Applications of next-generation sequencing in conservation genomics: kinship analysis and dispersal patterns

Lídia Escoda Assens

Applications of next-generation sequencing in conservation genomics: kinship analysis and dispersal patterns

Lídia Escoda Assens

Doctoral Thesis

Applications of next-generation sequencing in conservation genomics: kinship analysis and dispersal patterns

Lídia Escoda Assens

FACULTAT DE BIOLOGIA
DEPARTAMENT DE GENÈTICA
Programa de Doctorat en Genètica

Applications of next-generation sequencing in conservation genomics: kinship analysis and dispersal patterns

Memòria presentada per
Lídia Escoda Assens
per a optar al grau de Doctora per la
Universitat de Barcelona

Treball realitzat a l'Institut de Biologia Evolutiva (CSIC - Universitat Pompeu Fabra)

Doctoranda
Lídia Escoda Assens
Barcelona, març del 2018

Director
José Castresana Villamor
Investigador Científic
Institut de Biologia Evolutiva
(CSIC - Universitat Pompeu Fabra)

Tutora
Marta Riutort León
Professora Titular
Universitat de Barcelona

AGRAÏMENTS

Aquells que em coneixen saben que mai m'havia plantejat dedicar-me a la ciència. La gent se sorprèn quan els dic que sempre havia volgut estudiar història, i jo em sorprenc de trobar-me ara mateix escrivint els agraïments d'una tesi doctoral en genòmica quan sempre havia afirmat categòricament que la meva vocació eren les lletres. Encara no he tret l'entrellat de com he passat d'un extrem a l'altre, però ara que em trobo a les portes d'acabar el doctorat, puc afirmar que no em penedeixo pas d'haver-me endinsat en el món de la ciència.

Durant aquests tres anys de tesi he trobat un bon grapat de persones que m'han ajudat a créixer tant personalment com professionalment, i a les que m'agradaria dedicar unes ratlles d'agraïment.

Primer de tot al Jose Castresana, per haver confiat en mi abans, durant i després d'aquesta tesi. Gràcies per haver-me animat a seguir en el món de la ciència, per haver-me fet veure que la bioinformàtica no era quelcom a témer i, sobretot, per la seva infinita paciència en tots aquells moments en què he defallit durant aquest anys (que no han estat pas pocs).

A la Marta Riutort, per haver accedit a tutelar aquesta tesi, ajudar-me en el dur món de la burocràcia de la UB i pels seus inestimables comentaris sobre el manuscrit.

Als membres del p59, la meva llar durant aquests anys: a la Marina, per haver estat la primera a acollir-me al laboratori i per haver esdevingut un pilar fonamental durant el darrer any de tesi; al Joan, per haver-me explicat mil i una curiositats científiques; a l'Oliver, per la seva aportació sènior al grup i per i haver-me ajudat a millorar el meu anglès; a la Karla, per haver-nos descobert la generositat de la cultura xilena; i a l'Alfonso, per les nostres desvariejades diàries i per haver-me sofert pacientment aquests mesos d'escriptura. També a l'Ángel Fernández, per les seves idees i contribucions en els articles i per haver-se preocupat d'enviar-me a temps les fotos per a la tesi.

Al Carranza's lab en general, per haver fet de la meva estada a l'IBE molt més amena: al Salvi, per tots els seus consells i les seves constants injeccions d'ànims; al Marc, per compartir amb mi tantes opinions, converses i frustracions de la vida del doctorand; a l'Héctor, per la seva constant aportació cultural al grup; a la Karin, pels seus ànims i la seva paciència amb el meu anglès; i al Luis, per la seva preocupació constant vers als altres. També a l'Adrián, pel seu entusiasme i la seva crucial ajuda en l'edició de vídeos de tesi, i a l'Ana Fernández, pel seu riure contagiós i per haver fet des del primer moment l’estada a l’institut molt més suportable gràcies a les xerrades eternes pels passadissos de l’IBE.

A la Silvia, per totes aquelles converses encoratjadores que han aconseguit convèncer-me que vam fer bé d’escollir fer el doctorat, i juntament amb l’Anna Ullastres, la Míriam, la Vivien i la María per fer-me veure que hi ha vida més enllà de les quatre parets de l’institut.

A l’administració de l’IBE, en especial a la Blanca, l’Anna i el Vicente, per haver-me dedicat el seu temps sempre que he tingut algun problema de contractes, comandes i paperassa vària.

A l’Aida, per haver-hi estat des del primer dia de carrera i perquè sé que hi continuarà sent. A la Mar, la Sílvia, l’Anaís i la Tània, per haver continuat donant-nos suport les unes a les altres, tot i haver-nos dispersat del poble. A la Judit i la Rut, per ser d’aquelles amigues amb qui es pot arreglar el món davant d’una tassa de cafè.

A la meva família, en especial als de Premià. A la tata Montse, per haver-me transmès l’amor per la poesia, per haver-me animat sempre a escriure (espero que escriure sobre ciència també compti!) i per haver-me acollit a casa seva. Al Ferran, per les seves recomanacions literàries, pels seus concerts privats de piano i perquè la redacció de més d’un manuscrit s’hauria fet molt més tediosa sense les Voll-Damm que sempre em reservava quan arribava a casa després d’un dia llarg.

A l’Albert, per ser el millor germà que es pot tenir. Per haver compartit amb mi tantes afeccions i gustos i per les nostres interminables converses sobre gairebé qualsevol tema.

I per últim i més important, als meus pares, el Lluís i la Pinyeres, pel seu suport incondicional, la seva confiança plena en mi i per no haver-me deixat llençar mai la tovallola davant les adversitats. Per haver-me inculcat el valor de l’esforç, de la constància i la importància de la feina ben feta. Gràcies per haver estat tan comprensius durant aquests tres anys de nervis, inquietuds i inseguretats.

A tots vosaltres, gràcies de tot cor.

CONTENTS

I. INTRODUCTION

1. Next-generation sequencing

The appearance of "Next-Generation Sequencing (NGS) Technologies" has revolutionised biological research in general and genomic studies in particular. Since the completion of the human genome project (International Human Genome Sequencing Consortium et al., 2001; Venter et al., 2001), considerable progress has been made in genome sequencing technologies, leading to a decrease in the cost per base and a huge increase in the number of sequences retrieved (Goodwin et al., 2016). Some of the areas that have greatly benefited from the introduction of NGS technologies include the genetic diagnosis of common and rare diseases, cancer research, microbiome studies, personal identification in forensics, metagenomic studies in ecology, and evolutionary and population studies (McCormack et al., 2013; Morey et al., 2013).

Nevertheless, these emerging technologies have given rise to new methodological challenges. Data obtained in this way has higher error rates and shorter read lengths than those of traditional Sanger sequencing platforms, requiring innovative storage and complex bioinformatic analysis pipelines (Goodwin et al., 2016; Mardis, 2011).

1.1. NGS technologies

The NGS workflow has a number of common steps in the different platforms. Library preparation is generally accomplished by random fragmentation of DNA, followed by ligation of adapter sequences. Then, a PCR is performed to generate clonally clustered amplicons that will be employed as sequencing templates. The sequencing process consists in alternating cycles of enzyme-driven biochemistry and imaging-based data acquisition (Shendure & Ji, 2008). There are two approaches for NGS sequencing: sequencing by ligation (SBL), which rely upon the sensitivity of DNA ligase for base-pairing mismatches, and sequencing by synthesis (SBS), which use a DNA polymerase to synthesise a new strand and a signal, such as a fluorophore, identifies the incorporation of a nucleotide into the elongating strand (Goodwin et al., 2016).

Platforms commercially available vary in their use of the above-mentioned approaches, throughput, cost, error profile, and read structure (Table 1). The SOLiD platform uses an SBL approach, while all other sequencing platforms (Roche 454, Ion Torrent, and Illumina) are based on an SBS approach.

Table 1. Summary of NGS platforms.

Platform	Read length	Throughput	Reads	Runtime	Error profile
SOLiD 5500xl	1x50 bp 1x75 bp	160 Gb 240 Gb	1.4 B	~24 hours	0.1%, substitution
Roche 454 GS FLX+	Up to 1 kb	700 Mb	1 M	~23 hours	1%, indel
Ion Torrent S5 (540 chip)	200 bp	10-15 Gb	60-80 M	2.5 hours	1%, indel
Illumina MiSeq v3	2x75 bp 2x300 bp	3.3-3.8 Gb 13.2-15 Gb	22-25 M (SE) 44-50 M (PE)	~21 hours ~56 hours	0.1%, substitution
Illumina NextSeq 550 Mid output	2x75 bp 2x150 bp	16-20 Gb 32-40 Gb	400 M (SE) 800 M (PE)	~15 hours ~26 hours	<1%, substitution
Illumina NextSeq 550 High output	1x75 bp 2x75 bp 2x150 bp	25-30 Gb 50-60 Gb 100-120 Gb	130 M (SE) 260 M (PE)	~11 hours ~18 hours ~29 hours	<1%, substitution
Illumina HiSeq 4000	1x50 bp 2x75 bp 2x150 bp	210-250 Gb 650-750 Gb 1300-1500 Gb	4.3-5 B	1-3.5 days	0.1%, substitution
Illumina HiSeq X (dual flow cell)	2x150 bp	1.6-1.8 Tb	5.3-6 B	< 3 days	0.1%, substitution
Pacific BioSciences RS II	~20 kb	0.5-1 Gb	~55,000	4 hours	10-15%
Oxford Nanopore MinION	Up to 200 kb	1.5 Gb	~350,000	0.5-6 hours	5-10%

M, million; B, billion, bp, base pairs, kb, kilobase pairs; Mb, megabase pairs; Gb, gigabase pairs; Tb, terabase pairs; SE, single-end sequencing; PE, paired-end sequencing.

Illumina is currently the most used NGS platform. In this methodology, DNA is randomly fragmented and adapters are ligated to both ends of the fragments. These adapters are fixed to complementary adapters on a solid plate and PCR bridge amplification is carried out. After cluster formation, fluorescent-labelled nucleotides and the DNA polymerase are added. Each nucleotide incorporated into the new DNA strand is detected and identified by a camera (Figure 1). The terminator with the fluorescent label is then removed and a new cycle starts (Shendure & Ji, 2008). The main concern with this platform is the sample loading control, as overloading can result in overlapping clusters and poor sequencing quality. Nucleotide substitution due to a bad identification of the nucleotide incorporation is the most common type of error in this platform (Kchouk et al., 2017).

Illumina accounts for the largest market share with a wide range of instruments: from small, low-throughput benchtop units to large ultra-high-throughput instruments. The MiSeq and HiSeq are some of the most established platforms, both using a four-channel strategy in which each nucleotide is labelled with a different fluorophore. The

NextSeq platform was later introduced, and it employs a two-colour labelling system that increases speed and reduces costs by reducing scanning to two colour channels and reducing fluorophore usage. However, this system results in a higher error rate, requiring a larger number of reads for assemblies (Goodwin et al., 2016). The most recent platform is the HiSeq X Ten System, which is capable of outputting 1.8 Tb in 3 days or 18,000 human genomes at 30x coverage per year (Reuter et al., 2015).

Other technologies, known as "Third Generation Sequencing (TGS) Technologies", are starting to appear. Their main advantage is that they can sequence single molecules of DNA with no need for clonal amplification prior to sequencing, avoiding the introduction of PCR artefacts (Morey et al., 2013). These technologies are also able to produce long reads of several kilobases of great interest for the assembly and resolution of repetitive regions in complex genomes (Kchouk et al., 2017). Among the most used platforms are the Pacific Biosciences single-molecule real-time sequencer and the MinION device from Oxford Nanopore Technologies (Table 1).

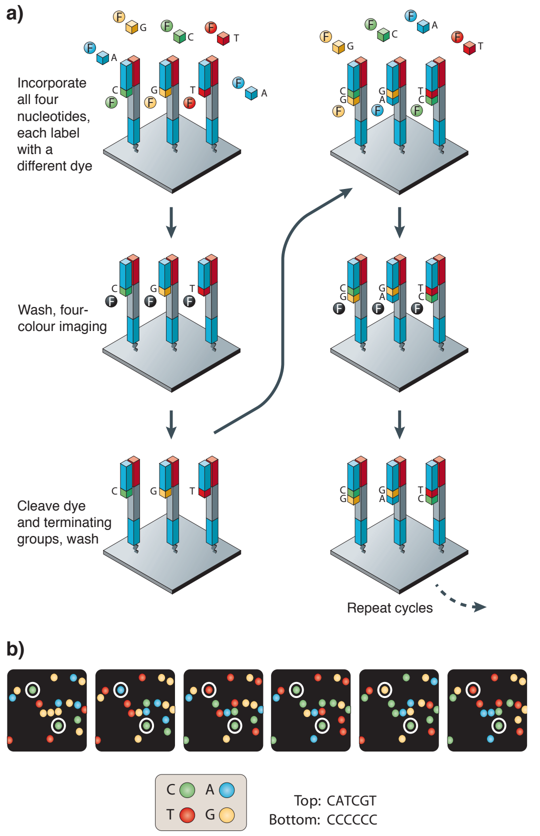
a)
Incorporate all four nucleotides, each label with a different dye
Wash, four-colour imaging
Cleave dye and terminating groups, wash
Repeat cycles
b)
C
A
T
G
Top: CATCGT
Bottom: CCCCCC

Figure 1. Illumina sequencing. (a) After bridge amplification, a mixture of primers, DNA polymerase and fluorescent-labelled nucleotides are added to the flow cell. In each sequencing cycle, DNA fragments incorporate only one nucleotide. The unincorporated bases are washed away and the slide is imaged. The dye is then cleaved and the 3'-OH is regenerated to start a new cycle. (b) The four-colour images highlight the sequencing data from two clonally amplified templates (from Metzker (2010)).

1.2. Library construction for NGS

Next-generation sequencing technologies can be used for whole-genome sequencing, for sequencing a fraction of a genome, for sequencing the transcriptome of an individual with RNAseq, or for sequencing multiplexed PCR products (McCormack et al., 2013). Regardless of the sequencing technology used, a DNA library must be first constructed. The basic steps for any library construction are the fragmentation of DNA, the ligation of adapters, the size-selection of the fragments, a PCR amplification of the fragments, and the purification and quantification of the final library (Linnarsson, 2010).

Fragmentation of genomic DNA can be achieved either by mechanical force (i.e. sonication or nebulisation) or by enzymatic digestion. After this, fragment ends are repaired and specific upstream and downstream adapters are ligated to them. Then, a size-selection step is performed to select DNA molecules in the desired size range. Size selection is performed by, for example, agarose gel electrophoresis or with solid-phase reversible immobilisation beads (Linnarsson, 2010; van Dijk et al., 2014).

Adapter-ligated fragments are amplified by PCR to generate sufficient quantities of template DNA and to add additional adapter sequences that are needed for sequencing. PCR amplification may introduce bias in sample composition as not all fragments in the mixture are amplified with the same efficiency. To minimise it, PCRs should be performed with as few cycles as possible (van Dijk et al., 2014). Finally, the library is purified to remove free adapters and adapter dimers that could have been generated during the PCR. For this, magnetic bead-based clean-up or agarose gel purification can be used (Head et al., 2014).

The library must be quantified, as the sequencing instruments are highly sensitive to its molar concentration. Too low or too high concentrations may lead to a waste of the sequencer capacity or an overlap between detection sites, respectively. For quantification, either a fluorescent approach or qPCR can be used. Once the mass concentration of the library is obtained in ng/µl, it can be converted to molar concentration (nM) with the average fragment length. The distribution of fragment sizes and the overall quality of the library can be estimated using capillary electrophoresis from a diagnostic agarose gel such in Agilent BioAnalyzer (Linnarsson, 2010).

The use of barcodes or indexes, which are short identifying DNA sequences, allows sample multiplexing. In-line barcodes occur at the end of the adapter that is immediately adjacent to the genomic DNA fragment and will appear at the beginning of the sequence reads. On the other hand, indexes are unique sequences of 6-8 bp length added at the PCR stage that do not appear in the sequence reads (Andrews et al., 2016). Many techniques use a combinatorial approach, in which DNA fragments from each sample are identified by a unique combination of in-line barcodes and Illumina indexes.

Whole-genome sequencing of large numbers of individuals is still unaffordable, especially for non-model organisms that do not have a reference genome. Sequencing a complex genome still has a great cost and take a large amount of time. To obtain genome information from a large number of samples at a reasonable cost, several methods of genome reduction and genome enrichment have been developed (Mamanova et al., 2010)

1.2.1. Genome reduction methods

Genome reduction methods involve the digestion of genomic DNA with one or more restriction enzymes, the selection of the resulting restriction fragments, and the sequencing of the final set of fragments. This type of methods is ideal for non-model organisms without a closely related reference genome, as no prior genomic information is required (Davey et al., 2011).

In Restriction-site Associated DNA sequencing (RADseq), DNA is digested using a chosen restriction enzyme that produces sticky end overhangs. Barcoded adapters are ligated to these overhangs to allow identification of each individual pooled in the library. Then, samples from different individuals are pooled and the DNA fragments are randomly sheared to a length suitable for the sequencing platform. Fragments are size-selected from an agarose gel and amplified by PCR (Davey et al., 2011).

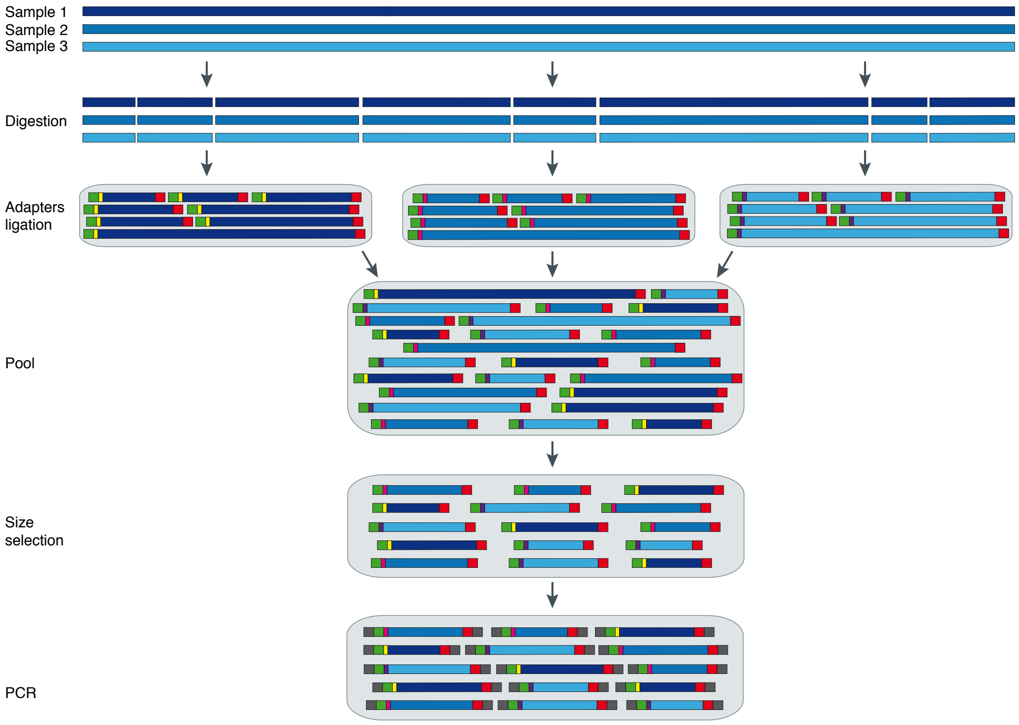

Figure 2. ddRAD protocol. Genomic DNA is digested with two different restriction enzymes. Fragments of each sample are ligated to P1 (green) and P2 (red) adapters. P1 adapters contain a barcode sequence different for each sample (yellow for sample 1, magenta for sample 2, and purple for sample 3). Adapter-ligated fragments of all samples are pooled and size-selected. Finally, the fragments are amplified by PCR using P1- and P2-specific primers, which add the platform-specific adapters (grey) needed for the sequencing reaction (modified from Davey et al., (2011)).

One of the RADseq methods that has rapidly popularised in the last few years is the Double Digest RADseq method, commonly known as ddRAD (Peterson et al., 2012). This method eliminates random shearing of genomic DNA by using simultaneously two different restriction enzymes with adapters specific to each enzyme. Then, a fraction of genomic fragments is size-selected, amplified by PCR, and sequenced (Figure 2).

1.2.2. Genome enrichment methods

Genome enrichment methods (also known as target enrichment or sequence capture methods) involve the selective capture of genomic regions using DNA or RNA probes that hybridise to the targeted genomic DNA. These methods require some prior genomic information to design the probes (McCormack et al., 2013). Some of the most widely used approaches are multiplex PCR, hybrid capture, and molecular inversion probes.

In multiplex PCR, several primer pairs are used in a single reaction, generating multiple amplicons. This approach has several limitations, including biased representation of some products and chimeric DNA sequences (Mamanova et al., 2010). Some alternatives of multiplex PCR have emerged, such as microdroplet PCR, where millions of PCR reactions occur in picoliter-sized droplets (Tewhey et al., 2009).

In enrichment by hybridisation, a DNA library is hybridised with probes that are complementary to the target region. Non-specific hybrids are removed and the targeted DNA obtained is eluted for sequencing. Compared with multiplex PCR, enrichment methods can capture a larger number of targets with more homogeneous coverage. There are two alternative methods: array hybridisation and in-solution hybridisation. In-solution hybridisation is more scalable than array hybridisation and requires much less starting DNA and no specialized equipment (Mamanova et al., 2010; Morey et al., 2013). Several commercial kits for in-solution hybridisation have been developed, such as MYbaits enrichment kit (MYcroarray) and SureSelect kit (Agilent).

Molecular Inversion Probes (MIPs) are single-stranded DNA molecules consisting of a common linker flanked by target-specific sequences. Following hybridisation of MIPs to the target, the generated gaps are filled by a polymerase and the double-stranded DNA molecules are circularised by a ligase. Uncircularised molecules are digested by exonucleases and circularised DNA molecules are amplified by PCR. Products can be pooled using barcodes nested in the post-capture amplification primers. The main disadvantage of this technique is that it has a low uniformity in coverage due to inefficiencies in the capturing reaction itself (Mamanova et al., 2010), although an adequate probe selection may help to circumvent this problem (Niedzicka et al., 2016).

1.3. Bioinformatic analysis of NGS data

The massive amount of data produced by NGS presents a significant challenge for data storage, analysis, and management that requires advanced bioinformatic tools. Analysis of NGS data include de-multiplexing and trimming of barcodes, filtering of reads based on sequence quality, alignment and assembly of sequences, and genotype and Single Nucleotide Polymorphism (SNP) calling (McCormack et al., 2013).

1.3.1. Filtering and quality control

The first step in processing raw NGS data is eliminating low quality reads. During the sequencing process, base-calling algorithms infer nucleotide information and produce per-base quality scores from image analysis (Nielsen et al., 2011). Nucleotide sequences and quality scores associated with them are stored in FASTQ files. Quality is represented with the Phred quality score (Ewing & Green, 1998), given by:

$$Q_{Phred} = -10 \log_{10} P(error),$$

where P is the estimated error probability for that base-call. For example, a base-call having a probability of 1/1,000 of being incorrect is assigned a quality value of 30. Sequences with mean quality smaller than a given value can be discarded.

The second step is to de-multiplex the sequences using their barcodes. Then, primer and barcode sequences are removed from the flanks of the reads.

1.3.2. Alignment and assembly

Calling genotypes requires alignments of homologous reads. There are two types of alignment methods, those that use a reference (either a reference genome or information on the output reads such as the probe sequences used for target enrichment) and *de novo* assemblies. Alignment results are saved in SAM format. Its compressed format, the BAM format, is easier processed in subsequent bioinformatic analysis.

Most alignment algorithms of NGS data against a reference genome are based on either "hashing" or an effective data compression algorithm called the "Burrows-Wheeler transform" (BWT). BWT-based aligners, such as BOWTIE (Langmead & Salzberg, 2012) and BWA (Li & Durbin, 2009), are fast and memory-efficient and particularly useful for aligning repetitive reads, but they tend to be less sensitive than hash-based algorithms.

In the case of not having a reference, *de novo* assemblers are used. If reads belong to a whole-genome sequencing experiment, programs like SOAPDENOVO2 (Luo et al., 2012), ABYSS (Simpson et al., 2009), or ALLPATHS (Butler et al., 2008) are used to assemble short reads into longer contigs or scaffolds. If the reads to analyse come from a genome reduction experiment, programs such as STACKS (Catchen et al., 2011) can be used, in which the reads are collected into groups within a given percent similarity and alignments are generated from these (Figure 3).

The accuracy of the alignment is crucial in variant detection, as incorrectly aligned reads may lead to errors in SNP and genotype calling. For this, alignment algorithms should be able to cope with sequencing errors and potentially real differences between the reference genome and the sequenced genome. The optimal choice of the tolerable number of mismatches between the sequences aligned and the reference genome may differ between different organisms (Nielsen et al., 2011).

1.3.3. Genotype and SNP calling

The last step of raw NGS data analysis consists in detecting variant positions in the assembled sequences and is divided into two steps: SNP calling and genotype calling. SNP calling (or variant calling) is the process of determining polymorphic positions and genotype calling is the process of determining the genotype of each individual for polymorphic positions (Nielsen et al., 2011).

The simplest method for performing genotype and SNP calling is by counting alleles at each site and using threshold values for calling a SNP or genotype. More recent methods incorporate uncertainty in a probabilistic framework, in which additional information regarding allele frequencies and patterns of linkage disequilibrium can be incorporated (McCormack et al., 2012; Nielsen et al., 2011).

There are several programs available for SNP and genotype calling, such as SAMTOOLS and BCFTOOLS (Li et al., 2009), GATK (McKenna et al., 2010), and STACKS (Catchen et al., 2011). SAMTOOLS is a package for manipulating NGS alignments in SAM format, which includes the computation of genotype likelihoods (SAMTOOLS) and SNP and genotype calling (BCFTOOLS). GATK has many options for improving the quality of SNP calling in genomes, and STACKS is commonly used for genotype calling from ddRAD data (Figure 3). These programs allow the generation of result files in several formats such as PED and VCF for further analyses.

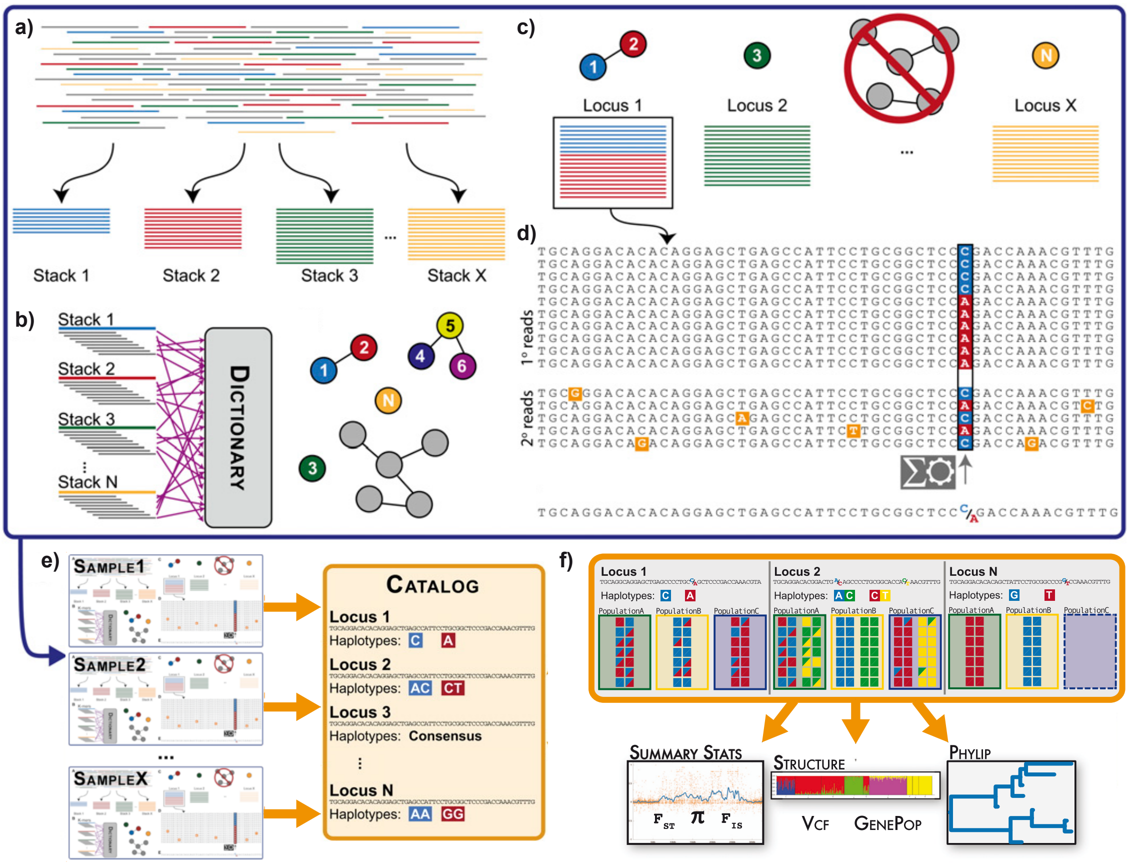

Figure 3. STACKS pipeline. (a) The PROCESS RADTAGS program de-multiplexes and cleans the reads and the USTACKS program forms stacks from reads that match exactly. (b) USTACKS breaks down the sequence of each stack into k-mers and loads them into a Dictionary, which uses to create a list of potentially matching stacks. (c) USTACKS merges matched stacks to form putative loci (d) and then matches secondary reads that were not initially placed in a stack against putative loci to increase stack depth, checks each locus for polymorphisms, calls a consensus sequence, and records SNP and haplotype data. (e) The CSTACKS program loads stacks from all individuals to create a catalogue of loci and the SSTACKS program matches loci from each individual against the catalogue to determine the allelic state at each locus in each individual. (f) The POPULATIONS program tabulates the state of loci within and among population, calculates population genetics statistics and exports to a number of additional formats (modified from Catchen et al. (2011) and Catchen et al. (2013)).

1.3.4. Sources of genotyping errors and bias

Genotyping errors occur when the inferred genotype of an individual does not correspond to its real genotype. Errors can be generated at every step of the genotyping process and can be due to human or equipment error, to the PCR, or to the DNA quality (Pompanon et al., 2005).

Human errors include the contamination of the samples with exogenous DNA, cross-contamination between samples, and errors in data handling (Pompanon et al., 2005), while equipment errors are mainly due to an incorrect base calling during the sequencing process (Nielsen et al., 2011).

Errors introduced by the PCR include *Taq* polymerase errors, which can cause false alleles. In the case of reduced-representation libraries, mutations in the DNA sequence close to a marker in enrichment methods can lead to the failure in the amplification of the affected allele, while mutations at a restriction enzyme recognition site in RADseq methods can lead to a failure to cut the genomic DNA at that location. In both cases, the affected alleles are not sequenced and are therefore null alleles.

Low quantity or quality of DNA, where a small number of target DNA are available due to an extreme dilution of the DNA or from degradation, can favour allelic dropouts, which are the preferential amplification of one allele over the other (Pompanon et al., 2005).

Finally, PCR duplicates are sequence reads resulting from sequencing two or more copies of the same DNA molecule in the library (Ebbert et al., 2016). PCR duplicates can interfere in some post-sequencing analyses and therefore may be eliminated for certain applications.

Genotyping errors, if they are abundant and biased, can have a strong impact on the conclusions derived from any study. For example, they can be very important in biodiversity research and particularly in conservation biology, as population size estimations based on the genotypes identified from non-invasive samples can be overestimated or parentage analysis can be affected by generating incorrect paternity or maternity exclusions (Pompanon et al., 2005).

2. Applications of NGS in conservation genomics

Conservation genomics can be defined as the use of new genomic techniques to solve problems in conservation biology. These new genomic techniques allow raising the number of neutral loci screened, increasing the power and accuracy of estimating important parameters in conservation by genotyping hundreds to thousands of markers in numerous individuals. In other cases, the large amount of sequence data allow estimations that were not possible before with a small number of markers such as the determination of kinship relationships that include distant relatives and the estimation of inbreeding coefficients (Allendorf et al., 2010).

2.1. Improved analysis of adaptation, genetic structure and diversity with NGS data

Genomic analyses can be used to improve the precision of recent and historical demographic events, genetic variation, and population structure estimations in wild species. It is especially important in the case of threatened species, for which natural selection and inbreeding can be inferred at the genome level. Genome-wide data can also improve designation of conservation units for protection, management, and recovery of endangered species (Steiner et al., 2013). It is important to take under consideration that markers are usually assumed to be independent in this kind of analyses, so as the number of markers increases, linkage disequilibrium should be taken into account (Allendorf et al., 2010).

Genomic approaches may allow the identification of adaptive genetic variation related to key traits, which is important for knowing the species fitness and population viability. Quantitative trait loci (QTL) mapping is frequently used to identify genetic regions associated with phenotypes, but it cannot be applied to species in which crosses are infeasible or from which pedigree data are not available. In these cases, genome-wide association studies (GWAS) can be used to detect them (Steiner et al., 2013). For example, Poissant et al. (2012) used a QTL mapping approach to study several fitness-related traits in wild pedigreed individuals of bighorn sheep.

Inference of recent and historical demographic events needs accurate estimates of genomic variation and effective population size. Genome-wide data can allow more precise estimates of the timing and extent of population bottlenecks and expansions, the estimation of current and ancestral population sizes, recombination rates, and speciation times between closely related species. For example, Locke et al., (2011) found great differences in genetic variation and evolutionary histories between Sumatran and Bornean orang-utans using allele frequency spectra analyses from genomic data.

Population structure and gene flow between different populations can be inferred with genome-wide data. Population structure analyses from large SNP data sets can be performed, for example, with principal component analysis (PCA) or with clustering methods such as STRUCTURE (Pritchard et al., 2000). For example, Riley et al., (2006) used population structure analyses to study the effect of a highway to dispersal of two carnivore species and found that even some individuals were able to cross the highway, it had an effect on the gene flow between the populations at both sides of the barrier.

Genetic diversity is critical in endangered species to maintain population fitness when facing threats to the survival of the species, such as loss of habitat, environmental changes, and diseases. Hendricks et al., (2017) studied the genetic diversity in the Tasmanian devil, an endangered species that has recently experienced a dramatic population decline due to an infectious disease, and found low genetic diversity throughout its geographic range. However, inbreeding provides a more direct assessment of genetic health (Keller & Waller, 2002).

Estimation of relatedness among individuals, individual inbreeding coefficients, and pedigree reconstruction becomes feasible in wild populations with the analysis of hundreds of loci. Methods for estimating these parameters and inferring pedigrees are explained in Introduction section 3.

2.2. Prospects from whole-genome sequencing

The number of sequenced genomes of non-model organisms has greatly increased in the last few years, contributing to more robust inferences in studies of adaptation, trait evolution, and demographic events in wild populations (Ellegren, 2014).

Detection of adaptive evolution can be achieved by comparing genome sequences from closely related species. For example, Qiu et al., (2012) compared the genome of the domestic yak with the cattle genome to study animal adaptation to high altitude and found adaptive evolution of several genes in yaks involved in the cellular response to hypoxia.

Genome sequence data from a single individual allows the inference of the past demography of the species. Using a coalescent-based method, the pairwise sequentially Markovian coalescent (PSMC) method, past demography can be inferred from the local density of heterozygous sites in diploid data (Heng Li & Durbin, 2011). The PSMC approach has already been applied to unravel the demographic history of several species, such as pandas (Zhao et al., 2013) and bears (Miller et al., 2012).

Genomic data can also provide crucial information for effective conservation of endangered species. For example, Benazzo et al., (2017) explored patterns of genomic variation and divergence, estimated inbreeding, and detected adaptation and maladaptation in the Apennine brown bear, an isolated population with a very small population size. Abascal et al., (2016) also used genomic data to infer the demographic history of the Iberian lynx and identified a series of severe population bottlenecks that had greatly impacted its genome evolution. They also found multiple signatures of genetic erosion and a low genome-wide genetic diversity.

3. Kinship analysis and pedigree reconstruction

Knowledge of the genealogical relationships among individuals in a population is essential to many studies in population genetics and conservation. For example, they provide information about kin selection and cooperative breeding, trait heritability, social behaviour and structure, and gene flow (Taylor, 2015). Methods of kinship analysis include relatedness estimation and assignment of pairs of individuals into categories of kinship. Relatedness (r) is a continuous measure of the identity-by-descent between individuals, while kinship categories are specific pedigree relationships such as parent-offspring or full siblings (Blouin, 2003).

3.1. Kinship coefficient

Characterisation of the relatedness between individuals is based on the probability that their alleles have descended from a single ancestral gene, that is the probability that they are identical-by-descent (IBD). Two individuals are related if they have alleles that are IBD (Weir et al., 2006).

The kinship coefficient or coancestry coefficient (θ_{XY}) of two individuals is the probability that two alleles, each randomly chosen from each individual at the same locus, are identical-by-descent (Blouin, 2003; Jacquard, 1972; Weir et al., 2006). This probability can range from 0 for unrelated individuals to 1 for two identical and completely inbred twins. The relatedness coefficient (r_{XY}), which is two times the kinship coefficient, is commonly used to describe the degree of shared ancestry (Milligan, 2003; Wang, 2011).

Relatedness can be estimated using genealogical relationships from a complete pedigree or using molecular markers, such as microsatellites or SNPs. Quantification of relatedness has been widely used in different aspects of wildlife studies such as the analysis of social organization and philopatry (Arora et al., 2012; Bonin et al., 2012; Watts et al., 2011).

3.1.1. Estimation of the kinship coefficient from pedigrees

Given a pedigree, the relatedness coefficient between two individuals X and Y can be calculated with the formula:

$$r_{XY} = \sum \left(\frac{1}{2}\right)^n$$

where n is the number of direct connecting links separating both individuals.

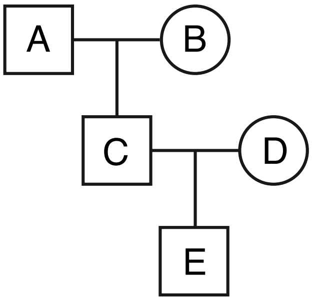

Figure 4. Pedigree showing an outbred mating. Males are represented with squares and females with circles.

Taking as example the pedigree from Figure 4, relatedness between different individuals can be calculated by counting the direct links between them. For example, for calculating the relatedness between grandparent A and grandchild E (r_{AE}), direct paths or connections between them should be counted, 2 in this case. Then, relatedness can be calculated by applying the formula above:

$$r_{AE} = \left(\frac{1}{2}\right)^2 = 0.25$$

Inbreeding, caused by matings between related individuals, increases the relationship between two individuals above the theoretical values. An example to this is the relationship between individual F from Figure 5, resulting from a mating between two half-siblings, and individual B, which is the grandmother of individual F through both maternal and paternal lines. For calculating relatedness between these two individuals (r_{BF}), both possible paths must be taken into account. Therefore,

$$r_{AF} = \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 = 0.5$$

showing a higher relatedness value than the expected for a grandparent-grandchild relationship of 0.25.

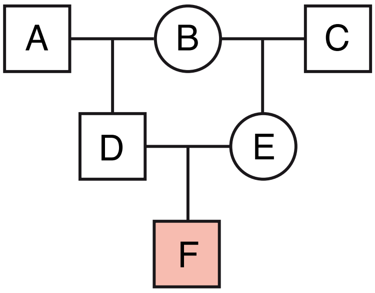

Figure 5. Pedigree showing a mating between two half siblings. Males are represented with squares, females with circles, and inbred individuals are represented in light red.

The main problem with pedigree-based relatedness is that founder individuals are assumed to be outbred and unrelated, which is not always true. The lack of a complete or an ideal pedigree makes that the resulting kinship values deviate from the theoretical values (Speed & Balding, 2015).

3.1.2. Estimation of the kinship coefficient with molecular markers

In the absence of an accurate pedigree, relatedness can be estimated from the genotypes of individuals. When comparing the two alleles at a locus for two individuals X and Y, there exist 15 IBD states among the four alleles if those alleles coming from the mother and those alleles coming from the father can be differentiated for each individual (Jacquard, 1972). When the network of ancestry between the two individuals is known, it is possible to determine the probability δ_i of each of the IBD states S_i (Figure 6).

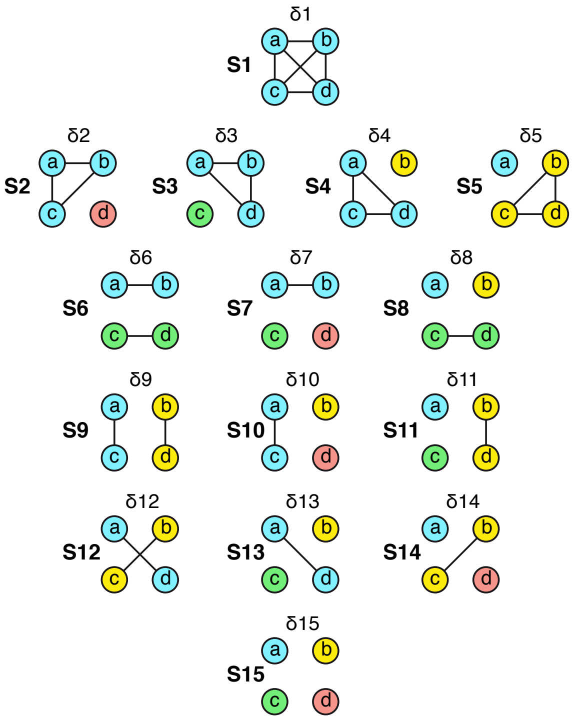

Figure 6. Detailed identity states. Given two individuals X and Y with alleles at a certain autosomal locus *a*/*b* and *c*/*d*, respectively, in which allele *a* has been transmitted to X by his father, *b* has been transmitted to X by his mother, *c* has been transmitted to Y by his father, and *d* has been transmitted to Y by his mother, 15 different configurations of identity-by-descent can be found between them. Each IBD configuration (denoted by S1~S15) has a corresponding probability (denoted by δ1~δ15). Alleles that are IBD are represented in the same colour and linked by lines, while alleles that are not IBD are represented in different colours and unlinked.

It is generally neither possible nor necessary to distinguish between paternal and maternal alleles, leading to a condensation of these 15 IBD states to 9 (Figure 7).

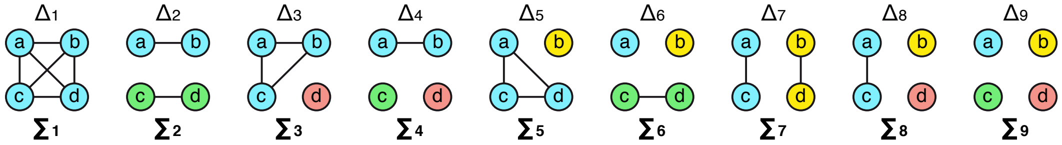

Figure 7. Condensed identity states. Given two individuals X and Y with alleles at a certain autosomal locus *a*/*b* and *c*/*d*, respectively, and no information about paternal origin of the alleles, 9 different configurations of identity-by-descent can be found between them. Each IBD configuration (denoted by Σ1~ Σ9) has a corresponding probability (denoted by Δ1~ Δ9). Alleles that are IBD are represented in the same colour and linked by lines, while alleles that are not IBD are represented in different colours and unlinked.

Given the probability of each of the nine IBD states for individuals X and Y, the kinship coefficient between them is:

$$\theta_{XY} = \Delta_1 + \frac{1}{2}(\Delta_3 + \Delta_5 + \Delta_7) + \frac{1}{4}\Delta_8,$$

where Δ_i is the probability of IBD state i (=1~9) with $\sum_{i=1}^{9} \Delta_i = 1$.

For non-inbred individuals, the 9 IBD states are collapsed to a set of three k coefficients: k_0, k_1, and k_2, which are the probabilities of having zero, one or two IBD alleles, respectively. These three coefficients correspond to the last three states of Figure 7:

$$k_0 = \Delta_9, \qquad k_1 = \Delta_8, \qquad k_2 = \Delta_7,$$

and the kinship coefficient in this case corresponds to

$$\theta_{XY} = \frac{1}{2}k_2 + \frac{1}{4}k_1.$$

The three IBD probabilities and theoretical relatedness coefficients for some common relationships are listed in Table 2.

Table 2. IBD probabilities for common, non-inbred relatives.

Relationship	k_2	k_1	k_0	r
Monozygotic twins, self	1	0	0	1
Parent-offspring	0	1	0	0.50
Full siblings	0.25	0.50	0.25	0.50
2nd degree	0	0.50	0.50	0.25
3rd degree	0	0.25	0.75	0.125
Unrelated	0	0	1	0

2nd degree relationships include half-sibling, grandparent-grandchild and avuncular relationships; 3rd degree relationships include first cousins, great-grandparent-great-grandchild, etc.

Genetic relatedness based on molecular markers can be estimated by methods of moments or by maximum likelihood. Methods of moments (Li et al., 1993; Lynch & Ritland, 1999; Queller & Goodnight, 1989; Ritland, 1996; Wang, 2002) are based on allele frequency moments. They were introduced earlier as an approximation for these estimations. In contrast, maximum-likelihood methods use the model with 9 different IBD states described above; they allow the estimation of both relatedness and inbreeding coefficients but need more data for the estimations (Milligan, 2003; Wang, 2007). These methods need a population background to estimate allele frequencies in the studied population, which are necessary for discriminating alleles shared between two individuals that have been inherited from the same ancestor (identical by descent), from those that two individuals share by chance because they are common in the population (identical by state).

3.2. Inbreeding coefficient

The inbreeding coefficient of an individual (F_X) is the probability that two alleles at a locus are IBD and is equivalent to the kinship coefficient of its parents (Jacquard, 1975). As with the relatedness coefficient, the inbreeding coefficient can either be estimated from a pedigree or using molecular markers.

3.2.1. Estimation of the inbreeding coefficient from pedigrees

Given a pedigree, the inbreeding coefficient of individual X can be calculated with the formula

$$F_X = \sum\left[\left(\tfrac{1}{2}\right)^{n+1}(1 + F_Y)\right],$$

where n is the number of connecting links between the two parents of X through common ancestors and F_Y is the coefficient inbreeding of the common ancestor Y.

Taking as example the pedigree from Figure 5, and assuming that the common ancestor B is not inbred (i.e. $F_B = 0$), the inbreeding coefficient of individual F (F_F), offspring resulting from the mating of two half siblings, can be calculated applying the formula above:

$$F_F = \left(\frac{1}{2}\right)^{2+1} = 0.125$$

In the case of an inbred individual resulting from the mating of two full siblings, as in the case of individual E from pedigree in Figure 8, the two common ancestors, A and B, must be considered and therefore the inbreeding coefficient of the individual is:

$$F_E = \left(\frac{1}{2}\right)^{2+1} + \left(\frac{1}{2}\right)^{2+1} = 0.25$$

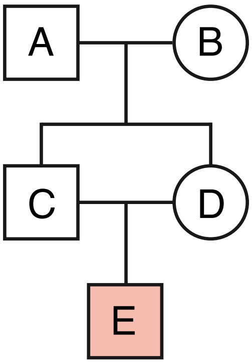

Figure 8. Pedigree showing a mating between two full siblings. Males are represented with squares, females with circles, and inbred individual is represented in light red.

3.2.2. Estimation of the inbreeding coefficient with molecular markers

In the absence of a pedigree, the inbreeding coefficient can be estimated from the genotypes of individuals using the 9 IBD states from Figure 7. Given the probability of each of the 9 IBD states for individuals X and Y, the inbreeding coefficients of each of them are:

$$F_X = \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4,$$

$$F_Y = \Delta_1 + \Delta_2 + \Delta_5 + \Delta_6.$$

The inbreeding coefficient can be estimated by maximum-likelihood estimators (Milligan, 2003; Wang, 2007) as indicated above, as well as by two moment estimators (Lynch & Ritland, 1999; Ritland, 1996).

3.3. Methodological issues in the estimation of relatedness and inbreeding coefficients

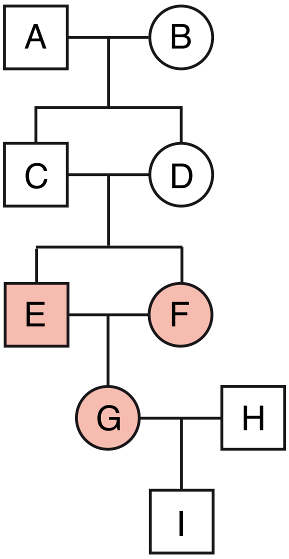

Figure 9. Pedigree showing several consecutive matings between full siblings. Males are represented with squares, females with circles, and inbred individuals are represented in light red.

As mentioned before, inbreeding increases the relatedness coefficient between two individuals above the standard values. This effect can be easily seen by using a pedigree with inbred matings, as in the pedigree from Figure 9. To recreate it, actual genotypes based on 912 SNPs of real individuals from a study (Escoda et al., 2017) were used for founder individuals A, B, and H. The rest of individuals were obtained by crossing the former individuals via computer simulations. Then, relatedness coefficients for all dyads, as well as inbreeding coefficients for all individuals, were estimated using a maximum-likelihood estimator (Milligan, 2003). By performing 100 simulations of the pedigree, distributions for each coefficient were obtained and compared with the expected values according to the theory outlined above.

The relatedness coefficient between two full siblings is 0.5 in absence of inbreeding, as in the case of individuals C and D of the pedigree in Figure 9 (Figure 10a). If the siblings tested are the product of a mating between two full siblings, as in the case of individuals E and F, the relatedness coefficient between them will be higher than the expected 0.5, as the simulations show (Figure 10b).

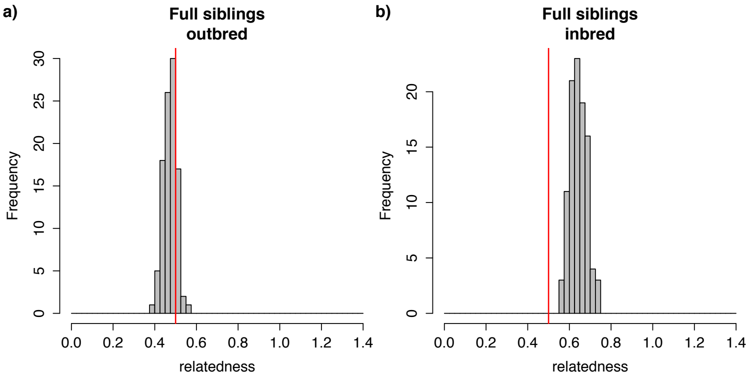

Figure 10. Relatedness coefficients between full siblings from pedigree in Figure 9. (a) Individuals C and D, resulting from a mating between unrelated individuals. (b) Individuals E and F, resulting from the mating between two full siblings.

Figure 11 shows how the relatedness coefficient increases with consecutive inbred matings between full siblings from the pedigree in Figure 9. The relatedness coefficient for a parent-offspring dyad in absence of inbreeding is 0.5, a value that shows very low standard deviation in the estimated values when enough markers are used for parent-offspring dyads (Figure 11a); if there is a mating between two full siblings, this value increases (Figure 11b); and if more inbred matings between full siblings occur in consecutive generations, the relatedness coefficient between parent-offspring pairs continues increasing (Figure 11c), as again shown in the simulations.

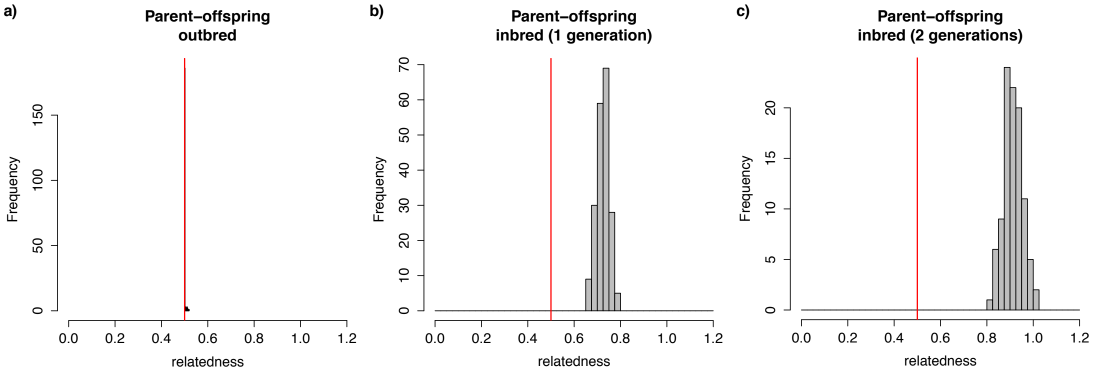

Figure 11. Relatedness coefficients between parent-offspring dyads from the pedigree in Figure 9. (a) Dyads A-C and A-D, outbred parent-offspring relationships, (b) dyads C-E and C-F, inbred parent-offspring relationships from a mating between two full siblings, and (c) dyad E-G, inbred parent-offspring relationship from two consecutives matings between two full siblings.

The inbreeding coefficient of individuals also increases if several inbred matings occurred in consecutive generations (Figure 9). The inbreeding coefficient of an individual whose parents are unrelated is 0 (Figure 12a). It increases to 0.25 if their parents are full siblings (Figure 12b). If a second consecutive inbred mating between full siblings occurs, this coefficient continues increasing (Figure 12c). Inbreeding can easily disappear if an inbred individual mates with an unrelated individual: the inbreeding coefficient of the offspring of such mating will be 0 again (Figure 12e).

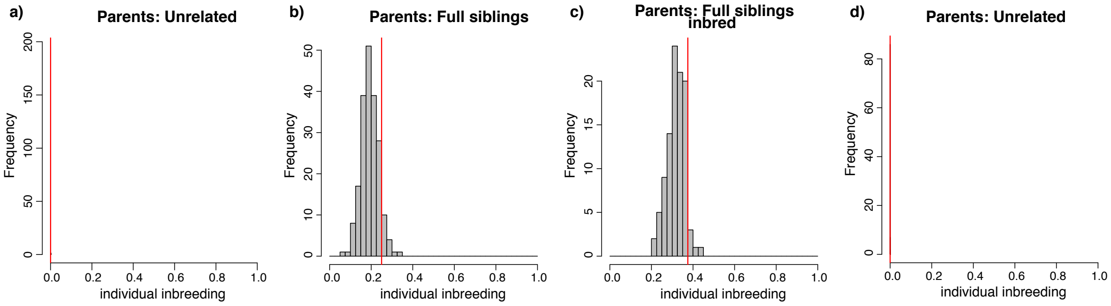

Figure 12. Individual inbreeding coefficients for individuals from pedigree in Figure 9. (a) Individuals C and D with unrelated parents, (b) individuals E and F with outbred full-sibling parents, (c) individual G with inbred full-sibling parents, and (d) individual I with unrelated parents.

Examining hundreds of loci is necessary for this kind of analysis, although the exact amount of SNPs needed could vary for each study. For example, Santure et al. (2010) found that 771 SNPs provided inconclusive results when estimating relatedness in zebra finch pedigrees, whereas Lopes et al. (2013) showed that 2,000 of these markers can give optimal estimates in pig pedigrees.

To show the effect of the number of markers used to estimate relatedness and inbreeding coefficients, a subset of 50 SNPs from the previous simulation was used with the same pedigree from Figure 9. With these new simulations, a decrease in the precision of relatedness estimations between outbred full-sibling and inbred full-sibling pairs can be observed in Figure 13a and Figure 13b, respectively, as compared with those obtained with 912 SNPs (Figure 10), indicating that low numbers of genetic markers may generate highly unreliable relatedness estimates.

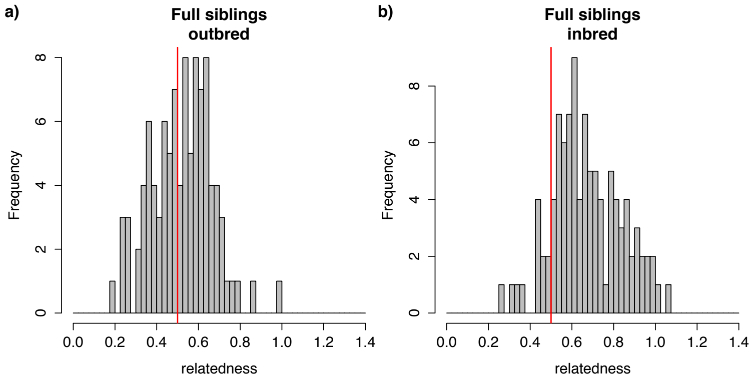

Figure 13. Relatedness coefficients between full siblings from pedigree in Figure 9 estimated from a data set of 50 SNPs. (a) Individuals C and D, resulting from a mating between unrelated individuals. (b) Individuals E and F, resulting from the mating between two full siblings.

3.4. Estimation of kinship categories

Inferring kinship relationships is an important step forward in the investigation of mating systems, inbreeding avoidance, kin recognition, kin selection and so on (Städele & Vigilant, 2016). The relatedness coefficient, in the absence of inbreeding, can help to identify close kinship categories (Table 2) provided that large amounts of markers are available. However, wild populations, and especially endangered populations, usually have some degree of inbreeding, rising relatedness estimates above the theoretical values, as shown in the previous section. This makes it difficult to directly use relatedness values to classify dyads of individuals into specific kinship categories in wild populations.

Microsatellites have been widely used for inference of kin relationships as they have a high variability, but only the closest kinship categories could be inferred with them. More recently, SNPs have been incorporated in pedigree reconstruction in natural populations. SNPs, although being individually much less informative than microsatellites, exist in large numbers and scoring is potentially less error-prone than with microsatellites (Pemberton, 2008). In addition, SNPs, together with maximum-likelihood methods based on the nine IBD-states model, allow determining more distant kinship categories (Städele & Vigilant, 2016).

In addition to genetic markers, information about the age of individuals or social status may be helpful for supporting or identifying kinship assignments (Städele & Vigilant, 2016).

3.4.1. Individual identification

Individual identification can be considered as the closest type of kinship category determination. Identification of duplicated samples from the same individual using genetic marker data is important in many research fields. When marker information is ample and with no genotyping errors, individual identification is straightforward (Wang, 2016). However, when genotyping errors are present, methods that are able to deal with them should be used.

Several programs for this purpose have been developed. GIMLET (Valière, 2002), for example, is a program that identifies identical genotypes, estimates genotyping errors, and estimates the kinship and several population parameters from a pool of samples.

Relatedness analysis can also be used to identify duplicated samples from the same individual. In this approach, duplicated samples are expected to have a relatedness coefficient of 1 if they are not inbred (Wang, 2007). However, it is important to be aware that monozygotic twins share all loci, making it difficult to distinguish them from two biological replicas of the same individual with any method.

3.4.2. Parentage assessment

Inference of parent-offspring relationships can be achieved with higher confidence than other relationships because the parent and the offspring must share at least one allele at every locus if there are no genotyping errors.

Parentage analysis techniques can be classified into several categories according to the methodology used: exclusion methods, based on Mendelian rules of inheritance (Chakraborty et al., 1974); categorical allocation, which use likelihood-based approaches (Meagher & Thompson, 1986); fractional allocation (Devlin et al., 1988); full probability parentage analysis (Hadfield et al., 2006); and parental reconstruction, that uses genotypes of offspring to reconstruct parental genotypes (Jones, 2001). Genotype errors, null alleles, and *de novo* mutations can cause apparent incompatibilities between true parents and their offspring, so care must be taken to accommodate such errors in parentage analysis (Jones & Ardren, 2003; Jones et al., 2010).

Two of the most popular programs for parentage assessment are CERVUS (Kalinowski et al., 2007) and COLONY (Jones & Wang, 2010). CERVUS is a categorical allocation program that calculates a likelihood-odds ratio score for each possible parent-offspring dyad and then assigns parentage across a group of offspring. COLONY uses maximum likelihood to assign both sibship and parental relationships by clustering offspring into paternal families and assigning candidate parents to the clusters.

Other commonly used parentage assessment programs are FAMOZ, that uses a categorical allocation approach similar to the one used by CERVUS (Gerber et al., 2003), and KINGROUP, a JAVA based program that uses maximum pairwise likelihood to group relatives (Konovalov et al., 2004).

3.4.3. Sibship reconstruction

Full siblings share, on average, half of their alleles, but at any locus they may share 0, 1, or 2 alleles, therefore making the detection of this relationship more complicated than parentage assignment.

Sibship reconstruction methods can be classified into two main groups: methods that require explicit pedigree reconstruction and pairwise methods that do not rely upon complete pedigree reconstruction. The second group can be subdivided into techniques based on pairwise relatedness estimation and likelihood-based techniques that allow the classification of pairs of individuals into different classes of relationships (Butler et al., 2004).

Some programs discussed above for parentage analysis are also able to infer sibship relationships, such as FAMOZ, KINGROUP, and COLONY. There are other programs available for sibship reconstruction, such as FRANZ (Riester et al., 2009). FRANZ is a program for pedigree inference that is able to detect parent-offspring and full and half siblings using a maximum-likelihood approach, that is robust in the case of errors, and that allows incorporation of prior information such as age or sex of individuals.

3.4.4. Distant relationships

Discrimination between kinship relationships of distantly related individuals is more difficult to achieve because several kinship categories have the same identical-by-descent proportions and, therefore, the same relatedness coefficient (Weir et al., 2006). For example, second-degree relationships, with a relatedness coefficient of 0.25, include half-sibling, grandparent-grandchild, and avuncular relationships. Furthermore, relatedness is not always exact due to inbreeding or a low number of genetic markers. Blouin et al. (1996) suggested using simulated distributions of the relatedness coefficient for certain kinship categories to define cut-off values that would allow classifying dyads as belonging to these kinship categories. The misclassification rate of this method is determined by the cut-off chosen. The resolution of such analyses can be improved by combining cut-offs for likelihood ratios with cut-offs for the relatedness coefficients (Langergraber et al., 2007).

These approaches could theoretically be used to classify dyads beyond the second degree of kinship, but as the overlap of genetic relatedness values increases with a decreasing degree of kinship, especially when not enough markers are used (see simulation above), no satisfactory trade-off between misclassifications and correct classifications can be reached for distant relationships (Städele & Vigilant, 2016).

3.4.5. Genome-wide analysis of kinship categories

Efforts to infer distant kinship categories with genome-wide data have been made in the last few years. One example is the method implemented in the software PRIMUS, which uses genome-wide estimates of pairwise identity-by-descent (IBD) to assign up to third-degree relatives that requires no prior knowledge on the pedigree structure and allows for missing individuals in the pedigree (Staples et al., 2014). However, if the sample to analyse consists mostly of distant relatives, relationship assignment becomes uncertain due to high variance in IBD sharing, leading to incorrect pedigree reconstruction (Ko & Nielsen, 2017).

High levels of inbreeding may affect relatedness and therefore pedigree inference. Recently, methods that use a likelihood-ratio test to discriminate various degrees of relatives, even for highly consanguineous families, have been developed to deal with the issue of inbreeding. Two examples are the programs VCF2LR (Heinrich et al., 2017) and SEQUOIA (Huisman, 2017), both being able to discriminate up to second-degree relatives.

More recently, Ko & Nielsen (2017) have developed CLAPPER, a method that uses pairwise likelihoods between pairs of individuals to infer pedigrees. This method can use a great amount of genome-wide markers, supports multi-generational pedigrees up to fifth-degree relationships and polygamous reproduction, and allows missing individuals in the sample. However, this method assumes that all individuals are outbred, making it difficult to be used with species with high levels of inbreeding such as many endangered species.

When data from thousands of SNPs or even whole-genome information can be obtained, more distant relationships can be inferred. For example, Manichaikul et al. (2010) identified in humans relatives up to third degree with low rates of misclassifications using 500k SNPs with the program KING, and Kling et al. (2012) detected up to fifth-degree relationships with high accuracy using thousands of unlinked SNPs with the programs FEST (Skare et al., 2009) and MERLIN (Abecasis et al., 2002).

A different kind of methods use the identification of chromosome fragments that are identical by descent, as inferred from whole-genome sequence data, to resolve distant relationships (Figure 14). This method, implemented in the program ERSA (Huff et al., 2011), accurately estimates the degree of relationship for up to eight-degree relatives and is able to detect twelfth-degree relatives. Li et al. (2014) have also developed a method that calculates the fraction of genomic blocks sharing identity and uses them to infer first to fifth-degree relatives.

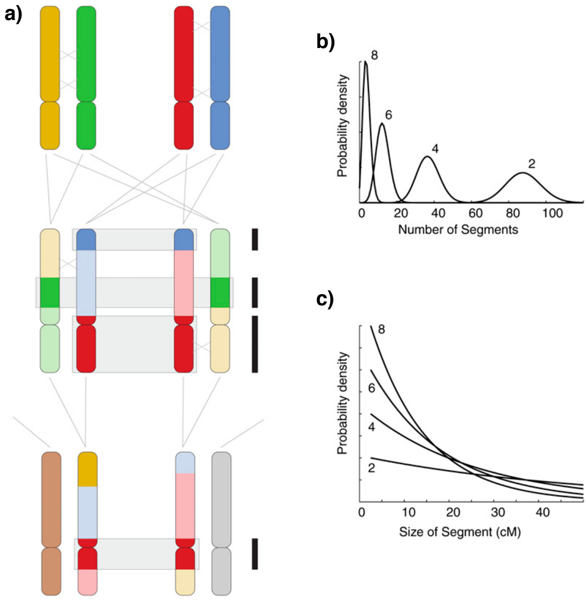

Figure 14. Expected distributions of IBD chromosomal segments between pairs of individuals. (a) Two chromosomes are shown for two parents, each coloured differently. Meiosis and recombination occur, and two sibling offspring inherit recombinant chromosomes. For some segments of the chromosome, the siblings share a stretch inherited from one of the four parental chromosomes. When the siblings mate with unrelated individuals, each offspring inherits an unrelated chromosome and one that is a recombinant patchwork of the grandparental chromosomes. (b) The number of IBD segments shared by a pair of individuals is approximately Poisson-distributed, with a mean that depends on the number of meiosis on the path relating the individuals. (c) The lengths of the IBD segments are approximately exponentially distributed, with mean length depending on the relationship between individuals (from Huff et al. (2011)).

3.5. The relevance of kinship analysis in conservation

3.5.1. Reproductive biology

Knowledge of kinship relationships between members of wild animal populations allows the inference of fundamental information about the species, especially regarding reproductive biology. Several studies have used either relatedness estimates or kinship category assignments for unravelling these crucial aspects of wild populations. For example, Bonin et al. (2012) studied mate infidelity and heteropaternity in Antarctic fur seals with the detection of half siblings in a single litter.

Relatedness estimates have also been used to infer kin structure and social organization in several wild species such as chimpanzees (Langergraber et al., 2007) and the spotted eagle ray (Newby et al., 2014). Other genetic studies have been used to evaluate the presence or absence of inbreeding avoidance in isolated populations, such as in the bighorn sheep (Rioux-Paquette et al., 2010).

Kin recognition plays an important role in the reproduction biology of many species in order to avoid mating with close relatives. Stenglein et al. (2011) described the pack social structure in grey wolves (*Canis lupus*) with NGS data from non-invasive samples, and Charpentier et al., (2005) used parentage assignments and relatedness coefficients estimates to study the male reproductive success in a population of mandrills. Charpentier et al., (2005) found that the probability of paternity by a dominant male decreased if it was related to the family group, demonstrating that female mandrills may exercise an active choice of partner to avoid inbreeding.

3.5.2. Inbreeding and inbreeding depression

Inbreeding depression is the reduction of fitness caused by mating between relatives, which increases the frequency of deleterious alleles in homozygosis in offspring. Inbreeding depression can be detected by the lower fertility, survival, and growth rates of individuals with high inbreeding coefficients (Charlesworth & Willis, 2009; Ouborg et al., 2010). It can be reduced by minimizing inbreeding in managed populations or with genetic rescue, that is the introduction of variation from outside the affected population. In natural populations, purging reduces in frequency deleterious mutations that result in inbreeding depression when they are in homozygosis. For example, brief bottlenecks may reduce the inbreeding depression of small populations by exposing deleterious alleles to selection (Hedrick & Garcia-Dorado, 2016; Leberg & Firmin, 2008).

Mean genome-wide heterozygosity has been used to evaluate individual inbreeding and the loss of genetic variation. Nevertheless, low levels of heterozygosity can be found in both natural populations with recent reduction of effective population size and in populations that have been small for a long time and that are not necessarily inbred (Kardos et al., 2016). Only the inbreeding coefficient can be used to infer inbreeding in the absence of direct pedigree information.

Care must also be taken to infer inbreeding depression via the inbreeding coefficient alone, as populations that have been small for a long time may had suffered a purge and, therefore, they may not be affected by inbreeding depression (Hedrick & Garcia-Dorado, 2016). Inbreeding depression can only be measured with fitness estimates.

Genome scans of large numbers of markers together with estimates of fitness can detect the signature of inbreeding depression. One example is the study of inbreeding depression in a wild population of red deer, where the authors used over 30,000 SNPs to estimate inbreeding and to test inbreeding depression in several fitness components and correlated traits (Huisman et al., 2016).

Inbreeding can also be detected with whole-genome data by identifying IBD chromosome segments. These can be used to estimate inbreeding by detecting runs of homozygosity (ROH) (Kardos et al., 2016). Distribution of ROH lengths can also contribute to evaluate the impact of purging in populations and to identify genes influencing inbreeding depression (Abascal et al., 2016; Hedrick & Garcia-Dorado, 2016; Kardos et al., 2016).

4. Dispersal and connectivity

The dispersal process, by which individuals move from their birthplace to a new settlement locality, has important consequences for the dynamics of genes, individuals, populations, and species. Dispersal knowledge is essential for conservation management, especially in the case of small and fragmented populations, which may become extinct due to catastrophic events, environmental stochasticity, or inbreeding depression if they remain isolated for a long time (Driscoll et al., 2014; Soulé, 1987).

Animal dispersal may have positive effect on the viability of fragmented populations, for example, by reducing the effects of inbreeding (Banks & Lindenmayer, 2014), but it may also have negative effects, for example, due to increased mortality of dispersing animals (Bowler & Benton, 2009). It is therefore important to know how dispersal affects the viability of populations. However, it is extremely difficult to determine whether some individuals can move from one population to another and breed in the new population, and more so for elusive species (Mills, 2013).

The migration rate is the proportion of individuals migrating from a population to another (Broquet & Petit, 2009). Different multilocus-based genetic techniques have been used extensively to estimate long-term migration rates between populations. They include the F-statistics (Slatkin, 1985), the isolation-with-migration models (Pinho & Hey, 2010), and the D-statistics (Durand et al., 2011) methods, but these techniques are best suited to the study of ancient dispersal. However, for endangered species, it is crucial to have information concerning current movements or those of the last few generations.

Traditionally, dispersal parameters were measured through direct observation of movements, capture-mark-recapture protocols, and radio tracking. Molecular markers were later introduced to the study of dispersal with two different approaches: indirect approaches, such as comparison of allele frequencies between populations, and direct approaches, such as the assignment of individuals to their population of origin (Broquet & Petit, 2009).

4.1. Methods to estimate recent dispersal

Classification of individuals in a sample into populations can be achieved with assignment methods. These methods can either classify individuals into predefined categories or cluster individuals into non-predefined categories that are constructed from the data (Manel et al., 2005).

Among methods that use predefined populations to classify individuals, BAYESASS is an example of program used for direct detection of migrants and to estimate migration rates. This program uses a Bayesian method and individual multilocus genotypes to estimate rates of recent migration among populations (Wilson & Rannala, 2003).

The most used program for clustering individuals into non-predefined populations is STRUCTURE (Pritchard et al., 2000), which infers the number of populations by comparing the posterior probability for different numbers of putative populations. STRUCTURE is also able to estimate the admixture of each individual by calculating the proportion of genomic loci belonging to each population cluster. Migrants can be then inferred from the genomic composition of individuals and their geographic location.

Finally, a number of other methods based on kinship analysis can be used to infer population structure and, indirectly, migration rates between the identified populations (Økland et al., 2010; Palsbøll et al., 2010; Watts et al., 2007). Nevertheless, most of these methods require the presence of populations with different allele frequencies to detect the individuals' source population. In fact, populations that only became isolated recently may have a similar genetic background and so these methods would not be able to detect movements between them. Furthermore, these population-based approaches require the previous assignation of individuals into populations, which is not always possible in populations with diffuse borders of permeable contact zones.

To overcome these difficulties, an additional class of genetic techniques based on paternity analysis has been proposed to analyse recent dispersal (Wang, 2014). These methods try to determine the individual's source population by identifying its parents. However, these methods require the analysis of a large number of specimens from different populations to identify a sufficient number of parental relationships, which may not be feasible for rare species. Similar kinship-based approaches try to infer the presence of dispersal through detection of kinship relationships between individuals. In these approaches, connectivity can be deduced when individuals with some type of kinship relationship are found in different localities or subpopulations. Several studies have explored this approach using parent-offspring and sibship relationships (De Woody, 2005; Melero et al., 2017; Norman & Spong, 2015; Oliver et al., 2016), but these are only a minor part of the relationships that can be found in a population. The detection of more distant relatives would be of great benefit in the study of connectivity patterns in the last few generations.

4.2. Sex-biased dispersal

Sex-biased dispersal occurs when one sex is more likely to disperse between populations than the other. In most mammals, males disperse from their birthplace and never return, while females are usually philopatric, tending to remain at their natal site for breeding (Freeland, 2005).

Sex-biased dispersal may be explained by three different hypotheses: resource-competition hypothesis, where the sex remaining at its natal site will take more profit from home-ground familiarity (Greenwood, 1980); local mate competition hypothesis, for which individuals disperse so that they will not have to compete with their relatives for mating (Perrin & Mazalov, 1984); and inbreeding avoidance hypothesis, for which the sex that incurs the greatest cost from inbreeding is the one that disperses (Pusey, 1987).

Quantification of sex-biased dispersal can be achieved by comparing population differentiation estimates based on autosomal nuclear markers versus mitochondrial markers, by comparing genetic differentiation calculated for each sex based on data from the same biparentally inherited loci, or using relatedness estimates (Freeland, 2005). In the last approach, relatedness coefficients are estimated between adult female dyads and adult male dyads within populations and compared: the sex that has lower dispersal should show higher relatedness levels for their dyads. With this approach, male-biased dispersal has been confirmed in several species such as in bobcats (Croteau et al., 2010), Bornean orang-utans (Arora, et al., 2012), and bats (Hua et al., 2013).

4.3. Networks and connectivity

Extending the paternity analysis to more distant relationships and using relatedness values estimated between all pairs of individuals in a populations can be used to infer dispersal and detect recent migration events. The idea is that when two individuals from different localities present significant relatedness values, then some of these individuals or their ancestors must have dispersed between the localities in the last few generations. Although the actual dispersal route cannot be determined and may involve intermediate, unknown localities, recent gene flow between localities can be inferred in this way. When applied to many individual pairs, the population's dispersal pattern over the last few generations emerges.

Network analysis is emerging as a useful tool for examining and quantifying the patterns of social relationships between individuals in animal populations (Hamede et al., 2009; Shizuka & Farine, 2016; Wey et al., 2008). Taking individuals as nodes and relationships as edges, patterns of connectivity between individuals can be estimated for specific discrete characteristics. If nodes are connected to a large number of other nodes with the same characteristic, the network is said to have assortative mixing; if the network connects nodes of different characteristics, it shows disassortative mixing (Newman, 2003).

The assortativity coefficient quantifies the level of assortative mixing in a network for a given characteristic. This coefficient can range from 1 (when there is perfect assortative mixing) to -1 (when the network is perfectly disassortative). When the nodes are randomly connected, the assortativity coefficient is 0 (Figure 15).

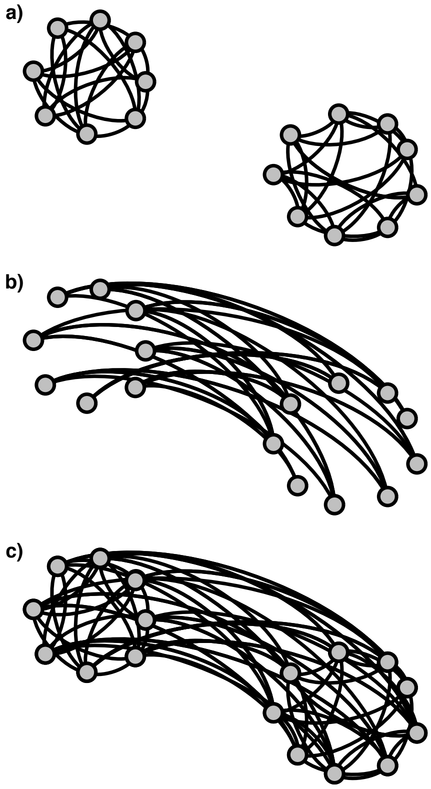

Figure 15. Networks showing different degrees of mixing. (a) Assortative mixing (r = 1), (b) dissassortative mixing (r = -1), and (c) an intermediate level of mixing (r = 0.28).

It is important to study the level of connectivity between individuals, as assortative mixing can have a profound effect on a population. For example, assortative mixing of a network by a discrete characteristic will lead to a fragmentation of the population into separate communities (Newman, 2003).

4.3.1. Calculation of the assortativity coefficient

Considering assortative mixing according to a discrete characteristic, such mixing can be characterized by a quantity e_{ij}, which is the fraction of edges in a network connecting a vertex of type i to one of type j. This quantity is symmetric on an undirected network and $e_{ij} = e_{ji}$. This quantity satisfies the sum rules

$$\sum_{ij} e_{ij} = 1, \quad \sum_{j} e_{ij} = a_i, \quad \sum_{i} e_{ij} = b_j,$$

where a_i and b_i are the fraction of each type of end of an edge that is attached to vertices of type i. On undirected graphs, $a_i = b_i$.

According to Newman (2003), the level of assortative mixing in a network can be quantified by the assortativity coefficient

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i}$$

The expected standard deviation on the value of r can be calculated using the jackknife method

$$\sigma_r^2 = \sum_{i=1}^{M} (r_i - r)^2$$

where M is the number of edges in the network and r_i is the value of r for the network in which the ith edge is removed.

4.3.2. Quantification of connectivity between populations

The assortativity coefficient can be used to estimate the level of connectivity between two adjacent sectors. Given a network of kinship relationships and considering the population in which the individual was found as the discrete characteristic, estimation of the assortativity coefficient for adjacent populations may provide an estimation of the connectivity between populations. Thus, if there is sufficient dispersal between populations, the kinship network should connect a large number of individuals from different populations, and the assortativity coefficient will be close to 0. If there are barriers to dispersal, the relationships between individuals of different populations will be low compared to the relationships found within a single population, and the assortativity coefficient will approach 1.

5. The Pyrenean desman as a model species to unravel the biology of an elusive and endangered species with NGS data

5.1. The Pyrenean desman (*Galemys pyrenaicus*)

The Pyrenean desman (*Galemys pyrenaicus*) is a small semi-aquatic mammal belonging to the order Eulipotyphla and the family Talpidae. It is placed within the sub-family Desmaninae together with the Russian desman (*Desmana moschata*), which are the only two extant representatives of this group. It is endemic from the northern part of the Iberian Peninsula (Palmeirim & Hoffmann, 1983). Due to recent contraction of its occupation area, it has a patchy and discontinuous distribution (Figure 16).

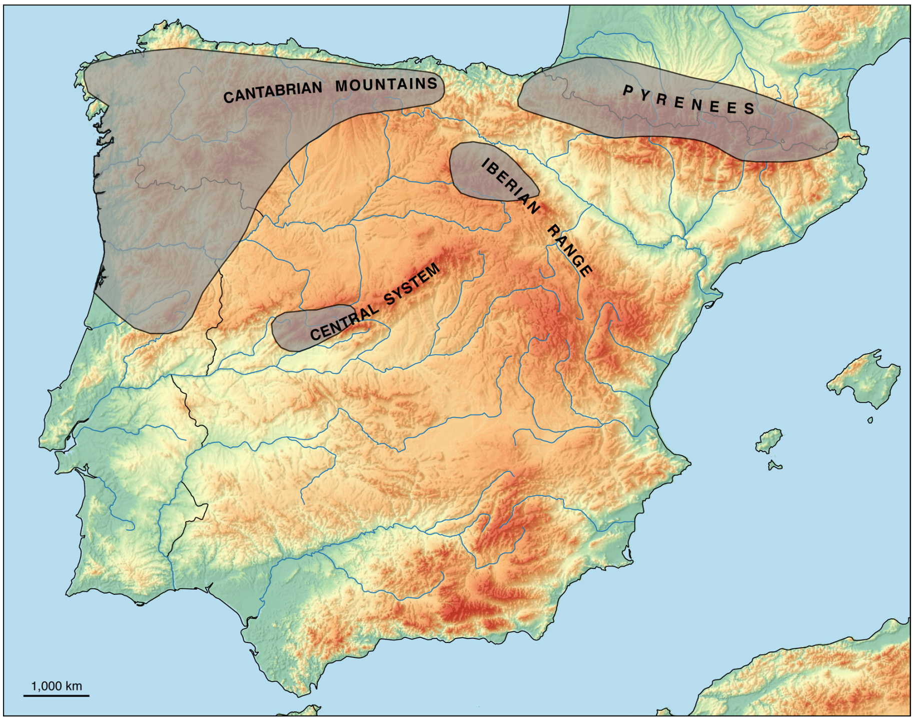

Figure 16. Distribution of the Pyrenean desman with the most recent data available. Mountain ranges mentioned in the text are indicated.

The Pyrenean desman has strong adaptations to the aquatic medium that include broad, partially webbed hind limbs with long claws on all digits (Palmeirim & Hoffmann, 1983). It has an elongated, highly sensory snout and a long tapering tail (Figure 17). Its diet consists in larvae of benthic macroinvertebrates that captures underwater such as Ephemoeroptera, Plecoptera, and Trichoptera (Biffi et al., 2017; Castien & Gosálbez, 1994). It inhabits clean, oxygenated rivers and streams, an habitat generally found in mountain areas (Figure 18).

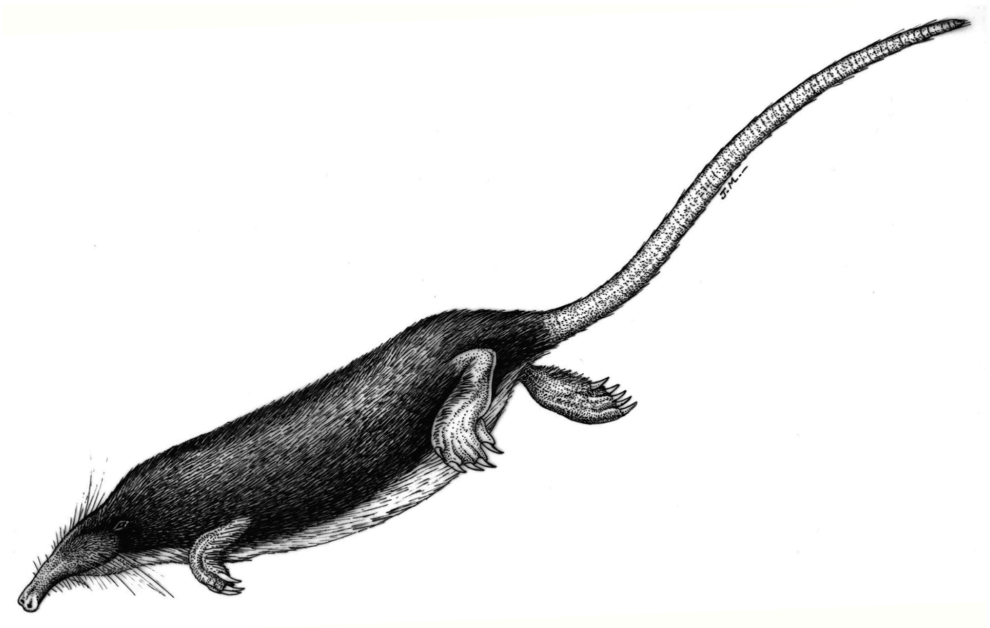

Figure 17. *Galemys pyrenaicus* (from Nores (2007)).

The species has suffered strong declines in the last few years for reasons not fully understood. The primary causes put forward to explain this regression include water pollution, the desiccation of rivers, and construction of dams (Fernandes et al., 2008). One of the most important consequences of the alteration of rivers where it lives is the reduction and fragmentation of populations. Also, the construction of dams in rivers may generate isolated populations upstream of a dam, although so far nothing is known about this thread.

The Pyrenean desman is legally protected in its four native countries (Spain, Portugal, France, and Andorra) and is classified as "vulnerable" in the IUCN Red List (Fernandes et al., 2008). It is also legally protected at the European level through the annexes II and IV of the European Habitats Directive 92/43/CEE and through the annex II of the Bern Convention. Furthermore, significant declines in populations

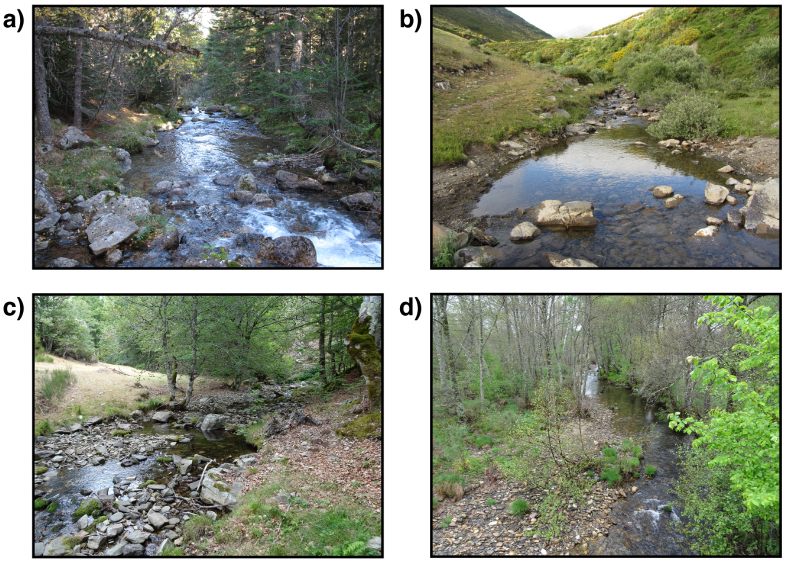

Figure 18. Typical habitats of the Pyrenean desman. (a) Escrita river (Aigüestortes; Lleida); (b) Puerma river (León); (c) Umbría river (La Rioja); and (d) Castro river (Tera basin; Zamora) (photo credit: Jose Castresana).

located in the Central System, in the southern part of the distribution (Gisbert & Garcia-Perea, 2014), led the Spanish Government to catalogue such populations as "in danger of extinction", which is the highest protection category.

5.2. Genomic structure

Igea et al. (2013) studied the phylogeography of the Pyrenean desman using mitochondrial genes and found a marked phylogeographic structure in which two large groups, A and B, were subdivided into two further groups to give a total of four mitochondrial lineages with parapatric distribution (A1, A2, B1, and B2). They detected two contact zones between the main mitochondrial groups (A and B), one located in the Cantabrian Mountains and the other in the Iberian Range (Figure 19).

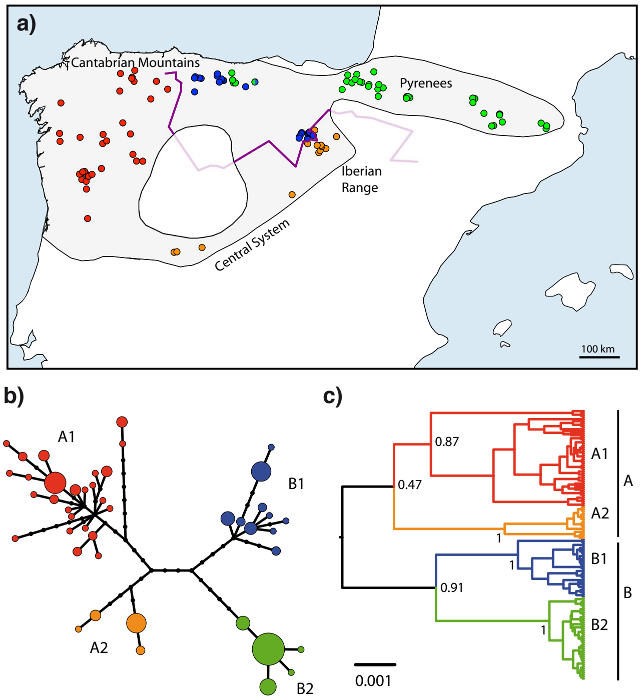

Figure 19. Phylogeography of the Pyrenean desman. (a) Map of the northern part of the Iberian Peninsula showing the distribution of the four mitochondrial lineages found with (b) the haplotype genealogy of the concatenated mitochondrial sequences of the individuals and (c) the Bayesian tree of the same sequences (from Igea et al. (2013)).

The contact zone of the Iberian Range was particularly striking because there was no apparent mixing between the two mitochondrial lineages present in the area (A2 and B1). A separation line was found in the valley of the Najerilla river, which seemed to restrict the dispersal of female desmans (Igea et al., 2013).

In a more recent study, Querejeta et al. (2016) analysed genomic data and found five clusters that were largely coincident with the mitochondrial lineages, although some differences were detected, particularly in the Iberian Range. Whereas Igea et al. (2013) detected two mitochondrial lineages in the Iberian Range, A2 in the southeast and B1 in the northwest, Querejeta et al. (2016) observed just one main genomic cluster, although subdivided into two subgroups with different genomic compositions.

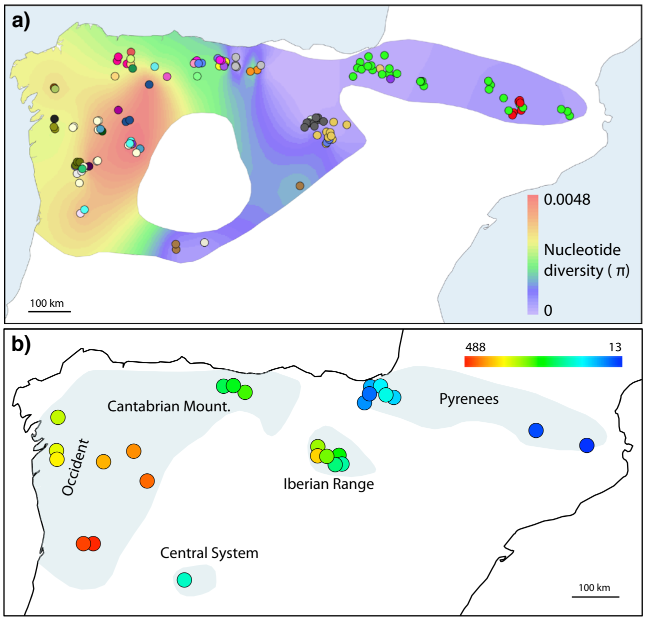

Figure 20. (a) Map of the genetic diversity of *Galemys pyrenaicus* from Igea et al., (2013) and (b) map plotting colour-coded heterozygosity rates in different specimens from Querejeta et al. (2016).

Regarding its genetic diversity, Igea et al. (2013) found that the overall mitochondrial genetic diversity of the Pyrenean desman was relatively low and with high differences between the lineages. The nuclear genetic diversity found was also very low as estimated from eight nuclear markers. Those individuals from the A1 lineage presented the highest levels of both mitochondrial and nuclear genetic diversity, while individuals from the B2 lineage were the ones with the lowest values (Figure 20a). The study suggested that the genetic structure of the Pyrenean desman had been more influenced by the history of the Pleistocene glaciations than by its current habitat distribution. Querejeta et al. (2016) found a similar pattern when studying heterozygosity from genomic data: the individuals with the highest values were those from the occidental zone, whereas the individuals from the Pyrenees had much lower values (Figure 20b). The heterozygosity rate found for the Pyrenean desman was one of the lowest reported in mammals so far, with a mean of 246 heterozygous positions per million bases.

More recently, Querejeta et al., (2017) studied the correlation between phylogenetic distances and geographical distances in the north-western region of the Iberian Peninsula using an isolation-by-distance approach. They found that the correlations obtained were consistent with an effect of overland dispersal due to a possible postglacial colonization of new territories using terrestrial corridors and a more extensive fluvial network that may have been present during the Holocene. Given the absence of current spatial mixing between the mitochondrial lineages in the studied contact zone, the authors suggested a reduction of contemporary inter-river dispersal of the Pyrenean desman after the postglacial colonization.

Igea et al. (2013) inferred four glacial refugia using a species distribution model based on the known-presence localities of *G. pyrenaicus*. The main glacial refugium was found in the northwestern area of the distribution, which is coincident with the area of contemporary greater genetic diversity of the species. The postglacial expansions from these refugia gave rise to the current distribution of the species (Figure 21).

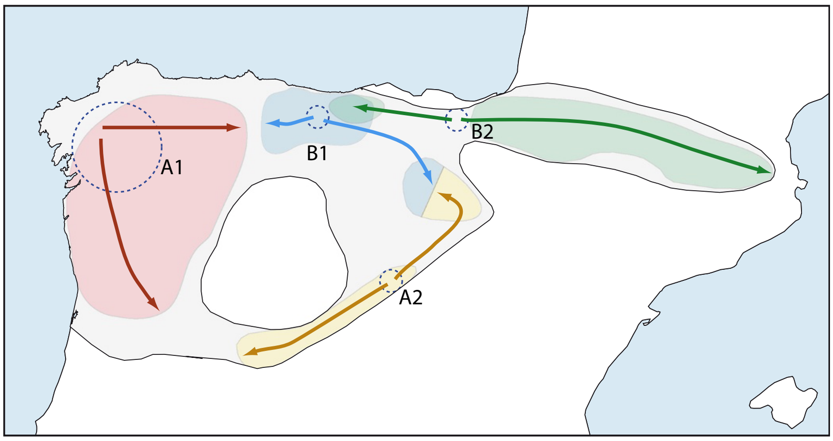

Figure 21. Schematic representation of the evolutionary history of the Pyrenean desman. Hypothetical positions of glacial refugia are illustrated with dotted circles within the current distribution of each mitochondrial lineage, represented by different colours (from Igea et al. (2013)).

5.3. Social behaviour and dispersal patterns

According to data obtained through radio tracking in a river of the Pyrenees (Stone, 1985), the Pyrenean desman was shown to be aggressive towards its conspecifics and largely solitary in nature. The species had primarily nocturnal activity (Stone & Gorman, 1985). Stone (1985) subdivided the Pyrenean desman population into sedentary members, consisting of single adult males and females, and transient members, consisting of juveniles and nomadic adult males and females. According to a subsequent study (Stone & Gorman, 1985), the fraction of sedentary members was shown to consist of pairs of adult desmans of opposite sex that occupied a similar stretch of river but using separate nest sites, with males occupying a larger length of home range than females (429 and 301 meters, respectively). Solitary adult specimens occupied larger home ranges than juvenile specimens (584 and 248 meters, respectively).

Recent studies also based on radio tracking of desmans in the Pyrenees arrived to highly different conclusions since they found a mostly non-territorial and non-aggressive behaviour of the Pyrenean desman (Melero et al., 2012, 2014). In general, the activity of desmans was mainly nocturnal with a short diurnal bout, whose duration increased in spring possibly due to the need to feed to meet energy requirements. They also found that individuals had more than one main resting site and that each of them may be shared by more than one individual, regardless of their sex and age. The estimated mean home range was 523 meters. Melero et al. (2014) determined that seasonality and daylight were crucial factors influencing the range use and movement pattern of desmans, as individuals used a higher percentage of their home ranges and travelled longer distances at night and in autumn.

Using microsatellite markers amplified from fresh tissue and faecal samples of desmans from the Pyrenees, Gillet et al. (2016) found that individuals were able to disperse longer distances than previously recorded. In particular, two individuals were found to have moved 16 and 18 km, respectively, after several months.

Population densities in rivers vary between 3 and 7 individuals/km (Nores et al., 1998), although these estimates could be higher if the species is less territorial than previously assumed (Melero et al., 2014).

5.4. Reproductive biology

Richard (1976) was the first to propose a method for age determination of the Pyrenean desman based on tooth wears and eruption, and taking into account a total of 18 measures of different tooth. Later, González-Esteban et al. (2002) developed a simpler method for age determination based only on the upper canine tooth wear. With this method, individuals can be classified into 5 classes: class 0, corresponding to the first year of life of individuals; class 1, second year of life; class 2, second to third year of life; class 3, third to fifth year of life; and class 4, third to sixth year of life.

The absence of sexual dimorphism in the Pyrenean desman makes it difficult to sex live animals. Sex determination is further hindered by the masculinisation of the external genitals in females, whose urinary papilla has an internal structure similar to the penis of young males (Peyre, 1955, 1962). Richard (1986) reported the possibility of distinguishing between adult males and females during the reproductive season, as the open, pigmented vaginal orifice is easily visible at this time of the year. Peyre (1957) made observations regarding the sexual dimorphism in the pelvic arch, which some researchers have used as a criterion to sex live desmans (González-Esteban et al., 2003; Palmeirim & Hoffmann, 1983). Nevertheless, this difference in the pelvic arch is not discernible in immature specimens.

Regarding the ontogeny and reproduction of the Pyrenean desmans, Peyre (1962) suggested that the species is polyestrous with three annual peaks in the number of pregnant females, taking place in February, March, and May. The gestation period is about 30 days and the number of embryos per female ranges from 1 to 5. It is thought that individuals reach maturity during their second year.

5.5. Conservation of the Pyrenean desman

As already mentioned, the Pyrenean desman is endangered and some of its populations have experienced strong declines in the last few years. Efforts being done for the conservation of the species include appropriate management of water courses, habitat restoration, and limitations to specific landscape interventions such as the construction of new dams and water extraction (Cabral et al., 2005; Nores, 2007).

In the last few years, national and international conservation plans have been implemented in the whole distribution of the species. It is of especial importance the development of several LIFE programs, which is the European Union's financial instrument supporting environmental, nature conservation, and climate action projects. In France, the program LIFE+ DESMAN (2014-2020, LIFE13 NAT/FR/000092) is aimed to improve the desman's long-term conservation status in 11 Natura 2000 sites. In Spain, four LIFE programs were dedicated to the Pyrenean desman and its habitat: the LIFE MetWetRivers (2012-2018, LIFE11 NAT/ES/00699), aimed at developing a management and monitoring program for 60 Natura 2000 sites in Castilla y León for Mediterranean wetlands and rivers; the LIFE IrekiBAI (2015-2020 LIFE14 NAT/ES/000186), aimed at the conservation status of river habitats and species in Navarra and Gipuzkoa; the LIFE DESMANIA (2012-2017 LIFE11 NAT/ES/00691), aimed at establishing the basis for a long-term strategy of recovery for the Iberian desman in Castilla y León and Extremadura; and the LIFE MARGAL-ULLA (2010-2016, LIFE09 NAT/ES/000514), for the recovery of populations of *Margaritifera margaritifera* and *Galemys pyrenaicus* in the Ulla river basin in Galicia.

Despite all these efforts, there is still a profound lack of knowledge on the biology of the species, with many studies performed decades ago. In addition, most of them were developed in specific areas, and therefore they cannot easily be extrapolated to other areas of the whole species range. Being a highly elusive animal, the species can only be rarely observed in the wild, making more difficult obtaining reliable data about the species. Many aspects of its reproductive biology are still unknown, being specially relevant the lack of knowledge about the breeding period in different areas, the age of first reproduction or the number of pups per litter. Also, dispersal has been based on radio tracking studies, but this technique can only detect short movements over a short period of time and therefore long-range dispersal movements go unobserved. In addition, little is known about how river barriers, especially dams, affect the dispersal of the desmans and their genetic health.

Therefore, it is highly necessary to promote research on the species, especially of those populations that are more threatened. Knowledge about the population structure of the species, their dispersal patterns, and the genetic health would be of great benefit for implementing more thorough conservation plans.

II. OBJECTIVES

Knowledge of the genealogical relationships among individuals in a population and the dispersal patterns of the species are essential to many studies of species of conservation concern, especially for small and fragmented populations. Obtaining genomic information from endangered species with next-generation sequencing techniques is especially challenging, and therefore new methods and protocols that are able to deal with low amounts of DNA from non-invasive samples should be developed.

The main objective of this thesis is to design new methodologies that use genomic data to infer contemporary dispersal patterns of species using relatedness networks, to classify pairs of individuals into specific kinship categories, to infer pedigrees from these assignments, and to quantify the effect of anthropogenic and geographic barriers on the dispersal of individuals, using as a model a semi-aquatic mammal, the Pyrenean desman (*Galemys pyrenaicus*).

More specifically, the objectives of this thesis are:

1. Study at the genomic level the contact zone between two lineages of the Pyrenean desman previously determined in the Iberian Range using SNPs derived from ddRAD genomic libraries. Inference of the population structure and admixture proportions of individuals. Estimation of relatedness coefficients based on individuals and individual inbreeding levels. Design and use of bioinformatic simulations of genomic data based on artificial pedigrees to assess the accuracy of the relatedness and inbreeding estimates. Inference of the contemporary dispersal patterns of the species in the area using relatedness networks that represent the relationships between individuals. Study of the variation of inbreeding levels across the studied area.

2. Determination of kinship relationships and pedigrees in Pyrenean desmans of two rivers of Zamora (NW of the Iberian Peninsula) using genomic data obtained mainly from hair samples. Evaluation of the use of this type of minimally invasive samples for genomic studies in comparison with fresh tissue samples. Estimation of kinship categories using different methods. Use of bioinformatic simulations of genomic data based on artificial pedigrees to assess the accuracy of the kinship category assignments. Reconstruction of pedigrees from the kinship categories and evaluation of the congruence of the pedigrees using different sources of data. Inference of preliminary data about the reproductive biology of the species. Estimation of the dispersal distance of the individuals per generation, as well as the study of philopatric patterns of the two sexes.

3. Quantification of the impact of anthropogenic and geographic barriers on the dispersal of the Pyrenean desman in two rivers of Zamora by estimating the assortativity coefficient between adjacent sectors of kinship networks. Use of this estimator to assess the level of connectivity between the two basins and between both sides of several dams present in the area. Comparison of the results with those obtained from clustering-based methods of population structure inference.

III. METHODS

1. Using relatedness networks to infer contemporary dispersal: application to the endangered mammal *Galemys pyrenaicus*

1.1. Samples of Pyrenean desman from the Iberian Range (La Rioja)

A total of 66 samples of Pyrenean desmans from different rivers of the Iberian Range were used for the study. Of these, 37 tissue samples were used to perform library construction and genomic analyses (Table S1, Figure 22). All tissue samples were obtained from specimens captured during a monitoring project performed in 2011 for the La Rioja Regional Government (Spain), with permit number A/2011/52. A small portion was taken from the tip of the tail before the captured specimens were released back in the wild. The work was carried out following national and international regulations, and all necessary steps were taken to prevent any damage to the specimens. The rest of samples, consisting of 26 faecal samples from faeces that desmans deposited in exposed rocks of the rivers (Igea et al., 2013) and three specimens found dead in the field, were used to complete the mitochondrial phylogeography in this area (Table S2). All samples were conserved in tubes containing absolute ethanol and stored at -20ºC (tissues) or 4ºC (faeces) in the laboratory.

1.2. Sample processing

1.2.1. DNA extraction

DNA from fresh tissue and faecal samples was extracted using the DNEasy Blood and Tissue Kit (QIAGEN), following manufacturer's instructions. DNA extractions from faecal samples were carried out in a separated UV-irradiated area with dedicated equipment.

1.2.2. PCR and phylogenetic analysis of mitochondrial sequences

The cytochrome *b* gene was amplified from all the samples using primers specific for *Galemys pyrenaicus* (ACTAATGACATGAAAAATCATCGTT and TTTTCGTTTTTGGTTTACAAGAC) as previously done in Igea et al. (2013). All PCR reactions were set up in a dedicated PCR clean-room that is physically separated from post-PCR working areas and regularly decontaminated by UV-irradiation. PCR reactions were performed in 25 μl of final reaction volume, containing 2 μl of DNA, 1 μM of each primer, 0.2 mM dNTPs, 0.75 units of Promega Go Taq DNA polymerase and 17.5 μM of bovine serum albumin under the following conditions: an initial denaturation of 3 min at 95ºC, followed by 35 cycles of denaturation (30 s at 95ºC), annealing (1 min at 54ºC), and extension (1 min at 72ºC). A final extension of 5 min at 72ºC was added. PCR products were revealed by electrophoresis in 1% agarose SYBR-Safe (Invitrogen) stained gel.

PCR products were purified using ExoSAP-It (Affymetrix) and sequenced in both directions using the original primers at Macrogen Inc. (Macrogen Europe, The Netherlands). An extra set of two internal primers (TACAAGATCAGTTCCGATGTAAG and GGATTATCATCCGACACTGATAA) was also used in the sequencing process of entire cytochrome *b* sequences. Sequences were trimmed, assembled and analysed using Geneious Pro 5.1.7 (Biomatters Ltd.).

The entire mitochondrial cytochrome *b* sequence was obtained for 61 samples. A 1,103 bp fragment was obtained for one sample and a 724 bp fragment for four samples using primers that amplify shorter fragments, as described in Igea et al. (2013). To assign the mitochondrial lineage to each specimen, a maximum-likelihood phylogenetic tree of the aligned cytochrome *b* sequences was reconstructed using RAXML version 8.0.17, with a GTR model of nucleotide substitution and a gamma distribution of evolutionary rates, as recommended in the program (Stamatakis, 2014).

1.2.3. Quantification of DNA concentration and sex determination by qPCR

To quantify the DNA concentration of the samples, a quantitative PCR was performed. A 76 bp fragment of the ultraconserved element among mammals Chr2_23668 (Faircloth et al., 2012) was amplified and quantified using the primers AAATGCAGCGATCAGCAGT and ACGGGTGCCACATGTTAAG (Querejeta et al., 2016). qPCR reactions were performed in a final volume of 20 μl, containing 10 μl of IQ SYBR Green Supermix (Bio-Rad), 300 nM of each primer, and 2 μl of genomic DNA (prediluted 100-fold). The reactions were set in triplicates and run on a MyiQ Single-Color Real-Time PCR Detection System (Bio-Rad) with the thermocycling protocol of 2 min at 95ºC for the initial denaturation followed by 40 cycles of denaturation (15 s at 95ºC) and annealing/extension (30 s at 60ºC). The standard curve for absolute quantification was prepared using a commercial *Bos taurus* DNA sample of known concentration (Sigma-Aldrich) consisting of a ten-fold dilution series ranging from 2 ng to 0.02 ng and a non-template control.

Visual determination of gender is difficult in desmans, especially for young animals. To avoid errors and reduce handling time and stress to the animals, a genetic sexing protocol was applied to all individuals. To sex the specimens, a 77 bp region of intron 42 of USP9Y of the Y chromosome was amplified and quantified by qPCR using the primers GACAGCTTCCAAAATAAAGAATT and GAACTGGCAGTAATTTTCAAAGTG (Querejeta et al., 2016). qPCR reactions were performed as described above. The standard curve for Y-chromosome quantification was prepared using a male *G. pyrenaicus* sample of known concentration consisting of a ten-fold dilution series ranging from 8.5 ng to 0.085 ng and a non-template control.

1.3. Construction and sequencing of genomic libraries

1.3.1. Library construction and Illumina sequencing

DNA libraries were constructed using the ddRAD protocol (Peterson et al., 2012) with some modifications from Querejeta et al. (2016). Each library was performed in series of 24 samples, repeating samples with a low sequence yield in subsequent experiments to get sufficient coverage for all samples. For each sample, 50 ng of genomic DNA (as estimated by qPCR) was double digested using EcoRI and MspI restriction enzymes, in a 30 μl final volume. After overnight incubation, the enzymatic reaction was heat inactivated at 80ºC for 30 min. A 70 μl ligation mix was then added, including T4 DNA ligase, P1 adapter (which binds to the EcoRI overhangs and has a 5-nucleotide barcode to identify each specimen), and P2 adapter (which binds to the MspI overhangs), and the solution was incubated for 5 hours at room temperature. Then all the ligation reactions were inactivated at 65ºC for 10 min

All the ligation reactions were mixed in a single tube and concentrated to 20 μl using the MinElute PCR Purification Kit (QIAgen). The entire pool was run in a precast EX 2% agarose gel using the E-Gel system (Invitrogen) and the fraction between 300 and 400 bp was cut and purified with the QIAquick Gel Extraction Kit (QIAgen) in 30 μl.

A PCR amplification of the size-selected sample was performed to add Illumina adapters. Phusion High-Fidelity DNA Polymerase (New Englands Biolabs) amplifications were carried out using 6 μl of the size-selected sample pool as template and employing 16-20 cycles (depending on the intensity of the initial PCR products). A total of five PCRs were performed to increase the concentration of the libraries and to minimize bias. Then the PCR products were combined and concentrated into a 30 μl volume using the MinElute PCR Purification Kit.

Finally, 400 ng of DNA library (as estimated by NanoDrop) were run in a precast E-Gel EX 2% agarose gel and the band corresponding to the library was extracted in 30 μl with the QIAquick Gel Extracti on Kit. The libraries were sequenced in single-read runs using NextSeq Sequencing System (Illumina) and the 150-cycles Mid Output kit in the Genomics Core Facility at the Pompeu Fabra University.

1.3.2. Sequence processing

The STACKS 1.35 package (Catchen et al., 2013) was used to process the sequences obtained. First, the PROCESS RADTAGS program was used to filter out reads with low-quality sequences and to separate reads belonging to different samples according to their barcodes. This program was used with the recovery option, which corrects isolated errors in the restriction cut site sequence or in the barcode, and with different values for the quality score limit (s: from 10 to 30) depending on the overall quality of the library.

Then reads from samples that were sequenced in different runs were combined. After this step, USTACKS was used to assemble loci in each sample, with a minimum number of three sequences per locus for each sample (minimum stack depth or m) and a maximum of two differences between reads (M). The mean sequence coverage for each specimen was then calculated as the number of assembled reads divided by the number of assembled loci. The loci of all the specimens were subsequently merged with CSTACKS, allowing for a maximum of two mismatches between reads (n), and a catalogue of loci and sequences was created using SSTACKS.

The POPULATIONS program was used to save output files with different filter combinations available in the program: minimum proportion of called individuals (r), minimum stack depth or coverage (m), and minimum minor allele frequency (MAF). SNPs were saved in PLINK flat format and sequences of loci in FASTA format for further analyses. For the SNP data set, only one SNP per locus was saved.

The quality of the library reads was verified by generating an initial data set with the parameters r = 1, m = 9, and MAF = 0 (data set 1 in Table S3). Additional SNP data sets were generated with different filters to analyse how they perform when estimating pairwise relatedness.

1.4. Relatedness estimation and simulations

1.4.1. Selection of the best estimator

Estimation of pairwise relatedness among individuals was performed with the program *RELATED* (Pew et al., 2015), which is an R implementation of COANCESTRY (Wang, 2011). The accuracy of the different relatedness estimators was assessed with simulations implemented in RELATED, using the first SNP data set (data set 1, Table S3). This step was necessary because it has been shown that the performance of different estimators is highly dependent on the specific characteristics of the data set used (Blouin, 2003; Russello & Amato, 2004; Van De Casteele et al., 2001). Individual genotypes were simulated using the allele frequencies of the population to calculate the relationship between 250 dyads of each of the following relationships: parent-offspring, full siblings, half siblings, and unrelated individuals. The relatedness coefficient was subsequently calculated for each dyad using five moment estimators: *lynchli* (Li et al., 1993), *lynchrd* (Lynch & Ritland, 1999), *quellergt* (Queller & Goodnight, 1989), *ritland* (Ritland, 1996), and *wang* (Wang, 2002); and two maximum-likelihood estimators: *dyadml* (Milligan, 2003) and *trioml* (Wang, 2007). The performance of the seven different estimators was then evaluated by calculating means and standard deviations from the estimates and correlating them against the expected values.

1.4.2. Selection of the best SNP data set

Filters to generate SNPs with the POPULATIONS program (r, m, and MAF) were optimized according to their best performance in relatedness-based individual identification using RELATED and the *dyadml* estimator, which was the best estimator for the data set (see Results section 1.1.2). For this purpose, two different DNA aliquots of four samples were processed, sequenced, and analysed independently. Then, it was tested whether duplicated samples from the same individual reproduced the expected relatedness value of one when using a model with no inbreeding. To do so, SNP data sets with different combinations of parameters in POPULATIONS were generated and relatedness values among individuals were calculated.

1.4.3. Construction of relatedness networks and estimation of individual inbreeding coefficients

Pairwise relatedness among the 37 Pyrenean desmans was calculated using the optimal estimator of RELATED (*dyadml*) and the optimal data set found (912 SNPs; data set 2, Table S3), using all specimens for allele frequency estimation. The full nine IBD-states model, that takes inbreeding into account, was selected in RELATED and therefore the inbreeding coefficient for each specimen was also estimated. The use of this model was justified as the preliminary relatedness analyses revealed that inbreeding was high in the studied population and the marker information was enough for this more complex model (Wang, 2007). Confidence intervals (95%) for the estimation of relatedness were calculated using bootstrapping over loci (100 replicates). To avoid false positives that may alter dispersal patterns, the smallest relatedness values were removed. Thus, only relatedness estimates where the lower 95% confidence limit of the bootstrap replicas was higher than 0 were considered.

To visualize the relationships between individuals in space, a network of relationships was plotted with the program GEPHI (Bastian et al., 2009) using individuals as nodes and relatedness values as edge thicknesses. Nodes were represented according to the geographic location of the individuals with the plug-in GEOLAYOUT and the network was superimposed on a map.

1.4.4. Simulations along pedigrees of pairwise relatedness and inbreeding coefficients

To test the reliability of the relatedness coefficients obtained, simulations were performed along artificial pedigrees in which individuals with different origins were computationally crossed. In 12 of the pedigrees, all founders were from the same river (Figure S1), and in 12 additional pedigrees a “migrant”, that is an individual from a different river, was included (Figure S2). Some pedigrees had four founders and others had five to be able to recreate different kinship categories of interest. Simulations of offspring for the different crosses were performed with the custom Perl script GetCrosses.pl, which randomly selects one allele for each locus from each parent to generate the alleles of a new individual. The new individuals generated in this manner were then added to the output file. Using the GetCrosses.pl and RELATED programs, 100 simulations were conducted for each pedigree and the relatedness values for all relationships within the artificial pedigree were estimated and compared with the expected values.

In the simulations along pedigrees with migrants, detection of second and third generation migrants was also determined by counting all relationships between the generated offspring (F101, F102, F201, and F202) and the individuals of the river from which the migrant came from, after excluding this migrant. Relatedness values above 0.0625 (corresponding to the lowest value detected in the real data set after bootstrap; see Results section 1.1.4) were counted for all simulated pedigrees and the average number of inter-river relationships detected per pedigree was computed. The expected average number of relationships depends on the strength of the relationships of the migrant with the individuals of its river of origin and therefore it can be different for each simulated pedigree.

Using the pedigrees with migrants, additional crosses between the generated offspring were simulated to obtain inbred individuals of known ancestry. Then the performance of the SNP data set to estimate the individual inbreeding coefficients of these inbred individuals was tested. These individuals were obtained from crossing full siblings, half siblings, first cousins and half-first cousins (Figure S3).

1.5. Proportion of heterozygous positions in each specimen and Hardy-Weinberg equilibrium test

The proportion of heterozygous positions for each individual should be estimated from all loci, including both variable and invariable loci, to be comparable with genome-wide estimates (Prado-Martinez et al., 2013; Robinson et al., 2016). In addition, a MAF filter should not be used to avoid altering this estimation. For this reason, a new data set was generated with the POPULATIONS program using filters r = 1, m = 12, and MAF = 0, which resulted in 7,583 total loci (all loci of data set 3, Table S3). The proportion of heterozygous positions was estimated from these sequences in FASTA format as the number of all heterozygous positions of the specimen divided by the total length of the loci.

Deviations from Hardy-Weinberg equilibrium were assessed using the exact test implemented in GENEPOP Version 4.6 (Rousset, 2008), using rivers as populations. Heterozygote deficiency and excess tests were also implemented.

1.6. Genomic tree

A distance phylogenetic tree was constructed as in Querejeta et al. (2016). The matrix of the average genomic divergence between specimens was constructed using the 1,262 variable loci generated with the same filters as above: r = 1, m = 12, and MAF = 0 (variable loci of data set 3; Table S3). Because nuclear genomes are diploid, a method is needed to summarize the divergence of the two alleles per individual in the calculation of the distance matrix. For this purpose, pairwise distances between all specimens were calculated using the formula 8.2 taken from Freedman et al. (2014). Basically, for each variable position being compared between two individuals, the average of the four possible matches between the two individuals is computed. The resulting pairwise distance matrix was used to construct a distance tree using the FITCH program in the PHYLIP package (Felsenstein, 1989).

1.7. Principal component analysis

Principal component analysis (PCA) was performed with the R program SNPRELATE (Zheng et al., 2012) and the genetic covariance matrix of data set 2 (Table S3). The axes of the plot were orientated to maximize the positional coincidence between the specimens in the plot and their geographical locations.

1.8. Population structure

Population structure and admixture proportions were estimated using the SNP data set 2 (Table S3) with the program STRUCTURE 2.3.4, which implements a Bayesian model-based clustering method (Pritchard et al., 2000). The analysis was performed using the admixture and correlated allele frequency models and with no prior information about population origin. A total of 500,000 iterations were run after a burn-in of 50,000 iterations and with a number of clusters (K) ranging from 1 to 10. A total of 10 independent runs were performed for each K value to plot the trend of the estimated posterior probability for the data, Ln P(D) (Pritchard et al., 2000). In addition, the optimal K value was assessed using the ΔK method (Evanno et al., 2005) as implemented in STRUCTURE HARVESTER (Earl & VonHoldt, 2012).

2. Reconstruction of pedigrees of an elusive mammal from genome-wide data

2.1. Samples of Pyrenean desman from Zamora

A total of 73 samples from Pyrenean desmans of two rivers in Zamora, the Tera and Tuela, were used (Table S4, Figure 27). The samples consisted of 18 tissue and 55 hair samples. All the samples were obtained from specimens captured between 2014 and 2016, during a project monitoring the effects of the construction of a high-speed railway line. All permits were issued by the regional government Junta de Castilla y León. Whenever possible, a transponder was placed under the skin in captured animals so that they could be identified in subsequent captures. Then, either a small piece of the tail tip (1 – 2 mm) or hair sample was taken before the captured specimens were released back into the wild.

Age class at time of the capture was determined for 48 specimens by taking a picture of the teeth and estimating the degree of dental wear according to a previous study relating dental wear with age (González-Esteban et al., 2002). The work was carried out following national and international regulations, and all necessary steps were taken to reduce stress levels and prevent any harm to the animals. All the samples were preserved in tubes containing absolute ethanol and stored at -20ºC in the laboratory.

2.2. Sample processing

2.2.1. DNA extraction

DNA from fresh tissue and hair samples was extracted using the DNEasy Blood and Tissue Kit (QIAGEN), following manufacturer's instructions. DNA extractions from hair samples were carried out in a separated UV-irradiated are with dedicated equipment.

2.2.2. PCR sequencing of mitochondrial sequences

The complete cytochrome *b* gene (1,140 bp) was amplified from all the samples using the methodology previously described in Methods section 1.2.2.

2.2.3. Quantification of DNA concentration

Quantification of DNA concentration for all samples was performed as previously described in Methods section 1.2.3.

2.3. Construction and sequencing of genomic libraries

2.3.1. Library construction and Illumina sequencing

DNA libraries were constructed using the ddRAD protocol (Peterson et al., 2012). In order to be able to recover enough information from as many samples as possible, a modification by which each sample was processed independently was introduced. This way, samples for which the PCR amplification was weak were amplified with more cycles, re-processed or discarded before making the library pool. This is an important advantage over having to wait for the sequencing run to determine the yield of each sample and repeating samples with low sequence yield in a new experiment.

Each library was performed in series of 24 samples. Genomic DNA concentration was estimated by qPCR. Thanks to the introduced modification, a quantity as low as 10 ng was processed for some samples, allowing for the possibility to use hair samples for which DNA yield was very low. For other, more concentrated samples, up to 200 ng were used. Each sample was processed as previously described in Methods section 1.3.1 for the digestion, ligation, and size-selection steps. PCR amplifications were performed independently for each size-selected sample in order to add Illumina adapters with the conditions previously described in Methods section 1.3.1 and employing initially 20 cycles. When the intensity of the initial PCR products was weak, a new PCR was performed with 25 cycles. Samples that did not have a clearly visible band were discarded. Then PCR products of different samples were combined at similar concentrations (as estimated by the intensity of the bands) and concentrated into a 30 μl volume using the MinElute PCR Purification Kit.

Finally, 400 ng of the DNA library (as estimated by NanoDrop) were run in a precast E-Gel EX 2% agarose gel and the band corresponding to the library was extracted in 30 μl with the QIAquick Gel Extraction Kit. The libraries were sequenced in single-read runs using the NextSeq Sequencing System (Illumina) and 150-cycles Mid Output kit in the Genomics Core Facility at the Pompeu Fabra University.

2.3.2. Sequence filtering and processing

The STACKS 1.35 package (Catchen et al., 2013) was used to process the sequences obtained. The parameters used for the initial processing were the same as in the Methods section 1.3.2: recovery option and quality score limit (s) between 10 and 25, depending on the overall quality of the run, for PROCESS RADTAGS; minimum stack depth (m) = 3 and maximum differences between reads (M) = 2 for USTACKS; and maximum mismatches between reads (n) = 2 for CSTACKS. Using the POPULATIONS program of the package, sequences of loci in FASTA format were saved with a minimum proportion of called individuals (r) of 0.51, a minimum stack depth or coverage (m) of 12, and a minimum minor allele frequency (MAF) of 0.

In order to obtain a loci set free from bacteria or parasites that could be present on hair samples, the subset of 18 tissue samples was processed separately. With the loci assembled this way and using one sequence per locus, a tissue sample database was built with the BOWTIE-BUILT tool from BOWTIE 2 v2.3.0 (Langmead & Salzberg, 2012). All the specimens were then processed by applying an additional filter that only kept reads with a match to the tissue samples database. For this, BOWTIE 2 was used with the default option "*—score-min L,-0.6,-0.6*".

The reads remaining after this filter were processed with STACKS, as described previously. The final SNP data set was generated with the filters r = 0.9, m = 12, and MAF = 0 of the POPULATIONS program, which preliminary experiments showed to be optimal parameters for the relatedness analysis. The SNPs were saved in PLINK flat and VCF formats, and loci sequences in FASTA format, for further analyses. Only one SNP per locus was saved.

2.3.3. Sex determination using ddRAD data

Sex determination was performed with a bioinformatic protocol developed here based on ddRAD sequences. Using the sequences of the 49 loci belonging to the Y chromosome found in Querejeta et al. (2016), a Y-chromosome loci database was built with BOWTIE-BUILD. Then BOWTIE 2 was used to map all the endogenous sequences of each sample to the Y-chromosome database in order to determine their sex. To perform a sensitive search and avoid matches to X-chromosome sequences, BOWTIE 2 was used with the option "*—score-min L,0,-0.05*".

2.4. Estimation of pairwise relatedness and individual inbreeding coefficients

Pairwise relatedness among Pyrenean desmans was calculated using the *dyadml* estimator (Milligan, 2003) of the program RELATED (Pew et al., 2015; Wang, 2011), as was previously shown to be the best choice for SNPs (see Results section 1.1.2) and that was confirmed through an equivalent analysis using this data set (see Results section 2.1.2). Inbreeding was taken into account, so the full nine identity-by-descent (IBD) states model was used and therefore the inbreeding coefficient of each specimen was also estimated. Confidence intervals (95%) for the estimation of relatedness were calculated using bootstrapping over loci (100 replicates). To avoid false positives, only relatedness estimates where the lower 95% confidence limit of the bootstrap replicas was higher than 0 were considered.

The genetic identification of initial samples was performed using the three IBD-states model of the RELATED program. In this model, replicas of the same individual are expected to have a relatedness coefficient of 1.

To test the reliability of the relatedness coefficients obtained, simulations of genotypes based on 7 artificial pedigrees (Figure S4), in which individuals with different origins were computationally crossed using the script GetCrosses.pl (Methods section 1.4.4.), were performed. The accuracy of the inbreeding coefficients was also assessed using 7 pedigrees in which additional crosses between full and half siblings from the previous pedigrees were performed (Figure S5). A total of 100 simulations were carried out for each pedigree.

2.5. Pedigree reconstruction from genomic data

2.5.1. Determination of kinship categories

In order to determine kinship categories between pairs of individuals, two different programs were used. The first one was PRIMUS version 1.9.0 (Staples et al., 2014). The input for this program are the IBD coefficients (IBD0, IBD1, and IBD2), which are the proportions of loci shared between two individuals on 0, 1, and 2 alleles, respectively. These coefficients were estimated for all pairwise comparisons of individuals with the RELATED program using the three IBD-states model. PRIMUS classifies the relationships into the following categories: parent-offspring, full siblings, second-degree relationship, third-degree relationship, and distantly related pairs. A number of relationships were classified as unrelated and not used in subsequent analyses.

The second program was VCF2LR, which takes inbreeding into account (Heinrich et al., 2017). The SNP file in VCF format generated using the *POPULATIONS* program was provided as input. VCF2LR also needs a file of genotype counts from the studied population. The custom script GenotpeCounts.pl was made to calculate the genotype counts from the SNP data set. Using this information in a maximum-likelihood framework, VCF2LR classifies the relationships into the following categories: parent-offspring, full siblings, and second-degree relationship. Pairs of individuals not classified by the program but with a significant relatedness value were designated as "others".

To test the performance of the kinship categories estimated by VCF2LR and PRIMUS, 100 simulations of genotypes based on an outbred pedigree were performed using the script GetCrosses.pl (pedigree 7 from Figure S4). In addition, to test a scenario with a high degree of inbreeding, 100 simulations based on a pedigree with several inbred matings were performed (Figure S6).

2.5.2. Pedigree inference

The large number of distant relationships found between specimens and the high degree of inbreeding in many of them gave rise to pedigrees that were too complex to be drawn by automatic scripts. Therefore, pedigrees were manually reconstructed using the closest kinship categories.

For the representation of the inferred pedigrees, pairs of individuals obtained into the parent-offspring category with the software VCF2LR were first taken into account. The software only gave the orientation of the parent-offspring pairs in families with several siblings. In order to determine the orientation of the other pairs, two sources of evidence were used. First, the age of the individuals was considered as estimated by the dental wear of the captured specimens (González-Esteban et al., 2002). Second, information about mitochondrial haplotypes was considered to detect incongruences between mothers and their children, as they should have the same haplotype since mitochondrial DNA is inherited through the maternal line.

Then, relationships between siblings of each family were verified by searching the pairs of putative siblings in the categories of full siblings and second-degree relationships for half-siblings. The inbreeding coefficient in full siblings was also compared, which should be similar as they share the same parents.

2.6. Estimation of dispersal distances

The matrix of distances between all pairs of individuals was calculated as the shortest distance along the course of the river. First, a shapefile of the river network in the studied area was obtained from the web page of the corresponding river authority, the Confederación Hidrográfica del Duero (www.chduero.es). Then, using the R package SECRLINEAR (Efford, 2017), the shapefile was converted to a *linearmask* object using the *read.linearmask* function and the pairwise distance matrix was then obtained with the *networkdistance* function. When two points belonged to the two different rivers, they were treated as if they were connected through the watershed divide, which is the most likely dispersal path in this case.

Using these distances, the average river distance between dyads for each kinship category and the mean dispersal distance per generation were calculated. As an estimation of generations, the minimum number of pedigree connections between two individuals was used: 1 for parent-offspring pairs, 2 for full siblings, 2 for second-degree relatives, and 3 for third-degree relatives. The remaining categories (“others” in VCF2LR, and “distantly related” and “unrelated” in PRIMUS) were not used. Given the uncertainties in how each member of a pedigree disperses, these calculations can give an approximate estimation of the effective dispersal distance per generation.

3. Quantitative analysis of connectivity between populations using kinship categories and network assortativity

3.1. Samples of Pyrenean from Zamora and SNP data set used

Genotype data of 70 Pyrenean desmans from the previous study (Tera and Tuela rivers of Zamora) were used (Table S4, Figure 33). The data set used consisted of 3,874 SNPs, as obtained with the POPULATIONS program from the STACKS 1.35 package (Catchen et al., 2013) with the filters r = 0.9, m = 12, and MAF = 0 (see Methods section 2.3.2).

3.2. Population structure

The population structure and admixture proportions were estimated with the program STRUCTURE 2.3.4 (Pritchard et al., 2000), using the admixture and correlated allele frequency models and no prior information on population origin. A total of 1,000,000 iterations were run after a burn-in of 100,000 iterations, and with a cluster number (K) ranging from 1 to 6. A total of 10 independent runs for each K value were performed, which were summarized using the CLUMPP software (Jakobsson & Rosenberg, 2007). Given the difficulties in obtaining a single appropriate K value (Janes et al., 2017), the admixture plots for several K values were generated.

3.3. Kinship networks

Determination of kinship categories among pairs of individuals was performed with VCF2LR (Heinrich et al., 2017), as previously described in Methods section 2.5.1. The input was the VCF file generated with the POPULATIONS program, as well as a genotype counts file generated with the script GenotypeCounts.pl.

Pairs of individuals were classified according to their kinship category determination and the networks for each kinship category were plotted using the program GEPHI and the plug-in GEOLAYOUT (Bastian et al., 2009), and superimposed on a map.

3.4. Assortativity coefficient

In order to provide a quantitative estimation of connectivity across a barrier in the kinship categories networks, the assortativity coefficient between different classes or sectors of a network was calculated. This coefficient is based on a comparison of the fraction of edges e_{ij} that connects nodes from the same or different sectors. Following the notation from Newman (2003), simplified for undirected networks, the assortativity coefficient is:

$$AC = \frac{\sum_i e_{ii} - \sum_i a_i^2}{1 - \sum_i a_i^2}$$

where e_{ij} is the proportion of edges in the network that connect two individuals from the same sector i and $a_i = \sum_j e_{ij}$ is the proportion of edges extending from sector i to nodes of sector j, for all values of i.

In this study, only two sectors were tested at a time, A and B, on either side of a putative barrier, so the formula can be reduced to:

$$AC = \frac{(e_{AA} + e_{BB}) - (a_A^2 + a_B^2)}{1 - (a_A^2 + a_B^2)}$$

Also, $a_A = e_{AA} + e_{AB}$ and $a_B = e_{BB} + e_{AB}$. Taking into account that:

$$e_{AA} + e_{BB} + 2e_{AB} = 1$$

$$a_A + a_B = 1$$

and therefore that

$$a_A^2 + a_B^2 + 2a_A a_B = 1$$

then AC can be expressed as:

$$AC = \frac{2a_A a_B - 2e_{AB}}{2a_A a_B}$$

where $2e_{AB}$ is the observed proportion of edges between sectors A and B, and $2a_A a_B$ represents the expected proportion of edges between sectors A and B if the nodes were randomly connected.

The assortativity coefficient therefore measures the proportion of missing edges between two sectors with respect to the expected proportion of edges between these sectors. When there are fewer connections than expected between two sectors (for example due to a barrier), then $AC > 0$ and it can reach 1. When two sectors are perfectly mixed, then $AC = 0$. Values of $AC < 0$ indicate more connections than expected between two sectors (Newman, 2003).

The expected standard deviation was calculated with the formula from Newman (2003):

$$\sigma_{AC}^2 = \sum_{i=1}^{n} (AC_i - AC)^2$$

where n is the number of edges in the network and AC_i is the value of AC for the network in which the ith edge is removed. Finally, to determine the p-value of the assortativity coefficients obtained, 1 million random networks were simulated using the same numbers of individuals available in each sector. All these calculations were performed with the custom script Assortativity.pl.

Locations of barriers more than 3.5 meters high were obtained from the Confederación Hidrográfica del Duero (www.chduero.es). Localities adjacent to the two sides of each barrier were used to calculate the assortativity coefficient. The same approach was used for the watershed divide.

IV. RESULTS AND DISCUSSION

1. Using relatedness networks to infer contemporary dispersal: application to the endangered mammal *Galemys pyrenaicus*

1.1. Results

1.1.1. Phylogeography of the Pyrenean desman in the contact zone of two ancient mitochondrial lineages in the Iberian Range

Cytochrome *b* sequences from 66 specimens of Pyrenean desman (Tables S1 and S2) were obtained from different localities in the Iberian Range. A maximum-likelihood phylogenetic tree was constructed from these sequences to assign each specimen to its mitochondrial lineage (not shown). Figure 22a shows an overview of the lineages distribution in the whole range according to Igea et al. (2013) and Figure 22b shows the map of the specimens used in this work coloured by lineage. This map revealed the presence of a spatial separation between lineages B1, restricted to the Oja and Najerilla rivers, and A2, mainly found in the Iregua, Tera and Duero rivers, and some areas of the Najerilla. Three different mitochondrial lineages were found in the Umbría river (A2, B1, and B2). Individuals of lineage B2 had not been detected in the Iberian Range so far (Igea et al., 2013). Interestingly, females from the Iregua river (lineage A2) appeared to have migrated to Najerilla and Umbría, but no B1 female was found in Iregua, indicating a certain element of asymmetric past migration.

Of these specimens, 37 present in four rivers (Iregua, Najerilla, Oja, and Umbría) were used for library construction and genomic analysis. The sex of each of these specimens was determined by qPCR, resulting in 20 females and 17 males (Figure 22b and Table S1).

1.1.2. Sequence assembly and filtering parameters

After sequencing the ddRAD libraries, a total of 183,708,284 Illumina reads passed the initial quality filters, with an average of 4,965,089 reads for each of the 37 *G. pyrenaicus* specimens analysed and a mean depth of coverage of assembled loci of 42.5 (Table S5). It should be taken into account that this coverage includes possible PCR duplicates, which cannot be detected in standard ddRAD libraries, and therefore real mean coverage may be lower. An initial data set was generated with filters r = 1, m = 9, and MAF = 0 in the POPULATIONS program, resulting in 1,651 SNPs (data set 1, Table S3).

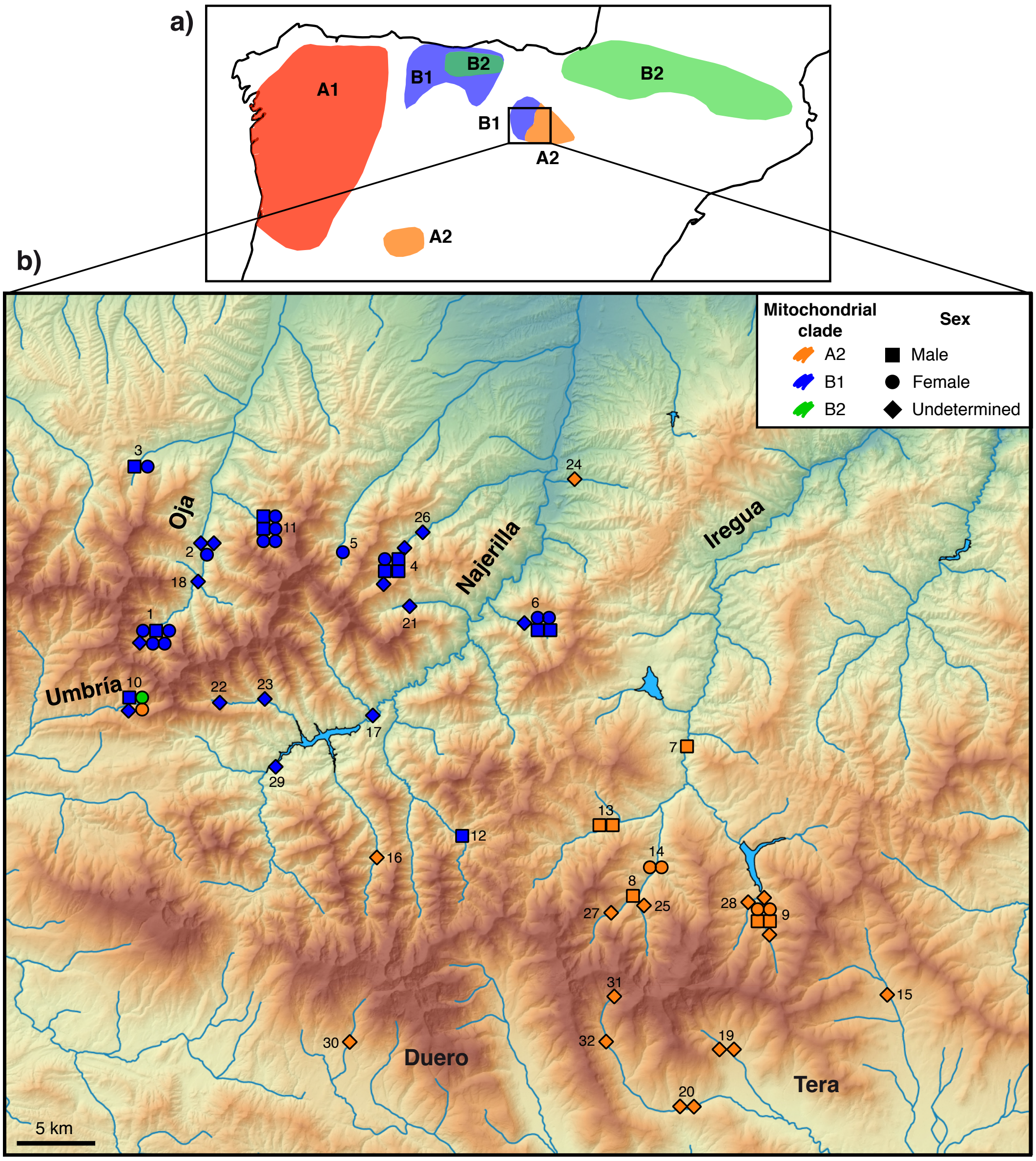

Figure 22. (a) Map of the northern part of the Iberian Peninsula showing the distribution of the main mitochondrial lineages found in Igea et al. (2013) and (b) map of Pyrenean desman specimens of the Iberian Range used in this study (b). Different colours reflect the mitochondrial lineage to which they belong (A1: red, A2: orange; B1: blue; B2: green). Samples used only for cytochrome *b* sequencing are shown with a diamond and samples used for ddRAD sequencing are shown with squares (males) or circles (females). Localities are indicated with numbers as in Tables S1 and S2.

Optimizing these filters was needed to generate a data set that would have an optimal performance in relatedness analysis. Therefore, the statistical properties of the different relatedness estimators implemented in the program RELATED were assessed. Using the initial SNP data set, the performance of different estimators was evaluated through simulations based on the allele frequencies. The distribution of relatedness values for different kinship relationships was obtained (Figure S7) and their means and standard deviations were calculated (Table S6). Although the means were close to the expected values for most estimators, the *dyadml* (Milligan, 2003) and *trioml* (Wang, 2007) maximum-likelihood estimators showed the lowest standard deviations. In addition, the maximum-likelihood estimations presented the best correlations with the expected values (R=0.994 in both cases; Table S7). Finally, relatedness values between unrelated individuals were mostly zero or close to zero for the maximum likelihood estimations (Figure S7), indicating a low level of false positives for them. Therefore, both maximum-likelihood estimators showed the best overall properties for the data set. The *dyadml* estimator was chosen for subsequent analyses, as it required less computational time than the *trioml* estimator.

Then, the filters of POPULATIONS were optimized to generate the SNP data set. For this purpose, four pairs of duplicated samples, for which a relatedness value of 1 is expected, were used. Using data sets generated with different filters and estimating relatedness from them with *dyadml* (Figure S8), filters r = 1 and MAF = 0.05 (solid blue line in Figure S8) were found to give the best overall relatedness values for values of minimum coverage of 12 and 15. The filter m = 12 was chosen because it yielded the highest number of SNPs. In this way, 912 SNPs were obtained (data set 2 of Table S3), which were used for the final relatedness estimations and other SNP-based analyses. Using these parameters, the average discrepancy in genotypes of the four replicated samples was 0.33%. As r was set to one, all samples were genotyped for all SNPs, and therefore there was no missing data. Additionally, the FASTA sequences with the same filters except that they were unfiltered for MAF (data set 3) were used for the genomic tree reconstruction and the heterozygosity estimation.

Hardy-Weinberg equilibrium was tested for data set 2. A total of 134 SNPs (15%) were found to have significant deviations in at least one of the four sampled populations (rivers), mostly due to heterozygote deficiency.

1.1.3. Genomic tree, PCA and STRUCTURE

The genomic tree of the specimens revealed the presence of four groups comprising individuals from the same river, although the individual IBE-C3745 from the Najerilla was grouped with those from the OJA (Figure 23a). Interestingly, all individuals from the Umbría river formed a single genomic group despite having different mitochondrial lineage origins. Additionally, the 11 specimens of mitochondrial lineage A1 belonged to two different genomic groups (10 to the Iregua group and one to the Umbría group). PCA basically showed the same results, grouping most individuals according to their river of origin and not by their mitochondrial lineage (Figure 23b).

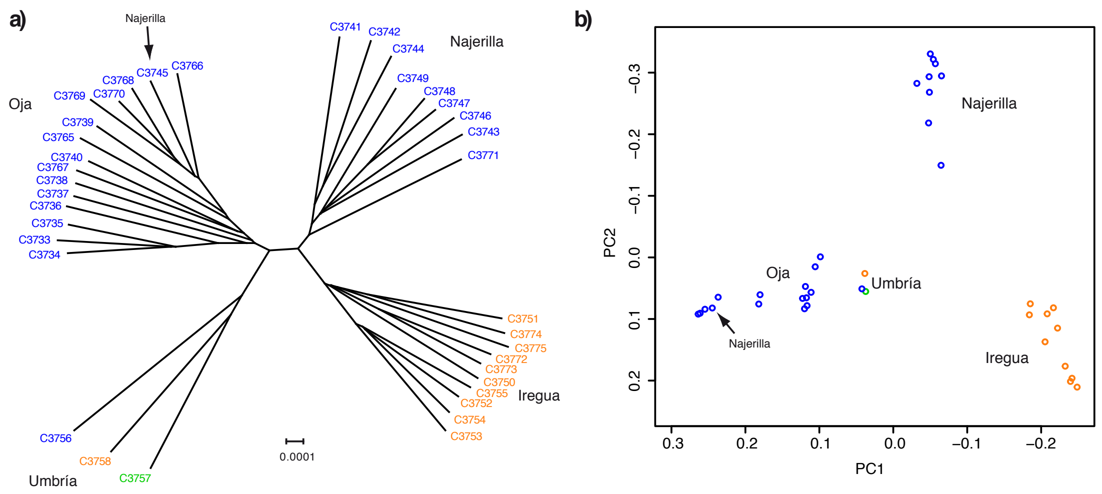

Figure 23. Maximum-likelihood phylogenetic tree (a) and principal component analysis (b), with specimens of Pyrenean desman colour-coded by mitochondrial lineage as in Figure 22 (A2: orange, B1: blue, B2: green). The scale of the tree is in substitutions per position.

The STRUCTURE analysis showed good convergence of the different runs for K values between 1 and 6 (Figure S9). The Evanno method (Evanno et al., 2005) was used to detect the point of inflection of the likelihood curve. According to this method, there were two peaks (Figure S9). The main one was at K = 2, but this peak is far from the saturation point of the likelihood curve, whereas the peak at K = 4 is closer to this point and may correspond to a better model. Additionally, when 134 SNPs that were not in Hardy-Weinberg equilibrium were removed, an additional peak appeared at K=6 (not shown). All this made it difficult selecting a single K value, and therefore several K values were considered to show the hierarchical nature of structure (Betto-Colliard et al., 2015). Figure 24 shows the admixture proportions from K = 2 to K = 6 (these proportions remained virtually identical after removing SNPs that were not in Hardy-Weinberg equilibrium). With K = 2, there is a subdivision basically between specimens of

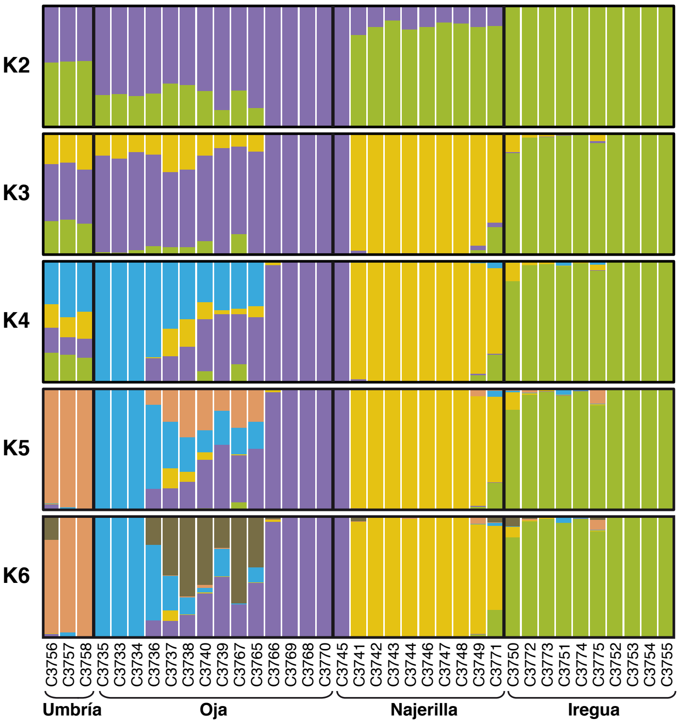

Figure 24. Bar plot of admixture proportions of each Pyrenean desman specimen as determined with STRUCTURE and different K values.

the Iregua and Najerilla on one side and those of the Oja and Umbría rivers on the other, with the exception of one individual (see below). With three clusters, the Iregua and Najerilla rivers become separated. With four, five, and six clusters, new subpopulations appear in the Oja and Umbría rivers, with the Umbría river basically becoming a single population in the models with five and six clusters. As previously noted in the genomic tree and PCA, the individual IBE-C3745, found in the river Najerilla, had the same genomic composition as some individuals in the Oja under all K values. Additionally, as seen from the mitochondrial data, certain asymmetric past migration can be appreciated, particularly in the models with 3 and 4 clusters: Iregua genome components were present in the river Najerilla and Najerilla components in the rivers Oja and Umbría, but no admixture was observed in the opposite direction.

1.1.4. Pairwise relatedness and relatedness networks

A total of 160 relatedness values were found between individuals for whom the lower 95% confidence limit of the bootstrap replicas was higher than 0. These values were used to construct relatedness networks (Figure 25). The minimum pairwise relatedness was 0.0625, the maximum 1.1866, and the average 0.3564. To simplify the networks, these relationships were subdivided into those with a relatedness value above 0.2018 (close kinship relationship) and those below this value (distant kinship relationships). This limit corresponds to the lower 95% confidence limit of the simulations of half siblings with the *dyadml* estimator (Figure S10). Therefore, the network of close kinship covers relationships equivalent to half siblings (second degree) and above. The minimum relatedness value considered (0.0625) corresponds to the equivalent of a fourth-degree relationship, and therefore both networks together provide a picture of dispersal occurred in the last four generations at most.

The network of close kinship relationships (Figure 25a) showed that most relationships were between individuals from the same river. In total, 112 intra-river relationships were found. In comparison with these, there were only 7 inter-river relationships. All of these inter-river relationships were due to the individual IBE-C3745, probably a first-generation migrant from the Oja river to the Najerilla as deduced from several relatives found in the former. The network of more distant kinship relationships (Figure 25b) also showed an overall pattern of mainly intra-river relationships, of which 36 were found. As for the inter-river relationships, apart from three more relationships between the Oja and Najerilla, two additional relationships were discovered between the Umbría and

Najerilla rivers. Relatedness networks showed a similar pattern when the estimation of pairwise relatedness was performed with a model without inbreeding (not shown).

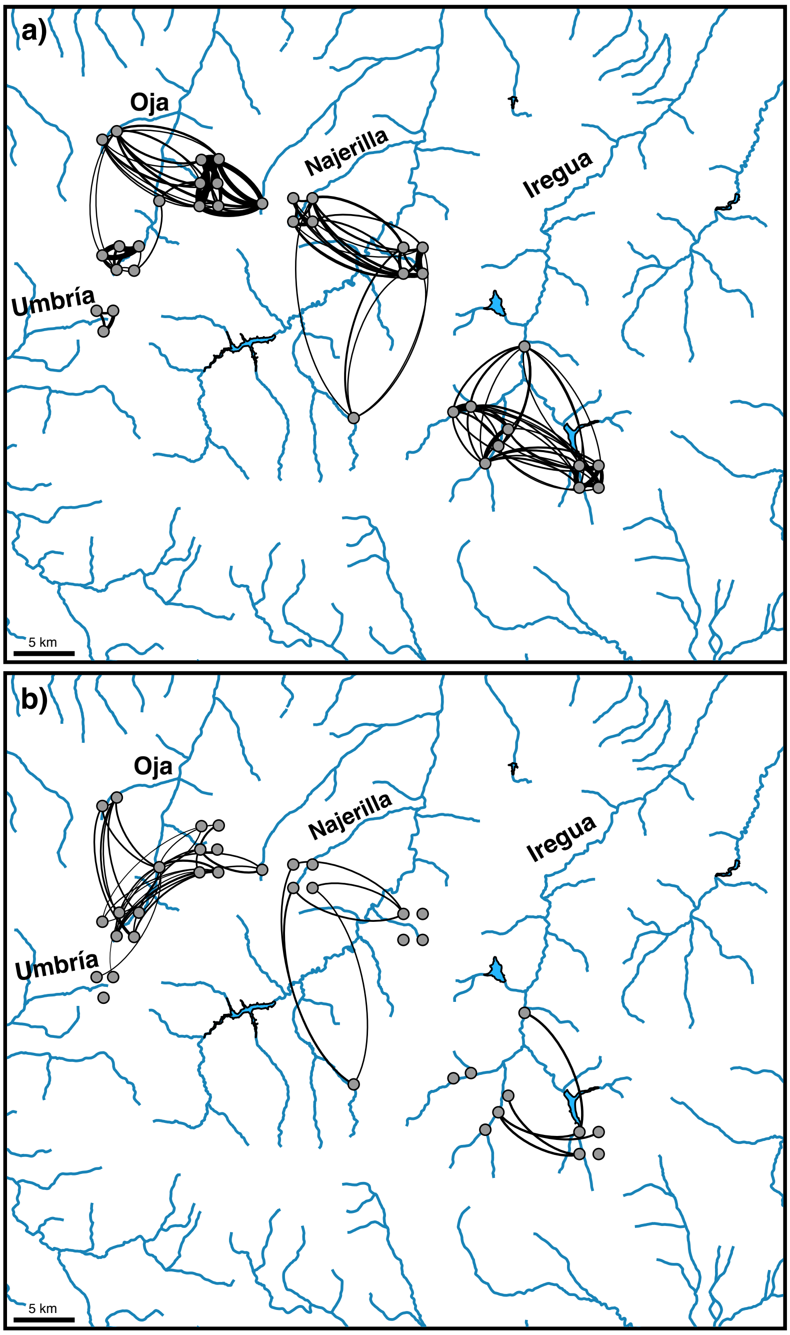

Figure 25. Map plotting relatedness networks with (a) values above 0.2018 and (b) under this value. Curved lines (edges) connect Pyrenean desman specimens (nodes) for which a relationship was found. The thickness of edges is proportional to the relatedness value of the value of the connected specimens.

To study possible male or female philopatry, the relatedness values between pairs from the same locality were analysed and classified by sex. A mean relatedness of 0.56 was obtained for female dyads (n = 16), 0.53 for male dyads (n = 7), and 0.55 for mixed dyads (n = 26). These values reflected a high level of kinship of the individuals of the same locality (with the mean equivalent to full siblings), but no significant differences were found between the three analysed classes with the Tukey-Kramer test and a p value of 0.05.

1.1.5. Individual inbreeding and heterozygosity

Individual inbreeding coefficients were high overall (mean value of 0.33), with 33 individuals of 37 having values over 0.1, indicating a close relationship between their parents (Table S8). When these coefficients were plotted on the map (Figure 26), it was observed that the specimens with the highest values were those from the Iregua (mean value of 0.42), followed by Najerilla (0.38), Umbría (0.30), and Oja (0.23). Interestingly, four individuals from the river Iregua sampled upstream of a dam (in the tributary La Vieja) presented the highest values of all, with an average inbreeding coefficient of 0.53.

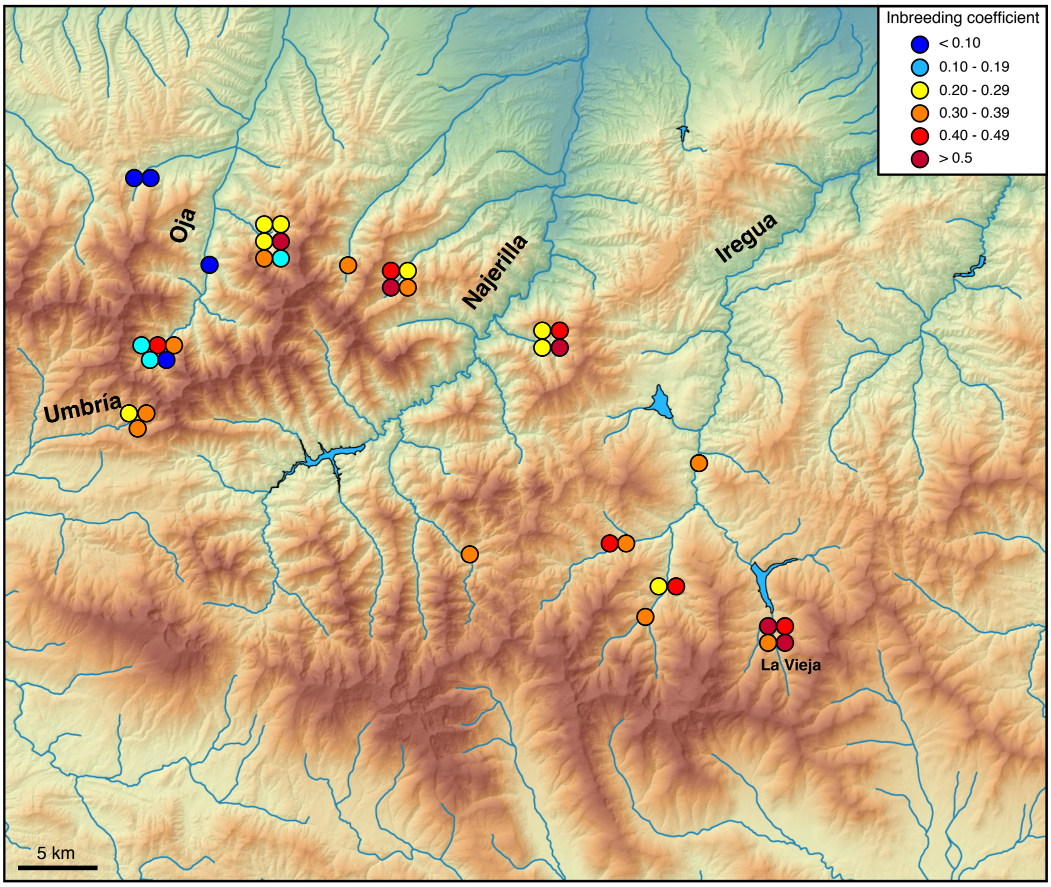

Figure 26. Map plotting colour-coded inbreeding coefficients of different specimens of Pyrenean desman.

Heterozygosity rates ranged from 103 to 322 heterozygous positions per million bases (Table S8). These values and individual inbreeding coefficients presented a strong negative correlation (R = -0.921).

1.1.6. Simulations along pedigrees using genotypes of actual specimens

The estimation of relatedness and inbreeding depends on the population used as a reference for allele frequency estimation (Milligan, 2003). Ideally, the closest possible population should be used for these estimates, which in this case could be each river. Therefore, it was tested whether the best reference for these estimations was the whole set of specimens or each of the three main rivers separately. For this purpose, simulations were performed using the actual genotypes from the data set, in which individuals were crossed computationally along specific artificial pedigrees with founders of each river (Figure S1).

When the whole set of specimens was used as reference, most relationships of the pedigree were detected (Table S9). Relatedness values were higher than theoretical ones for outbred individuals due to kinship between founders, leading to inbred offspring; these relationships could not be avoided as most desmans from each river were related. However, when each river was analysed separately, many relationships of the known pedigrees were zero, leading to means that were close to zero for most relationships (Table S9). These results may be due to the low sample size within each river, showing that in this case the best reference was the set of all specimens.

Relationships among individuals from different rivers were scarce according to relatedness analyses. In order to assess if these relationships were especially difficult to detect due to the presence of genetic structure, that is, to the comparison of specimens with different genomic compositions, further simulations were performed with artificial pedigrees in which a migrant from an adjacent river was added to the founders of the river of interest (Figure S2). When relatedness from the dyads in the pedigrees were estimated, basically all relationships were detected (Table S10, Figure S10), indicating that relatedness can be estimated using pairs of individuals with different genomic compositions. The estimated means were closer to the expected ones than in the previous simulations as average kinship relationships between the founders were lower due to the presence of a migrant (Table S10). Significantly, the descendants from the founders were found to be related to the individuals of the river from which the migrant came (Table S11), indicating that inter-river dispersal occurred in the last few generations, if it existed, can be detected with this approach.

The accuracy of the estimated inbreeding coefficients was also tested using artificial pedigrees with migrants (Figure S3). As expected for crossings of parents from different rivers, their simulated offspring showed basically no inbreeding (Table S12). When relatives of different degree from the pedigrees were computationally crossed, the inbreeding coefficients obtained for the offspring were close to the expected values (Table S12, Figure S11), suggesting that the SNPs used were valid for these estimations.

1.2. Discussion

1.2.1. Performance of relatedness estimators using SNPs

Of all available relatedness estimators, those based on maximum likelihood gave the best results for the data set according to simulations based on allele frequencies. Furthermore, the correlation between estimated and expected values for simulated dyads was much better than those reported for microsatellites (Taylor, 2015).

However, given the uncertainties that still have to be resolved about the estimation of relatedness, specific relatedness categories weren't used in this study. The high inbreeding observed in the samples would have made it even more difficult to determine these categories. Instead, relatedness networks were used to summarize the data and visualize the general pattern of relationships among the sampling sites (Figure 25).

Another problem faced was the presence of genetic structure in the data, meaning that the estimates of relatedness values were based on allele frequencies of individuals with different genomic compositions (Anderson & Weir, 2007; Thornton et al., 2012). To test if the genetic structure was strong enough to affect relatedness estimations, artificial pedigrees with individuals from different rivers were constructed and their offspring was simulated. Almost all inter-river and intra-river relationships could be detected in these pedigrees (Table 10, Figure S10). In addition, relationships between the migrants' descendants and individuals belonging to the source population were also detectable (Table S11), demonstrating that migration events between populations occurring in the last few generations can be identified.

A different strategy that could overcome potential problems caused by population structure would be to perform separate, more homogenous analyses for each population or river. However, the simulations showed that the relatively low sample size used as a background in each river meant that many relationships went undetected, indicating that the entire data set provided a better baseline in this case. Furthermore, this approach would have precluded the study of inter-river dispersal. Finally, even if the existence of population structure or the background used had altered some of the lowest relatedness values, then the networks that only used relatedness values above 0.2018 (Figure 25a) should be robust enough to perform an analysis of the dispersal patterns in the Pyrenean desman.

1.2.2. Contemporary dispersal patterns in the Pyrenean desman visualized with relatedness networks: inter- and intra-river dispersal

Networks that connected relatives were used to obtain a clear visualization of dispersal patterns. These networks first showed that most of the relationships corresponded to desmans sampled in the same locality or in different localities in the same river (148 relationships; Figure 25). The relatedness values between individuals sampled in the same locality were particularly high, with many values corresponding to the equivalent of full siblings or half siblings. Dyads of both males and females showed high relatedness values. Therefore, these results can be explained by the existence of close family groups due to a strong philopatry of both males and females. Nevertheless, sampling in each locality was small, so it was not possible to determine whether this trend was more predominant in one of the sexes. Data from radio tracking studies into this species in a specific river of the eastern Pyrenees did not reveal any differences in movements between males and females (Melero et al., 2014), but it is still unknown whether they exhibit different long-range movements. Therefore, further studies are necessary to better understand the philopatric behaviour of the Pyrenean desman.

The analysis performed also revealed some long distance relationships within rivers. For example, in the river Najerilla, the two most distant sampling localities with relatives that suggest a connection between them (Ormázal and Tobía), are separated by approximately 50 km of watercourse (Figure 25). It is important to note that this approach cannot determine the route of a dispersal event that must have occurred to give rise to a relationship between different localities. Another possibility is that individuals from an unknown locality independently dispersed to the two connected localities. Notwithstanding, when there are two localities with relatives it can be inferred that they are genetically interconnected by some route and that at least one migration event must have taken place through this route in the last few generations.

The river system studied included three rivers running in parallel (Oja, Najerilla, and Iregua) and a fourth river, Umbría, whose headwaters are near those of the Oja.

Importantly, the relatedness networks revealed that, despite the proximity of some sampling localities in adjacent rivers, there were few relationships between the four rivers (12 relationships; Figure 25). The few inter-river relationships observed connected Najerilla with Oja and Oja with Umbría, while no connection was observed between the adjacent Najerilla and Iregua rivers. It is therefore clear that Pyrenean desmans do not frequently move between different rivers. Connectivity between rivers further downstream is unlikely due to suboptimal conditions for the Pyrenean desman as rivers become larger (Palmeirim & Hoffmann, 1983). Consequently, it is likely that the few connections observed between rivers took place overland and across watershed divides. Further studies into the habitat suitability of the overland corridors will be vital in understanding why some accesses are more permeable than others.

In conclusion, the Pyrenean desman showed a low level of contemporary inter-river dispersal compared to intra-river dispersal, according to the relatedness networks. These results can be compared with those from the genomic tree, PCA and STRUCTURE analysis. These methods generally provide information about more ancient migrations but, interestingly, they can also detect some recent dispersal events. The most obvious case is that of individual IBE-C3745, which was sampled in the river Najerilla but grouped with desmans from the Oja in the three analyses, thus indicating a recent dispersal event from the Oja to the Najerilla. While the genomic tree and PCA cannot be used to infer additional recent events, earlier movements between rivers can be deduced from the STRUCTURE analysis. Crucially, these methods can only detect dispersal events from differentiated populations, which are mostly inter-river movements in the present study, whereas the same methods would not detect intra-river dispersal events. By contrast, relatedness networks provide an overview of recent dispersal events both between and within rivers.

1.2.3. High inbreeding in the Pyrenean desman

The inbreeding coefficient was exceedingly high in most individuals analysed (Table S8), suggesting that many of them are the product of several generations of mating between close relatives. Similar values of individual inbreeding can only be found in highly inbred species, such as Przewalski's horse (Liu et al., 2014), or critically endangered species, such as Attwater's Prairie-chicken (Hammerly et al., 2013). These high inbreeding values for the Pyrenean desman may partly be due to a strong philopatric behaviour of both males and females. In addition, the lack of connectivity between rivers revealed here can only worsen the inbreeding situation in this area.

Finally, the low species density in some rivers (Nores et al., 1998) may facilitate that juveniles occupy new territories close to their natal site, thus increasing the chances of inbreeding (Lambin, 1994; Matthysen, 2005).

In agreement with the high inbreeding coefficients, low heterozygosity levels were found in all individuals (Table S8). These low heterozygosity values explain why most SNPs that deviated from the Hardy-Weinberg equilibrium presented heterozygote deficiency. The heterozygosity observed in the Iberian Range (between 103 and 322 heterozygous positions per million bases) was intermediate compared with that of Pyrenean desmans from the whole distribution area, which ranged between 13 and 488 (Querejeta et al., 2016). In any case, those values are among the lowest recorded to date for animals (Prado-Martinez et al., 2013; Robinson et al., 2016).

Interestingly, the highest inbreeding values were found in individuals from the small tributary La Vieja, which is situated upstream of a dam (Figure 26). It is tempting to speculate that this additional artificial barrier to dispersal may have been responsible for the increase in inbreeding. However, more specimens and data would be required to study the dispersal behaviour of the Pyrenean desman in the presence of artificial barriers and to learn whether these obstacles exacerbate inbreeding.

Nevertheless, inbreeding is not necessarily detrimental in itself. Further studies should therefore be performed to determine whether inbreeding depression affects the viability of these populations due to the presence of recessive alleles in homozygous form (Charlesworth & Willis, 2009; Hedrick & Garcia-Dorado, 2016; Kardos et al., 2016). The Pyrenean desman may have presented fragmented populations in some rivers under natural conditions, for example, due to difficult dispersal downstream in large rivers (Palmeirim & Hoffmann, 1983). Theoretically, it then may be argued that this species is less susceptible to inbreeding depression because deleterious alleles were purged during previous inbreeding situations in isolated populations (Keller & Waller, 2002; Leberg & Firmin, 2008). However, for many species past inbreeding has been shown to have little effect when it comes to purging genetic load (Ballou, 1997; Crnokrak & Barrett, 2012). Clearly the data in this study highlights the need to compare fitness traits, such as brood size or first-year survival rate, in individuals with different inbreeding levels to asses the intensity of inbreeding depression (Charlesworth & Willis, 2009; Hedrick & Garcia-Dorado, 2016; Kardos et al., 2016). It would also be of interest to study the mating system of this species and statistically test whether there is inbreeding avoidance through kin recognition or, on the contrary, there is random mating as observed in other species (Rioux-Paquette et al., 2010; Szulkin et al., 2013).

So far, no data regarding any of these crucial aspects is available for the Pyrenean desman.

1.2.4. Population history of the Pyrenean desman in the contact zone of the Iberian Range

One of the most interesting aspects that emerged from the initial genetic studies of the Iberian desman was the discovery of two contact zones for the main mitochondrial lineages, with little spatial mixing between them after the postglacial colonization (Igea et al., 2013). Similar contact zones have been found in other species, although rarely with such a strict delimitation (Gómez & Lunt, 2007). The large amount of samples taken from the Iberian Range contact zone in the present study allowed the reconstruction of the evolutionary history of the Iberian desman in greater detail than previous studies. Thus, it has been possible to better determine the distribution of the mitochondrial lineages, with A2 located mainly in the southeast of the Iberian Range (rivers Iregua, Tera, and Duero) and B1 in the northwest (rivers Oja, Najerilla, and Umbría).

Genomic analyses revealed the existence of a strong genetic structure correlated with geographic origin. According to the STRUCTURE models with more than two populations (Figure 24), the area occupied by individuals with mitochondrial lineage A2 corresponded to one of the genomic clusters while the area of lineage B1 was subdivided into several additional clusters. These data also indicated that the populations of the main rivers studied here, the Iregua, Najerilla, Oja, and Umbría, were differentiated at the genome level. Additionally, both the spatial distribution of mitochondrial lineages and the admixture levels estimated for the different individuals suggested the existence of certain levels of gene flow between the detected populations. Interestingly, migration seems to have occurred more often from the southeast to the northwest than vice versa. This means that the Iberian Range contact zone is asymmetric, as observed for other species (Johnson et al., 2015). As a likely consequence of this, the Oja river presents a higher admixture whereas the other rivers are more homogeneous (Figure 24). It is possible that specific access routes between rivers have been more permeable in one direction than in the other due to specific geographical constraints, but further data and more in-depth analyses of relationship categories are necessary to shed light on these details.

The contemporary dispersal patterns of the Iberian desman revealed here through relatedness networks are congruent with the scenario outlined above. The low dispersal levels detected between rivers are probably the cause of the slow spatial mixing of the two mitochondrial lineages since postglacial colonization of the contact zone. This low dispersal would also explain the genomic differentiation detected between the populations of the different rivers. Finally, the similar philopatry behaviour that can be deduced for males and females agrees with the genetic differentiation observed for both the mitochondrial and nuclear genomes. In conclusion, the different molecular markers and methods used here, including the elucidation of the current dispersal patterns, contributed to a better understanding of the complex evolutionary history of the Pyrenean desman in this contact zone.

1.2.5. Implications of the relatedness networks for conservation and future prospects

Much remains to be learned about the dispersal behaviour of the Iberian desman, but the evidence uncovered in this study suggests that there are too few movements between rivers in the Iberian Range. This information, along with the observation of high inbreeding levels in some rivers, points to the existence of a previously unidentified conservation problem for this endangered species. Although it couldn’t be concluded from this study that fragmentation induced by dams exacerbates this problem, the data suggest that this may be an issue that deserves further attention, especially given the large number of dams present in rivers inhabited by the Pyrenean desman (Nores et al., 1998). Additionally, the endangered populations of the Central System are known to be highly fragmented (Gisbert & Garcia-Perea, 2014), and so their long-term survival could also be compromised by low dispersal between patches and inbreeding.

Still, as already pointed out, it cannot be discarded that high inbreeding levels are tolerated by this species due to the nature of its fragmented habitat without creating inbreeding depression, another topic that needs to be addressed. In any case, it is known from studies into other species that the arrival of just a few specimens can greatly help to alleviate inbreeding problems (Åkesson et al., 2016; Vilà et al., 2003). Therefore, simple actions designed to increase the natural connectivity between Pyrenean desman populations (e.g. by improving potential corridors between rivers or making artificial barriers more permeable) could be highly beneficial for the species’ conservation.

However, previous knowledge on the degree of connectivity between populations and their genetic health are necessary before proceeding to these or other management steps. The relatedness networks proposed here to study dispersal phenomena can then become a fundamental tool to evaluate the populations and monitor the effectiveness of these actions. Such measures, if effective, would lead with time to denser relatedness connections between populations, thus helping to reduce individual inbreeding. Further studies will be necessary in different areas of the distribution range of the Pyrenean desman, and with different landscape features, to gain a greater overall understanding of the conservation problems related to dispersal patterns.

2. Reconstruction of pedigrees of an elusive mammal from genome-wide data

2.1. Results

2.1.1. Sequence assembly of individual samples of Pyrenean desman from Zamora and filtering of exogenous reads

A total of 73 Pyrenean desman samples obtained from the rivers Tuela and Tera in Zamora were used to construct ddRAD libraries. Since a transponder could not be placed in all the specimens captured, the first necessary step was to genetically identify the individuals to avoid including recaptured specimens in the final data set. Using the three IBD-states model in the program RELATED, three pairs of samples were found to have a relatedness value close to 1 (0.9955, 0.9619, and 0.9582, respectively), which were assumed to come from the same specimen. The final set therefore included 70 individuals (Table S4, Figure 27).

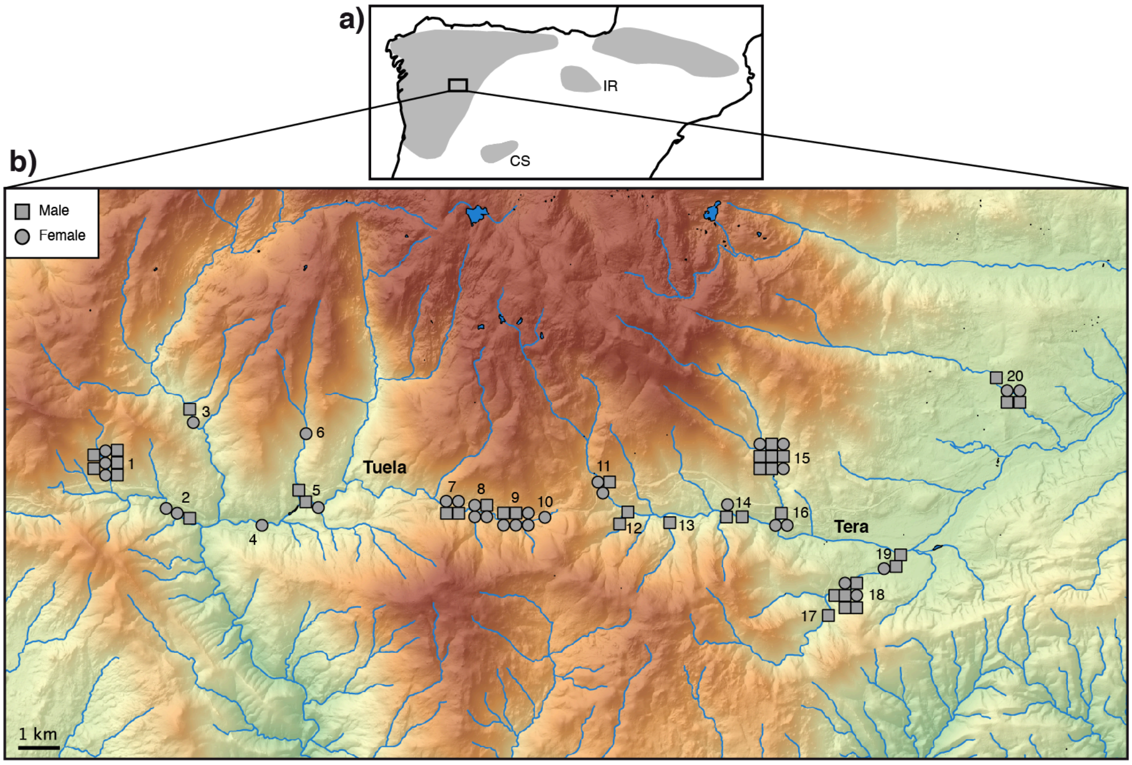

Figure 27. (a) Map of the northern part of the Iberian Peninsula showing the distribution of the Pyrenean desman. Two populations mentioned in the text, those from the Iberian Range (IR) and the Central System (CS), are indicated. (b) Map of the studied area with the Pyrenean desman specimens used in the study represented with squares (males) and circles (females). Localities are indicated with numbers as in Table S4.

A total of 336,033,963 Illumina reads passed the first quality filters (Table S13). The initial assembly rendered a much higher average number of loci per individual with hair samples (100,052) than with tail tip samples (72,898), indicating that exogenous sequences could be present in the former (Table 3). Exogenous sequences were filtered using a database of tail tip only samples, as described in the Methods section 2.3.2. The percentage of reads passing this filter was 81.6% for tail tip samples and 73.7% for hair samples, and the average number of assembled loci became similar for tail tips and hairs after filtering (44,468 and 42,824, respectively). The reduction in these numbers was due to the filtering out of exogenous sequences from different types of organisms (bacteria, parasites, etc.) that were probably present on the hairs, as well as loci not present in the library constructed for filtering.

Table 3. Statistics of the library reads obtained from tail tip and hair samples.

	Tail tip	Hair
Number of samples	18	52
Total reads per sample (before filtering)	5,085,177	4,701,938
Assembled loci per sample (before filtering)	72,898	100,052
Total reads per sample	4,023,557	3,316,289
Assembled loci per sample	44,468	42,824
Endogenous DNA (%)	81.6	73.7
Coverage	88.9	75
Heterozygosity rate per sample	0.000356	0.000352
Percentage of missing loci	5.5 %	5.0%

After filtering, the average heterozygosity was similar in tail tip and hair samples (Table 3), suggesting that the data quality was similar for both types of samples.

After merging the loci of all the samples, 16,805 loci were assembled and a data set of 3,874 SNPs was generated from the variable loci. As the proportion of called individuals r was set to 0.9, the individuals had different proportions of missing loci, ranging from 0.3% to 20.3%. No large differences in missing data were found between hair and tail tip samples (Table 3).

The genetic sexing of all specimens by mapping the reads of each sample to a 49 Y-chromosome loci catalogue (Figure 28) led to the unambiguous determination of 32 females and 38 males (Table S4).

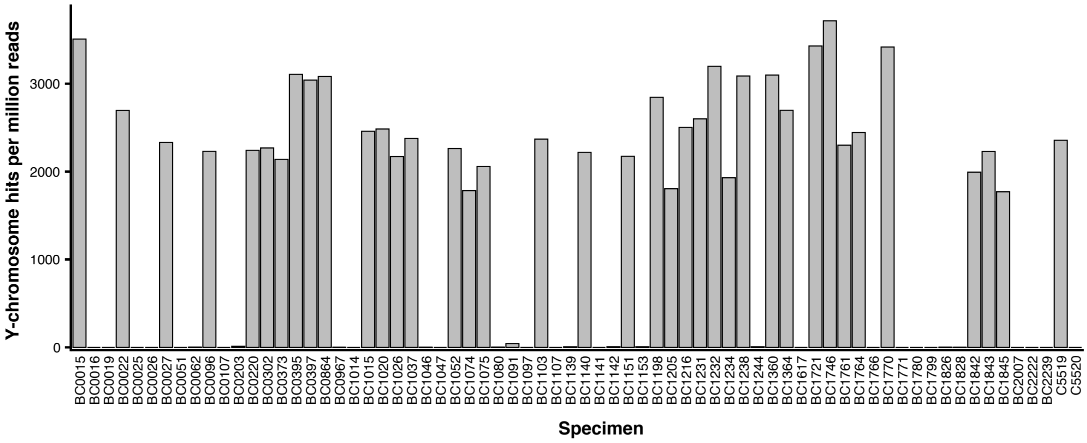

Figure 28. Plot of hits per million reads against Y-linked loci for each specimen as obtained with BOWTIE 2.

2.1.2. Pairwise relatedness and inbreeding coefficients

A preliminary analysis was performed to assess which was the best estimator for the SNP data set through simulations based on the allele frequencies of the individuals, using the 3,874 SNPs obtained. A distribution of relatedness values for different kinship relationships was obtained (Figure S12) and means and standard deviations were calculated (Table S14). As previously seen in Results section 1.1.2, the *dyadml* (Milligan, 2003) and *trioml* (Wang, 2007) maximum-likelihood estimators showed the lowest standard deviations and presented the best correlations with the expected values (Table S15). In the light of these results, the *dyadml* estimator was chosen for subsequent analyses.

A total of 697 relatedness values between individual pairs for which the lower 95% confidence limit of the bootstrap replicates was higher than 0 were found. The minimum pairwise relatedness was 0.0177 (close to a sixth-degree relationship, with a theoretical value of 0.015625), the maximum was 0.8809, and the average was 0.1652. The relatedness of dyads whose capture sites were separated by less than 500 meters of river (corresponding, basically, to specimens from the same locality) was twice as high for female pairs (0.33; n = 21) than for male pairs (0.15; n = 44). These differences were highly significant (p = 0.0007) according to a Tukey Honest Significant Differences test. However, when only adult individuals (age class 1 or higher) were considered, these values were 0.24 (n = 7) and 0.14 (n = 11) for female and male pairs, respectively. The number of pairs was too low and the difference was not statistically significant.

The mean value of the individual inbreeding coefficient was 0.11, with 31 individuals out of 70 having values over 0.1 (Table S16). The plot of the inbreeding coefficients over a map revealed higher inbreeding coefficients in the Tera river, particularly at the Mondera river (locality code 20). It was also very high in an individual of the Tuela (locality code 4), with a value of 0.3267 (Figure 29).

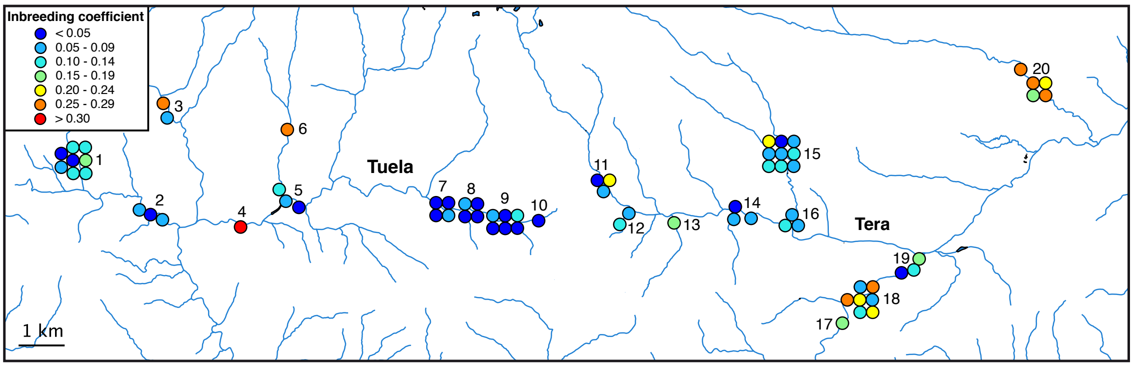

Figure 29. Map plotting colour-coded inbreeding coefficients of the specimens used. Locality codes are shown.

The reliability of the relatedness and inbreeding coefficients obtained was tested with simulations that used the actual genotypes from the data set and crossed them computationally in artificial pedigrees. Highly accurate estimates were obtained for all the relationships tested (Tables S17 and S18; Figures S13 and S14).

2.1.3. Kinship categories

Each of the 697 pairs of related individuals were assigned to a kinship category using the programs PRIMUS and VCF2LR. Using PRIMUS, 14 parent-offspring pairs, 36 full-sibling pairs, 121 pairs with a second-degree relationship, 327 pairs with a third-degree relationship, 81 distant pairs, and 118 unrelated pairs were found. When the analysis was carried out using VCF2LR, 15 parent-offspring pairs, 27 full-sibling pairs, 232 pairs with a second-degree relationship, and 423 pairs with no assigned kinship category (noted as "others") were found.

Both programs generated slightly different assignments for the different relatedness and IBD coefficients. To compare the assignment from both programs, the plot for each pair of relatedness value versus IBD1 obtained with the nine-states model in RELATED was plotted in Figure 30. Both PRIMUS and VCF2LR showed similar ranges of relatedness and IBD1 values for parent-offspring as 14 out of 15 pairs assigned into this category by *PRIMUS* were also assigned by VCF2LR, with relatedness values ranging between 0.5375 and 0.8809 in both programs. Full-sibling relationships were more different since *PRIMUS* detected 36 pairs and VCF2LR only detected 27 pairs into this category. Relatedness values for the assignment to this category had the same higher limit of 0.8491 for both programs, while the lower limit differed between PRIMUS (0.3676) and VCF2LR (0.4263). In the case of second-degree relationships, PRIMUS detected 121 pairs, with relatedness values ranging from 0.1514 to 0.5325, and VCF2LR detected 232 pairs, with relatedness values ranging from 0.1314 to 0.4823.

To test the reliability of the kinship category assignment for the two programs, simulations were performed using an artificial pedigree (pedigree 7 from Figure S4) and a pedigree with inbred matings (Figure S6); this pedigree leads to individuals with inbreeding coefficients similar to those found in the present study: an average of 0.18 for both F301 and F302.

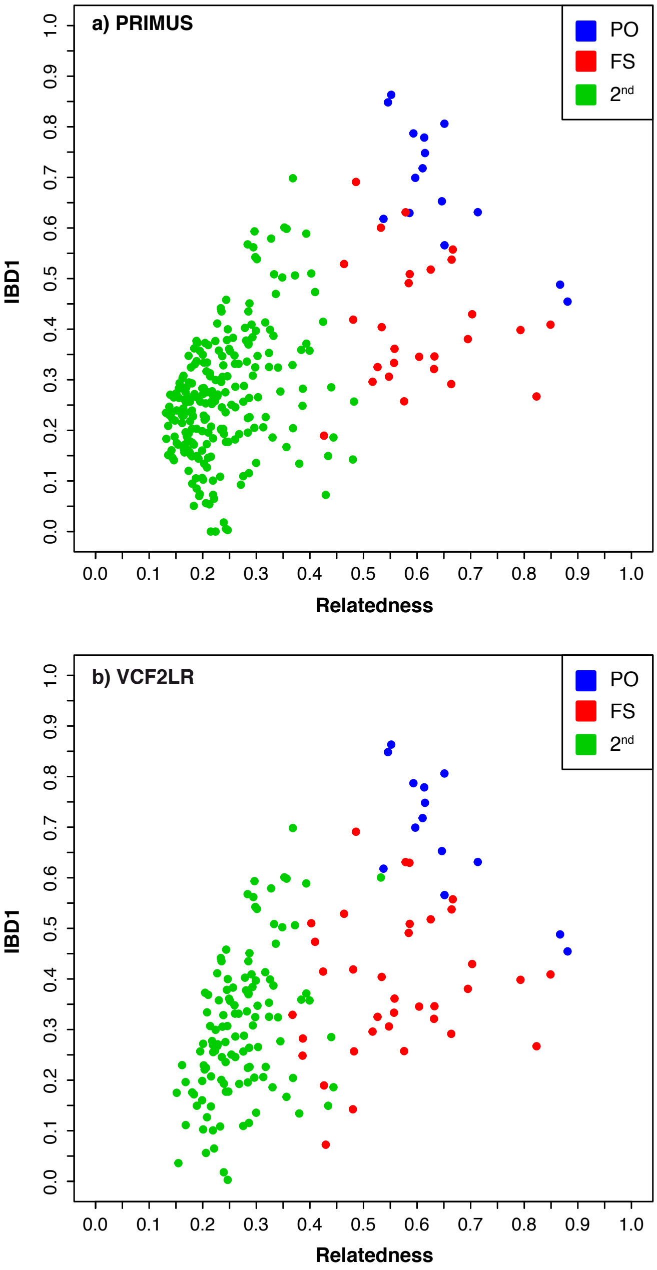

Figure 30. Plots of the relatedness coefficients vs. IBD1 values obtained with *RELATED* and colour-coded by their kinship category assignment using (a) *PRIMUS* and (b) *VCF2LR*. Kinship categories shown are as follows: PO (parent-offspring), FS (full siblings), and 2nd (second degree).

VCF2LR was able to correctly detect all kinship categories from the outbred pedigree, while PRIMUS had some problems detecting certain pairs of second-degree relationships (Table S19). For the pedigree with inbred matings (Table S20), both PRIMUS and VCF2LR were able to detect all parent-offspring and full-sibling relationships but, in the case of second-degree relationships, VCF2LR was able to correctly detect more pairs than PRIMUS. Thus, VCF2LR performed slightly better assigning kinship categories in both outbred and inbred pedigrees, especially for second-degree relationships.

2.1.4. Pedigree inference

Using the 15 parent-offspring relationships assigned by the VCF2LR program, 8 pedigrees containing at least one of these relationships were inferred (Figure 31). No pedigree contained both parents and, interestingly, more pedigrees involving the mother than the father were found (6 and 2 pedigrees, respectively). Four pedigrees (5, 6, 7, and 8) were automatically oriented by the program due to the existence of several siblings. In the other four pedigrees there was only one descendant and, without additional data, it is not possible to infer who is the parent and who is the descendants. In these cases, additional data was used to orient the pedigree. First, the age of the individuals as estimated by dental wear (González-Esteban et al., 2002) was used (Table S4). In all unoriented pedigrees, the age class was higher in one individual than the other, therefore allowing the assignation of the older individual to the parental generation. Second, information on the mitochondrial haplotypes of mothers and their children was considered, as these should be the same. This allowed the orientation of one pedigree, coinciding with the orientation provided by the individuals' ages (pedigree 1). Then the relationships between the siblings of each family and the internal consistency between them were verified. Finally, the inbreeding coefficient in full siblings was compared, which should be similar as they share the same parents. Specific information on each pedigree and how these different kinds of data were used to reconstruct and verify the pedigrees were as follows.

In pedigree 1 (Figure 31), individuals BC1037 (male) and BC0062 (female) were found to have a parent-offspring relationship. The pedigree was oriented, first, by the mitochondrial haplotype: both individuals had different mitochondrial haplotypes, indicating that individual B0062 could not be the mother of individual BC1037, as they should share the same mitochondrial haplotype. Therefore, BC1037 is the father. In addition, individual BC1037 was determined to belong to age class 4 (3-6 years old) in 2015, while individual BC0062 belonged to age class 2 (2-3 years old) in 2014, indicating that the former was more likely to be older and corroborating the father-daughter hypothesis. Both father and daughter were captured in Parada (locality code 18). The father was captured twice in the same locality, in 2015 and 2016, supporting the philopatric behaviour of the species.

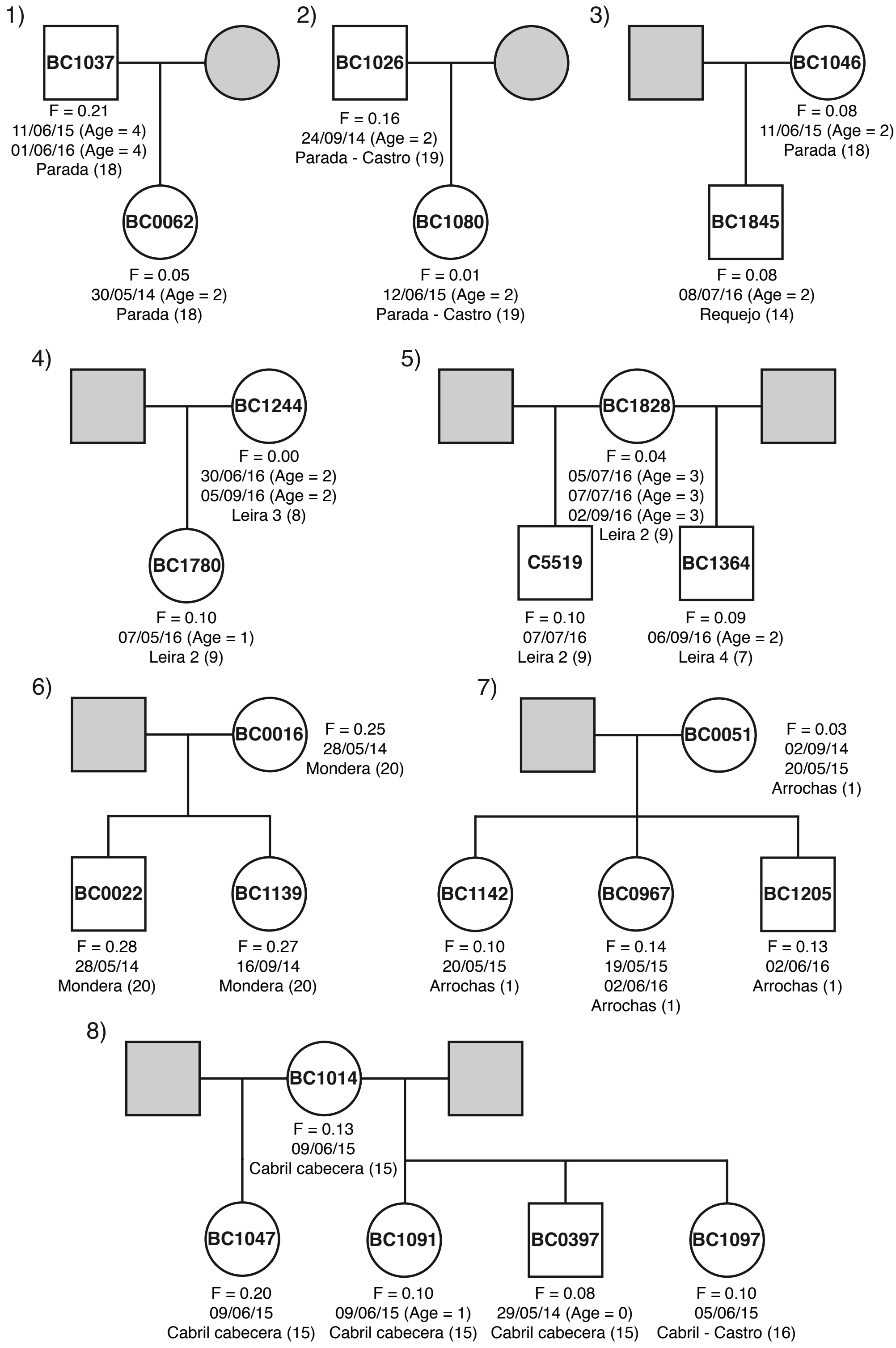

Figure 31. Pedigrees inferred from genomic data using the VCF2LR program. Males are represented with squares and females with circles. Unidentified individuals are represented in grey. For each individual, inbreeding coefficient (F), capture dates in format DD/MM/YY, age class (in parenthesis, coded as indicated in Table S4), and locality (with code in parenthesis) are given.

In pedigree 2 (Figure 31), individuals BC1026 (male) and BC1080 (female) had a parent-offspring relationship. The pedigree was oriented by the age of the individuals, as both were determined to belong to age class 2 (2-3 years old), but were captured in consecutive years, and therefore BC1026 is more likely to be the father. Both individuals were captured in the same locality, Parada-Castro (code 19). This pedigree indicates that the father had offspring at the age of one or two years.

In pedigree 3 (Figure 31), individuals BC1046 (female) and BC1845 (male) had a parent-offspring relationship. This pedigree was oriented by the age of individuals, as both mother and son belonged to age class 2 (2-3 years old) when they were captured in 2015 and 2016, respectively. Both mother and son were captured in different localities, Parada (code 18) and Requejo (code 14), separated by 7 km distance along the river, the longest dispersal distance detected in a single generation. This pedigree indicates that the mother had offspring at the age of one or two years.

In pedigree 4 (Figure 31), individuals BC1244 (female) and BC1780 (female) had a parent-offspring relationship. The pedigree was oriented by the age of the individuals, as both individuals were captured in 2016 and belonged to different age classes: individual BC1244 belonged to age class 2 (2-3 years old), while individual BC1780 was age class 1 (1 year old). Both mother and daughter were captured in different localities, Leira 3 (code 8) and Leira 2 (code 9), separated by 1 km distance along the river. The mother was captured twice in the same locality in June and September of 2016. This pedigree indicates that the mother had offspring at the age of one or two years.

In pedigree 5 (Figure 31), individual BC1828 (female) had a parent-offspring relationship with individuals C5519 (male) and BC1364 (male). The orientation of this pedigree was done automatically by the program and was supported by the age of the individuals, as individual BC1828 was determined to belong to age class 3 (3-5 years old), while BC1364 was age class 2 (2-3 years old); the age of C5519 could not be determined. The two descendants were found into the group of second-degree relationships, indicating that they were half siblings who shared a mother but had different fathers. The mother and one of the sons (C5519) were captured in Leira 2 (code 9) in the same trap in June 2016. The other son (BC1364) was found at 1,300 meters from her mother and half brother, in Leira 4 (code 7), indicating a probable small dispersal event of this individual from his birthplace. The mother was captured three times in the same locality, indicating that this female did not disperse during the summer of 2016. The presence of half-sibling relationships suggests that the species is not monogamous. Due to missing ages in this pedigree, it was not possible to know whether desmans mate with different partners in the same year.

In pedigree 6 (Figure 31), individual BC0016 (female) had a parent-offspring relationship with individuals BC0022 (male) and BC1139 (female). The pedigree was oriented automatically due to the existence of several descendants. Age classes could not be determined for any of the individuals. The two descendants were found in the full-sibling category and had similar inbreeding coefficients (0.28 ad 0.27, respectively), corroborating this hypothesis. An additional (not represented) individual, BC0015, a male, had a full-sibling relationship with all the integers of the pedigree, which is not possible. The most likely scenario is that this individual was a full sibling of the mother, BC0016, as they had similar inbreeding coefficients (0.29 and 0.25, respectively), whereas the other relationships were likely to be inflated due to inbreeding. All three individuals were captured in Mondera (locality code 20) in 2014. The mother and her son were captured together in Mondera in May 2014.

In pedigree 7 (Figure 31), individual BC0051 (female) had a parent-offspring relationship with individuals BC1142 (female), BC0967 (female), and BC1205 (male). The pedigree was oriented automatically due to the existence of several descendants. Age classes could not be determined for any of the individuals. The three pairwise relationships between the three descendants were determined to be full siblings. The three siblings had similar inbreeding coefficients (0.10, 0.14, and 0.13, respectively), supporting the fact that they shared the same parents. An additional (not represented) individual, BC2222, a female, had a full-sibling relationship with all the integers of the pedigree, which is not possible. This individual had an inbreeding coefficient of 0.33. The most probable scenario is that this individual was the half sibling of BC0051 and half aunt of the other three individuals. All four individuals in the pedigree were captured in Arrochas (locality code 1). The mother and the two daughters (BC1142 and BC0967) were captured together in May 2015. Siblings BC0967 and BC1205 were also captured together in June 2016. Multiple captures of the mother and daughter BC0967 in the same locality in different years support female philopatry.

In pedigree 8 (Figure 31), individual BC1014 (female) had a parent-offspring relationship with individuals BC1047 (female), BC1091 (female), BC0397 (male), and BC1097 (female). The pedigree was oriented automatically due to the existence of several descendants. Age classes could only be determined for two individuals: individual BC0397 was determined to belong to age class 0 in 2014 and BC1091 to age class 1 in 2015, supporting the fact that these putative full siblings were born in the same year (2014). Pairs of individuals BC1047/BC1091, BC1091/BC0397, BC1091/BC0397, and BC0397/BC1097 were determined to be full siblings, while BC1047/BC0397 and BC1047/BC1097 were assigned to a second-degree relationship.

Individuals BC1091, BC0397, and BC1097 probably belonged to the same litter as they have all been determined to have full-sibling relationships between them, and they also have similar inbreeding coefficients (0.10, 0.08, and 0.10, respectively). Individual BC1047, although assigned to a full-sibling relationship with individual BC1091, is likely to be a half sibling of the other 3 siblings. An additional (not represented) individual, BC1140, a male, had a second-degree relationship with individual BC0397 and full-sibling relationships with the rest of the individuals from the pedigree, which is not possible. This individual had a different mitochondrial haplotype from the rest of the individuals in the pedigree, which indicated that it cannot be full sibling to any of them. The most probable scenario is that this individual was half sibling of individual BC1014 and half uncle of the rest of individuals. Individual BC1047 had an incongruent full-sibling relationship with individual BC1091. The mother and three of the descendants (BC1047, BC1091, and BC0397) were all captured in 2015 in Cabril-upstream (locality code 15), while individual BC1097 was captured in 2014 in Cabril-Castro (code 16), situated 1.8 km downstream from Cabril-upstream, indicating dispersal of this individual from his likely birthplace. This pedigree indicates that at least two descendants (BC1091 and BC0397) were raised in the same year. The presence of half-sibling relationships suggests that the species is not monogamous. Due to missing ages, it was not possible to know whether desmans breed with different partners in the same year.

In contrast to the results obtained with VCF2LR, the pedigrees reconstructed with PRIMUS showed a higher number of inconsistencies, particularly regarding the congruence of full-sibling relationships, where the number was higher and many did not fit with each other (not shown). This once again demonstrated the overall better performance of VCF2LR for the data set, most probably due to the incorporation of inbreeding in this program.

2.1.5. Dispersal patterns

The mean distance between localities where related individuals were found increased with kinship category for both VCF2LR (Table 4, Figure 32) and PRIMUS (Table S21, Figure S15). When distances were weighted by the number of dyads in each category, the mean effective dispersal distance per generation was 1,401 and 1,364 meters for VCF2LR and PRIMUS, respectively. This distance was even smaller if only parent-offspring or full siblings were considered, but the number of these close relationships was relatively small for a robust estimation (Tables 4 and S21).

Table 4. Means and standard deviations (in parenthesis) of the distances in meters for the kinship categories estimated with the *VCF2LR* program.

Kinship category	Number of relationships	Mean distance (SD)	Distance per generation
Parent-offspring	15	822 (1,751)	822
Full siblings	27	2,348 (2,985)	1,174
2^{nd} degree	232	2,929 (2,716)	1,464
Others	423	5,665 (3,953)	-

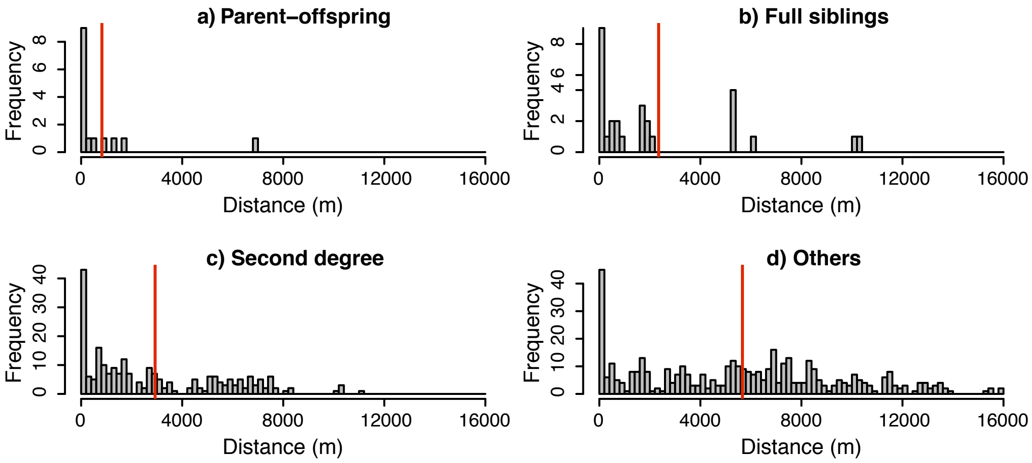

Figure 32. Frequency histograms of the distances between related individuals for each kinship category inferred with VCF2LR. Vertical red lines indicate the mean value for each category.

2.2. Discussion

2.2.1. Use of non-invasive samples as a source for the construction of NGS libraries

Obtaining kinship and dispersal information on a highly elusive and endangered species like the Pyrenean desman involved a number of methodological challenges. A first problem was imposed by the use of minimally invasive samples (hairs) to obtain genomic information: although not usually recognised, this type of samples contains a high number of exogenous sequences, and therefore they cannot be directly used with a non-selective technique like ddRAD.

Non-invasive or minimally invasive samples can be extremely useful to obtain genomic information from endangered species, but their analysis is not without problems. In this work, a hair sample was taken from most of the captured specimens of the Pyrenean desman with the purpose of causing them the least possible disturbance. However, their analysis and the comparison with the few tissue samples taken from tail tips evidenced that hair samples assembled a strikingly higher number of loci. This was probably due to different types of organisms (bacteria, parasites, etc.) present in the skin and hairs and the small proportion of tissue cells, likely to be mostly in the root.

The use of this type of samples therefore requires filtering out exogenous sequences. This can be achieved with a reference genome. In its absence, an endogenous reads catalogue can be built by using reads obtained from fresh tissue samples. After filtering reads with this catalogue, both hair and tissue samples rendered a similar number of loci and showed similar properties such as heterozygosity. Assembling loci from only non-invasive samples (or old museum samples) may give rise to artefactual results. Therefore, despite the extra bioinformatic efforts involved and the loss of many reads during the filtering process, these results validate and encourage the use of minimally invasive samples (hair, in this case) for library preparation and sequencing as long as the necessary precautions are taken.

2.2.2. Assessment of the accuracy of relatedness and inbreeding coefficients estimates from SNP data sets

Even when a large number of SNPs are available for estimating relatedness and inbreeding coefficients, simulations must be performed to assess not only the best estimator to be used, but also the accuracy of these estimates (Pew et al., 2015; Van De Casteele et al., 2001). Previous studies have demonstrated that the correlation between the triadic likelihood relatedness estimates and the pedigree relatedness coefficients increased as more loci were used (Santure et al., 2010). Here, two different simulations were performed to assess, first, which was the best estimator for the data set and, second, the accuracy of the relatedness and inbreeding coefficients in known pedigrees.

Simulations based on allele frequencies of the two data sets used in the study in La Rioja (Results section 1.1.2) and the study in Zamora (Results section 2.1.2) showed that maximum-likelihood estimators presented the best correlations with the expected values for the different kinship relationships tested (Tables S7 and S15). They also showed the lowest standard deviations, indicating that maximum-likelihood estimators provide better results for SNPs data sets (Tables S6 and S14). Regarding the number of individuals and loci used, simulations performed with the large data set (3,874 SNPs from 70 individuals from Zamora) were more accurate and had lower standard deviations than the simulations performed with the small data set (912 SNPs from 37 individuals in La Rioja), as indicated in Figures S7 and S12 and Tables S6 and S14, corroborating that a large amount of loci are needed to ensure a proper estimation of relatedness coefficients.

Simulations based on artificial pedigrees were also better in the study of Zamora than in the study of la Rioja. Relatedness coefficients in the simulations of known categories were slightly above the expected values in all cases in La Rioja (Table S10 and Figure S10), while they were closer to the expected values and had lower standard deviations for simulations in Zamora (Table S17 and Figure S13). Inbreeding coefficients were also more accurate and had lower standard deviations in the simulations performed in Zamora (Table S18 and Figure S14) than in La Rioja (Table S12 and Figure S11).

Differences in the accuracy of relatedness and inbreeding coefficients estimates in the two studies may be directly correlated with the number of markers used, but interference of the high inbreeding levels in La Rioja with the correct estimations of relatedness coefficients cannot be excluded, as inbreeding raises relatedness coefficients above theoretical values. All these results demonstrate the importance of performing simulations prior to estimations of relatedness and inbreeding coefficients in order to use the best estimator in each data set.

2.2.3. Methodological issues of pedigree reconstruction

A large number of methods to estimate the coefficient of relatedness are available and they have been tested with different types of genomic data (Ivy et al., 2016; Norman et al., 2013). However, methods to estimate kinship categories of different degrees have only been recently developed and they have been rarely used with wild animal populations (Malenfant et al., 2016; Schmidt et al., 2016).

Here, the VCF2LR (Heinrich et al., 2017) and PRIMUS (Staples et al., 2014) programs were used with samples of the Pyrenean desman to discriminate among possible kinship categories relating individuals. PRIMUS had the advantage of discriminating more types of kinship categories (up to third degree), whereas VCF2LR can take inbreeding into account, a phenomenon that undoubtedly affects the populations of the Pyrenean desman (see Results sections 1.1.5 and 2.1.2). The use of both methods with different types of simulated genotypes indicated a slightly better performance of VCF2LR, likely due to its capacity to consider inbreeding, although their application to other species and data sets should be tested through simulations, as they may have different requirements.

Another way to test the performance of these methods comes from the reconstruction of pedigrees using the assigned kinship categories. For this purpose, pedigrees corresponding to the 15 parent-offspring relationships determined by VCF2LR were reconstructed by hand (Figure 31) and the consistency of the reconstructions was then evaluated. A good level of consistency was found in all the pedigrees with the estimated age of the individuals, the sequenced mitochondrial haplotypes, and the estimated inbreeding coefficients, giving support to all the parent-offspring assignments, as well as most of the full-sibling relationships determined. However, some full-sibling relationships were inconsistent (for example, some pairs of full siblings were found not to share their mother). Simulations had shown that inbreeding rises the actual kinship category between individuals even with the VCF2LR program and therefore inconsistent full-sibling relationships are likely to be actually second-degree relationships.

Whether these limitations can be avoided with SNP data derived from complete genomes and improved methods to analyse genome-wide data (Ko & Nielsen, 2017) remains to be seen in the future. However, it is likely that the reconstruction of pedigrees will always be challenging when working with small and isolated populations, as several members of a pedigree may be related in more than one way due to inbred matings among individuals.

2.2.4. Preliminary data on the reproductive biology and social organization of the Pyrenean desman as provided by pedigrees

The obtained pedigrees were very helpful to extract fundamental information about the reproductive biology of a species that is only rarely observed in the wild. For example, they provided data on the age of first reproduction, a key parameter for modelling population viability (Ferrer et al., 2004). In fact, pedigrees 2, 3, and 4 (Figure 31) indicated that desmans reach sexual maturity when they are one or two years old. A greater degree of accuracy in this parameter was not possible due to the uncertainty in the age determination of older specimens (age class 2 and higher), but further studies along these lines may help to provide more certainty in this important parameter. The average generation time is another important parameter of interest for different calculations, but there is too little data at the moment to determine it with certainty. Given that desmans can live in their natural environment over 5 years, although it is difficult to find desmans older than 4 years-old (González-Esteban et al., 2002), the average generation time can be provisionally established at around 2 years.

Six mothers and only two fathers were found in the reconstructed pedigrees (Figure 31). Furthermore, the mothers of pedigrees 5, 6, 7, and 8 were captured together with one or two of their descendants during the same night and along the same stretch of river, and all in the months of May and June, during the likely reproduction period (Palmeirim & Hoffmann, 1983). In contrast, none of the fathers were captured together with their offspring. This suggests that the females take care of the offspring during their first months of life while males may move to other parts of the river after reproducing. The stronger philopatry found in females is consistent with these results.

Finally, the existence of half siblings in pedigrees 5 and 8 (Figure 31) suggest that the species is not monogamous. Due to missing ages, it is not possible to know if breeding with different couples can occur the same year. Pedigree 8 indicates that at least two pups were raised the same year, although surely this number can increase when additional data becomes available.

2.2.5. Overall dispersal of the species deduced from pedigrees

Another parameter of great interest that can be estimated with pedigrees is the dispersal distance per generation. The estimated average dispersal distance per generation for the Pyrenean desman was approximately 1.4 km (or even less if only the closest categories were considered for the estimation; Table 4). This low dispersal distance is supported by the high average relatedness found for individuals from the same locality. In addition, the fact that no desmans unrelated to any other were found is also an indication that long-distance dispersal events, at least towards this population, are rare.

Additionally, mean relatedness for female dyads from the same locality was higher than for males, pointing to a higher degree of female philopatry. In the previous study of the Pyrenean desman in the Iberian Range, similar relatedness means were found for female and male dyads (see Results section 1.1.4). These different results may be explained because overall relatedness within the same locality and inbreeding were much higher in the Iberian Range, probably due to more serious dispersal problems than in the area studied in Zamora. When only adults were used to avoid the effect of recent offspring in the relatedness at each locality, the pairs remaining for calculation were too low and the difference was not significant. More data from aged desmans will be necessary to settle this issue.

The strong philopatric behaviour of the Pyrenean desman does not exclude the existence of important exploratory movements in the species. In fact, it was previously shown that desmans have a home range of approximately 500 meters and a desman can move more than 2 km in a single day (Melero et al., 2014) and even 18 km after a few months (Gillet et al., 2016). In this study, the largest dispersal distance recorded in a single generation was that of a descendant found 7 km away from its mother (Figure 32). Therefore, not only the average dispersal distance per generation but also the distribution of distances and the extreme values (Figure 32) are important to understand the dispersal behaviour of the species.

2.2.6. Concluding remarks

Obtaining information about behavioural aspects of wild species is relatively easy for species that can be marked and observed in their natural habitat. However, for elusive species like the Pyrenean desman the information that can be obtained from direct observation and radio tracking is scarce. The rapid development of cost-effective tools to sequence hundreds of SNPs has allowed studying wild species at the genomic level. The results presented here illustrate the potential that robust pedigrees based on genome-wide data may have to unravel crucial information for highly elusive species regarding its reproductive biology, social organization, and dispersal events across generations.

Here, only the first steps towards the necessary methodological development to extract this information from genomic data have been taken. Further studies along these lines will allow generating enough data to unravel fundamental aspects of the biology of this and other endangered species.

3. Quantitative analysis of connectivity between populations using kinship categories and network assortativity

3.1. Results

3.1.1. Barriers to dispersal of the Pyrenean desman in the Tera and Tuela rivers of the Zamora area

The effect of river obstacles on the dispersal of the Pyrenean desman was tested for several potential barriers in the Tera and Tuela rivers of the Zamora area. Information about these barriers was obtained from the Confederación Hidrográfica del Duero. Seven dams higher than 3.5 meters were found in the area (Figure 33) and four of them had enough samples at both sides to test their effect on connectivity. From east to west, they were: the Pedro dam, two concatenated dams in Prados de la Fraga, and the Requejo dam. The Pedro dam (or dam 1 throughout this work) is the highest barrier in the area, with 8.16 meters high, and is used by a hydroelectric power station (Figure 33). It is the only studied dam that has a fish ladder. The two concatenated dams in Prados de la Fraga, in the Leira river (dam 2), are 4 and 5.7 meters high, respectively. The first one is dedicated to water supply, and the second one was used by a small hydroelectric power station. Finally, the Requejo dam (dam 3) is 3.7 meters high and is used for fish management. Only the Pedro dam creates an important water reservoir, whereas all other dams have minimal or non-existent water reservoir.

Additionally, the watershed divide between the two main rivers, the Tera and Tuela, was also included in the study. The two rivers are separated by a small stretch of land of a few hundred meters with a small elevation profile, which the desmans need to cross to go from one river basin to the other (Figure 33).

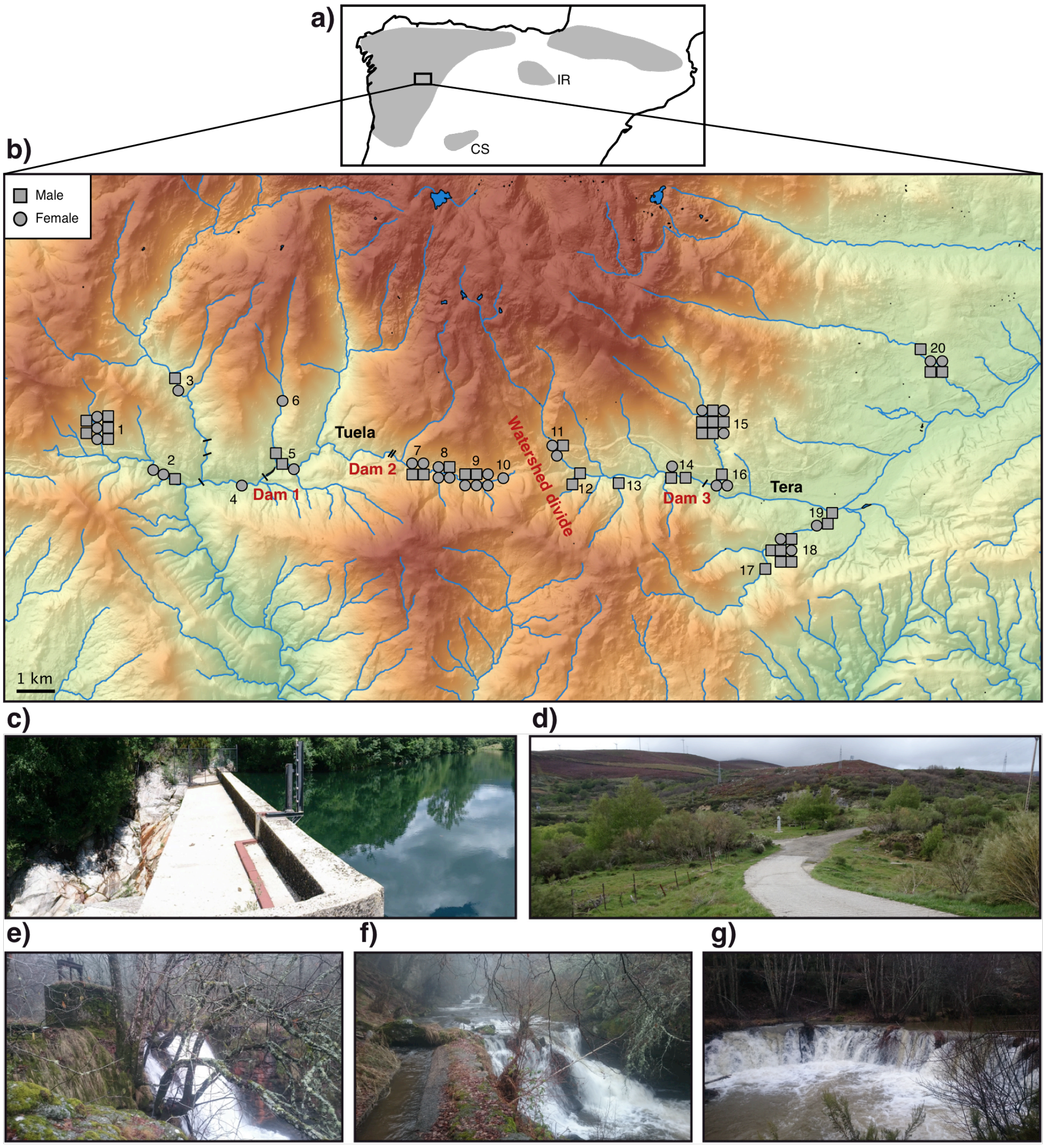

Figure 33. (a) Map of the studied area with the Pyrenean desman specimens used in the study represented with squares (males) and circles (females). Localities are indicated with numbers as in Table S4. Dams higher than 3.5 meters are represented with a bar. The barriers studied are depicted in red. The pictures of the main barriers in the area correspond to (c) Pedro dam, (d) watershed divide between the Tera and Tuela rivers, (e, f) two concatenated dams in Prados de la Fraga, and (g) Requejo dam (photo credit: Ángel Fernández-González (c, e, f, and g) and Jose Castresana (d)).

3.1.2. Construction of kinship networks

Using VCF2LR, 15 parent-offspring pairs, 27 full-sibling pairs, and 232 pairs with a second-degree relationship were found (see Results section 2.1.3). The networks for each of these kinship categories were constructed (Figure 34). Of the 15 parent-offspring pairs, 4 were found in different localities, indicating dispersal, likely of the descendant. Many of the other relationships were between individuals from different localities, but the directionality of dispersal was not possible to determine with the available data in these cases. It is evident from these networks that connections are not randomly distributed: they are abundant between some localities but very scarce between others.

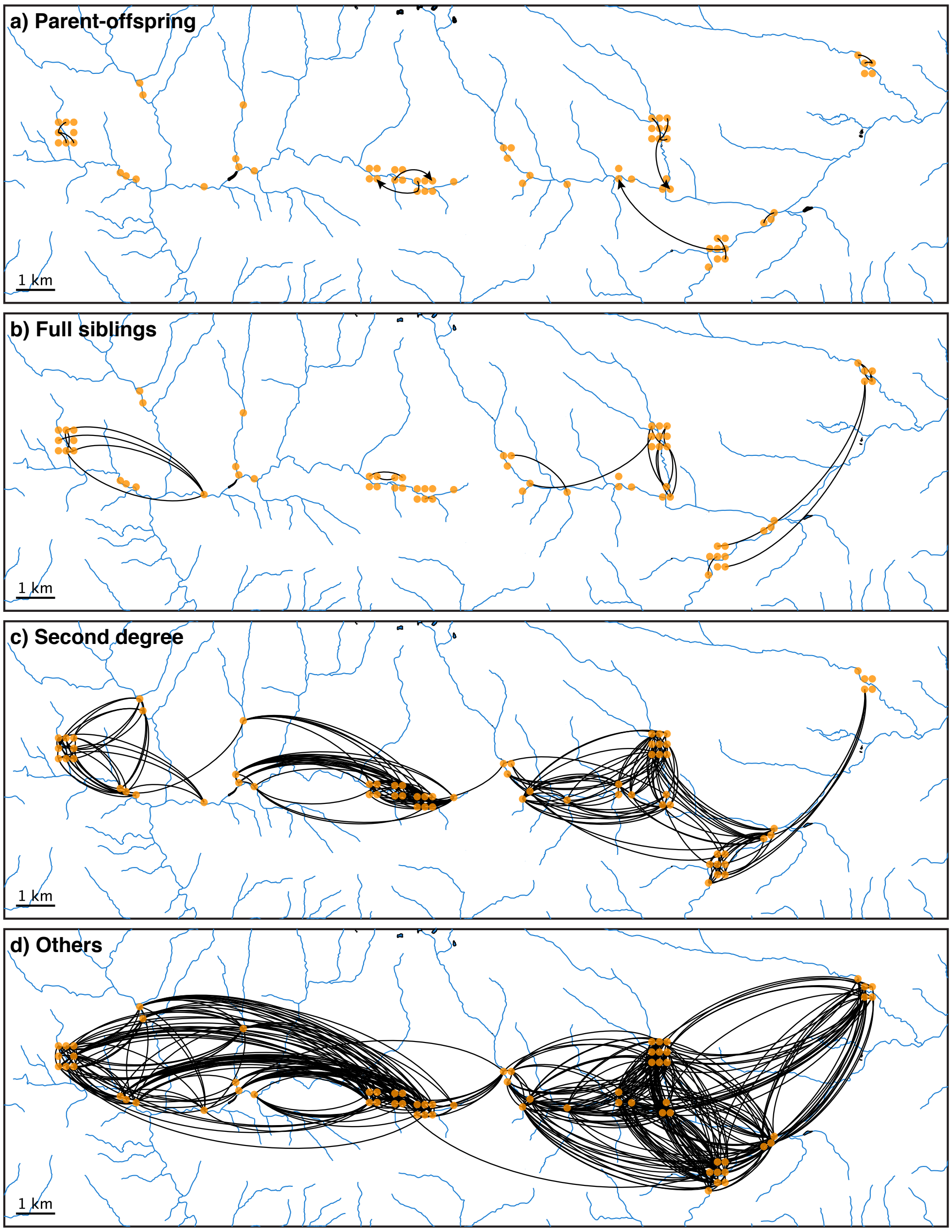

Figure 34. Maps plotting networks for each kinship category inferred between dyads of individuals as estimated with VCF2LR. Lines connecting individuals indicate a relationship between them. Map a) shows parent-offspring relationships, b) full-sibling relationships, c) second-degree relationships, and d) other relationships.

3.1.3. Quantification of network assortativity

The effect of several potential barriers in the area to dispersal of individuals was quantified with the assortativity coefficient. When the assortativity coefficients for the different barriers were estimated (Table 5), it was found that the barrier that caused the largest impact on dispersal was the watershed divide between the rivers. Of the 173 relationships expected between the two rivers if they were freely connected, only 10 were observed. This led to a highly significant assortativity coefficient of 0.94 (94% of missing edges due to the barrier). The effect of this barrier was also tested for different kinship categories. An assortativity coefficient of 0.99 was found for individuals with a second-degree relationship and 0.89 for individuals with more distant relationships, both being significant.

Table 5. Observed relationships, expected relationships and assortativity coefficients (with standard deviation in parenthesis) between adjacent sectors at both sides of specific barriers. Localities used for the estimations are as follows. Watershed divide: localities 5-10 vs. 11-16, Dam 1: 1-4 vs. 5-10; Dam 2: 5-6 vs. 7-10; Dam 3: 11-14 vs. 15-16.

Barrier	Observed relationships	Expected relationships	Assortativity coefficient (SD)	p-value
Watershed divide	10	173.1	0.94 (0.02)	0.0000 (***)
Watershed divide (2nd degree)	1	71.5	0.99 (0.01)	0.0000 (***)
Watershed divide (others)	9	84.7	0.89 (0.03)	0.0000 (***)
Dam 1	70	148.6	0.53 (0.05)	0.0000 (***)
Dam 1 (2nd degree)	1	57.6	0.98 (0.02)	0.0000 (***)
Dam 1 (others)	69	80.4	0.14 (0.08)	0.0322 (*)
Dam 2	54	48.8	-0.11 (0.07)	0.7938
Dam 2 (2nd degree)	24	24.9	0.04 (0.12)	0.3721
Dam 2 (others)	30	23.1	-0.30 (0.05)	0.9506
Dam 3	36	33.4	-0.08 (0.09)	0.6778
Dam 3 (2nd degree)	12	9.7	-0.24 (0.07)	0.7637
Dam 3 (others)	24	22.4	-0.07 (0.13)	0.6218

Regarding the effect of the dams, an important barrier effect was found for only one of them across the Tuela river (dam 1 in Figure 33), which had a highly significant assortativity coefficient of 0.53 when all relationships were considered. However, large differences were found in this case when the coefficient was calculated for different generations: it was 0.98 and highly significant for second-degree relationships, and only 0.14 and non-significant for more distant relationships. The other two barriers tested, two small concatenated dams across the Tuela (dam 2), and a small dam across the Tera (dam 3), showed non-significant assortativity coefficients, indicating that they do not affect the dispersal of individuals.

3.1.4. Population structure

STRUCTURE was used to test whether the various barriers corresponded with different genomic clusters. The analysis showed good convergence of the different runs for K values between 1 and 6. As determining only one optimum value of K is difficult (Betto-Colliard et al., 2015), several K values were represented to show the hierarchical nature of structure. Figure 35 shows the average admixture proportions with different K values. With 2 populations, the Tuela and Tera rivers became separated. With K = 3, the Tuela river was subdivided into two clusters predominantly located at each side of dam 1. Modelling 4 and 5 populations caused additional subdivisions along the Tera river. Higher K values rendered new populations that were not geographically delimited and therefore had no apparent biological significance.

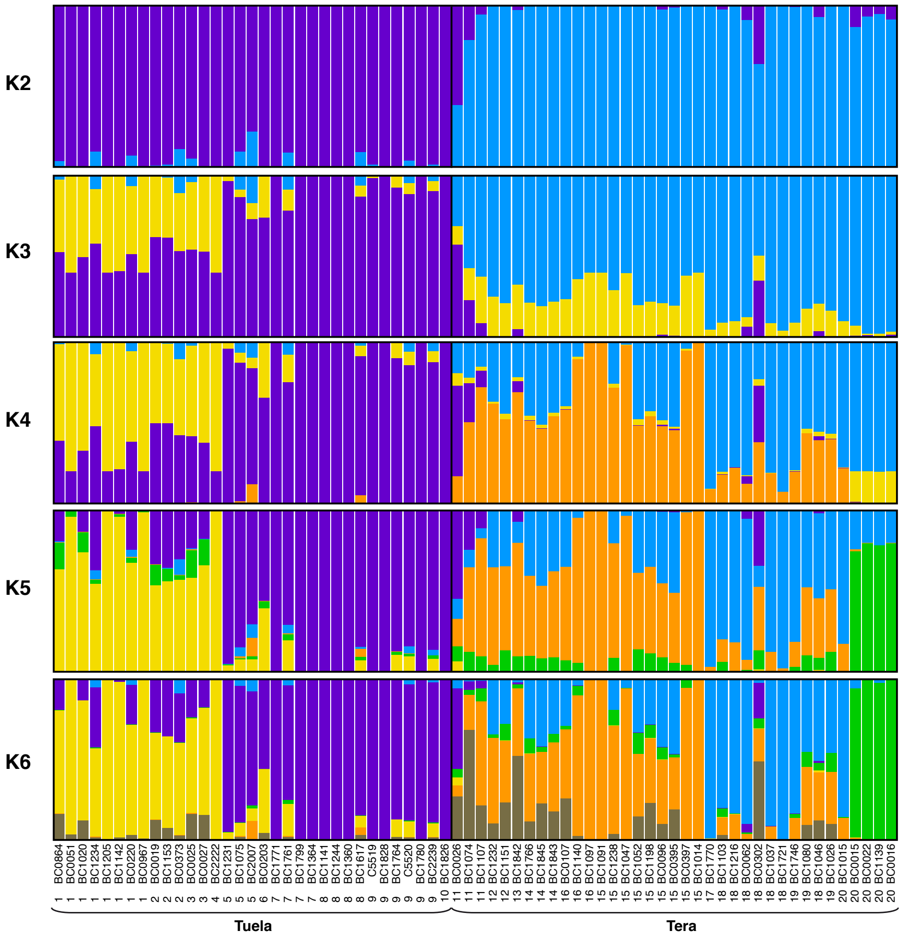

Figure 35. Bar plots of admixture proportions of each specimen as determined with STRUCTURE and different K values. Specimen code, locality code, and river are indicated for each individual.

In fact, a model of 5 populations (Figure 36) generated a geographic structure similar to that obtained with the kinship networks; the correspondence was particularly notorious with the network obtained from second-degree relationships (Figure 34c).

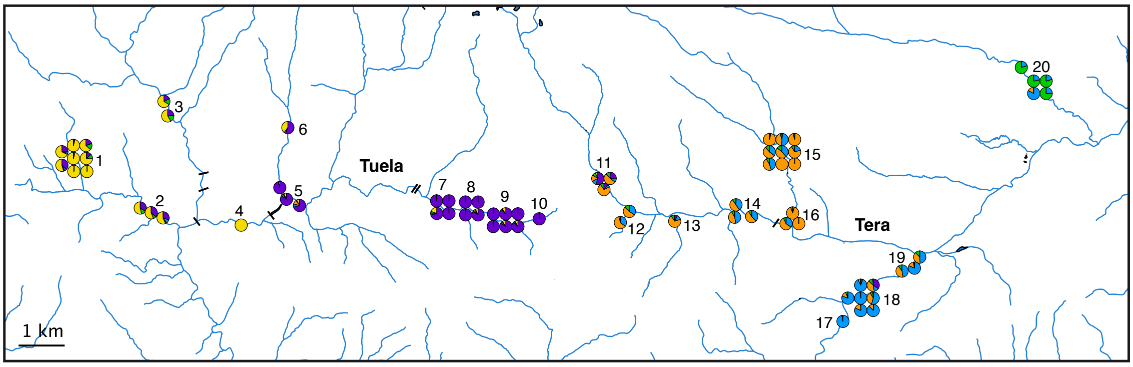

Figure 36. Map plotting admixture proportions of Pyrenean desman specimens from Zamora as determined with STRUCTURE and K = 5. Locality codes are indicated as in Table S4. Barriers are shown wit a line.

3.2. Discussion

3.2.1. Quantification of the effect of specific barriers in two rivers of Zamora: the assortativity coefficient

Both geographic and anthropogenic barriers present in the studied area are likely to affect the overall dispersal patterns of a semi-aquatic species like the Pyrenean desman. The anthropogenic barriers include dams of various sizes whose effect on the connectivity of this species is unknown. The most obvious geographic barrier in this area is the watershed divide between the two main rivers. To go from one river basin to the other, the desmans would have to cross a few hundred meters, a distance that may be reduced during rainy or snowmelt periods. How the particular features of a watershed divide affect the movements of a semi-aquatic species like the Pyrenean desman, with reduced terrestrial mobility (Palmeirim & Hoffmann, 1983), is completely unknown.

Clearly, it is important to determine how each of these barriers affects contemporary dispersal. Using genomic data from individuals captured at both sides of a barrier, it was possible to obtain the number of kinship relationships through it but, in order to quantify its effect, it is also necessary to know the number of expected relationships if no barrier existed, which can be estimated from the relationships present at each side of the barrier. Here, an estimator used in network theory, the assortativity coefficient (Newman, 2003), was adapted to perform these estimations. The resulting assortativity coefficient for the watershed divide was 0.94 (Table 5), indicating that this barrier is the most important in the area. The fact that 10 kinship relationships between desmans located on the two sides of the watershed divide were detected (Table 5) is also of high interest. Even if these relationships are only a small proportion of those expected, they demonstrate that terrestrial crossing is indeed feasible for this species. Downstream dispersal as a mechanism for explaining these relationships can be discounted due to the long distance to the point where the rivers meet.

The second most important barrier was the largest dam in the area (of 8 meters high), with an assortativity coefficient of 0.53 (Table 5). As expected, all other, smaller dams did not significantly impact the dispersal of the Pyrenean desman. Importantly, the assortativity coefficient can be broken down for different generations. When this coefficient was calculated from only second-degree relationships for the large dam, a value of 0.98 was obtained, while for other, more distant relationships this value was 0.15. This reduced barrier effect in older generations may be due to indirect crossing through small tributaries at both sides of the dam, which may happen after a few generations. Alternatively, the existence of particular conditions that caused a change in the barrier effect in the last few years cannot be ruled out. In contrast, the similarly high assortativity coefficients estimated for different generations for the watershed divide indicated a stronger and more constant effect for this barrier.

The kinship networks suggest that additional fragmentation could be present in the Tera river due to the reduced connectivity between locality 20 (Mondera, situated in the north-easternmost part) and the rest of localities, as well as between localities 17, 18, and 19 (the Parada tributary, in the south) and the rest (Figure 31). In fact, this river, and particularly the Mondera locality, shows high inbreeding coefficients, suggesting the existence of certain fragmentation (Figure 29 and Table S16). The assortativity coefficient between these sectors was not estimated because there is no clear and delimited barrier that could be established as a hypothesis for the test. Rather, the existence of these connectivity gaps may be due to diffuse ecological barriers in this part of the river (e.g. contamination, abundance of predators, etc.) that will need to be further investigated.

3.2.2. Comparison of network assortativity and STRUCTURE for the estimation of recent dispersal

The kinship networks and the assortativity coefficients estimated are highly congruent with the clusters obtained by STRUCTURE. At K = 2, the two clusters correspond with the fragmentation caused by the main barrier detected, the watershed divide, while the additional subdivision appearing at K = 3 coincides with the second most important barrier, the large dam (Figure 33). The additional groups appearing at K = 5 in the Tera river (Figure 35) are also consistent with the kinship networks, as indicated above. Furthermore, admixed specimens with components of the Tuela and Tera rivers in localities 11 and 18 (Figure 35) indicate a low but non-negligible dispersal level between these two rivers, in agreement with the small number of relationships (10) observed between them (Figure 34). This high level of congruence with STRUCTURE gives strong support to the network assortativity results.

Bayesian clustering algorithms like STRUCTURE have been used in previous studies of recent fragmentation (Blair et al., 2012; Estes-Zumpf et al., 2010; Riley et al., 2006). However, in certain populations, for example in contact zones of ancient lineages, clustering methods will reflect both ancient and recent population divergence, making it difficult to determine the genetic structure specifically due to recent barriers. The advantage of an approach based on close kinship relationships is that it ensures that the dispersal and fragmentation events inferred are recent (from the last few generations).

Given the high level of congruence between the clustering and kinship methods in this study, it seems that no additional ancient structure existed in this population of the Pyrenean desman, but this cannot be discarded for other populations or species. Additionally, the assortativity coefficient allows the quantification of the impact of each specific barrier of interest. Without doubt, the computation of the expected number of relationships through a barrier could be improved, for example, by means of simulations of desman movements that take into account the distance between localities at each side of the barrier. However, when the chosen localities are close enough to the barrier and within reach of typical exploratory movements of the species (Gillet et al., 2016; Melero et al., 2014), the assortativity coefficients estimated for the various barriers should have a relevant comparative value to detect the strongest obstacles. Further studies with this and other species will help to better understand the potential and limitations of this approach.

3.2.3. Conclusions and future potential for network assortativity in conservation

Obtaining information on current dispersal patterns and connectivity problems is a fundamental step in developing conservation plans for endangered species. In the case of aquatic species, conservation measures to improve connectivity may involve the permeability of river barriers, such as dams, through the construction of bypass facilities (e.g. fish passes, rocky ramps or artificial rivers). For semi-aquatic species, in addition to these actions, improving the habitat of terrestrial corridors (for example, by protecting an extensive vegetated area close to the watershed divides) may also be crucial to reduce mortality in these critical areas and thus increase the connectivity between basins.

Faced with a limited budget, information on the effect of specific barriers is essential for helping to prioritize measures directed at improving connectivity between populations. Using a network assortativity approach, it has been demonstrated that the greatest barrier to dispersal for the Pyrenean desman in the area studied was the watershed divide between the two sub-basins. It can therefore be concluded that actions directed at improving this natural pass would be extremely beneficial to the connectivity of the desmans living in the area. Only one of the dams in the area was shown to cause significant problems to connectivity, highlighting it as a preferential target for conservation efforts.

These studies may have important implications for the conservation of the most endangered populations of this species. Specifically, a critical area for the conservation of the Pyrenean desman is the Central System (Fernandes et al., 2008; Gisbert & Garcia-Perea, 2014). This population occupies a small area located at the southern limit of the range (see Figure 33a), and represents a distinct evolutionary unit (Igea et al., 2013; Querejeta et al., 2016). Fragmentation and reduced connectivity in this population may lead to inbreeding depression and, eventually, its extinction (Frankham et al., 2014; Kardos et al., 2016). Therefore, adequate connectivity within and between the few remaining sub-populations in this area may be crucial for their long-term survival. A network assortativity approach similar to that proposed in this study may be fundamental for detecting barriers to dispersal in this endangered population, and thus provide essential information for its conservation.

The methodological approach developed here based on kinship networks and assortativity coefficients can be applied to populations of any species. It can be extremely helpful to obtain information on dispersal of aquatic and semi-aquatic species, for which river barriers can be a serious threat (Ardren & Bernall, 2017; Gouskov et al., 2016), but also terrestrial species affected by specific barriers, such as roads (Estes-Zumpf et al., 2010; Riley et al., 2006). In all cases where the barriers are a problem for the viability of the populations, the information obtained with a network assortativity approach can help to develop conservation plans aimed at improving genetic exchange between populations, prioritize efforts, and take management decisions that can be crucial for the conservation of threatened species.

V. CONCLUSIONS

1. The contact zone between the two mithochondrial lineages of the Pyrenean desman in the Iberian Range was not as strict as previously thought, as determined with SNPs obtained from ddRAD libraries of 37 samples from La Rioja. According to the genomic tree, the principal component analysis, and the population structure analysis, the genetic variability in the area was structured by rivers instead of by mitochondrial lineages.

2. The mean relatedness coefficient was found to be very high in the area. Individuals also showed high inbreeding levels. In agreement with it, low heterozygosity levels were found in all individuals. The locality with the highest inbreeding coefficient was located upstream of a dam, suggesting that river barriers may have an important impact on recent dispersal and the genetic health of the species.

3. The relatedness and individual inbreeding coefficients estimated with this data set were robust according to bioinformatic simulations based on artificial pedigrees that included actual genotypes of the studied population as founders.

4. The relatedness networks of the Pyrenean desman in La Rioja showed a low level of contemporary inter-river dispersal compared to intra-river dispersal, indicating poor connectivity between rivers in the Iberian Range.

5. The ddRAD protocol was modified and optimized to allow processing each sample independently, which enabled the use of minimally invasive hair samples.

6. Mean relatedness and inbreeding coefficients calculated with SNPs obtained from 70 Pyrenean desmans of Zamora were much lower than those from La Rioja. Mean relatedness was higher for female dyads than for male dyads, suggesting a higher degree of female philopatry.

7. Kinship categories were determined for the Pyrenean desmans of Zamora and their reliability was assessed with bioinformatic simulations based on artificial pedigrees. Using the kinship category assignments, pedigrees were reconstructed and their congruence was evaluated by checking the age of the individuals, the mitochondrial haplotypes, and the inbreeding coefficients. Both bioinformatic simulations and assessment of the pedigrees' congruence

revealed that methods modelling inbreeding performed better than those without inbreeding.

8. Pedigree reconstruction based on SNPs has proven to be useful to unravel preliminary information on the reproductive biology and the social organisation of a highly elusive species like the Pyrenean desman. Pedigrees also allowed estimating the average dispersal distance per generation. The low dispersal distance determined was in agreement with the high average relatedness coefficient found for individuals from the same locality.

9. The assortativity coefficient was used to quantify the effect of specific barriers on the dispersal of Pyrenean desmans in the two studied rivers of Zamora, the Tera and the Tuela. The most important barrier found in the area was the watershed divide between both rivers, followed by a large dam in one of the rivers. Other smaller dams studied did not show a significant impact to the dispersal of individuals.

10. The kinship networks and the assortativity coefficients estimated for the different barriers were highly congruent with the results obtained from the population structure analysis, although only kinship networks allow inferring that the observed genetic structure has been recently generated.

11. The methodological approach developed to quantify the effect of specific barriers on dispersal of individuals can be applied to populations of any species. Information about the impact of different barriers in a certain area may help to design and prioritize conservation measures directed at improving connectivity between populations.

VI. REFERENCES

Abascal, F., Corvelo, A., Cruz, F., Villanueva-Cañas, J. L., Vlasova, A., Marcet-Houben, M., Martínez-Cruz, B., Cheng, J. Y., Prieto, P., Quesada, V., Quilez, J., Li, G., García, F., Rubio-Camarillo, M., Frias, L., Ribeca, P., Capella-Gutiérrez, S., … Godoy, J. A. (2016). Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered Iberian lynx. *Genome Biology*, 17, 251.

Abecasis, G. R., Cherny, S. S., Cookson, W. O., & Cardon, L. R. (2002). Merlin — rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, 30, 97–101.

Åkesson, M., Liberg, O., Sand, H., Wabakken, P., Bensch, S., & Flagstad, Ø. (2016). Genetic rescue in a severely inbred wolf population. *Molecular Ecology*, 25, 4745–4756.

Allendorf, F. W., Hohenlohe, P. A., & Luikart, G. (2010). Genomics and the future of conservation genetics. *Nature Reviews Genetics*, 11, 697–709.

Anderson, A. D., & Weir, B. S. (2007). A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. *Genetics*, 176, 421–440.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, 17, 81–92.

Ardren, W. R., & Bernall, S. R. (2017). Dams impact westslope cutthroat trout metapopulation structure and hybridization dynamics. *Conservation Genetics*, 18, 297–312.

Arora, N., Van Noordwijk, M. A., Ackermann, C., Willems, E. P., Nater, A., Greminger, M. P., Nietlisback, P., Dunkel, L. P., Utami Atmoko, S. S., Pamungkas, J., Perwitasari-Farajallah, D., van Shaik, C. P., & Krützen, M. (2012). Parentage-based pedigree reconstruction reveals female matrilineal clusters and male-biased dispersal in nongregarious Asian great apes, the Bornean orang-utans (*Pongo pygmaeus*). *Molecular Ecology*, 21, 3352–3362.

Ballou, J. D. (1997). Ancestral inbreeding only minimally affects inbreeding depression in mammalian populations. *Journal of Heredity*, 88, 169–178.

Banks, S. C., & Lindenmayer, D. B. (2014). Inbreeding avoidance, patch isolation and matrix permeability influence dispersal and settlement choices by male agile antechinus in a fragmented landscape. *Journal of Animal Ecology*, 83, 515–524.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*, 1–2.

Benazzo, A., Trucchi, E., Cahill, J. A., Maisano Delser, P., Mona, S., Fumagalli, M., Bunnefeld, L., Cornetti, L., Ghirotto, S., Girardi, M., Ometto, L., Panziera, A., Rota-Stabelli, O., Zanetti, E., Karamanlidis, A. A., Groff, C., Paule, L., ... Bertorelle, G. (2017). Survival and divergence in a small group: the extraordinary genomic history of the endangered Apennine brown bear stragglers. *Proceedings of the National Academy of Sciences*, 114, E9589–E9597.

Betto-Colliard, C., Sermier, R., Litvinchuk, S., Perrin, N., & Stöck, M. (2015). Origin and genome evolution of polyploid green toads in Central Asia: evidence from microsatellite markers. *Heredity*, 114, 300–308.

Biffi, M., Gillet, F., Laffaille, P., Colas, F., Aulagnier, S., Blanc, F., Galan, M., Tiouchichine, M. L., Némoz, M., Buisson, L., & Michaux, J. R. (2017). Novel insights into the diet of the Pyrenean desman (*Galemys pyrenaicus*) using next-generation sequencing molecular analyses. *Journal of Mammalogy*, 98, 1497–1507.

Blair, C., Weigel, D. E., Balazik, M., Keeley, A. T. H., Walker, F. M., Landguth, E., Cushman, S., Murphy, M., Waits, L. P., & Balkenhol, N. (2012). A simulation-based evaluation of methods for inferring linear barriers to gene flow. *Molecular Ecology Resources*, 12, 822–833.

Blouin, M. S. (2003). DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends in Ecology & Evolution*, 18, 503–511.

Blouin, M. S., Parsons, M., Lacaille, V., & Lotz, S. (1996). Use of microsatellite loci to classify individuals by relatedness. *Molecular Ecology*, 5, 393–401.

Bonin, C. A., Goebel, M. E., O'Corry-Crowe, G. M., & Burton, R. S. (2012). Twins or not? Genetic analysis of putative twins in Antarctic fur seals, *Arctocephalus gazella*, on the South Shetland Islands. *Journal of Experimental Marine Biology and Ecology*, 412, 13–19.

Bowler, D. E., & Benton, T. G. (2009). Variation in dispersal mortality and dispersal propensity among individuals: the effects of age, sex and resource availability. *Journal of Animal Ecology*, 78, 1234–1241.

Broquet, T., & Petit, E. J. (2009). Molecular estimation of dispersal for ecology and population genetics. *Annual Review of Ecology, Evolution, and Systematics*, 40, 193–216.

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I. A., Belmonte, M. K., Lander, E. S., Nusbaum, C., & Jaffe, D. B. (2008). ALLPATHS: *De novo* assembly of whole-genome shotgun microreads. *Genome Research*, 18, 810–820.

Butler, K., Field, C., Herbinger, C. M., & Smith, B. R. (2004). Accuracy, efficiency and robustness of four algorithms allowing full sibship reconstruction from DNA marker data. *Molecular Ecology*, 13, 1589–1600.

Cabral, M. J., Almeida, J., Almeida, P. R., Dellinger, T., Ferrand de Almeida, N., Oliveira, M. E., Palmeirim, J. M., Queiroz, A. I., Rogado, L., & Santos-Reis, M. (Eds.). (2005). *No Livro Vermelho dos Vertebrados de Portugal*. Lisboa: Instituto da Conservaçao da Natureza.

Castien, E., & Gosálbez, J. (1994). Diet of *Galemys pyrenaicus* (Geoffroy, 1811) in the North of the Iberian Peninsula. *Netherlands Journal of Zoology*, 45, 422–430.

Catchen, J. M., Amores, A., Hohenlohe, P. A., Cresko, W. A., & Postlethwait, J. H. (2011). Stacks: building and genotyping loci de novo from short-read sequences. *Genes|Genomes|Genetics*, 1, 171–182.

Catchen, J. M., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology*, 22, 3124–3140.

Chakraborty, R., Shaw, M., & Schull, W. J. (1974). Exclusion of paternity: the current state of the art. *American Journal of Human Genetics*, 26, 477–488.

Charbonnel, A., Buisson, L., Biffi, M., D'Amico, F., Besnard, A., Aulagnier, S., Blanc, F., Gillet, F., Lacaze, V., Michaux, J. R., Némoz, M., Pagé, C., Sanchez-Perez, J. M., Sauvage, S., & Laffaille, P. (2015). Integrating hydrological features and genetically validated occurrence data in occupancy modelling of an endemic and endangered semi-aquatic mammal, *Galemys pyrenaicus*, in a Pyrenean catchment. *Biological Conservation*, 184, 182–192.

Charlesworth, D., & Willis, J. H. (2009). The genetics of inbreeding depression. *Nature Reviews. Genetics*, 10, 783–796.

Charpentier, M., Peignot, P., Hossaert-McKey, M., Gimenez, O., Setchell, J. M., & Wickings, E. J. (2005). Constraints on control: factors influencing reproductive success in male mandrills (*Mandrillus sphinx*). *Behavioral Ecology*, 16, 614–623.

Crnokrak, P., & Barrett, S. C. H. (2012). Purging the genetic load: a review of the experimental evidence. *International Journal of Organic Evolution*, 56, 2347–2358.

Croteau, E. K., Heist, E. J., & Nielsen, C. K. (2010). Fine-scale population structure and sex-biased dispersal in bobcats (*Lynx rufus*) from southern Illinois. *Canadian Journal of Zoology*, 88, 536–545.

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter,

M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, 12, 499–510.

De Woody, J. A. (2005). Molecular approaches to the study of parentage, relatedness, and fitness: practical applications for wild animals. *Journal of Wildlife Management*, 69, 1400–1418.

Devlin, B., Roeder, K., & Ellstrand, N. C. (1988). Fractional paternity assignment: theoritical development and comparison to other methods. *Theoretical and Applied Genetics*, 76, 369–380.

Driscoll, D. A., Banks, S. C., Barton, P. S., Ikin, K., Lentini, P., Lindenmayer, D. B., Smith, A. L., Berry, L. E., Burns, E. L., Edworthy, A., Evans, M. J., Gibson, R., Heinsohn, R., Howland, B., Kay, G., Munro, N., Scheele, B. C., … Westgate, M. J. (2014). The trajectory of dispersal research in conservation biology. Systematic review. *PLoS ONE*, 9, e95053.

Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, 28, 2239–2252.

Earl, D. A., & VonHoldt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, 4, 359–361.

Ebbert, M. T. W., Wadsworth, M. E., Staley, L. A., Hoyt, K. L., Pickett, B., Miller, J., Duce, J., Kauwe, J. S. K., & Ridge, P. G. (2016). Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics*, 17, 239.

Efford, M. (2017). Secrlinear - Spatially Explicit Capture – Recapture for Linear Habitats.

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends in Ecology and Evolution*, 29, 51–63.

Escoda, L., González-Esteban, J., Gómez, A., & Castresana, J. (2017). Using relatedness networks to infer contemporary dispersal: application to the endangered mammal *Galemys pyrenaicus*. *Molecular Ecology*, 26, 3343–3357.

Estes-Zumpf, W. A., Rachlow, J. L., Waits, L. P., & Warheit, K. I. (2010). Dispersal, gene flow, and population genetic structure in the pygmy rabbit (*Brachylagus idahoensis*). *Journal of Mammalogy*, 91, 208–219.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, 14, 2611–2620.

Ewing, B., & Green, P. (1998). Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Research*, 8, 186–194.

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, 61, 717–726.

Felsenstein, J. (1989). PHYLIP - Phylogeny inference package. *Cladistics*, 5, 164–166.

Fernandes, M., Herrero, J., Aulagnier, S., & Amori, G. (2008). *Galemys pyrenaicus*. *Lista Roja de Especies Amenazadas de la UICN 2012.2*, e.T8826A12934876.

Ferrer, M., Otalora, F., & García-Ruiz, J. M. (2004). Density-dependent age of first reproduction as a buffer affecting persistence of small populations. *Ecological Applications*, 14, 616–624.

Frankham, R., Bradshaw, C. J. A., & Brook, B. W. (2014). Genetics in conservation management: revised recommendations for the 50/500 rules, Red List criteria and population viability analyses. *Biological Conservation*, 170, 56–63.

Freedman, A. H., Gronau, I., Schweizer, R. M., Ortega-Del Vecchyo, D., Han, E., Silva, P. M., Galaverni, M., Fan, Z., Marx, P., Lorente-Galdos, B., Beale, H., Ramirez, O., Hormozdiari, F., Alkan, C., Vilà, C., Squire, K., Geffen, E., … Novembre, J. (2014). Genome sequencing highlights the dynamic early history of dogs. *PLoS Genetics*, 10, e1004016.

Freeland, J. R. (2006). Molecular ecology. *Molecular Ecology*. West Sussex, England. John Wiley & Sons.

Gerber, S., Chabrier, P., & Kremer, A. (2003). FAMOZ: a software for parentage analysis using dominant, codominant and uniparentally inherited markers. *Molecular Ecology Notes*, 3, 479–481.

Gillet, F., Roux, B. Le, Blanc, F., Bodo, A., Fournier-Chambrillon, C., Fournier, P., Jacob, F., Lacaze, V., Némoz, M., Aulagnier, S., & Michaux, J. R. (2016). Genetic monitoring of the endangered Pyrenean desman (*Galemys pyrenaicus*) in the Aude River, France. *Belgian Journal of Zoology*, 146, 44–52.

Gisbert, J., & Garcia-Perea, R. (2014). Historia de la regresión del desmán ibérico *Galemys pyrenaicus* (É. Geoffroy Saint-Hilaire, 1811) en el Sistema Central (Península Ibérica). *Conservation and Management of Semi-Aquatic Mammals of*

Southwestern Europe. Munibe Monographs. Nature Series, 3, 19–35.

Gómez, A., & Lunt, D. H. (2007). Refugia within refugia: patterns of phylogeographic concordance in the Iberian Peninsula. In *Phylogeography in Southern European Refugia* (Vol. 11, pp. 155–188).

González-Esteban, J., Villate, I., & Castién, E. (2003). Sexual identification of *Galemys pyrenaicus*. *Acta Theriologica*, 48, 571–573.

González-Esteban, J., Villate, I., Castién, E., Rey, I., & Gosálbez, J. (2002). Age determination of *Galemys pyrenaicus*. *Acta Theriologica*, 47, 107–112.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17, 333–351.

Gouskov, A., Reyes, M., Wirthner-Bitterlin, L., & Vorburger, C. (2016). Fish population genetic structure shaped by hydroelectric power plants in the upper Rhine catchment. *Evolutionary Applications*, 9, 394–408.

Greenwood, P. J. (1980). Mating systems, philopatry and dispersal in birds and mammals. *Animal Behaviour*, 28, 1140–1162.

Hadfield, J. D., Richardson, D. S., & Burke, T. (2006). Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework. *Molecular Ecology*, 15, 3715–3730.

Hamede, R. K., Bashford, J., McCallum, H., & Jones, M. (2009). Contact networks in a wild Tasmanian devil (*Sarcophilus harrisii*) population: using social network analysis to reveal seasonal variability in social behaviour and its implications for transmission of devil facial tumour disease. *Ecology Letters*, 12, 1147–1157.

Hammerly, S. C., Morrow, M. E., & Johnson, J. A. (2013). A comparison of pedigree- and DNA-based measures for identifying inbreeding depression in the critically endangered Attwater's Prairie-chicken. *Molecular Ecology*, 22, 5313–5328.

Head, S. R., Kiyomi Komori, H., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R., & Ordoukhanian, P. (2014). Library construction for next-generation sequencing: overviews and challenges. *BioTechniques*, 56, 61–77.

Hedrick, P. W., & Garcia-Dorado, A. (2016). Understanding inbreeding depression, purging, and genetic rescue. *Trends in Ecology and Evolution*, 31, 940–952.

Heinrich, V., Kamphans, T., Mundlos, S., Robinson, P. N., & Krawitz, P. M. (2017). A likelihood ratio-based method to predict exact pedigrees for complex families from next-generation sequencing data. *Bioinformatics*, 33, 72–78.

Hendricks, S., Epstein, B., Schönfeld, B., Wiench, C., Hamede, R. K., Jones, M., Storfer, A., & Hohenlohe, P. A. (2017). Conservation implications of limited genetic diversity and population structure in Tasmanian devils (*Sarcophilus harrisii*). *Conservation Genetics*, 18, 977–982.

Hua, P., Zhang, L., Guo, T., Flanders, J., & Zhang, S. (2013). Dispersal, mating events and fine-scale genetic structure in the lesser flat-headed bats. *PloS ONE*, 8, e54428.

Huff, C. D., Witherspoon, D. J., Simonson, T. S., Xing, J., Watkins, W. S., Zhang, Y., Tuohy, T. M., Neklason, D. W., Burt, R. W., Guthery, S. L., Woodward, S. R., & Jorde, L. B. (2011). Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome Research*, 21, 768–774.

Huisman, J. (2017). Pedigree reconstruction from SNP data: parentage assignment, sibship clustering and beyond. *Molecular Ecology Resources*, 17, 1009–1024.

Huisman, J., Kruuk, L. E. B., Ellis, P. A., Clutton-Brock, T., & Pemberton, J. M. (2016). Inbreeding depression across the lifespan in a wild mammal population. *Proceedings of the National Academy of Sciences*, 113, 3585–3590.

Igea, J., Aymerich, P., Fernández-González, A., González-Esteban, J., Gómez, A., Alonso, R., Gosálbez, J., & Castresana, J. (2013). Phylogeography and postglacial expansion of the endangered semi-aquatic mammal *Galemys pyrenaicus*. *BMC Evolutionary Biology*, 13, 115.

International Human Genome Sequencing Consortium, Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., … Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409, 860–921.

Ivy, J. A., Putnam, A. S., Navarro, A. Y., Gurr, J., & Ryder, O. A. (2016). Applying SNP-derived molecular coancestry estimates to captive breeding programs. *Journal of Heredity*, 107, 403–412.

Jacquard, A. (1972). Genetic information given by a relative. *Biometrics*, 28, 1101–1114.

Jacquard, A. (1975). Inbreeding: one word, several meanings. *Theoretical Population Biology*, 7, 338–363.

Jakobsson, M., & Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, 23, 1801–1806.

Janes, J. K., Miller, J. M., Dupuis, J. R., Malenfant, R. M., Gorrell, J. C., Cullingham, C. I., & Andrew, R. L. (2017). The K = 2 conundrum. *Molecular Ecology*, 26, 3594–3602.

Johnson, B. B., White, T. A., Phillips, C. A., & Zamudio, K. R. (2015). Asymmetric Introgression in a Spotted Salamander Hybrid Zone. *Journal of Heredity*, 106, 608–617.

Jones, A. G. (2001). GERUD1.0: a computer program for the reconstruction of parental genotypes from progeny arrays using multilocus DNA data. *Molecular Ecology Notes*, 1, 215–218.

Jones, A. G., & Ardren, W. R. (2003). Methods of parentage analysis in natural populations. *Molecular Ecology*, 12, 2511–2523.

Jones, A. G., Small, C. M., Paczolt, K. A., & Ratterman, N. L. (2010). A practical guide to methods of parentage analysis. *Molecular Ecology Resources*, 10, 6–30.

Jones, O. R., & Wang, J. (2010). COLONY: a program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources*, 10, 551–555.

Kalinowski, S. T., Taper, M. L., & Marshall, T. C. (2007). Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Molecular Ecology*, 16, 1099–1106.

Kardos, M., Taylor, H. R., Ellegren, H., Luikart, G., & Allendorf, F. W. (2016). Genomics advances the study of inbreeding depression in the wild. *Evolutionary Applications*, 9, 1205–1218.

Kchouk, M., Gibrat, J. F., & Elloumi, M. (2017). Generations of sequencing technologies: from first to next generation. *Biology and Medicine*, 9.

Keller, L. F., & Waller, D. M. D. M. (2002). Inbreeding effects in wild populations. *Trends in Ecology and Evolution*, 17, 230–241.

Kling, D., Welander, J., Tillmar, A., Skare, Ø., Egeland, T., & Holmlund, G. (2012). DNA microarray as a tool in establishing genetic relatedness—Current status and future prospects. *Forensic Science International: Genetics*, 6, 322–329.

Ko, A., & Nielsen, R. (2017). Composite likelihood method for inferring local pedigrees. *PLOS Genetics*, 13, e1006963.

Konovalov, D. A., Manning, C., & Henshaw, M. T. (2004). KINGROUP: a program for pedigree relationship reconstruction and kin group assignments using genetic markers. *Molecular Ecology Notes*, 4, 779–782.

Lambin, X. (1994). Natal philopatry, competition for resources, and inbreeding avoidance in Townsend's voles (*Microtus townsendii*). *Ecology*, 75, 224–235.

Langergraber, K. E., Mitani, J. C., & Vigilant, L. (2007). The limited impact of kinship on cooperation in wild chimpanzees. *Proceedings of the National Academy of Sciences*, 104, 7786–7790.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9, 357–359.

Leberg, P. L., & Firmin, B. D. (2008). Role of inbreeding depression and purging in captive breeding and restoration programmes. *Molecular Ecology*, 17, 334–343.

Li, C. C., Weeks, D. E., & Chakravarti, A. (1993). Similarity of DNA Fingerprints due to chance and relatedness. *Human Heredity*, 43, 45–52.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25, 1754–1760.

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, 475, 493–496.

Li, H., Glusman, G., Huff, C., Caballero, J., & Roach, J. C. (2014). Accurate and robust prediction of genetic relationship from whole-genome sequences. *PLoS ONE*, 9, e85437.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Subgroup, 1000 Genome Project Data Processing. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25, 2078–2079.

Linnarsson, S. (2010). Recent advances in DNA sequencing methods - general principles of sample preparation. *Experimental Cell Research*, 316, 1339–1343.

Liu, G., Shafer, A. B. A., Zimmermann, W., Hu, D., Wang, W., Chu, H., Cao, J., & Zhao, C. (2014). Evaluating the reintroduction project of Przewalski's horse in China using genetic and pedigree data. *Biological Conservation*, 171, 288–298.

Locke, D. P., Hillier, L. W., Warren, W. C., Worley, K. C., Nazareth, L. V, Muzny, D. M., Yang, S. P., Wang, Z., Chinwalla, A. T., Minx, P., Mitreva, M., Cook, L., Delehaunty, K. D., Fronick, C., Schmidt, H., Fulton, L. A., Fulton, R. S., … Wilson, R. K. (2011). Comparative and demographic analysis of orang-utan genomes. *Nature*, 469, 529–533.

Lopes, M. S., Silva, F. F., Harlizius, B., Duijvesteijn, N., Lopes, P. S., Guimarães, S. E., & Knol, E. F. (2013). Improved estimation of inbreeding and kinship in pigs using optimized SNP panels. *BMC Genetics*, 14, 92.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., … Wang, J. (2012). SOAPdenovo2: an empirically improved memory efficient short-read *de novo* assembler. *GigaScience*, 1, 18.

Lynch, M., & Ritland, K. (1999). Estimation of pairwise relatedness with molecular markers. *Genetics*, 152, 1753–1766.

Malenfant, R. M., Coltman, D. W., Richardson, E. S., Lunn, N. J., Stirling, I., Adamowicz, E., & Davis, C. S. (2016). Evidence of adoption, monozygotic twinning, and low inbreeding rates in a large genetic pedigree of polar bears. *Polar Biology*, 39, 1455–1465.

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., Howard, E., Shendure, J., & Turner, D. J. (2010). Target-enrichment strategies for next-generation sequencing. *Nature Methods*, 7, 111–118.

Manel, S., Gaggiotti, O. E., & Waples, R. S. (2005). Assignment methods: matching biological questions with appropriate techniques. *Trends in Ecology & Evolution*, 20, 136–142.

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26, 2867–2873.

Mardis, E. R. (2011). A decade’s perspective on DNA sequencing technology. *Nature*, 470, 198–203.

Matthysen, E. (2005). Density-dependent dispersal in birds and mammals. *Ecography*, 28, 403–416.

McCormack, J. E., Faircloth, B. C., Crawford, N. G., Gowaty, P. A., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Research*, 22, 746–754.

McCormack, J. E., Hird, S. M., Zellmer, A. J., Carstens, B. C., & Brumfield, R. T. (2013). Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, 66, 526–538.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation

DNA sequencing data. *Genome Research*, 20, 1297–1303.

Meagher, T. R., & Thompson, E. (1986). The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction. *Theoretical Population Biology*, 29, 87–106.

Melero, Y., Aymerich, P., Luque-Larena, J. J., & Gosálbez, J. (2012). New insights into social and space use behaviour of the endangered Pyrenean desman (*Galemys pyrenaicus*). *European Journal of Wildlife Research*, 58, 185–193.

Melero, Y., Aymerich, P., Santulli, G., & Gosálbez, J. (2014). Activity and space patterns of Pyrenean desman (*Galemys pyrenaicus*) suggest non-aggressive and non-territorial behaviour. *European Journal of Wildlife Research*, 60, 707–715.

Melero, Y., Oliver, M. K., & Lambin, X. (2017). Relationship type affects the reliability of dispersal distance estimated using pedigree inferences in partially sampled populations: a case study involving invasive American mink in Scotland. *Molecular Ecology*, 26, 4059–4071.

Miller, W., Schuster, S. C., Welch, A. J., Ratan, A., Bedoya-Reina, O. C., Zhao, F., Kim, H. L., Burhans, R. C., Drautz, D. I., Wittekindt, N. E., Tomsho, L. P., Ibarra-Laclette, E., Herrera-Estrella, L., Peacock, E., Farley, S., Sage, G. K., Rode, K., … Lindqvist, C. (2012). Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. *Proceedings of the National Academy of Sciences*, 109, E2382–E2390.

Milligan, B. G. (2003). Maximum-likelihood estimation of relatedness. *Genetics*, 163, 1153–1167.

Mills, L. S. (2013). *Conservation of wildlife populations: Demography, genetics, and management*. Oxford: Wiley-Blackwell.

Morey, M., Fernández-Marmiesse, A., Castiñeiras, D., Fraga, J. M., Couce, M. L., & Cocho, J. A. (2013). A glimpse into past, present, and future DNA sequencing. *Molecular Genetics and Metabolism*, 110, 3–24.

Newby, J., Darden, T., Bassos-Hull, K., & Shedlock, A. M. (2014). Kin structure and social organization in the spotted eagle ray, *Aetobatus narinari*, off coastal Sarasota, FL. *Environmental Biology of Fishes*, 97, 1057–1065.

Newman, M. E. J. (2003). Mixing patterns in networks. *Physical Review E*, 67, 26126.

Niedzicka, M., Fijarczyk, A., Dudek, K., Stuglik, M., & Babik, W. (2016). Molecular Inversion Probes for targeted resequencing in non-model organisms. *Scientific Reports*, 6, 24051.

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12, 443–451.

Nores, C. (2007). *Galemys pyrenaicus. Atlas y Libro Rojo de los Mamíferos Terrestres de España*, 92–98.

Nores, C., Ojeda, F., Ruano, A., Villate, I., Gonzalez, J., Cano, J. M., & Garcia, E. (1998). Estimating the population density of *Galemys pyrenaicus* in four Spanish rivers. *Journal of Zoology*, 246, 454–457.

Norman, A. J., & Spong, G. (2015). Single nucleotide polymorphism-based dispersal estimates using noninvasive sampling. *Ecology and Evolution*, 5, 3056–3065.

Norman, A. J., Street, N. R., & Spong, G. (2013). *De novo* SNP discovery in the Scandinavian brown bear (*Ursus arctos*). *PLoS ONE*, 8, e81012.

Økland, J.-M., Haaland, Ø. A., & Skaug, H. J. (2010). A method for defining management units based on genetically determined close relatives. *ICES Journal of Marine Science*, 67, 551–558.

Oliver, M. K., Piertney, S. B., Zalewski, A., Melero, Y., & Lambin, X. (2016). The compensatory potential of increased immigration following intensive American mink population control is diluted by male-biased dispersal. *Biological Invasions*, 18, 3047–3061.

Ouborg, N. J., Pertoldi, C., Loeschcke, V., Bijlsma, R. K., & Hedrick, P. W. (2010). Conservation genetics in transition to conservation genomics. *Trends in Genetics*, 26, 177–187.

Palmeirim, J. M., & Hoffmann, R. S. (1983). *Galemys pyrenaicus*. *Mammalian Species*, 1–5.

Palsbøll, P. J., Peery, M. Z., & Bérubé, M. (2010). Detecting populations in the “ambiguous” zone: kinship-based estimation of population structure at low genetic divergence. *Molecular Ecology Resources*, 10, 797–805.

Pemberton, J. M. (2008). Wild pedigrees: the way forward. *Proceedings of the Royal Society B: Biological Sciences*, 275, 613–621.

Perrin, N., & Mazalov, V. (1984). Dispersal and inbreeding avoidance. *The American Naturalist*, 154, 676–678.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double Digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS ONE*, 7, e37135.

Pew, J., Muir, P. H., Wang, J., & Frasier, T. R. (2015). related: an R package for analysing pairwise relatedness from codominant molecular markers. *Molecular Ecology Resources*, 15, 557–561.

Peyre, A. (1955). Intersexualité du tractus génital femelle du Desman des Pyrénées (*Galemys pyrenaicus* G.). *Bulletin de La Société Zoologique de France*, 80, 132–138.

Peyre, A. (1957). Dimorphisme sexuel de la ceinture pelvienne d'un Mammifcre Insectivore, *Galemys pyrenaicus* G. *Compte Rendus de l'Académie Des Sciences de Paris*, 244, 118–120.

Peyre, A. (1962). Recherches sur l'intersexualité spécifique chez *Galemys pyrenaicus*, G. (Mammifcre Insectivore). *Archives de Biologie*, 73, 1–174.

Pinho, C., & Hey, J. (2010). Divergence with gene flow: models and data. *Annual Review of Ecology, Evolution, and Systematics*, 41, 215–230.

Poissant, J., Davis, C. S., Malenfant, R. M., Hogg, J. T., & Coltman, D. W. (2012). QTL mapping for sexually dimorphic fitness-related traits in wild bighorn sheep. *Heredity*, 108, 256–263.

Pompanon, F., Bonin, A., Bellemain, E., & Taberlet, P. (2005). Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics*, 6, 847–846.

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., Veeramah, K. R., Woerner, A. E., O'Connor, T. D., Santpere, G., Cagan, A., Theunert, C., Casals, F., Laayouni, H., Munch, K., Hobolth, A., Halager, A. E., … Marques-Bonet, T. (2013). Great ape genetic diversity and population history. *Nature*, 499, 471–475.

Pritchard, J. K., Stephens, M., & Donelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155, 945–959.

Pusey, A. E. (1987). Sex-biased dispersal and inbreeding avoidance in birds and mammals. *Trends in Ecology and Evolution*, 2, 295–299.

Qiu, Q., Zhang, G., Ma, T., Qian, W., Wang, J., Ye, Z., Cao, C., Hu, Q., Kim, J., Larkin, D. M., Auvil, L., Capitanu, B., Ma, J., Lewin, H. A., Qian, X., Lang, Y., Zhou, R., … Liu, J. (2012). The yak genome and adaptation to life at high altitude. *Nature Genetics*, 44, 946–949.

Queller, D. C., & Goodnight, K. F. (1989). Estimating relatedness using genetic markers. *Evolution*, 43, 258–275.

Querejeta, M., Fernández-González, A., Romero, R., & Castresana, J. (2017).

Postglacial dispersal patterns and mitochondrial genetic structure of the Pyrenean desman (*Galemys pyrenaicus*) in the northwestern region of the Iberian Peninsula. *Ecology and Evolution*, 7, 4486–4495.

Querejeta, M., González-Esteban, J., Gómez, A., Fernández-González, A., Aymerich, P., Gosálbez, J., Escoda, L., Igea, J., & Castresana, J. (2016). Genomic diversity and geographical structure of the Pyrenean desman. *Conservation Genetics*, 17, 1333–1344.

Reuter, J. A., Spacek, D. V., & Snyder, M. P. (2015). High-throughput sequencing technologies. *Molecular Cell*, 58, 586–597.

Richard, P. B. (1976). Determination de l'age et de la longevité chez les desman de Pyrenées (*Galemys pyrenaicus*). *Extrait de La Terre et Id Vie, Revull d'Ecologie Appliquée*, 30, 181–192.

Richard, P. B. (1986). *Le Desman des Pyrénées. Un mammifčre inconnu à découvrir. Editions Le Rocher, Monaco.*

Riester, M., Stadler, P. F., & Klemm, K. (2009). FRANz: reconstruction of wild multi-generation pedigrees. *Bioinformatics*, 25, 2134–2139.

Riley, S. P. D., Pollinger, J. P., Sauvajot, R. M., York, E. C., Bromley, C., Fuller, T. K., & Wayne, R. K. (2006). A southern California freeway is a physical and social barrier to gene flow in carnivores. *Molecular Ecology*, 15, 1733–1741.

Rioux-Paquette, E., Festa-Bianchet, M., & Coltman, D. W. (2010). No inbreeding avoidance in an isolated population of bighorn sheep. *Animal Behaviour*, 80, 865–871.

Ritland, K. (1996). Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetical Research*, 67, 175–185.

Robinson, J. A., Ortega-Del Vecchyo, D., Fan, Z., Kim, B. Y., VonHoldt, B. M., Marsden, C. D., Lohmueller, K. E., & Wayne, R. K. (2016). Genomic flatlining in the endangered Island Fox. *Current Biology*, 26, 1183–1189.

Rousset, F. (2008). GENEPOP'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*, 8, 103–106.

Russello, M. A., & Amato, G. (2004). *Ex situ* population management in the absence of pedigree information. *Molecular Ecology*, 13, 2829–2840.

Santure, A. W., Stapley, J., Ball, A. D., Birkhead, T. R., Burke, T., & Slate, J. (2010). On the use of large marker panels to estimate inbreeding and relatedness: empirical and simulation studies of a pedigreed zebra finch population typed at

771 SNPs. *Molecular Ecology*, 19, 1439–1451.

Schmidt, K., Davoli, F., Kowalczyk, R., & Randi, E. (2016). Does kinship affect spatial organization in a small and isolated population of a solitary felid: The Eurasian lynx? *Integrative Zoology*, 11, 334–349.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, 26, 1135–1145.

Shizuka, D., & Farine, D. R. (2016). Measuring the robustness of network community structure using assortativity. *Animal Behaviour*, 112, 237–246.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., & Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Research*, 19, 1117–1123.

Skare, Ø., Sheehan, N. A., & Egeland, T. (2009). Identification of distant family relationships. *Bioinformatics*, 25, 2376–2382.

Slatkin, M. (1985). Gene flow in natural populations. *Annual Review of Ecology and Systematics*, 16, 393–430.

Soulé, M. E. (1987). *Viable populations for conservation*. (M. E. Soulé, Ed.). Cambridge: Cambridge University Press.

Speed, D., & Balding, D. J. (2015). Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics*, 16, 33–44.

Städele, V., & Vigilant, L. (2016). Strategies for determining kinship in wild populations using genetic data. *Ecology and Evolution*, 6, 6107–6120.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30, 1312–1313.

Staples, J., Qiao, D., Cho, M. H., Silverman, E. K., Nickerson, D. A., & Below, J. E. (2014). PRIMUS: rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. *The American Journal of Human Genetics*, 95, 553–564.

Steiner, C. C., Putnam, A. S., Hoeck, P. E. A., & Ryder, O. A. (2013). Conservation genomics of threatened animal species. *Annual Review of Animal Biosciences*, 1, 261–281.

Stenglein, J. L., Waits, L. P., Ausband, D. E., Zager, P., & Mack, C. M. (2011). Estimating grey wolf pack size and family relationships using noninvasive genetic sampling at rendezvous sites. *Journal of Mammalogy*, 92, 784–795.

Stone, D. R. (1985). Home range movements of the Pyrenean Desman (*Galemys pyrenaicus*) (Insectivora: Talpidae). *Z. Angew. Zool.*, 72, 25–37.

Stone, D. R., & Gorman, M. L. (1985). Social organization of the European mole (*Talpa europaea*) and the Pyrenean desman (*Galemys pyrenaicus*). *Mammal Review*, 15, 35–42.

Szulkin, M., Stopher, K. V., Pemberton, J. M., & Reid, J. M. (2013). Inbreeding avoidance, tolerance, or preference in animals? *Trends in Ecology & Evolution*, 28, 205–211.

Taylor, H. R. (2015). The use and abuse of genetic marker-based estimates of relatedness and inbreeding. *Ecology and Evolution*, 5, 3140–3150.

Tewhey, R., Warner, J. B., Nakano, M., Libby, B., Medkova, M., David, P. H., Kotsopoulos, S. K., Samuels, M. L., Hutchison, J. B., Larson, J. W., Topol, E. J., Weiner, M. P., Harismendy, O., Olson, J., Link, D. R., & Frazer, K. A. (2009). Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nature Biotechnology*, 27, 1025–1031.

Thornton, T., Tang, H., Hoffmann, T. J., Ochs-Balcom, H. M., Caan, B. J., & Risch, N. (2012). Estimating kinship in admixed populations. *The American Journal of Human Genetics*, 91, 122–138.

Valière, N. (2002). GIMLET: a computer program for analysing genetic individual identification data. *Molecular Ecology Notes*, 2, 377–379.

Van De Casteele, T., Galbusera, P., & Matthysen, E. (2001). A comparison of microsatellite-based pairwise relatedness estimators. *Molecular Ecology*, 10, 1539–1549.

van Dijk, E. L., Jaszczyszyn, Y., & Thermes, C. (2014). Library preparation methods for next-generation sequencing: tone down the bias. *Experimental Cell Research*, 322, 12–20.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., … Zhu, X. (2001). The sequence of the human genome. *Science*, 291, 1304–1351.

Vilà, C., Sundqvist, A.-K., Flagstad, Ø., Seddon, J., Rnerfeldt, S. B., Kojola, I., Casulli, A., Sand, H., Wabakken, P., & Ellegren, H. (2003). Rescue of a severely bottlenecked wolf (*Canis lupus*) population by a single immigrant. *Proceedings of the Royal Society B: Biological Sciences*, 270, 91–97.

Wang, J. (2002). An estimator for pairwise relatedness using molecular markers. *Genetics*, 160, 1203–1215.

Wang, J. (2007). Triadic IBD coefficients and applications to estimating pairwise relatedness. *Genetical Research*, 89, 135–153.

Wang, J. (2011). COANCESTRY: a program for simulating, estimating and analysing relatedness and inbreeding coefficients. *Molecular Ecology Resources*, 11, 141–145.

Wang, J. (2014). Estimation of migration rates from marker-based parentage analysis. *Molecular Ecology*, 23, 3191–3213.

Wang, J. (2016). Individual identification from genetic marker data: developments and accuracy comparisons of methods. *Molecular Ecology Resources*, 16, 163–175.

Watts, H. E., Scribner, K. T., Garcia, H. A., & Holekamp, K. E. (2011). Genetic diversity and structure in two spotted hyena populations reflects social organization and male dispersal. *Journal of Zoology*, 285, 281–291.

Watts, P. C., Rousset, F., Saccheri, I. J., Leblois, R., Kemp, S. J., & Thompson, D. J. (2007). Compatible genetic and ecological estimates of dispersal rates in insect (*Coenagrion mercuriale*: Odonata: Zygoptera) populations: analysis of "neighbourhood size" using a more precise estimator. *Molecular Ecology*, 16, 737–751.

Weir, B. S., Anderson, A. D., & Hepler, A. B. (2006). Genetic relatedness analysis: modern data and new challenges. *Nature Reviews Genetics*, 7, 771–780.

Wey, T., Blumstein, D. T., Shen, W., & Jordán, F. (2008). Social network analysis of animal behaviour: a promising tool for the study of sociality. *Animal Behaviour*, 75, 333–344.

Wilson, G. A., & Rannala, B. (2003). Bayesian inference of recent migration rates using multilocus genotypes. *Genetics*, 163, 1177–1191.

Zhao, S., Zheng, P., Dong, S., Zhan, X., Wu, Q., Guo, X., Hu, Y., He, W., Zhang, S., Fan, W., Zhu, L., Li, D., Zhang, X., Chen, Q., Zhang, H., Zhang, Z., Jin, X., … Wei, F. (2013). Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. *Nature Genetics*, 45, 67–71.

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, 28, 3326–3328.

VII. ANNEX

Table S1. Specimens used for library construction and genotyping in the study carried out in La Rioja. Specimens used in previous studies are indicated.

Specimen code	Lineage	Sex	River	Locality	Locality code	Lat.	Long.	Ref.
IBE-C3733	B1	Male	Oja	Oja Cabecera	1	42.2	-3.1	
IBE-C3734	B1	Female	Oja	Oja Cabecera	1	42.2	-3.1	
IBE-C3735	B1	Female	Oja	Oja Cabecera	1	42.2	-3.1	
IBE-C3736	B1	Female	Oja	Oja Cabecera	1	42.2	-3.1	
IBE-C3737	B1	Female	Oja	Oja Cabecera	1	42.2	-3.1	
IBE-C3738	B1	Female	Oja	Oja Azarrulla	2	42.3	-3.0	
IBE-C3739	B1	Female	Oja	Ciloria	3	42.3	-3.1	(2)
IBE-C3740	B1	Male	Oja	Ciloria	3	42.3	-3.1	
IBE-C3741	B1	Female	Najerilla	Tobía	4	42.3	-2.9	
IBE-C3742	B1	Male	Najerilla	Tobía	4	42.2	-2.9	
IBE-C3743	B1	Male	Najerilla	Tobía	4	42.3	-2.9	
IBE-C3744	B1	Male	Najerilla	Tobía	4	42.2	-2.9	
IBE-C3745	B1	Female	Najerilla	Cárdenas	5	42.3	-2.9	
IBE-C3746	B1	Male	Najerilla	Roñas	6	42.2	-2.8	
IBE-C3747	B1	Female	Najerilla	Roñas	6	42.2	-2.8	
IBE-C3748	B1	Male	Najerilla	Roñas	6	42.2	-2.8	
IBE-C3749	B1	Female	Najerilla	Roñas	6	42.2	-2.8	
IBE-C3750	A2	Male	Iregua	Iregua	7	42.1	-2.7	(2)
IBE-C3751	A2	Male	Iregua	Iregua-Cabecera	8	42.1	-2.7	
IBE-C3752	A2	Male	Iregua	La Vieja	9	42.1	-2.6	
IBE-C3753	A2	Male	Iregua	La Vieja	9	42.0	-2.6	
IBE-C3754	A2	Female	Iregua	La Vieja	9	42.0	-2.6	(2)
IBE-C3755	A2	Female	Iregua	La Vieja	9	42.1	-2.6	
IBE-C3756	B1	Male	Umbria	La Soledad	10	42.2	-3.1	(1, 2)
IBE-C3757	B2	Female	Umbria	La Soledad	10	42.2	-3.1	
IBE-C3758	A2	Female	Umbria	La Soledad	10	42.2	-3.1	
IBE-C3765	B1	Female	Oja	Urdanta	11	42.3	-3.0	
IBE-C3766	B1	Female	Oja	Urdanta	11	42.3	-3.0	
IBE-C3767	B1	Male	Oja	Urdanta	11	42.3	-3.0	
IBE-C3768	B1	Female	Oja	Urdanta	11	42.3	-3.0	
IBE-C3769	B1	Male	Oja	Urdanta	11	42.3	-3.0	
IBE-C3770	B1	Female	Oja	Urdanta	11	42.3	-3.0	
IBE-C3771	B1	Male	Najerilla	Ormázal	12	42.1	-2.8	
IBE-C3772	A2	Male	Iregua	Mayor	13	42.1	-2.7	
IBE-C3773	A2	Male	Iregua	Mayor	13	42.1	-2.7	
IBE-C3774	A2	Female	Iregua	Iregua Achichuelo	14	42.1	-2.7	(2)
IBE-C3775	A2	Female	Iregua	Iregua Achichuelo	14	42.1	-2.7	

(1) Igea, J., Aymerich, P., Fernández-González, A., González-Esteban, J., Gómez, A., Alonso, R., Gosálbez, J., & Castresana, J. (2013). Phylogeography and postglacial expansion of the endangered semi-aquatic mammal *Galemys pyrenaicus*. *BMC Evolutionary Biology*, 13, 115.

(2) Querejeta, M., González-Esteban, J., Gómez, A., Fernández-González, A., Aymerich, P., Gosálbez, J., Escoda, L., Igea, J., & Castresana, J. (2016). Genomic diversity and geographical structure of the Pyrenean desman. *Conservation Genetics*, 17, 1333–1344.

Table S2. Specimens used for the cytochrome *b* phylogeny in the study carried out in La Rioja. Specimens used in a previous study are indicated.

Specimen code	Sample type	Lineage	River	Locality	Locality code	Lat.	Long.	Ref.	Length (bp)
IBE-C437	Tissue	A2	Tera	Barriomartín	15	42	-2.5	(1)	1140
IBE-C1053	Faeces	B1	Najerilla	Roñas	6	42.2	-2.8	(1)	1140
IBE-C1068	Faeces	A2	Najerilla	Portilla-Collado Grande	16	42.1	-2.9	(1)	1140
IBE-C1069	Tissue	B1	Najerilla	Calamantío	17	42.2	-2.9	(1)	1140
IBE-C1070	Faeces	B1	Oja	Oja Azarrulla	2	42.3	-3	(1)	1140
IBE-C1072	Faeces	B1	Oja	Oja Cabecera	1	42.2	-3.1	(1)	1140
IBE-C1075	Faeces	B1	Oja	Oja Altuzarra	18	42.2	-3	(1)	1140
IBE-C1077	Faeces	B1	Oja	Oja Azarrulla	2	42.3	-3	(1)	1140
IBE-C1132	Faeces	A2	Tera	Razoncillo	19	42	-2.6	(1)	1140
IBE-C1661	Faeces	A2	Tera	Razón-Sotillo del Rincón	20	41.9	-2.7	(1)	1140
IBE-C1671	Faeces	A2	Tera	Razón-Sotillo del Rincón	20	41.9	-2.7	(1)	1140
IBE-C1687	Faeces	A2	Tera	Razoncillo	19	42	-2.6	(1)	1140
IBE-C2596	Faeces	B1	Najerilla	Valvanera	21	42.2	-2.9	(1)	1140
IBE-C3426	Faeces	B1	Najerilla	Gatón Cabecera	22	42.2	-3		1140
IBE-C3428	Faeces	B1	Najerilla	Gatón	23	42.2	-3		1103
IBE-C3455	Faeces	A2	Najerilla	Pedroso	24	42.3	-2.7	(1)	1140
IBE-C3473	Faeces	A2	Iregua	Arroyo de Puente Ra	25	42.1	-2.7	(1)	1140
IBE-C3474	Faeces	B1	Najerilla	Tobía	4	42.2	-2.9	(1)	1140
IBE-C3475	Faeces	B1	Najerilla	Tobía-Las Minas	26	42.3	-2.9	(1)	1140
IBE-C3476	Faeces	A2	Iregua	Villoslada de Cameros	27	42	-2.7		1140
IBE-C3481	Faeces	B1	Najerilla	Tobía	4	42.3	-2.9	(1)	1140
IBE-C3482	Faeces	A2	Iregua	Lumbreras	28	42.1	-2.6	(1)	1140
IBE-C3486	Faeces	A2	Iregua	La Vieja	9	42	-2.6	(1)	1140
IBE-C3489	Faeces	A2	Iregua	La Vieja	9	42	-2.6	(1)	1140
IBE-C4486	Tissue	B1	Najerilla	Barranco de la Sabandija	29	42.1	-3		1140
VM96-21-D	Faeces	B1	Umbría	La Soledad	10	42.2	-3.1		724
WM04-02bis-C	Faeces	A2	Duero	Duero Cabecera	30	42	-2.9		724
WM24-11-Cg	Faeces	A2	Tera	Razón	31	42	-2.7		724
WM24-SO09-C	Faeces	A2	Tera	Razón Cabecera	32	42	-2.7		724

(1) Igea, J., Aymerich, P., Fernández-González, A., González-Esteban, J., Gómez, A., Alonso, R., Gosálbez, J., & Castresana, J. (2013). Phylogeography and postglacial expansion of the endangered semi-aquatic mammal *Galemys pyrenaicus*. *BMC Evolutionary Biology*, 13, 115.

Table S3. Data sets analysed in the study carried out in La Rioja and STACKS filters used to generate them. Specific data used in each data set is shown in bold.

	Data set 1	Data set 2	Data set 3

Parameters in *populations* of Stacks	r = 1
	m = 9
	MAF = 0
		r = 1
		m = 12
		MAF = 0.05
			r = 1
			m = 12
			MAF = 0

Total loci	10,000 loci	7,233 loci	**7,583 loci**
Variable loci / SNPs	**1,651 SNPs**	**912 SNPs**	**1,262 loci**
Analysis	- Selection of the best estimator of relatedness	- Relatedness and inbreeding estimations - Pedigree simulations - PCA - STRUCTURE	- Heterozygosity using total loci (1,099,535 bp) - Genomic tree using variable loci (182,990 bp)

Table S4. Specimens used in the study carried out in Zamora with estimated age, sex, and locality data. Age class 0 corresponds to the first year of life of individuals, age class 1 to 1 year of life, age class 2 to 2-3 years of life, age class 3 to 3-5 years of life, and age class 4 to 3-6 years of life.

Specimen code	Age class	Sample type	Sex	River	Locality	Locality code	Lat.	Long.
IBE-BC0051	2	Tissue	Female	Tuela	Arrochas	1	42	-7
IBE-BC0220	-	Tissue	Male	Tuela	Arrochas	1	42	-7
IBE-BC0864	0	Hair	Male	Tuela	Arrochas	1	42	-7
IBE-BC0967	-	Hair	Female	Tuela	Arrochas	1	42	-7
IBE-BC1020	0	Hair	Male	Tuela	Arrochas	1	42	-7
IBE-BC1142	-	Hair	Female	Tuela	Arrochas	1	42	-7
IBE-BC1205	-	Hair	Male	Tuela	Arrochas	1	42	-7
IBE-BC1234	2	Hair	Male	Tuela	Arrochas	1	42	-7
IBE-BC0019	-	Tissue	Female	Tuela	Tuiza	2	42	-6.9
IBE-BC0373	-	Tissue	Male	Tuela	Tuiza	2	42	-6.9
IBE-BC1153	0	Hair	Female	Tuela	Tuiza	2	42	-6.9
IBE-BC0025	3	Tissue	Female	Tuela	Tuela - upstream	3	42.1	-6.9
IBE-BC0027	-	Tissue	Male	Tuela	Tuela - upstream	3	42.1	-6.9
IBE-BC2222	2	Hair	Female	Tuela	Pedro	4	42	-6.9
IBE-BC1075	2	Tissue	Male	Tuela	Porto - Pedro	5	42	-6.9
IBE-BC1231	3	Hair	Male	Tuela	Porto - Pedro	5	42	-6.9
IBE-BC2007	4	Hair	Female	Tuela	Porto - Pedro	5	42	-6.9
IBE-BC0203	1	Hair	Female	Tuela	Porto - upstream	6	42.1	-6.9
IBE-BC1364	2	Hair	Male	Tuela	Leira 4	7	42	-6.9
IBE-BC1761	2	Hair	Male	Tuela	Leira 4	7	42	-6.9
IBE-BC1771	1	Hair	Female	Tuela	Leira 4	7	42	-6.9
IBE-BC1799	1	Hair	Female	Tuela	Leira 4	7	42	-6.9
IBE-BC1141	0	Hair	Female	Tuela	Leira 3	8	42	-6.8
IBE-BC1244	2	Hair	Female	Tuela	Leira 3	8	42	-6.8
IBE-BC1360	1	Hair	Male	Tuela	Leira 3	8	42	-6.8
IBE-BC1617	0	Hair	Female	Tuela	Leira 3	8	42	-6.8
IBE-BC1764	-	Hair	Male	Tuela	Leira 2	9	42	-6.8
IBE-BC1780	1	Hair	Female	Tuela	Leira 2	9	42	-6.8
IBE-BC1828	3	Hair	Female	Tuela	Leira 2	9	42	-7.4
IBE-BC2239	1	Hair	Female	Tuela	Leira 2	9	42	-6.8
IBE-C5519	-	Hair	Male	Tuela	Leira 2	9	42	-6.8
IBE-C5520	-	Hair	Female	Tuela	Leira 2	9	42	-6.8
IBE-BC1826	3	Hair	Female	Tuela	Leira 1	10	42	-6.9
IBE-BC0026	-	Tissue	Female	Tera	Los Tornos	11	42	-6.8
IBE-BC1074	-	Hair	Male	Tera	Los Tornos	11	42	-6.8
IBE-BC1107	1	Hair	Female	Tera	Los Tornos	11	42	-6.8
IBE-BC1151	1	Hair	Male	Tera	Tabolazas - Padornelo	12	42	-6.8
IBE-BC1232	1	Hair	Male	Tera	Tabolazas - Padornelo	12	42	-6.8
IBE-BC1842	3	Hair	Male	Tera	Tejedelo	13	42	-6.8
IBE-BC1766	-	Hair	Female	Tera	Requejo	14	42	-6.8
IBE-BC1843	0	Hair	Male	Tera	Requejo	14	42	-6.8
IBE-BC1845	2	Hair	Male	Tera	Requejo	14	42	-6.8
IBE-BC0096	1	Tissue	Male	Tera	Cabril - upstream	15	42	-6.8
IBE-BC0395	0	Tissue	Male	Tera	Cabril - upstream	15	42	-6.8
IBE-BC0397	0	Tissue	Male	Tera	Cabril - upstream	15	42	-6.8
IBE-BC1014	-	Hair	Female	Tera	Cabril - upstream	15	42	-6.8
IBE-BC1047	4	Hair	Female	Tera	Cabril - upstream	15	42	-6.8

IBE-BC1052	-	Hair	Male	Tera	Cabril - upstream	15	42	-6.8
IBE-BC1091	1	Hair	Female	Tera	Cabril - upstream	15	42	-6.8
IBE-BC1198	-	Hair	Male	Tera	Cabril - upstream	15	42	-6.8
IBE-BC1238	2	Hair	Male	Tera	Cabril - upstream	15	42	-6.8
IBE-BC0107	-	Tissue	Female	Tera	Cabril - Castro	16	42	-6.8
IBE-BC1097	-	Hair	Female	Tera	Cabril - Castro	16	42	-6.8
IBE-BC1140	0	Hair	Male	Tera	Cabril - Castro	16	42	-6.8
IBE-BC1770	2	Hair	Male	Tera	Parada - upstream	17	42	-6.7
IBE-BC0062	2	Tissue	Female	Tera	Parada	18	42	-6.7
IBE-BC0302	1	Tissue	Male	Tera	Parada	18	42	-6.7
IBE-BC1037	4	Hair	Male	Tera	Parada	18	42	-6.7
IBE-BC1046	2	Hair	Female	Tera	Parada	18	42	-6.7
IBE-BC1103	-	Tissue	Male	Tera	Parada	18	42	-6.7
IBE-BC1216	2	Hair	Male	Tera	Parada	18	42	-6.7
IBE-BC1721	2	Hair	Male	Tera	Parada	18	42	-6.7
IBE-BC1026	2	Tissue	Male	Tera	Parada - Castro	19	42	-6.7
IBE-BC1080	2	Hair	Female	Tera	Parada - Castro	19	42	-6.7
IBE-BC1746	1	Hair	Male	Tera	Parada - Castro	19	42	-6.7
IBE-BC0015	2	Hair	Male	Tera	Mondera	20	42.1	-6.7
IBE-BC0016	-	Tissue	Female	Tera	Mondera	20	42.1	-6.7
IBE-BC0022	0	Hair	Male	Tera	Mondera	20	42.1	-6.7
IBE-BC1015	-	Hair	Male	Tera	Mondera	20	42.1	-6.7
IBE-BC1139	-	Tissue	Female	Tera	Mondera	20	42.1	-6.7

Table S5. Basic statistics of the library sequences in the study carried out in La Rioja.

Specimen code	Total reads	Assembled reads	Assembled loci	Coverage
IBE-C3733	4,333,991	3,973,714	107,331	37.0
IBE-C3734	3,783,292	3,675,420	74,974	49.0
IBE-C3735	3,014,159	2,634,010	154,249	17.1
IBE-C3736	2,180,584	2,090,752	70,514	29.7
IBE-C3737	7,959,967	7,699,603	98,163	78.4
IBE-C3738	1,827,852	1,695,900	76,348	22.2
IBE-C3739	3,630,939	3,500,316	70,864	49.4
IBE-C3740	3,470,264	3,293,976	88,750	37.1
IBE-C3741	3,863,617	3,441,808	100,352	34.3
IBE-C3742	3,546,597	3,291,956	98,119	33.6
IBE-C3743	3,102,519	3,007,803	63,291	47.5
IBE-C3744	2,795,163	2,697,833	75,072	35.9
IBE-C3745	1,476,004	1,382,195	65,083	21.2
IBE-C3746	2,169,694	1,890,478	92,336	20.5
IBE-C3747	5,873,897	5,728,112	76,450	74.9
IBE-C3748	1,957,770	1,798,282	69,060	26.0
IBE-C3749	3,253,756	3,125,561	81,651	38.3
IBE-C3750	1,537,869	1,469,637	58,976	24.9
IBE-C3751	11,062,283	10,044,199	139,704	71.9
IBE-C3752	5,820,592	5,652,972	83,807	67.5
IBE-C3753	6,002,916	5,539,604	132,844	41.7
IBE-C3754	4,529,979	3,984,260	194,353	20.5
IBE-C3755	8,055,321	7,332,011	122,128	60.0
IBE-C3756	3,529,250	3,261,733	117,722	27.7
IBE-C3757	7,338,254	6,642,968	107,299	61.9
IBE-C3758	2,507,142	2,285,289	85,298	26.8
IBE-C3765	2,316,150	2,196,899	78,253	28.1
IBE-C3766	11,704,257	10,963,851	134,552	81.5
IBE-C3767	5,423,159	4,911,871	190,552	25.8
IBE-C3768	12,215,261	11,305,323	166,374	68.0
IBE-C3769	4,351,626	3,979,281	137,233	29.0
IBE-C3770	1,500,452	1,422,875	63,641	22.4
IBE-C3771	10,280,980	9,613,765	142,874	67.3
IBE-C3772	5,666,969	5,032,580	102,953	48.9
IBE-C3773	2,960,732	2,625,408	197,762	13.3
IBE-C3774	11,869,363	11,191,581	126,270	88.6
IBE-C3775	6,795,664	6,104,940	138,753	44.0
Average	4,965,089	4,607,804	107,674	42.5
Total	183,708,284	170,488,766	3,983,955	

Table S6. Means and standard deviations of the simulations results performed with the program RELATED for each relatedness estimator and relationship category in the study carried out in La Rioja. Expected values of relatedness are given in parenthesis for each relationship category.

	Parent-offspring	**Full siblings**	**Half siblings**	**Unrelated**
	(0.5)	(0.5)	(0.25)	(0)
Estimator	**Mean (SD)**	**Mean (SD)**	**Mean (SD)**	**Mean (SD)**
lynchli	0.4995 (0.0200)	0.5001 (0.0278)	0.2465 (0.0347)	-0.0047 (0.0426)
lynchrd	0.4981 (0.0383)	0.4955 (0.0399)	0.2444 (0.0371)	-0.0018 (0.0250)
quellergt	0.4992 (0.0236)	0.5003 (0.0275)	0.2462 (0.0323)	-0.0047 (0.0364)
ritland	0.5000 (0.0596)	0.4899 (0.0587)	0.2434 (0.0429)	-0.0017 (0.0252)
wang	0.4996 (0.0199)	0.5000 (0.0280)	0.2465 (0.0351)	-0.0044 (0.0432)
dyadml	0.5067 (0.0105)	0.4993 (0.0262)	0.2478 (0.0287)	0.0127 (0.0160)
trioml	0.5066 (0.0100)	0.4975 (0.0263)	0.2473 (0.0287)	0.0120 (0.0160)

Table S7. Correlation coefficients between the observed and expected values of the simulation results performed with the program RELATED for each relatedness estimator in the study carried out in La Rioja.

Estimator	**Correlation**
lynchli	0.988
lynchrd	0.986
quellergt	0.990
ritland	0.973
wang	0.988
dyadml	0.994
trioml	0.994

Table S8. Individual inbreeding coefficients and heterozygosity rates of individuals from the study carried out in La Rioja.

Specimen code	River	Locality	Inbreeding coefficient	Heterozygosity rate
IBE-C3733	Oja	Oja Cabecera	0.4105	0.000185
IBE-C3734	Oja	Oja Cabecera	0.3415	0.000196
IBE-C3735	Oja	Oja Cabecera	0.1635	0.000247
IBE-C3736	Oja	Oja Cabecera	0.1774	0.000256
IBE-C3737	Oja	Oja Cabecera	0.0921	0.000276
IBE-C3738	Oja	Azarrulla	0.0807	0.000296
IBE-C3739	Oja	Ciloria	0.0722	0.000280
IBE-C3740	Oja	Ciloria	0.0362	0.000291
IBE-C3741	Najerilla	Tobía	0.4562	0.000182
IBE-C3742	Najerilla	Tobía	0.6142	0.000131
IBE-C3743	Najerilla	Tobía	0.2833	0.000220
IBE-C3744	Najerilla	Tobía	0.3495	0.000205
IBE-C3745	Najerilla	Cárdenas	0.3604	0.000184
IBE-C3746	Najerilla	Roñas	0.2043	0.000257
IBE-C3747	Najerilla	Roñas	0.4669	0.000154
IBE-C3748	Najerilla	Roñas	0.5051	0.000145
IBE-C3749	Najerilla	Roñas	0.2719	0.000214
IBE-C3750	Iregua	Iregua	0.2967	0.000229
IBE-C3751	Iregua	Iregua Cabecera	0.3658	0.000196
IBE-C3752	Iregua	La Vieja	0.3822	0.000194
IBE-C3753	Iregua	La Vieja	0.6909	0.000103
IBE-C3754	Iregua	La Vieja	0.5472	0.000141
IBE-C3755	Iregua	La Vieja	0.4816	0.000162
IBE-C3756	Umbría	La Soledad	0.2141	0.000322
IBE-C3757	Umbría	La Soledad	0.3104	0.000256
IBE-C3758	Umbría	La Soledad	0.3664	0.000260
IBE-C3765	Oja	Urdanta	0.1844	0.000257
IBE-C3766	Oja	Urdanta	0.2009	0.000226
IBE-C3767	Oja	Urdanta	0.2503	0.000229
IBE-C3768	Oja	Urdanta	0.5432	0.000132
IBE-C3769	Oja	Urdanta	0.2383	0.000246
IBE-C3770	Oja	Urdanta	0.3901	0.000176
IBE-C3771	Najerilla	Ormazal	0.3223	0.000202
IBE-C3772	Iregua	Mayor	0.3406	0.000207
IBE-C3773	Iregua	Mayor	0.3983	0.000184
IBE-C3774	Iregua	Iregua Achichuelo	0.4021	0.000196
IBE-C3775	Iregua	Iregua Achichuelo	0.2767	0.000244

Table S9. Relatedness estimated with simulated pedigrees using different reference populations in the study carried out in La Rioja. Means and standard deviations (in parenthesis) are given.

Relationship	Theoretical value	Observed Value (SD)			
		Whole set	**Oja**	**Najerilla**	**Iregua**
Parent-offspring	0.5	0.6823 (0.1106)	0.5169 (0.1093)	0.5337 (0.1089)	0.5406 (0.1138)
Full siblings	0.5	0.6935 (0.0625)	0.4602 (0.0831)	0.4763 (0.0419)	0.4291 (0.0551)
Half siblings	0.25	0.4233 (0.0638)	0.0111 (0.0212)	0.0345 (0.0403)	0.0160 (0.0221)
Grandparent-grandchild	0.25	0.4302 (0.1143)	0.0460 (0.0525)	0.0689 (0.0755)	0.0705 (0.0910)
Uncle-nephew	0.25	0.4204 (0.0837)	0.0097 (0.0214)	0.0055 (0.0133)	0.0117 (0.0247)
Half uncle-half nephew	0.125	0.2928 (0.0721)	0.0002 (0.0016)	0.0001 (0.0010)	0.0000 (0.0000)
First cousins	0.125	0.3006 (0.0948)	0.0002 (0.0017)	0.0000 (0.0001)	0.0002 (0.0024)
Half-first cousins	0.0625	0.2425 (0.0816)	0.0000 (0.0001)	0.0000 (0.0000)	0.0000 (0.0000)

Table S10. Relatedness values estimated along simulated pedigrees with migrants in the study carried out in La Rioja. Means and standard deviations (in parenthesis) are given.

Relationship	**Theoretical value**	**Observed Value (SD)**
Parent-offspring	0.5	0.6189 (0.0998)
Full siblings	0.5	0.6227 (0.0558)
Half siblings	0.25	0.3035 (0.0804)
Grandparent-grandchild	0.25	0.2919 (0.1576)
Uncle-nephew	0.25	0.3101 (0.0811)
Half uncle-half nephew	0.125	0.1544 (0.0658)
First cousins	0.125	0.2029 (0.0767)
Half-first cousins	0.0625	0.1403 (0.0649)

Table S11. Average inter-river relationships per pedigree detected along simulated pedigrees with migrants in the study carried out in La Rioja.

River	**Source of the migrant**	**Average number of relationships detected**
Oja	Najerilla	7.7
Najerilla	Oja	1.6
Najerilla	Iregua	10.0
Iregua	Najerilla	9.8

Table S12. Individual inbreeding coefficients estimated with simulated pedigrees of offspring (in parenthesis) from different types of parental relationships in the study carried out in La Rioja.

Parental relationship (offspring)	**Theoretical value**	**Observed Value (SD)**
None (F101 and F102)	0	0.0001 (0.0004)
Full siblings (F203a)	0.25	0.2491 (0.0465)
Half siblings (F203b)	0.125	0.1039 (0.0492)
First cousins (F301a)	0.0625	0.0768 (0.0470)
Half-first cousins (F301b)	0.03125	0.0465 (0.0392)

Table S13. Basic statistics of the library sequences in the study carried out in Zamora before and after filtering for exogenous sequences.

	UNFILTERED			FILTERED			
Specimen code	**Total reads**	**Assembled loci**	**Coverage**	**Total reads**	**Assembled loci**	**Coverage**	**Endogenous DNA (%)**
IBE-BC0015	5,251,017	127,692	39.6	4,013,771	41,532	95.7	76.4
IBE-BC0016	5,812,258	79,574	69.2	4,702,632	46,008	100.0	80.9
IBE-BC0019	3,560,679	66,038	50.7	2,993,214	44,357	66.0	84.1
IBE-BC0022	4,667,116	75,788	60.0	3,878,203	43,355	88.9	83.1
IBE-BC0025	8,502,682	83,323	96.3	6,987,215	47,101	144.9	82.2
IBE-BC0026	4,810,691	66,780	68.0	4,092,996	44,550	89.7	85.1
IBE-BC0027	7,007,382	87,421	74.4	5,793,666	45,554	123.1	82.7
IBE-BC0051	2,453,350	55,438	41.4	2,084,169	40,111	50.6	85.0
IBE-BC0062	3,010,333	58,191	48.3	2,596,514	44,784	56.2	86.3
IBE-BC0096	5,231,051	82,805	58.4	4,036,229	46,162	85.4	77.2
IBE-BC0107	1,937,321	62,770	28.7	1,565,500	44,501	34.3	80.8
IBE-BC0203	6,760,819	93,169	67.7	5,501,953	46,194	115.1	81.4
IBE-BC0220	5,532,278	86,373	59.7	4,173,241	46,654	87.4	75.4
IBE-BC0302	6,781,989	78,548	81.4	5,570,055	46,610	116.8	82.1
IBE-BC0373	1,669,592	45,869	33.9	1,401,970	35,606	38.3	84.0
IBE-BC0395	6,694,642	84,584	74.2	5,043,160	46,139	106.6	75.3
IBE-BC0397	7,140,160	87,668	75.6	5,306,432	44,907	115.0	74.3
IBE-BC0864	5,169,138	81,684	61.2	4,173,202	44,355	93.3	80.7
IBE-BC0967	5,644,058	98,598	55.5	4,585,431	46,648	96.3	81.2
IBE-BC1014	2,910,306	96,126	28.4	2,069,628	41,858	49.1	71.1
IBE-BC1015	3,975,799	331,505	39.9	5,142,106	46,244	108.9	36.8
IBE-BC1020	8,474,364	164,940	47.1	6,396,292	46,501	131.8	75.5
IBE-BC1026	4,509,449	64,325	65.8	3,823,049	41,428	90.0	84.8
IBE-BC1037	3,834,566	182,889	70.3	0,497,667	48,732	206.6	75.9
IBE-BC1046	4,941,001	100,173	47.4	3,881,025	43,556	87.6	78.5
IBE-BC1047	0,784,217	108,058	93.6	8,209,219	46,182	170.4	76.1
IBE-BC1052	4,771,017	98,249	44.4	3,578,808	43,870	79.8	75.0
IBE-BC1074	4,976,547	103,498	44.8	3,739,345	43,393	83.4	75.1
IBE-BC1075	3,588,336	61,548	54.8	3,056,838	42,833	69.8	85.2
IBE-BC1080	3,120,907	87,302	33.5	2,395,869	41,144	57.4	76.8
IBE-BC1091	3,920,991	85,922	43.7	2,830,887	43,893	64.0	72.2
IBE-BC1097	6,859,438	138,362	46.2	4,965,011	43,957	109.4	72.4
IBE-BC1103	0,286,097	93,524	103.1	8,659,301	48,489	173.7	84.2
IBE-BC1107	3,102,842	76,950	39.5	2,568,102	42,296	60.0	82.8
IBE-BC1139	3,004,895	67,382	41.0	2,407,331	44,632	52.5	80.1
IBE-BC1140	2,495,330	70,507	34.1	1,977,245	41,076	47.8	79.2
IBE-BC1141	2,340,530	72,838	28.8	1,733,045	42,461	39.7	74.0
IBE-BC1142	1,314,698	63,180	19.4	1,024,383	38,990	25.9	77.9
IBE-BC1151	1,864,138	58,578	29.3	1,463,846	40,854	35.0	78.5
IBE-BC1153	4,943,757	97,367	47.1	3,835,615	44,563	83.6	77.6
IBE-BC1198	2,873,945	85,749	30.2	2,068,330	40,121	50.3	72.0
IBE-BC1205	1,584,681	67,501	21.2	1,141,738	36,182	30.6	72.0
IBE-BC1216	6,712,984	138,291	45.4	4,928,359	44,733	106.5	73.4
IBE-BC1231	4,223,351	74,178	52.9	3,376,430	43,125	76.3	79.9
IBE-BC1232	2,140,167	65,696	29.1	1,613,804	42,260	37.1	75.4
IBE-BC1234	7,336,657	116,810	58.8	5,527,081	46,176	116.6	75.3
IBE-BC1238	3,277,040	75,747	41.8	2,520,760	37,555	66.5	76.9
IBE-BC1244	3,918,320	161,242	22.4	2,248,590	42,673	51.3	57.4
IBE-BC1360	9,224,260	106,535	80.2	7,059,555	47,959	142.8	76.5
IBE-BC1364	4,924,900	109,819	40.2	3,514,794	43,270	78.5	71.4
IBE-BC1617	7,715,358	168,616	42.7	5,613,834	45,877	119.1	72.8

IBE-BC1721	4,961,492	122,203	35.9	3,039,278	41,894	70.7	61.3
IBE-BC1746	5,745,987	105,338	45.8	3,677,109	44,755	80.1	64.0
IBE-BC1761	3,241,155	105,192	28.5	2,137,830	42,050	49.7	66.0
IBE-BC1764	5,153,348	104,975	43.9	3,634,285	45,520	77.9	70.5
IBE-BC1766	1,665,356	58,064	26.3	1,265,211	39,023	31.5	76.0
IBE-BC1770	3,705,363	99,668	33.2	2,417,893	40,959	57.1	65.3
IBE-BC1771	2,791,881	64,095	40.7	2,232,572	42,414	51.4	80.0
IBE-BC1780	1,080,982	55,424	16.7	767,605	36,096	20.5	71.0
IBE-BC1799	1,616,831	60,414	24.3	1,229,883	39,395	30.3	76.1
IBE-BC1826	7,700,003	106,407	68.2	5,988,582	46,881	124.5	77.8
IBE-BC1828	8,581,212	123,830	63.4	6,292,587	46,995	129.5	73.3
IBE-BC1842	2,783,017	67,421	37.7	2,140,188	41,795	49.9	76.9
IBE-BC1843	1,448,369	56,978	23.0	1,116,664	40,213	27.0	77.1
IBE-BC1845	1,799,730	59,257	27.4	1,365,082	39,406	33.7	75.8
IBE-BC2007	5,142,408	86,162	55.4	3,891,646	45,549	83.2	75.7
IBE-BC2222	4,172,296	90,296	42.5	3,103,249	44,808	67.7	74.4
IBE-BC2239	2,855,240	132,381	19.5	1,710,038	41,068	40.7	59.9
IBE-C5519	1,766,269	56,073	29.0	1,359,161	37,143	35.6	77.0
IBE-C5520	2,235,580	64,945	31.3	1,663,937	39,314	41.2	74.4
Average	4,800,485	93,069	47.2	3,591,432	43,247	78.7	75.8

Table S14. Means and standard deviations of the simulation results performed with the program RELATED for each relatedness estimator and relationship category in the study carried out in Zamora. Expected values of relatedness are given in parenthesis for each relationship category.

	Parent-Offspring	**Full siblings**	**Half siblings**	**Unrelated**
	(0.5)	(0.5)	(0.25)	(0)
Estimator	**Mean (SD)**	**Mean (SD)**	**Mean (SD)**	**Mean (SD)**
lynchli	0.5010 (0.0140)	0.4998 (0.0180)	0.2487 (0.0238)	-0.0038 (0.0272)
lynchrd	0.4992 (0.0280)	0.4981 (0.0262)	0.2453 (0.0284)	-0.0023 (0.0150)
quellergt	0.5010 (0.0163)	0.4991 (0.0179)	0.2491 (0.0225)	-0.0042 (0.0236)
ritland	0.4987 (0.0449)	0.4974 (0.0423)	0.2449 (0.0330)	-0.0023 (0.0151)
wang	0.5010 (0.0137)	0.4998 (0.0184)	0.2487 (0.0240)	-0.0037 (0.0275)
dyadml	0.5051 (0.0076)	0.4995 (0.0167)	0.2503 (0.0203)	0.0064 (0.0089)
trioml	0.5051 (0.0074)	0.4986 (0.0167)	0.2501 (0.0203)	0.0063 (0.0088)

Table S15. Correlation coefficients between the observed and expected values of the simulation results performed with the program RELATED for each relatedness estimator in the study carried out in Zamora.

Estimator	**Correlation**
lynchli	0.9948
lynchrd	0.9929
quellergt	0.9953
ritland	0.9855
wang	0.9947
dyadml	0.9975
trioml	0.9975

Table S16. Individual inbreeding coefficients of the specimens in the study carried out in Zamora.

Specimen code	River	Locality	Locality code	Inbreeding coefficient
IBE-BC0051	Tuela	Arrochas	1	0.0289
IBE-BC0220	Tuela	Arrochas	1	0.0459
IBE-BC0864	Tuela	Arrochas	1	0.1451
IBE-BC0967	Tuela	Arrochas	1	0.1360
IBE-BC1020	Tuela	Arrochas	1	0.1618
IBE-BC1142	Tuela	Arrochas	1	0.1021
IBE-BC1205	Tuela	Arrochas	1	0.1326
IBE-BC1234	Tuela	Arrochas	1	0.0650
IBE-BC0019	Tuela	Tuiza	2	0.0865
IBE-BC0373	Tuela	Tuiza	2	0.0704
IBE-BC1153	Tuela	Tuiza	2	0.0030
IBE-BC0025	Tuela	Tuela cabecera	3	0.0833
IBE-BC0027	Tuela	Tuela cabecera	3	0.2575
IBE-BC2222	Tuela	Pedro	4	0.3267
IBE-BC1075	Tuela	Porto - Pedro	5	0.0817
IBE-BC1231	Tuela	Porto - Pedro	5	0.1039
IBE-BC2007	Tuela	Porto - Pedro	5	0.0001
IBE-BC0203	Tuela	Porto cabecera	6	0.2754
IBE-BC1364	Tuela	Leira 4	7	0.0868
IBE-BC1761	Tuela	Leira 4	7	0.0386
IBE-BC1771	Tuela	Leira 4	7	0.0038
IBE-BC1799	Tuela	Leira 4	7	0.0451
IBE-BC1141	Tuela	Leira 3	8	0.0580
IBE-BC1244	Tuela	Leira 3	8	0.0016
IBE-BC1360	Tuela	Leira 3	8	0.0135
IBE-BC1617	Tuela	Leira 3	8	0.0295
IBE-BC1764	Tuela	Leira 2	9	0.0489
IBE-BC1780	Tuela	Leira 2	9	0.1008
IBE-BC1828	Tuela	Leira 2	9	0.0399
IBE-BC2239	Tuela	Leira 2	9	0.0072
IBE-C5519	Tuela	Leira 2	9	0.0996
IBE-C5520	Tuela	Leira 2	9	0.0018
IBE-BC1826	Tuela	Leira 1	10	0.0433
IBE-BC0026	Tera	Los Tornos	11	0.0126
IBE-BC1074	Tera	Los Tornos	11	0.2025
IBE-BC1107	Tera	Los Tornos	11	0.0998
IBE-BC1151	Tera	Tabolazas - Padornelo	12	0.0637
IBE-BC1232	Tera	Tabolazas - Padornelo	12	0.1096
IBE-BC1842	Tera	Tejedelo	13	0.1667
IBE-BC1766	Tera	Requejo	14	0.0373
IBE-BC1843	Tera	Requejo	14	0.0907
IBE-BC1845	Tera	Requejo	14	0.0825
IBE-BC0096	Tera	Cabril cabecera	15	0.1046
IBE-BC0395	Tera	Cabril cabecera	15	0.0128
IBE-BC0397	Tera	Cabril cabecera	15	0.0799

IBE-BC1014	Tera	Cabril cabecera	15	0.1319
IBE-BC1047	Tera	Cabril cabecera	15	0.2018
IBE-BC1052	Tera	Cabril cabecera	15	0.0932
IBE-BC1091	Tera	Cabril cabecera	15	0.0985
IBE-BC1198	Tera	Cabril cabecera	15	0.0748
IBE-BC1238	Tera	Cabril cabecera	15	0.1423
IBE-BC0107	Tera	Cabril - Castro	16	0.1164
IBE-BC1097	Tera	Cabril - Castro	16	0.0969
IBE-BC1140	Tera	Cabril - Castro	16	0.0871
IBE-BC1770	Tera	Parada cabecera	17	0.1764
IBE-BC0062	Tera	Parada	18	0.0538
IBE-BC0302	Tera	Parada	18	0.2974
IBE-BC1037	Tera	Parada	18	0.2088
IBE-BC1046	Tera	Parada	18	0.0824
IBE-BC1103	Tera	Parada	18	0.2534
IBE-BC1216	Tera	Parada	18	0.1117
IBE-BC1721	Tera	Parada	18	0.2208
IBE-BC1026	Tera	Parada - Castro	19	0.1644
IBE-BC1080	Tera	Parada - Castro	19	0.0092
IBE-BC1746	Tera	Parada-Castro	19	0.1449
IBE-BC0015	Tera	Mondera	20	0.2924
IBE-BC0016	Tera	Mondera	20	0.2454
IBE-BC0022	Tera	Mondera	20	0.2786
IBE-BC1015	Tera	Mondera	20	0.1731
IBE-BC1139	Tera	Mondera	20	0.2668

Table S17. Relatedness coefficients estimated with simulated pedigrees in the study carried out in Zamora. Means and standard deviations (in parenthesis) are given. A total of 100 simulations of each of the 7 different pedigrees from Figure S4 were performed.

Relationship	**Theoretical value**	**Observed Value (SD)**
Parent-offspring	0.5	0.5405 (0.0403)
Full siblings	0.5	0.5295 (0.0263)
Half siblings	0.25	0.2733 (0.0384)
Grandparent-grandchild	0.25	0.2572 (0.0676)
Uncle-nephew	0.25	0.2585 (0.0421)
Half uncle-half nephew	0.125	0.1473 (0.0618)
Half-first cousins	0.0625	0.0754 (0.0412)

Table S18. Individual inbreeding coefficients estimated with simulated pedigrees of offspring (in parenthesis) from different types of parental relationships in the study carried out in Zamora. Means and standard deviations (in parenthesis) are given. A total of 100 simulations of each of the 7 different pedigrees from Figure S5 were performed.

Parental relationship (offspring)	**Theoretical value**	**Observed value (SD)**
None (F101, F102, F103 and F104)	0	0.0030 (0.0114)
Full siblings (F201 and F203)	0.25	0.2400 (0.0254)
Half siblings (F202)	0.125	0.1103 (0.0303)

Table S19. Results of kinship category assignments of a simulated outbred pedigree with PRIMUS and VCF2LR. Pedigree used was pedigree 7 from Figure S4 and the number of simulations performed was 100. Kinship categories are as follows: PO (parent-offspring), FS (full siblings), 2nd (second degree), and 3rd (third degree).

			Observed relationship						
			PRIMUS				**VCF2LR**		
Individual 1	**Individual 2**	**Simulated relationship**	**PO**	**FS**	**2nd**	**3rd**	**PO**	**FS**	**2nd**
F101	BC1842	PO	**100**	0	0	0	**100**	0	0
F102	BC1842	PO	**100**	0	0	0	**100**	0	0
F101	BC1046	PO	**100**	0	0	0	**100**	0	0
F102	BC1046	PO	**100**	0	0	0	**100**	0	0
F103	BC1046	PO	**100**	0	0	0	**100**	0	0
F104	BC1046	PO	**100**	0	0	0	**100**	0	0
F103	BC1026	PO	**100**	0	0	0	**100**	0	0
F104	BC1026	PO	**100**	0	0	0	**100**	0	0
F201	BC1080	PO	**100**	0	0	0	**100**	0	0
F101	F201	PO	**100**	0	0	0	**100**	0	0
F202	BC0027	PO	**100**	0	0	0	**100**	0	0
F104	F202	PO	**100**	0	0	0	**100**	0	0
F101	F102	FS	0	**100**	0	0	0	**100**	0
F103	F104	FS	0	**100**	0	0	0	**100**	0
F101	F103	2nd	0	0	**100**	0	0	0	**100**
F101	F104	2nd	0	0	**100**	0	0	0	**100**
F102	F103	2nd	0	0	**100**	0	0	0	**100**
F102	F104	2nd	0	0	**100**	0	0	0	**100**
F201	BC1842	2nd	0	0	**100**	0	0	0	**100**
F201	BC1046	2nd	0	0	**100**	0	0	0	**100**
F202	BC1046	2nd	0	0	**51**	49	0	0	**100**
F202	BC1026	2nd	0	0	**12**	88	0	0	**100**
F102	F201	2nd	0	0	**100**	0	0	0	**100**
F103	F202	2nd	0	0	**65**	35	0	0	**100**

Table S20. Results of kinship category assignments of a simulated inbred pedigree (Figure S6) with PRIMUS and VCF2LR. The number of simulations performed was 100. Kinship categories are as follows: PO (parent-offspring), FS (full siblings), 2nd (second degree), and 3rd (third degree). Inbred individuals are marked with an asterisk.

			Observed relationship						
			PRIMUS				**VCF2LR**		
Individual 1	**Individual 2**	**Simulated relationship**	**PO**	**FS**	**2nd**	**3rd**	**PO**	**FS**	**2nd**
F102	BC1843	PO	**100**	0	0	0	**100**	0	0
F102	BC1046	PO	**100**	0	0	0	**100**	0	0
F102	F202*	PO	**100**	0	0	0	**100**	0	0
F103	BC1046	PO	**100**	0	0	0	**100**	0	0
F103	BC1232	PO	**100**	0	0	0	**100**	0	0
F103	F202*	PO	**100**	0	0	0	**100**	0	2
F103	F203*	PO	**100**	0	0	0	**100**	0	1
F104	BC1232	PO	**100**	0	0	0	**100**	0	0
F104	BC1080	PO	**100**	0	0	0	**100**	0	0
F104	F203*	PO	**100**	0	0	0	**100**	0	0
F202*	F301*	PO	**100**	0	0	0	**100**	0	0
F203*	F301*	PO	**100**	0	0	0	**100**	0	0
F101	F102	FS	0	**100**	0	0	0	**100**	0
F104	F105	FS	0	**100**	0	0	0	**100**	0
F201*	F202*	FS	0	**100**	0	0	0	**100**	0
F203*	F204*	FS	0	**100**	0	0	0	**100**	0
F301*	F302*	FS	0	**100**	0	0	0	**100**	0
F102	F103	2nd	0	0	**100**	0	0	6	**94**
F103	F104	2nd	0	0	**99**	1	0	0	**100**
F202*	F203*	2nd	0	2	**98**	0	0	10	**97**
F202*	BC1843	2nd	0	0	**100**	0	0	0	**100**
F202*	BC1046	2nd	0	98	**2**	0	0	100	**100**
F202*	BC1232	2nd	0	0	**100**	0	0	0	**100**
F203*	BC1046	2nd	0	0	**93**	7	0	0	**100**
F203*	BC1232	2nd	0	95	**5**	0	1	99	**100**
F203*	BC1080	2nd	0	0	**87**	13	0	0	**100**
F301*	F102	2nd	0	18	**82**	0	0	62	**38**
F301*	F103	2nd	2	98	**0**	0	2	98	**0**
F301*	F104	2nd	0	3	**97**	0	0	10	**90**

Table S21. Means and standard deviations (in parenthesis) of the distances in meters for the kinship categories estimated with PRIMUS.

Kinship category	**Number of relationships**	**Mean distance (SD)**	**Distance per generation**
Parent-offspring	14	881 (1802)	881
Full siblings	36	2230 (2803)	1115
Second degree	121	2799 (2822)	1399
Third degree	327	4196 (3538)	1399
Distantly related	81	7047 (3873)	-

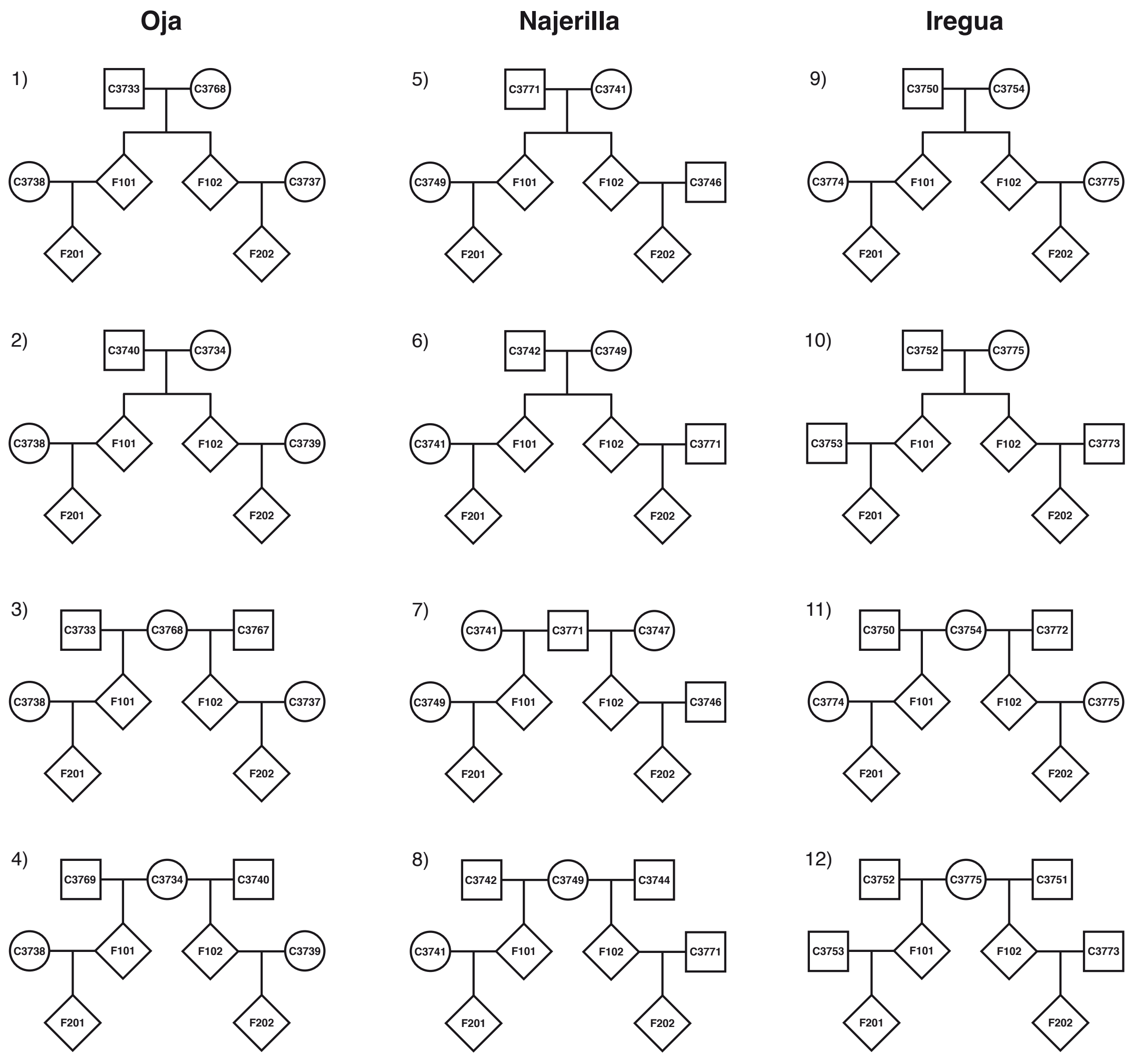

Figure S1. Pedigrees used for simulations of relationships among individuals belonging to the same river in the study carried out in La Rioja. Founder males are represented with squares, founder females with circles and simulated offspring with diamonds. The identity of the founders can be found in Table S1.

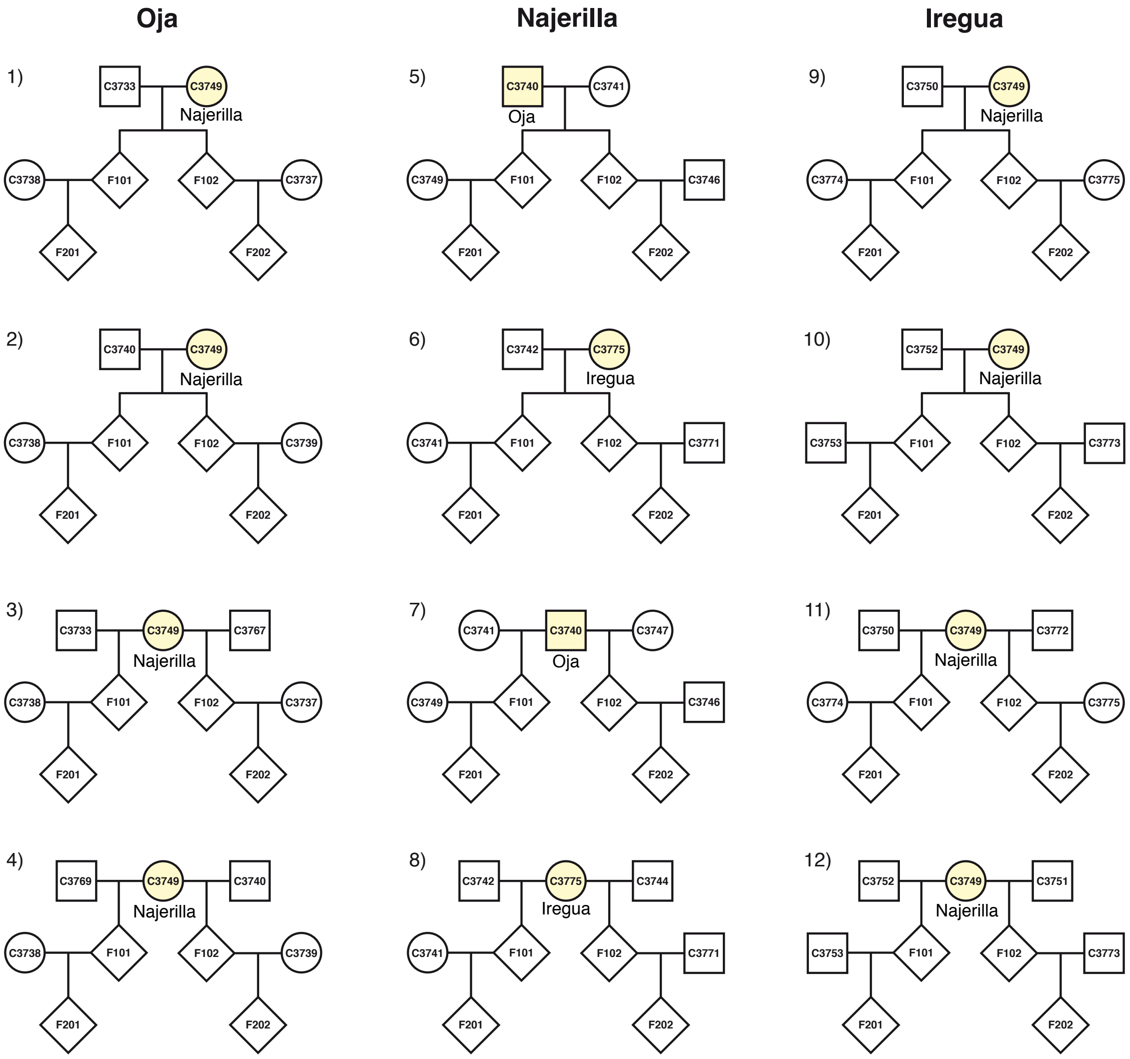

Figure S2. Pedigrees used for simulations of relationships among individuals belonging to the different rivers in the study carried out in La Rioja. Founder males are represented with squares, founder females with circles and simulated offspring with diamonds. Individuals from a different river (migrants) are represented in yellow. The identity of the founders can be found in Table S1.

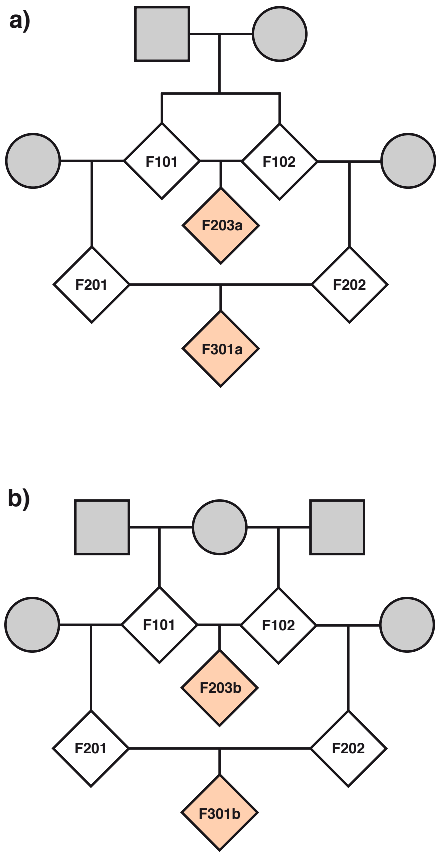

Figure S3. Additional crosses in the pedigrees used for simulations of individual inbreeding coefficients in the study carried out in La Rioja with a) 4 founders and b) 5 founders. Founder males are represented with squares, founder females with circles and simulated offspring with diamonds. Inbred individuals are represented in light red. The identity of the founders is the same as in Figure S2.

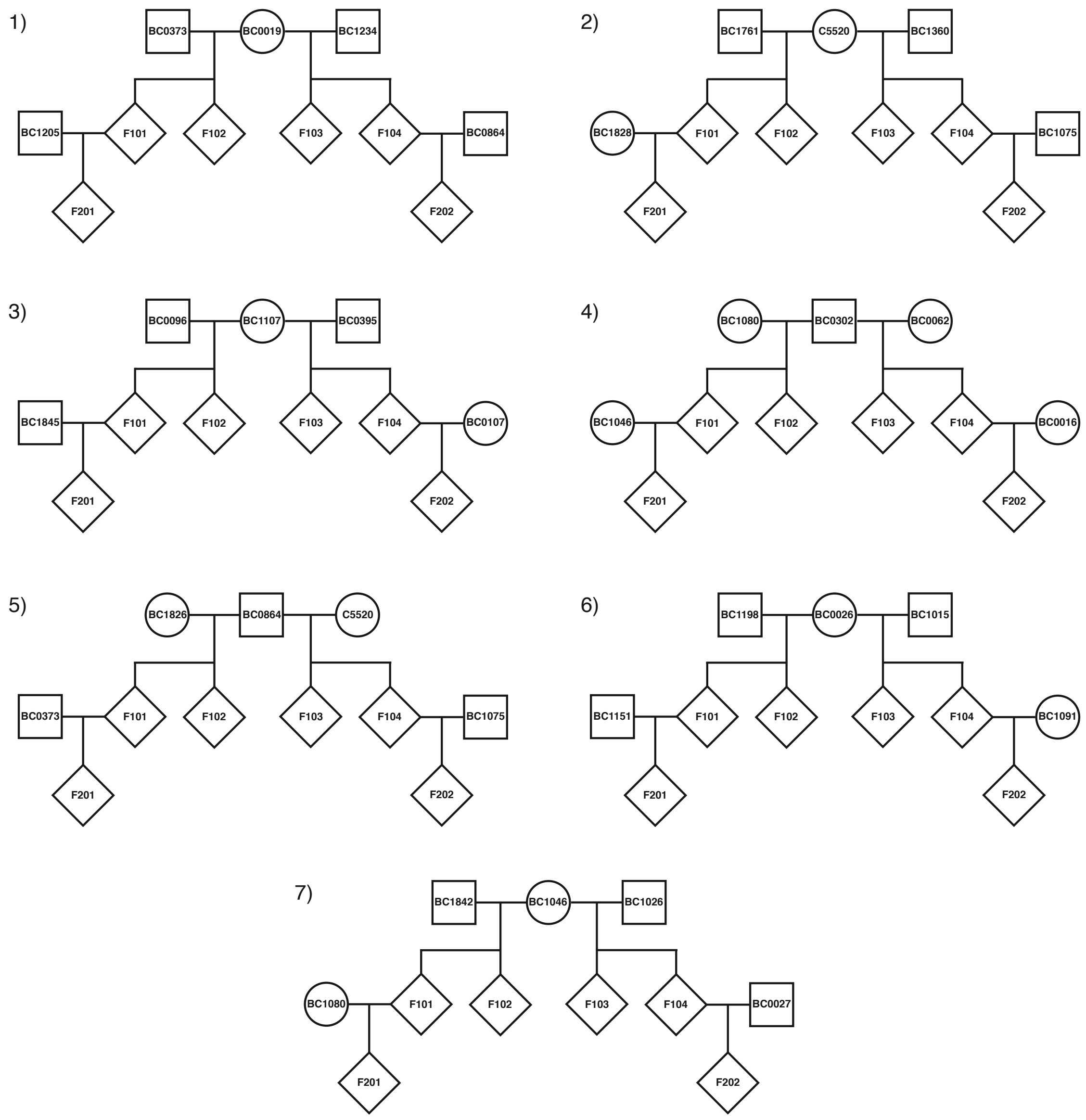

Figure S4. Pedigrees used for simulations of relatedness between individuals in the study carried out in Zamora. Founder males are represented with squares, founder females with circles, and simulated offspring with diamonds. The identity of the founders can be found in Table S4.

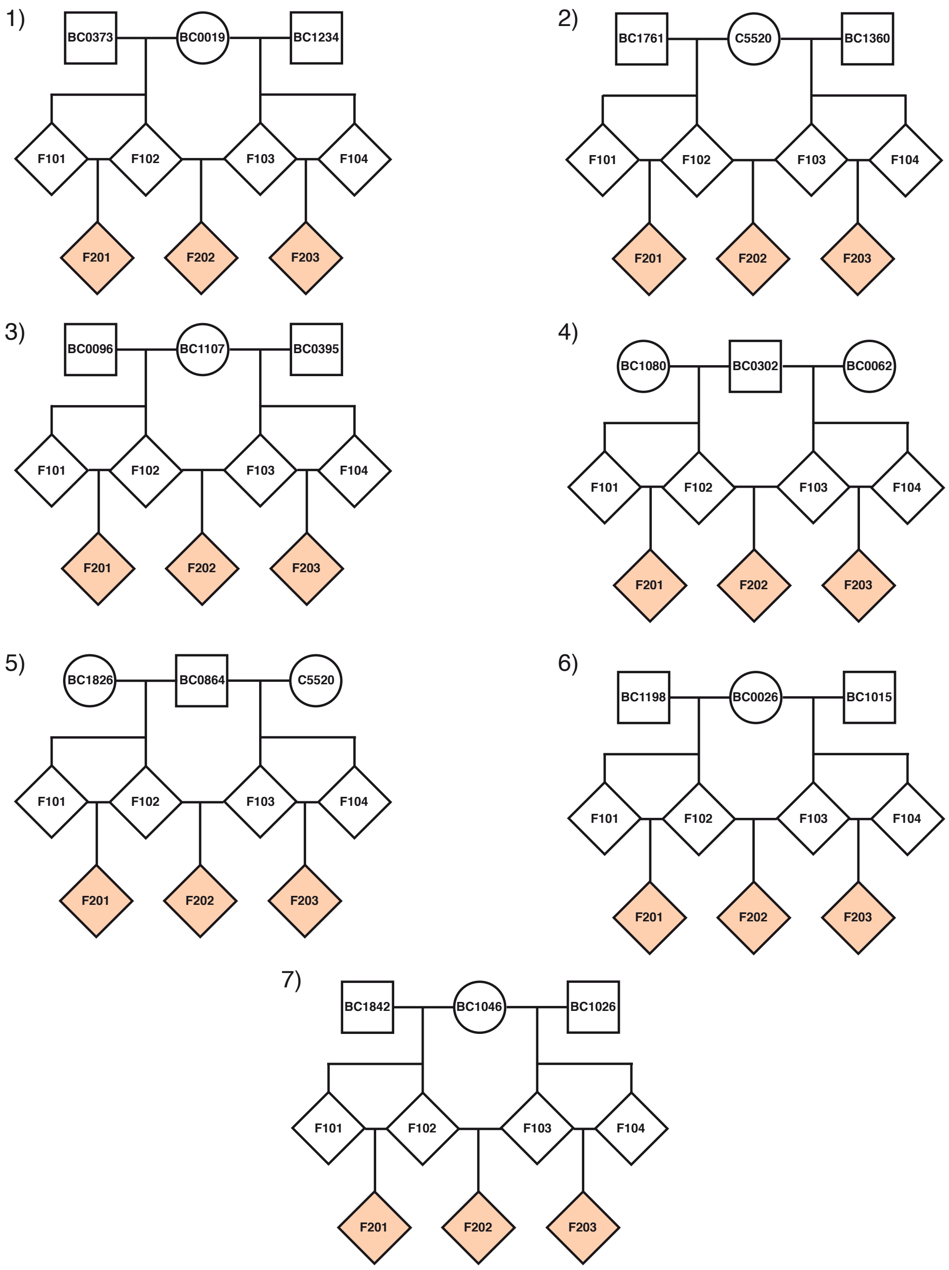

Figure S5. Pedigrees used for simulations of individual inbreeding coefficients in the study carried out Zamora. Founder males are represented with grey squares, founder females with grey circles, and simulated offspring with diamonds. Inbred individuals are represented in light red. The identity of the founders can be found in Table S4.

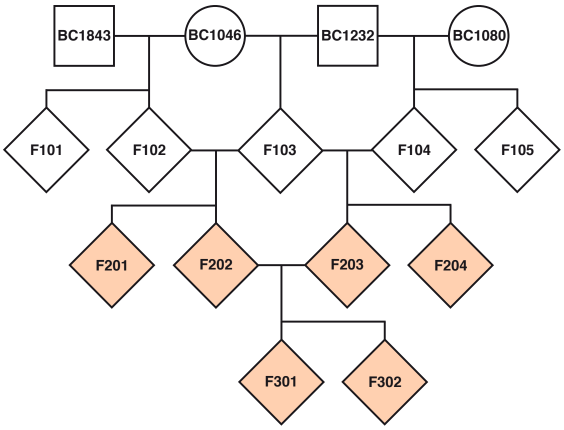

Figure S6. Pedigree used for simulations of kinship category determination in a scenario with high inbreeding in the study carried out in Zamora. Founder males are represented with squares, founder females with circles, and simulated offspring with diamonds. Inbred individuals are represented in light red. The identity of the founders can be found in Table S4.

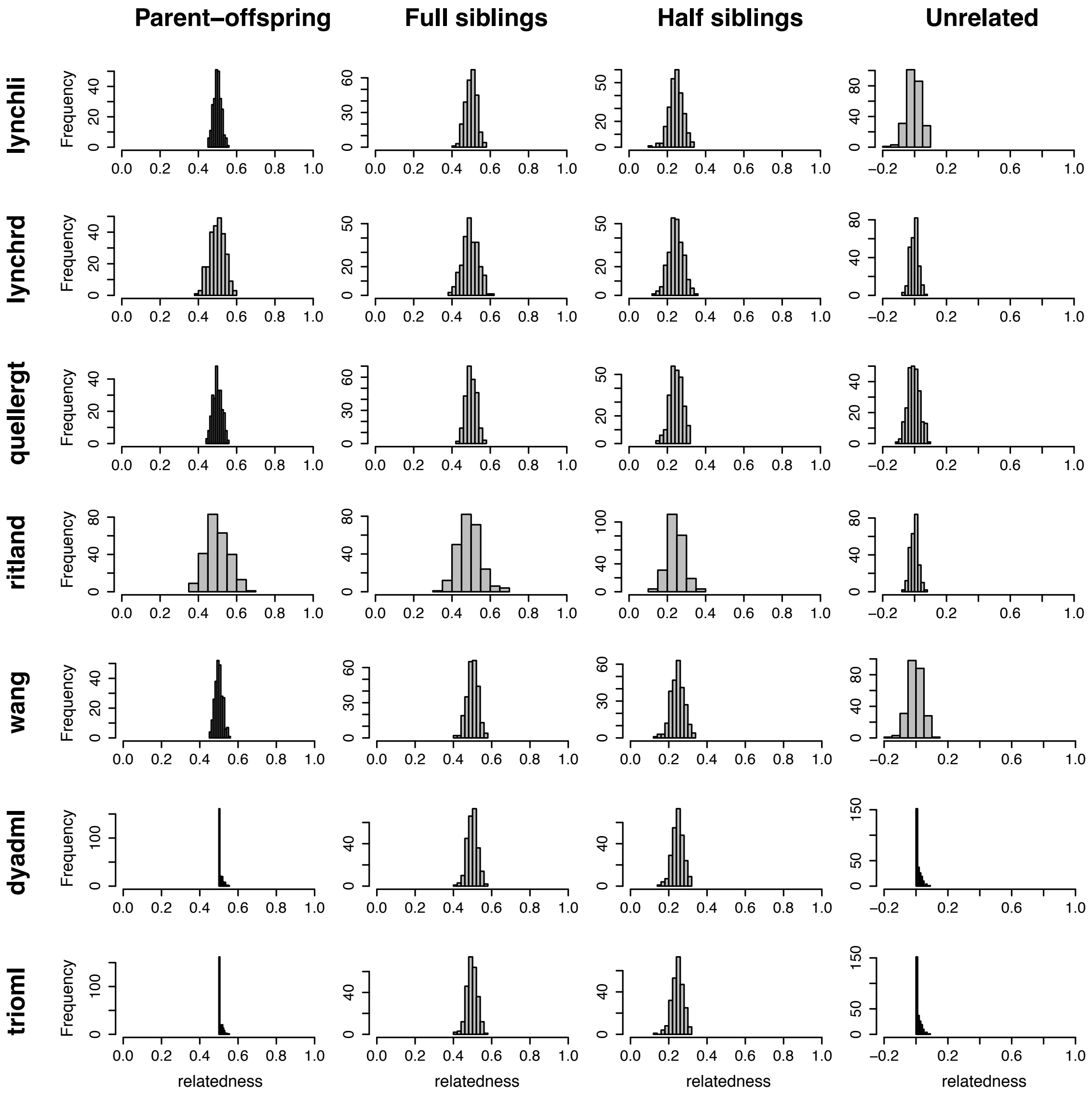

Figure S7. Frequency histograms of the relatedness values obtained from the simulations performed with the program RELATED for the individuals in the study carried out in La Rioja, using population allele frequencies and all seven estimators.

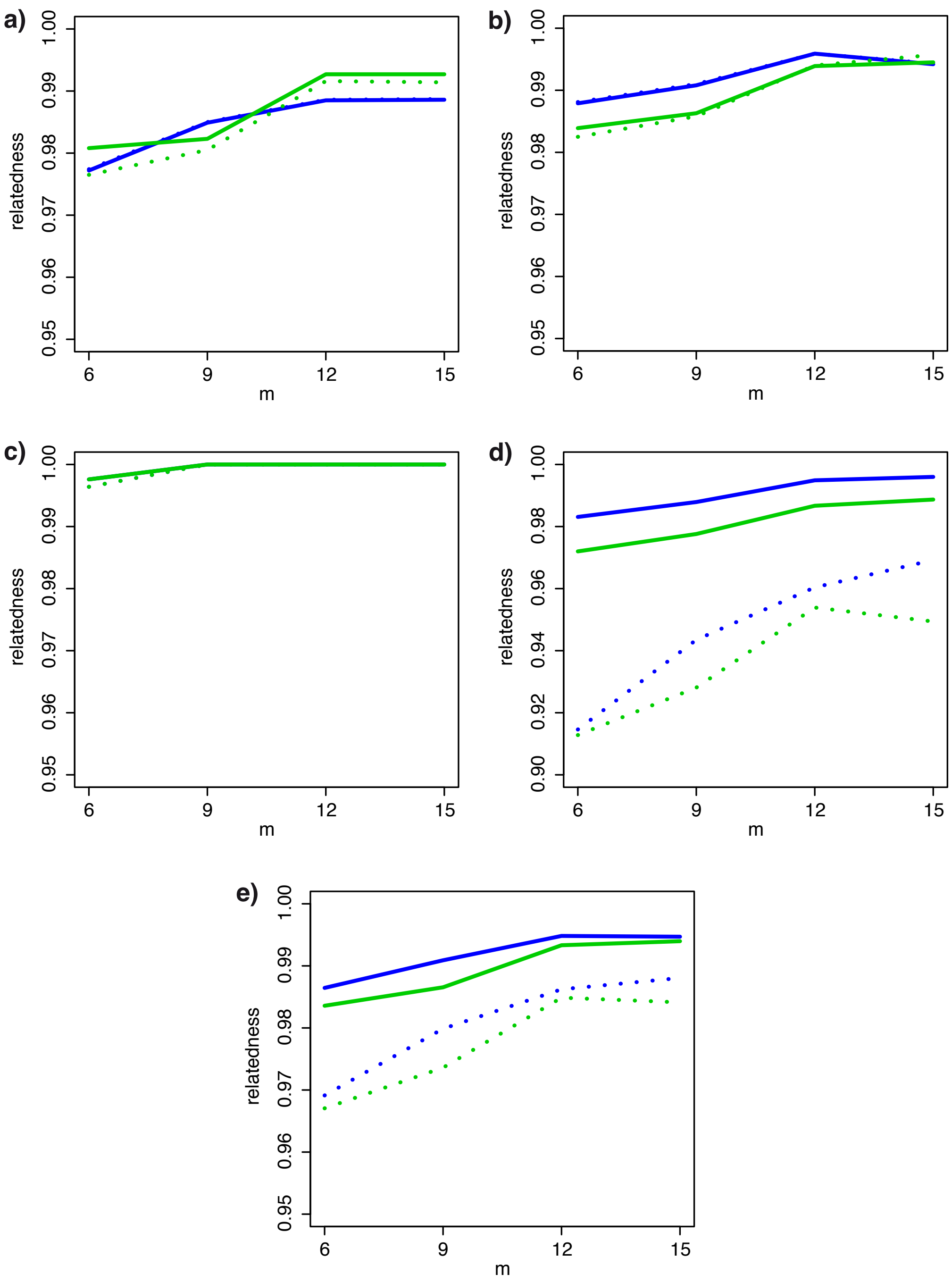

Figure S8. Plots of the relatedness values for the duplicated samples tested and mean values. Blue lines correspond to r = 1, green lines to r = 0.9, solid lines to MAF = 0.05, and dashed lines to MAF = 0. The graphics correspond to the analysis of samples (a) IBE-C3767, (b) IBE-C3765, (c) IBE-C3749, (d) IBE-C3742, and (e) mean values of the four pairs.

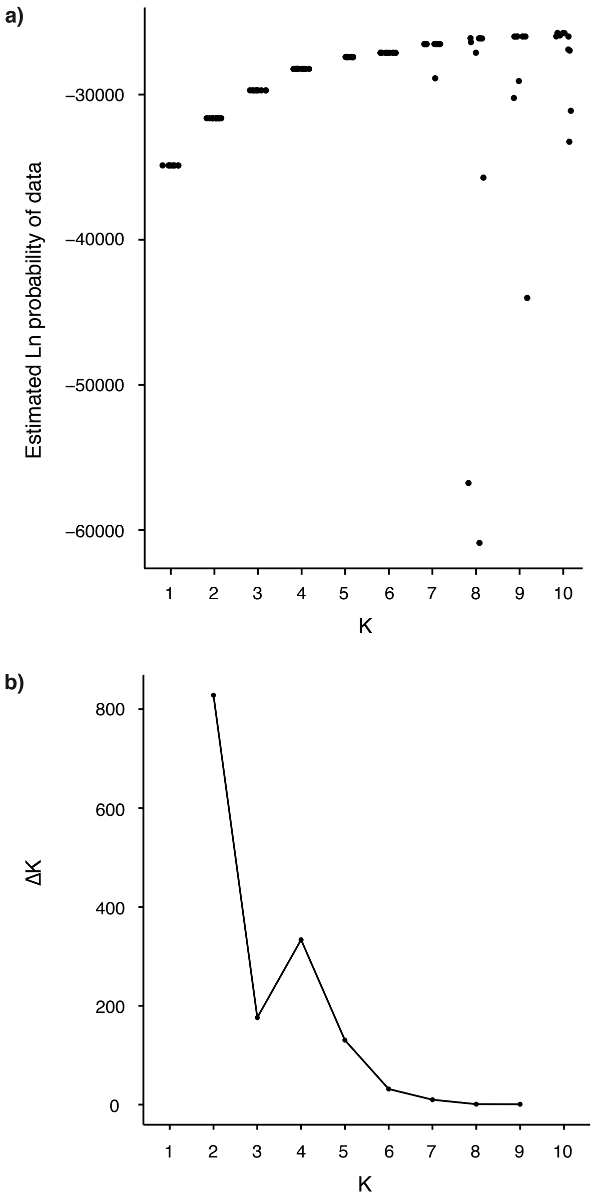

Figure S9. (a) Plot of the mean estimated posterior probability of the data for different K values in STRUCTURE in the study carried out in La Rioja. Values for each of 10 independent runs are shown. (b) Plot of the ΔK values for each K.

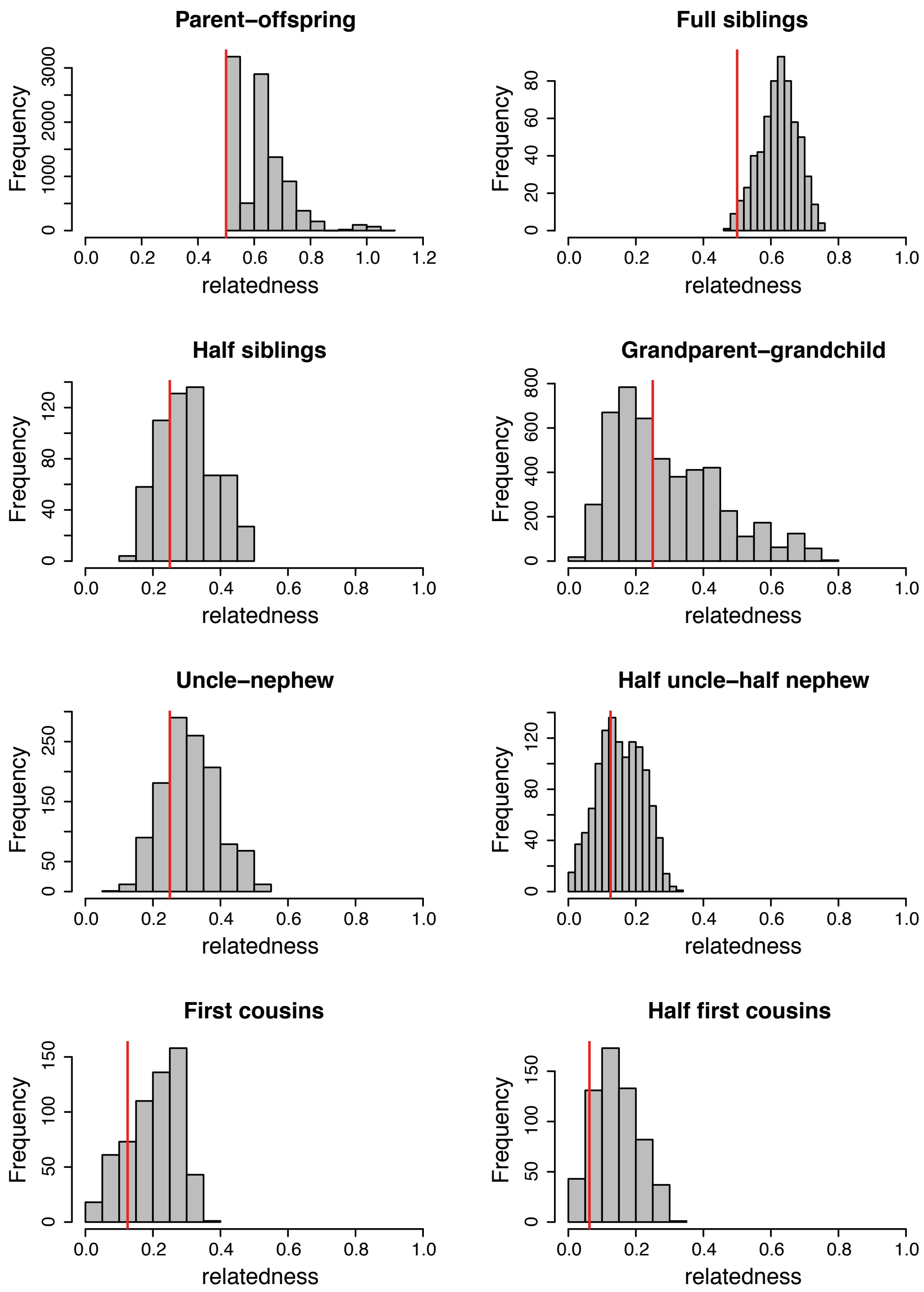

Figure S10. Frequency histograms of the relatedness values obtained from the simulations performed along artificial pedigrees with migrants from Figure S2 in the study carried out in La Rioja. Red lines indicate the expected values for each category.

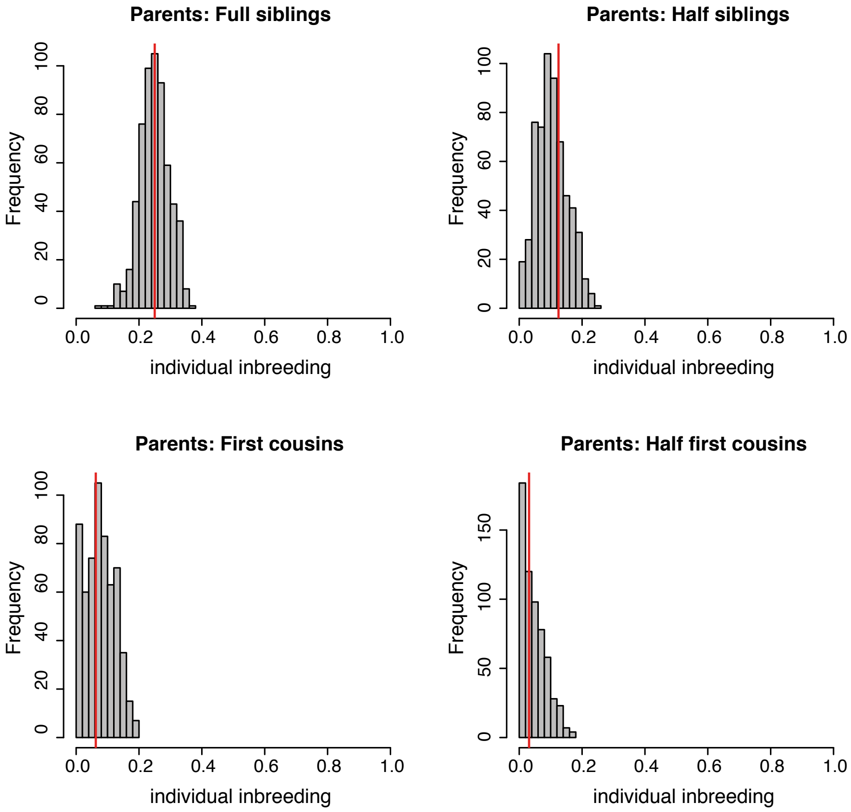

Figure S11. Frequency histograms of the individual inbreeding coefficients obtained from the simulations performed along artificial pedigrees with migrants and crosses of pairs with known parental relationships from Figure S3 in the study carried out in La Rioja. Red lines indicate the expected values for each category.

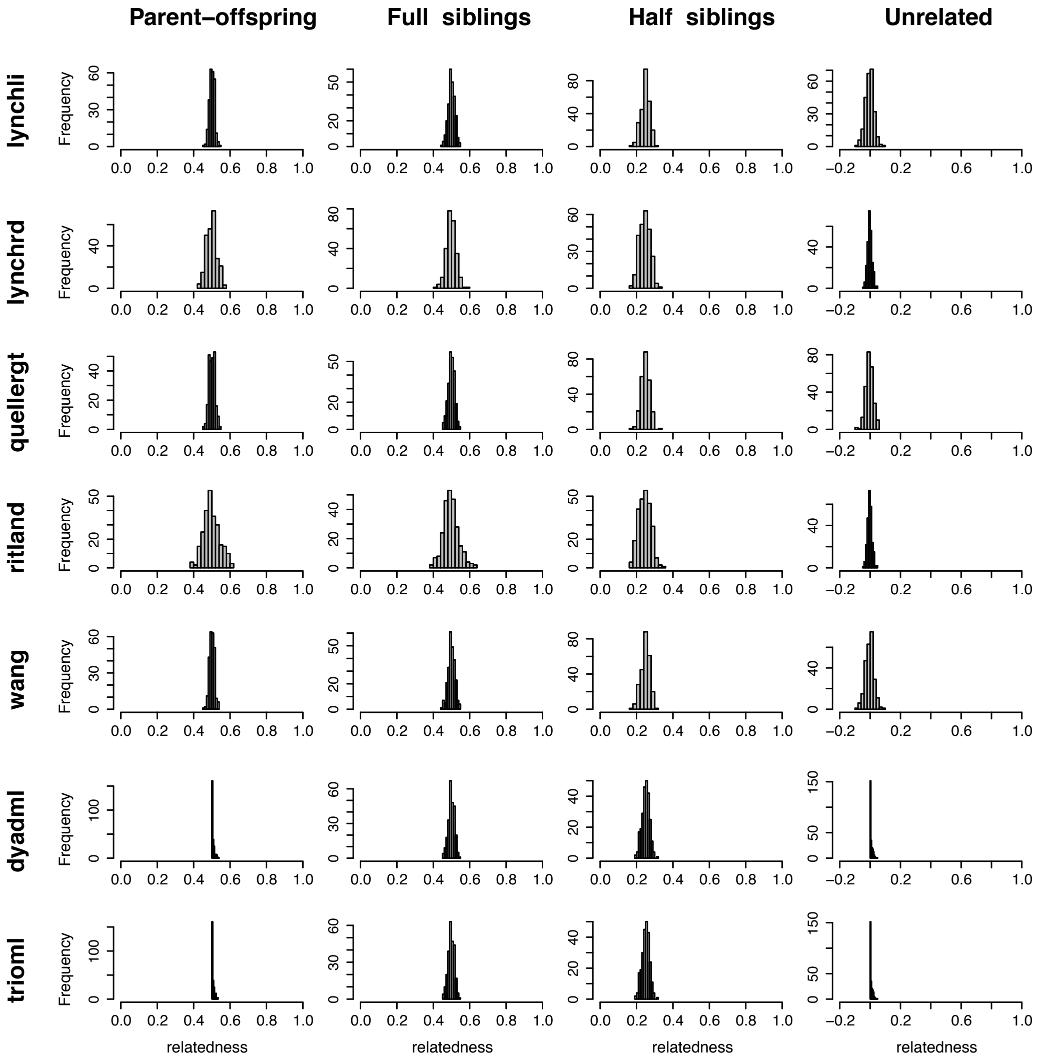

Figure S12. Frequency histograms of the relatedness values obtained from the simulations performed with the program RELATED in the study carried out in Zamora, using population allele frequencies and all seven estimators.

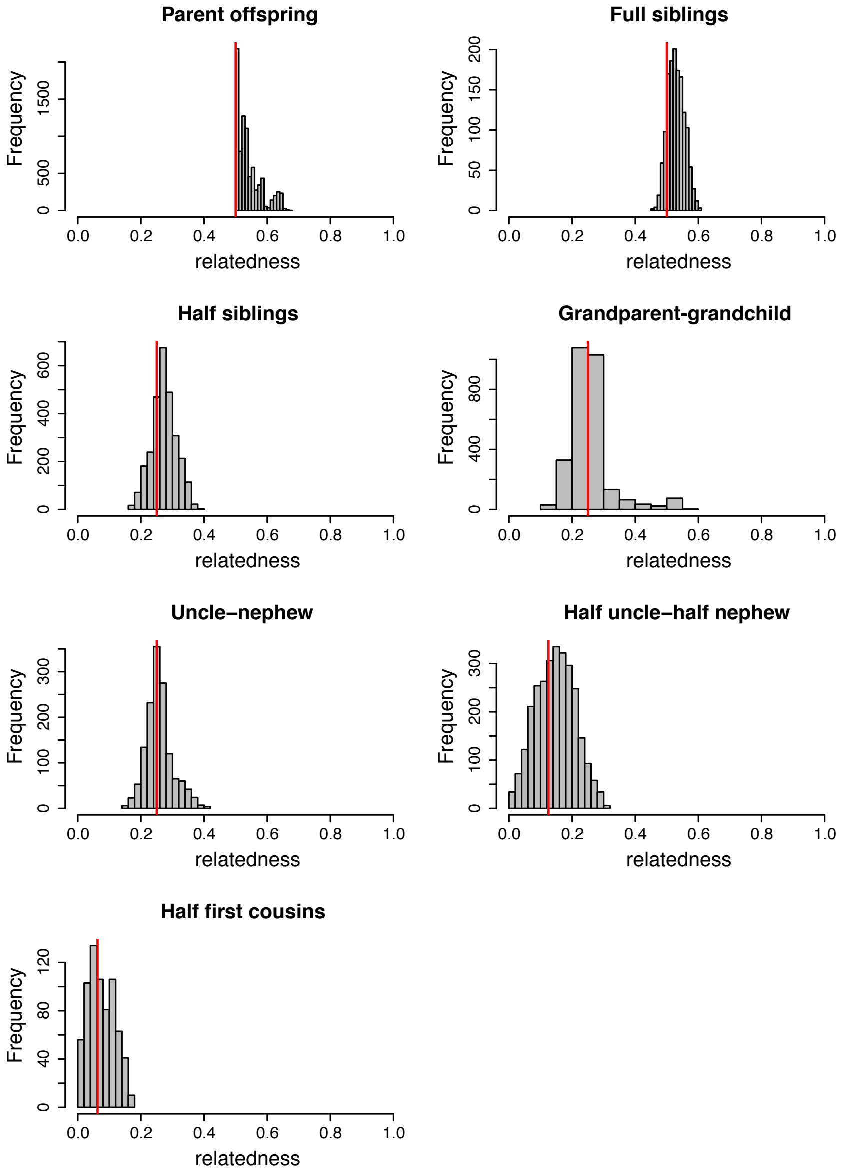

Figure S13. Frequency histograms of the relatedness values obtained form the simulations performed along artificial pedigrees from Figure S4 in the study carried out in Zamora. Red lines indicate the expected values for each category.

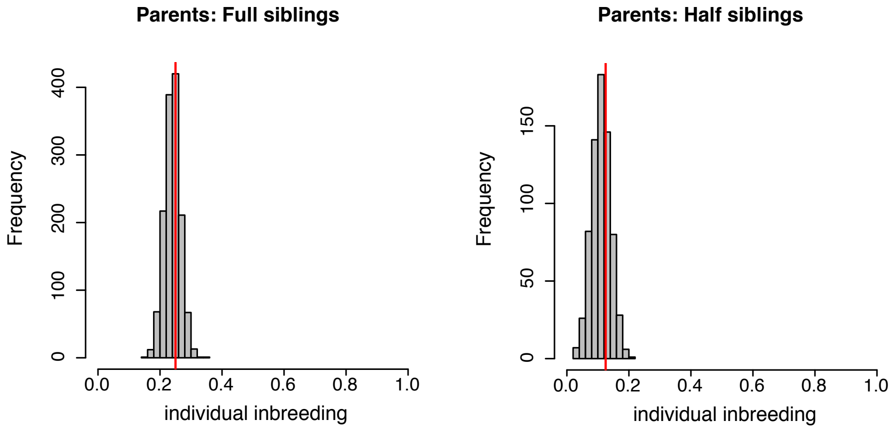

Figure S14. Frequency histograms of the individual inbreeding coefficients obtained from the simulations performed along artificial pedigrees with crosses of pairs with known parental relationships from Figure S5 in the study carried out in Zamora. Red lines indicate the expected values for each category.

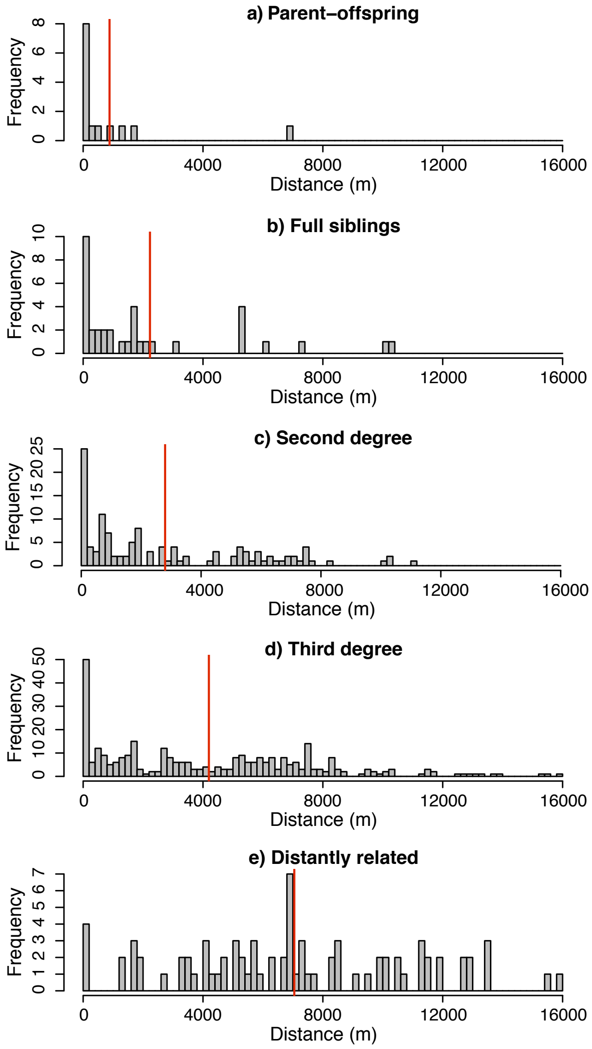

Figure S15. Frequency histograms of the distances between related individuals for each kinship category inferred with PRIMUS. Vertical red lines indicate the mean value for each category: (a) parent-offspring relationships, (b) full-sibling relationships, (c) second-degree relationship, (d) third-degree relationships, and (e) distantly related pairs.

PUBLICATION

Escoda L., González-Esteban J., Gómez A., Castresana J. (2017) Using relatedness networks to infer contemporary dispersal: Application to the endangered mammal *Galemys pyrenaicus*. *Molecular Ecology*, **26**, 3343–3357.

Received: 2 November 2016 | Accepted: 21 March 2017

DOI: 10.1111/mec.14133

ORIGINAL ARTICLE

WILEY MOLECULAR ECOLOGY

Using relatedness networks to infer contemporary dispersal: Application to the endangered mammal *Galemys pyrenaicus*

Lídia Escoda[1] | Jorge González-Esteban[2] | Asunción Gómez[3] | Jose Castresana[1]

[1]Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), Barcelona, Spain

[2]Desma Estudios Ambientales S.L., Sunbilla (Navarra), Spain

[3]Área de Biodiversidad, Tragsatec, Madrid, Spain

Correspondence
Jose Castresana, Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), Barcelona, Spain.
Email: jose.castresana@csic.es

Funding information
Ministerio de Economía y Competitividad, Grant/Award Number: CGL2014-53968-P; Generalitat de Catalunya, Grant/Award Number: 2015FI_B 00429

Abstract

Information about the degree of contemporary dispersal is important when trying to understand how populations interchange individuals and identify the specific barriers that prevent these movements. In the case of endangered species, this can represent crucial information when designing appropriate conservation strategies. Here we analyse relatedness between individuals from different localities and use these data to infer whether dispersal occurred in recent generations. We applied this approach to the Pyrenean desman (*Galemys pyrenaicus*), a semiaquatic and endangered species endemic to the Iberian Peninsula. We studied this species in four primary rivers in the Iberian Range, where two ancient mitochondrial lineages are separated by a strict contact zone, suggesting the existence of complex dispersal patterns. Using next-generation sequencing, we obtained 912 SNPs from each specimen and estimated relatedness values between them. While relatedness networks were dense within each river, we found surprisingly few relationships between individuals from different rivers despite their close proximity in some cases, indicating much lower dispersal between rivers compared to dispersal within a single river. In agreement with this result, the degree of inbreeding was exceedingly high in most individuals. These data show that relatedness information can be crucial to understand the contemporary dispersal patterns and conservation status of specific populations of endangered species.

KEYWORDS
conservation genetics, dispersal, inbreeding, mammals, relatedness

1 | INTRODUCTION

Knowledge of dispersal patterns is essential for the management of endangered species (Baguette, Blanchet, Legrand, Stevens, & Turlure, 2012; Woodroffe, 2003). This information is particularly important in the case of small and fragmented populations, which may become extinct due to catastrophic events, environmental stochasticity or inbreeding depression if they remain isolated for a long time (Soulé, 1987). Animal dispersal may have positive effects on the viability of fragmented populations, for example, by reducing the effects of inbreeding, but it may also have negative effects, for example, due to increased mortality of dispersing animals (Baguette et al., 2012; Banks & Lindenmayer, 2014; Woodroffe, 2003). It is therefore important to know how dispersal affects the viability of populations. However, it is extremely difficult to determine whether some individuals can move from one population to another and breed in the new population, and more so for elusive species (Mills, 2013). Different multilocus-based genetic techniques have been used extensively to estimate long-term migration rates between populations (Durand, Patterson, Reich, & Slatkin, 2011; Pinho & Hey, 2010; Slatkin, 1985), but these techniques are best suited to the study of ancient dispersal. However, for endangered species, it is vital to have information concerning current movements or those of the last few generations. A different kind of multilocus genetic method can be used to assign individuals of unknown origin to potential source populations and thus detect migration events that occurred in the last one or two generations (Piry et al., 2004; Wilson & Rannala, 2003). Nevertheless, all these techniques require the presence of populations with different allele frequencies to detect the individuals' source population. In fact, populations that only became isolated recently may have a similar genetic background and so these techniques would not be able to detect movements between them. Furthermore, these population-based approaches require that individuals are previously assigned to populations, which is not always possible in populations with diffuse borders or permeable contact zones. To overcome these difficulties, an additional class of genetic techniques based on paternity analysis has been proposed to analyse recent dispersal (Wang, 2014b). These methods try to determine the individual's source population by identifying its parents. However, these methods require the analysis of a large number of specimens from different populations to identify a sufficient number of parental relationships, which may not be feasible for rare species. Similar kinship-based approaches try to infer the presence of dispersal through individuals that do not have close relatives in their social group, but they also require comprehensive sampling of each group in the population (Rollins et al., 2012). Finally, a number of other methods based on kinship analysis can be used to infer population structure and, indirectly, migration rates between the identified populations (Økland, Haaland, & Skaug, 2010; Palsbøll, Zachariah Peery, & Bérubé, 2010; Watts et al., 2007).

It would undoubtedly be of great benefit if there were genetic methods available with a greater flexibility to detect recent migration. Here we explore the possibility of extending the paternity analysis to more distant relationships and using relatedness values estimated between all pairs of individuals to infer dispersal. The idea is that when two individuals from different localities present significant relatedness values, then some of these individuals or their ancestors must have dispersed between the localities in the last few generations. Although the actual dispersal route cannot be determined and may involve intermediate, unknown localities, recent gene flow between localities can be inferred. When applied to many individual pairs, the population's dispersal pattern over the last few generations emerges.

The kinship coefficient of two individuals is the probability that two alleles, each randomly chosen from each individual at the same locus, are identical by descent (Blouin, 2003; Jacquard, 1972; Weir, Anderson, & Hepler, 2006). This probability can range from 0 for unrelated individuals to one for two completely inbred (homozygous) twins. The relatedness coefficient, which is two times the kinship coefficient, is most commonly calculated to describe the degree of shared ancestry (Milligan, 2003; Wang, 2011). Relatedness can be estimated by methods of moments or by maximum likelihood. Methods of moments (Li, Weeks, & Chakravarti, 1993; Lynch & Ritland, 1999; Queller & Goodnight, 1989; Ritland, 1996; Wang, 2002) are based on allele frequency moments and allow negative values for relatedness, which could arise if the reference population is the same as the population being analysed (Wang, 2014a). By contrast, maximum-likelihood methods use a model with different identity states that reflect the mode of identity-by-descent between the four alleles compared (two per individual) at each locus (Jacquard, 1972; Milligan, 2003; Wang, 2007). When inbreeding is considered negligible, the model consists of three identity states corresponding to zero, one or two alleles shared between two individuals; genotype data can be used to estimate the probability of these states and subsequently calculate the coefficient of relatedness. When inbreeding is taken into account, a full model of nine identity states allows to calculate, not only the pairwise relatedness, but also the inbreeding coefficient of each individual, which is the probability that two alleles at a locus in an individual are identical by descent. The inbreeding coefficient of an individual is equivalent to the kinship coefficient of its parents and is a critical parameter when characterizing the genetic health of a population (Ellegren, 1999; Hammerly, Morrow, & Johnson, 2013; Liu et al., 2014). In maximum-likelihood methods, kinship, relatedness and inbreeding estimates are limited to values above 0 due to their probabilistic nature.

Relatedness can be estimated using genealogical relationships from a complete pedigree or using molecular markers, such as microsatellites or single nucleotide polymorphisms (SNPs). Relatedness values identify not only parent-offspring or full-sibling relationships (relatedness = 0.5), but they can also determine more distant relationship categories such as grandparent-grandchild (0.25), half-siblings (0.25) or first cousins (0.125) (Blouin, 2003; Milligan, 2003; Weir et al., 2006). However, the large variance in these estimates when calculated using traditional markers means the values obtained are only approximate for specific dyads and, generally, only the average relatedness of groups or populations is considered (Taylor, 2015). Quantification of relatedness has been widely used in different aspects of wildlife studies such as the analysis of social organization and philopatry (Arora et al., 2012; Bonin, Goebel, O'Corry-Crowe, & Burton, 2012; Watts, Scribner, Garcia, & Holekamp, 2011). In this study we plan to use the relatedness coefficient, rather than specific relationships categories, to infer dispersal.

Until recently, most studies using relatedness were based on microsatellites (Taylor, 2015). There is still a reduced number of studies that have used SNPs to estimate relatedness and only some of them evaluated SNP performance. For example, Santure et al. (2010) found that 771 SNPs provided inconclusive results when estimating relatedness in zebra finch pedigrees, whereas Lopes et al. (2013) showed that 2,000 of these markers can give optimal estimates in pig pedigrees. Although both studies used a high number of SNPs, there may be large variations in the information contained by different sample sets. Hence, further analysis of specific cases, particularly natural populations, is necessary for a better understanding of the optimal conditions under which SNPs can be used in relatedness analysis.

The Pyrenean desman (*Galemys pyrenaicus*) is a small eulipotyphlan mammal endemic to the northern part of the Iberian Peninsula. It is a semi-aquatic species with strong adaptations to the aquatic medium and inhabits clean, oxygenated rivers and streams (Charbonnel et al., 2015; Palmeirim & Hoffmann, 1983). The species has suffered strong declines in the last few years for reasons not fully understood. The primary causes put forward to explain this regression include water pollution, the desiccation of rivers and construction of dams (Fernandes, Herrero, Aulagnier, & Amori, 2011). The species currently has a patchy distribution and is generally found in mountain rivers (Nores, Queiroz, & Gisbert, 2007). Population densities in rivers vary between 3 and 7 individuals/km (Nores et al., 1998), although these estimates might be higher if the species is less territorial than previously assumed (Melero, Aymerich, Santulli, & Gosálbez, 2014). The Pyrenean desman is a legally protected species in its four native countries (Spain, Portugal, France and Andorra) and is classified as "vulnerable" in the IUCN Red List (Fernandes et al., 2011). Furthermore, significant declines in populations located in the southern part of the distribution (Gisbert & Garcia-Perea, 2014) led the Spanish Government to catalogue such populations as "in danger of extinction." As the species' habitat is restricted to the upper parts of certain rivers, generally in mountain areas, even its natural populations are fragmented. This fragmentation has probably been aggravated by the construction of dams and other physical barriers in many rivers inhabited by the species (Charbonnel et al., 2015; Fernandes et al., 2011). However, it is not known how these infrastructures affect the dispersal capabilities and genetic health of the Pyrenean desman. In fact, information on long-range movements of the Pyrenean desman is scarce. Radio-tracking data has provided valuable information regarding the behaviour of the species (Melero, Aymerich, Luque-Larena, & Gosálbez, 2012; Melero et al., 2014), but this technique can only detect short movements over a short period of time and therefore long-range dispersal movements, if they exist, go unobserved. Moreover, individual identification from genotyping faeces has produced little movement data to date (Gillet et al., 2016). This species is therefore an ideal model for studying the current interconnectivity of populations in different rivers and to test whether relatedness between individuals can be used to infer dispersal between localities and populations.

In previous work, Igea et al. (2013) studied the phylogeography of the Pyrenean desman using mitochondrial genes and found a marked phylogeographic structure comprised of four mitochondrial lineages (A1, A2, B1 and B2). They detected two contact zones between the main mitochondrial groups (A and B), one located in the Cantabrian Mountains and the other in the Iberian Range. In a more recent study, Querejeta et al. (2016) analysed genomic data and found five clusters that were largely coincident with the mitochondrial clades, although some differences were detected, particularly in the Iberian Range. Whereas Igea et al. (2013) detected two mitochondrial clades in the Iberian Range, A2 in the southeast and B1 in the northwest, Querejeta et al. (2016) observed just one main genomic cluster, although subdivided into two subgroups with different genomic compositions. However, these previous studies only counted on a small number of samples from the Iberian Range. A larger sample size, together with a more detailed analysis of genetic structure and a study of contemporary dispersal patterns, may help us better understand the evolutionary history of the species in this area and estimate the degree of intermixing that occurs across the contact zone.

Here we used next-generation sequencing techniques to obtain SNPs for Pyrenean desmans sampled from different rivers of the Iberian Range. We then estimated relatedness among all pairs and, to visualize dispersal patterns, constructed relatedness networks. As the use of SNPs to analyse relatedness and inbreeding has not been widely explored to date, we devised different simulation analyses to determine the accuracy of the results. We show that the information obtained from these analyses is of great use when inferring contemporary dispersal patterns in this species and helps us better understand the degree of isolation of different populations. We also demonstrate how this information may be important in the identification of critical conservation problems faced by the species.

2 | MATERIALS AND METHODS

2.1 | Samples

We used a total of 66 samples of Pyrenean desmans from different rivers of the Iberian Range. Of these, 37 tissue samples were used to perform library construction and genomic analysis (Table S1). All tissue samples were obtained from specimens captured during a monitoring project performed in 2011 for the La Rioja Regional Government (Spain), with permit number A/2011/52. A small portion was taken from the tip of the tail before the captured specimens were released back in the wild. The work was carried out following national and international regulations, and all necessary steps were taken to prevent any damage to the specimens. An additional set of 26 samples consisting of faeces that desmans deposit in exposed rocks of the rivers (Igea et al., 2013) as well as three specimens found dead in the field were used to complete the mitochondrial phylogeography in this area (Table S2). All samples were conserved in tubes containing absolute ethanol and stored at −20°C (tissues) or 4°C (faeces) in the laboratory.

DNA extraction, PCR of mitochondrial sequences and quantification of DNA concentration are explained in Appendix S1 of Supporting Information.

2.2 | Phylogenetic analysis of mitochondrial sequences

The mitochondrial cytochrome *b* gene was amplified from all the samples. To assign the mitochondrial clade to each specimen, a maximum-likelihood phylogenetic tree of the aligned cytochrome *b* sequences was reconstructed using RAXML version 8.0.17, with a GTR model of nucleotide substitution and a gamma distribution of evolutionary rates, as recommended in the program (Stamatakis, 2014).

2.3 | Sex determination by qPCR

To sex the specimens, we used a pair of primers that amplified a 77 bp region of intron 42 of USP9Y of the Y chromosome by qPCR (Querejeta et al., 2016). Further details of the qPCR conditions can be found in Appendix S1.

2.4 | Library construction and Illumina sequencing

DNA libraries were constructed using the ddRAD protocol (Peterson, Weber, Kay, Fisher, & Hoekstra, 2012) with some modifications from Querejeta et al. (2016). Each library was performed in series of 24 samples, repeating samples with a low sequence yield in subsequent experiments to get sufficient coverage for all samples. For each sample, 50 ng of genomic DNA (as estimated by qPCR) was double digested using EcoRI and MspI restriction enzymes, in a 30 μl final volume. After overnight incubation, the enzymatic reaction was heat inactivated at 80°C for 30 min. We then added a 70 μl ligation mix, including T4 DNA ligase, P1 adapter (which binds to the EcoRI overhangs and has a 5-nucleotide barcode to identify each specimen), and P2 adapter (which binds to the MspI overhangs), and the solution was incubated for 5 hr at room temperature. The ligation reaction was heat inactivated at 65°C for 10 min. Then all the ligation reactions were mixed in a single tube and concentrated to 20 μl using the MinElute PCR Purification Kit (QIAGEN). The entire pool was run in a precast EX 2% agarose gel using the E-Gel system (Invitrogen) and the fraction between 300 and 400 bp was cut and purified with the QIAquick Gel Extraction Kit (QIAGEN) in 30 μl. We subsequently performed a PCR amplification of the size-selected sample to add Illumina adapters. Phusion High-Fidelity DNA Polymerase (New England Biolabs) amplifications were carried out using 6 μl of the size-selected sample pool as template and employing 16–20 cycles (depending on the intensity of the initial PCR products). A total of five PCRs were performed to increase the concentration of the libraries and to minimize bias. Then we combined and concentrated the PCR products into a 30 μl volume using the MinElute PCR Purification Kit. Finally, 400 ng of DNA library (as estimated by NanoDrop) were run in a precast E-Gel EX 2% agarose gel and the band corresponding to the library was extracted in 30 μl with the QIAquick Gel Extraction Kit. The libraries were sequenced in single-read runs using the NextSeq Sequencing System (Illumina) and the 150-cycles Mid Output kit in the Genomics Core Facility at the Pompeu Fabra University.

2.5 | Sequence processing

We used the STACKS 1.35 package (Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013) to process the sequences obtained. First, the *PROCESS_RADTAGS* program was used to filter out reads with low-quality sequences and to separate reads belonging to different samples according to the barcodes. This program was used with the recovery option, which corrects isolated errors in the restriction cut site sequence or in the barcode, and with different values for the quality score limit (s: from 10 to 30) depending on the overall quality of the library. Then reads from samples sequenced in different runs were combined (this set of quality-filtered reads are available in Dryad; see Data Accessibility section). After this step, *USTACKS* was used to assemble loci in each sample, with a minimum number of three sequences per locus for each sample (minimum stack depth or *m*) and a maximum of two differences between reads (*M*). The mean sequence coverage for each specimen was then calculated as the number of assembled reads divided by the number of assembled loci. The loci of all the specimens were subsequently merged with *CSTACKS*, allowing for a maximum of two mismatches between reads (*n*), and a catalogue of loci and sequences was created using *SSTACKS*. We then used the *POPULATIONS* program to save output files with different filter combinations available in the program: minimum proportion of called individuals (*r*), minimum stack depth or coverage (*m*) and minimum minor allele frequency (MAF). SNPs were saved in PLINK format and sequences of loci in FASTA format for further analyses. For the SNPs, only one SNP per locus was saved.

The quality of the library reads was verified by generating an initial data set with $r = 1$, $m = 9$ and MAF = 0 (data set 1 in Table S3). Additional SNP data sets were generated with different filters to analyse how they perform when estimating pairwise relatedness (see below).

2.6 | Relatedness estimation: selection of the best estimator

Estimation of pairwise relatedness among individuals was performed with the program *RELATED* (Pew, Muir, Wang, & Frasier, 2015), which is an R implementation of COANCESTRY (Wang, 2011). We first assessed the accuracy of the different relatedness estimators with simulations implemented in *RELATED*, using our first SNP data set (data set 1). This step was necessary because it has been shown that the performance of different estimators is highly dependent on the specific characteristics of the data set used (Blouin, 2003; Gonçalves da Silva & Russello, 2011; Russello & Amato, 2004; Van de Casteele, Galbusera, & Matthysen, 2001). Individual genotypes were simulated using the allele frequencies of the population to calculate the relationship between 250 dyads of each of the following relationships: parent-offspring, full-siblings, half-siblings and unrelated individuals. The relatedness coefficient was subsequently calculated for each dyad using five moment estimators: lynchli (Li et al., 1993), lynchrd (Lynch & Ritland, 1999), quellergt (Queller & Goodnight, 1989), ritland (Ritland, 1996) and wang (Wang, 2002); and two maximum-likelihood estimators: dyadml (Milligan, 2003) and trioml (Wang, 2007). We then evaluated the performance of the seven different estimators by calculating means and standard deviations from the estimates and correlating them against the expected values.

2.7 | Relatedness estimation: selection of the best SNP data set

Filters to generate SNPs with the *POPULATIONS* program (*r*, *m* and MAF) were optimized according to their best performance in relatedness-based individual identification using *RELATED* and the dyadml estimator, which was the best estimator for our data set (see Results). For this purpose we used four samples, of which we processed and sequenced two different DNA aliquots and analysed them independently. We then tested whether duplicated samples from the same individual reproduced the expected relatedness value of one when using a model with no inbreeding (Wang, 2015). To do so, we generated SNP data sets with different combinations of parameters in POPULATIONS and calculated relatedness values.

2.8 | Construction of relatedness networks and estimation of individual inbreeding coefficients

Pairwise relatedness among the 37 Pyrenean desmans was then calculated using the optimal estimator of RELATED (dyadml) and the optimal data set (912 SNPs; data set 2) that we found, using all specimens for allele frequency estimation. We selected in RELATED the option that takes inbreeding into account (the full nine-states model) and therefore also estimated the inbreeding coefficient for each specimen. The use of this model was justified because our preliminary relatedness analyses revealed that inbreeding was high in the studied population and we had in principle ample marker information for this more complex model (Wang, 2007). Confidence intervals (95%) for the estimation of relatedness were calculated using bootstrapping over loci (100 replicates). To avoid false positives that may alter dispersal patterns, we removed the smallest relatedness values. Thus, we only considered relatedness estimates where the lower 95% confidence limit of the bootstrap replicas was higher than 0.

To visualize the relationship between individuals in space, we plotted a network of relationships with the program GEPHI (Bastian, Heymann, & Jacomy, 2009) using individuals as nodes and relatedness values as edge thicknesses. We represented nodes according to the geographic location of the individuals with the plug-in GeoLayout and then superimposed the network on a map.

2.9 | Simulations along pedigrees of pairwise relatedness and inbreeding coefficients

To test the reliability of the relatedness coefficients obtained, we performed simulations along artificial pedigrees in which individuals with different origins were computationally crossed. In 12 of the pedigrees all founders were from the same river (Fig. S1) and in 12 additional pedigrees we included a "migrant," that is, an individual from a different river (Fig. S2). Some pedigrees had four founders and others had five to be able to recreate different kinship categories of interest. Simulations of offspring for the different crosses were performed with the custom Perl script GetCrosses.pl (available in Dryad), which randomly selects one allele for each locus from each parent to generate the alleles of a new individual. The new individuals generated in this manner were then added to the output file. Using the GetCrosses.pl and RELATED programs, we conducted 100 simulations for each pedigree and estimated the relatedness values for all relationships within the artificial pedigree, which were then compared with the expected values.

In the simulations along pedigrees with migrants, we also determined if second and third generation migrants can be detected by counting all relationships between the generated offspring (F101, F102, F201 and F202) and the individuals of the river from which the migrant came from, after excluding this migrant. We counted for all simulated pedigrees relatedness values above 0.0625 (corresponding to the lowest value detected in the real data set after bootstrap; see Results) and computed the average number of inter-river relationships detected per pedigree. The expected average number of relationships depends on the strength of the relationships of the migrant with the individuals of its river of origin and therefore it can be different for each simulated pedigree.

Using the pedigrees with migrants, we simulated additional crosses between the generated offspring to obtain inbred individuals of known ancestry and tested the performance of our SNP data set to estimate their individual inbreeding coefficients. These individuals were obtained from crossing full-siblings, half-siblings, first cousins and half-first cousins (Fig. S3).

2.10 | Proportion of heterozygous positions in each specimen and Hardy-Weinberg equilibrium test

The proportion of heterozygous positions for each individual should be estimated from all loci, including both variable and invariable loci, to be comparable with genome-wide estimates (Prado-Martinez et al., 2013; Robinson et al., 2016). In addition, a MAF filter should not be used to avoid altering this estimation. We therefore generated a new data set with the POPULATIONS program using filters $r = 1$, $m = 12$ and MAF = 0, which resulted in 7,583 total loci (all loci of data set 3; Table S3). The proportion of heterozygous positions was estimated from these sequences in FASTA format as the number of all heterozygous positions of the specimen divided by the total length of the loci.

Deviations from Hardy-Weinberg equilibrium were assessed using the exact test as well as the heterozygote deficiency and excess tests implemented in GENEPOP version 4.6 (Rousset, 2008), using rivers as populations.

2.11 | Genomic tree

We constructed a distance phylogenetic tree from the matrix of the average genomic divergence between specimens, as in Querejeta et al. (2016), using the 1,262 variable loci generated with the same filters as above: $r = 1$, $m = 12$ and MAF = 0 (variable loci of data set 3; Table S3). Because nuclear genomes are diploid, a method is needed to summarize the divergence of the two alleles per individual in the calculation of the distance matrix. For this purpose, we calculated pairwise distances between all specimens using the formula 8.2 taken from Freedman et al. (2014). Basically, for each variable position being compared between two individuals, the average of the four possible matches between the two individuals is computed. The resulting pairwise distance matrix was used to construct a distance tree using the FITCH program in the PHYLIP package (Felsenstein, 1989).

2.12 | Principal component analysis

Principal component analysis (PCA) was performed with the R program SNPRELATE (Zheng et al., 2012) and the genetic covariance matrix of data set 2 (Table S3). The axes of the plot were orientated to maximize the positional coincidence between the specimens in the plot and their geographical locations.

2.13 | Population structure

Population structure and admixture proportions were estimated from the SNPs of data set 2 (Table S3) with the program STRUCTURE 2.3.4, which implements a Bayesian model-based clustering method (Pritchard, Stephens, & Donnelly, 2000) using the admixture and correlated allele frequency models, and with no prior information about population origin. A total of 500,000 iterations were run after a burn-in of 50,000 iterations and with a number of clusters (*K*) ranging from 1 to 10. We performed 10 different runs for each *K* value to plot the trend of the estimated posterior probability for the data, Ln P(D) (Pritchard et al., 2000). In addition, the optimal *K* value was assessed using the Δ*K* method (Evanno, Regnaut, & Goudet, 2005), as implemented in STRUCTURE HARVESTER (Earl & vonHoldt, 2012).

3 | RESULTS

3.1 | Phylogeography of the Pyrenean desman in the contact zone of two ancient mitochondrial lineages

Cytochrome *b* sequences from 66 specimens of Pyrenean desman (Tables S1 and S2) were obtained from different localities in the Iberian Range. We constructed a maximum-likelihood phylogenetic tree from these sequences to assign each specimen to its mitochondrial clade (not shown). Figure 1a shows an overview of the clades distribution in the whole range according to Igea et al. (2013) and Figure 1b the map of the specimens used in this work coloured by clade. This map revealed the presence of a spatial separation between clades B1, restricted to the Oja and Najerilla rivers, and A2, mainly found in the Iregua, Tera and Duero rivers, and some areas of the Najerilla. Three different mitochondrial clades were found in the Umbría river (A2, B1 and B2). Individuals of clade B2 had not been detected in the Iberian Range so far (Igea et al., 2013). Interestingly, females from the Iregua river (clade A2) appeared to have migrated to Najerilla and Umbría, but no B1 female was found in Iregua, indicating a certain element of asymmetric past migration.

Of these specimens, 37 present in four rivers (Iregua, Najerilla, Oja and Umbría) were used for library construction and genomic analysis. The sex of each of these specimens was determined by qPCR, resulting in 20 females and 17 males (Figure 1b and Table S1).

3.2 | Sequence assembly and filtering parameters

A total of 183,708,284 Illumina reads passed the initial quality filters, with an average of 4,965,089 reads for each of the 37 *G. pyrenaicus* specimens analysed and a mean depth of coverage of assembled loci of 42.5 (Table S4). It should be taken into account that this coverage includes possible PCR duplicates, which cannot be detected in standard ddRAD libraries, and therefore real mean coverage may be lower. An initial data set was generated with filters $r = 1$, $m = 9$ and MAF = 0 in the POPULATIONS program, resulting in 1,651 SNPs (data set 1; Table S3).

We were interested in optimizing these filters to generate a data set that would have an optimal performance in relatedness analysis. Therefore, we first needed to assess the statistical properties of different relatedness estimators. Using our initial SNP data set, we evaluated the performance of different estimators through simulations based on the allele frequencies. We obtained the distribution of relatedness values for different kinship relationships (Fig. S4) and calculated their means and standard deviations (Table S5). Although the means were close to the expected values for most estimators, the dyadml (Milligan, 2003) and trioml (Wang, 2007) maximum-likelihood estimators showed the lowest standard deviations. In addition, the maximum-likelihood estimations presented the best correlations with the expected values ($R = 0.994$ in both cases; Table S6). Finally, relatedness values between unrelated individuals were mostly zero or close to zero for the maximum-likelihood estimations (Fig. S4), indicating a low level of false positives for them. Therefore, both maximum-likelihood estimators showed the best overall properties for our data set; we chose the dyadml estimator for subsequent analysis as it required less computational time than the trioml estimator.

We then proceeded to optimize the filters of POPULATIONS to generate the SNP data set. For this purpose, we used four pairs of duplicated samples, for which a relatedness value of 1 is expected. Using data sets generated with different filters and estimating relatedness from them with dyadml (Fig. S5), we found that filters $r = 1$ and MAF = 0.05 (solid blue line in Fig. S5) gave the best overall relatedness values for values of minimum coverage of 12 and 15. We chose $m = 12$ because it yielded the highest number of SNPs. In this way, we obtained 912 SNPs (data set 2 of Table S3), which were used for the final relatedness estimations and other SNP-based analyses. Using these parameters, we found an average discrepancy in genotypes of the four replicated samples of 0.33%. As *r* was set to one, all samples were genotyped for all SNPs and therefore there was no missing data. Additionally, the FASTA sequences with the same filters except that they were unfiltered for MAF (data set 3) were used for genomic tree construction and heterozygosity estimation.

Hardy-Weinberger equilibrium was tested for data set 2. We found 134 SNPs (15%) with significant deviations in at least one of the four sampled populations (rivers), mostly due to heterozygote deficiency.

3.3 | Genomic tree, PCA and STRUCTURE

The genomic tree of the specimens revealed the presence of four groups comprising individuals from the same river, although the individual IBE-C3745 from the Najerilla was grouped with those from the Oja (Figure 2a). Interestingly, all individuals from the Umbría river formed a single genomic group despite having different

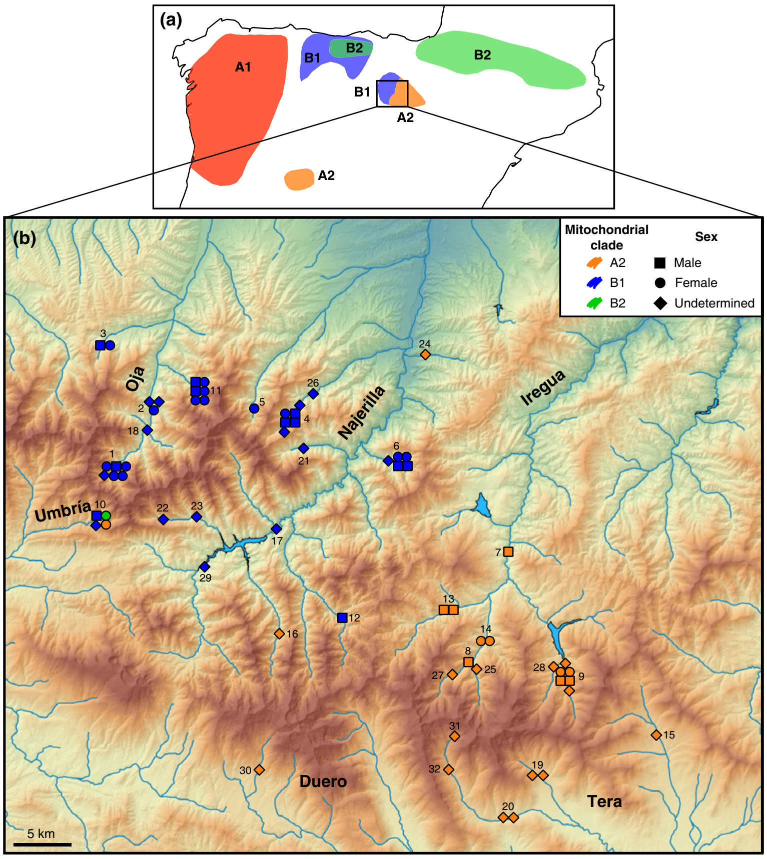

FIGURE 1 Map of the northern part of the Iberian Peninsula showing the distribution of the main mitochondrial clades found in Igea et al. (2013) (a) and map of Pyrenean desman specimens of the Iberian Range used in this study (b). Different colours reflect the mitochondrial clade to which they belong (A1: red; A2: orange; B1: blue; B2: green). Samples used only for cytochrome *b* sequencing are shown with a diamond and samples used for ddRAD sequencing are shown with squares (males) or circles (females). Localities are indicated with number as follows: (1) Oja Cabecera, (2) Oja Azarrulla, (3) Ciloria, (4) Tobía, (5) Cárdenas, (6) Roñas, (7) Iregua, (8) Iregua-Cabecera, (9) La Vieja, (10) La Soledad, (11) Urdanta, (12) Ormázal, (13) Mayor, (14) Iregua Achichuelo, (15) Barriomartín, (16) Portilla-Collado Grande, (17) Calamantío, (18) Oja Altuzarra, (19) Razoncillo, (20) Razón-Sotillo del Rincón, (21) Valvanera, (22) Gatón Cabecera, (23) Gatón, (24) Pedroso, (25) Arroyo de Puente Ra, (26) Tobía-Las Minas, (27) Villoslada de Cameros, (28) Lumbreras, (29) Barranco de la Sabandija, (30) Duero Cabecera, (31) Razón, and (32) Razón Cabecera. See also Tables S1 and S2 of Supplementary Information [Colour figure can be viewed at wileyonlinelibrary.com]

mitochondrial clade origins. Additionally, the 11 specimens of mitochondrial clade A1 belonged to two different genomic groups (10 to the Iregua group and one to the Umbría group). PCA basically showed the same results, grouping most individuals according to their river of origin and not their mitochondrial clade (Figure 2b).

The Structure analysis showed good convergence of the different runs for K values between 1 and 6 (Fig. S6). We then used the Evanno method (Evanno et al., 2005) to detect the point of inflection of the likelihood curve. According to this method there were two peaks (Fig. S6). The main one was at $K = 2$, but this peak is far from the saturation point of the likelihood curve whereas the peak at $K = 4$ is closer to this point and may correspond to a better model. Additionally, when we removed 134 SNPs that were not in Hardy-Weinberg equilibrium, an additional peak appeared at $K = 6$ (not shown). All this made it difficult selecting a single K and therefore we opted for considering several K values to show the hierarchical nature of structure (Betto-Colliard, Sermier, Litvinchuk, Perrin, & Stöck, 2015). Figure 3 shows the admixture proportions from $K = 2$ to $K = 6$ (these proportions remained virtually identical after removing SNPs that were not in Hardy-Weinberg equilibrium). With $K = 2$, there is a subdivision basically between specimens of the Iregua and Najerilla on one side and those of the Oja and Umbría rivers on the other, with the exception of one individual (see below). With three clusters, the Iregua and Najerilla rivers become separated. With four, five and six clusters, new subpopulations appear in the Oja and Umbría rivers, with the Umbría river basically becoming a single population in the models with five and six clusters. As previously noted in the genomic tree and PCA, the individual IBE-C3745, found in the river Najerilla, had the same genomic composition as some individuals in the Oja under all K values. Additionally, as seen from the mitochondrial data, certain asymmetric past migration can be appreciated, particularly in the models with 3 and 4 clusters: Iregua genome components were present in the river Najerilla and Najerilla components in the rivers Oja and Umbría, but no admixture was observed in the opposite direction.

3.4 | Pairwise relatedness and relatedness networks

We found 160 relatedness values between individuals for which the lower 95% confidence limit of the bootstrap replicas was higher than 0 and used them to construct relatedness networks (Figure 4). The minimum pairwise relatedness was 0.0625, the maximum 1.1866 and the average 0.3564. To simplify the networks, we subdivided these relationships into those with a relatedness value above 0.2018 (close kinship relationships) and those below this value (distant kinship relationships). This limit corresponds to the lower 95%

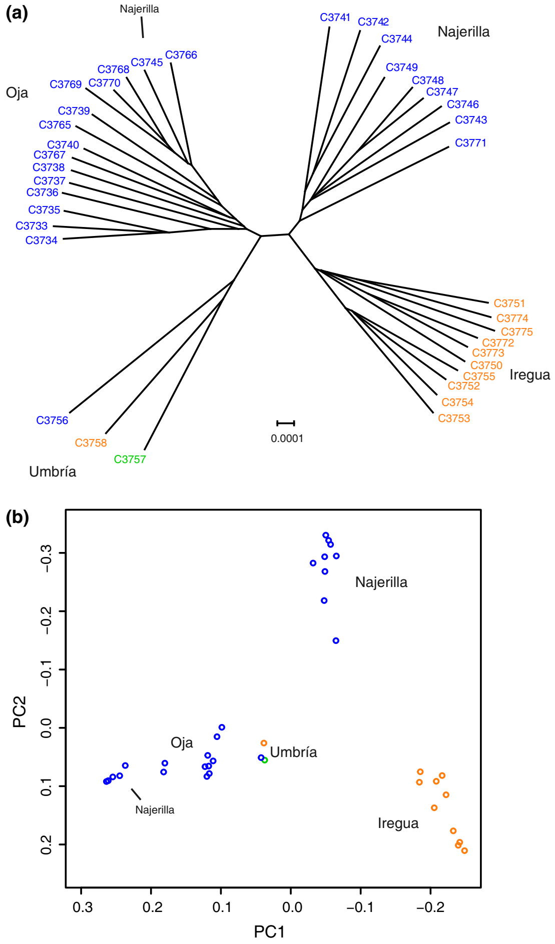

FIGURE 2 Maximum-likelihood phylogenetic tree (a) and principal component analysis (b), with specimens of Pyrenean desman colour-coded by mitochondrial clade as in Figure 1 (A2: orange; B1: blue; B2: green). The scale of the tree is in substitutions per position [Colour figure can be viewed at wileyonlinelibrary.com]

confidence limit of the simulations of half-siblings with the dyadml estimator (Fig. S4). Therefore, the network of close kinship covers relationships equivalent to half-siblings (second degree) and above. The minimum relatedness value considered (0.0625) corresponds to the equivalent of a fourth-degree relationship, and therefore both networks together provide a picture of dispersal occurred in the last four generations at most. The network of close kinship relationships (Fig. 4A) showed that most relationships were between individuals from the same river. In total, we found 112 intra-river relationships. In comparison with these, there were only 7 inter-river relationships. All of them were due to the individual IBE-C3745, probably a first-generation migrant from the Oja river to the Najerilla as deduced from several relatives found in the former. The network of more distant kinship relationships (Figure 4b) also showed an overall pattern of mainly intra-river relationships, of which we found 36. As for the inter-river relationships, apart from three more relationships between the Oja and Najerilla, two additional relationships were discovered between the Umbría and Najerilla rivers. Relatedness networks showed a similar pattern when the estimation of pairwise relatedness was performed with a model without inbreeding (not shown).

To study possible male or female philopatry, we analysed the relatedness values between pairs from the same locality and classified them by sex. We computed a mean relatedness of 0.56 for female dyads (n = 16), 0.53 for male dyads (n = 7) and 0.55 for mixed dyads (n = 26). These values reflected a high level of kinship of the individuals of the same locality (with the mean equivalent to full-siblings) but we did not find significant differences between the three analysed classes (Tukey–Kramer test; p = .05).

3.5 | Individual inbreeding and heterozygosity

Individual inbreeding coefficients were high overall (mean value of 0.33), with 33 individuals of 37 having values over 0.1, indicating a close relationship between their parents (Table S7). When these coefficients were plotted on the map (Figure 5), we observed that the specimens with the highest values were those from the Iregua (mean value of 0.42), followed by Najerilla (0.38), Umbría (0.30) and Oja (0.23). Interestingly, four individuals from the river Iregua sampled upstream from a dam (in the tributary La Vieja) presented the highest values of all, with an average inbreeding coefficient of 0.53.

Heterozygosity rates ranged from 103 to 322 heterozygous positions per million bases (Table S7). These values and individual inbreeding coefficients presented a strong negative correlation (R = −0.921).

3.6 | Simulations along pedigrees using genotypes of actual specimens

The estimation of relatedness and inbreeding depends on the population used as reference for allele frequency estimation (Milligan, 2003). Ideally, one should use the closest possible population for these estimates, which in our case could be each river. Therefore, we tested whether the best reference for these estimations was the whole set of specimens or each of the three main rivers separately. For this purpose, we performed a different type of simulation that used the actual genotypes from our data set and crossed them computationally along specific artificial pedigrees with founders of each river (Fig. S1). When we used the whole set as reference, most relationships of the pedigree were detected (Table S8). Relatedness values were higher than theoretical ones for outbred individuals due to kinship between founders, leading to inbred offspring; these relationships could not be avoided as most desmans from each river were related. However, when each river was analysed separately, many relationships of the known pedigrees were zero, leading to means that were close to zero for most relationships (Table S8). These results may be due to the low sample size within each river, showing that in this case the best reference was the set of all specimens.

Relationships among individuals from different rivers were scarce according to our relatedness analyses. We wanted to assess if these relationships were especially difficult to detect due to the presence of genetic structure, that is, to the comparison of specimens with different genomic compositions. To test this, we used artificial pedigrees in which a migrant from an adjacent river was added to the founders of the river of interest (Fig. S2). When we estimated relatedness from the dyads in the pedigrees, basically all relationships were detected (Table S9, Fig. S7), indicating that relatedness can be estimated using pairs of individuals with different genomic compositions. The estimated means were closer to the expected ones than in the previous simulations as average kinship relationships between the founders were lower due to the presence of a migrant (Table S9). Significantly, the descendants from the founders were found to be related to the individuals of the river from which the migrant came (Table S10), indicating that inter-river dispersal occurred in the last few generations, if it existed, can be detected with this approach.

We also tested the accuracy of the estimated inbreeding coefficients using artificial pedigrees with migrants (Fig. S3). As expected for crossings of parents from different rivers, their simulated offspring showed basically no inbreeding (Table S11). When we computationally crossed relatives of different degree from the pedigrees, we obtained inbreeding coefficients for the offspring that were close to the expected values (Table S11, Fig. S8), suggesting that the SNPs we used were valid for these estimations.

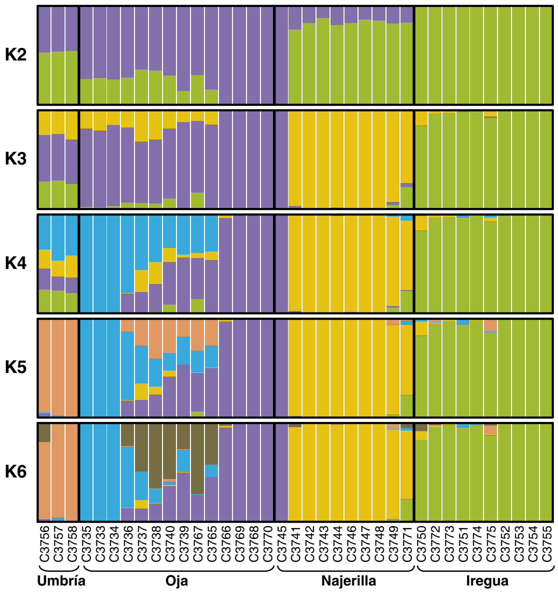

FIGURE 3 Bar plot of admixture proportions of each Pyrenean desman specimen as determined with STRUCTURE and different *K* values [Colour figure can be viewed at wileyonlinelibrary.com]

4 | DISCUSSION

We have explored the use of SNP-based relatedness estimates for Pyrenean desmans to infer recent dispersal phenomena. Several important points have emerged from this work. First, the SNP data set used allowed us to obtain robust relatedness estimates, as evidenced by simulations of kinship relationships along known pedigrees. Second, we proposed the use of relatedness network graphs to visualize dispersal patterns; these networks clearly showed that desmans frequently disperse within rivers and rarely between different rivers. Third, we detected through relatedness analysis an excessive level of inbreeding in most individuals, probably as a consequence of limited dispersal in this population. And finally, the dispersal patterns uncovered help explain the low levels of mixing between lineages that takes place in the Iberian Range contact zone.

4.1 | Performance of relatedness estimators using SNPs

Of all available relatedness estimators, those based on maximum likelihood gave the best results for our data set according to simulations based on allele frequencies. Furthermore, the correlation between estimated and expected values for simulated dyads was much better than those reported for microsatellites (Taylor, 2015). However, given the uncertainties that still have to be resolved about the estimation of relatedness, we did not use specific relatedness categories in this study. The high inbreeding observed in our samples would have made it even more difficult to determine these categories. Instead, we used relatedness networks to summarize the data and visualize the general pattern of relationships among our sampling sites (Figure 4).

Another problem we faced derived from the presence of genetic structure in our data, meaning that the estimation of relatedness values was based on allele frequencies of individuals with different genomic compositions (Anderson & Weir, 2007; Thornton et al., 2012). To test if the genetic structure was strong enough to affect relatedness estimations, we constructed artificial pedigrees with individuals from different rivers and simulated their offspring. We showed that almost all inter-river and intra-river relationships could be detected in these pedigrees (Table S9, Fig. S7). In addition, relationships between the migrant's descendants and individuals belonging to the source population were also detectable (Table S10), demonstrating that migration events between populations occurring in the last few generations can be identified.

A different strategy that could overcome potential problems caused by population structure would be to perform separate, more homogenous analyses for each population or river. However, our simulations showed that the relatively low sample size used as a background in each river meant that many relationships went undetected, indicating that the entire data set provided a better baseline in our case. Furthermore, this approach would have

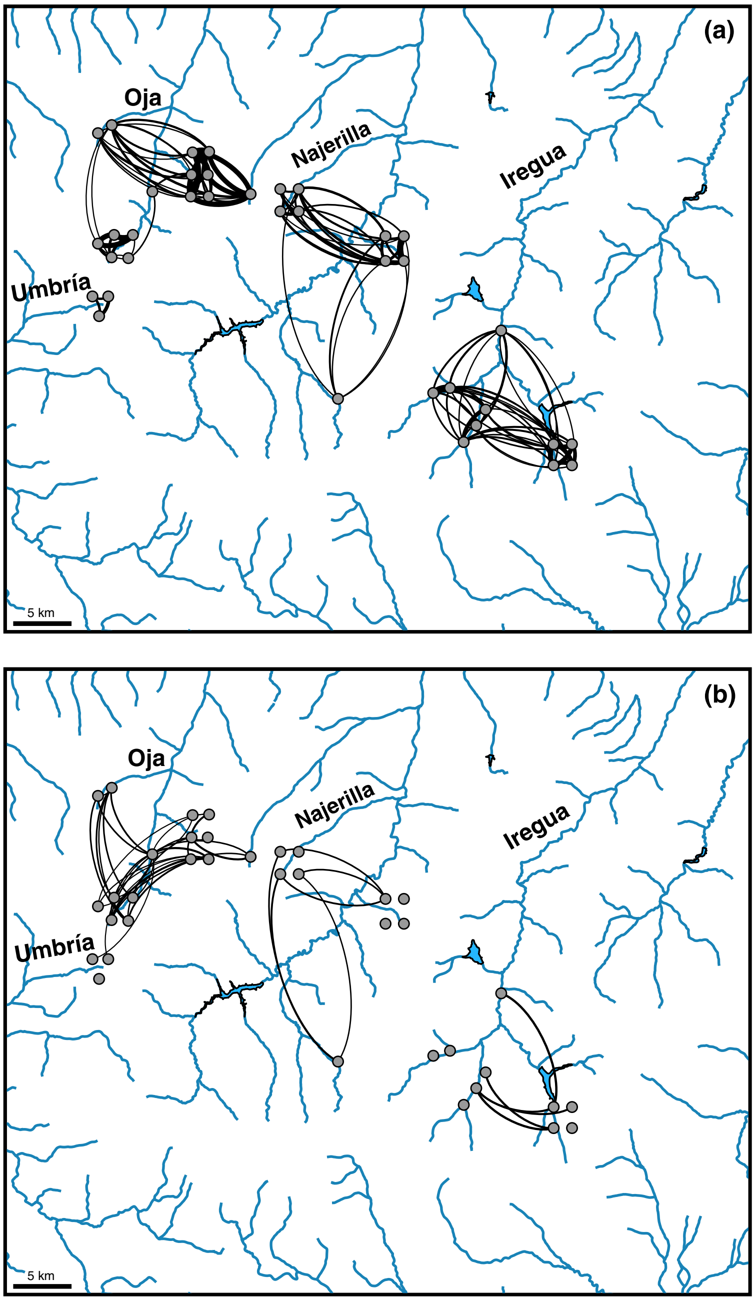

FIGURE 4 Map plotting relatedness networks with values above 0.2018 (a) and under this value (b). Curved lines (edges) connect Pyrenean desman specimens (nodes) for which a relationship was found. The thickness of edges is proportional to the relatedness value of the connected specimens [Colour figure can be viewed at wileyonlinelibrary.com]

precluded the study of inter-river dispersal. Finally, even if the existence of population structure or the background used had altered some of the lowest relatedness values, then the networks that only used relatedness values above 0.2018 (Figure 4a) should be robust enough to perform an analysis of the dispersal patterns in the Pyrenean desman.

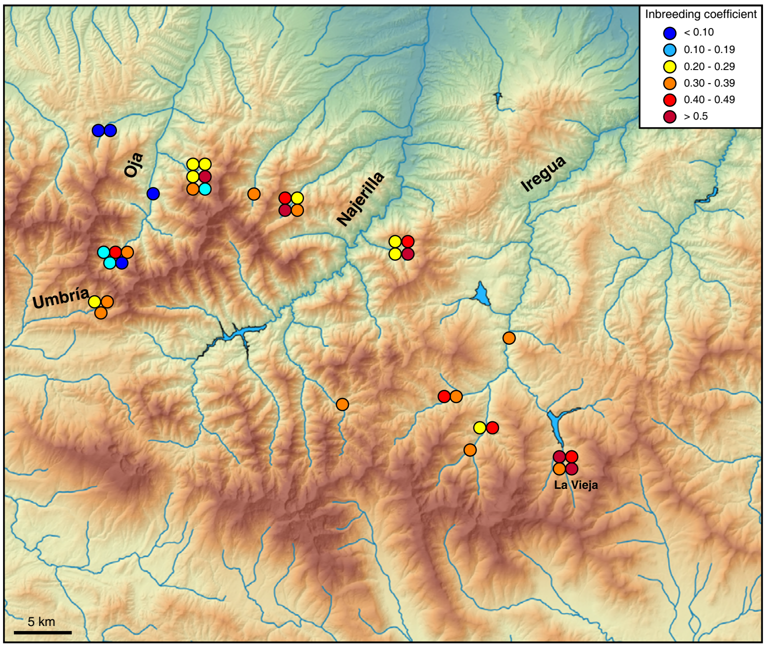

FIGURE 5 Map plotting colour-coded inbreeding coefficients of different specimens of Pyrenean desman [Colour figure can be viewed at wileyonlinelibrary.com]

4.2 | Contemporary dispersal patterns in the Pyrenean desman visualized with relatedness networks: inter- and intra-river dispersal

We used networks that connected relatives to obtain a clear visualization of dispersal patterns. These networks first showed that most of the relationships corresponded to desmans sampled from the same locality or different localities in the same river (148 relationships; Figure 4). The relatedness values between individuals sampled in the same locality were particularly high, with many values corresponding to the equivalent of siblings or half-siblings. Dyads of both males and females showed high relatedness values. Therefore these results can be explained by the existence of close family groups due to a strong philopatry of both males and females. Nevertheless, sampling in each locality was scarce and consequently the number of pairwise comparisons in the same locality was small, so we could not determine whether this trend was more predominant in one of the sexes. Data from radio-tracking studies into this species in a specific river of the eastern Pyrenees did not reveal any differences in movements between males and females (Melero et al., 2014), but it is still unknown whether they exhibit different long-range movements. Therefore, further studies are necessary to better understand the philopatric behaviour of the Pyrenean desman.

Our analyses also revealed some long distance relationships within rivers. For example, in the river Najerilla, the two most distant sampling localities with relatives that suggest a connection between them (Ormázal and Tobía), are separated by approximately 50 km of watercourse (Figure 4). It is important to note that this approach cannot determine the route of a dispersal event that must have occurred to give rise to a relationship between different localities. Another possibility is that individuals from an unknown locality independently dispersed to the two connected localities. Notwithstanding, when there are two localities with relatives we can infer they are genetically interconnected by some route and that at least one migration event must have taken place through this route in the last few generations.

The river system we studied included three rivers running in parallel (Oja, Najerilla and Iregua) and a fourth river, Umbría, whose headwaters are near those of the Oja. Importantly, the relatedness networks revealed that, despite the proximity of some sampling localities in adjacent rivers, there were few relationships between the four rivers (12 relationships; Figure 4). The few inter-river relationships we did observe connected Najerilla with Oja and Oja with Umbría, while no connection was observed between the adjacent Najerilla and Iregua rivers. It is therefore clear that Pyrenean desmans do not frequently move between different rivers. Connectivity between rivers further downstream is unlikely due to suboptimal conditions for the Pyrenean desman as rivers become larger (Palmeirim & Hoffmann, 1983). Consequently, it is likely that the few connections observed between rivers took place overland and across watershed divides. Further studies into the habitat suitability of the overland corridors will be vital in understanding why some accesses are more permeable than others.

In conclusion, the Pyrenean desman showed a low level of contemporary inter-river dispersal compared to intra-river dispersal, according to the relatedness networks. These results can be compared with those from the genomic tree, PCA and STRUCTURE analysis. These methods generally provide information about more ancient migrations but, interestingly, they can also detect some recent dispersal events. The most obvious case is that of individual IBE-C3745, which was sampled in the river Najerilla but grouped with desmans from the Oja in the three analyses, thus indicating a recent dispersal event from the Oja to the Najerilla. While the genomic tree and PCA cannot be used to infer additional recent events, earlier movements between rivers can be deduced from the Structure analysis (see section "Population history of the Pyrenean desman in the contact zone of the Iberian Range"). Crucially, these methods can only detect dispersal events from differentiated populations, which are mostly inter-river movements in the present study, whereas the same methods would not detect intra-river dispersal events. By contrast, relatedness networks provide an overview of recent dispersal events both between and within rivers.

4.3 | High inbreeding in the Pyrenean desman

The inbreeding coefficient was exceedingly high in most individuals analysed (Table S7), suggesting that many of them are the product of several generations of mating between close relatives. Similar values of individual inbreeding can only be found in highly inbred species, such as Przewalski's horse (Liu et al., 2014), or critically endangered species, such as Attwater's Prairie-chicken (Hammerly et al., 2013). These high inbreeding values for the Pyrenean desman may partly be due to the strong philopatric behaviour of both males and females. In addition, the lack of connectivity between rivers revealed here can only worsen the inbreeding situation in this area. Finally, the low species density in some rivers (Nores et al., 1998) may facilitate that juveniles occupy new territories close to their natal site, thus increasing the chances of inbreeding (Lambin, 1994; Matthysen, 2005).

In agreement with the high inbreeding coefficients, we found low heterozygosity levels in all individuals (Table S7). These low heterozygosity values explain why most SNPs that deviated from the Hardy-Weinberger equilibrium presented heterozygote deficiency. The heterozygosity observed in the Iberian Range (between 103 and 322 heterozygous positions per million bases) was intermediate compared with that of Pyrenean desmans from the whole distribution area, which ranged between 13 and 488 (Querejeta et al., 2016). In any case, these values are among the lowest recorded to date for animals (Prado-Martinez et al., 2013; Querejeta et al., 2016; Robinson et al., 2016).

Interestingly, the highest inbreeding values were found in individuals from the small tributary La Vieja, which is situated upstream of a dam (Figure 5). It is tempting to speculate that this additional artificial barrier to dispersal may have been responsible for the increase in inbreeding. However, more specimens and data would be required to study the dispersal behaviour of the Pyrenean desman in the presence of artificial barriers and to learn whether these obstacles exacerbate inbreeding.

Nevertheless, inbreeding is not necessarily detrimental in itself. Further studies should therefore be performed to determine whether inbreeding depression affects the viability of these populations due to the presence of recessive alleles in homozygous form (Charlesworth & Willis, 2009; Hedrick & Garcia-Dorado, 2016; Kardos, Taylor, Ellegren, Luikart, & Allendorf, 2016). The Pyrenean desman may have presented fragmented populations in some rivers under natural conditions, for example, due to difficult dispersal downstream in large rivers (Palmeirim & Hoffmann, 1983). Theoretically, it then may be argued that this species is less susceptible to inbreeding depression because deleterious alleles were purged during previous inbreeding situations in isolated populations (Keller & Waller, 2002; Leberg & Firmin, 2008). However, for many species past inbreeding has been shown to have little effect when it comes to purging genetic load (Ballou, 1997; Crnokrak & Barrett, 2002). Clearly our data highlight the need to compare fitness traits, such as brood size or first-year survival rate, in individuals with different inbreeding levels to assess the intensity of inbreeding depression (Charlesworth & Willis, 2009; Hedrick & Garcia-Dorado, 2016; Kardos et al., 2016). It would also be of interest to study the mating system of this species and statistically test whether there is inbreeding avoidance through kin recognition or, on the contrary, there is random mating as observed in other species (Rioux-Paquette, Festa-Bianchet, & Coltman, 2010; Szulkin, Stopher, Pemberton, & Reid, 2013). So far, no data regarding any of these crucial aspects is available for the Pyrenean desman.

4.4 | Population history of the Pyrenean desman in the contact zone of the Iberian Range

One of the most interesting aspects that emerged from the initial genetic studies of the Iberian desman was the discovery of two contact zones for the main mitochondrial lineages, with little spatial mixing between them after the postglacial colonization (Igea et al., 2013). Similar contact zones have been found in other species, although rarely with such a strict delimitation (Gomez & Lunt, 2007). The large sample taken from the Iberian Range contact zone in the present work allowed us to reconstruct the evolutionary history of the Iberian desman in greater detail than previous studies. Thus, we could better determine the distribution of the mitochondrial clades, with A located mainly in the southeast of the Iberian Range (rivers Iregua, Tera and Duero) and B in the northwest (rivers Oja, Najerilla and Umbría). Moreover, genomic analysis revealed the existence of a strong genetic structure correlated with geographic origin. According to the STRUCTURE models with more than two populations (Figure 3), the area occupied by individuals with mitochondrial clade A corresponded to one of the genomic clusters while the area of clade B was subdivided into several additional clusters. These data also indicated that the populations of the main rivers studied here, the Iregua, Najerilla, Oja and Umbría, were differentiated at the genome level. Additionally, both the spatial distribution of mitochondrial clades and the admixture levels estimated for the different individuals suggested the existence of certain levels of gene flow between the detected populations. Interestingly, this migration seems to have occurred more often from the southeast to the northwest than vice versa. This means that the Iberian Range contact zone is asymmetric, as observed for other species (Johnson, White, Phillips, & Zamudio, 2015). As a likely consequence of this, the Oja river presents a higher admixture whereas the other rivers are more homogeneous (Figure 3). It is possible that specific access routes between rivers have been more permeable in one direction than in the other due to specific geographical constraints, but further data and more in-depth analyses of relationship categories are necessary to shed light on these details.

The contemporary dispersal patterns of the Iberian desman revealed here through relatedness networks are congruent with the scenario outlined above. The low dispersal levels we detected between rivers are probably the cause of the slow spatial mixing of the two mitochondrial clades since postglacial colonization of the contact zone. This low dispersal would also explain the genomic differentiation detected between the populations of the different rivers. Finally, the similar philopatry behaviour that can be deduced for males and females agrees with the genetic differentiation observed for both the mitochondrial and nuclear genomes. In conclusion, the different molecular markers and methods used here, including the elucidation of the current dispersal patterns, contributed to a better understanding of the complex evolutionary history of the Pyrenean desman in this contact zone.

4.5 | Implications of the relatedness networks for conservation and future prospects

Much remains to be learned about the dispersal behaviour of the Iberian desman but the evidence uncovered in this work suggests that there are too few movements between rivers in the Iberian Range. This information, along with the observation of high inbreeding levels in some rivers, points to the existence of a previously unidentified conservation problem for this endangered species. Although we cannot conclude from this work that fragmentation induced by dams exacerbates this problem, our data suggest that this may be an issue that deserves further attention, especially given the large number of dams present in rivers inhabited by the Pyrenean desman (Nores et al., 2007). Additionally, the endangered populations of the Central System are known to be highly fragmented (Gisbert & Garcia-Perea, 2014), and so their long-term survival could also be compromised by low dispersal between patches and inbreeding. Still, as already pointed out, it cannot be discarded that high inbreeding levels are tolerated by this species due to the nature of its fragmented habitat without creating inbreeding depression, another topic that needs to be addressed. In any case, it is known from studies into other species that the arrival of just a few specimens can greatly help to alleviate inbreeding problems (Akesson et al., 2016; Vilà et al., 2003). Therefore, simple actions designed to increase the natural connectivity between Pyrenean desman populations (e.g., by improving potential corridors between rivers or making artificial barriers more permeable) could be highly beneficial for the species' conservation. However, previous knowledge on the degree of connectivity between populations and their genetic health are necessary before proceeding to these or other management steps. The relatedness networks we propose here to study dispersal phenomena can then become a fundamental tool to evaluate the populations and monitor the effectiveness of these actions. Such measures, if effective, would lead with time to denser relatedness connections between populations, thus helping to reduce individual inbreeding. Further studies will be necessary in different areas of the distribution range of the Pyrenean desman, and with different landscape features, to gain a greater overall understanding of the conservation problems related to dispersal patterns.

ACKNOWLEDGEMENTS

We acknowledge the La Rioja Regional Government (Spain) for permits to use samples obtained during field works to study the Pyrenean desman. In addition, the Junta de Castilla y León facilitated additional samples of faeces obtained in the LIFE+ project MedWetRivers. This work was financially supported by research Project CGL2014-53968-P of the "Plan Nacional I+D+I del Ministerio de Economía y Competitividad" (Spain), cofinanced with FEDER funds, to J.C. Support was also provided by grant 2014-SGR-1532 of the "Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya." L.E. was supported by fellowship 2015FI_B 00429 of the Generalitat de Catalunya, cofinanced with FEDER funds.

DATA ACCESSIBILITY

New cytochrome *b* sequences obtained in this work have been deposited in the European Nucleotide Archive/GenBank nucleotide database under Accession nos LT746127–LT746170. Quality-filtered reads for each sample, final genotypes (SNPs from data sets 1 and 2) and the custom script GetCrosses.pl to generate the alleles of offspring are stored in the Dryad Digital Repository (https://doi.org/10.5061/dryad.4dv48). Additional data and figures may be found in Supporting information.

AUTHOR CONTRIBUTIONS

J.C. designed and supervised the study. L.E. performed the laboratory work and analysed the data. J.G.-E. and A.G. collected the samples. L.E. and J.C. wrote the manuscript. All authors discussed the results and contributed to the preparation of the manuscript.

REFERENCES

Akesson, M., Liberg, O., Sand, H., Wabakken, P., Bensch, S., & Flagstad, Ø. (2016). Genetic rescue in a severely inbred wolf population. *Molecular Ecology*, *25*, 4745–4756.

Anderson, A. D., & Weir, B. S. (2007). A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. *Genetics*, *176*, 421–440.

Arora, N., Van Noordwijk, M. A., Ackermann, C., Willems, E. P., Nater, A., Greminger, M., Nietlisbach, P., ... Krützen, M. (2012). Parentage-based pedigree reconstruction reveals female matrilineal clusters and male-biased dispersal in nongregarious Asian great apes, the Bornean orang-utans (*Pongo pygmaeus*). *Molecular Ecology*, *21*, 3352–3362.

Baguette, M., Blanchet, S., Legrand, D., Stevens, V. M., & Turlure, C. (2012). Individual dispersal, landscape connectivity and ecological networks. *Biological Reviews of the Cambridge Philosophical Society*, *88*, 310–326.

Ballou, J. D. (1997). Ancestral inbreeding only minimally affects inbreeding depression in mammalian populations. *The Journal of Heredity*, *88*, 169–178.

Banks, S. C., & Lindenmayer, D. B. (2014). Inbreeding avoidance, patch isolation and matrix permeability influence dispersal and settlement choices by male agile antechinus in a fragmented landscape. *Journal of Animal Ecology*, *83*, 515–524.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*, pp. 1–2.

Betto-Colliard, C., Sermier, R., Litvinchuk, S., Perrin, N., & Stöck, M. (2015). Origin and genome evolution of polyploid green toads in Central Asia: Evidence from microsatellite markers. *Heredity*, *114*, 300–308.

Blouin, M. S. (2003). DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends in Ecology & Evolution*, *18*, 503–511.

Bonin, C. A., Goebel, M. E., O'Corry-Crowe, G. M., & Burton, R. S. (2012). Twins or not? Genetic analysis of putative twins in Antarctic fur seals, *Arctocephalus gazella*, on the South Shetland Islands. *Journal of Experimental Marine Biology and Ecology*, *412*, 13–19.

Catchen, J. M., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, *22*, 3124–3140.

Charbonnel, A., Buisson, L., Biffi, M., DAmico, F., Besnard, A., Aulagnier, S., ... Laffaille, P. (2015). Integrating hydrological features and genetically validated occurrence data in occupancy modelling of an endemic and endangered semi-aquatic mammal, *Galemys pyrenaicus*, in a Pyrenean catchment. *Biological Conservation*, *184*, 182–192.

Charlesworth, D., & Willis, J. H. (2009). The genetics of inbreeding depression. *Nature Reviews Genetics*, *10*, 783–796.

Crnokrak, P., & Barrett, S. C. H. (2002). Perspective: Purging the genetic load: A review of the experimental evidence. *Evolution*, *56*, 2347–2358.

Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, *28*, 2239–2252.

Earl, D. A., & vonHoldt, B. M. (2012). STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, *4*, 359–361.

Ellegren, H. (1999). Inbreeding and relatedness in Scandinavian grey wolves *Canis lupus*. *Hereditas*, *130*, 239–244.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology*, *14*, 2611–2620.

Felsenstein, J. (1989). PHYLIP-phylogeny inference package (version 3.4). *Cladistics*, *5*, 164–166.

Fernandes, M., Herrero, J., Aulagnier, S., & Amori, G. (2011). *Galemys pyrenaicus*. IUCN Red List of Threatened Species. Version 2011.2, pp. 1–3.

Freedman, A. H., Gronau, I., Schweizer, R. M., Ortega-Del Vecchyo, D., Han, E., Silva, P. M., ... Novembre, J. (2014). Genome sequencing highlights the dynamic early history of dogs. *PLoS Genetics*, *10*, e1004016.

Gillet, F., Le Roux, B., Blanc, F., Bodo, A., Fournier-Chambrillon, C., Fournier, P., ... Michaux, J. R. (2016). Genetic monitoring of the endangered Pyrenean desman (*Galemys pyrenaicus*) in the Aude River, France. *Belgian Journal of Zoology*, *146*, 44–52.

Gisbert, J., & Garcia-Perea, R. (2014). Historia de la regresión del desmán ibérico *Galemys pyrenaicus* (É. Geoffroy Saint-Hilaire, 1811) en el Sistema Central (Península Ibérica). In *Conservation and management of semi-aquatic mammals of South-western Europe. Munibe Monographs. Nature Series 3* (pp. 19–35). Aranzadi Society of Sciences, San Sebastian.

Gomez, A., & Lunt, D. H. (2007). Refugia within refugia: Patterns of phylogeographic concordance in the Iberian Peninsula. In *Phylogeography in Southern European Refugia* (pp. 155–188).

Gonçalves da Silva, A., & Russello, M. A. (2011). *iREL*: Software for implementing pairwise relatedness estimators and evaluating their performance. *Conservation Genetics Resources*, *3*, 69–71.

Hammerly, S. C., Morrow, M. E., & Johnson, J. A. (2013). A comparison of pedigree- and DNA-based measures for identifying inbreeding depression in the critically endangered Attwater's Prairie-chicken. *Molecular Ecology*, *22*, 5313–5328.

Hedrick, P. W., & Garcia-Dorado, A. (2016). Understanding inbreeding depression, purging, and genetic rescue. *Trends in Ecology & Evolution*, *31*, 940–952.

Igea, J., Aymerich, P., Fernández-González, A., González-Esteban, J., Gómez, A., Alonso, R., ... Castresana, J. (2013). Phylogeography and postglacial expansion of the endangered semi-aquatic mammal *Galemys pyrenaicus*. *BMC Evolutionary Biology*, *13*, 115.

Jacquard, A. (1972). Genetic information given by a relative. *Biometrics*, *28*, 1101–1114.

Johnson, B. B., White, T. A., Phillips, C. A., & Zamudio, K. R. (2015). Asymmetric introgression in a spotted salamander hybrid zone. *The Journal of Heredity*, *106*, 608–617.

Kardos, M., Taylor, H. R., Ellegren, H., Luikart, G., & Allendorf, F. W. (2016). Genomics advances the study of inbreeding depression in the wild. *Evolutionary Applications*, *9*, 1205–1218.

Keller, L. F., & Waller, D. M. (2002). Inbreeding effects in wild populations. *Trends in Ecology & Evolution*, *17*, 230–241.

Lambin, X. (1994). Natal philopatry, competition for resources, and inbreeding avoidance in Townsend's voles (*Microtus townsendii*). *Ecology*, *75*, 224–235.

Leberg, P. L., & Firmin, B. D. (2008). Role of inbreeding depression and purging in captive breeding and restoration programmes. *Molecular Ecology*, *17*, 334–343.

Li, C. C., Weeks, D. E., & Chakravarti, A. (1993). Similarity of DNA fingerprints due to chance and relatedness. *Human Heredity*, *43*, 45–52.

Liu, G., Shafer, A. B. A., Zimmermann, W., Hu, D., Wang, W., Chu, H., ... Zhao, C. (2014). Evaluating the reintroduction project of Przewalski's horse in China using genetic and pedigree data. *Biological Conservation*, *171*, 288–298.

Lopes, M. S., Silva, F. F., Harlizius, B., Duijvesteijn, N., Lopes, P. S., Guimarães, S. E., & Knol, E. F. (2013). Improved estimation of inbreeding and kinship in pigs using optimized SNP panels. *BMC Genetics*, *14*, 92.

Lynch, M., & Ritland, K. (1999). Estimation of pairwise relatedness with molecular markers. *Genetics*, *152*, 1753–1766.

Matthysen, E. (2005). Density-dependent dispersal in birds and mammals. *Ecography*, *28*, 403–416.

Melero, Y., Aymerich, P., Luque-Larena, J. J., & Gosálbez, J. (2012). New insights into social and space use behaviour of the endangered Pyrenean desman (*Galemys pyrenaicus*). *European Journal of Wildlife Research*, *58*, 185–193.

Melero, Y., Aymerich, P., Santulli, G., & Gosálbez, J. (2014). Activity and space patterns of Pyrenean desman (*Galemys pyrenaicus*) suggest non-aggressive and non-territorial behaviour. *European Journal of Wildlife Research*, *60*, 707–715.

Milligan, B. G. (2003). Maximum-likelihood estimation of relatedness. *Genetics*, *163*, 1153–1167.

Mills, L. S. (2013). *Conservation of wildlife populations: Demography, genetics, and management*. Oxford: Wiley-Blackwell.

Nores, C., Ojeda, F., Ruano, A., Villate, I., González, J., Cano, J., & García, E. (1998). Estimating the population density of *Galemys pyrenaicus* in four Spanish rivers. *Journal of Zoology*, *246*, 454–457.

Nores, C., Queiroz, A. I., & Gisbert, J. (2007). *Galemys pyrenaicus*. *Atlas y libro rojo de los mamíferos terrestres de España*, pp. 92–98.

Økland, J.-M., Haaland, Ø. A., & Skaug, H. J. (2010). A method for defining management units based on genetically determined close relatives. *ICES Journal of Marine Science*, *67*, 551–558.

Palmeirim, J. M., & Hoffmann, R. S. (1983). *Galemys pyrenaicus*. *Mammalian Species*, *207*, 1–5.

Palsbøll, P. J., Zachariah Peery, M., & Bérubé, M. (2010). Detecting populations in the "ambiguous" zone: Kinship-based estimation of population structure at low genetic divergence. *Molecular Ecology Resources*, *10*, 797–805.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS ONE*, *7*, e37135.

Pew, J., Muir, P. H., Wang, J., & Frasier, T. R. (2015). related: An R package for analysing pairwise relatedness from codominant molecular markers. *Molecular Ecology Resources*, *15*, 557–561.

Pinho, C., & Hey, J. (2010). Divergence with gene flow: Models and data. *Annual Review of Ecology, Evolution, and Systematics*, *41*, 215–230.

Piry, S., Alapetite, A., Cornuet, J.-M., Paetkau, D., Baudouin, L., & Estoup, A. (2004). GENECLASS2: A software for genetic assignment and first-generation migrant detection. *The Journal of Heredity*, *95*, 536–539.

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., … Bonet, T. (2013). Great ape genetic diversity and population history. *Nature*, *499*, 471–475.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*, 945–959.

Queller, D. C., & Goodnight, K. F. (1989). Estimating relatedness using genetic markers. *Evolution*, *43*, 258–275.

Querejeta, M., González-Esteban, J., Gómez, A., Fernández-González, A., Aymerich, P., Gosálbez, J., … Castresana, J. (2016). Genomic diversity and geographical structure of the Pyrenean desman. *Conservation Genetics*, *17*, 1333–1344.

Rioux-Paquette, E., Festa-Bianchet, M., & Coltman, D. W. (2010). Animal behaviour. *Animal Behaviour*, *80*, 865–871.

Ritland, K. (1996). Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetical Research*, *67*, 175–185.

Robinson, J. A., Ortega-Del Vecchyo, D., Fan, Z., Kim, B. Y., Vonholdt, B. M., Marsden, C. D., … Wayne, R. K. (2016). Genomic flatlining in the endangered island fox. *Current Biology*, *26*, 1183–1189.

Rollins, L. A., Browning, L. E., Holleley, C. E., Savage, J. L., Russell, A. F., & Griffith, S. C. (2012). Building genetic networks using relatedness information: A novel approach for the estimation of dispersal and characterization of group structure in social animals. *Molecular Ecology*, *21*, 1727–1740.

Rousset, F. (2008). genepop'007: A complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*, *8*, 103–106.

Russello, M. A., & Amato, G. (2004). Ex situ population management in the absence of pedigree information. *Molecular Ecology*, *13*, 2829–2840.

Santure, A. W., Stapley, J., Ball, A. D., Birkhead, T. R., Burke, T., & Slate, J. (2010). On the use of large marker panels to estimate inbreeding and relatedness: Empirical and simulation studies of a pedigreed zebra finch population typed at 771 SNPs. *Molecular Ecology*, *19*, 1439–1451.

Slatkin, M. (1985). Gene flow in natural populations. *Annual Review of Ecology and Systematics*, *16*, 393–430.

Soulé, M. E. (Ed.). (1987). *Viable populations for conservation*. Cambridge: Cambridge University Press.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*, 1312–1313.

Szulkin, M., Stopher, K. V., Pemberton, J. M., & Reid, J. M. (2013). Inbreeding avoidance, tolerance, or preference in animals? *Trends in Ecology & Evolution*, *28*, 205–211.

Taylor, H. R. (2015). The use and abuse of genetic marker-based estimates of relatedness and inbreeding. *Ecology and Evolution*, *5*, 3140–3150.

Thornton, T., Tang, H., Hoffmann, T. J., Ochs-Balcom, H. M., Caan, B. J., & Risch, N. (2012). Estimating kinship in admixed populations. *American Journal of Human Genetics*, *91*, 122–138.

Van de Casteele, T., Galbusera, P., & Matthysen, E. (2001). A comparison of microsatellite-based pairwise relatedness estimators. *Molecular Ecology*, *10*, 1539–1549.

Vilà, C., Sundqvist, A.-K., Flagstad, Ø., Seddon, J., Björnerfeldt, S., Kojola, I., … Ellegren, H. (2003). Rescue of a severely bottlenecked wolf (*Canis lupus*) population by a single immigrant. *Proceedings of the Royal Society B: Biological Sciences*, *270*, 91–97.

Wang, J. (2002). An estimator for pairwise relatedness using molecular markers. *Genetics*, *160*, 1203–1215.

Wang, J. (2007). Triadic IBD coefficients and applications to estimating pairwise relatedness. *Genetical Research*, *89*, 135–153.

Wang, J. (2011). COANCESTRY: A program for simulating, estimating and analysing relatedness and inbreeding coefficients. *Molecular Ecology Resources*, *11*, 141–145.

Wang, J. (2014a). Marker-based estimates of relatedness and inbreeding coefficients: An assessment of current methods. *Journal of Evolutionary Biology*, *27*, 518–530.

Wang, J. (2014b). Estimation of migration rates from marker-based parentage analysis. *Molecular Ecology*, *23*, 3191–3213.

Wang, J. (2015). Individual identification from genetic marker data: Developments and accuracy comparisons of methods. *Molecular Ecology Resources*, *16*, 163–175.

Watts, P. C., Rousset, F., Saccheri, I. J., Leblois, R., Kemp, S. J., & Thompson, D. J. (2007). Compatible genetic and ecological estimates of dispersal rates in insect (*Coenagrion mercuriale*: Odonata: Zygoptera) populations: Analysis of “neighbourhood size” using a more precise estimator. *Molecular Ecology*, *16*, 737–751.

Watts, H. E., Scribner, K. T., Garcia, H. A., & Holekamp, K. E. (2011). Genetic diversity and structure in two spotted hyena populations reflects social organization and male dispersal. *Journal of Zoology*, *285*, 281–291.

Weir, B. S., Anderson, A. D., & Hepler, A. B. (2006). Genetic relatedness analysis: Modern data and new challenges. *Nature Reviews Genetics*, *7*, 771–780.

Wilson, G. A., & Rannala, B. (2003). Bayesian inference of recent migration rates using multilocus genotypes. *Genetics*, *163*, 1177–1191.

Woodroffe, R. (2003). Dispersal and conservation: A behavioral perspective on metapopulation persistence. In M. Festa-Bianchet, & M. Apollonio (Eds.), *Animal behavior and wildlife conservation*. Washington, DC: Island Press.

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, *28*, 3326–3328.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Escoda L, González-Esteban J, Gómez A, Castresana J. Using relatedness networks to infer contemporary dispersal: Application to the endangered mammal *Galemys pyrenaicus*. *Mol Ecol*. 2017;26:3343–3357. https://doi.org/10.1111/mec.14133