Modeling and simulation of interlocus gene conversion

Diego Andrés Hartasánchez Frenk

TESI DOCTORAL UPF / ANY 2016

DIRECTOR DE LA TESI
Dr. Arcadi Navarro Cuartiellas
Departament de Ciències Experimentals i de la Salut

A Johannes Engelken

Thus, the task is not so much to see
what no one has yet seen, but to
think what nobody has yet thought,
about that which everybody sees.

Arthur Schopenhauer

Acknowledgements

Hace casi 6 años llegué a Barcelona, mi segunda casa, con la intención de empezar un doctorado en biología evolutiva. Mi interés en la biología venía desde hacía muchos años, desde la escuela secundaria y el bachillerato, y se había fortalecido a lo largo de los años, especialmente mientras estudiaba la carrera de física, en la grandiosa UNAM. Habiendo ya explorado la biología desde la perpectiva de la física y de los físicos, quería estudiar biología, en particular evolución molecular, al lado de biólogos y aportar un enfoque diferente a esta maravillosa disciplina.

Así llegué a BioEvo, al IBE, y a la UPF, a este gran grupo de personas que me acogieron desde el primer momento, que me dejaron hacer mis preguntas, que me enseñaron genética de poblaciones, que me enseñaron quiénes son los YRI y quiénes son los CEU, que me mostraron que el DNA antiguo es lo moderno, que me explicaron lo que son un PC-plot y un Manhattan plot, que me mostraron que existen los microRNAs, que me enseñaron qué son los reads y los contigs; que me enseñaron y me formaron en biología evolutiva, que de eso realmente se trata hacer un doctorado.

Por esto, por los maravillosos años que he vivido en esta magnífica ciudad y por haber sido parte de esta aventura, mi agradecimiento profundo y sincero a todos ustedes.

Gracias / gràcies / thank you:

A Arcadi, per haver confiat en mi des del primer moment, per la llibertat que m'has donat durant tots aquests anys, per ensenyar-me que sempre es pot anar més enllà, per la teva coherència, per mostrar-me que no basta amb voler canviar les coses, que les coses es poden canviar i per canviar-les s'ha de saber què dir i com dir-ho, per demostrar-me, un cop més, que ser interdisciplinari és una passada.

A Marina, perquè aquesta tesi només ha estat possible gràcies a tu, per ser la millor en tot el que fas, per ensenyar-me que amb els pronoms febles sí que cal estar-hi molt atent però que *no en cal-hi de massa per comprendre'n-los-hi*, per la teva companyia i els teus ànims constants, per ser curiosa i perfeccionista, pels enigmes, per l'Interplay i pel SeDuS, per tenir-me confiança i creure en mi, pel gran equip que som i pel millor equip que serem.

A Oriol, per la nostra amistat, perquè construint un *forward simulator* vam construir una relació que tira cap endavant i per mostrar-me que si es vol, no cal

renunciar a res.

A JuanMa, por la teva paciència i la teva disposició, perquè només una persona en aquest món ha acabat un Ironman i li ha posat cara a SeDuS.

A los *419ers*: a Urko, Juan, Nino, Òscar y Ferran, por todas las conversaciones, las risas y los viernes en compañía de la tortuga.

To Moose, you really are a great man, thanks for your kindness, your constant willingness to lend a hand and your support.

To Josephine, thanks for your stories and for accepting my metabolic rate.

A Ángel y a Txema, por haber curtido a Cadaqués y parido a Floquet.

A Jaume, David, Tomàs, Elena, Hafid, Carles, Francesc, Yolanda, Ricard, Luc y Pepi.

To Fyodor, Mar, Laurent, Toni and Tomàs.

A Judit, per fer tot possible, y por tu risa.

A todos aquellos que han pasado por BioEvo desde que pisé este lugar: en especial a Ixa, Laura, Paula, Mònica, Victoria, Brandon, Giovanni, Rui, Carlos M., Graciela, Neus, Koldo, Íñigo, Mayukh, Guillem, Raj, Xavi y a toda la nueva generación de estudiantes que me hacen darme cuenta que *it's time to move on*.

A los *tomasines*. A los de antaño, con los que me inicié en BioEvo: Javi, Belén, Irene H., Marcos, Marc D., Marta, Tiago y Xavi Q.; a los que ya llevan un rato, con los que continué en BioEvo: Jéssica, Guillem, Raquel, Marc de M., Irene L., Clàudia; a mis nuevos compañeros de despacho: Aitor, Lukas, Martin, Inna e Irune; y a Esther.

A la *llumenera d'Occident*: a Gabriel, per la teva parsimònia; a Urko, per la teva saviesa i les teves històries; a Gerard, por la teva llibertat; a Fernando, por llevar el color adentro; a Txema, pel que ets capaç de trobar a la internet; i a Oriol, pel que ets capaç d'escriure en una tarda; a tots, per la vostra intel.ligència i la vostra heterogènia.

A Javi, per totes les converses, la teva sinceritat i per sempre preguntar si tot anava bé. A Ignasi, per la teva claredat.

A Gabriel i a Urko, a Ludo i a Marta, per ser professors, companys i amics, per servir-me de guia.

A los del CSL, en especial a Adriano, Max, Sergi, Salva, Carlos y Raúl; a los del EGA, en especial a Ángel, Jordi, Óscar, Mauricio, Mario y Sabela; a los del servei, en especial a Roger, Núria y Raquel.

A los del IBE-CMIMA, en especial a Gisella, Cristina, Carlos, Joan, Jesús, Elena G. y Elena C., y con un recuerdo muy especial por Margarita, eres una

inspiración para mí.

A la gente que me acompaña en el día a día en este edificio, a los que saludo al llegar y cuando me voy y con los que comparto historias y anécdotas: a Mercedes, por las plantitas y los chiles, a Juan De Dios, a Jorge y a Manuel.

Al coro del PRBB, del que hasta me hicieron presidente, gracias a todos, especialmente a Òscar, Eva, Inma, Mònica y Estel y a todos los que me acompañado en el *low-end*.

A toda la gente de este maravilloso edificio con quien he forjado una amistad. No hay manera que aquí mencione a todos, pero gracias al *Bort* Barcénas, Lucho, Ramón, Jose, Tommaso, Michi, María Andree, Emre, Alba, Nieves, Leszek, Carla, Jesse, Joan Pau, Juanjo, Laura, Clara, Jürgen y tantísimos más. Gracias al PRBB, por los seminarios, por el volley, por las beer-sessions y por dejarme ver el mar cada día.

To all the people at SOKENDAI that showed me what hospitality really means and introduced me to a culture from which we have so much to learn: Hideki-sensei, Jeff-san, Metal-san, Ryuichi-san, Sayaka-san, Anand-san, Jun-san, Shohei-san, Quintin-san, Tetsuya-san, Kazuki-san, Akiko-san, Jonas-san, and all the other wonederful people at SOKENDAI.

A Osbaldo Resendis y a todo su grupo.

A Anomalocaris, una aventura basquetbolísitca que es un pequeño orgullo que llevo dentro. Gracias a los eternos luchadores: Pierre, Javi e Ignasi; a Carlos V., David, y Marc S.; a Gerard, Antonio, Joan, Edu, Jesse, Lukas, Tommaso, Jordi Majó, Jordi Mayol, Davide, Àlex, Miquel, Carlos, Manuel, Bryan, y a todos los demás que han formado parte de este gran equipo.

A los Beachbumbas y a los Naranjitos, por los veranos playeros y los trofeos. A todos con los que he compartido esas mañanas futboleras en la cancha del Marítim.

A *la bandita*, porque no tengo palabras para agradecerles lo que me han dado en estos años: por las pláticas, las fiestas, las discusiones, las preguntas y la amistad; porque los quiero. Gracias, a Pierre, por tu irreverencia, tu confianza, y porque aunque te vayas a la pampa nunca te librarás de mí; a Juan, por tu inocencia y por haber sido como mi hermano menor durante estos años; a Nino, por tu pragmatismo y por el catalán que sólo tu y yo hablamos; a Elenita, por querer traer el pelo suelto y llevarlo suelto y llevarlo a todos lados; a María, por los pomodoros, por Ruby y Jade, y por Poble Sec; a Fede, por ser un auténtico luchador, por el *FedeFest* y por completar el *Mexicombo*; a Marco, porque eres

increíble, que por algo existe *Project Telford* y por Australia; a Marc, per Les Pobles, pel gust que tens per cuinar i ser crític, i per l'ajuda amb el Collapsed; a Licha, por enseñarme que la optimización del tiempo es una utopía, que tirar cajas vacías a la basura deja espacio para poner una luz, por dejarme usar la Thermomix, por cagarte en todo y por ser la mejor compañera de piso que alguien puede tener.

A mi familia: a mis papás, José Miguel y Alicia, por haberme enseñado a perserguir mis sueños y por haberme dado las herramientas para lograrlo, por confiar siempre en mí y en mis decisiones, por apoyarme siempre y por impulsarme a llegar siempre más lejos; a mis hermanos, Josemi y Jan, es duro vivir lejos de ustedes, gracias por ser mis eternos acompañantes, con quines crecí y seguiré creciendo, siempre; a mi hermana, Sandra, por quererme tanto, porque te llevo siempre conmigo a todos lados, por tu sinceridad y tu confianza; a Dani, Gaby, Luis, Julia y Luisa, por el futuro.

A Abril, por haber sido mi compañera en esta aventura, por la diversión del juego inocente, por el poder de la salud consciente, por tu compromiso con la vida, por tu subversiva e incandescente luz, por los buenos tiempos y por nuestro porvenir.

A mis abuelos, Licha y Silvestre, Joselillo y Mamina. A Mutti y a Margit. A mis tíos, especialmente a Cuqui y a Jas, a Javier y a Maritere, siempre en mi recuerdo.

A mis primas, que por algo se me conoce como *Diego el de las primas*, especialmente a Daniela, María José, Irene y Sinaia. A Pamela, por haberle puesto a esta tesis una cara chidísima.

A mis primos, especialmente a Esteban, que tantos fines de semana me vino a visitar, y a Felipe, por ser un referente en mi vida.

Als meus amics de Barcelona: al Xavi i al Tino, per ser companys de vida, a la Cris i al Plèbot, a l'Here i al Scott, als d'Horta, al PiS, per aquell primer any a Barcelona, per aquella paret i per la Cala del Sr. Ramón, a l'Enric, al Xavi Normal i al Roger, per sa fresqueta d'un vespre a Menorca; a l'Equip B; a Marco y Marc T.; a Espe, Vale y a Osvaldo; a Jess; a Verena y a Pauet, por enseñarme que el futuro siempre es luminoso.

A mis amigos de México, especialmente a aquellos que me han visitado una y otra vez por estos rumbos: Monch, Nico, Barna, Abuelo, Alán, Fabián y Abraham.

A Mattea, por haberme aguantado este par de meses a pesar de estar yo en

modo tesis, por tu curiosidad y por tus preguntas.

A Cyntia, por tu cariño y tu apoyo permanente en este año, por tu risa tierna y tu practicidad, por la salsa y el raggeatón, por el café, por aceptarme como soy y ser siempre auténtica, porque ha sido maravilloso conocerte y que formes partes de mi vida.

Al *Taper Club*: a María, Marc, Joha y Fede, porque juntos le dimos vida; a Pierre, Ele, por darle alas; a Juan, Nino y Arturo por hacerlo grande (y a Marc Haber y Martino porque también contibuyeron a ello); a Ali y Marco por llevarlo a otro nivel; a María Niño y a Adriano, por hacerlo *sui generis*; a Katha y a Miruna, por menearlo; a Vero, Vicky y Lisa, por abrirle las puertas; y a Àlex, Lara y Marina por asegurarle el futuro que se merece. Gracias, *by far*, porque la vida sólo es posible compartiendo y compartiéndola.

A Joha. Más allá del dolor que perdura en mí desde que nos dejaste, me has enseñado más que nadie en estos años. Gracias, Joha, por tu amistad sincera, por tu ternura, por tu locura, por cada noche que cruzamos Barcelona pasándonos el frisbee, por cada paseo que dimos por la montaña, por cada charla que tuvimos cuando sólo quedábamos tú y yo en el laboratorio, por tu adoración a la ciencia, por tu adoración a la vida, y por tu rechazo a vivir sin poder ser tú mismo. Te extraño, amigo.

Abstract

Duplicated regions of the genome, such as Segmental Duplications (SDs), are a pervasive feature of eukaryotic genomes and have been linked to phenotypic changes. Given their evolutionary relevance, having a neutral model to describe their evolution is essential. In this thesis, I report the development of SeDuS, a forward-in-time computer simulator of SD neutral evolution. Duplications are known to undergo a recombination process, termed interlocus gene conversion (IGC), which is known to affect the patterns of variation and linkage disequilibrium within and between duplicates. Here I describe the effects of overlapping crossover and IGC susceptible regions and of incorporating sequence similarity dependence of IGC. Furthermore, since SDs are potential targets of natural selection, I report potential confounding effects of IGC on test statistics when these are applied to duplications. Finally, I explore the possibility of combining results of different test statistics applied genome-wide to detect the presence of collapsed duplications.

Resum

Les regions duplicades del genoma, com ara les duplicacions de segments (SDs), són una característica comuna dels genomes eucariotes i han estat associades a canvis fenotípics. Donada la seva rellevància evolutiva, tenir un model neutre per descriure la seva evolució és essencial. En aquesta tesi, descric el desenvolupament de SeDuS, un simulador computacional endavant en el temps de l'evolució neutra de SDs. Les duplicacions estan sotmeses a un procés de recombinació, anomenat conversió gènica interlocus (IGC), que afecta els patrons de variació i de desequilibri de lligament dins i entre duplicacions. Aquí descric els efectes de sobreposar regions susceptibles de recombinació homòloga amb regions susceptibles d'IGC i d'incorporar dependència d'IGC en la similitud de seqüències. Addicionalment, ja que les SDs són objectius potencial de la selecció natural, informo sobre possibles alteracions a proves estadístiques quan aquestes s'apliquen a regions duplicades sotmeses a IGC. Finalment, exploro la possibilitat de combinar resultats de diferents proves estadístiques aplicades al llarg de tot el genoma per detectar la presència de duplicacions col·lapsades.

Preface

I can see no other escape from this dilemma (lest our true aim be lost for ever) than that some of us should venture to embark on a synthesis of facts and theories, albeit with second-hand and incomplete knowledge of some of them -and at the risk of making fools of ourselves. So much for my apology.

Preface to *What is life?*, Erwin Schrödinger, 1944

All of life on Earth, as far as we know it, has been the product of a replication whereby information has been transferred from one individual to another. A great part of the mystery of life lies precisely in this auto-replicative nature of DNA. How beautiful life is!

It is beautiful indeed, and although there are still many open questions regarding the first steps in this auto-replicative adventure, once DNA was formed and established as the genetic material that organisms would transmit to their offspring, all DNA has been the result of replication and some modification. We'll get to the type of modifications later. For the moment, let us imagine an auto-replicative machinery that accumulates small modifications with time. How beautiful life is!

It is beautiful indeed, and many of us *evolutionary biologists*, among other scientists, are interested in understanding, to say it in big words, the *history of DNA on Earth*. That is, how has DNA changed throughout time and space. By space I mean not strictly *where on Earth* but *in which organism*, or more appropriately, *in which species*, because once sexual reproduction came along, the line of descent from one organism to its offspring became much more intricate, and replication attained a whole new meaning once sex joined the party. Sexual organisms replicate their DNA, and then recombine it with DNA from another individual (in principle, from the same species) in order to give birth to a new organism. In fact, recombination is not at all exclusive of sexual organisms; asexual organisms also recombine DNA. More so, replication and recombination, and the molecular mechanisms that control these processes, are fundamental in order to understand how DNA has evolved. Evolving DNA: how beautiful life is!

It is beautiful indeed, and more beautiful it is once we go back to the modifications I mentioned earlier. If DNA were only about replication and recombination, evolution would have had very limited possibilities. The third essential characteristic of DNA is its modifiability or, rather, its mutability. Mutations (point mutations and short insertion or deletions) allow the information transmitted by DNA to change with time, and so, organisms are capable of adapting to change in their environment. There are many other possible modifications, many of which are a product of a mistake in the replicative or recombinatorial activity of the cell (such as inversions, translocations or retrotranpositions) which can modify not only the sequence, but also the regulation of DNA expression in incredibly diverse ways. How beautiful life is!

It is beautiful indeed because duplication of genetic material is also one of these modifications. Duplications differ substantially from new DNA originated by replication because the latter is the essence of transmission of (reliable) genetic information from parent to offspring while the former is raw material for evolutionary exploration and innovation. Point mutations and deletions are fine because they can alter a previously existing function, but duplications can allow organisms to maintain (albeit with some exceptions and/or difficulties) the original function while having an additional copy that basically has the opportunity to explore and exploit the evolutionary landscape. How beautiful life is!

So, in my view, life is beautiful to a great extent because of the fundamental properties and characteristics of the essential molecule for life as we know it: DNA. DNA can replicate, DNA can recombine, DNA can mutate and DNA can duplicate. This thesis will center on how two of these properties, recombination and duplication, are intricately related. In particular, I will try to contribute to the understanding of the evolution of duplicated regions of the genome and how it is determined and affected by a type of recombination known as intelocus gene conversion.

Contents

Chapter 1

INTRODUCTION

It is said that "necessity is the mother of invention". To be sure, wheels and pulleys were invented out of necessity by the tenacious minds of upright citizens. Looking at the history of mankind, however, one has to add that "leisure is the mother of cultural improvement". Man's creative genius flourished only when his mind, freed from the worry of daily toils, was permitted to entertain apparently useless thoughts. In the same manner, one might say with regard to evolution that "natural selection merely modified, while redundanry created".

Susumu Ohno, 1970

We can't define anything precisely. If we attempt to, we get into that paralysis of thought that comes to philosophers...
One saying to the other: "You don't know what you are talking about!".
The second one says: "What do you mean by talking? What do you mean by you? What do you mean by know?"

Richard Feynman

1.1 Duplications

The duplication of genetic material was considered as possibly being of great importance even before the advent of molecular biology (Haldane, 1932). Already in 1936, the doubling of a chromosomal band in a *Drosophila melanogaster* mutant was recognized by Bridges (1936) and Muller (1936) as a gene duplication causing extreme eye-size reduction. Following this discovery, the potential role of gene duplication was explored by several authors (Stephens, 1951; Nei, 1969) and models began to be developed to explain their evolution. However, it was not until Ohno published his seminal book, *Evolution by Gene Duplication* (Ohno, 1970), that the idea became popular among biologists. Some more years had to pass until enough data was collected to prove that indeed gene duplication has been the source of new genetic material throughout the tree of life (Lynch and Conery, 2000; Zhang, 2003;

Conant and Wolfe, 2008).

Up to this point, I have used the term *gene duplication* to refer to duplication of genetic material in general. This makes sense given the historical relevance of the term *gene*. However, duplications are extremely diverse in terms of size, content, frequency and importance in evolution.

The largest type of duplication is a whole-genome duplication (or polyploidization), in which the entire genome of an individual is duplicated in one generation. Whole-genome duplications have been very common throughout eukaryotic evolution (Sémon and Wolfe, 2007), particularly in plants (Mühlhausen and Kollmar, 2013), but also in organisms such as *Saccharomyces cerevisiae* (Kellis et al., 2004). Whole-genome duplications have the great advantage of not altering the gene-dosage balance at first, although they are energetically costly given the amount of energy needed to maintain a genome double the size of what it used to be. Most of the DNA that originates from a whole-genome duplication will be lost relatively shortly after its appearance (Inoue et al., 2015), but some of it will be maintained as I will describe in the section 1.1.3.

The second type of duplication in terms of size are segmental duplications (SDs). In the following section I will define SDs along with copy-number variants (CNVs) since they are strongly related to each other.

1.1.1 Segmental duplications and copy-number variants

In biology, defining anything is both a sport and an art. It is a sport because whenever *something different* is found, the easiest way to describe it is by giving it a, say, *private definition*. It is an art, because in biology exceptions are always the rule, and finding something different is extremely common. So even though private definitions might be a sport, they are so abstract and intricate constructions that coming up with a particular definition is itself, in many cases, art.

Segmental duplications are not the exception. In a review paper on structural variation in the human genome Feuk et al. (2006) presented some structural variation definitions:

"Copy-number variant (CNV). A segment of DNA that is 1 kb or larger and is present at a variable copy number in comparison with a reference genome. Classes of CNVs include insertions, deletions and duplications. [...]

Copy-number polymorphism. A CNV that occurs in more than 1% of the population. Originally, this definition was used to refer to all CNVs.

Segmental duplication or low-copy repeat. A segment of DNA >1 kb in size that occurs in two or more copies per haploid genome, with the different copies sharing >90% sequence identity. They are often variable in copy number and can therefore also be CNVs."

The text is accompanied by a simple figure to represent them (see Figure 1.1). However, despite the definitions being rather intricate, they do not clarify some important details. What exactly is the difference between SDs and CNVs? Do CNVs need to be >90% similar? If SDs are present in a variable number, do they become CNVs and cease to be SDs?

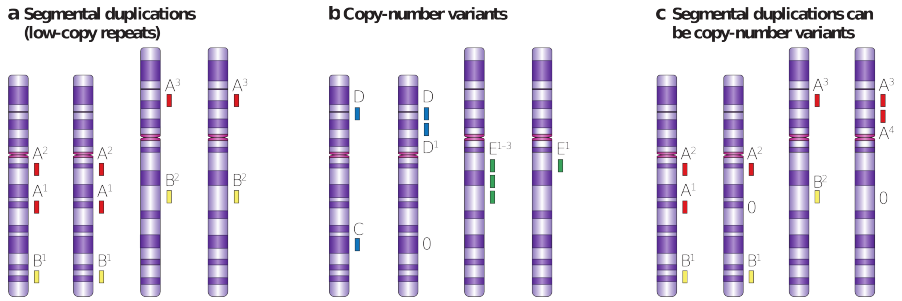

Figure 1.1: (a) Segmental duplications (SDs) can be intrachromosomal ($A^1 - A^2$) or interchromosomal ($A^1 - A^3$, $A^2 - A^3$ and $B^1 - B^2$). (b) Copy number variants (CNVs) are found in different copies with respect to a reference genome. The left chromosome in each pair represents a reference DNA sequence. Regions C and D are CNVs since they are deleted and duplicated, respectively, in the right chromosome, compared to the reference. CNV E is present in three copies in the reference and only once in the right chromosome. (c) The repetitive nature of SDs makes them have an increased tendency to vary in copy number. In this example, SDs A and B would also be categorized as CNVs. [Image taken from Feuk et al. (2006).]

Another set of definitions, by Campbell et al. (2011), states:

"Copy-number variants (CNVs) were originally defined as deletions or duplications greater than 1 kb in size. CNVs present at higher frequencies (>1%) in populations are distinguished as copy-number polymorphisms (CNPs). Both CNVs and CNPs are enriched in regions of the genome with

highly identical copies of paralogous sequence known as segmental duplications (SDs)."

This set of definitions is not very clear either. In any case, I have decided to come up with my own definition, or rather, my own extended description. I do not intend this definition to be better than the previously referenced, but rather to describe details that are important to take into account when referring to SDs and CNVs. Most of these details have become relevant in recent years given the data that has been produced by new technologies, showing that SDs and CNVs are a prevalent and diverse feature of many genomes.

An SD is a region of the genome, $\geq$1 kb in length, that has $\geq$90% identity (*i.e.* 10% or less base-pair mismatches, excluding gaps and insertions) when aligned to another region of the genome also $\geq$1 kb in length. SDs are therefore always classified as such if and only if they have a corresponding pair also classified as an SD and for which the alignment between both SDs is 90% or more identical. I will refer to the two copies that define an SD as an SD-pair. This does not imply, however, that there need be only two copies. There can very well be a 3-copy SD, say A, B, and C, where A-B, A-C and B-C are all classified as SD-pairs, but it can also be the case that A-B and A-C are identified as an SD-pair but B-C is not. There can also be fragmented copies, for example, a case in which A-B is an SD-pair, and C-D is another SD pair, but C happens to be a fragment of A, and indeed more highly complex regions dubbed *duplicons*. This nested and complex distribution of SDs throughout the genome can provide insight into their formation process (see Section 1.1.2).

SDs are sometimes considered to be fixed in the population (Kim et al., 2008). However, the perception on whether something is fixed or not in any population depends on the amount of individuals analyzed. In principle, by increasing sample size, an apparently fixed characteristic can be found to be a segregating one. So defining anything based on if it is fixed or not makes no sense if the sample size from which this status is inferred is not a defined number. It does makes sense, though, to define terms based on the reference genome. Fortunately, the academic community has agreed to have a reference genome for each species and even though a new reference genome appears periodically for each species, for every reference genome there is a corresponding annotation of SDs. Furthermore, reference genomes do not specifically include regions that are variable in number, that is, the reference genome is just a reference against which to evaluate if you have a difference or

not (SNP, deletion, insertion, inversion, translocation, etc.). It is, to say it clearly, an artificial genome constructed to serve as a reference. So, to complete my definition of SDs I will add that SDs can be defined for an individual or for a reference genome. They are at no point a population-based characteristic, but the result of a process (segmental duplication) that has taken place at least in one genome.

On the other hand, a CNV is a population-based concept. CNVs are precisely defined on the basis of being present in some, but not all, individuals of a population. If there is a reference genome (as is mostly the case for CNV studies) I would define CNV as a region of the genome for which one or more individuals have a different number of copies than that reported in the reference genome. In this sense, if an individual has a deletion or a duplication, or simply a different number of copies of any region with respect to the reference genome, this region can be considered a CNV. Since the reference genome is a haploid sequence, we could in fact consider a CNV a region which is present in a different number of copies in two chromosomes, even if the latter belong to the same individual (see Figure 1.1). One can expect CNVs to have a site-frequency spectrum similar to single nucleotide polymorphisms (SNPs) in the sense that most CNVs (supposing perfect detection power) will be private to one individual in the sample. However, there are many common CNVs, that is, regions that have a different number of copies in different individuals within a population (and therefore at least one of them has a different number from that of the reference genome). Common CNVs that appear in more than 1% of the sampled population are known as copy-number polymorphisms (CNPs), in analogy to SNPs being common single nucleotide variants.

To round up, SDs can be defined individually (or for a reference genome) while CNVs are always defined by comparing two or more individuals (where one of these acts as a reference, and in most cases it is the reference genome *per se*). In the case of humans, many regions of the reference genome annotated as SDs, according to the SegDups database at the University of California Santa Cruz (UCSC) Genome Browser (http://genome.ucsc.edu) are not fixed in the population. In fact, more than 50% of the nucleotides present within SDs overlap with CNVs (Cooper et al., 2007). Furthermore, duplicated regions of the genome have a 4 to 10-fold enrichment for copy-number variation compared to the genome average (Sebat et al., 2004; Sharp et al., 2005).

It would seem that with a detailed description it would be enough to *define*

SDs and CNVs, but there is another aspect of defining things in biology that is extremely relevant in this case, and I have so far not made explicit reference to it. Even though both definitions by Feuk et al. and Campbell et al. stated that CNVs are defined as regions larger than 1 kb, I made no mention of their length in my description above. The reason for this exclusion is that limits like these are usually included in definitions in the first place only due to technical reasons. As such, they are bound to change with time.

The definition of SDs was probably coined after the initial draft of the human genome became available in 2001 (International Human Genome Sequencing Consortium, 2001) given the BAC-based technology they were using at the time. This definition was maintained throughout the array-based technology which followed, that dominated the detection of structural variation. When read-depth approaches for next-generation sequencing technologies entered the CNV-detection stage, this limitation in terms of size was no longer needed and so, the 1 kb limit dropped to 50 bp (Baker, 2012). Regarding an upper limit in length for SDs and CNVs, if any, it is either 100 kb (Eichler, 2001) or 400 kb (Stankiewicz and Lupski, 2002).

1.1.2 Formation of segmental duplications and copy-number variants

As mentioned above, the location of SDs and CNVs is strongly correlated. The reason behind this correlation is that SDs are, in some instances, the birthplace of CNVs. The process through which SDs become CNVs is called non-allelic homologous recombination (NAHR) (Sharp et al., 2006). Given the high identity between SD-pairs, during recombination, non-allelic homologous pairing can occur between SD-pairs. The outcome of this recombination process will depend on the localization and relative orientation of the SD pair. As shown in Figure 1.2, NAHR can occur between SDs located on the same chromatid or on different chromatids. NAHR between SDs located within the same chromatid will result in an inversion if SDs are in reverse orientation, and in a deletion if they are in direct orientation, while NAHR between SDs located in different chromatids (homologous chromosomes or sister chromatids) will result in both a duplication and deletion event.

However, not all CNVs are originated by NAHR between SDs. Kim et al. (2008) found that less than 30% of their studied CNVs could have been formed by an SD-mediated mechanism. Rather, they hypothesized that random

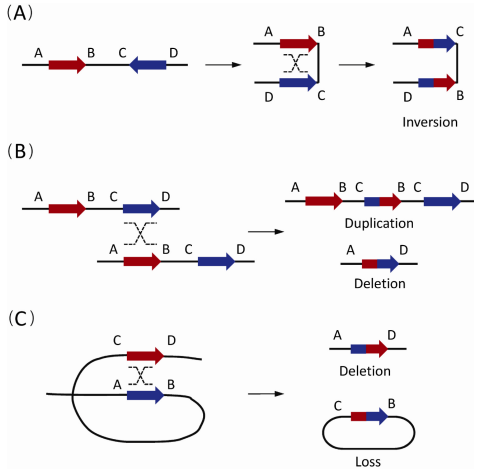

Figure 1.2: Possibilities of non-allelic homologous recombination (NAHR) between segmental duplications (SDs). SD-pairs are depicted as red and blue bold arrows with the orientation indicated by arrowheads. Capital letters refer to the flanking unique sequences. Dashed crossed lines represent a homologous recombination event. (A) The intrachromatid NAHR event between reversely oriented SDs can cause an inversion. (B) The interchromatid NAHR events between directly oriented SDs result in deletions and duplications. (C) The intrachromatid NAHR events between directly oriented SDs can generate deletions and ring-shaped DNA segments that will be lost in subsequent cell divisions. [Image taken and text adapted from Chen et al. (2014).]

breakage, followed by non-homologous end-joining (NHEJ) is one of the major mechanisms behind CNV formation. Although they did not present final evidence for such claims, they did find that 40% of CNV breakpoints contain microhomologies that can be a signature of NHEJ (Lieber et al., 2003). They also found that breakpoints lie in genomically unstable regions, consistent with NHEJ mechanisms.

In fact, there seems to have been not only different mechanisms of SD formation, but a change in the process of SD formation during the past 40 million years (Kim et al., 2008). Most of the old human SDs, measured by higher divergence between copies, seem to have been formed via NAHR between *Alu* elements. This high rate of formation fits in with the supposed peak of *Alu* activity around 40 million years ago (Kim et al., 2008), consistent

with earlier reports of highly significant enrichment of *Alu* elements near or within SD junctions, in particular for interspersed SDs separated by more than 1 Mb (Bailey et al., 2003).

Another process which seems to have been more common in the past is SD-mediated SD formation. According to Kim et al. (2008), the location of SDs follows a scale-free distribution that is consistent with a preferential attachment mechanism (Barabási and Albert, 1999), whereby *the rich get richer and the poor get poorer*, and in which there are a few places with many co-localized SDs and many regions with non-co-localized SDs. This implies that NAHR between SDs has been an important mechanism in SD formation. However, old SDs tend to co-localize with SDs of similar age more frequently than young SDs tend to co-localize (Kim et al., 2008), which points again at a change in SD formation mechanisms. This preferential attachment mechanism is consistent with the core-duplicon hypothesis (Jiang et al., 2007), which states that in primates, a set of 14 gene-rich regions were the focal point for the duplication expansion observed in African great apes (Marques-Bonet et al., 2009) and that this expansion occurred in a stepwise process (Marques-Bonet and Eichler, 2009).

1.1.3 Evolution of gene duplications

Despite the overarching recognition of the importance of gene duplication in the evolution of new gene functions, the mechanisms by which these new functions ultimately arise are not easily identifiable and there are still many open questions regarding the evolution and fate of gene duplications. There has been, however, an extensive amount of theoretical work describing the models by which this evolution might take place.

The most important model in historical terms is the *neofunctionalization* model, and it is in part so because, although not with the current name, this was the model assumed to be the most predominant by Ohno, who popularized the idea that most new gene functions arise from gene duplications (Ohno, 1970). The basic idea behind this model is that a single copy of a gene is enough to perform a particular function, so that once a gene duplication arises, there is a redundant copy that is under no purifying selection and is therefore free to accumulate mutations. In most cases, mutations will be slightly deleterious or might cause a loss of function, rendering the duplication non-functional, in a process commonly known as *pseudogenization*. In a few cases, one of these mutations might cause the gene to acquire a new function and, if beneficial,

positive selection acting on this mutation might favor its fixation. This is the neofunctionalization model that Ohno referred to.

There is a third common possibility (apart from pseudogenization and neofunctionalization), proposed by Force et al. (1999), described by the *duplication-degeneration-complementation* (DDC) model that consists in a say, division of labor schema, in which once duplicated, relaxed purifying selection can cause the duplicates to accumulate damaging mutations that affect the function of the gene, so that one copy alone cannot perform the original function. The affected copies might then further divide their function through changes in regulation or by accumulating further mutations in a process called subfunctionalization. There are other cases of subfunctionalization, such as the model proposed by Hughes (1994) and then named the *escape-from-adaptive-conflict* (EAC) model (Des Marais and Rausher, 2008). In this model, the original single-copy gene performed several functions but was unable to perfect any of them because of selective constraints. Once duplicated, each copy can specialize or subfunctionalize and escape from the now nonexistent adaptive conflict.

There is an additional common outcome, which was in fact also mentioned by Ohno (1970) and is referred to as the *selection for more of the same* or *selection for increased gene dosage* model. Basically, if a gene product is such that more of the same is better, then selection might act to maintain both copies with their original function, now doubled.

Aside from pseudogenization, neofunctionalization, subfunctionalization and increased gene dosage, other models have been proposed and surely more models will continue to appear. In order to systematize the existing and future possible models, Innan and Kondrashov (2010) came up with a classification based primarily not in the outcome of the duplication, but in the manner in which the duplication rises in frequency until fixation. The pseudogenization, neofunctionalization, and subfunctionalization models all assume that the new copy is fixed in the population by drift. In the case of beneficial increase of gene dosage, the duplication is fixed because there is positive selection acting on the presence of the duplication itself, which can be due to at least two reasons according to Innan and Kondrashov (2010): the masking of deleterious mutations, or the opportunity for the immediate emergence of a new function. Another possibility is that the duplication occurs in a gene in which there already exists variation in the population. In this case, several outcomes are

possible, from an adaptive radiation, to the resolution of balancing selection via a *permanent heterozygote* and including multiallelic diversifying selection. A final possibility is one in which the fixation of the duplication is a precondition, in the sense that it occurs as a by-product of large-scale events such as a whole-genome duplication. In this model, called the *dosage balance* model, if two or more genes that become duplicated have an optimum dosage that is dependent on the dosage of the each other, they will tend to be either maintained together or eliminated together because of negative selection against dosage imbalance (Papp et al., 2003). (For details of these models, refer to Innan and Kondrashov (2010)).

Now, evidence supporting one model over the other is of course very case-specific. However, there are important unanswered questions that are a matter of hot debate, for example, which are the processes driving the retention of mammalian duplicate genes. Lately, evidence has accumulated suggesting that after duplication, there is an acceleration in the rates of molecular evolution restricted to only one of the copies, in agreement with the neofunctionalization model and positive selection acting on one of the copies (Pegueroles et al., 2013; Pich I Roselló and Kondrashov, 2014). It appears that the functional divergence of duplicates might be associated with copies acquiring diverse tissue-specific biological roles (Assis and Bachtrog, 2015). Alternatively, Lan and Pritchard (2016) gather evidence consistent with the *dosage-sharing* hypothesis, whereby, in order to match the expression levels of the single-copy genes, most young duplicates are down-regulated and this allows for their initial survival. Once saved from rapid loss, a slower functional adaptation enables them to acquire novel gene functions and therefore, their long-term preservation.

Entia non sunt multiplicanda sine necessitate.

Attributed to William of Ockham

If the neutral or nearly neutral mutation is being produced in each generation at a much higher rate than has been considered before, then we must recognize the great importance of random genetic drift due to finite population number in forming the genetic structure of biological populations. [...] To emphasize the founder principle but deny the importance of random genetic drift due to finite population number is, in my opinion, rather similar to assuming a great flood to explain the formation of deep valleys but rejecting a gradual but long lasting process of erosion by water as insufficient to produce such a result.

Motoo Kimura, 1968

1.2 Modeling the neutral evolution of duplications

Given the evolutionary relevance of gene duplications and SDs, and the diverse selective scenarios that can drive their fate, modeling their evolution seems like an interesting and important endeavor. However, prior to the incorporation of selection, it is fundamental to figure out the factors that affect the neutral evolution of duplications, since without a neutral model, almost anything that could be said about their evolution would be lacking a solid ground.

In 1968, Kimura published a beautiful paper (Kimura, 1968) that revolutionized the fields of population genetics and evolutionary biology. Kimura gave simple arguments and presented clear evidence supporting the important contribution of neutral or nearly neutral mutations to evolutionary change.

The fact that randomness could introduce polymorphism in populations had been recognized even by Darwin himself in *On the Origin of Species* (Darwin, 1859):

"Variations neither useful nor injurious would not be affected by natural

selection, and would be left a fluctuating element, as perhaps we see in the species called polymorphic."

That these polymorphisms could in fact become ultimately fixed in the populations and that random genetic drift was important in the fixation of neutral mutation in particular for small populations (or highly inbreeding ones) was widely accepted and formalized mathematically by Fisher (1930), Wright (1931), and Haldane (1932). However, the evolutionary relevance of these neutral mutations compared to those fixed by selection was one of the main points of disagreement between Fisher and Wright. Charlesworth and Charlesworth (2016) recall an anecdote from Lewontin from a meeting in 1971:

"Dick's 'sermon' began by stating that 'our field is divided into two warring sects. These are the adherents to the Epistle of St. Sewall to the Japanese, who believe that *the race is not to the swiftest nor the battle to the strong... but time and chance happeneth to the all*, and the followers of St. Ronald, who believe that *many are called but few are chosen*'." (italics in original)

In my opinion, when Kimura presented his *neutral theory of molecular evolution* (Kimura, 1968), he was in some way bringing *Ockham's Razor* into the selection versus drift dilemma: random genetic drift is incredibly more simple and straightforward than selection, so, if random drift can explain most genetic differences between species, there is no need for selection. From Kimura onward, the field of population genetics and genomics has by no means stopped its search for signatures of selection in extant species, in particular, in humans. It has however, modified the search strategy by looking for signatures of selection that deviate from the neutral expectations. As such, the development of accurate models of neutral evolution, incorporating demography, for example, in the case of human population genetics, has been a very important field of research. However, to date, there is no accurate neutral model for the evolution of duplications.

Furthermore, despite important theoretical work during the 1980's on the neutral evolution of gene families by Ohta (1982, 1983), Nagylaki (1983, 1984a,b), and others, this work was not incorporated as a framework from which to analyze large-scale data on duplications once these became available. An example to illustrate my point is that in 2002, Bailey and colleagues from Eichler's group, presented the most thorough analysis at that time of recent SDs

in the human genome (Bailey et al., 2002). The originality behind their novel method of SD detection and the importance of this contribution (and many others from that same group that followed) is undeniable, and the knowledge of the field of many of the paper's authors is indisputable. However, within the main text of the paper came the following sentence:

"Because there is no reason to expect that polymorphic variation is increased within duplicated regions, the approximate doubling of SNP density suggests that roughly one of two SNPs is, in fact, a paralogous sequence variant rather than an allele."(Bailey et al., 2002)

I will not dispute the authors' conclusions following this argument, but the first part of the sentence is wrong. As Ohta (1982) and Nagylaki (1983) had clearly showed analytically, variation present in duplicated regions can be up to twice as high compared to a single-copy region even under strictly neutral evolution. Even though this sentence could have been simply overlooked, it could also be possible that the authors were unaware of this important body of work (Hurles, 2002).

Be it as it may, during my PhD, and as I will try report in this thesis, I have made an effort to put together the theoretical framework developed not only by Ohta and Nagylaki, but very importantly by Innan and collaborators (reviewed in (Innan, 2009), with analyses via simulations of the variation and *linkage disequilibrium* (LD) patterns present within and between duplications, and with the molecular mechanisms underlying the evolution of duplicated regions of the genome. Although one of the initial aims of my PhD was to incorporate selection in the evolution of duplicates, the neutral scenario proved to be much more complicated and indeed, unknown. Being so, I concentrated on exploring the neutral model of evolution of duplications, focusing in particular on a mechanism known as *interlocus gene conversion*.

1.2.1 Interlocus gene conversion

Most of the models describing the evolution of duplicated genomic regions assume that each of these regions evolves independently, in the sense that the mutations that appear in each region are not influenced by mutations present in the other region. However, gene duplications, in particular small multigene families, have been known to evolve in a non-independent manner called

concerted evolution since the 1980's (Baltimore, 1981; Nagylaki and Petes, 1982; Ohta, 1982).

The main mechanism responsible for concerted evolution of gene duplications is *interlocus gene conversion* (IGC). IGC has the unlucky quality (common in biology) of being addressed by many names and none of them has been successful enough to dominate the literature. IGC is also referred to as non-allelic, ectopic, interchromosomal, interparalog and intergenic gene conversion, but I believe that *interlocus* is the best adjective to describe this recombination process between two sequences in two different loci, so I will use IGC to refer to it throughout this thesis. IGC is believed to be caused by the same molecular mechanisms underlying *allelic gene conversion*, which occurs within a single locus between homologous chromosomes during mitosis and meiosis (Hastings, 2010). Interallelic gene conversion is sometimes used as a term to refer to allelic gene conversion occurring between highly differentiated alleles located in the same locus, which is possible in highly variable regions of the genome (Chen et al., 2007).

IGC can be best described as a *copy-paste* event between homologous regions within paralogous copies (Innan, 2009). Just to clarify, paralogous copies are those that were created from a duplication event and are found in a single genome (while orthologous copies were created from a speciation event and are therefore found in genomes of two different species), either within a single chromosome or in a different chromosome. Since IGC occurs during meiosis or mitosis, when chromosomes are duplicated, IGC can happen between copies on a single chromatid, on sister chromatids or on homologous chromosomes, in the case of intrachromosomal IGC (Ohta, 1983; Nagylaki, 1984b), or between copies on different chromosomes in the case of interchromosomal IGC (Ohta, 1983; Nagylaki, 1984a).

In an IGC event, a limited tract of DNA within a duplicated segment is effectively replaced by the sequence present in a highly homologous (*i.e.* with high sequence similarity) copy found elsewhere in the genome (see Figure 1.3). In a population genetics framework, the most widely used gene conversion model considers that a gene conversion event is initiated at a random position within the duplicated segment with rate g (Teshima and Innan, 2004). From this position, a gene conversion tract extends in either a 5' or 3' direction, and this elongation terminates at any position with a fixed probability q that does not depend on the current length of the gene conversion tract (Wiuf and Hein,

Figure 1.3: Interlocus gene conversion (IGC) is a non-reciprocal recombination process, usually described as a *copy-paste* event, in which a short tract of DNA, ranging from a few to thousands of base pairs, is transferred between paralogous regions. IGC can occur in either direction, thereby shuffling DNA variation between paralogous duplicated regions and driving the concerted evolution of duplicates. [Image taken and text adapted from Innan and Kondrashov (2010).]

2000). Given this model, the length of gene conversion tracts will follow a geometric distribution (whose continuous analogue is an exponential distribution) with parameter q, giving an average tract length of $\lambda = 1/q$. Then, the per-site gene conversion rate will be $c = g\lambda$, a product of the initiation rate of an event and the average length of the gene conversion tract.

Measurements of IGC rates have been performed in several species. Rates are usually given in very wide estimates such as $\sim 10^{-10}$ to $\sim 10^{-3}$ IGC per site per generation rate in *S. cerevisiae* (Mansai et al., 2011), and $\sim 10^{-4}$ to $\sim 10^{-3}$ in humans (Chen et al., 2007). In terms of the percentage of paralogous genes that undergo IGC, estimates are also variable, and tend to lie below 20%, for example, 2% in *C. elegans* (Semple and Wolfe, 1999), 7-13% in yeast (Drouin, 2002; Casola et al., 2012), 7-14% in *Drosophila* (Casola et al., 2010), and 8-19% in mammals (McGrath et al., 2009), although as high as 25.4% in SDs in humans (Dumont and Eichler, 2013), explaining at least 2.7% of single nucleotide variants within SDs (Dumont, 2015). Average gene conversion tract lengths range from 55 to 290 bp (Jeffreys and May, 2004; McGrath et al., 2009) in humans, and are around 100 bp por yeast and rodents (Mansai et al., 2011). A description of the dependence of IGC rates on factors such as location, orientation and similarity of duplications, as well as more details pertaining to the estimates of IGC rates and gene conversion tract lengths for several species can be found in the Appendix (Files S1 and S2).

Details of the mechanistic process through which IGC happens will be explained briefly in section 1.3. In section 3.1, I present a study in which variation and LD patterns within and between duplications are analyzed for a wide range of IGC rates in different crossover scenarios.

1.2.2 Measuring IGC

There are basically two approaches to measure IGC rates, the empirical approach and the evolutionary approach (Mansai et al., 2011). The empirical approach is in principle much more powerful given the flexibility of experimental setups but it is very limited to the species to which it can be applied (mainly, model organisms). This approach consists in mutation accumulation experiments performed in transgenic systems. In these systems, strains or cell lines are artificially modified in order to obtain highly homologous DNA sequences with interspersed markers along their length. One or more of these markers can be used as a reporter marker by associating it to a phenotypic effect such as a radioactive signal (Lichten et al., 1987). Strains can be easily screened for converted reporter makers and flanking markers can be checked for conversion. Through this approach, gene conversion rates can be studied in a straightforward manner but are highly dependent on the experimental setup (Mansai et al., 2011) (see Appendix, File S1). Tract lengths can also be measured, but for each conversion event only minimum and maximum tract lengths can be determined, and these will depend on the separation between markers.

The evolutionary approach is in principle applicable to any species for which one can obtain sequence data of duplicated regions from a population. Note that distinguishing between paralogous sequences is not always trivial and thus, as will be presented in section 3.2, many studies confound paralogs given their high sequence similarity. Therefore, obtaining clearly differentiated sequence data from paralogs might prove to be quite complicated. If this is accomplished experimentally, sequences from a sample of the population can be obtained, and by studying the pattern of polymorphisms (mainly SNPs) in the duplications one can in principle detect the footprints of past gene conversion events. The main limitation of this approach is that it relies heavily on the underlying population genetic models that determine what is the pattern of polymorphism expected in the absence of gene conversion (Mansai et al., 2011).

There are several algorithms and software implemented to measure gene conversion rates and gene conversion tract lengths. As evidenced by Mansai and Innan (2010), the most commonly used software, GENECONV (Sawyer, 1989) has very low power to detect gene conversion events if gene conversion rates are high. This happens because GENECONV looks for unusually long tracts of homozygosity. So, if gene conversion rates are high, these homozygous tracts

are seen as normal, even though the most parsimonious explanation for them is gene conversion. Another software that can detect gene conversion events between paralogous sequences is DnaSP (Librado and Rozas, 2009), that implements part of the algorithm by Betrán et al. (1997). This algorithm can estimate true tract lengths from observed tract lengths, but this feature is not implemented in DnaSP (Librado and Rozas, 2009). This algorithm considers *informative sites* and assigns each site a value depending basically on their frequency and their type, which can be *fixed* (*i.e.* a fixed difference between paralogs), *specific* (*i.e.* polymorphic only in one paralog) or *shared* (*i.e.* having the same polymorphism in both paralogs) (see Figure 1.4).

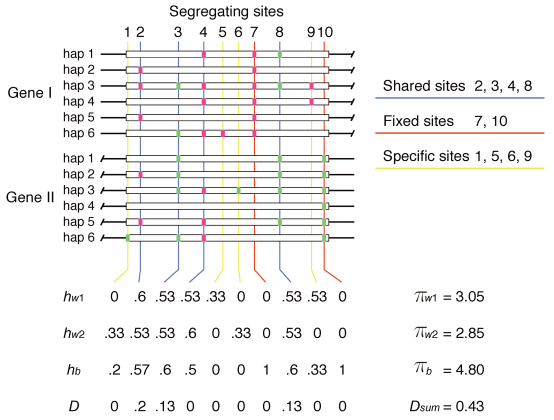

Figure 1.4: Example of variation present at two paralogous genes (Gene I and Gene II). A segregating site can be specific, fixed or shared, depending on the frequency of variants (pink and green boxes) in each gene. It will be specific to one gene if it is polymorphic (present in some but not all of the haplotypes) in only one of the genes (yellow line) and absent from the other gene. It will be a fixed site if it is present in all haplotypes of one gene and absent from all haplotypes of the other gene (red line). It will be a shared site if it is polymorphic in both genes (blue line). Heterozygozity within genes (h_{w1} and h_{w2}), heterozygozity between genes (h_b) and linkage disequilibrium between paralogous sites (D, and D_m in main text). From these values, average pairwise differences within genes (π_{w1} and π_{w2}), average pairwise differences between genes (π_b) and linkage disequilibrium between paralogs (D_{sum}) are calculated. [Image taken from Innan (2004).]

Both GENECONV and the algorithm of Betrán et al. (1997) were designed

to detect allelic gene conversion. The degree of identity between homologous sequences on the same locus is expected to be very high, and the algorithms are *tuned* accordingly. Paralogous sequences, however, usually have a lower degree of identity than homologous sequences. Although this could in principle imply that we could have more power to detect gene conversion, it does not mean that these specific algorithms will do a better job at it.

Another evolutionary approach, which has been used by Innan (2002, 2003b) is to consider exclusively shared polymorphic sites since they are strong candidates of being the result of gene conversion events (see Figure 1.4). Although exclusively considering shared polymorphic sites limits the power to detect gene conversion events if gene conversion rates are low, it performs much better than GENECONV for high gene conversion rates. Additional caveats are that the direction of gene conversion might not be clear if an ancestor reference is not provided, and gene conversion tract lengths might be difficult to identify. Also, private polymorphisms, in particular those in which the low-frequency variant is the same as the fixed variant in the other paralog, are putative gene conversion events that are ignored when only considering shared polymorphisms. Although they could very well be caused by point mutation, there might be cases in which this special kind of private polymorphisms appear contiguously in the same sequence. These cases most likely represent gene conversion events but would nevertheless still be ignored with this approach. Furthermore, using only shared polymorphic sites can be highly underpowered in cases in which there is a strong directionality (donor-acceptor) bias.

There is a final evolutionary approach which is based on an infinite-site model of a small multigene family undergoing mutation, IGC and crossover, developed by Innan (2003b) which can be used with population sequence data. By experimentally measuring variation within one of the copies of the duplication, variation between both copies of the duplication, and LD between all paralogous sites at equilibrium, one can obtain estimates for mutation, IGC and crossover rates.

Following previous work by Ohta (1983) and himself (Innan, 2002) (see Appendix, File S3, for details), Innan (2003b) obtained analytical expectations for the average pairwise differences within loci $E(\pi_w)$ and between loci $E(\pi_b)$:

$$E(\pi_w) = \frac{2\Theta\,(2C + R + 2)}{4C + R + 2} \tag{1.1}$$

$$E(\pi_b) = \frac{\Theta\left(4C^2 + 4C + 2CR + R + 2\right)}{C\left(4C + R + 2\right)} \tag{1.2}$$

Additionally, he defined D_m as a measure of the amount of LD between two paralogous sites, which is also a proxy of IGC since it is based also on the number and distribution of shared polymorphic sites. He defined $D_m = \frac{n_{AA}n_{aa} - n_{Aa}n_{aA}}{n(n-1)}$, where n_{xy} represents the number of samples with nucleotide x at site m in one of the paralogous copies and nucleotide y at site m in the other paralog. He then defined D_{sum} as the sum of D_m over all L sites along the paralogs:

$$D_{sum} = \sum_{n=1}^{L} D_m \tag{1.3}$$

His expectation for D_{sum} for an infinite-site model is:

$$E(D_{sum}) = \frac{2\Theta C}{4C + R + 2} \tag{1.4}$$

From equations 1.1, 1.2 and 1.4, Innan (2003) derived equilibrium values for the population-scaled rates of mutation Θ, IGC (C) and allelic crossover (R), where $\Theta = \theta L = 4N\mu L$, $C = 4Nc = 4Ng\lambda$, $R = 4Nr$ (and where μ is the per site per generation point mutation rate and r is the per generation crossover rate between paralogs):

$$\hat{\Theta} = \frac{\pi_w + 2D_{sum}}{2} \tag{1.5}$$

$$\hat{C} = \frac{\pi_w + 2D_{sum}}{2\left(\pi_b - \pi_w\right)} \tag{1.6}$$

$$\hat{R} = \frac{{\pi_w}^2 + 4{D_{sum}}^2 - 4\pi_b D_{sum}}{2\left(\pi_b - \pi_w\right) D_{sum}} \tag{1.7}$$

So, if one assumes that the sequenced sample is under mutation-conversion-drift equilibrium, by calculating π_w, π_b, and D_{sum} from the population sequence data (as shown in Figure 1.4), one can obtain equilibrium values for mutation, IGC and crossover rates.

1.2.3 Effect of interlocus gene conversion in the evolution of duplications

The most important effect of IGC is that *paralogs* (*i.e.* paralogous copies) diverge less than what it would be expected if they were evolving independently. After a duplication appears, divergence between copies increases but may reach an equilibrium determined by the independent mutational input (which increases divergence) and the homogenization of IGC (which decreases divergence). As illustrated in Figure 1.5, divergence will fluctuate around its equilibrium value for some time. However, since IGC is dependent on a certain degree of sequence similarity between copies, if the divergence between copies goes above a certain threshold, IGC will cease to occur, effectively ending concerted evolution.

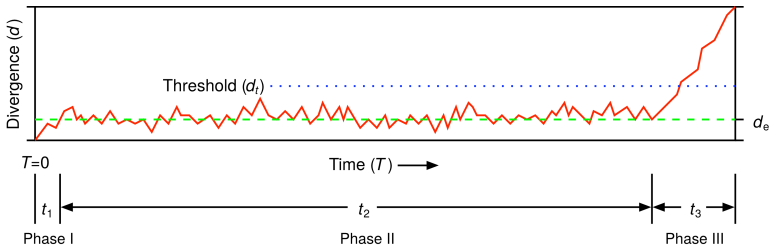

Figure 1.5: Simulation example of divergence (d) between duplicated genes through time (T). At $T = 0$, a duplication appears. Divergence between duplicates increases during Phase I until it reaches an equilibrium (d_e). Interlocus gene conversion (IGC) between duplications maintains the low levels of divergence below the threshold d_t during Phase II, or concerted-evolution phase. Under random drift, fluctuations in variation can be very high so there might come a moment in which the divergence between duplicates is higher that the threshold that limits the IGC process. If d goes above d_t, Phase II terminates and duplicates diverge from each other under the molecular clock, which involves independence between duplications (Phase III). [Image taken from Innan (2009), in turn modified from Teshima and Innan (2004).]

The duration of the concerted-evolution phase will be determined by several factors (Teshima and Innan, 2004; Innan, 2009). Under a neutral scenario it will primarily depend on the difference between the equilibrium and threshold values of divergence, since random fluctuations around the equilibrium divergence are more likely to reach the threshold if the difference between them is smaller. In principle, the higher the IGC rate, the longer the duration of

concerted evolution. It will also depend on the width of the variations around the equilibrium divergence which will be smaller if λ is smaller (for a fixed gene conversion rate). All the above considers a neutral scenario without selective pressures acting on duplications or on their genic content. However, the consequences of this homogenization can be different depending on the context in which it happens. Effects of homogenization can be either beneficial or adverse and can make the preservation of the duplication more or less likely (Innan and Kondrashov, 2010).

For example, frequent IGC might be beneficial and promote the maintenance of a duplication under the increased gene dosage or dosage-balance models since it would keep a high sequence identity between paralogs and in principle preserve two functional copies. Accordingly, in yeast, there appears to be a strong positive correlation between gene expression and the duration of concerted evolution (Sugino and Innan, 2006), suggesting that genes whose products are on high demand, such as ribosomal genes, are more prone to undergo long-term concerted evolution (Innan, 2009). Although not due to the same reasons, that is, even if there is no intrinsic beneficial effect of having increased dosage, duplicated genes under strong selective constraint might also undergo long-term concerted evolution. In this case, IGC might help to erase non-synonymous mutations from both copies (Teshima and Innan, 2004).

On the other hand, there might be selection acting to stop IGC between paralogs. This would be the case under the neofunctionalization model, in which one of the copies has acquired a novel function. IGC between copies would potentially eliminate either the original or the novel function and would therefore be disfavored. Selection against IGC would effectively reduce IGC rates, allowing for more mutations to accumulate independently in each copy which would eventually impede IGC between copies (Innan, 2009). A pattern of polymorphism consistent with selection favoring the accumulation of mutations and acting against IGC around the target site of selection (Teshima and Innan, 2008) has been measured in the human RH blood-type genes, RHCE and RHD (Innan, 2003a). A similar pattern has been observed in the human red-green opsin genes (Zhao et al., 1998; Verrelli and Tishkoff, 2004), in the pancreatic ribonuclease genes in the colobine monkeys (Schienman et al., 2006), and in heat-shock protein genes in yeast (Takuno and Innan, 2009).

Similarly, under the pseudogenization model, in which one of the copies has

acquired a loss-of-function mutation, IGC from the pseudogenized copy to the original one would have strong adverse effects. Indeed, in humans, this mechanism is the cause of several genetic diseases (Bischof et al., 2006) (see Table 1.1) and selection should act to stop IGC. Along the same lines, a recent paper reports a *de novo* IGC between a reduced-function green opsin allele and a red opsin gene (Buena-Atienza et al., 2016) that causes blue cone monochromacy. This case is interesting because although human opsin genes are prone to copy-number variation (Macke and Nathans, 1997), they are accompanied by a locus control-like element that allows the transcription of only a single (the most proximal) copy of the green opsin gene (Winderickx et al., 1992). This allows loss-of function variants, reduced-function variants, and green-red hybrid genes to persist since they do not have a deleterious effect unless they are in a position in the gene array that allows their expression in the retinal cone cells. So although in this case IGC might be deleterious, the fact that it happens in group of genes that has allowed humans to have trichromatic vision, illustrates the possibility for IGC to create new combinations of mutations, which can be exploited especially in a scenario in which allelic diversity is favored, such as in the major histocompatibility complex (Ohta, 1991; Takuno et al., 2008).

Finally, the rate with which beneficial mutations can be fixed or deleterious mutations eliminated from both copies might be increased due to IGC. Mano and Innan (2008) found that IGC in multigene families increases the effective population size in such a way that weak selection acts more efficiently and therefore accelerates rates of evolution.

There are, thus, multiple scenarios in which selection could be acting upon duplications. However, searching for signatures of selection in duplications is not straightforward since it requires specific methods that take into account the concerted evolution of duplications (Teshima and Innan, 2008; Osada and Innan, 2008). As such, genome-wide scans for selection must treat duplicated regions of the genome adequately.

1.2.4 Duplications in genome-wide scans for selection

In 1973, Lewontin and Krakauer (1973) conducted a seminal study in which they estimated effective inbreeding coefficients across many loci in humans. Since breeding structure should affect all loci in the same way, they argued that a significant heterogeneity in apparent inbreeding coefficients across loci would

Disease/phenotype	Donor gene	Acceptor gene	Chromosomal location
Atypical haemolytic syndrome	*CFHR1**	*CFH0*	1q32
Congenital adrenal hyperplasia	*CYP21A1P*	*CYP21A2*	6p21.3
Syndrome of corticosterone mehtyloxidase II deficiency	*CYP11B1**	*CYP11B2*	8q21-q22
Increased 18-hydroxycortisol production	*CYP11B1**	*CYP11B2*	8q21-q22
Autosomal dominant cataract	*CRYBP1*	*CRYBB2*	22q11.2-q12.1
Neural tube defects	*FOLR1P*	*FOLR1*	11q13.3-q14.1
Gaucher disease	*GBAP*	*GBA*	1q21
Short stature	*GH2**	*GH1*	17q22-q24
Mild microcytosis	*HBB**	*HBD*	11p15.5
Hereditary persistence of fetal haemoglobin	*HBG2*	*HBG1*	11p15.5
Agammaglobunlinaemia	*IGLL3*	*IGLL1*	22q11.23
Chronic granulomatous disease	*NCF1B* or *NCF1C*	*NCF1*	7q11.23
Blue cone monochromacy	*OPN1MW**	*OPN1LW*	Xq28
Autosomal dominant polycystic kidney disease	?	*PKD1*	16p13.3
Chronic pancreatitis	*PRSS2**	*PRSS1*	7q35
Shwachman-Bodian-Diamon syndrome	*SBDSP*	*SBDS*	7q11.22
Spinal muscular atrophy	*SMN2*	*SMN1*	5q13.2
von Willebrand disease	*VWFP*	*VWF*	22q11.22-q11.23/12p13.3
Congenital adrenal hyperplasia	*CYP21A1P*	*CYP21A2*	6p21.3
Increased CYP3A7 expression in adult liver and intestine	*CYP3A4*	*CYP3A7*	7q21-q22.1
Novel St glycophorin	*GYPE*	*GYPA*	4q28-q31
Microcytosis	*HBA2*	*HBA1*	16p13.3
Agammaglobulinemia	*IGLL3*	*IGLL1*	22q11.23
Sec1-FUT2-Sec1 hybrid allele	*FUT2*	*Sec1*	19q13.3
Atypical hemolytic uremic syndrome	*CR1L*	*CD46*	1q32
Pachyonychia congenita type 2	*KRT17P3*	*KRT17*	17q21.2
X-linked cone and cone-rod dystrophies	*OPN1M*	*OPN1LW*	Xq28

Table 1.1: Interlocus gene conversion events that cause human inherited disease. Functional donor genes are indicated by *, showing cases of pseudogene-mediated gene conversion events linked to disease. [Table adapted from Chen et al. (2007) and Chen et al. (2010).]

indicate evidence for selection (Lewontin and Krakauer, 1973). This study has served as an inspiration for what has been a prevalent field in population genetics for the past fifteen years: scanning genomes in search for deviations from expected patterns of variation (Haasl and Payseur, 2016). To that means, they rely on test statistics designed to identify regions with alterations in their expected levels of variation, site frequency spectra, levels of linkage disequilibrium and/or of interpopulation or interspecies divergence (Jensen et al., 2016).

There are many factors that are known to affect genome-wide scans for positive selection. Some factors are mainly technical, such as low-quality genome assemblies used as reference to map reads (Mallick et al., 2009; Manel et al., 2016), or different performance of aligners (Markova-Raina and Petrov, 2011), or ascertainment biases due to searching for unusual loci prior to selecting outliers (Thornton, 2007), while others have a more theoretical basis such as ignoring demographic history (Teshima et al., 2006), population structure (Excoffier et al., 2009), or background selection (Stephan, 2010).

These and additional factors contribute to there being abundant false positives in genome-wide scans for positive selection (see Haasl and Payseur (2016) and Jensen et al. (2016)).

Among these additional factors is the presence of duplications and copy-number variants in the genome. IGC and crossover between duplications generates alterations in variation and LD levels, and generates distortions in the site frequency spectrum. If duplicated regions are included in genome-wide scans they can render false-positive signals for selection. This fact has been addressed previously by Innan (2003b) and Thornton (2007) and indeed, most genome-wide studies mask the genome for duplications, eliminating these regions from the selection scans.

In section 3.2, I will present a study that explores the way in which test statistics designed to detect deviations from neutrality in single-copy regions are confounded by the effects of IGC and crossover when applied to duplications.

Recombination is very important stuff, by the way. At some point I will tell you that understanding recombination was actually the origin of the Human Genome Project.

Eric Lander, Fundamentals of Biology class, MIT 2012

1.3 Molecular mechanisms of interlocus gene conversion

Life is robust, and great part of this robustness relies on the cellular machinery that minimizes damage to DNA. One of the most common types of DNA damage are *double-strand breaks* (DSBs). If not repaired, a DSB will trigger a response within the cell, arresting the cell cycle or even causing apoptosis. Misrepair of DSBs can cause large-scale chromosomal changes, such as translocations, chromosomal fusions and deletions, causing genome instability (Shrivastav et al., 2008). To avoid these outcomes, cells have evolved signaling networks that detect DSBs and that activate DNA-repair pathways. These can be classified into two big groups: *homologous recombination* (HR) and *non-homologous end-joining* (NHEJ) (Haber, 2000; Shrivastav et al., 2008; Lieber, 2010). How the cell determines which pathway to use is an active area of investigation. It is thought to be determined in part by the causes of the DSB (Lieber, 2010), although there is also evidence for operational reasons, such as the presence of homologous regions close to the DSB (Sonoda et al., 2006).

There are four main types of HR pathways: *double-strand break repair* (DSBR), *synthesis-dependent strand annealing* (SDSA), *single-strand annealing* (SSA), and *break-induced repair* (BIR) (Sung and Klein, 2006). This classification has changed over time and most of the molecular mechanisms governing the choice of pathway are still unknown. However, substantial knowledge has been acquired over the past 30 years, about both the proteins involved in each one of these pathways and the way in which DSBs are repaired within each one (Haber, 2000; Krogh and Symington, 2004; Sung and Klein, 2006; Shrivastav et al., 2008). Even though all four types of HR are possible gene conversion pathways, of particular relevance for gene conversion are DSBR and SDSA.

1.3.1 Gene conversion as a consequence of DNA repair

The DSBR model was proposed by Szostak and collaborators (Szostak et al., 1983) after a series of observations regarding branch migration, mismatch correction and initiation placed several constraints on the then accepted model of recombination, the Meselson-Radding model (Meselson and Radding, 1975). It is curious that, at the time, the most important contribution of the DSBR model was not the model itself, but that it suggested that meiotic recombination is initiated, not by a single-strand nick as in previous models, but with a double-strand break.

The DSBR model is also commonly referred to as the *double Holliday junction* (dHJ) model. The DSBR mechanism is initiated by a DSB, which is processed by resection of the 5' ends, of what is called the *recipient* sequence. The 3' overhang invades the intact *donor* chromosome, forming a structure known as a *D-loop*. The donor sequence acts as a primer for the initiation of new DNA synthesis. Then, the end of the newly synthesized segment binds to the complementary 3' end from the other side of the break undergoing a second end capture, which leads to the formation of a dHJ after which branch-migration of the *Holliday junctions* (HJs) can proceed. Theoretically, each HJ can be cleaved by cutting either two crossed strands or two non-crossed strands (Pâques and Haber, 1999). If both HJs are cleaved in the same manner, there will be an exchange of genetic material between homologous chromosomes, but it will be limited to the region in between the HJs. This resolution is referred to as a non-crossover (NCO) resolution and basically consists in an allelic gene conversion event. The alternative is that each HJ is cleaved differently, which will result in the exchange of flanking markers, or what is commonly referred to as a crossover (CO) resolution (see Figure 1.6). Although for many years, CO and NCO resolutions were accepted outcomes of the DSBR model, evidence has accumulated (McMahill et al., 2007) indicating that once the dHJ is formed, the resolution will be a CO, while NCO resolutions are rather the outcome of another pathway, namely, SDSA. One final detail is that independently of the resolution, the DSBR model will always produce a region that will putatively be a gene conversion (I say putatively because it will always be dependent on the existence of differences between homologs). The SDSA pathway was proposed by Resnick (1976) although its current name was coined later (Nassif et al., 1994). SDSA starts in the same way as the DSBR pathway. After a DSB, 5' to 3' resection occurs, followed by *Rad51*-dependent strand invasion of the 3' end

and D-loop formation. After repair synthesis, accompanied by D-loop extension, the newly synthesized strand is dissociated from the donor sequence. After dissociation, ligation to the complementary 3' end from the other side of the break occurs, resulting in a NCO product (see Figure 1.6).

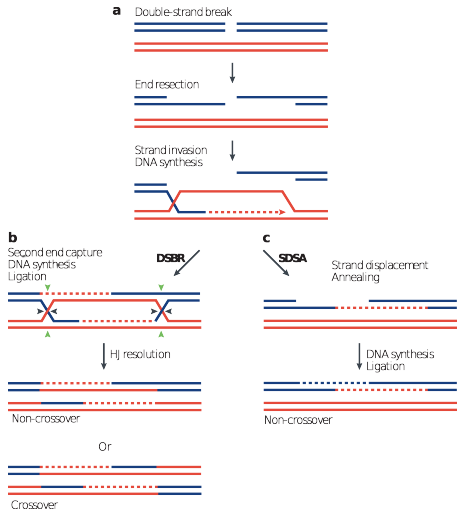

Figure 1.6: Double-strand breaks (DSBs) can be repaired by several pathways of homologous recombination, including double-strand break repair (DSBR) and synthesis-dependent strand annealing (SDSA). (a) In both pathways, repair is initiated by resection of 5' ends, revealing 3' single-stranded DNA overhangs. Strand invasion by these 3' overhangs into a homologous sequence is followed by DNA synthesis at the invading end. (b) After strand invasion and synthesis, the second DSB end can be captured to form an intermediate with two Holliday junctions (HJs). After gap-repair DNA synthesis and ligation, the structure is resolved at the HJs in a non-crossover (black arrow heads at both HJs) or crossover mode (green arrow heads at one HJ and black arrow heads at the other HJ). (c) Alternatively, the reaction can proceed to SDSA by strand displacement, annealing of the extended single-strand end to the single-stranded DNA on the other break end, followed by gap-filling DNA synthesis and ligation. The repair product from SDSA is always non-crossover. [Image taken and text adapted from Sung and Klein (2006).]

Both the DSBR and SDSA pathways involve the formation of *heteroduplex DNA* (hDNA), which can include mismatches. The correction of these mismatches will result in a gene conversion event and is carried out by a

mismatch-repair mechanism (MMR) (Pâques and Haber, 1999; Do and LaRocque, 2015). If we focus in the SDSA pathway, there will be three contiguous regions in which gene conversion can take place. One of them is the region that corresponds to newly synthesized DNA that bridges the DSB. Here, the direction of the gene conversion event is fixed given that the receptor strand will always correspond to the invading strand (in which the DSB happened) since it is being synthesized using the intact strand as template (donor) for the gene conversion event.

One of the other two regions that can undergo gene conversion corresponds to the hDNA formed by the 3' overhang and the homologous region it pairs to during strand-invasion. In principle, these two regions must be highly homologous (see section 1.3.2) but they can contain some mismatches. Evidence suggests that the MMR mechanism corrects these mismatches during the recombination process (Pâques and Haber, 1999; Do and LaRocque, 2015). The other region that can undergo gene conversion is the region in which the newly synthesized DNA binds to the other 5' end effectively bridging the DSB. Mismatches found in these regions might be corrected in either direction in what are referred to as either *conversion* events or *restoration* events. The consequences of conversion over restoration in either of these cases will have implications on the gene conversion tract lengths and on their position relative to the DSB. Since this is an area of debate and indeed one in which I would like to speculate on, I will leave the details of these consequences to the Discussion chapter, section 4.2.1.

1.3.2 Homology requirements

I have mentioned the need for homology between sequences for HR to take place, and I have mentioned also that mismatches might be present in these homologous regions. The questions I will focus on now are what are the length and the degree of homology that are needed for HR to take place and what is the molecular mechanism through which this control takes place.

Although the degree of homology is important for allelic gene conversion, its relevance in IGC is much more given that paralogous regions can in principle diverge much more than homologous regions. Paralogous sequences are referred to in many occasions as *homeologous* since this is a term used to refer to sequences whose similarity is below a certain threshold, although there is no real consensus on the threshold itself. Waldman (2008) considers homeologous

those sequences that share between 80 and 90% identity, while Hastings (2010) sets the threshold at 97%.

Irrespective of their definition, there is a correlation between the rate with which homeologous sequences undergo IGC and their overall degree of homology. Chen et al. (2007) found that the degree of homology between interacting sequences is always above 92% and usually above 95% with very little exceptions. Accordingly, Waldman and Liskay (1987) showed that relative to recombination between near-identical sequences, the rate of recombination between two closely linked repeats that shared 80% identity was reduced more than 1000-fold.

In 1986, Shen and Huang conducted a set of experiments to assay the degree of identity necessary for homeologous recombination in *Escherichia coli* (Shen and Huang, 1986). They introduced the concept of a *minimal efficient processing segment* (MEPS), to describe the minimal length of 100% identity needed for recombination to take place. They reported a value of 23 to 27 bp and 44 to 90 bp for two different pathways. They argued that if the length of perfect homology was larger than the reported MEPS, then the recombination rate would be proportional to the number of overlapping MEPS fragments contained within that length. Being so, the number of MEPS contained within a 100% identity substrate pair is: $N = L - M + 1$, where L is the length of the perfectly homologous substrate and M is the length of MEPS (Shen and Huang, 1989). Furthermore, the recombination frequency of a given substrate pair is proportional to the number of MEPS it contains: $F = fN$, where f is the recombination rate of one MEPS (Shen and Huang, 1989). MEPS lengths have been reported to be between 134 and 232 bp in mouse fibroblasts (Waldman and Liskay, 1988), and between 337 and 456 bp in humans (Reiter et al., 1998).

In section 3.3, I will present the application note reporting SeDuS, a forward-in-time simulator of SDs which includes, among other features, the inclusion of MEPS thresholds for IGC. In section 3.4, I will present preliminary results of the effect of applying MEPS thresholds to IGC on the levels of variation across duplications.

Chapter 2

OBJECTIVES

The main objective of this thesis is to contribute to the understanding of the evolution of duplicated regions of the genome, focusing in particular on modeling interlocus gene conversion.

More specifically, this work aims to:

1. Contribute to a better understanding of the patterns of variation and linkage disequilibrium in duplications by exploring the interplay between interlocus gene conversion and crossover.

2. Generate awareness on the potential confounding effects of interlocus gene conversion and the collapse of duplications in genome-wide scans for selection.

3. Present a software that can simulate the evolution of duplicated regions of the genome under a wide range of scenarios.

4. Present preliminary results about the effects that sequence similarity dependence of interlocus gene conversion can have in the patterns of variation along duplicated sequences.

Chapter 3

RESULTS

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

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 (Bethesda). 2014 Jun 6;4(8):1479–89. DOI: 10.1534/g3.114.012435

3.2 Collapsed duplications: what to expect and what to look for.

Diego A. Hartasánchez, Marina Brasó-Vives, Marc Pybus, and Arcadi Navarro

Manuscript in preparation

Collapsed duplications: what to expect and what to look for.

Diego A. Hartasánchez*, Marina Brasó-Vives*, Marc Pybus*, Arcadi Navarro* § †

*Institute of Evolutionary Biology (Universitat Pompeu Fabra – CSIC), PRBB, Barcelona, Catalonia, Spain, 08003.

§National Institute for Bioinformatics (INB), Barcelona, Catalonia, Spain.

†Centre for Genomic Regulation (CRG), Barcelona, Catalonia, Spain, 08003.

Abstract

The detection and characterization of Segmental Duplications (SDs) and Copy-Number Variants (CNVs) is of great importance in the field of genomics. Even though SDs and CNVs may be privileged targets of natural selection, they are usually eliminated from genome-wide selection scans due to their possible source of confounding factors. On the one hand, duplications are prone to be collapsed onto one single region due to high identity between duplicates when constructing genome assemblies. Furthermore, low frequency CNVs, which are not in the reference, will also be collapsed when aligning sequence data from single individuals to the reference, even if repeat regions are masked. On the other hand, concerted evolution between duplications alters their site frequency spectrum and linkage disequilibrium patterns compared to neutral single-copy regions. Therefore, summary statistics traditionally used to detect the action of natural selection on DNA sequences cannot be applied to SDs and CNVs. Here we have obtained expectation values for ten summary statistics for duplications evolving in concert, under a broad range of interlocus gene conversion and crossover rates. We have compared simulated data for single-copy, duplicated and collapsed regions evolving neutrally obtained with SeDuS (a forward simulator of segmental duplications) and simulated data for selective and neutral scenarios obtained with MSMS (a coalescent simulator of single-copy regions under selective scenarios). In some cases, values for known duplications mimic selective signatures, such as those characteristic of incomplete sweeps in the case of Fay and Wu's H. However, both known and collapsed duplications can be differentiated from single-copy regions or regions under selective pressures with test statistics that measure levels of nucleotide and haplotype diversity. Therefore, if we scan the genome for regions of high nucleotide and haplotype diversity we might expect to encounter some cases of collapsed duplications. Contrary to our expectations, we find that regions with low (and not high) nucleotide and haplotype diversity are enriched in duplications. This pattern might be due to the strict filtering applied by SNP calling algorithms.

Segmental duplications (SDs), defined as ≥ 1 kb blocks of DNA that are present at several sites within the genome and that have ≥ 90% sequence similarity between copies (Sharp et al., 2006), are an ubiquitous characteristic of eukaryotic genomes. There are several reasons for their complex evolution: first, SDs undergo interlocus gene conversion (IGC), also referred to as non-allelic or ectopic gene conversion, which drives their concerted evolution (Walsh, 1987) and is a source of variation (Innan, 2002); second, genes residing within SDs may suffer different selective pressures giving rise to subfunctionalization or neofunctionalization (He and Zhang, 2005; Teshima and Innan, 2008; Assis and Bachtrog, 2013); and, third, SDs are mediators of non-allelic homologous recombination (NAHR), a common source of copy number variants (CNVs), which in turn are associated with susceptibility to disease (Stankiewicz and Lupski, 2010).

The detection and characterization of SDs and CNVs is of great importance in the field of genomics. However, when constructing genome assemblies, duplications are prone to be collapsed onto one single region due to high identity between duplicates, which complicates their detection through sequencing methods. Collapsed duplications are known to be a particularly widespread problem of reference genomes constructed via whole genome shotgun assembly (Salzberg and Yorke, 2005; Kelley and Salzberg, 2010), particularly those constructed with low coverage, and next generation sequencing also has this problem given the short length of the reads it produces (Alkan et al., 2011; Hahn, et al., 2014; Ribeiro et al., 2015).

To circumvent this problem, algorithms based on depth of coverage have been developed (Bailey et al., 2002; Yoon et al., 2009). These algorithms align reads to a reference genome and identify regions that have more reads aligned to them than expected, implying either a CNV present in the sequenced individual or the absence of an SD in the reference sequence. Thanks to the application of these tools, reference genomes have improved considerably.

In the the human reference genome (the highest quality mammalian genome), probably most of the common SDs and CNVs have already been detected, even though some regions were still found to be missing from the reference as late as 2010 (Kidd et al., 2010). Nonetheless, most CNVs that are present at very low frequencies in the population are not present in the reference genome (van Ommen, 2005; Kidd et al., 2010; Chen et al., 2011). Therefore, when sequence data from a single individual are aligned to the reference genome, there will still be some collapsed sequences, even after repeat-masking. Strategies to avoid collapsing duplications range from the most stringent, such as not considering any read that maps to

more than one location in the reference genome, to the more elaborate, such as filtering with the aforementioned depth of coverage algorithms. In cases of highly divergent regions, stringent strategies might fail to merge two sufficiently different haplotypes (Zimin et al., 2012).

In any case, given that SDs may be privileged targets of natural selection (Bailey and Eichler, 2006), the increasing availability of databases identifying SDs and CNVs (Sudmant et al., 2015), and the existence of methods to avoid collapsing duplications, it would seem natural to assess the action of selection within SDs. However, most genome-wide scans for selection concentrate their efforts on filtering-out SDs from their analyses in order to avoid spurious signals coming, not only from collapsed duplications, but from perfectly identified SDs (*e.g.* Chen et al., 2009; Enard et al., 2014).

The reason for this is that summary statistics traditionally used to detect the action of natural selection on DNA sequences cannot be applied to SDs since the latter evolve in a concerted fashion. Theoretical results have shown that increased diversity within duplicates can occur as a consequence IGC between paralogous copies (Ohta, 1982; Innan 2002, 2003; Teshima and Innan, 2012; Hartasánchez et al., 2014). IGC recombines variants between SDs having strong effects on both the site frequency spectrum and the patterns of linkage disequilibrium within these regions (Teshima and Innan, 2004; Thornton, 2007; Hartasánchez et al., 2014).

Innan (2003) had already pointed out that test statistics that are based on the standard coalescent cannot be correctly applied to duplicated regions since the distributions of these tests (such as Tajima's D, Fu and Li's D*, and Hudson, Kreitman and Aguadé's test) differ for multigenes and single-copy genes. Thornton (Thornton, 2007) also reported deviations from single-copy expectations depending on the fixation time of the gene duplication event. The presence of duplicated regions (collapsed or not) are therefore recognized as strong confounding factors in genome scans for positive (Mallick et al., 2009) and balancing selection (Fijarczyk and Babik, 2015). However, to our knowledge, there has been no assessment on what is the expected outcome of standard statistical tests for natural selection if they are applied to duplicated regions and collapsed duplicated regions of the genome under a wide range or recombination parameters. We here confirm that test statistic values that indicate neutrality for single-copy regions cannot be applied to duplicated regions undergoing concerted evolution. Additionally, we show that even under neutrality, different gene conversion rates among duplications and crossover rates between them render large variations in test statistic values for duplications and, more importantly, for collapsed duplications.

In order to obtain neutral estimates of diversity present in duplicated regions, we ran SeDuS (Hartasánchez et al., 2016) under a broad range of IGC and crossover rates. We then computed a set of ten summary statistics (shown in Table 1) from the data generated by SeDuS. Most of our selected set are test statistics that have been developed to detect deviations from neutral expectations. Each statistic is more or less sensitive to different deviations from neutrality (*e.g.* more sensitivity to intermediate frequency variants) and are more or less robust to potential confounding factors such as population bottlenecks or expansions. For simplicity, we will refer to the whole set as test statistics although formally this is not the case.

Results are compared between single-copy regions, whose average neutral variation levels are in general not strongly affected by differences in crossover rates and, of course, not affected by IGC rate, and duplications, which are analyzed in two ways. What is termed *duplicated* are statistics applied exclusively to one of the duplicated copies (across the population). For this to happen with real data, paralogs would need to be differentiated by some method. In the case of paralogs of high similarity, long reads (spanning to flanking regions) would be necessary, for example. The second case, termed *collapsed*, refers to both paralogs analyzed as if they were only one copy. To measure the effect of collapsed duplications, in our calculations, we have taken half of the sequences from each paralog, which increases intermediate frequency variants, in particular for low IGC rates.

We have compared mean values from 1,000 simulation runs for each statistic for all the IGC and crossover rates explored (Figure 1A). Values for collapsed duplications vary considerably between low and high IGC rates and crossover rates for some tests statistics, such as Tajima's D. On the other hand, for some other tests, such as Fay and Wu's H, values remain constant and are very close to the single-copy expectations. If we compare the distribution of values for each statistic for low, intermediate, and high IGC rates (Figure 1B), we can observe a considerable overlap between single-copy and collapsed duplications in most cases. However, nucleotide diversity estimators (*i.e.* average pairwise differences (π) and Watterson's estimator) as well as the haplotype diversity estimator (dh) show little overlap and seem to be less dependent on IGC rates. In principle, these estimators could serve as candidates to detect collapsed duplications even though there can be multiple other histories for a single-copy region that can give this type of signal. In order to test if these distributions are attainable under simple selective scenarios we ran coalescent simulations under MSMS (Ewing and Hermisson, 2010) under four different models: hard sweep, soft sweep, balancing selection and neutrality (see Methods).

In Figure 3, we compare simulated data for selective and neutral scenarios from MSMS and simulated data for single-copy, duplicated and collapsed regions from SeDuS. These data show that in isolation, most test statistics cannot clearly distinguish selective from neutral scenarios as has been reported previously (Pybus et al., 2015). Furthermore, in some cases, values for known duplications mimic selective signatures, such as those characteristic of incomplete sweeps in the case of Fay and Wu's H. On the other hand, both known and collapsed duplications can be differentiated from single-copy regions or regions under selective pressures when focusing on test statistics given their high levels of nucleotide and haplotype diversity. Therefore, if we scan the genome for regions of high nucleotide and haplotype diversity we might expect to encounter some cases of collapsed duplications.

However, in order to measure diversity, variants must be previously called. Genome-wide calling of SNPs (Single-Nucleotide Polymorphisms) performed with sequences of low coverage must establish very strict quality filters to avoid spurious calls. Hence, duplications that are present in the population at low frequencies and are not annotated as repeats in the reference genome might in some cases be collapsed when performing the SNP calling. If the calling is done with strong filters, diversity in these regions might be underestimated. Should we therefore expect to find more collapsed duplications in regions with high diversity or regions of low diversity?

To answer this question we selected outlier regions (top and low 1%) from the distributions of π and dh from human populations. We looked for regions that belonged to both the top 1% of π and the top 1% of dh, and to both the low 1% of π and the low 1% of dh. These values were extracted from calculations (Pybus et al., 2013) performed on the vcf files of Phase I of the 1000 Genomes Project for three populations: YRI, CEU and CHB. We then checked for the copy-number of these regions in the 11 human individuals used in the Great Ape Genome Project (Prado-Martinez et al., 2013; Sudmant et al., 2013) (see Methods). No conclusions can be extracted for the CEU and CHB populations. However, for the YRI population, the set of regions of low diversity (655 regions) shows a slight increase of values around copy-number 4 (one duplicate) when compared to the set of regions of high diversity or randomly chosen regions (Figure 4).

From this observation, we conclude, first, that in principle, high nucleotide and haplotype diversity cannot be used as a means to detect the presence of collapsed duplications, at least with data that has been treated with filtering criteria similar to the data in the 1000 Genomes Project. Prior to SNP calling, any region of

the genome that is annotated as a repeat is masked, and so variation is only performed for putative single-copy regions. Furthermore, even in putative single-copy regions, SNP calling is not carried out if multiple reads map to these regions.

Second, we consider that a possible explanation for the observed increase in high copy-number values (around copy-number 4) in low diversity regions for African populations might be the presence of CNVs in close proximity to human fixed duplications (Monlong et al., 2015). While the latter are very well annotated and included in the RepeatMasker track of the UCSC Genome Browser, this is not the case for low-frequency CNVs. We argue that African populations are more likely to harbour CNVs not included in the duplication tracts. Reads from the duplication are collapsed onto the single-copy reference but are not called as if they were single-copy. Rather, both the original reads and the reads from the duplication are eliminated. The effect of this elimination is a strong decrease in the estimates for variation within these regions and this is only observed for the YRI population.

Despite the high quality of the human reference genome and its annotation of repeats in particular, we do observe the effect of putative collapsed duplications in African individuals (YRI). We consider that the underestimation of variation in African genomes might be a widespread problem for low-coverage genomes, both at the CNV and nucleotide level, due to stringent SNP calling. Although we have not assayed the frequency of collapsed duplications in other species, it is most surely a widespread problem given the low quality of many reference genomes. We have shown with simulations that IGC strongly alters the site-frequency spectrum and linkage disequilibrium patterns of duplications in such a way that neutral references for test statistics cannot be applied to them. Furthermore, we show that known or collapsed duplications can imitate signals of selection, emphasizing that test statistics should not be used in isolation as a means to detect natural selection.

Methods

All simulations involving duplicated regions were done with a slightly modified version of SeDuS (Hartasánchez et al., 2016). We chose a range of crossover and IGC values to show that summary statistics are very dependent on these parameters: low crossover: R=1; intermediate crossover: R=10; high crossover: R=100; low IGC: C=0.5; intermediate IGC: C=1; high IGC: C=10. SeDuS simulates the evolution of two-copy duplicates evolving under concerted evolution and a single-copy control region. It also outputs a collapsed sample between the original and duplicated copies. All our calculations are done once the simulated population has reached mutation-drift equilibrium.

Additionally to the results from SeDuS, we have run simulations with MSMS (Ewing and Hermisson, 2010). Simulated scenarios involve neutrality, a complete selective sweep (-SAA 40 -SaA 20), an incomplete selective sweep (-SAA 40 -SaA 20) and a case of balancing selection (-SAA 0 -Saa 0 -SaA 40).

Test statistic values were extracted from the data presented in Pybus et al. (2013). Values for each test statistic are provided genome-wide in contiguous 3 kb windows for African (Yoruba from Ibadan or YRI), European (Central European from Utah or CEU), and Asian (Han Chinese from Beijing or CHB) populations. These statistics were calculated from the vcf files of the Phase I release of the 1000 Genomes Project, which used reads aligned to the GRCh37 human assembly. Coordinates were translated to GRCh38 with the coordinate conversion tool (liftOver) at http://genome.ucsc.edu/.

Copy-number variation estimates were performed using the human reference genome GRCh38, with the algorithm by Sudmant et al. (2013). The 11 humans used to calculate copy-number genome-wide were sequenced as part of the Great Ape Genome Diversity Project (Prado-Martinez et al., 2013), the Orangutan Genome Project (Locke et al., 2011), and the Denisova Genome Project (Meyer et al., 2012). The copy-number that we report for our selected windows is the weighted average (by overlapping length) of the copy-number of the regions detected by the algorithm by Sudmant et al. (2013). The randomly-chosen regions used as a null distribution for genome-wide copy number were extracted from a subset of regions with copy number below 6, in order to avoid including repetitive regions.

Acknowledgments

We thank Marcos Fernandez-Callejo for running the pipelines to calculate copy-number calls on GRCh38. This work has been supported by the Spanish National Institute of Bioinformatics, a platform of the Instituto de Salud Carlos III (PT13/0001/0026) and the Spanish Government, grant BFU2012-38236 to A. N.; a grant to D. A. H. from Conacyt; and by the Fondo Europeo de Desarrollo Regional (FEDER) and the Fondo Social Europeo (FSE).

Conflict of interest

None.

References

Alkan, C., Sajjadian, S. & Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. Nature Methods 8: 61-65.

Assis, R. & Bachtrog, D. (2013). Neofunctionalization of young duplicate genes in Drosophila. Proc. Natl. Acad. Sci. USA 110: 17409-17414.

Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., et al. (2002). Recent segmental duplications in the human genome. Science 297: 1003-1007.

Bailey, J. A. & Eichler, E. E. (2002). Recent segmental duplications in the human genome. Science 297 (5583): 1003-1007.

Chen, C.-H., Chuang, T.-J., Liao, B.-Y. & Chen, F.-C. (2009). Scanning for the signatures of positive selection for human-specific insertions and deletions. Genome. Biol. Evol. 1: 415-419.

Chen, W., Hayward, C., Wright, A. F., Hicks, A. A., Vitart, V., et al. (2011). Copy number variation across European populations. PLoS ONE 6: e23087.

Enard, D., Messer, P. W. & Petrov, D. A. (2014). Genome-wide signals of positive selection in human evolution. Genome Res. 24: 885-895.

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

Fijarczyk, A. & Babik W. (2015). Detecting balancing selection in genomes: limits and prospects. Mol. Ecol. 24: 3529-3545.

Fay, J. C. & Wu, C.I. (2000). Hitchhiking under positive Darwinian selection. Genetics 155: 1405-1413.

Fu, Y. X. & Li, W. H. (1993). Statistical tests of neutrality of mutations. Genetics 133: 693–709.

Gautier, M. & Vitalis, R. (2012). rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. Bioinformatics 28: 1176-1177.

Hahn, M. W., Zhang, S. V. & Moyle, L. C. (2014). Sequencing, assembling, and correcting draft genomes using recombinant populations. G3 (Bethesda) 4: 669-679.

Hartasánchez, D. A., Vallès-Codina, O., Brasó-Vives, M. & Navarro, A. (2014). Interplay of interlocus gene conversion and crossover in segmental duplications under a neutral scenario. G3 (Bethesda) 4: 1479–1489.

Hartasánchez, D. A., Brasó-Vives, M., Fuentes-Díaz, J., Vallès-Codina, O. & Navarro, A. (2016). SeDuS: segmental duplication simulator. Bioinformatics 32:148-150.

He, X. & Zhang, J. (2005). Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics 169: 1157-1164.

Innan, H. (2002). A method for estimating the mutation, gene conversion and recombination parameters

in small multigene families. Genetics 161: 865-872.

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

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

Kidd, J. M., Sampas, N., Antonacci, F., Graves, T., Fulton, R., et al. (2010). Characterization of missing human genome sequences and copy-number polymorphic insertions. Nature Methods 7: 365-371.

Li H. (1993). A new test for detecting recent positive selection that is free from the confounding impacts of demography. Mol. Biol. Evol. 28: 365-375.

Lin, K., Li, H., Schlötterer, C. & Futschik, A. (2011). Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics. Genetics 187: 229-244.

Locke, D. P., Hillier, L. W., Warren, W. C., Worley, K. C., Nazareth, L. V., et al. (2011). Comparative and demographic analysis of orang-utan genomes. Nature 469: 529-533.

Mallick, S., Gnerre, S., Muller, P. & Reich, D. (2009). The difficulty of avoiding false positives in genome scans for natural selection. Genome Research 19: 922-933.

Meyer, M., Kircher, M., Gansauge, M. T., Li, H., Racimo, F., et al. (2012). A high-coverage genome sequence from an archaic Denisovan individual. Science 338: 222-226.

Monlong, J., Meloche, C., Rouleau, G., Cossette, P. Girard, S. L., et al. (2015). Human copy number variants are enriched in regions of low-mappability. bioRxiv. doi: http://dx.doi.org/10.1101/034165

Nei, M. & Li,W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA 76: 5269–5273.

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

Ohta, T. (1982). Allelic and nonallelic homology of a supergene family. Proc. Natl. Acad. Sci. USA 79: 3251-3254.

Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E. & Lercher, M. J. (2014). PopGenome: An efficient swiss army knife for population genomic analyses in R. Mol. Biol. Evol. 31:1929–1936.

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., et al. (2013). Great ape genetic diversity and population history. Nature 499: 471-475.

Pybus, M., Dall'Olio, G. M., Luisi, P., Uzkudun, M., Carreño-Torres, A., et al. (2014). 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. Nucleic Acid. Res. 42: D903-D909.

Pybus, M., Luisi, P., Dall'Olio, G. M., Uzkudun, M., Laayouni, H., et al. (2015). Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations.

Bioinformatics 31: 3946-3952.

Ramírez-Soriano, A., Ramos-Onsins, S.E., Rozas, J., Calafell, F. & Navarro, A. (2008). Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. Genetics 179: 555-567.

Ribeiro, A., Golicz, A., Hackett, C. A., Milne, I., Stephen, G., et al. (2005). Beware of mis-assembled genomes. Bioinformatics 21: 4320-4321.

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

Sharp, A. J., Cheng, Z. & Eichler, E. E. (2006). Structural variation in the human genome. Annu. Rev. Genomics Hum. Genet. 7: 407-442.

Stankiewicz, P. & Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. Annu. Rev. Med. 61: 437-455.

Sudmant, P. H., Huddleston, J., Catacchio, C. R., Malig, M., Hillier, L. W., et al. (2013). Evolution and diversity of copy number variation in the great ape lineage. Genome Res. 23: 1373-1382.

Sudmant, P. H., Mallick, S., Nelson, B. J., Hermozdiari, F., Krumm, N., et al. (2015). Global diversity, population stratification, and selection of human copy-number variation. Science 349: aab3761.

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

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

Teshima, K. M. & Innan, H. (2008). Neofunctionalization of duplicated genes under the pressure of gene conversion. Genetics 178: 1385-1398.

Teshima, K. M. & Innan, H. (2012). The coalescent with selection on copy number variants. Genetics 190: 1077-1086.

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

van Ommen, G.-J. B. (2005). Frequency of new copy number variation in humans. Nature Genetics 37: 333-334.

Voight, B.F., Kudaravalli, S., Wen, X. & Pritchard, J.K. (2006). A map of recent positive selection in the human genome. PloS Biol. 4: e72.

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

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7: 256–276.

Yoon, S., Xuan, Z., Makarov, V., Ye, K. & Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 19: 1586–1592.

Zeng, K., Fu, Y.-X., Shi, S. & Wu C.-I. (2006). Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics 174: 1431-1439.

Zimin, A. V., Kelley, D. R., Roberts, M., Marçais, G., Salzberg, S. L., et al. (2012). Mis-assembled "segmental duplications" in two versions of the *Bos taurus* genome. PloS ONE 7: e42680.

Test statistic	Type	Reference	Package
π	Diversity estimator	Nei and Li, 1979	Evolboosting
Watterson's θ	Diversity estimator	Watterson, 1975	Evolboosting
Tajima's D	Neutrality statistic	Tajima, 1989	PopGenome
Fu and Li's D	Neutrality statistic	Fu and Li, 1993	PopGenome
Fu and Li's F	Neutrality statistic	Fu and Li, 1993	PopGenome
Fay and Wu's H	Neutrality statistic	Fay and Wu, 2000	PopGenome
Zeng's E	Neutrality statistic	Zeng et al., 2006	PopGenome
Li's MFDM	Neutrality statistic	Li, 2011	Evolboosting
dh	Haplotype based	Nei, 1987	SScosi
iHS	Haplotype based	Voight *et al*., 2006	rehh

Table 1. Selected set of summary statistics applied to our data. To calculate these statistics we used four software programs: PopGenome (Pfeifer et al., 2014), Evolboosting (Lin et al., 2011), SSCosi (Ramírez-Soriano et al., 2008), and rehh (Gautier and Vitalis, 2011). Some of the statistics are implemented in several programs and results are reproducible between programs to a very large extent.

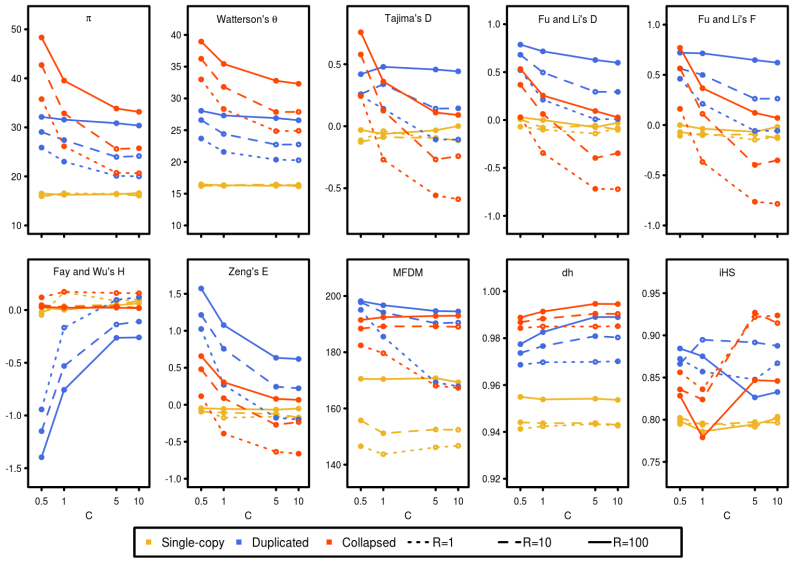

Figure 1. Average values for our selected set of summary statistics from 1000 SeDuS simulations. Values are shown for single-copy (yellow), duplicated (blue) and collapsed (red) for a range of crossover rates (R = 1, 10, 100) and interlocus gene conversion rates (C = 0.5, 1, 5, 10).

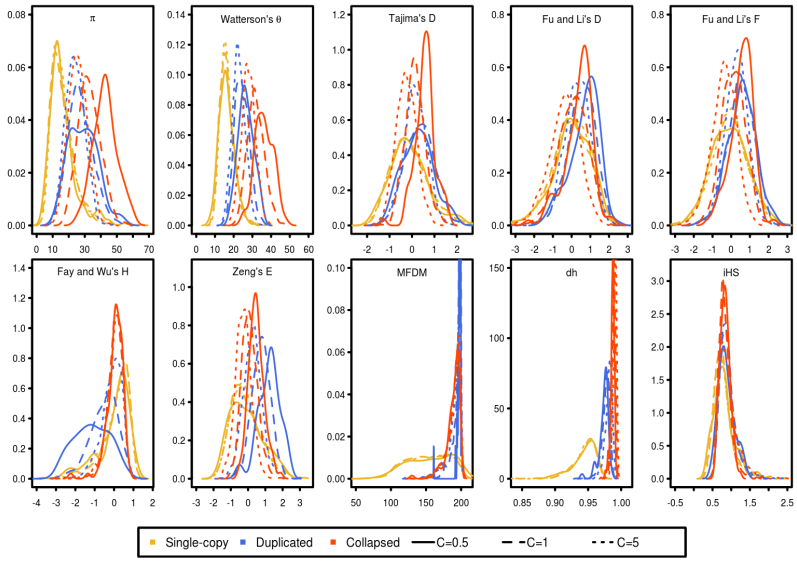

Figure 2. Distribution of values for our selected set of summary statistics from 1000 SeDuS simulations. Values are shown for single-copy (yellow), duplicated (blue), and collapsed (red) for a range of interlocus gene conversion rates ($C = 0.5$, 1, 5) and a crossover rate of $R = 10$.

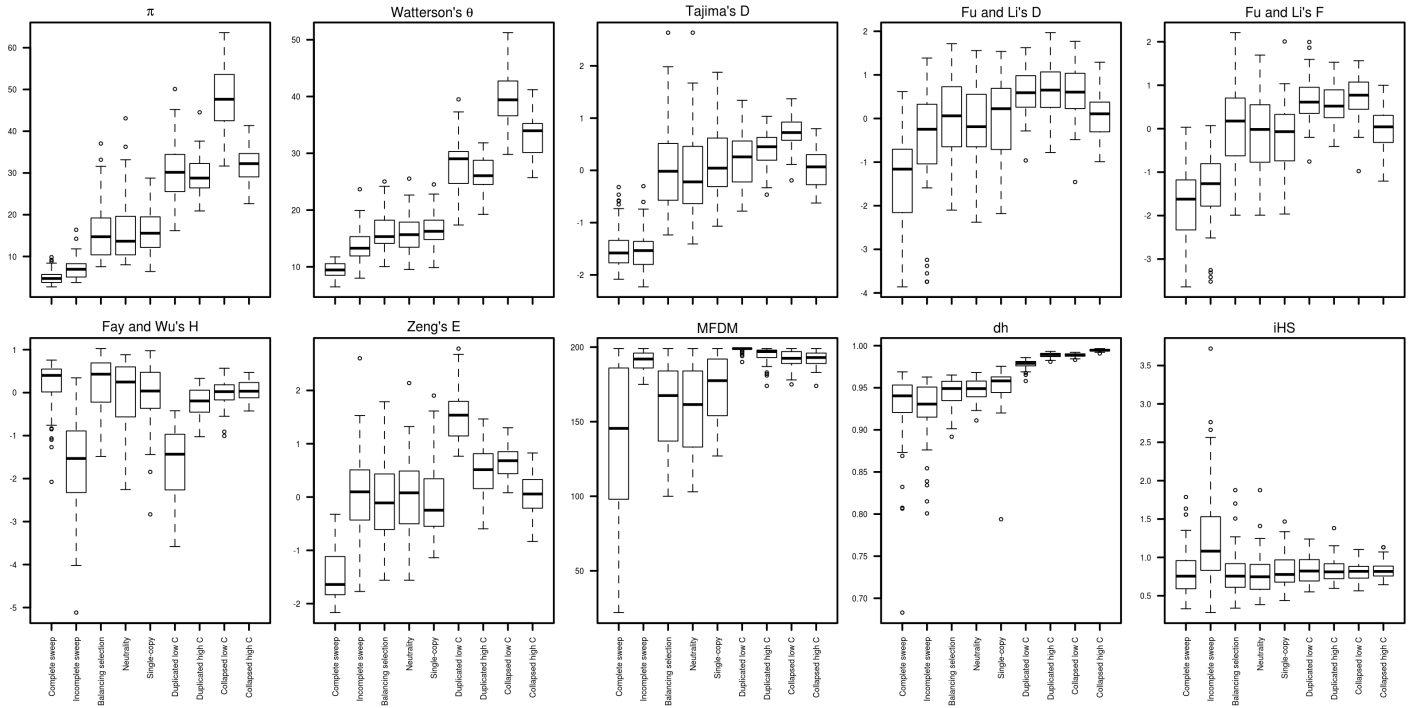

Figure 3. Boxplot comparison between simulation results from MSMS (complete sweep, incomplete sweep, balancing selection, and neutrality) and SeDuS (single-copy, duplicated, and collapsed) with low ($C = 0.5$) and high ($C = 10$) values of IGC rates. The length of the boxplot whiskers are 1.5 times the inter-quantile range. Distributions for Fay and Wu's H for an incomplete sweep resemble those from duplicates with low C. π and dh are two statistics that clearly differentiate between duplicates and collapsed regions from regions under selection.

3.3 SeDuS: segmental duplication simulator

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

3.4 Interlocus gene conversion dependence on sequence similarity

Marina Brasó-Vives, Diego A. Hartasánchez, and Arcadi Navarro

Work in progress

In this last results section I will present preliminary results on an analysis of the effect of introducing sequence similarity thresholds on IGC. In the IGC model presented in section 3.1, IGC acted regardless of the degree of identity between duplicates. Aware of this limitation, we included an IGC dependence on sequence similarity in version 1.0 of SeDuS, presented in section 3.3. Even though these results are very preliminary, they may help to illustrate the main consequences of introducing sequence similarity thresholds for IGC.

Even though the contents of this section should not be read as a finished manuscript, I have included a short summary and background information specific to the section along with a few results from simulations carried out in collaboration with Marina Brasó-Vives, under the supervision of Arcadi Navarro. The discussion related to these results is included in the general discussion of the thesis (section 4.2.1). The references used herein are included in the Bibliography of the thesis.

Summary

Interlocus gene conversion (IGC) has been recognized as the driver mechanism for the concerted evolution of gene duplications. A fundamental feature of IGC, which is ignored in most models of evolution of gene families, is that IGC is strongly dependent on the degree of similarity between the sequences exchanging information. Sequence similarity can be evaluated by measuring the length and/or the degree of homology between duplications. Regarding the length of homology, experiments in several species have confirmed that a *minimal efficient processing segment* (MEPS), consisting in a 100% identity tract between duplicates, is required for IGC. Regarding the degree of homology, a *minimal efficient sequence homology* (MESH) has been proposed: sequences that share less than perfect homology along the whole duplication are less prone to undergo IGC compared to sequences with perfect homology. We here present a preliminary exploration of the effect of introducing a restriction on IGC by MEPS. We have performed simulations with SeDuS and measured divergence between duplicates across the length of the duplication. Regions in which there is a positive feedback loop between the action of IGC decreasing divergence and the presence of MEPSs, are flanked by regions in which there is a negative feedback loop between the lack of IGC events causing and increase of divergence and a lack of MEPSs. This duality along the duplicates might generate *islands of divergence*. We show that introducing MEPS restrictions

limits the possibilities of concerted evolution to a narrower range of IGC rates. Additionally, we propose that MESH restrictions might in fact be a consequence of two MEPS restrictions, one for initiation and one for resolution, acting at different moments of the IGC event and at the two ends of the IGC tract, which is consistent with a few experimental observations.

Background

The functional role of gene duplication was first observed in *Drosophila* in the 1930's (Bridges, 1936; Muller, 1936). However, the importance of gene duplication in evolution did not become popular among biologists until after the 1970's, in part due to the influential book by Ohno, *Evolution by Gene Duplication* (Ohno, 1970). Since then, the pervasiveness of gene duplications across the tree of life has been recognized (Lynch and Conery, 2000; Kellis et al., 2004; Sémon and Wolfe, 2007) and multiple models to explain their evolution have been proposed (reviewed by Innan and Kondrashov (2010)).

Interlocus Gene Conversion (IGC) is the main mechanism for the concerted evolution of gene duplications or small multigene families (Ohta, 1983)). However, the molecular mechanisms underlying IGC are still largely unknown (Hastings, 2010). IGC rate has been shown to depend on many factors including the distance between duplicates, their relative orientation, the crossover rate in the regions in which they lie, among others (Chen et al., 2007; Benovoy and Drouin, 2009; McGrath et al., 2009; Casola et al., 2010; Mansai et al., 2011). However, there is a main factor that determines IGC rate which is the sequence similarity (or sequence homology) between duplicates. To evaluate sequence similarity one can measure two factors: the length of homology and the degree of homology (Waldman, 2008).

Regarding the degree of homology, in order for duplications to engage in IGC, there must be a minimum degree of homology between the sequences involved. Chen et al. (2007) evaluated this overall degree of homology and reached the conclusion that all IGC events (except one, in their study) involved duplicates with more than 92% homology while the large majority of them involved a 95% homology. Chen et al. (2010) defined the *minimal efficient sequence homology* (MESH) as the minimum overall homology between duplicates necessary for there to be IGC between them.

The name MESH was coined in analogy to the measurement proposed by Shen and Huang, in 1986. Shen and Huang (1986) proposed that for there to be

IGC there needs to be a *minimal efficient processing segment* (MEPS) consisting in a 100% identity fragment between duplicates of a given length. Once this requirement is satisfied, the rate of IGC is linearly correlated with the number of MEPS fragments found between duplicates. In other words, if one considers any fragment of DNA as a series of overlapping MEPS, each of which recombines at the same rate, then the recombination rate of the whole fragment should be directly proportional to the number of MEPS within that fragment (Datta et al., 1996). MEPS values have been experimentally measured in *E. coli* ($\sim$30) (Shen and Huang, 1986), in mouse ($\sim$200) (Waldman and Liskay, 1988) and in humans ($\sim$400) (Reiter et al., 1998) among other species.

There are only three theoretical models that incorporate IGC dependence on sequence similarity between duplications. Walsh (1987) developed two models: first, the *k-hit* model in which IGC is completely inactivated by a few (k) mutations, which were not thought to be point mutations, but rather large insertions or deletions; and second, a more general model in which IGC rate gradually declines as point mutations accumulate and stops once divergence goes above a certain threshold. Teshima and Innan (2004) considered that divergence between duplicates would arrive at a mutation-conversion-drift equilibrium and that IGC rate would remain constant. However, they considered that there would be random fluctuations around this equilibrium that could cause divergence to reach a certain threshold above which IGC would be terminated. Both Walsh (1987) and Teshima and Innan (2004) were therefore considering that IGC was determined not by a local threshold, such as MEPS, but by a general threshold such as MESH, although they do not mention it explicitly in their work.

Contrary to the aforementioned models, a threshold on IGC imposed by MEPS would not cause a generalized decrease of IGC rate with increasing divergence or a sudden end to concerted evolution when divergence surpasses a certain threshold. Rather, dependence on sequence similarity based on MEPS would cause differential IGC rates along the sequence depending on local sequence divergence. The overall rate decrease would only be a consequence of averaging regions of high and low IGC rates. For example, we can imagine that, by chance, two or more mutations appear in a region fracturing a MEPS fragment, therefore impeding IGC to act in that region. Since most duplications include several MEPS fragments, IGC will not be effectively terminated throughout the whole duplication. Instead, a negative feedback loop would be

established in the vicinity of the broken MEPS fragment: given that IGC cannot initiate in that region, the less likely it will be for an IGC tract to extend into that region and therefore the more likely it will be for divergent mutations to be fixed in that vicinity, increasing the divergence in the region and extending the width of what I will refer to as an *island of divergence*. Of course, if islands of divergence reach a density such that no MEPS fragment is left within the duplicated segment, IGC would be effectively terminated throughout the whole duplicated region.

Although MEPS and MESH values have been measured in different experiments and organisms, there has been, to my knowledge, no study that has measured both for the same experimental setup. However, there is some evidence that two independent inhibitory mechanisms for homeologous recombination exist (Datta et al., 1997). An inverted repeat assay using sequences ranging from 74% to 100% identity was implemented by Datta et al. (1997) to evaluate the role of the *mismatch repair* (MMR) machinery in inhibiting mitotic crossover in yeast. By comparing wild-type to MMR defective strains, they concluded that the MMR mechanism was responsible for inhibiting crossover between sequences with a single mismatch. Additional mismatches had a cumulative negative effect on the recombination rate that could be attributed exclusively to the MMR system for sequences with more than 90% identity. However, they noted that for sequences with more than 10% divergence, wild-type and MMR defective strains had similar inefficient recombination rates, suggesting that a factor other than the MMR machinery strongly impairs the recombination process. Although they refer to this factor as a "general limitation in the yeast recombination machinery" they do not make any reference to a MESH-like restriction. Rather, they suggest this limitation could correspond to an inability to have an efficient crossover resolution (Datta et al., 1997) and they invoke another MEPS segment in their theoretical modeling to account for this limitation. Evidence for two MEPS segments, one necessary for initiation and one for resolution, has also been presented by Yang et al. (2006) for homoelogous recombination between mammalian chromosomes.

It seems, thus, that even though the MESH concept may be valuable to describe the overall degree of homology necessary for recombination, it might just be a consequence of local restrictions in initiation and resolution of recombination. Additionally, there has been, to my knowledge, no mechanism

associated to a MESH-mediated inhibition, although the absence of evidence is of course, by no means, evidence of absence.

Simulations involving only initiation MEPS, simultaneous MESH and MEPS, and both initiation and resolution MEPS could be performed, among other possibilities, in order to compare their effects. The simulations performed so far consider only an initiation MEPS and the results I present here evidence the drastic effects that sequence similarity dependence of IGC has on the patterns of variation across duplications.

Preliminary results

For our preliminary exploration of the dependence of IGC on sequence similarity, we have performed computer simulations with a slightly modified version of SeDuS (see section 3.3). In the IGC model of version 1.0 of SeDuS, IGC tracts extend $l/2$ sites to the left and $l/2$ sites to the right of the initiation site, where l is the IGC tract length extracted from a geometric distribution (Wiuf and Hein, 2000) with an average length of λ (which is set to 100 bp throughout this study). Once the initiation site of the potential IGC event has been determined (with rate $g = C/\lambda$), the IGC event will only be carried out if both copies are identical along the MEPS length, which also extends to the left and to the right of the initiation site. In some cases, l will be shorter than the MEPS length and so the IGC event will not convert any mutation. In what follows, I will distinguish between potential and effective IGC rates. Potential IGC rate (C) corresponds to the IGC rate that would result if no sequence similarity threshold were set in the simulations and all IGC attempts took place. The effective IGC rate will be calculated *a posteriori* based on the IGC attempts that do take place, that is, the ones that pass the sequence similarity thresholds.

Additional to the MEPS threshold, the simulations performed with a MEPS restriction also included a MESH-type threshold requiring a 92% homology between donor and acceptor sequences throughout each potential IGC tract for IGC to be effective. However, the inclusion of this threshold did not modify the results to an observable extent.

Our results show that effective IGC rates change throughout the simulations and that this change is dependent on potential IGC rates. The ratio between effective and potential IGC rates is depicted in Figure 3.1A as a function of time. Time 0 corresponds to the moment in which the duplication first appears, when both copies are identical. At this time, there is no difference between the

potential and the effective IGC rates since there are no differences between duplicates and the MEPS requirement is satisfied regardless of the potential IGC rate. From this moment on, mutations appear and differences begin to accumulate between duplicates. These differences are represented by the variation between duplicates in Figure 3.1B.

The left image in Figure 3.1A corresponds to a MEPS of 20 bp. Red lines at the top correspond to very high potential IGC rates. In these cases, IGC erases almost all the differences between duplicated regions and the variation between duplicates remains very small (Figure 3.1B). For small potential IGC rates (bottom yellow line), the number of IGC events happening between duplicates is not enough to balance the increase in divergence between them. The effective IGC rate declines progressively as differences accumulate between duplicates. By comparing different MEPS lengths (20 bp, 50 bp and 200 bp) one can see that increasing MEPS lengths increases the threshold below which potential IGC rates are not enough to maintain effective IGC rates at a non-zero equilibrium value. For example, in the case of a MEPS length of 200 bp, the only potential IGC rate to attain a non-zero equilibrium is $C = 50$. The case of $C = 0.5$ for a MEPS length of 50 bp is also interesting since the effective IGC rate and the corresponding variation between duplicates attain an equilibrium value even though many IGC events do not satisfy the MEPS threshold.

Figure 3.2 shows the expectations for variation within one of the duplicated regions at equilibrium as a function of IGC rate for three theoretical models: first, 1 minus Ohta's identity coefficient $(1 - f)$ (Ohta, 1983); second, the expectation for heterozygosity $E(h_w)$, developed by Innan (2002); and third, the expectation for average pairwise differences $E(\pi_w)$ of the coalescent, infinite-site model also by Innan (2003a). [Details of the models can be found in the Appendix (File S3).]

None of these models include IGC dependence on sequence similarity. As can be observed in Figure 3.2, the simulation results with no restriction on sequence similarity follow the theoretical expectations (for $C = 0.01$, the simulation runs where not long enough for equilibrium to be achieved and therefore theoretical expectations for variation are above the simulation results). However, none of the models with restriction of IGC on sequence similarity, corresponding to MEPS lengths of 20, 50 and 200 bp, fit the theoretical expectations. First, the corresponding curves are shifted to the right with respect to the theoretical expectations. This is caused by the decrease of effective IGC rates compared to the potential IGC rate: the stronger the restriction (longer

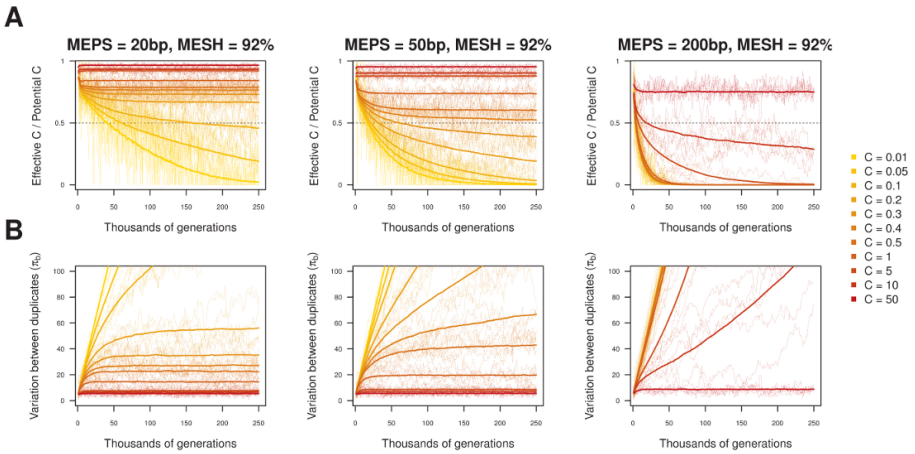

Figure 3.1: (A) Ratio between effective IGC rates and potential IGC rates against time and (B) variation between duplicates against time. Results are shown for three different MEPS lengths: 20 bp (left), 50 bp (middle), and 200 bp (right). Each plot shows results for potential IGC rates (C) ranging from 0.01 to 50, a fixed crossover rate (R) between duplicates of 10 and MESH values of 92%. Continuous lines are average values of 1000 simulations while discontinuous thin lines show 5 randomly chosen trajectories for each value of C. Generation 0 corresponds to the time of the duplication event in our simulations. The cases in which the ratio of effective over potential IGC rate tends to zero correspond to parameter sets in which IGC between duplicates is canceled by MEPS requirements. Accordingly, in these cases, which correspond to low C values, once concerted evolution ends, variation between duplicates increases linearly. For intermediate to high C values, non-zero equilibrium values are attained, and higher MEPS lengths imply a narrower range of C values for which this equilibrium is reached.

MEPS lengths), the higher the difference between effective and potential IGC rates. Second, the stronger the restriction, the narrower the corresponding curve because of a smaller range of IGC rates for which a non-zero equilibrium for effective IGC rates can be achieved.

I have shown so far that non-zero equilibrium is more restricted for models with MEPS. I will now explore what is the pattern of variation between duplicates (a proxy for divergence) along the sequence for different scenarios. Figures 3.3A, B and C show simulation results for a MEPS length of 50 bp and $C = 0.1$, 1 and 0.5 respectively and Figure 3.3D serves as a neutral comparison

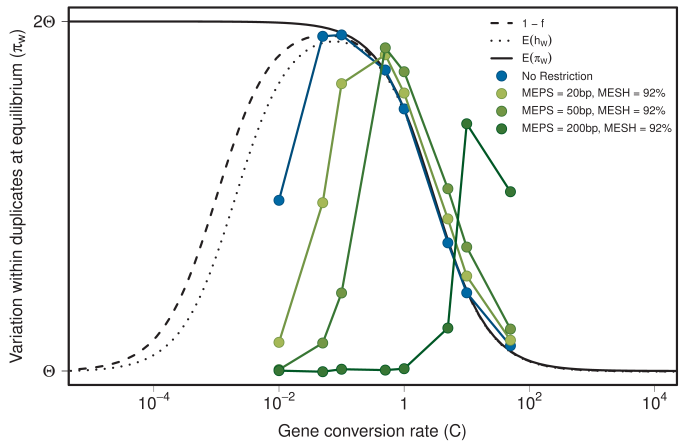

Figure 3.2: Variation within duplicated regions, showing theoretical expectations (black lines) and simulation results (filled circles) for different MEPS lengths. Theoretical expectations correspond to $(1 - f)$ Ohta (1983), $E(h_w)$ Innan (2002), and $E(\pi_w)$ Innan (2003a). The latter corresponds to an infinite site model like the one used in our simulations. The blue line corresponds to a model where IGC rate does not depend on sequence similarity and fits theoretical expectations except for the lowest values of potential IGC rates (C), in which the average time to reach equilibrium is longer that the simulations. Green lines, showing results for different MEPS lengths, display a narrowing of the bell-shaped curve with increasing MEPS length, implying a limitation of the range of C values for which increased variation can be attained. Curves are shifted to the right in all cases of IGC restriction showing that by increasing MEPS lengths, a higher C is needed to reach a given effective IGC rate and the corresponding level of variation.

with no restriction and $C = 0.5$. For each case, plots for variation within duplicates (π_w), variation between duplicates (π_b) and effective IGC rates are depicted for one example simulation run. Variation is calculated in 250 bp sliding windows every 50 bp during 30,000 generations in 1,000 generation intervals. Shades of green correspond to the 30 time points and means are shown in black. In the bottom plot, the corresponding effective IGC rates along the sequence are shown in shades of white (effective IGC equal to 0) to black (effective IGC rate equal to potential IGC rate).

Figure 3.3A is an example of a potential IGC rate that is too small to

maintain sequence similarity ($C = 0.1$). Variation between duplicates is very high in almost all the sequence and allows no IGC events. There is an exception of a small fragment in the beginning of the duplicated region in which some IGC events are still happening. Variation within the duplicate in this case is, as expected, around zero in all the sequence except for the small fragment that keeps undergoing some IGC events. In this small fragment, we find that variation within duplicates is increased due to the effect of these IGC events. In Figure 3.3B we find the opposite situation, in which a high potential IGC rate ($C = 10$) keeps variation between duplicates very low. This means that the effective IGC rate is almost equal to the potential IGC rate. In this case, variation within duplicates is between Θ and 2Θ, and its distribution along the sequence coincides with the distribution of variation between duplicates. IGC converts variation between duplicates to variation within duplicates by transferring new variants to a duplicate while at the same time reducing the divergence between duplicates. This is why in this example, in which there is presence of IGC, fragments with higher values of π_b will have higher levels of π_w and vice versa. Figure 3.3C ($C = 0.5$) shows an intermediate case. The effective IGC rate is lower than the potential IGC rate but different from zero and we find considerable variation within duplicates (see Figure 3.2). In this equilibrium, we find a non-homogeneous distribution of the variation between duplicates along the sequence. There are regions with low values of π_b that allow IGC events to happen (and therefore, increase π_w), interspersed with isolated regions with high values of π_b (the divergence islands) that show lower local levels of effective IGC rate (and therefore, low π_w). Figure 3.3D is an example of a simulation with the same potential IGC rate than Figure 3.3C ($C = 0.5$) but without any restriction on sequence similarity. We can see the consequences of having IGC rate independent on sequence similarity: there are no zones with increased π_b as seen in the restrictive model. We find a homogeneously-distributed IGC rate and some correlation between π_w and π_b.

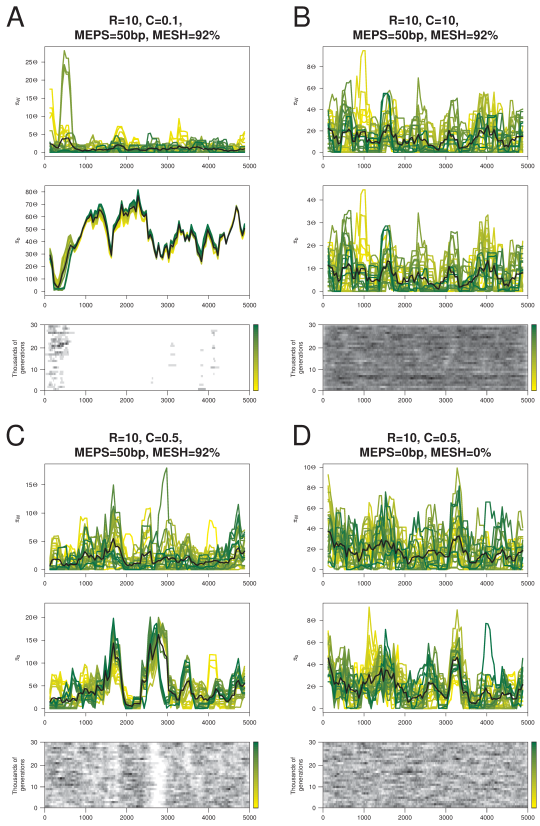

Figure 3.3: Variation within duplicates (top), variation between duplicates (middle), and effective IGC rate between duplicates (bottom) along the length of the duplicated sequence for a single simulation run in each case (A-D). A, B and C correspond to simulations with the same IGC restriction on sequence similarity (MEPS = 50 bp, MESH = 92%) but different potential IGC rates ($C = 0.1$, 10, and 0.5, respectively). As a means of comparison, D shows the results of the model in which no restriction on sequence similarity is applied to IGC (MEPS = 0 bp, MESH = 0%) and $C = 0.5$. Each case corresponds to a simulation run of 30,000 generations and data is shown for snapshots taken every 1,000 generations, corresponding to different green lines (top and middle) and position along the vertical axis (bottom). The horizontal axis represents the nucleotide positions along the duplicates in all cases. The black line in the top and middle plots for each case corresponds to the average values across the 30 measurements. Note that the vertical axis covers a different range of values in each plot. In the bottom plot, white represents absence of IGC and black corresponds to the maximum IGC rate in each case so rates are not comparable between cases.

Chapter 4

DISCUSSION

¡La gracia de lo imperfecto!
¡La bendición del error!
Cada cual es quien es, por
lo que hace de sus defectos.
La bruma de los afectos
que gobierna el alma humana,
nos libre de la tirana
fiebre de perfeccionar.
¡Que a veces sólo al errar
acierta uno en la diana!

Jorge Drexler

4.1 Main findings

4.1.1 Variation in duplications

In the study presented in section 3.1, we analyzed the effect that different crossover distributions can have in the variation and LD patterns within and between duplications. Theoretical models by Ohta (1982), Nagylaki (1983), and Innan (2002, 2003b) already predicted accurately that IGC between duplications could increase the amount of variation (diversity) present within each copy up to twice the amount expected under neutrality for single-copy regions. The rationale behind this effect is that since IGC is a copy-paste event, mutations arising in one copy can be transferred to the other copy. For certain IGC rates, the effective mutation rate of each copy can be twice as much as the mutation rate in a single-copy region. If recombination is restricted to IGC between duplications lying on the same chromosome, exchange of mutations can only occur in a back-and-forth manner between the same pair of copies. However, by allowing crossover to occur between the duplications, there is an additional shuffling that boosts the effect of IGC by putting new pairs of copies on the same chromosome and therefore increasing the probability that an IGC event will effectively transfer mutations between different copies. Being so, for the same IGC rate, an increase in the inter-duplication crossover rate always increases variation within duplications.

The shuffling power of crossover was also included in the aforementioned models. However, limiting crossover to act in the region between duplications is a condition that is not appropriate in many cases. To take a clear example, tandem duplications are duplications that lie in close proximity to one another and the probability of a crossover junction to fall in between the duplications is very small. A more realistic crossover model would be to allow crossover in a wider region including the duplicated regions themselves. We ran simulations under this crossover model and our results showed that the overall effect was a decrease in the shuffling power of crossover. The equivalent crossover rate R' in an inter-duplication crossover model (referred to as single-copy crossover (SCC) model in section 3.1) for a crossover rate R in a whole-region crossover (WRC) model is:

$$R' = \frac{(L_{dup} + L_{inter})\, R}{(2L_{dup} + L_{inter})}, \tag{4.1}$$

where L_{dup} is the length of the duplication and L_{inter} is the separation between the duplications. The reasoning for this fact can be found in the Appendix (File S6).

Additional to the WRC model, in the study presented in section 3.1 we incorporated a hotspot crossover (HSC) model. The fine-scale structure of recombination rates in humans is known to be dominated by hotspots (Jeffreys et al., 2001; McVean et al., 2004; Kong et al., 2010). Around 80% of all recombination events take place in less than 20% of the sequence (Myers et al., 2005) and these hotspots of recombination (defined as having at least a 5-fold increase of recombination rate compared to the background surrounding recombination rate) occur in average every 200 kb or less in the human genome (McVean et al., 2004). Importantly, gene conversion events are known to co-localize with crossover events (Ardlie et al., 2001; Jeffreys and May, 2004), however, the effect of crossover hotspots located within duplications has not been widely analyzed.

Thus, we simulated several scenarios under the HSC model in which a crossover hotspot was located inside one of the duplicated regions and we analyzed the effect it would have in the variation present within that copy and within the other copy. Our results show that the amount of variation within a region will depend on its location relative to the hotspot. If the hotspot lies in between the region and the corresponding paralogous region in the other copy,

the level of variation will be exactly the one expected under the SCC model for the same crossover rate. If the hotspot is not located between the paralogous regions, the level of variation will correspond to that with a null crossover rate. Regions overlapping the hotspot location will have intermediate levels of variation. As a consequence, the different levels of variation present along the copy without the hotspot mimic the variation present in the corresponding paralogous regions of the copy with the hotspot. This situation might cause important differences in the amount of variation present within a copy without there being any difference in the IGC rate or the crossover rate within that region. This illustrates the complex interplay between IGC and crossover that we highlight in this study.

4.1.2 Linkage disequilibrium in duplications

Crossover and gene conversion recombine DNA and therefore reduce the levels of LD over time. However, the effect that each one has is qualitatively different. Whereas the decay of LD due to crossover increases as the distance between markers increases, the decay of LD due to gene conversion is independent of the inter-marker distance when the latter is greater than the gene conversion tract length (Andolfatto and Nordborg, 1998; Wiehe et al., 2000). Therefore, while large-scale LD patterns are determined mostly by crossover events, short-scale LD patterning is caused by the added effect of gene conversion and crossover but mainly by gene conversion (Andolfatto and Nordborg, 1998; Ardlie et al., 2001; Frisse et al., 2001; Jeffreys et al., 2001; Padhukasahasram et al., 2004; Plagnol et al., 2006). Being so, it was not until LD could be measured at a very high resolution (*i.e.* between very closely placed markers) that the high frequency of gene conversion compared to crossover was realized (see Hellenthal and Stephens (2006).

It is important to keep in mind that all of the observations mentioned in the previous paragraph were made for allelic gene conversion and not for interlocus gene conversion. However, all evidence indicates that the molecular mechanisms controlling allelic gene conversion and those controlling IGC are the same (Jeffreys and May, 2004; Chen et al., 2007). Thus, we expect that all of the aforementioned characteristics will apply to IGC as well.

In the study presented in section 3.1, we analyzed the LD pattern from our simulations under the three crossover models: SCC, WRC and HSC. Given that our data is generated by simulations, we have power to measure LD between

every two positions for the whole simulated region. We are not only calculating LD within duplications, but also LD in the single-copy region, and more importantly LD between duplications on the same chromosome. Interestingly, since we are dealing with IGC, we have studied a rather overlooked aspect of LD, namely, LD between paralogous regions in duplications.

The results under the SCC model with $R = 50$ compared to $R = 0$ are as expected: crossover breaks down long-range LD. In our simulations this would correspond to all pairwise measurements involving two points from two different regions (that is, original, single-copy or duplicated, in the notation used in our study). When incorporating IGC, we also observe the expected breakdown of short-range LD corresponding to LD measurements within the same region. The higher the IGC rate, the lower the LD values. The novelty in our study is in the observation of high LD between paralogous positions in original and duplicated regions (what we refer to as a diagonal blue line in Figures 4 and 8 of the study presented in section 3.1). The higher the IGC rate, the stronger the LD between paralogous regions. With high-quality phased genomes, that is, genomes in which one can be sure of which variant is present in which chromosome, long-range LD could be accurately measured. With such data, one could potentially spot high IGC rates between duplications, or in any repetitive sequence for that matter, by looking for high LD between paralogous regions in duplications or simply by looking for high LD in diagonals analogous to the ones we report.

Under the WRC model we observe the same added effect of crossover and IGC in LD breakdown. Simulation results for high IGC rates show that LD is higher within the single-copy region than in the duplicated copies. Under the hotspot model with only one hotspot, analogous to our observation for variation, we observe strong differences in short-range LD across the duplicated regions despite it not containing a crossover hotspot itself. This corresponds to a *second-order* effect in which differences in LD along a sequence are not due to differences in recombination rates across the sequence itself but to crossover distributions elsewhere in the genome.

It would indeed be interesting to study crossover rates across paralogous regions in the genome and to check for the presence of shared or private hotspots within them. Given the fast turnover rate of hotspots (Winckler et al., 2005) and the strong effects that we have shown that their presence can have in the variation and LD patterns within duplications, the evolutionary fate of

duplications might be due in part to the recombination rates across them.

Regardless of the particular crossover models analyzed in section 3.1, the main conclusion that can be extracted from the paper is that variation and LD patterns within and between duplications can be affected to a large extent when compared to expectations in single-copy regions. Given that neutral expectations for single-copy regions are the basis for test statistics employed to search for signatures of natural selection, in the study presented in section 3.2, I addressed, along with my coauthors, some of the potential confounding factors that duplicated regions can have in genome-wide scans for natural selection.

4.1.3 Neutrality tests applied to duplications

In the study presented in section 3.2, we applied a series of test statistics to duplications. Our results demonstrate that IGC and crossover between duplications can cause important deviations in these statistics. Importantly, in many instances, these deviations are highly dependent on the specific rates of IGC and crossover even though simulations were only performed under the SCC model, therefore avoiding the most striking deviations reported in section 3.1. We also analyzed the effect of collapsing duplications, that is, taking both copies of the duplication and applying the set of test statistics to them as if they were only one copy. The variation in the results compared to single-copy expectations is even greater than for statistics applied to only one of the copies.

Additionally, we compared the distribution of values obtained with our simulator with those obtained with the coalescent simulator MSMS (Ewing and Hermisson, 2010) under three selective scenarios to see if the signatures of IGC could be confounded by signatures of hard sweeps, soft sweeps or balancing selection. Although we have not performed an exploration of a broad range of strengths of selection, we have confirmed, in agreement with previous results (Innan, 2003b; Thornton, 2007), that duplications undergoing IGC can have patterns of variation and LD akin to those characteristic of regions under selection for some test statistics. We have extended these conclusions to collapsed duplications.

It is clear that duplications (whether they are known duplications or unknown collapsed duplications) can be potential confounding factors in selection scans. However, if we look in the opposite direction there is an unaddressed question regarding the possibility of using test statistics to detect collapsed duplications. So instead of focusing on those test statistics whose

results are similar between selection scenarios and duplications, we focused on those in which the results for duplications were clearly different to results for single-copy regions. From our analysis, we concluded that the statistics that can best distinguish single-copy regions (regardless of them being under selection or neutral) and duplications were estimates of nucleotide diversity and haplotype diversity in which duplications render high values. This was expected since the main effect of IGC is the creation of novel short-range haplotypes and the accompanying increase in nucleotide variation. It is also in agreement with these statistics having been identified as less prone to being confounded in selection scans (Teshima et al., 2006; Thornton, 2007).

To prove the duplication detection power of these two test statistics, we took outlier regions of the genome for haplotype and nucleotide diversity. We kept those regions that were in the top 1% tail of haplotype diversity and overlapped them with the set in the top 1% tail of nucleotide diversity. We expected there to be an increased copy-number in those regions, probably due to unrecognized duplications or segregating copy-number variants. What we found was the opposite: we found an increase in copy-number in those regions that belonged to the bottom 1% tail of haplotype and nucleotide diversity. Curiously, we only detected this pattern when the set of windows with low nucleotide and haplotype diversity was extracted from African individuals.

It is important to mention that the data we used to select regions of high and low diversity were vcf files from Phase I of the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010) and that SNPs were called only for putatively single-copy regions, that is, only those regions not overlapping with previously identified simple repeats, tandem repeats, SDs or interspersed repeats. Despite this fact, we expected some population-specific duplications or copy-number variants segregating at low frequencies not to have been identified and as such, being potentially collapsed therefore displaying high nucleotide and haplotype diversity. The explanation for our observation comes from an additional technical aspect, namely, that the SNP calling for these genome sequences at low coverage was performed with stringent filtering criteria to avoid false positives (The 1000 Genomes Project Consortium, 2010). With stringent filtering, collapsed duplications or unidentified copy-number variants might cause an artificial depletion in variation measures and from there our enrichment of high copy number in regions of low variation and low haplotype diversity.

An additional observation that needs adequate explanation is the fact that we only observe this depletion when regions are selected from the YRI (African Yoruba) population and not from CEU (Central European) or CHB (Han Chinese). Our *ad hoc* explanation is that the human reference is constructed mostly from European individuals and the duplication databases are also constructed mainly with data from Europeans while most of the singleton duplications and low-frequency copy-number variants are likely to be found in African populations (Campbell and Tishkoff, 2008).

Although we are well aware that these results require appropriate testing and further investigation, we consider that our idea to use test statistics as a means to detect duplications is not a dead end. Using humans to test our hypothesis had both a weak and a strong points. Since humans are the species for which duplications have been more properly characterized and studied, the probability of there being collapsed duplications in the data was rather low. However, being aware of this point, we considered that if we were to find evidence for collapsed duplication in humans, given the quality of the genome assembly, this would represent a very clear signal that this was bound to be a common problem in many species. Given our results, we cannot extend our conclusion to any other species and further testing is required.

Mehr licht!

Attributed to Goethe

Who will believe my verse in time to come,
If it were fill'd with your most high deserts?

From Sonnet XVII, William Shakespeare

4.2 Future directions with SeDuS

The paper presented in section 3.3 is the application note of SeDuS, a forward-in-time simulator of the evolution of SDs coded in C++. SeDuS can be used to explore a wide variety of evolutionary scenarios and a broad range of rates, some of which escape the realm of coalescent simulators. Version 1.0 of SeDuS is built upon the in-house software that we developed for the study presented in section 3.1, which was also used to obtain the results in section 3.2. Compared to the previous version, the software presented in section 3.3 includes several features that are important to take into account when simulating the evolution of duplications (see the SeDuS Tutorial, included in the Appendix, for details). There are three additions that I consider of particular interest given the potential effects they can have. First, we have introduced biased directionality (*i.e.* donor-acceptor bias) in IGC, which has been observed experimentally in several instances (Chen et al., 2007). Second, we have included the possibility to simulate different fixation trajectories for the duplication. We consider that duplications that are under positive selection will rise in frequency at a rate proportional to the strength of selection and we simulate this rise in frequency through a deterministic linear trajectory akin to Teshima and Innan (2012). Third, we have incorporated sequence similarity dependence in IGC.

We have already obtained preliminary results for all of these additions. In particular, sequence similarity dependence of IGC has proven to be extremely interesting for a wide variety of reasons: because of its undeniable relevance in the evolution of duplications; because of the extensive amount of research

dedicated to it; because of the great variety of species on which experiments have been performed to test it; because of the inconclusive data produced by these experiments; because of the molecular mechanisms involved in this process; and because of the results that we have produced ourselves, which I will discuss in section 4.2.1.

Regarding further applications of SeDuS, its modular structure is intented for the straightforward implementation of new features, and projects in collaboration with other research groups are already underway to expand SeDuS. In particular, I have been supervising the expansion of SeDuS to include multiple copies (instead of only two) undergoing IGC. Additionally, we are considering the implementation of variable population sizes along simulations, which will allow SeDuS to be used in the exploration of many additional scenarios.

Regarding crossover hotspots, in the current version of SeDuS, their location is fixed and stable throughout the simulation, and it might be interesting to include motif-mediated hotspot locations. In an evolutionary context, the colocalization of crossover hotspots with allelic gene conversion events has led to the hotspot paradox (Boulton et al., 1997). If hotspots are localized at PRDM9-binding sites (Baudat et al., 2010; Myers et al., 2010) and the sequence around the DSBs is replaced by the sequence of an intact homologous chromatid as a result of allelic gene conversion, DSB-prone sequences are bound to be lost in favor of alleles less prone to DSBs. Including motif-mediated crossover hotspot locations in SeDuS would allow us to investigate this effect in the context of duplicated regions.

SeDuS also has some limitations which are due primarily to its underlying structure, such as a binary representation of DNA. G-C biased gene conversion, for example, has been shown to importantly affect genome content (Lesecque et al., 2013; Glémin et al., 2015; Williams et al., 2015) and it would be extremely useful to incorporate it in SeDuS. However, in order to implement it, we would at least have to distinguish between two types of mutations and this would require large modifications to SeDuS.

An additional contribution of this thesis is SeDuS' graphical user interface, which is not only intended for a quick exploration of the effect of parameter changes within the research community, but also, and very importantly, as a teaching application for graduate and post-graduate students of molecular evolution.

4.2.1 Sequence similarity dependence

In section 3.4, I presented preliminary results of an ongoing project. The aims of the project are as follows. First, we intend to analyze the effects of sequence similarity dependence of IGC compared to a model with no restriction. The preliminary results presented in section 3.4 address this specific issue. Second, we wish to compare between different plausible models of sequence similarity dependence. Third, we will explore the possibility of gene conversion tract lengths being a consequence of sequence similarity dependence of IGC.

Regarding the effects of introducing MEPS into our simulations, we observe that compared to the no-restriction scenario, the rates of IGC for which an elevated amount of variation can be found within copies is limited. The higher the restriction, the narrower the range. Additionally we observe the creation of islands of divergence. Our results show that for a given MEPS length, width and stability in time of islands of divergence is dependent on the potential IGC rate. We expect that increasing MEPS lengths for a fixed IGC rate will increase the mean width and mean-life of an island of divergence. The probability of disappearance of an island of divergence will be inversely correlated with its width. Since IGC is a stochastic process, and given the possibility of there being long IGC tracts that could span a whole island of divergence, there will be non-zero probability for an island of divergence to disappear if there is at least one MEPS segment in the duplication. Local peaks of divergence between paralogs have been theoretically analyzed for the process of neofunctionalization under the pressure of gene conversion (Teshima and Innan, 2008) and experimental evidence for divergence peaks around the target site of selection has been obtained in *Drosophila melanogaster* (Osada and Innan, 2008). Since we show evidence that islands of divergence might appear even in the absence of selective pressures, it would be interesting to incorporate sequence similarity dependence into these models and compare the width and height of the islands of divergence with different selective strengths.

Regarding other models of sequence similarity dependence, I described in section 3.4 the concept of MESH, which is an overall sequence similarity threshold that must exist in order for IGC to happen at considerable rates (Chen et al., 2010). In principle, IGC restriction by a global measure, such as MESH, is not consistent with a machinery, such as the MMR, which acts locally. Thus, it seems plausible that MESH thresholds are just a consequence of restriction by MEPS. However, a unique MEPS requirement for initiation of IGC would not

be consistent with experimental measurements of MESH (Datta et al., 1997; Pâques and Haber, 1999). An alternative would be that there are two independent inhibitory mechanisms for homeologous recombination. One would correspond to the commonly accepted initiation restriction by MEPS and the other would be a resolution restriction, which in principle could also be a 100% identity segment (Datta et al., 1997; Yang et al., 2006). Interestingly, in Yang et al. (2006), a paper by the group of Waldman, the authors suggest that the MEPS measurement for mammalian cells (134 to 232 bp) which was carried out by the same group (Waldman and Liskay, 1988) might in fact have been a measurement of the MEPS length for resolution and not for initiation as they considered at the moment. Indeed, it is still considered as the latter by many authors (Chen et al., 2007; Mansai et al., 2011).

Two MEPS requirements would be consistent also with gene conversion being a consequence of SDSA. As I mentioned in section 1.3.1, under the SDSA pathway, there are two clear instances of homology recognition (initiation and resolution). The MMR mechanism is not only responsible for rejection of the hDNA when it does not satisfy MEPS. It is also responsible for the correction of hDNA when it does satisfy MEPS and these corrections will be the actual gene conversion tracts. Being so, the gene conversion tract length would not only be a consequence of a stochastic mechanism of extension as suggested by Wiuf and Hein (2000), but also related to the homology present at initiation and resolution. In this regard, McVey et al. (2004) present evidence of multiple strand invasions during SDSA. They suggest that after an initial strand invasion, synthesis and dissociation, if the newly synthesized strand does not find the required homology for resolution, it might re-invade and proceed with additional synthesis.

Well it's all right, riding around in the breeze.
Well it's all right, if you live the life you please.
Well it's all right, doing the best you can.
Well it's all right, as long as you lend a hand.

"End of the line", The Traveling Willburys

4.3 Concluding remarks

Alike many phenomena in molecular biology, IGC might be seen as an accident. To correct DSBs the cell has evolved a machinery that consists in the search for an homologous region to the region in the vicinity of the DSB. Practically nothing is known about how a DNA strand is able to scan through the whole genome in search for an homologous region (Barzel and Kupiec, 2008), but the truth is that the cell is pretty successful in this *homology search*. In general, when DSBs occur in dipoid organisms, the homologous region used as a template is found in the homologous chromosome or in a sister chromatid, and the DSB is corrected. However, in some occasions, homology is found in a paralogous region in another locus. This process, in which a DSB is corrected using a paralogous region as a template, is precisely IGC. So, indeed, IGC seems to be a consequence of a failure by the cell to use the *adequate* homologous copy as a template to correct DSBs, that is, an accident.

We know that this accident might in fact cause disease, and there are several known examples of deleterious IGC events between pseudogenes and their functional predecessors (Chen et al., 2010). However, IGC might also allow for the long-term survival of duplicated copies of a gene whose product is beneficial when its dosage is increased (Innan and Kondrashov, 2010). Also, loss-of-function mutations appearing in functional copies can be repaired through IGC (Katju et al., 2008), and recently, the spread of a favorable mutation via IGC has been observed across transposable elements in *Drosophila miranda* (Ellison and Bachtrog, 2015). So it seems that this accidental process might in some cases be beneficial to the cell.

The evolution of duplicated copies and their ultimate fates under the pressure of gene conversion has been investigated in several selective scenarios (Mano and Innan, 2008; Takuno et al., 2008; Teshima and Innan, 2008; Takuno and Innan,

2009) with interesting results. However, there are some issues regarding their neutral evolution which have not been previously addressed.

In this thesis, I have presented a few studies whose aim is to shed light on the neutral evolution of duplicated sequences in the context of IGC:

1. A study on the complex interplay between IGC and crossover. I have shown the striking effect that overlapping IGC susceptible regions with crossover hotspots can have on the patterns of variation and LD within and between duplicated regions.

2. A study confirming evidence of the potential confounding effects of duplications in genome-wide scans for selection. I have suggested that stringent SNP-calling might cause an underestimation of allelic variation due to collapsed duplications.

3. A forward-in-time computer simulator that allows for the simulation of duplicated copies evolving under the action of IGC. SeDuS can simulate the evolution of SDs under a wide range of mutation, IGC and crossover rates, and it can potentially be used to explore a broad range of evolutionary scenarios.

4. Preliminary results on the effect of introducing restriction on the process of IGC by making it dependent on the degree and length of sequence similarity between the interacting sequences.

I have modeled IGC focusing on relevant factors, and I have explored the effects of these factors via simulations, providing a context to show the importance of IGC in the evolution of duplicated sequences.

Bibliography

Andolfatto, P. and Nordborg, M. (1998). The effect of gene conversion on intralocus associations. *Genetics*, 148(2):1397–1399.

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. (2015). Rapid divergence and diversification of mammalian duplicate gene functions. *BMC Evolutionary Biology*, 15(1):138.

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. (2002). 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. *American Journal of Human Genetics*, 73(4):823–834.

Baker, M. (2012). Structural variation: the genome's hidden architecture. *Nature Methods*, 9(2):133–137.

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

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Barzel, A. and Kupiec, M. (2008). Finding a match: How do homologous sequences get together for recombination? *Nature Reviews Genetics*, 9:27–37.

Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G., and de Massy, B. (2010). PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science*, 327(5967):836–840.

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

Betrán, E., Rozas, J., Navarro, A., and Barbadilla, A. (1997). The estimation of the number and the length distribution of gene conversion tracts from population DNA sequence data. *Genetics*, 146(1):89–99.

Bischof, J. M., Chiang, A. P., Scheetz, T. E., Stone, E. M., Casavant, T. L., Sheffield, V. C., and Braun, T. A. (2006). Genome-wide identification of pseudogenes capable of disease-causing gene conversion. *Human Mutation*, 27(6):545–552.

Boulton, A., Myers, R. S., and Redfield, R. J. (1997). The hotspot conversion paradox and the evolution of meiotic recombination. *Proceedings of the National Academy of Sciences of the United States of America*, 94(15):8058–8063.

Bridges, C. B. (1936). The Bar "gene" a duplication. *Science*, 83(2148):210–211.

Buena-Atienza, E., Ruther, K., Baumann, B., Bergholz, R., Birch, D., De Baere, E., Dollfus, H., Greally, M. T., Gustavsson, P., Hamel, C. P., Heckenlively, J. R., Leroy, B. P., Plomp, A. S., Pott, J. W., Rose, K., Rosenberg, T., Stark, Z., Verheij, J. B., Weleber, R., Zobor, D., Weisschuh, N., Kohl, S., and Wissinger, B. (2016). De novo intrachromosomal gene conversion from OPN1MW to OPN1LW in the male germline results in Blue Cone Monochromacy. *Scientific Reports*, 6:28253.

Campbell, C. D., Sampas, N., Tsalenko, A., Sudmant, P. H., Kidd, J. M., Malig, M., Vu, T. H., Vives, L., Tsang, P., Bruhn, L., and Eichler, E. E. (2011). Population-genetic properties of differentiated human copy-number polymorphisms. *American Journal of Human Genetics*, 88(3):317–332.

Campbell, M. C. and Tishkoff, S. A. (2008). African genetic diversity: Implications for human demographic history, modern human origins, and complex disease mapping. *Annual Review of Genomics and Human Genetics*, 9:403–433.

Casola, C., Conant, G. C., and Hahn, M. W. (2012). 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.

Charlesworth, B. and Charlesworth, D. (2016). Population genetics from 1966 to 2016. *Heredity*, pages 1–8.

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., Férec, C., and Cooper, D. N. (2010). 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.

Conant, G. C. and Wolfe, K. H. (2008). Turning a hobby into a job: how duplicated genes find new functions. *Nature Reviews Genetics*, 9(12):938–950.

Cooper, G. M., Nickerson, D. A., and Eichler, E. E. (2007). Mutational and selective effects on copy-number variants in the human genome. *Nature Genetics*, 39:S22–S29.

Darwin, C. (1859). *On the Origin of Species.* Murray, London.

Datta, A., Adjiri, A., New, L., Crouse, G. F., and Jinks Robertson, S. (1996). Mitotic crossovers between diverged sequences are regulated by mismatch repair proteins in Saccaromyces cerevisiae. *Molecular and Cellular Biology*, 16(3):1085–93.

Datta, A., Hendrix, M., Lipsitch, M., and Jinks-Robertson, S. (1997). Dual roles for DNA sequence identity and the mismatch repair system in the regulation of mitotic crossing-over in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 94(18):9757–9762.

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

Do, A. T. and LaRocque, J. R. (2015). The role of Drosophila mismatch repair in suppressing recombination between diverged sequences. *Scientific Reports*, 5:17601.

Drouin, G. (2002). Characterization of the gene conversions between the multigene family members of the yeast genome. *Journal of Molecular Evolution*, 55(1):14–23.

Dumont, B. L. (2015). Interlocus gene conversion explains at least 2.7 % of single nucleotide variants in human segmental duplications. *BMC Genomics*, 16(1):456.

Dumont, B. L. and Eichler, E. E. (2013). Signals of historical interlocus gene conversion in human segmental duplications. *PloS ONE*, 8(10):e75949.

Eichler, E. E. (2001). Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends in Genetics*, 17(11):661–669.

Ellison, C. E. and Bachtrog, D. (2015). Non-allelic gene conversion enables rapid evolutionary change at multiple regulatory sites encoded by transposable elements. *eLife*, 4:e05899.

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.

Excoffier, L., Hofer, T., and Foll, M. (2009). Detecting loci under selection in a hierarchically structured population. *Heredity*, 103(4):285–298.

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

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection.* Clarendon Press, Oxford.

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

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):831–43.

Glémin, 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.

Haasl, R. J. and Payseur, B. A. (2016). Fifteen years of genomewide scans for selection: Trends, lessons and unaddressed genetic sources of complication. *Molecular Ecology*, 25(1):5–23.

Haber, J. E. (2000). Partners and pathways. *Trends in Genetics*, 16(6):259–264.

Haldane, J. B. S. (1932). *The Causes of Evolution.* Longmans Green, London.

Hastings, P. (2010). Mechanisms of ectopic gene conversion. *Genes*, 1(3):427–439.

Hellenthal, G. and Stephens, M. (2006). Insights into recombination from population genetic variation. *Current Opinion in Genetics and Development*, 16(6):565–572.

Hughes, A. L. (1994). The evolution of functionally novel proteins after gene duplication. *Proceedings of the Royal Society B - Biological Sciences*, 256(1346):119–124.

Hurles, M. (2002). Are 100,000 "SNPs" useless? *Science*, 298(5598):1509.

Innan, H. (2002). A method for estimating the mutation, gene conversion and recombination parameters in small multigene families. *Genetics*, 161(2):865–72.

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. (2004). Theories for analyzing polymorphism data in duplicated genes. *Genes & Genetic Systems*, 79(2):65–75.

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.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409:860–921.

Jeffreys, A. J., Kauppi, L., and Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics*, 29(2):217–222.

Jeffreys, A. J. and May, C. A. (2004). Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nature Genetics*, 36(2):151–156.

Jensen, J. D., Foll, M., and Bernatchez, L. (2016). The past, present and future of genomic scans for selection. *Molecular Ecology*, 25(1):1–4.

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.

Katju, V., LaBeau, E. M., Lipinski, K. J., and Bergthorsson, U. (2008). Sex change by gene conversion in a Caenorhabditis elegans fog-2 mutant. *Genetics*, 180(1):669–672.

Kellis, M., Birren, B. W., and Lander, E. S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature*, 428(6983):617–624.

Kim, P. M., Lam, H. Y. K., Urban, A. E., Korbel, J. O., Affourtit, J., Grubert, F., Chen, X., Weissman, S., Snyder, M., and Gerstein, M. B. (2008). Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. *Genome Research*, 18(12):1865–1874.

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

Kong, A., Thorleifsson, G., Gudbjartsson, D. F., Masson, G., Sigurdsson, A., Jonasdottir, A., Walters, G. B., Jonasdottir, A., Gylfason, A., Kristinsson, K. T., Gudjonsson, S. A., Frigge, M. L., Helgason, A., Thorsteinsdottir, U., and Stefansson, K. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 467(7319):1099–1103.

Krogh, B. O. and Symington, L. S. (2004). Recombination proteins in yeast. *Annual Review of Genetics*, 38:233–271.

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

Lesecque, Y., Mouchiroud, D., and Duret, L. (2013). GC-biased gene conversion in yeast is specifically associated with crossovers: Molecular mechanisms and evolutionary significance. *Molecular Biology and Evolution*, 30(6):1409–1419.

Lewontin, R. C. and Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, 74:175–195.

Librado, P. and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, 25(11):1451–1452.

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.

Lieber, M. R. (2010). The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annual Review of Biochemistry*, 79:181–211.

Lieber, M. R., Ma, Y., Pannicke, U., and Schwarz, K. (2003). Mechanism and regulation of human non-homologous DNA end-joining. *Nature Reviews Molecular Cell Biology*, 4(9):712–720.

Lynch, M. and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155.

Macke, J. P. and Nathans, J. (1997). Individual variation in the size of the human red and green pigment gene array. *Investigative Ophthalmology and Visual Science*, 38(5):1040–1043.

Mallick, S., Gnerre, S., Muller, P., and Reich, D. (2009). The difficulty of avoiding false positives in genome scans for natural selection. *Genome Research*, 19(5):922–933.

Manel, S., Perrier, C., Pratlong, M., Abi-Rached, L., Paganini, J., Pontarotti, P., and Aurelle, D. (2016). Genomic resources and their influence on the detection of the signal of positive selection in genome scans. *Molecular Ecology*, 25(1):170–184.

Mano, S. and Innan, H. (2008). The evolutionary rate of duplicated genes under concerted evolution. *Genetics*, 180(1):493–505.

Mansai, S. P. and Innan, H. (2010). The power of the methods for detecting interlocus gene conversion. *Genetics*, 184(2):517–27.

Mansai, S. P., Kado, T., and Innan, H. (2011). The rate and tract length of gene conversion between duplicated genes. *Genes*, 2(4):313–331.

Markova-Raina, P. and Petrov, D. (2011). High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. *Genome Research*, 21(6):863–74.

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. (2009). A burst of segmental duplications in the genome of the African great ape ancestor. *Nature*, 457(7231):877–881.

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

McMahill, M. S., Sham, C. W., and Bishop, D. K. (2007). Synthesis-dependent strand annealing in meiosis. *PLoS Biology*, 5(11):2589–2601.

McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science*, 304(5670):581–584.

McVey, M., Adams, M., Staeva-Vieira, E., and Sekelsky, J. J. (2004). Evidence for multiple cycles of strand invasion during repair of double-strand gaps in Drosophila. *Genetics*, 167(2):699–705.

Meselson, M. S. and Radding, C. M. (1975). A general model for genetic recombination. *Proceedings of the National Academy of Sciences of the United States of America*, 72(1):358–361.

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.

Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–4.

Myers, S., Bowden, R., Tumian, A., Bontrop, R. E., Freeman, C., MacFie, T. S., McVean, G., and Donnelly, P. (2010). Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science*, 327(5967):876–879.

Nagylaki, T. (1983). Evolution of a finite population under gene conversion. *Proceedings of the National Academy of Sciences of the United States of America*, 80(20):6278–6281.

Nagylaki, T. (1984a). Evolution of multigene families under interchromosomal gene conversion. *Proceedings of the National Academy of Sciences of the United States of America*, 81(12):3796–3800.

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

Nagylaki, T. and Petes, T. D. (1982). Intrachromosomal gene conversion and the maintenance of sequence homogeneity among repeated genes. *Genetics*, 100(2):315–37.

Nassif, N., Penney, J., Pal, S., Engels, W. R., and Gloor, G. B. (1994). Efficient Copying. *Molecular and Cellular Biology*, 14(3):1613–1625.

Nei, M. (1969). Gene duplication and nucleotide substitution in evolution. *Nature*, 221:40–42.

Ohno, S. (1970). *Evolution by Gene Duplication*. Springer-Verlag, 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):3251–4.

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

Ohta, T. (1991). Role of diversifying selection and gene conversion in evolution of major histocompatibility complex loci. *Proceedings of the National Academy of Sciences of the United States of America*, 88:6716–6720.

Osada, N. and Innan, H. (2008). Duplication and gene conversion in the Drosophila melanogaster genome. *PLoS genetics*, 4(12):e1000305.

Padhukasahasram, B., Marjoram, P., and Nordborg, M. (2004). Estimating the rate of gene conversion on human chromosome 21. *American Journal of Human Genetics*, 75(3):386–97.

Papp, B., Pál, C., and Hurst, L. D. (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature*, 424(6945):194–197.

Pâques, F. and Haber, J. E. (1999). Multiple pathways of recombination induced by double-strand breaks in Saccharomyces cerevisiae. *Microbiology and Molecular Biology Reviews*, 63(2):349–404.

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.

Pich I Roselló, O. and Kondrashov, F. A. (2014). Long-term asymmetrical acceleration of protein evolution after gene duplication. *Genome Biology and Evolution*, 6(8):1949–1955.

Plagnol, V., Padhukasahasram, B., Wall, J. D., Marjoram, P., and Nordborg, M. (2006). Relative influences of crossing over and gene conversion on the pattern of linkage disequilibrium in Arabidopsis thaliana. *Genetics*, 172(4):2441–8.

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.

Resnick, M. A. (1976). The repair of double-strand breaks in DNA; a model involving recombination. *Journal of Theoretical Biology*, 59(1):97–106.

Sawyer, S. (1989). Statistical tests for detecting gene conversion. *Molecular Biology and Evolution*, 6(5):526–538.

Schienman, J. E., Holt, R. A., Auerbach, M. R., and Stewart, C. B. (2006). Duplication and divergence of 2 distinct pancreatic ribonuclease genes in leaf-eating African and Asian colobine monkeys. *Molecular Biology and Evolution*, 23(8):1465–1479.

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.

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

Semple, C. and Wolfe, K. H. (1999). Gene duplication and gene conversion in the Caenorhabditis elegans genome. *Journal of Molecular Evolution*, 48(5):555–564.

Sharp, A. J., Cheng, Z., and Eichler, E. E. (2006). Structural variation of the human genome. *Annual Review of Genomics and Human Genetics*, 7:407–442.

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.

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.

Shrivastav, M., De Haro, L. P., and Nickoloff, J. A. (2008). Regulation of DNA double-strand break repair pathway choice. *Cell Research*, 18(1):134–147.

Sonoda, E., Hochegger, H., Saberi, A., Taniguchi, Y., and Takeda, S. (2006). Differential usage of non-homologous end-joining and homologous recombination in double strand break repair. *DNA Repair*, 5(9-10):1021–1029.

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

Stephan, W. (2010). Genetic hitchhiking versus background selection: the controversy and its implications. *Philosophical Transactions Of The Royal Society Of London Series B-Biological Sciences*, 365(1544):1245–1253.

Stephens, S. G. (1951). Possible significance of duplication in evolution. *Advances in Genetics*, 4:247–265.

Sugino, R. P. and Innan, H. (2006). Selection for more of the same product as a force to enhance concerted evolution of duplicated genes. *Trends in Genetics*, 22(12):642–644.

Sung, P. and Klein, H. (2006). Mechanism of homologous recombination: mediators and helicases take on regulatory functions. *Nature Reviews Molecular Cell Biology*, 7(10):739–750.

Szostak, J. W., Orr-Weaver, T. L., Rothstein, R. J., and Stahl, F. W. (1983). The double-strand-break repair model for recombination. *Cell*, 33(1):25–35.

Takuno, S. and Innan, H. (2009). Selection to maintain paralogous amino acid differences under the pressure of gene conversion in the heat-shock protein genes in yeast. *Molecular Biology and Evolution*, 26(12):2655–2659.

Takuno, S., Nishio, T., Satta, Y., and Innan, H. (2008). Preservation of a pseudogene by gene conversion and diversifying selection. *Genetics*, 180(1):517–531.

Teshima, K. M., Coop, G., and Przeworski, M. (2006). How reliable are empirical genomic scans for selective sweeps? *Genome research*, 16(6):702–12.

Teshima, K. M. and Innan, H. (2004). The effect of gene conversion on the divergence between duplicated genes. *Genetics*, 166(3):1553–60.

Teshima, K. M. and Innan, H. (2008). Neofunctionalization of duplicated genes under the pressure of gene conversion. *Genetics*, 178(3):1385–98.

Teshima, K. M. and Innan, H. (2012). The coalescent with selection on copy number variants. *Genetics*, 190(3):1077–86.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073.

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

Verrelli, B. C. and Tishkoff, S. A. (2004). Signatures of selection and gene conversion associated with human color vision variation. *American Journal of Human Genetics*, 75(3):363–375.

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

Waldman, A. S. and Liskay, M. (1987). Differential effects of base-pair mismatch on intrachromosomal versus extrachromosomal recombination in mouse cells. *Proceedings of the National Academy of Sciences of the United States of America*, 84:5340–5344.

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.

Wiehe, T., Mountain, J., Parham, P., and Slatkin, M. (2000). Distinguishing recombination and intragenic gene conversion by linkage disequilibrium patterns. *Genetics Research*, 75(1):61–73.

Williams, A. L., Genovese, G., Dyer, T., Altemose, N., Truax, K., Jun, G., Patterson, N., Myers, S. R., Curran, J. E., Duggirala, R., Blangero, J., Reich, D., and Przeworski, M. (2015). Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *eLife*, 4:e04637.

Winckler, W., Myers, S. R., Richter, D. J., Onofrio, R. C., McDonald, G. J., Bontrop, R. E., McVean, G. A. T., Gabriel, S. B., Reich, D., Donnelly, P., and Altshuler, D. (2005). Comparison of fine-scale recombination rates in humans and chimpanzees. *Science*, 308(5718):107–111.

Winderickx, J., Battisti, L., Motulsky, a. G., and Deeb, S. S. (1992). Selective expression of human X chromosome-linked green opsin genes. *Proceedings of the National Academy of Sciences of the United States of America*, 89(20):9710–9714.

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

Wright, S. (1931). Evolution in Mendelian Populations. *Genetics*, 16(2):97–159.

Yang, D., Goldsmith, E. B., Lin, Y., Waldman, B. C., Kaza, V., and Waldman, A. S. (2006). Genetic exchange between homeologous sequences in mammalian chromosomes is averted by local homology requirements for initiation and resolution of recombination. *Genetics*, 174(1):135–144.

Zhang, J. (2003). Evolution by gene duplication: An update. *Trends in Ecology and Evolution*, 18(6):292–298.

Zhao, Z., Hewett-Emmett, D., and Li, W. H. (1998). Frequent gene conversion between human red and green opsin genes. *Journal of Molecular Evolution*, 46(4):494–496.

Appendix

Hartasánchez DA, Vallès-Codina O, Brasó-Vives M, Navarro A. Interplay of interlocus gene conversion and crossover in segmental duplications under a neutral scenario. Suplementary material. G3 (Bethesda). 2014 Jun 6;4(8):1479–89. DOI: 10.1534/g3.114.012435

Hartasánchez DA, Brasó-Vives M, Fuentes-Díaz J, Vallès-Codina O, Navarro A. SeDuS: segmental duplication simulator. Suplementary material. Bioinformatics. 2016 Jan 1;32(1):148–50. DOI: 10.1093/bioinformatics/btv481

G3 · Genes | Genomes | Genetics

Interplay of Interlocus Gene Conversion and Crossover in Segmental Duplications under a Neutral Scenario

Diego A. Hartasánchez*, Oriol Vallès-Codina*, Marina Brasó-Vives*, Arcadi Navarro* § † ‡

*Institute of Evolutionary Biology (Universitat Pompeu Fabra – CSIC), PRBB, Barcelona, Catalonia, Spain, 08003.
§National Institute for Bioinformatics (INB), Barcelona, Catalonia, Spain.
†Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain.
‡Centre for Genomic Regulation (CRG). Barcelona, Catalonia, Spain, 08003.

DOI: 10.1534/g3.114.012435