schirmer, M., R. D'Amore, U. Z. Ijaz, N. Hall, and C. Quince. 2016. Illumina error profiles: resolving fine-scale variation in metagenomics sequencing data. BMC Bioinformatics 17:125.

Schmidt, A. O., and H. Herzel. 1997. Estimating the entropy of DNA sequences. Journal of Theoretical Biology 3:369–377.

Siegenthaler, A., O. S. Wangensteen, C. Benvenuto, J. Campos, and S. Mariani. 2019. DNA metabarcoding unveils multiscale trophic variation in a widespread coastal opportunist. Molecular Ecology 28:232–249.

Sigsgaard, E. E., et al. 2016. Population characteristics of a large whale shark aggregation inferred from seawater environmental DNA. Nature Ecology and Evolution 1:4.

Song, H., J. E. Buhay, M. F. Whiting, and K. A. Crandall. 2008. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are amplified. Proceedings of the National Academy of Sciences USA 105:13486–13491.

Stat, M., M. J. Huggett, R. Bernasconi, J. D. DiBattista, T. E. Berry, S. J. Newman, E. S. Harvey, and M. Bunce. 2017. Ecosystem monitoring with eDNA: metabarcoding across the tree of life in a tropical marine environment. Scientific Reports 7:12240.

Stefanni, S., D. Stankovic, D. Borme, A. De Olazabal, T. Juretic, A. Pallavicini, and V. Tirelli. 2018. Multi-marker metabarcoding approach to study mesozooplankton at basin scale. Scientific Reports 8:12085.

Taberlet, P., A. Bonin, L. Zinger, and E. Coissac. 2018. Environmental DNA for biodiversity research and monitoring. Oxford University Press, Oxford, UK.

Taboada, S., and R. Pérez-Portela. 2016. Contrasted phylogeographic patterns on mitochondrial DNA of shallow and deep brittle stars across the Atlantic-Mediterranean area. Scientific Reports 6:32425.

Tang, C. Q., F. Leasi, U. Obertegger, A. Kieneke, T. G. Barraclough, and D. Fontaneto. 2012. The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. Proceedings of the National Academy of Sciences USA 109:16208–16212.

Uchii, K., H. Doi, and T. Minamoto. 2016. A novel environmental DNA approach to quantify the cryptic invasion of non-native genotypes. Molecular Ecology Resources 16:415-422.

Vamos, E. E., V. Elbrecht, and F. Leese. 2017. Short COI markers for freshwater macroinvertebrate metabarcoding. Metabarcoding and Metagenomics 1:e14625.

Vellend, M., G. Lajoie, A. Bourret, C. Murria, S. W. Kembel, and D. Garant. 2014. Drawing ecological inferences from coincident patterns of population- and community-level biodiversity. Molecular Ecology 23:2890–2901.

Wang, X. V., N. Blanes, J. Ding, R. Sultana, and G. Parmigiani. 2012. Estimation of sequencing error rates in short reads. BMC Bioinformatics 13:185.

Wangensteen, O. S. and X. Turon. 2017. Metabarcoding techniques for assessing biodiversity of marine animal forests. Pages 445–503 *in* S. Rossi, L. Bramanti, A. Gori, and C. Orejas, editors. Marine animal forests. The ecology of benthic biodiversity hotspots. Volume 1. Springer International Publishing, Switzerland.

Wangensteen, O. S., X. Turon, A. Garcia-Cisneros, M. Recasens, J. Romero, and C. Palacin. 2011. A wolf in sheep's clothing: carnivory in dominant sea urchins in the Mediterranean. Marine Ecology Progress Series 441:117–128.

Wangensteen, O. S., C. Palacín, M. Guardiola, and X. Turon. 2018*a*. DNA metabarcoding of littoral hard-bottom communities: high diversity and database gaps revealed by two molecular markers. Peer J 6:e4705.

Wangensteen, O. S., E. Cebrian, C. Palacín, and X. Turon. 2018*b*. Under the canopy: community-wide effects of invasive algae in Marine Protected Areas revealed by metabarcoding. Marine Pollution Bulletin 127:54–66.

Wares, J. P., and P. Pappalardo. 2016. Can theory improve the scope of quantitative metazoan metabarcoding? Diversity 8:1.

Wheeler, Q., P. H. Raven, and E. O. Wilson. 2004. Taxonomy: impediment or expedient? Science 303:285.

Zink, R. M. 2002. Methods in comparative phylogeography, and their application to studying evolution in the North American aridlands. Integrative and Comparative Biology 42:953–959.

Supporting Information

Additional supporting information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/eap.2036/full

Data Availability

Data are available on Mendeley Data at https://doi.org/10.17632/xpmtvn2k7m.2.

Antich *et al. BMC Bioinformatics* (2021) 22:177
https://doi.org/10.1186/s12859-021-04115-6

BMC Bioinformatics

RESEARCH ARTICLE Open Access

To denoise or to cluster, that is not the question: optimizing pipelines for COI metabarcoding and metaphylogeography

Adrià Antich[1], Creu Palacin[2], Owen S. Wangensteen[3*] and Xavier Turon[1*]

*Correspondence: owen.wangensteen@uit.no; xturon@ceab.csic.es
[1] Department of Marine Ecology, Centre for Advanced Studies of Blanes (CEAB-CSIC), Blanes (Girona), Catalonia, Spain[3] Norwegian College of Fishery Science, UiT The Arctic University of Norway, Tromsö, Norway
Full list of author information is available at the end of the article

Abstract

Background: The recent blooming of metabarcoding applications to biodiversity studies comes with some relevant methodological debates. One such issue concerns the treatment of reads by denoising or by clustering methods, which have been wrongly presented as alternatives. It has also been suggested that denoised sequence variants should replace clusters as the basic unit of metabarcoding analyses, missing the fact that sequence clusters are a proxy for species-level entities, the basic unit in biodiversity studies. We argue here that methods developed and tested for ribosomal markers have been uncritically applied to highly variable markers such as cytochrome oxidase I (COI) without conceptual or operational (e.g., parameter setting) adjustment. COI has a naturally high intraspecies variability that should be assessed and reported, as it is a source of highly valuable information. We contend that denoising and clustering are not alternatives. Rather, they are complementary and both should be used together in COI metabarcoding pipelines.

Results: Using a COI dataset from benthic marine communities, we compared two denoising procedures (based on the UNOISE3 and the DADA2 algorithms), set suitable parameters for denoising and clustering, and applied these steps in different orders. Our results indicated that the UNOISE3 algorithm preserved a higher intra-cluster variability. We introduce the program DnoisE to implement the UNOISE3 algorithm taking into account the natural variability (measured as entropy) of each codon position in protein-coding genes. This correction increased the number of sequences retained by 88%. The order of the steps (denoising and clustering) had little influence on the final outcome.

Conclusions: We highlight the need for combining denoising and clustering, with adequate choice of stringency parameters, in COI metabarcoding. We present a program that uses the coding properties of this marker to improve the denoising step. We recommend researchers to report their results in terms of both denoised sequences (a proxy for haplotypes) and clusters formed (a proxy for species), and to avoid collapsing the sequences of the latter into a single representative. This will allow studies at the cluster (ideally equating species-level diversity) and at the intra-cluster level, and will ease additivity and comparability between studies.

Keywords: Metabarcoding, Metaphylogeography, COI, Denoising, Clustering, Operational taxonomic units

Background

The field of eukaryotic metabarcoding is witnessing an exponential growth, both in the number of communities and substrates studied and the applications reported (reviewed in [1–4]). In parallel, technical and conceptual issues are being discussed (e.g., [5, 6]) and new methods and pipelines generated. In some cases, however, new practices are established after a paper reporting a technique is published and followed uncritically, sometimes pushing its application outside the context in which it was first developed.

A recently debated matter concerns the treatment of reads by denoising procedures or by clustering techniques [7]. Both methods are often presented as alternative approaches to the same process (e.g., [7–11]). However, both are philosophically and analytically different [12]. While denoising strives to detect erroneous sequences and to merge them with the correct "mother" sequence, clustering tries to combine a set of sequences (without regard to whether they contain or not errors) into meaningful biological entities, ideally approaching the species level, called OTUs or MOTUs (for Molecular Operational Taxonomic Units). Usually only one representative sequence from each MOTU is kept (but note that this is only common practice, not a necessary characteristic of the method). Thus, while both procedures result in a reduced dataset and in error correction (by merging reads of erroneous sequences with the correct one or by combining them with the other reads in the MOTU), they are not equivalent. More importantly, they are not incompatible at all and can (and should) be used together.

A recent paper [13] proposes that denoised sequences should replace MOTUs as the unit of metabarcoding analyses. We contend that it may be so for ribosomal DNA datasets such as the one used in that paper, but this notion has gained momentum also in other fields of metabarcoding for which it is not adequate. In particular, when it comes to highly variable markers such as COI. This proposal misses the fact that sequence clusters are a proxy for species-level entities, the basic unit in eukaryotic biodiversity studies. The 3′ half (also called Leray fragment) of the standard barcode fragment of COI (Folmer fragment) is becoming a popular choice for metabarcoding studies addressed at metazoans or at eukaryotic communities at large [14], reaching now 28% of all metabarcoding studies [15]. Metabarcoding stems from studies of microbes where 16S rRNA is the gene of choice, and the concept was then applied to analyses of the 18S rRNA gene of eukaryotes. With the recent rise of COI applications in metabarcoding, programs and techniques developed for rDNA are sometimes applied to COI without reanalysis and with no parameter adjusting given the highly contrasting levels of variation of these markers.

The idea that denoising should be used instead of clustering has been followed by some (e.g., [16–20]), while other authors have combined the two approaches (e.g., [21–23]). Indeed, denoising has the advantages of reducing the dataset and to ease pooling or comparing studies, which is necessary in long term biomonitoring applications. However, with COI there is a wealth of intraspecific information that is missed if only denoising is applied [24]. COI has been a prime marker of phylogeographic studies to date [25,

26], and these studies can be extended to metabarcoding datasets by mining the distribution of haplotypes within MOTUs (metaphylogeography [12]). The latter authors suggested to perform clustering first, and that denoising should be done within MOTUs to provide the right context of sequence variation and abundance skew. They also advised to perform a final abundance filtering step. In other studies, denoising is performed first, followed by clustering and refining steps (e.g., [22, 23]).

There are several methods for denoising (reviewed in [27]) and for clustering (reviewed in [28]). We will use two of the most popular denoising techniques, based on the DADA2 algorithm (Divisive Amplicon Denoising Algorithm, [29]) and the UNOISE3 algorithm [30]. The results of the former are called Amplicon Sequence Variants (ASVs) and those of the latter ZOTUs (zero-radius OTUs). In practice, the terminology is mixed and ASV, ZOTU, ESV (Exact Sequence Variant), sOTU (sub-OTU) or ISU (Individual Sequence Variant), among others, are used more or less interchangeably. For simplicity, as all of them are equivalent, we will use henceforth the term ESV. Clustering, on the other hand, can be performed using similarity thresholds (e.g., [31, 32]), Bayesian Methods (CROP, [33]), or methods based on single-linkage-clustering (SWARM, [34]), among others. We will focus on de novo clustering methods (i.e., independent of a reference database), while denoising is always de novo by its very nature [13]. We will here use SWARM as our choice of clustering program due to its good performance compared to other methods [28]. It is noteworthy that all these programs were originally developed and tested on ribosomal DNA datasets. When applied to other markers, often no indication of parameter setting is given (i.e., omega_A for DADA2, α for UNOISE3, d for SWARM), suggesting that default parameter values are used uncritically.

In this article, we aim to use a COI metabarcoding dataset of benthic littoral communities to (1) set the optimal parameters of the denoising and clustering programs for COI markers, (2) compare results of the DADA2 algorithm with the UNOISE3 algorithm, (3) compare the results of performing only denoising, only clustering, or combining denoising with clustering in different orders, and (4), suggest and test improvements in the preferred denoising algorithm to take into account the fact that COI is a coding gene. We implement these modifications in the new program DnoisE. Our aims are to provide guidelines for using these key bioinformatic steps in COI metabarcoding and metaphylogeography. The conceptual framework of our approach is sketched in Fig. 1.

Methods

The dataset

We used as a case study an unpublished dataset of COI sequences obtained from benthic communities in 12 locations of the Iberian Mediterranean. Information on the sampling and sample processing is given in Additional File 1. Sequences were obtained in a full run of an Illumina MiSeq (2 * 250 bp paired-end reads).

Bioinformatic analyses

The initial steps of the bioinformatic pipeline followed [12] and were based on the OBItools package [35]. Reads were paired and quality filtered, demultiplexed, and dereplicated. A strict length filter of 313 bp was used. We also eliminated sequences with only one read. Chimera detection was performed on the whole dereplicated

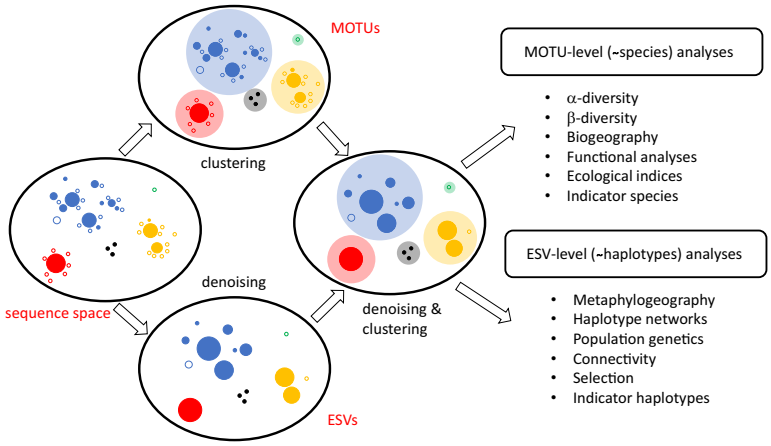


Fig. 1 Conceptual overview of the denoising and clustering processes. The oval on the left sketches a fragment of the sequence space with four biological species plus an artefact divergent sequence (denoted by colours). Correct sequences are indicated by filled circles and artefacts by empty circles, with indication of abundance (circle size). Denoising results in the detection of putatively correct sequences to which the reads of putatively incorrect sequences are merged (leading to a reduced dataset). The outcome of denoising should ideally approach the true haplotype composition of the samples. Clustering generates MOTUs without regard as to whether the grouped sequences are erroneous or not. This is usually accompanied by read pooling and keeping only one representative sequence per MOTU (leading to a reduced dataset). The outcome of clustering should ideally approach the species composition of the samples. Combining both processes results in a dataset that is reduced in size, comparable across studies, and amenable to analyses at the MOTU (species) and ESV (haplotype) levels. Note that errors likely persist in the final dataset both as artefact MOTUs and artefact ESVs within MOTUs, and carefully designed filters should be used to minimize them (abundance filtering, chimera filtering, numts removal)

dataset with uchime3_denovo as embedded in unoise3 (USEARCH 32-bit free version, [36]). We used minsize = 2 to include all sequences. Those identified as chimeras were recovered from the –tabbedout file and eliminated from the dataset. Sequences with small offsets (misaligned), identified as shifted in the output, were likewise deleted. The working dataset thus comprised well-aligned, chimera-free, unique sequences which had appeared with at least two reads in the samples.

Note that for this technical study we didn't consider the sample distribution of the reads. A complete biogeographic study of the samples is ongoing and will be published elsewhere. For the present analysis, for each unique sequence only the actual DNA sequence and the total number of reads were retained.

The denoisers: UNOISE3 and DADA2

Comparing denoising algorithms is challenging because each method comes with a different software suite with embedded features and recommendations [27]. For instance, uchime3_de novo is embedded in the unoise3 command as implemented in USEARCH, while a chimera removal procedure (removeBimeraDenovo) is an optional feature in the DADA2 pipeline. Furthermore, while UNOISE3 uses paired reads, DADA2 recommends denoising forward and reverse reads separately,

and then performing a merging step. We have tried to isolate the algorithms from their pipelines for comparability. This was done by generating a Python script [37] that implements the algorithm described in [30] and by using DADA2 from its R package v. 1.14.1 and not as embedded into the qiime2 pipeline [38].

For UNOISE3, our program (henceforth DnoisE) was compared on the working dataset described above with command unoise3 in USEARCH with minsize = 2, alpha = 5 and without the otutab step. That is, we recovered the ESV composition and abundance with an R script directly from the output of unoise3 (using the output files –tabbedout and –ampout), without a posterior re-assignment of sequences to ESVs via otutab. This step was not necessary as all sequences were included in the ESV calculations. The results of DnoisE and unoise3 were > 99.99% identical in ESVs recovered and reads assigned to them, so we continued to use our script for performing the comparisons and for further improvements of the algorithm (see below).

The recommended approach for DADA2 is to denoise separately the forward and reverse reads of each sequence. This complicates the technical comparison, as all initial filtering steps cannot be equally performed (e.g., we won't know if there is just one read of a particular sequence, or if the merged pair will be discarded for low quality of the assembly or for unsuitable final length) and thus we cannot have two identical starting datasets. More importantly, we cannot use this procedure when we test the effects of denoising at later steps (i.e., after clustering), so we would be unable to compare the denoisers at this level. Thus, for our comparative analysis we need to use DADA2 on paired reads. According to Callahan et al. [29], this can result in a loss of accuracy, but this point has never been tested to our knowledge. We addressed this issue by comparing denoising before and after pairing on half of the reads in the final dataset. After this analysis, we decided to continue our comparison of DADA2 and UNOISE3 on paired reads.

Additionally, denoising before pairing is not optimal if a PCR-free library preparation protocol is used, as in our case, because half of the reads are in one direction and the other half are in the opposite direction (hence the use of half of the reads in the above comparison). Forward and reverse reads can of course be recombined to generate new files with all reads in the same direction, but the quality of the reads with original forward and reverse orentation is different. Alternatively, two rounds of DADA2 (one per orientation) must be performed and combined at later steps.

To run DADA2 on paired reads, we entered them in the program as if they were the forward reads and did not use a merging step after denoising. In all DADA2 runs we did not perform the recommended chimera removal procedure as the input sequences were already chimera-free according to uchime3_de novo. Note that, when denoising was done after clustering, we used error rates calculated for the whole dataset, and not for each MOTU separately (most of them do not have enough number of sequences for a reliable estimation of error rates).

UNOISE3 relies heavily on the stringency parameter α, which weights the distance between sequences as a function of the number of differences between them [30]. In short, lower values of α tend to merge sequences more strongly, while higher values recovered higher numbers of ESVs. The default, and the value used in most studies with ribosomal DNA, is 2. However, for COI three independent approaches, based on

mock communities [39], entropy changes [12], and co-sequenced control DNA [40] suggested that for this marker $\alpha = 5$ is the optimal value. For DADA2 the key parameter is omega_A, which indicates the probability threshold at which a sequence *i* is considered an error derived from another sequence *j* given their abundance values and the inferred error rates. If the observed value is higher than omega_A, then sequence *i* is considered an error of sequence *j*. Omega_A is by default set to a very low value (10^{-40}), but no study has analysed the impact of changing this parameter for COI datasets. To our knowledge, only [41], based on a comparison of 3 values, concluded that the default value of omega_A was adequate for a marker based on the control region of the mitochondrial DNA.

The clustering algorithm

Our preferred clustering method is SWARM v3 [42], as it is not based on a fixed distance threshold and is independent of input order. It is a very fast procedure that relies on a single-linkage method with a clustering distance (*d*), followed by a topological refining of the clusters using abundance structures to divide MOTUs. As we were interested in keeping all sequences within MOTUs, and not just a representative sequence, we mined the SWARM output with an R script to generate MOTU files, each with its sequence composition and abundance.

The crucial parameter in this approach is *d*, the clustering distance threshold for the initial phase. The default value is 1 (that is, amplicons separated by more than one difference will not be clustered together), and this value has been tested in ribosomal DNA. However, Mahé et al. [42] pointed out that higher *d* values can be necessary for fast evolving markers (such as COI) and advised to analyse a range of *d* to identify the best fitting parameter (i.e., avoiding over- or under-clustering) for a particular dataset or scientific question. A *d* value of 13 (thus, allowing 13 differences over ca. 313 bp to make a connection) has been recently used for the Leray fragment of COI (e.g., [43–47]), but a formal study of its adequacy has not been published yet.

Setting the right parameters

With our dataset, we assessed the best-fitting parameters for UNOISE3, DADA2 and SWARM as applied to COI data. For the first two, we used changes in diversity values per codon position (measured as entropy, [48]), as calculated with the R package *entropy* [49]. Coding sequences have properties that can be used in denoising procedures [12, 41]. They have naturally a high amount of variation concentrated in the third position of the codons, while errors at any step of the metabarcoding pipeline would be randomly distributed across codon positions. Thus, examining the change in entropy values according to codon position can guide the choice of the best cleaning parameters. Turon et al. [12] suggested to use the entropy ratio (Er) between position 2 of the codons (least variable) and position 3 (most variable). In a simulation study these authors showed that Er decreased as more stringent denoising was applied until reaching a plateau, which was taken as the indication that the right parameter value had been reached.

Using the Er to set cut-points, we re-assessed the adequate value of α in UNOISE3 testing the interval of $\alpha = 1$ to 10. With the same procedure, we tested DADA2 for values of omega_A between $10^{-0.05}$ (ca. 0.9) and 10^{-90}.

For SWARM, we compared the output of SWARM with a range of values of *d* from 1 to 30 applied to our dataset (prior to denoising). We monitored the number of MOTUs generated and the mean intra- and inter-MOTU distances to find the best-performing value of *d* for our fragment.

The impact of the steps and their order

With the selected optimal parameters for each method, we combined the two denoising procedures and the clustering step in different orders. We therefore combined denoising (Du for UNOISE3 algorithm implemented in DnoisE, Da for DADA2) and clustering with SWARM (S) and generated and compared datasets of ESVs and MOTUs as follows (for instance, Da_S means that the dataset was first denoised with DADA2, then clustered with SWARM):

ESVs: Du, Da
MOTUs: Du_S, Da_S, S_Du, S_Da

For comparison of datasets, we used Venn diagrams and an average match index of the form

$$\text{Match Index (A,B)} = \left(N_{match\_A}/N_A + N_{match\_B}/N_B\right)/2$$

where N_{match_A} is the number of a particular attribute in dataset A that is shared with dataset B, and N_A is the total number of that attribute in dataset A. The same for N_{match_B} and N_B. The matches can be the number of ESVs shared, the number of MOTUs shared, the number of ESVs in the shared MOTUs, or the number of reads in the shared ESVs or MOTUs, depending on the comparison.

Improving the denoising algorithm

The preferred denoising algorithm (UNOISE3, see Results) has been further modified in two ways. Let i be a potential error sequence derived from sequence j. The UNOISE3 procedure is based on two parameters: the number of sequence differences between i and j (d, as measured by the Levenshtein distance) and the abundance skew (β, abundance i/abundance j) between them. These parameters are related by the simple formula [30]:

$$\beta(d) = 1/2^{\alpha d+1}$$

where β(d) is the threshold abundance skew allowed between two sequences separated by distance d so that below it the less abundant would be merged with the more abundant, and α is the stringency parameter. Thus, presumably incorrect "daughter" sequences are merged with the correct "mother" sequences if the number of sequence differences (d) is small and the abundance of the incorrect sequence with respect to the correct one (abundance skew) is low. The higher the number of differences, the lower the skew should be for the sequences to be merged.

For COI, however, the fact that it is a coding gene is a fundamental difference with respect to ribosomal genes. In a coding fragment, the amount of variability is substantially different among codon positions. This is not considered in the UNOISE3 formulation (nor in DADA2 or other denoising programs that we knew of, for that matter). We suggest to incorporate this information in DnoisE by differentially weighting the d values according to whether the change occurs in the first, second, or third codon position. Note that our sequences are all aligned and without indels, which makes this weighting scheme straightforward. The differences in variability can be quantified as differences in entropy values [48]; position 3 of the codons has the highest entropy, followed by position 1 and position 2. In other words, two sequences separated by n differences in third positions are more likely to be naturally-occurring sequences than if the n differences happen to occur in second positions, because position 3 is naturally more variable. To weight the value of d, we first record the number of differences in each of the three codon positions (d(1) to d(3)), we then correct the d value using the formula

$$d_{corr} = \sum_{i=1}^{3} d(i)*entropy(i)*3/\big(entropy(1) + entropy(2) + entropy(3)\big)$$

where i is the position in the codon, and d_{corr} is the corrected distance that will be used in the UNOISE3 formula instead of d.

With this formula, two sequences separated by just one difference in each codon position will continue to have a d of 3, but a change in a high entropy position (3) will translate in a higher d than the same change in a low entropy position (2), thus the program will tend to keep the former and to merge the later. The entropy of the three positions of the codons for the weighting was obtained from the original dataset prior to any denoising, thus entropy(1) = 0.473, entropy(2) = 0.227, and entropy(3) = 1.021. Note that d(i) is based on the number of differences occurring at each codon position. The Levenshtein distance used in the non-corrected d measures is not adequate for this purpose, as it cannot keep track of codon positions. However, for sequences of equal length, aligned, and without indels, as in our case, the number of differences is in practice equivalent to the Levenshtein distance.

The present algorithm of UNOISE3 gives precedence to the abundance skew over the number of differences (d) because sequences are considered in order of decreasing abundance. Thus, a very abundant sequence will form a centroid that can "capture" a rare one even if d is relatively high. Other, somewhat less abundant, sequences can be more similar (less d) to the rare sequence and can fulfil the conditions to capture it, but this will never happen as the rare sequence will be incorporated to the first centroid and will become unavailable for further comparisons. In our modification, DnoisE does not automatically join sequences to the first centroid that fits the condition. Rather, for each sequence the potential "mothers" are stored (with their abundance skew and d) and the sequences are left in the dataset. After the round of comparisons is completed, for each daughter sequence we can choose, among the potential mothers, the one whose abundance skew is lower (precedence to abundance skew, corresponding to the usual UNOISE3 procedure), the mother with the lowest

distance (precedence to d), or the one for which the ratio (abundance skew/max abundance skew for the observed d, β(d)) is lower, thus combining the two criteria.

We compared in our dataset the results of the different formulations of DnoisE: precedence to abundance skew, precedence to distance, combined precedence, and correcting distances according to codon position of the differences. A beta version of DnoisE is available from [37].

Benchmarking

Ground truthing is a difficult task in metabarcoding studies. Constructing mock communities is the most common method. However, mock communities, even the largest ones, are orders of magnitude simpler than complex biological communities. Thus, some technical aspects cannot be tested accurately. For instance, metabarcoding results of mock communities in general lack true sequences at very rare abundances (the most problematic ones). For complex communities, we need to rely on metrics that can evaluate the fit of denoising and filtering procedures. The coding properties of COI can help design useful parameters, such as the entropy ratio mentioned above. Another possible metric stems from the evaluation of the prevalence of incorrect ESVs (defined by having indels or stop codons) across denoising and filtering procedures [50].

In this work, we have performed two benchmarking procedures that rely on taxonomic assignment of the MOTUs. This assignment was done using the ecotag procedure in OBItools against the db-COI_MBPK database [51], containing 188,929 eukaryote COI reference sequences (available at [52]). Ecotag assigns a sequence to the common ancestor of the candidate sequences selected in the database, using the NCBI taxonomy tree. This results in differing taxonomic rank of the assignments depending on the density of the reference database for a given taxonomic group.

First, we checked the performance of the entropy correction of DnoisE by examining the percent of incorrect to total ESVs. To this end, we retained only the MOTUs assigned to metazoans and, following [12], examined the presence of stop codons and changes in the 5 aminoacids present in the fragment amplified that are conserved among metazoans [53]. To be on the conservative side, for a given MOTUs we evaluated the different genetic codes and selected the ones that produced the smaller number of stop codons. The five aminoacids were then checked using these codes and the minimal number of "wrong" aminoacids was recorded. The R package Biostrings [54] was used for the translations. The ESVs featuring stop codons and/or aminoacid changes in the five conserved positions were labelled as erroneous. The rationale is that a suitable denoising procedure would reduce the ratio of error vs total ESVs.

Second, we performed a taxonomic benchmarking. As MOTUs should ideally reflect species-level entities, we selected those sequences assigned at the species level as a benchmark for the MOTU datasets. We also enforced a 97% minimal best identity with the reference sequence. We traced these sequences in the output files of our procedures and classified the MOTUs containing them into three categories (following the terminology in [9]): closed MOTUs, when they contain all sequences assigned to a species and only those; open MOTUs, when they contain some, but not all, sequences assigned to one species and none from other species, and hybrid MOTUs. The latter

included MOTUs with sequences assigned to more than one species, or MOTUs with a combination of sequences assigned to one species and sequences not assigned (i.e., they don't have species-level assignment, or they do with less than 97% similarity).

This analysis was intended as a tool for comparative purposes, to benchmark the ability of the different MOTU sets generated to recover species-level entities. In other words, which procedure retains more ESVs with species-level assignment and places them in closed (as opposed to open or hybrid) MOTUs.

Results

The dataset

After pairing, quality filters, and retaining only 313 bp-long reads, we had a dataset of 16,325,751 reads that were dereplicated into 3,507,560 unique sequences. After deleting singletons (sequences with one read), we kept 423,164 sequences (totalling 10,305,911 reads). Of these sequences, 92,630 were identified as chimeras and 152 as misaligned sequences and eliminated. Our final dataset for the study, therefore, comprised 330,382 sequences and 9,718,827 reads (the original and the refined datasets were deposited in Mendeley Data, [55]).

For testing the performance of DADA2 on unpaired and paired reads on a coherent dataset, we selected the reads that were in the forward direction, that is, the forward primer was in the forward read (R1). As expected, they comprised ca. half of the reads (4,892,084). For these reads we compared the output of applying DADA2 before and after pairing, as detailed in Additional File 2. The results were similar, with most reads placed in the same ESVs in both datasets, albeit 21% more low-abundance ESVs were retained using the paired reads. Henceforth we will use DADA2 on paired sequences, as this was necessary to perform our comparisons.

Setting the right parameters

We used the change in entropy ratio (Er) of the retained sequences of the global dataset (330,382 sequences and 9,718,827 reads) for selecting the best performing α–value in UNOISE3 and the best omega_A in DADA2 across a range of values. We also assessed the number of ESVs resulting from the procedures.

For UNOISE3 as implemented in our DnoisE script, the Er diminished sharply for α–values of 10 to 7, and more smoothly afterwards (Fig. 2a). The number of ESVs detected likewise decreased sharply with lower α–values, but tended to level off at $\alpha = 5$ (Fig. 2a). The value of 5 seems a good compromise between minimizing the Er and keeping the maximum number of putatively correct sequences.

For the DADA2 algorithm we tested a wide range of omega_A from $10^{-0.05}$ to 10^{-90} (we set parameter omega_C to 0 in all tests, so all erroneous sequences were corrected). The results showed that, even at the highest value ($10^{-0.05}$, or ca. 0.9 p-value, thus accepting as new partitions a high number of sequences), there was a substantial drop in number of sequences (ca. 75% reduction) and in Er with respect to the original dataset (Fig. 2b). Both variables remained relatively flat with a slight decrease between omega_A 10^{-2} and 10^{-15}, becoming stable again afterwards (Fig. 2b).

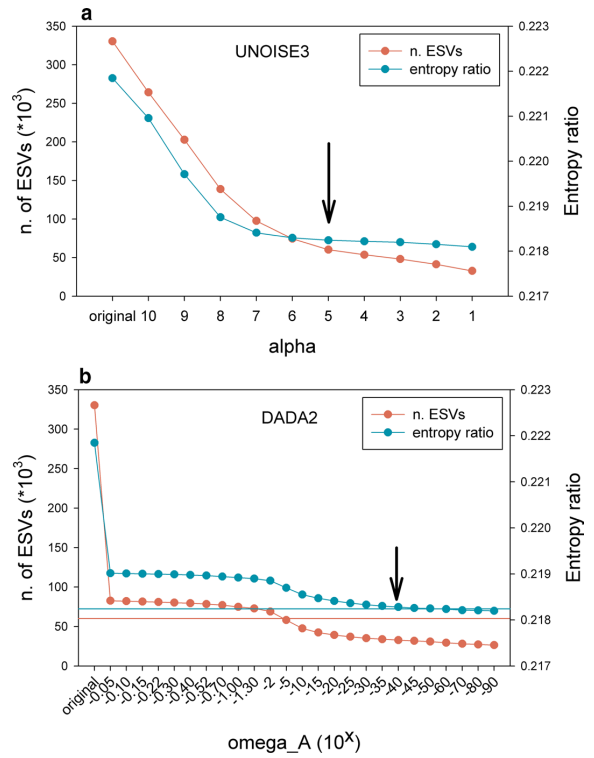


Fig. 2 Values of the Entropy ratio (Er) of the set of ESVs obtained with the UNOISE3 algorithm at decreasing values of α (**a**), and of those obtained with the DADA2 algorithm at decreasing values of omega_A (**b**). Arrows point at the selected value for each parameter. Horizontal blue line in (**b**) represents the Er value reached in (**a**) at α = 5, horizontal red line marks the number of ESVs detected in (**a**) at α = 5

The number of ESVs retained was considerably lower than for UNOISE3. In fact, the number obtained at α = 5 by the latter (60,198 ESVs) was approximately reached at omega_A = 10^{-5} (58,191 ESVs). On the other hand, the entropy value obtained at α = 5 in UNOISE3 (0.2182) was not reached until omega_A = 10^{-60}. As a compromise, we will use in this study the default value of the dada function (10^{-40}), while acknowledging that the behaviour of DADA2 with changes in omega_A for the parameters analysed was unexpected and deserves further research.

For the clustering algorithm SWARM v.2, we monitored the outcome of changing the *d* parameter between 1 and 30. For each value, we tracked the number of clusters formed (separately for all MOTUs and for those with 2 or more sequences), as well as the mean intra-MOTU and the mean inter-MOTU genetic distances (considering only the most abundant sequence per MOTU for the latter). The goal was to find the value that maximizes the intra-MOTU variability while keeping a sharp difference between both values (equivalent to the barcode gap).

The total number of MOTUs decreased sharply from 38,560 ($d = 1$) to around 19,000 with a plateau from $d = 9$ to $d = 13$, and then decreased again (Fig. 3a). If we only

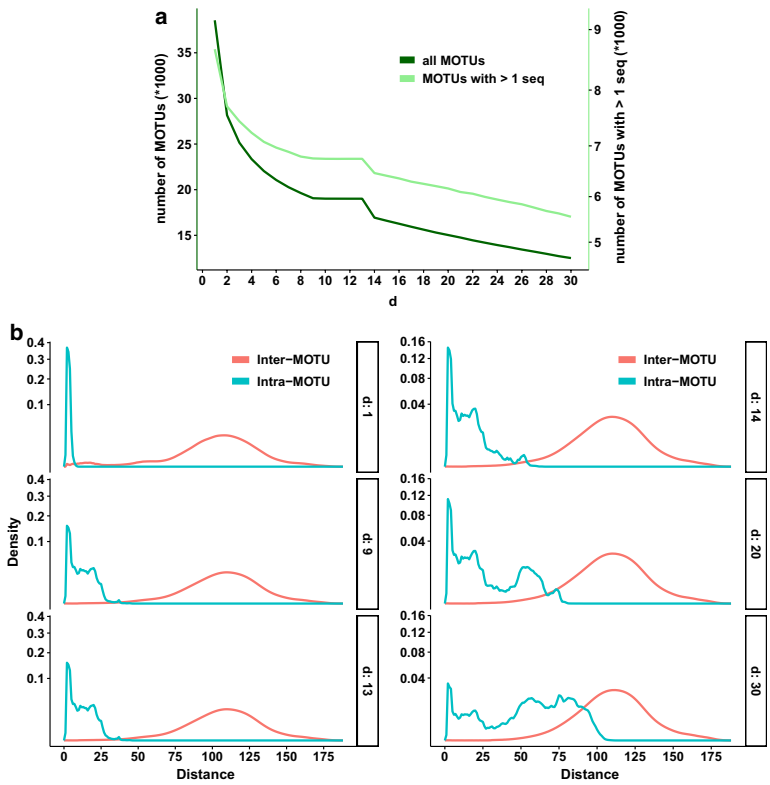


Fig. 3 **a** Number of MOTUs obtained at different values of *d* using SWARM. Total number of MOTUs (dark green) and of MOTUs with two or more sequences (light green) are represented (note different Y-axes). **b** Density plots (note quadratic scale) showing the distribution of number of differences between different clusters (inter-MOTU, red) and sequences within clusters (intra-MOTU, blue) obtained by SWARM for selected values of the parameter *d* (1, 9, 13, 14, 20 and 30)

consider the MOTUs with 2 or more sequences, the overall pattern is similar, albeit the curve is much less steep. The numbers decreased from 8684 for $d=1$ to 6755 at $d=12$ and 13, and decreased again at higher values (Fig. 3a).

Inter-MOTU distances had a similar distribution with all values of the parameter *d*, albeit with a small shoulder at distances of 10–20 differences with $d=1$ (selected examples in Fig. 3b). Intra-MOTU distances, on the other hand, became more spread with higher values of *d* as expected. Values from 9 to 13 showed a similar distribution of number of differences, but for *d* values higher than 14, intra-MOTU distances started to overlap with the inter-MOTU distribution (Fig. 3b). The value of $d=13$ seems, therefore, to be the best choice to avoid losing too much MOTU variability (both in terms of number of MOTUs and intra-MOTU variation), and at the same time keeping intra- and inter-MOTU distances well separated. The mean intra-MOTU distance in our dataset at $d=13$ was 9.10 (equivalent to 97.09% identity), and the mean inter-MOTU distance was 108.78 (65.25% identity).

Table 1 Main characteristics of the original and the generated datasets

	n. ESVs (*)	n. MOTUs	Single-ESV MOTUs	ESVs/MOTU (*)	Reads/MOTU
Original	330,382	–	–	–	–
Du (**)	60,198	–	–	–	–
Da	32,798	–	–	–	–
Du_e (***)	113,133	–	–	–	–
S	330,382	19,012	12,257	17.378	511.194
Du_S	60,198	19,058	12,471	3.159	509.961
S_Du	75,069	19,012	12,433	3.949	511.194
Da_S	32,798	19,167	15,565	1.711	507.060
S_Da	35,376	19,012	15,198	1.861	511.194
Du_d_S	60,198	19,058	12,471	3.159	509.960
Du_c_S	60,198	19,058	12,471	3.159	509.960
Du_e_S	113,133	19,016	12,365	5.949	511.087
Du_e_d_S	113,133	19,016	12,365	5.949	511.087
Du_e_c_S	113,133	19,016	12,365	5.949	511.087

All datasets had 9,718,827 reads. 1-ESV MOTUs refer to the number of MOTUs with just one ESV. Codes of the datasets: Du, denoised with UNOISE3 algorithm (unless otherwise stated, it refers to the original formulation giving precedence to abundance ratio); Da, denoised with DADA2 algorithm; S, clustered with SWARM algorithm; Du_S, denoised (UNOISE3) and clustered; S_Du, clustered and denoised (UNOISE3); Da_S, denoised (DADA2) and clustered; S_Da, clustered and denoised (DADA2); Du_d_S, denoised (UNOISE3) with precedence to distance and clustered; Du_c_S, denoised (UNOISE3) with combined precedence and clustered; Du_e _S, denoised (UNOISE3) with correction taking into account the entropy of the codon positions and clustered; Du_e_d_S, denoised (UNOISE3) with correction plus precedence to distance and clustered; Du_e_c_S, denoised (UNOISE3) with correction plus combined precedence and clustered

*For the original and S datasets the number of sequences instead of ESVs is used

**The same values apply to Du_d (distance precedence) and Du_c (combined precedence)

***The same values apply to Du_e_d (distance precedence) and Du_e_c (combined precedence)

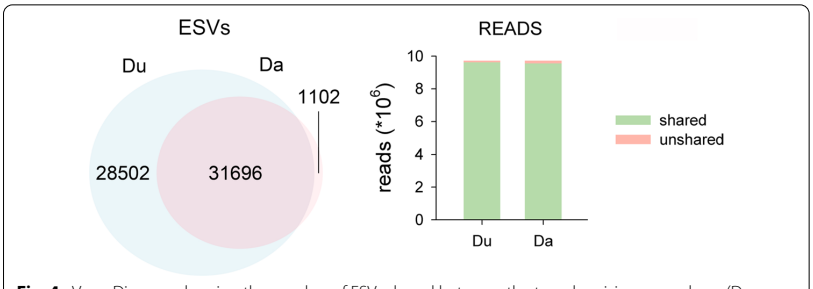


Fig. 4. Venn Diagram showing the number of ESVs shared between the two denoising procedures (Du vs Da). Bar chart shows the number of reads in the shared and unshared ESVs

The impact of the steps and their order

Table 1 shows the main characteristics of the original and the generated datasets, as well as the datasets obtained by modifying the UNOISE3 algorithm (see below). All datasets are available from Mendeley Data [55]).

We first compared the outcomes of denoising the original reads with UNOISE3 and DADA2 (Du vs Da), with the stringency parameters set as above. The error rates of the different substitution types as a function of quality scores were highly correlated in the DADA2 learnErrors procedure. The lowest Pearson correlation was obtained between

the substitutions T to C and A to G ($r = 0.810$), and all correlations (66 pairs of substitution types) were significant after a False Discovery Rate correction [56].

The main difference found is that the Du dataset retained almost double number of ESVs than the Da dataset: 60,198 vs 32,798. Of these, 31,696 were identical in the two datasets (Fig. 4), representing a match index of 0.746. Of the shared ESVs, 20,691 (65.28%) had exactly the same number of reads, suggesting that the same reads have been merged in these ESVs.

On the other hand, the shared ESVs concentrated most of the reads (Fig. 4): the match index for the reads was 0.986. This is coherent with the fact that most of the non-shared ESVs of the Du dataset had a low number of reads (mean = 3.66). Thus, the two denoising algorithms with the chosen parameter values provided similar results as for the abundant ESVs, but UNOISE3 retained a high number of low abundance ESVs as true sequences.

We then evaluated the output of combining denoising and clustering, using either of them as a first step. Thus, we compared the datasets Du_S, S_Du, Da_S, and S_Da. The results showed that the final number of MOTUs obtained was similar (ca. 19,000) irrespective of the denoising method and the order used (Table 1). Moreover, the shared MOTUs (flagged as MOTUs that have the same representative sequence) were the overwhelming majority (Fig. 5), with MOTU match indices over 0.96 in all comparisons.

As for the number of ESVs, clustering first results in a higher number of retained sequence variants than clustering last, ca. 25% more for Du and ca. 8% for Da. In all comparisons, the majority of ESVs were to be found in the shared MOTUs, and the same applies to the number of reads (Fig. 6, match indices for the ESVs, all > 0.95, match indices for the reads, all > 0.99). Ca. 2/3 of the MOTUs comprised a single ESV when using Du, and this number increased notably with Da (ca. 80% of MOTUs, Table 1). In both cases, clustering first resulted in a slight decrease of the number of single-ESV MOTUs.

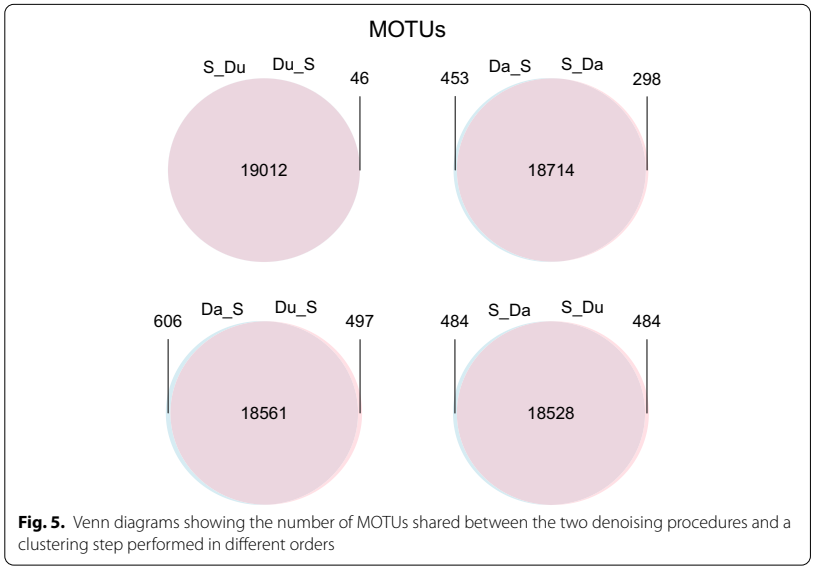


Fig. 5. Venn diagrams showing the number of MOTUs shared between the two denoising procedures and a clustering step performed in different orders

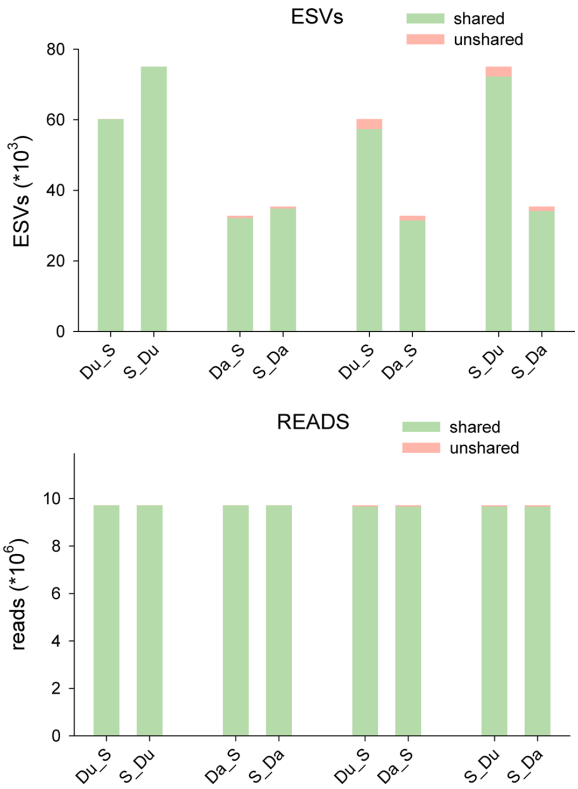


Fig. 6. Bar charts of the number of ESVs and the number of reads found in the shared and unshared MOTUs in the same comparisons as in Fig. 5

Improving the denoising algorithm

We tried different options of our DnoisE algorithm. The use of the Levenshtein distance without any correction and with priority to abundance skew corresponds to the original UNOISE3 algorithm (i.e., the Du dataset used previously). We also tried priority to distance and a combination of skew and abundance to choose among the potential "mother" ESVs to which a given "daughter" sequence will be joined. The same three options were applied when correcting distances according to the entropy of each codon position. In this case we used a pairwise distance accounting for the codon position where a substitution was found. We further applied a clustering step (SWARM) to the DnoisE results to generate MOTU sets (Du_S, Du_d_S, Du_c_S, Du_e_S, Du_e_d_S, Du_e_c_S, see Table 1 for explanation of codes) for comparison with those obtained previously.

The three ways to join sequences have necessarily the same ESVs, only the sequences that are joined under each centroid can vary and, thus, the abundance of each ESV and how these are clustered in MOTUs. However, this had a very small effect in our case. For the three datasets generated without distance correction, most MOTUs were shared, and the shared MOTUs comprised most ESVs. In turn most ESVs have the same number

of reads, suggesting that the same sequences have been grouped in each ESV. All match indices were ca. 0.99. The same was found for the three entropy-corrected datasets.

On the other hand, if we consider the entropy of codon positions the results change notably in terms of ESV recovered. The corrected datasets have 113,133 ESVs (against 60,198 of the uncorrected datasets). So, when considering the entropy in distance calculations the number of retained ESVs increased by 88%. This is the result of accepting sequences that have variation in third codon positions as legitimate. When comparing the entropy-corrected and uncorrected datasets 57,318 ESVs were found in common (ESV match index of 0.729). These ESVs comprise a majority of reads, though (read match indices of ca. 0.97 in all possible comparisons). Figure 7 illustrates one of these comparisons (Du vs Du_e_c).

When clustering the ESVs obtained with the different methods, the final number of MOTUs obtained was similar to those generated in the previous sections (ca. 19,000 in all cases, Table 1). This indicates that the entropy corrected datasets provided more intra-MOTU variability, but no appreciable increase in the number of MOTUs. As an example, the mean number of ESVs per MOTU was 3.159 for the Du_S dataset, and 5.949 for the Du_e_c_S dataset. The number of single-ESV MOTUs decreased slightly (12,471 for Du_S, 12,365 for Du_e_c_S). Taking this comparison as an example, most MOTUs (as indicated by identity in the representative sequence) were shared between datasets. In addition, most of the ESVs and most of the reads were found in the shared MOTUs (match indices for MOTUs, ESVs and reads > 0.99).

Benchmarking

We computed the percent of erroneous ESVs (either because they have stop codons or changes in the five conserved aminoacids) in the MOTUs assigned to metazoans for the datasets obtained with and without entropy correction. The original dataset clustered without any denoising (dataset S) had 9,702 erroneous ESVs (or 4.65% of the total number of ESVs). The denoised dataset Du_S had 559 erroneous ESVs (1.58%), while the dataset denoised considering the variability of the codon positions (Du_e_c_S) had 500 erroneous ESVs (0.70%). Thus, albeit the uncorrected UNOISE3 procedure reduced the proportion of errors to one third, when a correction for codon position is applied the absolute number of errors is reduced, out of almost double total number of ESVs, thus the relative number is cut by more than one half.

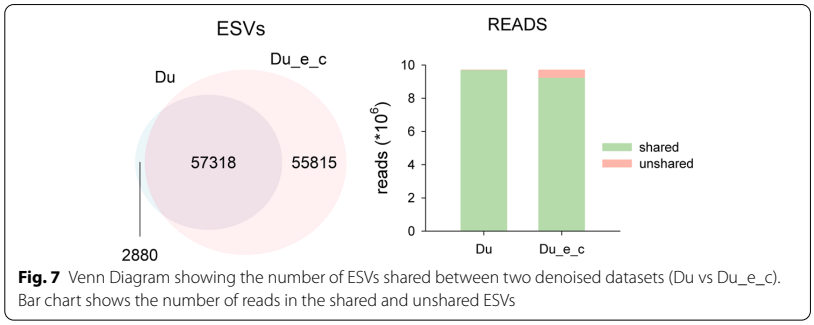


Fig. 7 Venn Diagram showing the number of ESVs shared between two denoised datasets (Du vs Du_e_c). Bar chart shows the number of reads in the shared and unshared ESVs

The results of the taxonomic benchmarking are given in detail in Additional File 3, while the obtained species-level dataset is available as Additional File 4. In short, all datasets recovered a majority of closed MOTUs, meaning that ESVs assigned to a given species were placed in the same MOTU. The proportion of hybrid MOTUs was lower for the more stringent DADA2 datasets. On the contrary, the proportion of species recovered and the proportion of ESVs with species-level assignment was lowest for the DADA2 datasets and highest for the entropy-corrected UNOISE3 datasets.

Discussion

After adjusting the different parameters of the algorithms based on ad hoc criteria for COI amplicons, between ca. 33,000 and ca. 113,000 ESVs were obtained depending on the denoising procedure used. Irrespective of the method, however, they clustered into ca. 19,000 MOTUs. This implies that there was a noticeable intra-MOTU variability even for the most stringent denoising method. The application of SWARM directly to the original dataset (without any denoising) generated likewise ca. 19,000 MOTUs. This suggests that the SWARM algorithm is robust in recovering alpha-diversity even in the presence of noisy sequences. Thus, denoising and clustering clearly accomplish different functions and, in our view, both are complementary and should be used in combination. The fact that some studies detect more MOTUs than ESVs when analysing datasets using clustering and denoising algorithms separately (e.g., [8, 57]) reflects a logical flaw: MOTUs seek to recover meaningful species-level entities, ESVs seek to recover correct sequences. There should be more sequences than species, otherwise something is wrong with the respective procedures. It has even been suggested that ESVs or MOTUs represent a first level of sequence grouping and that a second round using network analysis is convenient [9]. We contend that, with the right parameter settings, this is unnecessary for eukaryotic COI datasets.

We do not endorse the view of Callahan et al. [13] that ESVs should replace MOTUs as the standard unit analysis of amplicon-sequencing datasets. Using information at the strain level may be useful in the case of prokaryotes, and in low-variability eukaryote markers such as ribosomal 18S rDNA there may be correspondence between species and unique sequences (indeed, in many cases different species share sequences). But even in more variable nuclear markers such as ITS, a clustering step is necessary [58]. In eukaryotes the unit of diversity analyses is the species. MOTUs and not ESVs target species-level diversity and, in our view, should be used as the standard unit of analyses for most ecological and monitoring applications. Most importantly, that ESVs are organized into MOTUs is highly relevant information added at no cost. We do not agree that clustering ESVs into MOTUs eliminates biological information [29]. This only happens if only one representative sequence per MOTU is kept. We strongly advocate here for keeping track of the different sequences clustered in every MOTU and reporting them in metabarcoding studies. In this way analyses can be performed at the MOTU level or at the ESV level, depending on the question addressed.

Denoising has been suggested as a way to overcome problems of MOTU construction and to provide consistent biological entities (the correct sequences) that can be compared across studies [13]. We fully agree with the last idea: ESVs are interchangeable units that allow comparisons between datasets and can avoid generating too big datasets

when combining reads of, say, temporally repeated biomonitoring studies. But clustering ESVs into MOTUs comes as a bonus, provided the grouped sequences are kept and not collapsed under a representative sequence, thus being available for future reanalyses.

The denoising and clustering methods here tested have been developed for ribosomal markers and uncritically applied to COI data in the past, with default parameter values often taken at face value (in fact, parameters are rarely mentioned in methods sections). We confirm that the UNOISE3 parameter $\alpha = 5$ is adequate for COI data, in agreement with previous research using three independent approaches [12, 39, 40]. We also tested and confirmed the suitability of a d value of 13 for SWARM that has been used in previous works with COI datasets (e.g., [43–47]). As Mahé et al. [42] noted, higher d values can be necessary for fast evolving markers. They advised to track MOTU coalescing events as d increases to find the value best-fitting the sequence marker chosen. We have followed this approach, together with the course of the intra- and inter-MOTU distances, to select the d-value for the COI marker. In our view, fixed-threshold clustering procedures should be avoided, as even for a given marker the intra- and interspecies distances can vary according to the group of organisms considered. With SWARM, even if the initial clusters were made at $d = 13$ (for a fragment of 313 this means an initial threshold of 4.15% for connecting sequences), after the refining procedure the mean intra-MOTU distances obtained was 2.91%, which is in line with values suggested using the whole barcoding region of COI [59]. Furthermore, in our taxonomic benchmarking, we found a high proportion of closed MOTUs, irrespective of the denoising method used, indicating that the SWARM procedure adequately and robustly grouped the sequences with known species-level assignments.

Our preferred algorithm for denoising is UNOISE3. It is a one-pass algorithm based on a simple formula with few parameters, it is computationally fast and can be applied at different steps of the pipelines. It keeps almost double ESVs than DADA2 and, combined with a clustering step, results in less single-sequence MOTUs and a higher number of ESVs per MOTU, thus capturing a higher intra-MOTU diversity. It also produced 60% more closed group MOTUs than DADA2 in our taxonomic benchmarking. Edgar et al. [30], by comparing both algorithms in mock and in vivo datasets, also found that UNOISE had comparable or better accuracy than DADA2. Similarly, Tsuji et al. [41] found that UNOISE3 retained less false haplotypes than DADA2 in samples from tank water containing fish DNA. We also found that the entropy values of the sequences changed as expected when denoising becomes more stringent with UNOISE3, indicating that the algorithm performs well with coding sequences. We also suggest ways of improving this algorithm (see below).

DADA2, on the other hand, is being increasingly used in metabarcoding studies but its suitability for a coding gene such as COI remains to be demonstrated. We had to use paired reads (against recommendation) to be able to make meaningful comparisons, but our results indicate that with unpaired sequences the number of ESVs retained would have been even lower. The DADA2 algorithm, when tested with increasingly stringent parameters, did not progressively reduce the entropy ratio values that should reflect an adequate denoising of coding sequences. Further, the high correlation of error rates between all possible substitution types suggests that the algorithm may be over-parameterized, at least for COI, which comes at a computational cost. Comparisons based on

known communities (as in [41]) and using COI are needed to definitely settle the appropriateness of the two algorithms for metabarcoding with this marker.

In addition, PCR-free methods now popular in library preparation procedures complicate the use of DADA2 as there is no consistent direction (forward or reverse) of the reads. We acknowledge that our paired sequences still included a mixture of reads that were originally in one or another direction and, thus, with different error rates. However, the non-overlapped part is only the initial ca. 100 bp, and these are in general good quality positions in both the forward and reverse reads.

Another choice to make is to decide what should come first, denoising or clustering. Both options have been adopted in previous studies (note that clustering first is not possible with DADA2 unless paired sequences are used). Turon et al. [12] advocated that denoising should be made within MOTUs, as they provide the natural "sequence environment" where errors occur and where they should be targeted by the cleaning procedure. We found that clustering first retained more ESVs, because sequences that would otherwise be merged with another from outside its MOTU were preserved. It also resulted in less single-ESV MOTUs, retaining more intra-MOTU variability. It can also be mentioned that denoising the original sequences took approximately 10 times more computing time than denoising within clusters, which can be an issue depending on the dataset and the available computer facilities. We acknowledge, however, that most MOTUs are shared and most ESVs and reads are in the shared MOTUs when comparing the two possible orderings, irrespective of denoising algorithm. The final decision may come more from the nature and goals of each study. For instance, a punctual research may go for clustering first and denoising within clusters to maximize the intra-MOTU variability obtained. A long-term research that implies multiple samplings over time that need to be combined together may use denoising first and then perform the clustering procedure at each reporting period with the ESVs obtained in the datasets collected so far pooled.

There are other important steps at which errors can be reduced and that require key choices, but they are outside the scope of this work as we addressed only clustering and denoising steps. In particular, nuclear insertions (numts) may be difficult to distinguish from true mitochondrial sequences [50, 60]. Singletons (sequences with only one read) are also a problem for all denoising algorithms (as it is difficult to discern rare sequences from errors). Singletons are often eliminated right at the initial steps, as we did in this work. Likewise, a filtering step, in which ESVs with less than a certain amount of reads are eliminated, is deemed necessary to obtain biologically reliable datasets. A 5% relative abundance cut-off value was suggested by [39], while [12] proposed an absolute threshold of 20 reads. However, the procedure and the adequate threshold are best adjusted according to the marker and the study system, so, albeit we acknowledge that a filtering step is necessary, this has not been addressed in this paper.

We recommend that the different denoising algorithms be programmed as standalone steps (not combined, for instance, with chimera filtering) so anyone interested could combine the denoising step with the preferred choices for other steps. We also favour open source programs that could be customized if needed. For UNOISE3 algorithm we suggest that a combination between distance and skew ratio be considered to assign a read to the most likely centroid. This had little effect in our case, but can

be significant in other datasets. For DADA2 algorithm, we advise to weight the gain of considering the two reads separately vs using paired sequences. The advantages of the latter involve a higher flexibility of the algorithm as it does not need to be performed right at the beginning of the pipeline. For both algorithms, we think it is important to consider the natural variation of the three positions of the codons of a coding sequence such as COI, which can allow a more meaningful computation of distances between sequences and error rates. This of course applies to other denoising algorithms not tested in the present study (e.g., AmpliCI [27], deblur [61]). Our DnoiSE program, based on the UNOISE3 algorithm, includes the option of incorporating codon information in the denoising procedure. With this option, we found ca 50,000 more ESVs than with the standard approach. Importantly, this fact did not increase the proportion of erroneous sequences, as determined using aminoacid substitution patterns in metazoan MOTUs. Rather, this proportion was cut by one-half, and erroneous sequences were less even in absolute numbers. In our taxonomic benchmarking, a higher proportion of ESVs with species-level matches in the reference database were detected with the codon position-corrected method. We used a dataset of fixed sequence length and eliminated misaligned sequences. The correction for codon position would be more complicated in the presence of indels and dubious alignments. We also acknowledge the lack of a mock community to ground truth our method, but we contend that mock communities are hardly representative of highly complex communities such as those here analysed. We hope our approach will be explored further and adequately benchmarked in future studies on different communities.

Conclusions

COI has a naturally high intraspecies variability that should be assessed and reported in metabarcoding studies, as it is a source of highly valuable information. Denoising and clustering of sequences are not alternatives. Rather, they are complementary and both should be used together to make the most of the inter- and intraspecies information contained in COI metabarcoding datasets. We emphasize the need to carefully choose the stringency parameters of the different steps according to the variability of this marker.

Our results indicated that the UNOISE3 algorithm preserved a higher intra-cluster variability than DADA2. We introduce the program DnoisE to implement the UNOISE3 algorithm considering the natural variability (measured as entropy) of each codon position in protein-coding genes. This correction increased the number of sequences retained by 88%. The order of the steps (denoising and clustering) had little influence on the final outcome.

We provide recommendations for the preferred algorithms of denoising and clustering, as well as step order, but these may be tuned according to the goals of each study, feasibility of preliminary tests, and ground-truthing options, if any. Other important steps of metabarcoding pipelines, such as abundance filtering, have not been addressed in this study and should be adjusted according to the marker and the study system.

We advise to report the results in terms of both MOTUs and ESVs included in each MOTU, rather than reporting only MOTU tables with collapsed information and just a representative sequence. We also advise that the coding properties of COI should be

used both to set the right parameters of the programs and to guide error estimation in denoising procedures. We wanted to spark further studies on the topic, and our procedures should be tested and validated or refined in different types of community.

There is a huge amount of intra- and inter-MOTU information in metabarcoding datasets that can be exploited for basic (e.g., biodiversity assessment, connectivity estimates, metaphylogeography) and applied (e.g., management) issues in biomonitoring programs, provided the results are reported adequately.

Abbreviations
ASV: amplicon sequence variant; COI: cytochrome *c* oxidase subunit 1; ESV: exact sequence variant; MOTU: molecular operational taxonomic unit; OTU: operational taxonomic unit; ZOTU: zero-radius operational taxonomic unit.

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04115-6.

Additional file 1 Format: .pdf. Details of the dataset used in the analysis of the article, including information of the sampling localities in the Iberian Peninsula and the sample processing steps prior to sequencing.

Additional file 2 Format: .pdf. Comparison of DADA2 on paired and unpaired reads.

Additional file 3 Format: .pdf. Details of the taxonomic benchmarking.

Additional file 4 Format: .csv. Species-level dataset. Table with the ESVs identified at the species level with >97% similarity. The taxonomy assigned is indicated, as well as the best-match in the reference database, the taxid, and the sequence.

Acknowledgements
We thank Daniel Sanromán for his help with the field sampling.

Authors' contributions
XT and OSW conceived the study; XT, AA and CP performed the sampling; AA performed the laboratory work; AA and OSW wrote code; XT, CP, AA and OSW performed analysis; XT and AA drafted the ms; all authors contributed intellectual content and revised the ms.

Funding
This study has been funded by Project PopCOmics (CTM2017-88080, MCIU/AEI/FEDER/UE) and project BigPark (Autonomous Organism of National Parks, OAPN 2462/2017) from the Spanish Government.

Availability of data and materials
All the datasets supporting the conclusions of this article are available in the Mendeley Data repository in https://data.mendeley.com/datasets/84zypvmn2b/ and in Additional File 4. The Phyton program DnoisE is available in the Github repository in https://github.com/adriantich/DnoisE.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Department of Marine Ecology, Centre for Advanced Studies of Blanes (CEAB-CSIC), Blanes (Girona), Catalonia, Spain. [2] Department of Evolutionary Biology, Ecology and Environmental Sciences, University of Barcelona and Research Institute of Biodiversity (IRBIO), Barcelona, Catalonia, Spain. [3] Norwegian College of Fishery Science, UiT The Arctic University of Norway, Tromsö, Norway.

Received: 9 January 2021 Accepted: 30 March 2021
Published online: 05 April 2021

References

1. Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, de Vere N, Pfrender ME, Bernatchez L. Environmental DNA metabarcoding: transforming how we survey animal and plant communities. Mol Ecol. 2017;26:5872–95.
2. Aylagas E, Borja A, Muxika I, Rodríguez-Ezpeleta N. Adapting metabarcoding-based benthic biomonitoring into routine ecological status assessment networks. Ecol Ind. 2018;95:194–202.
3. Bani A, De Brauwer M, Creer S, Dumbrell AJ, Limmon G, Jompa J, von der Heyden S, Beger M. Informing marine spatial planning decisions with environmental DNA. Adv Ecol Res. 2020;62:375–407.
4. Compson ZG, McClenaghan B, Singer GAC, Fahner N, Hajibabaei M. Metabarcoding from microbes to mammals: comprehensive bioassessmenton a global scale. Front Ecol Evol. 2020;8:581835.
5. Mathieu C, Hermans SM, Lear G, Buckley TR, Lee KC, Buckley HL. A systematic review of sources of variability and uncertainty in eDNA data for environmental monitoring. Front Ecol Evol. 2020;8:135.
6. Rodríguez-Ezpeleta N, Morisette O, Bean CW, Manu S, Banerjee P, Lacoursière-Roussel A, Beng KC, Alter SE, Roger F, Holman LE, Stewart KA, Monaghan MT, Mauvisseau Q, Mirimin L, Wangensteen OS, Antognazza CM, Helyar SJ, de Boer H, Monchamp ME, Nijland R, Abbott CL, Doi H, Barnes MA, Leray M, Hablützel PI, Deiner K. Trade-offs between reducing complex terminology and producing accurate interpretations from environmental DNA: comment on 'Environmental DNA: What's behind the term?' by Pawlowski et al. (2020). EcoEvoRxiv. 2020. https://doi.org/10.32942/OSF.IO/KGNYD.
7. Porter TM, Hajibabaei M. Putting COI metabarcoding in context: the utility of exact sequence variants (ESV) in biodiversity analysis. Front Ecol Evol. 2020;8:248.
8. Macheriotou L, Guilini K, Bezerra TN, Tytgat B, Nguyen DT, Nguyen TXP, Noppe F, Armenteros M, Boufahja F, Rigaux A, Vanreusel A, Derycke S. Metabarcoding free-living marine nematodes using curated 18S and CO1 reference sequence databases for species-level taxonomic assignments. Ecol Evol. 2019;9:1211–26.
9. Forster D, Lentendu G, Filker S, Dubois E, Wilding TA, Stoeck T. Improving eDNA-based protist diversity assessments using networks of amplicon sequence variants. Environ Microbiol. 2019;21(11):4109–24.
10. O'Rourke DR, Bokulich NA, Jusino MA, MacManes MD, Foster JT. A total crapshoot? Evaluating bioinformatic decisions in animal diet metabarcoding analyses. Ecol Evol. 2020;10:9721–9.
11. Giebner H, Langen K, Bourlat SJ, Kukowka S, Mayer C, Astrin JJ, Misof B, Fonseca VG. Comparing diversity levels in environmental samples: DNA sequence capture and metabarcoding approaches using 18S and COI genes. Mol Ecol Resour. 2020;20:1333–45.
12. Turon X, Antich A, Palacín C, Praebel K, Wangensteen OS. From metabarcoding to metaphylogeography: separating the wheat from the chaff. Ecol Appl. 2020;30:e02036.
13. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 2017;11:2639–43.
14. Andujar C, Arribas P, Yu DW, Vogler AP, Emerson BC. Why the COI barcode should be the community DNA metabarcode for the Metazoa. Mol Ecol. 2018;27:3968–75.
15. van der Loos LM, Nijland R. Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. Mol Ecol. 2020. https://doi.org/10.1111/MEC.15592.
16. Tapolczai K, Keck F, Bouchez A, Rimet F, Kahlert M, Vasselon V. Diatom DNA metabarcoding for biomonitoring: strategies to avoid major taxonomical and bioinformatical biases limiting molecular indices capacities. Front Ecol Evol. 2019;7:409.
17. Holman LE, de Bruyn M, Creer S, Carvalho G, Robidart J, Rius M. Consistent marine biogeographic boundaries across the tree of life despite centuries of human impacts. bioRxiv. 2020. https://doi.org/10.1101/2020.06.24.169110.
18. Steyaert M, Priestley V, Osborne O, Herraiz A, Arnold R, Savolainen O. Advances in metabarcoding techniques bring us closer to reliable monitoring of the marine benthos. J Appl Ecol. 2020;57:2234–45.
19. Zamora-Terol S, Novotny A, Winder M. Reconstructing marine plankton food web interactions using DNA metabarcoding. Mol Ecol. 2020;29:3380–95.
20. Pearman JK, Chust G, Aylagas E, Villarino E, Watson JR, Chenuil A, Borja A, Cahill AE, Carugati L, Danovaro R, David R, Irigoien X, Mendibil I, Moncheva S, Rodríguez-Ezpeleta N, Uyarra MC, Carvalho S. Pan-regional marine benthic cryptobiome biodiversity patterns revealed by metabarcoding autonomous reef monitoring structures. Mol Ecol. 2020;29:4882–97.
21. Brandt MI, Trouche B, Quintric L, Wincker P, Poulain J, Arnaud-Haond S. A flexible pipeline combining bioinformatic correction tools for prokaryotic and eukaryotic metabarcoding. bioRxiv. 2020. https://doi.org/10.1101/717355.
22. Nguyen BN, Shen EW, Seemann J, Correa AMS, O'Donnell JL, Altieri AH, Knowlton N, Crandall KA, Egan SP, McMillan WO, Leray M. Environmental DNA survey captures patterns of fish and invertebrate diversity across a tropical seascape. Sci Rep. 2020;10:6729.
23. Laroche O, Kersten O, Smith CR, Goetze E. Environmental DNA surveys detect distinct metazoan communities across abyssal plains and seamounts in the western Clarion Clipperton Zone. Mol Ecol. 2020;29:4588–604.
24. Zizka VMA, Weiss M, Leese F. Can metabarcoding resolve intraspecific genetic diversity changes to environmental stressors? A test case using river macrozoobenthos. Metabarcoding Metagenom. 2020;4:23–34.
25. Avise JC. Phylogeography: retrospect and prospect. J Biogeogr. 2009;36:3–15.
26. Emerson BC, Cicconardi F, Fanciulli PP, Shaw PJA. Phylogeny, phylogeography, phylobetadiversity and the molecular analysis of biological communities. Philos Trans R Soc B. 2011;366:2391–402.
27. Peng X, Dorman K. AmpliCI: A high-resolution model-based approach for denoising Illumina Amplicon data. Bioinformatics. 2020. https://doi.org/10.1093/bioinformatics/btaa648.
28. Kopylova E, Navas-Molina JA, Mercier C, Xu ZZ, Mahé F, He Y, Zhou HW, Rognes T, Caporaso JG, Knight R. Open-source sequence clustering methods improve the state of the art. mSystems. 2020;1(1):e00003–15.
29. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13(7):581–3.
30. Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS aplicon sequencing. bioRxiv. 2016. https://doi.org/10.1101/081257.

31. Edgar RC. UPARSE: hihgly accurate OTU sequences from microbial amplicon reads. Nat Methods. 2013;10(10):996–1000.
32. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:e2584.
33. Hao X, Jiang R, Chen T. Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. Bioinformatics. 2011;27(5):611–8.
34. Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ. 2015;3:e1420.
35. Boyer F, Mercier C, Bonin A, Le Bras Y, Taberlet P, Coissac E. OBITOOLS: a UNIX-inspired software package for DNA metabarcoding. Mol Ecol Resour. 2016;16:176–82.
36. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;2010(26):2460–1.
37. Antich A. DnoisE, Distance denoise by Entropy. GitHub repository. https://github.com/adriantich/DnoisE. Accessed 20 November 2020.
38. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu YX, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS 2nd, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019;37:852–7.
39. Elbrecht V, Vamos EE, Steinke D, Leese F. Estimating intraspecific genetic diversity from community DNA metabarcoding data. PeerJ. 2018;6:e4644.
40. Shum P, Palumbi SR. Testing small-scale ecological gradients and intraspecific differentiation from hundreds of kelp forest species using haplotypes from metabarcoding. Mol Ecol. 2021. https://doi.org/10.1111/mec.15851.
41. Tsuji S, Miya M, Ushio M, Sato H, Minamoto T, Yamanaka H. Evaluating intraspecific genetic diversity using environmental DNA and denoising approach: a case study using tank water. Environ DNA. 2020;2:42–52.
42. Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. Swarm: robust and fast clustering method for amplicon-based studies. PeerJ. 2014;2:e593.
43. Siegenthaler A, Wangensteen OS, Soto AZ, Benvenuto C, Corrigan L, Mariani S. Metabarcoding of shrimp stomach content: Harnessing a natural sampler for fish biodiversity monitoring. Mol Ecol Resour. 2019;19:206–20.
44. Garcés-Pastor S, Wangensteen OS, Pérez-Haase A, Pèlachs A, Pérez-Obiol R, Cañellas-Boltà N, Mariani S, Vegas-Vilarrúbia T. DNA metabarcoding reveals modern and past eukaryotic communities in a high-mountain peat bog system. J Paleolimnol. 2019;62:425–41.
45. Bakker J, Wangensteen OS, Baillie C, Buddo D, Chapman DD, Gallagher AJ, Guttridge TL, Hertler H, Mariani S. Biodiversity assessment of tropical shelf eukaryotic communities via pelagic eDNA metabarcoding. Ecol Evol. 2019;9:14341–55.
46. Atienza S, Guardiola M, Praebel K, Antich A, Turon X, Wangensteen OS. DNA metabarcoding of deep-sea sediment communities using COI: community assessment, spatio-temporal patterns and comparison with 18S rDNA. Diversity. 2020;12:123.
47. Antich A, Palacin C, Cebrian E, Golo R, Wangensteen OS, Turon X. Marine biomonitoring with eDNA: Can metabarcoding of water samples cut it as a tool for surveying benthic communities? Mol Ecol. 2021. https://doi.org/10.1111/mec.15641.
48. Schmidt AO, Herzel H. Estimating the entropy of DNA sequences. J Theor Biol. 1997;3:369–77.
49. Hausser J, Strimmer K. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. J Mach Learn Res. 2009;10:1469–84.
50. Andújar C, Creedy TJ, Arribas P, López H, Salces-Castellano A, Pérez-Delgado AJ, Vogler AP, Emerson BC. Validated removal of nuclear pseudogenes and sequencing artefacts from mitochondrial metabarcode data. Mol Ecol Resour. 2021. https://doi.org/10.1111/1755-0998.13337.
51. Wangensteen OS, Palacin C, Guardiola M, Turon X. DNA metabarcoding of littoral hard-bottom communities: high diversity and database gaps revealed by two molecular markers. Peer J. 2013;6:e4705.
52. Wangensteen OS. Reference-databases Metabarpark. GitHub repository. http://github.com/metabarpark/Reference-databases. Accessed 23 December 2020.
53. Pentinsaari M, Salmela H, Mutanen M, Roslin T. Molecular evolution of a widely-adopted taxonomic marker (COI) across the animal tree of life. Sci Rep. 2016;6:35275.
54. Pagès H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: Efficient manipulation of biological strings. R package version 2.58.0. https://bioconductor.org/packages/Biostrings. Accessed 10 March 2021.
55. Antich A, Palacin C, Wangensteen OS, Turon X. Dataset for "To denoise or to cluster? That is not the question. Optimizing pipelines for COI metabarcoding and metaphylogeography". Mendeley Data. 2021. https://data.mendeley.com/datasets/84zypvmn2b/.
56. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B. 1995;55(1):289–300.
57. Nearing JT, Douglas GM, Comeau AM, Langille MGI. Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. PeerJ. 2018;6:e5364.
58. Estensmo EL, Maurice S, Morgado L, Martin-Sanchez P, Skrede I, Kauserud H. The influence of intraspecific sequence variation during DNA metabarcoding: a case study of eleven fungal species. Authorea. 2020. https://doi.org/10.22541/au.160071155.58915559.

59. Ratnasingham S, Hebert PDN. A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. PLoS ONE. 2013;8(8):6.
60. Porter TM, Hajibabaei M. Profile hidden Markov model sequence analysis can help remove putative pseudogenes from DNA barcoding and metabarcoding datasets. bioRxiv. 2021. https://doi.org/10.1101/2021.01.24.427982.
61. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Xu ZZ, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R. Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems. 2017;2(2):e00191-16.

Publisher's Note

DnoisE: distance denoising by entropy. An open-source parallelizable alternative for denoising sequence datasets

Adrià Antich[1], Creu Palacín[2], Xavier Turon[1] and Owen S. Wangensteen[3]

[1] Department of Marine Ecology, Centre for Advanced Studies of Blanes (CEAB- CSIC), Blanes (Girona), Catalonia, Spain

[2] Department of Evolutionary Biology, Ecology and Environmental Sciences and Biodiversity Research Institute (IRBIO), University of Barcelona, Barcelona, Catalonia, Spain

[3] Norwegian School of Fishery Science, UiT The Arctic University of Norway, Tromsø, Troms og Finnmark, Norway

ABSTRACT

DNA metabarcoding is broadly used in biodiversity studies encompassing a wide range of organisms. Erroneous amplicons, generated during amplification and sequencing procedures, constitute one of the major sources of concern for the interpretation of metabarcoding results. Several denoising programs have been implemented to detect and eliminate these errors. However, almost all denoising software currently available has been designed to process non-coding ribosomal sequences, most notably prokaryotic 16S rDNA. The growing number of metabarcoding studies using coding markers such as COI or RuBisCO demands a re-assessment and calibration of denoising algorithms. Here we present DnoisE, the first denoising program designed to detect erroneous reads and merge them with the correct ones using information from the natural variability (entropy) associated to each codon position in coding barcodes. We have developed an open-source software using a modified version of the UNOISE algorithm. DnoisE implements different merging procedures as options, and can incorporate codon entropy information either retrieved from the data or supplied by the user. In addition, the algorithm of DnoisE is parallelizable, greatly reducing runtimes on computer clusters. Our program also allows different input file formats, so it can be readily incorporated into existing metabarcoding pipelines.

Subjects Biodiversity, Bioinformatics, Ecology, Marine Biology

Keywords Metabarcoding, Bioinformatic pipelines, Metaphylogeography, Entropy correction, Denoising algorithms, Coding markers

Submitted 13 July 2021
Accepted 16 December 2021
Published 19 January 2022

Corresponding author
Owen S. Wangensteen,
owen.wangensteen@uit.no

Academic editor
Joseph Gillespie

Additional Information and Declarations can be found on page 13

DOI **10.7717/peerj.12758**

BACKGROUND

Biodiversity studies have experienced a revolution in the last decade with the application of high throughput sequencing (HTS) techniques. In particular, the use of metabarcoding in ecological studies has increased notably in recent years. For both prokaryotic and eukaryotic organisms, a large number of applications have been developed, ranging from biodiversity assessment (*Wangensteen et al., 2018*), detection of particular species (*Kelly et al., 2014*), analysis of impacts (*Pawlowski et al., 2018*), and diet studies (*Clarke et al., 2020*; *Sousa, Silva & Xavier, 2019*), among others. Also, different sample types have been used: terrestrial soil,

How to cite this article Antich Aà, Palacín C, Turon X, Wangensteen OS. 2022. DnoisE: distance denoising by entropy. An open-source parallelizable alternative for denoising sequence datasets. *PeerJ* **10**:e12758 http://doi.org/10.7717/peerj.12758

freshwater, marine water, benthic samples, arthropod traps, or animal faeces (*Creer et al., 2016*; *Deiner et al., 2017*). Many of these studies have direct implications on management and conservation of ecosystems and are thus providing direct benefits to society. They have also brought to light a bewildering diversity of organisms in habitats difficult to study with traditional techniques.

Metabarcoding studies have greatly contributed to so-called big community data (*Pichler & Hartig, 2020*) by generating an enormous amount of sequence data that, in most cases, is available online. Handling these datasets is memory intensive and filtering steps are required to analyze such information. Clustering and denoising are the two main strategies to compress data into Molecular Operational Taxonomic Units (MOTUs, aka OTUs) or Exact Sequence Variants (ESVs; also ASVs, Amplicon Sequence Variants, or ZOTUs, zero ratio OTUs) to extract biodiversity composition (*Antich et al., 2021*). Both methods rely on minimizing sequencing and PCR errors either by clustering sequences into purportedly meaningful biological entities (MOTUs) or by merging erroneous sequences with the correct ones from which they possibly originated, and keeping just correct amplicons (ESVs). Hence, both methods differ philosophically and analytically. Furthermore, they are not incompatible and can be jointly applied. Software development is crucial to create tools capable of performing these tasks in a fast and efficient way. The type of samples, the marker, and the target organisms are also instrumental in choosing the adequate bioinformatic pipelines to provide interpretable results.

Recent studies have explored the joint application of both methods to filter metabarcoding data (*Antich et al., 2021*; *Brandt et al., 2021*; *Elbrecht et al., 2018*; *Turon et al., 2020*). Importantly, the combination of clustering and denoising opens the door to the analysis of intraspecies (intra-MOTU) variability (*Antich et al., 2021*). *Turon et al. (2020)* proposed the term metaphylogeography for the study of population genetics using metabarcoding data, and *Zizka, Weiss & Leese (2020)* found different haplotype composition between perturbed and unperturbed rivers, both studies using a combination of clustering and denoising steps.

The software presented here focuses on the denoising step. There are currently several software programs developed to denoise sequencing and PCR errors, such as DADA2 (*Callahan et al., 2016*), AmpliCL (*Peng & Dorman, 2020*), Deblur (*Amir et al., 2017*), or UNOISE (*Edgar, 2016*). These programs have been widely used in metabarcoding studies to generate ESVs, using sequence quality information for the first two and simple analytical methods for the latter two. All were originally tested for ribosomal DNA (non-coding) and thus some adjustment is necessary for application to other markers (*Antich et al., 2021*).

Here we present DnoisE, a parallelizable Python3 software for denoising sequences using a modification of the UNOISE algorithm and tested for metabarcoding of eukaryote communities using mitochondrial markers (COI, Cytochrome Oxidase subunit I). We introduce a novel correction procedure for coding sequences using changes in diversity values per codon position. In coding genes, the natural entropy of the different positions is markedly different, with the third position being always the most variable. We therefore contend that differences in each position should have different weights when deciding whether a change in a given position is legitimate or is attributable to random PCR or

sequencing errors. DnoisE is also applicable to other markers due to the settable options and offers a fast and open source alternative to non-parallelizable closed source programs. Scripts for installation and example files to run DnoisE are provided in the GitHub repository: https://github.com/adriantich/DnoisE.

WORKFLOW

Structure of input files

DnoisE is designed to run with HTS datasets (after paired-end merging and de-replicating sequences) to obtain ESVs, or after clustering with SWARM (*Mahé et al., 2015*) to obtain haplotypes within MOTUs. Due to variability in format files, we have designed an algorithm that can read both fasta and csv files. In the present version, however, sample information (if present) is kept only for csv input.

Combining the UNOISE algorithm and the entropy correction

Sequences are stored as a data frame, with each row corresponding to a sequence record and the columns to the abundances (either total or per sample). The original Edgar's (2016) function used by UNOISE to determine whether two sequences should be merged is:

$$\beta(d) = 0.5^{\alpha \cdot d+1}$$

where β(d) is the threshold abundance ratio of a less abundant sequence with respect to a more abundant one (from which it differs by distance d) below which they are merged. The distance d is the Levenshtein genetic distance measured in DnoisE with the Levenshtein module (https://maxbachmann.github.io/Levenshtein/) and α is the stringency parameter (the higher α, the lower the abundance skew required for merging two sequences).

The UNOISE algorithm sorts sequences by decreasing abundance and each one is compared with the less abundant ones. At each comparison, the distance between sequences (d) is computed and, if the abundance ratio between the less abundant and the more abundant sequence is lower than β(d), the former is assumed to be an error. In UNOISE terminology, the sequences form clusters, of which the correct one is the centroid and the remaining members are inferred to derive from the centroid template but contain errors. In his original paper, *Edgar (2016)* suggests constructing a table of centroids excluding low abundance reads, and then constructing a ZOTU table by mapping all reads (before the abundance filtering) to the centroids table using the same merging criterion but without creating new centroids. So, the original formulation of this algorithm gives priority to the abundance ratio over the genetic distance. The first, very abundant, sequences will "capture" rare sequences even if d is relatively high. Other, less abundant sequences may be closer (lower d) and still fulfill Edgar's formula for merging the rare sequence, but this will never happen as the rare sequence will be joined with the very abundant one and will not be available for further comparisons. However, in the standard procedure of this algorithm implemented as UNOISE3 in the USEARCH pipeline (*Edgar, 2010*; https://drive5.com/usearch/), the reads are mapped to the centroid table using a similarity criterion (identity threshold in the otutab command), so in practice a distance criterion is used during the mapping.

DnoisE is a one pass algorithm, with no posterior mapping of reads to centroids (which is indeed repetitive, as reads have already been evaluated against the centroids when constructing the centroid table) and with a choice of merging criteria. If deemed necessary, low abundance reads can be eliminated previously or, alternatively, ESVs with one or a few reads can be discarded after denoising. Chimeric amplicons can likewise be eliminated before or after denoising. DnoisE follows previously used terminology (*Turon et al., 2020*; *Antich et al., 2021*) in which the correct sequences (centroids in UNOISE terms) are called "mother" sequences and the erroneous sequences derived from them are labelled "daughter" sequences. DnoisE provides different options for merging the sequences. Let PMS (potential "mother" sequence) and PDS (potential "daughter" sequence) denote the more abundant and the less abundant sequences that are being compared, respectively, and let d be the genetic distance between them. When the abundance ratio PDS/PMS is lower than $\beta(\mathrm{d})$, the PDS is tagged as an error sequence but is not merged with the PMS. Instead, a round with all comparisons is performed and, for a given PDS, all PMS fulfilling the UNOISE criterion for merging are stored. After this round is completed, the merging is performed following one of three possible criteria: (1) Ratio criterion, joining a PDS to its more abundant PMS (lowest abundance ratio, corresponding to the original UNOISE formulation); (2) Distance criterion, joining a sequence to the closest (least d value) possible "mother"; and the (3) Ratio-Distance criterion, whereby a PDS is merged with the PMS for which the quotient $\beta/\beta(\mathrm{d})$ (*i.e.,* between the abundance ratio PDS/PMS and the maximal abundance ratio allowed for the observed d), is lowest, thus combining the two previous criteria. For each criterion, the best PMS and the corresponding values (ratio, d and ratio skew values) are stored. The user then has the choice to select one or another for merging sequences. As an option, if the user wants to apply only the Ratio criterion, each PDS is assigned to the first (*i.e.,* the most abundant) PMS that fulfills the merging inequality and becomes unavailable for further comparisons, thus decreasing computing time. Figure 1 shows a conceptual scheme of this workflow process.

In addition, for coding markers such as COI, the codon position provides crucial additional information that must be taken into account. In nature, the third codon position is the most variable, followed by the first and the second position. This variation can be measured as entropy (*Schmitt & Herzel, 1997*) of the different positions. A change in third position is more likely to be a natural change (and not an error) than the same change in a second position, much less variable naturally. To our knowledge, no denoising algorithm incorporates this important information. We propose to use the entropy values of each codon position to correct the distance d in Edgar's formula as follows:

$$d_{\mathrm{corr}} = \sum_{i=1}^{3} d(i) \cdot \mathrm{entropy}(i).3/\big(\mathrm{entropy}(1)+\mathrm{entropy}(2)+\mathrm{entropy}(3)\big)$$

where i is the codon position and d is the number of differences in each position. The d_{corr} value is then used instead of d in the formula. This correction results in a higher d_{corr} when a change occurs in a third position than in the first or second position, thus a sequence with changes in third positions will be less likely to be merged. In practice, as many changes occur naturally in third positions, this correction will lead to a higher number of ESVs retained that would otherwise be considered errors. Careful choice of entropy values is crucial, and it is recommended that they are adjusted for each marker and particular study.

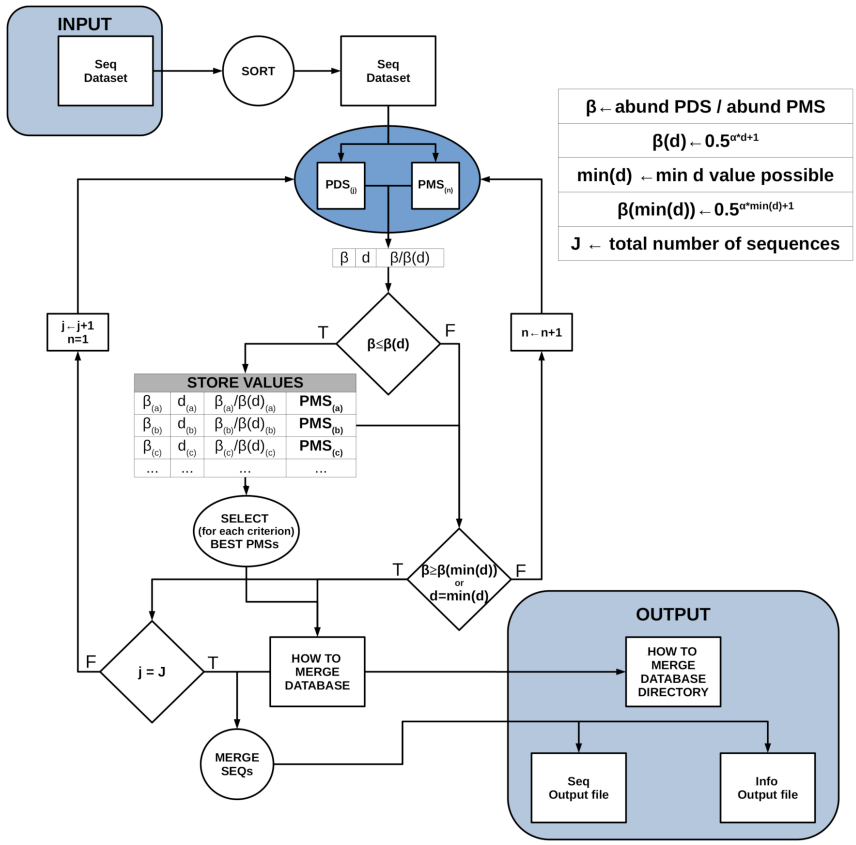


Figure 1 Scheme of the workflow of DnoisE. Starting from an abundance-sorted sequence dataset, subsets of possible daughter sequences (PDS) and possible mother sequences (PMS) are selected as detailed in Fig. 2. For each subset, all PDS are compared with all compatible PMS (in terms of MDA and MMA). If the merging inequality is met, the values of the main parameters are stored. After all subsets have been evaluated, for each merging criterion the best PMS for each PDS is chosen and a sequence file is generated, together with a file with information on the merging process.

Full-size DOI: 10.7717/peerj.12758/fig-1

The values of entropy for each position can be obtained from the data (computed directly by the program) or added manually by the user.

Note that, when applying this correction, the Levenshtein distance is not used as it cannot consider codon positions. Instead, the number of differences is used. In practice, in aligned sequences with no indels both distances are equivalent. In addition, with the entropy correction, lengths should be equal when comparing two sequences. The dataset is thus analysed separately by sequence length sets. These sets must differ from the modal length (the modal sequence length can also be set using the -m parameter) of the complete dataset by n number of codons (groups of three nucleotides), as in general indels in coding

sequences are additions or deletions of whole codons. A sequence differing from these accepted lengths is considered erroneous and removed. Sequences of the same length must be aligned for the algorithm to run properly.

Parallel processing

Parallel processing is a useful tool to increase speed when multicore computers are available. DnoisE implements parallel processing in the algorithm so the required time to run huge datasets decreases drastically as more cores are used. Parallel processing was applied using the multiprocessing module of Python3 (*McKerns et al., 2011*). A computational bottleneck of denoising procedures is their sequential nature, which is hardly parallelizable, and more so in the case of DnoisE that computes all comparisons before merging. In particular, a sequence that has been tagged as "daughter" (error) cannot be a "mother" of a less abundant sequence. Therefore, to compare a PDS to all its PMS requires that those more abundant sequences have been identified as correct before.

We incorporate two concepts, based on the highest skew ratio required for a sequence to be merged with a more abundant one. This is of course β(min(d)), where min(d) is one if entropy correction is not performed, and it equals the $d_{cor\,r}$ corresponding to a single change in the position with less entropy (position 2) if entropy is considered. From this maximal abundance ratio we can obtain, for a given potential "mother", the maximal "daughter" abundance (MDA, any sequence more abundant than that cannot be a "daughter" of the former). Conversely, for a given "daughter" sequence we can obtain the minimum "mother" abundance (MMS, any sequence less abundant than that cannot be the "mother" of the former). The formulae are:

$$\mathrm{MDA} = \mathrm{abundancePMS}/\beta(\min(d))$$

$$\mathrm{MMA} = \beta(\min(d))/\mathrm{abundancePDS}$$

$$\beta(\min(d)) = 0.5^{\alpha\cdot 1+1} OR\, \beta(\min(d)) = 0.5^{\alpha\cdot \min\left(\mathrm{entropy}(i)\cdot 3/\left(\mathrm{entropy}(1)+\mathrm{entropy}(2)+\mathrm{entropy}(3)\right)\right)+1}$$

The use of MDA and MMA simplifies the workload of the program as it greatly reduces the number of comparisons (a PMS will not be evaluated against sequences more abundant than the MDA, and a PDS will not be compared with sequences with less abundance than the MMA). Likewise, it allows for a parallel processing of sequences using the MDA as follows:

1- Sequences are ordered by decreasing abundance.

2- The first sequence is automatically tagged as a correct sequence.

3- MDA is calculated for this sequence (MDA_1).

4- All sequences with abundances between the first sequence and the MDA are, by definition, tagged also as correct sequences.

5- For the last sequence tagged as correct, the MDA is calculated (MDA_2).

6- Every sequence with abundance between the last correct sequence and MDA_2 is evaluated in parallel against all correct sequences that are more abundant than its MMA. Those for which no valid "mother" is found are tagged as correct, the rest are "daughter" (error) sequences.

7- Repeat steps 5 and 6 (*i.e.,* calculating MDA_3 to n) until all sequences have been evaluated.

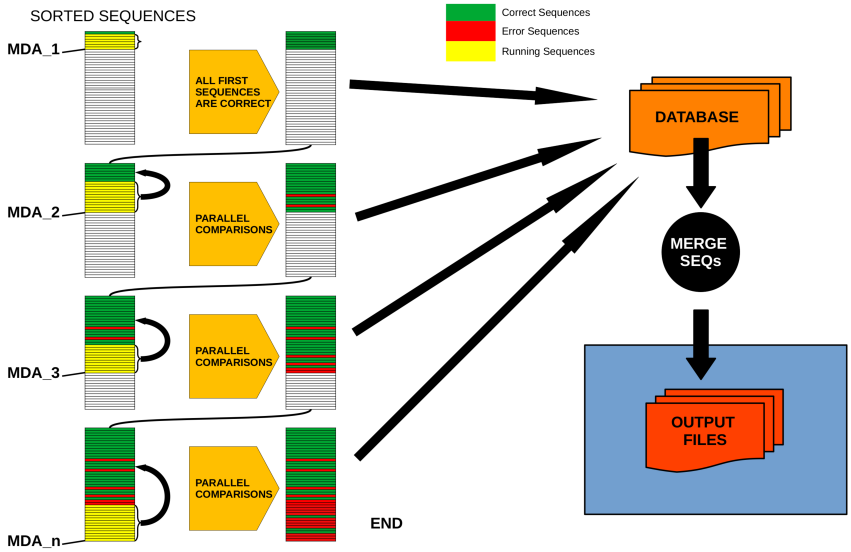


Figure 2 Schematic workflow of parallel processing of DnoisE. When running in parallel, comparisons between sequences are computed in sets of sequences defined by their abundances. Using the Maximum Daughter Abundance (MDA) value, computed from the last correct sequence of the previous step, we can define sets of sequences that are compared in parallel with the previously tagged correct sequences.
Full-size DOI: 10.7717/peerj.12758/fig-2

Figure 2 provides a conceptual scheme of this procedure. Note that, for each block of sequences that is evaluated in parallel, no comparisons need to be performed between them as they will never fulfill the merging inequality. After this process is completed, all sequences not labelled as "daughter" are kept as ESVs, and all "daughters" are merged to them according to the merging criterion chosen.

DNOISE PERFORMANCE

A previous version of DnoisE was tested in *Antich et al. (2021)* on a COI metabarcoding dataset of marine benthic communities. The version used in *Antich et al. (2021)* implemented the same basic algorithm but was not curated for general use. For the present version, we have corrected bugs, made the program user-friendly, and added more settable options and features. The dataset consisted of 330,382 chimera-filtered COI sequences of 313 bp (all sequences had more than one read). They came from benthic marine communities in 12 locations of the Iberian Mediterranean coast (see (*Antich et al., 2021*) for details), and are available as a Mendeley Dataset (https://data.mendeley.com/datasets/84zypvmn2b/). DnoisE was used in *Antich et al. (2021)* in combination with the clustering algorithm SWARM, and was compared with the results of DADA2 denoising algorithm. *Antich et al. (2021)* also compared DnoisE with and without entropy correction, and obtained twice the number of ESVs with correction, while

the proportion of erroneous sequences (defined as those having stop codons or substitutions in conserved positions) decreased to one half as compared with not correcting for codon position variation, as discussed in *Antich et al. (2021)*.

Comparison with UNOISE3

We benchmarked the current version of DnoisE (with alpha = 5) against the current implementation of the UNOISE algorithm: UNOISE3 (USEARCH 32-bit, free version, with alpha = 5 and minsize = 2) on this same dataset. To be able to make a direct comparison, for UNOISE3 we didn't perform an otutab step, rather, we recovered the ESVs and their abundance directly from the output files generated with -tabbedout and -ampout. As chimeric sequences were already removed from the dataset, and for the sake of comparability, we didn't exclude the few sequences flagged as such by the chimera filtering procedure embedded in UNOISE3. The number of ESVs obtained was almost the same: 60,198 and 60,205, respectively, if no entropy correction was performed. In addition, 60,196 ESVs were shared (comprising > 99.999% of the total reads) among the two programs, confirming that DnoisE (without correction) and UNOISE3 were practically equivalent. For further analyses of the effect of entropy correction we will therefore compare DnoisE with and without this correction.

Running performance

We compared the run speed of DnoisE with and without entropy correction for the same dataset of sequences. We used different numbers of cores, from 1 to 59, for parallelization. We applied the entropy correction values from *Antich et al. (2021)*.

Running DnoisE with just one core (without entropy correction) took about 29 h, decreasing sharply when using parallel processing with just a few cores. DnoisE took 4.5 h with 6 cores and 2.78 h with 10 cores. As a reference, the execution time of UNOISE3 (32-bit version, not parallelizable) without the otutab step was ca. 7 h, albeit this execution time is not directly comparable as UNOISE3 has a chimera filtering step embedded. Using entropy correction, run times increased (Fig. 3) as there is a higher number of comparisons needed because the MMA values are generally lower. This slows the process as any given PSD has more PMS to compare with. With entropy correction, DnoisE retrieved ca. twice the number of ESVs, further increasing run time. For the Ratio-Distance merging criterion, when entropy correction was performed, 16 cores were required for DnoisE to run at a similar time speed than 6 cores with no entropy correction (Fig. 3). Above 10 cores (without correction) or 20 cores (with correction), run times reached a plateau and did not further improve, while memory usage continued to increase steadily. A trade-off between both parameters should be sought depending on the cluster architecture and the dataset being run.

Merging performance

Due to the practical impossibility of building a mock community of the complexity required with known COI haplotypes for multiple species, in order to compare the merging performance of the original formula of UNOISE with the entropy correction available in DnoisE, we performed a simulation following the procedure described in

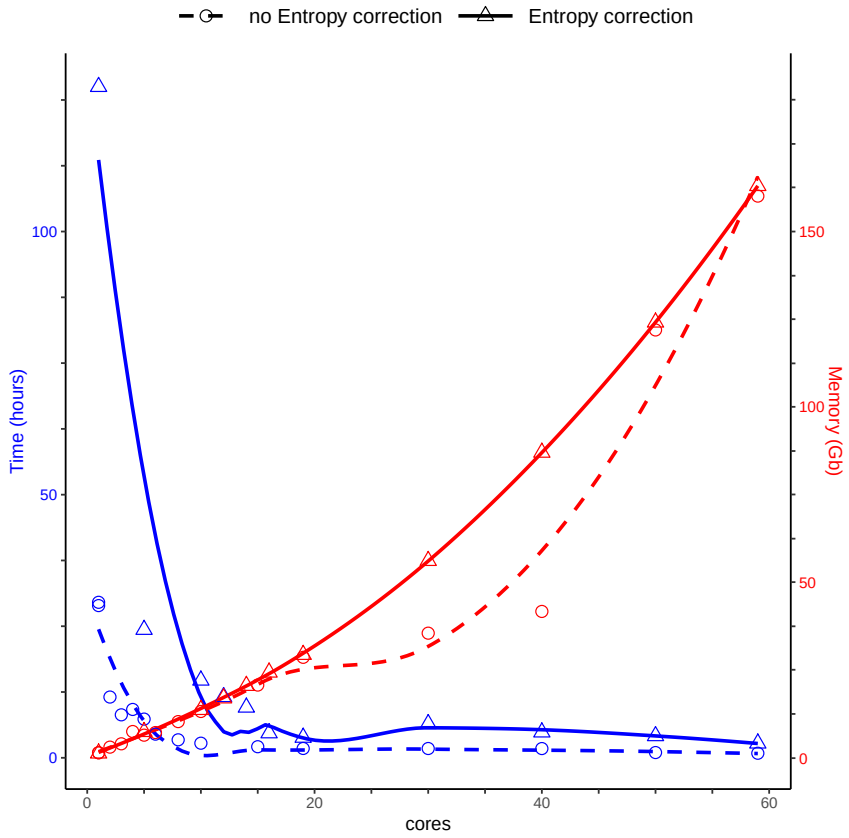


Figure 3 Time (blue) and memory (red) used by DnoisE to denoise and merge sequences with the Ratio-Distance criterion using different cores on a computer cluster. Denoising using entropy correction (triangles and dashed line) is compared against no correction (circles and dashed line). Lines are computed using the geom_smooth() function of the ggplot2 package with method = 'loess'.
Full-size DOI: 10.7717/peerj.12758/fig-3

Turon et al. (2020), and using the same dataset of 1,000 "good" sequences from marine samples used in that study. The rationale was to start with a dataset of good sequences with realistic read abundance distribution, simulate sequencing errors at a given error rate (henceforth "error" amplicons), and then denoise the resulting dataset to recover the original one. In addition, in the present study we kept track of which original sequences produced each error amplicon and used this information to check if error sequences are merged or not with their "true" mother. We applied a random error rate per base of 0.005, which is intermediate among reported values for Illumina platforms (*Pfeiffer et al., 2018*; *Schirmer et al., 2016*). After the simulation, we removed all sequences with only one read. This resulted in a dataset with the 1,000 original sequences and 265,297 error sequences.

We used the DnoisE software with and without entropy correction (the latter equivalent to the UNOISE3 results, see above) to denoise the simulated dataset. The entropy values were automatically computed from the data by the program and we tested alpha values from 10 to 1 (from lowest to highest stringency level). The results showed a decreasing number of total remaining sequences with more stringent (lower) alpha values (Fig. 4). There was also a drop in the number of good sequences remaining as alpha diminished. Except for the less stringent alpha values, however, data denoised with entropy correction kept a higher number of true sequences. With entropy correction, they remained almost constant for alpha values of 5 or higher, and decreased at lower values. Without entropy correction, the number of true sequences started to decrease at alpha values below 8. On the other hand, the entropy correction procedure also retrieved a higher number of false positives (*i.e.,* error sequences) at intermediate alpha values, but the vast majority of them could be removed by applying a minimum abundance filter of 10 reads (–min_abund 10).

We also computed the match ratio, which is the ratio of sequences that merged with their "true" mothers divided by the number of merged sequences (Fig. 5). For alpha values of 6 or higher, the match ratio was close to 1 irrespective of the use of entropy correction or not, albeit it was slightly better without correction. At lower values of alpha, the match ratio decreased markedly for the Ratio merging criterion, and more so without correction, reaching values of ca. 75% at alpha =1. There were also marked differences in the three joining criteria (compared only for the runs with entropy correction). While the abundance Ratio criterion resulted in a strong decrease of the match ratio, using the Distance or the Ratio-Distance joining criteria, the match ratios remained close to 1 until values of alpha 3 and decreased slightly at alpha 2 and 1. Note that the different joining criteria do not affect the number of ESVs produced, but the number of sequences merged with each ESV and, thus, their relative abundances. By keeping track of which original sequence produced each error sequence, we could compare how the relative performance of the different methods changed with alpha values.

While this simulated dataset may not be a perfect representative of true metabarcoding datasets, it nevertheless highlights the importance of choosing the correct parameters of both alpha and minimum abundance filtering values as well as the need of choosing the proper joining criterion, especially at more stringent denoising levels (lower values of alpha). Note also that the results can vary depending on the error rate (we acknowledge that applying an uniform error rate of 0.005 is a simplification). Alpha values of 5 have been proposed for datasets of this COI fragment (*Elbrecht et al., 2018*; *Shum & Palumbi, 2021*; *Turon et al., 2020*) using several lines of evidence, but none of these studies included entropy correction. In addition, a minimal abundance filtering step is deemed necessary (*Elbrecht et al., 2018*; *Turon et al., 2020*) but an adequate threshold should be determined in each case. With our dataset and the explored error rate, values of 4 for alpha and 10 for minimal abundance seem a good compromise between keeping ca. 95% good sequences and accepting only a few error sequences. Our results emphasize the importance of calibrating the parameters for each type of data using any available evidence, including mock community data when available. The flexibility of DnoisE can greatly facilitate this exercise in future studies.

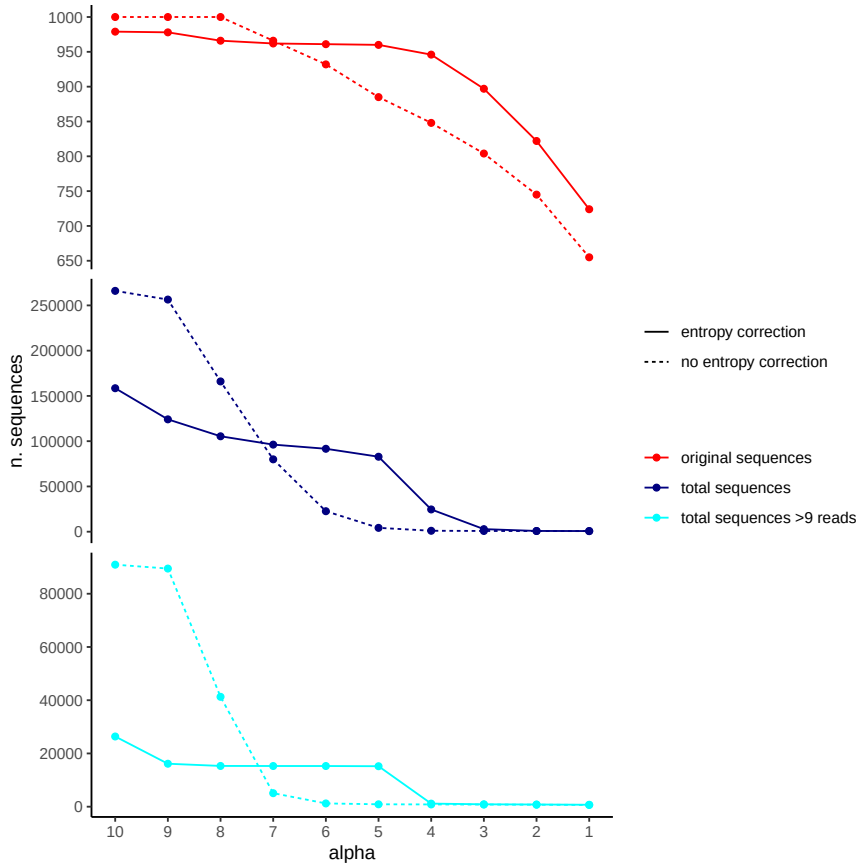


Figure 4 Number of original (correct) sequences (red), total sequences (dark blue) and total sequences filtered by read abundance (light blue) retrieved by DnoisE with entropy correction (solid line) and without entropy correction (equivalent to UNOISE). Values with abundance filtering were computed using a minimum abundance of 10 reads (–min_abund 10).

Full-size DOI: 10.7717/peerj.12758/fig-4

CONCLUSIONS

DnoisE is a novel denoising program that can be incorporated into any metabarcoding pipeline. It is a stand-alone program that addresses exclusively the denoising step, so that users can apply their favourite programs at all other steps (*e.g.*, chimera filtering, clustering…). Moreover, DnoisE is open-source code. Other programs used in metabarcoding pipelines also have open codes, such as DADA2 (*Callahan et al., 2016*), OBITOOLS (*Boyer et al., 2016*), SWARM (*Mahé et al., 2015*), or VSEARCH (*Rognes et al., 2016*). We strongly adhere to the open software concept for continuous and collaborative development of computing science and, in particular, in the metabarcoding field.

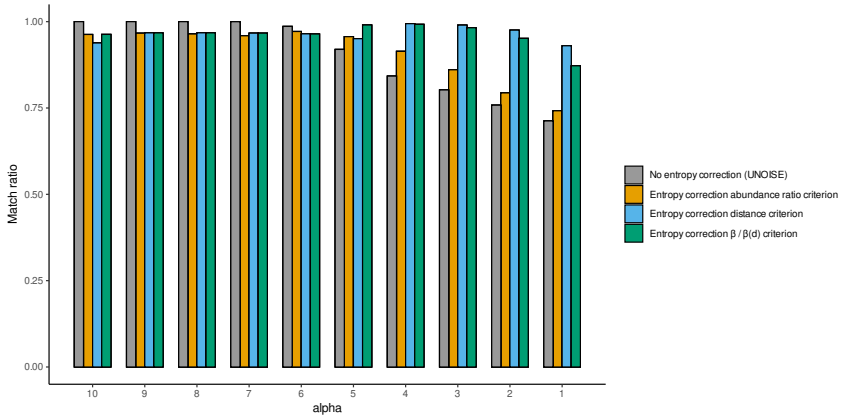


Figure 5 Match ratio (error sequences merged to their "true" mothers/total number of merged sequences) of DnoisE without entropy correction and abundance ratio joining criterion (equivalent to UNOISE) grey bars) and DnoisE with entropy correction. For DnoisE with entropy correction the three merging criteria were compared, abundance ratio criterion (orange bars), the genetic distance criterion (blue bars) and the criterion based on the cocient between the abundance ratio and the β(d) (green bars).
Full-size DOI: 10.7717/peerj.12758/fig-5

DnoisE is based on the UNOISE algorithm developed by *Edgar (2016)*, but with three main improvements: first, it allows to select among different criteria for joining sequences to optimize the match ratio; second, it incorporates the option to perform an entropy correction for coding genes, thus keeping more true sequences with high natural variability in third nucleotide positions in the codon; third, it is parallelizable to take advantage of the cluster architecture of modern computers.

Our correction by entropy opens a new field of analysis of coding genes, considering the different natural variability between codon positions. The flexibility of DnoisE with its settable options make this program a good tool for optimizing parameters in metabarcoding pipelines and for running the denoising step at any desired point of the pipeline (before or after clustering sequences into MOTUs).

In the next few years, processors are expected to reach the minimum size permitted by quantum laws. Parallel processing is needed to optimize future computer performance (*Gebali, 2011*; *Zomaya, 2005*). DnoisE offers a new parallel processing algorithm based on the MDA (maximum "daughter" abundance) to run analyses in parallel by groups of sequences that do not need to be compared between them. Parallel processing allows users to run huge datasets in a fast way using multithread computers. In our example, when running with 10 cores, DnoisE took about 2.78 h to compute a large dataset. On the other hand, memory management can be critical when running a high number of cores and large datasets and should be considered when setting the running parameters. DnoisE is written in Python3, one of the most popular languages, so it is a good option for users who want to modify or customize the code. We indeed encourage new developments of this software.

We consider that DnoisE is a good option to denoise metabarcoding sequence datasets from all kinds of markers, but especially for coding genes, given the entropy differences of codon positions. More details, sample files and complete instructions are available at GitHub (https://github.com/adriantich/DnoisE).

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This research was funded by the projects PopCOmics (CTM2017-88080, MCIN/AEI/10.13039/501100011033 and "ERDF A way of making Europe", EU), MARGECH (PID2020-118550RB, MCIN/AEI/10.13039/501100011033), and BigPark (OAPN, 2462/2017) from the Spanish Government. The publication charges for this article have been funded by a grant from the publication fund of UiT The Arctic University of Norway. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
PopCOmics: CTM2017-88080, MCIN/AEI/10.13039/501100011033.
MARGECH: PID2020-118550RB, MCIN/AEI/10.13039/501100011033.
BigPark from the Spanish Government: OAPN, 2462/2017.
UiT The Arctic University of Norway.

Competing Interests

Owen S. Wangensteen is an Academic Editor for PeerJ.

Author Contributions

- Adrià Antich conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, wrote the original code in Python, and approved the final draft.
- Creu Palacín conceived and designed the experiments, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Xavier Turon and Owen S. Wangensteen conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The open source code for DnoisE is publicly available at GitHub: https://github.com/adriantich/DnoisE.

The data set used to test the software is publicly available at Mendeley: Antich, Adrià; Palacin, Cruz; Wangensteen, Owen; Turon, Xavier (2021), "Dataset for "To denoise or to cluster? That is not the question. Optimizing pipelines for COI metabarcoding and metaphylogeography"", Mendeley Data, V3, doi: 10.17632/84zypvmn2b.3.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.12758#supplemental-information.

REFERENCES

Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R. 2017. Deblur rapidly resolves single-nucleotide community sequence patterns. *MSystems* **2(2)**:e00191–16 DOI 10.1128/msystems.00191-16.

Antich A, Palacín C, OS Wangensteen, Turon X. 2021. To denoise or to cluster, that is not the question: optimizing pipelines for COI metabarcoding and metaphylogeography. *BMC Bioinformatics* **22(1)**:177 DOI 10.1186/s12859-021-04115-6.

Boyer F, Mercier C, Bonin A, Le Bras Y, Taberlet P, Coissac E. 2016. Obitools: a unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources* **16**:176–182 DOI 10.1111/1755-0998.12428.

Brandt MI, Trouche B, Quintric L, Günther B, Wincker P, Poulain J, Arnaud-Haond S. 2021. Bioinformatic pipelines combining denoising and clustering tools allow for more comprehensive prokaryotic and eukaryotic metabarcoding. *Molecular Ecology Resources* **21(6)**:1904–1921 DOI 10.1111/1755-099813398.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. *Nature Methods* **13(7)**:581–583 DOI 10.1038/nmeth.3869.

Clarke LJ, Trebilco R, Walters A, AM Polanowski, Deagle BE. 2020. DNA-based diet analysis of mesopelagic fish from the southern Kerguelen Axis. *Deep Sea Research Part II: Topical Studies in Oceanography* **174**:104494 DOI 10.1016/J.DSR2.2018.09.001.

Creer S, Deiner K, Frey S, Porazinska D, Taberlet P, Thomas WK, Potter C, Bik HM. 2016. The ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology and Evolution* **7(9)**:1008–1018 DOI 10.1111/2041-210X.12574.

Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, De Vere N, Pfrender ME, Bernatchez L. 2017. Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Molecular Ecology* **26(21)**:5872–5895 DOI 10.1111/mec.14350.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26(19)**:2460–2461 DOI 10.1093/bioinformatics/btq461.

Edgar RC. 2016. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *BioRxiv*. 081257 DOI 10.1101/081257.

Elbrecht V, Vamos EE, Steinke D, Leese F. 2018. Estimating intraspecific genetic diversity from community DNA metabarcoding data. *PeerJ* **2018(4)**:e4644 DOI 10.7717/peerj.4644.

Gebali F. 2011. Algorithms and parallel Computing. In: *Algorithms and parallel computing*. Hoboken, New Jersey, USA: John Wiley and Sons DOI 10.1002/9780470932025.

Kelly RP, Port JA, Yamahara KM, Crowder LB. 2014. Using environmental DNA to census marine fishes in a large mesocosm. *PLOS ONE* **9(1)**:e86175 DOI 10.1371/journal.pone.0086175.

Mahé F, Rognes T, Quince C, De Vargas C, Dunthorn M. 2015. Swarmv2: highly-scalable and high-resolution amplicon clustering. *PeerJ* **2015(12)**:e1420 DOI 10.7717/peerj.1420.

McKerns MM, Strand L, Sullivan T, Fang A, Aivazis MAG. 2011. Building a framework for predictive science. In: *Proceedings of the 10th python in science conference*. 76–86.

Pawlowski J, Kelly-Quinn M, Altermatt F, Apothéloz-Perret-Gentil L, Beja P, Boggero A, Borja Á, Bouchez A, Cordier T, Domaizon I, Feio MJ, Filipe AF, Fornaroli R, Graf W, Herder J, Van der Hoorn B, Iwan Jones J, Sagova-Mareckova M, Moritz C, Kahlert M , et al. 2018. The future of biotic indices in the ecogenomic era: Integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. *Science of the Total Environment* **1295**:637–638 1310 DOI 10.1016/j.scitotenv.2018.05.002.

Peng X, Dorman KS. 2020. AmpliCI: a high-resolution model-based approach for denoising Illumina amplicon data. *Bioinformatics* **36(21)**:5151–5158 DOI 10.1093/bioinformatics/btaa648.

Pfeiffer F, Gröber C, Blank M, Händler K, Beyer M, Schultze JL, Mayer G. 2018. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Scientific Reports* **8**:10950 DOI 10.1038/s41598-018-29325-6.

Pichler M, Hartig F. 2020. A new method for faster and more accurate inference of species associations from big community data. ArXiv preprint. arXiv:2003.05331.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**:e2584 DOI 10.7717/peerj.2584.

Schirmer M, D'Amore R, Ijaz UZ, Hall N, Quince C. 2016. Illumina error profiles: resolving fine-scale variation in metagenomics sequencing data. *BMC Bioinformatics* **17**:125 DOI 10.1186/s12859-016-0976-y.

Schmitt AO, Herzel H. 1997. Estimating the entropy of DNA sequences. *Journal of Theoretical Biology* **188(3)**:369–377 DOI 10.1006/jtbi.1997.0493.

Shum P, Palumbi SR. 2021. Testing small-scale ecological gradients and intraspecific differentiation for hundreds of kelp forest species using haplotypes from metabarcoding. *Molecular Ecology* **30(13)**:3355–3373 DOI 10.1111/MEC.15851.

Sousa LL, Silva SM, Xavier R. 2019. DNA metabarcoding in diet studies: unveiling ecological aspects in aquatic and terrestrial ecosystems. *Environmental DNA* edn3.27 DOI 10.1002/edn3.27.

Turon X, Antich A, Palacín C, Præbel K, Wangensteen OS. 2020. From metabarcoding to metaphylogeography: separating the wheat from the chaff. *Ecological Applications* **30(2)**:e02036 DOI 10.1002/eap.2036.

Wangensteen OS, Palacín C, Guardiola M, Turon X. 2018. DNA metabarcoding of littoral hard-bottom communities: high diversity and database gaps revealed by two molecular markers. *PeerJ* **6**:e4705 DOI 10.7717/peerj.4705.

Zizka VMA, Weiss M, Leese F. 2020. Can metabarcoding resolve intraspecific genetic diversity changes to environmental stressors? A test case using river macrozoobenthos. *Metabarcoding and Metagenomics* **4**:23–34 DOI 10.3897/mbmg.4.51925.

Zomaya AY. 2005. Parallel computing for bioinformatics and computational biology. In: Zomaya AY, ed. *Parallel computing for bioinformatics and com-putational biology: models, enabling technologies, and case studies*. Hoboken: John Wiley & Sons Inc DOI 10.1002/0471756504.

Metabarcoding reveals high-resolution biogeographic and metaphylogeographic patterns through marine barriers

Antich A[1], Palacin C[2], Zarcero J[1], Turon X[1]*, Wangensteen OS[3]*

[1]Department of Marine Ecology, Centre for Advanced Studies of Blanes (CEAB- CSIC), Blanes (Girona), Catalonia, Spain

[2]Department of Evolutionary Biology, Ecology and Environmental Sciences and Biodiversity Research Institute (IRBIO), University of Barcelona, Barcelona, Catalonia, Spain

[3]Norwegian School of Fishery Science, UiT The Arctic University of Norway, Tromsø, Troms og Finnmark, Norway

* co-corresponding authors

Abstract

Aim The marine environment features oceanographic barriers that affect both the distribution and the connectivity of the marine biota. Biogeography can be extended by phylogeography, which analyses the distribution of genetic diversity within species. Metabarcoding can represent a leap forward in our ability to assess biogeographic and phylogeographic patterns, as it allows us to study many species at a time, including the often neglected small meio- and micro-organisms.

Location We tested the utility of the metabarcoding approach in one key biogeographic area, the Atlanto-Mediterranean transition along the E Iberian coast. This transition is marked by two barriers, the Almeria-Oran Front (AOF) and the Ibiza Channel (IC).

Time period Present

Major taxa studied Eukaryotes

Methods We sampled shallow hard-bottom communities at 12 sites over the littoral and performed community DNA metabarcoding using the cytochrome oxidase I (COI) marker. The resulting dataset was analysed at several levels: beta diversity of MOTUs (Molecular Operational Taxonomic Units, surrogate for species) and ESVs (Exact Sequence Variants, surrogate for haplotypes), and genetic differentiation within MOTUs (metaphylogeography).

Results In a context of high differentiation and isolation by distance, we nevertheless found a strong effect of the AOF at all levels, which marks the main boundary between the Atlantic and Mediterranean waters. The IC had a comparatively minor role. With the MOTU dataset we obtained more clear cut patterns than with ESVs, and we discourage the use of the latter as the unit of

biogeographic analyses. On the other hand, the metaphylogeographic approach provided the highest resolution in terms of differentiating localities and identifying geographic barriers.

Main conclusions Metabarcoding coupled with metaphylogeography provides a new tool to integrate the simultaneous analysis of beta diversity and genetic differentiation, unlocking a vast amount of information on the geographic distribution of biodiversity for basic and applied research.

Introduction

The marine environment, despite its apparent continuity, has physical and oceanographic barriers that determine the distribution of the different biota. The study of marine biogeography is a well-established field, and different regions and provinces have been proposed over the years, from Ekman's seminal review [1] to more recent accounts (e.g., [2–5]). These regions are usually defined by species turnover or changes in species abundances (beta-diversity) concomitant with geographic and oceanographic features. In addition, the advent of genetic techniques added a new component to the study of marine biogeography, thus giving rise to the field of phylogeography [6,7] which sought to assess how the present-day distribution of genetic diversity within species was reached [8,9]. Barriers may be reflected, not just in species change, but also in genetic divergence within species due to restricted connectivity coupled with drift/selection. Delimiting homogeneous biogeographic regions has relevance for management, marine reserves' delimitation, evolutionary approaches, and socio-economic issues [10,11].

Biogeographic breaks have been commonly studied on particular taxa, while studies with broad taxonomic coverage are rarer. Costello et al [10] provided the most comprehensive analyses of marine realms by compiling data from 65,000 marine species from public databases. Likewise, phylogeographic studies have usually addressed one species at a time, with few instances encompassing up to tens of species (e.g., [12–15]) or reviewing available information from multiple groups (e.g., [16–19]). Most often, however, biogeographic and phylogeographic studies concern macro-organismal components of biodiversity, while the small meio- and micro-eukaryotes have been comparatively neglected, in spite of their importance and evidence of genetic breaks in them (e.g., [20,21]). It is crucial to analyse patterns across macro- and micro-organisms to determine underpinning processes [22].

The rise of metabarcoding techniques during the last decade provided a new tool for assessing marine diversity in an integrative way, encompassing thousands of organisms (so-called MOTUs, or Molecular Operational Taxonomic Units), and including from micro- to macro-organisms to

efficiently detect biodiversity patterns and processes. Metabarcoding has become an invaluable tool for biomonitoring, impact assessment, and detection of introduced species, among others (reviewed in [23–28]). Likewise, metabarcoding datasets using highly variable markers can be mined for intraspecies genetic diversity [29–32] thereby opening the field for multispecies phylogeography (metaphylogeography, [33]). For metaphylogeographic analysis, stringent denoising of sequences to eliminate errors is necessary, generating Exact Sequence Variants (ESVs, e.g., [34–37]).

Coastal communities are among the most diverse marine habitats [38,39] and benthic species are likely to be more affected by oceanographic discontinuities than pelagic species [10]. The study of the benthos, therefore, is a powerful tool to assess biogeographic breaks. As in other environments, metabarcoding has boosted our ability to assess biodiversity in benthic communities, where most studies have been performed on soft-bottoms (e.g., [40–43]), with comparatively less work on rocky substrates, which are analysed either deploying artificial settlement units [44,45] or by collecting samples directly from the natural communities [46,47].

Metabarcoding has been commonly used for community analysis, but it has seldom been applied to the formal assessment of biogeographic breaks in coastal areas [48]. Some instances focused on particular groups of organisms (e.g., [49], protists; [50], ciliates; [51,52], vertebrates; [53], zooplankton), while other studies encompassed several groups [44,54] or even across-kingdom comparisons [55]. In all cases so far, however, these contributions were based on alpha- and beta-diversity changes. However, metabarcoding has the potential to uncover not only turnover rates and abundance changes of biotic components, but also to detect phylogeographic patterns of many species simultaneously as related to biogeographic breaks. Although it has been suggested that ESVs should be the unit of study instead of MOTUs for ribosomal markers [56], for markers with a high intraspecies variability such as the cytochrome oxidase I (COI) gene, this can lead to an overestimation of the alpha and beta biodiversity and to interpret as biogeographic breaks what in fact are phylogeographic discontinuities. The combined use of MOTUs (as surrogate of species) and ESVs (as surrogate of haplotypes) allows to extract both biogeographic and phylogeographic patterns [33,57,58], thus widening the scope of biogeographic studies for the assessment of marine discontinuities.

The Mediterranean is a well-known sea from the point of view of oceanographic features and biogeographic regions [59,60]. The Atlanto-Mediterranean transition is one of the most important biogeographic boundaries worldwide. Albeit the geographic border lies in the Gibraltar Strait, the main barrier is considered to be eastwards in the nearby Almería-Oran Front (AOF) [61,62], a density front where the inflowing Atlantic water is deflected southeastward. The AOF poses an effective limitation for the dispersion of marine organisms, and it is the effective genetic barrier between both seas for diverse groups of organisms (e.g., [16,63,64]). The westernmost Mediterranean Sea features a

sharp transition from Atlantic to Mediterranean waters, both along the N African coast and along the Iberian Peninsula. In this work we apply metabarcoding to characterise the biotic component of hard bottom benthic communities along the Iberian Mediterranean coast. We seek to analyse previously defined biogeographic breaks in these important transitional waters using a multilevel approach encompassing beta diversity analysis using both MOTUs (species level) and ESVs (haplotype level), and phylogeographic structures within MOTUs. Our aim is to test the potential of the metabarcoding approach to capture biogeographic and metaphylogeographic patterns across established oceanographic breaks.

Material and Methods

Sampling sites

We collected samples from 12 localities along the Mediterranean coast of the Iberian Peninsula. From South to North: Tarifa (TAR), Costa del Sol (SOL), La Herradura (LHE), Granada coast (GRA), Carboneras (CAR), Azohia (AZO), Cape Palos (PAL), Villajoyosa (JOY), Cullera (CLL), Calafat (CAL), Tossa de Mar (TOS) and Roses (ROS). These localities encompass two well-known oceanographic discontinuities: the Almeria-Oran front (AOF), between GRA and CAR [61,62,65], commonly considered the true boundary between the Atlantic and the Mediterranean, and the Ibiza Channel (IC), between JOY and CLL [66,67].

Accordingly, we grouped locations into three regions separated by these potential barriers: southern (TAR, SOL, LHE and GRA), central (CAR, AZO, PAL and JOY) and northern (CLL, CAL, TOS and ROS) regions (Fig. 1 and Supp. Table 1).

Sample collection and laboratory procedures

We targeted the eukaryote component of the photophilous community found between 4 and 8 m depth in subvertical rocky walls. These communities are dominated by seaweeds with a highly diverse understorey of macro- and meio-organisms. Sampling and laboratory processing were performed as described in [46]. In short, three sample replicates per locality were collected by scraping to bare rocky quadrats of 25 x 25 cm using a hammer and chisel. The material was collected in zip bags underwater, fixed with 95% ethanol within the hour, and stored at -20ºC. Sample processing included a size fractionation step in two sizes, large (L, >1mm) and small (S, between 1mm and 63µm) using stainless steel sieves. The two fractions were then homogenised separately with a blender, and 10 g of each were used for DNA extraction with the DNeasy PowerMax Soil Kit (Qiagen). Our initial dataset had thus a total of 72 samples (2 fractions x 3 replicates x 12

localities). All laboratory hardware was rinsed and bleached between samples. Negative controls were prepared by processing charred sand [68] instead of actual samples.

A fragment of the COI mitochondrial gene (Leray fragment) was amplified with the degenerated primer set Leray-XT from [46] with PCR conditions as indicated in that work. Amplification blanks were obtained using the PCR mix without addition of DNA template. Primers were tagged (as in [46]) to allow sample demultiplexing after sequencing. Library preparation was done with the BIOO NEXTFLEX PCR-Free DNA-Seq Kit (Perkin-Elmer) and sequencing was performed in an Illumina MiSeq V3 run with 2 x 250 bp paired-ends.

Bioinformatics pipeline

We processed the sequencing reads following a pipeline based on the OBITools package [69]. Illuminapairedend was used to align paired-end reads keeping only those with >40 quality score. Reads were demultiplexed using ngsfilter. Those with mismatched primer tags at any end were discarded. Obigrep and obiuniq were used to perform a length filter (retaining reads 299-320 bp long) and dereplicate sequences. Uchime-denovo algorithm from VSEARCH [70] was used to remove chimeric amplicons.

The downstream processing included clustering sequences into MOTUs with SWARM (with d=13 following [57]). We removed all MOTUs with less than 5 reads and used ecotag for taxonomic assignment against a local reference database, which is available at https://github.com/uit-metabarcoding/DUFA/ and contains 185,015 COI sequences. We then ran LULU [71] to remove potentially erroneous MOTUs and manually filtered the MOTU dataset to retain only the marine eukaryotes.

We then generated a sequence table for each MOTU using the output information of SWARM that contains a list of all sequences clustered in each MOTU. We denoised the sequences within each MOTU using DnoisE [37] to generate a table of exact sequence variants (ESV, [57]) for each MOTU. DnoisE takes into account the natural variability (measured as entropy values) of each codon position for coding genes (such as COI) to improve the denoising algorithm. The entropy values (0.4812, 0.2407, 1.0285 for the first, second, and third codon position, respectively) were obtained from the whole dataset before clustering using DnoisE. The stringency parameter (alpha) was set to 4 following [37]. Final filtering steps were as follows: i) we removed any ESV for which the abundance in the blanks or negative controls was higher than 10% of its total read abundance; ii) in each sample, we applied a minimum relative abundance threshold, setting to zero the reads of any ESV with abundance below 0.005% of the total reads of this sample (this was done to eliminate tag-switching between samples); iii) we eliminated all remaining ESVs with <5 total reads; iv) from the

whole ESV table we removed sequences with lengths deemed as incorrect: as for most species the length of the fragment used is 313, a correct sequence is expected to have 313 ± 3·n, being n the number of codons added or removed in indels; v) we finally removed sequences with stop codons and (for Metazoans) sequences with changes in conserved amino acids, since they probably arise from NUMTs, as described in [33]. The relative read abundances of each ESV in the two fractions of each sample were averaged for downstream analyses.

After these filtering steps, we obtained a dataset of MOTUs with taxonomic information and a dataset of ESVs (including all ESVs of all MOTUs). This allowed us to perform analyses at both levels: MOTUs (as surrogate of species) and ESVs (as surrogate of haplotypes) using relative read abundances as the analysed variable. Rarefaction curves and sample accumulation plots for both datasets were done using rarecurve and specaccum functions from the R package vegan [72].

Metaphylogeography dataset

The ESVs obtained in the previous analysis can be used to construct haplotype tables for phylogeographic inference for each MOTU [57]. To be able to capture potential patterns we selected only MOTUs that were present in at least two localities of two adjacent regions and with at least two ESVs each. As a proxy for haplotype abundances, ESV read abundances were converted to semi-quantitative values (following [33]): for each MOTU all ESVs were sorted in each sample in order of increasing abundance and ranked from 0 to 4 following percentiles of the ordered distribution: rank 0 for sequences with 0 reads; rank 1, for sequences that fell below the 51 percentile of the distribution; rank 2, sequences in percentiles >50 ≤ 75; rank 3, sequences in percentiles >75 ≤ 90; rank 4, sequences in the top >90 percentiles. The fractions of the same sample were ranked separately and then averaged to obtain the final semiquantitative abundance of each sequence in each sample.

Analyses

To assess community composition, MOTUs and ESVs were grouped into taxonomic super-groups (as in [73]) and, for metazoans, into phyla.

For biogeographic inference, Bray-Curtis (BC, with four-root transformation of relative read abundance per sample) and Jaccard (with presence-absence data) dissimilarities between samples were calculated using either the MOTU and the ESV dataset. These dissimilarities were used to ordinate samples in non-metric multidimensional scaling (nmMDS) using the metaMDS function from vegan package.

For the analysis of metaphylogeographic patterns, we computed a genetic differentiation matrix using the D estimator [74] with the function pairwise_D from the mmod R package [75]. D values ranged from 0 to 1 (maximal dissimilarity). D values were obtained for each MOTU selected for phylogeographic analysis (see above) by performing pairwise comparisons of all samples in which the MOTU was present. Finally, for each pair of samples, the average D values across all shared MOTUs was computed and used to construct a genetic dissimilarity matrix. This matrix was used to create a nmMDS, and a network analysis with EDENetworks [76]. For the latter, we used the mean D values across all-shared MOTUs among localities as the dissimilarity matrix and then transformed them into a network. The program automatically computes the percolation threshold (at which the all-including network breaks down into its main components), and we plotted the network just below this threshold. Finally, we plotted haplotype networks for all selected MOTUs, using the function haplonet of the R package pegas [77].

Mantel tests were performed with the three dissimilarity measures (BC for MOTUs and ESVs, D for genetic differentiation) and the logarithm of the shortest distances by sea among localities with the mantel function of the vegan package. These analyses were repeated separately for the localities within each of the three regions defined. As localities separated by fronts tended to be also more distant geographically, to disentangle the effects of geographic distance from those of the fronts, the different dissimilarities between adjacent localities were calculated to assess whether there is a peak in dissimilarity associated with the transition between fronts.

We also assessed the pattern of distributional breaks of MOTUs and ESVs in adjacent localities using a randomization approach (partly based on [78]). For each pair of adjacent communities, the number of breaks (defined as the number of MOTUs or ESVs present at only one of the two localities) was assessed and compared with the number found when the matrix of presence-absence was randomised across samples independently for each MOTU or ESV, thus effectively removing any geographic structure. This process was repeated 10,000 times and the distribution of breaks was determined and compared with the observed value. The variable used was number of breaks / number of MOTUs (ESVs) present at the two localities being compared, and significance was assessed when the observed value fell outside the generated distribution or at its extremes (using a two-tailed test with Bonferroni correction).

To further separate the effect of differentiation among localities and of potential breaks, we performed permutational analysis of variance (PERMANOVA) on the three dissimilarity matrices. We compared adjacent regions (South-Center and Center-North) using region and locality (nested within region) as factors. In this way the effect of the two discontinuities could be assessed once the contribution of differences between localities was factored out. The PERMANOVA module

incorporated in the Primer v6 statistical package [79] was used. Tests of multivariate dispersions (permdisp) were run when the main factors were significant to determine whether this outcome was a result of different multivariate means or different heterogeneity (spread) of the groups. A second PERMANOVA, followed by permutational pair-wise tests, was run with just the locality factor (12 levels) on the three dissimilarity matrices to assess the degree of differentiation between localities.

Results

We obtained 16,096,788 reads comprising 4,149,955 unique COI sequences after demultiplexing, quality filtering and chimera removal. The original raw sequences have been deposited in the NCBI SRA archive (accession numbers pending)

Sequences were clustered with SWARM followed by LULU, resulting in 257,719 MOTUs of which only 17,944 had 5 or more reads. We filtered taxonomically all MOTUs to retain only those assigned to marine eukaryotes, 8,696 MOTUs in total. We then obtained the ESVs using DnoisE within MOTUs. After all filtering steps we retained 18,026 ESVs, 3,392 MOTUS and 9,423,471 reads. The list of MOTUs and ESVs is provided as Supp File 1, and the taxonomic assignment of MOTUs in Supp File 2. As per sample, we had 588 ± 20 (mean±SE) ESVs, 263 ± 10 MOTUs and 130,882 ± 6,138 reads. At the locality level (combining samples), there is a significant correlation between the number of MOTUs and the number of ESVs (Pearson's r=0.641, p=0.025). From the retained MOTUs only 339 (indicated in Supp File 2) fulfilled the conditions to be used for metaphylogeographic analyses. They had a median of 11 haplotypes (ESVs) each, with 3 and 60 as 10% and 90% percentiles, respectively.

The rarefaction curves showed that all samples reached an asymptote (Fig. S1). However, for the species accumulation curve no clear plateau was reached as more samples were added (Fig. S2).

Community composition

Metazoans were the dominant group in all localities both in number of MOTUs and ESVs (Fig. S4). They were also the most abundant in relative number of reads, except in JOY, where Rhodophyta were dominant (Fig. 2). The latter group was the second most abundant in relative read abundance in all other localities except in TAR where Stramenopiles was the second group (Fig. 2a). For metazoans (Fig. 2b), a similar distribution in the number of reads across samples was found, with Porifera, Annelida, Arthropoda and Mollusca being the most abundant groups. MOTU composition across localities was homogeneous at the phylum level, but the composition in terms of ESV was

more variable (Fig. S3). The abundance of unidentified metazoans was higher in ROS and AZO, (over 30% of metazoan reads unassigned).

Biogeography

We computed non-metric Multidimensional Scaling using the Bray-Curtis (BC) (Fig. 3) and Jaccard dissimilarities (Fig. S4) to obtain a reduced space representation of the samples for MOTUs and ESVs. BC and Jaccard dissimilarities distances provided highly congruent results. In general, the different localities appeared well separated, with no overlap of the inertia ellipses in the nmMDS plots of MOTUs for both BC and Jaccard indices, while some overlap was found for ESV data between TOS, CAL and AZO. A geographic distribution was apparent, with a differentiation of the southern region from the other two along the first axis. The central and northern region did not form clearly separated clusters for MOTUs, and even less so for ESV data.

PERMANOVA analyses (Table 1) of Bray-Curtis dissimilarities showed for MOTUs a significant effect of the differentiation between southern and central regions, and not between central and northern regions. For the ESVs, no significant differentiation associated to regions was detected. In all cases, the nested locality factor explained most of the variation and was highly significant ($p<0.001$). No dispersion differences were detected for levels of significant factors (permdist tests).

PERMANOVAs for the locality factor alone were highly significant for both MOTUs and ESVs ($p<0.001$, while permdisp tests were not significant), and pairwise tests revealed that all pairs of localities were significantly differentiated in the MOTU dataset (with the exception of the two northernmost localities, TOS and ROS), while for the ESV dataset the following pairs of populations were not significantly different: TAR-SOL, SOL-LHE, SOL-GRA, GRA-AZO, AZO-JOY, AZO-CLL, TOS-ROS.

The values of BC and Jaccard dissimilarities were higher for the analysis of ESVs (means of 0.893 and 0.942, respectively) compared to MOTUs (means of 0.757 and 0.841), which was expectable as localities should share less ESVs than MOTUs. For both MOTUs and ESVs, using the Jaccard distance the low number of shared taxonomic units between samples of different localities was evident, especially for TAR (which had the highest BC and Jaccard dissimilarities with other localities (Fig. S4). TOS and ROS had the lowest values of both dissimilarities (Fig. 3 and Fig. S4). Overall, BC values for MOTUs spanned a wider range of values (inter-locality comparisons, from 0.472 to 0.946) than for ESVs (from 0.665 to 0.988).

When comparing dissimilarity values from adjacent localities (Fig. 4 and Fig. S5), the transition associated with the Almeria-Oran Front (GRA-CAR) had the highest mean values both for MOTUs and ESVs, followed by the comparisons between TAR and SOL and PAL and AZO. The Ibiza

Channel (JOY-CLL) came next, with relatively high values but lower than some intra-region comparisons. The same results were obtained using the Jaccard distance (not shown).

The study of distributional breaks through permutation (Fig. 5) showed that most transitions had significantly more breaks than expected if breaks were randomly distributed, again revealing a strong structure between populations. However, for MOTUs the two highest values were found in the AOF transition (GRA-CAR, 26% more breaks than expected) and the IC transition (JOY-CLL, 22%), while there were significantly less breaks than expected between the two northernmost localities (TOS-ROS). For ESVs, the values were in general lower, with again the highest deviation (11%) from random expectation in the AOF transition, while the IC did not show any increase in associated breaks.

Metaphylogeography

A total of 339 MOTUs were selected for the metaphylogeographic analysis. Of these, 160 were found in at least 2 localities of each region, of which 12 were found in all localities. Of the 339 MOTUs, 85 were tagged with a species name, 8 had genus and 13 family assignments, and the remaining 233 MOTUs were assigned at order or higher taxonomic rank. 240 of the MOTUs were Metazoa, 53 Rhodophyta, 10 Stramenopiles, 3 Viridiplantae, 2 Alveolata and the remaining 31 were unassigned eukaryotes. The best represented metazoan phylum was Annelida (53 MOTUs), followed by Arthropoda (51 MOTUs), Cnidaria (28 MOTUs), Porifera (20 MOTUs) and Mollusca (17 MOTUs). Haplotype networks for these 339 MOTUs are presented in Supp. File 3. They show a variety of patterns, but are predominantly star-shaped with a few dominant haplotypes. The latter are in general shared among regions, albeit with different proportions.

We computed a dissimilarity matrix of 36x36 (3 samples x 12 localities) with the average D values for each pair of samples computed from the shared MOTUs. These values were used to map samples in a nmMDS (Fig. 3) that showed a sharp separation between localities, with no overlap of the inertia ellipses. The first axis separated the southern region from the other two, which in turn formed distinct clusters along the second dimension. PERMANOVA analyses showed a significant effect of the Region factor, in both the differentiation between southern and central, and between central and northern regions ($p=0.011$ and $p=0.033$, respectively). The nested locality factor again explained most of the variance and was highly significant ($p<0.001$) with significant differences also in dispersion levels (permdisp tests, Table 1). The analysis of the factor locality as the main factor showed a highly significant effect ($p<0.001$) but also a significant difference in dispersion values (permdisp $p=0.003$). All pairwise comparisons were significant except between the two southernmost localities TAR and SOL.

The analysis of D dissimilarities from adjacent localities (Fig. 4) showed that GRA and CAR (corresponding to the AOF) had the highest average differentiation, followed by TAR and SOL and PAL and AZO. The lowest differentiation between adjacent localities was found in the northern region (TOS and ROS followed by CAL and TOS). No clear differentiation was detected associated with the IC break.

The network analysis using EDENetworks detected the percolation threshold at a D value of 0.51. The network obtained just below this threshold (D=0.50, Fig. 6) showed a separation between the southern region and the central and northern regions corresponding to the AOF. In turn, the central and northern regions were connected by a few weak links involving mostly the northernmost central region locality (JOY). Only the link between the localities at both sides of the break, JOY and CLL, was relatively strong, which is consistent with the pattern shown in Fig. 4. The northern region showed strong internal links, particularly between CAL, TOS, and ROS. The node with the highest betweenness centrality (indicating its importance in connecting other nodes, [76]) is JOY, which also has the highest number of links.

The Mantel tests (Fig. S6) showed that, for the three variables considered (BC dissimilarity between localities calculated with the MOTU and ESV, and genetic distance D) there was a highly significant correlation with geographic distance (all Mantel $r>0.803$, $p<0.001$). The same result was obtained within regions (all Mantel $r>0.748$, $p<0.001$), indicating a clear signal of isolation by distance.

Finally, we computed the relationship between the Bray-Curtis dissimilarities between localities for the MOTU and ESV datasets and also plotted the D value of the MOTUs selected for the metaphylogeographic analysis (Fig. 7). The results showed a general good correlation of the three measures. Overall, pairwise differentiation values are higher for ESVs than for MOTUs, and the difference is reduced for highly differentiated localities (i.e., with values close to 1 for both datasets).

Discussion

Metabarcoding of highly diverse shallow benthic communities, using a broadly used mitochondrial marker (COI), retrieved both biological and genetic diversity from the Atlanto-Mediterranean transition along the eastern Iberian coast, marked by two known discontinuities. The present study is the first to explore the effects of barriers to gene flow in the marine realm simultaneously with biogeographic patterns using metabarcoding data and encompassing different groups of eukaryotes.

Both the biogeographic and the phylogeographic perspectives showed similar patterns of community differentiation but with different resolution. The different approaches reveal important information at several levels of biological organisation.

Biogeographic (MOTUs and ESVs) patterns

Along the 1,200 km of the Iberian coast we retrieved a high diversity of taxa in all localities. Rarefaction curves showed an adequate number of reads per sample but more replicated samples seem necessary to capture all diversity present in such complex assemblages. Communities were dominated by metazoans in both number of MOTUs and relative read abundance, with Porifera, Cnidaria, Annelida and Arthropoda being the most abundant phyla. Across all samples, taxa composition was similar in terms of relative number of MOTUs, but TAR showed a higher relative abundance of Stramenopiles than other localities. It had a community dominated by brown algae, unlike the other sites. PAL was also different in the read abundance of metazoan groups compared to other localities, with a higher abundance of Cnidaria. The benthic community in PAL visually differed from other locations, with low abundance of algae and high abundance of Anthozoa and Hydrozoa. Overall, about 25% of metazoan MOTUs did not match with any phyla, emphasising the importance of completing current reference databases [46,80]. However, even if low-level taxonomic assignments are lacking, unassigned MOTUs can still be used to calculate diversity metrics.

In the present study three regions were considered, separated by two previously described fronts. We ordered samples of these areas in nmMDS plots using both MOTUs and ESVs. Localities from the southern region were well separated from those of central and northern regions. In addition, the localities of GRA and CAR, separated by the AOF, showed the highest values of BC dissimilarity of all comparisons of adjacent localities. This emphasises the importance of this barrier. The AOF is a geostrophic front that separates Atlantic waters entering through the Gibraltar Strait from Mediterranean waters, thus marking the main boundary in the Atlanto-Mediterranean transition [61,62,65]. However, although its role in the genetic structure of many species has been investigated (f.i., [16,19,81]), there is to date no comprehensive analysis of its effect in species beta diversity. The role of this front is a clear-cut feature of our analyses. In fact, of the MOTUs present in the southern and central regions (separated by the AOF), 57.1% were present at only one side of this divide. In terms of ESVs, this figure was 81.8%. The IC was not as strong a break in terms of regional differentiation. The nMDS analyses showed an imperfect separation of the northern and central regions, and the BC dissimilarity of JOY and CLL (separated by the IC) was relatively high but smaller than values obtained from other comparisons of adjacent localities. Overall, dissimilarity values show that AOF is a strong biogeographic barrier but the IC is not and, thus, the central and northern regions are not well differentiated in biotic composition.

We also note that TAR appeared separated from all other localities in our nmMDS analyses of both MOTUs and ESVs. As mentioned earlier, this community was the only one dominated by brown algae, but being located just on the Gibraltar Strait, this result could also indicate an effect of this geographic boundary. Unfortunately, our sampling scheme was not designed to test this effect, which should be considered in future studies. PAL was also somewhat offset from the other central region localities, which is attributable to a different community with less algae and higher invertebrate dominance.

While both descriptors (MOTUs and ESVs) provided basically the general pattern, there were nevertheless differences. The PERMANOVA analyses showed a significant differentiation between the southern and central localities with the MOTU dataset, which was not found with ESVs. Likewise, almost all pairwise comparisons between localities (PERMANOVA) revealed significant differences with MOTUs (except TOS-ROS), while seven comparisons were not significant with ESVs. This is also reflected in the overlap of some localities in the nmMDS analyses. When considering the pattern of distributional breaks between adjacent localities, in general more breaks than expected in a random simulation were found, supporting the idea that localities were well differentiated. However, values were higher for MOTUs than for ESVs, and for the former the AOF and IC transitions had the highest values, while IC was not significant for ESVs. The narrower range of dissimilarities obtained with ESVs (0.524 to 0.988) than with MOTUs (0.343 to 0.946) may have hampered resolution when using the ESV dataset.

The analysis of MOTU-level turnover is the metabarcording equivalent to the standard biogeographic species-level analysis. On the other hand, using ESVs instead is equivalent to analyse haplotype turnover, whose interpretation is unclear. In particular, if we miss the MOTU information we would be giving the same weight to a distributional change in a haplotype from a species with high genetic variability (many ESVs) than from a species with just a few haplotypes. Diversity calculations based on ESVs are thus driven and biased by species with high intraspecies variability. We would be lumping together biogeographic (interspecific) and phylogeographic (intraspecific) information. On the contrary, if ESV information is grouped into meaningful biological units (MOTUs), then the well-developed and tested tools for analysis of geographic genetic differentiation can be applied in a metaphylogeographic context. We used in this work only basic analyses (population differentiation and network analyses), but the whole panoply of phylogeographic analytical tools [82] can be applied depending on the question of interest.

Recent articles have discussed the relative merits of using MOTUs and/or ESVs for metabarcoding studies [57,58]. These works emphasise that using ESVs as a standard unit of analysis, as suggested previously [56], may be valid for ribosomal markers but not when studying eukaryotes with highly

variable markers such as COI. In our view, in these cases diversity patterns must be studied using MOTUs as a proxy for species, while ESVs must be used within MOTUs as a proxy for haplotypes and intraspecies variability (metaphylogeographic approach), allowing a hierarchical analysis of the distribution of diversity.

Phylogeographic perspective

Geophysical barriers play a crucial role in population fragmentation even in apparently continuous marine environments. Phylogeography analyses the geographic distribution of genetic lineages, linking geography and genealogy, and has been developed since the eighties of the last century [6,7]. Phylogeographic studies rely on species that are easy to sample, being therefore restricted in general to macro-organisms, commercially interesting species, or flagship iconic species. Small organisms are only rarely studied due to the difficulty of sampling individuals. There is therefore a lack of information on whether phylogeographic patterns of marine macro-organisms are coherent with those of meio- and micro-organisms.

From all the MOTUs that we found, only 10% could be used for genetic dissimilarity analysis. This is caused by a high number of low abundance MOTUs found in few samples compared to those selected for the analyses, which had a broader distribution. Our results show high values of genetic dissimilarity when comparing samples from different localities, with almost all comparisons being significant in PERMANOVA analyses. However, dissimilarities between localities of the same region were smaller (0.470 ± 0.004) than those between regions (0.560 ± 0.002) meaning that gene flow is higher within than between regions. The three regions appeared well separated in nmMDS ordinations, and PERMANOVA analyses showed significant differences between southern and central and between central and northern regions. Furthermore, a network analysis reflected disconnected networks in the southern and the central plus northern regions. Among the latter, the links between regions were feeble with the exception of the edge between JOY and CLL. The JOY locality, on the other hand, had links with all other localities in the central and northern regions and the highest betweenness centrality in the whole network, thus constituting a hotspot for genetic connectivity in the area. If we perform the network analysis without JOY, the central and northern regions appear disconnected (results not shown). Rather than a clearcut divide, the network analysis indicated that the IC is placed in a transition zone connected to both sides.

Phylogeographic structure and species beta diversity are two complementary dimensions of integrative biogeography in a broad sense [8]. However, the former is much harder information to acquire. Phylogeographic marine breaks have been usually studied on a single species basis, sampling populations and analysing a set of genetic markers, depending on the study. Multispecies

studies are rare and include only a handful of species (e.g., [13,14]). Alternatively, meta-analyses of published data can be used to make inferences [19,78,83,84]. Metaphylogeography is a new way to study population genetic differentiation for the whole community using metabarcoding data. This new tool has the potential to detect subtle patterns and study genetic connectivity with a relatively low sampling effort and targeting a huge amount of taxa of any size. As pointed out by [85], the study of haplotypic diversity can provide crucial information on the state of the ecosystem and predict which populations are more sensitive to environmental changes. Moreover, the study of barriers affecting gene flow is mandatory to manage not only biodiversity but also genetic diversity [86], and metaphylogeography can become a key tool to achieve integrated management programs.

Conclusions

The simultaneous study of biogeographic and phylogeographic patterns captured important information at different levels of biological organisation. There was an overall pattern of high structure between localities and a significant relationship with geographic distance. Superimposed to this pattern, the Almeria-Orant Front (AOF) had a strong structuring effect in most analyses, confirming expectations. On the other hand, the Ibiza Channel (IC) barrier had a minor effect, detected only with the genetic differentiation analyses (metaphylogeography) and the distribution of breaks using MOTU data, but not with ESV data. The distribution of species can be determined by a broad range of biotic and abiotic factors, leading to differences in community composition. However, both isolation and local adaptation can have an effect not only on the species distribution but also determine shifts in haplotype frequencies within species. We do not favour the study of ESVs alone without the species (MOTU) context. The haplotypes are not independent units, they are distributed, adapt, and evolve encapsulated in biological units (species), which have biological, historical, and demographic traits that determine their haplotype richness and distribution. We therefore suggest using MOTUs as the unit for species turnover analysis and ESVs within MOTUs for phylogeographic analysis when using metabarcoding data.

Metabarcoding coupled with metaphylogeography provide a new tool to integrate the simultaneous analysis of species turnover and genetic differentiation, unlocking a vast amount of information on the geographic distribution of biodiversity for basic and applied research.

Bibliography

1. Ekman, S. *Zoogeography of the Sea*. *Animal Biology Series* (Sidgwick & Jackson, 1953). doi:10.1080/00222935308654417
2. Briggs, J. C. Global biogeography. *Dev. Palaeontol. Stratigr.* **14**, i–xvii, 1–452 (1995).
3. Longhurst, A. *Ecological geography of the sea*. (Longhurst, Alan, 1998).
4. Spalding, M. D. *et al.* Marine Ecoregions of the World: A Bioregionalization of Coastal and Shelf Areas. *Bioscience* **57**, 573–583 (2007).
5. Toonen, R. J., Bowen, B. W., Iacchei, M. & Briggs, J. C. Biogeography, Marine. in *Encyclopedia of Evolutionary Biology* (ed. Kliman, R. M.) 166–178 (Elsevier Inc., 2016). doi:10.1016/B978-0-12-800049-6.00120-7
6. Avise, J. C. *et al.* INTRASPECIFIC PHYLOGEOGRAPHY: The Mitochondrial DNA Bridge Between Population Genetics and Systematics. *Ann. Rev. Ecol. Sys* **18**, 489–522 (1987).
7. Avise, J. C. Phylogeography: retrospect and prospect. *J. Biogeogr.* **36**, 3–15 (2009).
8. Riddle, B. R. *et al.* The role of molecular genetics in sculpting the future of integrative biogeography. *Prog. Phys. Geogr. Earth Environ.* **32**, 173–202 (2008).
9. Vellend, M. *et al.* Drawing ecological inferences from coincident patterns of population- and community-level biodiversity. *Mol. Ecol.* **23**, 2890–2901 (2014).
10. Costello, M. J. *et al.* Marine biogeographic realms and species endemicity. *Nat. Commun. 2017 81* **8**, 1–10 (2017).
11. Thiel, M. *et al.* The Humboldt Current System of Northern and Central Chile. in *Oceanography and Marine Biology : An annual Review* (eds. Gibson, R. N., Atkinson, R. J. A. & Gordon, J. D. M.) **45**, 195–344 (Taylor & Francis, 2007).
12. Ayre, D. J., Minchinton, T. E. & Perrin, C. Does life history predict past and current connectivity for rocky intertidal invertebrates across a marine biogeographic barrier? *Mol. Ecol.* **18**, 1887–1903 (2009).
13. Haye, P. A. *et al.* Phylogeographic Structure in Benthic Marine Invertebrates of the Southeast Pacific Coast of Chile with Differing Dispersal Potential. *PLoS One* **9**, e88613 (2014).
14. Kelly, R. P. & Palumbi, S. R. Genetic Structure Among 50 Species of the Northeastern Pacific Rocky Intertidal Community. *PLoS One* **5**, e8594 (2010).
15. Cahill, A. E. *et al.* A multispecies approach reveals hot spots and cold spots of diversity and

connectivity in invertebrate species with contrasting dispersal modes. *Mol. Ecol.* **26**, 6563–6577 (2017).

16. Patarnello, T., Volckaert, F. A. M. J. & Castilho, R. Pillars of Hercules: is the Atlantic–Mediterranean transition a phylogeographical break? *Mol. Ecol.* **16**, 4426–4444 (2007).

17. Hardy, S. M. *et al.* Biodiversity and phylogeography of Arctic marine fauna: insights from molecular tools. *Mar. Biodivers.* **41**, 195–210 (2011).

18. Teske, P. R., Von der Heyden, S., McQuaid, C. D. & Barker, N. P. A review of marine phylogeography in southern Africa. *S. Afr. J. Sci.* **107**, (2011).

19. Pascual, M., Rives, B., Schunter, C. & Macpherson, E. Impact of life history traits on gene flow: A multispecies systematic review across oceanographic barriers in the Mediterranean Sea. *PLoS One* **12**, e0176419 (2017).

20. Derycke, S. *et al.* Phylogeography of the Rhabditis (Pellioditis) marina species complex: evidence for long-distance dispersal, and for range expansions and restricted gene flow in the northeast Atlantic. *Mol. Ecol.* **17**, 3306–3322 (2008).

21. Tulchinsky, A. Y., Norenburg, J. L. & Turbeville, J. M. Phylogeography of the marine interstitial nemertean Ototyphlonemertes parmula (Nemertea, Hoplonemertea) reveals cryptic diversity and high dispersal potential. *Mar. Biol.* **159**, 661–674 (2012).

22. Shade, A. *et al.* Macroecology to Unite All Life, Large and Small. *Trends Ecol. Evol.* **33**, 731–744 (2018).

23. Deiner, K. *et al.* Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Mol. Ecol.* **26**, 5872–5895 (2017).

24. Cordier, T. & Pawlowski, J. BBI: an R package for the computation of Benthic Biotic Indices from composition data. *Metabarcoding and Metagenomics* **2**, e25649 (2018).

25. Bowers, H. A. *et al.* Towards the optimization of edna/erna sampling technologies for marine biosecurity surveillance. *Water* **13**, 1113 (2021).

26. Rodríguez-Ezpeleta, N. *et al.* Biodiversity monitoring using environmental DNA. *Mol. Ecol. Resour.* **21**, 1405–1409 (2021).

27. Miya, M. Environmental DNA Metabarcoding: A Novel Method for Biodiversity Monitoring of Marine Fish Communities. *Ann. Rev. Mar. Sci.* **14**, 161–185 (2022).

28. Pawlowski, J. *et al.* Environmental DNA metabarcoding for benthic monitoring: A review of sediment sampling and DNA extraction methods. *Sci. Total Environ.* **818**, 151783 (2022).

29. Elbrecht, V., Vamos, E. E., Steinke, D. & Leese, F. Estimating intraspecific genetic diversity from community DNA metabarcoding data. *PeerJ* **2018**, e4644 (2018).

30. Adams, C. I. M. *et al.* Beyond Biodiversity: Can Environmental DNA (eDNA) Cut It as a Population Genetics Tool? *Genes (Basel).* **10**, 192 (2019).

31. Sigsgaard, E. E. *et al.* Population-level inferences from environmental DNA—Current status and future perspectives. *Evol. Appl.* **13**, 245–262 (2020).

32. Andújar, C. *et al.* Community assembly and metaphylogeography of soil biodiversity: insights from haplotype-level community DNA metabarcoding within an oceanic island. *Authorea Prepr.* (2022).

33. Turon, X., Antich, A., Palacín, C., Præbel, K. & Wangensteen, O. S. From metabarcoding to metaphylogeography: separating the wheat from the chaff. *Ecol. Appl.* **30**, e02036 (2020).

34. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).

35. Edgar, R. C. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv* 081257 (2016). doi:10.1101/081257

36. Andújar, C. *et al.* Validated removal of nuclear pseudogenes and sequencing artefacts from mitochondrial metabarcode data. *Mol. Ecol. Resour.* **21**, 1772–1787 (2021).

37. Antich, A., Palacín, C., Turon, X. & Wangensteen, O. S. DnoisE: distance denoising by entropy. An open-source parallelizable alternative for denoising sequence datasets. *PeerJ* **10**, e12758 (2022).

38. Reaka-Kudla, M. L. The global biodiversity of coral reefs: a comparison with rainforests. Biodiversity II: understanding and protecting our natural resources. in *Biodiversity II: Understanding and Protecting Our Biological Resources* 83–108 (Joseph Henry/National Academy Press, 1997).

39. Agardy, T. *et al.* Coastal systems. in *Millennium ecosystem assessment: ecosystems and human well-being* (ed. Reid, W. V.) 513–549 (Island Press, 2005).

40. Guardiola, M. *et al.* Spatio-temporal monitoring of deep-sea communities using metabarcoding of sediment DNA and RNA. *PeerJ* **4**, e2807 (2016).

41. Fonseca, V. G. *et al.* Revealing higher than expected meiofaunal diversity in Antarctic sediments: a metabarcoding approach. *Sci. Reports 2017 71* **7**, 1–11 (2017).

42. Brannock, P. M., Learman, D., Mahon, A., Santos, S. & Halanych, K. Meiobenthic community composition and biodiversity along a 5500 km transect of Western Antarctica: a metabarcoding analysis. *Mar. Ecol. Prog. Ser.* **603**, 47–60 (2018).

43. Atienza, S. *et al.* DNA Metabarcoding of Deep-Sea Sediment Communities Using COI: Community Assessment, Spatio-Temporal Patterns and Comparison with 18S rDNA. *Diversity* **12**, 123 (2020).

44. Cahill, A. E. *et al.* A comparative analysis of metabarcoding and morphology-based identification of benthic communities across different regional seas. *Ecol. Evol.* **8**, 8908–8920 (2018).

45. David, R. *et al.* Lessons from photo analyses of Autonomous Reef Monitoring Structures as tools to detect (bio-)geographical, spatial, and environmental effects. *Mar. Pollut. Bull.* **141**, 420–429 (2019).

46. Wangensteen, O. S., Palacín, C., Guardiola, M. & Turon, X. DNA metabarcoding of littoral hard-bottom communities: high diversity and database gaps revealed by two molecular

markers. *PeerJ* **6**, e4705 (2018).

47. Shum, P., Barney, B. T., O'Leary, J. K. & Palumbi, S. R. Cobble community DNA as a tool to monitor patterns of biodiversity within kelp forest ecosystems. *Mol. Ecol. Resour.* **19**, 1470–1485 (2019).

48. Gaither, M. R., DiBattista, J. D., Leray, M. & Heyden, S. Metabarcoding the marine environment: from single species to biogeographic patterns. *Environ. DNA* **4**, 3–8 (2022).

49. Pagenkopp Lohan, K. M., Fleischer, R. C., Torchin, M. E. & Ruiz, G. M. Protistan Biogeography: A Snapshot Across a Major Shipping Corridor Spanning Two Oceans. *Protist* **168**, 183–196 (2017).

50. Santoferrara, L. F., Rubin, E. & McManus, G. B. Global and local DNA (meta)barcoding reveal new biogeography patterns in tintinnid ciliates. *J. Plankton Res.* **40**, 209–221 (2018).

51. Closek, C. J. *et al.* Marine vertebrate biodiversity and distribution within the central California Current using environmental DNA (eDNA) metabarcoding and ecosystem surveys. *Front. Mar. Sci.* **6**, 732 (2019).

52. Czachur, M. V., Seymour, M., Creer, S. & von der Heyden, S. Novel insights into marine fish biodiversity across a pronounced environmental gradient using replicated environmental DNA analyses. *Environ. DNA* **00**, 1–10 (2021).

53. Pitz, K. J. *et al.* Zooplankton biogeographic boundaries in the California Current System as determined from metabarcoding. *PLoS One* **15**, e0235159 (2020).

54. DiBattista, J. D. *et al.* Environmental DNA reveals a multi-taxa biogeographic break across the Arabian Sea and Sea of Oman. *Environ. DNA* **4**, 206–221 (2022).

55. Holman, L. E. *et al.* Animals, protists and bacteria share marine biogeographic patterns. *Nat. Ecol. Evol.* **5**, 738–746 (2021).

56. Callahan, B. J., McMurdie, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* **11**, 2639–2643 (2017).

57. Antich, A., Palacín, C., Wangensteen, O. S. & Turon, X. To denoise or to cluster, that is not the question: optimizing pipelines for COI metabarcoding and metaphylogeography. *BMC Bioinformatics* **22**, 177 (2021).

58. Brandt, M. I. *et al.* Bioinformatic pipelines combining denoising and clustering tools allow for more comprehensive prokaryotic and eukaryotic metabarcoding. *Mol. Ecol. Resour.* **21**, 1904–1921 (2021).

59. Bianchi, C. N. & Morri, C. Marine Biodiversity of the Mediterranean Sea: Situation, Problems and Prospects for Future Research. *Mar. Pollut. Bull.* **40**, 367–376 (2000).

60. Bianchi, C. N. Biodiversity issues for the forthcoming tropical Mediterranean Sea. *Hydrobiol. 2007 5801* **580**, 7–21 (2007).

61. Tintore, J., Violette, P. E. La, Blade, I. & Cruzado, A. A Study of an Intense Density Front in the Eastern Alboran Sea: The Almeria–Oran Front. *J. Phys. Oceanogr.* **18**, (1988).

632 62. Folkard, A. M., Davies, P. A. & Prieur, L. The surface temperature field and dynamical
633 structure of the Almeria-Oran front from simultaneous shipboard and satellite data. *J. Mar.*
634 *Syst.* **5**, 205–222 (1994).

635 63. Naciri, M., Lemaire, C., Borsa, P. & Bonhomme, F. Genetic study of the
636 Atlantic/Mediterranean transition in sea bass (Dicentrarchus labrax). *J. Hered.* **90**, 591–596
637 (1999).

638 64. Carreras, C. *et al.* East is East and West is West: Population genomics and hierarchical
639 analyses reveal genetic structure and adaptation footprints in the keystone species
640 Paracentrotus lividus (Echinoidea). *Divers. Distrib.* **26**, 382–398 (2020).

641 65. L'Helguen, S., Le Corre, P., Madec, C. & Morin, P. New and regenerated production in the
642 Almeria-Oran front area, eastern Alboran Sea. **49**, 83–99 (2002).

643 66. Bouffard, J., Pascual, A., Ruiz, S., Faugère, Y. & Tintoré, J. Coastal and mesoscale dynamics
644 characterization using altimetry and gliders: A case study in the Balearic Sea. *J. Geophys.*
645 *Res. Ocean.* **115**, 10029 (2010).

646 67. Pinot, J. M., López-Jurado, J. L. & Riera, M. The CANALES experiment (1996-1998).
647 Interannual, seasonal, and mesoscale variability of the circulation in the Balearic Channels.
648 *Prog. Oceanogr.* **55**, 335–370 (2002).

649 68. Wangensteen, O. S. & Turon, X. Metabarcoding Techniques for Assessing Biodiversity of
650 Marine Animal Forests. in *Marine Animal Forests* 445–473 (Springer International
651 Publishing, 2017). doi:10.1007/978-3-319-21012-4_53

652 69. Boyer, F. *et al.* Obitools: a unix-inspired software package for DNA metabarcoding. *Mol.*
653 *Ecol. Resour.* **16**, 176–182 (2016).

654 70. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open
655 source tool for metagenomics. *PeerJ* **4**, e2584 (2016).

656 71. Frøslev, T. G. *et al.* Algorithm for post-clustering curation of DNA amplicon data yields
657 reliable biodiversity estimates. *Nat. Commun.* **8**, 1–11 (2017).

658 72. Oksanen, J. *et al.* vegan: Community Ecology Package. (2019).

659 73. Guardiola, M. *et al.* Deep-Sea, Deep-Sequencing: Metabarcoding Extracellular DNA from
660 Sediments of Marine Canyons. *PLoS One* **10**, (2015).

661 74. Jost, L. GST and its relatives do not measure differentiation. *Mol. Ecol.* **17**, 4015–4026
662 (2008).

663 75. Winter, D., Green, P., Kamvar, Z. & Gosselin, T. Modern Measures of Population
664 Differentiation. (2017).

665 76. Kivelä, M., Arnaud-Haond, S. & Saramäki, J. EDENetworks: A user-friendly software to
666 build and analyse networks in biogeography, ecology and population genetics. *Mol. Ecol.*
667 *Resour.* **15**, 117–122 (2015).

668 77. Paradis, E. pegas: an R package for population genetics with an integrated--modular

approach. *Bioinformatics* **26**, 419–420 (2010).

78. Arranz, V., Fewster, R. M. & Lavery, S. D. Geographic concordance of genetic barriers in New Zealand coastal marine species. *Aquat. Conserv. Mar. Freshw. Ecosyst.* **31**, 3607–3625 (2021).

79. Anderson, M. J., Gorley, R. N. & Clarke, K. R. PERMANOVA+ for PRIMER: guide to software and statistical methods. in 214 (Plymouth: Primer-E Ltd., 2008).

80. Mugnai, F. *et al.* Are well-studied marine biodiversity hotspots still blackspots for animal barcoding? *Glob. Ecol. Conserv.* **32**, e01909 (2021).

81. El Ayari, T., Trigui El Menif, N., Hamer, B., Cahill, A. E. & Bierne, N. The hidden side of a major marine biogeographic boundary: a wide mosaic hybrid zone at the Atlantic–Mediterranean divide reveals the complex interaction between natural and genetic barriers in mussels. *Hered. 2019 1226* **122**, 770–784 (2019).

82. Knowles, L. L. The burgeoning field of statistical phylogeography. *J. Evol. Biol.* **17**, 1–10 (2004).

83. Dawson, M. N. Natural experiments and meta-analyses in comparative phylogeography. *J. Biogeogr.* **41**, 52–65 (2014).

84. Arranz, V., Fewster, R. M. & Lavery, S. D. Genogeographic clustering to identify cross-species concordance of spatial genetic patterns. *Divers. Distrib.* **00**, 1–13 (2022).

85. Zizka, V. M. A., Weiss, M. & Leese, F. Can metabarcoding resolve intraspecific genetic diversity changes to environmental stressors? A test case using river macrozoobenthos. *Metabarcoding and Metagenomics* **4**, 23–34 (2020).

86. Sandström, A., Lundmark, C., Andersson, K., Johannesson, K. & Laikre, L. Understanding and bridging the conservation-genetics gap in marine conservation. *Conserv. Biol.* **33**, 725–728 (2019).

Tables

Table 1. PERMANOVA results obtained from the three distance matrices (MOTUs and ESVs based on Bray-Curtis distances and Jost's D distances) using adjacent Regions (southern vs central and central vs northern) and Locality (nested in Region) as factors. The p-values of the permutational

Residuals	16	1.446			

analysis of multivariate dispersions (permdisp) are also given for significant factors.

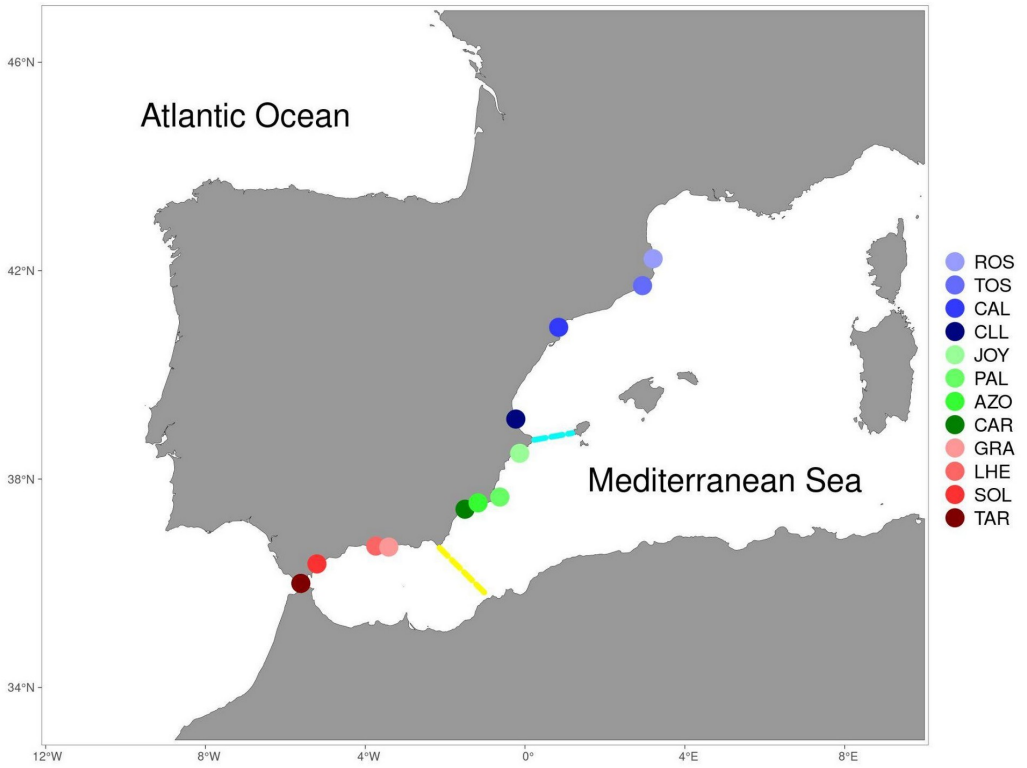
ESVs	df	SS	pseudo-F	p-value	Permdisp p
Region (S vs C)	1	1.255	1.647	0.088	
Locality(Region)	6	4.572	3.596	<0.001	0.352
Residuals	16	3.390			

Region (C vs N)	1	1.080	1.493	0.116	
Locality(Region)	6	4.340	3.674	<0.001	0.786
Residuals	16	3.150			

D	df	SS	pseudo-F	p-value	Permdisp p
Region (S vs C)	1	0.818	3.071	0.011	0.137
Locality(Region)	6	1.599	5.978	<0.001	0.002
Residuals	16	0.713			

Region (C vs N)	1	0.584	2.386	0.033	0.536
Locality(Region)	6	1.469	6.911	<0.001	0.046
Residuals	16	0.567			

Figures

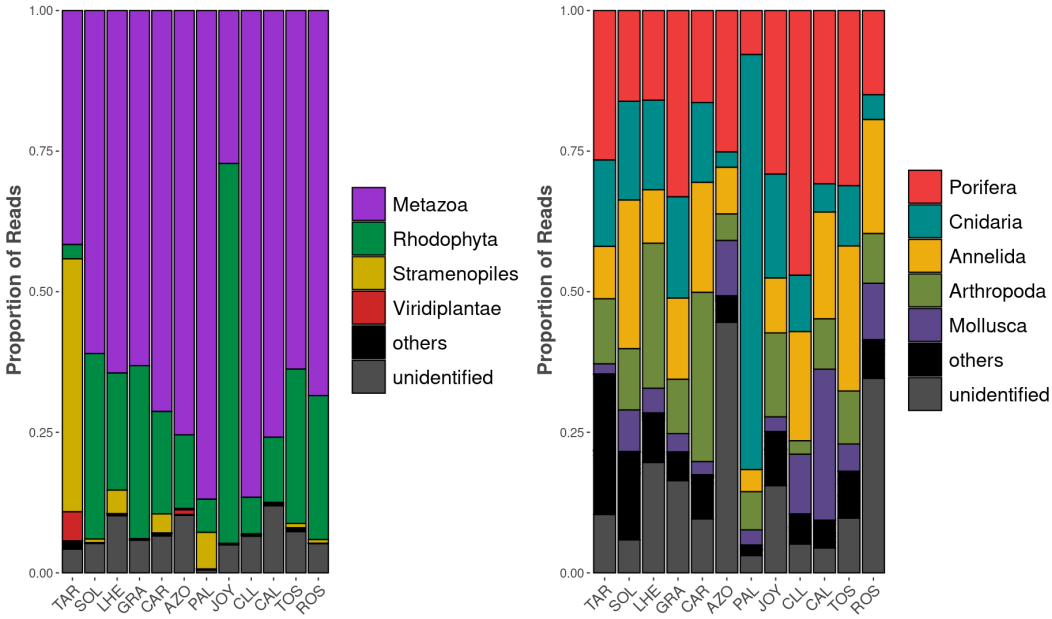


709

710 **Figure 1. Map of the Iberian Mediterranean coast with the sampling localities and the two fronts**
711 **studied : Ibiza Channel (IC, light blue) and Almeria Oran Front (AOF, yellow).**

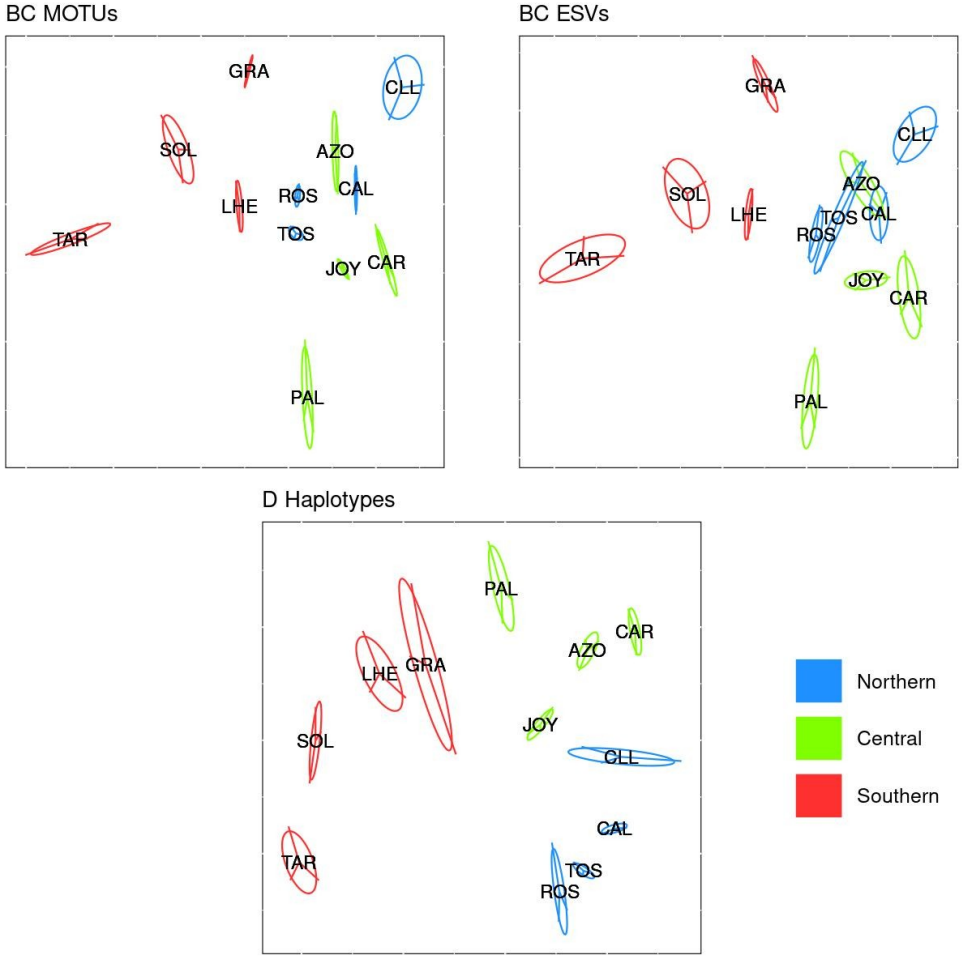
712

713



714

715 **Figure 2. Supergroup (a) and metazoan phyla (b) composition in relative read abundance for each**
716 **locality (averaging the three replicates).**



717

718

719 **Figure 3. Non-metric Multidimensional Scaling of samples using Bray-Curtis dissimilarities for**
720 **MOTUs and ESVs and mean D dissimilarity for haplotypes within MOTUs. Samples grouped by**
721 **locality. Factor region is represented by colours (northern, blue; central, green; southern, red).**

722

723

724

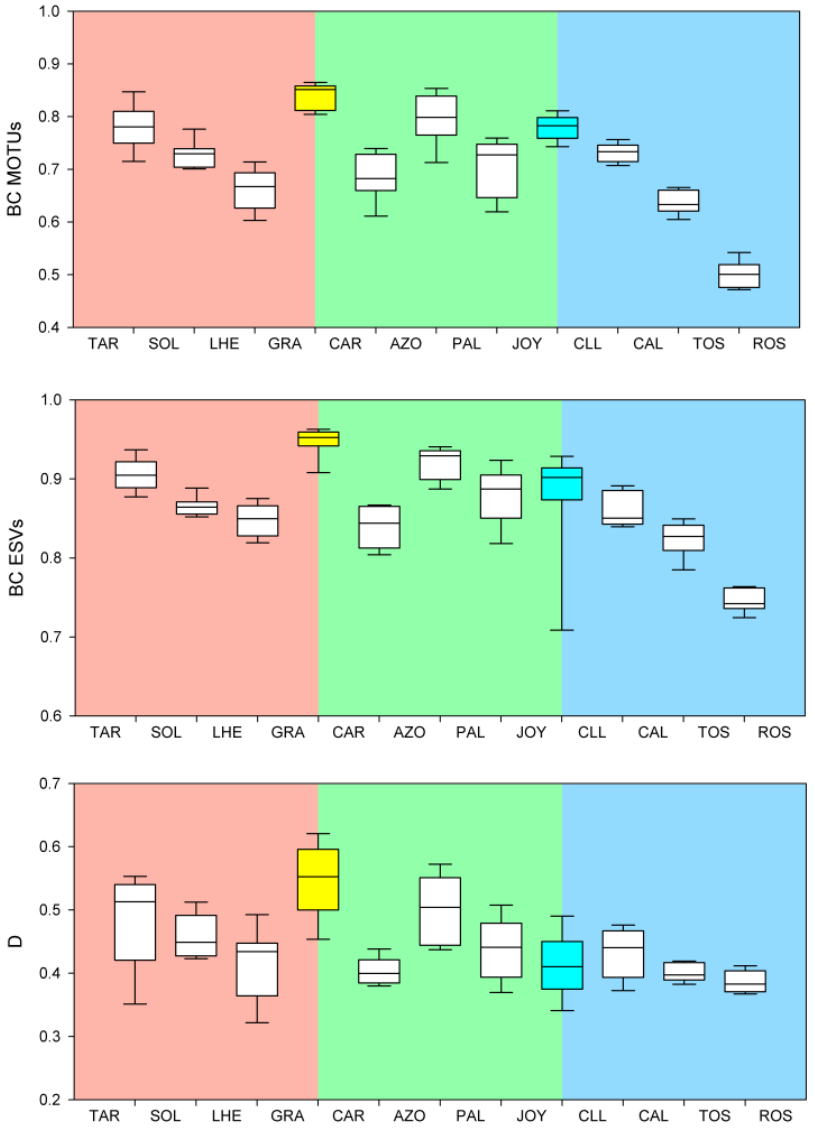


Figure 4. Bray-Curtis dissimilarities of MOTUs (left) and ESVs (central) and average D dissimilarities of haplotypes from adjacent localities. Fronts are represented in yellow for the AOF and light blue for IC.

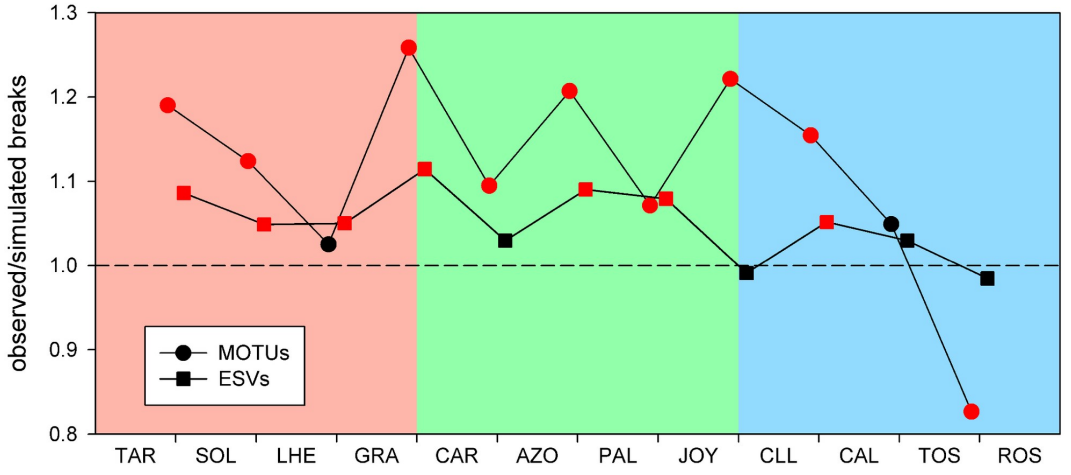


Figure 5. Number of observed vs simulated (through randomization) breaks between each pair of adjacent localities for MOTUs and ESVs. In red are transitions with significantly more (or less) breaks than expected.

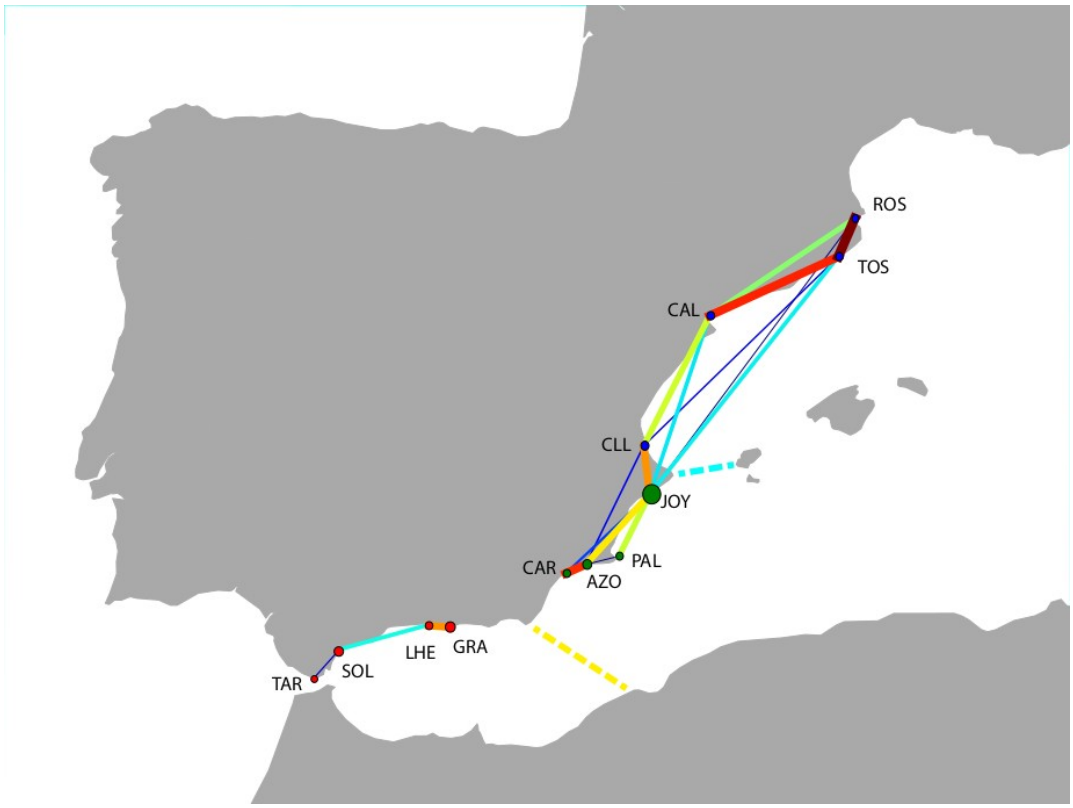
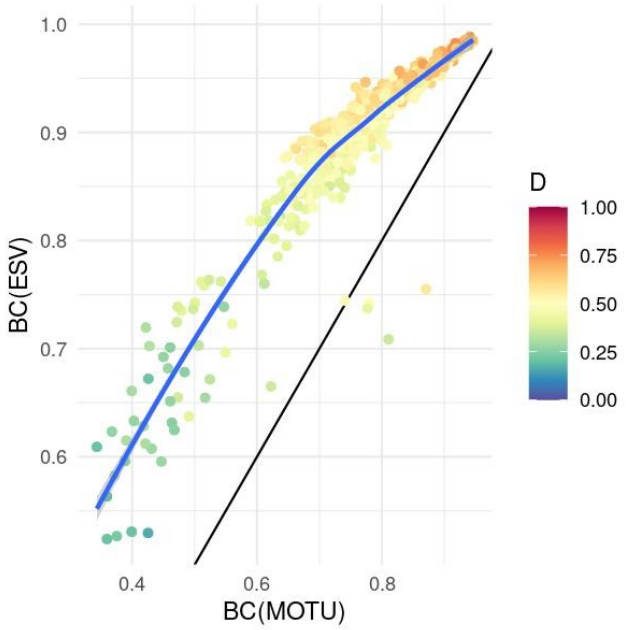


Figure 6. Network analysis using EDENetworks with D values. Width and warmer colours represent stronger connections and thinner and colder colours represent weaker connections. The size of the locality symbols is proportional to the betweenness centrality of the nodes. The two breaks are represented in dashed lines; Almeria-Oral Front (AOF) in yellow and Ibiza Channel (IC) in light blue.



738

739 **Figure 7. Relation between Bray-Curtis dissimilarities of MOTUs (x-axis) and ESV data (y-axis). D**
740 **values computed for the selected MOTUs are represented by colours. Each point represents a pairwise**
741 **comparison between localities and the blue line is the trend line obtained by the 'loess' method of**
742 **ggplot2.**

743