Dynamical Aspects of the Evolution of Segmental Duplications in the Human Genome

Dissertation zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.) am
Fachbereich Mathematik und Informatik der Freien Universität Berlin

VORGELEGT VON
ELDAR ABDULLAEV

Berlin, 2022

Erstgutachter: Prof. Dr. Martin Vingron
Zweitgutachter: Prof. Dr. Fyodor Kondrashov
Datum der Disputation: 05.07.2022

Preface

Publications

This thesis is built on a research project that was recently published in BMC Genomics by Abdullaev, Umarova, and Arndt (2021) under the title "Modelling segmental duplications in the human genome". It is available at https://doi.org/10.1186/s12864-021-07789-7. Parts of the text in the thesis were adopted from this research paper. The publication covers our results described in Chapters 4 and partly 5. On the other hand, Chapter 6 contains our yet unpublished results about the reconstruction of duplication events from the SD network and genomic features associated with duplicated regions. We expect to submit this part of the thesis for publication soon.

Acknowledgements

First of all, my work was possible because of doctoral program of the International Max Planck Research School for Computational Biology and Scientific Computing. The Max Planck Institute for Molecular Genetics and the Free University provided me with all facilities for the research I needed.

I would like to thank my supervisor Martin Vingron for his support and great discussions that were organized among students of Computational molecular biology department. Some of his lectures are especially memorable for me: the lecture (during retreat) about what can be considered a science was very inspiring, the one about precision matrices (surprisingly, inverse of a covariance matrix nullifies elements corresponding to conditionally independent variables) was quite similar to what is ordinarily taught in Hogwarts. All TAC committee members (Ralf Herwig, Knut Reinert, Jotun Hein, Martin Vingron and Peter Arndt) along with Georgii Bazykin and the group of evolutionary genomics in MSU were very helpful for my project development. Moreover, I would say, that the second half of my research project (Chapter 6) came out as a result of discussions about "what would be nice to add". This resulted in a separate story about segmental duplications which we hope to publish soon. Thanks to our collaborator Iren Umarova for her help in math related tasks and overall support. I would like to thank all fellow students from the Max Planck Institute for Molecular Genetics for our scientific discussions and good time together.

So much assistance I received from Kirsten Kelleher for which I am extremely grateful. She helped with myriads of organizational questions and supported me in hard times. I sincerely wish her all the best in her new working position.

Most of all, I would like to thank my advisor Peter Arndt for his involvement in the project, his ideas, very fruitful discussions that sometimes took many hours in a row, for sharing his experience (or even wisdom!), especially, when it came up to IT or math related questions, being always in contact, very encouraging in a course of my PhD program and sharing the same Tao of the research we did. He suggested taking a look at some characteristics of systems we studied which I missed or did not consider important. Eventually, many of those observations were vital for our research. Other than that, Peter Arndt also helped me a lot in various non-research related issues which made my work comfortable during the PhD program. I am very grateful for his supervision, it was a great great pleasure working with Peter in a course of these years!

Finally, I would like to thank my family, friends and all who were around and supported me.

Contents

List of Figures

List of Tables

List of Abbreviations

SD - Segmental duplication;
HR - Homologous recombination;
NAHR - Non-allelic homologous recombination;
SDSA - Synthesis-dependent strand annealing;
BIR - Break-induced replication;
LOH - Loss of heterozygosity;
DSB - DNA double-strand break;
FoSTeS - Fork stalling and template switching;
MMBIR - Microhomology-mediated break-induced replication;
NHEJ - Non-homologous end joining;
CNV - Copy-number variation;
SNP - Single nucleotide polymorphism;
MAC - Minor allele count;
UCM - Uniform Copying Model;
PCM - Preferential Copying Model;
WGAC - Whole-genome assembly comparison;
WSSD - Whole-genome shotgun sequencing detection;
WGS - Whole-genome shotgun sequencing;
KMC - Kinetic Monte Carlo;
ABC - Approximate Bayesian computation;
MST - Minimum spanning tree;
CGH - Comparative genomic hybridization;

To Marina, Teymur, Ruslan and Iren;
Hope you'll always be around!

Chapter 1

Introduction

The diploid human genome consists of 23 pairs of chromosomes that encode heritable information about an individual. There are 22 pairs of autosomes and a pair of sex chromosomes that differ in male (X, Y) and female (X, X). Almost exact sets of chromosomes (except for somatic mutations accumulated) are present in all somatic cells of a body. Even though the genomic information is the same, the difference in epigenetic states and regulatory mechanisms effect the cell fate thus leading to distinct cell morphologies. The DNA sequences that comprise genomes vary between unrelated human individuals and between human and other species in their content. Per average, two individual human haploid genomes differ by 4.1 to 5 million variants which cover approximately 20 million basepairs (Consortium et al., 2015). This is close to 0.6% of the overall genome sequence since the human genome is about 3.1 billion basepairs long. About 3 million sites or 0.1% of genomic sequence correspond to single nucleotide polymorphisms or differences in one nucleotide. The rest are structural variants that include: duplications, deletions, insertions (when DNA sequence is added without copying), translocations (when DNA fragment is transferred from another locus), inversions (when an order of nucleotide sequence changes to the opposite) and complex events. Long duplications or deletions are called copy-number variations or CNVs. Copy-number variations correspond to about 0.4% of difference between individuals. Long duplications (longer than 1 kbp) that are fixed in the human population are called segmental duplications or SDs. In this thesis we analyze properties of segmental duplications in the human genome.

When a new mutation appears in a population, it, eventually, either reaches fixation (presence in all individuals) or vanishes. Three main forces affect this process: the mutation rate, random drift (stochastic effects of sampling) and selection. In this thesis we study large genomic duplications that are fixed in the human population. Some of them can be explained by positive selection (when, for example, a functional gene is duplicated) or by increased mutation rate resulting in recurrent duplications, however, in most cases, it is hard to find out the reason for fixation. In terms of duplication mechanisms it is also not always easy to predict the responsible one. Thus the question of why a specific duplication is present, which sequences get copied, and where such copies are inserted into a genome is a complicated one. In this thesis, a similar question was studied: why duplications that we see in the human genome are distributed in such a way relative to each other. We handled this task by studying possible dynamic scenarios of duplications expansion in genomes.

1.1 Research objective

This thesis presents an analysis of long genomic duplications, known as segmental duplications (or SDs), in the human genome. The main focus is put upon the dynamic aspects of expansion of segmental duplications. The distribution of segmental duplications in the genome is highly non-uniform. But why segmental duplications happen non-uniformly in the genome? Are there any universal "rules" behind the process of SD expansion or is it just a sum of independent mechanisms acting in different parts of the genome. If there are some, can we model this process? To answer those questions we analyzed special distributions and sequence similarities between SDs to reject certain naive scenarios for SD evolution and came up with a minimal model for the expansion of SDs, which complies with the observed data. The general processes of SD propagation were rarely modelled mathematically. In this thesis we attempted to change this situation, and, eventually, accumulated new facts about SD evolution. We suggest a universal propagation model for segmental duplications in the human genome. Moreover, we predicted characteristics of duplicated regions associated with high duplication rate. Our research project gives an insight into segmental duplications dynamics both by modelling duplication process and by analyzing genomic features affecting duplication rates.

1.2 Thesis outline

In the second chapter we introduce segmental duplications (SDs) and present known facts about them. It starts with a historic overview, discussion about different sources of redundancy in the human genome, detection methods, dynamic properties of SD expansion in the genome and the role in gene duplications and human genome evolution. This thesis describes an interdisciplinary project, which includes mathematical algorithms applied to answer biological questions. Thus a chapter dedicated to the network theory was added to the thesis. The Chapter 3 gives a brief glossary of complex networks related terminology. In Chapter 4 our results on dynamical aspects of SD evolution are presented and discussed. Human genome segmental duplications were studied as a graph which allowed to suggest a model of SD expansion in the genome. The same analysis was performed in genomes of other species - results are presented in Chapter 5. Then in Chapter 6 the duplication events are reconstructed from the complex network of segmental duplications and genomic features associated with actively duplicating regions are suggested. Conclusions and a summarizing discussion can be found in Chapter 7.

Chapter 2

Genomics Background

2.1 Human Genome Project and genomic sequence redundancy

It makes sense to start our introduction with the Human Genome Project (HGP). This seems relevant when introducing segmental duplications or SDs (the main object of our research project). Even though earlier studies observed large genomic duplications often associated with disease (Tomlinson et al., 1994; Eichler et al., 1997; Wong, Royle, and Jeffreys, 1990), still large-scale annotation of segmental duplications became possible when the reference genome was ready. Technically, the HGP was the moment of a phase transition in computational biology which affected its paradigms: the tools we use, the questions we ask, the scale of data we work with etc. So even if not considering our topic of interest, one would not call that a beginning *in medias res*.

Historically, the Human Genome Project was established in 1990, it was planned for 15 years with an overall costs of $3 billion. It ended up two years earlier than expected in April 2003 (50th anniversary of Watson and Crick discovery) and costed less than expected ($2.7 billion dollars (1991)). The first draft version of human genome was announced even earlier in 2000. The project was done by the International Human Genome Sequencing Consortium (IHGSC) which united scientists from 20 sequencing centers in US, China, France, Germany (including the Max Planck Institute for Molecular Genetics), UK and Japan. In parallel, a private biotechnology company called Celera Genomics was competing with IHGSC to produce a first sequence of the human genome.

Eventually, approximately 92% of human genome was sequenced except for heterochromatic parts which are mostly comprised of highly repetitive subtelomeric and pericentromeric regions. The estimated sequence quality of the hg17 reference (2004) exceeded 99.99% accuracy and only 341 assembly gaps remained unresolved (IHGSC, 2004).

The way human genome was sequenced by IHGSC is known as hierarchical shotgun and it differs from regular NGS whole-genome sequencing that became widespread later. The sequencing pipeline started with genomic DNA fragmentation into large pieces ($150 - 200$ kbp) which were further cloned in bacterial artificial chromosome (BAC) vectors (Fig. 2.1). These BACs were transformed in *E. coli* culture, so that, as a result, each clone carries one BAC which is amplified with the means of bacterial DNA replication machinery. Such a collection of genomic fragments is called BAC library.

Since the price of sequencing in late 90^s was much higher than nowadays, only most informative BACs were picked for it. However, it is not a straightforward task when you do not know the sequence. To disentangle this vicious circle, additional sources of information were considered. Sequence-tagged sites (STSs) are short ($200 - 500$ bps) nucleotide sequences that are present in a genome in one copy and location of a corresponding locus is known in advance. Presence of specific STSs in BACs was detected by PCR amplification. This allowed to tag BACs to specific genomic regions and thus pick non-redundant set of BACs covering largest fraction of the genome (Fig. 2.1). Other methods like fluorescence in situ hybridization (FISH) and DNA fingerprinting were also applied to map BAC library fragments on chromosomes, but less extensively (IHGSC, 2001). Genomic inserts from selected BACs were fragmented into smaller DNA pieces ($\sim$ 2 kbp) which were than sequenced with a Sanger-based sequencing machine and assembled (IHGSC, 2001). For resulting contigs and scaffolds the BACs of origin were known which in its turn were approximately mapped on chromosomes. This knowledge, substantially eased the process of genome assembly from a set of contigs. At late stages of the project $147,480$ assembly gaps were closed and complicated regions of the assembly were finalized (IHGSC, 2004).

The resulting genome is approximately 3.1 billion base pairs long and includes approximately 21 thousand protein-coding genes. This was one of the surprising observations of the HGP, because it was expected that there are much more (over 40 thousand) genes. The fraction of protein-coding sequences in the genome was also surprisingly low (1.5%), while the rest of the genome is comprised of high-copy repeats, introns, regulatory regions and more that altogether was prematurely stamped as "junk" DNA (Fig. 2.3a). The level of redundancy in human genome was higher than expected, in particular, it was estimated that segmental duplications cover $\sim$ 5% of human reference genome (Bailey et al., 2002). In the next sections we will discuss origins and types of genomic redundancy and a large fraction of the text is dedicated to segmental duplications. In conclusion of this section something rather obvious has to be said: it is hard to overestimate an importance of the Human Genome Project and especially its impact on bioinformatics, biotechnology and medicine. It is quite illustrative that if we iterate through big human-related bioinformatics projects that came after: the HapMap, ENCODE, TCGA, 1000 Genomes project, Telomere-to-Telomere Consortium project – none of them would be possible without proper assembly of the human genome.

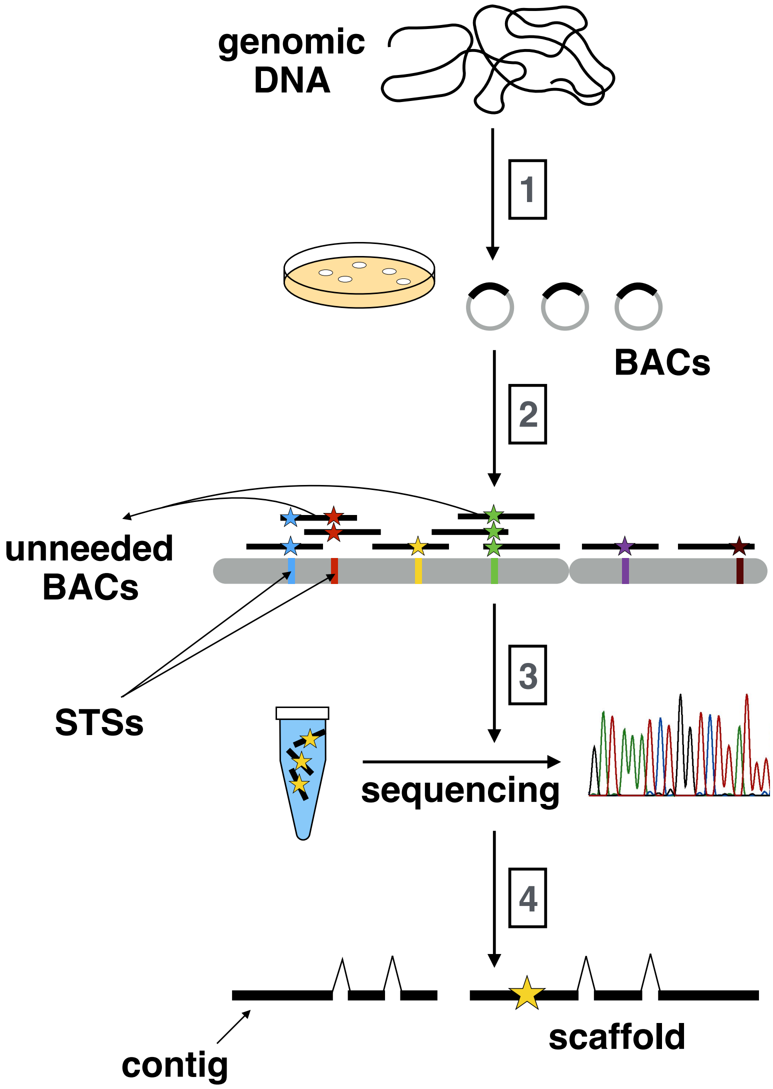


FIGURE 2.1: **The workflow of human genome sequencing by IHGSC.** At the first step the genome was fragmented and cloned into BACs. On the second step approximate locations on chromosomes were determined for DNA fragments based on sequence-tagged sites or STSs (drawn as coloured stars). Non-redundant BACs were further picked for sequencing and, on the last step, contigs corresponding to specific BAC are assembled.

2.1.1 Perfect matches and stick-breaking process

We would like to start a discussion about redundant sequences in human genome from stretches of perfect matches observed in it. This view does not imply knowledge on duplication mechanisms or biology behind those sequences, but study them in a light of mathematical models. When we align genomes of various species against themselves we can observe that not only one trivial alignment is present (a perfect match between full genomic sequence and its copy), but also additional shorter perfect matches (identical sequences) are observed between non-allelic loci of a genome. Surprisingly, a length distribution of such perfect matches is dramatically different from what we expect if we imagine a genome as a random string of nucleotides. In this case the expected length distribution is exponential. It comes up if considering mismatched basepairs as events originating from the Poisson distribution, while distance between such events (or length of a perfect match) is distributed exponentially then. However, in real life the length distribution follows a power-law $n(l) \propto l^{\alpha}$ where $n(l)$ is a number of perfect matches of length l (Fig. 2.2). Power-law slopes are observed both in genomes with masked and unmasked high-copy simple repeats. The exponent values α are quite consistent among eukaryotic species: $\alpha \sim -3$ as it was first observed by Gao and Miller (2011). This distribution of perfect matches was further explained by Massip and Arndt (2013) with a so-called "stick-breaking" model. This model was inspired a model from the field of polymer chemistry (Kuhn, 1930). Based on stick-breaking process, duplications of constant length L happen in a genome with the rate γ. Single-nucleotide substitutions happen in a genome in parallel with the rate μ and fragment perfect matches observed between duplications into shorter matching intervals until complete dissolution of duplicated sequence. Thus we can assume that over time the number of perfect matches of specific length l can either decrease when substitutions disrupt them or grow when substitutions break longer sequences of length $\hat{l}$ where $\hat{l} \in [l+1;\ L]$ into fragments l and $\hat{l} - l$. Simulations and analytical solutions for such a process predict that in equilibrium state the following distribution is present $n(l) = Cl^{-3}$ where $C = \frac{\gamma L}{\mu}$ which agrees with observed distribution of perfect matches in the masked version of human genome. Curiously, the value of C is close to one for human genome which in a context of stick-breaking model suggests that probability of a genomic position to be duplicated is approximately equal to its probability of being substituted (Massip and Arndt, 2013). The stick-breaking process with modifications was further successfully applied to explain other match length distributions, such as exponent $\alpha = -5$ for matches between protein–coding genes, $\alpha = -4$ for retroduplicated loci and k-mer distributions in *Alu* repeats sequences (Sheinman et al., 2016; Massip et al., 2015; Massip et al., 2016). Simple mathematical model of random duplications and substitutions is sufficient to explain perfect matches length distribution in genomes. However, to get a deeper look into nature of redundant sequences in the human genome we have to consider different types of repeats with their mechanisms of propagation.

Finally, we wanted to mention that this story is illustrative to the fact that prediction of simple universal mechanism for a process could arise even in presence of multiple facts on the complexity of its specific cases. There are many possible mechanisms involved in long genomic duplications and short repeats propagation in a genome. Accumulation of knowledge about it is often an accumulation of multiple complex scenarios which makes the overall picture more detailed, but less clean as a whole. Sometimes it is reasonable to make a step back from a frontier of such analysis by reduction or even deconstruction of considered objects in order to make a

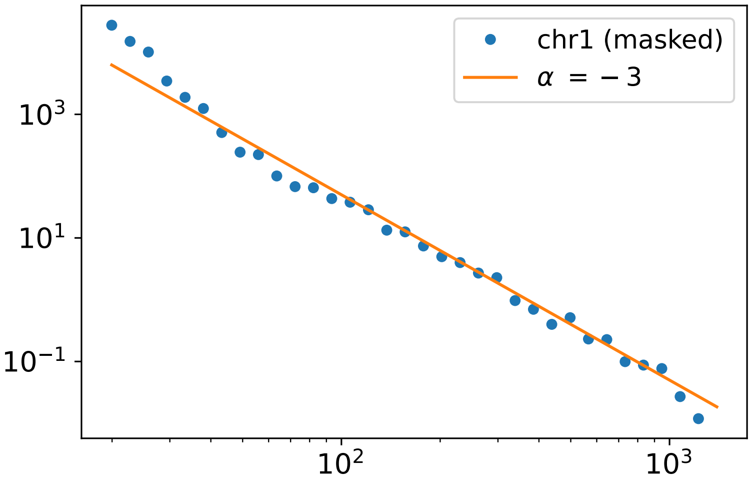


FIGURE 2.2: **The perfect matches length distribution.** Plotted on a log-log scale with the logarithmic binning. The perfect matches were detected by aligning repeat masked chromosome 1 of the *hg19* reference against itself with MUMmer tool (Kurtz et al., 2004). One can see a clear power-law distribution with the slope $\alpha = -3$ as discussed in the main text.

forward step with a generalized prediction which naturally arises then. This view is also consonant with what was done in our research project on segmental duplications.

2.1.2 Repeat classes

Redundant sequences in the human genome are mostly represented with high copy repeats. Based on annotation of a human reference these repeats cover around half of it (Fig. 2.3a) (IHGSC, 2001). When this fraction was estimated based on analysis of high-abundance k-mer clusters it grew up to about two thirds of the genome (Koning et al., 2011). Unknown repeats and highly divergent ones are included in this set along with previously annotated ones. This large group of redundant sequences can in most cases be assigned to different repeat classes. Based on a distribution throughout a genome repeats are divided into tandem ones: those where copies are located in arrays one after another, and interspersed: when copies are located in a genome more or less at random positions. As we will see further, the difference between two groups is more dramatic than just positioning of copies relative to each other.

Tandem repeats

Tandem repeats are mostly represented with so-called satellite repeats. The name "satellite" originates from early experiments on genomic DNA centrifugation in a density gradient. Low-complexity satellite repeats have biased nucleotide frequencies in comparison with the rest of the genome thus corresponding DNA is moving in a gradient test tube as a separate "satellite" band. Satellite DNA is composed of large arrays of tandemly repeating non-coding sequences. Most of it is concentrated in centromeric and telomeric regions of the genome, moreover, satellite repeats are important for centromere and telomere formation along with organization of heterochromatin (Shatskikh et al., 2020). The main mechanism responsible for satellite repeats formation is a DNA polymerase slippage that happens during replication in the S phase of cell cycle (Tautz and Schlötterer, 1994). Highly repetitive sequences form DNA loops that either increase or decrease resulting number of repeat units added by a polymerase. Satellite repeats are classified into three groups: microsatellites, minisatellites and satellites in a narrow sense.

- **Microsatellites** or short tandem repeats are arrays of small (2 - 9 bps) units that are located in multiply loci in the human genome. Arrays of microsatellite repeats are highly mutable and differ in their number between different individuals. Overall, microsatellites cover 3% of the human genome sequence (Subramanian, Mishra, and Singh, 2003). Changes in microsatellite composition can effect gene regulation, expression or lead to disease, for instance, there are several triplet expansion disorders, such as fragile X syndrome, Friedreich's Ataxia or Huntington's disease (Pearson, Nichol Edamura, and Cleary, 2005; Bidichandani, Ashizawa, and Patel, 1998). The genome-wide difference in microsatellite composition accounts for $\sim 10 - 15\%$ of heritable variation in gene expression (Gymrek et al., 2016). Human telomeres themselves are composed of tandemly repeating blocks: $\ldots TTAGGG - TTAGGG - TTAGGG \ldots$ which are classified as microsatellites (Thakur, Packiaraj, and Henikoff, 2021).

- **Minisatellites** represent tracks of repetitive DNA where longer units (usually, 10 - 60 bps long) are organized in tandem arrays (typically, repeated 5 - 50 times). Similarly to microsatellites, these repeats are highly polymorphic in human population and are enriched around centromeric and telomeric regions. Minisatellites are often associated with fragile genomic sites and genomic rearrangement hotspots (Vergnaud and Denoeud, 2000).

- **Satellite repeats** are sometimes considered as a separate class. Satellite repeats include several tandem repeat groups of variable length which constitute centromeres, pericentromeric and subtelomeric regions (as we said, telomeres are usually included in microsatellites). Centromeres of human chromosomes consist of specific α-satellites which are one of the longest (repeat unit of 170 bps) and most widespread among satellites ($\sim$ 10% of all repeats number) (Aldrup-Macdonald and Sullivan, 2014). Subcentromeric regions differ among human chromosomes and are predominantly occupied by satellites I-III and β-satellites (Thakur, Packiaraj, and Henikoff, 2021).

Curiously, satellite repeats are fairly well conserved in terms of their sequence, while their copy number changes are common even between individuals of the same specie (Plohl, Meštrović, and Mravinac, 2012). This characteristic of microsatellite repeats is exploited in DNA fingerprinting for population genetics, crime investigations and kinship tests. Also, microsatellites were widely used as genetic markers allowing to locate genes responsible for a specific phenotype or disease based on co-inheritance observations, however, later SNPs became more applicable for this role.

Interspersed repeats

Overwhelming fraction of interspersed repeats is represented by transposable elements (TEs) which exploit specific enzymes for their propagation in a genome. Transposable elements are mobile genetic elements that encode enzymes needed for their propagation. Some transposable elements encode all needed enzymes and thus are able to independent transposition in a genome while other TEs miss some enzymatic machinery and exploit one encoded in other TEs. These two groups are called autonomous and non-autonomous TEs, respectively. Transposable elements and tandem repeats are quite distinct: units of transposable elements are longer, satellite repeats do not encode proteins and have low complexity repetitive sequence which it is not the case for TEs. Based on a mechanism of propagation, transposable elements are divided into two classes.

- **Retrotransposons** or class I TEs need an RNA intermediate for their propagation. Common mechanism for retrotransposition include the following generalized steps. Retrotransposons are first transcribed with one of RNA polymerases. A reverse transcriptase enzyme is than translated from resulting RNA intermediate or exploited from the host cell in order to synthesize complementary DNA or cDNA. This DNA fragment is inserted back into genome with the help of integrase or endonuclease enzymes. This mechanism is similar to retroviral life cycles because class I elements and retroviruses have common evolutionary origin while transitions between these two groups and horizontal gene transfer is possible (Koonin, Dolja, and Krupovic, 2015; Hayward, 2017). Two big groups of autonomous class I TEs include: LTR retrotransposons which are structurally characterized by long terminal repeats and non-LTR retrotransposons, which are usually referred to as LINEs (long interspersed elements). Non-autonomous ones are represented with short interspersed elements (SINEs) which in primate genomes are dominated by *Alu* repeats. Without going into details, all listed groups have rather different mechanisms of propagation, patterns of insertion, evolutionary origin and encode different gene sets. Overall, class I transposable elements cover a huge (more than 40%) fraction of the human genome (Fig. 2.3a) (Gregory, 2005). Among them *Alu* ($\sim$ 11%) and LINE-1 ($\sim$ 17%) are the most abundant representatives (IHGSC, 2001; Batzer and Deininger, 2002).

- DNA **transposons** or class II transposable elements are mobile elements that use single- or double-stranded DNA intermediate in a process of transposition. Class II TEs recruit transposase enzyme to recognize genomic sequence of transposon, cut one from the genome and paste it elsewhere. This mechanism is referred to as "cut-and-paste" in contrast to "copy-and-paste" one of retrotransposons. Overall, DNA transposons account for 3% of genomic sequence. Most of class II transposable elements observed in the human genome were active early in primate evolution ($>$ 37 mya), while currently there are no active DNA transposons in our genome (Pace and Feschotte, 2007).

Transposable elements shaped human genome throughout primate evolution by affecting both linear distances between variable genetic elements and the spatial structure. This influenced various aspects, from regulation of expression, genes evolution and DNA repair mechanisms to large-scale evolutionary effects, such as speciation process, origin of cellular life, sexual reproduction (for example, germline cells defence from TEs), virus evolution and so on (Koonin, 2016; Tóth et al., 2016; Serrato-Capuchina and Matute, 2018). What is more relevant to our project, transposable elements create hotspots for genomic rearrangements. As we will see in further sections, burst of segmental duplications in primate evolution was associated with excess of *Alu* repeats. Finally, mobile elements are associated with multiple human diseases, such as: hemophilia A and B (L1 retrotransposon insertion into anti-hemophilic factor (AHF) gene), cause of colon cancer (disruption of APC gene by LINE-1), porphyria (insertion of *Alu* into PBGD gene) etc. (Miki et al., 1992; Kazazian et al., 1988; Mustajoki et al., 1999).

2.1.3 Segmental duplications

Segmental duplications (SDs) or low copy repeats (LCRs) are long duplications of genomic sequence that are fixed in a genome. A fixation in population genomics is a process when specific allele frequency reaches 100% or, in other words, it gets present in all individual genotypes in a population. By the common definition segmental duplications are longer than 1 kbp with the level of sequence identity higher than 90%. Let's dive into the given definition. Firstly, the threshold for the length and the level of sequence identity between copies are imposed artificially, because of technical reasons. The process of sequence alignment of a human reference against itself was extremely time consuming in early 2000^s (weeks of computation) so there were some restriction for reported hits. In parallel, alternative approach for SD prediction (WSSD) that we will discuss in detail later used another technical threshold ($\geq$ 95% and $\geq$ 15 kbp), which, eventually, did not become conventional.

Based on neutral sequence evolution, duplications with sequence identity $>$ 90% appeared around 40 mya or after, which roughly corresponds to the split of the New

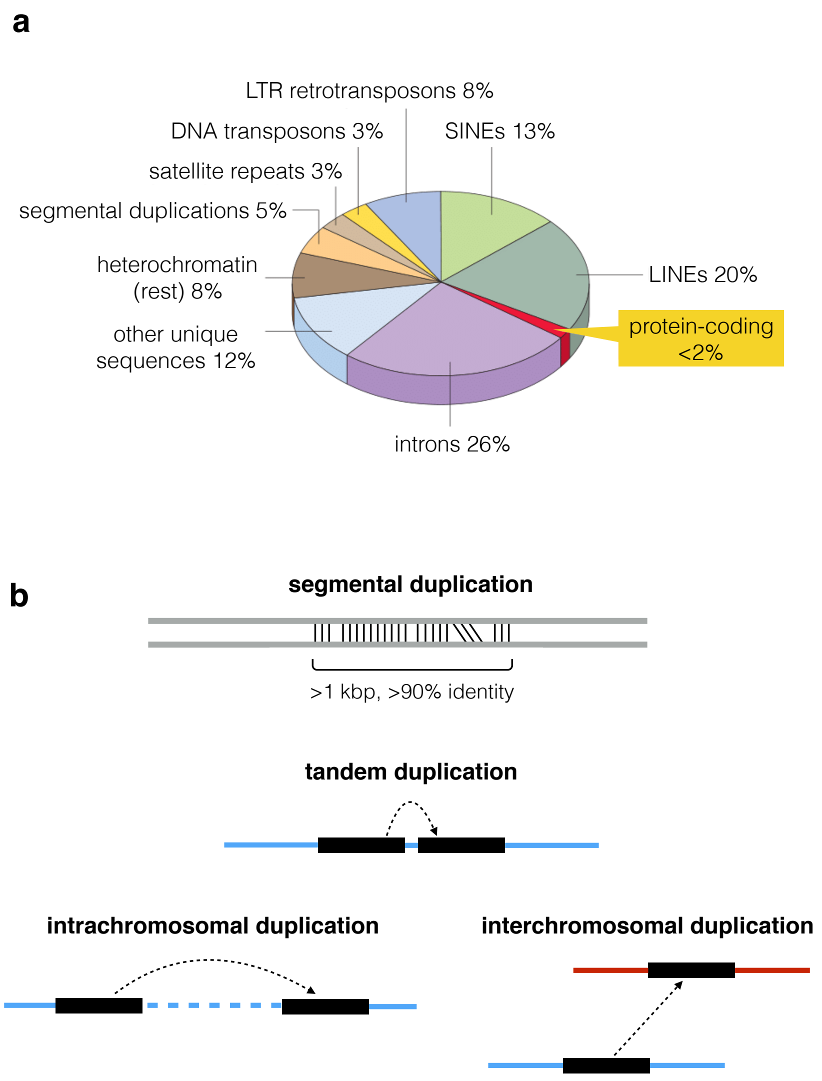


FIGURE 2.3: **Redundant sequences in the human genome.** The piechart (**a**) represents types of genomic sequences and their fraction in the human genome. One can see that protein-coding exons cover minority of the genome, while about half of the genome is represented with redundant sequences. Adopted from Gregory (2005). **b**. Segmental duplications are recognized as longer than 1 kbp alignments of $> 90\%$ sequence identity between non-allelic loci of a genome. Based on location of duplicated sequences relative to each other SDs are classified into inter- and intrachromosomal, while intrachromosomal ones can be tandem or not (interspersed). Coloured lines correspond to different chromosomes.

and Old World monkeys (Bailey and Eichler, 2006). This means that SDs that originated during the burst of duplication activity early in hominid evolution are included in SD annotation. It also means that human shares annotated SDs with other Old World monkeys, but not with less related species.

Segmental duplications are evolutionary fixed in a genome according to definition, otherwise duplications are classified as copy-number polymorphic or CNVs.

Technically, we can not say for sure if duplication is fixed in a population, especially when it comes to early annotations when only one human genome was sequenced. This could lead to reconsidering some SDs as CNVs when more individual samples are studied. However, the nature of CNVs and SDs is similar, that is why some conclusions about SDs are drawn from observations made on CNVs. Now, as we discussed the definition of SDs we can talk about ones in more details.

2.2 Segmental duplications analysis

2.2.1 Segmental duplications detection methods

Detection of segmental duplications in a genome is not a straightforward task: all methods have their own limitations and even now, after about 20 years of SD research, we can not say that a golden standard method has been established. Widespread application of long-read sequencing could, in theory, solve this problem in the future when this technique becomes more accessible, but, as for now, most of our data on SDs comes from other sources. In this section we will focus on two main computational approaches for SD prediction: WGAC and WSSD, and briefly mention other methods.

Among experimental methods for detection of SDs, fluorescence *in situ* hybridization (FISH) is a widespread technique for validation of SDs. Usually a subset of SDs predicted computationally is validated with FISH to assess an accuracy of predictions. In FISH fluorescently labeled DNA probes are hybridized against metaphase chromosomes in order to find all complementary DNA sequences. This allows to detect long duplications and to map them approximately on chromosomes. Alternatively, copy-number variable loci can be detected with comparative genomic hybridization arrays (aCGH) with high resolution. Two samples of genomic DNA: test and reference one, are first labeled with red Cyanine 5 (Cy5) and green Cyanine 3 (Cy3) fluorophores, respectively. Fragmented genomic DNA from two samples in equal quantities is mixed and cohybridized against DNA microarray covered with single-stranded DNA oligonucleotides complementary to genomic loci of interest. The fluorescence colour or the Cy3/Cy5 fluorescence ratio is proportional to the copy-number ratio of a specific piece of genome in test and reference samples. Increased copy-number in test sample would lead to red fluorescence, while decreased one would result in green light. This method allows to detect relative copy-number changes between individuals of the same or related species, but not the location of copied DNA sequences in a genome. Finally, quantitative PCR (qPCR) can be applied to measure copy-number of specific genomic region in order to validate possible duplication.

The first method is called whole-genome assembly comparison or WGAC. It is based on the most intuitive solution for segmental duplications prediction: in a nutshell, a reference genome is split into genomic intervals (400 kbp) that are further aligned against each other (Bailey et al., 2001). Segmental duplications are detected as long stretches of alignment observed between different genomic intervals. In WGAC approach high-copy simple repeats are first excluded from genomic segments prior to alignment. This step is called "fuguization", it makes WGAC more sensible to SDs riddled with repeats that would otherwise be missed. Alignment of genomic intervals is performed via BLAST with consequent refinement of alignment ends after returning repeat sequences back (Altschul et al., 1990).

Another approach termed whole-genome shotgun sequencing detection (WSSD) distinguishes unique and duplicated sequence on the basis of the depth of read coverage and the level of sequence identity of reads alignment to a reference (Bailey et al., 2002). The idea is the following, if we allow reads to map in multiple positions on a reference genome, those sites with increased copy number in the genome will have higher depth of read coverage, moreover, we expect that read coverage grows proportionally to copy-number. The level of identity of reads to reference alignment is used as another evidence, because copies of a duplicated region often differ in their sequence composition. For instance, segmental duplications, by a common definition, differ by up to 10% of their sequence. Thus long genomic regions where decreased read mapping quality is observed together with increased depth of read coverage are considered as potential loci involved in duplications.

The WSSD approach is appropriate to detect SDs with sufficient level of conservation and length (usually, $\geq$ 15 kbp and $\geq$ 94%, respectively). Lower sequence identity level could affect mapping of reads originating from diverged copies, while short regions of increased coverage depth could be unrecognized. This is where WGAC algorithm has an advantage over WSSD. However, one important advantage of WSSD approach is that it does not depend on quality of a given reference genome assembly, which, as we will further discuss in detail, is often quite low. The WGAC approach can only recognize duplications that are present in a reference genome and can not validate their correctness per se. On the other hand, even SDs collapsed by assembler into single genomic region can be recognized by WSSD. Moreover, WSSD algorithm can be applied in a cross-specie manner when WGS reads from one specie are mapped to a reference genome of closely related specie. Both WGAC and WSSD have their own strong and weak sides. That is why *de novo* SD annotations are often done with both methods applied to make SD predictions more accurate (for example, Bailey et al. (2004) or Liu et al. (2009)).

Long-read sequencing technology has a great potential for producing *de novo* genome assemblies of a quality exceeding *GRCh38* reference (Miga et al., 2020). For example, long reads are more likely to cover breakpoints of SDs with deeper insight into surrounding unique DNA sequence, which allows to proper anchoring of duplication copies. Long reads also allow resolving redundant low-complexity sequences surrounding SDs which were often unresolved and left as assembly gaps before. However, there is one major technical issue that limits opportunities of long-read sequencing. The error rate of both PacBio and Oxford Nanopore Technologies

(ONT) long reads is quite high ($\sim$ 15% of bps) which is comparable to the level of sequence identity between copies of SDs (Weirather et al., 2017). Such a high error rate limits opportunities of genome assemblers to recognize separate copies of a duplicated region, which leads to collapses where paralogous sequences are erroneously merged into one. For example, it was found that only about 30% of SDs were properly resolved in recent *de novo* long-read assemblies *CHM1* and ONT assembly of *NA12878*, where proper resolution means that duplication is correctly placed in the genome and at least 50 kbp extended into unique sequence on both sides of SD (Vollger et al., 2019).

The Segmental Duplication Assembler (SDA) is a computational tool that is often applied to solve this problem (Vollger et al., 2019). It takes advantage of polyploid phasing in order to resolve collapsed regions in assemblies. First, genomic intervals with elevated read coverage are detected and studied as potential collapsed regions. For each of them paralogous sequence variants (PSVs) are suggested as those SNVs that appear at the same threshold as unique sequencing depth. This threshold allows to distinguish paralogous or allelic variants from read errors. Then the graph is constructed where PSVs represent nodes, edges are added if a pair of PSVs is observed in the same read (thus reads themselves define paths in the graph). Clusters of connected nodes identified in this PSV graph correspond to paralogous copies of a genomic region. Reads belonging to these clusters are reassembled again independently in order to get separate sequences of copies (Fig. 2.4). The SDA approach successfully resolved many collapsed sequences in long-read assemblies.

2.2.2 Segmental duplications in the human genome

Segmental duplications were first annotated in the draft human reference genome in 2001 with the WGAC algorithm (Bailey et al., 2001). The first estimate of a duplicated sequence content was at 13.2% of the genome with a low level of recall by FISH. It was clear that this fraction is inflated by the misassemblies of the reference. Application of the WSSD method in 2002 re-estimated this fraction to be at around 5% which agrees with a nowadays knowledge on SDs (Bailey et al., 2002). Anyway, the amount of segmentally duplicated sequence was higher than what was expected before the HGP. Moreover, it was found that segmental duplications account for higher fraction of nucleotide difference between human and chimp than single-base substitutions ($>$ 2.7% and 1.2%, respectively) and strongly contribute to the variation between human individuals (Cheng et al., 2005; Redon et al., 2006). Based on positioning relative to each other segmental duplications are classified into inter- and intrachromosomal ones. Interchromosomal SD is the one where copies are located at different chromosomes while copies of intrachromosomal SD are on the same. The last group includes tandem SDs, those where copies are located adjusted to each other ($<$ 1 Mbp between copies). Human genome is enriched with interchromosomal and interspersed SDs (non-tandem ones) in comparison with other

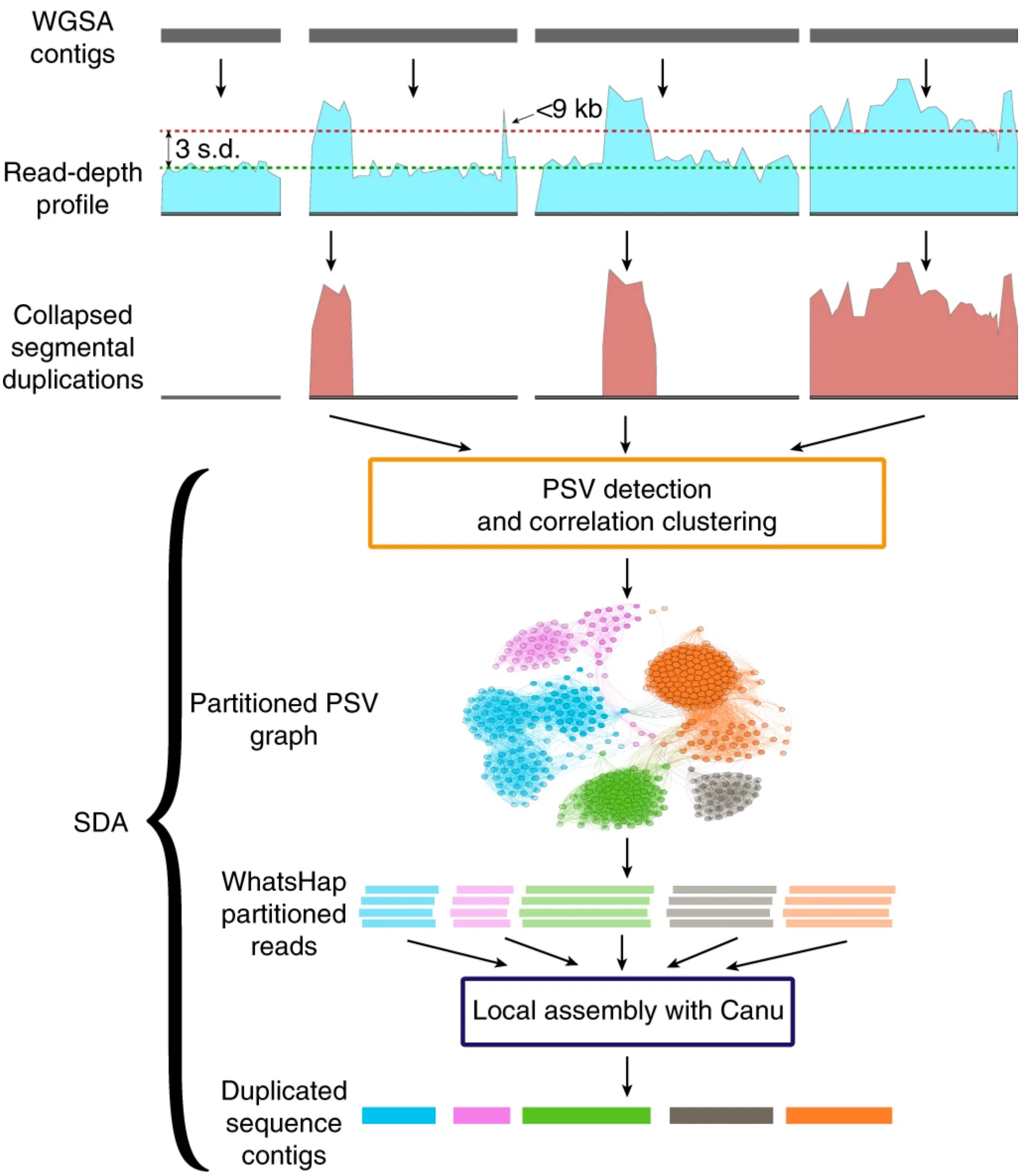

WGSA contigs
Read-depth profile
3 s.d.
<9 kb
Collapsed segmental duplications
PSV detection and correlation clustering
SDA
Partitioned PSV graph
WhatsHap partitioned reads
Local assembly with Canu
Duplicated sequence contigs

FIGURE 2.4: **The schematic pipeline of the SDA tool.** Regions suspicious for assembly collapse are defined based on read coverage depth profiles. Green and red dotted lines represent normal coverage of unique sequence and a threshold for elevated one (normal coverage +3 s.d.), respectively. PSVs are identified and the PSV graph is constructed as described in the main text. Clusters of PSVs are detected; then reads are partitioned accordingly and reassembled independently. Colours on the scheme illustrate these clusters and resulting contigs. The source of the figure: Vollger et al. (2019).

non-hominid genomes which are dominated by tandem duplications (Bailey and Eichler, 2006).

SDs are distributed very non-uniformly in the human genome (Fig. 2.5). Hotspots for segmental duplications are concentrated in pericentromeric and subtelomeric regions of chromosomes. From $\sim$ 150 Mbp of genomic sequence covered by SDs, 31% is located in pericentromeric regions, 2% in subtelomeric ones and 67% are interstitial (i.e. lie between pericentromeric and subtelomeric regions) (Koszul and Fischer, 2009). The dominating mechanisms of duplications are distinct in all these areas. For example, segmental duplications in subtelomeric regions are mainly formed by NHEJ or non-homologous end joining. NHEJ is one of the pathways for repair of double-strand breaks in DNA. In contrast to homology-mediated repair this mechanism does not rely on long stretches of homologous sequences. Erroneous NHEJ can lead to unequal translocations of telomeric fragments between chromosomes which result in accumulation of interchromosomal duplications. Detailed analysis of 41 subtelomeric duplicons found that 92% of breakpoints were consistent with the NHEJ mechanism (Linardopoulou et al., 2005). It also agrees with the fact that pericentromeric regions are enriched with interchromosomal SDs (2.7-fold enrichment over the genome average) (Bailey et al., 2001), especially high sequence identity ones (Fig. 2.6).

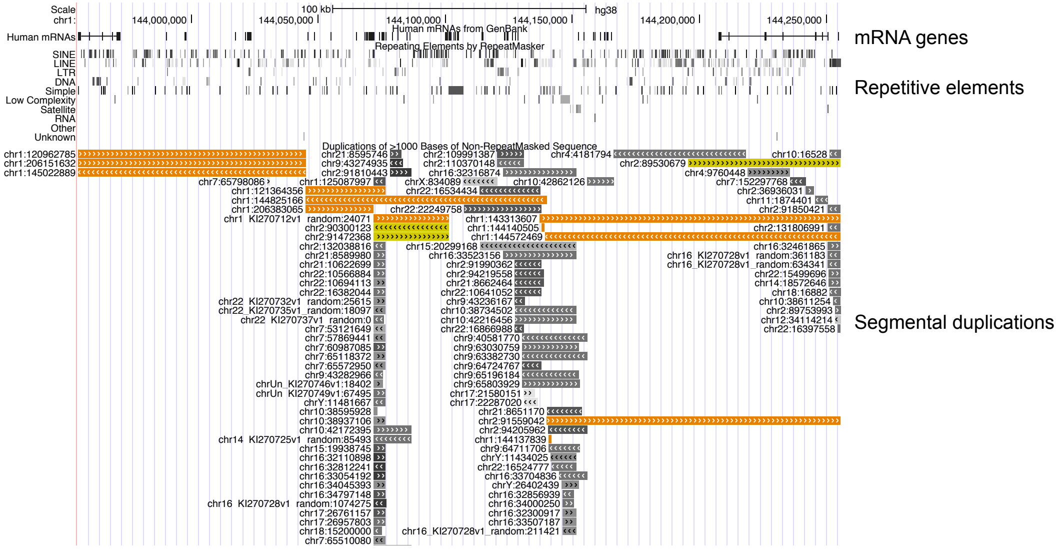


FIGURE 2.5: **The UCSC genome browser annotation of an example genomic region (chr1:143,955,418 - 144,255,418).** The illustrated region shows a complex structure of a genomic locus enriched with SDs. Every element of the "Segmental duplication" track represents a long alignment observed between corresponding genomic region and its copy located elsewhere (indicated by the coordinates aside of each element). Colours of the alignments represent the level of sequence identity: light to dark grey (90 - 98%); yellow (98 - 99%); orange (> 99%). Repetitive elements like LINEs, SINEs, LTRs etc. are usually much shorter and present in high number of copies in the genome (tracks on the top). These were added to illustrate the difference between SDs and high-copy repeats.

Even stronger enrichment of SDs is observed in pericentromeric regions: both inter- (4.5-fold) and intrachromosomal duplications (3.1-fold) are enriched there in comparison with the rest of the genome (Bailey et al., 2001). Within closest to centromere 0.5 Mbp there are 6 times more interchromosomal duplications than intrachromosomal ones while further away this fraction gradually descends to the genome average. Big fraction of duplicated sequence in pericentromeric regions can be traced back to ancestral fragments of interstitial euchromatin (She et al., 2004). To explain this, special "two-step model" was proposed on how pericentromeric regions evolved in terms of duplications (Eichler et al., 1997). In a nutshell, on the first step, euchromatic segments are copied into pericentromeric loci where these blocks are concentrated. On the second step, resulting complex loci are copied with juxtaposed duplicons from diverse euchromatic regions in them. This process gives rise to mosaic SDs in pericentromeric regions (Fig. 2.10a). We will come back to this model in later sections.

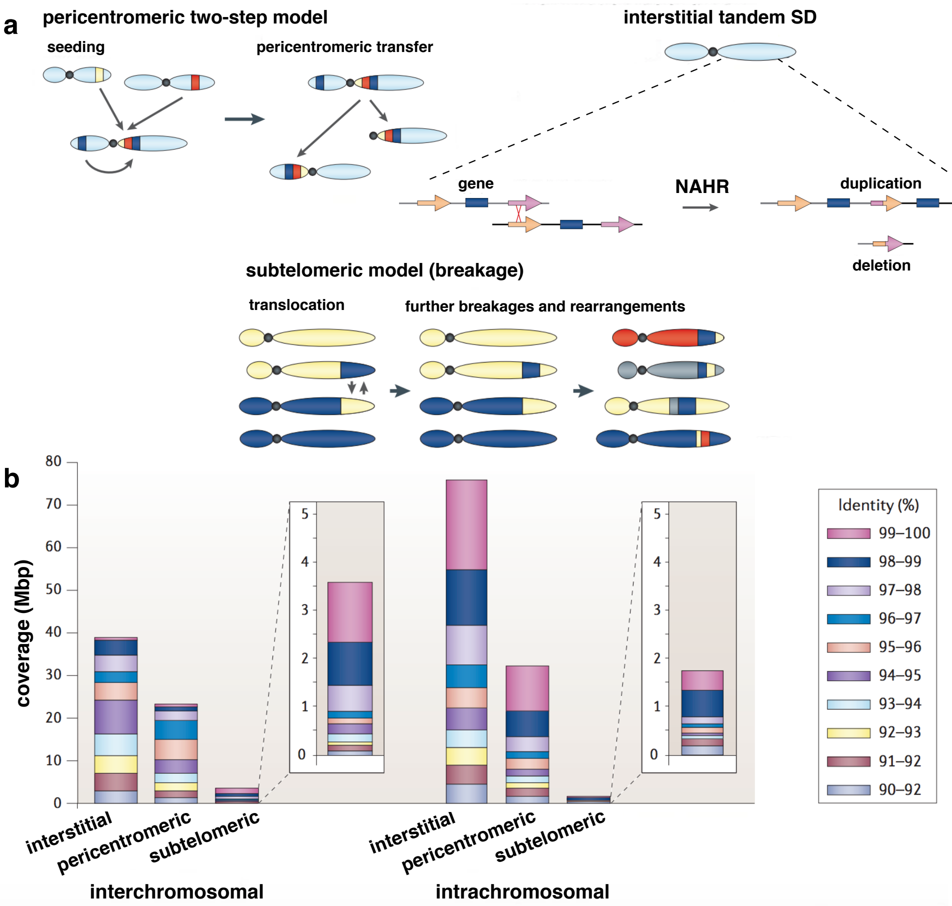


FIGURE 2.6: **Segmental duplications mechanisms overview. a**. Subtelomeric SDs are formed by NHEJ-mediated translocations between chromosome arms. Pericentromeric SDs are often complex and appear as a result of two-step model of SDs expansion. Euchromatic fragments from multiple genomic locations are first seeded in one locus and then duplicated as mosaic SDs in pericentromeric parts. Interstitial SDs, covering rest of chromosomes are formed by homology mediated mechanisms as illustrated on the figure where NAHR creates tandem duplication. **b**. Quantitative characteristics of SDs in different chromosome parts. *Y* axis represents overall length of SDs, coloured bars represent sequence identity levels. One can see that both in pericentromeric and interstitial regions intrachromosomal duplications are, on average, younger (have higher sequence identity level), while the opposite is true for subtelomeric ones. Adopted from Bailey and Eichler (2006)

Interstitial segmental duplications are dominated by intrachromosomal duplications in comparison with the regions discussed. When comparing with non-primate genomes, the average distance between copies of intrachromosomal SDs is higher in human, because of lower frequency of tandem SDs. Interstitial segmental duplications are often associated with repeated sequences at the boundaries. Repeated sequences make a locus more susceptible to all homology-driven mechanisms of segmental duplications formation: both non-allelic homologous recombination (NAHR) and DNA polymerase slippage during replication or repair. *Alu* repeats are an important group of sequences that make genomic regions unstable. An analysis of a sequence content of pairwise SDs boundaries showed that *Alu* repeats are significantly enriched there (covering 24% of the sequence as compared to 10% elsewhere). The enrichment was observed for younger *Alu* subfamilies (*AluY* and *AluS*) which emerged recently in primate evolution, whereas the oldest primate subfamily (*AluJ*) showed no enrichment (Bailey, Liu, and Eichler, 2003). It is suggested that the burst of *Alu* retroposition activity during the primate evolution around 35 – 40 mya sensitized the ancestral human genome for *Alu*–mediated segmental duplication events both via NAHR and ploymerase slippage mechanisms. Based on sequence identity levels and phylogenetic reconstruction (we will discuss one in detail in Chapter 5), the peak of SD expansion was observed early in hominoid evolution (Samonte and Eichler, 2002). *Alu*-mediated segmental duplications are mostly characterized by long-range translocations, because of relatively universal distribution of *Alu* repeats in the genome.

Alternatively, segmental duplications themselves are the source of genomic instability leading to further duplications and rearrangements. Segmental duplications are long highly identical sequences that increase the probability of non-allelic homologous recombination or template switches during DNA replication. One widespread scenario of a duplication driven by previously duplicated sequences is illustrated at (Fig. 2.6a). NAHR between copies of a tandem SD can lead to deletion along with duplication of the copies number. It explains an elevated duplication dynamics of tandem SDs and, as it was noted by Seymour Fogel in 1983: "It is conceivable that the rate limiting step in tandem gene amplification is the very first event leading to the initial duplication. Once two or more copies of the segment are present, the likelihood of tandem amplification is much greater." (Fogel et al., 1983). Findings in comparative genomics analysis showed that recent SDs tend to happen in already duplicated regions (this phenomenon is known as "duplication shadowing") (Cheng et al., 2005). CNVs in the human population also preferentially happen in already duplicated loci (Korbel et al., 2007; Kim et al., 2008).

2.2.3 Mechanisms for segmental duplications

In this section we will discuss the mechanisms of segmental duplications in detail, their abundance and conditions needed. In general, the mechanisms were formulated as theoretical models which were tested based on indirect experimental evidence. CNVs accumulation was tested in various stress conditions or cell cultures where crucial enzymes of DNA repair were nonfictional. An information on the mechanism responsible for duplication can also be extracted from DNA sequences in its junctions. For example, NHEJ leaves characteristic "scars" proximal to breakpoints: short deletions or insertions. NAHR-mediated duplications are surrounded by long highly homologous sequences which provided the source for homologous recombination. Replication-mediated mechanisms often imply shorter homologous sequences. These events can be detected by microhomology stretches at CNVs/SDs junctions.

The CNVs can be classified into recurrent and non-recurrent based on their ability to be formed as independent evolutionary events. Recurrent events are the same CNVs that appear independently in unrelated individuals, i.e. have the same size, junctions coordinates and sequence content. On the other hand, non-recurrent CNVs are those appearing in the same locus, but having nothing in common in terms of sequence characteristics. At least 40 genomic disorders are caused by recurrent rearrangements and more than 70 are attributed to non-recurrent events (Vissers and Stankiewicz, 2012). The same terminology can be applied for SDs, where recurrent means homoplasic event: when the same duplications are fixed in several species, but originated independently in ancestral lineages. We introduce this difference here, because different mechanism are prone to either recurrent or non-recurrent events.

Genomic rearrangements appear when erroneous DNA replication, DNA repair or homologous recombination take place. The mechanisms of rearrangements are further classified based on homology needed to proceed: homology mediated or micro-/non-homology mediated ones. Homologous recombination (HR) is a mechanism involved in repair of dsDNA breaks and in processes of genomic exchange between chromosomes (gene-conversion and crossing-over). HR between chromosomes does not lead to mutagenesis if the same chromosomal region is involved. On the other hand, non-allelic homologous recombination can result in duplications, deletions or inversions of genomic sequence. When a double-strand DNA break is detected in the interphase, the HR is applied to fix the break based on homologous chromosome or sister chromatid sequence template (Krejci et al., 2012). HR can go either via double Holliday junction DSB repair (the pass that could lead to crossing over or gene conversion) or synthesis-dependent strand annealing (SDSA) - both these mechanisms can result in genomic rearrangements (an illustrated scheme of HR can be found in Appendix chapter). Non-allelic HR between tandem repeats in direct orientation leads to duplication along with deletion (Fig. 2.7a). NAHR between inverted repeats results in inversion of the fragment between repeat units.

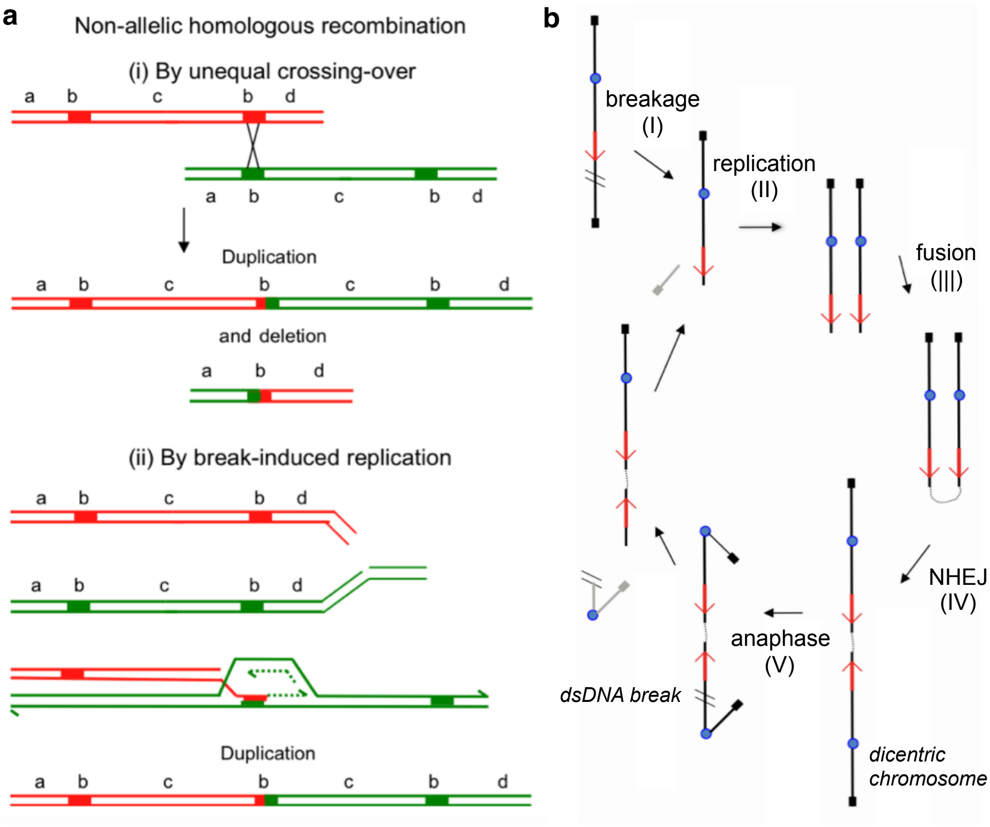


FIGURE 2.7: **Duplication mechanisms not associated with DNA polymerase template switching.** **a**. An example of non-allelic homologous recombination between tandem repeats that leads to the duplication and deletion of genetic material. The break-induced replication (BIR) is a pathway to resolve one ended dsDNA breaks. A missing arm of a chromosome is restored by replicating homologous chromosome sequence. Similarly, when tandem repeats are not properly aligned during BIR the genomic region between two repeats can be duplicated or lost. The scheme (**b**) illustrates the breakage–fusion–bridge cycle. Loss of telomeric parts by two chromosomes can lead to them being merged into one dicentric chromosome by the NHEJ (telomeres play the protective function). Such a situation can happen when a dsDNA break is not fixed upon replication thus leading to a pair of chromatids without telomeric caps. During anaphase the dicentric chromosome is split by two kinetochores pulling it in opposite direction. Random double-strand break leads to genomic DNA translocation, while the chromosome without a telomere can initiate another round of the breakage–fusion–bridge cycle. Adopted from Hastings et al. (2009).

When HR happens through both either double Holliday junction or SDSA scenarios it requires presence of both fragments of DNA that were split by a double-strand DNA break. However, HR happens differently when only one fragment is present. Then a whole missing chromosome arm is restored based on homologous or sister chromosome. Thus one-ended dsDNA break repair results in prolong regions of loss of heterozygosity (LOH) (Hastings et al., 2009). This type of DNA repair is often associated with DNA replication, because when helicase reaches a nick in DNA template a single dsDNA fragment is released (Fig. 2.7). This replication error is resolved with break-induced replication (BIR) pathway (Hastings et al., 2009). When homologous sequence is encountered the BIR pathway allows DNA replication to proceed from that place, however, non-allelic stretches of homology lead to genomic rearrangements in this case.

In any of discussed scenarios the homologous recombination requires long stretches of homology. Minimum of 134 - 232 bps of matching sequence for mammalian species is needed for HR (Waldman and Liskay, 1988). That is why NAHR-mediated SDs/CNVs are often found in regions that underwent segmental duplications in a past. The rate of HR between homologous sequences correlates positively with the sequence identity level, length and G/C content and correlates negatively with the distance between copies (Dittwald et al., 2013; Carvalho and Lupski, 2016). The fact that HR-based duplications can appear as a result of interaction between specific loci (those with sufficient sequence homology) explains the fact that recurrent CNVs mostly resulted from NAHR. Recurrent duplications happen in unstable genomic sites where precise positions of breakpoints are determined by homology borders and the sequence is unambiguously defined by the interval between them (Carvalho and Lupski, 2016).

Another group of duplication scenarios requires short or no homology. As opposed to NAHR, non-homology mediated scenarios give rise to non-recurrent CNVs, because various sites can be involved in genomic rearrangement, which itself can be complex and associated with extra deletion/insertion "scars" at junctions (Carvalho and Lupski, 2016). The first such mechanism is NHEJ which, as we discussed in the previous section, is responsible for subtelomeric SD propagation. Non-homologous end joining is one of the pathways involved in repair of dsDNA breaks (along with HR and microhomology-mediated end joining). Homologous recombination is a preferred mechanism for DNA breaks repair, because of its accuracy, however, if homologous template is not accessible (usually, during the G1 phase), NHEJ is a mechanism for dsDNA breaks repair. Even microhomology of 1 - 3 bps is sufficient for NHEJ to merge two DNA fragments: hybridize, fill in the gaps and ligate them (Pannunzio et al., 2014). If there are several double-stranded breaks present in a moment: NHEJ can result in chromosome arm translocations because of its low specificity. We observe many such translocations and associated SDs in subtelomeric regions of mammalian genomes (Fig. 2.6), because exchange of telomeres between non-homologous chromosomes is less deleterious than large-scale chromosome arms translocations. More complex scenario of NHEJ-mediated genomic rearrangements was described in cancerogenesis. The so-called "breakage–fusion–bridge cycle" happens when loss of telomeric region happens in a pair of chromosomes at the same time (Fig. 2.7b) (Murnane, 2012). For example, this condition appears when a chromosome loses one of telomeres and is replicated whereas dsDNA break is not fixed. As a result, a pair of chromatids lacking telomere on one side appears, which are not protected from fusion by the cap of telomere repeats. These chromatids will likely be fused into one dicentric chromosome by the NHEJ pathway. During mitosis the centromeres of dicentric chromosome are pulled in opposite directions with a new dsDNA break emerging at random position. New cells will inherit the chromosomes with missing telomeres that could start a new cycle. Since the dsDNA break happens at random position during the anaphase of mitosis, this event is associated with large translocations.

Aphidicolin is a reversible inhibitor of DNA replication in eukaryotic cells. It blocks the replicative DNA polymerases $Pol\alpha$ and $Pol\delta$ (DeFilippes, 1984). When CNVs characteristics were analyzed in cells subjected to aphidicolin, it was observed that 65% of CNVs had microhomologies or no homologous sequences at breakpoints, showing a limited impact of homologous recombination (Arlt et al., 2009). This illustrates the fact that non-homology mediated replicative mechanisms can be responsible for the CNVs formation. One such widespread scenario is called "replication slippage" (Hastings et al., 2009). It happens in replication when a DNA polymerase slips over the DNA template to another position and continues the complementary strand synthesis from another locus (Fig. 2.8a). Depending on a slippage direction this can result either in tandem duplication (backward transition) or deletion of a genetic material (forward transition). The slippage not necessarily happens over the same DNA template. The process when a replication fork stalls and a DNA polymerase jumps to a ssDNA fragment belonging to another replication fork to continue the synthesis is called fork stalling and template switching or FoSTeS (Fig. 2.8b) (Lee, Carvalho, and Lupski, 2007). Earlier we discussed the homology mediated break-induced replication pathway, however, one-sided dsDNA breaks that can appear during DNA replication can be resolved without HR. The suggested mechanism is called microhomology-mediated BIR (MMBIR). It is similar to BIR in its general principles, however, because of difference in enzymatic machinery it does not rely on long homologous sequences and utilizes microhomologies for DNA replication to proceed (Fig. 2.8c) (Hastings et al., 2009). Downregulation of *Rad51* enzyme in stress conditions which is essential for homology detection in HR, hinders BIR pathway and leads to MMBIR where 3′ end overhang anneals to a ssDNA template sharing minor homology. Low processivity polymerase *Pol32* carrying out replication in MMBIR is prone to multiple template switches until a fully functional replication fork is established (as illustrated at Fig. 2.8c). This leads to a complex nature of resulting non-recurrent CNVs which share microhomologies at junctions.

2.3 Insights into segmental duplications propagation and evolution

This section covers several research projects that aimed to suggest some generalized models of SD evolution. With the focus on quantitative or dynamic sides of it. In two

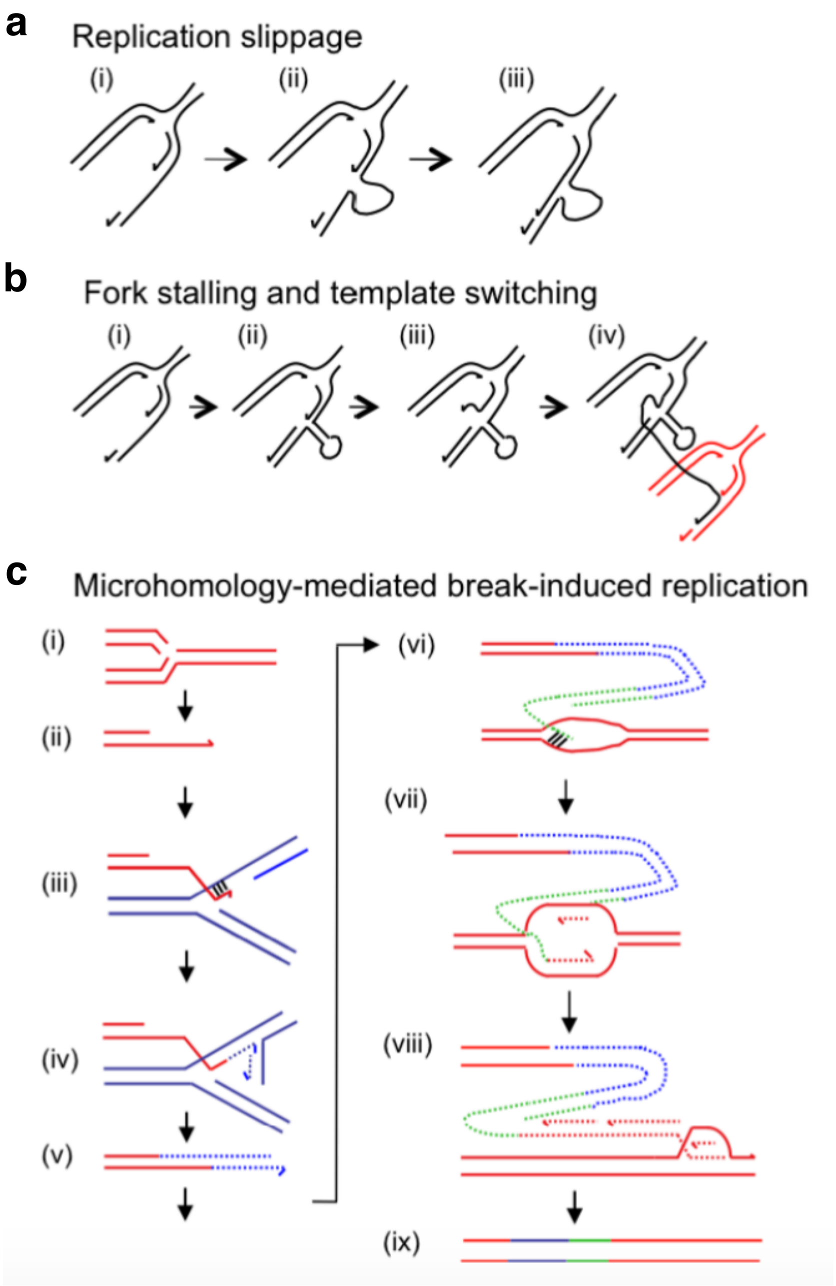


FIGURE 2.8: **Duplication mechanisms associated with DNA polymerase template switching.** Several scenarios of non-homology mediated rearrangements. **a**. Replication slippage over the template leads to deletions or insertions. **b**. Stalling of a replication fork can lead to temporary DNA polymerase template switching to a DNA template in another replication fork (mediated by microhomology). This results in genetic material being copied from the non-homologous locus. **c**. The microhomology-mediated BIR (MMBIR) differs from the standard BIR because it does not rely on homologous sequence. It happens that DNA polymerase undergoes several rounds of template switching until reaching the correct homologous template. DNA fragments copied from multiple loci form complex duplication event composed of sequences of different ancestry. The figure source: Hastings et al. (2009).

projects a modified version of an A-Bruijn graph was utilized to study SDs ancestral origin, evolution and expansion principles.

2.3.1 Markov process for segmental duplications propagation

The following research project proposes a dynamic model of the duplication process as early as in 2005 (Zhou and Mishra, 2005). Several mammalian genomes were studied, but we will focus on the human genome related findings. It was observed by earlier studies that *Alu* repeats, especially, young subfamilies are enriched in flanking regions of SDs in comparison to the genomic background (Bailey, Liu, and Eichler, 2003). However, it is hard to answer quantitatively, which fraction of SDs were formed by repeat-mediated homology mechanisms, because the presence of repeats in SD flanking regions does not necessarily mean involvement of repeats in their formation. A special Markov process was proposed to answer this question.

Only those SDs copied once, not overlapping other SDs and non-tandem ones were studied to be sure that alignment corresponds to an actual duplication event. All such SDs were classified according to presence of repeats in flanking regions. The expectation is that repeats have to belong to the same family, be on the same side, same orientation and distance from the breakpoints if they mediated the duplication. SDs are denoted as (+/+), (+/-) or (-/-) if such repeats are present in both, one or no flanks respectively. However, in the course of evolution, repeats can be added or lost in the flanks, thus making the picture more complicated. More than that, SDs were binned according to the level of sequence divergence (8 divergence or age groups). So the Markov process included transitions between various states. Overall, there were $3 * 8 = 24$ combinations of flanking repeats and sequence identity states (Fig. 2.9). To measure the probabilities of transitions between states in one time step Δt realistic parameters were included from other literature sources: mutation rate, repeats insertion rate, mutation rates in repeats etc. (see Zhou and Mishra (2005) for details). The only unknown parameters that the authors wanted to infer were h - the fraction of SDs originated in repeat-mediated homologous recombination and the fractions f_x attributed to different repeat families x. For example, based on the model, hf_{Alu} of all SDs originated via *Alu*-mediated recombination.

The inference of unknown parameters is possible under two assumptions. Firstly, all parameters were conserved over a long evolutionary period and, secondly, the stationary state was reached in the system. An expected distribution of SDs in a stationary state was calculated analytically (see Zhou and Mishra (2005) for details). The parameters (h, f_x) were fitted by minimizing χ^2 statistic between observed and expected distributions of SD states. Cross-validation was applied to evaluate the accuracy of models. Only *Alu* and L1 repeats were considered. The second group did not show any significant enrichment in flanking regions. Moreover, the value hf_{L1} was not significantly different from zero, while hf_{Alu}, on the other hand, was. The highest accuracy of the model was achieved when repeat-mediated recombination explains $h \sim 30\%$ of SDs, while $hf_{Alu} \sim 12\%$ of SDs are *Alu*-mediated ones. Finally,

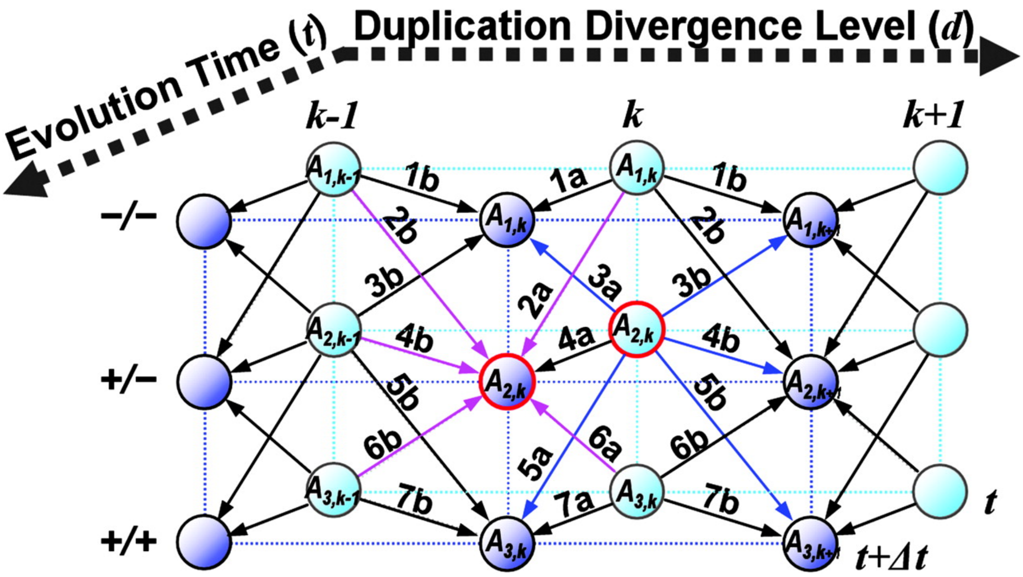


FIGURE 2.9: **The Markov process scheme.** The Markov process that we discussed in the main text includes 3 states for flanking sequences: (-/-), (-/+), (+/+) and 8 levels for sequence identities between copies ($k-1$, k, $k+1$, etc.). Each circle on the scheme represents a state, arrows represent possible transitions that can happen in a time interval $[t, t+\Delta t]$ with various probabilities. When the stationary state is reached, SD distribution over all conditions stays unchanged, i.e. all states in-flows and out-flows are equal. The figure source: Zhou and Mishra (2005).

it was observed that DNA helix stability was lower, while flexibility was higher in flanking regions without repeats. It was suggested that lower mechanic stability of DNA in flanking regions might be one of the factors providing duplications when SD can not be explained by recombination.

Some common suggestions used in this research are now outdated. The duplication rates and other parameters varied substantially in a course of primate evolution, the reference genome quality improved substantially over last decades, non-homology mediated mechanisms were not considered and lower mechanic stability of DNA helix in SDs breakpoints can not explain the rest of duplications. However, this was an early attempt to model the process of segmental duplications expansion in the genome. Considering the Markov process formalization of SD evolution allowed authors to quantify the impacts of duplication mechanisms.

2.3.2 A-Bruijn graphs and core duplicons

One approach to systematically study SDs in their complexity was suggested at (Jiang et al., 2007). A modified version of A-Bruijn graph was suggested as a way of representation for complex events. All alignments of SDs were split into duplicons which are continuous segments of synteny not interrupted by any breakpoints (see Fig. 2.10a for illustration). It makes sense to introduce our notion of a duplicated region here: a duplicated region is a genomic interval that covers a maximal set of overlapping alignments (i.e. duplicated region equals to a union of all overlapping alignment intervals). If a new SD happens in a genomic region covered by one or several SDs in most cases the number of duplicons will increase by 2. This happens because every new breakpoint divides previous duplicons if it falls between the borders or adds a duplicon outside of the farmost breakpoint (if it falls outside of a duplicated region). Only when a new breakpoint hits an earlier alignment breakpoint no increase is expected there. The A-Bruijn graph was constructed in the following way: edges correspond to duplicons (uninterrupted segments), nodes are breakpoints. Then segmental duplication (same as duplicated region) is a path through the graph; the path includes only two nodes if SD and duplicon borders match. If a segmental duplication covers n junctions, this path covers n+2 nodes.

Overall there were 11,951 non-redundant duplicons of length $>$ 100 bps. For each of these duplicons an ancestral locus was predicted by comparative genomics. The logic is the following: among all possible ancestral loci for a specific duplicon, the one that shares the longest homologous syntheny block in an outgroup specie is likely the ancestral region (see Fig. 2.10b). This method allowed to establish 4,692 ancestral duplicons out of 11,951 duplicon segments. The comparative FISH and validation with known datasets proved the consistency of predicted ancestral loci. The analysis of duplicons distribution showed that chromosomes 1q, 7, 9p, 10q, 15q, 16p, 17, 19, 22q, X and Yq are substantially enriched with both ancestral and derivative duplicons, while 2, 3p, 4, 5q, 6q, 8q, 12 and 18q ones are depleted (Jiang et al., 2007). To some extent it can be explained by more active intrachromosomal duplication activity in the first group of chromosomes.

Special attention was paid to those duplicated regions where multiple duplications happened in a course of evolution. Overall, 437 duplicated regions were termed complex duplication blocks when included more than 10 duplicons. To answer the question on how these regions evolved, the duplicon-wise sequence divergence from an ancestral locus, their coordinates in the genome and pairwise similarities between complex duplication blocks were analyzed. Hierarchical clustering of complex duplication blocks (based on shared duplicons) showed that there are 24 clusters present, 10 and 14 of which are dominated by inter- and intrachromosomal events, respectively. Detailed analysis of ancestral duplicon composition identified 14 duplicons characteristic for specific clusters. These were termed "core" duplicons and defined as those present in $>$ 67% of complex duplication blocks within a cluster (Fig. 2.10c). It was suggested that core duplicons are associated with evolutionary important regions. For example, the genes embedded in 4 out of 14 core duplicons are associated with human gene innovations. The fraction of RefSeq genes and spliced ESTs in core duplicons was about twice higher than in non-core duplicons, however, little enrichment over unduplicated genome parts was observed. In further studies the maximum parsimony evolutionary history was modelled for duplicons and the concept of core duplicons was revisited (Kahn, Hristov, and Raphael, 2010). Some non-core duplicons turned out to be as promising in defining mosaic

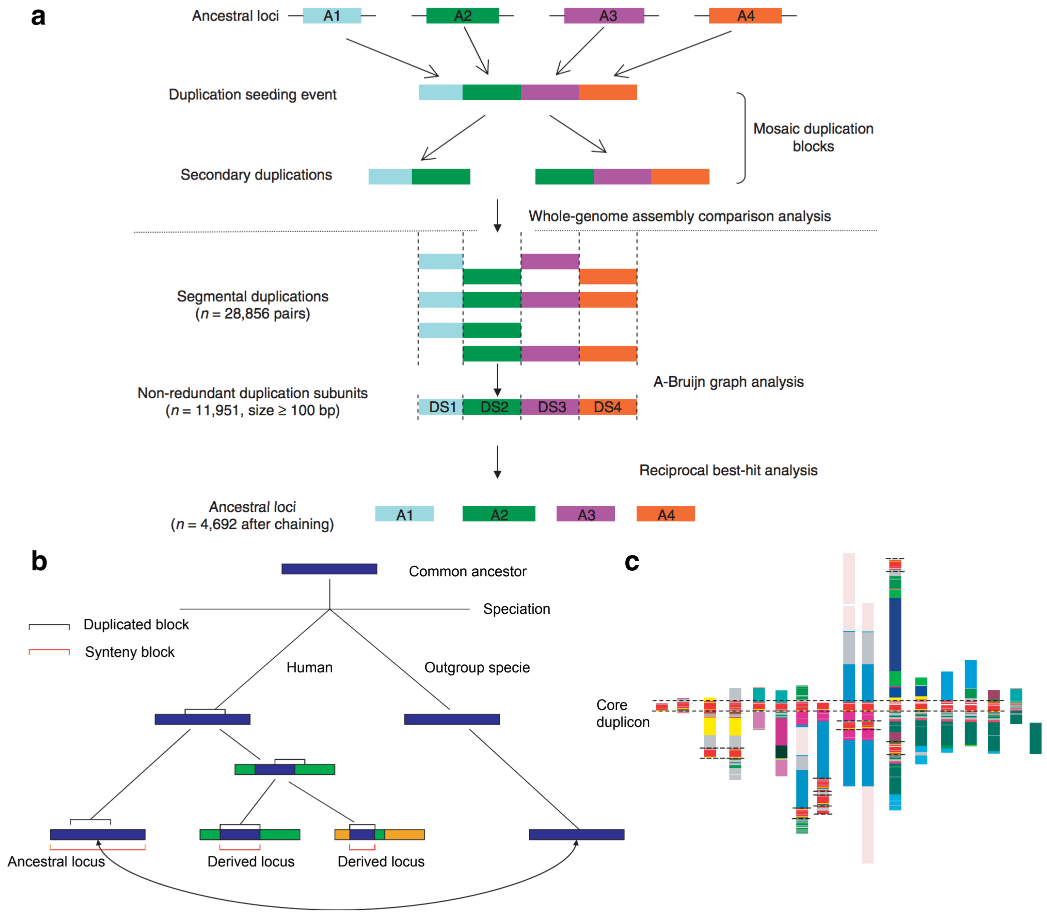


FIGURE 2.10: **The workflow and approaches to study mosaic SDs.** **a**. The scheme illustrates the two-step model for mosaic SDs formation and the overall pipeline of analysis from Jiang et al. (2007). First, a complex region is formed by duplications from ancestral loci (so-called seeding events), then mosaic SDs are duplicated from it. Colored blocks represent duplicons of different ancestry, breakpoints are dotted lines on the scheme. When all duplicons are detected an ancestral locus for each one is predicted by reciprocal best hit method as illustrated at (**b**). Several duplications in a specie of interest resulted in one genomic region being present in three copies (marked with blue brackets). By aligning those regions to an outgroup specie genome where no duplications happened, one can detect an ancestral locus in the specie of interest. It is present as the longest homologous synteny block. **c**. A set of complex duplicated regions that share a core duplicon. Coloured blocks represent duplicons of different ancestry. Adopted from Jiang et al. (2007).

duplications clusters as the core ones.

In this analyzes complex alignments observed in duplicated regions were divided into blocks not interrupted by breakpoints (duplicons). This formalization simplified the task of systematic ancestry reconstruction otherwise only possible for specific loci with intensive manual inspection (Locke et al., 2005; Horvath et al., 2005; Lupski and Stankiewicz, 2005). Global predictions of how ancestry and non-ancestry duplicons are distributed, how they cluster relative to each other, reconstruction of duplicon-wise phylogeny etc. - many such questions can be answered with this formalization. Probably, the most prominent observation was that some duplicons (core duplicons) are involved in duplications especially often which likely points out to their evolutionary importance. This research illustrates one of the ways we can study general principles of SD expansion.

2.3.3 Mosaic model

Here we will discuss another use of an A-Bruijn graph modification to systematically study SDs (Pu, Lin, and Pevzner, 2018). The main focus of the research is on the SDquest tool which allows fast prediction of SDs in genomes. It utilizes techniques based on k-mers and A-Bruijn graphs to detect SDs, even those falling below the conventional threshold of $>90\%$ sequence identity. It might as well be discussed in the section dedicated to methods of SD prediction, however, this method is more specific than widespread WSSD and WGAC. Secondly, we are more interested in the way segmental duplications were mathematically formalized to draw some conclusions about their evolution.

As we said, the SDquest tool allows fast characterization of SDs even those originated earlier than 40 mya. When defining SDs as duplications longer than 500 bps and $>70\%$ sequence identity it turned out that 6.05% of *hg19* reference genome is covered with SDs in comparison with 5.2% for conventional thresholds. Let's discuss the algorithm principles in detail.

- A preliminary identification of segmental duplications is carried out as a search for repetitive k-mers. Repetitive k-mers are those present in several copies in a genome. The length of a k-mer is picked so that it is unlikely to see more than one copy at random sequence of a genome length (by default $k = 25$). Copies of segmental duplications share matching sequences thus can be detected by increased number of repetitive k-mers in comparison with not duplicated regions. After filtering for high-copy repeats repetitive k-mers are identified in a genome. In the human genome more than 19 million distinct repetitive k-mers were identified, 90% of which fall in known SD regions.

- Two distinct repetitive k-mers are called d-paired if the distance between them is less than d (default $d = 500$ bps). In the graph where repetitive k-mers are vertices, d-paired k-mers are connected by edges, putative SDs are detected as connected components. There were 20,009 such putative SDs detected in the human genome after filtering for length and repetitive k-mers frequency.

- Putative SDs are aligned all against all to validate them and refine the breakpoints. Then pairwise alignments are merged into mosaic SDs (synonym of our duplicated region) if overlap. There were 16,231 mosaic SDs detected.

- Similarly, the A-Bruijn graph is constructed by the SDquest with minor modifications. Mosaic SDs are divided into shared alignment blocks (or SD-blocks)

> not interrupted by breakpoints in any mosaic SD. The notions of SD-block and duplicon from the previous section are very similar. SD-blocks are edges in the A-Bruijn graph, vertices correspond to clusters of breakpoints. This graph is similar to the one described in the previous section, however, at Pu, Lin, and Pevzner (2018) more attention is paid to its formation or, in other words, the network growth. The resulting graph represents a footprint of SD evolution. Segmental duplications and duplicated regions are paths in the graph, connected components are independent units of evolution: SD-blocks belonging to different connected components share no common duplication events.

Large branching connected components in the A-Bruijn graph correspond to "bursts" of duplications with multiple loci involved. Linear connected components (simple tree graphs without branches) represent those cases where likely all sequences involved in duplications are from the same ancestry region. So the A-Bruijn graph topology can give some clues on segmental duplications evolution which can not be seen from alignments coordinates alone. There are 4002 connected components in the human genome, 2836 of them are trivial SDs made of one SD-block, 75 of the other 1166 are composed of more than 10 SD-blocks. There is a node of a degree 137 and two edges with multiplicities 262 and 232. These are the signs of duplication hotspots or even emerging common repeats.

Curiously, out of 1166 non-trivial components 169 included cycles (see an example at Fig. 2.11a). These cycles can not be explained by duplications only and therefore more complex evolutionary scenarios were considered (Fig. 2.11b). A cycle can be caused by an SD-block present in several copies in SDs from the same connected component. This type of cycles are observed in 91 out of 169 cyclic components. Alternative scenario suggests that either SD jumped into an already duplicated region or deletion happened inside of a duplicated region. These cases explained cycles in 74 out of 78 (78 = 169 - 91) components. The most complicated explanations characterized last 4 components. These cycles resulted from duplications of genetic material from extrachromosomal circular DNA elements (ecDNA) also known as amplisomes (Raphael and Pevzner, 2004). This type of genomic translocations via ecDNA is characteristic for tumor cells (Turner et al., 2017), however, it was suggested that this mechanism plays a role in human genome evolution.

In conclusion, similar use of the A-Bruijn graph by Pu, Lin, and Pevzner (2018) allowed to systematically study SD evolution. For example, independent groups of SD-blocks were detected that coevolved together, based on connected components topology some predictions can be made on a duplication process that corresponding loci underwent. Cyclic components were studied in detail. This allowed to assign specific mechanisms for each case and thus evaluate quantitatively the prevalence of each mechanism. It also make sense to mention that both described research projects and the one that we did utilized a graph representation for annotated SD alignments. However, as we will see later, the ways these graphs were constructed are completely different. That is why most conclusions of described research projects relate

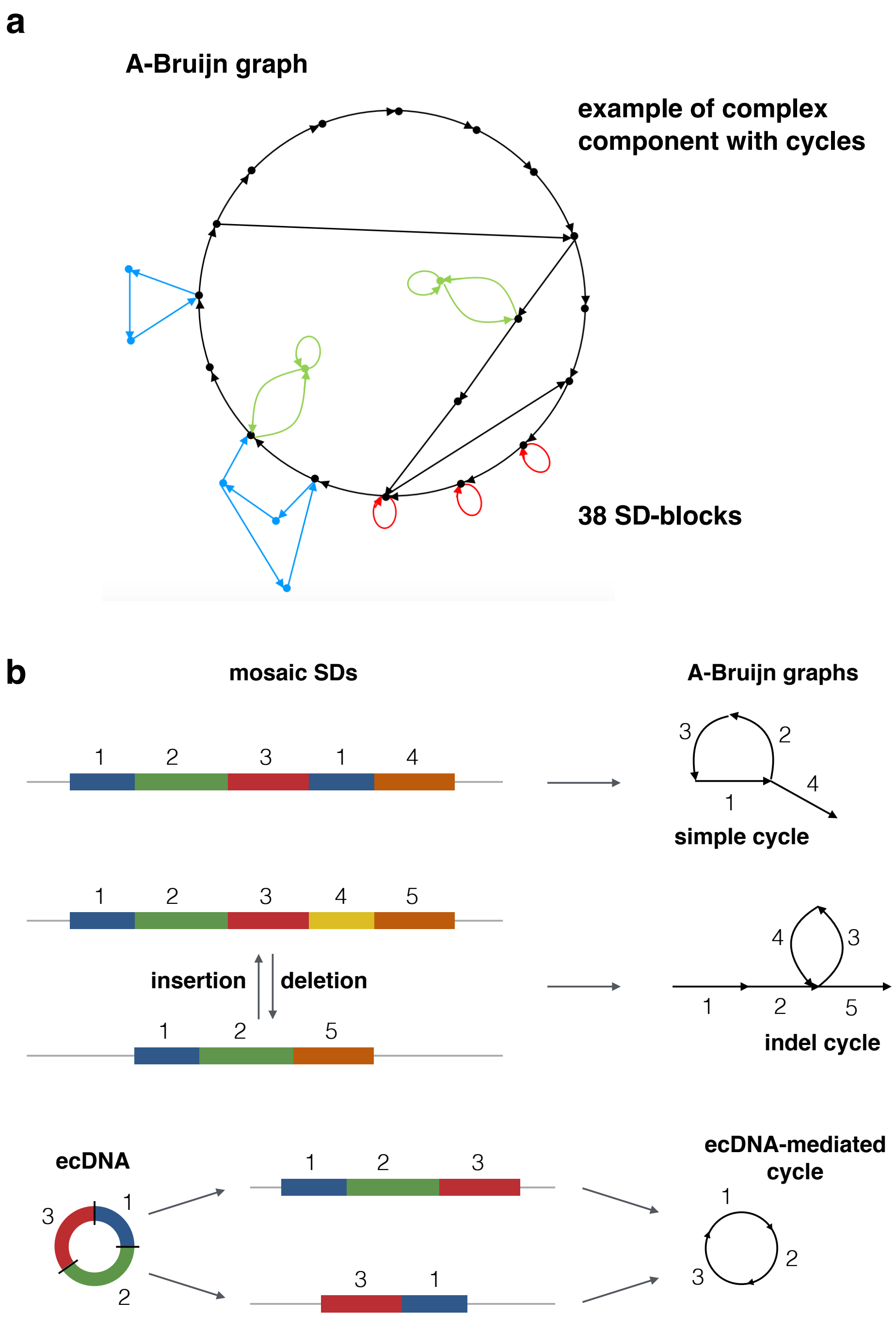


FIGURE 2.11: **Cycles in the A-Bruijn graph.** **a**. An example of A-Bruijn graph connected component. It originated as a result of complex pattern of duplications which also gave rise to the cycles in the graph (coloured for convenience). Adopted from Pu, Lin, and Pevzner (2018). The scheme (**b**) shows three possible scenarios on how cycles can appear in the A-Bruijn graph. The simple cycles appear when an SD-block is repeated, indel cycle when insertion or deletion disrupted mosaic SD and ecDNA-mediated cycles are formed when genomic fragment is copied from extra-chromosomal circular DNA elements.

to SD-blocks/duplicons which were nodes in the networks, while, in our research project evolutionary dynamics of duplicated regions was in the focus.

2.3.4 Evolutionary role of SDs: gene duplications

According to the well-known Ohno's model the primary mechanism for emergence of genes with novel functions is by duplication and divergence (Ohno, 1970). A gene duplication creates redundancy thus liberating one copy from a purifying selection characteristic for a single gene state. Thus one copy can accumulate changes and gain a new function without harmful consequences for the organism. Segmental duplications are large enough to duplicate entire or large fractions of genes thus making SDs an important source of gene innovations. Segmental duplications can affect the genes in various manners: by duplicating an entire gene, shuffling exons, making gene fusions and altering expression profiles by modifying regulatory sequences.

Gene duplications are mostly formed by interstitial SDs. Some gene families were propagated by segmental duplications in subtelomeric regions, including olfactory receptors family members and FOXD4 gene paralogues (Trask et al., 1998; Wong et al., 2004). On the other hand, the pericentromeric SDs are depleted with genes in comparison with other duplications (Dennis and Eichler, 2016). There are several characteristic features of genes duplicated by SDs. Firstly, those are often under positive selection (especially ones overlapping core duplicons) (Jiang et al., 2007). Moreover, these genes are 5 to 10 times as likely being copy-number variable between species and among individuals of the same specie than single copy genes (Cheng et al., 2005; Tuzun et al., 2005). Finally, the genes duplicated in the human lineage are enriched with some functional categories: neural system development (synaptogenesis, neuron migration and expansion of the prefrontal cortex), foreign chemical substances detection (olfactory reception) and metabolism, immune response, cell differentiation and spermatogenesis (Jiang et al., 2007; Sudmant et al., 2013; Dennis and Eichler, 2016).

Gene duplication events observed in the human genome can be mapped on its ancestral lineage. Overall, there are 27 human genes that were duplicated after the divergence of the gorilla lineage but before the split of human and chimp lineages (5 - 7 mya), duplications of 80 genes happened in the human and African great apes lineage (7 - 13 mya), 124 genes were duplicated in the human and great ape lineage (13 – 18 mya) and 105 genes were involved in duplications in the human lineage, while shared by all ape species (18 - 24 mya) (Dumas et al., 2007). Human-specific duplication events gave rise to 80 paralogous genes belonging to 33 gene families (Dennis et al., 2017). We will consider the most prominent examples of human-specific gene duplications.

- The gene *SRGAP2A* which encodes SLIT-ROBO Rho GTPase activating protein 2A was duplicated in the human lineage twice. Firstly, the duplication of

SRGAP2A 3.4 mya resulted in the truncated copy *SRGAP2B*. The second duplication of *SRGAP2B* gave rise to the paralogous gene *SRGAP2C* 2.4 million years ago (Dennis et al., 2012). This time interval coinsides with paleontological estimates on when the neurocortex expansion happened in the human lineage (Jobling, Hurles, and Tyler-Smith, 2019). This gene family, likely, plays an important role in a brain development. Overexpression of human-specific paralog *SRGAP2C* in cell cultures and *in vivo* regulates dendritic spine maturation and radial migration of neurons (Charrier et al., 2012).

- The duplication of *ARHGAP11A* gene (encoding Rho-type GTPase-activating protein 11A) 5.2 mya resulted in paralogous *ARHGAP11B* copy (Dennis et al., 2012). The *ARHGAP11B* gene seem to be involved in progenitor cells differentiation into the basal radial glial and neuronal cells (Florio et al., 2015). Moreover, the corresponding human-specific SD could mediate further 15q13.3 microdeletion which causes cognitive disorders and schizophrenia (Sharp et al., 2008).

- The *NBPF15* (or protein domain DUF1220) gene belongs to the neuroblastoma breakpoint family (*NBPF*). This gene duplicated multiple times recently in the human evolution, its paralogous sequences populated the chromosome 1. The gene copy number alterations observed in the region 1q21.1 are associated with various neurologic diseases, such as microcephaly and autism spectrum disorder (Dumas et al., 2012).

- The *NPIP* gene family encodes nuclear pore complex interacting proteins of unknown function. The *NPIP* genes show signs of strong positive selection (Johnson et al., 2001). Moreover, about 10% of euchromatin sequence of 16p chromosome arm is comprised of SDs referred as LCR16. The corresponding core duplicon LCR16a embeds *NPIP* genes which means that *NPIP* gene expansion was evolutionary important (Jiang et al., 2007).

- Several duplications of *AMY1* gene that encodes salivary amylase also played a role in human evolution. Increased copy number of *AMY1* gene likely lead to higher level of amylase in the saliva of modern human (Sudmant et al., 2015; Groot et al., 1989). This expansion was beneficial for early humans because it allowed to digest starch-rich diet associated with a hunter-gatherer to an agricultural life transition. Curiously, the copy-number of *AMY1* gene is highly variable among modern human populations. The copy-number is correlated with the amount of starch in diets worldwide (Perry et al., 2007).

Finally, SDs can provide long stretches of homology thus leading to recurrent genomic rearrangements. Many CNVs associated with disease take plays at segmental duplication sites. This list includes CNVs associated with cognitive disorders (Stankiewicz and Lupski, 2002; Sharp et al., 2006), multiple cancer types (Lahortiga et al., 2007; Weir et al., 2007), epilepsy (Bonaglia et al., 2005), autism spectrum disorder (Ullmann et al., 2007), Alzheimer disease (Rovelet-Lecrux et al., 2006), glomerulonephritis (Aitman et al., 2006) and many more. Also, as we saw earlier for human-specific gene duplications, those innovations are often associated with corresponding disorders when paralogous copies are lost.

Chapter 3

Introduction to Network Analysis

In this section we will introduce all the terminology needed to understand our analysis of complex networks. The field of science that analyses complex networks is very comprehensive with numerous applications, methods and algorithms. For this thesis, we will leave a big part of it out of the scope to be more focused on the relevant terminology.

3.1 Fundamental terminology in graph theory

- A **graph** is a mathematical object that represents a system of **nodes** (or vertices) and **edges** that link pairs of nodes. More formally, a graph is a pair $G = (\{V\}, \{E\})$, where $\{V\}$ and $\{E\}$ are sets of nodes and edges, respectively. Each edge can be considered as a tuple of two connected nodes $[v_i, v_j]$ where nodes v_i and v_j are called endpoints. In directed graphs (see the definition below) endpoints are ordered according to an edge orientation. **Neighbors** are all nodes connected to a specific node by edges.

  Graphs can be mathematical abstractions and studied as such, however, often graphs represent some real life systems where nodes correspond to objects and edges are interactions between them. Typically, graphs are visualized as a set of points for nodes which are connected by lines corresponding to edges.

- Graphs can be either **directed** when edges are oriented or **undirected** when not. Directed edges, usually, represent asymmetric interactions with a source and a target (forces, streams, citations etc.) while undirected edges are symmetric (phone calls, distances, routes etc.).

- Edges that connect a node to itself are called **loops** or self-loops. **Multi-edges** or multiple edges are two or more edges that link the same pair of vertices. **Simple graph** is the one which includes no **loops** or **multi-edges** as opposed to a **multigraph** where these structures are permitted.

- **Weighted edges** are those ones for which a number (weight) is assigned. A corresponding graph is also called weighted.

- A **connected component** of an undirected graph is a subgraph where all pairs of nodes are linked by a path of edges and it is not part of any larger connected subgraph. Connected components partition a graph into isolated subgraphs not connected between each other.

- A graph is called **complete** or fully connected when edges between all pairs of nodes are present. A complete graph is characterized by N nodes and $\frac{N(N-1)}{2}$ edges composition, because there are maximum $\frac{N(N-1)}{2}$ edges possible for a simple graph with N nodes.

- A **clique** is a subset of nodes such that every pair of distinct nodes in the clique are adjacent (linked with an edge). In other words, the subgraph that corresponds to a clique is fully connected.

- A **path** is a sequence of distinct edges that joins a sequence of vertices. A **cycle** in a graph is such a path where the first and the last nodes are the same. A connected undirected graph with no cycles is called a **tree** graph.

- A **minimum spanning tree** (MST) is a path going through all nodes of the edge-weighted connected graph without any cycles and with the minimal possible sum of edge weights. According to the definition this path has to be a tree covering all nodes. There are several algorithms that can find the MST of a graph, for example, Prim's and Kruskal's algorithms. Both find the minimal overall weight tree with the run-time $\propto O(E \log N)$, where N and E are numbers of nodes and edges, respectively.

- Edges of a graph define relations between nodes or adjacency relations. Thus a graph can be specified by a matrix of node-to-node relations or **adjacency matrix**. The adjacency matrix A is a square matrix of size N (number of nodes), where $A_{ij} > 0$ if i^{th} and j^{th} nodes are connected, otherwise $A_{ij} = 0$. The A_{ij} value equals to the number of edges between corresponding nodes. Thus adjacency matrices of simple graphs consist of 0 and 1 values. The adjacency matrix of an undirected graph is symmetric ($A_{ij} = A_{ji}$).

- The **node degree** (k_i) of a node is the number of edges that are connected to it. In the complete graph with N nodes all nodes are of degree $k_i = N - 1$.

- The local **clustering coefficient** of a node is a quantitative characteristic of how close its neighbours are to a clique, in other words, how connected the neighbors are between each other. The local clustering coefficient of a node v_i equals:

$$C_i = \frac{2|\{e_{jk} : v_j, v_k \in \mathcal{N}(v_i)\}|}{k_i(k_i - 1)}, \quad C_i \in [0, 1],$$

  where e_{jk} is an edge between nodes v_j and v_k that both belong to the neighborhood of v_i (denoted as $\mathcal{N}(v_i)$ in the formula), k_i is a node degree of v_i. Often, the mean clustering coefficient is calculated as a graph topology characteristic.

- The **average path length** is the average number of edges along the shortest paths between all possible combinations of node pairs.

3.2 Fundamental terminology in network science

- A **complex network** is similar in its meaning to a graph. The difference between those two is not formally defined and usually depends on questions one tries to answer. Generally, a complex network is a large graph, usually representing some real-life system, where nodes and/or edges have attributes assigned (e.g. ids, weights, flow capacities etc.). Many real-life complex networks display characteristic **network topologies** (specific arrangements of vertices and edges) that differ them from random graphs, with structural patterns that are neither completely regular nor random. Such features include heavy tails in the node degree distribution, high clustering coefficients, community and hierarchical structures (see the definitions below). These topological features often reflect specific mechanisms of networks formation.

- The **network science** is a field of the graph theory that studies complex networks. Complex networks are applied to study various systems in physics, computer science, sociology, logistics, omics data analysis, neurobiology, ecology, epidemiology and so on. Predictions on network growth, community structure, structural robustness, network flow etc. are, typically, made in complex network analysis tasks.

- Structures of biological networks are often non-uniform in their adjacency. Nodes often cluster together into denser **modules** (or communities) with higher level of connectivity than in the rest of the network. Modularity is a measure of how easily the network can be divided into modules. Networks with high modularity have more edges between the nodes within modules but sparse connections between nodes in different modules. It is also said that these networks have a community structure.

- A **giant component** is a connected component of a network that contains a significant proportion of nodes and edges in the network.

- **Centrality** measures are scalar values given to each node of a graph to quantify its importance. Which nodes are considered important depends on a specific centrality measure and their position in a graph. Some examples of centrality measures we discussed already: the local clustering coefficient and a node degree, however, there are more existing measures. The **betweenness** centrality of a node, for example, is the fraction of all possible shortest paths in a network that pass through it. Similarly, one can define betweenness of edges.

- A **random** graph is a term that refers to a graph generated according to some probability function. The probability distribution can describe its expected characteristics and topology. The Erdős–Rényi model is the model for generating random graphs. In the Erdős-Rényi model, all graphs of a fixed vertex set with a fixed number of edges are taken equally likely. Similarly, in related model which is called the Erdős–Rényi–Gilbert model, each theoretically possible edge is either present with probability p or absent otherwise. As a result, a probability of generating specific graph with N nodes and E edges equals:

$$p^E(1-p)^{\binom{N}{2} - E}, \quad E \leq \binom{N}{2},$$

where $\binom{N}{2}$ equals to the maximal number of possible edges in a graph of N nodes.

- The **scale-free** network is another network topology that is often observed in real-life complex networks (citation network, World Wide Web, protein-protein interaction network etc.). The scale-free network is a network whose node degree distribution asymptotically follows a power-law. Thus the probability of a node to have k neighbors is proportional to $p(k) \propto k^{\alpha}$, where the slope α is negative and often in the interval $-3 \leq \alpha \leq -2$ in real-life networks. Networks of such type have numerous edges with small node degrees and several "hubs" linked to multiple nodes. The mean shortest path length in these networks is substantially lower than in random graphs because of these hub nodes.

  There are several models that allow to generate scale-free networks, most of them are associated with various types of preferential attachment. It means that the probability of a new node to be adjusted to an existing one is higher for those nodes with higher node degree. This principle is also known as the "rich gets richer" principle. The most well-known model for scale-free networks generation is the Barabási–Albert (BA) model which also utilizes the preferential node attachment (Albert and Barabási, 2002). In the BA model nodes are added to the network one at a time. When added the new node is linked to m existing nodes. It happens in a preferential manner: the probability for an existing node v_i to be connected to the new one linearly dependence on its node degree k_i. When the network is large enough:

  $$p_i = m\frac{k_i}{\sum_j k_j},$$

  where the summation goes over all existing nodes. Networks resulting from the BA model have a scale-free topology with the slope $\alpha = -3$.

- The **configuration model** is a method for generating random networks of a given node degree sequence.

Chapter 4

Network Analysis of Segmental Duplications

4.1 Network construction

We based our analysis on the already annotated segmental duplications (SDs) in the reference human genome (Bailey et al., 2001). A corresponding list of *GRCh38* annotated SDs was downloaded from the UCSC genome browser website (Kent et al., 2002). Basically, we start with a list of pairwise local alignments longer than 1 kbp with at least 90% identity between different regions of the human reference genome. There are 27,348 autosomal alignments in this list. For our analysis we disregard the sex chromosomes because we expect different evolutionary forces acting on these chromosomes (recombination and mutation rates and their consequences). However, not every reported alignment refers to a unique segmental duplication event, because, when a new duplication overlaps with an older one, the new copy aligns not only to the ancestral region, but also can be aligned to other copies of the ancestral region. We call such an alignment "secondary" if it appears as a result of an overlap between a new duplication and an already duplicated region. These alignments do not represent a duplication event between aligning regions.

To study this puzzling system of segmental duplications (Fig. 2.5) we generated a network of SDs in the following way. Each node represents a duplicated region: a duplicated region is a genomic interval that covers a maximal set of overlapping alignments (i.e. duplicated region equals to a union of all overlapping alignment intervals). Undirected edges link nodes if an alignment between two regions exists (the construction process is illustrated at Fig. 4.1). In general we used this network after trimming multiple edges between any pair of nodes (multiple edges) and self-loop edges. In the remainder of the text we will denote genomic regions that correspond to nodes of the SD network as duplicated regions and will associate network characteristics to those regions directly, for example, we consider a node degree of a duplicated region (meaning a node degree of a node corresponding to a duplicated region of interest).

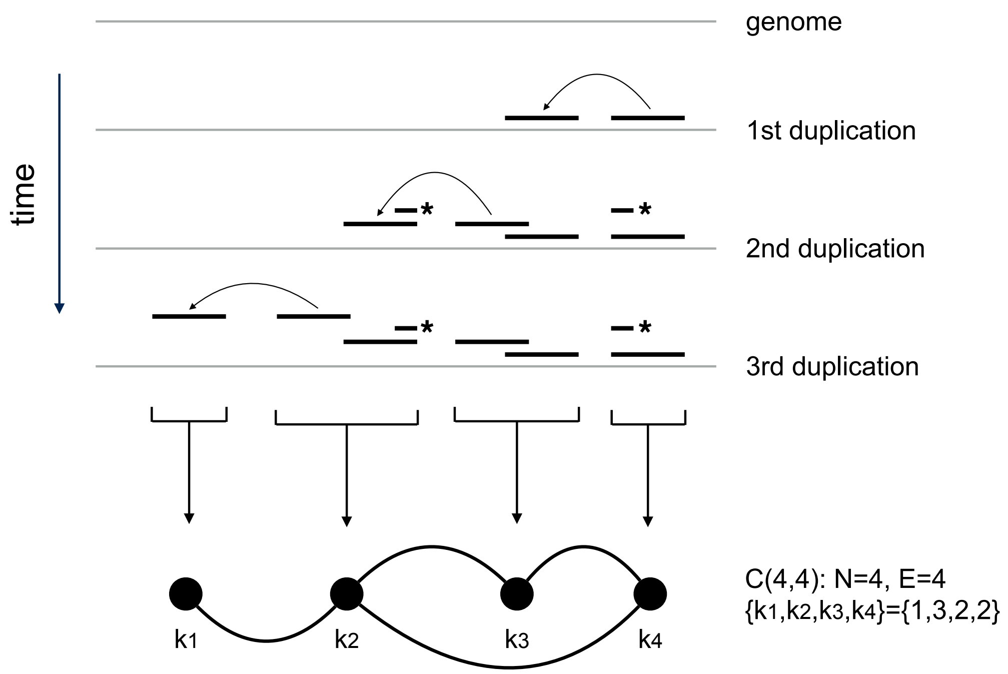


FIGURE 4.1: **An example of network construction from SDs.** The scheme illustrates an example of several duplication events in the genome, the resulting alignments and the network constructed based on those alignments. In every time step one duplication happens in the genome and a second copy is inserted in the genome nearby. Alignments appear not only between a copied region and its copy as expected, but also when a duplication overlaps one of existing duplicated regions (the second duplication event on the scheme). We refer to those alignments as "secondary". For the network construction we grouped sets of overlapping alignments into separate duplicated regions. Each duplicated region is represented with a node in the SD network. Edges are added if there exists an alignment between duplicated regions.

4.2 Network characteristics

The resulting SD network has 6656 nodes and 16,042 edges (Fig. 4.2). The network can be decomposed into 1999 connected components, i.e. isolated subgraphs where any pair of nodes is connected by a path of edges. One distinctive feature of the SD network is that it includes a giant component with 1325 nodes (19.9% of all nodes) and 9678 edges (60.3% of all edges) that corresponds to multiple overlapping duplication events enriched in some genomic loci.

The number of edges that a node has (node degree) represents a number of copies of a corresponding duplicated region. This network can be further described considering topological network characteristics, including a component size and a node degree distributions (Fig. 4.3). The connected component size distribution decreases following a power-law distribution $p(N) \propto N^{-2.7}$ while the giant component is well separated from this distribution. The distribution of node degrees has a mean of 4.8 and follows an exponential tail for large node degrees (Fig. 4.3b). Interestingly,

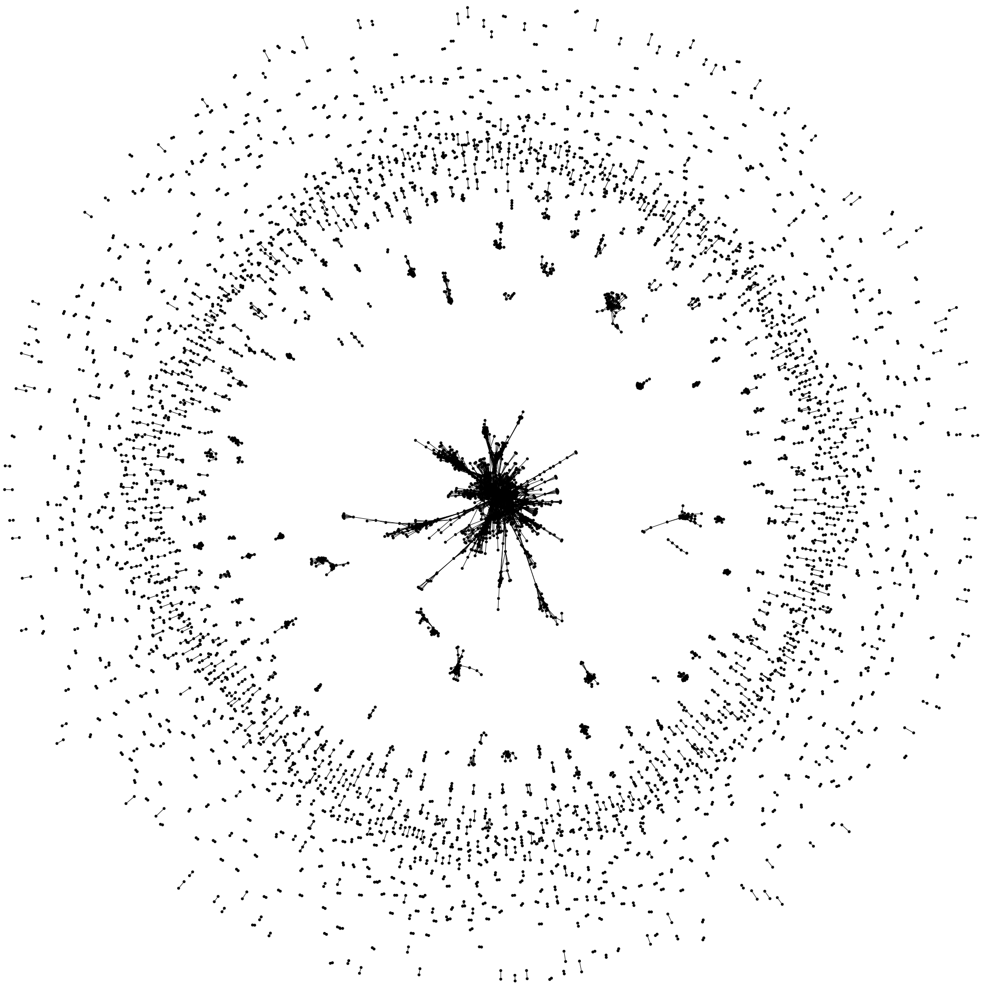


FIGURE 4.2: **The SD network.** The network constructed based on SDs of the reference Human genome (SD network). The black circles and lines represent nodes and edges of the SD network. There are 6656 nodes and 16,042 edges in the SD network in total. One can see that the SD network includes multiple small connected components and a distinctive giant component with 1325 nodes and 9678 edges in it (located in a center of the figure).

the average number of edges E in a component with N nodes follows a power-law: $E(N) \propto N^{1.47}$ (Fig. 4.3c). Later we will come back to this observation and give an interpretation of it.

Due to its size we can study the giant component in more detail. The clustering coefficient of a node is the number of edges between vertices in the neighborhood of a specific node divided by the overall number of possible edges between those neighbors. The mean clustering coefficient calculated over all nodes in the giant component was equal to $\overline{C} = 0.57$ in the SD network. The average shortest path length $l = 4.93$. The modular structure of the giant component was also investigated using the label propagation algorithm (Raghavan, Albert, and Kumara, 2007). It was found that the giant component is enriched with dense clusters of nodes or modules. The majority of network modules were enriched with intrachromosomal

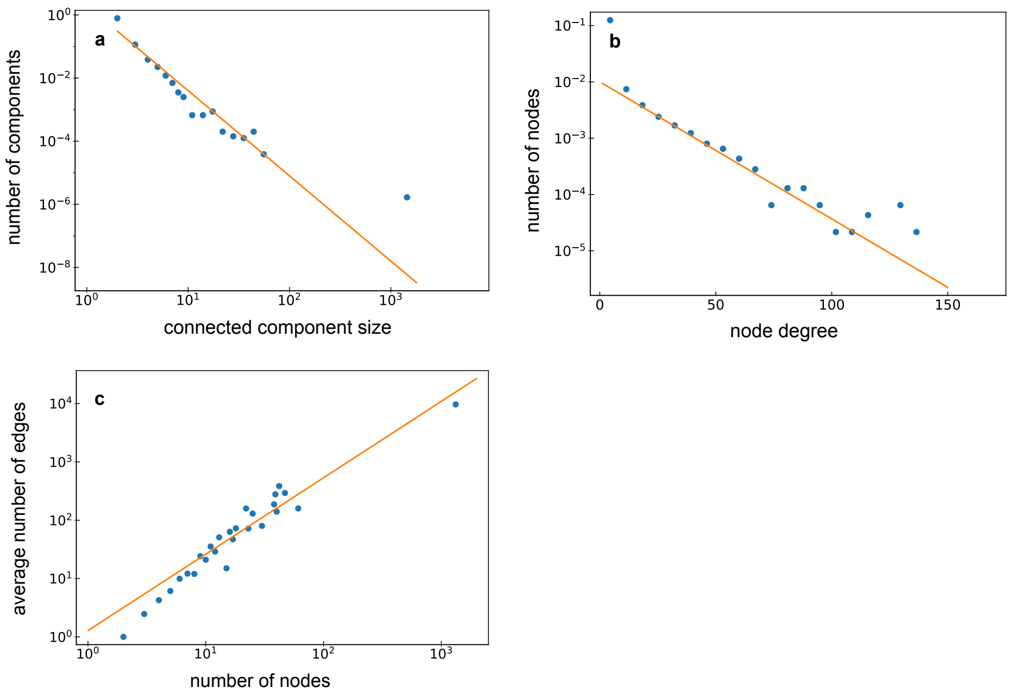


FIGURE 4.3: **Characteristics of the SD network.** **a**. The SD network component size distribution plotted on a log-log scale using logarithmic binning (to reduce stochastic noise in the heavy tail of distribution). The number of connected components decreases with their size comparable to a power-law distribution $p(N) \propto N^{-2.7}$ which is represented as a straight orange line added as a guide to the eye. One distinctive feature of this distribution is the presence of a giant component which shows up as a single dot on the right of the distribution. **b**. The node degree distribution of the SD network plotted on a log-linear scale. An exponential tail of the node degree distribution is stressed with the orange guide to the eye line. **c**. For each component size observed in the SD network the average number of edges in corresponding components is plotted on a log-log scale. An average number of edges in components grows as a power-law of a component size: $E(N) \propto N^{1.47}$ dependence (orange line) was fitted with linear regression $\log(E) \sim \log(N)$.

duplications (most of nodes in a module belong to the same chromosome). Additional figures illustrating characteristics of the SD network can be found in Appendix chapter.

Even though the SD network can be described by general topological features we want to remark that the observed topology does not coincide with one of the well-studied network topologies (like scale-free or random networks). We therefore decided to simulate the dynamics of a network growth based on some predefined "rules" inspired by our knowledge on genome evolution to see if such a synthetically generated network might reflect the same network topology as the observed SD network.

4.3 Dynamical processes

In order to study dynamical aspects of the propagation of SDs in the human genome we first constructed SD network and then asked what dynamical process could generate such a network. We decided to simulate possible network growth models that were inspired by copying models (Chung et al., 2003). Finding a simple network growth model that would generate similar features and topology as the SD network would shed light on the dynamical processes of how the SD network evolved.

Our network growth model includes two processes:

- The first process represents novel duplications that do not overlap any older ones. In the context of network growth this results in *de novo* addition of a connected components $C(2,1)$, i.e. the connected component with 2 nodes and 1 edge with rate π to the network (Fig. 4.4).

- The second process represents duplication events that overlap existing duplicated regions and thus new copies acquire not only alignments with an ancestral duplicated regions, but can also give rise to secondary alignments with other copies of a duplicated region (Fig. 4.1a). If the overlap is long enough we expect it to be annotated as a segmental duplication even though it corresponds to a secondary alignment. In the context of network growth this process is represented by a duplication of an existing "mother" node (that by definition has copies elsewhere in the genome) and the new "daughter" node inheriting some fraction of neighbors from the "mother" node in addition to the edge between the "mother" and the "daughter" nodes that is added by default (Fig. 4.4). In our probabilistic model we added a parameter f that represents the probability of each edge connected to the "mother" node to be inherited by the "daughter" node. After a duplication the node degree of a "daughter" node $k_d \in 1, 2, \ldots, k_m + 1$ where k_m is the node degree of a "mother" node (Fig. 4.4). Node duplications happen with the rates proportional to a second parameter δ. However, since only the ratio of the two rate parameters δ/π matters for simulations we assume $\pi = 1$ in the remainder of the text.

4.3.1 The Uniform Copying Model (UCM)

In the previous section we formulated universal principles of our copying models. However, the copying models that we use in practice differ in the way we define duplication rates of nodes. In the simplest model, we assume that duplication rates for all nodes i are the same: $\delta_i = \delta$. We will further refer to this model as the Uniform Copying Model or UCM (Fig. 4.4). The connected component size distributions in networks grown using the UCM follows a power-law distribution $p(N) \propto N^{-1}$. Although disguised by finite-size effect in Fig. 4.5a, this can be more clearly seen in longer simulations in (Fig. 4.6a). In the later "Analytical solutions" section we also

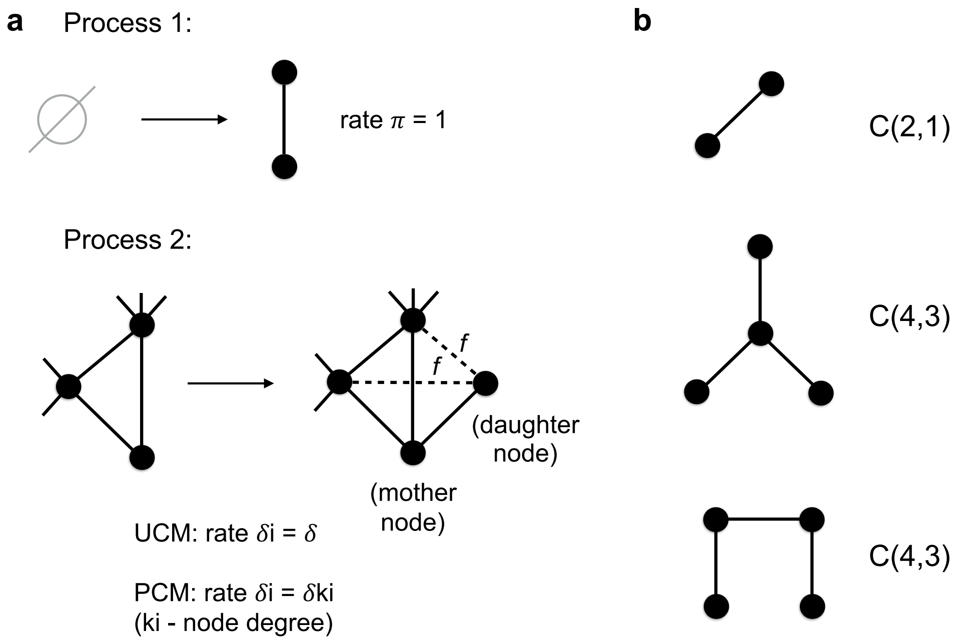


FIGURE 4.4: **Schematic representation of network growth processes and connected components.** **a**. The scheme illustrates two processes of a network growth in our growth models. One can find a biological explanation for these two processes in the main text. Process 1: The component $C(2,1)$ is added to a network with the rate π. Process 2: Each node i in the network can be duplicated with the rate δ_i. A "daughter" node gets at least one edge linked to a "mother" node by default and inherits connections from the "mother" node to its neighbors each with the probability f. In other words, each neighbor of a "mother" node can become a neighbor of a "daughter" node with the probability f. The difference between the Uniform Copying Model (UCM) and the Preferential Copying Model (PCM) is in defining the duplication rates of nodes. These are constant $\delta_i = \delta$ in the UCM, while in the PCM the duplication rates grow linearly with a node degree $\delta_i = \delta k_i$ where k_i is a node degree of corresponding node. **b**. We denote components with N nodes and E edges as $C(N,E)$. This notation does not always correspond to a unique possible graph topology, for example, there is only one topology for $C(2,1)$ while there are two for $C(4,3)$. Components with N nodes and any possible number of edges are denoted as $C(N,*)$ which is the same as all components of size N.

derive this behavior analytically (4.4). To reduce a noise in distributions associated with synthetic networks, here and for the next copying model, we run 500 simulations with the same parameters, aggregated all networks and plotted distributions of resulting pooled networks. Since the connected component size distributions of synthetic networks are different from the one of the SD network (the power-law exponents are different and they lack prominent giant components) we assume that the SD network evolved according to another network growth model.

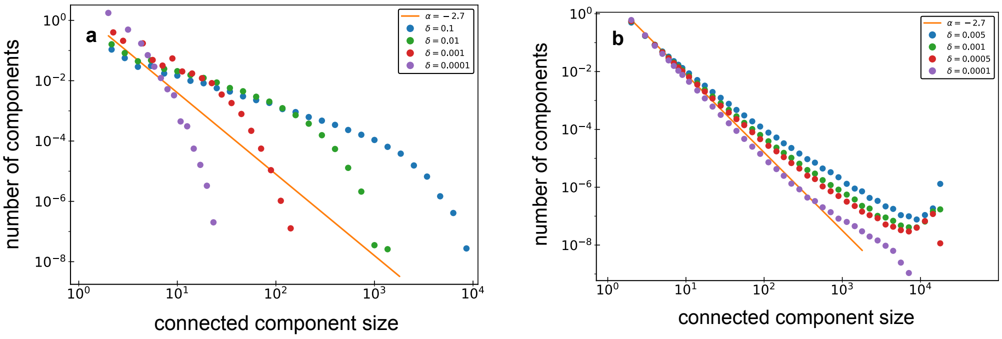


FIGURE 4.5: **The component size distributions observed in pooled simulations of network growth based on UCM and PCM growth models.** In all cases we used the parameter $f = 0.5$, δ values as indicated in the legends and simulated a network growth until resulting network reaches the size of the SD network. An orange guide to the eye line is added to illustrate the slope observed in the connected component size distribution of the SD network ($p(N) \propto N^{-2.7}$). **a**. The component size distributions observed in the UCM simulations differ from the one observed in the SD network. Both slopes are different and no peaks that correspond to giant components are observed in the UCM simulated networks. **b**. The component size distributions observed in the PCM simulated networks are similar to the one of the SD network. All distributions independently of δ value follow a similar slope on a log-log scale to the one of the SD network for component sizes observed in the SD network. Moreover, PCM synthetic networks and the SD network include giant components. It can be seen as a peak at the right side of the component size distributions of the PCM simulated networks.

4.3.2 The Preferential Copying Model (PCM)

The UCM was not sufficient to explain the SD network topology. This motivated us to study a different dynamics of copying models. The next simplest copying model is the one where a duplication rate of a node i depends linearly on a node degree k_i and the parameter δ value: $\delta_i = \delta k_i$. In this copying model highly connected nodes are duplicated with preference and we will further refer this model as the Preferential Copying Model or PCM (Fig. 4.4).

Our analytical solution predicts the power-law distribution $p(N) \propto N^{-1-f}$ for the connected component size distribution (see the "Analytical solutions" section 4.4). This behaviour is also observable in simulations of the PCM. The power-law tail gets obvious for pooled and long simulations (see Fig. 4.6).

There is no reason to reject the PCM based on the connected component size distributions of synthetic networks. We simulated a network growth according to the PCM until the number of nodes of synthetic networks reaches the one of the SD network. The connected component size distributions observed in PCM synthetic networks follow a similar slope on a log-log scale to the one of the SD network (Fig. 4.5b). Moreover, giant components appear in PCM simulations, similarly to the SD network (peaks on the right side of the distributions at Fig. 4.5b).

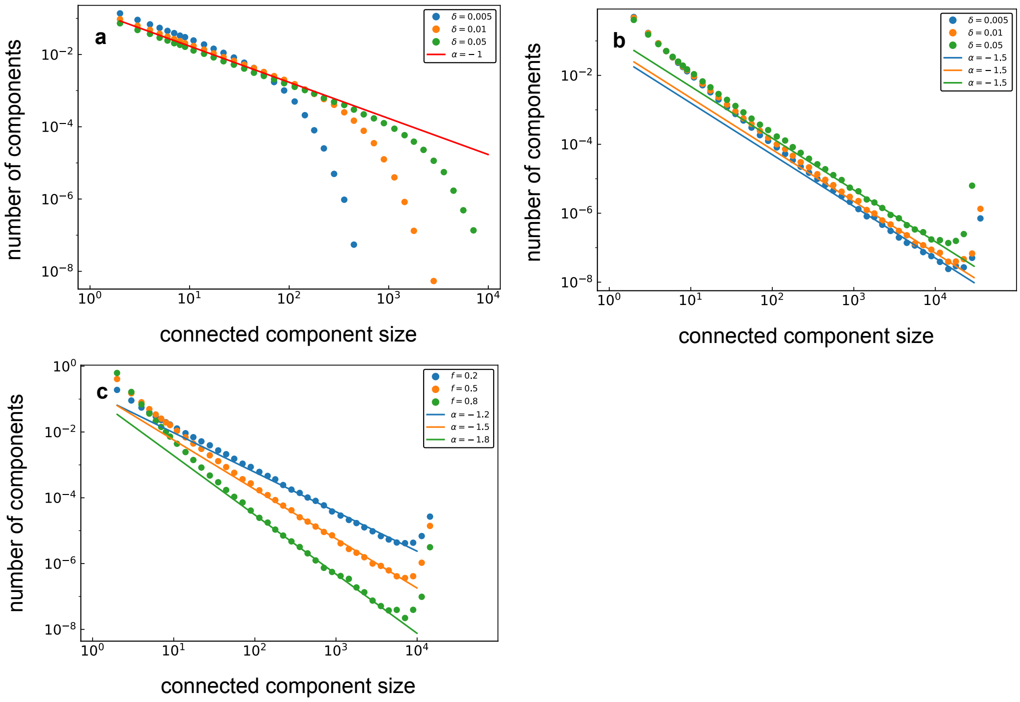


FIGURE 4.6: **The connected component size distributions observed in simulations of network growth (log-log scale).** The distributions observed in synthetic networks follow our analytically predicted slopes. **a**. The UCM simulated networks follow the power-law $p(N) \propto N^{\alpha}$ where $\alpha = -1$ for all parameter δ values (parameter f values do not effect the distributions in the UCM). **b-c**. The PCM simulated networks follow the power-law $p(N) \propto N^{\alpha}$ where $\alpha = -1 - f$ for all parameter δ and f values. Straight lines represent analytically predicted slopes in all panels.

4.3.3 Additional information on network growth

We construct our models of network growth based on specific copying mechanism as described in the previous sections. There are two types of processes happening during the network growth: an addition of a new connected component $C(2,1)$ to a network or duplication of an existing node and inheritance of some fraction of its edges. Our assumption is that all genomic loci can be duplicated independently of other duplication events. Thus we used the Kinetic Monte Carlo (KMC) method to run a simulation of network growth where all events happen independently of each other (Young and Elcock, 1966).

For a graph with N nodes and E edges at time point t, a total of $N + 1$ possible processes have to be considered. First the addition of a new component $C(2,1)$, with the rate π, and the duplications of any one of the existing nodes, with rates δ_i. The rates of all possible processes are represented as a vector $\vec{r}(t)$ of length $N + 1$. For the UCM we use $\delta_i = \delta$ thus the rates vector:

$$\vec{r}_{\text{UCM}}(t) = \begin{pmatrix} \pi \\ \delta \\ \vdots \\ \delta \end{pmatrix}$$

For the PCM we use $\delta_i = \delta k_i$ where k_i is a node degree of node i thus the rates vector:

$$\vec{r}_{\text{PCM}}(t) = \begin{pmatrix} \pi \\ \delta k_1 \\ \vdots \\ \delta k_N \end{pmatrix}$$

One of $N + 1$ possible processes at time point t is picked at random with probabilities proportional to the given rates $\vec{r}(t)$. An average waiting time before this event happens is exponentially distributed. It can be calculated as $\Delta t = -\ln(u)/(\sum_i \vec{r}_i(t))$, where u is sampled randomly from the $(0,\ 1]$ interval. Since only relative rates matter in the KMC we used $\pi = 1$ in all simulations and fitted only the δ value.

All network growth simulations terminate when the number of nodes in a network reaches some predefined threshold (in most cases the number of nodes in the SD network).

4.4 Analytical solutions

In this section we present analytical solution for key distributions of our models introduced earlier in the main text.

- In the UCM (Uniform Copying Model) each component grows with the rate proportional to its size (number of nodes). So a component size N as a function of an absolute time t and the time when a component was added to a network s changes in the following way:

  $$\frac{\partial N(s,t)}{\partial t} = \delta N,$$

  This differential equation can be solved by the following ansatz:

  $$N(s,t) = 2e^{\delta(t-s)}, \qquad N(s=t,t) = 2$$

  which can be inverted and solved for s:

  $$s(N,t) = t - \frac{1}{\delta}\ln(\frac{N}{2})$$

This finally gives us the power-law distribution $p(N) \propto N^{\alpha}$ where $\alpha = -1$:

$$p(N,t) \propto \frac{\partial s(N,t)}{\partial N} \quad \propto \quad N^{-1}$$

This result agrees with observations from UCM simulations (Fig. 4.6a).

- In the PCM (Preferential Copying Model), on the other hand, each component $C(N,E)$ grows with the rate proportional to the sum of node degrees of all N nodes in a component which equals $2E$. According to our simulations $E \propto N^{1+f}$ dependence is characteristic for components in the PCM growth thus the size $N(s,t)$ of PCM components changes in the following way:

  $$\frac{\partial N(s,t)}{\partial t} \propto \delta N^{1+f},$$

  This differential equation can be solved in the following way:

  $$N(s,t) \propto (C - \delta f(t-s))^{-1/f},$$

  it can be inverted:

  $$s - t \propto \frac{N^{-f} - C^{*}}{\delta f},$$

  This leads to the power-law distribution $p(N) \propto N^{\alpha}$ where $\alpha = -1-f$:

  $$p(N,t) \propto \frac{\partial s(N,t)}{\partial N} \quad \propto \quad N^{-1-f}$$

  C and C^{*} are constants. This result agrees with observations from long PCM simulations (Fig. 4.6b,c).

- When a node is duplicated in UCM component $C(N,E)$ an expected number of edges increases by $1 + f(2E)/N$, i.e. one edge to the daughter node plus an additional fraction of f edges of the average node degree $2E/N$, since all nodes are duplicated equally likely. Therefore the number of edges E in UCM components changes with N in the following way:

  $$\frac{dE(N)}{dN} = 1 + f\frac{2E}{N}$$

  Firstly, the homogeneous differential equation is rearranged:

  $$\frac{dE}{E} = 2f\frac{dN}{N},$$

  and solved:

  $$E(N) = CN^{2f}$$

Now we return to solving the original non-homogeneous equation for $C(N)$ using the variation of parameters method:

$$\frac{dC}{dN}N^{2f} + 2fCN^{2f-1} = 1 + 2fCN^{2f-1},$$

which can be reduced to:

$$dC = N^{-2f}dN,$$

and solved for $C(N)$:

$$C(N) = \begin{cases} \frac{1}{1-2f}N^{1-2f} + C^*, & f \neq 0.5 \\ \log N + C^*, & f = 0.5 \end{cases}$$

which leads to the following solution of the original differential equation:

$$E(N) = \begin{cases} \frac{1}{1-2f}N + N^{2f}C^*, & f \neq 0.5 \\ N\log N + NC^*, & f = 0.5 \end{cases}$$

Thus when $N \to \infty$ the number of edges E in the UCM components follows:

$$E(N) \propto \begin{cases} N, & 0 \leq f < 0.5 \\ N\log N, & f = 0.5 \\ N^{2f}, & 0.5 < f \leq 1 \end{cases}$$

C and C^* are constants. This result agrees with a dependence observed in UCM synthetic networks (Fig. 4.7a)

4.5 Estimation of the parameters for the PCM

To make further conclusions on relatedness of the PCM to the evolution of the SD network, we inferred values for the parameters f and δ such that a PCM generated network matches the characteristics of the observed SD network. In the next subsections we will discuss two strategies that were applied to infer the parameters values.

4.5.1 The parameters inference from edges to nodes ratio and ABC

In the first approach, the average fraction of neighbors f inherited from a "mother" node was predicted using one empirical observation. We found that the average number of edges E in connected components generated by the PCM grows with the number of nodes N according to $E \propto N^{1+f}$ when $N \to \infty$ (see Fig. 4.7b). This is in contrast to a more complicated dependence that can be analytically predicted for simpler UCM growth (see the "Analytical solutions" section 4.4) and observed in simulations (see Fig. 4.7a). We therefore used a linear regression of $\log(E) \sim \log(N)$ to estimate the power-law exponent and find that the power-law $E \propto N^{1.47}$ fits best to the observations, thus suggesting the value $f_{\text{reg}} = 0.47$ (Fig. 4.3c).

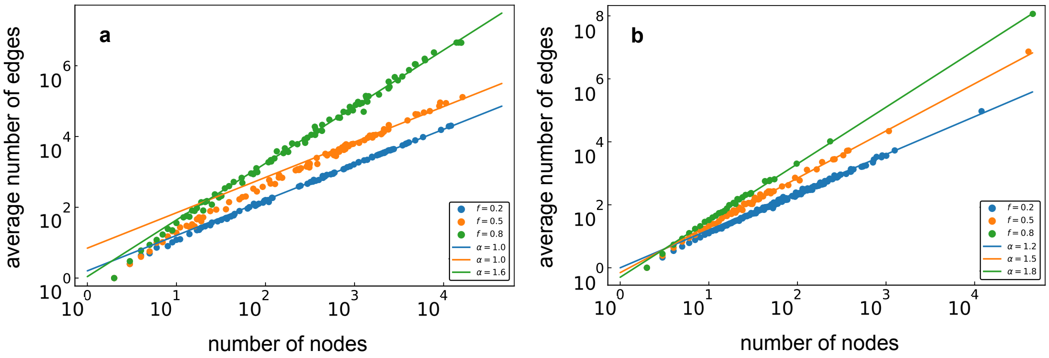


FIGURE 4.7: **The average number of edges in components is plotted against a component size (log-log scale).** Different colours correspond to network growth simulations with different f values using the UCM (**a**) and the PCM growth (**b**). Straight lines represent the slopes of a power-law growth predicted analytically for the UCM (see the "Analytical solutions" section 4.4) and the ones observed in the PCM simulations ($E \propto N^{1+f}$).

The parameter δ value was predicted using the Approximate Bayesian Computation method. The Approximate Bayesian Computation (ABC) is a Bayesian method to approximately predict posterior parameter distributions when an analytical formula for a likelihood function can not be derived (Rubin, 1984). To apply ABC a rejection criteria (specific distance measure) and a tolerance level (distance threshold) are needed that allow to say if the resulting outcome of a simulation is similar to a real observation or not. In our case we compared the connected component size distributions in the SD and the PCM simulated networks. As a rejection criterion we used the Bray-Curtis dissimilarity (D_{BC}) from Bray and Curtis (1957). The Bray-Curtis dissimilarity between a sorted arrays of N biggest connected component sizes is calculated in the following way:

$$D_{\text{BC}}(X, Y) = \frac{\sum_{i=1}^{N} |X_i - Y_i|}{\sum_{i=1}^{N} (X_i + Y_i)}.$$

We limited the number of components to $N = 500$ because the Bray-Curtis dissimilarity can only be calculated for arrays of the same length (which can not be guaranteed if we take all components). We applied the ABC method by running 5000 simulations of the PCM with $f = 0.47$ and δ values taken uniformly from the interval $[5 * 10^{-5};\ \ 9 * 10^{-4}]$. The rejection criterion is satisfied when the Bray-Curtis dissimilarity between component size vectors of simulated and the SD networks $D_{\text{BC}}(\text{simulated, SD}) < 0.2$ (tolerance level). Based on the ABC the parameter $\delta_{\text{ABC}} = 5.1 * 10^{-4}$ with the 95% confidence interval for the parameter value: $\delta_{\text{ABC}} \in [3 * 10^{-4};\ \ 6.6 * 10^{-4}]$.

4.5.2 The parameters inference from small components dynamics

Independent of the above methods, an alternative method was applied to infer the values of f and δ parameters. Based on the PCM we expect that when a duplication happens in a component $C(2,1)$ we get either a component $C(3,3)$ or $C(3,2)$ with probabilities f and $1-f$, respectively. Moreover, according to the PCM an overall rate of further duplications in $C(3,3)$ is 1.5 times higher than in $C(3,2)$ components because the sum of node degrees equals 6 and 4, respectively. All bigger components $C(>3,*)$ appear as a result of one or more duplications in $C(3,*)$ components. New $C(2,1)$ components appear with the rate $\pi = 1$. For a mathematical analysis of the temporal dynamics we will denote the expected numbers of such components at time t as $n_t(2,1), n_t(3,2), n_t(3,3)$ and $n_t(>3,*)$ respectively. As described above their time dependence is given by the following set of partial differential equations:

$$\begin{aligned}
\frac{\partial n_t(2,1)}{\partial t} &= 1 - 2\delta n_t(2,1), \\
\frac{\partial n_t(3,2)}{\partial t} &= 2\delta(1-f)n_t(2,1) - 4\delta n_t(3,2), \\
\frac{\partial n_t(3,3)}{\partial t} &= 2\delta f n_t(2,1) - 6\delta n_t(3,3), \\
\frac{\partial n_t(>3,*)}{\partial t} &= 4\delta n_t(3,2) + 6\delta n_t(3,3)
\end{aligned}$$

This system of equations was solved by the Wolfram (*Wolfram Alpha*):

$$\begin{aligned}
n_t(2,1) &= \frac{(1-e^{-2\delta t})}{2\delta}, \\
n_t(3,2) &= \frac{(1-f)(1-2e^{-2\delta t}+e^{-4\delta t})}{4\delta}, \\
n_t(3,3) &= \frac{f(1-1.5e^{-2\delta t}+0.5e^{-6\delta t})}{6\delta}, \\
n_t(>3,*) &= \frac{f-9+3(4-f)e^{-2\delta t}-3(1-f)e^{-4\delta t}-fe^{-6\delta t}}{12\delta}
\end{aligned}$$

There are 4 equations and 3 unknown variables f, δ and t in this system. Therefore the goal is to find f, δ, t values that minimize a certain loss function. Here we used the weighted city block distance L:

$$L = \sum_{i=1}^{4} \frac{|\vec{n}_{t,i} - \vec{n}_{\mathrm{sd},i}|}{\vec{n}_{\mathrm{sd},i}}$$

between the following vectors:

$$\vec{n}_t = \begin{pmatrix} n_t(2,1) \\ n_t(3,2) \\ n_t(3,3) \\ n_t(>3,*) \end{pmatrix} \quad \text{and} \quad \vec{n}_{\text{sd}} = \begin{pmatrix} n_{\text{sd}}(2,1) \\ n_{\text{sd}}(3,2) \\ n_{\text{sd}}(3,3) \\ n_{\text{sd}}(>3,*) \end{pmatrix}$$

where $n_{\text{sd}}(N,E)$ is a number of components $C(N,E)$ in the SD network. Minimization of a loss function L over f, δ and t variables was performed with the Nelder–Mead method (Nelder and Mead, 1965) which converged to its minimum at $f_{\min} = 0.52$; $\delta_{\min} = 3.2 * 10^{-4}$; $t_{\min} = 1320$.

4.6 Evaluation with PCM simulations

Both independent methods that we considered in the previous subsections result in consistent predictors for the model parameters ($f_{\text{reg}} = 0.47$, $\delta_{\text{ABC}} = 5.1 * 10^{-4}$) and ($f_{\min} = 0.52$, $\delta_{\min} = 3.2 * 10^{-4}$). However, from now on we will consider only the pair: $f = f_{\text{reg}} = 0.47$ and $\delta = \delta_{\text{ABC}} = 5.1 * 10^{-4}$. These parameter values were used for PCM simulations. Topological characteristics of the PCM simulated network ($f = 0.47; \delta = 5.1 * 10^{-4}$) were compared with ones of the SD network (Fig. 4.8). Those networks are very similar in both connected component size and node degree distributions.

Moreover, characteristics of the giant component in the SD network were compared with other randomly generated networks, i.e. the configuration model network (random network of a given degree sequence), random graph, scale-free network and the giant component of the PCM synthetic network. All these networks were of the same or comparable size (number of nodes and edges) as the SD network. This was achieved by specifying the size or applying proper parameters during a network growth. The giant component of the SD network is more similar to the giant components observed in the PCM simulations than to the other networks that we used in comparison (Table 4.1).

Moreover, we have no reason to reject the hypothesis that the giant component of the SD network comes from the distribution of the biggest components of the PCM synthetic networks (empirical p-value = 0.21). This p-value was measured in the following way: 500 PCM simulations were run ($f = 0.47, \delta = 5.1 * 10^{-4}$), each time the size of the biggest connected component was saved. Based on the empirical biggest component size distribution there is no reason to think that the giant component is of "unexpected" size given the PCM model (Fig. 4.9).

We found convincing evidence that the PCM growth results in networks topologically similar to the SD network. Our predicted f and δ parameter values were both consistent between two methods and accurate in reflecting the SD network topology when used in PCM simulations. Overall, this means that the PCM model or the network growth model with preferential duplication rates reflects the growth principles of the SD network. In other words, the SD network during its evolution grew

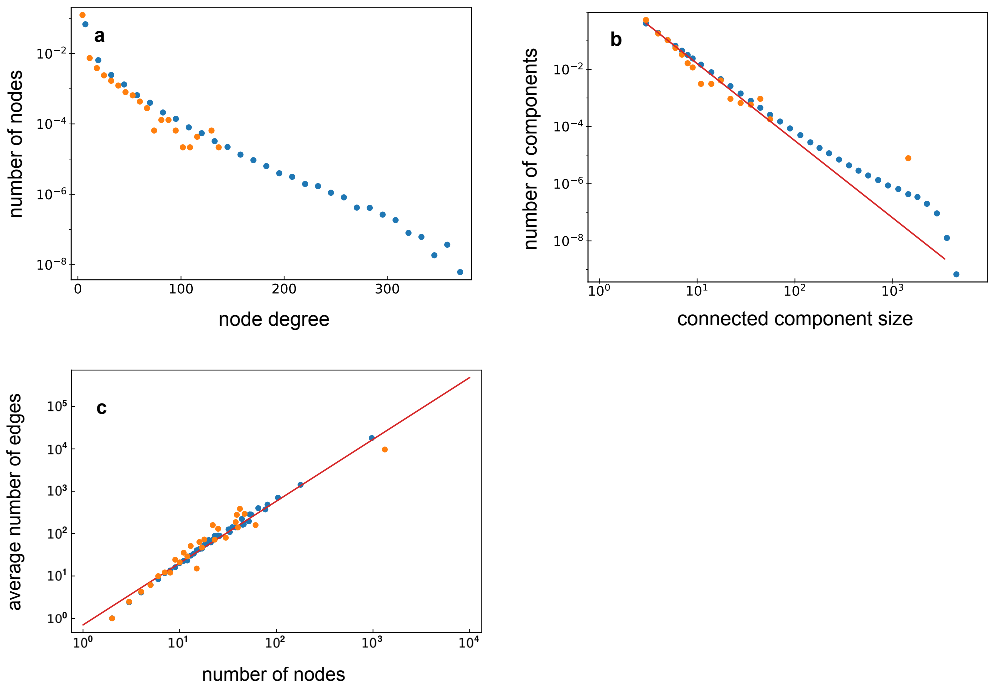


FIGURE 4.8: **Comparison of the SD network and PCM synthetic ones.** Topological characteristics of the SD network (orange dots) and the PCM simulated networks with inferred parameters $f = 0.47, \delta = 5.1 * 10^{-4}$ (blue dots) are compared. Multiple PCM simulations were pooled together to get a better resolution for the distributions. **a**. The node degree distribution is plotted (log-linear scale). We can see that the exponential tail is observed in both synthetic and the SD network and the power of exponents is the same. **b**. The connected component size distributions (log-log scale). The slopes observed in distributions are the same (where SD network components are present). The peak that corresponds to the giant component is also present where expected. **c**. The average number of edges in components of different sizes is plotted against a component size on a log-log scale. In both cases the dependence: $E \propto N^{1.47}$ (red line) is observed.

Type	Clustering coefficient	Shortest path
SD network GC	0.57	4.93
PCM network GC	0.18	3.5
Random network	0.012	2.95
Scale-free network	0.031	2.83
Configuration network	0.08	3.02

TABLE 4.1: **Comparison with other synthetic networks.** Different characteristics of Erdős–Rényi random graph, scale-free network, configuration model network (the same node degrees as in the giant component of the SD network) and the giant components (GC) observed in the SD and PCM simulated networks are compared. These characteristics include: a mean clustering coefficient and an average shortest path length. Among the networks we studied the PCM synthetic network is the most similar to the SD network (even though these are rather distinct).

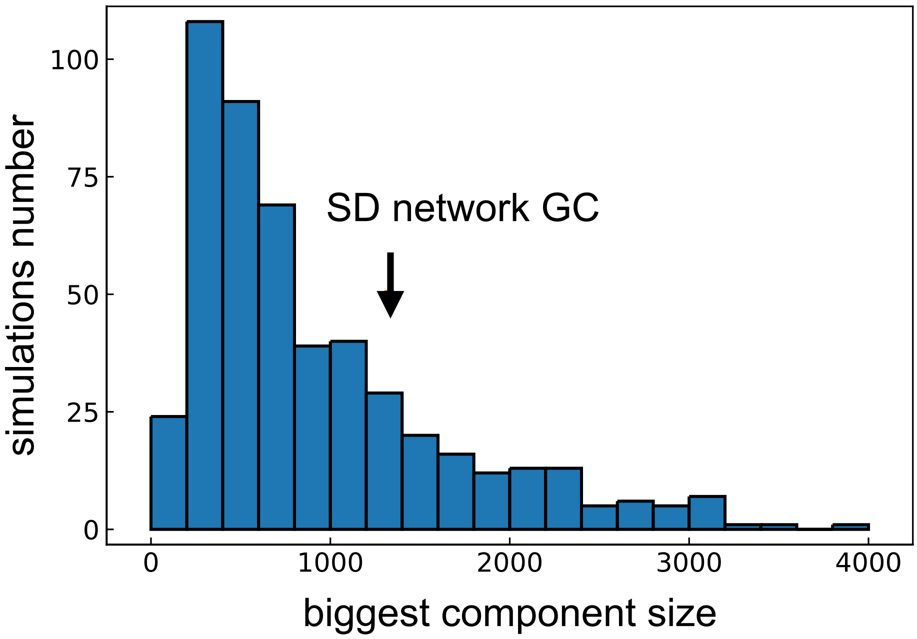


FIGURE 4.9: **The histogram of sizes of the biggest components.** These components were obtained in 500 PCM simulations. The arrow points to the size of the giant component of the SD network. According to the size distribution, there is no reason to assume that the giant component (GC) of the SD network comes from another distribution.

similarly to the PCM network growth. But what is the biological meaning of the preferential duplication rate? It means that the probability of a duplicated region to be duplicated again grows linearly with the number of copies (node degree) of that region. More precisely, the duplication probability grows linearly with the number of loci that share long homologous sequences with the region (including secondary alignments). This seems to be a fundamental "rule" for SD propagation in the human genome. In the next section we will discuss some biological scenarios explaining this propagation dynamics.

4.7 Reasons for the preferential copying model

Firstly, let's start with the most trivial explanation. The length of duplicated regions could be a major factor explaining why duplication rates grow linearly with node degree. We may expect that the probability of a duplicated region to overlap a new SD would grow with the length of the duplicated region. Simply speaking, if new SDs randomly "fall" on the genome, the longer duplicated regions are more likely to get a new duplication than shorter ones. If high node degree regions are longer - this could be an explanation for preferential duplication rates. To check this hypothesis we studied factors affecting the length of duplicated regions and, especially, effect of node degree.

One can do this by selecting those features of duplicated regions that are important in prediction of the length of duplicated regions. We used the random forest regression algorithm where the length of duplicated regions (response variable) was predicted from several characteristics of duplicated regions (predictor variables) from the untrimmed SD network. These characteristics include: a node degree, the size of a connected component a node belongs to, a mean copy number of a duplicated region, the fraction of intrachromosomal edges from all edges of a node, the number of self-loop edges and multi-edges among edges of a node. The last two characteristics can only be retrieved from the untrimmed SD network (the one where self-loops and multi-edges are not excluded). The percent of variance explained by the random forest using 10-fold cross-validation was $R^2 \sim 67\%$.

Permutation based importance values that are assigned to predictor variables by the random forest algorithm are usually affected by a number of categories and a scale of a variable. To overcome this problem the response variable was shuffled 1000 times while keeping predictor variables intact. Each time the random forest algorithm was trained on the data and all feature-specific importance values were measured. Then for each predictor variable i an empirical p-value was calculated in the following way:

$$p = \frac{\sum_{j=1}^{N_p} I(\text{imp}_j^p[i] \ > \ \text{imp}^r[i])}{N_p}$$

where N_p is the number of permutations, $I()$ is an indicator function, $\text{imp}^p[i]$ and $\text{imp}^r[i]$ are the ith feature permutation based importance values observed with and without the response variable shuffling respectively (Altmann et al., 2010). At significance level $\alpha = 0.01$ a node degree, a mean copy number of a region, the number of multiple edges and self-loops are significant in a duplicated region length prediction.

For these significant factors we can reason why they affect the length. With every new duplication of a duplicated region (which effects its node degree and mean copy-number) or duplication that "jumps" into an already duplicated region (effects the number of self-loops and multiple edges) we expect an increase of a duplicated region length. So the length of a duplicated region is influenced by the interplay of several factors, including a node degree. The corresponding node degree dependence is plotted at the Fig. 4.10. Thus we can assume a mechanistic explanation: the preferential duplication rates appear because the probability of a new SD to overlap a duplicated region is higher for longer duplicated regions.

The node degree represents the number of long sequences in other genomic loci homologous to a corresponding duplicated region. These stretches of long homology increase the probability for genomic rearrangements (including duplications). Thus with growing node degree the probability of a duplicated region to be involved in homology-mediated genomic rearrangements also grows and grows linearly. That might be another factor explaining the preferential duplication rates of the PCM.

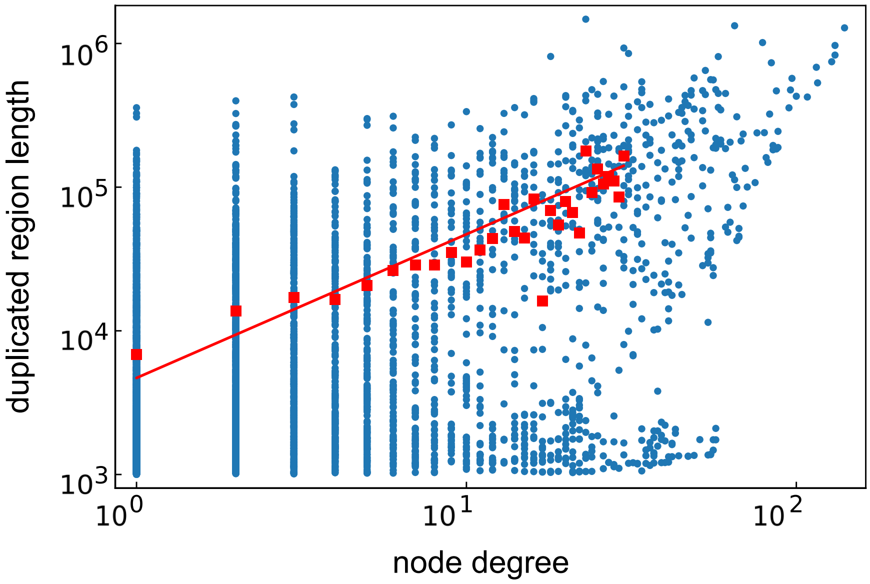


FIGURE 4.10: **Length against node degree dependence for duplicated regions.** The length (in bps) of all duplicated regions (nodes of the SD network) plotted against the node degree on a log-log scale (blue dots). Even though the observed dependence is complicated and not linear the average length of a duplicated region grows linearly with a node degree (red dots). The red line represents a linear growth on a log-log scale.

In the previous sections we studied only the SDs that were fixed in the human genome. However, the fixation process of new duplications can also be affected by the SDs that were duplicated before. To study this effect copy-number variations (CNVs) observed in 2504 individuals were downloaded from the 1000 Genomes project (Sudmant et al., 2015). All autosomal CNVs were split into 3 groups based on their frequency in the human population. There were rare, medium and high frequency CNVs with corresponding minor allele counts (MACs) in three ranges: [1; 3], [4; 15] and [16; 2504] (overall, there are 5008 haplotypes). The duplicated regions (nodes) were also split into 4 groups according to their node degree in the SD network: [1; 1], [2; 5], [6; 30] and [31; 140]. In both cases the intervals were chosen such that the number of observations in each interval is comparable. Since both distributions are highly skewed towards small values the intervals get longer for larger values.

For duplicated regions that belong to each group we studied frequencies of all CNVs that overlap those regions (Fig. 4.11). We can see that medium and high frequency CNVs are enriched in duplicated regions in comparison with the rest of the genome. Moreover, the fraction of high frequency CNVs grows with a node degree of a duplicated region, while the fraction of rare CNVs decreases. This can be explained by higher probability of recurrent duplication events in high node degree regions, variation in recombination rates or decreased purifying selection in those sites. All of these factors can affect the probability of fixation of new duplications in a population. In our case, it seems that the probability of a CNV to be fixed in a population is higher if it overlaps high node degree duplicated regions. This might be another factor explaining the preferential duplication rates of the PCM.

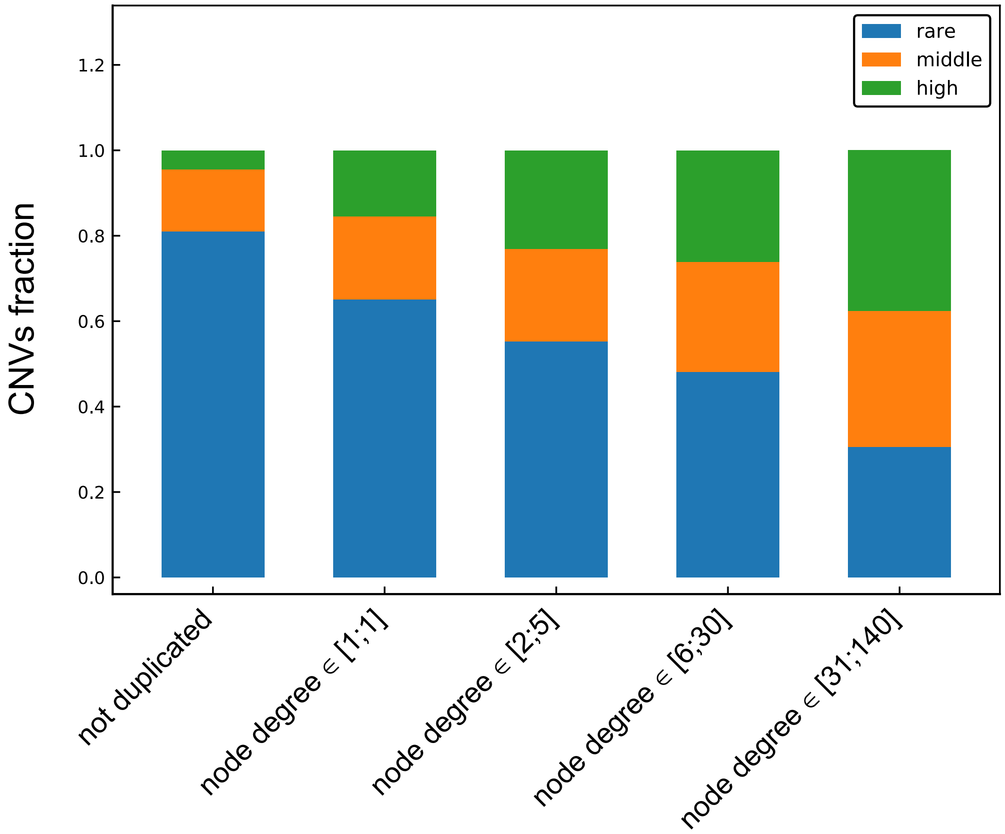


FIGURE 4.11: **Characteristics of CNVs that overlap different genomic regions.** These genomic regions include duplicated sequences of different node degrees (specified on the X axis) and the remaining not duplicated parts of the genome. All CNVs are divided into 3 groups: rare CNVs ($1 \leq$ MAC ≤ 3), medium frequency CNVs ($4 \leq$ MAC ≤ 15) and high frequency CNVs ($16 \leq$ MAC ≤ 2504) which are colored in blue, orange and green, respectively. The fraction of high frequency CNVs is higher in all duplicated regions than in the rest of the genome and this fraction grows with the node degree of duplicated regions.

4.8 Stability of our predictions

In this section we convinced ourselves that qualitative and to some extend also quantitative properties of our network analysis stay invariant under slight changes of the used cut-offs or considering uncertainties in the definition of the exact borders of segmental duplication. We considered the following stability tests addressing some critical aspects of the SD network construction.

We constructed the SD network based on duplications with reduced length and sequence identity cut-offs (length > 500 bps, sequence identity > 70%). These SDs were predicted by the SDquest tool at Pu, Lin, and Pevzner (2018) and include older duplications which are otherwise missed. This allowed us to overcome the issue that the common definition of segmental duplications filters out those duplications that are either shorter than 1000 bps or less than 90% identical.

Moreover, we parameterized the process of merging SDs into duplicated regions to see if our SD network and its characteristics are stable under different strategies of its construction. To do this we considered padded SDs, i.e. we increased the annotated length of SDs by P padding bps on both sides. Negative or positive values of P resulted in shorter or longer SDs, respectively, while $P = 0$ corresponds to our original merging process. Considering padded SDs will generate slightly different networks, since SDs will overlap less or more often, respectively.

We also checked if our predictions about network growth models are still valid if we add a process of edges loss to UCM and PCM simulations. To do this, at each time step of a network growth process we removed each edge of a synthetic network with pre-defined probability r. Only reasonable values of r were considered that do not abrupt the network growth completely, however, affect it. If addition of edges loss process makes UCM more relevant in reproducing the SD network topology or the PCM less applicable we would have to reconsider our predictions on the SD network growth.

Finally we considered the SD network constructed based on annotated SDs except for those belonging to pericentromeric regions. We defined those regions as 3 Mbp regions around centromeres. As we mentioned earlier, for complex duplicated regions in pericentromeric loci special two-step formation model was proposed (Eichler et al., 1997). Complex duplicated regions appear when a genomic locus, firstly, accumulates copies of other genomic loci and, secondly, duplicates as mosaic SDs. This duplication process is characteristic for pericentromeric parts of the genome and is not explicitly included in our copying models. We checked if the topology of the SD network stays invariant when excluding pericentromeric SDs. This would mean that the PCM is still valid as a general model of segmental duplications propagation even though we know that some fraction of duplicated regions propagated differently. If the topology changes, this would mean that the PCM seemed reasonable only when we included duplicated regions that likely propagated differently. In this case we would need to reject the PCM as a model for the SD network growth.

None of the factors above affected our results substantially or changed our conclusions about dynamics of duplication process (Table 4.2). The characteristics of SD networks stayed unchanged when we used padded SDs in construction (Fig. 4.12) and did not change substantially when we excluded pericentromeric SDs (Fig. 5.3b). Similarly, the UCM and PCM network growth where edges are lost in a course of simulation result in synthetic networks we expect from corresponding models without edges loss (Fig. 4.13). The SD network constructed on SDquest annotated SDs is larger (as expected with reduced cut-offs), however similar in all characteristics except for a mean clustering coefficient (Fig. 4.12). It is much smaller than the one observed in the normal SD network (0.17 and 0.57, respectively). The mean clustering coefficient might be higher in the normal SD network because that network includes more confident alignments than the SDquest based one.

Overall, our observations show the stability of our predictions given some technical variations in the SD network construction process.

SD networks:	Nodes	Edges	Intra- (%)	Tandem (%)	Shortest path	Clustering
original SDs ($P = 0$)	6656	16,042	29	9	4.93	0.57
SDquest SDs	9605	34,986	22	4	5.62	0.17
no centrom. SDs	5771	10,860	32	11	5.17	0.63
padding ($P = -100$)	7281	17,166	30	10	4.89	0.58
padding ($P = -50$)	7266	17,155	30	10	4.89	0.58
padding ($P = 50$)	6322	15,550	27	8	4.83	0.56
padding ($P = 100$)	6213	15,423	29	8	4.87	0.56

TABLE 4.2: **Characteristics of several alternatively constructed SD networks.** The normal SD network that we used everywhere by default ($P = 0$), several SD networks with different paddings P used in construction, SD networks built from SDquest predicted SDs and SDs excluding pericentromeric ones (see the main text). The characteristics include: number of nodes and edges, fraction of intrachromosomal and tandem edges among all edges, a mean clustering coefficient and an average shortest path length. We can see that all characteristics of the SD networks are stable when using different paddings P. The SD network constructed on SDquest annotated SDs is larger (as expected with reduced cut-offs), however similar in other characteristics except for the mean clustering coefficient.

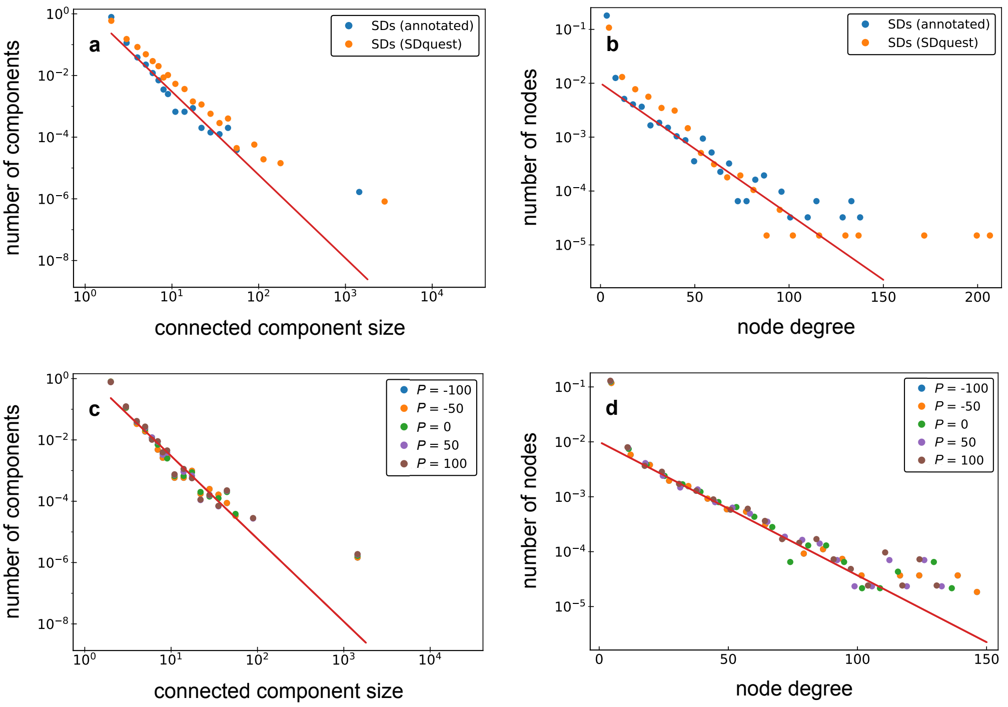


FIGURE 4.12: **The SD networks constructed differently are compared to check a stability of predictions.** There connected component size (**a**) and a node degree (**b**) distributions of original SD network and the one constructed based on duplications predicted with SDquest ($>$ 500 bps, sequence identity $>$ 70%) are plotted. Even though the size of alternative SD network is larger we can see that distributions are similar both in terms of slopes and giant component presence. The connected component size (**c**) and node degree (**d**) distributions of multiple SD networks with paddings P are plotted. This parameter represents number of bases added to extend (if $P > 0$) or shorten (if $P < 0$) each SD interval on both sides before constructing an alternative SD network. The value of $P = 0$ corresponds to our original SD network. We can see that even with quite large absolute values of P parameter both distributions stay pretty much unchanged in all networks.

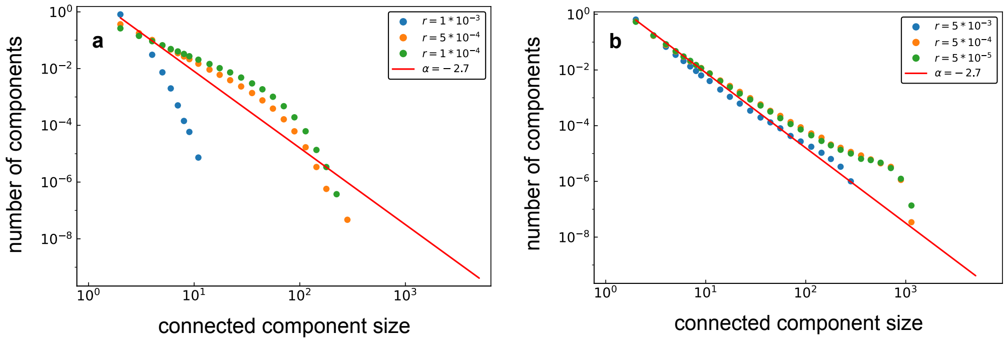


FIGURE 4.13: **The network growth models are studied when edges loss process is added.** The connected component size distributions observed in simulations of network growth are plotted on a log-log scale. Additional process of edges loss was added in both UCM (**a**) and PCM (**b**) to check how it affects the topology of resulting synthetic networks. At each time step each edge is removed with the probability r. Red line represents the slope of the distribution observed in the SD network. One can see that when using reasonable values of r both models of network growth behave as expected in standard UCM and PCM simulations. Too large values of r can hinder any network growth (like in UCM simulation with $r = 1 * 10^{-3}$).

4.9 Summary

In this chapter we studied segmental duplications in the human genome using approaches from complex network theory. We first constructed a network of segmental duplications which we called the SD network. Every node in it corresponds to a duplicated genomic region, edges are added if homology between a pair of duplicated regions exists. We studied topological characteristics of the SD network, and found that these were distinct from those observed in other well-known networks in the field of complex network theory. So we decided to simulate a network growth according to some predefined "rules" to find a model that reproduces the SD network topology in simulation. We used copying models of network growth where node duplication rates were defined differently. The trivial model with constant duplication rates (UCM) was unable to explain the SD network topology, while more complex preferential copying model or PCM was good for this task. This means that SDs, likely, evolved according to a simple dynamical principle: the probability of a duplicated region to be duplicated again grows linearly with its number of copies. Several biological scenarios were suggested to explain preferential duplication rates. Firstly, this effect can be explained mechanistically: duplicated regions with high node degree are usually longer, thus more likely to overlap new duplications. Secondly, the probability of fixation of new duplications (studied on population CNVs) is, likely, higher in high node degree regions. This might be because of reduced purifying selection, recurrent duplication events or other reasons, however, the probability of fixation seems to be higher in those sites. Finally, every edge of the SD network represents a long stretch of homology which can be a source for further genomic rearrangements. We also considered the stability of our prediction: whether alternative strategies of the network construction, different sets of SDs, changes in the UCM or PCM settings etc. affect our predictions. It turned out that our predictions about SD propagation principles are reproduced independently of technical variations.

Chapter 5

Segmental Duplications in Genomes of Other Species

5.1 Limited accuracy of predicted duplications

Even though there were many attempts to predict SDs in reference genomes of non-human species with WGAC algorithm - the overall consensus is that the quality of such predictions is quite limited. The main reason is that genome assemblies (especially old outdated ones) of non-human species are less accurate than the human reference. The last one was generated with hierarchical shotgun sequencing which allowed to partly overcome the main limitation of short-read sequencing: poor performance in low-complexity redundant sequences (IHGSC, 2004). There is a tendency to collapse repeated sequences into single contigs or completely miss those sequences when short-read *de novo* assemblies are constructed (Salzberg and Yorke, 2005).

On the other hand, the opposite error could arise when assembling diploid (or polyploid) genomes. When a locus is too heterozygous it is often erroneously assembled into duplicated sequence. Two contigs are formed from reads belonging to respective copies of chromosomes, which look like a duplicated sequence in resulting assembly. The last type of misassemblies was extensively studied and quantified by (Kelley and Salzberg, 2010). Paired-end reads that map to contigs with duplicated sequences were studied in 4 reference genome assemblies (chimpanzee (*panTro2*), domestic cow (*UMD1.6*), chicken (*galGal3*) and cow (*canFam2*)). The distances between mates in read pairs were compared with what we expect in corresponding whole-genome sequencing platform setup as illustrated at Fig. 5.1. If merging contigs into one resolves unrealistic distances between reads in pair, this likely means that observed duplications are erroneous. The analysis revealed that large fraction: 75% (14.4 Mbp) of chicken, 56% (2.3 Mbp) of cow, 81% (16.7 Mbp) of chimp and 10% (9.7 kbp) of dog contigs with duplicated sequences are result of erroneous assembly of divergent loci. This illustrates the fact that non-human reference genomes, especially their old versions, include multiple assembly errors associated with duplicated or variable DNA sequences.

Finally, the human reference genome was extensively studied and corrected which

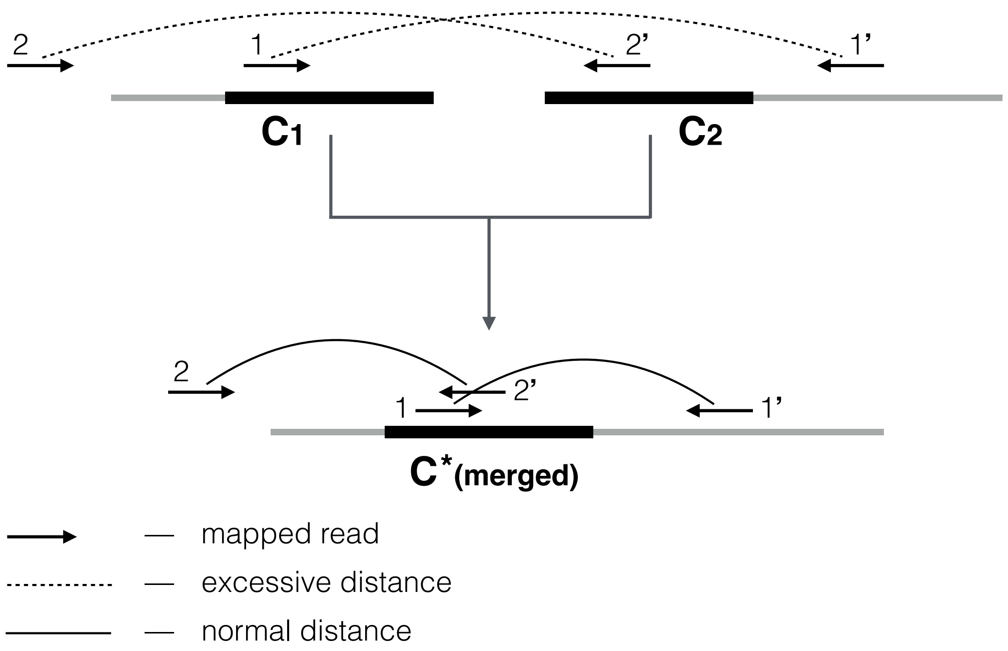


FIGURE 5.1: **The scheme illustrates the principle of erroneous duplications detection.** Two contigs with copies of a duplicated region C_1 and C_2 are given. Grey lines represent unique sequences. If merging contigs into one results in more realistic distances between read mates mapped to an assembly, this likely means that unique C^* sequence was misassembled. The figure was adopted from Kelley and Salzberg (2010).

is usually not the case for other species. Many assembly errors and gaps in the human reference genome were resolved in later *de novo* assemblies (Jain et al., 2018; Miga et al., 2020).

The advance of long-read sequencing allowed to construct more accurate genome assemblies for some non-human species (He et al., 2019; Hon et al., 2020; Jagannathan et al., 2021). High copy-number duplicated regions, long SDs and long stretches of low-complexity DNA sequence were, as expected, especially often misassambled in previous short-read assemblies. Anyway, even though the process of genome assemblies refinement got a giant bust when long-read sequencing technology became accessible, this process is far from being over (even in relatively accurate human genome (Nurk et al., 2021)).

5.2 Reconstruction of lineage-specific duplication events

In this section we will concentrate on the research project which provided us with the data on duplicated regions comparative genomics (Sudmant et al., 2013). We want to describe it in more detail and discuss how the data correlates with our analysis of the SD network.

The attempt to predict segmental duplications in great ape genomes and to assign corresponding duplication events to ancestral phylogenetic lineages was done by Sudmant et al. (2013). The dataset included 97 individual WGS samples, 75 of which originated from the Great Ape Genome Diversity Project (Prado-Martinez et al., 2013) and the remaining 22 were from the Orangutan Genome Project and the Denisova Genome Project (Locke et al., 2011; Meyer et al., 2012). These samples included: Bornean orangutans (9 samples), Sumatran orangutans (8 samples), humans (10 human and 1 Denisovan samples), bonobo (14 samples), western chimpanzee (5 samples), Nigerian–Cameroon chimpanzee (10 samples), eastern chimpanzee (6 samples), central chimpanzee (2 samples), western gorilla (29 samples), eastern gorilla (3 samples) and cross-river gorilla (1 sample). Even though some of these taxa are classified as species while some are subspecies, for simplicity, we will further refer them all as the great ape species. Reads from the listed ape samples were mapped on the human reference genome and the copy-number variable regions were detected based on coverage analysis. This allowed to determine the absolute copy number of loci and the breakpoints at an individual genome level which were further reported in human genome based coordinates. All predicted copy-number variable regions were classified as fixed in specific lineage, copy number polymorphic or private (present in one individual only). This method has an advantage over CGH arrays that were earlier used for the task of CNVs annotation in great ape genomes (Dumas et al., 2007; Gazave et al., 2011; Fortna et al., 2004). The breakpoints and a copy-number of CNVs can be predicted more accurately in the approach based on read coverage. Moreover, the biases associated with the fact that great ape samples are hybridized against DNA probes originating from the human genome further limits the resolution of CGH arrays.

There are 11,836 fixed duplications (325 Mbp), 5528 fixed deletions (47 Mbp) and 6406 CNVs (96.2 Mbp) detected by Sudmant et al. (2013) which overall comprises around 16% of hominid genome. It was found that long duplications that are fixed in great ape species (or segmental duplications) are distributed non-uniformly and tend to happen close to already duplicated genomic regions. On the other hand, fixed deletions are distributed independently with respect to each other. The fact that new SDs tend to happen in already duplicated regions is called "duplication shadowing" and was already observed earlier (Cheng et al., 2005; Marques-Bonet and Eichler, 2009; Newman et al., 2005). For example, approximately half of SDs shared between human and chimp are mapped within 5 kbp of SDs shared among human, chimpanzee and orangutan, while third of human-chimpanzee-orangutan duplications map adjacent to human-chimpanzee-orangutan-macaque shared SDs (Marques-Bonet and Eichler, 2009). Ancestral SDs are often prone for consequent duplications thus leading to duplication shadowing and recurrent duplication events. It could happen in several scenarios, for example, because of long homologous stretches of duplicated regions or because of repeats in flanking sequences increasing the NAHR probability. This observation indirectly agrees with our prediction that duplication rates grow with a number of copies of a region. This behaviour can be seen as duplication events that happen with a preference to already duplicated loci. In other words, duplications of an already duplicated region are more often than what we expect if SDs fall randomly on the genome. If we imagine this scenario in terms of network growth, we would expect that only components $C(2,1)$ are added to a network with hardly any node duplication events happening.

By the means of comparative genetics, duplications and deletions were assigned to specific lineages. Such a phylogenetic reconstruction of SDs is complicated by the recurrent duplication events. The rate of duplication events homoplasy in great apes is estimated around 20% (Marques-Bonet and Eichler, 2009). Overall, the number of duplicated nucleotides is ~ 3 times higher than deleted ones. Lineage-specific duplication rates normalized by a lineage time span (measured in Mbp per million years) are quite variable across the great ape phylogenetic lineages. The fastest rate was in ancestral African great ape lineage (6.61 Mbp/Mya) with further decay in gorilla ancestor and human-chimpanzee ancestor to 4.46 and 3.02 Mbp/Mya, respectively (Fig. 5.2a). This means a burst of duplications early in African hominid evolution which was also observed by Marques-Bonet and Eichler (2009) and contradicts earlier suggestions that excess of duplication events is specific for the human ancestor lineage (Olson, 1999; Varki, Geschwind, and Eichler, 2008). Deletions, on the other hand, happened in a relatively clock-like manner with some acceleration in chimpanzee-bonobo ancestor lineage (Fig. 5.2b).

There are 407 and 340 lineage-specific duplication and deletion events, respectively, that, at least partly, overlapped genes in the course of great ape evolution. Specifically, there are 33 gene duplication events that are private for human genome. The highest rates of gene duplications were characteristic to African great ape and human-chimpanzee ancestors, while the highest gene deletion rate was in chimpanzee-bonobo ancestor. One possible explanation for increased deletion rate in the chimpanzee-bonobo ancestor lineage is that it went through several population bottlenecks that reduced purifying selection acting on large deletions (Prado-Martinez et al., 2013).

The considered data was also used in our project. This will be further described in a more comprehensive way. If, as for now, we ignore details, the copy-number of a locus is similar in its meaning to a node degree of a corresponding node: both represent a number of homologous sequences. Thus we can study how a node degree of a specific node changes in different species or even, informally speaking, in parallel evolutionary experiments. Unfortunately, this reconstruction does not always allow to assign the ancestral copy-numbers, because many SDs originated before hominid evolution. However, we overcame this issue by studying how evolutionary dynamic a locus is instead of how many times it duplicated after its hominid ancestral state (see below).

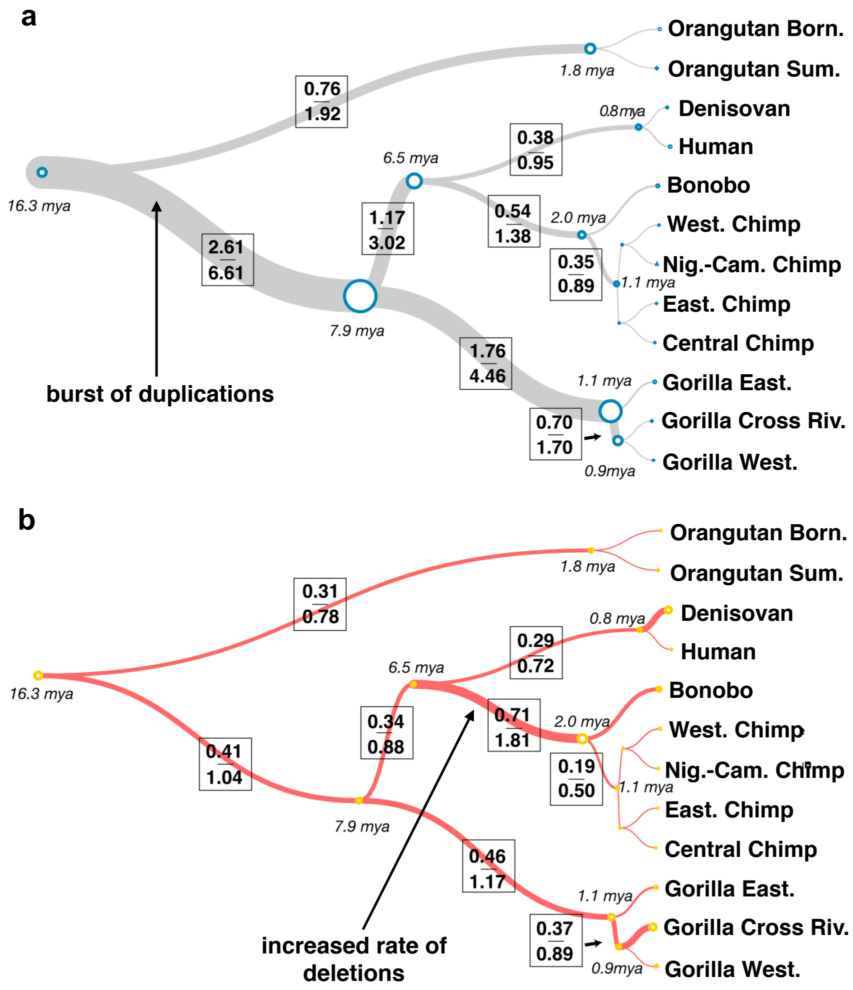


FIGURE 5.2: **Duplication (a) and deletion (b) rates are plotted along the great ape phylogenetic tree.** The width of branches is scaled proportionally to duplication or deletion rates, respectively. Two rates are assigned to the branches: the number of duplicated (**a**) or deleted (**b**) bps normalized by the number of substituted ones (above in pairs of values) and Mbp/Mya (below). The highest duplication rate was characteristic for the African great ape ancestor lineage, which further dropped in the chimpanzee–human and gorilla ancestor lineages. Deletions happened in a relatively clock-like manner in the course of great ape evolution. The figure was adopted from Sudmant et al. (2013).

5.3 SD networks of other species

In this section we tried to answer the following question, whether the SD networks of other species which evolved independently from humans are similar to the human SD network or may have resulted from other growth scenarios. We therefore downloaded the latest reference genomes of 8 additional species from the UCSC genome browser. The list of reference genomes includes: human (*hg38*), gorilla (*gorGor4*), gibbon (*nomLeu3*), mouse (*mm10*), rat (*rn6*), dog (*canFam3*), chicken (*galGal6*), zebrafish (*danRer11*) and C. elegans (*ce11*). Not for all considered genomes SD annotations exit. To overcome this problem and to make all annotations provided by the same tool we did *de novo* SD prediction ourselves. The SEDEF tool was used to annotate segmental duplications in the genomes (Numanagic et al., 2018); again, only autosomes were included in the analysis. For an accurate comparison we also used the same tool to *de novo* predict SDs in the human genome. Based on SDs identified by SEDEF tool SD networks were constructed for the above species including human. The human SD network built from the SEDEF predicted SDs was compared with the original one (Fig. 5.3a). The SEDEF predicted SD network is larger both in terms of the number of nodes and edges (Table 5.1), however, almost all duplicated regions from the original SD network were present in it. In fact, the number of SEDEF predicted SDs was higher than in the original UCSC annotation. This is due to the fact that SEDEF by default reports duplications satisfying less strict length and sequence identity criteria plus the UCSC annotation underwent additional filtering steps, such as: agreement of WGAC and WSSD methods predictions, FISH validation etc. (Bailey et al., 2001; Bailey et al., 2002).

	Human	Gorilla	Gibbon	Mouse	Rat	Dog	Chicken	Zebrafish	Worm
GS (10^9 bp)	2.88	2.78	2.65	2.46	2.62	2.2	0.96	1.35	0.083
Num. of nodes	12,579	30,935	29,376	14,766	35,919	18,438	3,169	34,445	1,572
Num. of edges	37,319	42,643	443,916	166,145	183,618	62,308	17,102	601,289	2,199
Intra- (%)	19	46	7	8	24	7	36	9	37
Tandem (%)	6	25	2	2	10	3	8	1	17
f value	0.48	0.43	0.57	0.42	0.42	0.33	0.29	0.35	0.32

TABLE 5.1: **Characteristics of the SD networks of different species.** These include: genome size (GS) excluding sex chromosomes, number of nodes, number of edges, fraction of intrachromosomal edges and tandem edges among all edges of a network and regression-based predicted f values. An edge is denoted as tandem if both duplicated regions linked with the edge are located at the same chromosome at the distance < 0.5 Mbp. One can see that the sizes and characteristics of the SD networks are quite distinct. Moreover, the human SD network constructed with SEDEF predicted SDs is substantially larger than the one built on UCSC annotated SDs (includes 6,656 nodes and 16,042 edges).

The resulting SD networks of different species are quite distinct in their sizes and other network characteristics (Table 5.1). The component size distributions, on the other hand, are similar both in terms of the slope of the distributions and in the presence of a giant component (Fig. 5.4). Similarly to the human SD network we also observed a clear power-law growth of the average number of edges with component size in all species. Corresponding regression-based predicted values of the parameter f are listed in the Table 5.1. The genome and the SD network of C. elegans are the smallest ones, thus we do not see a prominent giant component as we observe in other species. We would like to note that the SDs shared by the species are very unlikely to be responsible for such a similarity in the SD networks topologies. As we already said, based on sequence identity levels most of predicted segmental duplications appeared only after the divergence of the New and Old World monkeys.

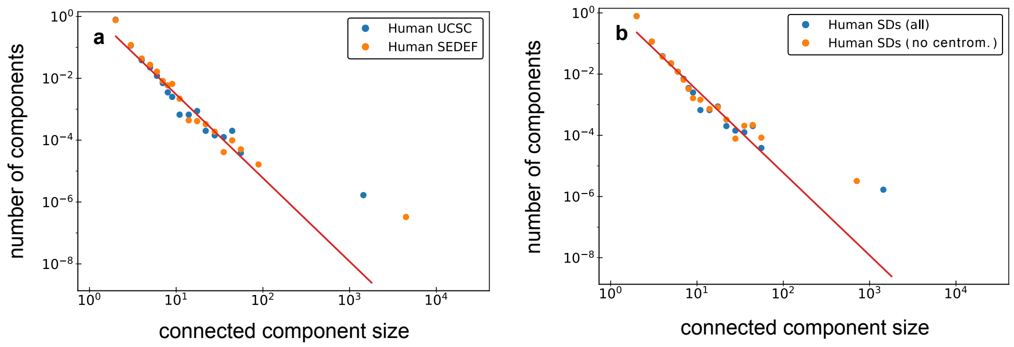


FIGURE 5.3: **SEDEF based, original SD network and the one constructed without pericentromeric regions.** The connected component size distributions plotted on a log-log scale with logarithmic binning. **a**. Comparison of the SD networks constructed from the UCSC annotated and SEDEF predicted SDs. **b**. The SD networks constructed based on all SDs of the human genome and all SDs excluding pericentromeric ones. In both cases we observe similar distributions.

When we calculated the Bray-Curtis pairwise dissimilarities between component size vectors we found that the species cluster similarly to their phylogenetic relationships, for example, primate and mammalian clusters were observed (Fig. 5.4d). Thus, based on our data, the topology of the SD network seems to be reflective of phylogenetic relationships among species which might be indicative of a shared slowly evolving molecular mechanism responsible for the continuous spread of segmental duplications in genomes.

5.4 Analysis of CNVs in ape genomes

We used a data from Sudmant et al. (2013) (which was earlier introduced in the "Reconstruction of lineage-specific duplication events" section 5.2) to see if evidence from comparative genomics of great ape species supports the preferential model (or PCM) of SD evolution. In this study reads from whole-genome sequencing samples of 12 great ape species were mapped on the human reference genome and copy-numbers of corresponding homologous loci were measured based on a read coverage depth. For each ape specie several individual samples were used: Bornean orangutans (9 samples), Sumatran orangutans (8 samples), humans (10 human and 1 Denisovan samples), bonobo (14 samples), western chimpanzee (5 samples), Nigerian–Cameroon chimpanzee (10 samples), eastern chimpanzee (6 samples), central chimpanzee (2 samples), western gorilla (29 samples), eastern gorilla (3 samples) and cross-river gorilla (1 sample). The resulting data we used includes copy-numbers of multiple genomic loci (each is defined in human reference genome coordinates) observed in all studied samples, from them only duplications were further used.

For each of 11,262 loci which were involved in segmental duplications during the ape evolution and fixed in at least one specie we calculated the metrics representing how dynamic (in terms of duplications) a region was. We first calculated

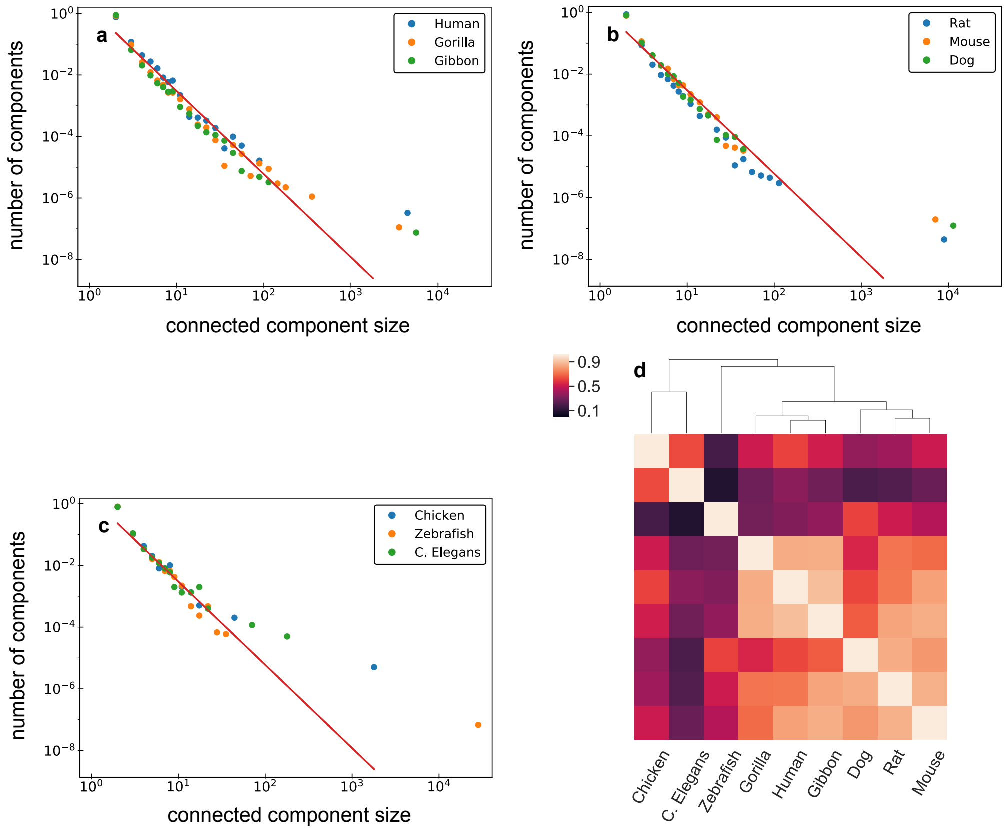


FIGURE 5.4: **The connected component size distributions plotted for the SD networks of different species (log-log scale).** The red lines in panels (**a**) - (**c**) represent the slope observed in the SD network of human. Observed distributions follow this slope on a log-log scale and a giant component is observed in most species. **a**. The group of primate species that includes human, gorilla and gibbon. **b**. The group of other mammalian species that includes rat, mouse and dog. **c**. The group of distinct species that do not belong to mammals: chicken, zebrafish and C. elegans. **d**. The heatmap of 1− Bray-Curtis dissimilarities between connected component size vectors in all the species. We can see that the hierarchical clustering dendrogram on top, to some extent, reflects phylogenetic relationships between the species (for instance, presence of primate and mammalian branches).

a mean copy-number of each genomic region over all samples of a specific specie. This was done to overcome the bias associated with different number of samples per specie. Then we used the first central moment of a copy-number or the following value: $\Delta_i = \sum_{j=1}^{12} \frac{|X_{ij} - \overline{X_i}|}{12}$ where X_{ij} is a copy-number of ith genomic region in jth specie (there are 12 species in the dataset), while $\overline{X_i}$ is a mean copy-number of a corresponding region. One of the reasons why we did not use a copy-number gain over the ancestral state is that for those loci which started to duplicate early in ape evolution the ancestral copy-number can not be accurately suggested.

Then we measured how dynamics Δ depends on characteristics of nodes from the human SD network (Fig. 5.6). The borders of duplications in ape species not necessarily match the borders of duplicated regions we used so we calculated the

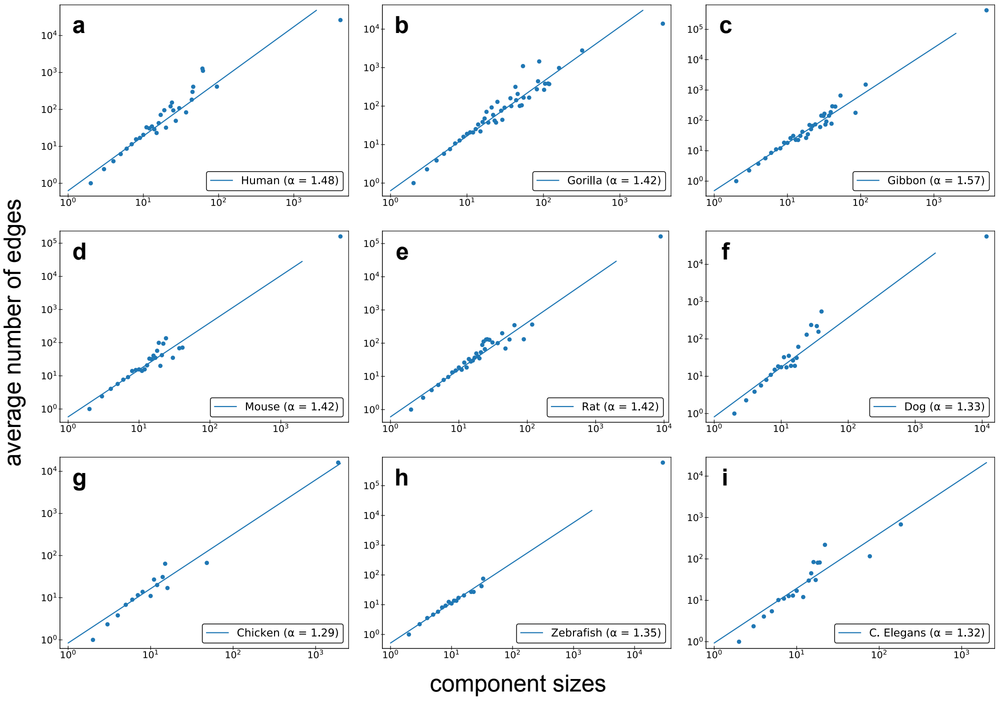


FIGURE 5.5: **The E against N dependence in connected components of different species.** The plots (**a**) - (**i**) show how the number of edges E in connected components grow with their size N in different species. For all sizes the mean value of E is plotted. One can see that $E \propto N^{\alpha_s}$ dependence is present in all species, where α_s values are specie-specific and listed in the legends.

mean Δ value over all intervals overlapping specific duplicated region. One can see that the duplication dynamics Δ of duplicated regions clearly grows with a node degree and a connected component size. This means that the giant component and especially high node degree duplicated regions in it are the most actively duplicating loci in apes. This observation agrees with the PCM growth that we suggested. This trend can be reproduced if we use another measures for duplication dynamics, like the second central moment or normalized $\Delta_i/\overline{X_i}$ (see Appendix). Moreover, the genomic loci that were duplicated recently in the human lineage mostly overlap the duplicated regions belonging to the giant component while older duplications that are shared by several primate species overlap smaller components more often (Fig. 5.6c). This again agrees with the PCM growth. In early phases of PCM network growth a giant component is not yet formed thus connected components grow in a relatively uniform manner, however, when a prominent giant component appears, majority of duplication events start to concentrate there (similarly to the "rich gets richer" principle). Thus the fact that recent (or human specific) SDs are enriched in the giant component agrees with such growth.

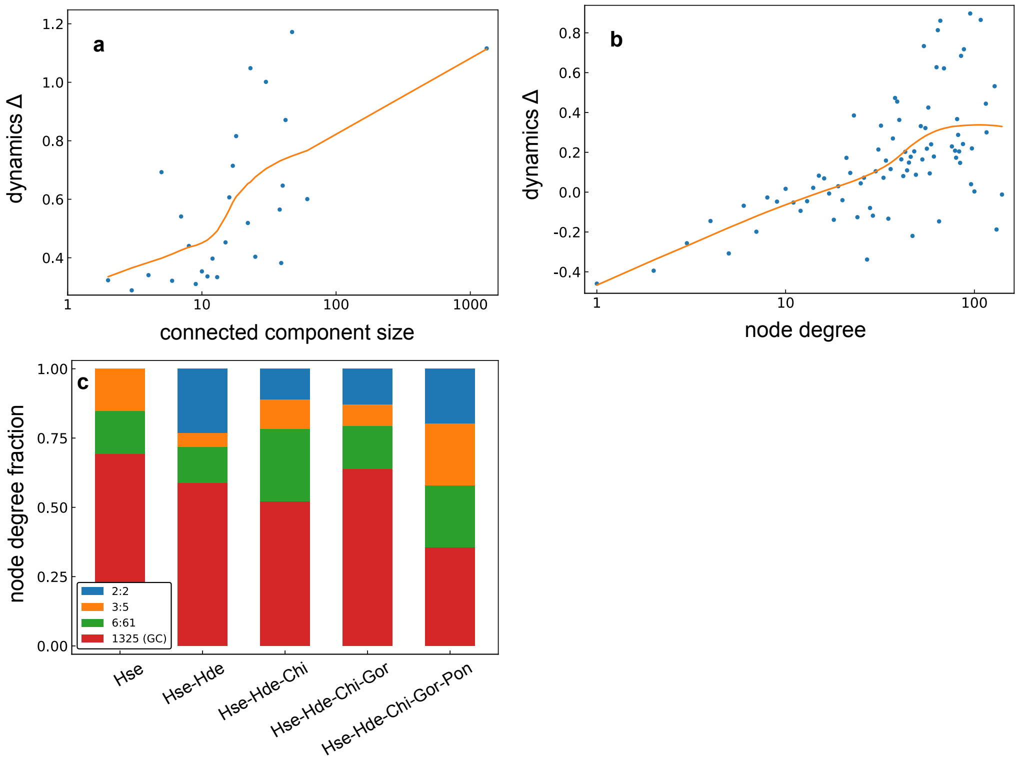


FIGURE 5.6: **The plot represents how the mean duplication dynamics Δ depends on network characteristics.** The average dynamics grows with a connected component size (**a**) and a node degree (**b**) which are plotted on a log-linear and a log-log scales respectively. This agrees with the PCM growth that we suggested for the SD network. To illustrate the growing trend the locally estimated scatterplot smoothing function loess($\alpha = 0.75$, $degree = 1$) was added (orange line). **c**. All nodes of the SD network were divided into groups based on a component size while human lineage SDs based on their age. We grouped SDs according to a phylogenetic lineage of origin (marked by a corresponding set of ape descendant species on X axis) and visualized nodes that overlap those groups. The ids we used to refer the species: *Hse* (human), *Hde* (Denisovan), *Chi* (chimpanzee), *Gor* (gorilla) and *Pon* (orangutan). One can see that the SDs exclusively observed in human genomes (evolutionary recent SDs) mostly happened in the giant component while older ones are distributed in components of various sizes.

5.5 Summary

In this chapter we studied segmental duplications in various non-human genomes. We *de novo* annotated segmental duplications in 9 reference genomes (human, gorilla, chimp, rat, mouse, dog, chicken, zebrafish and C. elegans). It turned out that even though the SD networks were quite distinct in their size, the network topologies were similar in all cases, except for the genome of C. elegans (which is, probably, too small). This supports our believe that the PCM network growth is characteristic not only for human genome, but likely for the clade of vertebrate species.

We also took advantage of the data from Sudmant et al. (2013) where copy-numbers of human duplicated regions were studied in several ape species. We inferred duplication activity of human duplicated regions by looking at intensity of copy-number changes in homologous sequences in ape species. We tested the correctness of the PCM in the following way: if high node degree regions are more evolutionary dynamic in other ape species - this would support the PCM, otherwise we would have to reject one. As a result, we found that high node degree regions are more evolutionary dynamic in ape species than low node degree ones.

Chapter 6

Duplication Events Reconstruction from the SD Network

6.1 Articulation of the question

The preferential copying model (PCM) for growth of the SD network was suggested to explain its topology and gives us a potential to reconstruct duplication events responsible for its formation. Edges of the SD network represent alignments of homologous sequences that share common origin, however, the information on the temporary order and direction of duplications (i.e. which genomic region among two copies is ancestral) is missing. Moreover, edges of the SD network represent either real duplication events or secondary alignments that appear because of overlaps between independent duplications. Reconstruction of real duplication events from the whole network of duplicated regions will allow us to further look into biological factors responsible (or at least associated) with segmental duplications.

This task can be illustrated with the following analogy: let's say we have a graph where each node represents one individual and edges are added if two individuals know each other. All individuals know their parents, but also have some more social connections either inherited from their parents or generated throughout their lives. Then our task is to find a real family tree (or reconstruct all parent to child edges) in the graph with unspecified relationships between individuals. In our case, this "recollection" of existing duplication events from the network is possible if reasoning of PCM is taken into account as a key.

Accurate reconstruction of duplication events in a complex genomic locus is a complicated bioinformatics task. It answers the question *how* some specific region was formed, however, more global question of *why* some genomic regions duplicate more often than others is obscured by such a reductionist approach. Even though our network approach disregards some information related to duplicated regions when it is studied as a node, it allows studying segmental duplications *as a whole* thus revealing some biological principles of SD creation from the SD network. In a nutshell, we reconstructed duplication events from the SD network to predict the number of duplications that each genomic region did in the course of evolution.

Then we solved a machine-learning task of predicting important genomic features associated with number of duplications of genomic region (see details below).

6.2 The origin of cycles in the SD network

Before going into our method of reconstruction, firstly, we have to understand the nature of the cycles in the SD network. Cycles appear as a result of duplication when a new "daughter" node inherits the edge to the neighbor of the "mother" node from it (Fig. 6.1a). One conclusion from this is that only cycles of size 3 can appear as a result of such secondary edges acquisition in PCM. Indeed in a cycle of size N where $N > 3$ we expect at least one event of secondary edge acquisition (as the only way to get cyclic structure). On the other hand, it is impossible to acquire a neighbor from the "mother" node that does not have one (Fig. 6.1b), thus only cycles of size 3 appear as a result of node duplications in the PCM. In agreement with this we observe that out SD network is depleted with cycles of size > 3 in comparison with other networks of well-known topology (Table 6.1). These bigger cycles are likely observed in our network because of superposition of several cycles of size 3 (Fig. 6.1c). Thus it is more informative to look at shortest self-paths of nodes (shortest paths from a specific node to itself if it exists). In dramatic contrast to other networks only for 1 out of 1325 nodes we observed a self-path longer than 3 edges, which is substantially less than what we see in other networks (Table 6.1). These observations show us that the SD network likely evolved according to one of our copying models.

Type	Clustering coef.	Shortest path	$>$ 3-cycles (%)	$>$ 3 self-paths (%)
SD network GC	0.57	4.93	8	0.1
Random network	0.012	2.95	96	37
Scale-free network	0.031	2.83	94	25
Configuration network	0.08	3.02	70	36

TABLE 6.1: **Cycles composition in various networks.** The characteristics of several networks that include: random graph, scale-free network, the giant component of the SD network and corresponding configuration network (random graph with node degrees of original network preserved) are listed in the table. First three columns represent the mean clustering coefficient, the mean shortest path between 2 nodes and the fraction of cycles of size larger than 3 among all cycles in the network. The SD network is strongly depleted with large cycles (or enriched with the cycles of size 3) in comparison with other network types. We also measured the length of a shortest self-path from a specific node to itself where possible (column 4) and calculated the fraction of self-paths that cover more than 3 edges. The depletion of long self-paths is even more prominent than in case of long cycles (0.1% of all self-paths). In all cases we used networks of the size comparable to the giant component of the SD network.

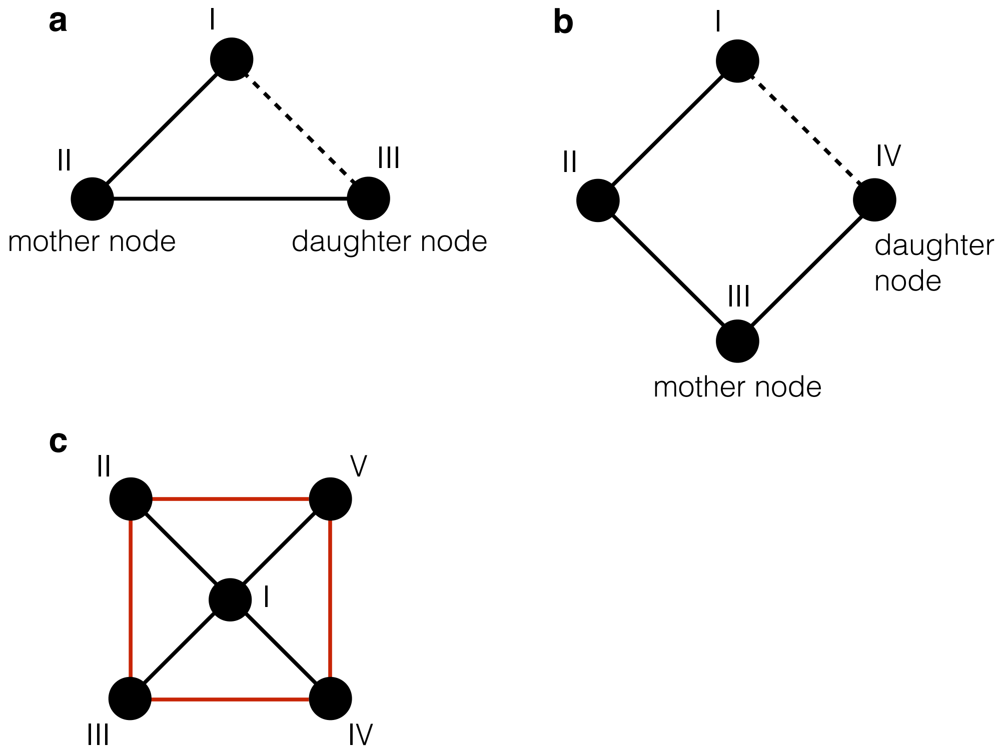


FIGURE 6.1: **Origin of cycles in the PCM. a**. The scheme illustrates that cycles of size 3 can appear as a result of a duplication event when a daughter node inherits an edge from a mother node (the dotted line represents the secondary edge). **b**. On the other hand, cycles of size 4 or more can not appear in our copying model. The node IV can not inherit the edge $(I,\ IV)$ from its mother node III, because the last one does not have the node I among its neighbors. Thus we do not expect cycles of size 4 or larger in our PCM network except for those cases when superposition of several cycles of size 3 forms cycles of larger size. Such an example is present at the scheme **c**. The cycle of size 4 (coloured in red) appears as a result of 4 cycles superposition: $(I,\ II,\ III)$, $(I,\ II,\ V)$, $(I,\ IV,\ V)$ and $(I,\ III,\ IV)$.

6.3 Principles of duplication event reconstruction

6.3.1 Reconstruction examples: graphs with and without cycles

If we disallow any inheritance of neighbors from the "mother" node (e.g. by simulating PCM with $f = 0$) we will observe a network without any cycles where each edge represents a real duplication event, while connected components unite duplicated regions of the same ancestry. Directionality of duplication events can be reconstructed in unique way if we assume that we know which node was the ancestral one in a connected component (Fig. 6.2). If we compare two figures (Fig. 6.2a and Fig. 6.2b) we will see that directionality of edges changes when taking another ancestral nodes while number of duplications that happened in each node is almost invariant to this choice. Thus for our task it is enough to find edges in the SD network that correspond to real duplications while assigning directionality is not that important.

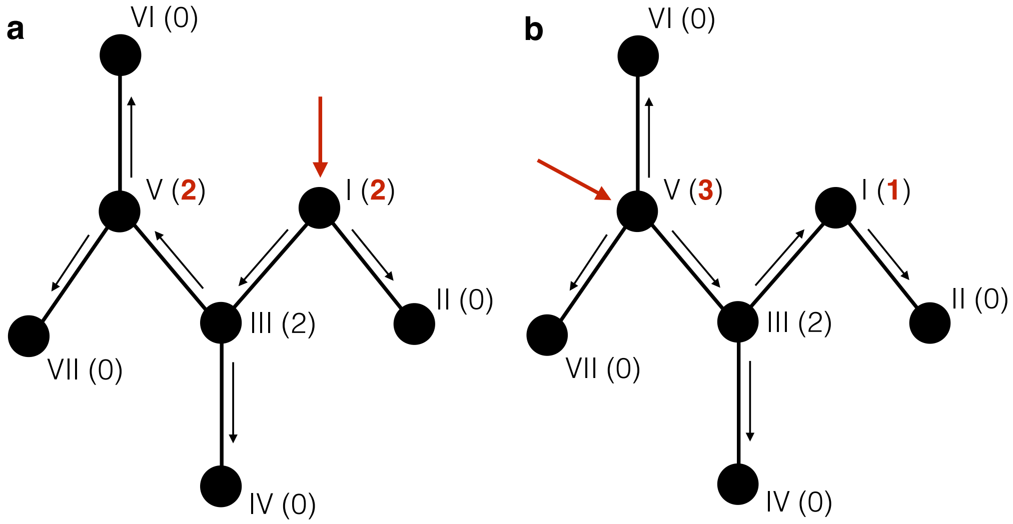


FIGURE 6.2: **Reconstruction of duplication events in a graph without cycles. a**. The scheme illustrates the fact that reconstruction of duplication events is a straightforward task when we study a network without any cycles. This type of network can only appear if no secondary edges were inherited during the network growth (for example, if we run PCM simulation with $f = 0$). If we know *a priori* which node was the first in a simulation, there is only one solution reconstructing duplication events responsible for such a network formation. We can reconstruct all "mother" to "daughter" nodes relationship, however, information on duplications timing is missing. On both schemes black arrows represent duplication events (where an arrow points from "mother" to "daughter" nodes), red arrows point to the first node in a simulation while numbers in parentheses denote the number of duplications each node made in a course of a network growth. One can see that by assigning different first nodes (**a**) and (**b**) we get quite a distinct pattern of duplications, however, overall number of duplications each node made stays almost the same except for those nodes used as a starting ones (numbers coloured in red). This means that to get an information on how many times each node of the SD network duplicated in the course of its evolution - we do not have to reconstruct directionality in the network. It is enough to distinguish edges representing real duplication events (primary edges) from secondary ones.

An example network with one cycle of size 3 can only appear if a node in a cycle inherited one neighbor in the cycle. However, this could happen in several scenarios (Fig. 6.3a) so reconstruction of duplication events from a network with cycles is not a straightforward task anymore and with every new cycle in a network number of alternative solutions grows. So algorithmically this task can be formulated as a search for a spanning tree that goes through all nodes of the SD network and covers only those edges that correspond to duplication events (excluding secondary alignments) (Fig. 6.3b). The spanning tree does not have any cycles thus in the SD network with $N_{SD} = 6656$ nodes, $E_{SD} = 16,042$ edges and $C_{SD} = 1999$ components we have to find a spanning tree with $N_{STree} = N_{SD} = 6656$ nodes,

$E_{STree} = N_{SD} - C_{SD} = 4657$ edges and $C_{STree} = C_{SD} = 1999$ components by resolving cycles. There are several existing algorithms that can find the minimum spanning tree (MST) from a graph that goes through the edges of minimal overall weight. To use these algorithms on our SD network we first have to assign weights to the edges so that the lower the weight of an edge – the higher the probability that this edge corresponds to a duplication event between two respective nodes. Thus resulting MST would mostly cover those edges representing real duplication events as opposed to secondary alignments.

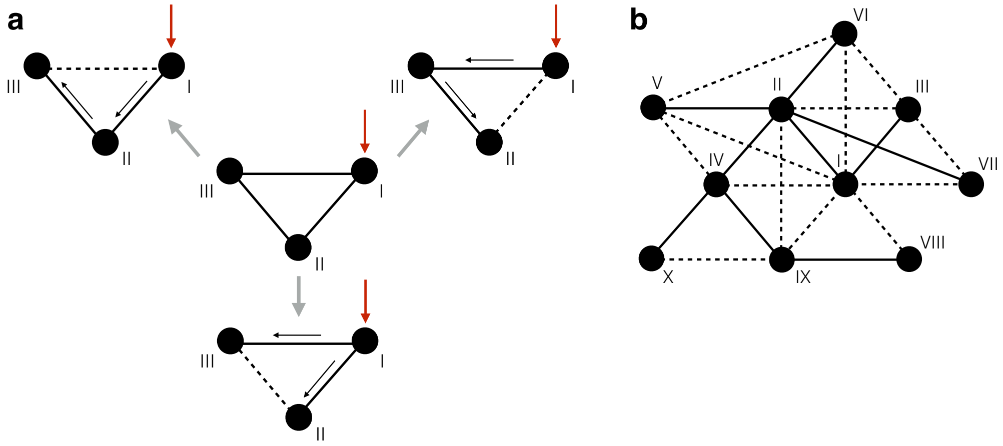


FIGURE 6.3: **Reconstruction of duplication events in graphs with cycles.** The black arrows represent duplication events (pointing from "mother" to "daughter" nodes), red arrows represent the starting node known *a priori* while the dotted lines represent secondary edges. One can see that the cyclic graph in the center of the scheme (**a**) can be reconstructed in three different manners even when the first node stays the same. This results from the fact that cycles can appear in a network if one of its nodes inherits an edge from its mother node. And each of three edges in the cycle can become secondary. The number of possible configurations grows fast with the number of secondary edges (or cycles) in the network of interest. The scheme (**b**) illustrates one possible reconstruction of duplication events in a more complex network. The graph of reconstructed duplications has to be of a tree structure (connected and acyclic).

6.3.2 Edge weight assignment

We used an heuristic approach to assign weights to the edges of the SD network. Each node in the network we independently studied and tried to predict its mother node. Based on the PCM each node of a connected component except for the ancestral pair of nodes appeared as a result of duplication of its mother node. Moreover, the mother node has to be among the neighbors of a node of interest. The mother node is the one that on average shares most of the neighbors with its daughter node.

Indeed, we can divide all potential neighbors of our node of interest D into five categories based on their origin relative to it (Fig. 6.4). The first two categories include those nodes that were already present at the moment of D birth and could potentially become its neighbors at that moment. These include one mother node and its neighbors at the moment of D node birth (nodes M and I, II, III, IV at Fig. 6.4a respectively). The other three categories include those nodes that appeared after the moment of D birth: the progeny of the mother node that appeared after the birth of our node (nodes $M_1^{'}$, $M_2^{'}$), the progeny of the neighbors of our node that appeared after the birth of our node (nodes $I_1^{'}$, $I_2^{'}$, $II_1^{'}$) and all the progeny of our node D itself (nodes $D_1^{'}$, $D_2^{'}$ and $D_3^{'}$). Also, for convenience, several functions were introduced to define sets of nodes: $\mathcal{N}()$ is a set of neighbors of a specific node, $\mathcal{N}_{birth}()$ has the same meaning as $\mathcal{N}()$, but the set includes only those neighbors that were present at the moment of D node birth, $\mathcal{P}()$ is a progeny of a specific node that appeared after the D node birth. Let's give several examples of those functions use:

$$\begin{aligned}
\mathcal{N}(M) &= \{I, II, III, IV, D, M_1^{'}, M_2^{'}\}, \\
\mathcal{N}_{birth}(M) &= \{I, II, III, IV\}, \\
\mathcal{P}(M) &= \{M_1^{'}, M_2^{'}\}, \\
\mathcal{N}(I) &= \{M, D, I_1^{'}, I_2^{'}\}, \\
\mathcal{N}_{birth}(I) &= \{M\}, \\
\mathcal{P}(I) &= \{I_1^{'}, I_2^{'}\}.
\end{aligned}$$

As one will see in further calculations, on average, a mother node of any chosen node shares the highest fraction of its daughter node neighbors. We used this observation in our approach of assigning weights to edges of the SD network. For each edge e_i between our node D and i^{th} neighbor n_i we assign the weight $w_i = 1 - \frac{\|\mathcal{N}(D) \cap \mathcal{N}(n_i)\|}{k_D}$, where k_D is a node degree of D while $\| \cdot \|$ brackets denote the number of elements in a set (Fig. 6.4b). Values of w_i lie in the interval $[0 < w_i \leq 1]$ where $w_i = 1$ when no neighbors of D are shared with n_i, while $w_i > 0$ because at least one neighbor of D (n_i itself) is never among neighbors of n_i.

Now we can estimate the expected number of shared neighbors between n_i and D, i. e. $\|\mathcal{N}(D) \cap \mathcal{N}(n_i)\|$ for neighbors belonging to all five described categories. For simplicity, in the last three categories which include neighbors inherited after the birth of D only "nearest" progeny (daughter nodes) will be included (because "further" progeny shares even less neighbors with D). Let's start by defining the following sums:

$$\mathbf{I} = \sum_{i \in \mathcal{N}_{birth}(M)} 1(i)$$

The indicator function $1()$ here and in all formulas below is defined as: $1(i) = 1$ if i is among neighbors of D and $1(i) = 0$ otherwise.

$$\mathbf{P} = \sum_{i \in \mathcal{P}(I)} 1(i) + \sum_{j \in \mathcal{P}(II)} 1(j) + ... + \sum_{l \in \mathcal{P}(IV)} 1(l)$$

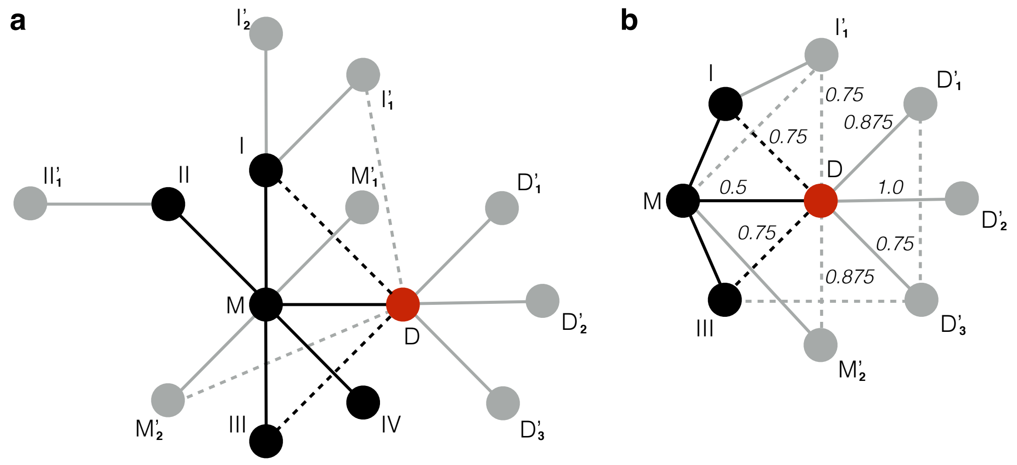


FIGURE 6.4: **All possible node to node relationships in our copying model from specific node's point of view. a**. All nodes in the nearest neighborhood of the node D include: the mother node M, the neighbors of the mother node that were present before the birth of the node D (I, II, III, IV), the daughters of the mother node that appeared after the birth of D (denoted as $M_i^{'}$), the progeny of the nodes $I - IV$ denoted with apostrophes and the daughters of the node D itself (denoted as $D_i^{'}$). The nodes and edges coloured in black were present at the moment of the birth of the node D, while gray ones appeared after. Dotted lines represent those secondary edges inherited by D either at the moment of its birth (black ones) or when some other nodes were duplicated after the birth of D (grey ones). For convenience we include on the first scheme only those secondary edges connected to D while other ones are not shown. The scheme (**b**) includes the node D and its neighborhood with all edges (as we said, some of secondary edges are missing on the previous scheme). The weights of the edges connected to the node D represent weights w_i calculated according to the formula described in the main text. As expected, the lowest weight, which corresponds to the node sharing the highest number of neighbors with D, was assigned to the edge connecting D to its mother node M.

where **P** represents the number of nodes that are connected to D among the after birth progeny of $\mathcal{N}_{birth}(M)$ nodes (i.e. among $I_1^{'}$, $I_2^{'}$, $II_1^{'}$ nodes in our case).

$$\mathbf{M} = \sum_{i\in\mathcal{P}(M)} 1(i)$$

where **M** represents the number of nodes that are connected to D among $M_i^{'}$ (the progeny of M that appeared after the birth of D).

$$\mathbf{D} = \sum_{i\in\mathcal{P}(D)} 1(i) = \|D^{'}\|$$

where **D** represents the number of nodes that are connected to D among $D_i^{'}$ (or equivalently among $\mathcal{P}(D)$). This value just equals to the number of $D^{'}$ nodes.

Then an expected number of shared neighbors $\|\mathcal{N}(D) \cap \mathcal{N}(n_i)\|$ for neighbors

n_i belonging to all five categories of nodes described above can be calculated this way:

$$\begin{aligned}
\|\mathcal{N}(D) \cap \mathcal{N}(M)\| &= \mathbf{I} + f\mathbf{P} + \mathbf{M} + f\mathbf{D} \\
\|\mathcal{N}(D) \cap \mathcal{N}(M_i^{'})\| &= 1 + f\mathbf{I} + f^2\mathbf{P} + f(\mathbf{M} - 1) + \hat{f}\mathbf{D} \\
\|\mathcal{N}(D) \cap \mathcal{N}(l)\| &= 1 + \dot{f}(\mathbf{I} - 1) + \dot{f}f(\mathbf{P} - \sum_j 1(l_j^{'})) + \sum_j 1(l_j^{'}) + f\mathbf{M} + f\mathbf{D} \\
\|\mathcal{N}(D) \cap \mathcal{N}(l_i^{'})\| &= 1 + \dot{f}f(\mathbf{I} - 1) + \dot{f}f^2(\mathbf{P} - \sum_j 1(l_j^{'})) + f\sum_j 1(l_j^{'}) + f^2\mathbf{M} + \hat{f}\mathbf{D} \\
\|\mathcal{N}(D) \cap \mathcal{N}(D_i^{'})\| &= f\mathbf{I} + \hat{f}\mathbf{P} + \hat{f}\mathbf{M} + f\mathbf{D}
\end{aligned}$$

where $l \in \mathcal{N}_{birth}(M)$, $\dot{f}$ – is a probability of an edge being present between two nodes from the $\mathcal{N}_{birth}(M)$ set. It is proportional to the local clustering coefficient. The value $\hat{f} \in [f^2, f]$ and depends on the order of duplication events (for example, whether the node $I_1^{'}$ appeared before or after the $D_1^{'}$). One can check the correctness of these equations by going over all possible pairwise relationships between node types (which is $5 * 5 = 25$ pairs) and calculate the probabilities of sharing neighbors from corresponding node categories. These probabilities would represent the coefficients associated with each term of the above sums. For example, if we consider only $\|\mathcal{N}(D) \cap \mathcal{N}(M)\|$ equation, the mother node M shares all neighbors of D from the set $\mathcal{N}_{birth}(M)$ because all these nodes are connected to M by definition; shares all neighbors of D from $\mathcal{P}(M)$ set because M is their mother node and thus connected to all of them; shares fraction f of $\mathcal{P}(D)$ nodes because these nodes inherit the edge to M each with probability f and, finally, shares the fraction f of neighbors D belonging to the $\{\mathcal{P}(I), \mathcal{P}(II), \mathcal{P}(III), \mathcal{P}(IV)\}$ set, because each node in it inherits the edge to M from its mother node with probability f.

Now one can see that a mother node is expected to share the highest number of neighbors with a daughter node in comparison with other nodes. However, the bias could appear if one of the neighbors n_i is actively duplicated in unbalanced manner after the birth of D. This, in our equations, means a high value of $\sum_{i \in \mathcal{P}(l)} 1(i)$ sum which inflates the $\|\mathcal{N}(D) \cap \mathcal{N}(l)\|$ value where $l \in \mathcal{N}_{birth}(M)$.

So the overall logic is the following: we know (from the PCM) that a mother node, on average, shares the highest number of neighbors of a daughter node in comparison with other nodes in a neighborhood of a daughter node. Thus after assigning weights in the described manner we expect that edges connecting mother and daughter nodes would, on average, have lower weights, thus minimum spanning tree covering all nodes in the graph and going through edges of minimal overall weight would be enriched with primary alignment edges (as opposed to secondary alignment edges) that represent real duplication events. Finally, let's note that an edge weight depends on a node we pick: for an edge e between vertices a and b the weight w_e can be calculated as $w_e = 1 - \frac{\|\mathcal{N}(a) \cap \mathcal{N}(b)\|}{k_a}$ or $w_e = 1 - \frac{\|\mathcal{N}(a) \cap \mathcal{N}(b)\|}{k_b}$ and those values are ordinarily not the same. In practice we assigned the least of two values as a weight of specific edge which equals $w_e = 1 - \frac{\|\mathcal{N}(a) \cap \mathcal{N}(b)\|}{min(k_a, k_b)}$.

6.3.3 Comments on the algorithm choice

In this section we would like to make some comments on how we formulated the algorithm and why our solution was the most reasonable among other potential approaches. Our algorithm is based on an MST search in a network where weights are reversely proportional to the number of shared neighbors between linked vertices. An accuracy of our algorithm was evaluated on synthetic networks generated using the PCM and compared with other networks using other trivial models (use of node degrees, use of random spanning tree) and several centrality measures.

We considered alternative approaches to reconstruct duplication events that take the SD network as an input and/or take advantage of the fact that the network likely grew according to PCM. Formulating the task as a probabilistic model is not practical because one can not estimate the probability of a specific MST without the knowledge about the directionality of edges and the temporal order of all events relative to each other. In theory, this information can be obtained by comparative genomics means, but this task is out of the scope of our analysis and is far from being simple (because of a complex nature of duplication events, assembly gaps etc.). Otherwise, one cannot suggest both the probability of an "observation" and the original probability distribution. Out of the same reason, even less relevant approach would be considering probabilities of a node making n duplications and inheriting l edges from other nodes without any suggested spanning tree of duplications. Getting advantage of the ensemble of all possible spanning trees (or even looking for a sample of a meaningful size) does not seem like a promising solution given enormous number of all possible spanning trees. Can we take a sample of random spanning trees, optimize them in some way and see if they reach some local maxima? Yes, but we did not come up to any alternative optimizational criteria except for the one based on shared neighbors, which in theory leads to the same solution as we see in our MST approach. Overall, because of these limitations we decided to switch from more intuitive "holistic" approaches based on trees, ensembles of trees and probabilities to the ones studying elements of the network (nodes or edges) independently.

As described in the previous section the weights reversely proportional to the number of shared neighbors were assigned to all edges independently and a MST was constructed based on them. In the next section we will compare the performance of our algorithm with other alternatives on PCM synthetic networks. These alternatives include two trivial "background" models: a random spanning tree and a node degree vector as a proxy of the number of duplications. And additionally multiple centrality measures which were formulated in the field of complex networks for diverse tasks. We checked if any of these metrics is more relevant than our method in reconstruction task we formulated.

6.4 Accuracy evaluation

6.4.1 Accuracy evaluation based on PCM simulations

We tested the accuracy of our approach on PCM simulated networks. All PCM simulations were done with the parameters inferred before: $\delta = 5.1 * 10^{-4}$, $f = 0.47$, and resulting networks were of a size of the giant component of the SD network ($N = 1325$ nodes). To do this we simulated the PCM based network growth and kept information on edges status (primary edges representing duplication events or secondary ones). Then we assigned weights to edges of synthetic networks in the manner discussed above and run Kruskal's algorithm to reconstruct the minimum spanning tree on the weighted network (Kruskal, 1956). Two metrics were used to measure the accuracy of our predictions: the fraction of correct (primary) edges in the MST and a variance explained in duplications number that each node underwent during the simulation (Table 6.2). We compared our method with several alternative simple strategies of predicting number of duplication events per node (listed in the Table 6.2). These include multiple types of node centrality measures. In the most trivial model a node degree of a node in a PCM synthetic network was used as a predicted number of duplications that the node did during the network growth. This model was used as a baseline to access an accuracy of our method (to check if it gives better predictions than the trivial model). One can see that our heuristic approach performs the best among other methods and explains 77% of variance in number of duplications while the trivial model reaches only 58%. This proves an accuracy of our method, which was then applied on the real SD network. From the resulting MST that covers putative duplication events in the SD network we calculated the number of duplications that happened in each node. Since edges of a node in MST represent duplication events (or primary alignments) we can say that almost every node i (except for the first one) in the network was duplicated $k_i - 1$ times where k_i is a node degree of i^{th} node in the MST. The first or "oldest" node is the one that was duplicated k_0 times (based on PCM), where k_0 is a node degree of the first node. Predicting a first node in each component of the SD network is not a straightforward task (and likely biologically irrelevant because of complex nature of some duplication events) we can approximate a number of duplication events as a node degree $k_i - 1$ for all nodes of the SD network. This way we generated an integer vector $\overrightarrow{D}_{sd}$, where each element $\overrightarrow{D}_{sd}[i]$ represents a number of duplication events that happened in i^{th} node (or node degree of i^{th} node in MST -1).

6.4.2 Additional validation of predicted MST

In a previous section we validated our method based on PCM simulations where we knew a correct order of duplication events. In this section we wanted to use some indirect evidence to check if our MST predicted for the SD network gives a reasonable reconstruction of real duplication events. To do this we first applied the

Method	Edges match (%)	Variance explained (R^2)
Node degrees vector	NA	0.58
Random spanning tree	20	0.44
Kruskal's MST	64	0.77
Betweenness (edges)	17	0.64
Betweenness (nodes)	NA	0.66
Closeness centr.	NA	0.17
Eigenvector centr.	NA	0.48
PageRank centr.	NA	0.67
Radiality centr.	NA	0.02

TABLE 6.2: **Comparison of several methods for predicting the number of duplications that each node made during PCM simulations.** To evaluate an accuracy we measured the fraction of correctly predicted primary edges (where possible) and a variance explained for the vector of real duplications number (R^2). We measured an accuracy of our MST based algorithm described in the main text ("Kurskal's MST" row in the table). We used multiple centrality measures for nodes in resulting PCM synthetic networks as listed; one can find more information about these measures, for example, in description of centralities in LightGraphs module documentation (Seth Bromberger and contributors, 2017; Brin and Page, 1998). In case of edges betweenness we, similarly to our algorithm, constructed the MST based on betweenness values of edges as weights. Finally, we considered the vector of node degrees and a random spanning tree covering nodes of a synthetic network as a trivial models to estimate a baseline quality of predictions. One can see that our algorithm performs the best in both criteria among all alternatives with, per average, 64% of primary edges correctly predicted and 77% of variance in number of duplications explained.

described algorithm to the SD network and checked if the resulting MST is enriched with real duplication events based on some features of edges in it.

We can expect that if two or more alignments have their breakpoint coordinates matching, it is quite unlikely that this match happened as a random coincidence. This could happen because of some features of the sequence around the breakpoint (mechanic instability, NAHR hotspot etc.) or simply because this alignment is secondary (appears as a result of overlap of duplication events). One can find an illustration of how this matching breakpoints appear at the Fig. 6.5a. It means that an alignment nested inside of another longer alignment with one of the breakpoints matching between them is suspicious for being secondary (thus representing a secondary edge). Not necessarily all secondary alignments are satisfying this criterion (the number of suspicious ones is much smaller than the number of secondary edges in the SD network), but the alignment that satisfies it is likely secondary. In practice, we considered inexact matches where the distance between breakpoints is less than 5 bps. The longer alignment in a pair or a pair of unnested alignments with matching borders are on the other hand not suspicious. So we collected all "suspicious" edges that correspond to alignments satisfying the described criterion and check if those edges are depleted in the MST predicted for the SD network. These "suspicious" edges with matching breakpoints were depleted among edges of our predicted MST in comparison with 1000 random samples of edges taken from the rest of the SD network (empirical p-value < 0.001).

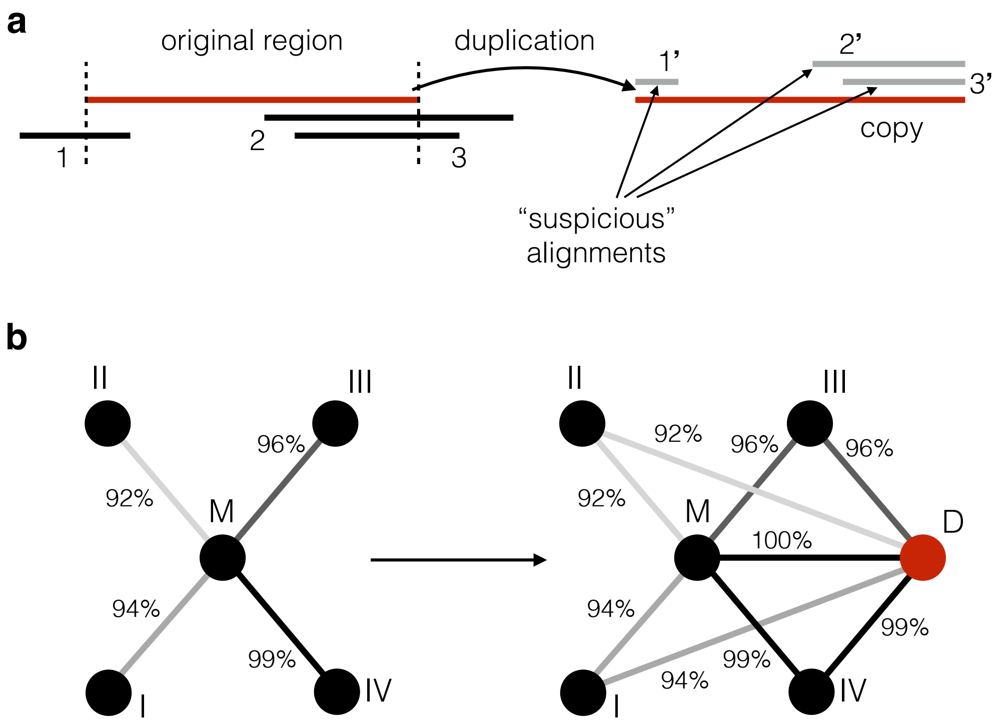


FIGURE 6.5: **The criteria we used to validated our predicted MST of duplications.** We suggested that the correct MST is depleted with suspicious alignments (nested ones with matching breakpoints) and the scheme (**a**) illustrates this. The duplication event overlaps the alignments marked by 1, 2 and 3 and copies them incompletely (the loci to which 1, 2 and 3 align are not present on the scheme). Thus the resulting copy aligns to the "mother" locus plus three additional loci (the alignments are marked as 1′, 2′ and 3′ respectively) and since the original alignments were copied in abrupted manner, the alignments 1′, 2′ and 3′ are suspicious according to our criterion. **b**. We suggest that at the moment of duplication two copies of a genomic sequence are almost identical (at least in a simple scenario of copy-paste process). Thus a "daughter" node inherits alignments (or secondary edges) with a sequence identity observed for corresponding alignments (edges) of a "mother" node which is illustrated on the scheme. The sequence identity levels of corresponding alignments are assigned to the edges on the scheme and shown by the colour.

Secondly, we expect that a correct MST covering real duplication events includes more edges of higher sequence identity than a random set of edges from the SD network. This might sound counter-intuitive, but edges with higher sequence identity are expected to be enriched in the correct MST of duplications in comparison with other secondary edges. In a simple scenario, we can assume that a level of sequence identity between two copies is the highest right after the moment of duplication and it decreases in a process of accumulation of neutral mutations with time. Secondly, when a duplicated region is duplicated again we expect that corresponding edge between "mother" and "daughter" node is of high sequence identity shortly after the duplication. On the other hand, edges inherited by a "daughter" node are of the same level of sequence identity as we observe for "mother" node (Fig. 6.5b). This can be illustrate with an example where a set of words represent duplicated regions while hamming distances are measures of sequence identity between them. If we make a copy of one of the words it will inherit all hamming distances to other words observed for the original word while the hamming distance between the original word and the new one will be equal to zero. Same behavior is expected for the sequence identity levels between duplicated regions. Thus at the moment when secondary edges appear, the level of their sequence identity is the same as the one observed between the "mother" node and the corresponding neighbor (they inherit corresponding level of identity), while primary edges are of high sequence identity at the moment of their formation (identical or close to it). This levels of sequence identity drop in time in a process of mutation accumulation, however, we expect that highly identical alignments more likely correspond to real duplication events than those of lower sequence identity. We found that in agreement with our suggestions the predicted MST for the SD network is enriched with highly identical edges (alignments). There are more alignments of sequence identity higher than 0.99 in the predicted MST in comparison with random samples of edges from the rest of the SD network (empirical p-value = 0.001 based on 1000 permutation tests).

In this subsection we studied some features of alignments in order to make some conclusions about the accuracy of reconstructed duplication events. Even though we can not make quantitative evaluation of the accuracy - we can see that for both criteria our predicted MST "behaves" as expected for the spanning tree enriched with primary edges (real duplications). This and the tests made on PCM synthetic networks (see previous subsection 6.4.1) give us a reason to believe in our MST of duplication events reconstructed for the SD network and allow us to move further into studying biological features of the SD propagation process.

6.5 Associations with genomic features

6.5.1 Observed associations

In this section we will discuss the current knowledge about genomic features associated with SDs, i.e. those genome properties that are often observed in SD sites or their flanking regions. We start with the fact that we already discussed: segmental duplications tend to lie in subtelomeric and pericentromeric parts of human chromosomes. A slight positive correlation of segmental duplications distribution with the gene density was reported by Zhang et al. (2005) along with negative correlation with recombination rates. On the other hand, when considering duplicons which constitute duplicated regions, core duplicons are enriched, while non-core duplicons are depleted with exonic sequences in comparison with the rest of the genome (see the "A-Bruijn graphs and core duplicons" section (2.3.2) or Jiang et al. (2007)). Moreover, large duplications tend to be enriched in heterochromatic parts of the genome or, according to other observations, in hetero- to eurochromatin transition regions (Grunau et al., 2006; Kirsch et al., 2008). Duplicated regions are, on average, of higher G/C content in comparison with the rest of the genome (Bailey, Liu, and Eichler, 2003; Zhang et al., 2005). In agreement with that, CNVs breakpoints seem to be enriched with G/C-rich sequences predicted to form G-quadruplexes (Bose et al., 2014). The listed factors are observed irregularly and not necessarily reproduced in other experimental settings. It was even observed that those associations could be specific to some chromosomes, but absent in others (Zhang et al., 2005).

High-copy repeats can cause segmental duplications through homology-mediated or other mechanisms. Thus it is not surprising that some repeat classes are enriched at breakpoints of SDs. Some cases of duplication causing repeats were reported even before SD annotation in the human genome (Eichler et al., 1996; Eichler, Archidiacono, and Rocchi, 1999; Guy et al., 2000). The systematic study by Bailey, Liu, and Eichler (2003) identified those repeat classes that are significantly overrepresented at SDs breakpoints when compared with the genome average. Only those segmental duplications not overlapping other SDs and accurately annotated were considered. As a result, two repeat classes are significantly overrepresented at breakpoints: *Alu* repeats and satellites (specifically, HSATII, GSAT, and TAR1), while many repeat classes are even underrepresented in flanking regions (Fig. 6.6). Among *Alu* subfamilies, younger ones (*AluY* and *AluS*) account for the enrichment, while the oldest primate subfamily (*AluJ*) does not (Bailey, Liu, and Eichler, 2003). Similarly, the flanking regions of the core duplicon (LCR16a) were studied by Cantsilieris et al. (2020). The chromosome 16 is especially enriched with interspersed SDs, while most of the chromosome 16 short arm duplications are associated with the 20-kbp core duplicon LCR16a. The sequence composition of the LCR16a flanks was studied in human and primate species. The flanks of the core duplicon are enriched with *Alu* repeats and are of higher G/C content (Cantsilieris et al., 2020).

Replication timing is another factor associated with duplications. For example, early replicating genomic regions are more gene rich, genes are more transcriptionally active, the G/C content is higher in early replicating regions etc. CNVs are also associated with replication timing, but the dependence seems to be complex. For example, it was observed that recurrent CNVs are enriched in early replicating genomic parts, whereas non-recurrent CNVs are more frequent in the late ones (Koren et al., 2012; Chen et al., 2015). The same question was addressed from a different angle when induced pluripotent stem cells (iPSC) were compared with their parent fibroblasts with respect to replication timing and CNVs accumulation. Specifically, it was found that CNVs gains are preferentially located in the genomic regions that became early replicating during pluripotent cell transition (Lu et al., 2014). Another evidence from the field of cancer genomics suggests a different pattern of CNVs accumulation. Often during carcinogenesis extensive changes in replication timing

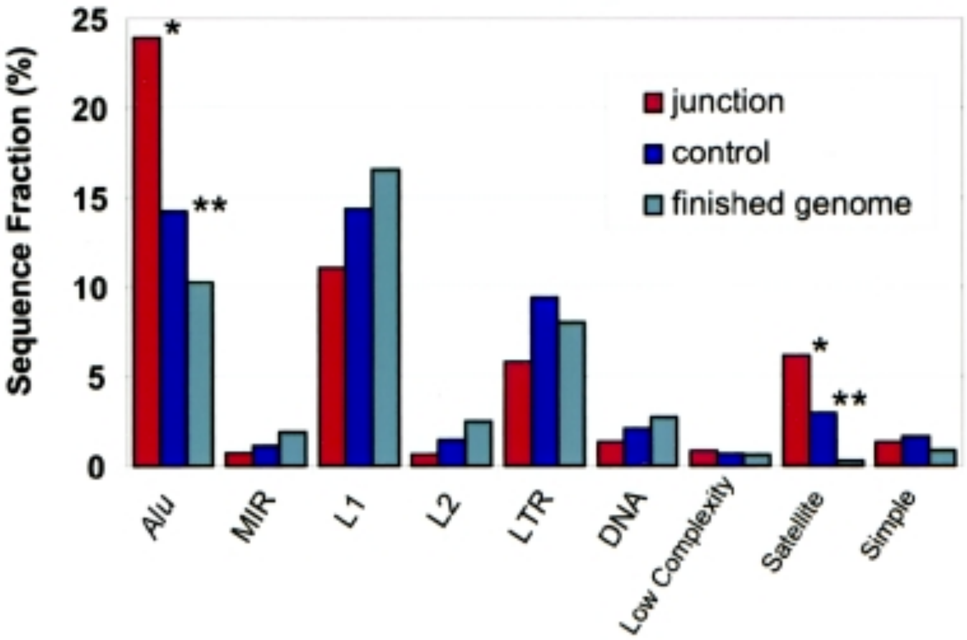


FIGURE 6.6: **Repeats at breakpoints of SDs.** The barchart represents frequencies of various repeat families in breakpoint regions ("junction" in the plot legend), inside of SDs and in their flanking regions ("control" in the plot legend) and in the rest of the genome ("finished genome"). Single and double asterisks highlight those distributions where significant differences: junctions vs control and control vs finished genome, respectively, are observed. One can see a significant enrichment of *Alu* and satellite repeats at SDs breakpoints. The figure source: Bailey, Liu, and Eichler (2003).

happen along with accumulation of somatic CNVs. An analysis of more than 330,000 somatic copy number alterations showed that these events are more frequent in late replicating regions in tumor cells (De and Michor, 2011). Finally, it was suggested that division into early and late replicating regions is not sufficient to explain CNVs distribution. The hotspots of CNVs are strongly associated with sites of reduced DNA polymerase velocity. These can be detected as ones where the difference in replication timing (or replication timing derivative) is the largest. In other words, a genomic site located between early and late replicating regions is the one where DNA polymerase progression slows down. This, in theory, leads to increased probability of replication template switching. Association of such genomic sites with CNVs was observed by Chen et al. (2015). After all we can say that replication timing could play a role in SD formation. It is not something unexpected given the fact that many segmental duplication mechanisms are associated with DNA replication. However, effect of replication timing on CNVs is not completely clear and likely complex.

6.5.2 Genomic features as predictor variables

For each duplicated region (node) of the SD network we collected multiple genomic features associated with that region. These genomic features include: replication timing, recombination rates, openness of chromatin, genome assembly gaps, CTCF sites, coordinates of a duplicated region, G/C nucleotides content, number of gene overlaps, repeat overlaps and CpG island overlaps (see Table 6.3). The fraction of intrachromosomal edges from all edges of a corresponding node is the only feature in the analysis that originates from the SD network description while all other genomic features were extracted from the UCSC genome browser (`https://genome.ucsc.edu`). UCSC LiftOver tool was used to transfer coordinates from *hg37* to *hg38* where needed (Kent et al., 2002). The feature values were measured either inside of a duplicated region (between its breakpoints) or in flanking regions of length 50 bps padding the duplicated regions on both sides ("Position" column at Table 6.3). The feature types include: counts (repeats, assembly gaps etc.), mean values in corresponding genomic intervals (replication timing, recombination rate etc.) and fractions (fraction of G/C nucleotides, intrachromosomal edges from all neighbors of a node). Other than that, a span of replication timing, i.e. the difference between maximal and minimal timing values in a genomic interval, was measured and used as a proxy for the replication pausing ("Replication pausing" at Table 6.3). For those features where flanking regions are studied we did not distinguish between flanks and used the sum (or mean) of two values. All genomic features described so far were used to find associations with high duplication rates. In other words, we formulated a machine-learning task of predicting the number of duplications (response variable) given the matrix of genomic features associated with duplicated regions (predictor variables). By assessing how important each genomic feature is in predicting the number of duplications we can say which features are associated with high duplication rates in the genome.

6.5.3 Genomic features associated with duplicated regions

First of all, we studied if various characteristics of duplicated regions differ from those observed in random genomic sites. We compared duplicated regions against random genomic sequences without, as for now, taking into account the number of duplications that happened in each region. In this chapter we do not distinguish actively and rarely duplication regions, however, the next section is dedicated to that task. We randomly shuffled duplicated regions positions in the human genome and measured the genomic features we discussed in the previous section. We did multiple rounds of shuffling to get a background distribution. Mean values observed for the original duplicated regions were compared to the background distributions and for most features these values were either significantly larger or smaller than expected from the null hypothesis. This is quite predictable given the fact that duplications are distributed very non-uniformly in the human genome and some of

Genomic feature	Type	Position	Cell line
Coordinates and length	value	—	—
Telo-/centromere dist.	value	—	—
Intrachrom. edges (%)	fraction	—	—
Replication timing	mean	inner + flanks	GM12878
Replication pausing	span	inner + flanks	GM12878
Recombination rate	mean	inner + flanks	—
DNAse hypersens.	mean	inner + flanks	master (125 cell types)
Assembly gaps	count	flanks	—
CTCF sites	count	flanks	GM12878
GC nucleotides (%)	fraction	inner + flanks	—
Genes	count	inner	—
CpG islands	count	inner + flanks	—
DNA transposons	count	flanks	—
LTR retrotransposons	count	flanks	—
LINEs	count	flanks	—
SINEs	count	flanks	—
Satellite repeats	count	flanks	—

TABLE 6.3: **Technical information on how each listed genomic feature was processed.** The columns "Type" and "Position" specify how and where a feature was measured (either inside of a duplicated region or in two flanking regions of 50 bps). Some characteristics were estimated both inside and in flanks of duplicated regions. Those were added into the analysis as two separate columns. Some of the features are cell line specific, however, the cell lines where segmental duplications happened belong to the germline. In absence of a relevant data from the germline, we picked the source of the data as listed in the "Cell line" column. We used either a data from the GM12878 lymphoblastoid cell line (picked arbitrary) or the master track for DNAse hypersensetivity. The master track represents an integrated DNAse hypersensetivity data for 125 separate cell lines (one can read more on that at http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeAwgDnaseMasterSites and Thurman et al. (2012)).

the associations we see were reported earlier. For example, assembly gaps are enriched at flanking regions (because large complex duplications are hard to properly map), duplicated regions are located in late replicating regions or/and in those where DNA polymerase slows down, recombination rates are lower in duplicated regions, while the G/C content is higher (Fig. 6.7).

Next we wanted to study in detail how repeats are distributed relative to duplicated regions breakpoints (inside, outside or at borders of duplicated regions). To do this we used sliding non-overlapping windows of length 50 bps that started at the breakpoint of a duplicated region (coordinates from a border: [0; 49]) and moved 30 steps away ([0; 49], [50; 99], . . . , [1450; 1499], respectively) and 5 steps inside ([−250; −201], [−200; −151], . . . , [−50; −1]). These sliding windows moved symmetrically (as a pair) relative to breakpoints on both sides of a duplicated region and counted the number of repeat overlaps in each pair of windows. Moreover, to study

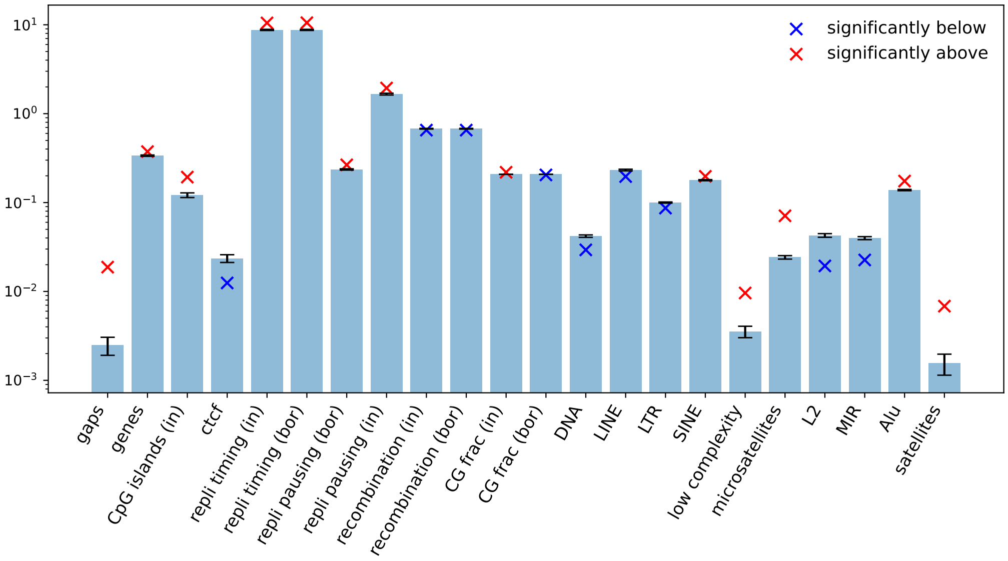


FIGURE 6.7: **Characteristics of duplicated regions in comparison with the genome average.** The barchart represents distributions of various genomic features (listed on the X axis) measured in randomly shuffled genomic intervals. The crosses are values that we observe for original duplicated regions (red ones are above and blue ones are below the background mean). Only those features where the difference is significant are listed out of the whole list of features (Table 6.3).

the background distribution the sliding windows coordinates were randomly shuffled throughout the human genome 10 more times. As a result the distributions were different for different repeat classes. These could be divided into two groups. Repeats in the first group include: DNA transposons, LTRs, L1 and L2 repeats from the LINE family and MIR repeats from SINE. These are depleted inside of a duplicated region and their numbers grow when we move further outside from a duplicated region borders (Fig. 6.8). Repeats in the second group are enriched in a close proximity outside of duplicated regions, especially at the very breakpoints. This group includes: satellites, microsatellites, *Alu* and low complexity repeats (illustrated at Fig. 6.8 subplots, starting from the satellites on). The low complexity repeats are composed of polypurine or polypyrimidine repeated stretches, or regions of high AT or GC content. The second group of repeats is likely responsible for genomic instability that leads to duplication events and is unlikely attributed to a non-uniform SD distribution in the genome (concentration of SDs in subtelomeric or subcentromeric regions, proximity to assembly gaps etc.). Firstly, because repeats from the second group are depleted inside of duplicated regions and at a relatively small distance from them and, secondly, because exclusion of duplicated regions flanked by the assembly gaps did not change the observed distributions. The differences in repeats frequency between proximal flanking windows ([0; 49]) and shuffled windows are significant for all repeat families, except for L1 (Fig. 6.7).

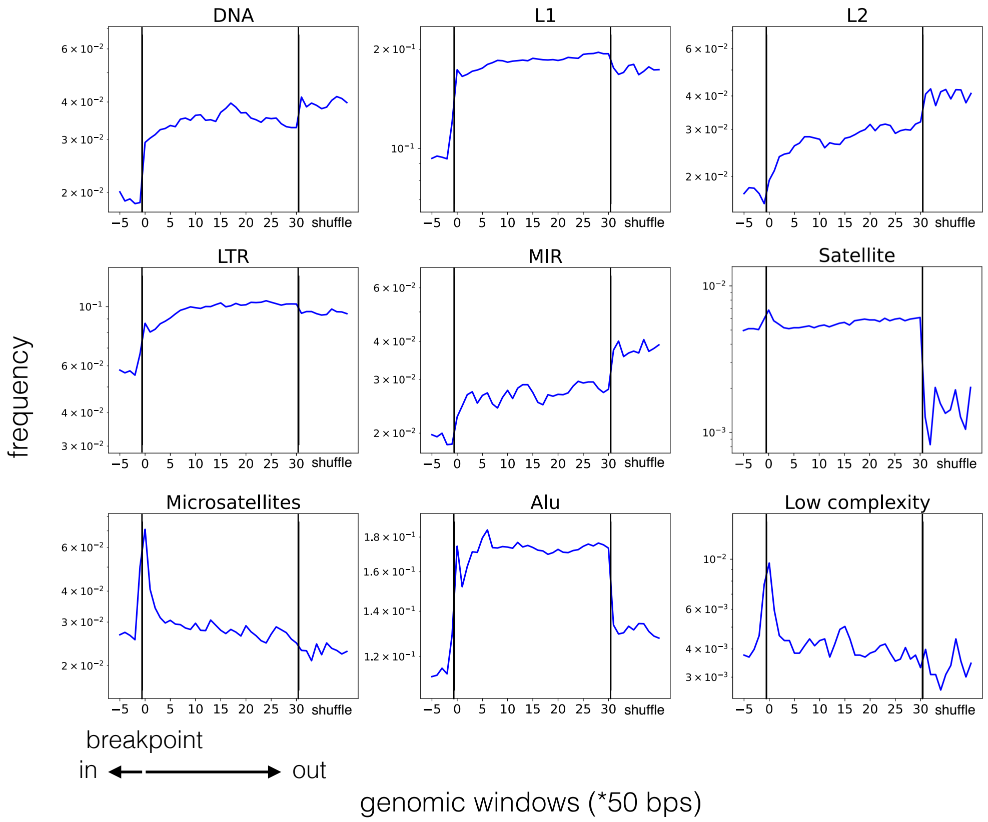


FIGURE 6.8: **The distribution of different repeat families relative to the breakpoints of duplicated regions**. Technically, we took two genomic windows of 50 bps upstream and downstream of all duplicated regions, calculated the number of repeats falling into those windows and divided it by the overall number of those windows (the frequency values on the Y axes). We did 5 measurement steps inside of duplicated regions, 30 measurements outside and 10 times in random genomic positions. The X axes represent the genomic positions relative to breakpoints (marked with vertical line); last 10 points separated by another vertical line correspond to counts in randomly thrown windows. The first group of repeats (from DNA transposons to MIR repeats) includes those depleted inside of duplicated regions with no enrichment observed around the breakpoints in comparison with the rest of the genome. On the other hand, the second group (from satellites to low complexity repeats) includes repeats that are enriched at the very breakpoints of duplicated regions or in nearby windows (satellites, microsatellites, *Alu* and low complexity repeats).

6.5.4 Prediction of genomic features associated with increased duplication rates

We applied several machine learning algorithms to estimate the accuracy of predictions as a percentage of response variable variance explained. Various genomic features were predictor variables, while number of duplications that happened in a region - response variable. The applied algorithms include: the linear regression, the support vector regression (SVR), decision trees and the random forest reached a similar quality of predictions with the maximal % of variance explained $R^2 = 30.5\%$

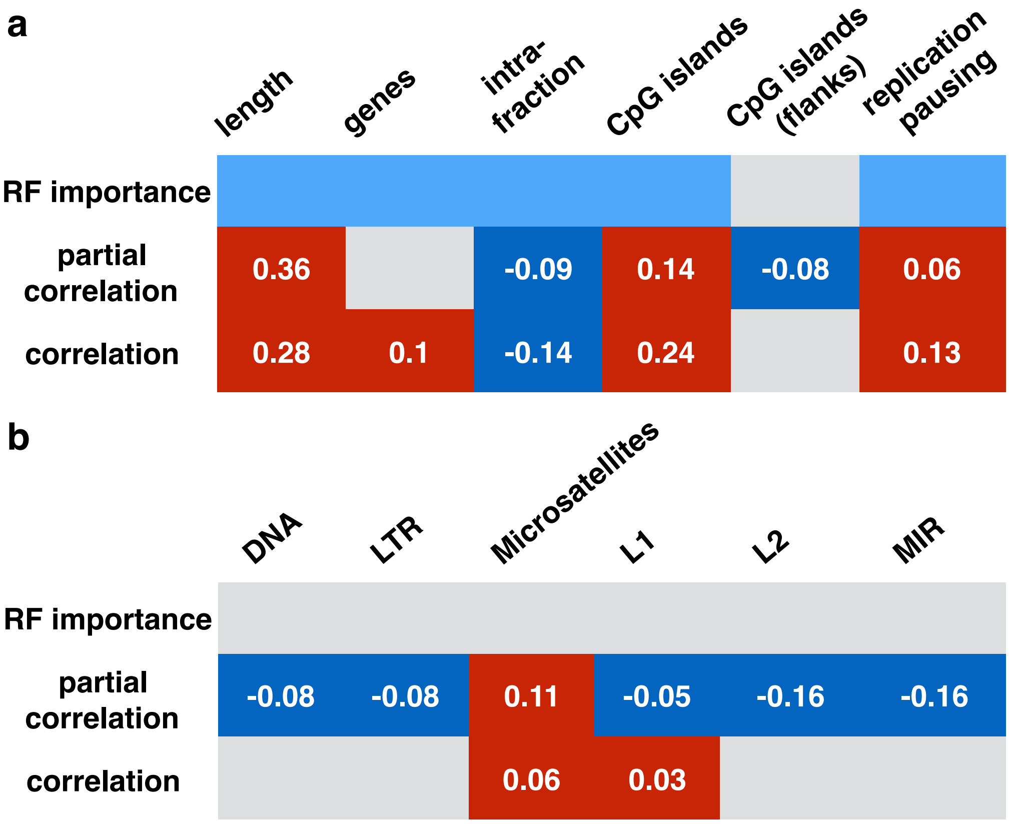
observed for the random forest (with 5-fold cross-validation). Our main interest was to find genomic features that are important in the prediction and thus associated with number of duplications that happen in duplicated region. So we first estimated feature importances that are assigned to predictor variables by the random forest algorithm. Then we estimated the statistical significance of those importance values in a special permutation manner described by Altmann et al. (2010). Multiple rounds of permutation of a response variable with consequent runs of random forest algorithm allows to approximate a null distribution of importance values for each feature when no interaction between a predictor and response variables exist. The resulting empirical p-value (the fraction of importance values in permutations that are larger than observed one) allows overcoming the biases characteristic for a regular importance value. As a result, the following genomic features were important in number of duplications prediction: length of a duplicated region (emp. p-value < 0.001), fraction of intrachromosomal edges (emp. p-value < 0.001), number of overlapping genes (emp. p-value $= 0.023$) and CpG islands (emp. p-value $= 0.0025$), replication pausing (emp. p-value $= 0.025$) (the last two measured inside of a duplicated region). Significant importance of those features cannot be attributed to the fact that some of those features are correlated with each other. To clarify the type of dependence between the features and response variable we calculated partial correlation coefficients (Spearman's) for each predictor variable (genomic feature) with the response variable (number of duplications or jumps) controlling for all other predictors as possible confounding variables (Fig. 6.9). Additionally, we measured the regular Spearman's rank correlation coefficient and estimated its significance by permutations. One can interpret these coefficients in the following way: the regular correlation coefficient represents the dependence we would observe without considering other correlated features. However it could be a result of other confounding correlations that are taken into account when the partial correlation is calculated. These two values could be dramatically different (even of an opposite sign), because the partial correlation coefficient shows the affect of one variable to another independent of other features (i.e. specific association). For example, if features are positively correlated, but are of negative partial correlation, it can be interpreted as the features are associated in the nature, but in the same conditions the presence of one decreases the probability to observe another.

The partial and regular correlation coefficients are significant and positive for the length of a duplicated region, number of CpG islands (inside) and replication pausing (inside), while negative for intrachromosomal edges fraction. Other correlation coefficients were not significantly different from zero according to at least one of the tests. These results agree with our observations. The length of a duplicated region is positively correlated with the node degree of a region (see Fig. 4.10 above) and thus with the number of duplications that a region underwent. As we discussed earlier, the loci with reduced DNA polymerase velocity (the "Replication pausing" track in our analysis) could be associated with genomic duplications. Our results supported this hypothesis: the association of replication pausing and duplication frequency is reproduced by several our approaches (Fig. 6.9a). Association with the number of gene overlaps also makes sense given the fact that segmental duplications were responsible for gene families propagation and evolution in the human lineage (see the "Evolutionary role of SDs: gene duplications" section 2.3.4). To our knowledge, only increased G/C content was observed in segmental duplications, however, enrichment of CpG islands in highly duplicated regions is not described yet.

Several high-copy repeat families had significant non-zero (mostly negative) partial correlation coefficients with the number of duplications that duplicated regions underwent (Fig. 6.9b). However, most of these associations were not supported by significant Spearman's correlation coefficients or the random forest importance estimates. Only microsatellites are positively correlated according to both correlation coefficients. However, it seems unlikely that any of considered repeat families are strongly associated with duplication rates of genomic loci. Overall, we predicted the genomic features that are associated with highly duplicating regions in the human genome and reasoned the type of association.

6.6 Summary

In this chapter we reconstructed duplication events from segmental duplications alignments. We did this based on the SD network without any additional sources of information. Earlier we suggested the preferential copying model for the SD network growth. This allowed us to come up with criteria for primary edges corresponding to real duplication events: if a node n_i shares a big fraction of neighbors with its neighbor n_j, it is likely that n_j is a mother node of n_i. We used this principle to assign weights to edges (those mother to daughter edges are of lower weight in comparison with secondary ones). We then constructed a minimum spanning tree (MST) that covers all nodes and goes through edges of minimal overall weight. We validated our approach on PCM simulations and other indirect evidence tests, such as distribution of sharp border alignments and high sequence identity ones. Our method of reconstruction turned out to be more accurate than its alternatives. An accurate MST allows to answer, how many times each node (or a duplicated region) was duplicated in the course of evolution. We used this knowledge to find those genomic features associated with actively duplicating loci. Moreover, we studied characteristics of duplicated regions and their flanks in comparison with the genome average. We found multiple skews associated with duplicated regions and their breakpoints, non-uniform repeats distributions and some associations with highly duplicating loci. These results in detail are described in this chapter.



	length	genes	intra-fraction	CpG islands	CpG islands (flanks)	replication pausing
RF importance	🟦	🟦	🟦	🟦		🟦
partial correlation	0.36		-0.09	0.14	-0.08	0.06
correlation	0.28	0.1	-0.14	0.24		0.13

	DNA	LTR	Microsatellites	L1	L2	MIR
RF importance						
partial correlation	-0.08	-0.08	0.11	-0.05	-0.16	-0.16
correlation			0.06	0.03		

FIGURE 6.9: **Different characteristics of duplicated regions significantly associated with their number of duplications.** Only significant associations (after Bonferroni correction) predicted by the random forest, the partial correlations or both are listed. **a**. These include: length of a duplicated region, fraction of intrachromosomal edges from all edges of a node, number of CpG islands overlapping a duplicated region or its flanks, replication pausing and number of overlapping genes. The cells coloured in light blue represent those features with high permutation importance in the random forest algorithm. Dark blue and red cells correspond to features with significant negative or positive correlation coefficients (specified in the cells) while unfilled cells represent non-significant ones. **b**. High copy repeats that showed significant partial correlation with the number of duplications. Most of them are not supported by the random forest or significant non-zero Spearman's correlation coefficient.

Chapter 7

Discussion

Segmental duplications (SDs) are long ($>$ 1 kbp) duplications of genomic sequence that have high sequence identity ($>$ 90%) and are fixed in a genome. Segmental duplications play an important role in evolution by increasing the probability of genomic rearrangements, creating new gene paralogs, affecting the speciation and so on. In this thesis we study SD evolution with the main focus on duplications in the human genome. Specifically, we address the dynamic properties of duplication process as our main topic of interest.

We studied the dynamics of a globally acting propagation process for segmental duplications in the human genome. To do this a mathematical formalization in terms of networks and network growth processes was applied. The SD network was generated from annotated SDs. In this approach network nodes represent genomic regions involved in duplications and edges indicate the presence of an alignment between two regions. This gave us the opportunity to investigate several network growth models and reason about their relevance in describing the nature of SD evolution. The simplest copying model with equal probabilities of node duplications (UCM) is not sufficient to explain the SD network topology. However, a more complicated preferentially copying model (PCM) with preferential node duplication rates nicely fits all topological characteristics of the SD network, especially if taking into account that the growth model includes only 2 parameters (f and δ). Based on the PCM the duplication rate of already duplicated regions grows linearly with the number of copies of those regions (more precisely, with the number of loci that share long homologous sequences with those regions).

The PCM was accurate in predicting the SD network characteristics even without inclusion of additional processes that reflect real life events, such as: deletions of duplicated regions, decrease of homology below the detection threshold in time, duplication where a new copy jumps into an already duplicated region etc. Moreover, we consciously did not include separate processes that correspond to different duplication mechanisms as they are described in the literature (duplication dynamics is different in pericentromeric, subtelomeric and other genomic regions) or different processes corresponding to intra- and interchromosomal duplications (see the "Genomics Background" chapter 2). The explicit addition of such processes to the model would make the parameterization heavier while biological conclusions more vague.

Models with multiple parameters can overfit the data even if irrelevant dynamics is suggested. The PCM, on the other hand, has little number of parameters thus the good fit that we observed, likely, results from its relevance to the SD network growth and not overfitting. However, the two-step model explaining the mosaic structure of pericentromeric SDs assumes that genomic segments are first translocated into one genomic locus and this process is not included in the PCM. As we observed, the exclusion of pericentromeric SDs from the analysis does not change the SD network topology substantially thus keeping our global predictions valid. Moreover, the topology of the SD network does not change significantly if we try alternative strategies of the SD network construction or add new processes to network growth models (see the "Stability of our predictions" section 4.8). This means that our prediction about preferential nature of duplication rates is stably reproduced and is not an artifact of the SD network construction strategy.

We suggested some interpretations of why the network evolution of the SD network follows a model with preferential node duplications. A mechanistic explanation could be that with growing node degree the length of the corresponding duplicated region also grows, thus the probability that the next duplication will overlap that duplicated region also grows. Secondly, growing node degree of a duplicated region is also associated with growing probability of genomic rearrangements (including duplications) in that locus. Finally, we observed that the frequency of copy-number variations in the human population grows with the node degree of a duplicated region it overlaps with. This might be attributed to different scenarios: overall genomic instability of a duplicated region that has multiple copies, recurrent duplications happening in unstable genomic sites, decreased purifying selection against new duplications in those regions or positive selection for beneficial gene duplications and decreased recombination rates that reduce an efficiency of the purifying selection. In all such cases, CNVs in high node degree duplicated regions are more likely to be fixed in human population. This might explain the preferential duplication rates in the PCM as well.

One more observation that comes from the PCM is that the number of nodes in the network at some point starts to grow hyperbolically and nodes accumulate almost exclusively in the giant component. This means that in the PCM we do not have an equilibrium steady state. Thus the overall length of duplicated regions in the genome should also reach a hyperbolic growth at some point leading to an "SD explosion". If this prediction is correct it is curious how this problem is addressed in natural evolution of genomes without any notable signs of this effect in the topology of the SD network. One possible scenario is that selection might act only on high node degree duplicated regions (very right tail of the node degree distribution) by decreasing the probability for further duplications. However, a constant rate of nodes/edges loss would slow down but not prevent the hyperbolic growth. To stop this growth more complex scenarios have to be considered, for instance, a time dependent rate of duplications or loss, when periods of increased relative loss rate compensate nodes accumulation. In principle, this agrees with estimates of duplication/deletion rates in primate phylogenetic lineages. Those rates changed quite substantially in the human ancestor lineage from the split of New and Old world monkeys to modern human (see the "Reconstruction of lineage-specific duplication events" section 5.2).

The network formalization was also applied to study SDs in other (non-human) species genomes. We predicted SDs in genomes of 9 phylogenetic distinct species which include well-studied ape genomes, genomes of domestic animals and model organisms. We decided not to use known SD annotations because these were often assigned to outdated reference genome versions and were constructed a bit differently, moreover, not all annotations that we needed were open-access or existed. For all latest reference genomes (including the human one) we *de novo* annotated SDs. Having the same experimental setting was more important for us than the accuracy of SD prediction which is likely higher in specialized studies dedicated to SD annotation in a specific genome. The reason is that we only considered topological characteristics of SD networks which have to be constructed in the same manner, otherwise biases resulting from different approaches could make them incomparable.

We found that even though sizes of SD networks were quite different, there was a clear similarity in the SD networks topologies. This can be observed in the same power-law slope in all connected component size distributions, presence of giant components in almost all cases, $E \propto N^{\alpha_s}$ dependence for components $C(N, E)$ in the SD networks etc. (α_s values are specie-specific). This similarity is unlikely to result from SDs shared by the species. A big fraction of predicted SDs consists of duplications that originated after the New and Old world monkeys split as we mentioned earlier. Thus we expect that distinct vertebrate species (like human and chicken) are far from sharing enough duplicated regions to explain similarities in networks characteristics. It seems like all considered vertebrate species follow similar evolutionary dynamics of duplications or the "rules" according to which SDs propagate in the genomes. More than that, we suggest that SDs in vertebrate genomes likely propagated according to the PCM (where duplication rate of a genomic region grows linearly with its number of copies). This can be seen from the characteristic topology of SD networks which, at least to our knowledge, was not met in other complex networks in the field (like internet, citation networks etc.) and can be reproduced by PCM simulations. The described topology is not recognized in the SD network constructed based on C. elegans duplications. This means either different mechanism of SD propagation in non-vertebrate species or simply results from the fact that the genome of C. elegans is too small.

Curiously, when we measured the Bray-Curtis pairwise dissimilarities between the connected component sizes vectors we found that to some extent these are reflective of phylogenetic relationships between species. In other words, related species seem to have more similar SD networks (at least when it comes to connected component size distributions) than less related ones. However, the accuracy of such phylogenetic reconstruction is in fact relatively limited (see Fig. 5.4).

We again addressed the question of the preferential copying model validation in the human genome with the means of comparative genomics. We took advantage of, informally speaking, more "dynamical" comparative genomics as opposed to the previous "static" one. Data from Sudmant et al. (2013) provided us with some metric of how dynamic (or actively duplicating) different human genome loci were in the course of ape species evolution (both in human and other ape lineages). In a nutshell, for nodes of the human SD network we got the information on how evolutionary active corresponding genomic locus was. Overall, our observations agreed with the PCM: high node degree duplicated regions are more evolutionary dynamic, duplicated regions belonging to bigger components are also duplicating more actively and, finally, recent duplications that happened in a course of human evolution are enriched among duplicated regions of the giant component. To make the last observation more clear: when we simulate network growth according to the PCM, we expect that at the beginning of simulations all components grow in a comparable rate, while at some point one starts to outgrow other (a future giant component) with accelerating rate. Thus at later stages of the PCM growth we expect that duplications almost exclusively happen in a giant component which is similar to what we observe in our genomic data.

Overall, we find our results of the comparative genomics tests supporting the preferential copying model of segmental duplications propagation. Moreover, it seems likely that the PCM shaped landscape of segmental duplication not only in human, but also in other vertebrate species.

In Chapters 4 and 5 we predicted the universal rules of how segmental duplications propagate in genomes. Duplication rates grow with the number of copies of a specific duplicated region, which we called a preferential duplication rate. However, we continued analysis of dynamical properties of SD evolution, but this time we paid attention to biological characteristics that could increase duplication rates. We approximated the number of duplications that happened in each genomic region. This, as we already mentioned, is not equal to the number of alignments, because some of them are secondary. According to the PCM network growth, each node in a connected component (except for the first pair) originated as a result of other node duplication. Moreover, a daughter node inherits a fraction of edges from its mother node. We can use this knowledge to approximately predict mother to daughter nodes relationships and, consequently, which edges are primary (real duplication events) and which are secondary.

We constructed an algorithm that predicts the number of duplications (which we also call "jumps") that each node did in a course of network growth. This was done in two steps. First we assigned weights to edges of the SD network which are inversely proportional to the number of neighbors shared by a pair of connected nodes. Secondly, we searched for the minimum spanning tree (MST) that goes through edges of minimal overall weight, because lower weight means more shared neighbors, means higher probability that an edge is a primary duplication edge (see the "Edge weight assignment" section for details 6.3.2). There is no *a priori* information on how many times each genomic region was duplicated in real life. Thus the algorithm was validated on PCM synthetic networks and some additional indirect evidence was used to estimate its quality. For example, it was suggested that primary edges have to be depleted with sharp borders alignments and enriched with high sequence identity ones. This turned out to be true for our predicted MST. Moreover, it was substantially more accurate than both random spanning trees and MSTs constructed on various centrality measures when tested on PCM synthetic networks (see the "Accuracy evaluation" section for details 6.4). A manual analysis of duplication events (usually quite a complicated task) can give more accurate reconstruction of ones and uncover their complexity. However, in this thesis we are interested, firstly, in systematic predictions for all duplicated regions not limited to specific loci. Secondly, there is no need in detailed reconstruction of events for our task, we only need a number of duplications that happened in a locus. Finally, we wanted to take advantage of the SD network approach as a way to study SDs. There could be alternative strategies to study genomic features associated with actively duplicating sites, but these approaches are out of the thesis scope.

Given the MST of duplications one can predict the number of times each node was duplicated. For majority of nodes it is equal to the node degree -1, so we used such a vector of $k_i - 1$ values as a response variable. Various genomic features measured inside of duplicated regions or in flanking windows were utilized as predictor variables. Our main interest was in finding those features associated with the number of duplications, however, before solving this machine learning task we considered whether genomic characteristics are different at duplicated regions when comparing to the rest of the genome. It turned out that the difference is quite dramatic which can be explained by highly non-random distribution of SDs in the human genome. Some of the biases were expected given earlier reports on SD distribution. Then we used various techniques to extract those features associated with duplications number. This is not a trivial task, because many features are correlated with each other. We used random forest importance values assigned to predictor variables and partial correlation coefficients to find significantly associated features.

Assembly gaps were enriched at duplicated regions breakpoints. This follows from the fact that complex duplication events are harder to assemble than non-redundant sequences, thus ones are often not fully "embedded" in a genome assembly. The gene content was higher in duplicated regions which agrees with earlier report (Zhang et al., 2005), moreover, it was positively correlated with the number of duplications of a region. This seems reasonable given the fact that core duplicons are parts of SDs that are duplicated especially intensely. These cores are enriched with genes, especially those important for human evolution. The number of overlapping CpG islands, to our knowledge, was not reported as associated feature, however, based on our analysis, it was both enriched in duplicated regions and positively correlated with its duplication rates. We also wanted to note that even though CpG islands are correlated with both the G/C content and gene density, the partial correlation coefficients and the random forest importance values are not subjected to corresponding biases. Thus we, likely, see a real association not reported before. We observed that duplicated regions are located in late replicating regions. Moreover, we also found that DNA replication is slower in duplicated regions in comparison with the genome average and this affect is more prominent for actively duplicating loci. This evidence supports earlier report of Chen et al. (2015) and additionally suggests a new correlation with duplication rates. Recombination rates seem to be lower in duplicated regions, while G/C content is higher. The fraction of intrachromosomal neighbors (the same chromosome duplications) among all neighbors of a node can be considered as a characteristic of duplicated region's duplication dynamics. Some duplicated regions are prone to intrachromosomal duplications, while other duplicate to all chromosomes (see Fig. A.7). Our analysis showed negative correlation between the fraction of intrachromosomal edges and the number of jumps.

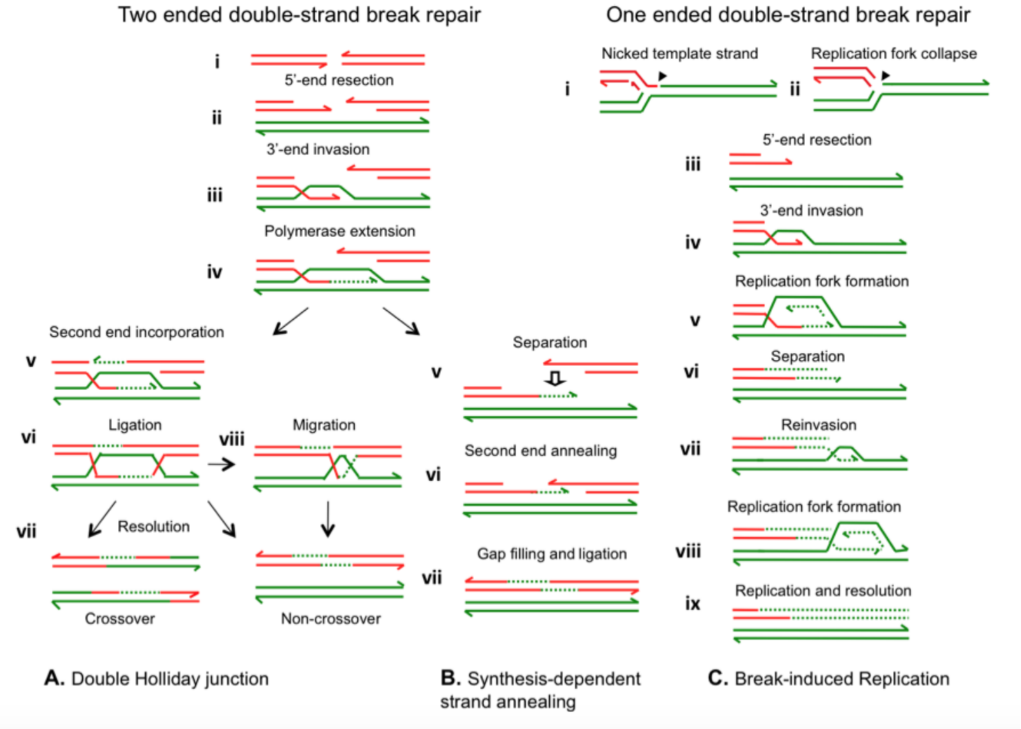
Another big group of features that we studied are high-copy repeats. We did not find strong associations between the composition of repeats and duplication rates. However, this composition was quite distinct in genomic windows located at different positions relative to breakpoints (windows inside of duplicated regions, outside ones, proximal to breakpoints and at random genomic positions). Those distributions are plotted in Fig. 6.8. We divided all repeats into two groups based on their distribution relative to duplicated regions. Those depleted inside of duplicated regions and around the breakpoints, while their fraction grows when moving away from breakpoints. These include DNA transposons, L1, L2, LTR and MIR repeat families. We can say that these repeats are rarely involved in duplication events. Another group includes repeats that are enriched at breakpoints and thus, ones making genome susceptible to duplications. These are satellite, microsatellite, *Alu* and low complexity repeats. Overall, these observations agree with earlier report by Bailey, Liu, and Eichler (2003) (also see Fig 6.6), however, we additionally observed significant enrichment of microsatellites and low complexity repeats at breakpoints in comparison with the genome average. Our analysis is different from the one done by Bailey, Liu, and Eichler (2003), because we considered all duplicated regions (regions that underwent any number of duplication events), while in earlier analysis only trivial duplications (single non-overlapping ones) were studied.

In conclusion, we think that the network formalization for the analysis of SDs is a good way to study the evolution of SDs in human or other species genomes. We illustrated how it can be applied to infer dynamical properties of SD propagation, compare them between distinct species and to reconstruct (approximately) real duplication events among all observed alignments. It allowed us to suggest a universal model of segmental duplications propagation: a duplication rate grows with the number of copies of a duplicated region. More than that, we predicted several associations between genomic features and high duplication rates, a task which was not systematically studied before. Overall, we think that this research projects gives a broad picture of factors affecting duplication rates and suggests a method (a network of segmental duplications) that can further be applied to other problems in computational biology. Glad someone read the thesis that far:)

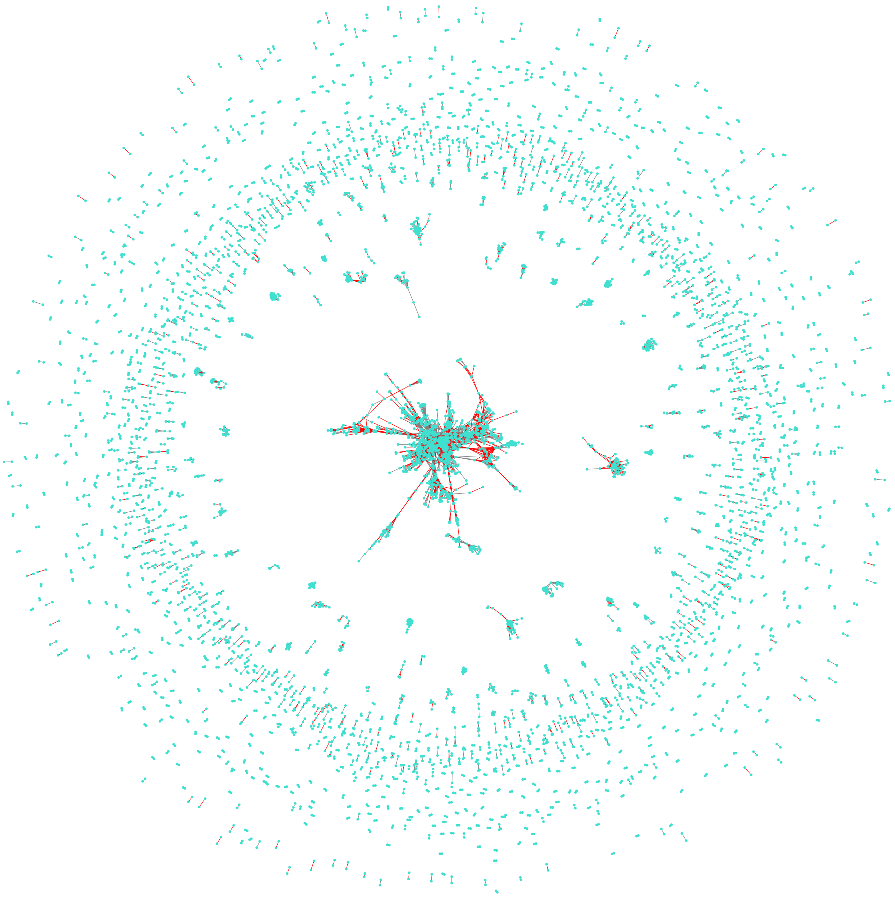
Appendix A

Supplementary materials

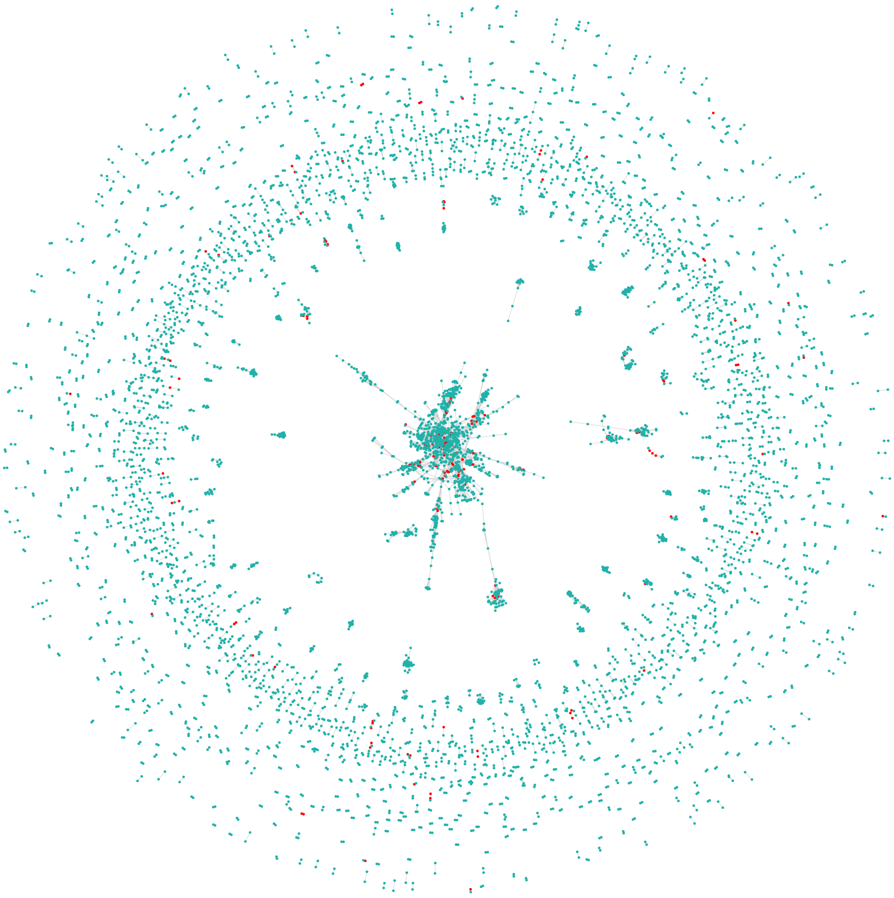
A.1 Supplementary figures



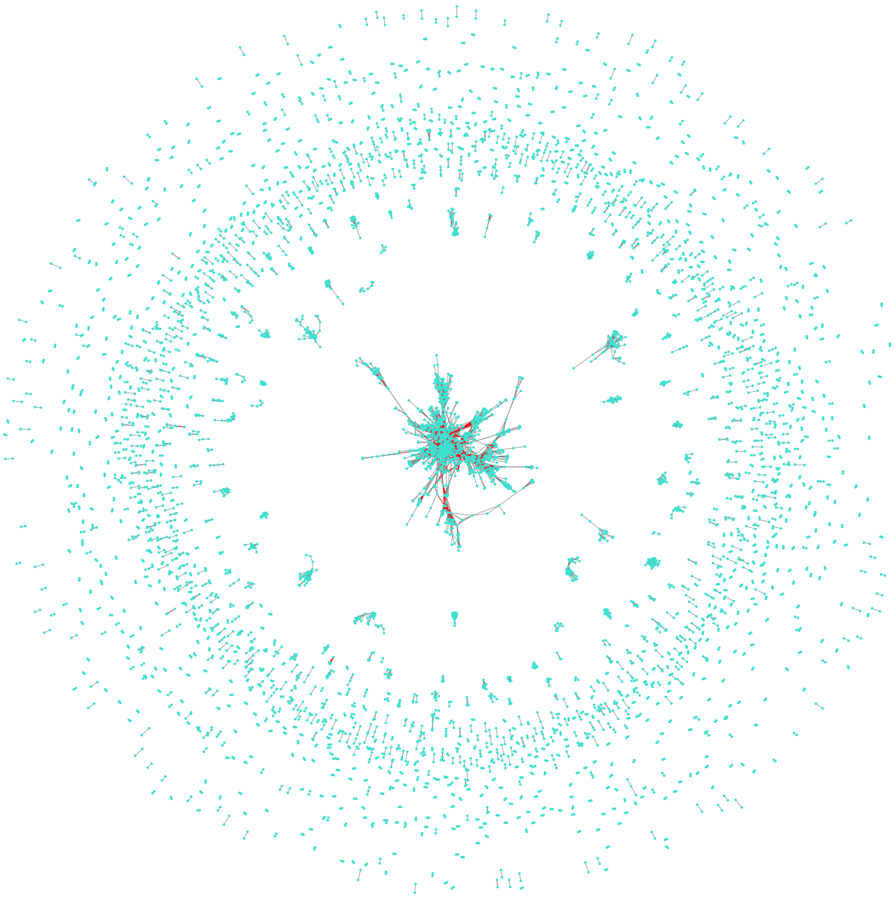
SUP. FIGURE A.1: **Homologous recombination pathways.** Double-strand breaks are fixed with homologous recombination pathways. If both fragments of DNA are detected by reparation machinery, either double Holliday junction pathway or synthesis-dependent strand annealing is activated. Double Holliday junction can be resolved in crossing-over or not depending on how DNA strands are cut. The one ended pathway is often activated during DNA replication. A missing chromosome arm is restored via break-induced replication based on homologous chromosome sequence. The figure source: Hastings, Ira, and Lupski, 2009



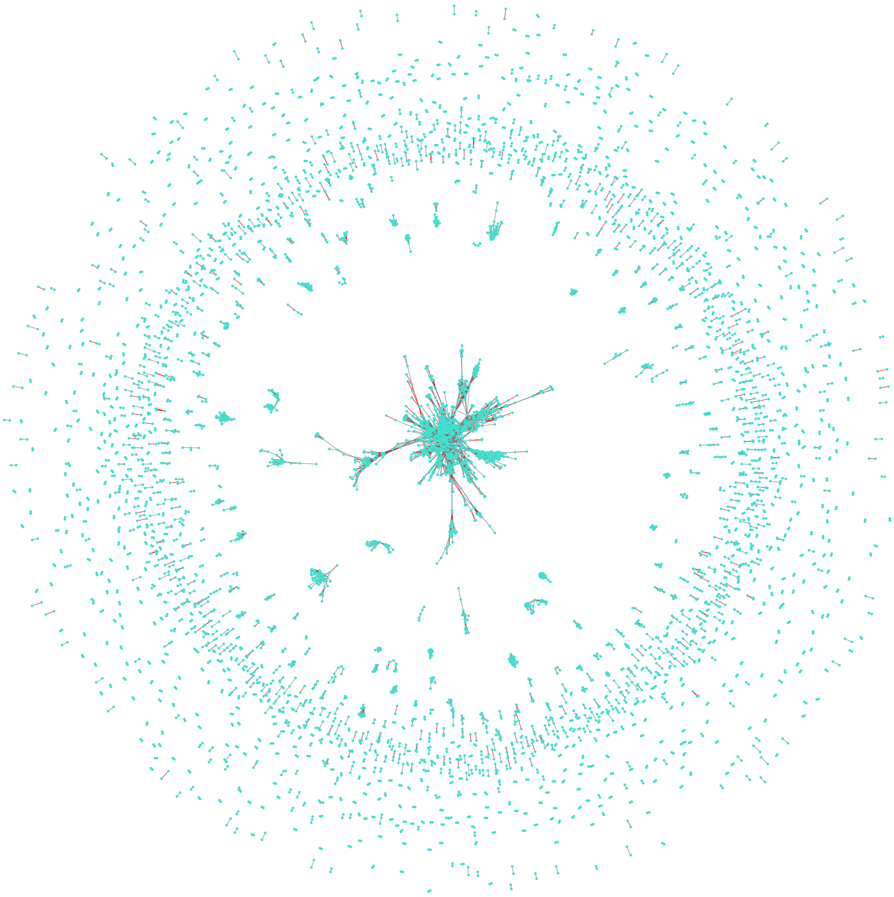
SUP. FIGURE A.2: **The SD network with intrachromosomal edges specifically coloured in red is illustrated.**



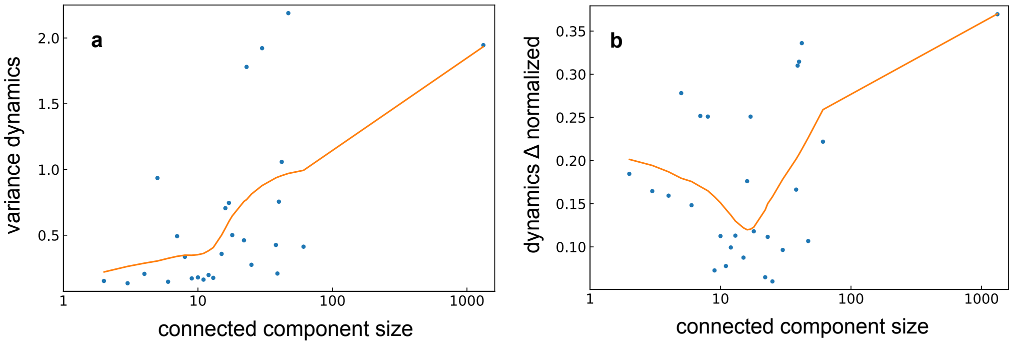
SUP. FIGURE A.3: **The SD network where nodes having self-loop edges are coloured in red.** The self-loop edges are those edges that connect a node to itself. One can see that these edges are mostly concentrated in the giant component. In the main text we consider the trimmed version of the SD network: the one where all self-loop edges are removed.



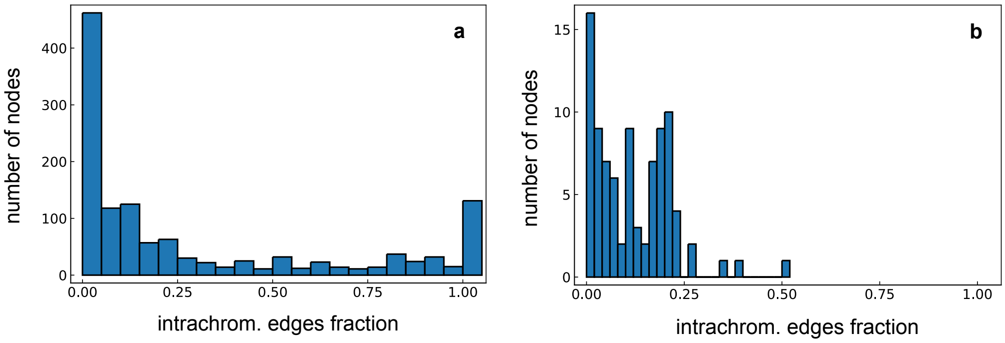
SUP. FIGURE A.4: **The SD network with multiple edges specifically coloured in red is illustrated.** Multiple edges appear when more than one edge connects a pair of nodes. One can see that these edges are mostly concentrated in the giant component. In the main text we consider the trimmed version of the SD network: the one where all redundant edges are removed.



SUP. FIGURE A.5: **The SD network with tandem edges specifically coloured in red is illustrated.** Tandem edges correspond to intrachromosomal SDs where the distance between two copies is less than $5 * 10^5$ bps.



SUP. FIGURE A.6: **The plot represents how different metrices of duplicational dynamics grow with a connected component size.** These include the second central moment of copy-number (**a**) and normalized dynamics $\Delta_i / \overline{X_i}$ (**b**). The locally estimated scatterplot smoothing function was added to make it more visible (orange line).



SUP. FIGURE A.7: **The histogram illustrates the distribution of fractions of intrachromosomal edges from all edges of nodes.** Nodes that are prone to intrachromosomal duplications have values close to 1, while those with predominantly interchromosomal duplications are closer to 0. In (**a**) we include nodes with $>$ 5 neighbors, while in (**b**) actively duplicating nodes with $>$ 50 neighbors are included. We can see that the first distribution is bi-modal with two peaks around 0 and 1. Nodes with high node degree from the second histogram are, on the other hand, depleted with intrachromosomal hotspots.

A.2 Computational Tools

The analysis and simulation described in this thesis were performed using the Julia programming language Bezanson et al., 2017. We used the following packages:

- All the steps of our network analysis are performed using the LightGraphs.jl package in the Julia programming language Seth Bromberger and contributors, 2017. Except for simple feature extraction and modification of the SD network we used this package to calculate the mean clustering coefficient, the average shortest path length, study the modularity of the network, use configuration model, generate random networks using the Erdős–Rényi model and scale-free networks using the Barabási–Albert model Newman, 2010, Erdös and Rényi, 1959, Albert and Barabási, 2002. We applied the Barabási–Albert model with $m = 7$ (a number of edges that a new node forms) to get the size of a resulting network close to the size of the SD network giant component.

- The network visualization is done with the GraphPlot.jl package in Julia.

- We used the Approximate Bayesian Computation method (ABC) from the ApproxBayes.jl Julia package.

- To estimate the PCM parameter values the loss function was minimized with the Nelder–Mead method from Optim.jl Julia package Nelder and Mead, 1965, Mogensen and Riseth, 2018.

Appendix B

Abstract

Segmental duplications (SDs) are long DNA sequences that are repeated in a genome and have high sequence identity. In contrast to repetitive elements they are often unique and only some have multiple copies in a genome. There are several well-studied mechanisms responsible for segmental duplications: non-allelic homologous recombination, non-homologous end joining and replication slippage that act in more complex scenarios. Overall, SDs comprise around 5% of the human genome and play an important evolutionary role. For example, the expansion of segmental duplications in the early human lineage, likely, affected the human brain evolution. SDs are sites of recurrent genomic rearrangements, including those responsible for genetic disorders and so on. However, we do not have a full understanding of the dynamic properties of the duplication process. Can we suggest a universal scenario of SDs propagation in genomes and which genomic characteristics affect this process? This thesis is dedicated to answer those questions.

We study segmental duplications through a graph representation where nodes represent genomic regions and edges represent duplications between them. The resulting network (the SD network) has distinct features which allow us to make inference on the evolution of segmental duplications. We propose a network growth model that explains features of the SD network thus giving us insights on dynamics of segmental duplications in the human genome. Based on our analysis of genomes of other species the network growth model seems to be applicable for multiple vertebrate genomes. Our model suggests that duplication rates of genomic loci grow linearly with the number of copies of a duplicated region. Finally, we studied genomic features associated with duplicated regions. Our evidence supports earlier observations and gives new insights about a duplication process.

Appendix C

Zusammenfassung

Segmentale Duplikationen (SD) sind lange DNA-Sequenzen, die in einem Genom wiederholt werden und eine hohe Sequenzidentität aufweisen. Im Gegensatz zu repetitiven Elementen sind sie oft einzigartig und nur einige haben mehrere Kopien in einem Genom. Es gibt mehrere gut untersuchte Mechanismen, die für segmentale Duplikationen verantwortlich sind: nicht-allelische homologe Rekombination, nicht-homologes End Joining und Replikationsschlupf, die in komplexeren Szenarien wirken. Insgesamt machen SD etwa 5% des menschlichen Genoms aus und spielen eine wichtige evolutionäre Rolle. So hat beispielsweise die Ausbreitung segmentaler Duplikationen in der frühen menschlichen Abstammungslinie wahrscheinlich die Entwicklung des menschlichen Gehirns beeinflusst. SD sind Orte wiederkehrender genomischer Umlagerungen, einschließlich derer, die für Krankheiten verantwortlich sind. Die dynamischen Eigenschaften des Duplikationsprozesses sind uns jedoch noch nicht vollständig bekannt. Können wir ein universelles Szenario für die Ausbreitung von SD in Genomen vorschlagen und welche genomischen Merkmale beeinflussen diesen Prozess? Diese Arbeit ist der Beantwortung dieser Fragen gewidmet.

Wir untersuchen segmentale Duplikationen anhand eines Graphen, in dem Knoten genomische Regionen und Kanten Duplikationen zwischen ihnen darstellen. Das sich daraus ergebende Netzwerk (das SD-Netzwerk) weist bestimmte Merkmale auf, die es uns ermöglichen, Rückschlüsse auf die Entwicklung segmentaler Duplikationen zu ziehen. Wir schlagen ein Netzwerkwachstumsmodell vor, das die Merkmale des SD-Netzwerks erklärt und uns somit Einblicke in die Dynamik segmentaler Duplikationen im menschlichen Genom gewährt. Basierend auf unserer Analyse von Genomen anderer Spezies scheint das Netzwerkwachstumsmodell für mehrere Wirbeltiergenome anwendbar zu sein. Unser Modell legt nahe, dass die Duplikationsraten genomischer Loci linear mit der Anzahl der Kopien einer duplizierten Region wachsen. Schließlich haben wir genomische Merkmale untersucht, die mit duplizierten Regionen in Verbindung stehen. Unsere Ergebnisse bestätigen frühere Beobachtungen und geben neue Einblicke in den Verdopplungsprozess.

Bibliography

Abdullaev, Eldar T, Iren R Umarova, and Peter F Arndt (2021). “Modelling segmental duplications in the human genome.” In: *BMC Genomics* 22.1, p. 496. DOI: 10.1186/s12864-021-07789-7.

Aitman, Timothy J et al. (2006). “Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans.” In: *Nature* 439.7078, pp. 851–855. ISSN: 1476-4687. DOI: 10.1038/nature04489.

Albert, Réka and Albert-László Barabási (2002). “Statistical mechanics of complex networks”. In: *Reviews of Modern Physics* 74.1, pp. 47–97. ISSN: 0034-6861. DOI: 10.1103/{RevModPhys}.74.47.

Aldrup-Macdonald, Megan E and Beth A Sullivan (2014). “The past, present, and future of human centromere genomics.” In: *Genes* 5.1, pp. 33–50. DOI: 10.3390/genes5010033.

Altmann, André et al. (2010). “Permutation importance: a corrected feature importance measure.” In: *Bioinformatics* 26.10, pp. 1340–1347. DOI: 10.1093/bioinformatics/btq134.

Altschul, S F et al. (1990). “Basic local alignment search tool.” In: *Journal of Molecular Biology* 215.3, pp. 403–410. DOI: 10.1016/S0022-2836(05)80360-2.

Arlt, Martin F et al. (2009). “Replication stress induces genome-wide copy number changes in human cells that resemble polymorphic and pathogenic variants.” In: *American Journal of Human Genetics* 84.3, pp. 339–350. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2009.01.024.

Bailey, J A et al. (2001). “Segmental duplications: organization and impact within the current human genome project assembly.” In: *Genome Research* 11.6, pp. 1005–1017. DOI: 10.1101/gr.gr-1871r.

Bailey, Jeffrey A and Evan E Eichler (2006). “Primate segmental duplications: crucibles of evolution, diversity and disease.” In: *Nature Reviews. Genetics* 7.7, pp. 552–564. DOI: 10.1038/nrg1895.

Bailey, Jeffrey A, Ge Liu, and Evan E Eichler (2003). “An Alu transposition model for the origin and expansion of human segmental duplications.” In: *American Journal of Human Genetics* 73.4, pp. 823–834. DOI: 10.1086/378594.

Bailey, Jeffrey A et al. (2002). “Recent segmental duplications in the human genome.” In: *Science* 297.5583, pp. 1003–1007. ISSN: 1095-9203. DOI: 10.1126/science.1072047.

Bailey, Jeffrey A et al. (2004). "Analysis of segmental duplications and genome assembly in the mouse." In: *Genome Research* 14.5, pp. 789–801. DOI: 10.1101/gr.2238404.

Batzer, Mark A and Prescott L Deininger (2002). "Alu repeats and human genomic diversity." In: *Nature Reviews. Genetics* 3.5, pp. 370–379. DOI: 10.1038/nrg798.

Bezanson, Jeff et al. (2017). "Julia: A fresh approach to numerical computing". In: *SIAM review* 59.1, pp. 65–98.

Bidichandani, S I, T Ashizawa, and P I Patel (1998). "The GAA triplet-repeat expansion in Friedreich ataxia interferes with transcription and may be associated with an unusual DNA structure." In: *American Journal of Human Genetics* 62.1, pp. 111–121. DOI: 10.1086/301680.

Bonaglia, Maria Clara et al. (2005). "A 2.3 Mb duplication of chromosome 8q24.3 associated with severe mental retardation and epilepsy detected by standard karyotype." In: *European Journal of Human Genetics* 13.5, pp. 586–591. DOI: 10.1038/sj.ejhg.5201369.

Bose, Promita et al. (2014). "Tandem repeats and G-rich sequences are enriched at human CNV breakpoints." In: *Plos One* 9.7, e101607. DOI: 10.1371/journal.pone.0101607.

Bray, J. Roger and J. T. Curtis (1957). "An Ordination of the Upland Forest Communities of Southern Wisconsin". In: *Ecological monographs* 27.4, pp. 325–349. ISSN: 00129615. DOI: 10.2307/1942268.

Brin, Sergey and Lawrence Page (1998). "The anatomy of a large-scale hypertextual Web search engine". In: *Computer Networks and ISDN Systems* 30.1-7, pp. 107–117. ISSN: 01697552. DOI: 10.1016/S0169-7552(98)00110-X.

Cantsilieris, Stuart et al. (2020). "An evolutionary driver of interspersed segmental duplications in primates." In: *Genome Biology* 21.1, p. 202. DOI: 10.1186/s13059-020-02074-4.

Carvalho, Claudia M B and James R Lupski (2016). "Mechanisms underlying structural variant formation in genomic disorders." In: *Nature Reviews. Genetics* 17.4, pp. 224–238. DOI: 10.1038/nrg.2015.25.

Charrier, Cécile et al. (2012). "Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation." In: *Cell* 149.4, pp. 923–935. ISSN: 00928674. DOI: 10.1016/j.cell.2012.03.034.

Chen, Lu et al. (2015). "CNV instability associated with DNA replication dynamics: evidence for replicative mechanisms in CNV mutagenesis." In: *Human Molecular Genetics* 24.6, pp. 1574–1583. DOI: 10.1093/hmg/ddu572.

Cheng, Ze et al. (2005). "A genome-wide comparison of recent chimpanzee and human segmental duplications." In: *Nature* 437.7055, pp. 88–93. DOI: 10.1038/nature04000.

Chung, Fan et al. (2003). "Duplication models for biological networks." In: *Journal of Computational Biology* 10.5, pp. 677–687. DOI: 10.1089/106652703322539024.

Consortium, 1000 Genomes Project et al. (2015). “A global reference for human genetic variation.” In: *Nature* 526.7571, pp. 68–74. ISSN: 0028-0836. DOI: 10.1038/nature15393.

De, Subhajyoti and Franziska Michor (2011). “DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes.” In: *Nature Biotechnology* 29.12, pp. 1103–1108. DOI: 10.1038/nbt.2030.

DeFilippes, F M (1984). “Effect of aphidicolin on vaccinia virus: isolation of an aphidicolin-resistant mutant.” In: *Journal of Virology* 52.2, pp. 474–482. DOI: 10.1128/{JVI}.52.2.474-482.1984.

Dennis, Megan Y and Evan E Eichler (2016). “Human adaptation and evolution by segmental duplication.” In: *Current Opinion in Genetics & Development* 41, pp. 44–52. DOI: 10.1016/j.gde.2016.08.001.

Dennis, Megan Y et al. (2012). “Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication.” In: *Cell* 149.4, pp. 912–922. DOI: 10.1016/j.cell.2012.03.033.

Dennis, Megan Y et al. (2017). “The evolution and population diversity of human-specific segmental duplications.” In: *Nature Ecology & Evolution* 1.3, p. 69. DOI: 10.1038/s41559-016-0069.

Dittwald, Piotr et al. (2013). “NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits.” In: *Genome Research* 23.9, pp. 1395–1409. DOI: 10.1101/gr.152454.112.

Dumas, Laura et al. (2007). “Gene copy number variation spanning 60 million years of human and primate evolution.” In: *Genome Research* 17.9, pp. 1266–1277. DOI: 10.1101/gr.6557307.

Dumas, Laura J et al. (2012). “DUF1220-domain copy number implicated in human brain-size pathology and evolution.” In: *American Journal of Human Genetics* 91.3, pp. 444–454. DOI: 10.1016/j.ajhg.2012.07.016.

Eichler, E E, N Archidiacono, and M Rocchi (1999). “CAGGG repeats and the pericentromeric duplication of the hominoid genome.” In: *Genome Research* 9.11, pp. 1048–1058. DOI: 10.1101/gr.9.11.1048.

Eichler, E E et al. (1996). “Duplication of a gene-rich cluster between 16p11.1 and Xq28: a novel pericentromeric-directed mechanism for paralogous genome evolution.” In: *Human Molecular Genetics* 5.7, pp. 899–912. DOI: 10.1093/hmg/5.7.899.

Eichler, E E et al. (1997). “Interchromosomal duplications of the adrenoleukodystrophy locus: a phenomenon of pericentromeric plasticity.” In: *Human Molecular Genetics* 6.7, pp. 991–1002. DOI: 10.1093/hmg/6.7.991.

Erdös, P. and A. Rényi (1959). “On Random Graphs (part 1)”. In: *Publicationes Mathematicae Debrecen* 6, p. 290.

Florio, Marta et al. (2015). “Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion.” In: *Science* 347.6229, pp. 1465–1470. DOI: 10.1126/science.aaa1975.

Fogel, S et al. (1983). “Gene amplification in yeast: CUP1 copy number regulates copper resistance.” In: *Current Genetics* 7.5, pp. 347–355. DOI: 10.1007/{BF00445874}.

Fortna, Andrew et al. (2004). “Lineage-specific gene duplication and loss in human and great ape evolution.” In: *PLoS Biology* 2.7, E207. DOI: 10.1371/journal.pbio.0020207.

Gao, Kun and Jonathan Miller (2011). “Algebraic distribution of segmental duplication lengths in whole-genome sequence self-alignments.” In: *Plos One* 6.7, e18464. DOI: 10.1371/journal.pone.0018464.

Gazave, Elodie et al. (2011). “Copy number variation analysis in the great apes reveals species-specific patterns of structural variation.” In: *Genome Research* 21.10, pp. 1626–1639. DOI: 10.1101/gr.117242.110.

Gregory, T Ryan (2005). “Synergy between sequence and size in large-scale genomics.” In: *Nature Reviews. Genetics* 6.9, pp. 699–708. DOI: 10.1038/nrg1674.

Groot, P C et al. (1989). “The human alpha-amylase multigene family consists of haplotypes with variable numbers of genes.” In: *Genomics* 5.1, pp. 29–42. DOI: 10.1016/0888-7543(89)90083-9.

Grunau, Christoph et al. (2006). “Mapping of the juxtacentromeric heterochromatin-euchromatin frontier of human chromosome 21.” In: *Genome Research* 16.10, pp. 1198–1207. DOI: 10.1101/gr.5440306.

Guy, J et al. (2000). “Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10q.” In: *Human Molecular Genetics* 9.13, pp. 2029–2042. DOI: 10.1093/hmg/9.13.2029.

Gymrek, Melissa et al. (2016). “Abundant contribution of short tandem repeats to gene expression variation in humans.” In: *Nature Genetics* 48.1, pp. 22–29. ISSN: 1061-4036. DOI: 10.1038/ng.3461.

Hastings, P J, Grzegorz Ira, and James R Lupski (2009). “A microhomology-mediated break-induced replication model for the origin of human copy number variation.” In: *PLoS Genetics* 5.1, e1000327. DOI: 10.1371/journal.pgen.1000327.

Hastings, P J et al. (2009). “Mechanisms of change in gene copy number.” In: *Nature Reviews. Genetics* 10.8, pp. 551–564. DOI: 10.1038/nrg2593.

Hayward, Alexander (2017). “Origin of the retroviruses: when, where, and how?” In: *Current opinion in virology* 25, pp. 23–27. DOI: 10.1016/j.coviro.2017.06.006.

He, Yaoxi et al. (2019). “Long-read assembly of the Chinese rhesus macaque genome and identification of ape-specific structural variants.” In: *Nature Communications* 10.1, p. 4233. DOI: 10.1038/s41467-019-12174-w.

Hon, Ting et al. (2020). “Highly accurate long-read HiFi sequencing data for five complex genomes.” In: *Scientific data* 7.1, p. 399. ISSN: 2052-4463. DOI: 10.1038/s41597-020-00743-4.

Horvath, Julie E et al. (2005). “Punctuated duplication seeding events during the evolution of human chromosome 2p11.” In: *Genome Research* 15.7, pp. 914–927. DOI: 10.1101/gr.3916405.

IHGSC (2001). "Initial sequencing and analysis of the human genome". In: *Nature* 409.6822, pp. 860–921. ISSN: 0028-0836. DOI: 10.1038/35057062.

— (2004). "Finishing the euchromatic sequence of the human genome." In: *Nature* 431.7011, pp. 931–945. ISSN: 1476-4687. DOI: 10.1038/nature03001.

Jagannathan, Vidhya et al. (2021). "Dog10K_Boxer_Tasha_1.0: A Long-Read Assembly of the Dog Reference Genome." In: *Genes* 12.6. DOI: 10.3390/genes12060847.

Jain, Miten et al. (2018). "Nanopore sequencing and assembly of a human genome with ultra-long reads." In: *Nature Biotechnology* 36.4, pp. 338–345. ISSN: 1087-0156. DOI: 10.1038/nbt.4060.

Jiang, Zhaoshi et al. (2007). "Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution." In: *Nature Genetics* 39.11, pp. 1361–1368. ISSN: 1546-1718. DOI: 10.1038/ng.2007.9.

Jobling, Mark A., Matthew Hurles, and Chris Tyler-Smith (2019). *Human Evolutionary Genetics*. Garland Science. ISBN: 9780203487211. DOI: 10.1201/9780203487211.

Johnson, M E et al. (2001). "Positive selection of a gene family during the emergence of humans and African apes." In: *Nature* 413.6855, pp. 514–519. ISSN: 0028-0836. DOI: 10.1038/35097067.

Kahn, Crystal L, Borislav H Hristov, and Benjamin J Raphael (2010). "Parsimony and likelihood reconstruction of human segmental duplications." In: *Bioinformatics* 26.18, pp. i446–52. DOI: 10.1093/bioinformatics/btq368.

Kazazian, H H et al. (1988). "Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man." In: *Nature* 332.6160, pp. 164–166. DOI: 10.1038/332164a0.

Kelley, David R and Steven L Salzberg (2010). "Detection and correction of false segmental duplications caused by genome mis-assembly." In: *Genome Biology* 11.3, R28. DOI: 10.1186/gb-2010-11-3-r28.

Kent, W James et al. (2002). "The human genome browser at UCSC." In: *Genome Research* 12.6, pp. 996–1006. DOI: 10.1101/gr.229102.

Kim, Philip M et al. (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." In: *Genome Research* 18.12, pp. 1865–1874. DOI: 10.1101/gr.081422.108.

Kirsch, Stefan et al. (2008). "Evolutionary dynamics of segmental duplications from human Y-chromosomal euchromatin/heterochromatin transition regions." In: *Genome Research* 18.7, pp. 1030–1042. DOI: 10.1101/gr.076711.108.

Koning, A P Jason de et al. (2011). "Repetitive elements may comprise over two-thirds of the human genome." In: *PLoS Genetics* 7.12, e1002384. DOI: 10.1371/journal.pgen.1002384.

Koonin, Eugene V (2016). "Viruses and mobile elements as drivers of evolutionary transitions." In: *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 371.1701. DOI: 10.1098/rstb.2015.0442.

Koonin, Eugene V, Valerian V Dolja, and Mart Krupovic (2015). “Origins and evolution of viruses of eukaryotes: The ultimate modularity.” In: *Virology* 479-480, pp. 2–25. DOI: 10.1016/j.virol.2015.02.039.

Korbel, Jan O et al. (2007). “Paired-end mapping reveals extensive structural variation in the human genome.” In: *Science* 318.5849, pp. 420–426. ISSN: 1095-9203. DOI: 10.1126/science.1149504.

Koren, Amnon et al. (2012). “Differential relationship of DNA replication timing to different forms of human mutation and variation.” In: *American Journal of Human Genetics* 91.6, pp. 1033–1040. DOI: 10.1016/j.ajhg.2012.10.018.

Koszul, Romain and Gilles Fischer (2009). “A prominent role for segmental duplications in modeling eukaryotic genomes.” In: *Comptes Rendus Biologies* 332.2-3, pp. 254–266. DOI: 10.1016/j.crvi.2008.07.005.

Krejci, Lumir et al. (2012). “Homologous recombination and its regulation.” In: *Nucleic Acids Research* 40.13, pp. 5795–5818. DOI: 10.1093/nar/gks270.

Kruskal, Joseph B. (1956). “On the shortest spanning subtree of a graph and the traveling salesman problem”. In: *Proceedings of the American Mathematical Society* 7.1, pp. 48–48. ISSN: 0002-9939. DOI: 10.1090/S0002-9939-1956-0078686-7.

Kuhn, Werner (1930). “Über die Kinetik des Abbaues hochmolekularer Ketten”. In: *Berichte der deutschen chemischen Gesellschaft (A and B Series)* 63.6, pp. 1503–1509. ISSN: 03659488. DOI: 10.1002/cber.19300630631.

Kurtz, Stefan et al. (2004). “Versatile and open software for comparing large genomes.” In: *Genome Biology* 5.2, R12. DOI: 10.1186/gb-2004-5-2-r12.

Lahortiga, Idoya et al. (2007). “Duplication of the MYB oncogene in T cell acute lymphoblastic leukemia.” In: *Nature Genetics* 39.5, pp. 593–595. DOI: 10.1038/ng2025.

Lee, Jennifer A, Claudia M B Carvalho, and James R Lupski (2007). “A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders.” In: *Cell* 131.7, pp. 1235–1247. DOI: 10.1016/j.cell.2007.11.037.

Linardopoulou, Elena V et al. (2005). “Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication.” In: *Nature* 437.7055, pp. 94–100. DOI: 10.1038/nature04029.

Liu, George E et al. (2009). “Analysis of recent segmental duplications in the bovine genome.” In: *BMC Genomics* 10, p. 571. DOI: 10.1186/1471-2164-10-571.

Locke, D P et al. (2005). “Molecular evolution of the human chromosome 15 pericentromeric region.” In: *Cytogenetic and Genome Research* 108.1-3, pp. 73–82. DOI: 10.1159/000080804.

Locke, Devin P et al. (2011). “Comparative and demographic analysis of orang-utan genomes.” In: *Nature* 469.7331, pp. 529–533. DOI: 10.1038/nature09687.

Lu, Junjie et al. (2014). “The distribution of genomic variations in human iPSCs is related to replication-timing reorganization during reprogramming.” In: *Cell reports* 7.1, pp. 70–78. DOI: 10.1016/j.celrep.2014.03.007.

Lupski, James R and Pawel Stankiewicz (2005). “Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes.” In: *PLoS Genetics* 1.6, e49. DOI: 10.1371/journal.pgen.0010049.

Marques-Bonet, T and E E Eichler (2009). “The evolution of human segmental duplications and the core duplicon hypothesis.” In: *Cold Spring Harbor Symposia on Quantitative Biology* 74, pp. 355–362. DOI: 10.1101/sqb.2009.74.011.

Massip, Florian and Peter F Arndt (2013). “Neutral evolution of duplicated DNA: an evolutionary stick-breaking process causes scale-invariant behavior.” In: *Physical Review Letters* 110.14, p. 148101. DOI: 10.1103/{PhysRevLett}.110.148101.

Massip, Florian et al. (2015). “How evolution of genomes is reflected in exact DNA sequence match statistics.” In: *Molecular Biology and Evolution* 32.2, pp. 524–535. DOI: 10.1093/molbev/msu313.

— (2016). “Comparing the statistical fate of paralogous and orthologous sequences.” In: *Genetics* 204.2, pp. 475–482. DOI: 10.1534/genetics.116.193912.

Meyer, Matthias et al. (2012). “A high-coverage genome sequence from an archaic Denisovan individual.” In: *Science* 338.6104, pp. 222–226. DOI: 10.1126/science.1224344.

Miga, Karen H et al. (2020). “Telomere-to-telomere assembly of a complete human X chromosome.” In: *Nature* 585.7823, pp. 79–84. ISSN: 0028-0836. DOI: 10.1038/s41586-020-2547-7.

Miki, Y et al. (1992). “Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer.” In: *Cancer Research* 52.3, pp. 643–645.

Mogensen, Patrick Kofod and Asbjørn Nilsen Riseth (2018). “Optim: A mathematical optimization package for Julia”. In: *Journal of Open Source Software* 3.24, p. 615. DOI: 10.21105/joss.00615.

Murnane, John P (2012). “Telomere dysfunction and chromosome instability.” In: *Mutation Research* 730.1-2, pp. 28–36. DOI: 10.1016/j.mrfmmm.2011.04.008.

Mustajoki, S et al. (1999). “Insertion of Alu element responsible for acute intermittent porphyria.” In: *Human Mutation* 13.6, pp. 431–438. DOI: 10.1002/({SICI})1098-1004(1999)13:6\textless431::{AID}-{HUMU2\textgreater3}.0.{CO};2-Y.

Nelder, J. A. and R. Mead (1965). “A Simplex Method for Function Minimization”. In: *The Computer Journal* 7.4, pp. 308–313. ISSN: 0010-4620. DOI: 10.1093/comjnl/7.4.308.

Newman, Mark (2010). *Networks*. Oxford University Press. ISBN: 9780199206650. DOI: 10.1093/acprof:oso/9780199206650.001.0001.

Newman, Tera L et al. (2005). “A genome-wide survey of structural variation between human and chimpanzee.” In: *Genome Research* 15.10, pp. 1344–1356. DOI: 10.1101/gr.4338005.

Numanagic, Ibrahim et al. (2018). “Fast characterization of segmental duplications in genome assemblies.” In: *Bioinformatics* 34.17, pp. i706–i714. DOI: 10.1093/bioinformatics/bty586.

Nurk, Sergey et al. (2021). “The complete sequence of a human genome”. In: *BioRxiv*. DOI: 10.1101/2021.05.26.445798.

Ohno, Susumu (1970). *Evolution by gene duplication*. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-642-86659-3. DOI: 10.1007/978-3-642-86659-3.

Olson, M V (1999). “When less is more: gene loss as an engine of evolutionary change.” In: *American Journal of Human Genetics* 64.1, pp. 18–23. DOI: 10.1086/302219.

Pace, John K and Cédric Feschotte (2007). “The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage.” In: *Genome Research* 17.4, pp. 422–432. ISSN: 1088-9051. DOI: 10.1101/gr.5826307.

Pannunzio, Nicholas R et al. (2014). “Non-homologous end joining often uses microhomology: implications for alternative end joining.” In: *DNA Repair* 17, pp. 74–80. DOI: 10.1016/j.dnarep.2014.02.006.

Pearson, Christopher E, Kerrie Nichol Edamura, and John D Cleary (2005). “Repeat instability: mechanisms of dynamic mutations.” In: *Nature Reviews. Genetics* 6.10, pp. 729–742. DOI: 10.1038/nrg1689.

Perry, George H et al. (2007). “Diet and the evolution of human amylase gene copy number variation.” In: *Nature Genetics* 39.10, pp. 1256–1260. ISSN: 1546-1718. DOI: 10.1038/ng2123.

Plohl, M, N Meštrović, and B Mravinac (2012). “Satellite DNA evolution.” In: *Genome dynamics* 7, pp. 126–152. DOI: 10.1159/000337122.

Prado-Martinez, Javier et al. (2013). “Great ape genetic diversity and population history.” In: *Nature* 499.7459, pp. 471–475. DOI: 10.1038/nature12228.

Pu, Lianrong, Yu Lin, and Pavel A Pevzner (2018). “Detection and analysis of ancient segmental duplications in mammalian genomes.” In: *Genome Research* 28.6, pp. 901–909. ISSN: 1088-9051. DOI: 10.1101/gr.228718.117.

Raghavan, Usha Nandini, Réka Albert, and Soundar Kumara (2007). “Near linear time algorithm to detect community structures in large-scale networks”. In: *Physical Review E* 76.3. ISSN: 1539-3755. DOI: 10.1103/{PhysRevE}.76.036106.

Raphael, Benjamin J and Pavel A Pevzner (2004). “Reconstructing tumor amplisomes.” In: *Bioinformatics* 20 Suppl 1, pp. i265–73. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/bth931.

Redon, Richard et al. (2006). “Global variation in copy number in the human genome.” In: *Nature* 444.7118, pp. 444–454. ISSN: 1476-4687. DOI: 10.1038/nature05329.

Rovelet-Lecrux, Anne et al. (2006). “APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy.” In: *Nature Genetics* 38.1, pp. 24–26. ISSN: 1061-4036. DOI: 10.1038/ng1718.

Rubin, Donald B. (1984). “Bayesianly Justifiable and Relevant Frequency Calculations for the Applies Statistician”. In: *The Annals of Statistics* 12.4, pp. 1151–1172.

Salzberg, Steven L and James A Yorke (2005). “Beware of mis-assembled genomes.” In: *Bioinformatics* 21.24, pp. 4320–4321. DOI: 10.1093/bioinformatics/bti769.

Samonte, Rhea Vallente and Evan E Eichler (2002). "Segmental duplications and the evolution of the primate genome." In: *Nature Reviews. Genetics* 3.1, pp. 65–72. DOI: 10.1038/nrg705.

Serrato-Capuchina, Antonio and Daniel R Matute (2018). "The role of transposable elements in speciation." In: *Genes* 9.5. DOI: 10.3390/genes9050254.

Seth Bromberger, James Fairbanks and other contributors (2017). *JuliaGraphs/LightGraphs.jl: an optimized graphs package for the Julia programming language*. DOI: 10.5281/zenodo.889971.

Sharp, Andrew J et al. (2006). "Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome." In: *Nature Genetics* 38.9, pp. 1038–1042. ISSN: 1061-4036. DOI: 10.1038/ng1862.

Sharp, Andrew J et al. (2008). "A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures." In: *Nature Genetics* 40.3, pp. 322–328. DOI: 10.1038/ng.93.

Shatskikh, Aleksei S et al. (2020). "Functional significance of satellite dnas: insights from drosophila." In: *Frontiers in cell and developmental biology* 8, p. 312. DOI: 10.3389/fcell.2020.00312.

She, Xinwei et al. (2004). "The structure and evolution of centromeric transition regions within the human genome." In: *Nature* 430.7002, pp. 857–864. ISSN: 1476-4687. DOI: 10.1038/nature02806.

Sheinman, Michael et al. (2016). "Evolutionary dynamics of selfish DNA explains the abundance distribution of genomic subsequences." In: *Scientific Reports* 6, p. 30851. DOI: 10.1038/srep30851.

Stankiewicz, Paweł and James R Lupski (2002). "Genome architecture, rearrangements and genomic disorders." In: *Trends in Genetics* 18.2, pp. 74–82. DOI: 10.1016/s0168-9525(02)02592-1.

Subramanian, Subbaya, Rakesh K Mishra, and Lalji Singh (2003). "Genome-wide analysis of microsatellite repeats in humans: their bla-bla abundance and density in specific genomic regions." In: *Genome Biology* 4.2, R13. DOI: 10.1186/gb-2003-4-2-r13.

Sudmant, Peter H et al. (2013). "Evolution and diversity of copy number variation in the great ape lineage." In: *Genome Research* 23.9, pp. 1373–1382. DOI: 10.1101/gr.158543.113.

Sudmant, Peter H et al. (2015). "An integrated map of structural variation in 2,504 human genomes." In: *Nature* 526.7571, pp. 75–81. ISSN: 0028-0836. DOI: 10.1038/nature15394.

Tautz, D and Schlötterer (1994). "Simple sequences." In: *Current Opinion in Genetics & Development* 4.6, pp. 832–837. DOI: 10.1016/0959-{437X}(94)90067-1.

Thakur, Jitendra, Jenika Packiaraj, and Steven Henikoff (2021). "Sequence, chromatin and evolution of satellite DNA." In: *International Journal of Molecular Sciences* 22.9. DOI: 10.3390/ijms22094309.

Thurman, Robert E et al. (2012). "The accessible chromatin landscape of the human genome." In: *Nature* 489.7414, pp. 75–82. DOI: 10.1038/nature11232.

Tomlinson, I M et al. (1994). "Human immunoglobulin VH and D segments on chromosomes 15q11.2 and 16p11.2." In: *Human Molecular Genetics* 3.6, pp. 853–860. DOI: 10.1093/hmg/3.6.853.

Trask, B J et al. (1998). "Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes." In: *Human Molecular Genetics* 7.1, pp. 13–26. DOI: 10.1093/hmg/7.1.13.

Turner, Kristen M et al. (2017). "Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity." In: *Nature* 543.7643, pp. 122–125. DOI: 10.1038/nature21356.

Tuzun, Eray et al. (2005). "Fine-scale structural variation of the human genome." In: *Nature Genetics* 37.7, pp. 727–732. ISSN: 1061-4036. DOI: 10.1038/ng1562.

Tóth, Katalin Fejes et al. (2016). "The piRNA Pathway Guards the Germline Genome Against Transposable Elements." In: *Advances in Experimental Medicine and Biology* 886, pp. 51–77. DOI: 10.1007/978-94-017-7417-8_4.

Ullmann, Reinhard et al. (2007). "Array CGH identifies reciprocal 16p13.1 duplications and deletions that predispose to autism and/or mental retardation." In: *Human Mutation* 28.7, pp. 674–682. DOI: 10.1002/humu.20546.

Varki, Ajit, Daniel H Geschwind, and Evan E Eichler (2008). "Explaining human uniqueness: genome interactions with environment, behaviour and culture." In: *Nature Reviews. Genetics* 9.10, pp. 749–763. DOI: 10.1038/nrg2428.

Vergnaud, G and F Denoeud (2000). "Minisatellites: mutability and genome architecture." In: *Genome Research* 10.7, pp. 899–907. DOI: 10.1101/gr.10.7.899.

Vissers, Lisenka E L M and Paweł Stankiewicz (2012). "Microdeletion and microduplication syndromes." In: *Methods in Molecular Biology* 838, pp. 29–75. DOI: 10.1007/978-1-61779-507-7_2.

Vollger, Mitchell R et al. (2019). "Long-read sequence and assembly of segmental duplications." In: *Nature Methods* 16.1, pp. 88–94. ISSN: 1548-7091. DOI: 10.1038/s41592-018-0236-3.

Waldman, A S and R M Liskay (1988). "Dependence of intrachromosomal recombination in mammalian cells on uninterrupted homology." In: *Molecular and Cellular Biology* 8.12, pp. 5350–5357. DOI: 10.1128/mcb.8.12.5350-5357.1988.

Weir, Barbara A et al. (2007). "Characterizing the cancer genome in lung adenocarcinoma." In: *Nature* 450.7171, pp. 893–898. ISSN: 1476-4687. DOI: 10.1038/nature06358.

Weirather, Jason L et al. (2017). "Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis." In: *F1000Research* 6, p. 100. DOI: 10.12688/f1000research.10571.2.

Wolfram, Research Inc. *Wolfram Alpha*. 2021.

Wong, Andrew et al. (2004). "Diverse fates of paralogs following segmental duplication of telomeric genes." In: *Genomics* 84.2, pp. 239–247. DOI: 10.1016/j.ygeno.2004.03.001.

Wong, Z, N J Royle, and A J Jeffreys (1990). "A novel human DNA polymorphism resulting from transfer of DNA from chromosome 6 to chromosome 16." In: *Genomics* 7.2, pp. 222–234. DOI: 10.1016/0888-7543(90)90544-5.

Young, W M and E W Elcock (1966). "Monte Carlo studies of vacancy migration in binary ordered alloys: I". In: *Proceedings of the Physical Society* 89.3, pp. 735–746. ISSN: 0370-1328. DOI: 10.1088/0370-1328/89/3/329.

Zhang, Liqing et al. (2005). "Patterns of segmental duplication in the human genome." In: *Molecular Biology and Evolution* 22.1, pp. 135–141. DOI: 10.1093/molbev/msh262.

Zhou, Yi and Bud Mishra (2005). "Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling." In: *Proceedings of the National Academy of Sciences of the United States of America* 102.11, pp. 4051–4056. DOI: 10.1073/pnas.0407957102.

Declaration of Authorship

I, Eldar Abdullaev, declare to Freie Universität Berlin that I have prepared this dissertation independently and without using any sources or aids other than those specified. This work is free of plagiarism. I have marked all explanations that are taken literally or in terms of content from other writings as such. This dissertation has not been submitted in the same or a similar form in any previous doctoral procedure.
I agree to have my thesis examined by a plagiarism examination software.

Signed:

Date: