Evolution and genomic impact of duplications

Human and rhesus macaque genomes as case studies

Marina Brasó-Vives

TESI DOCTORAL UPF / ANY 2018

DIRECTORS DE LA TESI

Dr. Arcadi Navarro i Cuartiellas
Dr. Tomàs Marquès-Bonet

Departament de Ciències Experimentals i de la Salut

upf. Universitat Pompeu Fabra *Barcelona*

A elles, les que volen viure lliures

After centuries of dormancy, young women can now look toward a future moulded by their own hands.

Rita Levi-Montalcini

All we have to decide is what to do with the time that is given us.

Gandalf, John Ronald Reuel Tolkien

Fins i tot en el cas de la matemàtica, la ciència "pura" per excel·lència, la comunitat científica, que per molt temps ha estat orgullosa de fer ciència tancada a la seva torre d'ivori, ha canviat de paradigma.

La necessitat de millorar la percepció social de la utilitat de les matemàtiques és una preocupació general.

Enric Brasó Campderrós

Agraïments

I don't know half of you half as well as I should like; and I like less than half of you half as well as you deserve.

Bilbo, John Ronald Reuel Tolkien

Durant aquests anys he après molt. He après i m'heu ensenyat molt. I no només sobre duplicacions. He après molt sobre mi mateixa, sobre qui sóc i qui vull ser. Aprendre és una cosa que m'apassiona així que no puc deixar d'agrair profundament a tots els que m'heu brindat l'oportunitat d'aprendre.

També a aquells que s'entossudeixen a ser feliços i compartir felicitat. No hi ha objectiu més valuós en aquesta vida.

Gràcies,

Als meus pares, na Maria Rosa i l'Enric, els culpables de tot. Per haver-me transmès la vostra passió i la vostra inexorable curiositat per la ciència.

A l'Arcadi per haver vist en mi el que fos que et va fer donar-me l'oportunitat de fer el doctorat. Per confiar, creure en mi i respectar-me. Per ensenyar-me la teva manera minuciosa de fer ciència. Per ser proper, sincer i coherent. Al Tomàs, per acceptar el repte d'empènyer el carro en moments de canvi. Per ensenyar-me que, en ciència, la practicitat i la resolució són dues grans virtuts.

A David por tu interminable capacidad de debate y tu inmenso interés en todo. Por haberme ayudado tanto con los macacos. Por estar ahí siempre y por recordarme que los genes también molan.

To Jeff for the macaque samples, for your time and dedication.

To Inna for helping me when I was new into the WSSD world.

Als companys. IBEs passats i presents i a l'IBE en si per ser una gran família.

Al Txema i al Xavi, please don't cry. Al Gerard i al Gabriel per donar al grup una mica de la vostra maduresa revolucionaria. A l'Aitor per la teva ajuda sempre

amable. A la Clàudia per la teva determinació i tendresa. A la Neus per repartir alegria allà on vas. A tots els altres companys per tot.

Als Tapers, Lets i Pulis. Per ser la família que s’escull. Per compartir felicitat, germanor i amor cada dia.

Als tapers passats presents i futurs, per donar-li vida. Als que el van fer néixer, als que mai li han donat vida i als que n’hi donen cada dia. Al Juanillo per la teva autenticitat. A Marco por ser tan increíble. Al Nino per ser tan divertit. A la Lara per la teva sororitat. A Lukas por tu complicidad y tierna brusquedad. A Vicky por ser divertida y curiosa. A Lisa por ser resolutivamente directa (y por los disfraces). A l’Àlex per tirar endavant tots els plans. A la Carla pel teu amor i la teva complicitat. A l’Anna per corregir-me la tesi, per ser tan valenta i per les mil i una cerveses. A Adriano por ser un genuino y fenomenal caos. A Ali por ser tan magníficamente estupenda. A Maria por el yoga, por tu determinación y por los proyectitos. Al Marc per la teva practicitat i senzillesa. Als dos pel Nico (i per més). A les noves generacions: Ester, Marina, David, Anna, Antonio i als que vindran.

Als magnífics i eterns Lets. A la Sílvia per ser tan autèntica, encantada i divertida. Per ser tan detallista, delicatessen i escandalitzable. Per tots els moments a RDF. M’encanta viure amb tu. A la Júlia per la teva grandesa, tendresa i complicitat. Pel plaer de debatre-ho tot infinitament. Per RDF i per deixar-me fer-hi casa meva. A la Juna per tenir tantes variables dins el cap. Elles et fan ser com ets: interessant i imprevisible. Per tants moments. Per la complicitat tot i el temps i la distància. A la Mariona per la teva maduresa i fortalesa. Per ser tan dolça i pràctica. A la Poi perquè tens l’habilitat que, quan ens veiem, sembli que no ha passat ni un sol dia des de l’últim cop que ens vam veure. Per ser carinyosa i propera. A la Carlota perquè no hi ha persona al món amb les coses més clares que tu. Per ser tan resolutiva i motivada. Per ensenyar-me que l’assaig-error és una manera d’anar per la vida força eficient. A la Cuca per la dignitat mai perduda tot i els cuques. Per ser serena i esbojarrada al mateix temps. A la Pano per la teva sinceritat i la teva contundència puntual. Per ser una combinació perfecte entre sàvia, empàtica i divertida. A l’Ali per la teva energia. Per viure amb alegria. De gran vull ser com tu. A l’Anne per ser valenta i per voler ser com ets. Per les birres passades i futures. Se’t troba a faltar. Al Marc per ser carinyós i pacient. Per saber com gaudir de la vida. Al Pepe per la teva bogeria i determinació perfectament

coordinades. Pel teu ampli i sorollós riure. Per fer-m'ho passar sempre tan bé. A l'Igor per ser una perfecta incorporació al grup. Per fer sempre que em rigui de mi mateixa. Per ser carinyós després de tot. A l'Anna per la teva calidesa. A l'Alba per la teva manera de riure per sota el nas. A tots per la vostra amistat. És un privilegi i un honor tenir-la.

A l'Abel per ser autènticament meravellós. Per saber viure tan bé la vida. Per ser savi i despistat a la vegada. Per respectar-me i valorar-me sempre. Per ensenyar-me a fer-ho jo també. Per obrir-me les portes del teu fantàstic món i de la teva fantàstica gent. Per cada moment. Per ser-hi.

Als pulis i al vostre amor etern. Per ser una extraordinària i perfecta mescla de personatges i personalitats entranyables i desconcertants. Per ser tan extremadament divertits. Per acollir-me i abraçar-me sempre.

A la família Pichot per tractar-me com a una filla. Per ensenyar-me el que és bo en aquesta vida. Per ser tan intensos, divertits, generosos, hospitalaris i propers sempre.

Als meus germans, la Neus i l'Enric, per ser els meus *alter egos*. A la Neus per haver-li donat a aquesta tesi una espectacular portada digne del teu talent. M'encanta que porti una mica de tu. A l'Enric per ser el meu còmplice de sempre i per sempre. Als dos per fer-me sentir orgullosa de vosaltres cada dia.

Al Diego per la teva passió per la vida i per la ciència. Per ser totes les altres mil i una cares d'aquesta tesi. Pel meu màster, l'Interplay, el SeDuS, el Collapsed i pels futurs MEPS i SDRs. Per aquesta tesi tan teva. Per teva tenacitat i intensitat. Per valorar-me tant. Per saber tant i per ensenyar-me tant. Pel 419, pel Tàper, pels Naranjitos i Beachbumbas, per l'anglès, per Vienna, Australia, Asturies, Lió, Mèxic, Nova York, Washington, Japó i, sobretot, per més Lió i més Mèxic. Per fer-me feliç. Pel futur. Per compartir la vida que vull viure.

Abstract

Duplication is the main mechanism for the formation of new genetic material and functional innovation. Understanding the way duplications arise, evolve, co-evolve and give rise to new functions is essential. In this thesis, I present my contributions to the pursuit of this goal. I investigate the recombination process driving the concerted evolution of duplicates: interlocus gene conversion (IGC). In particular, I explore how IGC and crossover interplay, how IGC dependence on sequence similarity between duplicates influences their concerted evolution, and how IGC and the collapse of duplications in genome assemblies alters test statistics. In addition, I characterize the diversity of highly similar duplications in the human genome to elucidate their duplication mechanisms, their time of appearance and their contribution to the formation of new genes. Finally, I describe the duplicated and copy-number variant regions in the rhesus macaque genome and identify therein gene copy-number differences of functional relevance with humans.

Resum

La duplicació és el principal mecanisme de formació de nou material genètic i d'innovació funcional. Entendre com les duplicacions sorgeixen, evolucionen, co-evolucionen i engendren noves funcions és essencial. Aquesta tesi recull les meves contribucions a aquest objectiu. Investigo la recombinació responsable de la co-evolució dels duplicats: conversió gènica entre loci (IGC). Específicament, exploro com l'IGC i la recombinació per entrecreuament interactuen; com la dependència de l'IGC de la similitud entre duplicats influencia la seva co-evolució i com el col·lapse de duplicats en el muntatge de genomes altera proves estadístiques. També caracteritzo la diversitat de les duplicacions altament similars del genoma humà per aclarir els seus mecanismes de duplicació, moment d'aparició i contribució a la formació de nous gens. Finalment, descric les regions duplicades i variants en nombre de còpia del genoma del macaco rhesus i hi identifico diferències gèniques en nombre de còpies de rellevància funcional amb el genoma humà.

Prefaci

La recerca

Una macaca curiosa, un bon dia, va decidir que havia arribat el moment d'encaminar-se en cerca del fruit suculent del bosc frondós de què tant la seva mare li havia parlat.

Entrant al bosc frondós els seus ulls ja miraven inquiets en totes direccions buscant el famós fruit. Miraven enlaire buscant un enorme fruit suculent penjant d'un arbre ben gros. De quina altra manera podia ser el seu somiat fruit?

Buscant, buscant, va trobar una bonica móra. Una móra petitíssima. I, per què no? Se la va posar dins la boca. 'Mmmmh! Que dolça!', va pensar sense deixar de mirar enlaire per si divisava el gran fruit suculent.

Animada, va reprendre la seva recerca.

Mentre buscava l'enorme fruit, la macaca curiosa anava trobant móres dolces que anava menjant per fer passar l'espera i la gana. 'Mmmmh! Que dolça!', pensava cada cop que se'n posava una dins la boca.

Va buscar i buscar incessant i alegre durant tot el dia. Sense èxit. On era l'enorme fruit suculent?

Cap al tard, cansada de donar voltes i més voltes mirant enlaire, va seure en una grossa roca. 'Què se n'ha fet del fruit suculent de què parlava la mare?', es preguntava mentre allargava la mà per agafar una altra móra dolça i endur-se-la a la boca. 'Mmmmh! Que dolça!'

'De tanta móra dolça al final he quedat ben tipa! Que dolces que són!', pensava tot ajeient-se, cansada i satisfeta, al peu de la grossa roca.

'Demà més!'

Contents

1. INTRODUCTION

1.1 Evolution by duplication 3

1.2 Genome variation: from single-nucleotide variants to structural variants 6

1.2.1 Segmental duplications and copy-number variants . . . 7

1.2.2 Detection of SDs and CNVs 9

1.3 Evolution of duplications 13

1.3.1 Concerted evolution 13

1.3.2 Birth of duplications 21

1.3.3 Fates of duplicates 27

1.4 What do we know about primate SDs and CNVs? 30

1.5 Implications in disease and phenotype 34

2. OBJECTIVES

3. UNDERSTANDING NEUTRAL CONCERTED EVOLUTION IN SEGMENTAL DUPLICATIONS

3.1 Rationale 43

3.2 Objectives 45

3.3 Results and Discussion 46
3.3.1 Fundamental effects of IGC 47
3.3.2 Interplay of IGC and crossover 51
3.3.3 IGC and sequence similarity dependence and reciprocity 58
3.3.4 Neutrality tests on duplications and collapsed duplications 65
3.4 Methods 69
4. HUMAN SEGMENTAL DUPLICATIONS REVISITED
4.1 Rationale 75
4.2 Objectives 77
4.3 Results and Discussion 78
4.3.1 Tandem vs. Isolated intrachromosomal SDRs 81
4.3.2 Span and distribution 84
4.3.3 Age classification 88
4.3.4 Characterization 94
4.4 Methods 106
4.4.1 SD enrichment 106
4.4.2 Annotations of centromeres, genes and retrotransposons 106
4.4.3 Great ape copy number calling based on read depth (WSSD) 107
4.4.4 Phylostratification 107
4.4.5 Duplication rate calculation 108
5. COPY-NUMBER VARIANTS AND DUPLICATIONS IN RHESUS MACAQUE AFFECTING HUMAN DISEASE GENES
5.1 Rationale 113

5.2 Objectives 115

5.3 Results and discussion 116

5.3.1 High-quality genome-wide maps of fixed duplications and CNVs 116

5.3.2 Copy-number profile of rhesus macaque and human protein-coding genes 120

5.3.3 Comparing rhesus macaque and human genes copy-number profile 124

5.4 Methods 128

6. DISCUSSION

6.1 Interplay of IGC, sequence similarity and natural selection . . . 139

6.2 Studying duplications: what to take into account 141

6.3 Perspectives on human duplication types and dynamics 143

6.4 WSSD and WGAC 146

6.5 Difficulties and efforts to determine the function of duplications 147

6.6 Concluding remarks 148

7. APPENDIX

7.1 Interplay of interlocus gene conversion and crossover in segmental duplications under a neutral scenario 177

7.2 SeDuS: segmental duplication simulator 191

7.3 Effect of collapsed duplications on diversity estimates: what to expect 197

7.4 Supplementary figures 219

Abbreviations

AGC	Allelic gene conversion
array-CGH	Microarray-based comparative genomic hybridization
BIR	Break-induced replication
bp	Base pair
C	Gene conversion rate
CNV	Copy-number variant
DNA	Deoxyribonucleic acid
DSB	Double-strand break
DSBR	Double-strand break repair
FISH	Fluorescence in situ hybridization
FoSTeS	Fork stalling and template switching
gBGC	Guanine-cytosine biased gene conversion
GC	Gene conversion
HMM	Hidden Markov model
HR	Homologous recombination
IGC	Interlocus gene conversion
kbp	Kilobase pair
MEPS	Minimal efficient processing segment
MESH	Minimal efficient sequence homology
MMEJ	Microhomology-mediated end joining
My	Million years
Mya	Million years ago
NAHR	Non-allelic homologous recombination
NGS	Next-generation sequencing
NHEJ	Non-homologous end joining
R	Crossover rate
RNA	Ribonucleic acid
SD	Segmental duplication
SDSA	Synthesis-dependent strand annealing
SNV	Single-nucleotide variant
SSA	Single strand annealing
SV	Structural variant
WGD	Whole-genome duplication
WGS	Whole-genome sequencing

Chapter 1

Introduction

The true definition of science is this: the study of the beauty of the world.

Simone Weil

The more we know, the more we realize there is to know.

Jennifer Doudna

1.1 Evolution by duplication

Genetic drift and natural selection modulate the fate of genomic material through evolution. Although they do not generate new genomic variants, they are the forces influencing which of them will endure from one generation to the other. Genomic variants need to exist before genetic drift and natural selection can act upon them (Darwin, 1859; Kimura, 1968).

Although point mutations are a common source of novel genomic variants with potential influence on phenotype and capacity for functional innovation, the huge variety of functions in life is not only the product of point mutations. Instead, the main mechanism responsible for the complexity of genomes across the tree of life is duplication (Ohno et al., 1968; Ohno, 1970). In 1980, Tomoko Ohta began her book entitled *Evolution and variation of multigene families* with these clarifying sentences (Ohta, 1980):

> *"In evolution of higher organisms, gene duplication has apparently played a very important role. For more complex organization, more genetic information is needed, and gene duplication seems to be the only way to achieve it."*

In 1970, Susumu Ohno published *Evolution by gene duplication*, a book that inspired not only this section's name but gave rise to a new vision on genome evolution. In *Evolution by gene duplication*, Ohno presented his theory about the evolution of genomes through duplication, which is today widely accepted in the field. According to Ohno's ideas, functional regions in our genomes have little space for innovation because almost all variation appearing in such regions is deleterious and erased by natural selection. Duplication resolves this

limitation by generating redundant genetic material leading to a relaxation of selective constraints and leaving space for genomic innovation (see Section 1.3.3 for more specific scenarios). Ohno himself explained it like this (Ohno, 1970):

> *"As long as a particular function of an organism is under the control of a single gene locus, natural selection does not permit perpetuation of mutations which result in affecting the functionally critical site of a peptide chain specified by that locus. Hence, allelic mutations are incapable of changing the assigned function of genes.*
>
> *Gene duplication emerged as the major force of evolution. Only when a redundant gene locus is created by duplication is it permitted to accumulate formerly forbidden mutations and emerge as a new gene locus with a hitherto unknown function."*

Ohno and Ohta, in their aforementioned work, talked about gene duplications. We now know that duplications are not only restricted to gene duplications but vary in content and in size, ranging from *whole-genome duplications* (WGDs) to small insertions (see Section 1.2).

Duplications are a pervasive feature of eukaryotic genomes (Ohno et al., 1968; Kaessmann, 2010; McGrath and Lynch, 2012). Particularly, two WGD events in the early stages of the diversification of vertebrates seem to have driven the transition from invertebrates to vertebrates (Ohno et al., 1968; Cañestro, 2012; Cañestro et al., 2013). Moreover, the architecture of vertebrate genomes is also the result of a long local-duplication history, exceptionally active in the great ape lineage (Lynch, 2007; Marques-Bonet et al., 2009a; Kaessmann, 2010; see Section 1.4). Thus, a large part of the sequence of our genomes orginitated from duplications, although only a fraction is currently considered duplicated or even can be detected as such.

After duplication, copies are identical. Differences between duplicates emerge progressively being exposed to evolutionary forces such as genetic drift, natural selection (see Section 1.3.3) and *interlocus gene conversion* (IGC; see Section 1.3.1). With time, these differences accumulate and make duplicates more and more divergent. Margaret O. Dayhoff, Winona C. Barker and Lois T. Hunt (Dayhoff et al., 1983) describe it like this:

> *"When gene pools become isolated, through either a separation of interbreeding populations or a duplication of genetic material within a species, the copies gradually acquire changes independently of one another. At first the sequences are so similar that there is no question about their common origin. With increasing time more and more change occurs until it may no longer be possible to recognize the similarity."*

Given its long history of duplication events, our genome is a landscape of duplications in a wide range of divergence stages: from highly diversified old duplications that can almost no longer be recognized as such, to highly similar duplications that are hard to distinguish one from each other (see Section 1.2.2). In practice, a particular region is considered a duplication when it has a considerable amount of similarity with respect to other parts of the genome (see Section 1.2.1).

Duplication is at the core of this thesis and as such I have considered it necessary to begin by stating its importance in evolution. I will now step back and describe genome variation in all its forms emphasizing the role of duplication[1].

[1]Through this thesis I will use the term *duplication* for both the event of an appearance of a genomic copy (or copies) of a prior genomic region, and the resulting redundant genomic regions themselves. Moreover, I will use the term *duplicates* to refer to duplicated copies or group of regions that share identity. Opposite to duplication, I will refer to a genomic region without significant similarity to other parts of the genome as a *non-duplicated* or *diploid* region.

We wish to discuss a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biologic interest.

Rosalind Franklin

1.2 Genome variation: from single-nucleotide variants to structural variants

Genomic variants are differences between the genomes of individuals within a given population or species. They take many forms: *single-nucleotide variants* (SNVs); short insertions and deletions; microsatelites and short tandem repeats; presence or absence of transposable elements; and larger genomic changes named *structural variants* (SVs).

Since the publication of the first draft of the human genome in 2001 (Lander et al., 2001; Venter et al., 2001), population genetics has focused on SNVs (Cann et al., 2002; The International HapMap Consortium, 2005; Haussler et al., 2009; Prado-Martinez et al., 2013; Auton et al., 2015; de Manuel et al., 2016; Mallick et al., 2016; Xue et al., 2016). Nevertheless, for several reasons, which include the failure of SNVs to account for all the observed heritability, an issue known as the *missing heritability* problem (Beckmann et al., 2007; Eichler et al., 2010), and the improvement and cost-minimization of sequencing technologies, during the last decade SVs have become the center of a large body of research (Itsara et al., 2009; Conrad et al., 2010; Mills et al., 2011; Sudmant et al., 2013, 2015a,b; Zarrei et al., 2015; Kronenberg et al., 2018).

SVs are changes in number of copies or location of a genomic region. The involved genomic region can be as big as the whole genome. This is the case of differences in ploidy (result of a WGD) among the individuals of the same species. These differences in ploidy are very frequent in plants (Jaillon et al., 2007; Mühlhausen and Kollmar, 2013) and have been repeatedly occuring in eukaryotic evolution (Sémon and Wolfe, 2007; Cañestro, 2012; McGrath and Lynch, 2012; Mühlhausen and Kollmar, 2013). Although WGDs have the initial

advantage of respecting the gene dosage balance, they are energetically very expensive and most of the new genetic material is finally lost (Inoue et al., 2015). In addition to WGDs, SVs include trisomies and other chromosomal rearrangements which can be frequently found between closely related species despite generally being strongly deleterious (Ventura et al., 2012). Finally, the smallest and more frequent type of SVs is *copy-number variants* (CNVs), which are discussed in the following section (Feuk et al., 2006a).

1.2.1 Segmental duplications and copy-number variants

CNVs are defined as differences in the number of copies of a genomic region of more than 1 kbp between individuals within a population or species (Feuk et al., 2006a). They can be the result of deletions, which decrease the number of copies of the involved region, or duplications, which increase it[2]. CNVs were discovered after observing that some patients had different number of copies of a given genomic region compared to healthy individuals (Lupski, 1998), being this the genetic cause of their disease. For this reason, like SNVs, CNVs can only be considered when comparing genomes within a given population (Iafrate et al., 2004; Sebat et al., 2004; Feuk et al., 2006a,b).

The definition of CNVs partially overlaps with another definition, the definition of *segmental duplications* (SDs). SDs are genomic duplications of 1 kbp or more in length presenting at least 90% of identity[3] between copies (Bailey et al., 2001). SDs were defined when the first draft of the reference human genome came out, to represent the redundancy of sequence found within it (Bailey et al., 2001; Lander et al., 2001; Venter et al., 2001). Thus, SDs are duplications present within one single haploid genome and, contrary to CNVs, their definition does not consider a population perspective. In other words, SDs are described in one individual regardless of being present or not in other

[2]Although keeping in mind that CNVs can also be the product of deletions, for the sake of simplicity and because of my particular interest in duplications, throughout this thesis I will mainly use the term CNV to refer to duplications segregating in number of copies in the population.

[3]During the whole of this thesis I will use the term *homologous* as the quality of descending from the same ancestral genomic region. In the same way, I will use *paralogs* for homologs generated by duplication and *orthologs* for homologs separated by speciation. On the other hand, I will use the terms *identity* and *similarity* as synonyms and as antonyms of *divergence*, and all three of them as quantitative characteristics of homologs.

individuals of the same population.

From the definitions of SD and CNV we could state that an SD having variable number of copies within a population is a CNV. And the other way around, if a given CNV involves a duplication in an individual, and its copies have more than 90% identity, this CNV would be an SD in such individual. Note that, on the one hand, the definition of SD does not specify whether they are variable or not in the population (although copies have to be at least 90% identical) and, on the other hand, the definition of CNV implies variability in the population but not a minimum identity between copies. Moreover, all SDs, regardless of being fixed or not in the population, arose in a single individual and, thus, were CNVs at the time they appeared (see Section 1.3.2). It is therefore not surprising to find that CNVs are enriched in SDs (Sharp et al., 2005; Itsara et al., 2009).

In practice, like many entities in biology, what is considered an SD or a CNV depends on the available data. In this case, it basically depends on the sample size and the method used to detect them. First, what is considered to be an SD or a CNV depends on the sample size because what is detected as fixed or variable depends on the amount analyzed individuals (Sandelowski, 1995). An apparently fixed SD in a given sample size could potentially be regarded as a CNV if a larger sample size is considered. Second, it will largely depend on the method used to detect them[4]. To detect SDs and CNVs is not an easy task. I will elaborate on the existing detection methods and on the limitations of duplications detections in Section 1.2.2.

SDs and CNVs are long stretches of DNA sequence that have a high degree of similarity with other parts of the genome. Given the high similarity between duplicates, we can roughly[5] say that SDs and CNVs are long duplications that appeared during the last 35-40 million years (Bailey and Eichler, 2006). Moreover, SDs and CNVs are long enough to include functional elements such

[4] In many cases, a CNV (when involve a duplication) is just an SD varying in number of copies in the population. This is so because the methods used to detect CNVs have a maximum sensibility on divergence and this is generally not bigger than 10% (see Section 1.2.2).

[5] High identity between duplicates implies either that the duplication is quite recent in evolution and that point differences between duplicates have still not appeared or that the point differences between duplicates that appeared during the past have not endured through evolution (see Section 1.3.1 and Section 1.3.3 for possible reasons behind the maintenance of similarity between duplicates through time).

as genes, and thus, they are the birthplace of potentially new functional elements. In other words, they are the recently created redundancy of sequence that will potentially result in functional innovation according to Ohno's theory (Ohno, 1970). Moreover, some SDs and CNVs have been found to be responsible for disease (see Section 1.5). These are two of the main reasons that motivate the study of SDs, CNVs and how they appear and evolve.

1.2.2 Detection of SDs and CNVs

As mentioned in Section 1.2.1, SD and CNV detection methods are a central factor for determining what is and what is not an SD and/or a CNV. Three main groups of SD and CNV detection methods dominate the field: hybridization-based CNV detection methods, assembly-based duplication detection methods, and methods based on *next-generation sequencing* (NGS) technologies.

Hybridization-based CNV detection methods

There are many tools and methodologies designed for patient-genotyping of specific CNVs causing disease (Vandeweyer and Kooy, 2013; Martin et al., 2015; Weckselblatt and Rudd, 2015). The most important of these targeted techniques is *fluorescence in situ hybridization* (FISH; Langer-Safer et al., 1982). With this technique, specific fluorescent DNA probes hybridize cellular DNA in order to visualize the genomic loci (and their number) that have the corresponding sequence.

On the other hand, *microarray-based comparative genomic hybridization* (array-CGH; Lucito et al., 2003) is a generalization of FISH. In array-CGH, genome-wide copy-number differences between two genomes (*e.g.* a patient's genome compared to a reference genome) are detected with high resolution. Typically, with an array-CGH, only differences in number of copy between the two tested genomes can be detected, whereas duplications having the same number of copies in the two compared genomes cannot be distinguished from non-duplicated (diploid) regions.

Section 1.2

Assembly-based duplication detection methods

During the first attempt to decode the human genome, the presence of highly similar long duplications was evidenced (Lander et al., 2001; Venter et al., 2001). Shortly after the release of the first draft of the human genome, Bailey et al. (2001) presented a new method designed to detect duplications in the assembly called *whole-genome assembly comparison* (WGAC), although this term was not coined until later (Bailey et al., 2002a).

The WGAC approach subjects a previously repeat-masked genome assembly to generalized inward BLAST similarity searches (Altschul et al., 1990). Previously masked repeat sequences are reincorporated to the found alignments and end-trimming algorithms are applied to optimize the resulting alignments. Only those of a minimum size of 1 kbp and a minimum identity between sequences of 90% (excluding gaps) are reported (Bailey et al., 2001).

The SD term was coined with the design of WGAC and its implementation in the first human assembly (Bailey et al., 2001). Since then, WGAC has been commonly used to annotate SDs in multiple species genome assemblies (UCSC Table Browser; Karolchik, 2004) and to study SDs (Bailey et al., 2002a, 2004b; She et al., 2006; Nicholas et al., 2009; Jiang et al., 2014; Feng et al., 2017). In Chapter 4, I study human SDs through an in depth analysis based on an SD database constructed from WGAC data.

WGAC provides very useful information for the study of SDs. Its complete list of alignments across the genome specifies, not only which are the duplicated regions in the genome, but also where they are located. Moreover, having the sequence alignment between duplicated pairs allows the study of their point differences. However, the WGAC approach has two main limitations as a result of being based on a genome assembly. First, as long as one single haploid genome assembly is used, WGAC is blind to CNVs. And, second, the quality of the WGAC SD annotation will depend on the quality of the genome assembly used.

As previously noted, DNA sequences from two highly similar duplicates are difficult to distinguish from each other. This fact entails serious problems in resolving duplications (that is, defining their limits, sequence

and locating them adequately) when constructing genome assemblies. Sequencing reads coming from duplications are hard to unambiguously assign to a given duplicate, being impossible in cases of highly similar duplicates and short read lengths. Many genome assembly algorithms and technologies intend to tackle this problem and distinguish highly similar duplicates to avoid collapsing them in a single assembly locus (Nagarajan and Pop, 2013; Simpson and Pop, 2015; Lu et al., 2016). Despite there being cases of recent *de novo* genome assemblies constructed with new higher-quality long-read technologies (Chaisson et al., 2015; Lu et al., 2016), there are many species' reference genomes essentially constructed with NGS short reads that contain collapsed duplications (Salzberg and Yorke, 2005; Kelley and Salzberg, 2010). In Section 3.3.4, I will discuss and present results regarding the consequences of collapsed duplications on whole-genome selection scans (see also Hartasánchez et al., 2018 in Appendix 7.3).

NGS-based duplication detection methods

Other approaches to detect genome-wide SDs and CNVs use NGS data. Despite the existence of methods that use local discordant read mappings, such as paired-end reads mapping in an unusual span or orientation, the use of NGS reads for SDs and CNVs detection is dominated by read-depth-based methods (Abel and Duncavage, 2013; Tattini et al., 2015). These methods are based on the principle that, if the sequencing process is regular, the number of reads mapping to a given genomic region is proportional to the number of genomic regions generating these reads or, in other words, to the number of copies of such genomic region.

The first time read depth was used to detect non-diploid regions genome-wide was in 2002 by Bailey et al. (2002a), who named the technique WSSD, for *whole-genome shotgun sequence detection*. Nowadays, we can find several improved WSSD algorithms although the initial steps are essentially common in all of them (Alkan et al., 2009; Yoon et al., 2009; Magi et al., 2012; Serres-Armero et al., 2017). To perform this technique, first, NGS reads are mapped to a repeat-masked genome assembly (normally with relaxed identity requirements around $\geq$95%).

Second, depth of coverage is corrected for biases in mapping and sequencing (especially GC-content biases). Third, mean read depth is computed in windows that cover the genome. Fourth, the copy number of each window is calculated taking as reference the read depth of diploid control regions. In Chapters 4 and 5, I will apply WSSD to study human and rhesus macaque SDs and CNVs. For more details on WSSD methods see Sections 4.4 and 5.4.

WSSD also allows the study of the number of copies along the whole genome of multiple individuals at the same time, and it has been widely used to study SDs and CNVs in different species and populations (Bentley et al., 2008; Campbell et al., 2008; Wang et al., 2008; Wheeler et al., 2008; Chiang et al., 2009; Marques-Bonet et al., 2009a; Sudmant et al., 2013, 2015a,b; Serres-Armero et al., 2017). Moreover, unlike WGAC, WSSD has the advantage of detecting a duplicated region as such regardless of it being resolved or collapsed in the assembly. Nevertheless, WSSD has the disadvantage of only indicating the number of copies of a given region and not providing information on the copies' location nor sequence (see Chapter 6 for further discussion on advantages and disadvantages of WGAC and WSSD).

Considering that the advantages and limitations of WGAC and WSSD are complementary to some extent, using both of them at the same time can allow an adequate study of the architecture and distribution of duplications within the genome (WGAC) and its variation within and between species and populations (WSSD). n Chapter 4, I use this combined approach akin to previous work (Bailey et al., 2004a; Nicholas et al., 2009; Jiang et al., 2014; Feng et al., 2017).

New promising long-read and nanopore-based sequencing technologies allow discerning duplicates more accurately and a better construction of genome assemblies (Nagarajan and Pop, 2013; Simpson and Pop, 2015; Lu et al., 2016; Wajid et al., 2016). High quality assemblies of several species are already available being highly valuable tools for studying SDs and CNVs (Huddleston et al., 2014; Chaisson et al., 2015; Gordon et al., 2016; Shi et al., 2016; Kronenberg et al., 2018). Additionally, innovative SD and CNV detection methods represent interesting new perspectives for the study of duplications (*e.g.* Pu et al., 2018).

I am incapable of conceiving infinity, and yet I do not accept finity.

Simone de Beauvoir

1.3 Evolution of duplications

Ohno's publication in 1970 (Ohno, 1970) and his proposal of the way new genetic material is created through duplication left many open issues. They can be summarized in three general questions:

1. How do duplications arise?
2. How do duplications evolve?
3. How does natural selection act upon duplicates?

During the almost 50 years that have passed since Ohno's publication, a substantial amount of research has been done devoted to these questions. In this section, I will give an overview of the current knowledge regarding the evolution of duplications. First I will explain how duplications can co-evolve, then how they arise and finally how they can result in new functions.

1.3.1 Concerted evolution

Duplicates do not always evolve independently. The co-evolution of duplicates is named *concerted evolution*. Although Ohno did not mention this in his 1970 book (Ohno, 1970), cases of concerted evolution of duplicates started to be reported short after this was published (Gally and Edelman, 1970; Brown et al., 1972; Brown and Sugimoto, 1973; Hood et al., 1975; Tartof, 1975).

As an example, let us consider a duplication event that occurs prior to a speciation event. In the absence of concerted evolution, duplicates will evolve independently (Figure 1.1 A) and thus, paralogs will, in principle, be more

divergent than orthologs, which will retain similarity given their *a priori* shared function (Figure 1.1 C). On the contrary, under concerted evolution, duplicates co-evolve within each one of the species after speciation (Figure 1.1 B) and paralogs will be more similar than orthologs (Figure 1.1 D). Under the latter scenario, the molecular clock (Zuckerkandl and Pauling, 1965) cannot be used to date duplications since they will appear to have a more recent origin than the actual time of duplication.

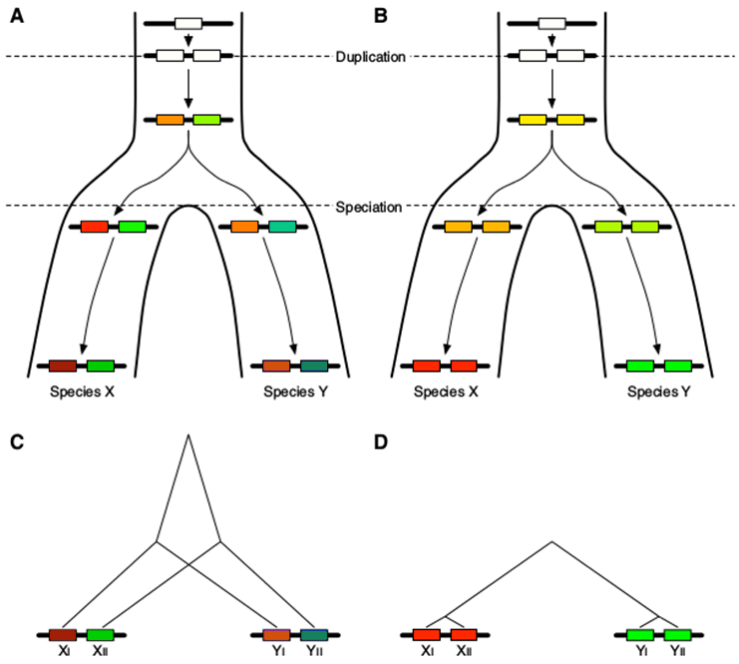


Figure 1.1: Presence and absence of concerted evolution of duplicates. Evolution of sequence similarity (represented by differences in colors) between regions after duplication and posterior to speciation under a model without concerted evolution (A) and with concerted evolution (B). Corresponding resultant trees are represented in C and D. [Image adapted from Innan (2009).]

Although many explanations to concerted evolution have been proposed since the first observed cases in the 1970s, only three are commonly accepted nowadays (Hood et al., 1975; Tartof, 1975; Ohta, 1980). First, a recombination

process can erase differences between duplicates (see below in Section 1.3.1). Second, a repeated birth and death process of duplications can confound the age estimates of duplications (see Section 1.3.2). Third, selection can prevent differences between duplicates to endure through evolution (see Section 1.3.3). These three mechanisms of concerted evolution are not mutually exclusive but can coexist and interplay (Graur and Li, 2000; Ohta, 2010).

Interlocus gene conversion

One of the main mechanisms underlying the concerted evolution of duplicates is IGC (Ohta, 2010). To present IGC, how it works and what are its effects on the molecular evolution of duplicates, I will first explain the general process of *gene conversion* (GC).

In *homologous recombination* (HR) a DNA *double-strand break* (DSB) is repaired using as a template an homologous DNA segment (Figure 1.2; Symington et al., 2014; Haber, 2018; Sung, 2018). DSBs can be either due to DNA damage or part of the normal recombination process during cell division, mainly meiosis (Murti et al., 1992). GC is one of the possible byproducts of HR characterized by the non-reciprocal exchange of genetic material between both involved DNA strands (Duret and Galtier, 2009; Hastings, 2010; Dwivedi and Haber, 2018). In other words, in GC the broken strand is fixed by copying the homologous information from the other strand (Figure 1.2). HR resolved in both synthesis-dependent strand annealing (SDSA) and *double-strand break repair* (DSBR) processes can result in GC events (Figure 1.2).

Allelic gene conversion (AGC) is the most common form of GC. It occurs when the broken strand in a DSB and the homologous template used to repair it are in the same locus, either in an homologous chromosome or in a sister chromatid (Waldman, 2008). AGC leads to non-mendelian proportions of meiotic products, was observed and described even before the structure of DNA was discovered in 1953 (Watson and Crick, 1953). According to Mendel's model (Mendel, 1866), of the four haploid meiotic products of an F1, two carry one trait and the other two carry the other trait. Nevertheless, studies of meiosis in *Saccharomyces cerevisiae* showed exceptions to this premise as proportions of 1:3 were sometimes observed

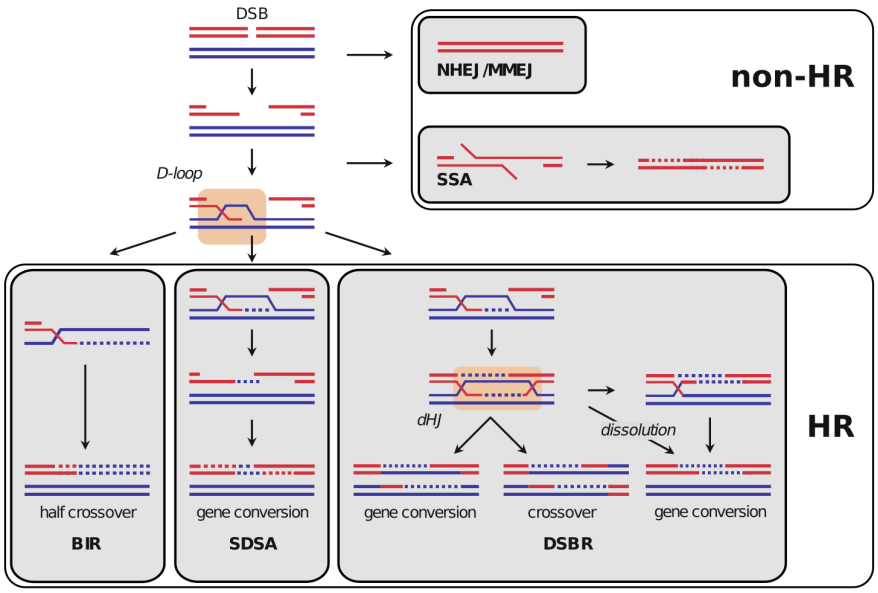


Figure 1.2: Repair pathways of DSBs. After a DSB, HR is the less mutagenic and, thus, preferred strategy. In HR after DSB, there is a 5' to 3' resection of the DNA ends. 3' overhangs carry out an homology search, with a posterior D-loop formation. Three mechanisms of HR are known. First, *break-induced replication* (BIR) which happens when one of the DNA extremes result of the DSB is missing. Second, SDSA which happens when the strand invasion is reverted after DNA synthesis. And, third, the classical DSBR which involves the formation and resolution of a *double Holliday junction*. SDSA results in GC and DSBR can result in GC or in crossover including GC. *Non-homologous end joining* (NHEJ), *microhomology-mediated end joining* (MMEJ) and *single strand annealing* (SSA) are alternative resolutions of the DSB that do not involve HR and, thus, are potentially more mutagenic. [Image adapted from Sebesta and Krejci (2016).]

(Lindegren, 1953). In 1964, Robin Holliday (Holliday, 1964) proposed a model of the mechanism through which GC could occur. Holliday set the foundations of the molecular mechanisms of HR and his model became the reference model for DSB resolution. Since then, AGC has been widely studied, especially because of the consequences of its bias for purine nucleotides, named gBGC, for *guanine-cytosine biased gene conversion* (Galtier et al., 2001; Duret and Galtier, 2009; Necşulea et al., 2011).

Differently, IGC, also named non-allelic, ectopic, interchromosomal,

interparalog or intergenic GC, occurs when the two homologous strands involved in a DSB repair are not in the same locus (Hastings, 2010). IGC has been observed in multiple species and between many different types of non-allelic homologous sequences (Chen et al., 2007; McGrath et al., 2009; Casola et al., 2010, 2012a; Kijima and Innan, 2010; Mansai et al., 2011; Assis and Kondrashov, 2012; Dumont and Eichler, 2013; Nuttle et al., 2013; Dumont, 2015; Ellison and Bachtrog, 2015; Trombetta et al., 2016; Harpak et al., 2017).

In general, HR happening between different loci is referred to as *non-allelic homologous recombination* (NAHR). When NAHR is resolved in a crossover, it can lead to chromosomal rearrangements, duplications, losses or even to the creation of aberrant chromosomes (Figure 1.3; see Section 1.3.2). The term NAHR has been largely used to refer only to its specific resolution in crossover without taking IGC into account. Despite this, IGC is an alternative resolution of NAHR (Figure 1.3) that, in fact, is not entirely incompatible with crossover (Figure 1.2). I will use NAHR to refer to both, IGC and NAHR resolved in crossover.

The main consequence of IGC is the concerted evolution of duplicates (Figure 1.1). After duplication, differences between both duplicates start to appear through point mutation. If IGC happens between them, these mutations are transferred from one duplicate to the other in both directions. This situation can lead to the stabilization of the divergence between the two duplicates depending on the IGC and the mutation rates. When this happens, it is said that the two duplicates are in a concerted evolution equilibrium (Figure 1.4). In Chapter 3, I will elaborate on the conditions in which concerted evolution equilibrium is possible.

Importantly, the impact of IGC on the divergence between duplicates is not its only effect. Already in the first models of IGC, an increase in sequence diversity within duplicates was predicted (Ohta, 1982, 1983; Innan, 2003a,b). It is also known that IGC affects *linkage disequilibrium* (LD) between duplicates (Ardlie et al., 2001; Frisse et al., 2001; Innan, 2002; Ptak et al., 2004; Hartasánchez et al., 2014). Predictions of divergence, diversity and LD between duplicates at equilibrium have been obtained under several models with and without selection (Baltimore, 1981; Ohta, 1982, 1983; Nagylaki, 1984; Innan, 2002, 2003a,b). According to these models, divergence, diversity and LD between duplicates not only depend

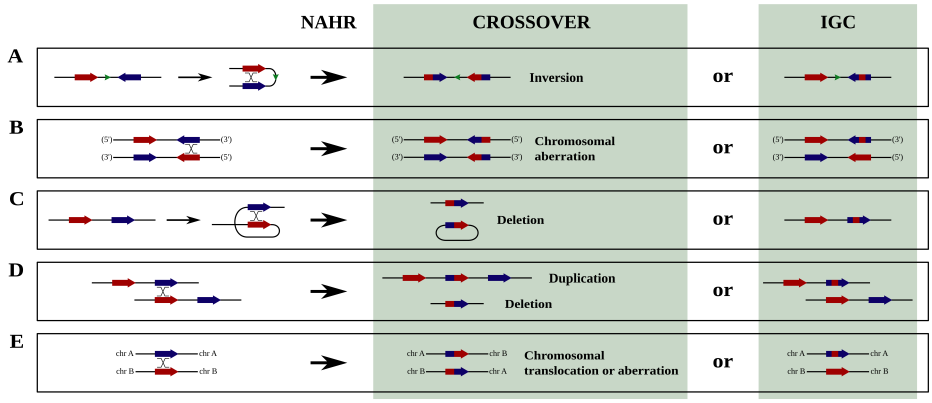


Figure 1.3: Products of NAHR between two different loci resolved in crossover or in IGC. Colored arrows represent homologous sequences in different genomic loci. Crossover and IGC products in different NAHR situations: (A) two duplicates in opposite orientations and in the same DNA molecule; (B) two duplicates in opposite orientations and in different DNA molecules (homologous chromosomes or sister chromatids); (C) two equally oriented duplicates in the same DNA molecule; (D) two equally oriented duplicates in different DNA molecules (homologous chromosomes or sister chromatids); (E) two duplicates in different chromosomes. [Image inspired in Chen et al. (2007, 2010b, 2014).]

on the IGC and the mutation rates but also on the crossover rate happening between duplicates. In Chapter 3, I will describe, in detail, the models and effects of IGC, and the interplay of crossover and IGC in the concerted evolution of duplicates.

IGC rate has been seen to negatively correlate with distance between duplicates (Lichten and Haber, 1989; Schildkraut et al., 2005; Chen et al., 2007; Zhi, 2007; Benovoy and Drouin, 2009; Casola et al., 2010); to be higher if duplicates are in the same chromosome than if they are in different chromosomes (Lichten et al., 1987; Lichten and Haber, 1989; Benovoy and Drouin, 2009; McGrath et al., 2009; Casola et al., 2010); to positively correlate with the crossover rate in the involved regions (Benovoy and Drouin, 2009); to be higher in meiosis than in mitosis (Jinks-Robertson and Petes, 1986); to have deletion bias (Assis and Kondrashov, 2012); and to have a donor-acceptor bias (Chen et al., 2007). In the case of the donor-acceptor bias, proximal-to-distal (relative to the

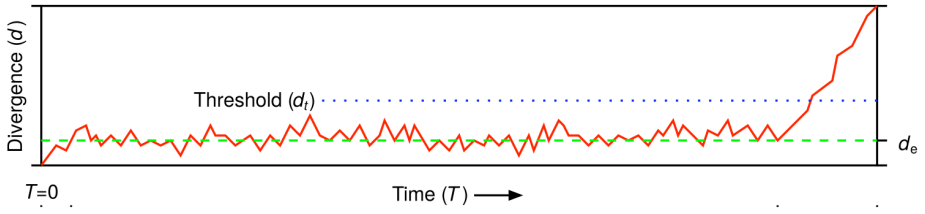


Figure 1.4: Divergence between duplicates through time. Duplication appears at $T = 0$. Divergence increases until it reaches equilibrium, around d_e, due to the action of IGC. At a given point in time, the fluctuating divergence can reach a threshold above which IGC can no longer happen and divergence starts to accumulate linearly between both duplicates. [Image adapted from Innan (2009), in turn adapted from Teshima and Innan (2004).]

centromere) IGC rate has been reported to be more frequent than distal-to-proximal IGC rate (Bosch et al., 2004); and if happening between paralogous genes, IGC directionality has been seen to depend on the relative expression levels of the involved genes (being the most expressed gene the most common donor; Papadakis and Patrinos, 1999). Moreover, like AGC, IGC has a bias towards purine nucleotides (gBGC) promoting them in an heteroduplex situation (Duret and Galtier, 2009).

Sequence similarity requirements

HR needs a certain degree of identity between the two involved sequences (Shen and Huang, 1986). After a DSB, an homology search has to be successful in order for HR to actually happen (Figure 1.2). If this homology search fails, the DSB would be resolved via a more mutagenic non-HR pathway. In the case of allelic HR (*e.g.* AGC), the degree of identity between strands is normally very high except in the case of highly diverged hybrids (Davies et al., 2016). However, in the case of NAHR, the degree up to which duplicate sequences have diverged determines to a large extent the possibility for these sequences to undergo NAHR (Walsh, 1987; Kijima and Innan, 2010).

Two different sequence similarity requirements have been proposed for HR: a *minimal efficient processing segment* (MEPS; Shen and Huang, 1986), and

a *minimal efficient sequence homology* (MESH; Chen et al., 2010b):

- MEPS is a minimal length of uninterrupted (100%) identity needed for HR to happen (Shen and Huang, 1986). MEPS has been measured in several species and with multiple techniques resulting in a large range of estimates. Among other, it has been seen to be >20 bp in *E. coli* (Shen and Huang, 1986), >200 bp in mouse cells (Liskay et al., 1987), between 337-456 bp in a pathological NAHR crossover in humans (Reiter et al., 1998), to fit within human LTR retrotransposons of an average size of 350 bp (Kijima and Innan, 2010; Trombetta et al., 2016) and within human Alu elements of about 300 bp (Zhi, 2007), and to be as short as 26 bp in yeast (Ahn et al., 1988; Mézard et al., 1992).
- MESH is a minimum degree of overall identity between duplicates for HR to happen. A meta-analysis of human pathogenic IGCs showed that they happen almost always between sequences with >92% and usually >95% sequence identity (Chen et al., 2007).

However, how does HR actually depend on sequence identity? There is a positive correlation between the length of uninterrupted identity and the rate of HR (Rubnitz and Subramani, 1984; Waldman and Liskay, 1988; Shen and Huang, 1989). This correlation is independent of the overall mean sequence identity (Waldman and Liskay, 1988) suggesting that it is MEPS rather than MESH that actually determines the rate of HR. The longer the length of the 100% identity segment, the higher the number of available MEPS within it and the higher the HR rate (Shen and Huang, 1989). According to these results MESH can only be an indirect consequence of MEPS.

IGC, as a form of HR, depends on the sequence identity between duplicates (Chen et al., 2007; Benovoy and Drouin, 2009; Casola et al., 2010). Nevertheless, unlike other types of HR, IGC dependence on sequence similarity leads to a feedback loop. If there is enough sequence similarity between duplicates for IGC to occur, IGC events will maintain high levels of identity between duplicates, laying the ground for future IGC events. Yet, if there is low sequence identity between duplicates, IGC cannot occur and more divergence will be generated with time. In Section 3.3.3, I will elaborate on the consequences of the feedback loop between sequence similarity and IGC on the molecular evolution of duplicates.

1.3.2 Birth of duplications

Duplications shape the architecture of eukaryotic genomes and result in genetic innovation but, how do they come to be? Alike all the other types of genomic variants, duplications arise in an individual genome and segregate in the population until they either reach fixation or disappear. The known duplication mechanisms include NAHR, abnormalities during DNA replication and retrotransposition (Babushok et al., 2007; Hastings et al., 2009a,b; Zhang et al., 2009; Cooper et al., 2011; Chen et al., 2014).

Duplication through NAHR

As stated in Section 1.3.1, NAHR is an HR happening between homologous sequences that are not in the same locus (Figure 1.2). When NAHR is resolved in crossover it can lead to duplications.

A crossover resolution of NAHR between different chromosomal loci can lead to either tandem duplications (Figure 1.3 D), deletions (Figure 1.3 C and D) or other chromosomal rearrangements (Figure 1.3 A, B and E). Many disease-causing duplications have been seen to be tandem duplications and are thought to be generated by NAHR between genomic loci (Stankiewicz and Lupski, 2002; Carvalho and Lupski, 2016; see Section 1.5).

NAHR, as all HR, needs a certain amount of homology between the involved DNA fragments to occur (see Section 1.3.1). Tandem duplications are a perfect context for NAHR to happen repeatedly, either by generating deletions or additional duplication (Figure 1.3 C and D). This process is normally referred to as the *birth and death of duplications* and consists on the expansion and contraction of chains of multiple tandem duplications. This recursive nature of NAHR can lead to parallel duplication events in close species that resemble paralogous products of a single duplication event before speciation. This might cause parallel duplications to be mistakenly assigned as orthologs. This birth and death of duplications is one of the mechanisms that can lead to paralogs having more similarity than (mistakenly-assigned) orthologs, which is the typical signature of concerted evolution (Ohta, 1983; see Section 1.3.1).

NAHR can also happen between a chromosomal locus and a free DNA

molecule (Figure 1.5). In such a situation, if the NAHR is resolved in crossover, it can lead to the insertion of the involved free DNA molecule into the genome (Figure 1.5). This free DNA can be viral DNA, retrotranscribed cellular RNA (including retrotransposons), product of a previous NAHR (Figure 1.3 C) or a molecule resulting from DNA damage. The insertion of a free DNA molecule in a genomic locus represents a duplication if this molecule has a cellular origin (see below in *Duplication by retrotransposition*).

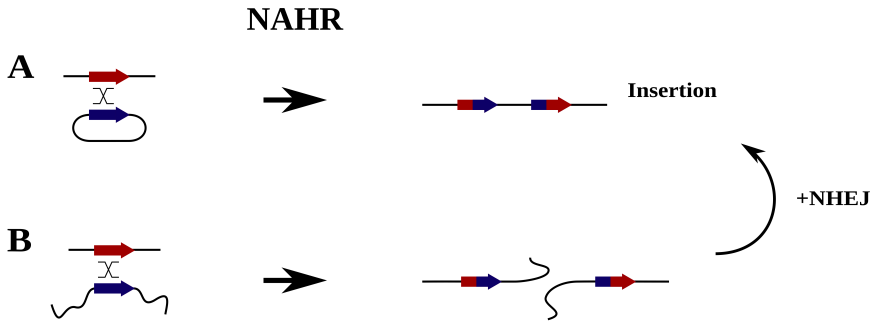


Figure 1.5: NAHR between a genomic locus and a free DNA molecule resolved by crossover. Colored arrows represent homologous sequences in a genomic locus (red) and a free DNA molecule (blue) either circularized (A) or not (B). Both cases can result in an insertion of the free DNA molecule in the genome. In B, a NHEJ recombination process after NAHR is needed. [Image inspired in Bailey et al. (2003).]

In 2003, Bailey et al. (2003) proposed a model of Alu-Alu-mediated NAHR for the generation of SDs in the human genome. In their model, NAHR generating duplicates was driven by Alu sequences which are extremely abundant in the genome and have highly conserved sequences. In fact, specific Alu-Alu mediated NAHRs have been identified as sources of genetic disease in humans (Rouyer et al., 1987; Hu and Worton, 1992; Brooks et al., 2001; Frühmesser et al., 2013; Gu et al., 2015). In Section 4.3.4, I will investigate and elaborate on the model of origin of human SDs by NAHR between Alu sequences proposed by Bailey et al. (2003).

Duplication during DNA replication

When DNA replication happens before cellular division, chromosomes are used as templates for the generation of the new genetic material. DNA synthesis is generated simultaneously in multiple replication forks. Alterations in the DNA replication process can lead to duplication or deletion of particular regions in the resulting new genome. There are two main mechanisms for this to occur:

1. *Replication slippage* happens when DNA polymerase stops, dissociates from the template DNA molecule and the posterior reannealing is done, not in the exact point where DNA synthesis stopped but in another close point in the same replication fork (typically with homology to the stop point). Depending on where DNA synthesis is resumed, either posterior or prior to the stop point (in the direction of the synthesis), replication slippage leads respectively to deletion or duplication (Figure 1.6).
2. *Fork stalling and template switching* (FoSTeS) also happens after a local stop of DNA synthesis and a dissociation of the DNA polymerase. In this case, the reannealing happens in an homologous fragment located at a distant point in the same chromosome or even in a different chromosome. Depending on its resolution, FoSTeS can result in a duplication with a possible associated deletion (Figure 1.6).

The stop of the polymerase activity during DNA synthesis has been related to the presence of repetitive sequences within a replication fork (Hastings et al., 2009b). These repetitive sequences generate substructures when double stranded DNA is opened for DNA synthesis (Figure 1.6). These substructures difficult polymerase activity resulting in situations like those depicted in Figure 1.6.

Duplication by retrotransposition

Retrotransposons are transposable elements that replicate by transcription, reverse transcription and insertion to a new genomic site (normally by

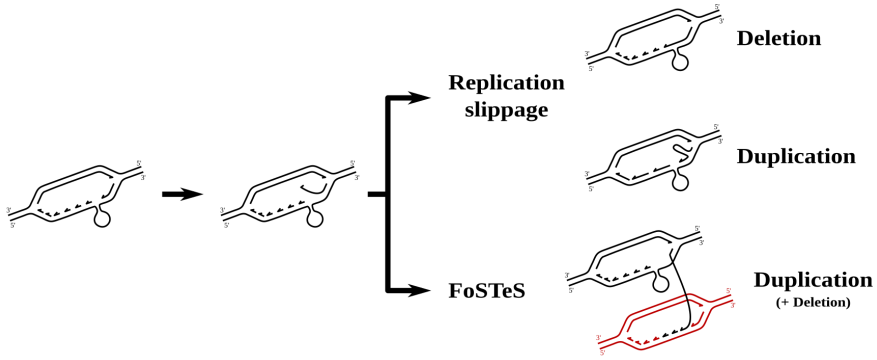


Figure 1.6: Duplication and deletion during DNA replication. In a replication fork, DNA synthesis is stopped frequently due to substructures of single strand DNA caused by local homologies. This stop and disassociation of the DNA polymerase can result in replication slippage leading to either a duplication or a deletion, or in FoSTeS that result in a duplication and a possible deletion. [Image adapted from Hastings et al. (2009b).]

NAHR resolved in crossover; Figure 1.5). There are autonomous retrotransposons that codify for a functional reverse transcriptase and non-autonomous retrotransposons that depend on the former to transpose (Kazazian Jr., 2004). The retrotransposition of a mobile element is itself a duplication, although retrotransposons are frequently not treated as duplications but as repetitive regions due to their typically high number of copies and specific internal content.

Interestingly, because of the reverse transcriptase activity of an autonomous retrotransposon, other cellular single stranded RNA molecules can be retrotranscribed to DNA and potentially be inserted into the genome, again resulting in a duplication (Figure 1.5). These RNA molecules can be mRNAs (processed or not) of expressed genes that, if retrotranscribed and inserted in the genome would result in a gene duplication (Babushok et al., 2007; Kaessmann, 2010; Richardson et al., 2014). Most of these gene duplicates or retrogenes are non-functional gene copies (pseudogenes) because the retrotranscribed and inserted mRNA lacks the promoter and other regulatory regions necessary for its expression (Vanin, 1985). Nevertheless, more than 3,500 functional retrogenes have been identified in

the human genome (Marques et al., 2005; Vinckenbosch et al., 2006). A frequent characteristic of retrogenes (although not ubiquitous) is the absence of introns due to the processed nature of the retrotranscribed and inserted mRNA.

As seen along this section, all known duplication mechanisms are driven by the presence of homologous fragments of DNA that are not in the same locus (*i.e.* non-allelic homologous regions). These non-allelic homologous regions are recurrently involved in duplication generation and can range in size from large duplications, such as SDs, to repetitive sequences, such as retrotransposons, or even smaller stretches of homology (Hastings et al., 2009a).

Close to 50% of the human genome (48.49% in hg38; Smit et al., 2013) is composed of mobile-element derived sequences (including both retrotransposons and DNA elements) that suppose an abundant source of non-allelic homologies for future duplication events. Additionally, in the human genome there is an active autonomous retrotransposon named *long interspersed element 1* (LINE1; Cordaux and Batzer, 2009). It is actually due to LINE1 that there is the possibility of reverse transcription activity in the human genome. Together the availability of non-allelic homologies and the activity of LINE1 make the human genome fertile ground for new duplications.

SDs are also the birthplace of duplications because they are highly similar homologous sequences in different genomic loci (non-allelic homologous regions). The presence of a duplication is a predisposing factor for the appearance of other duplications in the same genomic site. This recurrent duplication process leads to *duplication shadowing* (Cheng et al., 2005) and is thought to be the cause of the characteristic clusterized distribution of SDs observed in mammalian genomes (Bailey et al., 2001; Perry et al., 2008a,b; She et al., 2008; Marques-Bonet and Eichler, 2009; Ventura et al., 2011; Serres-Armero et al., 2017). In Chapter 5, I will present a genome-wide map of SDs and CNVs in the rhesus macaque (*Macaca mulatta*) genome showing this same clusterization pattern.

SDs are often not isolated in a given genomic site but organized in regions with mosaic patterns of duplications (Jiang et al., 2007; Figure 1.7). These mosaic regions are product of several rounds of duplications and often also involve

deletions, inversions and other chromosomal rearrangements. Mosaic regions can be decomposed in segments, termed *duplicons*, which retain homology to other segments in the genome (Figure 1.7). Some of these duplicons are found in multiple duplication mosaic regions across the genome and are then termed *core duplicons* (Jiang et al., 2007; Marques-Bonet and Eichler, 2009). In the human genome, the recurrence and organization of these core duplicons in several mosaic SD regions denotes their participation in multiple rounds of NAHR and other duplication mechanisms (Marques-Bonet and Eichler, 2009). In Chapter 4, I will present a study of the different types of SDs in the human genome including mosaic duplicated regions.

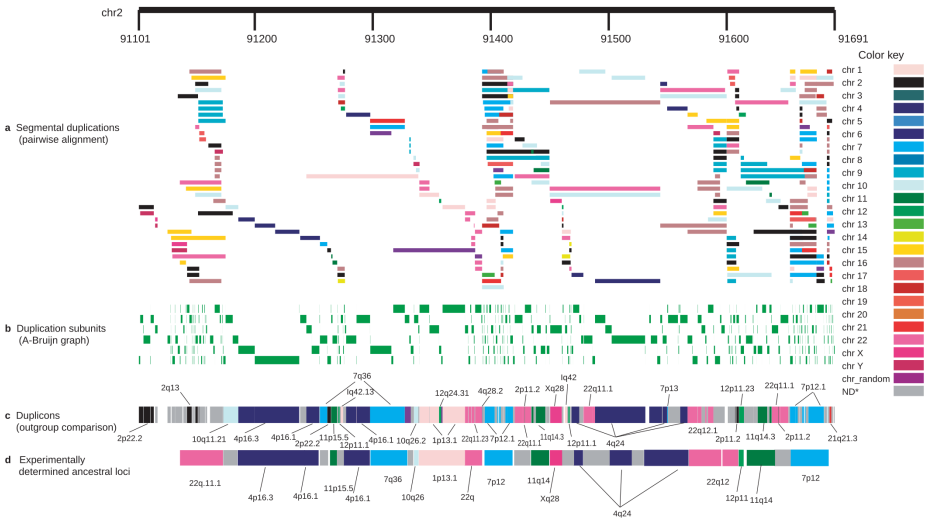


Figure 1.7: Complexity representation and fragmentation to basic ancestral units of a big region formed by mosaic SDs in the human chromosome 2. A. SDs are represented in colored rectangles along the 2p11 genomic region. Color code corresponds to the chromosome where the copies are located. Duplication subunits are indicated in B and organized in ancestral original sequence segments or duplicons in C. These results were validated with FISH (D; see Section 1.2.2). [Figure from Jiang et al. (2007).]

1.3.3 Fates of duplicates

Ohno's theory on the creation of new genetic material through gene duplication (Ohno, 1970) was based on gene duplications because he understood the gene as the functional unit in the genome. We know today that duplications do not necessarily involve full functional units. In fact, duplications normally happen independently of the location of functional units (see Section 1.3.2). For the sake of simplicity, in this section I will introduce existing models of evolutionary fates of duplications encompassing full functional elements, including their regulatory regions. In any case, these models could be adapted to partial duplications of functional units although general models might be difficult to develop.

Selection on the number of duplicates

A duplication of an entire gene increases its gene dosage. However, this may not have the same consequences for all genes (Innan, 2009). In this regard, we can roughly distinguish three types of genes. First, those genes in which increased dosage is favored. In these cases, natural selection will promote the increase in the frequency of the duplication within the population. If more duplications appear, they will also potentially be advantageous and, thus, promoted to fixation. These include genes for which more product is favored, and genes for which multiple copies are favored by diversifying selection (see below). Genes coding for structural and regulatory proteins have been found to be frequent in this category (Kondrashov and Koonin, 2004). Second, those genes for which duplicated copies are neutral (or nearly neutral). In these cases, gene duplication will segregate neutrally in the population and either reach fixation or disappear. This situation is common in genes coding for enzymes (Kondrashov and Koonin, 2004). Finally, those genes whose duplicated copies are deleterious, for example, when an increase in gene dosage creates a deleterious imbalance of gene-product concentration. In these cases, purifying selection disfavors the presence of the duplication in the population. This situation is the genetic basis of many diseases caused by CNVs (see Section 1.5).

Selection on the sequence of duplicates

Besides of its influence on the number of copies, natural selection is also sensitive to the specific content of the duplication. Selection on duplicated sequences is not independent of selection on the number of copies and, in fact, they frequently interplay. Several models have been proposed for the action of selection based on the content of gene duplicates (Innan, 2009; Innan and Kondrashov, 2010):

- *Pseudogenization* happens when one of the duplicated copies loses its function (due to inactivating mutations, for example), becoming a pseudogene. Once inactivated, the pseudogenized copy will, in principle, segregate neutrally in the population. A pseudogenized copy of a gene can act as a reservoir for genetic diversification of this gene through IGC (Hayakawa et al., 2005; Chen et al., 2007).
- In the *more of the same* model (referred to above as cases of positive selection on increased gene dosage), purifying selection acts upon both duplicates at the same time because having two copies of the same gene is advantageous. In this situation, natural selection disfavors mutations appearing in either one or the other duplicates resulting in a conservation of the identity between duplicates or, in other words, their concerted evolution (Samonte and Eichler, 2002; Hess et al., 2018; see Section 1.3.1).
- *Neofunctionalization* is the case that Ohno envisaged. It occurs when, after duplication, a beneficial mutation appears in one of the duplicates, thus changing its function to a new function that is favored by selection (Assis and Bachtrog, 2013; Qian and Zhang, 2014; Renaud et al., 2014). In this model, selection promotes the fixation of the new mutation (and the corresponding duplication if not previously fixed).
- *Subfunctionalization* is a flexible version of neofunctionalization. It happens when the original function of the gene previous to its duplication is split between the duplicates (Marques et al., 2008; Proulx, 2012; Lan and Pritchard, 2016). At the beginning both genes retain the original function but, with time, each gene gets specialized in one of its particular aspects (*e.g.* substrate, tissue or cellular

specialization). Examples of more specific subfunctionalization models are *duplication-degeneration-complementation* (Force et al., 1999) in which there is a degeneration of the two genes and both are finally necessary to perform the original function; and *escape-from-adaptive-conflict* (Des Marais and Rausher, 2008) where the original gene had two or more subfunctions but selective constraints restrained it from specializing in any of them. Under this last model, a duplication resolves the constraint and each duplicate specializes in one subfunction.

- *Multiallelic diversifying selection* happens when high levels of diversity are favored in a given gene. In this case, having multiple copies of the gene is advantageous because they can code for higher variety of gene products. This type of selection has been seen in gene fragments and protein subunits. The light chain of immunoglobulins and the zinc-finger array of PRDM9 are examples of this type of selection (Darlow and Stott, 2006; Buard et al., 2014).

Independently of specific models of selection on duplicates, positive selection has been measured in recently duplicated genes in mammals (Han et al., 2009). Moreover, several authors have found an increase in the evolutionary rate after duplication, either in one of the copies or in both of them, followed by a posterior return to pre-duplication evolutionary rates in rodent and great ape genes (Pegueroles et al., 2013; Pich i Rosello and Kondrashov, 2014). The first acceleration is believed to have been due to a relaxation of purifying selection and putative positive selection acting on new mutations in gene duplicates. A posterior recovery of the selective constraints is proposed to explain the return to pre-duplication evolutionary rates. Another study observed higher expression levels in recently duplicated genes compared to non-duplicated genes, suggesting an important role of gene dosage just after duplication (Vinogradov, 2012). Concerning gene dosage of duplicated genes, tandem duplicates in placental mammals have decreased expression levels that match the expression levels of the original gene prior to duplication ensuring their persistence through time and allowing for their later innovation (Lan and Pritchard, 2016). All these observations fit with the aforementioned models of selection on duplicates and clearly indicate an active role of duplication in the evolution of genomes, just like Ohno predicted back in the 1970s.

... what we know is really very, very little
compared to what we still have to know.

Fabiola Gianotti

1.4 What do we know about primate SDs and CNVs?

SDs and CNVs are of extreme relevance in human genetics for several reasons: first, because SDs represent around 5% of the human genome (Bailey et al., 2001) and around 7% of it is variable due to CNVs (Sudmant et al., 2015b); second, because SDs and CNVs are among the genetic basis of certain human disease (see Section 1.5); third, because human SDs contain genes and are enriched in exons (Bailey et al., 2002a; Samonte and Eichler, 2002; Zhang et al., 2005; She et al., 2006; Sudmant et al., 2013; Dennis et al., 2017); and fourth, because there was a burst of duplication activity in the great ape lineage leading to humans (Marques-Bonet et al., 2009a; Sudmant et al., 2013).

In 2009, Marques-Bonet et al. (2009a) compared the amount of human SDs shared with chimpanzees, gorillas, orangutans and rhesus macaques. They calculated the duplication rate per million year and found that it was exceptionally high in the time of the common ancestor of humans, chimpanzees, bonobos, and gorilla and was still high in the human, chimpanzee and bonobo common ancestor (Figure 1.8, Figure 1.9). In other words, humans, chimpanzees, bonobos and gorillas share more SDs than expected. Later studies gave further support to this finding (Gazave et al., 2011; Sudmant et al., 2013). Given the role of duplications in genome evolution (see Section 1.1 and Section 1.3.3), one can speculate that these new SDs may have resulted in new functions in these species. In fact, in Lorente-Galdos et al. (2013) the authors find accelerated evolution of some exons located in human and rhesus macaque SDs.

For all the reasons mentioned above, human SDs and CNVs have been extensively studied during the last decade. After the release of the first draft of the human genome, great effort was dedicated to understanding the diversity and organization of SDs (Bailey et al., 2001; Eichler, 2001; Bailey et al., 2002a;

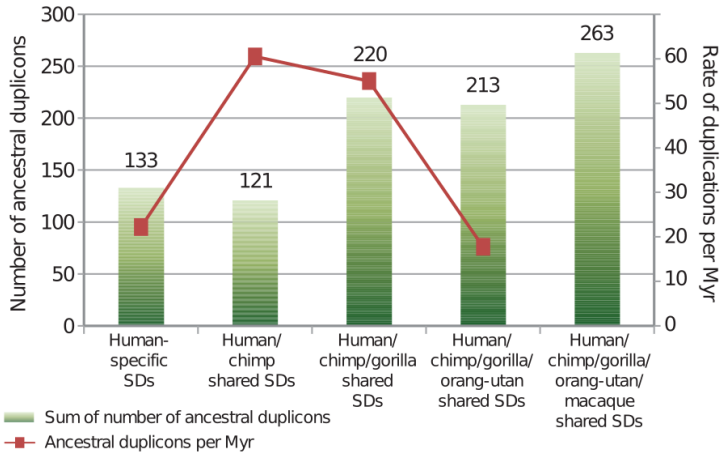


Figure 1.8: Burst of SDs in the great ape phylogeny. Number of duplication events and duplication rate per million years (Myr in the image) in the great ape phylogeny based in array-CGH data. [Image adapted from Marques-Bonet et al. (2009a).]

Samonte and Eichler, 2002; Bailey et al., 2003; Bailey and Eichler, 2006; Locke et al., 2006). Later, the discovery of the burst of duplications (Marques-Bonet et al., 2009a) and the persistent missing heritability problem (Eichler et al., 2010) brought more interest to the comparison of SDs and CNVs among great ape genomes (Gazave et al., 2011; Lorente-Galdos et al., 2013; Sudmant et al., 2013; Kronenberg et al., 2018). More recently, the NGS technologies and the access to more human *whole-genome sequencing* (WGS) data have permitted a deep characterization of human CNVs (Sudmant et al., 2015a,b; Dennis et al., 2017). Despite this, still many issues on the burst of SDs in great apes remain unknown. What triggered it? Which duplication mechanisms were involved? What kind of SDs appeared? Are there new functional regions in the human genome result of the great ape burst in SDs? In Chapter 4, I will address some of these questions.

Nevertheless, excluding the great apes, which have been mostly studied in comparison to humans (Cheng et al., 2005; Perry et al., 2006; Marques-Bonet et al., 2009a; Gazave et al., 2011; Ventura et al., 2011; Lorente-Galdos et al.,

Section 1.4

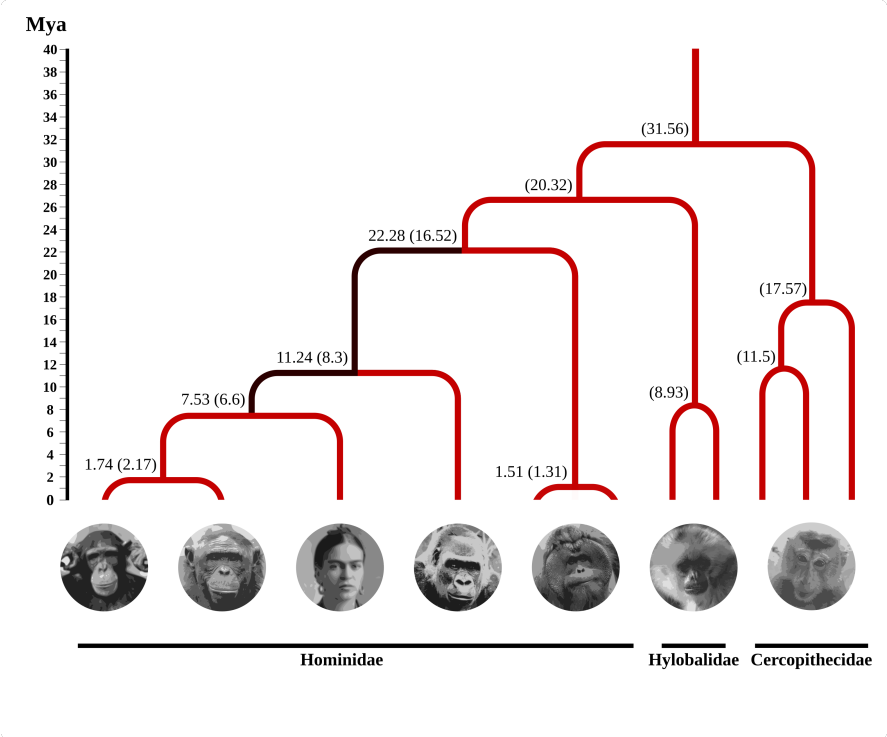


Figure 1.9: Burst of SDs contextualized in the Hominidae (great ape), Hylobatidae and Cercopithecidae phylogeny represented in an orientative time scale focalizing in the great ape phylogeny. Divergence times between lineages in the tree correspond to the estimated diverge times in Prado-Martinez et al. (2013) and in Perelman et al. (2011) (in parentheses). Tree fragments known to be involved in the burst of SDs in the great ape lineage are depicted in black.

2013; Sudmant et al., 2013), little is known about the other primates' SDs and CNVs (Lee et al., 2008; Gschwind et al., 2017). Rhesus macaque is a species of the family Cercopithecidae (Old World monkeys; Figure 1.9) that is mainly known for its large geographic range and its use in biomedical research as a model organism. It is employed as a model organism because of its physiological and genetic closeness to humans (Xue et al., 2016; Figure 1.9) and has been extensively used to understand human genetic disease (Gibbs et al., 2007; Vallender et al., 2008, 2010; Valentine et al., 2009; Rogers et al.,

2013; Vinson et al., 2013; Madlon-Kay et al., 2018). Still, there is limited knowledge regarding SDs and CNVs in macaques. Rhesus macaque SDs and CNVs studies are limited to a comparison with the great apes (Marques-Bonet et al., 2009a; Gokcumen et al., 2011; Lorente-Galdos et al., 2013) and an array CGH (focusing in CNVs) analysis with a low sample size (10 individuals; Lee et al., 2008). These studies show general similarities between human and rhesus macaque SDs and CNVs besides the burst of SDs in great apes (Marques-Bonet et al., 2009a; Gokcumen et al., 2011), although some functionally important specific differences were also detected by Lee et al. (2008). Considering the important role of duplications in shaping the human genome and the relevance of SV in great ape evolution, it seems appropriate and necessary to consider copy-number differences between humans and rhesus macaques if rhesus macaque is to be used as a model organism to study the genetic basis of human disease. In Chapter 5, I will present a genome-wide map of the SVs in the rhesus macaque genome with a sample size of close to 200 individuals, with special attention paid to differences in number of copies between human and macaque genes.

... evolution is a tinkerer, an ad-hocker, and a jury-rigger. It works with what it has on hand, not with what it has in mind. Some of its inventions prove elegant, while in others you can see the seams and dried glue.

Natalie Angier

1.5 Implications in disease and phenotype

As already mentioned, SDs and CNVs can be the genetic causes of certain human diseases (Beckmann et al., 2007; Higgins et al., 2008; Zhang et al., 2009; Casola et al., 2012b). Disease associated to SDs and CNVs can be caused by three main mechanisms: by generation of aberrant forms of a genomic functional structure (*e.g.* by disrupting a gene); by the creation of a gene imbalance; or by the transfer of disease-causing mutations between duplicates through IGC.

A new duplication can potentially result in disease causing anomalous functional units (see Section 1.3.2). The insertion of a retrotransposon, for example, can interrupt the sequence of a gene, or a replication slippage event can duplicate an important part of it leading to abnormal forms of the corresponding gene product (Stankiewicz and Lupski, 2002, 2010; Girirajan et al., 2011; Almal and Padh, 2012; Carvalho and Lupski, 2016).

Even if a given duplication does not result in aberrant forms of particular functional molecules, it can alter the level of expression of a gene (or genes) by changing its number of copies or altering its regulatory regions. An imbalance in gene expression of particular genes can cause disease (see Section 1.3.3). Moreover, a duplication can affect more than one contiguous gene resulting in multiple gene dosage imbalances (Stankiewicz and Lupski, 2002, 2010; Girirajan et al., 2011; Almal and Padh, 2012; Carvalho and Lupski, 2016).

Additionally, IGC (see Section 1.3.1) between a gene and its duplicate (either pseudogenized or not) can transfer mutations from one to the other, leading to frameshifts, splicing modifications, nonsense or missense mutations, among other types of alterations in the acceptor gene duplicate. Neutral or even beneficial mutations in one gene copy might be inherited from generation to

generation without showing any effect (especially if the gene copy is pseudogenized), but can lead to a disease if transferred to the other gene copy. In fact, recurrent *de novo* disease-causal point mutations observed in unrelated trios are candidates to be produced by IGC (Chen et al., 2007, 2010a).

The presence of CNVs and SDs has been linked to a huge variety of diseases and phenotypes including schizophrenia, autism, mental retardation, Parkinson's disease, Alzheimer's disease, epilepsy, systemic lupus erythematosus, Crohn's disease, pancreatitis, susceptibility to HIV and other infectious diseases, and differences in drug metabolism (for specific references, refer to Stankiewicz and Lupski, 2002, 2010; Almal and Padh, 2012). In addition, somatic copy-number variation has a central role in another highly prevalent disease: cancer (Chen et al., 2010a; Yang et al., 2013; Tubio, 2015).

Chapter 2

Objectives

This thesis aims to shed light on some of the aspects mentioned in the previous chapter. To do so, I have structured the specific objectives in:

General:

1. To contribute to a better understanding of the way duplicates undergo concerted evolution and its consequences.
2. To contribute to understand the duplication content, diversity and evolution of the duplications in the human and rhesus macaque genomes.

Specific:

1. To contribute to a better understanding of how specific crossover distributions in duplicated regions interplay with IGC affecting the concerted evolution of duplications.
2. To study the effects of IGC dependence on sequence similarity between duplicates on their concerted evolution.
3. To determine how IGC and collapsed duplications alter summary statistics and might confound results in genome-wide selections scans.
4. To identify and characterize different types of human SDs according to their genomic distribution, relative location and distinct features.
5. To provide insights into how and when different types of human SDs arose, and the way in which they have evolved since then.
6. To detect and characterize SDs and CNVs in the rhesus macaque genome.
7. To identify genes with relevant copy-number differences between human and rhesus macaque genomes with possible implications in biomedical research.

Chapter 3

Understanding neutral concerted evolution in segmental duplications

Basically, I have been compelled by curiosity.

Mary Leakey

...the more diversity you have in a field means more ways of solving a problem as well as more creativity and originality.

Eileen Pollack

3.1 Rationale

Duplications are a key element of genome evolution (see Section 1.1). They are an open door to genetic innovation and to the rise of functional novelties. This gives us a strong motivation to investigate how they evolve and how these new functions originate. Duplications can evolve together, *i.e.* in concert, which has huge implications in their functionality (see Section 1.3.1). Therefore, to be able to understand the way in which duplications give rise to new functions we need to understand the way in which this concerted evolution of duplicates happens.

There are several mechanisms through which duplicates can undergo concerted evolution. The main one, as explained in Section 1.3.1, is through a mechanism of NAHR named IGC. Several models of concerted evolution of duplicates through IGC have predicted alterations on basic sequence properties of duplicates undergoing IGC under neutrality (Ohta, 1982, 1983; Walsh, 1987; Innan, 2003a,b; Thornton, 2007). Despite major progress arising from these models in understanding the molecular evolution of duplicates undergoing IGC, many issues remained unaddressed and unstudied. After some years of work, we have contributed to the comprehension of such issues. Such is the main leitmotif of this chapter.

In this section, I will review the current knowledge that we have of the molecular evolution of duplications while presenting and interpreting our contributions to the field. Most of the results included in this section are part of the three published pieces of work added as supplementary materials to this thesis. This is not meant to be an exhaustive compilation of results and methods.

Section 3.1

My objective here is to present our results in a global, digested and ordered manner. For further information, please refer to the corresponding original published articles Hartasánchez et al. (2014, 2016, 2018) in Appendices 7.1, 7.2 and 7.3.

3.2 Objectives

General:

1. To contribute to the understanding of the neutral evolution of duplications.

Specific:

1. To simulate the evolution of duplicates undergoing IGC to validate variation and *linkage disequilibrium* (LD) expectations of previous models and explore and comprehend the influence of IGC in local maps of LD.
2. To understand the effects of specific configurations of crossover on the influence of IGC on sequence variation and LD.
3. To model IGC dependence on sequence similarity. Simulate and understand concerted evolution under this more complete model of IGC and explore its effect on local IGC rate, sequence variation and LD.
4. To describe deviations in summary statistics used to test for neutrality in duplications and collapsed duplications and determine if they may be confounded by or with selection.

Section 3.3

3.3 Results and Discussion

Immediately after a duplication event, duplicates are identical. If the duplication persists in the population, with time, a point mutation may appear in a given site in one of the duplicates of a given individual in the population having the duplication. This mutation will represent a difference between the two duplicates in such individual. If the individual has offspring, the difference can pass to the next generation and, potentially, increase in frequency within the population. But if an IGC event happens between the duplicates including the region where the difference is located, such event will erase the difference between copies either by restoring the original variant still present in the other duplicate or, and more interestingly, by transferring the new variant to the other duplicate. If the second case happens, the new variant will become a segregating site, not only in the duplicate where it first arose, but also in its paralogous copy. In time, many more mutations will appear and, if IGC is active, duplicates will share variants segregating in both, the original and the duplicated copy, at the same time. Thus is the way in which SDs share information and undergo concerted evolution.

The process of concerted evolution implies a trade-off between the rate with which mutations appear (mutation rate; μ) independently in each copy, increasing the differences between them, and the rate with which IGC erases these differences between duplicates (if there are any). The appearance and the disappearance of differences can reach an equilibrium in which there is a relatively constant number of differences between duplicates that endures through time. I will refer to this equilibrium as an *equilibrium in concerted evolution* or *concerted evolution equilibrium* (see Section 1.3.1).

It is known that the concerted evolution between duplicates due to IGC affects not only divergence between duplicates but also diversity, LD and other basic molecular properties of genomic regions (Innan 2009; Ohta 2010; see Section 1.3.1). Moreover, IGC is known to interact with other biological processes such as crossover between duplicated copies and to depend on other factors, mainly the sequence divergence between duplicated regions. Here I will explore and discuss all of these aspects of the concerted evolution between duplicates due to IGC.

All of the results presented in this chapter were obtained using forward-in-time

simulations of the neutral sequence evolution of duplications undergoing IGC. I have extracted and exposed the results in a way that their understanding is independent of the knowledge of the internal structure of the simulations. For further details and clarifications please refer to methods in Section 3.4 for a general contextualization and to Hartasánchez et al. (2014) in Appendix 7.1 and Hartasánchez et al. (2016) in Appendix 7.2 for more extensive methods.

3.3.1 Fundamental effects of IGC

Before diving into the subtler aspects of IGC and its interplay with other phenomena, we need to understand the fundamental effects that IGC has on the divergence between duplicates and on other elemental molecular properties of genomic regions, in this case, diversity in duplicated regions and LD. To do so and for the sake of clarity, in this first section I will go through the effects of IGC in absence of crossover between duplicates. Later on, in Section 3.3.2, I will discuss on the interplay between IGC effects and crossover.

Even though the terms *divergence* and *diversity* are useful because they are intuitive, when dealing with duplications they might be a source of confusion. So, from now on I will use the following terms: π_w or *variation within* a given region of the genome to refer to the within-population diversity present within this region (in our case, regions of interest happen to have a duplicate copy in another region of the genome); π_s or *variation between duplicates in the same chromosome* to refer to the divergence of duplicates located in the same chromosome; and π_b or *variation between duplicates* to refer to the divergence of duplicates in different chromosomes in the population (see Table 3.1 and Figure 3.1 for clarification). π_w, π_s and π_b are calculated by measuring the average pairwise differences statistic (π; Nei and Li, 1979).

π_w	*Variation within a given region of the genome*
π_s	*Variation between duplicates in the same chromosome*
π_b	*Variation between duplicates on different chromosomes*
D_{sum}	*LD between duplicates*

Table 3.1

Tomoko Ohta, already in 1982, modeled IGC happening between two duplicates (what she referred to as small multigene families; Ohta, 1982, 1983). Ohta derived analytical equations of expectations of the three mentioned types of variation at equilibrium. Hideki Innan revisited these analytical solutions in his work of 2002 and, in 2003, presented the coalescent and infinite-site model (Innan, 2002, 2003b). In addition, in his work, Innan derived the expectation for *average LD between duplicates* at equilibrium (Innan, 2003b). Innan used the D_{sum} statistic to measure average LD between duplicates (Table 3.1; Innan, 2003b; Hartasánchez et al., 2014; see methods in Section 3.4).

Figure 3.2 shows the expectation values for all variation measures and D_{sum} at equilibrium according to the models of Ohta (1982) and Innan (2003b) in absence of crossover (R = 0).

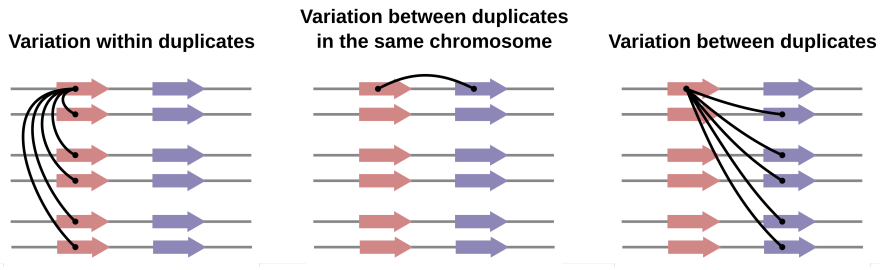


Figure 3.1: Variation measures within and between two duplicates undergoing IGC. Six chromosomes (three diploid individuals) are represented. Colored arrows represent paralogous duplicated regions undergoing IGC. Black lines exemplify the duplicates compared to the first duplicate of the first chromosome in each case. [Image inspired by Ohta (2010) and Hartasánchez et al. (2014).]

Predictions for variation between duplicates (π_s and π_b) at equilibrium in absence of crossover between duplicates are quite intuitive (Figure 3.2 A green lines). On the one hand, when IGC is very low, homogenization of duplicates is rare and π_s and π_b are expected to be very high. On the other hand, when IGC is very high, duplicates behave as one and harbor almost identical variation. In the latter situation, π_b approximates the neutral expectation of variation in a population at equilibrium (*i.e.* Θ, the neutral parameter of molecular evolution) and π_s approximates 0 as it does not take into account variation in other chromosomes in the population (in absence of crossover).

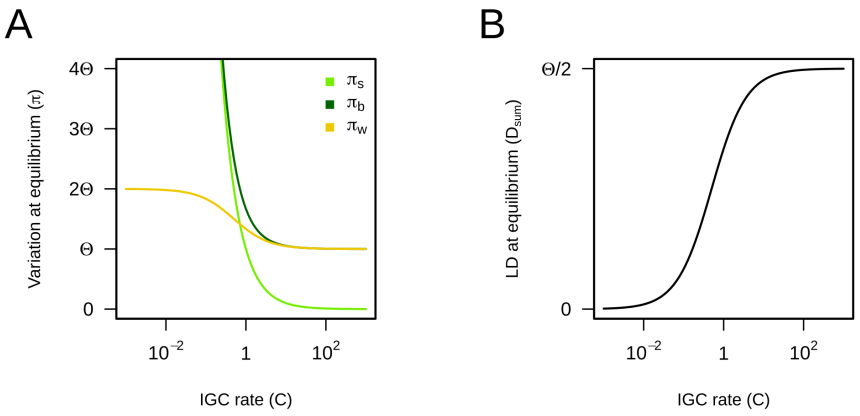


Figure 3.2: Theoretical expectations for π_s, π_b, π_w and D_{sum}. π_b, π_w and D_{sum} expectations are from Innan (2003b) and π_s expectation corresponds to the analytical solutions of Ohta (1982) since Innan did not use this measure in his work. All cases are in absence of crossover (R = 0). [Figure inspired by Hartasánchez et al. (2014).]

The impact of IGC on variation within duplicates (π_w) at equilibrium is neither as straightforward or well known as its impact on variation between duplicates (Figure 3.2 A yellow line). Variation within duplicates is expected to approximate Θ when IGC rate is very high because all the mutations that appear in any of the two copies are either erased or copied to the other copy immediately (again, duplicates behave as one). However, when IGC is less frequent, equilibrium is attained with non-zero variation between duplicates. In this case, variants are occasionally transferred from one copy to the other generating new segregating sites and, thus, more variation within each one of the copies.

Variation within duplicates can be up to 2Θ for the case of two duplicates of the same length. This is because mutations appear along the length of the two duplicates (two times the length of each duplicate) and can potentially be transferred to the other duplicate through IGC. This can cause an effect in variation within each duplicate as if the mutation rate or the population size were up to twice as much as the real one.

For very low IGC rates, Innan's expectations of π_w are not realistic since the time to reach equilibrium in concerted evolution with such IGC rates is extremely long (Nagylaki, 1984). Moreover, alike all HR mechanisms, IGC needs a certain level of sequence identity between duplicates to be able to act (see Section 1.3.1). For sufficiently low IGC rates this premise is possibly not satisfied and thus, equilibrium in concerted evolution may not be attained in such conditions (see Section 3.3.3 for modeling, results and discussion on this issue). For these two reasons, in practice, the increase in π_w at equilibrium due to IGC activity will only be observed for a reduced window of IGC rates (see Section 3.3.3).

IGC, like crossover, alters LD in the genomic regions where it is active. In particular, IGC rate affects average LD between duplicates (see Table 3.1) even in absence of crossover (see Figure 3.2 B). As one might expect, average LD between duplicates increases when IGC is very active and paralogous variants segregate in a coordinated manner in both duplicates. Specifically, D_{sum} grows up to $\Theta/2$ when IGC rate is high enough (Figure 3.2 B). On the contrary, when IGC rate is low, variants are not expected to segregate in a coordinate manner and LD between duplicates tends to 0. Long distance LD can be misinterpreted as a signal of epistasis, drift or selection (Wei et al., 2014). In the case of duplicates undergoing IGC, LD can be high under neutrality and should not be misinterpreted as having been caused by epistasis, drift or selection.

Innan's LD measure (D_{sum}) was designed to summarize LD between all paralogous sites in one single statistic. Inspired by this measure, in Hartasánchez et al. (2014), we suspected that IGC could leave a special imprint in local LD maps. Figure 3.3 shows the LD pattern product of IGC in absence of crossover. When IGC is active, a horizontal line with higher LD values appears connecting all paralogous windows. LD between paralogous regions in duplicates increases with IGC rate. Additionally, intermediate levels of IGC break short distance LD within duplicates (see C = 1 in Figure 3.3). This decrease of LD within duplicated regions does not happen under high rates of IGC (see C = 50 in Figure 3.3).

In this section, I have reviewed the basic properties of the evolution of duplicates under simple IGC models. In the following sections I will discuss how these basic properties change under more realistic scenarios. In particular, in Section 3.3.2, I will explore the influence that crossover has on IGC when they both act in the same context.

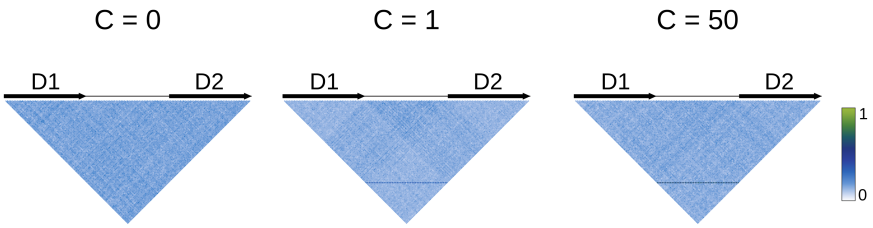


Figure 3.3: Patterns of local LD between duplicates. Three different IGC rates are represented (C = 0, 1, 50) from left to right. Crossover is absent in all three cases (R = 0). Black arrows indicate where duplicates are located along the sequence. Color codes average D' values of 1000 simulations for each pair of non-overlapping 100 bp windows. [Figure adapted from Hartasánchez et al. (2014).]

3.3.2 Interplay of IGC and crossover

Early models of the evolution of duplicates already described that the impact of IGC on region properties is modified by the effect of crossover acting between duplicates (Ohta, 1982, 1983). Although crossover is known to happen genome-wide and especially in crossover hotspots, all previous models of crossover interacting with IGC only considered IGC happening between duplicates and never within them (Ohta, 1982, 1983; Innan, 2002, 2003b; Thornton, 2007). Literature on this matter describes a tremendous influence of crossover on concerted evolution due to IGC. Acknowledging this fact and considering that other configurations of crossover happening in the context of IGC had not been explored, in Hartasánchez et al. (2014), we decided to tackle this issue and further explore the impact of crossover on IGC.

For consistency, in Hartasánchez et al. (2014), we recovered the model of crossover used in the work cited above, which I will refer to here as *crossover between duplicates model* (see SCC model in Hartasánchez et al., 2014). In this model, crossover happens in the region between duplicates (Figure 3.4 A).

In order to account for the variety of duplications found in eukaryotic genomes (see Chapters 4 and 5), in Hartasánchez et al. (2014) (in Appendix 7.1) we modeled other scenarios of crossover acting within and between duplicates undergoing IGC (Figure 3.4; see Chapter 4). In particular, we considered three

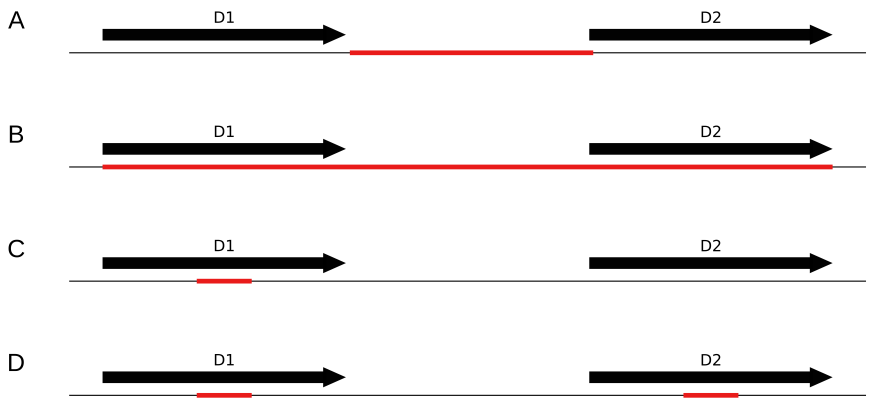


Figure 3.4: Crossover models. Black arrows represent duplicates, red region indicates where crossover is active. (A) Crossover between duplicates model, crossover happens only between duplicates; (B) whole-region crossover model, crossover happens within and between duplicates; (C) crossover hotspot model, crossover happens in a crossover hotspot, in this particular example with a hotspot located in the middle of one of the duplicates (see Hartasánchez et al., 2014 in Appendix 7.1 for further versions of this model); (D) double crossover hotspot model, crossover happens in two crossover hotspots located in paralogous regions within duplicates.

additional crossover models. First, the *whole-region crossover model*, in which crossover happens with a constant rate per bp along the whole region (Figure 3.4 B). Second, the *crossover hotspot model*, in which crossover happens only in a defined crossover hotspot. It is known that, in many species, including all yeast, plant and vertebrate species studied to date (Baker et al., 2017), crossover happens mainly in crossover hotpots in specific genomic sites. And, third, the *double crossover hotspot model* (Figure 3.4 D). Since crossover hotspots are characterized by the presence of a given sequence motive (Myers et al., 2008), one could consider that, if variation between duplicates is low, when a crossover hotspot is present in one of the duplicates, it might also be present in the same paralogous position in the other duplicate.

Figure 3.5 summarizes the interplay between IGC and crossover under different crossover rates and the four considered crossover models. As observed in previous work (Ohta, 1982; Innan, 2003b), in the crossover between duplicates model, crossover increases variation within and between duplicates and

decreases LD (solid lines and squares in grades of green in Figure 3.5 A). This is, if without crossover IGC increases variation between and within duplicates and decreases LD between duplicates, with crossover between duplicates, the increase in variation (between and within duplicates) and the decrease in LD between duplicates by IGC are larger. In other words, in the presence of crossover, the effect of IGC is boosted.

In Hartasánchez et al. (2014), we observed that for the same crossover rate, the whole-region crossover model had less impact in the IGC consequences compared to the crossover between duplicates model (diamonds in Figure 3.5 A). In fact, we noticed that whole-region crossover simulations fit the curve for $R' = \frac{2R}{3}$ in the crossover between duplicates model (dashed lines in Figure 3.5 A). In order to understand this observation, we must understand how the interplay between crossover and IGC happens.

If an IGC event overwrites an already converted pair of paralogous sites (having the same variant), it will have null effects in these sites. Otherwise, if a pair of already converted paralogous sites gets separated by crossover, there is a chance that, at least one of them, falls in a chromosome having a different variant in the other copy. In such case, this pair of paralogous sites in the new chromosome will be a potential substrate of further IGC events erasing the difference. It is by separating paralogous sites that crossover has an impact on the effect of IGC.

A crossover happening in the region between duplicates (Figure 3.6 A) will always separate all pairs of paralogous sites. On the contrary, when a crossover falls within one of the duplicates, it will only separate paralogous sites external to the crossover point (Figure 3.6 B, C and D). Sites towards the outer part of the region will be separated from their paralogous sites by crossover more frequently than sites in the inner part of the duplicate. When duplicates are equally oriented (as in A, B and C of Figure 3.6), the sites that are in the outer part in one duplicate are in the inner part in the other duplicate and, if crossover happens equally in both duplicates, differences between sites are compensated. In the end, all paralogous pairs have the same probability of being separated by crossover.

Given that only crossover events separating paralogous sites modify the effect of IGC, we can estimate the IGC-boosting effective crossover rate (rate at which a crossover separates paralogous sites; R'). For example, in the case of the whole-

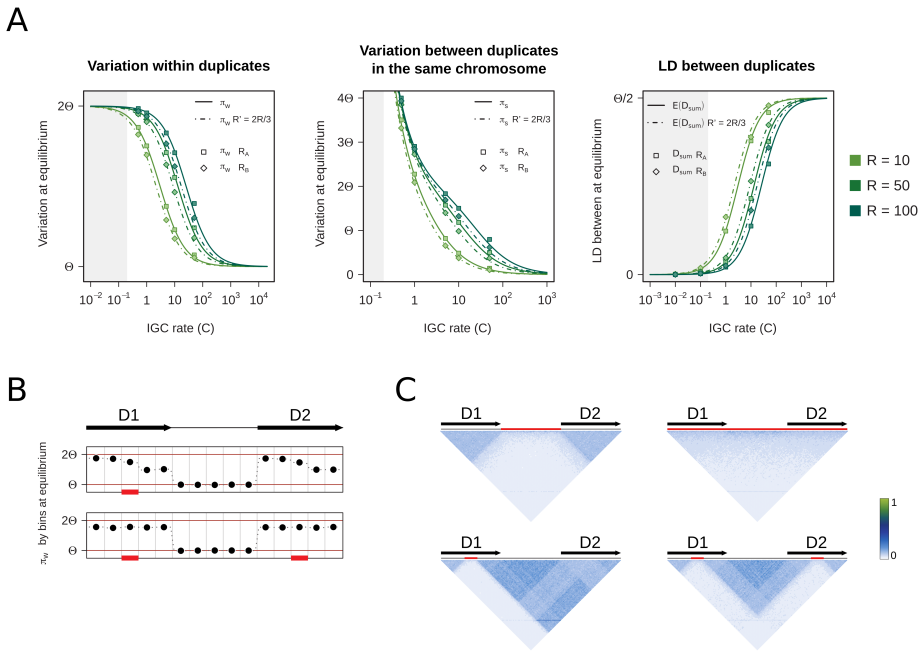


Figure 3.5: Interplay of IGC and crossover under four crossover models. (A) Lines represent theoretical expectations of π_w, π_s and D_{sum} (Innan, 2003b; Ohta, 1982 and Innan, 2003b respectively) under six different crossover rates under the crossover between duplicates model; R = 10, 50 and 100 (solid lines) and the corresponding $R' = \frac{2R}{3}$ (dashed lines). Points correspond to the average simulated values for 10,000 simulation runs under the crossover between duplicates model (squares; R_A) and the whole-region crossover model (diamonds; R_B). Theoretical expectations for the crossover between duplicates model with $R' = \frac{2R}{3}$ fit simulations of whole-region crossover model with crossover rate equal to R. Grey area ($C < 0.2$) indicates the IGC rate values for which, according to Walsh (1987), the prevalence of concerted evolution is not guaranteed and, thus, predictions might not be realistic (see Section 3.3.3). (B) π_w is calculated in bins (of 1,000 bp) along the duplicates and the region in between and represented in points. Black arrows indicate duplicate positions. Crossover has a rate of R = 10 and is reduced to the area indicated in red (crossover hotspot model and double crossover hotspot model). (C) Local patterns of LD (D') under the four crossover models. Average D' values of 1,000 simulations for each pair of 100 bp bins are coded in color. Black arrows indicate the position of duplicates. Regions undergoing crossover are depicted in red. C = 1 and R = 50 in all cases.

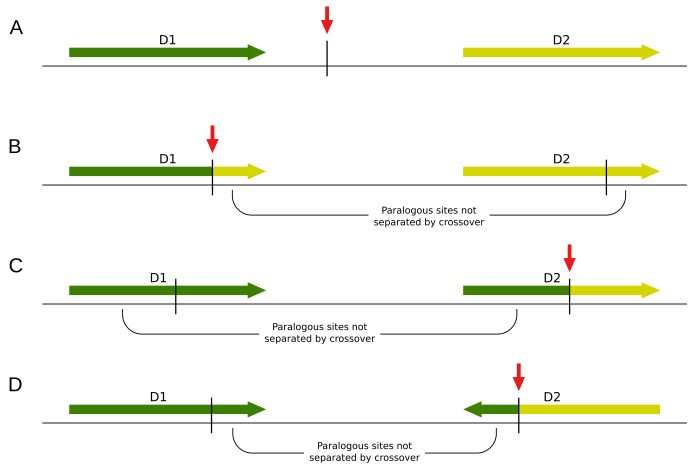


Figure 3.6: Crossover separating or not paralogous sites. Within a duplicate, each site has its paralogous site in the corresponding site in the other duplicate. When crossover (red arrow) happens in the region between duplicates (A) all the sites along each of the duplicates are separated by crossover from their corresponding paralogous sites (different shades of green in the figure). When crossover happens within a duplicate (B, C, D), not all sites get separated from their paralog site, only the ones external to the crossover point. The ones that are internal to the crossover point remain in the same chromosome than their paralogous site in the other duplicate (same shade of green in the figure). In the case where duplicates are in the same direction (A, B and C), crossovers happening in one duplicate compensate the effect from the ones happening in the other duplicate. This results in all the sites along duplicates having the same probability of being separated from their paralogous site. This probability is equal to 1/2 of the times a crossover falls within one of the duplicates. Otherwise, when duplicates are in inverse orientation (D), sites in the internal part of the region have less probability of being separated from their paralogous site than sites in the outer part of the region although the average number of times a crossover separates a site from its paralogous site remains the same.

region crossover model, if we know the duplicates' lengths, the distance between them and the crossover rate (R), we can extract an equation to calculate the rate at which a crossover will separate two paralogous sites:

$$R' = \frac{(\frac{L_A+L_B}{2} + L_{A\to B})R}{L_A + L_B + L_{A\to B}} \tag{3.1}$$

where L_A and L_B are the duplicate lengths (normally $L_A \simeq L_B$) and $L_{A \to B}$ is the distance between duplicates. R' represents IGC-boosting effective crossover rate for the case of the whole-region crossover model. In this model, when a crossover happens in a random position within either one of the two duplicates, it has 50% chances of separating a given site from its paralogous site, thus, only half of the crossovers happening within the duplicates actually separate a given site of its paralog site.

Once we have R' we could use it on the equations of different parameter expectations for the between duplicates crossover model (Ohta, 1982; Innan, 2003b; Thornton, 2007) to obtain the predictions for the whole-region crossover model. In the simulated model, duplicates are separated by a distance equal to their size and, thus, curves for $R' = \frac{2R}{3}$ fit the whole-region crossover model.

Notice that, in Equation 3.1, when distance between duplicates is very large compared to their length ($L_{A \to B} >>> L_A$ and L_B), then R' tends to R because a huge majority of crossovers will occur in the region between duplicates and not within them. In this case, the crossover between duplicates model is quite realistic. On the contrary, if the length of the duplicates is comparable to or higher than the distance between them, R' cannot be approximated to R. In such cases, the effect of crossover occurring within duplicates is not negligible and the crossover between duplicates model is no longer realistic.

The distance between intrachromosomal duplications is, in fact, a distinctive characteristic of specific types of intrachromosomal duplications. For example, tandem duplications, which are a frequent feature of eukaryotic genomes, are duplications that are next to each other (see Section 4.3.1 for further details and discussion on the distance between tandem duplications). In the case of tandem duplications, the crossover between duplicates model is not realistic. On the other hand, non-tandem intrachromosomal duplications, might be good candidates for this crossover model. Nevertheless, if these duplications are large (very common for non-tandem intrachromosomal duplications, see Chapter 4) or contain a recombination hotspot, other crossover models should be considered.

In the case in which duplicates are in inverted orientation (Figure 3.6 D), paralogous sites in the inner part of one duplicate have their paralogous sites also in the inner part of the other duplicate and, thus, less probability of being

separated by a crossover than those in the outer part. This means that the effective crossover rate between duplicates will be higher in the latter. We would therefore expect a diminishment of IGC effects (product of an increase in effective crossover rate) from the internal to the external part of the duplicates. Nevertheless, the average number of times a crossover separates a pair of paralogous sites remains the same and we can apply Equation 3.1 to calculate average R' for the whole region.

In summary, only crossovers that separate paralogous sites have influence in the effect of IGC in such paralogous sites. With this in mind, in the case of the crossover hotspot model, the position of the hotspot and not only its intensity shapes the pattern of variation and LD along the duplicates. Figure 3.5 B (top) shows how the fragment of the duplicate outside the crossover hotspot has more increase in variation within duplicates than the fragment internal to the crossover hotspot. Even more interestingly, the same pattern appears in the other duplicate despite there being no crossover hotspot active within it. This is due to the differences in crossover rate separating paralogous sites between internal and external parts.

With the double crossover hotspot model, all paralogous sites have, again, the same probability of being separated from each other by crossover and the differential impact of IGC due to differences in crossover rate is reversed (Figure 3.5 B bottom). Of course, this happens only in the case of duplicates being equally oriented. In the case of duplicates being inverted, the differential contribution of crossover between the internal and the external parts will be preserved, and the same pattern observed in the single crossover hotspot model will be expected.

Crossover is well known to break LD. As explained in Figure 3.3, IGC also changes local patterns of LD within and between duplicates. In Figure 3.5 C, we see crossover and IGC interplaying to shape local LD. Despite the action of crossover diminishing the effect of IGC, we can see the horizontal line of LD binding paralogous windows characteristic of IGC. Moreover, although crossover has stronger power to break up LD between consecutive regions, IGC also shows certain power to soften short and long-distance LD. This effect can be appreciated in all 4 crossover models, but it is more evident in the hotspot models (bottom).

These results reveal that local patterns of LD are not always and not only shaped by crossover or selection. The presence of duplications undergoing IGC can perfectly construct non-expected patterns of LD, especially in interplay with crossover. Moreover, it is important to consider that the patterns that we see in Figure 3.3 and Figure 3.5 C are the average values for 1,000 simulations. Real measurements of LD in regions undergoing IGC between duplicates might not reflect the signals here described since they will depend on the specific history of IGC and crossover events. In any case, the results here presented imply that unexpected patterns of LD in regions with duplications should not be interpreted as being the consequence of selective pressure acting upon the region.

3.3.3 IGC and sequence similarity dependence and reciprocity

It is well established that for HR to occur, a certain level of sequence similarity (absence of variation) is required between the two DNA fragments involved (see Section 1.3.1). In the case of IGC, its dependence on sequence similarity results on a positive-feedback loop because IGC precisely increases sequence similarity between duplicates. One can imagine fragile feedback dynamics between IGC rate and sequence similarity in which if one exists the other is boosted, but if any of them lacks, the other disappears. This kinetics determines IGC dynamics and its effects.

Knowing the precise mechanism through which HR depends on sequence similarity is, of course, fundamental to understand its consequences on the dynamics of IGC between duplicates. Two different sequence similarity requirements have been proposed to determine the viability of HR. These measures are MEPS and MESH. MEPS is a local threshold of complete sequence identity around the DSB point (Shen and Huang, 1986). MESH is a general threshold of sequence similarity between the involved regions (in our case, duplicates; Chen et al., 2010b). For more information and estimates of MEPS and MESH please, refer to Section 1.3.1.

Although the dependence of HR on sequence similarity has been known for a long time, none of the existing models of IGC deeply explores its dependence on local sequence similarity and the consequences that it has in IGC dynamics

(Ohta, 1982; Walsh, 1987; Innan, 2003b; Teshima and Innan, 2004; Thornton, 2007; Hastings, 2010). Only Walsh (1987) and Teshima and Innan (2004) implemented a simple overall sequence similarity threshold similar to MESH. Under the Teshima and Innan (2004) model, fluctuations of variation between duplicates at equilibrium on concerted evolution reach the limit imposed by MESH at some point. When this happens, IGC stops. These authors implemented this threshold in order to render their model a bit more realistic. Nevertheless, they left the dynamic feedback between IGC and local sequence similarity unexplored.

There are several open questions regarding the implementation of a local sequence similarity threshold such as MEPS. Is it possible to reach equilibrium on concerted evolution with dynamic IGC dependence on local sequence similarity? If so, under what conditions? What would be the variation between duplicates' equilibrium value at every case? Is variation within duplicates also increased in such equilibria? Under what conditions? Which would be its value at equilibrium? Do local patterns of variation and LD change?

In order to elucidate these questions, we designed a model of IGC depending on a fragment of total identity between sequences around the DSB point (MEPS) additional to the general sequence similarity threshold (MESH) already implemented in the previously mentioned models (Walsh, 1987; Teshima and Innan, 2004). MESH for humans is relatively well established at 92% (Chen et al., 2007, 2010b; Wolf et al., 2009) but MEPS is more difficult to measure. Some studies have concluded that MEPS is generally over 200 bp in humans and mouse (Waldman and Liskay, 1988; Reiter et al., 1998) but IGC with lower MEPS lengths have been observed (Waldman, 2008; see Section 1.3.1 for further information about MESH and MEPS estimates). For this reason, in our model, we use a fixed MESH of 92% and explore the effect of three different MEPS lengths (20, 50 and 200 bp). Figure 3.7 compiles the implementation of IGC, MESH and MEPS in our model.

As represented in Figure 3.7, only potential IGC events that accomplish both local and overall sequence similarity requirements result in effective IGC events (those that actually happen). In this manner, I will refer to *potential IGC rate* (or just C) and *effective IGC rate* as the rates in which potential IGCs and effective IGCs happen. The first measure is not dependent on local sequence similarity between duplicates while the second one is.

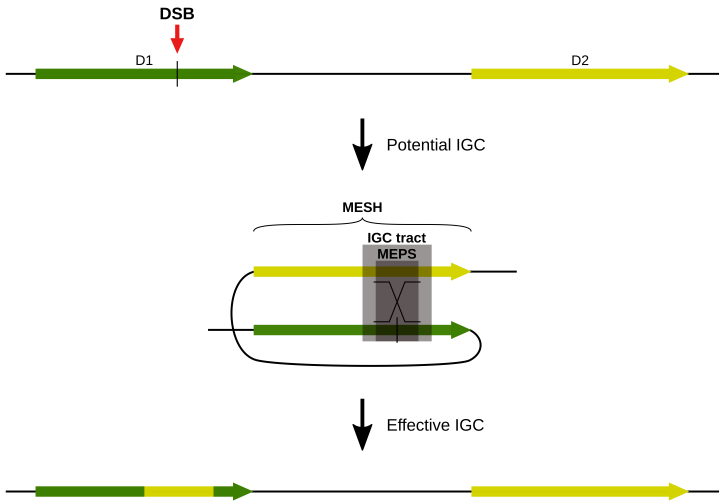


Figure 3.7: IGC dependence on local sequence similarity model. After a DSB, for a potential IGC to actually happen, two sequence similarity thresholds have to be satisfied. First, a minimum amount of overall sequence similarity between duplicates (MESH) of 92% has to be accomplished. And, second, a minimum length of 100% identity between duplicates centered in the DSB point (MEPS) has to be present. In our model we explore MEPS lengths of 20, 50 and 200 bp. Only if both requirements for sequence similarity are satisfied, IGC actually happens (Effective IGC). Each IGC tract is centered at the DSB point and its length is determined by a geometric distribution of mean 100 bp (Wiuf and Hein, 2000; Hartasánchez et al., 2014).

During the evolution of duplicates, the random accumulation of differences might lead, by chance, to regions of the duplicates having slightly more similarity than others. This might result in a situation in which there are parts of the duplicates' length where MEPS requirements are satisfied and parts where they are not. In the former, IGC will be possible while in the latter, it will not. This is, there will be differences in the effective IGC rate along the duplicates' sequence. Moreover, IGC will erase the few differences in IGC-viable regions while mutations will accumulate in IGC-impeded regions perpetuating and incrementing the differences in similarity along the sequence.

When IGC depends on local sequence similarity, equilibrium on concerted evolution is not always reached (Figure 3.8). In some cases, the action of IGC

erasing differences is not powerful enough to maintain enough sequence similarity for IGC to continue at the same rhythm. Variation between duplicates increases while progressively impeding IGC. IGC definitely stops when the variation between duplicates reaches any of the thresholds on similarity.

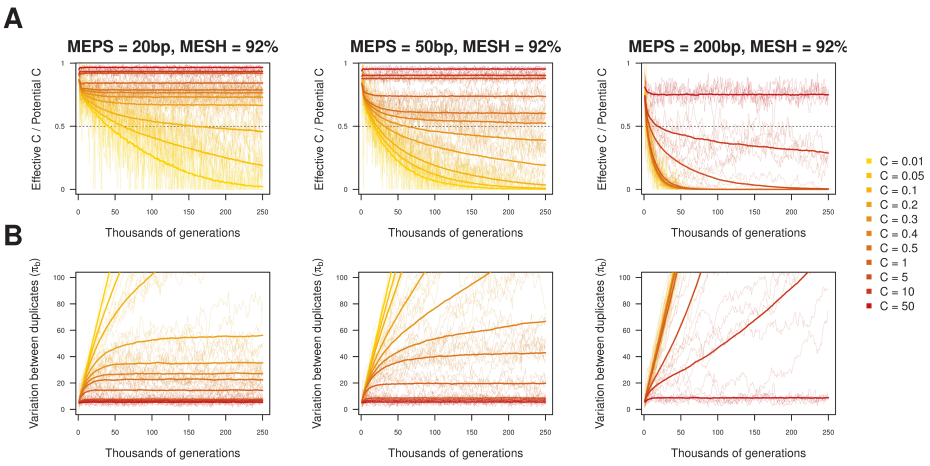


Figure 3.8: Progression of IGC rate and variation between duplicates since duplication under IGC dependent on sequence similarity through time. Top: proportion of potential IGC events that actually happen (become effective IGC events) through time since duplication. Bottom: variation between duplicates through time since duplication (π_b). MESH = 92% in all cases, MEPS = 20 bp (left), 50 bp (middle) and 200 bp (right). Solid lines correspond to average values of 1,000 simulations. Dashed lines are individual simulations trajectories and are intended to visualize fluctuations of particular cases. A big range of potential IGC rate values is represented (C = 0.01 to C = 50) and coded in colors. Crossover rate between duplicates is equal to 10 in all cases.

Although in some cases equilibrium is never reached, IGC will still delay the accumulation of differences between duplicates. For low IGC rates (left panel in Figure 3.9), IGC-viable regions will be gradually lost, with IGC events happening sporadically in less and less regions of the duplicates. In these cases, there is a period in which a non-equilibrium concerted evolution exists. Its duration depends on the strength of IGC and the level of restriction on sequence similarity.

Our results show that it is actually possible to reach an equilibrium in concerted evolution with IGC dependence on sequence similarity (Figure 3.8). This

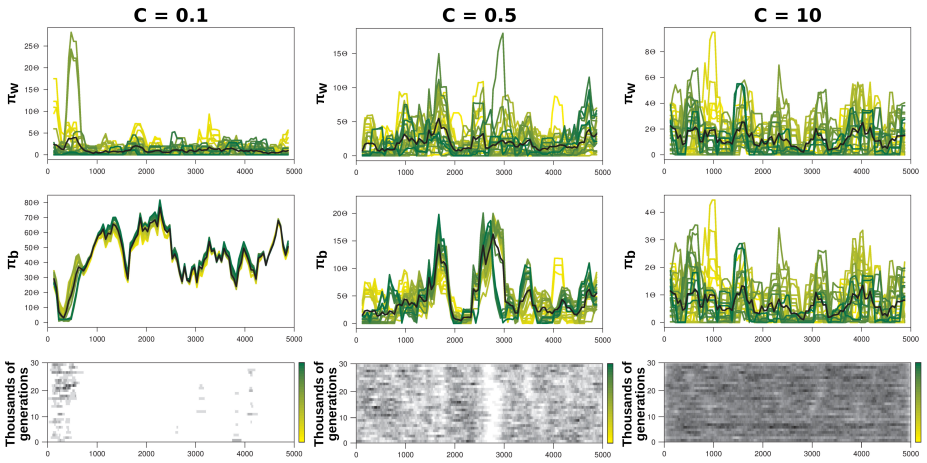


Figure 3.9: Example of IGC depending on local sequence similarity along duplicates sequence through time. Horizontal axes represent nucleotide positions along duplicate sequence in all cases. Plots represent variation within duplicates (top), variation between duplicates (middle) and IGC rate (bottom) along the sequence through 30,000 generations. Crossover rate between duplicates is equal to 10 in all cases. Three potential IGC rates are represented; C = 0.1 (left), 0.5 (middle), 10 (right). Yellow to green lines represent π_w (top plots) and π_b (middle plots) every 1,000 generations. Black lines in each plot represent the average of all generations. Note the differences in scale of the vertical axes. In the bottom plots, IGC rate in each point is represented for the same periods of time represented in the above plots. White to black shades represent absence of IGC to maximum IGC rate in each plot along the sequence (grey intensities are not comparable between plots).

equilibrium is possible when the action of IGC is strong enough to maintain sequence similarity within certain limits, preserving the IGC rate. A higher potential IGC rate is needed to reach equilibrium with more restrictive IGC dependence on sequence similarity (*e.g.* longer MEPS).

When IGC is very frequent, it keeps similarity very high along the whole sequence (see right panel in Figure 3.9). Otherwise, there are differences in effective IGC rate and variation between duplicates along the sequence in equilibrium (middle panel in Figure 3.9). They take the form of *islands of divergence* surrounded by regions of active IGC and much higher sequence similarity. Such islands of divergence are regions where, by chance, a

particularly high number of mutations occurred and/or where there has been a very low number of IGC events and, thus, differences between duplicates have accumulated. These islands can change in size and position or even disappear (regain sequence similarity) during equilibrium thanks to IGC events starting in IGC-viable regions next to them being long enough to convert part of the island length.

Differential effective IGC rate along the sequence, either in equilibrium or not, also has consequences on LD (Figure 3.10). Regions with higher IGC rate will show higher LD between duplicates than regions with lower IGC rate. Moreover, as explained in Section 3.3.1, IGC breaks short distance LD what will result in LD being higher in regions where IGC is not active. These two things together might result in case-specific patterns of LD along the duplicates' sequence being different from those normally expected to arise by crossover.

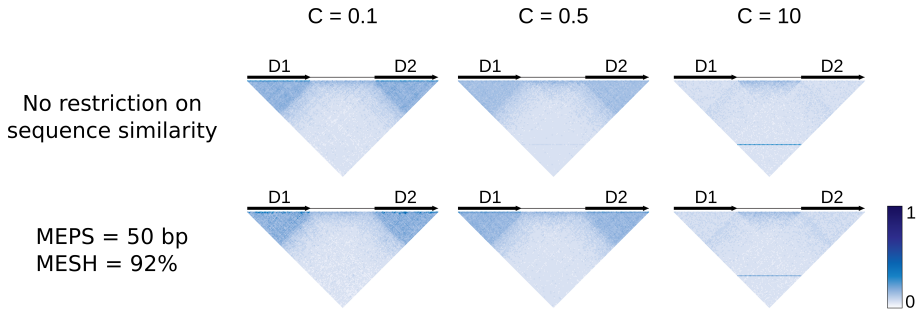


Figure 3.10: Effect of IGC dependence on sequence similarity on LD along the duplicate sequence. Three different IGC rates are represented (C = 0.1, 0.5, 10) from left to right. Crossover rate under the between duplicates crossover model is equal to 10 all cases. Black arrows indicate where duplicates are located along the sequence. Color codes average D' values for each pair of non-overlapping 100 bp windows for 1,000 simulations. Top row shows a case without IGC dependence on sequence similarity and bottom row shows cases with MEPS = 50 bp and MESH = 92%.

The expectation for variation within duplicates at equilibrium is also altered by the incorporation of IGC dependence on sequence similarity (Figure 3.11). The most drastic change it supposes is that the theoretical expectation of π_w (black solid line in Figure 3.11; Innan, 2003b) is not valid for low IGC rates. As

discussed above, cases with very low IGC rate or very restrictive constraint on sequence similarity, do not reach equilibrium on concerted evolution. In these cases, duplicates reach mutation-drift equilibrium as single regions and π_w ends up being Θ.

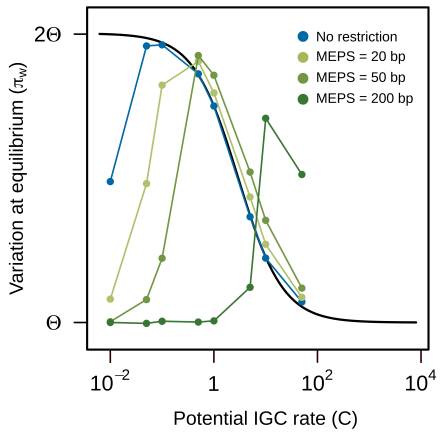


Figure 3.11: Variation within duplicates at equilibrium under four levels of sequence similarity requirements for IGC. Black line corresponds to the expectation of π_w according to the Innan (2003b) model. Points and link lines correspond to average values for 1,000 simulations (100 thousand generations long) with the crossover between duplicates model with R = 10. Several potential IGC values are represented; C = 0.01, 0.05, 0.1, 0.5, 1, 5, 10 and 50. Blue line corresponds to a model without IGC restriction on sequence similarity. Blue points fit Innan's theoretical expectations except for the two lowest IGC rates where simulations did not have time to reach equilibrium, despite long running times, due to the extremely long waiting-times to reach equilibrium. In any case, equilibria that take more than 100 thousand generations to be reached are not realistic. Green lines correspond to simulations under three levels of restriction of IGC for sequence similarity (MESH is 92% in all cases; MEPS is 20, 50, and 200 bp long).

With IGC depending on sequence similarity there is a reduction of the range of IGC rate values in which we would expect increased variation within duplicates at equilibrium. Still, equilibrium with increased π_w is possible even with very restrictive sequence similarity requirements (MEPS = 200 bp in Figure 3.11). Dependence on sequence similarity increases the amount of π_w at equilibrium for a given potential IGC (curves with restriction in Figure 3.11 are shifted to the right). In an equilibrium where IGC depends on sequence similarity, the actual

IGC rate (effective IGC rate) is lower than the potential IGC rate and, thus, the expected π_w is bigger.

IGC dependence on sequence similarity has huge implications in duplicates' evolution. First, equilibrium in concerted evolution is not always reached but concerted evolution can still determine duplicates' fate before ending. Second, concerted evolution (at or out of equilibrium) can be active in some parts of the duplicates even though other parts have escaped its effects. Big differences in sequence similarity along the sequence do not necessarily imply natural selection differentially acting along the sequence. In Chapter 6, I expand on the potential implications of IGC dependence on sequence similarity in duplicates' fate and its interplay with selection.

3.3.4 Neutrality tests on duplications and collapsed duplications

Knowing some of the consequences that IGC has in the sequences of duplicates, a natural question is whether this can affect the summary statistics commonly used to test for neutrality when searching for selection across the genome. Moreover, given the difficulties in properly separating paralogous reads in genome assemblies (and thus, properly identifying paralogous variants), one might wonder if the sequences of duplicates that we obtain with basic sequencing and variant calling methods show deviations from neutral expectations in summary statistics.

In order to tackle these questions, we calculated a set of popular neutrality tests and statistics on duplicated sequences and collapsed duplicated sequences, all undergoing IGC (Hartasánchez et al., 2018, in Appendix 7.3). Here I only show results for 4 statistics (average pairwise differences π_w, Tajima's D, Fay and Wu's H, and Nei's haplotype diversity dh; see methods in Section 3.4) but more can be found in supplementary files of Hartasánchez et al. (2018).

As explained in Section 3.3.1 and Section 1.3.1, IGC increases π_w (compare yellow and blue lines in the top left panel in Figure 3.12). What would be the effect of collapsing duplications on π_w? In Figure 3.12 we observe that the increase in π_w is magnified when collapsing duplicated sequences, especially with low IGC rates. This is because variants that are fixed (or almost fixed) in one copy and not present in the other copy will appear as intermediate

frequency variants if duplicates are collapsed, showing an increase in π_w. This type of variants is more frequent in duplicates undergoing concerted evolution with a low IGC rate.

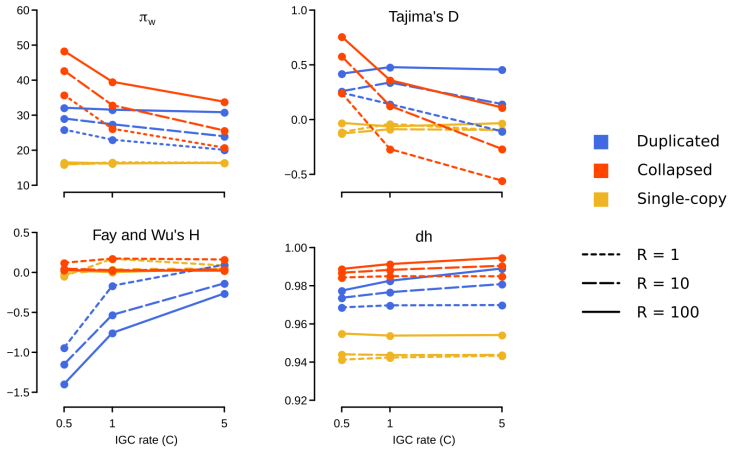


Figure 3.12: Effect of IGC and collapsing duplicates in neutrality tests. Average of 1000 simulation values of average pairwise differences (π_w), Tajima's D, Fay and Wu's H and Nei's haplotype diversity (dh) are shown for a duplicate sequence, two collapsed duplicated sequences and a control single-copy region. Three crossover rates (R = 1, 10, 100) and three IGC rates (C = 0.5, 1, 5) are represented. [Figure taken from Hartasánchez et al. (2018).]

Apart from the known effect on π_w, IGC between duplicates also reshapes the expectations of Tajima's D, Fay and Wu's H, and Nei's haplotype diversity in such regions (Figure 3.12). IGC increases Tajima's D and Nei's haplotype diversity and decreases Fay and Wu's H. The distortion is dependent on the IGC rate and the crossover between duplicates rate.

Collapsing paralogous variants of duplicates undergoing IGC distorts the statistics differently. On the one hand, Tajima's D values in collapsed duplicated sequences show a big range of values depending on IGC rate and crossover rate. On the other hand, Fay and Wu's H, and Nei's haplotype diversity values in collapsed duplications show narrower values compared to simple duplicated sequence values. Fay and Wu's H collapsed values surprisingly resemble control values and Nei's haplotype diversity values are even above the simple

duplicated ones.

Distortion on statistics in collapsed paralogous sequences could be due to the collapsing itself and/or due to the IGC acting between them. Both causes have interplaying effects and each statistic might be influenced differently. Moreover, one should consider that duplications will only be collapsed when divergence between duplicates is low and short sequencing reads coming from different duplicates cannot be distinguished. Being so, in the absence of IGC we would only observe collapsed duplications for very recent duplication events.

Given the big alterations that duplicates and collapsed duplicates undergoing IGC have on neutrality tests and taking into account that these statistics are commonly used to find non-neutral regions genome wide, we wanted to find out whether such values could be mistakenly taken for regions under positive or balancing selection or not. We ran simulations of complete sweeps, incomplete selective sweeps and balancing selection with MSMS (Ewing and Hermisson, 2010). We compared these results with the duplication and collapsed duplication results (Figure 3.13).

Values of Fay and Wu's H statistic for duplicated regions undergoing low levels of IGC mimic the values for this statistic in the case of an incomplete sweep. In all the other cases, test values under selective scenarios can be distinguished from the values of duplicates and collapsed duplicates.

Even though duplications (collapsed or not) evolving neutrally are unlikely to be confused with regions under selection, our results show that they could potentially be confused if summary statistics are not used in combination. Several summary statistics must be taken into account in order to fully uncover individual cases (Pybus et al., 2015) and potential unannotated or new duplications might be considered when observing alterations of such statistics.

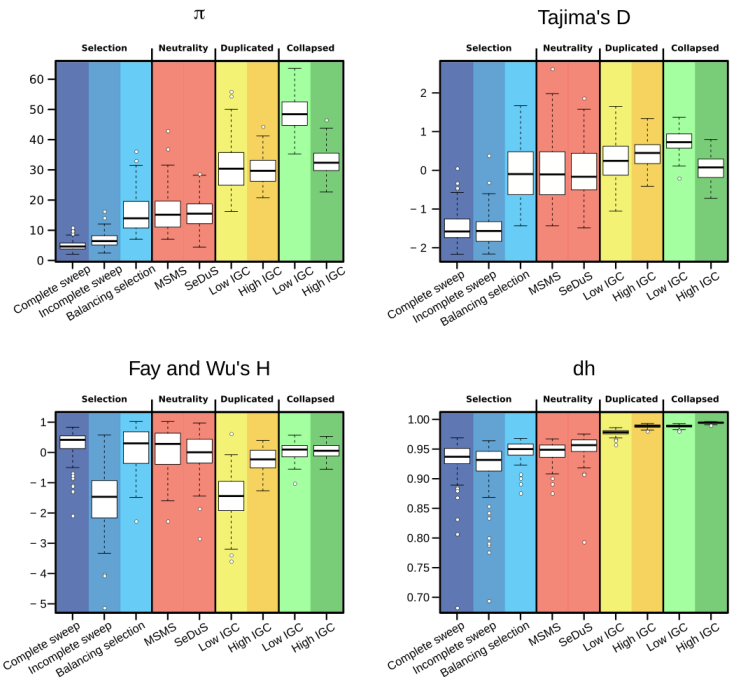


Figure 3.13: Comparison of values expected under selection and duplicated and collapsed duplicated sequences undergoing IGC for 4 summary statistics. Results obtained with simulations from MSMS (complete sweep, incomplete sweep, balancing selection and neutrality; Ewing and Hermisson, 2010) and SeDuS (neutrality, duplicated, collapsed; Hartasánchez et al., 2016). In the case of duplicated sequences and collapsed sequences, R = 10, low IGC rate is 0.5 and high IGC rate is 5. Boxplot whiskers correspond to 1.5 times the inter-quantile range. [Figure taken from Hartasánchez et al. (2018).]

3.4 Methods

The results presented in this chapter were obtained performing simulations. We used two different simulators: SeDuS (Hartasánchez et al., 2016) and MSMS (Ewing and Hermisson, 2010). First, we designed and developed SeDuS for the purpose of studying the evolution of duplicated sequences. It is an efficient and flexible tool to obtain genomic simulated values of duplicates undergoing IGC for a variety of conditions. With it we obtained the vast majority of the results presented in this chapter. Second, we used MSMS in Section 3.3.4 to simulate selection.

SeDuS is a forward-in-time simulator of the neutral evolution of duplications (Hartasánchez et al., 2014, in Appendix 7.1 and Hartasánchez et al., 2016, in Appendix 7.2). It is fast and built in a modular structure. This modular structure is designed to accommodate new implementations very easily, which makes SeDuS very flexible and a perfect tool for studying different models and specific cases. Moreover, apart from the command-line version, SeDuS has a user-friendly graphical user interface (GUI) version for more direct simulations (Hartasánchez et al., 2016, in Appendix 7.2).

SeDuS implements a Wright-Fisher model with a population of size N where the 2N chromosomes consist in two or three genomic regions, named blocks, of equal length. A single simulation of SeDuS consists in three phases. It starts with a burn-in phase in which all the chromosomes have two blocks (*original* and *single-copy*) and start undergoing random mating, mutation and crossover. At the end of this first phase, a duplication event happens. One of the blocks (*original*) gets duplicated on one of the chromosomes resulting in a new block in this chromosome (*duplicated*). This duplication reaches fixation through a neutral fixation trajectory (Kimura, 1980). This phase is named structured or CNV phase. The third phase, named concerted evolution phase, begins with the fixation of the duplication and it endures until the end of the simulation. All the chromosomes with three blocks, during the structured or the concerted evolution phase, experience IGC with a given rate between the original and the duplicated blocks (for further details see in Appendix 7.1 and in Appendix 7.2).

SeDuS provides genomic data for all the chromosomes in a periodic manner (every 1,000 generations). From this genomic data, one can calculate whatever

statistic of interest. In this chapter we used the average pairwise differences statistic (Nei and Li, 1979) to measure variation within (π_w) and between duplicates (π_b, π_s); the D_{sum} statistic[1] (Innan, 2003b) to measure average LD between duplicates; the mean D' statistic (Lewontin, 1964) to measure LD between pairs of windows along the simulated chromosomes sequence; and a set of neutrality tests including Tajima's D statistic (Tajima, 1989) , Fay and Wu's H statistic (Fay and Wu, 2000), and Nei's haplotype diversity statistic (Nei, 1987).

[1]

$$D_{sum} = \sum_{m=1}^{L} D_m \tag{3.2}$$

where D_m corresponds to the D statistic between paralogous sites at position m:

$$D_m = \frac{n_{AA}n_{aa} - n_{Aa}n_{aA}}{n(n-1)} \tag{3.3}$$

where n is the number of chromosomes in the sample and n_{xy} corresponds to the number of chromosomes that have the variant x at site m in the first duplicate and the variant y at site m in the other duplicate.

This chapter is a review of all the work performed in my group during the last years about the neutral evolution of duplications. I did a significant contribution to the project that was led by Diego A. Hartasánchez and supervised by Arcadi Navarro. Oriol Vallès-Codina, Juanma Fuentes-Díaz, Marc Pybus and Jose Maria Heredia-Genestar also contributed to different parts of this project.

This project has resulted in to three publications so far. First, an article describing the interplay between IGC and crossover and their effects on variation and linkage disequilibrium (Hartasánchez et al., 2014, in Appendix 7.1). In this article I contributed to the software writing and optimization, to the experiment design, to the analysis and visual representation of the results and to manuscript writing. Second, another article presenting a computationally improved, extended and user-friendly version of the forward-in-time simulator of the neutral evolution of SDs we used in the first article (Hartasánchez et al., 2016, in Appendix 7.2). My contribution to this second piece of work was software documentation, optimization of the code, development of a user-friendly graphical user interface and manuscript writing. Third, a manuscript recently accepted (Hartasánchez et al., 2018, in Appendix 7.3) about the deviations in summary statistics in duplications and collapsed duplications. I contributed to this article in the original experiment design, software adaptation and implementation, analysis and visualization of results and manuscript writing.

Apart from these published contributions, this chapter includes results form a piece of work in progress about the modeling and simulation of IGC dependence on sequence similarity. I contributed to this piece of work in the original idea, design, software writing, simulation implementation and analysis and visualization of results.

Chapter 4

Human segmental duplications revisited

I didn't want to just know names of things. I remember really wanting to know how it all worked.

Elizabeth Blackburn

All sorts of things can happen when you're open to new ideas and playing around with things.

Stephanie Kwolek

4.1 Rationale

Eukaryotic genomes have been known to present duplicated segments since the 1930s (Bridges, 1936; Muller, 1936). However, the relevance of duplication events as a major source of evolutionary innovation was not recognized until after Susumu Ohno's seminal work (Ohno et al., 1968; Ohno, 1970). Since then, extensive effort has been dedicated to elucidate, first, the duplication mechanisms; second, the evolution of duplicates (including concerted evolution); third, the fate of duplicated functional elements such as genes; and, finally, the presence of duplications and CNVs in the genomes of many species (see Section 1).

Highly similar duplications, such as SDs, were only identified as being pervasive in our genome when the first draft of the human genome was completed (Bailey et al., 2001; Lander et al., 2001; Venter et al., 2001; see Section 1.2.1). Later, comparative studies showed that duplications arose at an unusually high rate in the human lineage at the time of the African great ape ancestor (the most recent common ancestor of gorilla, chimpanzee, bonobo and human species; Marques-Bonet et al., 2009a; Sudmant et al., 2013; see Section 1.4).

Unlike other duplicated regions of the genome, such as transposable elements, which are quite well characterized in terms of sequence, mechanistic origin, implications on disease, and function (Cordaux and Batzer, 2009; Hancks and Kazazian, 2012; Kaer and Speek, 2013), there are still many open questions regarding the origin, function and evolution of SDs. During the last decade,

human SDs have been the subject of many studies: comparative studies (Marques-Bonet et al., 2009a; Gokcumen et al., 2013; Sudmant et al., 2013; Hafeez et al., 2016; Dennis et al., 2017), studies describing CNVs in human populations (Sudmant et al., 2015a,b) and because they are the genetic basis of specific diseases or the acquisition of new functions (Ballif et al., 2010; Marotta et al., 2012; Girirajan et al., 2013; Ebert et al., 2014; Zagaria et al., 2014; Zielinski et al., 2014; Bekpen et al., 2017). Still, human SDs have not been systematically classified and described since the early versions of the human reference genome (Bailey et al., 2001; Lander et al., 2001; Venter et al., 2001; Bailey et al., 2002a; see Section 1.4).

In this work I aim to analyze the diversity of SDs in the human genome in order to understand the distinct mechanisms through which duplications arise, the characteristic evolutionary time of different kinds of duplication events, the possible evolutionary paths duplications may have undergone, and the potential functional consequences and innovations brought about by duplications in our genome.

4.2 Objectives

General:

1. To contribute to the understanding of the birth, evolution and functional impact of the SDs in the human genome.

Specific:

1. To generate an accurate database of SDs in the human genome.
2. To propose a classification of SDs according to criteria relevant to their evolution.
3. To date the most recent human SDs.
4. To describe the distinct characteristics of human SDs.
5. To contribute to understand the reasons behind the distinctive characteristics of human SDs to help unraveling the mechanisms through which human SDs arose, evolve and contribute to genetic innovation.

4.3 Results and Discussion

In 2001, Jeffrey Bailey and collaborators (Bailey et al., 2001) presented the first approach to systematically identify recent and long human genomic duplications right after the first draft of the human genome was published (Lander et al., 2001; Venter et al., 2001). Their method was based on genome-wide pairwise alignments and searches for long tracts of identity between different parts of the genome assembly. This duplication detection method is nowadays called *whole-genome assembly comparison* (WGAC; see Section 1.2.2). Bailey et al. (2001) defined SDs as duplicated regions of the genome of 1kb or more in length and with 90% or more similarity between copies (Bailey et al., 2001; see Sections 1.2.1 and 1.2.2).

The SegDups database of the human reference genome hg38/GRCh38 (UCSC Table Browser; Karolchik, 2004) is the result of applying the WGAC approach to the hg38/GRCh38 human reference genome assembly. It reports every pair of regions (of 1 kbp or more in length) identified as being copy of one another (with 90% or more of identity) in the human reference genome. This database provides not only the exact genomic position, length, and sequence of all duplicated regions but also the genomic location, length and sequence of the corresponding copies in a pairwise manner. In other words, it provides high resolution information on the distribution of homology across the genome and, thus, affords the opportunity to better understand how duplications happen and how their identity evolves (see Section 1.2.2).

The SegDups database and, more generally, WGAC, are very valuable tools to study SDs. Nevertheless, they must be used considering the following details:

- First, WGAC (and the SegDups database) is based on genome-assemblies and, as such, it only retrieves information from a single, haploid genome. In other words, the SegDups database does not provide population diversity information (CNV data).
- Second, the quality of all WGAC databases depends on the quality of the corresponding reference genome. The adequate resolution of recent and long duplications, such as SDs, is known to be one of the most complicated problems for genome assembly methods, especially those

based on whole-genome shotgun sequences (see Section 1.2.2). Highly identical duplicates longer than the read length are very difficult to distinguish from each other and, thus, are candidates to being collapsed into a single region in the final assembly (Salzberg and Yorke, 2005; Kelley and Salzberg, 2010; Hartasánchez et al., 2018). The WGAC approach and the SegDups database are blind to duplications not resolved in the assembly. Fortunately, the human reference genome is clone-based and of high quality (see Section 1.2.2 and Section 5). Only some very recent CNV regions are expected to be absent in the hg38 assembly (Kidd et al., 2010; Sudmant et al., 2015a).

- Third, SDs in the human genome are frequently organized in clusters of duplications forming mosaic patterns organized around ancient core duplications named *core duplicons* (Jiang et al., 2007; Marques-Bonet and Eichler, 2009; see Section 1.3.2). Mosaic segmentally duplicated regions, here termed *duplication mosaics*, are the result of a complex history of duplication events at the very least, since they can also involve deletions, inversions, translocations and other complex genome rearrangements. Duplication mosaics, sometimes named duplication hubs (She et al., 2006) or duplication blocks (Jiang et al., 2007), appear in the SegDups database as multiple overlapping annotations mapping to the same mosaic region and reflecting the internal architecture and complexity of such region. The presence of mosaic duplications with a highly complex architecture of redundant and non-redundant homologies makes the SegDups database hard to handle.

- Fourth, only alignments accomplishing both criteria, $\geq$90% identity and $\geq$1 kbp in length, are listed in the SegDups database. For example, region A can be listed as being copy of both region B and region C, while regions B and C might not be listed as being copies of each other. This does not mean that regions B and C are not actual duplicates but that the identity between them is either too low or too short. Therefore, when working with the SegDups database, one should keep in mind that the copies of the concerning regions are not necessarily limited to those listed in the database.

SD pairs within mosaic duplications are only part of the whole duplication

picture of the region and, thus, not fully informative. Instead, taking into account all SD pairs within the same mosaic duplication is a better approach. For this reason, I decided to base this work in regions of our genome comprised of SDs, here named SDRs (for SD regions), instead of SD pairs. A similar approach was used in Redon et al. (2006), She et al. (2006), Jiang et al. (2007) and Pu et al. (2018).

SDRs are classified according to their distinct features; first, depending on their mosaic or isolated nature and, second, depending on the relative location of their copies in the following manner (Figure 4.1 A):

- *Non-mosaic SDRs* are genomic regions with one or more copies in other genomic sites sharing identity with the entire SDR as one. As opposed to mosaic SDRs which have several, partial identities to other parts of the genome, non-mosaic SDRs copies share identity to the entire length of the SDR. They are classified according to the relative genomic position of their copy (or copies):
  - *Tandem SDRs* are non-mosaic SDRs whose copy is located at less than 1 kbp of distance (see Section 4.3.1 for explanation for the 1kb criterion).
  - *Isolated intrachromosomal* SDRs are non-tandem non-mosaic SDRs with their copy (or copies) located in the same chromosome where the SDR is located.
  - *Isolated interchromosomal* SDRs are non-mosaic SDRs with their copy (or copies) located in a different chromosome where the SDR is located.

  There are only 23 cases of non-mosaic SDRs with copies both in the same and in other chromosomes and they were excluded from the analysis.

- *Mosaic SDRs* are genomic regions presenting clear evidence of more than one duplication event[1] and are comprised by overlapping fragments with

[1] Although non-mosaic SDRs could also be the result of more than one duplication event, mosaic SDRs are a clear product of more than one duplication comprising different but overlapping fragments of sequence. They are a clear case of duplication shadowing where at least one duplication has happened on top of an already duplicated region (see Section 1.3.2).

copies in different parts of the genome (see Section 1.3.2). They present a big range of complexity, from relatively simple mosaic SDRs product of two overlapping duplication events to very large and intricate mosaic SDRs typically found around core duplicons (Jiang et al., 2007; see Section 1.3.2). They are classified, similarly to the non-mosaic regions, according to the relative position of their copies:

- *Intrachromosomal mosaic SDRs* are duplication mosaics with all their copies located in the same chromosome where the SDR is located.
- *Interchromosomal mosaic SDRs* are duplication mosaics with all their copies located not in the same chromosome where the SDR is located.
- *Complex mosaic SDRs* are duplication mosaics with copies located at both the same and different chromosomes where the SDR is located. Complex mosaics include, as its name implies, the most conglomerated and intricate SDRs in our genome (Figure 4.1 B).

It is important to remark that SDRs are classified according to their mosaic or non-mosaic nature and according to the location of their copies, not according to the classification of their copies. In this way, for example, a complex mosaic can have an isolated intrachromosomal copy as long as this copy has a non-mosaic nature and that both SDRs are located in the same chromosome (Figure 4.1 A).

4.3.1 Tandem vs. Isolated intrachromosomal SDRs

Distinguishing between tandem and non-tandem intrachromosomal SDRs is not trivial. During the early stages of the human reference genome, the knowledge that tandem duplications had distinct characteristics compared to the rest of SDs was already clear (Bailey et al., 2002a, 2003; She et al., 2006). What was not clear at the time and is still a largely open question today is what are their defining characteristics. In principle, the distinction between tandem and non-tandem duplications is that non-tandem are further apart, but is there a maximum distance between tandem duplications for them to be considered as such?

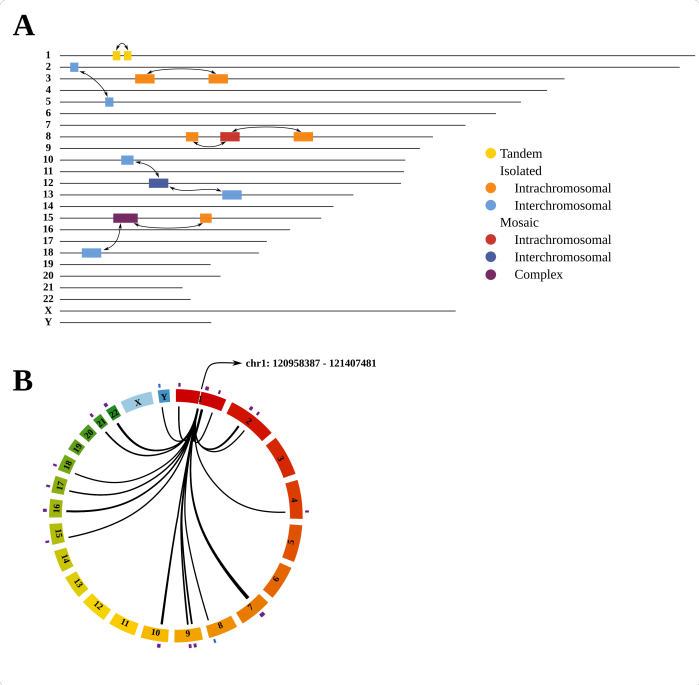


Figure 4.1: Classification of SDRs in the human genome. A. Schematic representation of the SDR categories. The most simple example is used for each case for the sake of clarity. Human chromosomes are represented as black lines and labeled to their left. Colored boxes represent SDRs. Their lengths are exaggerated in the image for visualization purposes. Colors correspond to specific SDR types: yellow for tandem duplications (less than 1 kbp apart from their copy, in the same chromosome), orange for isolated intrachromosomal duplications (isolated duplications with their copy in the same chromosome, more than 1 kbp apart), light blue for isolated interchromosomal duplications (isolated duplications with their copy in another chromosome), red for intrachromosomal duplication mosaics (mosaic duplication with copies only in the same chromosome), dark blue for interchromosomal duplication mosaics (mosaic duplication with copies only in other chromosomes) and purple for complex mosaics (mosaic duplications with copies in both, the same chromosome and other chromosomes). Arrows link duplicated pairs. Note that isolated SDRs can be copies of mosaic SDRs and vice versa. B. Circos plot of the complex duplication mosaic located in chr1:120958387-121407481. Human haploid genome chromosomes are represented as color boxes placed circularly. Location of the represented complex duplication mosaic is indicated as a white line in chromosome 1. Black lines link duplicated pairs. Color squares outside the chromosomes indicate the type of SDR of each copy (color coded as in A).

Bailey et al. (2003) defined the between-copies distance threshold between tandem and isolated intrachromosomal SDs (there termed interspersed SDs) in 1Mb with no explicit criterion. The distribution of intrachromosomal, non-mosaic SDRs (tandem and non-tandem) across the intervening distance between duplicates clearly shows two distinctive groups of such regions (Figure 4.2 A). However, the intervening distance between copies that distinguish these two groups is not close to the 1 Mbp set by Bailey et al. (2003). Rather it is close to 1 kbp. I have performed a series of analyses comparing different parameters against the intervening distance between copies. The results show, consistently, two distinct groups of intrachromosomal, non-mosaic SDRs and the between-copies distance threshold which better distinguishes them is close to 1 kbp (Figure 4.2). Tandem and non-tandem intrachromosomal duplications are two distinct classes of duplications and a good measure to classify them is a between-copies distance threshold of 1 kbp.

Tandem SDRs are shorter, tend to be more identical and have more GC content, less transposable element presence in their borders and less contribution to protein coding genes (Figure 4.2 B-F) compared to isolated intrachromosomal SDRs. The difference in length, GC content and retrotransposons content in their borders suggests that tandem SDRs and isolated intrachromosomal SDRs have different mechanisms of origin (see Section 4.3.4). The difference in identity suggests that tandem SDRs might be more recent than isolated intrachromosomal SDRs or that they have undergone more IGC (see Chapter 3).

Although I originally did not differentiate between tandem SDRs and isolated intrachromosomal SDRs for consistency with the literature, it has resulted a relevant approach for this work. It is so not only because tandem SDRs and isolated intrachromosomal SDRs show distinct characteristics, but because tandem SDRs appear to be a very peculiar type of duplicated regions. While all other types of SDRs, including duplication mosaics, share many features, tandem SDRs, persistently dissociate from all the other SDRs in many aspects. They are actually a very unique type of duplications and must be treated separately. Through all this chapter I will be pinpointing the peculiarities of tandem SDRs.

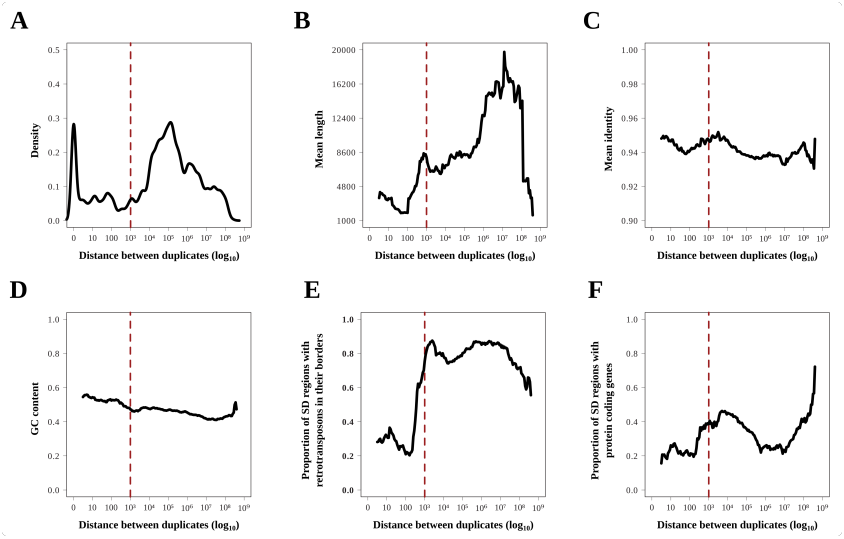


Figure 4.2: Distinct characteristics of intrachromosomal SDRs (tandem and isolated intrachromosomal). The horizontal axis represents the distance between duplicates (*i.e.* distance between SDRs and their copies) in all cases. The vertical line depicts the 1 kbp threshold used to distinguish between tandem SDRs and isolated intrachromosomal SDRs. A. Density of SDRs cross distance between copies. B-F. Different measures across distance between duplicates. Lines depict the mean value of a sliding window of size 1 and step equal to 0.05 (log10 of the distance). The use of a sliding window smooths the line but creates a distortion of the actual point where SD-region differences are present. The parameters represented are mean length (B), mean identity between SDRs and their copies (C), CG content (D), proportion of SDRs having transposable elements in their borders (window of 20 bp around junction; E), and fifth, proportion of SDRs with protein coding genes (F).

4.3.2 Span and distribution

SDRs span more than 166 Mbp, covering around 5.13% of the length of the human reference genome (Table 4.1, Figure 4.3 A, Bailey et al., 2002a, She et al., 2004). Most of this length, specifically 82.44%, is composed by mosaic duplications in contrast to 16.53% of isolated SDRs and 1.01% of tandem SDRs (Table 4.1). This difference in percentage of duplicated sequence is explained by the differences in length between mosaic and non-mosaic SDRs, not by differences in the number of regions (Table 4.1). In other words, mosaic SDRs

span more proportion of the genome than non-mosaic SDRs because they are longer, but not for being more abundant.

		Number	*Cumulative length (kbp)*	*Mean length*	*% of the genome*
Tandem		503	1686.203	3352.29	0.05201
Isolated	***Intrachromosomal***	1579	16364.583	10363.89	0.50478
	Interchromosomal	2228	11137.181	4998.73	0.34353
Mosaic	***Intrachromosomal***	949	25692.847	27073.6	0.79251
	Interchromosomal	1198	14706.032	12275.49	0.45362
	Complex	1778	96771.345	54427.08	2.98497
Total		8258	166394.521	20149.49	5.13254

Table 4.1

One remarkable observation from the classification of SDRs is the large amount of intrachromosomal and interchromosomal mosaic SDRs relative to complex mosaic SDRs. The former categories comprise regions of high complexity in terms of number of duplication events and distribution of identities with other copies. Despite this complexity, there are many duplication mosaics exclusively intrachromosomal (949) and exclusively interchromosomal (1198; see discussion in Chapter 6).

There are important differences in length between different types of SDRs (Figure 4.3 B). Complex mosaic duplications are the longest mosaic SDRs. Interestingly, for both, mosaic and isolated SDRs, intrachromosomal regions are longer than interchromosomal regions (Figure 4.3 B). This observation suggests that intrachromosomal and interchromosomal SDRs might be biologically distinct entities (*e.g.* have different origins) regardless of their mosaic or isolated nature.

Duplicated regions are known to be distributed non-randomly in our genome. Although one can find SDRs everywhere along the human genome, they tend to group in regions with high duplication content (Figure 4.3 A). Visual inspection of Figure 4.3 shows that SDRs are more frequent around centromeres and close to telomeres. In fact, pericentromeric and subtelomeric enrichment in SDs has been previously reported. She et al. (2004) describes differential clustering between interchromosomal and intrachromosomal SDs with the former being

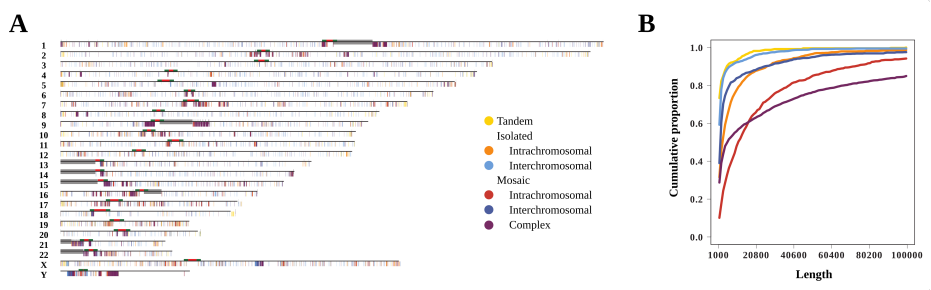


Figure 4.3: Distribution and length of SDRs. A. Distribution of SDRs along human chromosomes. Human chromosomes are represented as black lines and labeled to their left. Genomic gaps are depicted in grey squares, centromeres are represented as red marks and the pericentromeric regions (2 Mbp around the centromere) as green marks. SDRs are represented as color-coded marks below the corresponding chromosome line. SDRs marks are in scale, except for regions shorter than 50 kbp which are augmented to this size for visualization purposes. B. Cumulative proportions of the length of each SDR category.

more prevalent than the latter in pericentromeres. Bailey and Eichler, in their review of 2006 (Bailey and Eichler, 2006) state that the enrichment of SDs in pericentromeric regions is composed mainly by mosaic duplications (although they do not use this term). Moreover, in both papers the authors suggest the high Alu content of pericentromeres as responsible for their SD enrichment (Bailey et al., 2003; see Section 4.3.4 for results in the Alu-mediated origin of SDs).

Pericentromeres are statistically enriched in all types of SDRs except for tandem SDRs (Figure 4.4 B and methods in Section 4.4). This enrichment is extremely high for complex mosaic SDRs. A more detailed visualization of the distribution of SDRs along the chromosome reveals more information (Figure 4.4 A). First, tandem duplications show no preference for any specific chromosomal location. Second, although both intrachromosomal and interchromosomal SDRs, either isolated or mosaic, show statistical enrichment around centromeres, interchromosomal SDRs are more specific to these regions, consistent with the findings by She et al. (2004) discussed above (Figure 4.4 A).

The distribution of the different kinds of SDRs in particular human chromosomes shows interesting features (Figure 4.5 and Figure 7.3 from Chapter 5 for comparison). The first characteristic that draws attention is the

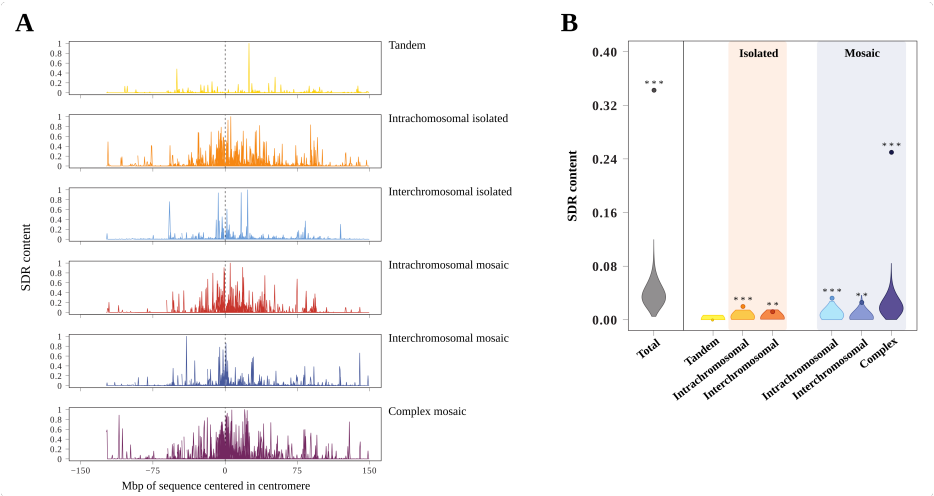


Figure 4.4: SDR content along the human chromosomes and enrichment in pericentromeres. SDR content corresponds to the proportion of the length of 0.5 Mbp windows (non-sliding) covered by a given type of SDR. A. SDR content along the chromosomal sequence. All chromosomes' SDR content per window is collapsed and placed with the centromere in the centre for each plot. B. SDR enrichment in pericentromeres. Violin distribution for each type of SDR corresponds to the pericentromeric SDR content in 1,000 circularly randomized positions of SDRs across the genome (see methods in Section 4.4 for further details on the enrichment test). Points correspond to the actual duplication content of pericentromeres. Stars correspond to the significance of the enrichment test (see methods in Section 4.4): *** for p-values ≤ 0.0001, ** for p-values ≤ 0.001 and * for p-values ≤ 0.01.

large diversity on SD content that different human chromosomes have. Not all chromosomes contribute equally to the pericentromeric enrichment of SDRs. In fact, there are chromosomes with no special accumulation of SDRs in their pericentromeres (for example, chromosomes 3, 4, 5, 8 or 12 in Figure 4.3 A) and chromosomes with massive accumulation of SDRs around centromeres (for example, chromosomes 1, 2, 7, 9, 15 or 20 in Figure 4.3 A). Particularly, chromosome 9 shows a specially high clustering of SDRs around its centromeres (Figure 4.3 A), consistent with previous reports (Bailey et al., 2001; Sudmant et al., 2013). This chromosomal diversity can be explained by the differential evolutionary history of each human chromosome (Bailey and Eichler, 2006; Chen et al., 2010b; Liu et al., 2012; Weckselblatt and Rudd,

2015). The second thing to be highlighted is that, although chromosome 9 has a number of SDRs similar to other chromosomes, they cover a larger amount of sequence. This difference is explained by the length covered by its complex mosaic SDRs (Figure 4.5). The large pericentromeric region of chromosome 9 is composed by a few, very long complex mosaic SDRs. A similar situation happens, although in smaller scale, in other chromosomes (for example, chromosome 2, 7, 15, 16, 21 or 22 in Figure 4.5). The third important observation to be extracted from Figure 4.5 is the peculiarity of sex chromosomes. Unlike autosomes, chromosome X and Y have an important contribution of interchromosomal SDRs (both isolated and mosaic) to their length. These chromosomes contain long interchromosomal SDRs that do not exist in autosomes. The peculiarities on the evolution of sex chromosomes might explain their particular SDR profile. However, to my knowledge, interchromosomal duplication enrichment in sex chromosomes has not been resolved to date despite substantial work on sex chromosome duplication content, especially in chromosome Y (Thornton and Long, 2002; Rozen et al., 2003; Kirsch et al., 2005, 2008; Betrán et al., 2012; Hallast et al., 2013; Veeramah et al., 2014; Trombetta and Cruciani, 2017; Kuderna et al., 2018).

4.3.3 Age classification

Sequence identity between copies of a given duplication has been used as a proxy for the age of the duplication event (Bailey et al., 2002b; Zhang et al., 2005). SDs, by definition, share at least 90% of identity between copies and, thus, it is common to assume that SDs are the result of duplication events that took place starting 35 to 40 Mya (Bailey and Eichler, 2006). Using identity between copies as a proxy for age has the underlying assumption of constant and homogeneously distributed mutation rate, selective neutrality and absence of IGC between copies (see Chapter 3, particularly Section 3.3.3). One can accept all these suppositions when talking about general trends of the whole set of SDs in the human genome together but when one tries to date the birth of particular SDs, these effects may not be negligible (see Section 4.3.4). Using identity between copies as a proxy for the age of a given duplication is not accurate and should be avoided.

Another commonly used strategy to infer the age of a given duplication is based on phylogeny (*i.e.* phylostratification). Marques-Bonet et al. (2009a) and

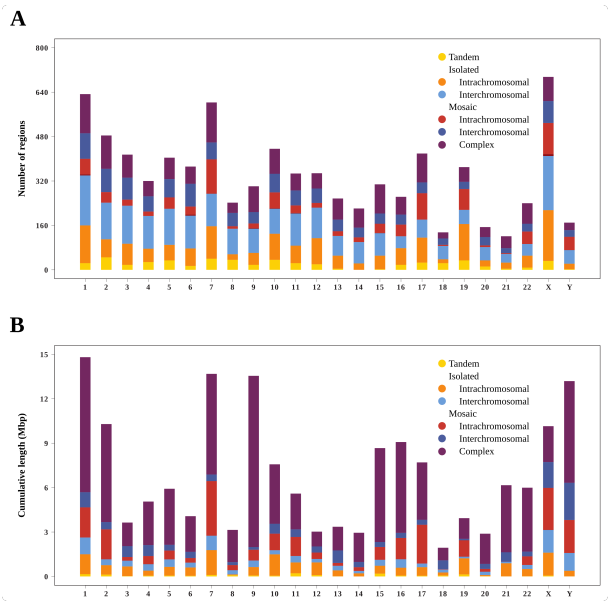


Figure 4.5: Distribution of SDRs across human chromosomes. Number of different types of SDRs (A) and their cumulative length (B) in all human chromosomes.

Sudmant et al. (2013) used this approach to date great ape SDs. Here I use whole-genome high-quality copy-number estimates based on read depth in SDRs in each of the great ape lineages genomes to phylostratify and assign a time-window to each SDR. More precisely I use the method presented in Serres-Armero et al. (2017) to estimate with high confidence the copy-number in 1 kbp windows that cover the whole genome for a sample of 10 individuals for each great ape species (see methods in Section 4.4). Estimating the copy number of a given genomic region based in read depth is the basis of the WSSD method (Bailey et al., 2002a), a very widely used SD-detection approach (see Section 1.2.2).

Importantly, by using read depth based copy-number estimates to perform phylostratification of SDRs, in this work I am taking advantage of two duplication detection methods at once. one based on assembly (WGAC; Bailey et al., 2001) and the other, based on depth of coverage (WSSD; Bailey et al.,

2002a). I use WGAC to precisely delimit SDRs, assess their internal structure and classify them according to it and to the relative location of their copies. Additionally, WSSD allows me to determine the presence or absence of each SDR in great apes and, thus, to infer their time window of appearance.

I, therefore, assign to every SDR a given window of time of appearance according to the species at which it is (or is not) present (see methods in Section 4.4 and Figure 4.6). I distinguish between, first, *Homo specific SDRs* that are human SDRs that are not present in any other great ape species genome, second, *Homo-Pan shared SDRs* or SDRs present exclusively in human and Pan genus species genomes, third, *Homo-Pan-Gorilla shared SDRs* corresponding to SDRs present in all the great ape species genomes except in the orangutan genome, and, fourth, *SDRs shared by all great apes* group conformed by SDRs that appeared before the lineage leading to orangutans split from the African great ape lineage (Figure 4.6). The SDRs in this last group appeared at some point in time before the great ape species diverged and could potentially be very ancient.

When assigning a given window of age to a given SDR I am assuming that the whole SDR was created at once, although this might not be the case for some SDRs, specially mosaic duplications. Thus, the age I assign to a given SDR refers to the age at which the SDR suffered the first main duplication covering at least 70% of the SDRs length. New duplications on the top of old mosaic duplication are not considered in my age classification model (see methods in Section 4.4).

For the sake of clarity, from now on I will use *type* of SDRs to refer to the different qualitative categories of SDRs (*i.e.* tandem SDRs, isolated intrachromosomal SDRs, isolated interchromosomal SDRs, intrachromosomal mosaic SDRs, interchromosomal mosaic SDRs and complex mosaic SDRs) and SDR *time class* to refer to the degree of sharedness across the great apes (*i.e.* Homo specific SDRs, Homo-Pan shared SDRs, Homo-Pan-Gorilla shared SDRs and SDRs shared by all great apes), which is a proxy of their time of origin.

Marques-Bonet et al. (2009a) showed that the rate of emergence of new SDs was particularly high during the African great ape ancestor period. I confirm this finding by detecting an especially high duplication rate relative to the point-mutation rate corresponding to this period of time compared to more recent times (Table 4.2 and Figure 4.7). This increase in duplication rate happened both at the

level of newly duplicated regions and at the level of number of bp of newly duplicated sequence.

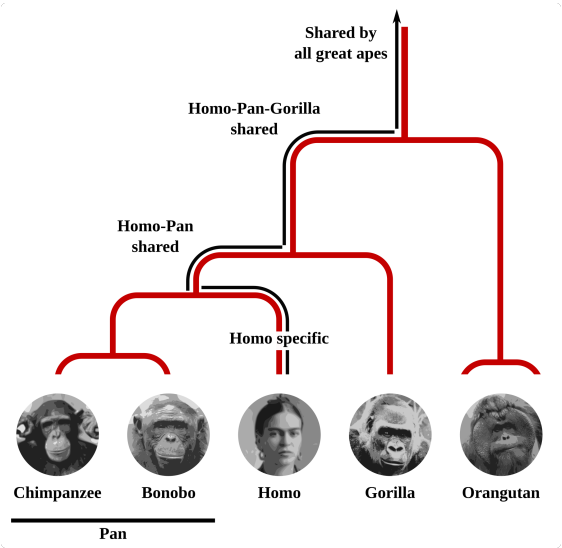


Figure 4.6: Phylostratification of SDRs. Great ape phylogeny is represented in orientative time scale. SDRs in the human genome are grouped according to their presence or absence in other great apes genomes. Labels indicate the four considered time windows and black lines specify the part of the phylogeny corresponding to each one of them. The first and most recent group of SDRs corresponds to Homo specific SDRs. These SDRs are not shared with any of the other great apes and, unless a very specific case of loss, they appeared after the split of the human lineage with the Pan lineage (leading to chimpanzees and bonobos). The second group of SDRs are the regions that are shared between human and Pan genus but that are absent in gorilla. I term this group as Homo-Pan shared SDRs. The third group includes all the regions that are present in all the great ape genomes except in the orangutan. I will refer to this third group of SDRs as Homo-Pan-Gorilla shared SDRs. Orangutan is the most distant species included in the analysis and it is used as an outgroup. The fourth and most ancient group of SDRs includes all the regions that are shared by all the great ape species, including the orangutan.

			Homo specific			*Homo-Pan shared*			*Homo-Pan-Gorilla shared*			*Shared by all great apes*	
		SDRs	SDR per My	Length (kbp)	kbp per My	SDRs	SDR per My	Length (kbp)	kbp per My	SDRs	SDR per My	Length (kbp)	kbp per My	SDRs	Length (kbp)
Tandem		29	3.85	133.02	17.66	40	10.78	79.67	21.47	75	6.79	174.49	15.81	252	1105.42
Isolated	***Intrachromosomal***	23	3.05	368.13	48.89	34	9.16	1087.68	293.17	181	16.39	2334.67	211.47	1160	10543.93
	Interchromosomal	64	8.50	1011.98	134.39	45	12.13	252.24	67.99	207	18.75	1685.84	152.70	1578	7212.93
Mosaic	***Intrachromosomal***	13	1.73	1326.72	176.19	17	4.58	714.19	192.50	64	5.80	2348.31	212.71	784	18444.67
	Interchromosomal	7	0.93	88.59	11.77	11	2.96	118.97	32.07	62	5.62	1432.32	129.74	1049	11809.45
	Complex	21	2.79	1619.58	215.08	13	3.50	611.95	164.95	104	9.42	8076.33	731.55	1457	67518.36
Total		157	20.85	4548.01	603.99	160	43.13	2864.69	772.15	693	62.77	16051.96	1453.98	6303	116671.09

Table 4.2: Duplication rate per type of SDR and time window. Number of SDRs, number of new duplicated regions per million years (My), cumulative SDR length in kbp and number of new duplicated kbp per My are presented for each type of SDR and each time class (except for the SDRs shared by all great apes for which I have no outgroup to calculate duplication rates). I use divergence estimates based on point-mutation between humans and the other great ape species from Prado-Martinez et al. (2013) to calculate duplication rates in each case. Divergence estimates are $7.53\frac{mut}{kbp}$ (between human and chimpanzee), $11.24\frac{mut}{kbp}$ (between human and gorilla) and $22.28\frac{mut}{kbp}$ (between human and orangutan). I calculate the duplication rate based in single nucleotide divergence assuming a mutation rate of $5 \cdot 10^{-10}\frac{mut}{bp \cdot year}$ (see Section 4.4).

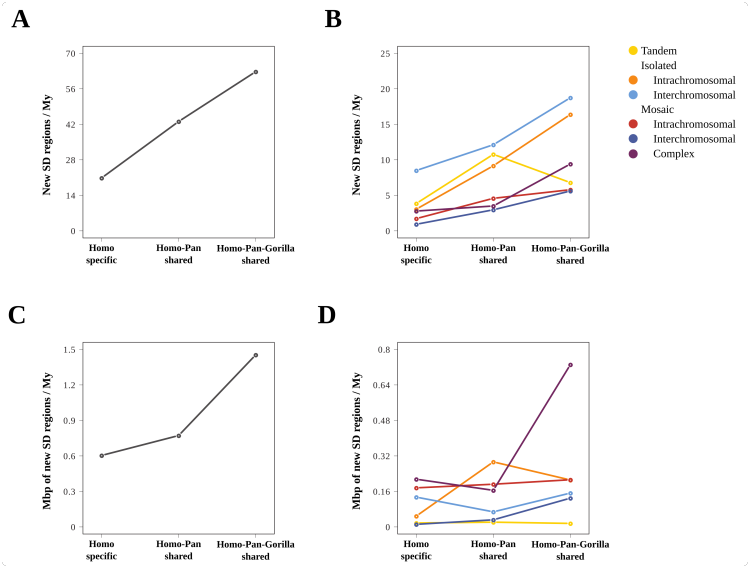


Figure 4.7: Duplication rate during great ape diversification. The horizontal axis represents the three most recent time windows under consideration (Figure 4.6). The vertical axis represent the duplication rate in terms of number of new SDRs that appeared per million years (My; A for all SDRs and B for SDR type) and in terms of amount of new duplicated sequence (Mbp) per My (C for all SDRs and D for SDR type) for each one of the time-windows considered.

When analyzing the duplication rate by type of SDR I find that all types of SDRs except tandem SDRs contributed to the increase in duplication number of the Homo-Pan-Gorilla ancestor (Figure 4.7 B and D). Tandem SDRs had an increase in duplication activity later, during the time of the Homo-Pan ancestor, although given their typically small size, the contribution in terms of Mbp of this duplication activity is negligible compared to the duplicated Mbp generated during the older SDR burst. Additionally, even though complex mosaic SDRs were the type of SDR that contributed the most to generate new duplicated sequence during the SDR burst (because they are longer), isolated SDRs, both intrachromosomal and interchromosomal, generated higher number of new duplicated regions.

4.3.4 Characterization

Mean identity between duplicates

As stated in Section 4.3.3, identity between duplicates is not always a good proxy for the age of the specific duplications specially if they undergo IGC and/or are under selective pressure (see Section 1.3.1 and Section 1.3.3). Even though identity between duplicates correlates negatively with age (Figure 4.8 A) some deviations from this general trend appear when breaking down SDRs into their types (Figure 4.8 B). Since mosaic SDRs involve several overlapping duplication events, statistics such as mean identity between duplicates might be confounded by the shadowing effect (Cheng et al., 2005). However, tandem and isolated SDRs have not confounding shadowing effect and mean identity between copies is very informative of these regions evolution.

It has been frequently stated in the literature that intrachromosomal SDs have more identity between copies than interchromosomal SDs, suggesting a difference in age or IGC rate between these two types of SDs (Samonte and Eichler, 2002; Hillier et al., 2003; Bailey et al., 2004a; Zhang et al., 2005; Bailey and Eichler, 2006; She et al., 2006). I find some, although low differences in mean identity between copies when comparing intrachromosomal and interchromosomal SDRs (for both isolated and mosaic duplications; Figure 4.8 C). In addition, intrachromosomal SDRs are not, on average, younger than interchromosomal SDRs (Figure 4.8 D).

Instead, it is tandem SDRs that drive the main difference in mean identity between intrachromosomal duplication and interchromosomal duplications reported by several authors. In none of the papers cited above were tandem and intrachromosomal SDRs distinguished. Tandem SDRs show higher sequence identity between copies than the rest of SDRs (Figure 4.8 C and Figure 4.2 C). To understand why tandem SDRs are more identical we have to look at their patterns of identity through time and their average age (Figure 4.2 B and C). Homo specific tandem SDRs are more identical than older tandem SDRs but no differences in average identity are observed between the three older time classes (Figure 4.8 B). This stabilization of divergence between duplicates along time is a typical signature of concerted evolution between duplicates through IGC (see Chapter 3) or selective constraints (see Section 1.3.3). IGC rate has been

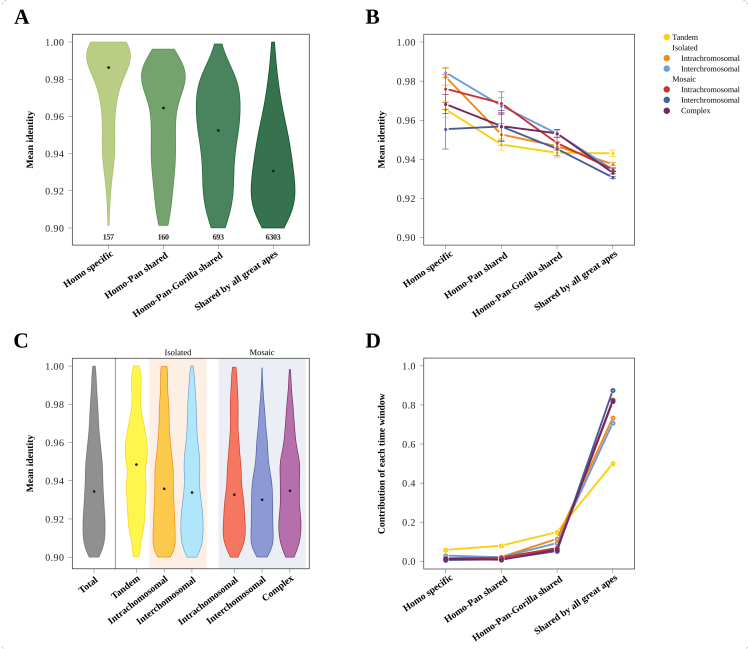


Figure 4.8: SDRs mean identity and average age. A. Violin distributions of the mean identity between copies of Homo specific, Homo-Pan shared, Homo-Pan-Gorilla shared and Shared by all Great Apes SDRs. B. Average mean identity between SDRs and their copies through time windows for each SDRs time class. Standard error bars are represented. C. Violin distributions of the mean identity between of SDRs and their copies for all SDR types. D. Proportion of SDRs in each time window relative to the total number of SDRs of each type.

previously reported to have a negative correlation with distance between duplicates (Lichten and Haber, 1989; Schildkraut et al., 2005; Chen et al., 2007; Zhi, 2007; Benovoy and Drouin, 2009; Casola et al., 2010). Further tests for the presence of specific signals of IGC would be needed to confirm or discard IGC activity. Interestingly, tandem SDRs also appear to be on average younger than the other SDRs (Figure 4.8 D). This is, tandem SDRs have more mean identity than the other SDRs not only because they maintain identity through time but also because they are younger.

GC content

GC content is another property of genomic sequences that has been studied in SDs. Zhang et al. (2005) find a weak positive correlation between SD content and GC content. We confirm this observation by finding significantly higher than expected GC content in all types of SDRs, specially tandem SDRs (Figure 4.9 C), which have, in average, more GC content than the other types of SDRs (Figure 4.9 A). Intrachromosomal SDRs (both isolated and in mosaic) also show slightly more GC content than interchromosomal SDRs. This pattern of mean GC content across SDR classes resembles the pattern observed for identity between copies suggesting a correlation between SDR identity and GC content (Figure 4.8 C and Figure 4.9 A). However, I find no correlation between identity and GC content within SDR classes (Figure 4.9 D); and neither do I find a correlation between GC content and age (Figure 4.9 B).

A possible explanation for higher GC content in tandem duplication might be gBGC. In Section 4.3.4 I already discussed the strong possibility of IGC in tandem SDRs. Since IGC is known to be GC biased (Duret and Galtier, 2009; Glemin et al., 2015; see Section 1.3.1), if it is active in tandem SDRs, I would expect to see a corresponding increase in GC content in these regions. Another possible explanation for GC content in tandem SDs would be that they tend to appear in high GC context although in this case, another explanation would be needed to account for their higher identity. Again, further IGC specific tests would be needed to confirm IGC activity and its GC bias in tandem SDRs.

Genes

Human SDs have been reported to contain an excess of genes (Bailey et al., 2002a; Samonte and Eichler, 2002; Zhang et al., 2005; Sudmant et al., 2013; Dennis et al., 2017) and to be enriched in exons (She et al., 2006). Taking into account the recent age of SDs, any relevant overlap between a gene and an SDR implies a relatively recent duplication of the corresponding part of the gene. Duplication of genic sequence, especially if the duplication involves the complete gene, can entail the formation of novel functional genetic material, modify the dynamics of expression or cause aberrant and/or deleterious forms of the implicated gene (see Sections 1.3.3 and 1.5).

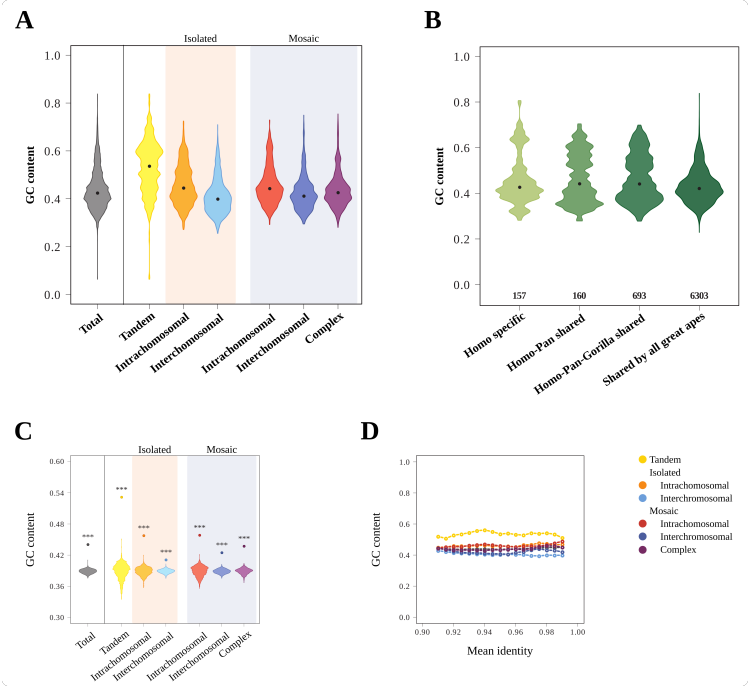


Figure 4.9: GC content of SDRs. A. Violin distributions of the mean GC content of all SDR types. B. Violin distributions of the mean GC content of all SDRs time classes. C. SDR enrichment in GC content. Violin distribution for each type correspond to the mean GC content calculated for 1,000 circularly randomized positions of SDRs across the genome (see methods in Section 4.4). Points correspond to the actual mean GC content of SDRs. Stars correspond to the significance of the enrichment (see methods in Section 4.4): *** for p-values ≤ 0.0001, ** ≤ 0.001 and * ≤ 0.01.

I find that 24.56% (4895/19,932) of all protein coding genes in the human genome overlap with an SDR. This number is within what is expected given the SDRs and protein coding genes span and distribution (see methods in Section 4.4). Differently, the subset of them that are completely within SDRs (883 protein coding genes or 4.43% of the total; Figure 4.10) is more than expected given SDRs and protein coding genes span and distribution (p-value = 0.017). Moreover, SDRs overlap with 4.79% of protein coding exons in our genome with a great majority of them (4.53% of the total) completely within an SDR (Figure 4.10). Both numbers are expected given the SDRs and protein coding

exons span and distribution.

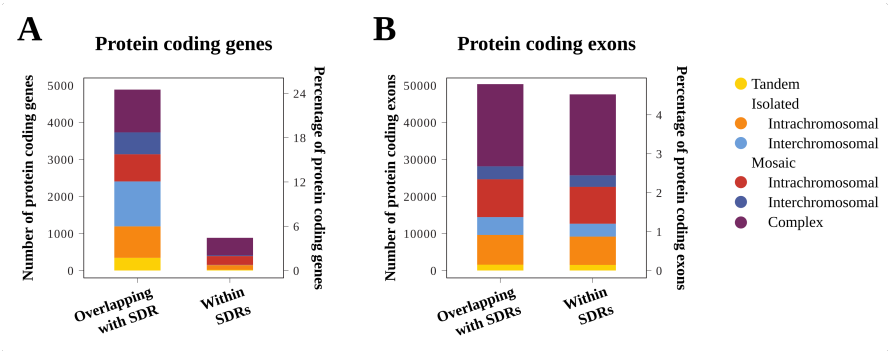


Figure 4.10: Amount of protein coding genes and exons overlapping and within SDRs. Number (and percentage of the total) of protein coding genes (A) and exons (B) that overlap with an SDR and from these, those that are completely within an SDR of a given type.

To understand when and through what type of SDR new entire genes arose during the last 35 - 40 My I formally test for enrichment of protein coding genes entirely encompassed by SDR of different types and time classes (Figure 4.11; see methods in Section 4.4). As state above, the total number of entire protein coding genes contained within SDRs is significantly higher than expected (p-value = 0.017; bottom right corner in Figure 4.11). Most of these genes are within complex mosaic SDRs despite them containing as much entire protein coding genes as expected given their length and distribution (472 genes; right column in Figure 4.11). In fact, are tandem and intrachromosomal (isolated and mosaic) SDRs that are enriched in entire genes, specially tandem SDRs. Interchromosomal SDRs (isolated and mosaic), on the contrary, are significantly depleted in entire protein coding genes (right column in Figure 4.11).

When looking at the time window at which these entire protein coding gene duplications happened it appears that in general, SDRs shared by all great apes, have more entire protein coding genes than expected (Figure 4.11). Indeed, these old SDRs contain most of the genes within SDRs. This enrichment is again driven by tandem and intrachromosomal (isolated and mosaic) SDRs that are highly enriched in entire protein coding genes and compensate the depletion of old interchromosomal (isolated and mosaic) SDRs in entire protein coding

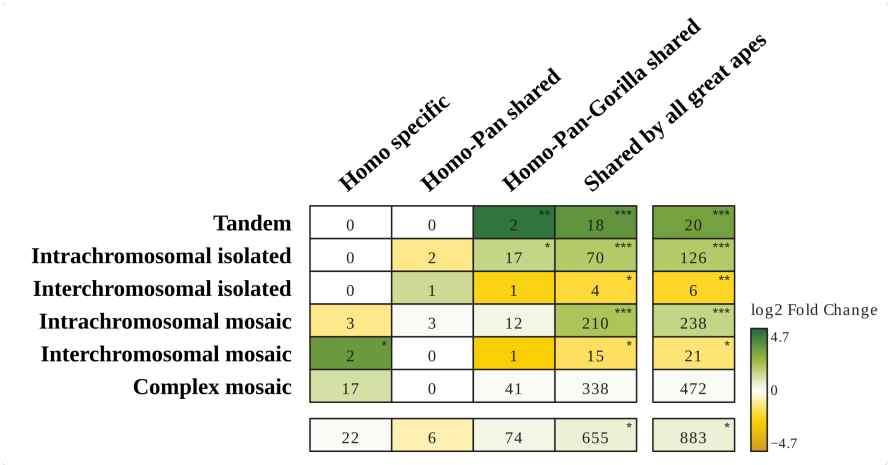


Figure 4.11: Enrichment for entire protein coding genes within all SDRs types and time classes. Rows correspond to the SDR types categories and columns to the SDR time classes. Bottom row and extreme right column correspond to the totals in each column and row, respectively. Bottom-right cell corresponds to the absolute total. Numbers indicate the number of entire genes found within the corresponding SDRs in each cell. Color corresponds to the log 2 fold change with respect to the expected number of protein coding genes calculated as the average entire gene content of 1,000 circular randomizations of the SDRs positions in the genome (see methods in Section 4.4). Cells with green colors correspond those with more entire genes than the random expectation and yellow color cells to those with less entire genes than expected. Stars correspond to the significance of the enrichment test (see methods in Section 4.4): *** for p-values ≤ 0.0001, ** ≤ 0.001 and * ≤ 0.01.

genes. Younger SDRs are not in general enriched in entire genes but Homo-Pan-Gorilla shared tandem and isolated intrachromosomal SDRs also show enrichment in entire protein coding genes as well as Homo specific interchromosomal mosaic SDRs. Nevertheless, the amount of new protein coding genes generated by these most recent duplications is very small.

Entire genes within SDRs, are relevant because they are potential new functional unitsbut to understand the way SDRs arise, it is more informative to look at it the other way around. In Figure 4.12 I show the amount of SDRs within, and overlapping with, protein coding genes and exons. First, although both isolated and mosaic interchromosomal SDRs show depletion in entire

protein coding genes, more than 40% of them are fully covered by protein coding genes (Figure 4.12 A). This means that interchromosomal SDRs tend to be within protein coding genes not the other way round. Moreover, around 10% of isolated interchromosomal SDRs are within exons (Figure 4.12 A). These observations suggest that interchromosomal SDRs (isolated and mosaic) are frequently the result of duplication by retrotranscription and insertion of long RNAs (either processed or not; see Section 4.3.4 and Section 1.3.2). This type of duplication mechanism frequently results in mono-exonic genes, which, in fact, are very frequent among interchromosomal SDRs. Second, almost 60% of tandem SDRs are covered by genes (Figure 4.12 A). Despite this, tandem SDRs show little exon content (Figure 4.12 A), suggesting that tandem SDRs, although being enriched in entire genes (Figure 4.11), are frequently intronic. That tandem SDRs are enriched in entire genes because they contain more entire genes than expected given their short length but the amount of entire genes that they encompass is quite small and they are more frequently find in introns (Figure 4.11). Third, around 50% of both, Homo specific and Homo-Pan shared SDRs are within protein coding genes (Figure 4.12 B) despite being the older categories that show enrichments in entire genes (Figure 4.11). In other words, older SDRs tend to contain more entire protein coding genes than expected while younger SDRs tend to be more frequently within these genes.

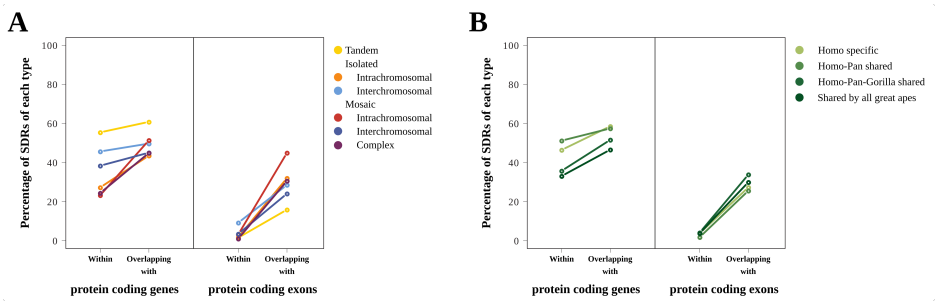


Figure 4.12: SDRs within and overlapping with protein coding genes and exons. Percentage of SDRs in each SDR type (A) and time class (B) within protein coding genes, overlapping with protein coding genes, within protein coding exons and overlapping with protein coding exons.

Natural selection, when acting on the content of duplicates, plays a relevant role in the identity of copies (see Section 1.3.3). Actually, there is a direct

correlation between SDR protein coding gene (and exon) content and identity between copies (Figure 4.13) which is consistent for all types of SDRs (Figure 4.13 A and B). This trend would suggest that either the SDRs with protein coding genes are younger or that there is a selection towards the maintenance of the identity between copies when there are protein coding genes involved but, when breaking it down in time classes, the direct correlation between mean identity and protein coding gene (and exon) content is clearly observed for the two oldest time classes. These results imply that selection and not young age is the most probable cause for the maintenance of identity between copies in SDRs with high gene (and exon) content (Figure 4.13 C and D).

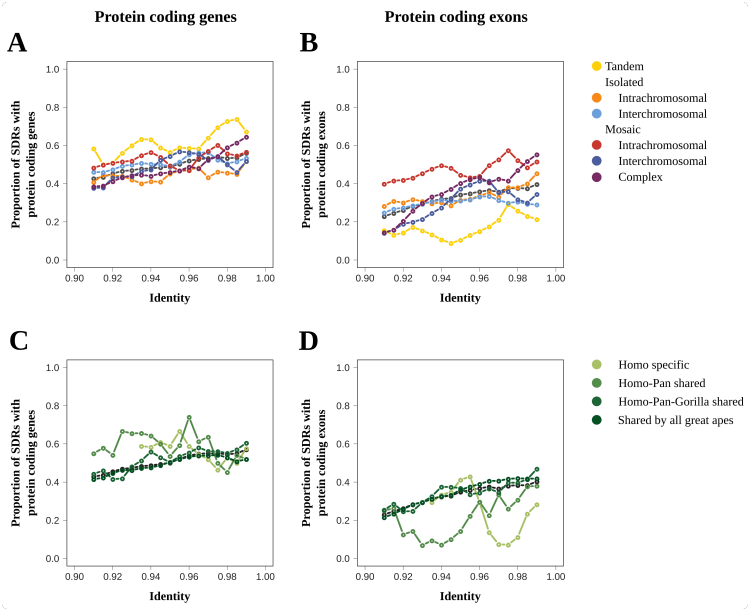


Figure 4.13: Direct correlation between SDR mean identity and protein coding gene and exon content. Horizontal axes represent average mean identity between SDRs and their copies. Each point depicts the mean value per sliding windows of mean identity (window size 0.02 identity points and step of 0.005 identity points). Vertical axes represent the proportion of SDRs overlapping with protein coding genes (A and C) or exons (B and D). Results shown for SDRs type (A and B) and time class (C and D).

Retrotransposons

Retrotransposons and other repeat elements in the human genome are known to be related to the formation of SDs (Bailey and Eichler, 2006; Marques-Bonet et al., 2009b; Monlong et al., 2018). In 2003, Bailey and collaborators presented a model of human SD formation based in retrotransposon-mediated duplication (Bailey et al., 2003). According to their results, in the human genome the main type of retrotransposons responsible for SD formation are Alu elements, specially Alu S elements. Retrotransposon-mediated duplication should not be confounded with duplication through retrotranscription and insertion of a free nuclear RNA molecule (see Section 1.3.2). Retrotransposon-mediated duplication corresponds exclusively to a duplication process produced by NAHR between two homologous retrotransposons located in different loci (Figure 4.14 A-C, see Section 1.3.1 for clarification on NAHR). Differently, duplication by retrotranscription and insertion of a free nuclear RNA molecule do not necessarily imply NAHR (Figure 4.14 D).

According to Bailey's model, a retrotransposon-mediated duplication process results in a duplication surrounded by retrotransposons. I examined the presence of retrotransposons around the borders of SDRs (Figure 4.15). When taking into account all human retrotransposons, duplication borders appear to spatially coincide with a clear transition in retrotransposon content (Figure 4.15 A). SDRs have less retrotransposon content than the genome average, specially tandem SDRs and interchromosomal SDRs (both isolated and mosaic). On the contrary, the flanking sequence of SDRs is slightly enriched in retrotransposons, with the exception of tandem SDRs that seem to be surrounded by regions with low retrotransposon content. Moreover, a sharp increase in retrotransposon content specifically at the SDR border appears for intrachromosomal SDRs (both isolated and mosaic) and complex mosaic SDRs. This peak is likely to be caused by both, retrotransposons that mediated the duplication and that are, thus, located just inside the border of the duplicated sequence, and retrotransposon insertions that disrupted the duplicated sequence breaking the SDR in two and that are, thus, just outside the annotated duplicated sequence. These two situations lead to retrotransposons being just outside or just inside the duplication junction and result in these narrow increase in the duplication border.

Both isolated and mosaic intrachromosomal SDRs, together with complex mosaic SDRs show high levels of Alu S elements in their borders compared to the genome average (Figure 4.15 B). Moreover, for these SDR types there appears to be a sharp drop in Alu S content at around 300 bp inside from the border, which is precisely the characteristic size of Alus (Cordaux and Batzer, 2009). This observation suggests that many of these SDRs are flanked by Alu S elements and provide further support for retrotransposon-mediated duplication being the at the origin some, albeit not all, SDR types. For example, interchromosomal SDRs (isolated and mosaic) show higher than genome average levels of Alu S elements on the flanking region outside the duplication but lower than genome average levels of these elements within the duplication. Similarly, tandem SDRs have low levels of Alu S elements within the duplication and genome average levels in the flanking region. These observations provide an important clarification to the report by Bailey et al. (Bailey2003) of an enrichment in Alu elements in the flanking regions of intrachromosomal duplications. In that work, the authors did not differentiate between intrachromosomal isolated duplications and tandem duplications. In fact, only intrachromosomal isolated duplications are enriched in Alu S elements, whereas tandem duplications are depleted in all types of transposable elements (Figure 4.15 C).

One of Bailey et al.'s (2003) retrotransposon-mediated duplication models, here extended in Figure 4.14, leads to tandem duplication (Figure 4.14 A). My results show that tandem duplications are not frequently duplicated through this process. On the contrary, the presence of Alu S elements on different types of SDRs shows a frequent retrotransposon-mediated origin of intrachromosomal SDRs (both isolated and mosaic). Intrachromosomal SDRs are more probably generated by the retrotransposon-mediated duplication models represented in Figure 4.14 B and C that result in a long distance duplication although other Alu-mediated duplication mechanisms, like Alu mediated FoSTeS, should be considered (see discussion in Chapter 6).

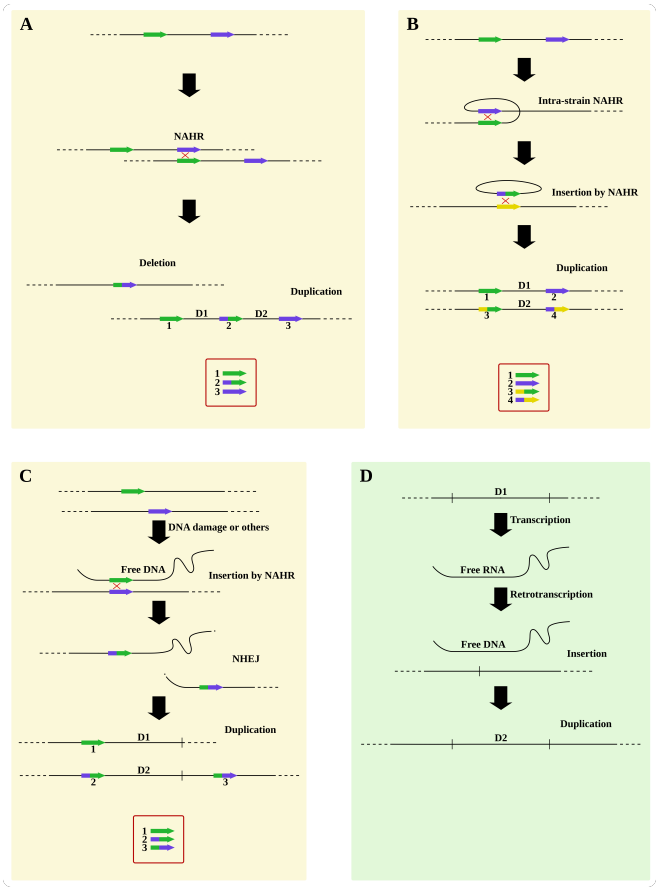


Figure 4.14: Models of retrotransposon-mediated duplication and duplication through retrotranscription. Three models of retrotransposon-mediated duplication (A-C; yellow) are shown together with a model of duplication through retrotranscription and insertion of a free nuclear RNA molecule (D; green). Genomic sequence is represented by horizontal black lines dashed at the ends. A-C. Differently colored horizontal arrows represent retrotransposon elements located in different loci. Arrows of two colors represent hybrid retrotransposon elements product of NAHR between two original retrotransposon elements. Black points indicate the two DNA extremes joined by NHEJ in C. Red boxes summarize the sequence pattern expected in retrotransposon elements after duplication in each retrotransposon-mediated duplication model. Note that D and C are not mutually exclusive. Insertion in D can be performed through NAHR as in C if the free DNA molecule contains a retrotransposon sequence. [Figure based on Bailey et al. (2003).]

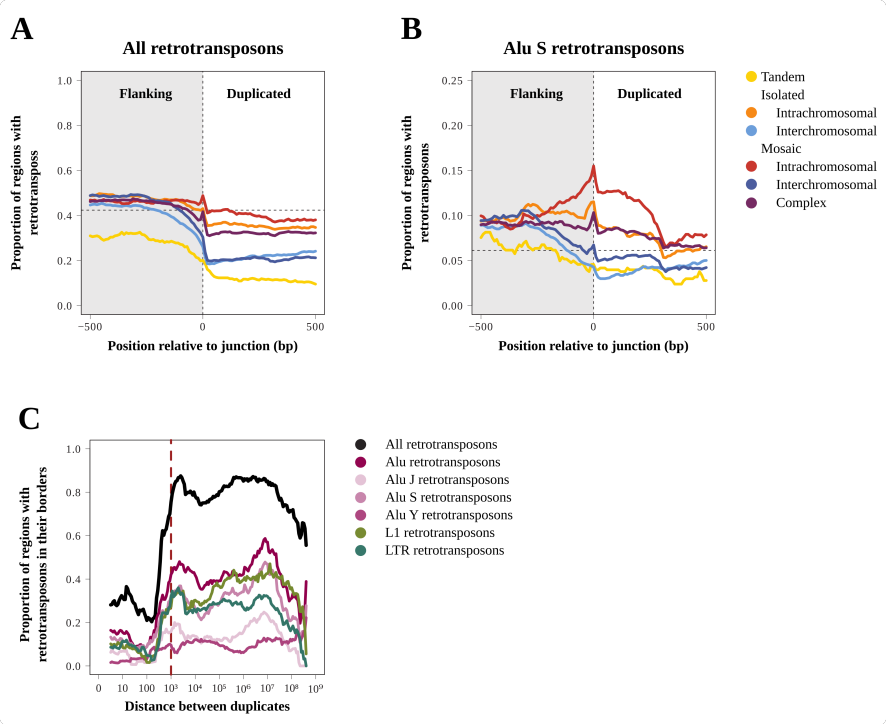


Figure 4.15: Retrotransposons around SDR borders. A, B. Proportion of SDRs having retrotransposons in general (A) or Alu S elements specifically (B) along SDRs borders (position relative to duplication border junction). Results are shown for sliding windows (windows size of 20 bp and window step of 10 bp). The horizontal dashed line represents the genome-wide proportion of all repeat elements (A) and Alu S elements (B). Results are shown for all SDR classes. C. Proportion of isolated intrachromosomal and tandem SDRs with retrotransposons in their borders (window of 20 bp around junction) depending on the distance between copies. Results are shown for all retrotransposons together, for all Alu elements together, Alu J elements, Alu S elements, Alu Y elements, L1 elements and LTR elements separately. Lines show the mean value of an sliding window of size 1 and step equal to 0.05 (log_{10} of the distance). The vertical dashed line corresponds to the distance-between duplicates threshold differentiating tandem SDRs and isolated intrachromosomal SDRs.

4.4 Methods

From SDs to SDRs

All the annotated SDs in the SegDups database of the human reference genome hg38/GRCh38 (UCSC Table Browser; Karolchik, 2004) are grouped in non-overlapping regions: SDRs. That is, all the SegDups annotations mapping to the same genomic region forming a mosaic SD region are grouped into a single mosaic SDR region and considered together. If there is only one SegDups annotation in one genomic region it will be considered an isolated SDR. SDRs one just next to each other are considered separately.

4.4.1 SD enrichment

Calculating SDRs enrichment in other genomic features of interest is not an easy task because human SDRs have a non-random distribution across the genome. A random iteration of the SD position was not considered a good reference for testing the SDR enrichment in different features. Instead, simulated concatenation of the human chromosomes, circularization of the genome and rotation of the SDRs positions was used (1000 different shifts of the position of SDRs covering the entire circularized genome were used). Circularized randomizations respect the relative position of SDRs and were considered a good reference. In each one of these circularized randomizations of the SDR position, a given parameter summarizing the overlap between SDRs and another feature of interest was calculated (e. g. proportion of overlapping sequence with protein coding genes). SDRs enrichment p-values were calculated as the fraction of circular randomizations with higher (or lower) estimates than the actual SDR position estimate. Log 2 fold changes were calculated comparing the actual SDR position estimate with the mean of the circular randomizations estimates.

4.4.2 Annotations of centromeres, genes and retrotransposons

Centromere annotations from UCSC Table Browser (Karolchik, 2004), human gene annotations form Ensembl hg38 (Zerbino et al., 2018) and RepeatMasker

on hg38 for the annotations of retrotransposons (Smit et al., 2013) were used.

4.4.3 Great ape copy number calling based on read depth (WSSD)

I used WGS data of 50 samples: 10 humans, 10 chimpanzees, 10 bonobos, 10 gorillas and 10 orangutans (including *Pongo pygmaeus* and *Pongo abelii* samples) from Prado-Martinez et al. (2013). The procedure applied to estimate copy number for every great ape individual along hg38 was first applied in Serres-Armero et al. (2017) and explained in detail in methods Section of Chapter 5 (Section 5.4).

With this procedure of copy-number calling I obtained copy-number estimates for a sample of individuals of all great ape lineages for each non-repetitive 1 kbp window along the human reference hg38. Moreover, from these estimates I obtained the probability of integer copy numbers in each window and summarized them in 99% confidence copy-number intervals. This is, for any given window of hg38, each individual has a copy number within the defined copy-number interval with a 99% probability.

Note: I do not expect all SDRs to be detected in WSSD. SDRs are based in the SegDups database that considers duplications of $\geq$ 90% of mean identity while WSSD mappings were done with 0.06% of divergence. Moreover, WSSD and WGAC are very different approaches and have very different sensibilities (see Section 1.2.2).

4.4.4 Phylostratification

I used WSSD 99% confidence copy-number intervals for all sampled great ape lineages in WSSD windows along hg38 to determine the duplicated status of each species in human SDRs. Every SDR is phylostratified as follows:

1. Extract all WSSD windows overlapping with the SDR. Only SDRs with 90% of their sequence covered by WSSD windows are considered informative for the phylostratification part of the analysis.

2. Each individual is considered to have either a non-duplicated copy number (normally diploid) or a duplicated copy number including a big variety of copy-number estimates above the diploid copy number (2). An individual is considered to have a non-duplicated copy-number in an SDR if $\geq 70\%$ of the SDR length is estimated to have a copy-number interval including the diploid copy-number (2) or lower. In all the other cases the individual is classified as being duplicated in that SDR.

3. If all individuals of a given great ape lineage, in a given SDR, are considered non-duplicated, the SDR in such great ape lineage is considered non-duplicated. All other situations are considered duplicated.

4. If the SDR was classified as being non-duplicated in humans according to WSSD, it was not included in the phylostratification part of the analysis. These regions are WGAC sensitive regions that are not sensitive for WSSD.

5. The SDR is considered Homo-Pan shared if, apart from humans, it only is duplicated in the Pan lineage. Otherwise, the SDRs is classified as Homo-Pan-Gorilla if it is considered duplicated in the tree lineages but not in Pongo. Finally, all the SDRs that are duplicated in all lineages are classified as being shared among all great apes. Incomplete lineage sorting cases are not considered in the phylostratification part of the analysis.

All the previously described procedure is designed to be very conservative in calling non-duplicated SDRs in the different great ape lineages. This is so to minimize the false positives in recent time-windows.

4.4.5 Duplication rate calculation

I used point-mutation based estimates of the divergence between humans and chimpanzees (7.53 mut/kbp), humans and gorillas (11.23 mut/kbp) and humans and orangutans (22.27 mut/kbp) from Prado-Martinez et al. (2013). Estimation of time was based on point-mutation divergence assuming a constant point mutation rate of $5 \cdot 10^{-10} \frac{mut}{(bp \cdot year)}$:

$$7.53\frac{mut}{kbp \cdot 2branches} \cdot \frac{1bp \cdot year}{5 \cdot 10^{-10} mut} \cdot \frac{1kbp}{10^3 bp} \cdot \frac{1My}{10^6 year} = 7.53My \quad (4.1)$$

Duplication rate was calculated both, in terms of number of new duplications (duplications/My) and in the number of new duplicated base pairs (duplicated bp/My).

This chapter includes the results of an unfinished piece of work describing the variety of SDs in the human genome aimed at understanding how they arise, evolve and contribute to new genic sequences. It has been supervised by Diego A. Hartasánchez and Arcadi Navarro.

Chapter 5

Copy-number variants and duplications in rhesus macaque affecting human disease genes

A ship in port is safe, but that's not what ships are built for.

Grace Hopper

For the things we have to learn before we can do them,
we learn by doing them.

Hannah Arendt

5.1 Rationale

Rhesus macaque is an abundant species of the family Cercopithecidae (Old World monkeys) that diverged from the superfamily Hominoidea (including gibbons or Hylobatidae and great apes or Hominidae) about 25 Mya. They are the non-human primates with the largest geographic range, expanding most of the South of Asia, and with exceptional levels of adaptability to a great variety of conditions (see Section 1.4).

Because of their abundance and adaptability, and given their genetic and physiological closeness to humans, rhesus macaques are the non-human primates most extensively used in biomedical research (Xue et al., 2016). They have largely been studied as a model organism for understanding human metabolism, physiology and disease, with special emphasis in infectious diseases and alcohol addiction (Schwandt et al., 2010; Uno et al., 2011, 2016; Wiseman et al., 2013; Walter and Ansari, 2015; Chong et al., 2018). Rhesus macaques have also been used to model and understand the genetic background of certain diseases (Champoux et al., 2002; Barr et al., 2004; Gibbs et al., 2007; Loffredo et al., 2007; Vallender et al., 2008, 2010; Valentine et al., 2009; Rogers, 2013; Vinson et al., 2013).

Considering the impact that studies on these animals might have on human well-being, acquiring knowledge on the genome and the genomic diversity of rhesus macaques is of major relevance in order to identify and understand the genetic and physiological differences between them and humans. This knowledge would help to evaluate the advantages and limitations of using this

species as a model organism. In accordance, great efforts have been performed during the last decade to describe the genetic diversity of rhesus macaques and have importantly contributed to this goal (Smith and Mcdonough, 2005; Lee et al., 2008; Xue et al., 2016; Bimber et al., 2017; Liu et al., 2018).

Nevertheless, so far systematic studies on duplications and CNVs in the rhesus macaque genome have been performed with low sample sizes, only through array-CGH and/or focusing in other species genomes (Lee et al., 2008; Marques-Bonet et al., 2009a; Gokcumen et al., 2011). This is so with the notable exception of Lorente-Galdos et al. (2013) where the authors focused on finding accelerated evolution in human and rhesus macaque duplicated exons. Therefore, there is a lack of a high resolution picture of rhesus macaque duplications and CNVs performed with a large sample size. Moreover, specific CNVs in the species have been reported to have functional consequences and potential implications for biomedical research (Degenhardt et al., 2009; Hellmann et al., 2011, 2012, 2013; Taormina et al., 2012; Ottolini et al., 2014; de Groot et al., 2015). With this in mind, having genome-wide information on the number of copy in the rhesus macaque genome covering its species variation would be of great use in biomedical research.

In this chapter, I present the first genome-wide map of the copy-number architecture and variation found in the rhesus macaque genome. In this collaborative project, I identified genes with high probability of having their function modified by copy-number alterations compared to human and, more importantly, genes that present different copy number or distinct copy-number diversity between both species. These genes are strong candidates for having divergent genotype-phenotype mechanisms between humans and macaques and, as such, should be handled with care in studies using rhesus macaque as a model organism since the results may not be comparable to humans.

5.2 Objectives

General:

1. To describe the copy-number architecture and copy-number variation of the rhesus macaque genome.
2. To identify and describe potential functional differences between humans and macaques due to differences in copy number.

Specific:

1. To create a genome-wide map of copy number of the rhesus macaque genome encompassing a great part of the diversity in copy-number in the species.
2. To identify CNV regions and non-diploid fixed regions in the rhesus macaque genome.
3. To recover non-diploid genes with copy-number profiles potentially influencing their function by altering the number of functionally competent copies of the gene in the genome.
4. To create a comparable map of copy number of the human genome and recover the human non-diploid genes with different number of functional copies.
5. To compare the copy-number profiles of human-rhesus macaque gene orthologs.
6. To identify genes which differences in copy number within and/or between the two species might lead to phenotypically relevant changes and explore the potential impact of these differences in biomedical research.

5.3 Results and discussion

5.3.1 High-quality genome-wide maps of fixed duplications and CNVs

The aim is to better understand the copy-number architecture of the rhesus macaque genome, and the implications that the gene copy-number variation within and between the human and macaque species might have on biomedical research. To tackle this issue I generated a fine-scale genome-wide map of copy number for a sample of 198 rhesus macaques. This panel of samples was obtained after applying a very stringent quality-control filtering to whole-genome sequences from 315 samples integrating 202 newly sampled individuals and 213 samples from Xue et al. (2016) (see methods in Section 5.4).

To be able to call non-diploid copy number regions with high confidence, specially CNV regions, I used a copy-number calling procedure that provides a 99% confidence interval based on raw read depth for each non-repetitive 1 kbp window in every sample (Serres-Armero et al., 2017, Figure 5.1 A; see methods in Section 5.4). I used these copy-number confidence intervals to differentiate diploid from non-diploid regions, out of noise-related intra- and inter-sample read-depth variability. Among the non-diploid regions, I confidently distinguished between *fixed duplications* and *CNV regions* (including gains, losses and gain/losses). Fixed duplications are regions with the same non-diploid copy number in all the individuals of my sample, while CNV regions are regions in which I could distinguish, with high confidence, more than one copy-number allele within my sample. Non-diploid regions with uncertain copy number distribution were labeled as *unclassified regions*. Typically, these regions contain different alleles with similar copy number that could not be distinguished from inter-sample noise with my confidence interval. Accordingly, the unclassified category contained many actual CNV regions for which I could not distinguish two clear copy-number alleles alongside read-depth noisy regions (see methods in Section 5.4).

Working with copy-number intervals provides high confidence in copy-number calls. However, it imposes specific restrictions that must be considered when

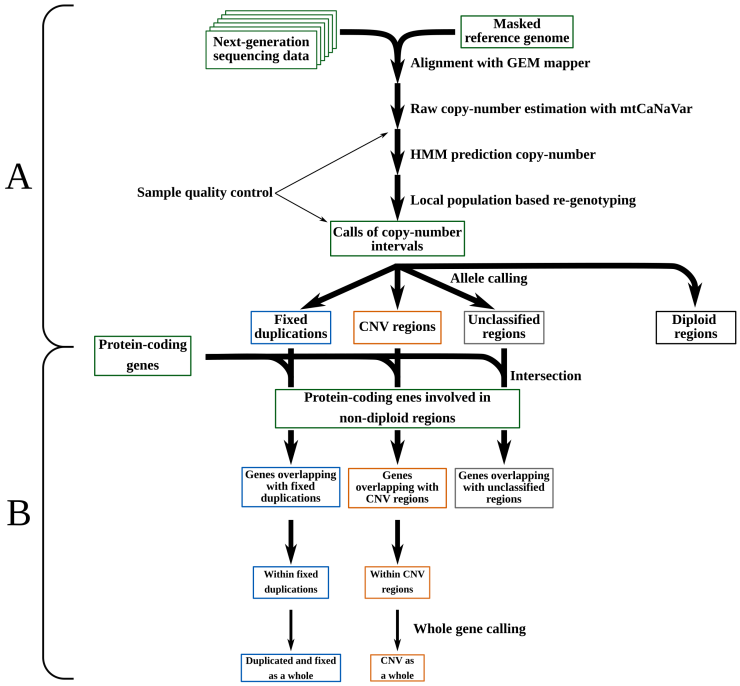


Figure 5.1: Summary of the methodology. A. Starting with NGS data and a masked genome reference, I followed a series of steps in order to call copy-number intervals for each non-repetitive 1 kbp window and sample (see methods in Section 5.4). I performed a copy-number allele calling per window to classify all non-diploid windows in fixed duplications, CNV regions or unclassified non-diploid regions. B. To assess the functional implications of these non-repetitive windows, I crossed the protein coding genes in the corresponding genome with the three types of non-diploid windows (see methods in Section 5.4). From all the genes that were related to non-diploid windows, I distinguished between those overlapping with the three different categories of non-diploid windows. Among those genes overlapping with fixed duplications I focused on those that had 90% or more of their coding region duplicated and fixed and, among these, I focused on the ones that were duplicated and fixed as a whole (see methods in Section 5.4). Equally, among those genes overlapping with CNV regions, I focused on those genes within CNV regions (90% or more of the coding region) and those that were CNV as a whole. The same pipeline was applied to both the rhesus macaque and the human set of samples.

compared to copy-number calls obtained with other methods. In order to perform reliable inter-species comparison between human and rhesus

macaques, I applied the same algorithm to a set of 35 WGS human samples from the Simons Genome Diversity Project (Mallick et al., 2016; 50 samples before quality control, see methods in Section 5.4). The individuals in this sample were selected as a balanced representation of human populations, providing an ideal dataset for retrieving inter-populations copy-number variability. Additionally, fine-scale maps of copy number in the human genome were compared and validated using the annotation of SDs in the human genome (UCSC Table Browser; Karolchik, 2004) and the CNV calls performed by Peter Sudmant et al. with the 1000 Genomes Project phase 3 WGS data (Sudmant et al., 2015a; see methods in Section 5.4).

Non-diploid regions are known to be spread along all the human genome but to show regions of high clustering (Marques-Bonet and Eichler, 2009) and to be enriched in pericentromeric and subtelomeric regions (Bailey et al., 2002a; She et al., 2004; Bailey and Eichler, 2006). A genome-wide overview of fixed duplications, CNV regions and unclassified non-diploid regions in the rhesus macaque genome shows a similar distribution (Figure 5.2 A). Non-diploid regions are common (12.25% in bp) and widespread along the rhesus macaque reference genome. My human copy-number maps show, as expected, high clustering of non-diploid regions in pericentromeric and subtelomeric regions (Supplementary Figure 7.1 A in Section 7.4). The same clustering is appreciated in my macaque copy-number maps but cannot be seen clearly in the case of pericentromeric regions because of the lack of resolution and annotation of centromeres and telomeres in the genome assembly.

Genomic regions consisting in fixed duplications appear to be the longest type of non-diploid regions (Figure 5.2 B). The same pattern appears in the human copy-number maps (Supplementary Figure 7.1 B in Section 7.4). The difference in length between fixed duplicated and CNV regions can be explained by two reasons. First, duplications are known to be organized in big mosaic clusters around ancestral duplicons (She et al., 2004; Bailey and Eichler, 2006; Jiang et al., 2007; Marques-Bonet and Eichler, 2009). These regions will mainly show up as fixed duplications in my classification due to their old age. Second, as stated above, CNV regions are more difficult to call, resulting in some of them being fragmented by unclassified segments that might be actually unresolved CNV segments (see methods in Section 5.4). Also, among CNV regions, CNV losses and gain/losses are shorter than CNV gains (Figure 5.2).

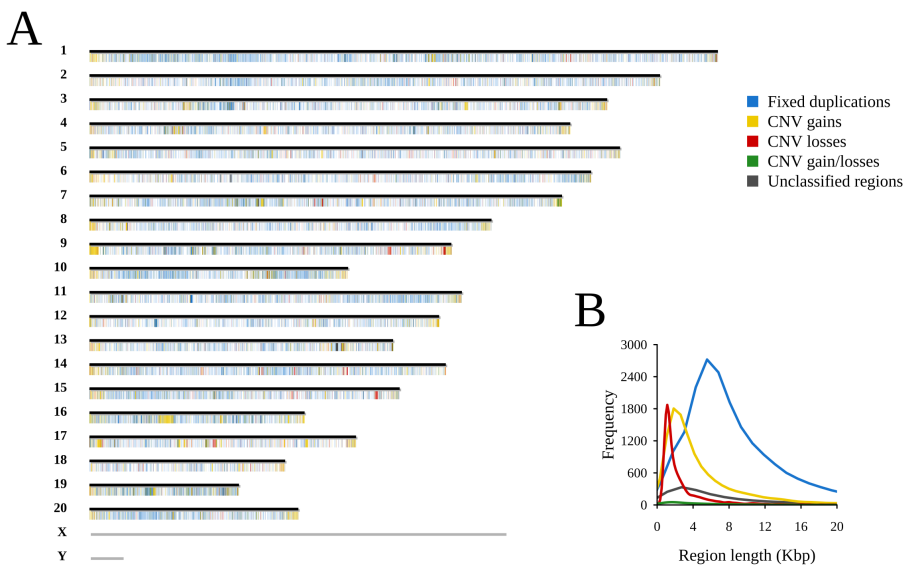


Figure 5.2: Landscape of the non-diploid regions in the rhesus macaque reference genome. A. Genome-wide map of the non-diploid regions. Chromosomes are represented by horizontal bars. Fixed duplications, CNV gains, CNV losses, CNV gain/losses and unclassified regions are represented as color-coded marks. Mark width is proportional to the region length with a minimum mark width of 25 kbp for short regions for the sake of visualization. Sex chromosomes were not included in the analysis. B. Length distribution of the non-diploid region types.

In the rhesus macaque reference genome, fixed duplications cover a 8.59% (in bps) of the rhesus macaque reference genome, CNV gains a 2.18%, CNV losses a 0.4%, CNV gain/losses a 0.07% and unclassified non-diploid regions a 1.03%. The lower quality of this reference genome compared to the human one suggests that it might contain unresolved duplications. Fortunately, the read-depth approach is independent of the resolution of duplications in the reference genome and I was able to call the non-diploid regions anyway (Bailey et al., 2002a; Marques-Bonet and Eichler, 2009). For this reason, the collapse of duplicated regions does not affect my estimation of the number of copies but it can affect the estimation of the number of non-diploid regions (*i.e.* I might be underestimating the number of non-diploid regions). Comparison of copy-number maps shows higher prevalence in humans of fixed duplications

and unclassified regions compared to the prevalence of CNVs (10.2% in bps of fixed duplications, 2.3% of unclassified regions, 1.6% of CNV gains, 0.18% of losses and 0.03% of gain/losses). My results are in agreement with rhesus macaque having higher copy-number diversity than humans mirroring SNVs distribution, what would point to a higher effective population size for macaques (Xue et al., 2016). However, I cannot disregard an underestimation of CNV in humans because of the difference in the used sample sizes.

Interestingly, there are differences in the chromosomal density of non-diploid regions (Figure 7.2 A). Rhesus macaque chromosome 19 shows a particularly high density of all non-diploid types of regions (fixed duplications, CNV gains and unclassified non-diploid regions) compared to the rest of the chromosomes. This observation coincides with the genomic distribution of macaque SDs (UCSC Table Browser; Karolchik, 2004) across chromosomes (Figure 7.2 B). Additionally, I independently retrieved information on mammal, primate and rhesus-specific gene duplications (Juan et al., 2013; see methods in Section 5.4) and explored their distribution across chromosomes. I did not observe higher density of genes specifically duplicated in the macaque lineage in chromosome 19 compared to other chromosomes (Figure 7.2 B). The reason behind this observation can be that these genes are the most recent and similar and, thus, the most likely to be collapsed in the genome assembly. Nevertheless, I observed a higher density of other genes duplicated in primates and in mammals in chromosome 19. A great part of these genes are expected to be shared with humans, which also show especially high contribution of non-diploid regions (basically fixed duplications) in chromosome 19 (Supplementary Figure 7.3 in Section 7.4). Additionally, human chromosome 9 is known to have huge pericentromeric non-diploid regions which here appear mainly as CNV regions (Bailey et al., 2001; Sudmant et al., 2013).

5.3.2 Copy-number profile of rhesus macaque and human protein-coding genes

To explore phenotypic differences between humans and rhesus macaques due to differences in copy number, I first had to detect those copy-number alterations within each species with a higher functional impact. To do so, I cross-examined my genome-wide maps of copy number for the two species and the annotation

of protein-coding genes in the corresponding reference genomes (see methods in Section 5.4).

However, assessing the functional impact of copy-number genotypes is not straightforward. A duplication of a small part of a protein-coding gene has a low probability to have a functional impact even if it is found in the protein-coding region (see Chapter 6 for further discussion on this issue). Now, whole-gene duplications, encompassing all the protein-coding length of a given gene suggest a higher number of potentially functional gene copies that can affect its expression levels with functional consequences. For this reason I focused on the protein-coding region (exons) of the main isoforms of protein-coding genes in the two species and distinguished between three levels of relationship with non-diploid regions. First, I identified all genes with at least part of their exons showing non-diploid copy number (overlapping with a non-diploid region). Second, among these, I focused on those genes having 90% or more of their protein-coding sequence duplicated and/or varying in copy number (within fixed duplications or CNV regions). And, finally, among the latter, I found those with a constant copy number per sample along the whole protein-coding region (Figures 5.3 and 5.1 B; see methods in Section 5.4). These two last groups of genes (termed *duplicated and fixed as a whole* and *CNV as a whole*) are protein-coding genes that have more than one genomically competent copy or a variant number of them within my samples (respectively; Figure 5.3).

According to these categories, I classified every protein-coding gene in the rhesus macaque and human genome (Figure 5.4 and Supplementary Figure 7.5 in Section 7.4; see methods in Section 5.4). My maps show that a great part (32.12%) of the rhesus macaque protein-coding genes overlaps with either fixed duplications (21.05%, 4409 genes), CNV regions (10.81%, 2264 genes) or other unclassified non-diploid regions (4.42%, 925 genes). Gene length is inversely correlated to the proportion of the protein-coding gene sequence having non-diploid copy number (Supplementary Figure 7.6 in Section 7.4), suggesting that most of these copy-number changes in genes are neutral and happening randomly throughout the genome. Nevertheless, in the rhesus macaque genome, there is a significant amount of protein-coding genes completely within non-diploid regions (Supplementary Figure 7.6 C in Section 7.4). These genes tend to be small (Supplementary Figure 7.6 D in Section 7.4) because they are more easily entirely encompassed by a duplication.

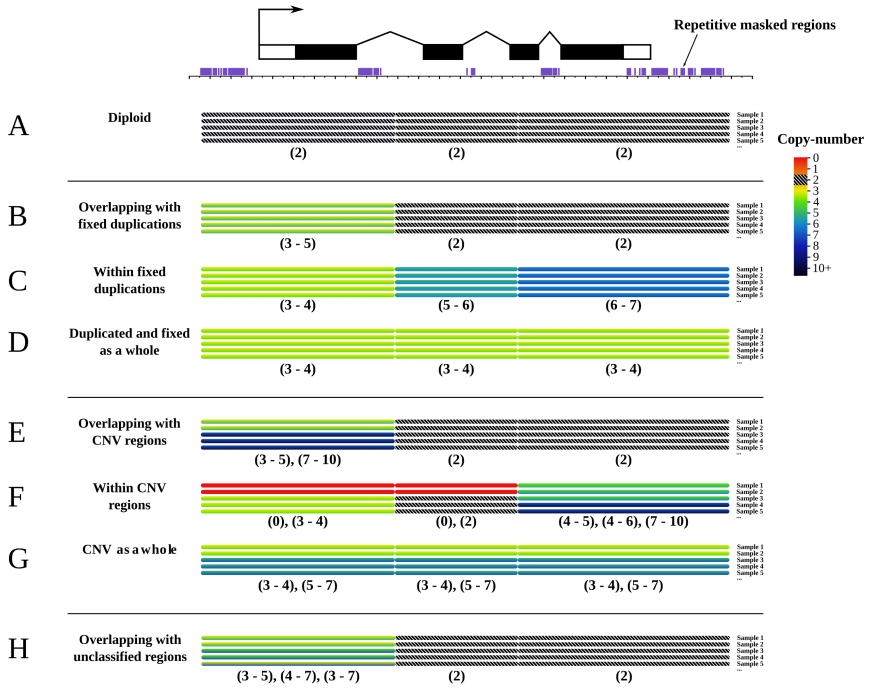


Figure 5.3: Schematic example of my gene classification according to its copy-number context. Top: schematic representation of a protein-coding gene (exons shown as black boxes) over the track of repetitive regions (in purple) that are masked from the reference genome before mapping (see methods in Section 5.4). Panels A-H: three consecutive non-repetitive 1 kbp windows covering the length of the gene are represented in all the panels. 99% confidence copy-number intervals for 5 samples are depicted in color transitions (plus a black and white pattern for the diploid copy number) as examples. All the copy number intervals present in each window are listed below each case. The 8 gene-copy-number relationships considered are exemplified (A-H). First, the gene will be classified as being diploid if all of its protein-coding length is covered by diploid windows (A). Second, the gene will be classified as overlapping with fixed duplications (B), CNV regions (E) or other unclassified non-diploid regions (H) if at least part of its length is covered by the respective type of non-diploid region. Third, the gene will be considered within fixed duplications (C) or within CNV regions (F) if 90% or more of its coding length is covered by the corresponding type of non-diploid region. Finally, the gene will be classified as being fixed and duplicated as a whole (D) or being CNV as a whole (G) if there is a constant copy number per sample along all its protein coding region (see methods in Section 5.4).

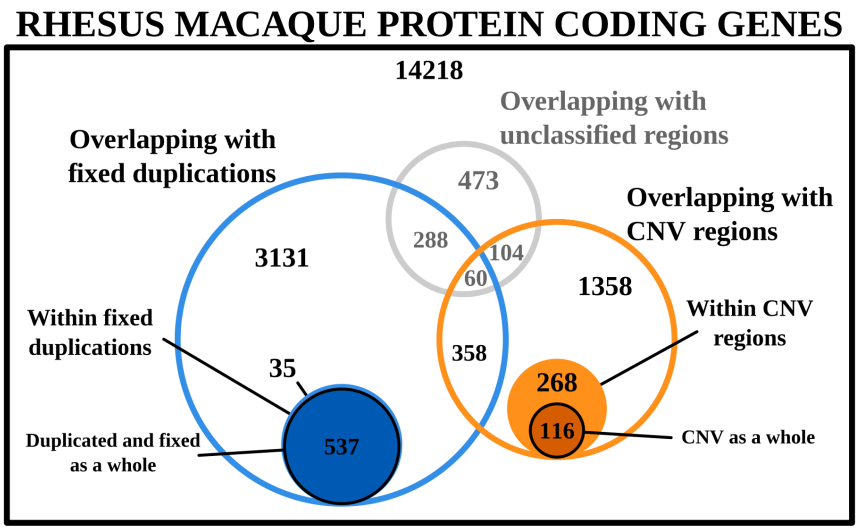


Figure 5.4: Copy-number profile of rhesus macaque protein-coding genes. Venn diagram representing, first, the entire set of rhesus macaque protein-coding genes considered (black square; a total of 20,946 protein-coding genes, 6,728 of which overlap with a non-diploid region); second, the number of genes overlapping (at least 1bp) with fixed duplications (blue; a total of 4,409 protein-coding genes), CNV regions (orange; a total of 2,264 protein-coding genes) and unclassified regions (grey; a total of 925 protein-coding genes) in open circles; third, genes that have at least 90% of their coding sequence within fixed duplications (blue; a total of 572 protein-coding genes) and CNV regions (orange; a total of 384 protein-coding genes) light solid circles and; fourth, genes that are fixed as a whole (blue) or CNV as a whole (orange) in dark solid circles. Note that numbers in the figure correspond to the number of genes in the area in which they are located excluding genes in other areas. All areas are approximations.

As shown in Figure 5.4, a 2.73% (572) of the protein-coding genes have 90% or more of the total length of their exons overlapping with fixed duplications and a 1.83% (384) with CNV regions. These genes have more probability of having their function affected by their copy-number profile. Moreover, among these genes, 537 are duplicated and fixed as a whole and 116 are CNV as a whole.

Comparison between human and macaque copy-number profile of protein-coding genes (Supplementary Figure 7.5 in Section 7.4) shows a higher

proportion of genes overlapping with fixed duplications in humans (25.52%). This is in agreement with the higher proportion of this type of regions in the human reference genome (Figures 5.2 and Supplementary Figure 7.1 in Section 7.4). There are fewer genes overlapping with CNV and unclassified regions in humans. This can be explained probably because of my lower sample size and samples distribution (4.37% for CNV regions and 3.46% for unclassified regions). The same happens with genes within CNV regions (128 genes) and CNV as whole genes (37 genes). There are less genes within fixed duplications (1.32% or 225 genes) or duplicated and fixed as a whole (134 genes) in human than in rhesus macaque. There are also fewer genes with all the exons with a change in copy number and less whole gene duplications in humans compared to rhesus macaque.

5.3.3 Comparing rhesus macaque and human genes copy-number profile

Once having cataloged the potential functional impact of the number of copies of genic regions in the two species, I could accomplish my final goal which was to identify differences in copy-number between the two species with high probabilities of having functional consequences. To tackle this I compared the copy-number profile of human and rhesus macaque orthologous gene pairs (see methods in Section 5.4).

Orthologies are not always one to one relationships but they can contain one to many or many to many relationships depending on the duplication history of the specific gene in both branches and on the quality of the corresponding genome assemblies. This aspect of orthologies is especially important to consider when studying changes in copy number in genes. In my case, high identity of the detected duplicates leaves little space for functional divergence between them. Nevertheless, I decided to implement the comparison between the two species with all pairs of orthologs being aware that not all of the annotated orthologous relationships were pairs of genes that conserved the same function. In parallel, I performed the comparison with only one to one orthologies. Private human genes and private rhesus macaque genes are not considered in this part of the analysis because they lack orthologs in the other species. For the sake of completeness, within this analysis I reported human genes within the Sudmant et al. (2015a)

CNV dataset or within SD annotations in the human genome (see methods in Section 5.4).

As expected, most of the orthologous pairs were diploid in both species. I found some genes with functionally disruptive different copy-number profile between human and macaques (Figure 5.5) and I also identified some rhesus macaque genes with orthologs within Sudmant et al. (2015a) CNV calls and within human SDs.

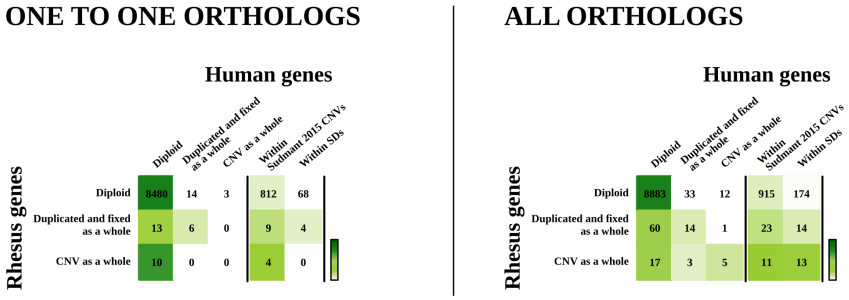


Figure 5.5: Copy-number profile comparison of orthologous pairs between human and rhesus macaque. A. Orthologous pairs with one to one relationship between human and macaque. B. All protein-coding genes in the rhesus macaque genome that have an annotated orthologous pair in humans (see methods in Section 5.4). Numbers correspond to the number of orthologous pairs in each of the represented categories: diploid genes, genes duplicated and fixed as a whole, genes CNV as a whole and genes within Sudmant et al. (2015a) human CNV calls and within human SDs for humans. Shades of green represent the proportion in each cell of the total genes with orthologs in each row (form dark green being equal to 1 and white being equal to 0). See Supplementary Figures 7.8 and 7.7 in Section 7.4 for the complete version of these tables.

Next, I studied the relevance for human disease of the few one to one pairs with different predicted number of copies in the two species. I explored complex disease related genes by intersecting my genes of interest with the GWAS catalog database (MacArthur et al., 1895), a compilation of all GWAS hits related to genes (see methods in Section 5.4). Moreover, to have a complete view of the human gene-disease associations I used the DisGeNET database (Bauer-Mehren et al., 2010), a compilation of gene-disease databases (see methods in Section 5.4). Table 5.1 includes a list of the identified associations of human disease and genes having relevant differences in copy-number profile

between human and macaque.

Interestingly, several important complex diseases related genes appear to have relevant differences between human and rhesus macaque (Table 5.1). These genes are of great relevance for the corresponding diseases biomedical research using rhesus macaque as a model organism. Among them, we found many genes related to complex neuropsychiatric and neurodegenerative disorders[1]. These results are of great relevance for the study of these diseases that frequently use rhesus macaque as a model organism because of its proximity to human neuro-physiology.

Of special interest is PPP2R5C. PPP2R5C encodes for a regulatory subunit of the protein phosphatase-2A (PP2A), an intracellular serine/threonine phosphatase (a tumor suppressor protein). It is diploid in humans and duplicated and fixed as a whole in rhesus macaque. In other words, my analysis shows that PPP2R5C probably has more functional copies in the rhesus macaque genome than in the human genome. Besides being involved in cancer, my results on PPP2R5C are of special relevance because it has been related to autism (Anney et al., 2010), a very commonly studied disease using rhesus macaque as model organism (Bauman and Schumann, 2018).

[1] This does not mean that humans and rhesus macaques have more differences in copy number in genes related to this type of diseases because the diseases appearing in the GWAS Catalog (MacArthur et al., 1895) might be biased in this regard.

Human gene	Disease	Database	Human copy-number profile	Rhesus copy-number profile
TXN	*Late-onset Alzheimer's disease*	*GWAS Catalog*	*Diploid & within Sudmant et al. (2015a) CNVs*	*Duplicated and fixed as a whole*
ENY2	*Multiple system atrophy*	*GWAS Catalog*	*Diploid*	*Duplicated and fixed as a whole*
MPC2	*Schizophrenia*	*GWAS Catalog*	*Diploid*	*Duplicated and fixed as a whole*
PPP2R5C	*Autism*	*GWAS Catalog*	*Diploid & within Sudmant et al. (2015a) CNVs*	*Duplicated and fixed as a whole*
ILF2	*Multiple myeloma*	*GWAS Catalog*	*Diploid*	*Duplicated and fixed as a whole*
SYCP3	*Azoospermia, Arrest of spermatogenesis*	*DisGeNET*	*Diploid*	*Duplicated and fixed as a whole*
POU4F1	*Alzheimer's disease*	*GWAS Catalog*	*Diploid*	*CNV as a whole*
CLC	*Schizophrenia*	*GWAS Catalog*	*Diploid*	*CNV as a whole*
VAPB	*Amyotrophic lateral sclerosis, spinal muscular atrophy*	*DisGeNET*	*Diploid*	*CNV as a whole*
CCDC115	*Congenital disorder of glycosylation*	*DisGeNET*	*Diploid & within Sudmant et al. (2015a) CNVs*	*CNV as a whole*
APOL2	*Chronic kidney disease*	*GWAS Catalog*	*Duplicated and fixed as a whole & within Sudmant et al. (2015a) CNVs*	*Duplicated and fixed as a whole*
HIST1H2AC	*Narcolepsy with cataplexy, Lung carcinoma, Schizophrenia*	*GWAS Catalog*	*Duplicated and fixed as a whole*	*Diploid*
NAT8	*Chronic kidney disease*	*GWAS Catalog*	*Duplicated and fixed as a whole*	*Diploid*
JUND	*Asthma*	*GWAS Catalog*	*Duplicated and fixed as a whole*	*Diploid*
LILRA5	*Prostate carcinoma*	*GWAS Catalog*	*Duplicated and fixed as a whole*	*Diploid*
HSD17B7	*Dupuytren Contracture*	*GWAS Catalog*	*Duplicated and fixed as a whole*	*Diploid*
SPDYE4	*Attention deficit hyperactivity disorder*	*GWAS Catalog*	*Duplicated and fixed as a whole*	*Diploid*
PVRIG	*Acute coronary syndrome*	*GWAS Catalog*	*Duplicated and fixed as a whole*	*Diploid*
HSPB8	*Trypanosoma cruzi seropositivity, Charcot-marie-tooth Disease, Neuronopathy, Distal Hereditary Motor*	*GWAS Catalog*	*Duplicated and fixed as a whole*	*Diploid*

Table 5.1

5.4 Methods

Sampling and data generation

I used sequence data from a panel of 315 rhesus macaques. This set includes 113 previously published samples (Xue et al., 2016) and 202 newly sequenced samples. WGS was performed using the Illumina HiSeq 2000 platform generating 100-bp paired-end reads.

For human comparison I used a panel of WGS data from 50 human samples especially selected from the Simons Genome Diversity Project (Mallick et al., 2016) that covered a large part of the human diversity.

Copy-number calling

In order to call copy number genome-wide with high confidence I used a *Hidden Markov Model* (HMM; Stratonovich, 1960) algorithm (first applied for copy-number estimation and described in detail in Serres-Armero et al., 2017) on top of continuous copy-number inferences based on read depth (Alkan et al., 2009).

I here describe the procedure adapted to the data in this study:

1. Repeat- and over-represented k-mer- masking of the assembly. I started masking all of the interspersed and simple repeats annotated in the assembly Mmul_8.0.1/rheMac8 (GRCh37/hg19 for humans) with RepeatMasker (Smit et al., 2013) and Tandem Repeat Finder (Benson, 1999). Then, in order to mask over-represented k-mers, I divided the assembly into 78 bp sliding k-mers. These k-mers were mapped to the assembly using GEM mapper (Marco-Sola et al., 2012). All positions with more than 20 k-mer placements were masked.

2. Mapping of WGS reads to the masked assembly. WGS reads were mapped to the previously masked version of rheMac8 using GEM mapper (Marco-Sola et al., 2012). In order to prevent underestimation of the copy number in the surroundings of the masked repetitive regions, I masked the 78 bps flanking all masked regions after the mapping.

3. Defining genome windows. I divided the assembly in 1 kbp non-overlapping windows of non-masked sequence and I calculated the read depth in these windows using mrCaNaVaR (Alkan et al., 2009).

4. Raw copy-number estimation per window. Mean read depth per window for a set of control diploid regions (form mrCaNaVaR) was taken as the read depth corresponding to the diploid copy number (copy number 2). This read depth was used as corresponding to copy-number 2 to scale the mean read depth in all the other windows to a raw copy number (continuous value directly proportional to the mean read depth). This procedure takes into account GC content and corrects for it with mrCaNaVaR.

5. Integer copy-number probability estimation with HMM (Stratonovich, 1960). I used a HMM to estimate the probability of a given window having an integer copy number from its raw copy number. In this HMM, the observed values are the raw copy-number estimates per window, the hidden states are real integer copy numbers (0, 1, 2, 3, ..., 20), and the transitions between states are changes in copy number between adjacent windows. The HMM hidden states for high copy numbers were set as intervals (21-100, 101-500 and 501-1000) instead of integers due to the correlation between copy number and noise. Emission probabilities were extracted from the corresponding read-depth distributions. For low copy-number states, these distributions were defined as normal distributions with mean equal to the corresponding copy number ($\mu_N = N$) and standard deviation $\sigma_N = \sigma_{CR}\sqrt{\frac{N}{2}}$, where σ_{CR} is the standard deviation in control regions. For copy numbers above 20, emission distributions were a mixture of the corresponding normal distributions weighted proportionally to the estimated frequencies of each copy number. I used the Baum-Welch algorithm (Rabiner, 1989; Pomegranate Python package) to train the transition matrix of this HMM until convergence with a random set of samples that were excluded from further analysis. Finally, in order to predict the probability of each of the hidden states for each window in each individual, I used the forward-reverse algorithm coded in the Pomegranate Python package.

6. Local population based re-genotyping. I applied a population-based

correction for noisy local copy-number estimates. I used Bayes' theorem (Bayes, 1763) to estimate the probability of observing a given raw copy-number estimate in a given window for a given individual given the different distributions corresponding to the different copy-number states. In this correction, the prior probability of a given copy number in a given window $(p(N))$ was the only parameter that changed locally. The prior probability was set as the average probability of this copy number across all individuals in 5 consecutive windows centered at the window of interest. See Serres-Armero et al. (2017) for further information.

7. Defining copy-number intervals with a confidence level of 99%. For each window and each individual I ranked all the possible copy-number states by their probability, starting with the top copy-number state, I added states to the copy-number interval, for example, from (3-4) to (3-5) until the cumulative probability of a given window in a given individual having a copy number within the copy-number interval was 99%. The underlying actual copy number will belong to this interval with 99% confidence.

Sample quality control

There are many factors that can affect the coverage of a given sample apart from its copy number and GC content. Cell immortalization, library preparation, sequencing biases and divergence to the reference, among other factors, can alter copy-number inferences from the depth of coverage (Abel and Duncavage, 2013; Tattini et al., 2015). To avoid confounding factors in my copy-number estimations as much as possible, I performed a very strict sample quality control which is described below.

I excluded samples according to several criteria. First, I filtered out all the samples that showed a standard deviation of the coverage higher than 0.5. Second, I performed a Kolmogorov-Smirnov test (Massey Jr., 1951) for normality of the distribution of coverage and excluded all the samples with a ks statistic lower than 0.03. Third, I calculated the Pearson correlation coefficient (Pearson, 1895) between neighboring windows and excluded the samples that showed a correlation greater than 0.15. Fourth, I excluded the samples that separated from all the others in a PCA according to visual inspection. Fifth, I

excluded those samples having an unusually high contribution to the minor allele compared to the others.

From the original set of 315 rhesus macaque samples, I filtered out 117 samples (including those used for the HMM training). Finally, my working set consisted in 198 rhesus macaque samples all with high-quality and fine-scale genome-wide maps of copy-number intervals. From the initial set of 50 human samples I filtered out 18 samples (including HMM training samples) and end up with a final set of 32 human samples.

Allele calling

As a result of the copy-number calling, for each window in the genome and each individual, I had a copy-number interval containing (with 99% of probability) the actual integer copy-number that this sample had in this window. Or, in other words, for each window in the genome I had a collection of copy-number intervals expressing with 99% of confidence the number of copy of all individuals in my sample in the window. I identified *copy-number alleles* in each of these windows through an allele calling algorithm. I here distinguished between *individual copy-number intervals*, being each one of the copy-number intervals that contain the copy-number of each sample in each window and *allele copy-number intervals*, being the copy-number interval representing a group of samples in each window belonging to a copy-number allele. The algorithm was applied individually to all windows as follows:

1. It identified the most common integer copy-number among all individual copy-number intervals of the window.
2. All the samples having this most common copy number(s) in their individual copy-number interval were classified as belonging to the *major copy-number allele*.
3. An allele copy-number interval was assigned to the major copy-number allele. This interval contained all integer copy numbers shared by 90% or more of the samples already classified as belonging to the major allele.

4. All the samples having in their individual copy-number interval any of the integrer copy-numbers of the major copy-number allele interval were classified as belonging to the major copy-number allele.

5. Once the major allele was identified, the same procedure (1 to 4) was repeated iteratively with the remaining samples (if any) to identify the second, third, and successively most frequent alleles until no samples were left.

The aim of this algorithm is to retrieve the different alleles present in each sample. The windows that showed more than one allele were classified as being CNV. They were classified as *CNV gains* when containing copy-number alleles with a copy number higher than diploid, as *CNV losses* when there was loss of copy number (allele copy-number with less than the diploid copy number) or as *CNV gain and losses* when there were alleles with copy numbers both, higher and lower than the diploid copy number.

This allele-calling algorithm groups most of the samples into the major allele, and was consequently very conservative in calling more than one allele. Only samples with a 99% copy-number interval clearly differentiable from the main allele were classified as having the minor allele (Supplementary Figure 7.4 in Section 7.4). This allele-calling algorithm groups most of the samples into the major allele. This procedure is very conservative in calling more than one allele because it requires that the individual copy-number intervals of each allele do not overlap. Therefore, for the windows for which I identified a single allele, I decided to, first, classify as *diploid* all the windows whose copy-number allele interval contained the copy-number 2 despite being very noisy, second, classify as *fixed duplications* all the windows in which none of the individual sample intervals spanned more than three (only one allele with very narrow individual copy-number intervals, maximum of 3 numbers) and, third, classify as *unclassified regions* all the other noisy windows for which I could not confidently distinguish more than one allele but where the intervals were not narrow.

After classifying each non-diploid window in a given category, I grouped consecutive windows of the same category into longer non-diploid regions of a given category. Of note, these non-diploid regions might be the result of more

than one duplication or deletion event as consecutive windows in the same category could have a different duplication/deletion history. Moreover, my resolution was limited by the 1kb of non-consecutive windows. For this reason, with this approach I was blind to differences in copy number within these windows. In practice, these differences appear as noise of the copy-number estimates.

Comparison to human copy-number maps

Copy-number calls in the human genome were compared to the annotation of SDs in the hg19 human reference genome (UCSC Table Browser; Karolchik, 2004 and the CNV calls in humans from Sudmant et al. (2015a).

I merged overlapping annotations of SDs in order to avoid redundant counts. On the one hand, 75.52% of the SD regions in the human genome overlapped with at least one of my non-diploid regions. My copy-number calls were more stringent than the SD calls regarding the identity between copies needed to be detected. For this reason, I did not retrieve part (24.48%) of the SD regions. On the other hand, 57.03% of my non-diploid windows overlapped with an SD region. The non-overlapping percentage might do so because of two reasons. First, SDs are genome-assembly dependent. If a given duplication (CNV or not) is not present (or resolved) in the assembly, it will not be annotated as an SD. Second, SDs are defined as regions that share 90% or more identity with another region of the assembly for 1 kbp or more length. With the read-depth based copy-number calls I was sensitive to smaller identity tracks.

The validation with CNV calls in Sudmant et al. (2015a) is not ideal. First, in Sudmant et al. (2015a), CNV calls were performed with the 1,000 Genomes Project phase 3 data WGS data including 2,504 individuals from 26 populations. This means a sample size considerably bigger than the 32 human samples that I used to perform human maps. Second, in Sudmant et al. (2015a), CNV calls were performed per population in order to see the within-population copy-number variability. For the purpose of comparing rhesus macaque copy-number maps with copy-number variability present in human populations, I designed my human sample set to detect the major inter-population variability. Despite that, I detected as non-diploid 27.81% of their CNV calls (41.44% or

their >5,000 bp CNV calls) and to see variation in 5.39% of their CNV calls (9.45% of their >5,000 bp CNV calls). A 36.06% of my inter-population CNV calls that overlaps with intra-population CNV calls from Sudmant et al. (2015a).

Duplicated genes by age and SDs

I compared my estimated chromosomal content of all types of non-diploid windows with independent proxies: duplicated genes and the SDs annotated in the rhesus macaque genome. I used data of precise phylostratification of rhesus macaque duplicated genes (Juan et al., 2013) to classify duplicated genes by three different age categories: first, macaque specific genes only present in the macaque species; second, other primate genes present in other primates but not only in macaques; and third, other mammalian duplicated genes found also duplicated in rhesus macaque that appeared before the primate diversification within the mammal class. For SDs annotations I used rheMac2 annotations from the UCSC table browser.

Definition of genes implicated in fixed duplications, CNV regions or unclassified regions

I used Ensembl annotations of rhesus macaque (rheMac8) protein-coding genes and a list of reliable protein-coding genes in the human genome from (Abascal et al., 2018). For those genes with more than one protein-coding transcript annotated in Ensembl, I selected only the exons of the transcripts used in the Ensembl Compara multi-species database (considered the main isoform). Genes with more than one annotated protein-coding transcript and no annotation in Ensembl Compara multi-species were not considered. According to this criteria I considered 20,946 rhesus macaque protein coding genes and 17,000 human protein coding genes. The difference in number of considered protein-coding genes reflects the differences in the quality of the annotations of these two species reference genomes (Abascal et al., 2018).

I divided the considered genes in several categories according to their copy-number profile to identify the ones with copy-number differences between human and rhesus macaque. First, I considered the genes that had part of their

protein-coding region (at least 1 bp) overlapping with a non-diploid window. According to this I considered genes that were overlapping with *fixed duplications*, with *CNV regions* and with *unclassified regions*. Among these genes, I considered those that had 90% or more of their protein-coding region within fixed duplications or CNV regions as being *within* these regions. Among the genes within fixed duplications or CNV regions, I searched for *genes duplicated and fixed as a whole* and *genes CNV as a whole*. To be considered duplicated and fixed as a whole or CNV as whole, at least half of the samples had to have a constant pattern of copy-number along the gene length. This is, in half of the samples, the copy-number intervals in all the windows that cover the gene had to contain the same copy-number (at least one of the integer copy numbers in the intervals has to be common for all windows along the gene). I then ran the allele calling algorithm for the whole gene in order to identify the possible gene alleles. Genes with more than one gene allele were considered CNV as a whole and genes with only one gene allele were considered duplicated and fixed as a whole.

Comparison of rhesus macaque and human gene copy-number profiles

I extracted the list of rhesus-human orthologous genes from Ensembl (rheMac8 - hg38). I considered those orthologous pairs that were within both the list of considered rhesus macaque protein-coding genes and the list of considered human ones.

I crossed the gene calls for rhesus macaque with the gene calls for their orthologs in humans. Moreover, I took human copy-number information from Sudmant et al. (2015a) and the annotation of human SDs in hg19 (UCSC table browser; Karolchik, 2004).

Disease association exploration

The GWAS Catalog (downloaded 26/09/18; MacArthur et al., 1895) was parsed in order to retrieve genes of interest with variants associated to a complex disease or phenotype. I recovered the associations considering both *Reported genes* (gene names reported by the paper) and the *Mapped genes* (mapped gene

given the position of the SNV).

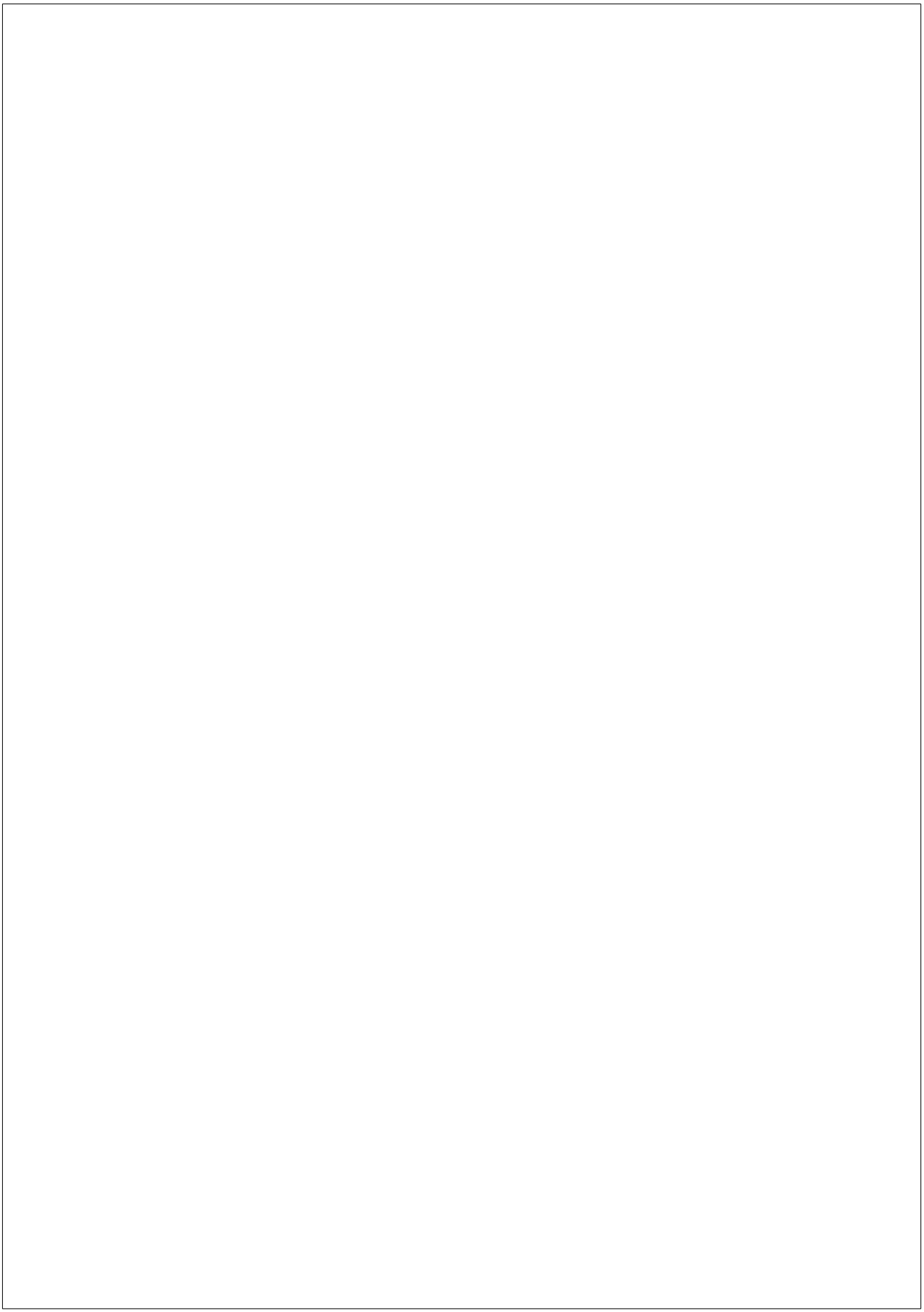
In order to obtain even more possible associations between the genes and diseases I used DisGeNET tool (Bauer-Mehren et al., 2010) that compiles information from some public databases and uses the power of text mining to find links between the gene and a diseases in the literature.

In this chapter I present a genome-wide map of duplications and CNVs in the rhesus macaque genome. Moreover, in it, I explore the potential functional consequences of duplications and CNVs in the rhesus macaque species and compare them to those in humans and identify functionally relevant regions with different copy-number profiles in the two species. These regions have be taken with care when using macaque as a model organism for biomedical research.

I did this work in collaboration with the group led by Jeffrey Rogers from the Baylor College of Medicine in Houston, Texas, USA. I performed the whole project under the supervision of David Juan, Tomàs Marquès-Bonet and Jeffrey Rogers. People from Jeffrey Roger's group contributed with the collection and sequencing of the samples and Xavier Farré conducted the gene-disease association analyses.

Chapter 6

Discussion



Science is voiceless; it is the scientists who talk.

Simone Weil

Reserve your right to think, for even to think wrongly is better than not to think at all.

Hypatia of Alexandria

Theory without practice is just as incomplete as practice without theory. The two have to go together.

Assata Shakur

The results presented in Chapters 3, 4 and 5 altogether represent three different but complementary points of view of the study of duplications. With those in mind, I have articulated this discussion around four transversal aspects of the evolution of duplications which I find relevant to remark and discuss: first, a discussion on how IGC, divergence between duplicates and natural selection interplay and determine their fate; second, a statement on the most relevant aspects of duplication evolution to be considered in genomic studies; third, perspectives on the mechanisms in which duplications arise; fourth, some considerations on the application and limitations of the duplication detection method used in this thesis; and finally, some observations regarding the possibilities to determine the function of duplications.

6.1 Interplay of IGC, sequence similarity and natural selection

The incorporation of IGC dependence on sequence similarity in our model has provided new insight into the way concerted evolution actually occurs. Through IGC, concerted evolution can potentially reach equilibrium around a specific value of divergence between duplicates if the IGC rate is high enough to counteract the point mutations accumulating independently on each duplicate

(Figure 6.1). However, some cases of concerted evolution never attain equilibrium. In these cases, divergence increases inexorably, albeit at a slower rate due to sporadic IGC events that become less and less frequent, until a certain threshold is surpassed and then concerted evolution stops (Figure 6.1). The local character of IGC dependence on sequence similarity allows for this progressive loss of concerted evolution by conserving IGC activity in some regions of the duplicates' sequence while being lost in other regions. So, even though concerted evolution can be temporary and may not reach equilibrium, it can nevertheless determine the duplicates' evolutionary path and fate.

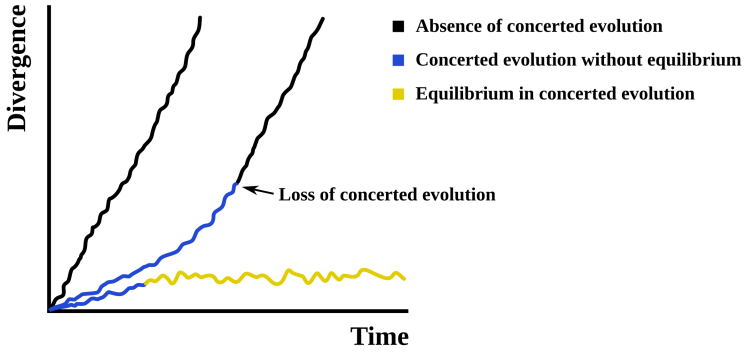


Figure 6.1: Concerted evolution through time. Graphic representation of the trajectories of divergence between duplicates along time in three different scenarios: absence of concerted evolution (black line), concerted evolution that never reaches equilibrium and is finally lost (blue to black line), and concerted evolution reaching equilibrium around a particular divergence value (blue to yellow line).

Natural selection can affect local sequence similarity between duplicates and, thus, interplay with IGC. For instance, if purifying selection acts independently in a specific paralogous fragment of both duplicates at the same time (*e.g.* an exon) non-synonymous mutations in such fragment will be disfavored. In this context, purifying selection will maintain local sequence similarity between duplicates (in non-synonymous sites), which will in turn allow the local action of IGC. Purifying selection and IGC will therefore act simultaneously and will maintain high sequence similarity in fragments which will tend to become islands of identity between duplicates. A completely different outcome will be expected if, for example, a newly appeared mutation in one of the duplicates is

favored by selection if it is present in one but not in both duplicates (*e.g.* a neofunctionalizing mutation). In this case, IGC between duplicates, in particular those events prone to erase the favored mutation reverting it to the original allele, will be disfavored. This will decrease the effective IGC rate in the region surrounding the focal mutation and creating a desert of identity around it.

Following this premise, one might want to search for islands or deserts of identity between duplicates to find cases of selection. Nevertheless, my results show that sharp differences in degrees of identity along duplicate sequences can arise also under neutrality due to the positive feedback loop between IGC and sequence similarity. For this reason, although selection can favor the persistence of islands or deserts of identity between duplicates through time, their presence cannot be taken as an unequivocal sign of selection.

6.2 Studying duplications: what to take into account

IGC has considerable effects on the evolution of duplicates. Multiple cases of concerted evolution through IGC have been observed (Chen et al., 2007, 2010b) and relevant effects of IGC have been measured in the human genome (Assis and Kondrashov, 2012; Dumont and Eichler, 2013; Dumont, 2015; Harpak et al., 2017). However, IGC and its effects are not always considered in genomic studies, including studies concerning SDs and CNVs in which IGC is prone to be more important. In order to have a complete picture of the diversity and extent of IGC effects, I will list the most relevant aspects of IGC and its interplay with other phenomena that affect sequence properties and may confound genomic studies' results:

- IGC decreases divergence between duplicates and, thus, duplication age can be underestimated when applying the molecular clock.
- IGC can increase diversity within duplicated regions. Consequently, one should consider this when measuring evolutionary rate and other parameters related to diversity in duplications (see Sections 3.3.1 and 3.3.4).

- IGC breaks short-distance LD and might create long-distance LD between duplicates. This long-distance LD has been interpreted as a product of epistatic selection or population admixture (Koch et al., 2013). However, IGC also has to be considered when interpreting long-distance LD involving paralogs (see Section 3.3.1).
- Crossover between paralogous sites boosts the effect of IGC acting between them. For example, the presence of a recombination hotspot in a duplicated region can indirectly alter sequence properties through IGC. The consequences of IGC will be different on either side of the crossover hotspot and will create patterns that would not be expected in the absence of IGC (see Section 3.3.2).
- Under neutrality, IGC and its feedback with sequence similarity can create big differences in sequence properties (*e.g.* divergence between duplicates, diversity or LD) along the length of duplicates. These large differences often take the form of islands or deserts of identity and can mimic natural selection (see Section 6.1).

Besides the aforementioned effects of IGC, other important aspects regarding the sequencing and study of duplications are important to take into account in genomic studies, especially in studies concerning duplications:

- Collapsed (non-resolved) duplications exist in the genome assemblies of many species. Even in high quality genomes, very recent duplications, especially CNV duplications, can be collapsed and/or not be annotated.
- Collapsed duplications can alter neutrality tests. To avoid drawing wrong conclusions, summary statistics should be combined in selection scans (see Section 3.3.4).
- In those duplications resolved in genome assemblies, reads from one duplicate can map on to the other duplicate and, thus, SNVs that are exclusive to one duplicate can be mistakenly called into the other duplicate. A very commonly used solution to this problem is to discard all reads that map to more than one region of the genome. This is only a partial solution for highly similar duplications because a large part of the

reads coming from either copy will map to both regions and will be discarded. Moreover, kept reads will be biased towards the most divergent parts of the duplicates and towards the most divergent alleles (*i.e.* alleles implying more identity between duplicates are more frequently discarded than the more divergent ones).

6.3 Perspectives on human duplication types and dynamics

Duplications are not an homogeneous group of regions. Their diversity is mostly due to the different existing mechanisms that give rise to different types of duplications and to the recursive nature of the latter. During this thesis I have developed some hypotheses on how different types of duplications arise and evolve.

Duplications can arise either isolated (*i.e.* in a genomic context without other duplications), or within a duplication mosaic. These are hard to distinguish from each other because of duplication shadowing. Since duplications within mosaics have common dynamics, they should not be studied individually.

On the other hand, non-mosaic duplications are simpler. They are frequently single duplication units and their peculiarities might help in determining how they arose. In humans there are three very distinct types of non-mosaic duplications:

- Tandem duplications: these are typically short, their copies are highly similar, they have a high GC-content and are frequently intronic. Contrary to expectation (Bailey et al., 2003), their borders are depleted in Alu elements meaning that Alu-mediated duplication by NAHR will not be frequently involved in their formation. In view of the latter, and given their typical small length, replication slippage is likely to be the most common mechanism leading to tandem duplications (see Section 1.3.2). These types of human duplications appear not to have had an increase in activity during the general SD burst in the great ape lineage (Marques-Bonet et al., 2009a) but did have a burst of activity later, after

the Homo and Pan lineages split from Gorilla. Moreover, their high between-copy similarity and high GC-content might be product of IGC (see Chapter 4).

- Non-tandem intrachromosomal duplications: these, in contrast to tandem duplications, are typically longer than the other non-mosaic duplications and frequently show the presence of Alu elements in their borders suggesting a clear contribution of these elements in their formation. How an Alu-mediated duplication can lead to intrachromosomal duplications with their characteristic long length is not clear. In Section 4.3.4, I exposed two different models of Alu-mediated duplication that led to long distance duplications through the insertion of free DNA molecules (inspired by Bailey et al., 2003). In addition to these models, another Alu-driven duplication mechanism, for example, Alu-driven FoSTeS, could also lead to non-tandem intrachromosomal duplications (see Section 1.3.2).

- Interchromosomal duplications: these are in general short and frequently within genes (often mono-exonic). According to these characteristics, the duplication by retrotranscription might be a common duplication mechanism leading to interchromosomal duplications. This hypothesis is reinforced by the presence of Alu elements in intrachromosomal duplication borders that can be the byproduct of Alu-driven retrotranscribed RNAs insertions through NAHR.

Both non-tandem intrachromosomal and interchromosomal duplications are enriched around centromeres. This clustering has already been related to the abundant presence of pericentromeric Alu elements (Bailey and Eichler, 2006) and strengthens the hypothesis of Alu-mediated duplication mechanisms for these two duplication types. However, the large difference in length between both suggests that they have been generated by different mechanisms. On the one hand, interchromosomal duplications show clear signals of being the product of Alu-mediated insertions of retrotranscribed RNAs. On the other hand, intrachromosomal duplications could be the product of Alu-mediated FoSTeS or Alu-mediated insertions of free DNA molecules without an RNA intermediate.

Apart from the exact mechanism giving rise to an individual duplication, the recursive nature of most mechanisms determines the architecture of duplicated regions in the human genome and their frequent mosaic structure. It is known that the presence of a duplication is in itself a predisposition for the appearance of new duplications, which leads to shadowing and mosaics (Cheng et al., 2005; Jiang et al., 2007). In the human genome, there is a considerable amount of exclusively intrachromosomal and exclusively interchromosomal mosaics that share many features with their isolated respective duplications (see Chapter 4 for specific numbers and results). This observation suggests that duplications of a certain type predispose to future duplications, frequently of the same type and in the same region. In other words, a given duplication is more prone to overlap with a preexisting duplication of its same type than with a duplication of a different type, giving rise, in time, to numerous exclusive duplication mosaics. Eventually however, a duplication of a different type might appear on top of a preexisting exclusive mosaic and that will lead to a mixed mosaic region (*i.e.* an intrachromosomal duplication on an exclusively interchromosomal mosaic or vice versa). In accordance with this hypothesis, mixed mosaic regions are frequent in the human genome and usually involve the larger and most intricate mosaic patterns.

Duplicated regions in the human genome are exceptional due to the burst of duplication activity in the great ape lineage leading to humans, although the causes behind this burst of SD formation are still unknown (Marques-Bonet et al., 2009a; Sudmant et al., 2013). In Section 4.3.3, I provided some insight into the type of SDs that participated in the great ape lineage SD expansion and, thus, into the possible duplication mechanisms that had a temporary increased activity. The formation of novel duplicated regions was driven by non-tandem intrachromosomal and interchromosomal non-mosaic SDs but most of the newly generated duplicated length (bp of new duplicated sequence) was in mosaic duplications. This means that the burst of SDs mainly created long mosaic duplicated regions, especially intra- and interchromosomal mixed mosaics.

6.4 WSSD and WGAC

WGAC and WSSD are both very useful tools for the study of duplications, but should be used in consonance with their limitations (see Section 1.2.2). Distinguishing between mosaic regions and non-mosaic duplications can be easily done with WGAC. Instead, with WSSD, the clustered architecture of some duplicated regions is not evident. In WSSD data, a complex mosaic region is seen as a long chain of contiguous non-diploid regions (*i.e.* windows) normally with variable number of copies. This pattern is not easily distinguishable from other types of genomic regions such as non-mosaic duplications or regions with high read-depth noise.

When using WSSD, it is important to keep in mind the frequent mosaic architecture of SD regions in mammalian genomes and the consequences of WSSD having a limited resolution (*i.e.* the used window size). In the same way, when using WGAC, one should always consider the genome assembly quality[1] and the possibility of collapsed duplications and should be prudent in drawing conclusions for the whole population, especially for specific duplication cases.

Chapters 4 and 5 are two good examples of the distinct potentiality of WGAC and WSSD. In Chapter 4, I combined WGAC and WSSD to construct a complete picture of SDRs in the human genome. I used WGAC to obtain high resolution information on human duplications and their distinct types and at the same time, I used WSSD to compare different great ape species' genomes and phylostratificate human duplications. In Chapter 5, I used WSSD to retrieve information of genome-wide copy number for a big sample of rhesus macaque genomes, despite the low quality of this species' genome assembly. Therein, I focused on the distribution of non-duplicated regions and their fixed or variable nature without dwelling on the details of specific types of duplications.

Having WGAC of high quality assemblies for a large sample of individuals would provide a complete and accurate picture of the duplication content and its architecture in a population. This scenario is not a reality nowadays but new long-read and nanopore-based sequencing technologies and the resultant high

[1] Although WSSD is more robust than WGAC for low quality assemblies, its results can also be affected by the genome assembly quality (see Chapter 5). This should also be considered when working with WSSD.

quality genome assemblies are really promising in this regard (Nagarajan and Pop, 2013; Simpson and Pop, 2015; Lu et al., 2016; Wajid et al., 2016). Meanwhile, to combine WSSD and WGAC is definitely the best option to have a complete view of duplications within genomes.

6.5 Difficulties and efforts to determine the function of duplications

The role of duplications in the generation of new functions in genome evolution is widely accepted. There are multiple described examples of functional innovation in pairs of paralogous genes (Kondrashov and Koonin, 2004; Innan, 2009; Innan and Kondrashov, 2010). However, to identify duplications that could be candidates for having functional implications (*i.e.* for harboring new functions) in genome-wide scans is a big challenge. Even more challenging is to link the presence of a duplication with an specific functional change (*i.e.* to link genotype with phenotype).

Duplications that do not involve genes or other functional units are mostly neutral and the cases producing functional changes are very hard to identify. Duplications involving parts of genes (not entire genes) are more likely to have functional effects[2] but these effects are in general hard to predict unless focalizing in particular cases. Differently, duplications involving entire genes, especially if they also include the corresponding regulatory regions, are strong candidates for causing functional changes and for harboring functional innovations. For these reasons, in both Chapter 4 and Chapter 5 I focused on the functional consequences of entire gene duplications.

Entire gene duplications are not always easy to identify. WGAC data consists in a list of pairs of paralogous regions with high resolution of duplication boundaries. As such, it is clear when a duplication involves an entire gene (see Chapter 4). Instead, with WSSD, because of the lack of information of specific duplication events, it is hard to assess if a gene has been duplicated as a whole or if it has been involved in multiple partial duplication events. In Chapter 5, I presented a method

[2]Partial gene deletions, unlike partial gene duplications, are extremely likely to have functional effects because they normally result in abnormal forms of the corresponding gene product.

to overcome this problem and found genes duplicated as a whole (either fixed or CNV) with WSSD. This method can be easily applied to other WSSD data to find duplicated genes being strong candidates to have functional implications.

6.6 Concluding remarks

Duplication is at the base of the generation of new genetic material and new functions. Given its role in evolution, understanding how duplications appear, evolve and give rise to new functions is fundamental to comprehend the transition from simple to complex organisms and genomes. In this thesis I have presented, in three chapters, the result of my work aimed at understanding such processes. Overall, this thesis brings together many aspects of the study of duplications, starting from a theoretical standpoint and the use of simulations to explore their evolution, going through an effort to characterize the variety of duplication types and mechanistic origins, and ending with two different case studies, human and rhesus macaque, which implied the use of two distinct duplication detection methods.

I now summarize the main contributions of this thesis:

1. I have studied how distinct realistic crossover distributions in duplicated regions interplay with IGC effects.
2. I have modeled and simulated IGC dependence on sequence similarity between duplicates and its consequences on the concerted evolution of duplicated regions of the genome.
3. I have studied the confounding effects that duplications performing IGC and their collapse in genome assemblies can have in genome-wide scans for selection.
4. I have contributed to the understanding of the diversity of SDs in the human genome.
5. I have generated novel hypotheses regarding how different duplications arise and lead to their frequent clusterized conformation.

6. I have provided insight into the time at which and the mechanisms through which human SDs arose and evolved.

7. I have described and characterized SDs and CNVs in the rhesus macaque genome in a big sample of individuals.

8. I have identified rhesus macaque genes with relevant differences in copy-number relative to the human genome, representing candidates for having implications in human disease.

Bibliography

Abascal, F., Juan, D., Jungreis, I., Martinez, L., Rigau, M., Rodriguez, J., Vazquez, J., and Tress, M. L. (2018). Loose ends: almost one in five human genes still have unresolved coding status. *Nucleic Acids Research*, 46(14):7070–7084.

Abel, H. J. and Duncavage, E. (2013). Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. *Cancer Genet*, 206(12):432–440.

Ahn, B. Y., Dornfeld, K. J., Fagrelius, T. J., and Livingston, D. M. (1988). Effect of limited homology on gene conversion in a Saccharomyces cerevisiae plasmid recombination system. *Molecular and Cellular Biology*, 8(6):2442–2448.

Alkan, C., Kidd, J. M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., Kitzman, J. O., Baker, C., Malig, M., Mutlu, O., Sahinalp, S. C., Gibbs, R. A., and Eichler, E. E. (2009). Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genetics*, 41(10):1061–1067.

Almal, S. H. and Padh, H. (2012). Implications of gene copy-number variation in health and diseases. *Journal of Human Genetics*, 57(1):6–13.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Anney, R. et al. (2010). A genome-wide scan for common alleles affecting risk for autism. *Human Molecular Genetics*, 19(20):4072–4082.

Ardlie, K., Liu-Cordero, S. N., Eberle, M. A., Daly, M., Barrett, J., Winchester, E., Lander, E. S., and Kruglyak, L. (2001). Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. *American Journal of Human Genetics*, 69(3):582–589.

Assis, R. and Bachtrog, D. (2013). Neofunctionalization of young duplicate genes in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, 110(43):17409–17414.

Assis, R. and Kondrashov, A. S. (2012). A strong deletion bias in nonallelic gene conversion. *PLoS Genetics*, 8(2):1–5.

Auton, A. et al. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

Babushok, D. V., Ostertag, E. M., and Kazazian, H. H. (2007). Current topics in genome evolution: Molecular mechanisms of new gene formation. *Cellular and Molecular Life Sciences*, 64(5):542–554.

Bailey, J. A., Baertsch, R., Kent, W. J., Haussler, D., and Eichler, E. E. (2004a). Hotspots of mammalian chromosomal evolution. *Genome Biology*, 5(4):R23.

Bailey, J. A., Church, D. M., Ventura, M., Rocchi, M., and Eichler, E. E. (2004b). Analysis of Segmental Duplications and Genome Assembly in the Mouse. *Genome Research*, 14(5):789–801.

Bailey, J. A. and Eichler, E. E. (2006). Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature Reviews Genetics*, 7(7):552–564.

Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M. D., Myers, E. W., Li, P. W., and Eichler, E. E. (2002a). Recent segmental duplications in the human genome. *Science*, 297(5583):1003–1007.

Bailey, J. A., Liu, G., and Eichler, E. E. (2003). An Alu Transposition Model for the Origin and Expansion of Human Segmental Duplications. *The American Journal of Human Genetics*, 73(4):823–834.

Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J., and Eichler, E. E. (2001). Segmental duplications: organization and impact within the current human genome project assembly. *Genome Research*, 11(6):1005–1017.

Bailey, J. A., Yavor, A. M., Viggiano, L., Misceo, D., Horvath, J. E., Archidiacono, N., Schwartz, S., Rocchi, M., and Eichler, E. E. (2002b). Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. *American Journal of Human Genetics*, 70(1):83–100.

Baker, Z., Schumer, M., Haba, Y., Bashkirova, L., Holland, C., Rosenthal, G. G., and Przeworski, M. (2017). Repeated losses of PRDM9-directed recombination despite the conservation of PRDM9 across vertebrates. *eLife*, 6:1–34.

Ballif, B. C. et al. (2010). Identification of a Recurrent Microdeletion at 17q23.1q23.2 Flanked by Segmental Duplications Associated with Heart Defects and Limb Abnormalities. *American Journal of Human Genetics*, 86(3):454–461.

Baltimore, D. (1981). Gene conversion: some implications for immunoglobulin genes. *Cell*, 24(3):592–594.

Barr, C. S., Newman, T. K., Lindell, S., Shannon, C., Champoux, M., Lesch, K. P., Suomi, S. J., Goldman, D., and Higley, J. D. (2004). Interaction between serotonin transporter gene variation and rearing condition in alcohol preference and consumption in female primates. *Archives of General Psychiatry*, 61(11):1146–1152.

Bauer-Mehren, A., Rautschka, M., Sanz, F., and Furlong, L. I. (2010). Disgenet: a cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. *Bioinformatics*, 26(22):2924–2926.

Bauman, M. D. and Schumann, C. M. (2018). Advances in nonhuman primate models of autism: Integrating neuroscience and behavior. *Experimental Neurology*, 299(Part A):252–265.

Bayes, F. R. S. l. R. M. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions*, 53:370–418.

Beckmann, J. S., Estivill, X., and Antonarakis, S. E. (2007). Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nature Reviews Genetics*, 8(8):639–646.

Bekpen, C., Künzel, S., Xie, C., Eaaswarkhanth, M., Lin, Y. L., Gokcumen, O., Akdis, C. A., and Tautz, D. (2017). Segmental duplications and evolutionary acquisition of UV damage response in the SPATA31 gene family of primates and humans. *BMC Genomics*, 18(1):1–12.

Benovoy, D. and Drouin, G. (2009). Ectopic gene conversions in the human genome. *Genomics*, 93(1):27–32.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, 27(2):573–580.

Bentley, D. R. et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59.

Betrán, E., Demuth, J. P., and Williford, A. (2012). Why Chromosome Palindromes? *International Journal of Evolutionary Biology*, 2012:207958.

Bimber, B. N., Ramakrishnan, R., Cervera-Juanes, R., Madhira, R., Peterson, S. M., Norgren, R. B., and Ferguson, B. (2017). Whole genome sequencing predicts novel human disease models in rhesus macaques. *Genomics*, 109(3-4):214–220.

Bosch, E., Hurles, M. E., Navarro, A., and Jobling, M. A. (2004). Dynamics of a human interparalog gene conversion hotspot. *Genome Research*, 14(5):835–844.

Bridges, C. B. (1936). The Bar "Gene" A Duplication. *Science*, 83(2148):210–211.

Brooks, E. M., Branda, R. F., Nicklas, J. A., and Patrick O'Neill, J. (2001). Molecular description of three macro-deletions and an Alu-Alu recombination-mediated duplication in the HPRT gene in four patients with Lesch-Nyhan disease. *Mutation Research*, 476(1-2):43–54.

Section 6.6

Brown, D. D. and Sugimoto, K. (1973). 5 S DNAs of Xenopus laevis and Xenopus mulleri: Evolution of a gene family. *Journal of Molecular Biology*, 78(3):397–415.

Brown, D. D., Wensink, P. C., and Jordan, E. (1972). A comparison of the ribosomal DNA's of Xenopus laevis and Xenopus mulleri: the evolution of tandem genes. *Journal of Molecular Biology*, 63(1):57–73.

Buard, J., Rivals, E., Dunoyer De Segonzac, D., Garres, C., Caminade, P., De Massy, B., and Boursot, P. (2014). Diversity of Prdm9 zinc finger array in wild mice unravels new facets of the evolutionary turnover of this coding minisatellite. *PLoS ONE*, 9(1).

Campbell, P. J. et al. (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genetics*, 40(6):722–729.

Cañestro, C. (2012). Two Rounds of Whole-Genome Duplication: Evidence and Impact on the Evolution of Vertebrate Innovations. In Soltis, P. S. and Soltis, D. E., editors, *Polyploidy and Genome Evolution*, pages 309–339. Springer-Verlag, Berlin Heidelberg.

Cañestro, C., Albalat, R., Irimia, M., and Garcia-Fernàndez, J. (2013). Impact of gene gains, losses and duplication modes on the origin and diversification of vertebrates. *Seminars in Cell and Developmental Biology*, 24(2):83–94.

Cann, H. M., de Toma, C., Cazes, L., Legrand, M. F., and Morel, V. (2002). A human genome diversity cell line panel. *Science*, 296(5566):261–262.

Carvalho, C. M. and Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews Genetics*, 17(4):224–238.

Casola, C., Conant, G. C., and Hahn, M. W. (2012a). Very low rate of gene conversion in the yeast genome. *Molecular Biology and Evolution*, 29(12):3817–3826.

Casola, C., Ganote, C. L., and Hahn, M. W. (2010). Nonallelic gene conversion in the genus Drosophila. *Genetics*, 185(1):95–103.

Casola, C., Zekonyte, U., Phillips, A. D., Cooper, D. N., and Hahn, M. W. (2012b). Interlocus gene conversion events introduce deleterious mutations into at least 1% of human genes associated with inherited disease. *Genome Research*, 22(3):429–435.

Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., Landolin, J. M., Stamatoyannopoulos, J. A., Hunkapiller, M. W., Kogornrlach, J., and Eichler, E. E. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, 517(7536):608–611.

Champoux, M., Bennett, A., Shannon, C., Higley, J. D., Lesch, K. P., and Suomi, S. J. (2002). Serotonin transporter gene polymorphism, differential early rearing, and behavior in rhesus monkey neonates. *Molecular Psychiatry*, 7(10):1058–1063.

Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C., and Patrinos, G. P. (2007). Gene conversion: mechanisms, evolution and human disease. *Nature Reviews Genetics*, 8(10):762–775.

Chen, J. M., Cooper, D. N., Férec, C., Kehrer-Sawatzki, H., and Patrinos, G. P. (2010a). Genomic rearrangements in inherited disease and cancer. *Seminars in Cancer Biology*, 20(4):222–233.

Chen, J.-M., Férec, C., and Cooper, D. N. (2010b). Gene conversion in human genetic disease. *Genes*, 1(3):550–563.

Chen, L., Zhou, W., Zhang, L., and Zhang, F. (2014). Genome Architecture and Its Roles in Human Copy Number Variation. *Genomics & Informatics*, 12(4):136–144.

Cheng, Z., Ventura, M., She, X., Khaitovich, P., Graves, T., Osoegawa, K., Church, D., DeJong, P., Wilson, R. K., Pääbo, S., Rocchi, M., and Eichler, E. E. (2005). A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature*, 437(7055):88–93.

Chiang, D. Y., Getz, G., Jaffe, D. B., O'Kelly, M. J. T., Zhao, X., Carter, S. L., Russ, C., Nusbaum, C., Meyerson, M., and Lander, E. S. (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature Methods*, 6(1):99–103.

Chong, H., Xue, J., Zhu, Y., Cong, Z., Chen, T., Guo, Y., Wei, Q., Zhou, Y., Qin, C., and He, Y. (2018). Design of Novel HIV-1/2 Fusion Inhibitors with Efficient Therapeutic Efficacy in Rhesus Monkey Models. *Journal of virology*, 92(16):e00775–18.

Conrad, D. F. et al. (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–712.

Cooper, D. N., Bacolla, A., Férec, C., Vasquez, K. M., Kehrer-Sawatzki, H., and Chen, J. M. (2011). On the sequence-directed nature of human gene mutation: The role of genomic architecture and the local DNA sequence environment in mediating gene mutations underlying human inherited disease. *Human Mutation*, 32(10):1075–1099.

Cordaux, R. and Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, 10(10):691–703.

Darlow, J. M. and Stott, D. I. (2006). Gene conversion in human rearranged immunoglobulin genes. *Immunogenetics*, 58(7):511–522.

Darwin, C. (1859). *On the Origin of Species*. D. Appleton and Company, New York.

Davies, B., Hatton, E., Altemose, N., Hussin, J. G., Pratto, F., Zhang, G., Hinch, A. G., Moralli, D., Biggs, D., Diaz, R., Preece, C., Li, R., Bitoun, E., Brick, K., Green, C. M., Camerini-Otero, R. D., Myers, S. R., and Donnelly, P. (2016). Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. *Nature*, 530(7589):171–176.

Dayhoff, M. O., Barker, W. C., and Hunt, L. T. (1983). Establishing Homologies in Protein Sequences. *Methods in Enzymology*, 91:524–545.

de Groot, N. G., Blokhuis, J. H., Otting, N., Doxiadis, G. G., and Bontrop, R. E. (2015). Co-evolution of the MHC class I and KIR gene families in rhesus macaques: Ancestry and plasticity. *Immunological Reviews*, 267(1):228–245.

de Manuel, M. et al. (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science*, 354(6311):477–481.

Degenhardt, J. D., De Candia, P., Chabot, A., Schwartz, S., Henderson, L., Ling, B., Hunter, M., Jiang, Z., Palermo, R. E., Katze, M., Eichler, E. E., Ventura, M., Rogers, J., Marx, P., Gilad, Y., and Bustamante, C. D. (2009). Copy number variation of CCL3-like genes affects rate of progression to simian-AIDS in rhesus macaques (Macaca mulatta). *PLoS Genetics*, 5(1).

Dennis, M. Y. et al. (2017). The evolution and population diversity of human-specific segmental duplications. *Nature Ecology and Evolution*, 1(3):69.

Des Marais, D. L. and Rausher, M. D. (2008). Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature*, 454(7205):762–765.

Dumont, B. L. (2015). Interlocus gene conversion explains at least 2.7 % of single nucleotide variants in human segmental duplications. *BMC Genomics*, 16(1):1–11.

Dumont, B. L. and Eichler, E. E. (2013). Signals of Historical Interlocus Gene Conversion in Human Segmental Duplications. *PLoS ONE*, 8(10):e75949.

Duret, L. and Galtier, N. (2009). Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. *Annual Review of Genomics and Human Genetics*, 10:285–311.

Dwivedi, G. and Haber, J. E. (2018). *Assaying Mutations Associated With Gene Conversion Repair of a Double-Strand Break*, volume 601. Elsevier Inc., 1 edition.

Ebert, G., Steininger, A., Weißmann, R., Boldt, V., Lind-Thomsen, A., Grune, J., Badelt, S., Heßler, M., Peiser, M., Hitzler, M., Jensen, L. R., Müller, I., Hu, H., Arndt, P. F., Kuss, A. W., Tebel, K., and Ullmann, R. (2014). Distribution of segmental duplications in the context of higher order chromatin organisation of human chromosome 7. *BMC genomics*, 15:537.

Eichler, E. (2001). Recent Duplication and the Dynamic Mutation of the Human Genome. *Trends in Genetics*, 17(11):661–669.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., and Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, 11(6):446–450.

Ellison, C. E. and Bachtrog, D. (2015). Non-allelic gene conversion enables rapid evolutionary change at multiple regulatory sites encoded by transposable elements. *eLife*, 2015(4):3–5.

Ewing, G. and Hermisson, J. (2010). MSMS: A coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, 26(16):2064–2065.

Fay, J. C. and Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, 155(3):1405–1413.

Feng, X., Jiang, J., Padhi, A., Ning, C., Fu, J., Wang, A., Mrode, R., and Liu, J. F. (2017). Characterization of genome-wide segmental duplications reveals a common genomic feature of association with immunity among domestic animals. *BMC Genomics*, 18(1):1–11.

Feuk, L., Carson, A. R., and Scherer, S. W. (2006a). Structural variation in the human genome. *Nature Reviews Genetics*, 7(2):85–97.

Feuk, L., Marshall, C. R., Wintle, R. F., and Scherer, S. W. (2006b). Structural variants: changing the landscape of chromosomes and design of disease studies. *Human Molecular Genetics*, 15 Spec No(1):57–66.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y.-l., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary degenerative mutations. *Genetics*, 151(4):1531–1545.

Frisse, L., Hudson, R. R., Bartoszewicz, A., Wall, J. D., Donfack, J., and Di Rienzo, A. (2001). Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *American Journal of Human Genetics*, 69(4):831–843.

Frühmesser, A., Blake, J., Haberlandt, E., Baying, B., Raeder, B., Runz, H., Spreiz, A., Fauth, C., Benes, V., Utermann, G., Zschocke, J., and Kotzot, D. (2013). Disruption of EXOC6B in a patient with developmental delay, epilepsy, and a de novo balanced t(2;8) translocation. *European Journal of Human Genetics*, 21(10):1177–1180.

Gally, J. A. and Edelman, G. M. (1970). Somatic translocation of antibody genes. *Nature*, 227(5256):341–348.

Galtier, N., Piganeau, G., Mouchiroud, D., and Duret, L. (2001). GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics*, 159(2):907–911.

Gazave, E., Darre, F., Morcillo-Suarez, C., Petit-Marty, N., Carreno, A., Marigorta, U. M., Ryder, O. A., Blancher, A., Rocchi, M., Bosch, E., Baker, C., Marques-Bonet, T., Eichler, E. E., and Navarro, A. (2011). Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. *Genome Research*, 21(10):1626–1639.

Gibbs, R. A. et al. (2007). Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, 316(5822):222–234.

Girirajan, S., Campbell, C. D., and Eichler, E. E. (2011). Human Copy Number Variation and Complex Genetic Disease. *Annual Review of Genetics*, 45(1):203–226.

Girirajan, S., Dennis, M. Y., Baker, C., Malig, M., Coe, B. P., Campbell, C. D., Mark, K., Vu, T. H., Alkan, C., Cheng, Z., Biesecker, L. G., Bernier, R., and Eichler, E. E. (2013). Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder. *American Journal of Human Genetics*, 92(2):221–237.

Glemin, S., Arndt, P. F., Messer, P. W., Petrov, D., Galtier, N., and Duret, L. (2015). Quantification of GC-biased gene conversion in the human genome. *Genome Research*, 25:1215–1228.

Gokcumen, O., Babb, P. L., Iskow, R. C., Zhu, Q., Shi, X., Mills, R. E., Ionita-Laza, I., Vallender, E. J., Clark, A. G., Johnson, W. E., and Lee, C. (2011). Refinement of primate copy number variation hotspots identifies candidate genomic regions evolving under positive selection. *Genome Biology*, 12(5):R52.

Gokcumen, O., Tischler, V., Tica, J., Zhu, Q., Iskow, R. C., Lee, E., Fritz, M. H.-Y., Langdon, A., Stütz, A. M., Pavlidis, P., Benes, V., Mills, R. E., Park, P. J., Lee, C., and Korbel, J. O. (2013). Primate genome architecture influences structural variation mechanisms and functional consequences. *Proceedings of the National Academy of Sciences of the United States of America*, 110(39):15764–15769.

Gordon, D. et al. (2016). Long-read sequence assembly of the gorilla genome. *Science*, 352(6281):aae0344.

Graur, D. and Li, W.-H. (2000). *Fundamentals of Molecular Evolution*. Sinauer Associates, Inc., Sunderland, Massachusetts.

Gschwind, A. R., Singh, A., Certa, U., Reymond, A., and Heckel, T. (2017). Diversity and regulatory impact of copy number variation in the primate Macaca fascicularis. *BMC Genomics*, 18(1):1–10.

Gu, S., Yuan, B., Campbell, I. M., Beck, C. R., Carvalho, C. M., Nagamani, S. C., Erez, A., Patel, A., Bacino, C. A., Shaw, C. A., Stankiewicz, P., Cheung, S. W., Bi, W., and Lupski, J. R. (2015). Alu-mediated diverse and complex pathogenic copy-number variants within human chromosome 17 at p13.3. *Human Molecular Genetics*, 24(14):4061–4077.

Haber, J. E. (2018). DNA Repair: The Search for Homology. *BioEssays*, 40(5):1–12.

Hafeez, M., Shabbir, M., Altaf, F., and Abbasi, A. A. (2016). Phylogenomic analysis reveals ancient segmental duplications in the human genome. *Molecular Phylogenetics and Evolution*, 94:95–100.

Hallast, P., Balaresque, P., Bowden, G. R., Ballereau, S., and Jobling, M. A. (2013). Recombination Dynamics of a Human Y-Chromosomal Palindrome: Rapid GC-Biased Gene Conversion, Multi-kilobase Conversion Tracts, and Rare Inversions. *PLoS Genetics*, 9(7):e1003666.

Han, M. V., Demuth, J. P., McGrath, C. L., Casola, C., and Hahn, M. W. (2009). Adaptive evolution of young gene duplicates in mammals. *Genome Research*, 19(5):859–867.

Hancks, D. C. and Kazazian, H. H. (2012). Active human retrotransposons: Variation and disease. *Current Opinion in Genetics and Development*, 22(3):191–203.

Harpak, A., Lan, X., Gao, Z., and Pritchard, J. K. (2017). Frequent nonallelic gene conversion on the human lineage and its effect on the divergence of gene duplicates. *Proceedings of the National Academy of Sciences of the United States of America*, 114(48):12779–12784.

Hartasánchez, D. A., Brasó-Vives, M., Fuentes-Díaz, J., Vallès-Codina, O., and Navarro, A. (2016). SeDuS: Segmental duplication simulator. *Bioinformatics*, 32(1):148–150.

Hartasánchez, D. A., Brasó-Vives, M., Heredia-Genestar, J. M., Pybus, M., and Navarro, A. (2018). Effect of collapsed duplications on diversity estimates: what to expect. *Genome Biology and Evolution*, evy223(https://doi.org/10.1093/gbe/evy223):1–20.

Hartasánchez, D. A., Vallès-Codina, O., Brasó-Vives, M., and Navarro, A. (2014). Interplay of Interlocus Gene Conversion and Crossover in Segmental Duplications Under a Neutral Scenario. *G3 (Bethesda)*, 4(8):1479–1489.

Hastings, P. J. (2010). Mechanisms of Ectopic Gene Conversion. *Genes*, 1(3):427–439.

Hastings, P. J., Ira, G., and Lupski, J. R. (2009a). A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genetics*, 5(1).

Hastings, P. J., Lupski, J. R., Rosenberg, S. M., and Ira, G. (2009b). Mechanisms of change in gene copy number. *Nature Reviews Genetics*, 10(8):551–564.

Haussler, D. et al. (2009). Genome 10K: A proposal to obtain whole-genome sequence for 10000 vertebrate species. *Journal of Heredity*, 100(6):659–674.

Hayakawa, T., Angata, T., Lewis, A. L., Mikkelsen, T. S., Varki, N. M., and Varki, A. (2005). A Human-Specific Gene in Microglia. *Science*, 309(5741):1693.

Hellmann, I., Letvin, N. L., and Schmitz, J. E. (2013). KIR2DL4 Copy Number Variation Is Associated with CD4+ T-Cell Depletion and Function of Cytokine-Producing NK Cell Subsets in SIV-Infected Mamu-A*01-Negative Rhesus Macaques. *Journal of Virology*, 87(9):5305–5310.

Hellmann, I., Lim, S. Y., Gelman, R. S., and Letvin, N. L. (2011). Association of activating KIR copy number variation of NK cells with containment of SIV replication in rhesus monkeys. *PLoS Pathogens*, 7(12):1–12.

Hellmann, I., Schmitz, J. E., and Letvin, N. L. (2012). Activating KIR Copy Number Variation Is Associated with Granzyme B Release by NK Cells during Primary Simian Immunodeficiency Virus Infection in Rhesus Monkeys. *Journal of Virology*, 86(23):13103–13107.

Hess, K., Oliverio, R., Nguyen, P., Le, D., Ellis, J., Kdeiss, B., Ord, S., Chalkia, D., and Nikolaidis, N. (2018). Concurrent action of purifying selection and gene conversion results in extreme conservation of the major stress-inducible Hsp70 genes in mammals. *Scientific Reports*, 8(1):1–16.

Higgins, A. W. et al. (2008). Characterization of Apparently Balanced Chromosomal Rearrangements from the Developmental Genome Anatomy Project. *American Journal of Human Genetics*, 82(3):712–722.

Hillier, L. W. et al. (2003). The DNA sequence of human chromosome 7. *Nature*, 424(6945):157–164.

Holliday, R. (1964). A mechanism for gene conversion in fungi. *Genetics Research Cambridge*, 89(5-6):285–307.

Hood, L., H, C. J., and Elgin, S. C. (1975). The oganization, expression, and evolution of antibody genes and other multigene families. *Annual Review of Genetics*, 9:305–353.

Hu, X. and Worton, R. G. (1992). Partial gene duplication as a cause of human disease. *Human Mutation*, 1(1):3–12.

Huddleston, J., Ranade, S., Malig, M., Antonacci, F., Chaisson, M., Hon, L., Sudmant, P. H., Graves, T. A., Alkan, C., Dennis, M. Y., Wilson, R. K., Turner, S. W., Korlach, J., and Eichler, E. E. (2014). Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Research*, 24(4):688–696.

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W., and Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature Genetics*, 36(9):949–951.

Innan, H. (2002). A method for estimating the mutation, gene conversion and recombination parameters in small multigene families. *Genetics*, 161(2):865–872.

Innan, H. (2003a). A two-locus gene conversion model with selection and its application to the human RHCE and RHD genes. *Proceedings of the National Academy of Sciences of the United States of America*, 100(15):8793–8798.

Innan, H. (2003b). The coalescent and infinite-site model of a small multigene family. *Genetics*, 163(2):803–810.

Innan, H. (2009). Population genetic models of duplicated genes. *Genetica*, 137(1):19–37.

Innan, H. and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*, 11(2):97–108.

Inoue, J., Sato, Y., Sinclair, R., Tsukamoto, K., and Nishida, M. (2015). Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling. *Proceedings of the National Academy of Sciences of the United States of America*, 112(48):14918–14923.

Itsara, A., Cooper, G. M., Baker, C., Girirajan, S., Li, J., Absher, D., Krauss, R. M., Myers, R. M., Ridker, P. M., Chasman, D. I., Mefford, H., Ying, P., Nickerson, D. A., and Eichler, E. E. (2009). Population analysis of large copy number variants and hotspots of human genetic disease. *American Journal of Human Genetics*, 84(2):148–161.

Jaillon, O. et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449(7161):463–467.

Jiang, J., Wang, J., Wang, H., Zhang, Y., Kang, H., Feng, X., Wang, J., Yin, Z., Bao, W., Zhang, Q., and Liu, J. F. (2014). Global copy number analyses by next generation sequencing provide insight into pig genome variation. *BMC Genomics*, 15(1):1–18.

Jiang, Z., Tang, H., Ventura, M., Cardone, M. F., Marques-Bonet, T., She, X., Pevzner, P. A., and Eichler, E. E. (2007). Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nature Genetics*, 39(11):1361–1368.

Jinks-Robertson, S. and Petes, T. D. (1986). Chromosomal translocations generated by high-frequency meiotic recombination between repeated yeast genes. *Genetics*, 114(3):731–752.

Juan, D., Rico, D., Marques-Bonet, T., Fernandez-Capetillo, O., and Valencia, A. (2013). Late-replicating CNVs as a source of new genes. *Biology Open*, 2(12):1402–1411.

Kaer, K. and Speek, M. (2013). Retroelements in human disease. *Gene*, 518(2):231–241.

Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. *Genome Research*, 20(10):1313–1326.

Karolchik, D. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, 32(90001):D493–D496.

Kazazian Jr., H. H. (2004). Mobile Elements : Drivers of Genome Evolution. *Science*, 303(5664):1626–1632.

Kelley, D. R. and Salzberg, S. L. (2010). Detection and correction of false segmental duplications caused by genome mis-assembly. *Genome Biology*, 11(3):R28.

Kidd, J. M., Sampas, N., Antonacci, F., Graves, T., Fulton, R., Hayden, H. S., Alkan, C., Malig, M., Ventura, M., Giannuzzi, G., Kallicki, J., Anderson, P., Tsalenko, A., Yamada, N. A., Tsang, P., Kaul, R., Wilson, R. K., Bruhn, L., and Eichler, E. E. (2010). Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nature Methods*, 7(5):365–371.

Kijima, T. E. and Innan, H. (2010). On the estimation of the insertion time of LTR retrotransposable elements. *Molecular Biology and Evolution*, 27(4):896–904.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626.

Kimura, M. (1980). Average time until fixation of a mutant allele in a finite population under continued mutation pressure: Studies by analytical, numerical, and pseudo-sampling methods. *Proceedings of the National Academy of Sciences of the United States of America*, 77(1):522–526.

Kirsch, S., Münch, C., Jiang, Z., Cheng, Z., Chen, L., Batz, C., Eichler, E. E., and Schempp, W. (2008). Evolutionary dynamics of segmental duplications from human Y-chromosomal euchromatin / heterochromatin transition regions. *Genome Research*, 18(7):1030–1042.

Kirsch, S., Weiß, B., Miner, T. L., Waterston, R. H., Clark, R. A., Eichler, E. E., Münch, C., Schempp, W., and Rappold, G. (2005). Interchromosomal segmental duplications of the pericentromeric region on the human Y chromosome. *Genome Research*, 15(2):195–204.

Koch, E., Ristroph, M., and Kirkpatrick, M. (2013). Long range linkage disequilibrium across the human genome. *PLoS One*, 8(12):e80754.

Kondrashov, F. A. and Koonin, E. V. (2004). A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends in Genetics*, 20(7):287–291.

Kronenberg, Z. N. et al. (2018). High-resolution comparative analysis of great ape genomes. *Science*, 360(6393):aar6343.

Kuderna, L. F., Lizano, E., Julia, E., Gomez-Garrido, J., Serres-Armero, A., Kuhlwilm, M., Alandes, R. A., Alvarez-Estape, M., Alioto, T., Gut, M., Gut, I., Schierup, M. H., Fornas, O., and Marques-Bonet, T. (2018). Selective single molecule sequencing and assembly of a human Y chromosome of African origin. *bioRxiv*.

Lan, X. and Pritchard, J. K. (2016). Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science*, 352(6288):1009–1013.

Lander, E. S., Linton, L. M., Birren, B. W., Nusbaum, C., and Zody, M. C. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Langer-Safer, P. R., Levine, M., and Ward, D. C. (1982). Immunological method for mapping genes on Drosophila polytene chromosomes. *Proceedings of the National Academy of Sciences of the United States of America*, 79(14):4381–4385.

Lee, A. S., Gutiérrez-Arcelus, M., Perry, G. H., Vallender, E. J., Johnson, W. E., Miller, G. M., Korbel, J. O., and Lee, C. (2008). Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. *Human Molecular Genetics*, 17(8):1127–1136.

Lewontin, R. C. (1964). The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics*, 49(1):49–67.

Lichten, M., Borts, R. H., and Haber, J. E. (1987). Meiotic gene conversion and crossing over between dispersed homologous sequences occurs frequently in Saccharomyces cerevisiae. *Genetics*, 115(2):233–246.

Lichten, M. and Haber, J. E. (1989). Position effects in ectopic and allelic mitotic recombination in Saccharomyces cerevisiae. *Genetics*, 123(2):261–268.

Lindegren, C. C. (1953). Gene conversion in Saccharomyces. *Journal of Genetics*, 51:625–637.

Liskay, R. M., Letsou, A., and Stachelek, J. L. (1987). Homology requirement for efficient gene conversion between duplicated chromosomal sequences in mammalian cells. *Genetics*, 115(1):161–167.

Liu, P., Carvalho, C. M., Hastings, P., and Lupski, J. R. (2012). Mechanisms for recurrent and complex human genomic rearrangements. *Current Opinion in Genetics & Development*, 22(3):211–220.

Liu, Z., Tan, X., Orozco-terWengel, P., Zhou, X., Zhang, L., Tian, S., Yan, Z., Xu, H., Ren, B., Zhang, P., Xiang, Z., Sun, B., Roos, C., Bruford, M. W., and Li, M. (2018). Population genomics of wild Chinese rhesus macaques reveals a dynamic demographic history and local adaptation, with implications for biomedical research. *GigaScience*, 7(9):giy106.

Locke, D. P., Sharp, A. J., Mccarroll, S. A., Mcgrath, S. D., Newman, T. L., Cheng, Z., Schwartz, S., Albertson, D. G., Pinkel, D., Altshuler, D. M., and Eichler, E. E. (2006). Linkage Disequilibrium and Heritability of Copy-Number Polymorphisms within Duplicated Regions of the Human Genome. *The American Journal of Human Genetics*, 79:275–290.

Loffredo, J. T., Maxwell, J., Qi, Y., Glidden, C. E., Borchardt, G. J., Soma, T., Bean, A. T., Beal, D. R., Wilson, N. A., Rehrauer, W. M., Lifson, J. D., Carrington, M., and Watkins, D. I. (2007). Mamu-B*08-Positive Macaques Control Simian Immunodeficiency Virus Replication. *Journal of Virology*, 81(16):8827–8832.

Lorente-Galdos, B., Bleyhl, J., Santpere, G., Vives, L., Ramírez, O., Hernandez, J., Anglada, R., Cooper, G. M., Navarro, A., Eichler, E. E., and Marques-Bonet, T. (2013). Accelerated exon evolution within primate segmental duplications. *Genome Biology*, 14(1):R9.

Lu, H., Giordano, F., and Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics and Bioinformatics*, 14(5):265–279.

Lucito, R., Healy, J., Alexander, J., Reiner, A., Esposito, D., Chi, M., Rodgers, L., Brady, A., Sebat, J., Troge, J., West, J. A., Rostan, S., Nguyen, K. C. Q., Powers, S., Ye, K. Q., Olshen, A., Venkatraman, E., Norton, L., and Wigler, M. (2003). Representational Oligonucleotide Microarray Analysis: A High-Resolution Method to Detect Genome Copy Number Variation. *Genome Research*, 13(10):2291–2305.

Lupski, J. R. (1998). Genomic disorder: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends in Genetics*, 14(10):417–422.

Lynch, M. (2007). The Origins of Genome Architecture. *Journal of Heredity*, 98(6):633–634.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., Pendlington, Z. M., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F., and Parkinson, H. (1895). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, 45(Database issue):D896–D901.

Madlon-Kay, S., Montague, M. J., Brent, L. J. N., Ellis, S., Zhong, B., Snyder-Mackler, N., Horvath, J. E., Skene, J. H. P., and Platt, M. L. (2018). Weak effects of common genetic variation in oxytocin and vasopressin receptor genes on rhesus macaque social behavior. *American Journal of Primatology*, 80(10):e22873.

Magi, A., Tattini, L., Pippucci, T., Torricelli, F., and Benelli, M. (2012). Read count approach for DNA copy number variants detection. *Bioinformatics*, 28(4):470–478.

Mallick, S. et al. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206.

Mansai, S. P., Kado, T., and Innan, H. (2011). The Rate and Tract Length of Gene Conversion between Duplicated Genes. *Genes*, 2(2):313–331.

Marco-Sola, S., Sammeth, M., Guigó, R., and Ribeca, P. (2012). The GEM mapper: Fast, accurate and versatile alignment by filtration. *Nature Methods*, 9(12):1185–1188.

Marotta, M., Chen, X., Inoshita, A., Stephens, R., Thomas Budd, G., Crowe, J. P., Lyons, J., Kondratova, A., Tubbs, R., and Tanaka, H. (2012). A common copy-number breakpoint of ERBB2 amplification in breast cancer colocalizes with a complex block of segmental duplications. *Breast Cancer Research*, 14(6):R150.

Marques, A. C., Dupanloup, I., Vinckenbosch, N., Reymond, A., and Kaessmann, H. (2005). Emergence of young human genes after a burst of retroposition in primates. *PLoS Biology*, 3(11):e357.

Marques, A. C., Vinckenbosch, N., Brawand, D., and Kaessmann, H. (2008). Functional diversification of duplicate genes through subcellular adaptation of encoded proteins. *Genome Biology*, 9(3):R54.

Marques-Bonet, T. and Eichler, E. E. (2009). The evolution of human segmental duplications and the core duplicon hypothesis. *Cold Spring Harbor Symposia on Quantitative Biology*, 74:355–362.

Marques-Bonet, T., Kidd, J. M., Ventura, M., Graves, T. A., Cheng, Z., Hillier, L. W., Jiang, Z., Baker, C., Malfavon-Borja, R., Fulton, L. A., Alkan, C., Aksay, G., Girirajan, S., Siswara, P., Chen, L., Cardone, M. F., Navarro, A., Mardis, E. R., Wilson, R. K., and Eichler, E. E. (2009a). A burst of segmental duplications in the genome of the African great ape ancestor. *Nature*, 457(7231):877–881.

Marques-Bonet, T., Ryder, O. A., and Eichler, E. E. (2009b). Sequencing Primate Genomes: What Have We Learned? *Annual Review of Genomics and Human Genetics*, 10(1):355–386.

Martin, C. L., Kirkpatrick, B. E., and Ledbetter, D. H. (2015). Copy Number Variants, Aneuploidies, and Human Disease. *Clinics in Perinatology*, 42(2):227–242.

Massey Jr., F. J. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46(253):68–78.

McGrath, C. L., Casola, C., and Hahn, M. W. (2009). Minimal effect of ectopic gene conversion among recent duplicates in four mammalian genomes. *Genetics*, 182(2):615–622.

McGrath, C. L. and Lynch, M. (2012). Evolutionary Significance of Whole-Genome Duplication. In Soltis, P. S. and Soltis, D. E., editors, *Polyploidy and Genome Evolution*, pages 1–21. Springer-Verlag, Berlin Heidelberg.

Mendel, G. (1866). Versuche über Plflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr 1865*, Abhandlungen:3–47.

Mézard, C., Pompon, D., and Nicolas, A. (1992). Recombination between similar but not identical DNA sequences during yeast transformation occurs within short stretches of identity. *Cell*, 70(4):659–670.

Mills, R. E. et al. (2011). Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332):59–65.

Monlong, J., Cossette, P., Meloche, C., Rouleau, G., Girard, S. L., and Bourque, G. (2018). Human copy number variants are enriched in regions of low mappability. *Nucleic Acids Research*, 46(14):7236–7249.

Mühlhausen, S. and Kollmar, M. (2013). Whole genome duplication events in plant evolution reconstructed and predicted using myosin motor proteins. *BMC Evolutionary Biology*, 13:202.

Muller, H. J. (1936). Bar Duplication. *Science*, 83(2161):528–530.

Murti, J. R., Bumbulis, M., and Schimenti, J. C. (1992). High-frequency germ line gene conversion in transgenic mice. *Molecular and Cellular Biology*, 12(6):2545–2552.

Myers, S., Freeman, C., Auton, A., Donnelly, P., and McVean, G. (2008). A common sequence motif associated with recombination hot spots and genome instability in humans. *Nature Genetics*, 40(9):1124–1129.

Nagarajan, N. and Pop, M. (2013). Sequence assembly demystified. *Nature Reviews Genetics*, 14(3):157–167.

Nagylaki, T. (1984). The evolution of multigene families under intrachromosomal gene conversion. *Genetics*, 106(3):529–548.

Necşulea, A., Popa, A., Cooper, D. N., Stenson, P. D., Mouchiroud, D., Gautier, C., and Duret, L. (2011). Meiotic recombination favors the spreading of deleterious mutations in human populations. *Human Mutation*, 32(2):198–206.

Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nei, M. and Li, W.-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, 76(10):5269–5273.

Nicholas, T. J., Cheng, Z., Ventura, M., Mealey, K., Eichler, E. E., and Akey, J. M. (2009). The genomic architecture of segmental duplications and associated copy number variants in dogs. *Genome Research*, 19(3):491–499.

Nuttle, X., Huddleston, J., O'Roak, B. J., Antonacci, F., Fichera, M., Romano, C., Shendure, J., and Eichler, E. E. (2013). Rapid and accurate large-scale genotyping of duplicated genes and discovery of interlocus gene conversions. *Nature Methods*, 10(9):903–909.

Ohno, S. (1970). *Evolution by Gene Duplication*. Springer Science + Business Media, New York.

Ohno, S., Wolf, U., and Atkin, N. B. (1968). Evolution From Fish To Mammals By Gene Duplication. *Hereditas*, 59(1):169–187.

Ohta, T. (1980). *Evolution and Variation of Multigene Families*. Springer-Verlag, Berlin Heidelberg New York.

Ohta, T. (1982). Allelic and nonallelic homology of a supergene family. *Proceedings of the National Academy of Sciences of the United States of America*, 79(10):3251–3254.

Ohta, T. (1983). On the evolution of multigene families. *Theoretical Population Biology*, 23(2):216–240.

Ohta, T. (2010). Gene Conversion and Evolution of Gene Families: An Overview. *Genes*, 1(3):349–356.

Ottolini, B., Hornsby, M. J., Abujaber, R., MacArthur, J. A., Badge, R. M., Schwarzacher, T., Albertson, D. G., Bevins, C. L., Solnick, J. V., and Hollox, E. J. (2014). Evidence of convergent evolution in humans and macaques supports an adaptive role for copy number variation of the β-defensin-2 gene. *Genome Biology and Evolution*, 6(11):3025–3038.

Papadakis, M. N. and Patrinos, G. P. (1999). Contribution of gene conversion in the evolution of the human β-like globin gene family. *Human Genetics*, 104(2):117–125.

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242.

Pegueroles, C., Laurie, S., and Albà, M. M. (2013). Accelerated evolution after gene duplication: A time-dependent process affecting just one copy. *Molecular Biology and Evolution*, 30(8):1830–1842.

Perelman, P., Johnson, W. E., Roos, C., Seuánez, H. N., Horvath, J. E., Moreira, M. A. M., Kessing, B., Pontius, J., Roelke, M., Rumpler, Y., Schneider, M. P. C., Silva, A., O'Brien, S. J., and Pecon-Slattery, J. (2011). A molecular phylogeny of living primates. *PLoS Genetics*, 7(3):1–17.

Perry, G. H., Ben-Dor, A., Tsalenko, A., Sampas, N., Rodriguez-Revenga, L., Tran, C. W., Scheffer, A., Steinfeld, I., Tsang, P., Yamada, N. A., Park, H. S., Kim, J. I., Seo, J. S., Yakhini, Z., Laderman, S., Bruhn, L., and Lee, C. (2008a). The fine-scale and complex architecture of human copy-number variation. *The American Journal of Human Genetics*, 82(3):685–695.

Perry, G. H., Tchinda, J., McGrath, S. D., Zhang, J., Picker, S. R., Cáceres, A. M., Iafrate, A. J., Tyler-Smith, C., Scherer, S. W., Eichler, E. E., Stone, A. C., and Lee, C. (2006). Hotspots for copy number variation in chimpanzees and humans. *Proceedings of the National Academy of Sciences of the United States of America*, 103(21):8006–8011.

Perry, G. H., Yang, F., Marques-Bonet, T., and Murphy, C. (2008b). Copy number variation and evolution in humans and chimpanzees. *Genome Research*, 18(11):1698–1710.

Pich i Rosello, O. and Kondrashov, F. A. (2014). Long-Term Asymmetrical Acceleration of Protein Evolution after Gene Duplication. *Genome Biology and Evolution*, 6(8):1949–1955.

Prado-Martinez, J. et al. (2013). Great ape genetic diversity and population history. *Nature*, 499(7459):471–475.

Proulx, S. R. (2012). Multiple routes to subfunctionalization and gene duplicate specialization. *Genetics*, 190(2):737–751.

Ptak, S. E., Voelpel, K., and Przeworski, M. (2004). Insights into recombination from patterns of linkage disequilibrium in humans. *Genetics*, 167(1):387–397.

Pu, L., Lin, Y., and Pevzner, P. A. (2018). Detection and analysis of ancient segmental duplications in mammalian genomes. *Genome Research*, 28(6):901–909.

Pybus, M., Luisi, P., Dall'Olio, G. M., Uzkudun, M., Laayouni, H., Bertranpetit, J., and Engelken, J. (2015). Hierarchical boosting: A machine-learning framework to detect and classify hard selective sweeps in human populations. *Bioinformatics*, 31(24):3946–3952.

Qian, W. and Zhang, J. (2014). Genomic evidence for adaptation by gene duplication. *Genome Research*, 24(8):1356–1362.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Redon, R. et al. (2006). Global variation in copy number in the human genome. *Nature*, 444(7118):444–454.

Reiter, L. T., Hastings, P. J., Nelis, E., De Jonghe, P., Van Broeckhoven, C., and Lupski, J. R. (1998). Human meiotic recombination products revealed by sequencing a hotspot for homologous strand exchange in multiple HNPP deletion patients. *American Journal of Human Genetics*, 62(5):1023–1033.

Renaud, M., Praz, V., Vieu, E., Florens, L., Washburn, M. P., L'Hôte, P., and Hernandez, N. (2014). Gene duplication and neofunctionalization: POLR3G and POLR3GL. *Genome Research*, 24(1):37–51.

Richardson, S. R., Salvador-Palomeque, C., and Faulkner, G. J. (2014). Diversity through duplication: Whole-genome sequencing reveals novel gene retrocopies in the human population. *BioEssays*, 36(5):475–481.

Rogers, J. (2013). In transition: Primate genomics at a time of rapid change. *ILAR Journal*, 54(2):224–233.

Rogers, J., Raveendran, M., Fawcett, G. L., Fox, A. S., Shelton, S. E., Oler, J. A., Cheverud, J., Muzny, D. M., Gibbs, R. A., Davidson, R. J., and Kalin, N. H. (2013). CRHR1 genotypes, neural circuits and the diathesis for anxiety and depression. *Molecular Psychiatry*, 18(6):700–707.

Rouyer, F., Simmler, M. C., Page, D. C., and Weissenbach, J. (1987). A sex chromosome rearrangement in a human XX male caused by Alu-Alu recombination. *Cell*, 51(3):417–425.

Rozen, S., Skaletsky, H., Marszalek, J. D., Minx, P. J., Cordum, H. S., Waterston, R. H., Wilson, R. K., and Page, D. C. (2003). Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature*, 423(6942):873–876.

Rubnitz, J. and Subramani, S. (1984). The minimum amount of homology required for homologous recombination in mammalian cells. *Molecular and Cellular Biology*, 4(11):2253–2258.

Salzberg, S. L. and Yorke, J. A. (2005). Beware of mis-assembled genomes. *Bioinformatics*, 21(24):4320–4321.

Samonte, R. V. and Eichler, E. E. (2002). Segmental duplications and the evolution of the primate genome. *Nature Reviews Genetics*, 3(1):65–72.

Sandelowski, M. (1995). Sample size in qualitative research. *Research in Nursing & Health*, 18(2):179–183.

Schildkraut, E., Miller, C. A., and Nickoloff, J. A. (2005). Gene conversion and deletion frequencies during double-strand break repair in human cells are controlled by the distance between direct repeats. *Nucleic Acids Research*, 33(5):1574–1580.

Schwandt, M. L., Lindell, S. G., Chen, S., Higley, J. D., Suomi, S. J., Heilig, M., and Barr, C. S. (2010). Alcohol response and consumption in adolescent rhesus macaques: Life history and genetic influences. *Alcohol*, 44(1):67–80.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T. C., Trask, B., Patterson, N., Zetterberg, A., and Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. *Science*, 305(5683):525–528.

Sebesta, M. and Krejci, L. (2016). Mechanism of Homologous Recombination. In Hanaoka, F. and Sugasawa, K., editors, *DNA Replication, Recombination, and Repair*, pages 73–109. Springer, Japan.

Sémon, M. and Wolfe, K. H. (2007). Consequences of genome duplication. *Current Opinion in Genetics and Development*, 17(6):505–512.

Serres-Armero, A. et al. (2017). Similar genomic proportions of copy number variation within gray wolves and modern dog breeds inferred from whole genome sequencing. *BMC Genomics*, 18(1):1–15.

Sharp, A. J., Locke, D. P., McGrath, S. D., Cheng, Z., Bailey, J. A., Vallente, R. U., Pertz, L. M., Clark, R. A., Schwartz, S., Segraves, R., Oseroff, V. V., Albertson, D. G., Pinkel, D., and Eichler, E. E. (2005). Segmental duplications and copy-number variation in the human genome. *American Journal of Human Genetics*, 77(1):78–88.

She, X., Cheng, Z., Zöllner, S., Church, D. M., and Eichler, E. E. (2008). Mouse segmental duplication and copy number variation. *Nature Genetics*, 40(7):909–914.

She, X., Horvath, J. E., Jiang, Z., Liu, G., Furey, T. S., Christ, L., Clark, R., Graves, T., Gulden, C. L., Alkan, C., Bailey, J. A., Sahinalp, C., Rocchi, M., Haussler, D., Wilson, R. K., Miller, W., Schwartz, S., and Eichler, E. E. (2004). The structure and evolution of centromeric transition regions within the human genome. *Nature*, 430(7002):857–864.

She, X., Liu, G., Ventura, M., Zhao, S., Misceo, D., Roberto, R., Cardone, M. F., Rocchi, M., Green, E. D., Archidiacano, N., and Eichler, E. E. (2006). A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-ape expansion of intrachromosomal duplications. *Genome Research*, 16(5):576–583.

Shen, P. and Huang, H. V. (1986). Homologous recombination in Escherichia coli: dependence on substrate length and homology. *Genetics*, 112(3):441–457.

Shen, P. and Huang, H. V. (1989). Effect of base pair mismatches on recombination via the RecBCD pathway. *Molecular & General Genetics*, 218(2):358–360.

Shi, L. et al. (2016). Long-read sequencing and de novo assembly of a Chinese genome. *Nature Communications*, 7:12065.

Simpson, J. T. and Pop, M. (2015). The Theory and Practice of Genome Sequence Assembly. *Annual Review of Genomics and Human Genetics*, 16(1):153–172.

Smit, A. F. A., Hubley, R., and Green, P. (2013). RepeatMasker Open-4.0. *"http://www.repeatmaker.org"*.

Smith, D. G. and Mcdonough, J. (2005). Mitochondrial DNA variation in Chinese and Indian Rhesus macaques (Macaca mulatta). *American Journal of Primatology*, 65(1):1–25.

Stankiewicz, P. and Lupski, J. (2010). Structural variation in the human genome and its role in disease. *Annual Review of Medicine*, 61:437–455.

Stankiewicz, P. and Lupski, J. R. (2002). Genome architecture, rearrangements and genomic disorders. *Trends in Genetics*, 18(2):74–82.

Stratonovich, R. L. (1960). Conditional Markov Processes. *Theory of Probability and its Applications*, 5(2):156–178.

Sudmant, P. H. et al. (2015a). An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81.

Sudmant, P. H. et al. (2015b). Global diversity, population stratification, and selection of human copy number variation. *Science*, 349(6253):aab3761.

Sudmant, P. H., Huddleston, J., Catacchio, C. R., Malig, M., Hillier, L. W., Baker, C., Mohajeri, K., Kondova, I., Bontrop, R. E., Persengiev, S., Antonacci, F., Ventura, M., Prado-Martinez, J., Project, G. A. G., Marques-Bonet, T., and Eichler, E. E. (2013). Evolution and diversity of copy number variation in the great ape lineage. *Genome Research*, 23(9):1373–1382.

Sung, P. (2018). Introduction to the thematic minireview series: DNA double-strand break repair and pathway choice. *Journal of Biological Chemistry*, 293(27):10500–10501.

Symington, L. S., Rothstein, R., and Lisby, M. (2014). Mechanisms and regulation of mitotic recombination in Saccharomyces cerevisiae. *Genetics*, 198(3):795–835.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595.

Taormina, P. L., Satkoski Trask, J. A., Smith, D. G., and Kanthaswamy, S. (2012). Variation in CCL3L1 copy number in rhesus macaques (Macaca mulatta). *Comp Med*, 62(3):218–224.

Tartof, K. D. (1975). Redundant genes. *Annual Review of Genetics*, 9:355–385.

Tattini, L., D'Aurizio, R., and Magi, A. (2015). Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Frontiers in Bioengineering and Biotechnology*, 3:Article 92.

Teshima, K. M. and Innan, H. (2004). The effect of gene conversion on the divergence between duplicated genes. *Genetics*, 166(3):1553–1560.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299–1320.

Thornton, K. and Long, M. (2002). Rapid divergence of gene duplicates on the Drosophila melanogaster X chromosome. *Molecular Biology and Evolution*, 19(6):918–925.

Thornton, K. R. (2007). The neutral coalescent process for recent gene duplications and copy-number variants. *Genetics*, 177(2):987–1000.

Trombetta, B. and Cruciani, F. (2017). Y chromosome palindromes and gene conversion. *Human Genetics*, 136(5):605–619.

Trombetta, B., Fantini, G., D'Atanasio, E., Sellitto, D., and Cruciani, F. (2016). Evidence of extensive non-allelic gene conversion among LTR elements in the human genome. *Scientific Reports*, 6:1–11.

Tubio, J. M. C. (2015). Somatic structural variation and cancer. *Briefings in Functional Genomics*, 14(5):339–351.

Uno, Y., Iwasaki, K., Yamazaki, H., and Nelson, D. R. (2011). Macaque cytochromes P450: Nomenclature, transcript, gene, genomic structure, and function. *Drug Metabolism Reviews*, 43(3):346–361.

Uno, Y., Uehara, S., and Yamazaki, H. (2016). Utility of non-human primates in drug development: Comparison of non-human primate and human drug-metabolizing cytochrome P450 enzymes. *Biochemical Pharmacology*, 121:1–7.

Valentine, L. E., Loffredo, J. T., Bean, A. T., León, E. J., MacNair, C. E., Beal, D. R., Piaskowski, S. M., Klimentidis, Y. C., Lank, S. M., Wiseman, R. W., Weinfurter, J. T., May, G. E., Rakasz, E. G., Wilson, N. A., Friedrich, T. C., O'Connor, D. H., Allison, D. B., and Watkins, D. I. (2009). Infection with "escaped" virus variants impairs control of simian immunodeficiency virus SIVmac239 replication in Mamu-B*08-positive macaques. *Journal of Virology*, 83(22):11514–11527.

Vallender, E. J., Priddy, C. M., Hakim, S., Yang, H., Chen, G. L., and Miller, G. M. (2008). Functional variation in the 3' untranslated region of the serotonin transporter in human and rhesus macaque. *Genes, Brain and Behavior*, 7(6):690–697.

Vallender, E. J., Rüedi-Bettschen, D., Miller, G. M., and Platt, D. M. (2010). A pharmacogenetic model of naltrexone-induced attenuation of alcohol consumption in rhesus monkeys. *Drug and Alcohol Dependence*, 109(1-3):252–256.

Vandeweyer, G. and Kooy, R. F. (2013). Detection and interpretation of genomic structural variation in health and disease. *Expert Review of Molecular Diagnostics*, 13(1):61–82.

Vanin, E. F. (1985). Processed Pseudogenes: Characterisitcs and Evolution. *Annual Review of Genetics*, 19:253–272.

Veeramah, K. R., Gutenkunst, R. N., Woerner, A. E., Watkins, J. C., and Hammer, M. F. (2014). Evidence for increased levels of positive and negative selection on the X chromosome versus autosomes in humans. *Molecular Biology and Evolution*, 31(9):2267–2282.

Venter, J. C. et al. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351.

Ventura, M., Catacchio, C. R., Alkan, C., Marques-Bonet, T., Sajjadian, S., Graves, T. A., Hormozdiari, F., Navarro, A., Malig, M., Baker, C., Lee, C., Turner, E. H., Chen, L., Kidd, J. M., Archidiacono, N., Shendure, J., Wilson, R. K., and Eichler, E. E. (2011). Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. *Genome Research*, 21(10):1640–1649.

Ventura, M., Catacchio, C. R., Sajjadian, S., Vives, L., Sudmant, P. H., Marques-Bonet, T., Graves, T. A., Wilson, R. K., and Eichler, E. E. (2012). The evolution of African great ape subtelomeric heterochromatin and the fusion of human chromosome 2. *Genome Research*, 22(6):1036–1049.

Vinckenbosch, N., Dupanloup, I., and Kaessmann, H. (2006). Evolutionary fate of retroposed gene copies in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, 103(9):3220–3225.

Vinogradov, A. E. (2012). Large scale of human duplicate genes divergence. *Journal of Molecular Evolution*, 75(1-2):25–33.

Vinson, A., Prongay, K., and Ferguson, B. (2013). The value of extended pedigrees for next-generation analysis of complex disease in the rhesus macaque. *ILAR Journal*, 54(2):91–105.

Wajid, B., Sohail, M. U., Ekti, A. R., and Serpedin, E. (2016). The A, C, G, and T of genome assembly. *BioMed Research International*, 2016.

Waldman, A. S. (2008). Ensuring the fidelity of recombination in mammalian chromosomes. *BioEssays*, 30(11-12):1163–1171.

Waldman, a. S. and Liskay, R. M. (1988). Dependence of intrachromosomal recombination in mammalian cells on uninterrupted homology. *Molecular and Cellular Biology*, 8(12):5350–5357.

Walsh, J. B. (1987). Sequence-dependent gene conversion: can duplicated genes diverge fast enough to escape conversion? *Genetics*, 117(3):543–557.

Walter, L. and Ansari, A. A. (2015). MHC and KIR polymorphisms in rhesus macaque SIV infection. *Frontiers in Immunology*, 6:Article 540.

Wang, J. et al. (2008). The diploid genome sequence of an Asian individual. *Nature*, 456(7218):60–65.

Watson, J. D. and Crick, F. H. (1953). Molecular structrue of nucleic acids. *Nature*, 171(4356):737–738.

Weckselblatt, B. and Rudd, M. K. (2015). Human Structural Variation: Mechanisms of Chromosome Rearrangements. *Trends in Genetics*, 31(10):587–599.

Wei, W. H., Hemani, G., and Haley, C. S. (2014). Detecting epistasis in human complex traits. *Nature Reviews Genetics*, 15(11):722–733.

Wheeler, D. A. et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452(7189):872–876.

Wiseman, R. W., Karl, J. A., Bohn, P. S., Nimityongskul, F. A., Starrett, G. J., and O'Connor, D. H. (2013). Haplessly hoping: Macaque major histocompatibility complex made easy. *ILAR Journal*, 54(2):196–210.

Wiuf, C. and Hein, J. (2000). The coalescent with gene conversion. *Genetics*, 155(1):451–462.

Wolf, A., Millar, D. S., Caliebe, A., Horan, M., Newsway, V., Kumpf, D., Steinmann, K., Chee, I. S., Lee, Y. H., Mutirangura, A., Pepe, G., Rickards, O., Schmidtke, J., Schempp, W., Chuzhanova, N., Kehrer-Sawatzki, H., Krawczak, M., and Cooper, D. N. (2009). A gene conversion hotspot in the human growth hormone (GH1) gene promoter. *Human Mutation*, 30(2):239–247.

Xue, C. et al. (2016). The population genomics of rhesus macaques (Macaca mulatta) based on whole genome sequences. *Genome Research*, 26(12):1651–1662.

Yang, L., Luquette, L. J., Gehlenborg, N., Xi, R., Haseley, P. S., Hsieh, C. H., Zhang, C., Ren, X., Protopopov, A., Chin, L., Kucherlapati, R., Lee, C., and Park, P. J. (2013). Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell*, 153(4):919–929.

Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research*, 19(9):1586–1592.

Zagaria, A., Anelli, L., Coccaro, N., Tota, G., Casieri, P., Cellamare, A., Minervini, A., Minervini, C. F., Brunetti, C., Cumbo, C., Specchia, G., and Albano, F. (2014). 5'RUNX1-3'USP42 chimeric gene in acute myeloid leukemia can occur through an insertion mechanism rather than translocation and may be mediated by genomic segmental duplications. *Molecular Cytogenetics*, 7(1):1–8.

Zarrei, M., MacDonald, J. R., Merico, D., and Scherer, S. W. (2015). A copy number variation map of the human genome. *Nature Reviews Genetics*, 16(3):172–183.

Zerbino, D. R. et al. (2018). Ensembl 2018. *Nucleic Acids Research*, 46(D1):D754–D761.

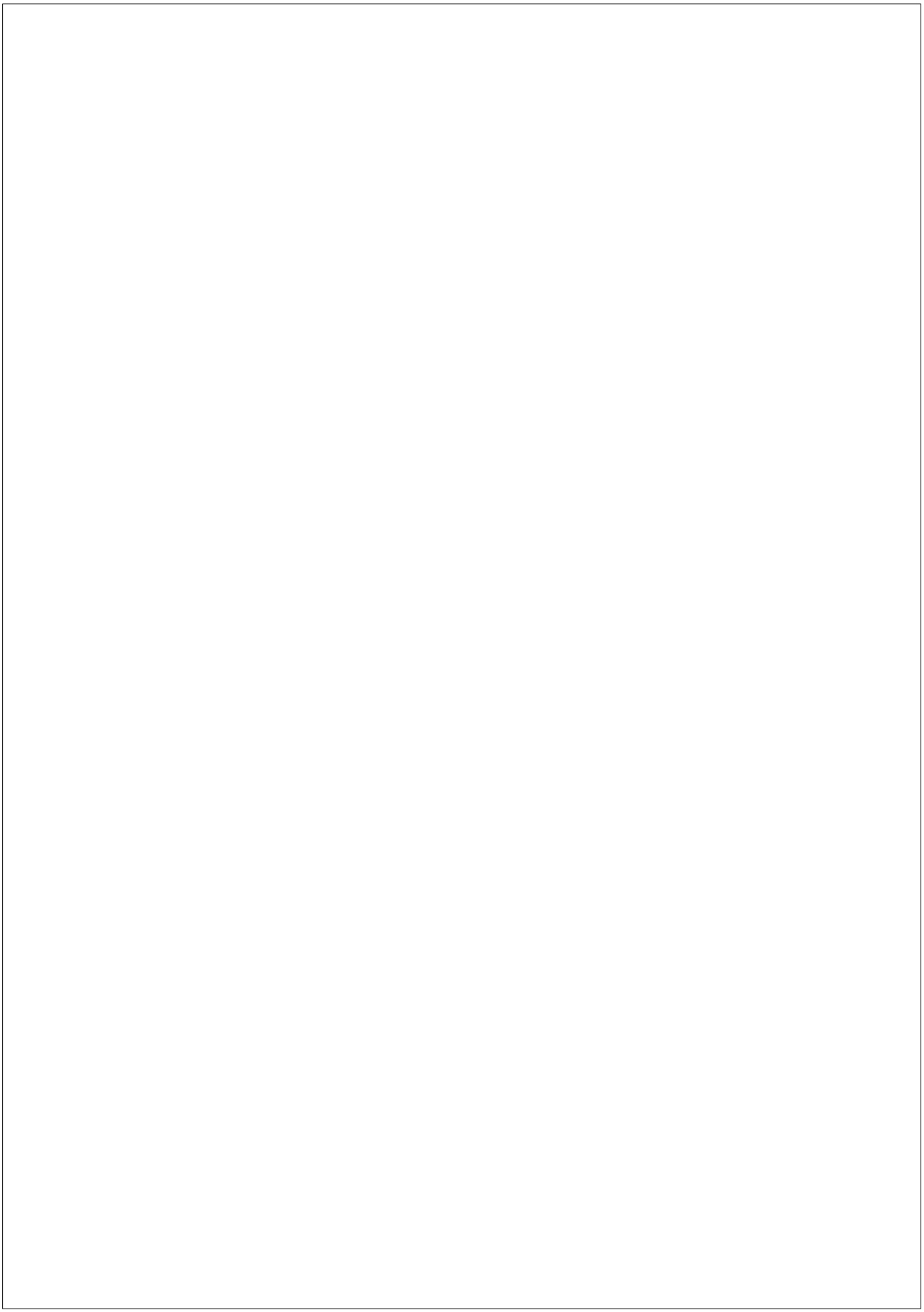
Zhang, F., Gu, W., Hurles, M., and Lupski, J. (2009). Copy Number Variation in Human Health, Disease, and Evolution. *Annual Review of Genomics and Human Genetics*, 10:451–481.

Zhang, L., Lu, H. H. S., Chung, W. Y., Yang, J., and Li, W. H. (2005). Patterns of segmental duplication in the human genome. *Molecular Biology and Evolution*, 22(1):135–141.

Zhi, D. (2007). Sequence correlation between neighboring Alu instances suggests post-retrotransposition sequence exchange due to Alu gene conversion. *Gene*, 390(1-2):117–121.

Zielinski, D., Markus, B., Sheikh, M., Gymrek, M., Chu, C., Zaks, M., Srinivasan, B., Hoffman, J. D., Aizenbud, D., and Erlich, Y. (2014). OTX2 duplication is implicated in hemifacial microsomia. *PLoS ONE*, 9(5):e96788.

Zuckerkandl, E. and Pauling, L. B. (1965). Evolutionary divergence and convergence in proteins. In Bryson, B. and J, V. H., editors, *Evolving Genes and Proteins*, pages 97–166. Academic Press, New York.



Chapter 7

Appendix

7.1 Interplay of interlocus gene conversion and crossover in segmental duplications under a neutral scenario

Diego A. Hartasánchez, Oriol Vallès-Codina, **Marina Brasó-Vives**, and Arcadi Navarro. (2014) G3 (Bethesda) 4(8): 1479-1489.

Article published in G3: Genes | Genomes | Genetics

Hartasánchez DA, Vallès-Codina O, Brasó-Vives M, Navarro A. Interplay of interlocus gene conversion and crossover in segmental duplications under a neutral scenario. G3 Genes, Genomes, Genet. 2014;4(8):1479–89. DOI: 10.1534/g3.114.012435

7.2 SeDuS: segmental duplication simulator

Diego A. Hartasánchez, **Marina Brasó-Vives**, Juanma Fuentes-Díaz, Oriol Vallès-Codina, and Arcadi Navarro. (2016) Bioinformatics 32 (1): 148-150.

Article published in Bioinformatics

Hartasánchez DA, Brasó-Vives M, Fuentes-Díaz J, Vallès-Codina O, Navarro A. SeDuS: Segmental duplication simulator. Bioinformatics. 2016 Jan 1;32(1):148–50. DOI: 10.1093/bioinformatics/btv481

7.3 Effect of collapsed duplications on diversity estimates: what to expect

Diego A. Hartasánchez*, **Marina Brasó-Vives***, Jose Maria Heredia-Genestar*, Marc Pybus, and Arcadi Navarro

Article published in Genome Biology and Evolution

* Equal contributions.

Hartasanchez DA, Braso-Vives M, Heredia-Genestar JM, Pybus M, Navarro A. Effect of collapsed duplications on diversity estimates: What to expect. Genome Biol Evol. 2018;10(11):2899–905. DOI: 10.1093/gbe/evy223