HOW CHROMOSOMAL REARRANGEMENTS SHAPE GENOMES

JULIETTE LUISELLI

Under the supervision of

GUILLAUME BESLON NICOLAS LARTILLOT

A COMPUTATIONAL AND MATHEMATICAL STUDY.

Beagle Team
LIRIS
INSA Lyon & INRIA

April 2025

ABSTRACT

The origins of genome complexity, as well as the determinants of genome size, remain widely debated. This thesis shows that chromosomal rearrangements are a key factor in the evolution of genome architecture in terms of size and complexity. In particular, it shows that genome size and coding fraction are closely linked to the selection for robustness to chromosomal rearrangements, which is notably modulated by population size and mutation rate.

We first study the impact of chromosomal rearrangements on the evolution of bacterial genome architecture. To this end, the thesis relies on computer simulations and mathematical modeling. In particular, for the simulations, it relies on Aevol, a software designed to study prokaryote genome structure evolution, that allows for chromosomal rearrangements to act directly on the genomic sequence of individuals. Using Aevol, we are able to show that chromosomal rearrangements are essential for sustaining long-term adaptation, but also for stabilizing genome size. This result enables us to show, through large-scale simulation campaigns, that the pressure imposed by rearrangements on genome size is modulated by both mutation rate (which modifies genome robustness) and population size (which modifies the efficiency of selection for robustness). This result is then confirmed by a mathematical model showing how these two parameters determine an equilibrium proportion of non-coding genome.

The second part of the thesis focuses on generalizing the previous results to eukaryotic genomes. First, it presents a new version of Aevol developed specifically for the project that entails diploid organisms with linear chromosomes that reproduce sexually and undergo a mandatory meiotic recombination event. Using this model, we then show that eukaryote-like genomes react to changes in mutation rate and population size in the same way as prokaryote-like genomes. In the last chapter, we show that the reproductive mode is also an important determinant of genome architecture, as self-fertilization leads to more streamlined genomes.

Overall, this PhD thesis presents a new globally coherent and self-contained framework for understanding fundamental aspects of genome size evolution, focused on the direct and indirect impact of mutations, especially chromosomal rearrangements, and how they affect the future of each lineage. We also show how other parameters, such as the population size and the reproduction mode (asexual, sexual, self-fertilization), interact with these mutations and modulate their impact on genome size evolution. Taken together, these results

contribute to a unifying view of the evolution of genome architecture and complexity along the tree of life.

RÉSUMÉ

Les origines de la complexité des génomes, ainsi que les déterminants de la taille des génomes, restent largement débattus. Cette thèse montre que les réarrangements chromosomiques sont un facteur clé de l'évolution de l'architecture du génome en termes de taille et de complexité. Elle montre en particulier que la taille et la fraction codante des génomes sont étroitement liées à la sélection de la robustesse aux réarrangements chromosomiques, et que celle-ci est notamment modulée par la taille de la population et le taux de mutation.

Dans un premier temps, nous avons étudié l'impact des réarrangements chromosomiques sur l'évolution de l'architecture des génomes bactériens. Pour cela, la thèse s'appuie sur des simulations informatiques et des modélisations mathématiques. En particulier, pour les simulations, elle s'appuie sur Aevol, un logiciel conçu pour étudier l'évolution de la structure des génomes procaryotes, qui permet aux réarrangements chromosomiques d'agir directement sur la séquence génomique des individus. En utilisant Aevol, nous avons pu montrer que les réarrangements chromosomiques sont essentiels pour soutenir l'adaptation à long terme, mais aussi pour stabiliser la taille du génome. Ce résultat nous a permis de montrer, par des campagnes de simulation à grande échelle, que la pression imposée par les réarrangements sur la taille du génome est modulée à la fois par le taux de mutation (qui modifie la robustesse des génomes) et par la taille de la population (qui modifie l'efficacité de la sélection pour la robustesse). Ce résultat a ensuite été confirmé par un modèle mathématique qui met en évidence comment ces deux paramètres déterminent une proportion d'équilibre du génome non codant.

La deuxième partie de la thèse se concentre sur la généralisation des résultats précédents aux génomes eucaryotes. Tout d'abord, elle présente une nouvelle version d'Aevol développée spécifiquement pour le projet, qui modélise des organismes diploïdes avec des chromosomes linéaires se reproduisant sexuellement et subissant un événement de recombinaison méiotique obligatoire. Avec ce modèle, nous montrons que les génomes de type eucaryote réagissent aux changements du taux de mutation et de la taille de la population de la même manière que les génomes de type procaryote. Dans le dernier chapitre, nous montrons que le mode de reproduction est également un déterminant important de l'architecture du génome, car l'auto-fécondation conduit à des génomes réduits.

En résumé, cette thèse de doctorat présente un nouveau cadre permettant de comprendre les aspects fondamentaux de l'évolution de la taille des génomes, en se concentrant sur l'impact direct et indirect des mutations, et en particulier des réarrangements chromosomiques, et sur la manière dont elles affectent l'avenir de chaque lignée. Nous montrons également comment d'autres paramètres, tels que la taille de la population et le mode de reproduction (asexué, sexué, autofécondation), interagissent avec ces mutations et modulent leur impact sur l'évolution de la taille du génome. L'ensemble de ces résultats contribue à une vision unifiée de l'évolution de l'architecture et de la complexité des génomes le long de l'arbre de la vie.

ACKNOWLEDGMENTS

I want to thank not only the people who helped me complete this PhD thesis, but also all those who supported me in getting started on this incredible journey and contributed to making my world a better place, and myself a better person. Every step along the way has been important to me.

First and foremost, I want to thank Guillaume for welcoming me into research with open arms, sharing his endless passion for knowledge, and showing me that there is no need to grow up to become a researcher. I can never express how grateful I am. Then, I thank Nicolas for his steady guidance and for keeping me on track.

I extend my thanks to my reviewers, Eduardo Rocha and Céline Scornavacca, for their time and very encouraging remarks. My thanks also go to Christine Dillmann and Guillaume Achaz for agreeing to be part of my jury and evaluate my work.

On a more personal note, I also want to thank:

- Paul, for putting me at ease early on in the office, for his invaluable scientific contribution to several chapters of this thesis, for the countless breaks, deep thoughts, board games, and video games we shared, and for bringing more greenery into my world.
- Jonathan, for his key help in navigating heavy computations, his cats, and his advice.
- Lisa C, Romain, Sofia, Thibaut, Nathan, Charlotte, Arsène, Arnaud, Théotime, Marco, Julie, Lisa B, for their camaraderie, for making our lunches and gatherings enjoyable with shared random facts, and for always taking care of each other.
- The Beagle Team as a whole, past and present, for creating a great working atmosphere.
- Basile, for proofreading parts of this manuscript and for taking care of my two cats, which allowed me to travel to Canada.
- Manuel and the Lab-fond, for welcoming me there and sharing lunch with me despite the apparent oddity of the concept to North American people.
- Diala, for welcoming me as a peer and sharing her enthusiasm.
- James, Isaac, Paulien, Sam, Ivan, Flora, and Olivier, for their support and faith in my early research endeavours, and all their useful advice.

- Marie, for having supported me through the most terrible stages of our teenage years and still being there for me every day.

- Flore, for her contagious motivation and sports drive, and her exemplary character that brings the best out of me.

- Guillaume and Benoît, for ensuring that I will always remember the prépa with a giant smile.

- Sarah, Manon, Ludo, the Ernestophone, the DG, the Turlurons, and the K-Fêt for making me love the ENS, and for making me leave it behind me when it was time.

- Ms Prévier, Ms Piquet, M. Huet, M. Kameche, Ms Lepage, Ms Thor, Ms Butigieg, and all the teachers and associates who pushed me to give the best of myself, believed in me, and supported me when I needed it.

- My volleyball teammates and my psychologists, for keeping me afloat all these years.

- My parents, for providing me the means to undertake my scientific journey, and for never questioning my career choices.

- Yris and Moccha, for their clueless love and unknowing support.

- My husband, Valentin, for reading this thesis in its entirety and providing valuable feedback, and above all for his unwavering faith, love, and support throughout everything.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

CR chromosomal rearrangements only – duplications, deletions and inversion

CRLM chromosomal rearrangements and local mutations

DFE Distribution of Fitness Effects

LM local mutations only – substitutions and InDels

MHH Mutational Hazard Hypothesis

SNP Single Nucleotide Polymorphism

SUB substitutions only

TE Transposable Element

1 INTRODUCTION

1.1 THE PARADOXES OF EVOLUTION

Evolution is the process by which populations and species change over time: as mutations create heritable variation, natural selection — or chance — can fix some of this variation. As such, evolution may seem easy to understand: some individuals have an innate advantage over others and are better equipped to survive and reproduce, overtaking the place of less adapted ones. Most high school students have heard about the evolution of peppered moths during the Industrial Revolution (Cook et al., 2012) and how the melanin phenotype progressively replaced the white phenotype due to it being a better camouflage from predators on soot-blackened trees.

While this evolutionary adaption seems like common sense, evolution often yields unexpected and counterintuitive results that our minds struggle to grasp and understand. Indeed, evolution is the result of only three factors: heritable variation (mutations), selection, and drift — *i.e.* random change in allele frequency due to chance. However, each of these three factors is very complex and counterintuitive.

First, there are many types of mutations (substitutions, recombinations, short insertions and deletions, large chromosomal rearrangements, etc.), and they are not always straightforward to detect and study. Mutations have a gigantic combinatorics, such that the space of accessible genotypes is huge, and an unexpected innovation is virtually always possible. This has been recently illustrated by Banse et al. (2024a), showing that even in a constant environment and with already adapted organisms, chromosomal rearrangements can occasionally bring new innovations. Second, selection is a continuous process over time and generations: at a given generation, some individuals are more susceptible to successfully reproduce than others, but their offspring have to also be able to reproduce to avoid evolutionary dead-ends. As such, there is selection for phenotypical adaptation, but there can also be sexual selection pushing in an opposite direction and second-order selection for robustness or evolvability (Wilke et al., 2001; Wagner, 2008; Liard et al., 2020). Third, drift also challenges our common sense, since deleterious mutations can actually reach fixation. The conditions under which this happens are complex and require a rigorous theoretical framework. Indeed, mean field approximations are not sufficient to understand how mutations can reach fixation in a population, and population genetics relies on complex mathematical models (Charlesworth and Charlesworth, 2017). Finally, these phe-

nomena act on vast timescales that are very distant from our human understanding of time. Thus, many remain in awe at the perfection of the human eye — often while wearing glasses — or are surprised that our chickens are the distant relatives of mighty dinosaurs. The pace at which species and organs evolve is challenging to grasp.

As a result, evolution is not a straightforward process: it pushes in opposing directions and constantly challenges our intuition. Evolution even probably shaped the human mind to make us think that evolution is wrong: cognitive biases are very helpful to make quick decisions and improve survival, but they make it more complicated for us to stop and think, and to apprehend complex reasoning. Consequently, deciphering the effects of evolution requires rigorous modeling and reasoning

More details in the book L'ironie de l'évolution, *by Thomas C. Durand*

As biologists delved deeper into the study of genomes, paradoxes and apparent inconsistencies accumulated, keeping them unsettled for years. First, with the discovery of DNA (deoxyribonucleic acid) and its first measurements emerged the C-value paradox (Thomas, 1971): why is the apparent complexity of an organism not related to the amount of DNA, which supposedly represents the amount of information it bears? This paradox was first resolved by pointing out that, in eukaryotes at least, the genome is mostly non-coding, so the genome size is not correlated with the number of genes — which would be the true measure of quantity of information. But then, as sequencing technologies advanced and after the completion of the first complete human genome sequence in 2001, the G-value paradox rose (Hahn and Wray, 2002): why is the number of genes not correlated with complexity either? How can humans have approximately the same number of genes as *C. elegans*? It turns out, complexity is not easy to define in biology (Adami, 2002), and humans might not be the most perfect, most complex product of evolution — that would obviously be the cat.

Other paradoxes are still under discussion, such as Lewontin's Paradox (Lewontin, 1974; Buffalo, 2021; Charlesworth and Jensen, 2022), which states that the levels of DNA sequence variation in natural populations are much lower than what would be expected given their known population sizes. More specifically, quantifying genetic drift is still a difficult matter. While many proxies can be used, such as the rate of coalescence, the level of inbreeding, or the variance in the number of offspring (Waples, 2022), each probably gives different information on a given population and should be interpreted carefully. More specifically, quantifying drift is still a difficult matter. Its measure is the effective population size (N_e), defined as the size of an idealized population that would experience the same level of genetic drift. However, the debate on ways to estimate N_e and the information actually provided by each estimator is still ongoing (Waples, 2022). Therefore, both genome and population studies show the need for solid

theoretical work accounting for these complex and counterintuitive phenomena.

These paradoxes illustrate how biology, and particularly evolution, can be difficult to grasp. It makes us think against ourselves and question pre-established beliefs or common-sense knowledge to reach new understandings of complex phenomena. To me, that is the beauty of evolution: it teaches us a disciplined way of thinking that we should impose on all our interactions with the outside world.

In this introductory chapter, I will first describe the main concepts of evolution (mutation, selection, and drift), then turn to genome studies and address the evolution of genome architecture. Finally, I will present chromosomal rearrangements as a major force of genome architecture evolution and discuss the necessity of models to study them and their impact on genome architecture.

1.2 MUTATIONS, SELECTION AND DRIFT, THE THREE FORCES OF EVOLUTION

Evolution is the product of variation in the heritable genomic information through mutations, and fixation of variants through natural selection (when adaptive) or drift.

1.2.1 *Genetic variations*

Genetic variation can occur through many mechanisms. First and foremost, mutations alter the DNA sequence. They can be local (a substitution changes just one letter of the sequence, a small insertion or deletion adds or removes a few nucleotides by polymerase slippage), or very large (chromosomal rearrangements can invert, delete, or duplicate whole segments of the DNA sequence, following double-strand breaks of the DNA). Mutations can also change the way the sequence will be read, without altering directly the sequence itself: these are called epigenetic mutations. For example, a nucleotide can be methylated, which typically will prevent gene expression in this region. While genomic studies have often been restricted to the study of polymorphism, *i.e.* to local mutations, this thesis will focus on chromosomal rearrangements, which are susceptible to bringing very large variations to populations.

Other mechanisms also contribute to increasing the genetic variation in a population, such as sexual reproduction — or any exchange of genetic material between individuals —, and recombination. While homologous recombination does not create new mutations, it reassorts existing mutations and can thus form a new genotype that was not present in the population. Additionally, illegitimate recombinations are a source of chromosomal rearrangements, *i.e.* mutations.

1.2.2 *Natural selection*

Given that several genotypes coexist in a population, some of them can be linked to more adapted phenotypes and hence have a greater chance of reproducing and getting fixed in the population: that is the process of natural selection. It can be positive — fixing mutations that bring an advantage —, or negative — removing deleterious mutations from the population. A selective advantage is not necessarily a better phenotypical adaptation to the environment in the sense of a better capability to survive, it can also entail differences that hinder survival but enhance reproductive success through sexual selection — as would, for example, be the case for the tail of the Indian peafowl. Additionally, selection acts upon several generations, and there can be second-order selective effects: a mutation granting a huge reproductive success but that yield only sterile offspring would not be selected ultimately. An individual must not only be adapted itself, but also be able to produce offspring that are at least equally adapted. The evolutionary race of species is not a sprint but a marathon.

As such, selection can act in different directions at once and its result can be profoundly counter-intuitive.

1.2.3 *Genetic drift*

Finally, non-adaptive variations can also get fixed in the population by chance — that is the process of genetic drift. It depends on the population size: the smaller the population, the more probable it is for a neutral or slightly deleterious mutation to go to fixation. Population genetics show that, in an idealized panmictic haploid population with only neutral mutations, each mutation has a probability $\frac{1}{N}$ to be fixed in a population of size N.

In practice, populations generally do not follow an idealized model. Their level of genetic drift can be compared using the effective population size N_e, which is the size of a population following a Wright-Fisher model (Fisher, 1923; Wright, 1931) that would display the same evolutionary characteristics and, in particular, the same level of genetic drift. However, N_e is an abstract value that changes over time (Brevet and Lartillot, 2021), and it cannot be measured directly but only approximated with different proxies such as the standing genetic variation or the coalescence time (Wang, 2005; Waples, 2010, 2022). It can be influenced by many factors, such as the population structure, mutation rate, or environmental changes, that all change the probability of fixation of new mutations. Computational simulations and mathematical modelling allow us to distance ourselves from the problem of measuring N_e to focus exclusively on the relationship between N_e and genome evolution.

These three forces together shape evolution, and in particular genome architecture evolution. To understand how genomes evolve, we will now look at how information is structured on them.

1.3 STUDYING GENOME: A MULTISCALE PROBLEM

Genomes are central to the study of biology and evolution. They are the carriers of information, evolving through mutations, selection, and drift. Their multiscale nature makes them complex to study: while part of the information they carry is directly encoded in genes, the way genes are distributed along the genome or the varying accessibility of different parts of the genome is also a piece of information that can evolve (Smolke and Keasling, 2002; Holder and Hartig, 2014). Two genes sharing a promoter are necessarily transcribed and expressed together, as is the case in prokaryotic operons (Koonin, 2009). Genes on the same eukaryotic chromosome have a higher chance of being transmitted together, and their different alleles can be more or less linked with one another (Ardlie et al., 2002). As such, genomes are multiscale and encapsulate information in their sequence, but also in their structure and macro-organization. While all genomic scales and their interactions are worth being studied in detail, they received varying interest from the scientific community over the years.

1.3.1 *A historical chromosome scale*

Before DNA sequencing, genome structure and mutations could be investigated by looking at karyotypes and the number of chromosome pairs (see Figure 1.1). This has been studied in detail at the beginning of the 20th century, first highlighting mutations in the number of chromosomes (Bridges, 1916, 1921). While those are caused by a simple non-disjunction of chromosomes at replication, more complex chromosomal rearrangements involving sections of chromosomes were also discovered: translocations (Dobzhansky, 1930), inversions (Dobzhansky and Sturtevant, 1938), duplications (Morgan, 1938), and deletions (Rick, 1940).

As such, the chromosome scale was already informative and gave important insights into genetics (Haldane, 1936). Yet, possibilities were limited without a finer understanding of genetic information. Moreover, most chromosomal rearrangements are lethal, which prevents them from being studied: only very few of them could be documented at the time. Sequencing technologies opened a wide range of new opportunities and, unsurprisingly, shifted the research focus towards DNA sequence

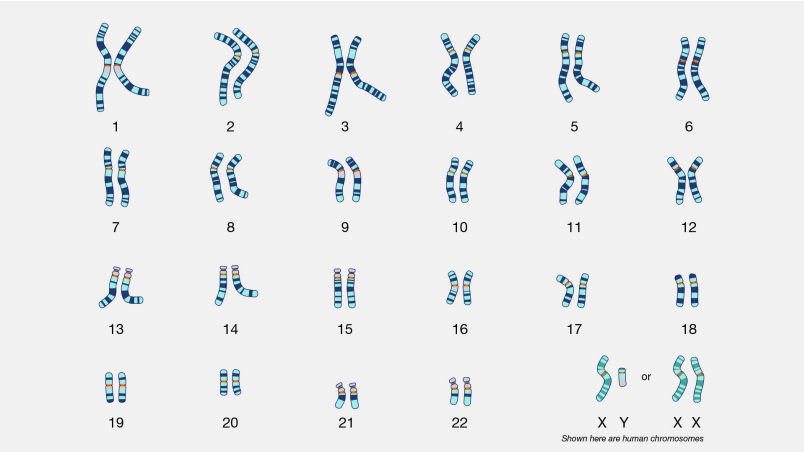

Figure 1.1: Example human karyotype. Retrieved March 3, 2025 from the "Talking Glossary of Genetic Terms.", National Institute of Health.

1.3.2 *The sequencing revolution*

Sequencing DNA first allowed the study of parts of genomes, a few bases at a time. One founding technique that became rapidly widespread is the Sanger sequencing method (Sanger et al., 1977). It allows detecting Single Nucleotide Polymorphisms (SNPs) in genes (Kreitman, 1983; Ravetch and Perussia, 1989), and it has been widely used until the 21^{st} century, drastically changing how biology was studied. However, it is limited to short DNA sequences (below 1,000 bases), which kept the focus on genes and alleles, overlooking non-coding DNA and genome architecture.

At the beginning of the 21st century, the first complete human genome (IHGSC, 2001) and next-generation sequencing methods again revolutionized biology (Schuster, 2008) by enabling the accumulation of more and more data on species' genomes and their phylogenies, variability within populations, or the discovery of new species from metabarcoding data (Coissac et al., 2012; Deiner et al., 2017). New sequencing technologies (long-read sequencing, nanopore, etc.) facilitated genome assemblies and the study of gene distribution along the chromosomes in addition to the gene sequences.

1.3.3 *Genome architecture scale*

While initial efforts were focused on increasing the speed and reducing the cost of gene sequencing (development of high-throughput sequencing techniques), a new and more recent focus has been on elongated reads (Schadt et al., 2010), facilitating genome assembly. These new technologies allow a more extensive study of genome architecture and notably repeated non-coding DNA or duplicated sequences (Treangen

and Salzberg, 2012; Lin et al., 2021; Liao et al., 2023). The study of genome architecture is thus still a rising field

Part of the renewed interest in the organization of genomes has been on the so-called junk DNA (Ohno, 1972; Doolittle, 2013; Palazzo and Gregory, 2014; Fagundes et al., 2022): the actual content, determinants, and evolutionary origins of non-coding DNA are still debated (Ahnert et al., 2008; Gil and Latorre, 2012), and yet it builds up a large part of genomes. In most eukaryotes, the largest part of non-coding DNA is composed of Transposable Elements (TEs) (Wessler, 2006). TEs are mobile selfish DNA sequences that are — or were — able to copy/-paste themselves within genomes. They are abundant in mammalian genomes and especially in the human genome (around 45% of the human genome is composed of TEs (IHGSC, 2001)). TEs seem to drive genome size with their number (Elliott and Gregory, 2015; Marino et al., 2024) . They also shape genome architecture as they are not uniformly distributed along the genome (Quesneville, 2020), although the forces shaping their distribution are still being debated (Sultana et al., 2017; Langmüller et al., 2023).

Another important force shaping genome architecture, and probably itself influenced by genome architecture, is the distribution of recombination breakpoints. Recombination events break blocks of linkage, *i.e.* alleles that are clustered together. The distribution of recombination points, also called the recombination landscape, is not random and varies between species (Zelkowski et al., 2019). That distribution can also evolve: some mutations modify it (as the *rec-1* loss of function in *C. elengans* (Parée et al., 2024)), and this variation could be selected (Parée et al., 2025). Recombination events can also be illegitimate — *i.e.* between non-homologous regions — and thus provoke changes in gene distribution and genome content through inversions, duplications, or deletions. An inversion of part of a chromosome not only changes the sequence around its breakpoints, but it also changes the wider vicinity of genes, sometimes with unexpected effects. For example, it could be deleterious if it breaks a gene's sequence, but also advantageous if it enhances the transcription of another gene by changing its location on the sequence.

The distribution of content in the genome is very important, as is its physical location in the nucleus, since it can also impact its transcription and is heavily regulated (Bickmore and Van Steensel, 2013). Indeed, the chromatins of different genome sections can interact physically together or, in eukaryotes, with other elements of the nucleus (Pombo and Dillon, 2015), and each chromosome has a specific location inside the cell. As such, there are several scales of genome organization: information within the sequence, neighborhood on the sequence, and physical neighborhood in the cell or nucleus.

In short, genome architecture study is a very wide field, and many aspects of genome organization can be studied. They all are in interaction with one another, as gene distribution could influence the distribution of recombination breakpoints and linkage blocks, TEs could drive illegitimate recombinations or phenotypic changes (Schrader and Schmitz, 2019) or accumulate in recombination-poor regions (Kejnovsky et al., 2009; Kent et al., 2017), etc. Genome architecture should also be studied in interaction with the sequence scale of genome information since, for example, inversions change both the macro-distribution of genes and the sequence around the breakpoints. More generally, mutations affecting the genome architecture also bring local modifications to the DNA sequence, and *vice versa* as any local mutations could change the affinity of a sequence with other parts of the genome. In addition, genome architecture influences the range of possible effects of mutations, as it influences the risk of a mutation to affect a gene, as well as the probability of a mutation to occur. Thus, there are numerous complex interactions between the genome architecture scale and the sequence scale.

Due to the complexity and intricacy of genome architecture descriptors, a theoretical framework to study genome architecture evolution should first focus on a few main variables. In this thesis, I will focus on the coding and non-coding sizes of genomes. They can be directly linked to the total genome size and coding fraction of genomes, and they are powerful descriptors of the wide diversity of genome architectures, as shown in the following section.

1.4 THE DIVERSITY OF GENOME ARCHITECTURES

Even when restricting genome architecture studies to the study of coding and non-coding genome sizes, there is a wide diversity of genome architectures in the Tree of Life, ranging from very small and compact genomes to large expanded genomes with only a low coding fraction. This has been notably reviewed by Koonin (2009). The most obvious dichotomy between two majorly different types of genome architecture seems to be between eukaryotes and prokaryotes.

1.4.1 *Prokaryotes and eukaryotes, differences and similarities*

Prokaryotes (archaea and bacteria) and eukaryotes diverged around two billion years ago (Craig et al., 2023), after around 1.7 billion years of common evolution (Ohtomo et al., 2014). This means they share a common ancestor and many similarities — such as the standard genetic code and the usage of DNA, RNA, and proteins. They all grow, disperse, reproduce, and mutate. However, they also had much time to evolve since their phylogenetic divergence, and eukaryotes

and prokaryotes display distinct genome architectures and apparent complexity (multicellularity, tissue differentiation, etc.).

Prokaryotes and eukaryotes display unique and different genome organization features (Koonin, 2009). Prokaryotes' genomes are organized in operons, *i.e.* groups of co-transcribed genes often encoding for interacting proteins. They are generally short in total genome size and have a high percentage of coding regions. In contrast, eukaryotes' genomes are generally big, with a high fraction of non-coding genome, and they generally do not have operons: RNAs are monocistronic and are rarely grouped by function. They also display a very conserved feature: the presence of introns that fragment protein-coding genes, which are essentially absent in prokaryotes.

At least some of these broad differences could be explained by a founder effect at their divergence or their discrete differences in characteristics. Indeed, many of the eukaryote/prokaryote differences could have a considerable impact on their genome evolution. Eukaryotes undergo meiotic recombinations, during which they provoke double-strand breaks in their genomes (Cao et al., 1990), reassorting existing mutations but also triggering new ones (Arbeithuber et al., 2015). Most eukaryotes also have proper sexual reproduction, which allows TEs to easily colonize new genomes and which thus changes their dynamics. Finally, some authors proposed that there is an energetic barrier to genome complexity that could explain the fundamental differences between prokaryotes and eukaryotes (Lane and Martin, 2010; Craig et al., 2023), although this remains heavily disputed (Lynch and Marinov, 2017; Chiyomaru and Takemoto, 2020).

While these profound differences between eukaryotes and prokaryotes could explain part of the difference in their genome architectures, it must not be forgotten that the genome architectures of these two clades also display a significant overlap. Some prokaryotes have self-splicing intron-like structures, while some eukaryotes are virtually intron-free. As we restrict our study to the coding and non-coding genome sizes, it is also worth noting that some eukaryotes are smaller and have fewer genes than some prokaryotes. Indeed, there is a large variation in genome size and coding fraction within prokaryotes and eukaryotes, hinting that the discrete differences between prokaryotes and eukaryotes cannot entirely explain the differences in genome size and density. As such, other hypotheses have been proposed to explain the evolution of genome size, as exposed in the next section.

1.4.2 *Evolutionary causes of genome architecture evolution*

As stated, genomes evolve through mutations, selection, and drift. Any of these factors could largely affect genome architecture evolution, and particularly genome size and coding density.

Adaptive hypotheses

Selection can act on genome size evolution: according to adaptive hypotheses, genome size is a trait under selection. It would be limited to increase the replication efficacy (Kang et al., 2015; Malerba et al., 2020), and is also tightly linked to phenotypical characteristics, such as cell and nucleus size (Knight and Beaulieu, 2008), that could be under direct selection. Yet, the adaptive hypotheses struggle to explain the vast diversity of genome sizes and coding densities, as they tend to narrow the range of optimal sizes and have arguably little empirical support (Lynch, 2007a). Yet, the neutralist/selectionist debate is still ongoing (Galtier, 2024).

Mutational mechanisms

Mutational patterns undoubtedly have a substantial impact on long-term genome architecture evolution. The Mutational Equilibrium Hypothesis (MEH) (Petrov, 2001) proposes that mutational biases in opposite directions — towards deletions for short indels and insertions for larger events — could determine an equilibrium genome size. More generally, variations in underlying mutational biases between species could account for the observed diversity in genome sizes. In particular, eukaryotes are prone to TEs invasions and, therefore, have a strong insertion bias (Ratcliff, 2024), coherent with their larger genome sizes. More generally, differences in mutation mechanisms and frequencies can push genome size evolution in one direction or another (Kuo et al., 2009; Kuo and Ochman, 2009; He et al., 2019; Loewenthal et al., 2022).

The mutational hazard hypothesis

Another central hypothesis of genome size evolution is that of the increase of genome size through genetic drift: the Mutational Hazard Hypothesis (MHH) (Lynch and Conery, 2003; Lynch, 2006b, 2007b). It postulates that any increase in genome complexity — for example, through introns, an added layer of gene regulation, or duplicated genes — is inherently dangerous as it increases the number of targets for deleterious mutations while keeping the same phenotype (Lynch, 2006b). Thus, genome size increase would be caused by a range of slightly deleterious mutations that can go to fixation through genetic drift. With population genetics arguments, we can conclude that prokaryotes are protected from substantial genome complexification — thus from major increases in genome size — thanks to their very large population sizes, hence low genetic drift. On the other hand, eukaryotes would have entered a complexity ratchet and are stacking non-adaptive complexity because their population sizes are too low and the purifying selection is not strong enough. In this view, there is a continuity between the streamlined prokaryotes, the bacteria with

large genomes, the unicellular eukaryotes, and the animals and plants (Koonin, 2009). Although proposed almost 20 years ago, the MHH has rarely been tested empirically. While some data seem to support it (Yi and Streelman, 2005; Kelkar and Ochman, 2012; Smith et al., 2013), others dispute it (Ai et al., 2012; Mohlhenrich and Mueller, 2016; Marino et al., 2024). Indeed, as it relies on the comparison of very different clades, the signal is relatively low once phylogenetic inertia is accounted for.

Finally, genome size evolution mechanisms based on mutations, selection, or drift are probably not mutually exclusive, further complicating the picture. For example, drift could lead to genome size reduction instead of genome expansion in prokaryotes (Bobay and Ochman, 2017), diminishing the role of the different population sizes of prokaryotes and eukaryotes as an explication factor for their differences in genome architecture.

The effect of population genetics mechanisms on genome architecture evolution (Lynch and Conery, 2003; Lynch, 2007b), and their interaction with mutational pressures (Schaack, 2006; Kuo et al., 2009; Kuo and Ochman, 2009), have been extensively studied. Yet, major mutational operators that directly change the genome structure have been less studied: chromosomal rearrangements.

1.5 CHROMOSOMAL REARRANGEMENTS AS KEY MUTATIONAL EVENTS OPERATING ON THE GENOME STRUCTURE

1.5.1 *State of the art on the study of chromosomal rearrangements*

While genome architecture and its evolution have been quite extensively documented (Lynch, 2007b; Koonin, 2009), a crucial determinant of genome structure has been widely overlooked: chromosomal rearrangements — also called structural mutations. They refer to mutations larger than 50 base-pairs that generally act on a segment of sequence by duplicating, inverting, or deleting it. Chromosomal rearrangements clearly change the structure of the genome, and yet their long-term impact on genome architecture in various conditions is mostly uncharted territory — despite rearrangements being the first observed mutations at the genomic level (see Section 1.3.1). The sequencing era relegated chromosomal rearrangements to the background due to the difficulty of detecting them and the few available variations to study: most rearrangements are vastly deleterious or lethal and are thus invisible in the genomic data. Consequently, genomic research has largely focused on substitutions and their combinations through recombinations.

Nevertheless, chromosomal rearrangements are ubiquitous throughout biological organisms: viruses, bacteria, unicellular eukaryotes, and multicellular life. Nowadays, long-read sequencing enables one to

detect structural variations between sequences more accurately (Guan and Sung, 2016; Ho et al., 2020; Ahsan et al., 2023; Chen et al., 2025), and interest in their study is rising worldwide (Mérot et al., 2020; Augustijnen et al., 2024). It has also been discovered that chromosomal rearrangements are much more frequent than previously thought: there could be as many, if not more, rearrangements as base pair substitutions in prokaryotes (Wei et al., 2018; Molari et al., 2025). Chromosomal rearrangements are also common in eukaryotes, reaching at least 10% of the per-base substitution rate (Weissensteiner et al., 2020; Saxena and Baer, 2025) — not accounting for lethal rearrangements that would not allow a cell to survive and be sequenced. Yet, chromosomal rearrangements are still absent from most models, and there is little to no theoretical understanding of their impact on genome evolution (Mérot et al., 2020). They are already known to impact genome evolution through, for example, the loss of synteny due to inversions in prokaryotes (Eisen et al., 2000; Koonin, 2009) or through long-term adaptation (Trujillo et al., 2022), but a cohesive and more general study integrating their impact on genome architecture evolution is lacking.

1.5.2 *How chromosomal rearrangements shape genomes*

Chromosomal rearrangements display unique properties that make them particularly challenging to study and theorize. First, they can be very large, changing wide parts of the DNA sequence. This forbids assuming a constant genome size, a neutral effect of mutations, or other simplifying hypotheses. Moreover, the space of potential mutations is huge, as virtually any size of mutation is possible between any two base pairs on the genome. Consequently, an extensive study of potential chromosomal rearrangements is not possible — especially since most of the possible rearrangements are potentially very deleterious, if not lethal, preventing studying them in lineages or living organisms. As a matter of fact, chromosomal rearrangements are scarce in lineages, which makes it difficult to generalize behaviors and impacts on evolution from observations.

Finally, chromosomal rearrangements are highly dependent on the history. Contrary to substitutions, they are not commutative (Trujillo et al., 2022). Indeed, the order in which inversions occur can change the final sequence and, even more obviously, it is not possible to duplicate a segment that has been previously deleted. Due to these difficulties, chromosomal rearrangements pose a real conceptual challenge, and we need modeling tools to grasp how they affect genome evolution (Mérot et al., 2020) and, more specifically, how they impact genome architecture evolution.

1.6 MODELING CHROMOSOMAL REARRANGEMENTS AND THEIR IMPACT ON GENOME STRUCTURE EVOLUTION

The very properties of chromosomal rearrangements that make them hard to theorize also make them hard to model: they can widely change the genome in a few events, hence also changing the probability and the effect of future events taking place. Contrary to substitutions, they are not commutative (Trujillo et al., 2022), and it is also impossible to model them backward in time along a coalescence process; they can only be modeled forward in time. Finally, their exact impact on fitness and evolution is *a priori* unknown, and potential models cannot easily rely on existing theories. Consequently, while several modeling frameworks dedicated to the study of structural variants exist (Lei et al., 2022), most of them focus on detecting rearrangements in data or on testing algorithms developed to detect rearrangements. As such, they do not model the long-term effect of structural variants or their interaction with fitness and drift (Bartenhagen and Dugas, 2013; Mu et al., 2015; Qin et al., 2015; Xia et al., 2017).

Other evolutionary simulators not focused on structural variants are used to study genome evolution more broadly, such as Avida (Adami, 2006) or SLiM (Haller and Messer, 2023). Both have been used to study genome structure evolution. For example, Avida has been used to study gene overlapping (Gerlee and Lundh, 2008) or the evolution of genome size (Gupta et al., 2016). Avida implements complex mutational operators that could, to some extent, be compared to chromosomal rearrangements, but the genome structure of Avidian organisms is very difficult to compare with biological genome structure. The genome of an Avidian is a computer program, and its phenotype is the execution of the program. Some parts of the genome are skipped when located after a `jump` instruction, hence, they do not carry any phenotypic information, as they are never read when the program is executed. However, the "non-coding" nature of these genome parts is not intrinsic but encoded in the preceding coding part. This is more comparable to a protein sequence, where certain parts are folded and not in contact with the substrate, than to junk DNA. In Avida, no mutation within a "non-coding" part can make it coding, and thus the Mutational Hazard Hypothesis (MHH) cannot be easily tested in this framework.

SLiM is quite different as it focuses on biological realism and allows testing a wide variety of evolutionary scenarios (Haller and Messer, 2023) (with population structure, continuous space, context-dependent selection, etc.). SLiM can model an explicit genomic sequence, with coding and non-coding parts, on which mutations will accumulate forward in time. However, the genome structure in itself cannot evolve as mutations are point mutations with a pre-defined distribution of fitness effects. While there are chromosomal inversions, SLiM does not

model duplications or deletions, and a non-coding part of the genome cannot become coding and *vice versa* (Haller and Messer, 2024).

One of the very few models that do account for chromosomal rearrangements is Aevol — see www.aevol.fr—, a forward-in-time artificial evolution platform developed by the Beagle Team in Lyon (France). It proposes an explicit genome sequence, on which RNAs and genes can be identified and the latter decoded into proteins. This complex genotype-to-phenotype map allows for great liberty in the evolution of genome structure: the coding and non-coding parts of the genome can evolve freely. Duplicating a promoter would create a new RNA, turning non-coding bases into coding bases, while mutating an existing one could turn the following coding bases into non-coding bases. Mutations happen directly on the sequence, without any predefined Distribution of Fitness Effects (DFE): they can destroy existing genes, duplicate or mutate them, change their expression, create new genes *de novo*, or just change the content of the non-coding genome without any impact on the phenotype. More details on the functioning of the model and its parameters can be found in Appendix A.

The distinction between the genotype and the phenotype — and the possibility for any given genotype to be interpreted as a phenotype — allows for complex mutational operators such as chromosomal rearrangements: they act on the sequence, and the new sequence is then decoded to compute the new phenotype. That phenotype can then be selected if it is associated with a good fitness: Aevol is individual-based, each individual having its own genome — hence its own phenotype and fitness value — and they compete with one another to populate the next generation. As a result, complex evolutionary behaviors that take into account the population dynamics can emerge, *e.g.* genome size can respond to various parameters such as the population size, the mutation rates of the different types of mutations, or the environment. As a variation in the genome architecture, *e.g.* induced by a chromosomal rearrangement, changes the range of possible future mutations and their Distribution of Fitness Effects (DFE), there can be complex feedback loops influencing the evolution of genome architecture, which is itself under selection.

While Aevol has already been used to study genome structure evolution (Knibbe et al., 2007a,b; Liard et al., 2020), or chromosomal rearrangements in themselves (Banse, 2023), the specificities of the interactions between chromosomal rearrangements and genome architecture were yet unexplored.

As such, this thesis uses Aevol to model and form theories around the impact of chromosomal rearrangements on genome architecture evolution. The aim is to characterize the evolutionary forces imposed on genome evolution through the mere presence of chromosomal rearrangements.

1.7 OVERVIEW OF THE THESIS

In this PhD thesis, I will use computational and mathematical modeling to study how chromosomal rearrangements determine genome architecture evolution and how their impact is modulated by different parameters.

The first part of the thesis will focus on prokaryotes-like genome evolution. Chapter 2 presents a simple experiment with Aevol, highlighting the importance of chromosomal rearrangements for genome architecture evolution. We compare evolutionary trajectories with and without chromosomal rearrangements and show that these mutations are essential to sustain long-term adaptation as they reduce the effect of diminishing return epistasis. Chromosomal rearrangements also significantly impact genome architecture evolution, as the genome size grows indefinitely in their absence, while their presence enables a stabilization of the genome size. As such, chromosomal rearrangements impact both the phenotypical adaptation and the genome structure evolution. Chapter 3 specifically focuses on the effect of chromosomal rearrangements on genome size evolution. Experiments in Aevol with different mutation rates and/or population sizes reveal that these mutations impose a robustness cost to the genomes: as bigger genomes are more prone to rearrangements, they are less robust. Selection for robustness to chromosomal rearrangements thus governs genome size evolution. Since robustness selection is driven by both the population size (determining the efficacy of selection) and the mutation rate (influencing the robustness cost of each additional base pair), this result opens the way to a formal link between these two parameters and the evolution of genome structure. Following this, Chapter 4 presents a mathematical model we derived from what we understood based on previous simulation experiments. This model links chromosomal rearrangements, robustness selection, and genome size evolution. It generalizes our results beyond Aevol and formally demonstrates how chromosomal rearrangements can set an equilibrium non-coding genome fraction. While the mathematical model focuses on prokaryote-like genomes, we attempted to apply it to eukaryote-like ranges of parameters. This yields coherent results, which raises the question of the generalization of our observations on selection for robustness to chromosomal rearrangements to eukaryotes.

The second part of the thesis will, therefore, turn to eukaryote-like genome evolution. To this end, Chapter 5 presents a new version of Aevol developed specifically for the thesis, which is closer to eukaryotes: this version of the simulator entails diploid organisms with linear chromosomes that reproduce sexually and undergo a mandatory meiotic recombination event (as opposed to haploid organisms with one circular chromosome that reproduce asexually). Chapter 6 then uses this new model to run a set of experiments similar to the ones of

Chapter 3 within a eukaryotic framework. It shows that eukaryote-like genomes respond to changes in population size or mutation rate in the same way as prokaryote-like genomes. This confirms that the mathematical model of Chapter 4 can be applied both to prokaryote-like and eukaryote-like values of population sizes and mutation rates and yield meaningful results. Finally, Chapter 7 studies the impact of the selfing rate on genome architecture evolution and shows how the reproductive mode affects the evolution of non-coding genome size and its variability both within and between populations.

To conclude, this PhD thesis presents a new globally coherent and self-contained framework that makes it possible to understand fundamental aspects of genome size evolution. It focuses on the direct and indirect impact of chromosomal rearrangements and how they affect the future of a lineage. It also shows how many parameters, such as the population size, the mutation rates, and the reproduction mode (asexual, sexual, selfing), interact with these mutations and modulate their impact on genome size evolution. Altogether, these results contribute to a unifying view of the evolution of genome architecture and complexity along the Tree of Life.

Part I

PROKARYOTES

"Animals may be evolution's icing, but bacteria are the cake." (Knoll, 2015)

2

FORWARD-IN-TIME SIMULATION OF CHROMOSOMAL REARRANGEMENTS: THE INVISIBLE BACKBONE THAT SUSTAINS LONG-TERM ADAPTATION

FOREWORD

The following work is published in *Molecular Ecology* (Banse et al., 2024b) and authored by Paul Banse and Juliette Luiselli (co-first authors), David P Parsons, Théotime Grohens, Marco Foley, Leonardo Trujillo, Jonathan Rouzaud-Cornabas, Carole Knibbe, and Guillaume Beslon. The paper itself has not been altered to stay true to the citation. Supplementary Materials have been added as the Appendix A.

In this chapter, we use Aevol to study the impact of chromosomal rearrangements on genome evolution. Aevol proposes a realistic genome structure, with a complex genotype-to-phenotype map and complex mutations happening on the sequence without an *a priori* distribution of fitness effects; it is thus perfectly appropriate for studying chromosomal rearrangements. It also allows impossible evolutionary experiments (O'Neill, 2003). Here, we compare evolutionary trajectories with and without chromosomal rearrangements to isolate their impact on fitness and genome architecture evolution, an experiment that is obviously impossible in real life.

We show that chromosomal rearrangements allow a fast initial expansion of the gene repertoire through gene duplication but also reduce the effect of diminishing-returns epistasis in the long term, as they open new possibilities and largely expand the genotype's neighborhood. Chromosomal rearrangements also seem to be linked to genome size limitation. Indeed, genome size grows indefinitely in their absence, while their mere presence seems to force a convergence towards an equilibrium genome size. As such, chromosomal rearrangements appear to be key mutations to understanding genome architecture evolution.

2.1 INTRODUCTION

Genomic structural variations occur in all domains of life, including viruses, prokaryotes and the full range of eukaryotic taxa (Darling et al., 2008; Alkan et al., 2011; Gao et al., 2017; Cao et al., 2022). These structural variations include insertions of transposable elements, recombinations, and chromosomal rearrangements. Although the precise definition of chromosomal rearrangements varies across references (Audrézet et al., 2004; Alkan et al., 2011; Mérot et al., 2020), they generally refer to inversions, translocations, duplications, and deletions of DNA segments. Chromosomal rearrangements have classically been a blind spot of molecular evolution, mainly due to technical issues linked to short-reads sequencing, but also due to their strong deleterious effects that can rapidly eliminate them from the population (Campo et al., 2004; Rocha, 2006; Kara et al., 2014; Connallon and Olito, 2022). Nevertheless, recent improvements in sequencing techniques have strongly increased our ability to detect them (Wala et al., 2018; Ho et al., 2020; Hanlon et al., 2022), and more and more data is being accumulated regarding their decisive impact on evolution, as highlighted in the 2019 special issue published by *Molecular Ecology* (Wellenreuther et al., 2019). It appears that duplications and deletions are far from rare in eukaryotes. In some cases, the per locus gene duplication rate can be higher than the per nucleotide substitution rate (Katju and Bergthorsson, 2013), resulting in one gene duplication per haploid genome every 50 generations in the yeast *S. cerevisiae* (Lynch et al., 2008), and every 500 generations in the fruit fly *D. melanogaster* (Schrider et al., 2013). In the human genome, many duplications and large deletions have been identified as causes of genetic diseases or cancers (Nattestad et al., 2018). In prokaryotes, Richard Lenski's Long Term Evolution Experiment (LTEE) has shown the importance of large scale rearrangements as drivers of genomic plasticity (Raeside et al., 2014) and innovation (Blount et al., 2012).

While new sequencing techniques and discoveries have shed a new light on chromosomal rearrangements (Quandt et al., 2015; Ho et al., 2020), theoretical frameworks have been slow to adapt. Indeed, the effect of chromosomal rearrangements is generally not addressed in theoretical articles and textbooks. In most models of evolution, substitutions are still the sole source of variation, with recombination merely expected to shuffle these variations among individuals (Weissman et al., 2010). In the rare cases where ectopic recombination is considered in evolutionary models, its effect is generally limited to gene permutations or variation of copy number, excluding *a priori* any effect on gene sequences themselves (Yancopoulos et al., 2005; Bhatia et al., 2018). Similarly, inversions are often viewed as just an evolutionary pathway that prevents recombination, hybridization, and introgression (Noor et al., 2001), thus keeping specific alleles together (Hoffmann

et al., 2004; Kirkpatrick, 2010). Nevertheless, the ubiquity of these rearrangements (Raeside et al., 2014; Wellenreuther and Bernatchez, 2018) calls for more in-depth studies of their potential other effects.

There are several reasons for chromosomal rearrangements not to be accounted for in classical evolutionary models. First, contrary to substitutions and InDels that act at the allelic scale, chromosomal rearrangements are multiscale events that can modify both the micro- and the macro-structure of the genome (*i.e.* the allelic sequences and the global organization of the genome), while most models simulate genes as unbreakable units, with different alleles but no explicit sequences (Yancopoulos et al., 2005; Weissman et al., 2010; Bhatia et al., 2018).

Second, chromosomal rearrangements entail a wide diversity of complex effects, notably due to their length distribution which spans several orders of magnitude, from a few base pairs to a substantial fraction of the genome (Darling et al., 2008), contrary to *e.g.* InDels, which length distribution is narrower. As a consequence, rearrangements can significantly modify the genome size, thus changing the overall probability of another rearrangement, as bigger chromosomes generally undergo more rearrangements (Kaback et al., 1992; Jensen-Seaman et al., 2004). As a consequence, successive chromosomal rearrangements should not be considered independent: the occurrence of a rearrangement is likely to change the rate and DFE of upcoming events.

The variety and complexity of chromosomal rearrangements makes it challenging to build a theoretical understanding of their effect on evolution. In this context, forward-in-time simulations are a promising tool to observe the effect of rearrangements and unravel their importance in adaptation to new environments (Mérot et al., 2020). However, in forward-in-time models – like the well-known SLiM (Haller and Messer, 2017) –, the effect of mutations is often either an allelic change, drawn from a predefined DFE, or a positional change of the gene. This prevents these models from considering any combination of small- and large-scale effects, and makes it difficult to account for non-independent events (where some kinds of events modify the DFEs of others). To overcome these difficulties, a model designed to study rearrangements should not rely on explicit *a priori* DFEs. On the opposite, the mutations should affect the pre-existing genome sequence, without regards for the phenotypic effect, which is computed after the mutation. In this way, the effect of a mutation depends on its characteristics (type, location, length), but also on the current genomic structure, the environment and the genotype-to-phenotype map.

Hence, a model designed to study chromosomal rearrangements should provide an explicit genome with both coding and non-coding regions, in which rearrangements can happen blindly and have both direct (when altering coding regions) and indirect (when modifying

the DFE of the different mutational operators – including rearrangements themselves) effects on fitness.

In this article, we use Aevol, a model addressing these requirements. Aevol is a forward-in-time simulation platform that emulates the evolution of prokaryotic-like organisms and enables repeated evolution experiments with adjustable parameters (Knibbe et al., 2007a). Although the model has been presented before (Parsons, 2011; Batut et al., 2013; Rutten et al., 2019; Liard et al., 2020), recent computational and methodological improvements have opened up a wide range of new possibilities for the software. Aevol allows for both local mutations and chromosomal rearrangements of the genetic sequence, without an *a priori* DFE. We propose a use-case of the software to highlight the importance of chromosomal rearrangements in genome evolution. To this end, we simulate evolution under multiple mutational scenarios of increasing complexity: with substitutions only, with local mutations only (mutations that can only alter the sequence at the allelic scale: substitutions, small Insertions and small Deletions), and with a full range of mutational operators, including local mutations and chromosomal rearrangements (duplications, deletions, and inversions). Also, in order to test whether chromosomal rearrangements can generate enough diversity on their own to enable efficient adaptation, we added a fourth scenario where only chromosomal rearrangements are present, without any kind of local mutation. These scenarios are repeated with two types of populations, one starting far from the fitness optimum and one starting close to it.

Our simulations first show that, when far from the optimum, chromosomal rearrangements are an essential component of evolution, and even more important than local mutations. Indeed, by the end of the simulation, populations evolved with solely chromosomal rearrangements are far better adapted than populations evolved with local mutations or substitutions only. Moreover, the simulations also show that the evolution of genetic structure – including the genome size – is very different when rearrangements are allowed, emphasizing their role in the regulation of the amount of DNA (Knibbe et al., 2007a). Simulations starting close to the fitness optimum confirm the latter effect, but also demonstrate that, on the long term, chromosomal rearrangements reduce the effect of diminishing-returns epistasis, defined as the speed at which the marginal improvement of beneficial mutations decreases at each improvement (Wiser et al., 2013). Taken together, these simulations emphasize the decisive contribution of chromosomal rearrangements to long-term evolution, and show the potential of the Aevol platform to study their evolutionary impact.

2.2 MATERIAL AND METHODS

2.2.1 *Aevol: a forward-in-time evolutionary simulator with complex mutations*

Aevol (https://www.aevol.fr) is a forward-in-time evolutionary simulator that simulates the evolution of a population of haploid organisms through a process of variation and selection (Knibbe et al., 2007a; Beslon et al., 2010; Parsons et al., 2010; Frenoy et al., 2013; Batut et al., 2013). Each artificial organism, similarly to prokaryotes, is asexual, haploid, and owns a single circular chromosome. The design of the model focuses on the realism of the genome structure and of the mutational process. Aevol can therefore be used to decipher the effect of chromosomal rearrangements on genome evolution, including their interactions with other types of mutational events.

In short, Aevol is made of three components (Figure 2.1A):

- A mapping that decodes the genomic sequence of an individual into a phenotype. The genomic sequence of each organism is a double-stranded cicular binary sequence. Reading this sequence enables us to identify start-stop locus of transcription and translation, thus delimiting Open-Reading Frames. These are genes that are then decoded into proteins, represented by mathematical functions which sum represents the phenotype. Finally, the phenotype is compared to an environmental target, and their difference is used to compute the individual's fitness value.

- A population of organisms, each owning its own genome, hence its own phenotype and fitness. These individuals are located on a grid with one individual per grid cell. At each generation, the organisms are selected according to their fitness to populate the next generation. By default the competition is local (each organism competing with its neighbors), although other selection modes are possible.

- A genome replication process during which genomes can undergo several kinds of mutational events. These include chromosomal rearrangements and local mutations, but no recombination in the current version. The seven modelled types of mutation are depicted in Figure 2.1B and comprise three local mutations: substitutions, small insertions, and small deletions; two balanced rearrangements (which conserve the genome size), inversions and translocations; and two unbalanced rearrangements, duplications and deletions. This allows the user to study the effect of chromosomal rearrangements and their interaction with other kinds of events such as substitutions and InDels. The position of the mutations and the breakpoints of the rearrangements are chosen uniformly along the genome. Hence, longer chromosomes

		SUB	LM	CR	CRLM
Local Mutations	Substitutions (per bp)	3×10^{-5}	1×10^{-5}	0	5×10^{-6}
	Small insertions (per bp)	0	1×10^{-5}	0	5×10^{-6}
	Small deletions (per bp)	0	1×10^{-5}	0	5×10^{-6}
Chromosomal Rearrangements	Duplications (per bp)	0	0	1×10^{-5}	5×10^{-6}
	Deletions (per bp)	0	0	1×10^{-5}	5×10^{-6}
	Inversions (per bp)	0	0	1×10^{-5}	5×10^{-6}
Total per base pair per generation event rate		3×10^{-5}	3×10^{-5}	3×10^{-5}	3×10^{-5}

Table 2.1: Mutation rates per base pair per generation for the four mutational scenarios: SUB, LM, CR and CRLM. For mutations affecting subsequences (*i.e.* all mutations but substitutions), this rate corresponds to the probability to initiate an event at a given locus. Note that the total mutation rate (per base pair, per generation) is constant across experiments. An additional scenario (CRLMx2) has been tested to have equal mutation rates for all kind of events (1×10^{-5}) between CR , LM and CRLMx2 (see Supplementary Material, Figure A.4).

can undergo longer rearrangements. By contrast, InDels have a predefined lengh distribution (1 to 6 bp by default).

Supplementary Materials for this paper are included in the Appendix A.

A detailed presentation of the model is available in the Supplementary Materials (Figure A.1).

2.2.2 *In silico experimental setup: Evolution with limited mutations*

Experiment starting from naive individuals

We run 11 replicate simulations for four types of conditions: substitutions only (SUB), local mutations only – substitutions and InDels (LM), chromosomal rearrangements only – duplications, deletions and inversion (CR), and both chromosomal rearrangements and local mutations (CRLM). Note that translocations, although possible in Aevol, are excluded here to have as many local mutations as chromosomal rearrangements, and so a constant per base mutation rate in our different setups. The median (in terms of final fitness) CRLM run will be used to start the second set of simulations. The simulations begin with naive individuals owning a single gene. We want to study lineages for $1,000,000$ generations, which is enough to reach a stable genome with no more large variations in genome size and structure – although there is still room for adaptation. To this end, we run the simulations for $1,100,000$ generations, the last $100,000$ being used to ensure the survival of the lineage we retrieve.

All replicates share the same population size ($1,024$ individuals on a 32×32 square grid), the same environment (a sum of three Gaussian lobes, see Figure 2.2 and Supplementary Material, Figure A.3)

Supplementary Materials for this paper are included in the Appendix A.

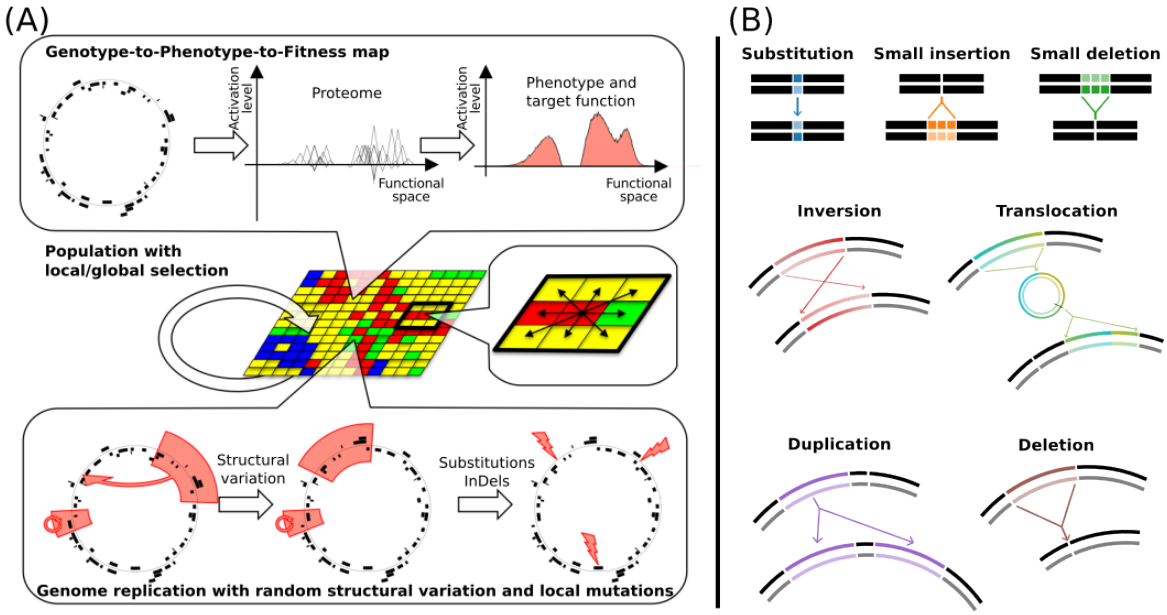

Figure 2.1: The Aevol model. The left panel (A) shows all steps of a generation in Aevol. *(top)* Overview of the genotype-to-phenotype map. Note that the organism shown here is a real organism evolved within Aevol for $1,000,000$ generations with a typical target. It contains many Open-Reading Frames on both strands, a large proteome (the set of proteins), and it is well adapted to its environment (*i.e.* its phenotypic function — black curve — is very close to the target function — light red area). *(middle)* Population on a grid is fully renewed every generation. Example of a local selection process occurring with a 3×3 neighborhood. *(bottom)* Mutation operators include chromosomal rearrangements (duplications, deletions, translocations and inversions – here a translocation and an inversion are shown) and local mutations (substitutions and InDels).
These mutations are described more precisely in the right panel (B): *(top)* Local mutations: substitution (one base pair is mutated to another), small insertion and small deletion (a few base pairs are inserted or deleted). *(middle)* Balanced chromosomal rearrangements: inversion (two points are drawn and the segment in between is rotated) and translocation (a segment is excised, circularized, re-cut and inserted elsewhere in the genome). *(bottom)* Unbalanced chromosomal rearrangements: duplication (copy-paste of a segment in the genome) and deletion (suppression of a segment of the genome).

and the same selection mode (local competition against the direct neighbors). The only difference lays in the mutation rates, as shown in Table 2.1. Importantly, for each condition, mutation rates are equally balanced between all mutation types and adjusted such that the overall mutation probability per locus is constant throughout all experiments. An example parameter file for the CRLM setup is provided in the Supplementary Material (Figure A.3).

For every simulation, we reconstruct the final lineage by tracking the ancestry of an individual from the final generation. We then retrieve the fitness, genome size, coding and non-coding sizes, and number of genes of all the individuals in this lineage. We also extract all mutations along the lineage, and record their type and effect on fitness.

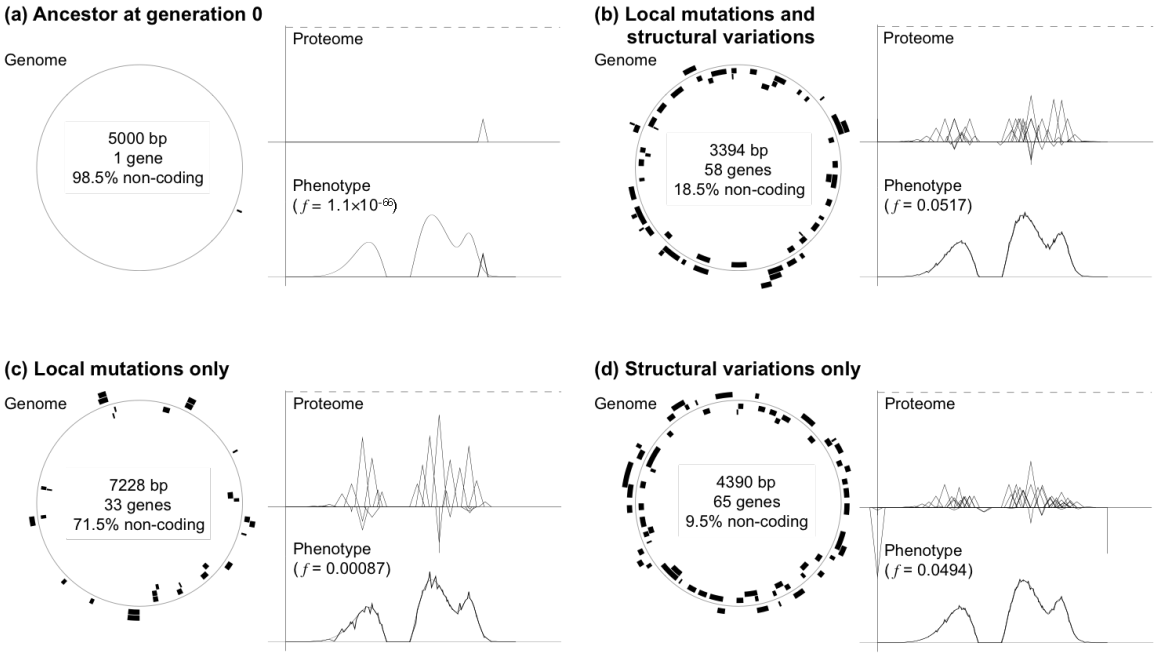

Figure 2.2: Initial ancestor (a) and examples of evolved organisms in the CRLM (b), LM (c) and CR (d) conditions after 1,000,000 generations. The organism presented in (b) corresponds to the Wild-Type used for the second step of the experiments. For each organism, there is on the left a visualisation of its genes localised on the genome. On the right, the proteome shows all the single proteins, and the phenotype (black curve) is their sum. The grey curve plotted in addition to the phenotype is the environmental target function, a sum of 3 Gaussian lobes (2 positives and 1 negative) – see Supplementary Material, Fig. S3. Finally, f is the absolute fitness value computed from the difference between the phenotype and the target function.

Finally, along the line of descent of the 11 CRLM experiments we extracted the 11 individuals at generation $1,000,000$ and selected the median one (in term of fitness) to estimate its Distribution of Fitness Effects (DFE) for each type of mutation. This allows to better understand the differences between local mutations and chromosomal

rearrangements in terms of impact on the fitness and chances of fixation. Note that this individual is the same that to one used to initiate the second run of experiments (see below).

Evolution from Wild-Types

After 1,000,000 generations, individuals are well-adapted to their environment, especially in the CRLM experiments. They can be used as Wild-Types to start new experiments. Here, the median CRLM experiment (in terms of final fitness) is used to initialize new clonal populations to test evolution from a well-adapted genome in the four mutational scenarios (SUB, LM, CR and CRLM). These populations are then evolved for another 3,100,000 generations to study the impact of chromosomal rearrangements when individuals are already well adapted to the environmental conditions. The same processing as for the first part of the experiments is then performed: reconstruction of the lineage for 3,100,000 generations and analysis of the genomes and mutations from generation 0 to 3,000,000 along this lineage (generations 3,000,001 to 3,100,000 being removed to ensure coalescence).

Fitting fitness trajetories

In order to estimate diminishing-returns epistasis, *i.e.* how fast the advantage provided by each new beneficial mutation reduces over time, for each mutational condition, we fit the mean fitness values along the 11 lines of descent with power laws of type $f = (bt + 1)^a$ where f is the fitness, t is the time in generations (Wiser et al., 2013). a and b are the parameters to be fitted with a corresponding to the diminishing-returns epistasis when $0 \leq a \leq 1$ ($a = 1$ corresponding to linear fitness growth without diminishing-returns epistasis) and b corresponding to an initial fitness growth parameter.

To compute the fit, we use the `lmfit` Python package with the least squares method. In order to ease the fitting process, the data points were sampled once every 1,000 generations.

2.3 RESULTS

To investigate the contribution of chromosomal rearrangements to evolutionary innovation, we compare the evolutionary dynamics of four sets of runs: SUB, with only substitutions; LM, with only local mutations; CRLM, with both local mutations and chromosomal rearrangements; and CR , with only chromosomal rearrangements. As we suspect that the relative contribution of chromosomal rearrangements versus local mutations depends on the distance to the fitness optimum, we repeated these experiments in two conditions: starting with naive individuals (see Section 2.3.1) or with pre-evolved ones (WT – see Section 2.3.2).

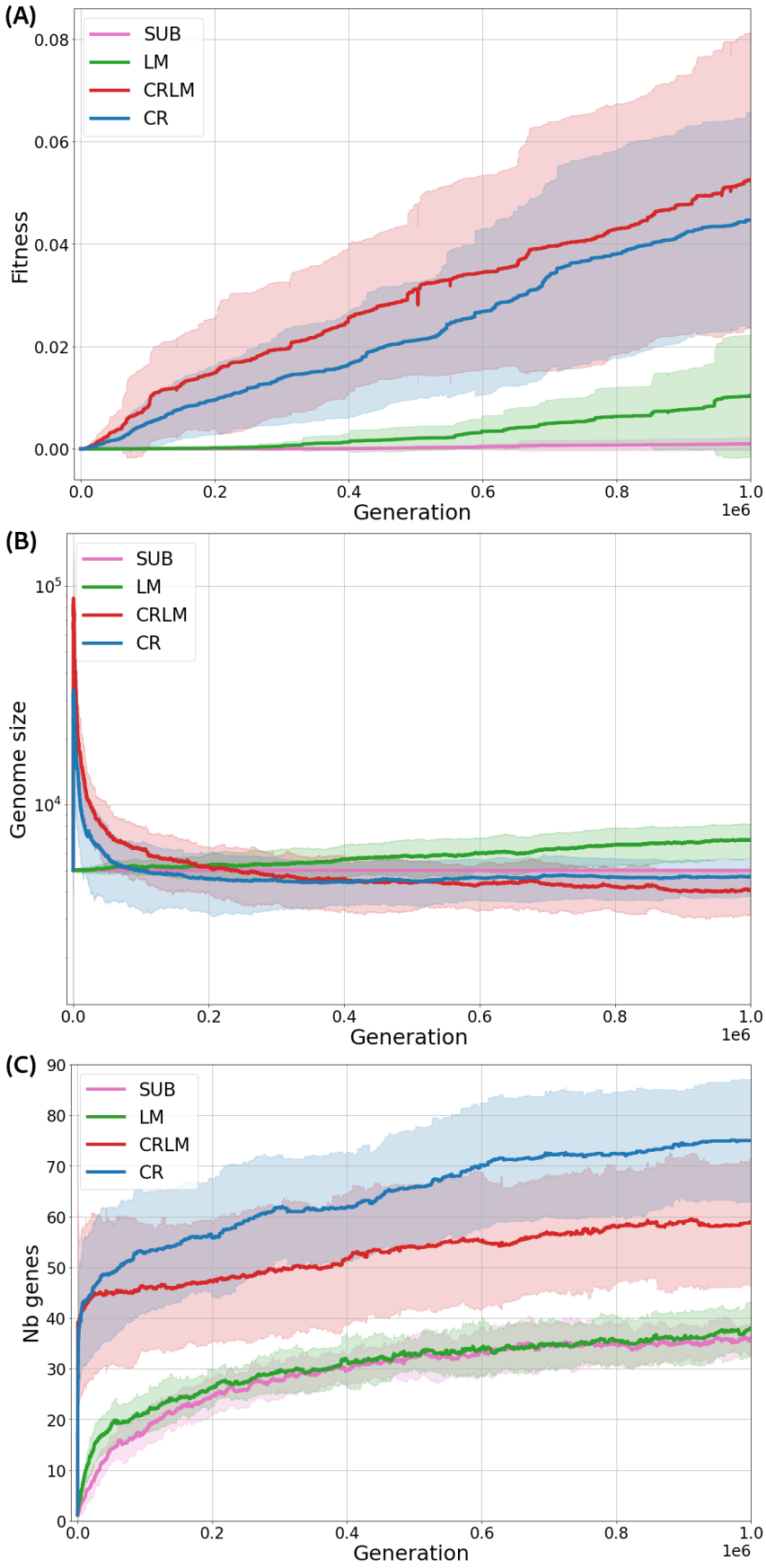

Figure 2.3: Mean variation of fitness (A), genome size (B) and gene number (C) on the line of descent of the final populations, starting from a naive individual for the four mutational scenarios. The shaded areas indicate the variability across the 11 repetitions (standard deviation).

2.3.1 *Local mutations are dispensable when far from the optimum*

As shown in Figure 2.2 and Figure 2.3, the evolutionary trajectories in terms of fitness, genome size and number of genes without local mutations (CR) are similar to the evolutionary trajectories with both rearrangements and local mutations (CRLM), whereas the simulations without rearrangements (SUB and LM) produce significantly less adapted organisms, with fewer genes and a smaller coding genome size despite a greater total genome size.

Strikingly, the end fitness in the CRLM setup is not statistically different from the CR setup (Mann-Whitney U test, p-value = 0.65), while both values are highly different from those in the cases without chromosomal rearrangements (Mann-Whitney U test, $p = 5 \times 10^{-4}$). This result is surprising, given that local mutations are usually thought to be a major evolutionary force, and would therefore be expected to provide a boost in fitness when present.

There are also structural differences in the genomes depending on the set of allowed mutations. First, the dynamics of gene creation is much slower in the SUB and LM simulations, as could be expected in the absence of gene duplication. Indeed, in the CRLM setup, a fixed duplication adds on average 2.58 genes to the genome (for a total across repetitions of 1,241 new genes), while all other mutations stand below 0.05 per fixed mutation (for a total of 388 new genes for all other mutations). However, we observe that the genomes evolved in the CRLM setup achieve a similar fitness but with fewer genes than the ones in the CR setup, highlighting that local mutations are better than chromosomal rearrangements at fine-tuning existing genes. Chromosomal rearrangements and local mutations also have different effects on genome size. Indeed, in the presence of chromosomal rearrangements (CR and CRLM), genome size sharply increases at first, before slowly reducing and stabilizing around 3,000 bp. On the contrary, in the LM setup, genome size never ceases to grow all along the experiment, although at a slow pace. This is caused by the fixation of more small insertions than small deletions (see Figure 2.4B). Ultimately, genomes evolved under the LM setup are longer than genomes evolved under the CR and CRLM setups but they contain much fewer genes, resulting in a larger proportion of non-coding DNA (see Figure 2.2).

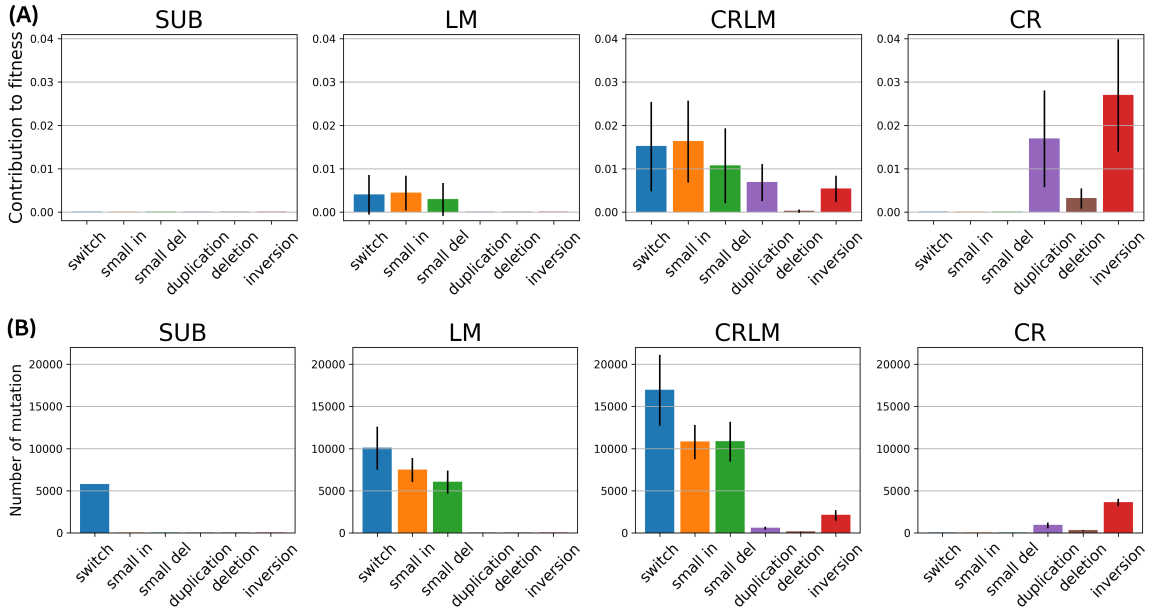

Figure 2.4: Fitness contribution and number of mutations fixed during the initial evolution from naive individuals (A) Contribution of each type of mutation to the total fitness gains, measured as the sum of the change in fitness of each mutation on the line of descent of the final best individuals, starting from naive individuals. Histograms show the mean values across the 11 repetitions, and the bars show their standard deviation. Reverted mutations (mutations which effect on fitness was exactly compensated by the following one) were filtered out to reduce noise. Fitness increase in the SUB simulations are negligible at this scale. (B) Number of non-neutral and non-reverted mutations fixed for the different mutation types and for the four conditions, normalized by the number of mutations occurring ($L \times \mu$, with L the genome size), on the line of descent of the final best individuals, starting from naive individuals. Histograms show the mean values across the 11 repetitions, and the bars show their standard deviation.

Finally, comparing the SUB and LM setups shows that the dynamics of *de novo* gene creation is similar in both conditions, but that the fitness of the LM simulations increases much faster than the fitness of the SUB ones. This shows that InDels do not facilitate *de novo* gene creation but that once a gene is present on the genome, they facilitate its evolution, hence reaching higher fitness.

To better understand the origin of these differences, we first look at the contribution of each mutation type to the end fitness. We computed the total gain of fitness per mutation type along the ancestral lineage during the $1,000,000$ generations of each experiment (Figure 2.4A). Interestingly, although CRLM are much fitter than LM, it is still the local mutations that contribute the most to the overall fitness gain in CRLM. Local mutations are crucial to evolution, and it is not surprising that they are the most impactful. However, the difference between SUB, LM and CRLM shows that their potential is only fully unleashed when

chromosomal rearrangements are also present and create a substrate that local mutations can then finely tune.

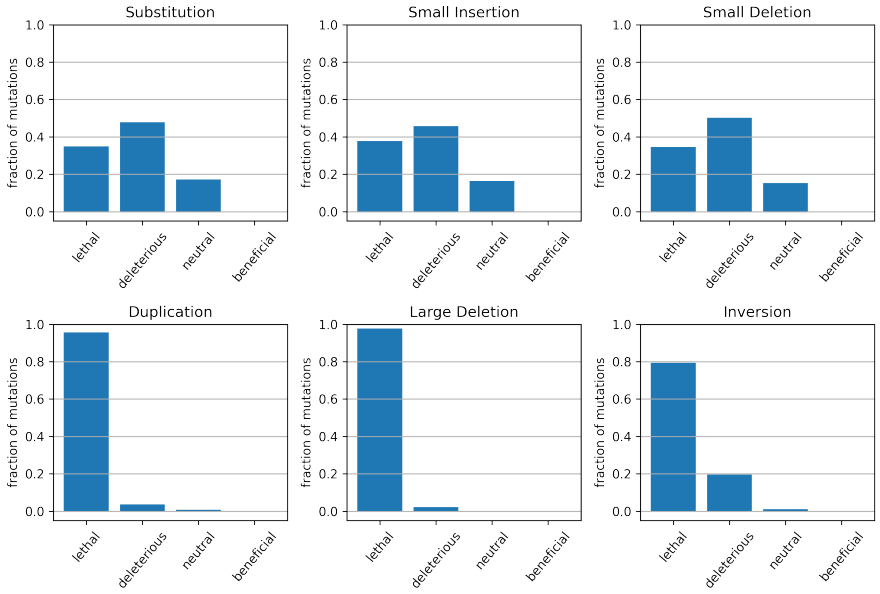

Figure 2.5: Distribution of Fitness Effects (DFE) of the different types of mutation, on the median individual of the CRLM experiment, after $1,000,000$ generations when starting from a naive individual. For each mutation type, $1,000,000$ mutants were generated, except for the substitution, which were exhaustively tested. The selection coefficient is computed as $s = \frac{f_{mutant}}{f_{parent}} - 1$. Lethality is defined as $s < -0.999$, and neutrality as $s \in [-0.001, 0.001]$. The detailed DFE is presented in Supplementary Material (Fig. S2). Interestingly, there is no advantageous substitutions available, showing that the population has reached a local fitness optimum for these mutations. However, as shown by supplementary Fig. S2, a few beneficial InDels and a few beneficial segmental duplications are available, although they are not frequent enough to be visible here.

The number of non-neutral mutations fixed along the line of descent (Figure 2.4B) shows that rearrangements, although rarely fixed compared to local events and hence almost invisible in the phylogeny, favor the fixation of beneficial local mutations. This is consistent with the dynamics of gene number shown on Figure 2.3C: by allowing for the recruitment of more genes, rearrangements increase the number of potential mutational targets on which local events can have an effect, hence favoring the fixation of more favorable local events.

The very rare fixation of rearrangements compared to the fixation rate of local mutations can be better understood by looking at the Distribution of Fitness Effects (DFE) for each type of mutation (see Figure 2.5). Duplications and deletions have a very broad effect and can disturb, delete or imbalance essential genes: they are therefore very often lethal (in approximately 95% of cases here). Local mutations, on the other hand, have a smaller chance of disrupting an essential gene,

as they affect a restricted section of the genome. They are more often neutral or "simply" deleterious, and lethal only in less than 40% of cases. Finally, inversions have two breakpoints while local mutations have only one, and are therefore more lethal than local mutations (80%), but, as inversions are balanced rearrangements, they are less likely to be deleterious than duplications or deletions.

2.3.2 *Chromosomal rearrangements sustain long-term adaptation*

When starting from a wild-type individual, whose gene repertoire has already evolved, the advantage of gene duplication over *de novo* gene creation vanishes, and we can study more subtle interactions between local mutations and chromosomal rearrangements. Here we initiate experiments from clonal populations of the median CRLM individual evolved in the previous set of experiments and follow their evolution for 3,000,000 generations in SUB, LM, CRLM and CR conditions.

Figure 2.6A shows that the four conditions result in very different dynamics of genome size. While the genome size of CR and CRLM experiments is quite stable, as observed at the end of the previous experiments, in LM conditions the genome size increases continuously during the 3 million generations of the experiment. At first sight, this result may seem contradictory, as the genome size is much more likely to vary in the presence of long segmental duplications/deletions than in the sole presence of small InDels. This shows the complex effect of chromosomal rearrangements in regulating genome size, and highlights the difference between InDels and rearrangements in doing so.

As expected, when looking at the fitness gain along the 3 million generations of the experiment (Figure 2.6B) the difference between the mutational scenarios is not as marked as what was observed when far from the optimum, at least for the LM, CR and CRLM scenarios. Yet, the SUB scenario still clearly lags behind in terms of fitness, showing again that substitutions alone are not sufficient in fine-tuning genes. In the four conditions, fitness improves all along the experiment, albeit with a clear diminishing-returns epistasis in the SUB, LM and CRLM conditions. Following Wiser et al. (2013), we used power-law curve fitting to estimate the amount of diminishing-returns epistasis in the four conditions (black lines on Figure 2.6B). Results show that diminishing-returns epistasis is higher in the SUB and LM conditions than in the CRLM conditions ($a_{SUB} = 0.2$; $a_{LM} = 0.4$; $a_{CRLM} = 0.5$ – see Methods, Section 2.2.2) which, in the long run, advantages the CRLM over the other scenarios. Strikingly, when evolving only with chromosomal rearrangements (CR scenario), populations show no diminishing-returns epistasis throughout the duration of the experiment ($a_{CR} = 1.5 > 1$). This contrasts with the other conditions and

allows the CR populations to catch up with the SUB and LM ones, despite an initial disadvantage.

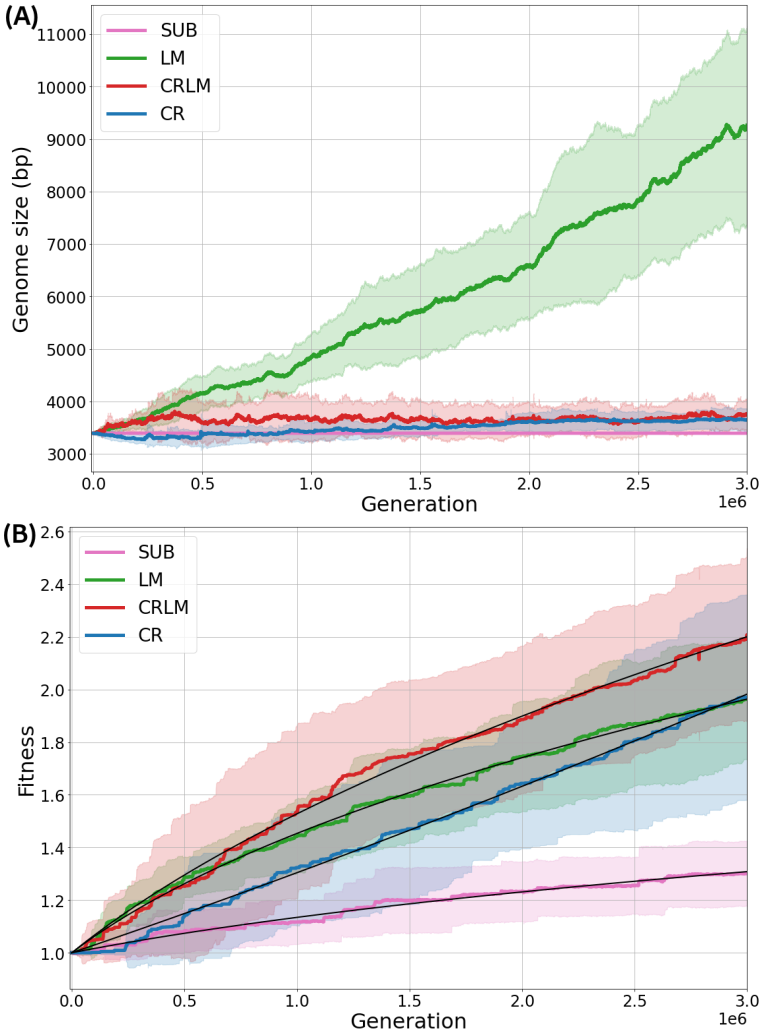

Figure 2.6: Temporal changes in genome size and fitness in evolution started from the WT. (A) Mean change in genome size on the line of descent of the final populations, for the 11 repetitions and the 3 conditions. All simulations started from the same Wild Type with a genome length of 3394 bp (Figure 2.2.b) and evolved for 3,000,000 generations. The shaded areas indicate the variability across repetitions (standard deviation). (B) Relative fitness variation on the line of descent of the final population, starting from a Wild Type. The shaded areas indicate the variability across repetitions (standard deviation). Black curves show the fitted power laws for the mean fitness values of the four sets of simulations (see Methods, Section 2.2.2). The fitted parameters are: $a_{SUB} = 0.2$, $b_{SUB} = 7.0 \times 10^{-7}$, $a_{LM} = 0.4$, $b_{LM} = 1.8 \times 10^{-6}$; $a_{CR} = 1.5$, $b_{CR} = 2.0 \times 10^{-7}$, $a_{CRLM} = 0.5$, $b_{CRLM} = 1.5 \times 10^{-6}$.

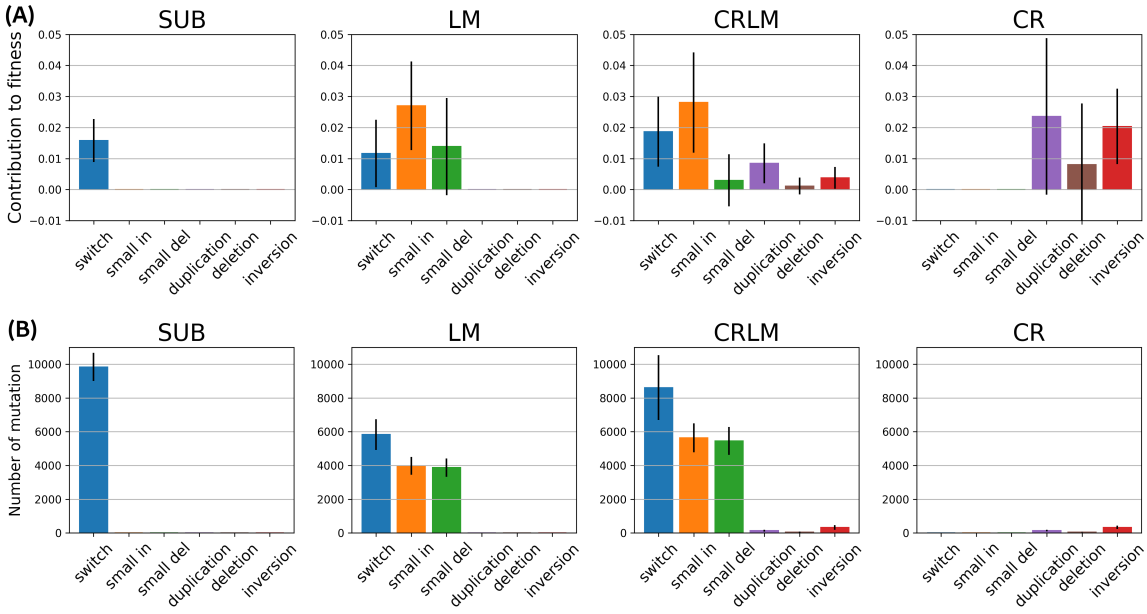

Figure 2.7: Fitness contribution and number of mutations during the evolution from WT individuals (A) Contribution of each type of mutation to the total fitness gains, measured as the sum of the change in fitness of each mutation on the line of descent of the final best individuals, starting from WT individuals. Histograms show the mean values across the 11 repetitions, and the bars show their standard deviation. Reverted mutations (mutations which effect on fitness was exactly compensated by the following one) were filtered out to reduce noise. (B) Number of fixed non-neutral and non-reverted mutations per generation for the different mutation types per million generation, normalized by the number of mutations occurring ($L \times \mu$, with L the genome size), on the line of descent of the final best individuals, starting from WT individuals. Histograms show the mean values across the 11 repetitions, and the bars show their standard deviation.

As previously, we measured the total fitness effect and the number of non-neutral mutations fixed along the lineage for the different types of mutation and for the four mutational scenarios (Figure 2.7A and Figure 2.7B respectively). As already noticed when starting far from the optimum, this shows that chromosomal rearrangements, although very rarely fixed in the lineage, have a dual contribution to fitness. While, in the CRLM, fixed rearrangements have a small impact on fitness on their own (Figure 2.7A), they contribute to increasing the number of favorable substitutions. Indeed, substitutions and InDels are more likely to be favorable and fixed in the CRLM populations than in the LM populations and almost as likely – for the substitutions – as in the SUB ones (Figure 2.7B). This leads to a sustained evolutionary dynamics, despite rearrangements being almost invisible in the phylogeny owing to their very low fixation probability.

2.4 DISCUSSION

It is widely admitted that genomes evolve under the combined pressure of a large variety of mutational operators, including of course substitutions and InDels but also chromosomal rearrangements (Mérot et al., 2020; Berdan et al., 2021b). However, models of genome evolution almost exclusively focus on the former, the latter being generally ignored owing to their difficult modelling and their apparent low frequency in phylogenies that could suggest a moderate impact compared to other events. A direct consequence is that the contribution of chromosomal rearrangements to the evolutionary dynamics is largely overlooked. Indeed, while substitution-based epistasis is largely recognized and quantified in several model systems (Olson et al., 2014; Bank et al., 2015; Starr and Thornton, 2016; Diss and Lehner, 2018), the epistatic effect of rearrangements is, with very few exceptions (Blount et al., 2012), *terra incognita*.

Here we used Aevol to simulate genome evolution under several conditions characterized by an increased mutational diversity but a constant overall mutational rate (see Table 2.1). We completed these experiments by testing evolution under the exclusive pressure of chromosomal rearrangements, in order to estimate their capacity to generate enough variation to allow sustained evolution. This enables an experimental (though simulated) exploration of the consequences of chromosomal rearrangements on the evolutionary dynamics. Specifically, we analyzed the results of the simulations with a focus on two levels: genome structure, which is likely to be largely impacted by rearrangements, and individuals' fitness.

Regarding the evolution of genome structure, our results show two clear differences when genomes evolve with (CRLM and CR simulations) or without (SUB and LM simulations) chromosomal rearrangements. First, they confirm the well-established theory of evolution by gene duplication (Zhang, 2003; Kalhor et al., 2023): in our simulations, rearrangements are essential for the rapid acquisition of a large gene repertoire, and duplications are the main cause of increase in gene number (see Section 2.3.1). Indeed, gene number rapidly increases in the very first thousands of generations for CR and CRLM (Fig. Figure 2.3C), and this process of gene recruitment is maintained throughout the simulation, though at a lower pace. On the opposite, lineages evolving without rearrangements only acquire a limited gene repertoire (see Figure 2.3C).

In a less intuitive way, our simulations show an important contribution of chromosomal rearrangements to the stabilization of genome length during evolution. Indeed, Figure 2.3B and Figure 2.6A show that, after an initial burst of genome size at the very beginning of the evolution (corresponding to the phase of fast gene acquisition through duplications), CR and CRLM lineages quickly undergo a reduction of

their genome size (while preserving their gene repertoire – see Figure 2.3B and C). Continuing the simulation for 3 million generations, we see that genome size varies very little thereafter (Fig. Figure 2.6A). This dynamic contrasts sharply with that of the LM lineages, which show a steady increase in genome size, both when starting far or close to the optimum. This sustained growth of genome size under the sole pressure of InDels advocates in favor of the mechanism of border-induced selection, which has been recently conceptualized by Loewenthal et al. (2022). Indeed, despite their spontaneous mutation rates being equal, the probability of fixation of neutral insertions is slightly higher than the probability of fixation of neutral deletions, due to interference with gene borders (Loewenthal et al., 2022): a small insertion close to a gene is most often harmless, while a small deletion at the same point can impact a gene if the size of the deletion is larger than the distance to this gene. In the absence of other constraints on the genome size, this bias leads to a steady genome growth, as we observe on Figure 2.3B and Figure 2.6A. Strikingly, in the presence of chromosomal rearrangements, this bias is not visible anymore, showing that rearrangements generate an evolutionary pressure that prevents genome growth. As already proposed by Knibbe et al. (2007a), deleterious chromosomal rearrangements lead to selection for robustness, favoring smaller genomes as these undergo fewer rearrangements than longer ones. This hypothesis is sustained by the low rate of fixation of chromosomal rearrangements (Figure 2.7B): they are largely filtered-out by purifying selection, suggesting that they have a strong robustness effect. The low number of fixed rearrangements, due to their high lethality, (Figure 2.5) questions the concept of mutation rate. Indeed, by measuring mutation rates on a live population, a bias is introduced towards non-lethal mutations. This bias has been observed in the case of substitutions (Wang et al., 2012) but we hypothesize that this could be even more important in the case of genome rearrangements, and models should take into account that spontaneous mutation rates could be very different from observed and fixed ones.

The influence of chromosomal rearrangements on fitness evolution is also very different depending on whether the simulations start far from the optimum (hence requiring them to acquire new genes) or close to the optimum (with a gene pool already acquired but that can still be optimized). In the former situation, lineages evolving in the presence of chromosomal rearrangements have a much higher fitness than those evolving with only substitutions, or even with all local mutations (Figure 2.3A). This confirms that, in such a situation, gene duplication has a decisive contribution (Zhang, 2003; Kalhor et al., 2023), enabling both the CR and CRLM lineages to largely overcome the LM and the SUB lineages. Strikingly, lineages evolving with chromosomal rearrangements only (CR) perform almost as well as those evolving

with both chromosomal rearrangements and local mutations (CRLM). This illustrates the multiscale nature of chromosomal rearrangements that can both enlarge the gene repertoire through large duplications but also optimize gene sequences by reorganizing them through *e.g.* inversions. This is coherent with Trujillo et al. (2022) which modeled inversions in simpler evolutionary setting and showed that, given enough time, inversions allow reaching higher fitness peaks than substitutions. Interestingly, the fitness of the SUB lineages (that evolved under the sole pressure of substitutions) is much lower than the fitness of the LM lineages (that evolved through substitutions and InDels) despite a very similar dynamic of gene recruitment. This confirms that small insertion and small deletions are decisive operators when the evolution of protein sequence is concerned, as they can add/remove codons when substitutions can only mutate existing ones (Vakhrusheva et al., 2011; Leushkin et al., 2012).

When starting close to the fitness optimum, the differences between the experiments are more subtle, except when substitutions are the sole mutational operator, in which case fitness gains are much lower than in the three other conditions (SUB curve on Figure 2.6B), highlighting the importance of the diversity of mutational operators (Berdan et al., 2021b). In all experiments, the dynamics of fitness is similar to what can be observed *in vitro*, for example in experimental evolution with bacteria (Wiser et al., 2013; Wang et al., 2016b), or yeast strains (Wei and Zhang, 2019): simulations show a sustained fitness gain all along the experiment albeit with a more or less pronounced diminishing-returns epistasis. Inspired by Wiser et al. (2013), we estimated the diminishing-returns epistasis in these different conditions, and showed that, in the long run, chromosomal rearrangements reduce diminishing-returns epistasis, hence enabling sustained evolutionary dynamics. It is known that clonal interference could also induce diminishing return (Wiser et al., 2013). However, as the population size and global mutation rates are the same in all our simulations (CR , CRLM, LM and SUB), we assumed clonal interference had similar effect in all simulations Moreover, as shown by Figure 2.7A, the effect of rearrangements is mainly indirect: they have a small effect by themselves but potentiate other factors. Indeed, in the CRLM lineage, substitutions have a larger impact than in the SUB and LM lineage. This suggests that rearranged sequences open new targets to substitutions, hence increasing the probability to fix beneficial local events (Figure 2.7B). Finally, as Figure 2.7B also shows, this effect is due to a very low number of fixed rearrangements. Hence, while rearrangements sustain long-term adaptation by reducing the effect of diminishing-returns epistasis, they are almost invisible in the phylogeny.

When quantifying the diminishing return, a striking result was the apparent accelerating evolution in the CR populations ($a_{CR} > 1$). We hypothesize that this is due to the low fixation rate of chromosomal

rearrangements (Figure 2.7B). As CR populations undergo only rearrangements, fitness comparatively evolve by bigger steps but with longer waiting times between mutations, and this creates an initial lag in the fitness gain (Figure 2.6B), hence the appearance of acceleration. Now, the number of possible rearrangements for a given genome is much larger than the number of possible local events (it is indeed mainly linked to the number of breakpoints to be chosen for a given type of event: one for local mutations, two for inversions and deletions, three for duplications – see Figure 2.1B). A direct consequence is that, contrary to substitutions and InDels, rearrangements neighborhood cannot be explored in a reasonable time, hence the lower diminishing-returns epistasis observed on the duration of our simulations when rearrangements are allowed. Further, exploring this question, *e.g.* by estimating the contribution of each type of rearrangement to the phenomenon, is a very promising research direction opened by our results.

Overall, our simulations show that chromosomal rearrangements have both a direct (through gene duplications) and an indirect (by potentiating the effect of local mutations) contribution to the evolutionary dynamics. They seem to also act as regulators of genome size, due to purifying selection against long genomes which undergo too many mutational events, as already proposed by Knibbe et al. (2007a). This inverse correlation between mutation rates and genome size has already been observed in prokaryotes (Drake, 1991; Lynch, 2010), but for substitutions only. Our results suggest that its main determinant could be the rearrangement rates. Interestingly, this hypothesis implies that the regulation of genome size is due to the events that *do not* go to fixation in the winning lineage. Hence, despite them being almost invisible in the phylogeny, chromosomal rearrangements act as a major player of evolution by regulating genome size, limiting the effect of diminishing-returns epistasis, and sustaining long-term adaptation. Our results also illustrate the potential power of forward-in-time simulators like Aevol to unravel the effect of "non-conventional" mutational operators. Despite their artificial nature, models mimicking genome structures and the genotype-to-phenotype map allow deciphering the impact of the different types of mutation with a limited set of *a priori* hypotheses.

All models rely on simplifying assumptions, and ours makes no exception. However, the interest of modelling is precisely to reduce the complexity of the system to be studied. Here, studying only a limited number of mutational operators has enabled us to identify effects that could have been blurred in a more complex setting. Indeed, our experimental strategy, which relies on a progressive complexification of the mutational repertoire, has enabled us to uncover profound differences between chromosomal rearrangements and small InDels, both in the evolution of genome size and in the adaptation of organisms. Both

kinds of events may seem rather similar at first sight, but they differ on two important aspects: first, contrary to duplications that copy pre-existing genomic sequences, small insertions add random sequences to the genome. Hence, they cannot duplicate genes, while this process is central in evolution (Zhang, 2003). Second, even though both types of mutation add/remove genomic segments to the chromosome, the distribution of the size of these segments is different: in the case of In-Dels, this distribution is fixed while in the case of rearrangements, the distribution depends on the size of the genome. A direct consequence of this property is that larger genomes undergo more deleterious rearrangements, leading to a lower robustness (Knibbe et al., 2007a). In our simulations, large duplications and deletions, far from randomly shuffling the genome size as could have been expected, impose a tight constraint on it.

In the development of the model, we chose to stay close to prokaryotic genomics. This means that genomes are haploid and circular, and undergo no recombination. This obviously prevents us from studying the interplay between structural variation and recombination and its potential effect on speciation and on the fate of chromosomal rearrangements (Berdan et al., 2021a). We also chose to study a limited set of chromosomal rearrangements (duplications, deletions, and inversions), while many other types of events could be added to the model (*e.g.*, transposable elements, horizontal gene transfer, etc.). As for the rearrangements we model, breakpoints are chosen uniformly on the chromosome, leading to a uniform distribution of rearrangement lengths. This distribution is difficult to estimate in real organisms, as a large fraction of chromosomal rearrangements are likely to be lethal (Rocha, 2006). However, experimental studies show that the rearranged segments can reach lengths of the same order of magnitude as the size of the genome (Raeside et al., 2014), hence supporting our simplifying hypothesis, although the shape of the distribution in more likely to be geometric (Darling et al., 2008). However, we choose the simplest hypothesis of random breakpoints so as not to add additional parameters. We conjecture that our main results hold even with a geometric distribution of rearrangements, as the tail of the distribution will indeed grow with genome length. Yet, this could partly relax the robustness constraints, as they are mostly due to the longest rearrangements. We therefore expect that the effect of chromosomal rearrangements on genome size would hold, although it might be less pregnant with another distribution.

Our conclusions are drawn from the comparison of the evolutionary trajectories of different experiments and open up several interesting perspectives. For example, Aevol also includes several analysis tools, such as the computation of the distribution of fitness effect for all mutation types and for all genomes along a lineage, as illustrated by Figure 2.5. Taking advantage of the perfect record of the mutational

events, these measures help quantify the evolutionary forces at work, as well as the relative contribution of the different types of mutation to these forces. As exemplified on Figure 2.4A and Figure 2.7A, the impact of the different types of mutation on the fitness can easily be quantified, allowing to estimate the direct contribution of each type of mutation. Although it would be very computationally demanding, it could be interesting to also quantify the consequences of each mutation type on robustness and evolvability as this could allow to estimate their indirect effect and explain how the different types of mutations interact. Finally, as long as chromosomal rearrangements are concerned, an obvious prospect is to extend the model to diploid eukaryote-like genomes with recombination. This would enable exploring the interplay between rearrangements and recombination (Berdan et al., 2021a).

Supplementary Materials for this paper are included in the Appendix A.

The experiments we presented here only scratch the surface of what can be done with Aevol. Indeed, as Table S1 of the Supplementary Material shows, many other experiments can be done, including testing the effect of mutation rates, mutation biases or population size. Aevol is available to any team that would like to test hypotheses regarding the effect of these parameters on the evolutionary dynamics and on genome structure. Moreover, as the code is open and freely available, any team can modify it to test some specific mutation type that would not already be implemented (see Section 2.6). Notably, there are many ways to be far from the optimum. Here we choose to start with naive individuals but another approach would be to force environmental changes. In Aevol this could easily be done by moving the target function after having adapted organisms to a first environment. This would enable studying the contribution of rearrangements to evolutionary rescue. Indeed, a previous study with Aevol has shown that, in the case of an environmental change, the frequency of gene duplications is positively correlated with the distance to the optimum (Kalhor et al., 2023), but the impact of all chromosomal rearrangements could be studied more in details by limiting the number of possible mutation types, as we do in the present study. The role of chromosomal rearrangements when organisms are confronted to a perpetually moving target, and so always relatively far from the optimum, could also be further studied.

Despite the highly artificial nature of our model, our simulations are consistent with the classical view of evolution: among the variety of mutational operators, substitutions and small InDels are by far the most visible adaptive events both in terms of their number (Figure 2.4B and Figure 2.7B) and their contribution to the fitness (Figure 2.4A and Figure 2.7A). However, our simulations also show that the scarcity of rearrangements that we observe in the phylogenies masks an important contribution to adaptation. While the vast majority of models and simulators of molecular evolution still implements a solely allelic view

of evolution, where rearrangements can modify gene organization but cannot create new gene sequences, our results suggest that the innovative potential of rearrangements is not marginal, and that it is essential to integrate them into population genetics models.

2.5 ACKNOWLEDGEMENTS

M.F. is funded by the French Agence Nationale pour la Recherche (Evoluthon grant). L.T. thanks the Institut National des Sciences Appliquées de Lyon (INSA-Lyon) as well as the Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS) for hospitality while part of this research was done. J.L. and G.B. would like to thank the Rhône-Alpes Institute for Complex Systems (IXXI) for funding. All authors thank the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr), for computational support. J.L and P.B. thank Lisa Chabrier for instructive discussions.

2.6 DATA ACCESSIBILITY AND BENEFIT-SHARING

The code of the Aevol software is available on gitlab (https://gitlab.inria.fr/aevol/aevol). More documentation is available on the website https://www.aevol.fr. All relevant final data, as well as parameters files to redo the simulations, are available on Zenodo (https://zenodo.org/record/8307916).

2.7 AUTHOR CONTRIBUTIONS

GB, CK, JRC, DP, TG and MF developed the model. GB, PB and JL conceived, realized and analyzed the experiments. All authors discussed the results and contributed to the final manuscript.

3

GENOME STREAMLINING: EFFECT OF MUTATION RATE AND POPULATION SIZE ON GENOME SIZE REDUCTION

FOREWORD

The following work is published in *Genome Biology & Evolution* (Luiselli et al., 2024) and authored by Juliette Luiselli, Jonathan Rouzaud-Cornabas, Nicolas Lartillot, and Guillaume Beslon. The paper itself has not been altered to stay true to the citation. Supplementary Materials have been added as the Appendix B.

The previous chapter demonstrated that chromosomal rearrangements are determinant factors of genome architecture evolution, as their presence was necessary in our experiments to have an equilibrium genome size. We can conclude from that that chromosomal rearrangements introduce a unique evolutionary force that acts upon genome size evolution. To better understand this force, we now try to modulate it by changing the mutation rate (μ) and the population size (N). Indeed, as this force is tied to mutations, μ probably influences it. Additionally, genome size reached a stable equilibrium, which could therefore be influenced by selection — hence by N.

This chapter thus focuses on the effect of increases in N and μ on genome size evolution, and shows that both lead to a genome size reduction. Taking advantage of the approach with simulations, we highlight that this behavior is explained by the selection for robustness to chromosomal rearrangements: the evolutionary force contributed by these mutations is tightly linked to a second-order selection on the fidelity of the genomic information transmitted to the offspring along generations. As a result, a larger population size leads to a more efficient selection and is associated with a reduced genome size that increases the robustness of the individuals. On the other hand, a higher mutation rate increases the risk that chromosomal rearrangements pose, which increases the selective factor of genome size and also leads to reduced genomes. Both mechanisms are robust to potential mutational biases, including deletion biases, showing that the relationship between the strength of drift, underlying mutational biases, and genome size evolution is not straightforward.

Interestingly, our results also shed light on a new regularity in genome architecture evolution: while the total genome size depends on both the population size and the mutation rate separately, the final coding fraction seems to depend solely on their product. Indeed,

similar coding fractions have been reached by multiplying either N or μ by 4, or by multiplying each by 2. This regularity is explored in more depth in the next chapter.

3.1 INTRODUCTION

Genome size was one of the first studied genome characteristics (Leth Bak et al., 1969; Bachmann, 1972), yet its dynamic and causal factors are still poorly understood. Genome size is hugely variable across life: from less than 10^4 base pairs (bp) for viruses (Sanjuán, 2009), to more than 10^{11} bp for some plants (Pellicier et al., 2010). It does not correlate reliably with the number of genes or other variables throughout the different branches of life (Barow and Meister, 2002; Westoby et al., 2021).

The observed range of genome sizes is more restricted when studying only bacterial organisms (Westoby et al., 2021), ranging from 10^5 bp for intracellular endosymbiotic bacteria (Chong et al., 2019) to 10^7 bp for some myxobacteria (Schneiker et al., 2007). Bacterial genomes are mostly dense, and within this domain of life, genome size is loosely correlated with the number of coding genes (Konstantinidis and Tiedje, 2004; Almpanis et al., 2018). However, the precise determinants of bacterial genome size are still unknown, as it is still impossible to accurately predict the total genome size from the number of coding genes or from other genomic characteristics (Petrov, 2001; Barow and Meister, 2002; Choi et al., 2020). Part of the determinants of genome size are likely to be highly lineage-specific and linked to the ecological or evolutionary history of the lineages (Martinez-Gutierrez and Aylward, 2022). Nevertheless, it has been argued that at least a part of the observed variation may be due to universal mechanisms, linked to population genetics and molecular evolutionary processes (Lynch and Conery, 2003; Lynch, 2007a). In particular, it has been suggested that population genetics mechanisms could explain the reductive evolution observed in several bacterial strains (Lynch, 2006a). However, among the shortest bacterial genomes, one can find two types of bacteria which have very different ecological environments and evolutionary history: endosymbionts such as *Buchnera aphidicola* (Moran and Mira, 2001) and free-living marine bacteria such as *Prochlorococcus marinus* (Dufresne et al., 2005) or *Pelagibacter ubique* (Mathur, 2005). Strikingly, both types of bacteria lie at the two extremes of bacterial population sizes, questioning the mechanisms that led to genome reduction (Batut et al., 2014; Martínez-Cano et al., 2015; Wernegreen, 2015).

Buchnera aphidicola, and endosymbionts more generally, are characterized by very small effective population sizes (N_e) and high mutation rates (μ). Endosymbiosis also generally entails the introduction to a new stable environment and very close interactions with the host (Moran, 1996; Mira and Moran, 2002). These many complex factors

result in decaying genomes, smaller in total size and with fewer coding genes than those of average bacteria (Heddi et al., 1998). Endosymbionts have typically lost both coding and non-coding genomic content (Moran and Mira, 2001; Wernegreen, 2002), maintaining a coding fraction on the order of 85% (Ham et al., 2003), which is quite typical for bacteria (Kuo et al., 2009).

In sharp contrast, free-living marine bacteria such as *Prochlorococcus marinus* or *Pelagibacter ubique* also have reduced genomes (Mathur, 2005; Batut et al., 2014), but are believed to have very large effective population sizes (Marais et al., 2008; Martiny, 2013; Giovannoni et al., 2014), although that is an ongoing debate (Chen et al., 2022; Filatov and Kirkpatrick, 2024). Noticeably, in their case, genome size reduction is primarily contributed by the loss of non-coding sequences rather than coding sequences (Mathur, 2005; Batut et al., 2014). This phenomenon is called streamlining and could indicate a very effective selection (Wolf and Koonin, 2013; Giovannoni et al., 2014). Many hypotheses have been proposed to account for genome size reduction and the associated changes in genome architecture in such free-living organisms: adaptation to a nutrient-poor environment or to other abiotic factors, the Black Queen hypothesis, or high mutation rates (Koskiniemi et al., 2012; Morris et al., 2012; Batut et al., 2014; Ngugi et al., 2023).

Both endosymbionts and free-living marine bacteria thus show a marked reduction in genome size, linked to an increase in mutation rate (Bourguignon et al., 2020) but, strikingly, also linked to either an increase or a decrease in effective population size N_e. Indeed, while some observations link the decrease in genome size to the increase in random drift (Moran, 2002; Andersson, 2005; Kuo et al., 2009), this is not consensual among the scientific community since a long-term reduction in N_e is also thought to increase genome complexity and genome size: the increase in genetic drift would cause the fixation of slightly deleterious duplications, which would be more frequent than slightly deleterious deletions (Lynch and Conery, 2003; Lefebure et al., 2017). The balance between insertion and deletion rates and spectra may also play a role in genome size evolution (Petrov, 2002) and deletion biases in particular are believed to contribute to the small genome size of prokaryotes (Ratcliff, 2024). Overall, this suggests that a specific study focusing on the interaction between various mutational biases, variations in mutation rate and variations in effective population size is needed.

In this study, we focus on determining the impact of both an increased mutation rate and a change in population size on genome size evolution. However, mutation rates and population sizes are difficult to estimate. The effective population size is also highly variable through time, such that it is not totally obvious which long-term average is relevant at the macro-evolutionary scale (Brevet and Lartillot, 2021; Müller et al., 2022). For that reason, many comparative analyses

have relied on somewhat indirect proxies, such as life-history traits (Gunbin, 2007; Romiguier et al., 2012; Figuet et al., 2016). However, the precise quantitative relation between these proxies and effective population size is difficult to assess. Moreover, the very different living conditions and potential mutational biases of the bacterial species that have undergone genome reduction introduce many confounding factors. To avoid these pitfalls, we choose to turn to simulation, which allows us to control all the parameters (population size, mutation rate, and mutational biases) and the magnitude of their variation. It also ensures that no other factor than the ones investigated will impact the phenomenon under study. Hence, we can gain a theoretical understanding of the relationship between the different factors at stake and genome size reduction.

In silico experimental evolution provides tools to study genomic architecture in detail (Adami, 2006; Hindré et al., 2012; Batut et al., 2013). For our study, we need a framework that provides coding and non-coding genomic compartments which can vary independently, and with arbitrary underlying mutational biases for the deletion/insertion balance. Then, running simulations in a perfectly controlled environment covering a broad range of population sizes N and mutation rates μ makes it possible to investigate the conditions and mechanisms leading to genome size reduction. We will hence use Aevol, a simulation platform that provides an explicit genomic structure where both the coding and non-coding genome can evolve freely. Aevol emulates the evolution of bacteria and enables replicated and controlled *in silico* evolution experiments with known and fixed parameters (Knibbe et al., 2007a; Banse et al., 2024b). It provides an ideal tool to uncover links between genome size and either population size or mutation rate, as the experimenter perfectly controls these parameters. Throughout the experiments, fitness, genome size, and amounts of coding and non-coding bases are monitored to study the evolution of genome architecture and the response of genome size to changes in μ and N.

Our results show that both an increase in N or μ lead to genome size reduction, regardless of the underlying mutational bias. However, both conditions lead to very different genome structures, as a high μ reduces both the coding and non-coding compartments while a high N reduces only the non-coding compartment. Surprisingly, they both lead to a similar coding proportion when increased by the same factor, such that $N \times \mu$ appears as a key compound parameter determining this proportion. To understand this result, we measured both the phenotypical adaptation and the replicative robustness of the genomes, *i.e.* their capacity to transmit faithfully their phenotypes to their offspring. Indeed, while the per-base mutation rate is constant within each of our experiments, the genome-wide mutation rate varies with genome size, and the impact of the mutations depends on the genome structure and the type of mutation. Therefore, replicative robustness is tightly linked

with genome size and coding proportion. We show that the observed variations in genome size and structure are due to the interaction between selection for phenotypical adaptation to the environment and selection for robustness.

3.2 RESULTS

We perform our experiments using Aevol, a forward-in-time evolutionary simulator (Knibbe et al., 2007a; Banse et al., 2024b). Aevol is an individual-based model which includes an explicit population and in which every organism owns a double-stranded genome. It uses an explicit genome decoding algorithm directly inspired by the central dogma of molecular biology to compute the phenotype, and thus the fitness, of each individual based on its genomic sequence. As Aevol also includes a large variety of mutational operators (including substitutions, InDels, and chromosomal rearrangements), this non-parametric genotype-to-phenotype map allows for changes in the genome architecture (genome size, coding density, overlapping genes or operons, etc.), without assuming a predefined distribution of fitness effects. Indeed, in the model, it is possible to reach similar fitnesses in many ways, by adjusting the number of genes, their loci, their lengths, or the intergenic distances, hence the total amount of non-coding DNA. In Aevol, genes are typically created by duplication-divergence (Kalhor et al., 2024), but they can also be deleted, and some may emerge *de novo*. Hence, the impact of a given mutation highly depends on the preexisting genome structure, which can in turn be indirectly selected (Knibbe et al., 2007a). Aevol therefore allows studying changes in size and structure of genomes in response to changes in population size and mutation rates.

Our experiments start from five "Wild-Type" (WT) lines, each having evolved for 10 million generations within a population of $1,024$ individuals and a mutation rate of 10^{-6} mutations per base pair for each mutation type: substitutions, small insertions, small deletions, duplications, deletions, translocations, and inversions. There is no underlying mutational bias: the insertion and deletion of bases are equally probable. The five WTs display stable genome structures (with small random variations, as exemplified by cases N_0 and μ_0 on Figure 3.1 and Figure 3.2) although they still slowly gain fitness by fixing rare favorable mutations (see case N_0 on Figure 3.5A). Their fitness and genomic characteristics are displayed in Section 3.4.2, Table 3.1. In our experiments, these WTs are used as founders of new populations, which are confronted with new evolutionary conditions for 2 million generations. In parallel, these same WTs were evolved in the same conditions they first evolved in, providing perfect control experi-

ments. We compare the fitness, genome size, and genome structure of populations that evolved in new conditions with those of the control populations. Finally, we repeat part of these experiments with WTs that evolved with either an insertion or a deletion bias to understand how an underlying mutational bias might impact our findings.

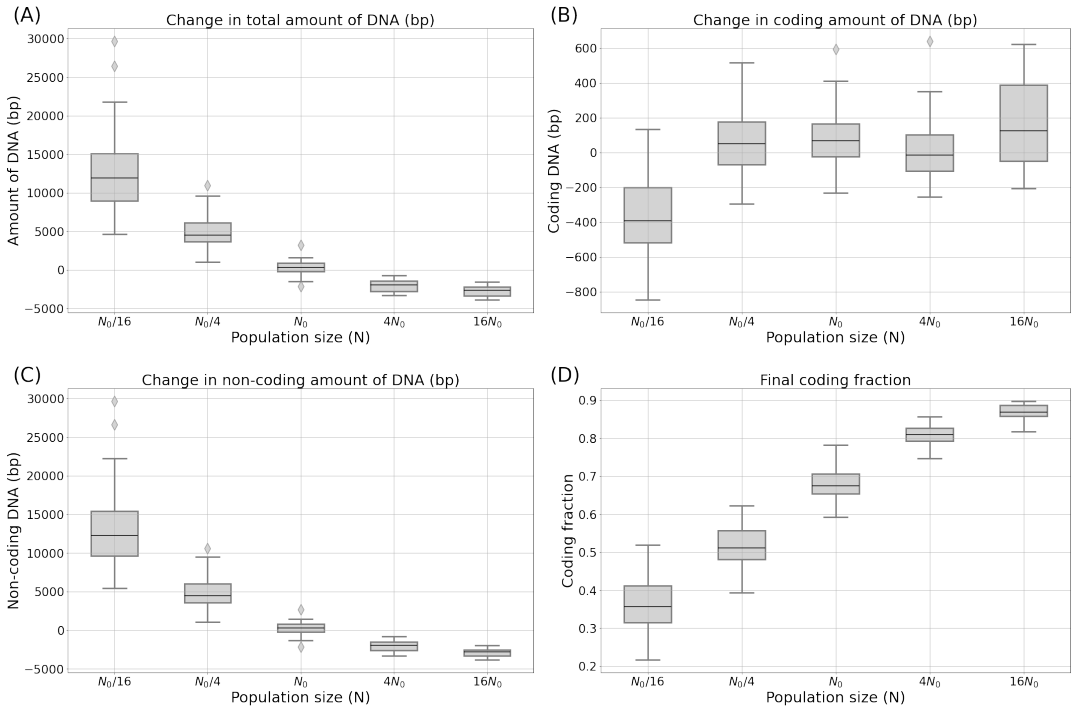

Figure 3.1: Total (A), coding (B) and non-coding (C) genome size variation, and final coding fraction (D), after 2 million generations. For each of the 5 WTs, 10 replicas were performed under a constant mutation rate ($\mu_0 = 10^{-6}$ per base pair for each type of mutation) with 5 different population sizes ($N_0 = 1,024$ being the control population size).

3.2.1 *Genome size evolution following a change in population size and mutation rate.*

Change in population size

In the absence of mutational bias, increasing the population size by a factor of 4 or 16 results in a reduction in the total genome size (see Figure 3.1A). Yet, this change does not impact the coding and non-coding parts of the genome proportionally: while the size of the coding compartment is barely affected (see Figure 3.1B), the non-coding genome size is greatly reduced (see Figure 3.1C). As a result, the coding proportion of the genome increases (see Figure Figure 3.1D). Conversely, reducing the population size by a factor of 4 or 16 increases the total genome size (Figure 3.1A) by increasing greatly the non-coding genome size (Figure 3.1C). In the extreme condition $N_0/16$,

the coding genome size is also slightly reduced (Figure Figure 3.1B). As a result, the coding fraction of the genome is drastically reduced (Figure 3.1D).

Change in mutation rate

In the absence of mutational bias, increasing the mutation rate drastically reduces the total genome size (see Figure 3.2A). Thus, at first sight, population size and mutation rate seem to have a similar effect on genome evolution. However, in the details, the effect of these two variables on genome structure appears to differ, as the reduction now occurs in both the coding and non-coding genomic compartments (see Figure 3.2B and C). Both are nevertheless not proportionally affected by the decrease in mutation rate, which affects more strongly the non-coding part of the genome, such that the final coding fraction of the genome increases with μ (see Figure 3.2D). Altogether, these results show that streamlined genomes, denser and shorter than their ancestors, can result from either an increase in population size or in mutation rate.

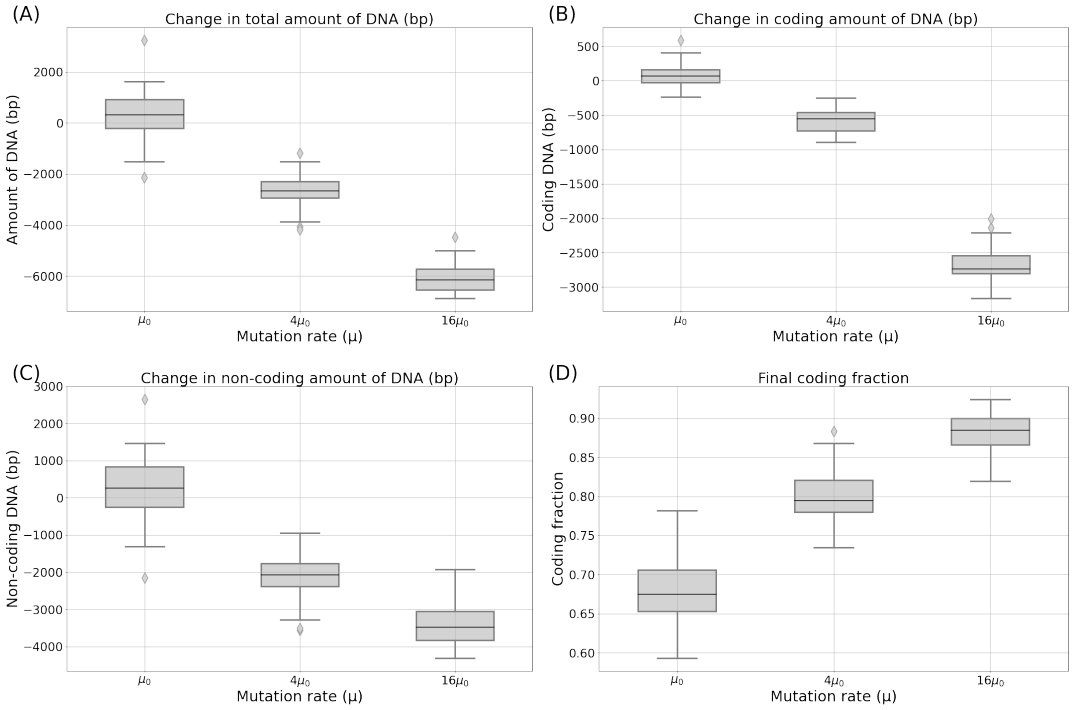

Figure 3.2: Total (A), coding (B) and non-coding (C) genome size variation, and final coding fraction (D), after 2 million generations. For each of the 5 WTs, 10 replicas were performed under a constant population size ($N_0 = 1,024$ individuals) with 3 different mutation rates: the control $\mu_0 = 10^{-6}$ mutations per base pair for each type of mutation, $4 \times \mu_0$ and $16 \times \mu_0$.

Notably, and despite the very different dynamics displayed in the two experiments, a 4-fold increase in N or in μ results in the same

final coding proportion of approximately 80%. The same is true for a 16-fold increase (88%). To further investigate this result, we conducted additional experiments to observe the combined effects of a simultaneous modification in both N and μ.

Linked effect of population sizes and mutation rates

Figure 3.3 shows the variation in the total amount of DNA, coding size, and non-coding size, as well as the variation in coding fraction for several combinations of changes in N and μ (note that, in the panels of Figure 3.3, the bottom line and the central column respectively correspond to the values presented in Figure 3.1 and Figure 3.2).

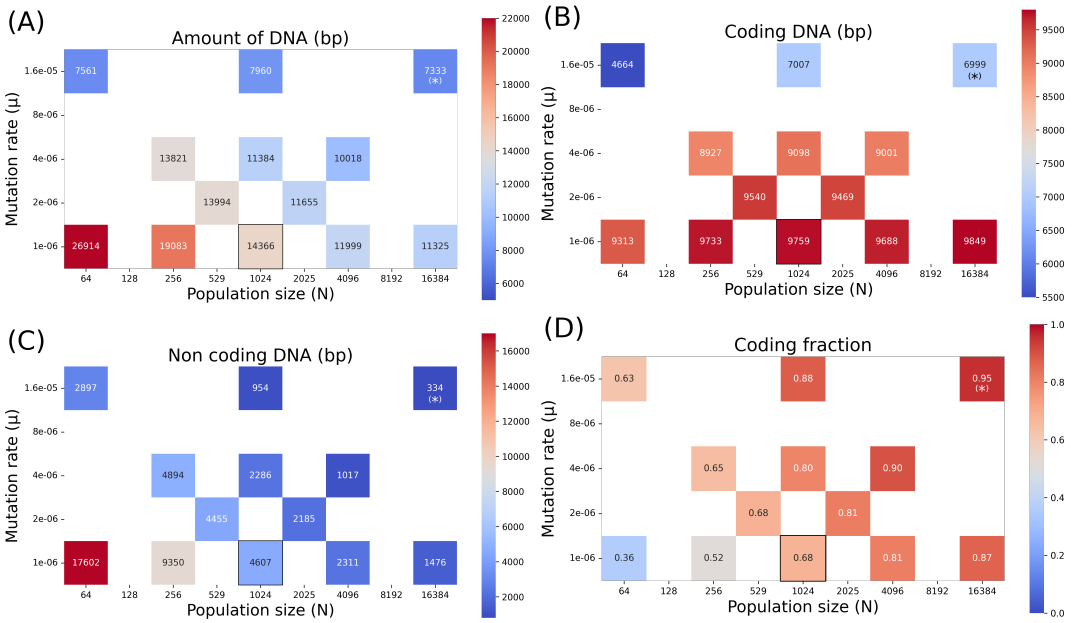

Figure 3.3: Amount of DNA (A), coding size (B), non-coding size (C) and coding fraction (D) for the different combinations of μ and N tested, after 2 million generations. For each of the 5 WTs, 10 replicas were performed for each tested set of conditions. Control conditions ($N = 1,024$ and $\mu = 1.10^{-6}$) are outlined in black. For the combination of both the highest mutation rate and the largest population size, only the median was tested due to computational limitations, which is indicated by a (*).

Overall, as N increases, the total amount of DNA decreases, whatever the value of μ (see Figure 3.3A). A higher μ also leads to a reduction in the total genome size, whatever the value of N. However, the effect of population size and mutation rate differ when considering the coding size of the genome: specifically, the coding size increases with N but decreases with μ (see Figure 3.3B). This is countered by the change in the non-coding size of the genomes (see Figure 3.3C), which strongly decreases with both N and μ and drives the overall change in genome size.

The interplay between N and μ results in a surprisingly constant coding fraction across the different constant values of $N \times \mu$ (see Figure 3.3D). Indeed, we observe that under constant $N \times \mu$, and although these two factors taken individually have changed in different proportions, the coding fraction remains constant: 80% when $N_0 \times \mu_0$ is multiplied by 4 compared to the control conditions, and 88% when $N_0 \times \mu_0$ is multiplied by 16 (see Figure 3.3D). Although the coding fraction does slightly vary (from 68% to 63%) for the most extreme tested configuration ($N_0/16$ and $16\mu_0$), the diagonal of constant $N_0 \times \mu_0$ also displays an almost constant coding fraction (Figure 3.3D).

However, strikingly, the total genome size as well as the coding and non-coding genome sizes vary greatly, even for similar coding densities (Figure 3.3B, C, and D). For densities of 63% and 65%, the total amount of DNA can be almost halved (from 13,821 bp to 7,561 bp) by going from $N_0/4$ and $4\mu_0$ to $N_0/16$ and $16\mu_0$ on the same diagonal of constant $N \times \mu$. Conversely, we can reach similar values of genome size (11,300 bp) despite important differences in the coding percentage (80% when μ is multiplied by 4, and 87% when N is multiplied by 16). Altogether, these results show that a large range of genome sizes and structures (here corresponding to coding densities) can result from a combined variation in both the population size N and the mutation rate μ.

3.2.2 *Mutational biases change the equilibrium genome size, but not the role of N and μ*

As genome sizes are generally thought to be heavily impacted by mutational biases, we control whether the effect of population size and mutation rate we observed is affected by either a deletion or an insertion bias. To this end, we evolved 5 Wild-Type organisms with either an insertion bias (twice as many duplications than large deletions), or a deletion bias (twice as many large deletions than duplications). The rates of all other types of mutations, as well as the sum of all mutation rates, are the same as in the previous experiments. As expected, the equilibrium genome sizes and coding proportions of these Wild-Types is affected by the balance between large deletions and duplications, with an average genome size of 11,623 bp in the presence of a deletion bias and 16,350 in the presence of a duplication bias (instead of 14,046 bp without any bias). The coding proportion is also affected: 0.78 and 0.61 respectively, instead of 0.69. This shows that the genome size and structure are, as expected, strongly influenced by the underlying mutation biases (Kuo and Ochman, 2009).

We then confronted the median (in terms of genome size) WT of each condition to changes in population size (multiplied or divided by 4) or mutation rate (multiplied by 4) for 10 replicas. Similarly to what is observed without bias, an increase in N reduces the non-coding

genome size only, while an increase in μ reduces both the coding and non-coding genome (see Figure 3.4). Notably, a decrease in N increases the non-coding genome size even in the case of a deletion bias, although an insertion bias greatly amplifies this effect. As a result, and despite the strong mutational biases, we observe that multiplying either the population size or the mutation rate by the same factor leads to a genome compaction in similar proportions (the final coding fraction being 0.85 vs. 0.88 in the case of the deletion bias, and 0.78 vs. 0.77 in case of the insertion bias respectively). Therefore, although mutational biases influence the equilibrium genome sizes and structures, they do not fundamentally change how the genomes react to variations in population size or mutation rate. In other words, our simulations show that mutational biases only determine the equilibrium set point around which population size and the overall mutation rate then modulate the genome size and structure. Similar experiments were run with biases in InDels and are presented in the Section B.2.

Supplementary Materials are included in the Appendix B.

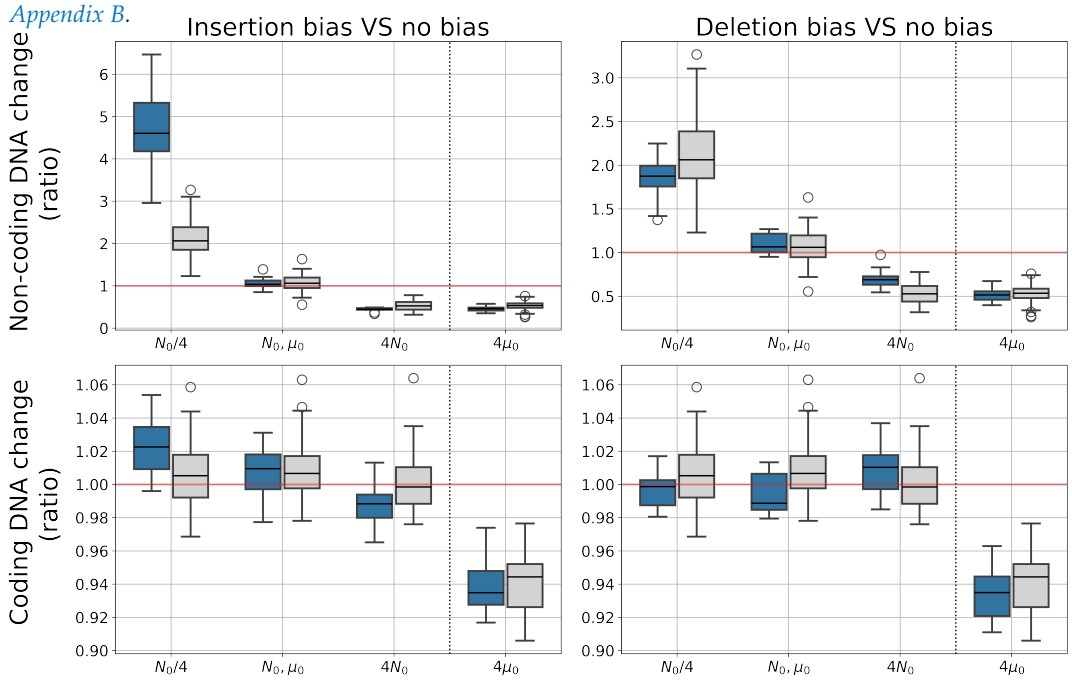

Figure 3.4: Change in coding and non-coding genome sizes in reaction to changes in N or μ for the different mutational biases. Blue boxes show the results with a mutational bias (left: insertion bias, right: deletion bias), and gray boxes show the results without mutational bias. Depicted values are the ratio of the coding/non-coding sizes at the final generation over the value at generation 0.

3.2.3 *Robustness selection as the explanatory mechanism*

We observed that two distinct processes, triggered by an increase in either population size or mutation rate, can lead to genome size

reduction in our experiments. However, both have different effects on coding and non-coding sequences: while an increased μ reduces both the coding and non-coding genome sizes, increasing N reduces only the non-coding genome size.

We propose that these observations can be explained by an interplay between selection for phenotypic adaptation to the environment (hereafter called *direct selection*), and selection for replicative robustness (hereafter referred to as *indirect selection*). More specifically, we define the replicative robustness of an individual as its ability to transmit its fitness to its offspring. It hence corresponds to the proportion of offspring that did not acquire new deleterious mutations. This depends both on the number of mutations occurring at replication (which in turn depends on genome size) and on the probability for a given mutation to be deleterious (usually called mutational robustness (Wilke and Adami, 2003)), which depends on the intertwining between the kind of mutation and the genomic architecture. In our case, wild-type organisms are very well adapted to their environment, thus most mutations will be deleterious if they affect the coding part of the genome. This is particularly true for chromosomal rearrangements, which can affect large genomic segments (Knibbe et al., 2007a; Banse et al., 2024b). Conversely, beneficial mutations are extremely rare. We therefore approximate the robustness of our organisms by measuring the proportion of their offspring that have the exact same fitness, *i.e.* that underwent no mutations or only neutral mutations.

A more robust individual has more chances to pass on its genomic information accurately than a less robust one, thus enabling its lineage to better maintain its fitness in the long term and to outcompete other lineages in which deleterious mutations would accumulate at a higher rate. This results in an indirect selection for replicative robustness. We recall that replicative robustness depends both on the probability for a given mutation to be neutral (hence on the fraction of non-coding sequences in the genome) and on the mean number of mutations undergone by the genome at each generation (hence on the genome-wide mutation rate). Here, while the per base mutation rate is constant within each experiment, the total amount of DNA, and hence the genome-wide mutation rate, varies and can thus be indirectly selected. By contrast, direct selection depends only on the content of the coding compartment, the size of which is likely to be positively correlated with the level of phenotypical adaptation (at least in our model). As a result, indirect selection for robustness favors shorter genomes with a lower coding fraction, while direct selection for phenotypical adaptation maintains or even increases the coding size of the genome.

The efficacy of both direct and indirect selection increases with population size, since some deleterious mutations that were quasi-neutral for a low N can become effectively counter-selected in the context of a high N, changing the balance of beneficial *vs* deleterious fixed

mutations. To quantify this effect, we measured the robustness of the individuals at time 2,000,000 in the simulations without mutational biases. Figure 3.5A and Figure 3.5B show that the increase in selection efficacy induced by the increase in population size indeed induces both an increase in fitness (due to direct selection) and an increase in replicative robustness (due to indirect selection). In terms of genomic structure, a more efficient direct selection (*i.e.* a weaker random drift) is thus expected to increase the coding genome size, and a more efficient indirect selection is expected to decrease the overall genome size. The combination of both these effects leads to a decrease in the non-coding genome size, and maintenance of the coding genome size, as exemplified by Figure 3.1B and C. Conversely when the population size is reduced, the increased drift leads to the loss of coding sequences and inflation of the non-coding compartment (Figure 3.1B and C). This reorganization of the genome structure is associated with a loss in robustness (Figure 3.5B).

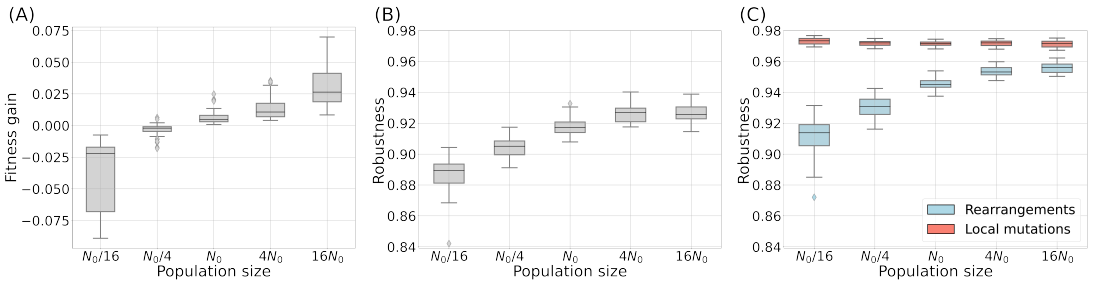

Figure 3.5: Fitness gain (A) and Robustness (B: overall and C: by mutation type) at the end of the simulations, for different population sizes N and without mutational biases. Robustness is defined as the proportion of neutral offspring. The mutation rate is fixed to 10^{-6} per base pair for each type of mutation.

In Aevol, genomes undergo different types of mutations that can be roughly grouped into local mutations (substitutions, InDels) and chromosomal rearrangements (duplications, deletions, inversions, translocations). Both kinds of events don't have the same effect on robustness. Figure 3.5C shows the change in robustness induced by the different types of events. It shows that the loss and gain in robustness are driven by chromosomal rearrangements. In contrast, local mutations (substitutions and InDels) do not have a significant effect on robustness.

In the case of an increased mutation rate, things are very different: a sudden increase in μ results in an immediate drop in robustness at the beginning of the experiments (Figure 3.6A). As the proportion of offspring that bears mutations rises with μ, we go from an initial robustness of 92% for μ_0, to 71% for $4\mu_0$, and only 26% for $16\mu_0$. In these new conditions, organisms are no longer able to transmit their genome to the next generation without deleterious mutations,

and thus the indirect selection for robustness becomes temporarily stronger than the direct selection for phenotypical adaptation. Indeed, features that would not be accurately inherited cannot be selected. This indirect selection for robustness leads to the fixation of mutations that drastically decrease genome size, even at the cost of a loss of fitness for the individuals (see Figure 3.6B): the only lineages that survive in the long term are those that have undergone a decrease in genome size, allowing them to reduce their per-genome mutation rate, thus regaining some robustness (see Figure 3.6C). Once the robustness has increased sufficiently, direct selection for phenotypical adaptation can resume and the fitness starts to increase again (see Figure 3.6B). Interestingly, organisms manage here to continue to lose some coding base pairs while increasing their fitness, probably thanks to global genome restructuring allowing for a more compact encoding of the phenotype, for example through overlapping genes. This dynamic is very different from when N is increased (and so the initial robustness is unaffected), as shown by Figure 3.6D, E, and F.

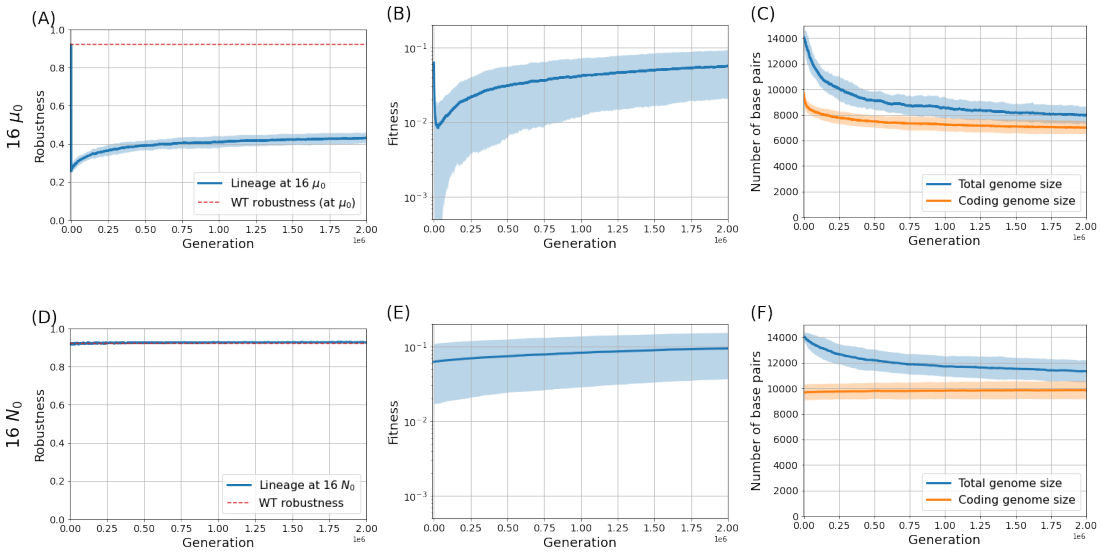

Figure 3.6: Robustness, fitness, and genome architecture across generations for $\mu = 1.6 \times 10^{-5}$ (16 μ_0) per base pair for each mutation type and $N = 1,024$ (N_0) (top row, panels A, B, and C) and $N = 16,384$ (16 N_0) and $\mu = 1 \times 10^{-6}$ (μ_0) per base pair for each mutation type (bottom row, panels D, E and F). Lines represent the mean values across the 50 simulations, and the shaded areas represent the standard deviations.

Notably, robustness does not reach values as high as that observed before the increase in mutation rate and stays below 50%. Indeed, the genome size could not be divided by 16 while keeping a good enough phenotypical adaptation, and the selection for phenotypical adaptation becomes stronger than the selection for robustness as soon as some organisms can pass on their genomic information reliably enough.

The interplay between direct and indirect selection can therefore explain both types of genome size reduction: affecting both coding and non-coding compartments (although not proportionally) when caused by an increased mutation rate, and restricted to the non-coding compartment when caused by an increased population size.

3.3 DISCUSSION

We found that, in our experiments, genome size reduction can be caused by an increase in population size, mutation rate, or both, even in case of mutational biases. These two factors can nevertheless be distinguished, as they have different effects on the coding and non-coding sequences of the genome. Their combination in various proportions can create a broad range of alternative patterns of genome size and coding density. In particular, by playing independently on mutation rate and population size, our model can reproduce the two extreme but different cases of genome size reduction that are seen in some endosymbionts and cyanobacteria. As an example, *Prochlorococcus marinus* is known to have lost both some parts of its coding and non-coding genome, although in different proportion such that its coding density has increased (Dufresne et al., 2005; Batut et al., 2014; Giovannoni et al., 2014). In our model, this would correspond to a population undergoing an increase in population size and a slight increase in mutation rate, which is coherent with the scientific literature on *Prochlorococcus marinus* (Hu and Blanchard, 2008; Marais et al., 2008), although the large effective population size of this species has been recently debated (Chen et al., 2022; Filatov and Kirkpatrick, 2024). On the other hand, *Buchnera aphidicola* has conserved its coding proportion but greatly reduced its total genome size (Moran and Mira, 2001), which could be explained in our model by an increase in mutation rate and a decrease in population size, in similar proportions. This suggests that indirect selection for shorter genomes through robustness selection could be a key factor playing on genome evolution (Wilke et al., 2001; Gabzi et al., 2022), and especially on the evolution of genome size and structure.

Our observations confirm those made by Lynch and Conery (2003), namely that an increased genetic drift, here associated with a decreased population size, increases the genome size. Our results also point towards an equilibrium genome size: a sufficient number of genes makes it possible to fine-tune the phenotype to the environment, but the genome also has to be short enough to prevent the degeneration caused by an excess of chromosomal rearrangements (Knibbe et al., 2007a; LaBar and Adami, 2020). Increasing the mutation rate or the population size displaces this equilibrium toward shorter genomes, either through a more efficient genome purification of non-coding sequences (when increasing N) or a loss of both coding and non-coding sequences to recover a minimal level of robustness (when increasing

μ). Of course, mutational biases (regarding the balance between insertions and duplications versus deletions) also play an important role in determining the equilibrium genome size. In particular, deletion biases have been suggested as one main reason explaining why bacterial genomes remain small (Mira et al., 2001). However, we show here that, because of the indirect selection for robustness, a deletion bias is not needed to prevent a runaway inflation in the size of genomes. Instead, selection for robustness provides a counteracting force that increases with genome size, eventually offsetting any underlying bias in favor of insertions or duplications. Importantly, this indirect selection was not postulated in the model but emerged spontaneously in the simulations.

We propose an evolutionary mechanism consisting of a trade-off between direct selection for phenotypical adaptation and indirect selection for replicative robustness. In this respect, mutations appear to be a weak selective force, as pointed out by Lynch (2007b). However, the emphasis was previously on the mutational targets contributed by genomic features, such as introns. Here, we emphasize another aspect, which seems to have been overseen thus far: any non-functional DNA represents an additional target for initiating macroscopic mutational events that can eventually impact the coding genome. This mechanism requires no additional hypotheses and is very general. It should therefore be pervasive in the living world.

Sung et al. (2012) have observed that, in real populations, the mutation rate scales negatively with both the population size and the amount of coding DNA. They propose that this is a consequence of selection for lower per-base mutation rates induced by the amount of coding DNA. Here, thanks to the use of fixed mutation rates, we have shown that the mutation rate can select the amount of DNA, including both the coding and non-coding compartments. This points towards the per-genome mutation rate being the relevant value, which can evolve due to changes in genome size and per-base mutation rate. This calls for further experiments in which both the genome size and the per-base mutation rate would be allowed to evolve, to study their relative speed of adaptation and their contribution to the variation of the per-genome mutation rate.

Although our main focus was on the final equilibrium reached by the populations after a change in N or μ, our observations are broader than the end equilibrium as we can observe the temporal dynamics (Figure 3.6 and S3 to S15). In particular, we observe that, when the mutation rate increases strongly, the fitness immediately drops drastically (Figure 3.6B). This can be related to an error-threshold crossing mechanism (Eigen, 1971; Takeuchi and Hogeweg, 2007; Boer and Hogeweg, 2010): individuals can no longer pass on to their descendants all the

information contained in their genome. They therefore lose fitness, and the lineage that survives in the long term is the one where genomes greatly reduced in size in the early phase of the experiment, thus reducing the number of mutations per replication event and finally reaching a point at which the information can be passed on reliably. The detailed aspects of these temporal dynamics could be the focus of future work. Indeed, it has been shown that genome reduction in endosymbionts occurred very quickly after the endosymbiosis became effective (Moran, 2003; Wernegreen, 2015), which is also what we observed in our data (Figure 3.6).

In our experiments, $N \times \mu$ stands out as a determining factor of some (although not all) aspects of genome structure, as isoclines of identical $N \times \mu$ values display similar coding densities, even in the case of reduced genomes or mutational biases. Understanding this invariant is one of the most exciting perspectives opened by our work. Its importance has already been highlighted by Schaack (2006) in organelles, but our results suggest that this joined factor of drift and mutational pressure is a determinant of genome evolution throughout the tree of life. Notably, there is a small variation in coding fraction along N isoclines, which could be due to our use here of population size (N) instead of effective population size (N_e). Indeed, in our setup, the competition is local and thus N_e is slightly greater than N, but this relationship is not linear (see Section B.1). Further versions of the model could rely on various measures of the effective population size to reach more accurate predictions, but we believe that our results can be interpreted nonetheless, as changes in population size and in effective population size are very similar over the range of population sizes tested here (see Section B.1).

Supplementary Materials are included in the Appendix B.

In order to allow for a fair quantitative comparison between the effect of mutation rates and population size, the amplitudes of the variations applied to the two parameters were similar in our experiments. In biological species, the range of variation in mutation rates is much narrower than the range of variation in effective population size, as shown by Lynch et al. (2023). Hence, given our explanatory mechanism, the observed range of variations in genome size is likely to be driven mainly by changes in N. However, our results show that μ and N do not play an identical role. Indeed, variations in N change solely the non-coding size of the genome, while the variation in μ impacts both the coding and the non-coding sizes. Therefore, even a small variation in μ compared to a variation in N could be significant in determining genome architecture trajectories. This highlights that the correlation of N and genome size is not enough to understand genome evolution and that μ, as well as any underlying mutational bias, also needs to be taken into account as a determining factor.

In this paper, we specifically focused on the effect of the variation in population size and mutation rates on genome size. Of course, it does not imply that the mechanism we identified is the only one, and various additional ones can also impact genome size evolution. For instance, there can be a limitation in available resources for nucleotide production, constraining the total genome size (Ngugi et al., 2023). In the case of endosymbiosis, exchanges can also happen between the host and the endosymbiont genomes, hence contributing to its streamlining (Bock, 2017). Recombination could also further complicate the picture by adding a new type of mutation with unexpected interactions. More importantly, mobile genetic elements, and Transposable Elements (TE) in particular, are often proposed as one of the main drivers of genome expansion (Marino et al., 2024), especially in populations with small effective population sizes that could not eliminate them efficiently due to the low selective pressure (Lynch and Conery, 2003). TE invasions have been shown to increase dramatically genome size in eukaryotes (Kidwell, 2002; Oggenfuss et al., 2021), although Dijk et al. (2022) have demonstrated that they can also lead to streamlining in prokaryotes because genome reduction prevents their invasion. We did not test their impact here, but our results show that the effect of the variations in population size and mutation rate is conserved, even in case of a strong insertion bias (Figure 3.4 and Figure B.2). This enables us to conjecture that mobile elements would change the equilibrium genome size (as observed in our simulations, Figure 3.4 and Figure B.2), and probably drastically increase the variance of observed sizes, but that they are unlikely to change the response of genome size evolution to changes in μ or N. This remains however to be tested.

To conclude, our experiments show that genome size reduction can occur in two very different conditions for bacteria. On the one hand, a very large population size promotes a more efficient selection in the face of random drift, which in turn enhances the robustness of genomes by decreasing their non-coding load. This corresponds to streamlining and leads to genomes with a high coding density. On the other hand, a higher mutation rate results in an instantaneous decrease in the robustness of genomes in the entire population, making the selection for robustness transiently stronger than the selection for phenotypical adaptation. The genome then shrinks rapidly, with both coding and non-coding sequences being discarded until a new robustness equilibrium is reached, all this at a substantial initial cost in phenotypical adaptation. This corresponds to a decaying genome and is compatible with empirical observations in endosymbiotic bacteria (Moran, 2003). Strikingly, this remains true even in the presence of a mutational bias. Although the model that we propose here, of a balance between selection for robustness and selection for phenotypical adaption, can explain the tendencies we observe and the final genome

structures in our populations, further work is needed to understand the transient regimes and the mechanisms behind the constant coding fraction along the $N \times \mu$ isoclines.

3.4 MATERIALS AND METHODS

3.4.1 *The Aevol framework*

Aevol (Knibbe et al., 2007a; Banse et al., 2024b) is an individual-based forward-in-time simulation software that has been specifically designed to study the evolution of genome structure. It emulates a population that is composed of a fixed number of individuals on a grid (Figure 3.7A). Each individual owns a double-stranded circular genomic sequence, composed of 0s and 1s. To compute the phenotype, sequences on the genome are recognized as promoters and mark the start of transcription, which stops when a sequence able to form a hairpin structure is encountered. On RNAs, Shine-Dalgarno-like sequences followed by a START codon mark the beginning of translation. The RNA sequence is then read 3 bases at a time until a STOP codon is encountered on the same reading frame. An artificial genetic code allows for each sequence of codons to be converted into a mathematical function, and the sum of all functions encoded on the genome defines the phenotype of the individual (Figure 3.7B). The distance between this function and a target function, which represents the ideal phenotype in the specified environment, gives the fitness of the individual with a scaling factor k that tunes the strength of the selection. A detailed explanation can be found on the dedicated website www.aevol.fr.

All individuals are replaced at each generation following a spatialized Wright-Fisher model. The number of descendants of each individual depends on its fitness difference with its neighbors. At each reproduction event, point mutations or genomic rearrangements can occur (Figure 3.7C). They create diversity in the genomes, hence in the phenotypes, and allow the genome size and structure to change. These changes can be neutral or not, depending on whether mutations alter coding and/or non-coding sequences. These changes do not have a predefined effect on the fitness of the offspring as their genomes will be decoded thereafter, thus the model does not impose an *a priori* genome structure and allows us to study the evolution of genome architecture in various experimental conditions.

The mutation rate (in bp^{-1}) is set for each type of mutation independently. When all mutation rates are equal, there is in an equal probability of losing or gaining base pairs. The size distribution of InDels is uniform in $[1, 6]$, and the size distribution of large deletions and duplications is uniform in $[1, L]$ (with L the genome length).

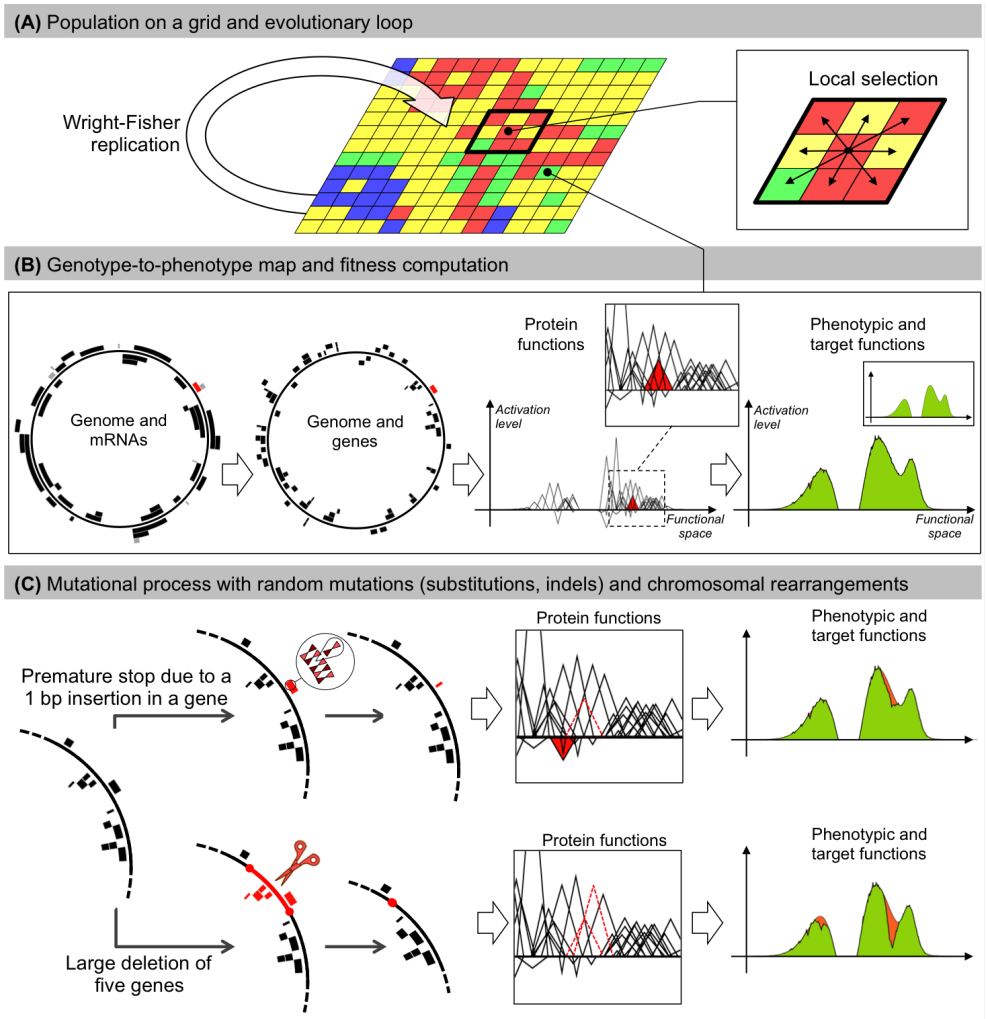

Figure 3.7: The Aevol model. (A) Individuals are distributed on a grid. At each generation, the whole population replicates according to a Wright-Fisher replication model, in which selection operates locally within a 3×3 neighborhood. **(B)** Each grid cell contains a single organism described by its genome. Genomes are decoded through a genotype-to-phenotype map with four main steps (transcription, translation, computation of protein functions, and computation of the phenotype). Here, for illustration purposes, a random gene and the corresponding mRNA are colored in red. The red triangle represents the function of this gene in the mathematical world of the model. The phenotypic function is calculated by summing all protein functions. The phenotype is then compared to a predefined target (in green) to compute the fitness. The individual presented here has evolved in the model during $500,000$ generations. **(C)** Individuals may undergo mutations during replication. Two example mutations are shown: A small insertion (top) and a large deletion (bottom). Top: A 1 bp insertion occurs within a gene. It causes a frameshift, creating a premature stop codon. The ancestral function of the gene is lost (dashed triangle) and the truncated protein has a deleterious effect (red triangle). This leads to a greater divergence between the phenotype and the target (orange area on the phenotype). Bottom: The deletion removes five genes. The functions of two of them can be seen in the box (dotted triangles). This results in a large discrepancy between the phenotype and the target (orange area on the phenotype).

WT id	Fitness (arbitrary unit)	Total genome size (bp)	Coding size (bp)	Non-coding size (bp)	Coding fraction
1	0.014903	13,599	9,395	4,204	0.69
2	0.103795	13,660	8,828	4,832	0.65
3	0.128472	14,171	9,507	4,664	0.67
4	0.035369	14,507	10,003	4,504	0.69
5	0.029588	14,290	10,644	3,646	0.74
Average	0.0624254	14,045.5	9,675.4	4,370	0.69

Table 3.1: Characteristics of the 5 Wild-Types at the start of our experiments.

3.4.2 *Experimental design*

Wild Types

In order to observe changes in genome architecture induced by changes in the population size and/or mutation rates, we begin our experiments from pre-evolved organisms, which are called "Wild Types" (WT). Having already evolved for millions of generations under constant conditions, WTs are very stable in genome structure and well adapted to their environment (although the fitness never stops increasing). 5 different WTs were used for our experiments, all having evolved for 10 million generations at the basal conditions of $N_0 = 1,024$ individuals and a mutation rate of $\mu_0 = 10^{-6}$ mutations per base pair per generation for each type of mutations (point mutations, small insertions, small deletions, inversions, duplications, large deletions, and translocations). Importantly, in this experiment, all types of mutations are equally probable: there is no mutational bias towards the insertion or deletion of base pairs. Bacterial populations are very large and cannot be directly modeled owing to computational load. We hence limit the population sizes in our experiments, but compensate by increasing the mutation rates such that the $N \times \mu$ parameter is of the same order of magnitude as for real bacterial populations. Finally, to limit the effect of drift, we used a selection strength $k = 1,000$, which is relatively high and guarantees an efficient selection. The fitnesses and genome structures of the WTs are listed in Table 3.1.

Experimental conditions

A range of population sizes increases or decreases and mutation rates increases, as well as some combinations of both, are tested. All conditions are listed in Table 3.2 below. For each combination of conditions, 10 replications of each of the 5 WTs are run. Initial

populations are always clonal: all individuals are identical to the specific WT used for the run.

Population size	Mutation rate (per base pair, per mutation type)	$N \times \mu$ product
64 ($N_0/16$)	10^{-6} (μ_0)	$1/16 N_0 \times \mu_0$
256 ($N_0/4$)	10^{-6} (μ_0)	$1/4 N_0 \times \mu_0$
1024 (N_0)	**10^{-6} (μ_0)**	$N_0 \times \mu_0$
529 ($\approx N_0/2$)	2×10^{-6} ($2 \times \mu_0$)	$\approx N_0 \times \mu_0$
256 ($N_0/4$)	4×10^{-6} ($4 \times \mu_0$)	$N_0 \times \mu_0$
64 ($N_0/16$)	16×10^{-6} ($16 \times \mu_0$)	$N_0 \times \mu_0$
2,025 ($\approx 2 \times N_0$)	2×10^{-6} ($2 \times \mu_0$)	$\approx 4 N_0 \times \mu_0$
4,096 ($4 \times N_0$)	10^{-6} (μ_0)	$4 N_0 \times \mu_0$
1,024 (N_0)	4×10^{-6} ($4 \times \mu_0$)	$4 N_0 \times \mu_0$
4,096 ($4 \times N_0$)	4×10^{-6} ($4 \times \mu_0$)	$16 N_0 \times \mu_0$
16,384 ($16 \times N_0$)	10^{-6} (μ_0)	$16 N_0 \times \mu_0$
1,024 (N_0)	16×10^{-6} ($16 \times \mu_0$)	$16 N_0 \times \mu_0$
16,384 ($16 \times N_0$)	16×10^{-6} ($16 \times \mu_0$)	$256 N_0 \times \mu_0$

Table 3.2: Experimental conditions tested. The control condition is in bold. Note that, as the simulations take place on a squared grid, population sizes could not be exactly divided or multiplied by 2.

Data analyses

To analyze the simulations, we reconstruct the ancestral lineages of the final populations. To this end, simulations are run for 2,100,000 generations, and we identify all the ancestors of a random individual of the final population. We then study the data from generation 0 to generation 2,000,000 and ignore the last 100,000 to ensure that the final population has coalesced and that we study the lineage of the whole final population.

On this lineage, we retrieve the fitness, coding and non-coding genome size at each generation, as well as the replicative robustness every 1,000 generations. The replicative robustness is measured as the proportion of the offspring of an individual that has the exact same fitness as its parent, *i.e.* that underwent no mutation at all, or only purely neutral mutations. To estimate replicative robustness for

a given individual of the lineage, we generate 10,000 offsprings and compare them to their parent.

To compare experimental conditions, we retrieve the individual at generation 2,000,000 in each lineage. This individual is the common ancestor of the final population (at generation 2,100,000), thus ensuring that its genome structure has been conserved by evolution. A visualization of the temporal lineage data (fitness, coding fraction and total, coding, and non-coding genome sizes) for the 50 replicas of each experimental condition is provided in the Section B.3 (Figure B.3 to Figure B.15).

Supplementary Materials are included in the Appendix B.

Effect of mutational biases

As it is often assumed that mutational biases – towards deletions for bacteria and towards insertions for eukaryotes – are very important for genome size evolution (Petrov, 2002), we also tried to confront our experiments to the impact of mutational biases. We tested four mutational biases: twice as many large deletions than duplications, twice as many small deletions than small insertions, twice as many duplications than large deletions, and twice as many small insertions than small deletions. In all cases, the sum of all mutation rates is conserved, such that the overall mutational pressure is the same as in the previous experiments.

For each mutational condition, 5 Wild-Types evolved for 10,000,000 generations. Then, the median-sized WT of each mutational condition was extracted and confronted with either an increase or decrease in population size ($4 \times N_0$, $N_0/4$) or an increase in all mutation rates proportionally ($4 \times \mu_0$ – note that, in case of bias, μ_0 may be different for the different types of mutation) for 2,100,000 generations. By extracting the ancestor of the lineage at generation 2,000,000, we could compare these experiments to the control conditions (where the population size and mutation rates remained stable for 2,100,000 generations).

Data availability

The code of Aevol is available on GitLab at `https://gitlab.inria.fr/aevol/aevol`. WTs sequences to reproduce the experiments, as well as the full lineages data and robustness data, are available on Zenodo: `https://doi.org/10.5281/zenodo.10669479`.

3.5 ACKNOWLEDGMENTS

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR), under grant ANR-20-CE02-0008 (NeGA project). J.L., G.B. and N.L. would like to thank the Rhône-Alpes Institute for Complex Systems (IXXI) for funding. All authors thank the

Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr), for computational support. The authors would like to thank Laurent Duret and David P. Parsons for fruitful comments on the manuscript.

4

STRUCTURAL MUTATIONS SET AN EQUILIBRIUM NON-CODING GENOME FRACTION

FOREWORD

The following work is published in *BioRxiv* (Luiselli et al., 2025a) and authored by Juliette Luiselli and Paul Banse (co-first authors), Olivier Mazet, Nicolas Lartillot, and Guillaume Beslon. The paper itself has not been altered to stay true to the citation. Supplementary Materials have been added as the Appendix C.

Chapter 2 showed that the presence of chromosomal rearrangements creates a unique evolutionary force that prevents an uncontrolled genome size growth. Chapter 3 showed that this force is modulated by the population size and the mutation rate. Indeed, there is a selection for robustness to chromosomal rearrangements that creates a pressure for genome size reduction, as larger genomes are more prone to undergo chromosomal rearrangements than smaller genomes. As such, the population size modulates the strength of selection and the mutation rate the cost of being bigger. These conclusions were drawn from experiments with Aevol, but the mechanism by which chromosomal rearrangements drive genome size regulation is likely to be more general than this framework: the fact that chromosomal rearrangements act on *segments* of DNA instead of a single base is a biological reality that leads to an increased mutational hazard for bigger genomes. Indeed, while an additional non-coding base pair is basically invisible to selection against substitutions, it increases the risk of double-strand breaks and thus of large chromosomal rearrangements that could affect genes.

In the present chapter, we present a mathematical model of genome evolution that integrates chromosomal rearrangements and the possibility of a second-order selection (selection on the potential offspring, at constant phenotypical adaptation). This allows for spontaneous selection for robustness to chromosomal rearrangements and a pressure towards genome size reduction. The model also shows that this pressure is counterbalanced by an intrinsic bias in the neutrality of mutations that pushes towards an increase in genome size. As a result, there is an equilibrium non-coding genome size for any given combination of effective population size N_e, mutation rate μ, and coding genome structure (size and number of contiguous segments). Interestingly, the model predicts that the equilibrium coding *fraction* is determined by the product $N_e \times \mu$ and does not depend on the coding *size*, as was the case in Chapter 3.

4.1 INTRODUCTION

Genome size varies greatly throughout the Tree of Life: from 10^5 base pairs (bp) for some bacteria (Riley et al., 2017), to more than 10^{11} bp for some plants (Pellicier et al., 2010). Coding sequences contribute to this variation through adaptive changes, but some parts of the genome seem devoid of phenotypic function and yet are highly variable in size (Liu et al., 2013). While non-coding DNA contains functional sequences, including regulatory regions (Rinn and Chang, 2012), large stretches seem to bear no function whatsoever. This "junk" DNA (Ohno, 1972; Doolittle, 2013; Palazzo and Gregory, 2014; Fagundes et al., 2022) is ubiquitous in all domains of life, regardless of genome sizes (Ahnert et al., 2008; Gil and Latorre, 2012). However, there is currently no consensus on the reasons behind the existence and maintenance of junk DNA (Fagundes et al., 2022).

In this work, we address the determinants of the amount of non-coding non-functional DNA. Several hypotheses have been proposed to address these issues, notably reviewed in Blommaert, 2020.

In adaptive hypotheses, genome size itself is under selection due to its phenotypic impact on *e.g.* nucleus size or replication time (Malerba et al., 2020). In this view, genome size would be selectively limited (Kang et al., 2015; Bales and Hersch-Green, 2019). Furthermore, the position of genes relative to each other or the centromere influences their expression (El Houdaigui et al., 2019). As such, it represents a potential selective pressure on the amount of intergenic DNA (Freeling et al., 2015). However, it can be argued that the variation in the proportion of non-coding DNA between species might be too high to be explained by these mechanisms (Petrov, 2002; Blommaert, 2020). More fundamentally, there is little direct evidence that selection induced by these phenotypes is strong enough to modulate the fate of mutations changing genome size.

Non-adaptive hypotheses have also been developed to decipher the mechanisms by which non-coding DNA could vary and stabilize. First, mutational explanations emphasize the impact of mutational patterns on the long-term evolution of genome size. In particular, the mutational equilibrium hypothesis (MEH)(Petrov, 2002) suggests that two different mutational biases of opposite directions — a negative bias on short indels and a positive long insertion/deletion bias that decreases with genome size — could mechanistically explain the existence of an equilibrium genome size. The equilibrium itself would be modulated between species by the variation in the strength of those biases.

The mutational hazard hypothesis (MHH)(Lynch and Conery, 2003), on the other hand, proposes an explanation in terms of *fixation* biases acting on mutations, related to second-order selective effects. According to the MHH, the non-coding genome expands by mutation,

drift, and the insertion of selfish elements. However, this expansion increases the number of targets for deleterious mutations — *e.g.*, such as gain-of-function mutations or loss of accurate splicing (Lynch, 2007b). In other words, non-coding DNA presents a mutational liability. As a result, genome expansion entails a slight selective cost, which could provide a sufficient force counteracting the growth of genome size (Lynch and Conery, 2003; Lynch, 2007b). The efficacy of this selective force is inversely related to effective population size, while the intensity of the force itself is directly proportional to the mutation rate. Thus, genome size should be inversely correlated with each of these two factors (Lynch, 2007b; Knibbe et al., 2007a).

Both theories receive support from some observations (Yi and Streelman, 2005; Kelkar and Ochman, 2012; Smith et al., 2013; Canapa et al., 2015; Sung et al., 2016; Mueller and Jockusch, 2018; Luiselli et al., 2024), but are also challenged by others (Ai et al., 2012; Sloan et al., 2012; Mohlhenrich and Mueller, 2016; Marino et al., 2024). Importantly, they are not mutually exclusive, as a combination of mutational biases and second-order selective effects due to the mutational liability of non-functional DNA could act together to determine an equilibrium genome size. This calls for an integrated explanation for what determines the amount and variation of non-coding DNA in genomes. In this direction, previous studies on simulated data (Banse et al., 2024b; Luiselli et al., 2024) suggest that structural mutations, *i.e.* chromosomal rearrangements or more generally any mutation larger than 50 bp, could be a key element linking both the MEH and the MHH. Indeed, structural mutations significantly affect genome size and are also a huge mutational liability in themselves due to their large-scale effect. It is therefore essential to examine their impact on genome size evolution.

Here, we propose a minimal probabilistic model of genome evolution, with the following assumptions: (1) genomes are composed of a coding component made of essential genes and a non-coding component that has strictly no phenotypic effect; (2) mutations occur at random uniformly over the genome. Our analysis of this model reveals non-trivial patterns: (1) structural mutations do not have the same probability of being neutral and this results in a trend towards increasing genome size; (2) as larger genomes are more susceptible to double-strand breaks — and thus to structural mutations —, changes in genome size change the probability of future, possibly lethal, structural mutations; (3) this increased risk of having a larger genome modulates the fixation probability of structural mutations in a way that favors deletions over insertions or duplications. Together, these mechanisms ensure a stable evolutionary equilibrium for non-coding genome size. More precisely, the equilibrium non-coding *fraction* depends on the product of the effective population size and mutation rate of a species ($N_e \times \mu$), while the non-coding *size* is determined by

this product plus the coding architecture of the genome (coding size and distribution). Notably, the equilibrium is a robust outcome of our model, even in the presence of mutational biases towards insertions or deletions: arbitrary mutational biases merely shift the equilibrium.

Altogether, our model integrates key aspects of the MEH and MHH to provide a general mechanistic explanation for genome size evolution. It highlights structural mutations as a major mutational hazard susceptible to driving genome size evolution under general conditions.

4.2 MODEL AND RESULTS

4.2.1 *Model overview and existence of an equilibrium non-coding genome size*

To address the question of non-coding genome size evolution, we study the effect of mutations on a population of N individuals with simplified, circular genomes.

As shown in Figure D.1, we consider a circular haploid genome of length L base pairs (bp), composed of g coding segments (and thus g non-coding segments). Let us note the number of non-coding base pairs z_{nc} and the number of coding base pairs z_c. We have $z_{\mathrm{c}} + z_{\mathrm{nc}} = L$. We assume that:

- Coding segments are of the size $\frac{z_c}{g}$ and represent "genes". Genes are non-overlapping and all oriented in the same direction. The non-coding segments are equally distributed between the g genes and are each of size $\frac{z_{\mathrm{nc}}}{g} \geqslant 0$. We assume this remains true after any change in non-coding size, as neutral inversions will reshuffle the genome. This ensures that a genome can be fully described with just g, z_{c} and z_{nc}.
- Deleting any base of a gene inactivates it and is always lethal. Genes are assumed to have a promoter, here represented by their first base. As a result, a partial duplication not including this first base is not expressed and is thus neutral, *i.e.* it does not affect viability, as long as it is not inserted within a gene. Conversely, a partial or complete duplication including the promoter results in a new expressed gene and is lethal, regardless of its insertion point.
- Non-lethal mutations are assumed to be perfectly neutral for the viability of the individual. Thus, fitness is binary: it is either 1 or 0.

Different types of mutations occur at different mutation rates. We note μ the basal per base mutation rate of the organism, and $\lambda_i \mu$ is the per base mutation rate for mutation type i. Throughout the manuscript, we analyze the evolution of the non-coding genome size z_{nc}, under the assumption that the coding genome size z_{c} and the number of coding segments g remain fixed.

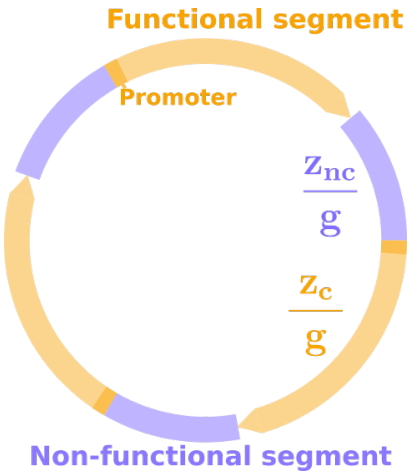

Figure 4.1: Representation of a genome, with $g = 3$. Non-coding segments are of the same size z_{nc}/g, and coding segments are of the same size z_c/g. Each coding segment starts with a promoter that can create a new coding segment if duplicated.

Neutral genome growth

We compute the probability of different types of mutations to be neutral and fixed in a population of size N. In the following, a mutation is said to be neutral when it does not affect the coding genome and thus does not alter the viability of the individual.

For the sake of clarity, we consider here (Section 4.2.1 and Section 4.2.1) a simple version of the model including only two types of structural mutations: duplications (dupl) and deletions (del), occurring at the same per bp rate ($\lambda_{dupl} = \lambda_{del} = 1$). A duplication copies a random segment of the genome and inserts it elsewhere, while a deletion removes a random segment of the genome. The breakpoints are chosen uniformly at random on the genome, such that both mutations have the same size distribution and the expected change of size of the genome upon one mutation is 0.

We calculate the probability ν for each of these two mutations to be neutral (in terms of viability), recalling that duplicating a promoter, inserting a segment within a gene, or deleting any base of a gene is always deleterious. Detailed computations are provided in Supplementary Materials (Section C.1).

$$\begin{aligned} \nu_{del}(g, z_c, z_{nc}) &= \frac{z_{nc}(z_{nc}+g)}{2gL^2} \\ \nu_{dupl}(g, z_c, z_{nc}) &= \frac{(z_{nc}+z_c-g)(z_{nc}+g)}{2gL^2} \end{aligned} \tag{4.1}$$

Notably, we have : $\frac{\nu_{del}(g,z_c,z_{nc})}{\nu_{dupl}(g,z_c,z_{nc})} = \frac{z_{nc}}{z_{nc}+z_c-g} \leqslant 1$, as z_c is obviously much larger than g. Thus, duplications are more often neutral than deletions. Similarly, we show that neutral duplications are also on average larger than neutral deletions (see Supplementary Materials Section C.6). As

illustrated by Figure 4.2A, we can also consider the probability for a mutation of a given size k to be neutral. It highlights that increasing z_{nc} increases both the probability for mutations of a given size to be neutral and the range of possible neutral mutations (note that, above a certain size, mutations are always lethal due to constraints from the genome architecture: larger mutations would necessarily delete part of a gene or duplicate a promoter). Consequently, genomes should grow indefinitely if we assume that only neutral mutations are fixed with an equal probability. However, a mutation that is neutral in terms of fitness for the individual is not necessarily neutral in terms of fitness for the lineage. The next subsection will explore the effect of this second-order selective force.

Robustness selection

By definition, a neutral duplication or deletion does not change the viability of an individual. However, it changes the non-coding genome size z_{nc}. Now, the probability for a mutation to be neutral depends on z_{nc} (see Equation 4.1), and so a neutral mutation changes the probability for future mutations to also be neutral. Changing the genome size also changes the probability for a mutation to occur at replication, as bigger genomes will naturally undergo more mutations for the same per base mutation rate. Therefore, mutations that are neutral in terms of their immediate effect on the viability still change the probability for the individual to have future offspring that are equally fit — their robustness. For the rest of the manuscript, we call the *effective fitness* f_e of an individual the average fitness of its potential offspring. This can also be viewed as the fecundity of an individual once the viability of the offspring has been taken into account. In our model, supposing that at most one mutation of each type can occur upon replication, we have:

$$f_e(\mu, g, z_c, z_{nc}) = \left(1 - \mu(1 - \nu_{del}(g, z_c, z_{nc}))\right)^L \left(1 - \mu(1 - \nu_{dupl}(g, z_c, z_{nc}))\right)^L \tag{4.2}$$

Since genome size modifies the effective fitness and affects a lineage survival probability in the long term, it can be selected. In particular, while increasing the non-coding genome size z_{nc} increases the probability for mutations to be neutral (see Figure 4.2A), it also increases the probability for a mutation to happen. As a result, the effective fitness f_e actually decreases as the non-coding genome size increases (see Figure 4.3), and selection can then act against the genome size increase described in paragraph A1.

We characterize this effect more precisely using a population genetics argument. We consider a haploid population of wild-type individuals of size N in which a mutant appears and bears a neutral mutation that adds k bases to its non-coding genome, with $k \in \mathbb{Z}^*$. k can be

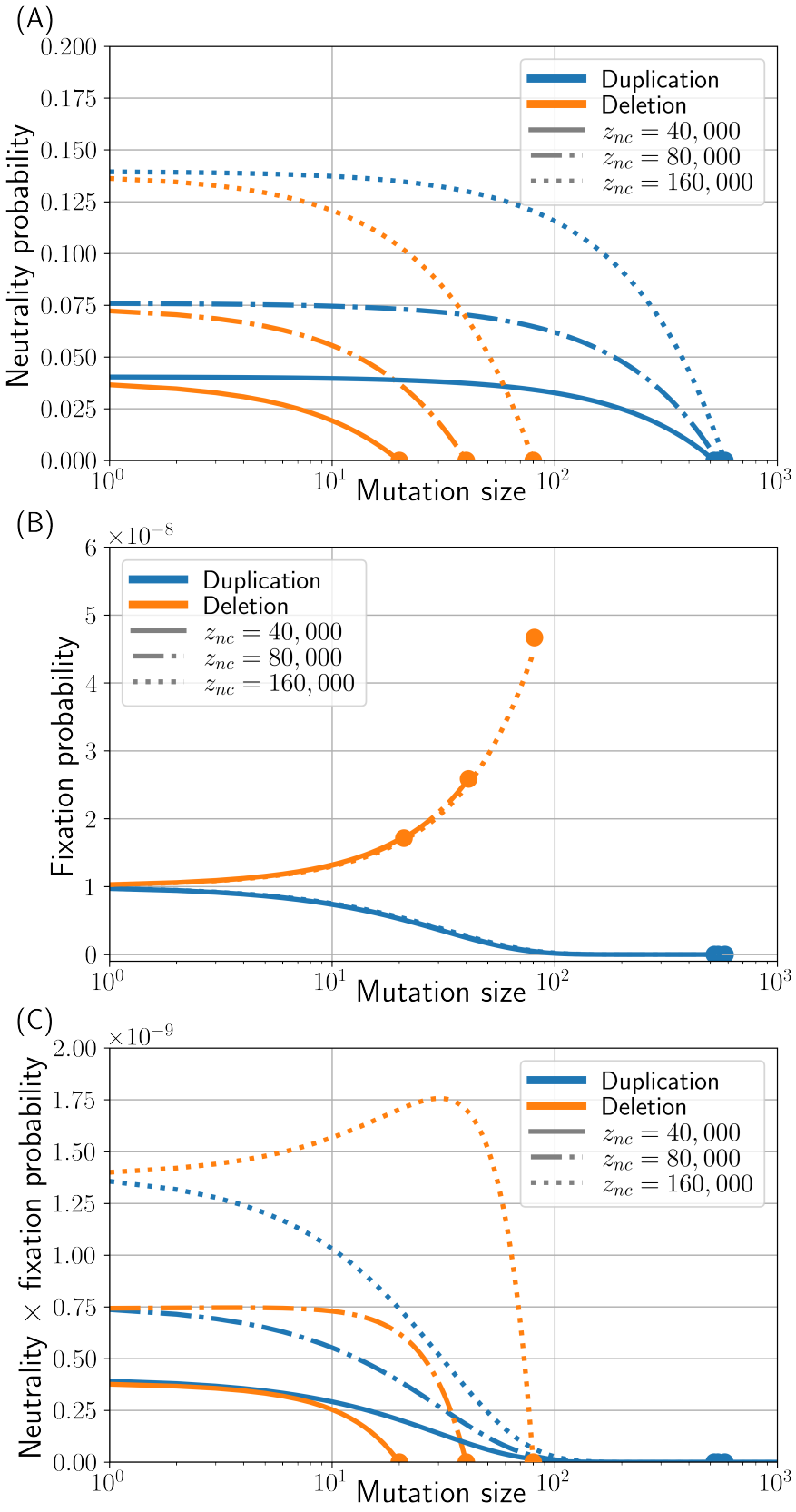

Figure 4.2: Effect of mutation type and genome size on neutrality and fixation. The dots mark points above which mutations cannot be neutral nor fixed due to constraints from the genome architecture. *(continued next page)*

(continued caption) **(A)** Probability of a duplication (blue) or a deletion (orange) to not affect the coding genome for different mutation sizes and non-coding sizes. The number of genes g is fixed at 2,000, and the coding genome size at $z_c = 1,000,000$ bp. Mutations are more likely to be neutral in bigger and more non-coding genomes, and neutral duplications are larger and more frequent than neutral deletions. **(B)** Probability of fixation of a neutral duplication (blue) or a neutral deletion (orange) for different mutation sizes and non-coding sizes. The number of genes g is fixed at 2,000, the coding genome size at $z_c = 1,000,000$ bp, and the population size is $N = 10^8$. **(C)** Probability of being neutral *and* fixed for different mutation sizes and different non-coding sizes. The number of genes g is fixed at 2,000, the coding genome size at $z_c = 1,000,000$ bp, and the population size is $N = 10^8$. For the shortest non-coding size (plain line), duplications are more often neutral and fixed than deletions for any mutation size, indicating that the non-coding genome size would increase. On the contrary, for the biggest non-coding size (dotted line) deletions are more often neutral and fixed than duplications for any mutation size, indicating that the non-coding genome size would decrease: there must be an equilibrium non-coding size between these values.

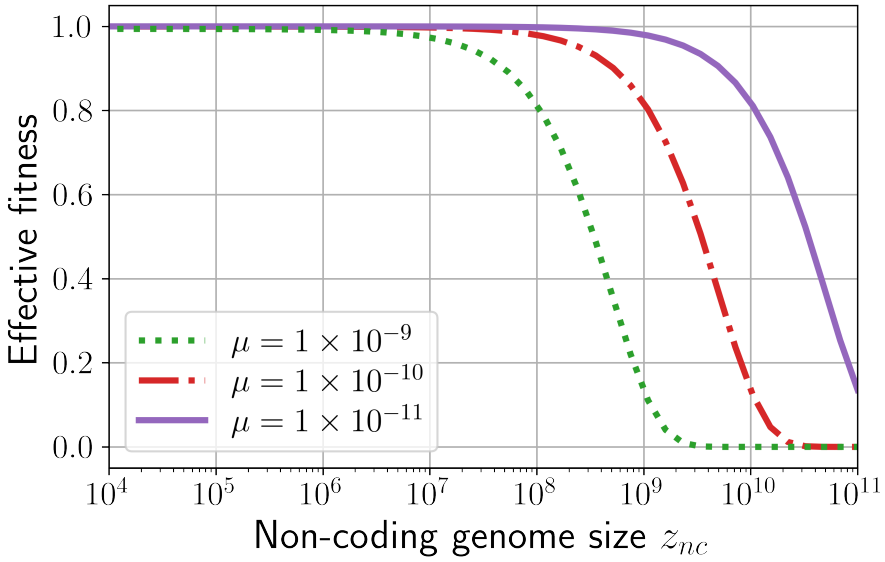

Figure 4.3: Effective fitness f_e for different non-coding sizes z_{nc} and different mutation rates μ. Genome architecture is fixed at $z_c = 1,000,000$ and $g = 2,000$, and $\lambda_{del} = \lambda_{dupl} = 1$. Notably, the effective fitness decreases with both z_{nc} and μ.

either positive (duplication) or negative (deletion). We consider that the population follows a Wright-Fisher model (Fisher, 1923; Wright, 1931), and we compute the probability for this mutant to go to fixation (Hirsh, 2005):

$$\mathbb{P}_{\text{fix}}(k, \mu, N, g, z_c, z_{nc}) = \frac{1 - \left(\frac{f_e(z_{nc})}{f_e(z_{nc}+k)}\right)^2}{1 - \left(\frac{f_e(z_{nc})}{f_e(z_{nc}+k)}\right)^{2N}}, \tag{4.3}$$

where we note $f_e(\mu, g, z_c, z_{nc})$ simply $f_e(z_{nc})$, as we consider that other parameters are fixed. As illustrated by Figure 4.2B, mutations that increase genome size are less likely to be fixed than mutations that decrease genome size. This is the direct consequence of an increase in genome size being tied to an increase in the per genome mutation rate, and hence a decrease in effective fitness (see Figure 4.3). Hence, while neutral duplications are more frequent and larger than neutral deletions, they are also more rarely fixed. When considering the combination of these two tendencies (the opposing biases in the immediate probability of being lethal and in the ultimate probability of being fixed), we can see that the shortest genomes are more likely to fix neutral duplications while the longest genomes are more likely to fix neutral deletions, as demonstrated in Figure 4.2C. This intuitively results in an equilibrium genome size at which the two effects cancel out.

Computing the equilibrium non-coding genome size

To formalize this equilibrium genome size, we compute the average contribution of duplications (δ_{dupl}) and deletions (δ_{del}) to changes in non-coding genome size in the population. δ_{dupl} and δ_{del} are expressed in bp per generation per mutation event and represent the average length of fixed mutations per time unit. They are computed under the origination-fixation approximation, meaning there is no clonal interference, and we consider the probability for each mutation individually to go to fixation in the absence of any other mutant in the population. Each δ thus depends on the mutation's probability of being neutral, its size, and its fixation probability:

$$\begin{aligned}
\delta_{\text{dupl}}(\mu, N, g, z_c, z_{nc}) &= \frac{g(z_{nc}+g)}{L^3} \sum_{j=1}^{\frac{z_{nc}+z_c}{g}-1} \left(\frac{z_{nc}+z_c}{g} - j\right) j\mathbb{P}_{\text{fix}}(j) \\
\delta_{\text{del}}(\mu, N, g, z_c, z_{nc}) &= \frac{g}{L^2} \sum_{j=1}^{z_{nc}/g} \left(\frac{z_{nc}}{g} - j + 1\right) j\mathbb{P}_{\text{fix}}(-j)
\end{aligned}, \tag{4.4}$$

where we denote $\mathbb{P}_{\text{fix}}(k, \mu, N, g, z_c, z_{nc})$ as $\mathbb{P}_{\text{fix}}(k)$, as other parameters are supposed fixed. Detailed derivations are presented in the Supplementary Materials, Section C.2. From Equation 4.4, we can derive the bias towards increasing or decreasing genome size as the ratio between the sum of the contributions of all deletions over the sum of the

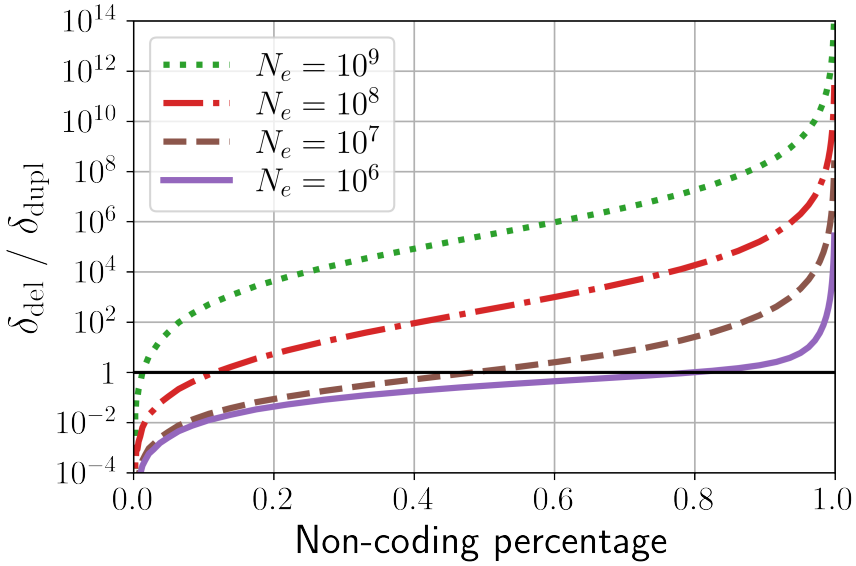

Figure 4.4: Measured bias for different non-coding proportions. Genome architecture is fixed at $z_c = 1,000,000$ and $g = 2,000$, the mutation rate is fixed at $\mu = 1 \times 10^{-10}$ and $\lambda_{\text{del}} = \lambda_{\text{dupl}} = 1$. z_{nc} varies in a logspace from 10^3 to 10^9, and four different values of N are depicted, showing a progression in the equilibrium non-coding percentage. The black horizontal line shows the equilibrium at $B = 1$.

contributions of all duplications for a given genome size, population size, and mutation rate.

$$\begin{aligned} B(\mu, N, g, z_c, z_{nc}) &= \frac{\mu\ L\ N\ \delta_{\text{del}}(\mu, N, g, z_c, z_{nc})}{\mu\ L\ N\ \delta_{\text{dupl}}(\mu, N, g, z_c, z_{nc})} \\ &= \frac{\delta_{\text{del}}(\mu, N, g, z_c, z_{nc})}{\delta_{\text{dupl}}(\mu, N, g, z_c, z_{nc})} \end{aligned} \tag{4.5}$$

The non-coding genome size is at equilibrium when $B = 1$ (see Figure 4.4). When the bias is above 1, deletions contribute more to genome size changes and the non-coding proportion shrinks. On the other hand, when the bias is below 1, duplications contribute more to genome size change and the non-coding proportion increases.

Joint impact of population size and mutation rate

B is a function of the genome architecture (z_c, z_{nc}, and g), the population size N, the mutation rate μ. However, we can show that B depends on N and μ only through their product, as previously observed in simulation data (Luiselli et al., 2024).

Indeed, N and μ only appear in f_e (Equation 4.2) and $\mathbb{P}_{fix}$ (Equation 4.3). Let us start with the expression of the effective fitness f_e, and consider that the mutation rate μ is negligible compared to 1 ($\mu \ll 1$).

$$\begin{aligned} f_e(z_{nc}) &= \left(1-\mu(1-\nu_{del}(g,z_c,z_{nc}))\right)^L \left(1-\mu(1-\nu_{dupl}(g,z_c,z_{nc}))\right)^L \\ &\underset{\mu\to 0}{\sim} \exp\Big(-L\mu\big[(1-\nu_{del}(g,z_c,z_{nc}))+(1-\nu_{dupl}(g,z_c,z_{nc}))\big]\Big) \\ &= \exp\Big(\mu A(g,z_c,z_{nc})\Big) \end{aligned}$$

Where $A(g,z_c,z_{nc}) = -(z_c+z_{nc})\big[(1-\nu_{del}(g,z_c,z_{nc}))+(1-\nu_{dupl}(g,z_c,z_{nc}))\big] < 0$. Then, the ratio of effective fitnesses used in the computation of $\mathbb{P}_{fix}$ (Equation 4.3) can be written as:

$$\begin{aligned} \frac{f_e(z_{nc})}{f_e(z_{nc}+k)} &\underset{\mu\to 0}{\sim} \frac{\exp\Big(\mu A(g,z_c,z_{nc})\Big)}{\exp\Big(\mu A(g,z_c,z_{nc+k})\Big)} \\ &\underset{\mu\to 0}{\sim} \exp\left(\mu\Delta(g,z_c,z_{nc},k)\right) \end{aligned}$$

Where $\Delta(g,z_c,z_{nc},k) = A(g,z_c,z_{nc}) - A(g,z_c,z_{nc}+k)$ is a function that depends solely on genome architecture (z_c, z_{nc} and g) and mutation size k. The probability of fixation of a mutation changing the genome size by k (positive or negative) is thus:

$$\begin{aligned} \mathbb{P}_{fix}(z_{nc},k) &= \frac{1-\left(\frac{f_e(z_{nc})}{f_e(z_{nc}+k)}\right)^2}{1-\left(\frac{f_e(z_{nc})}{f_e(z_{nc}+k)}\right)^{2N}} \\ &\underset{\mu\to 0}{\sim} \frac{1-\exp\left(2\mu\Delta(g,z_c,z_{nc},k)\right)}{1-\exp\left(2N\mu\Delta(g,z_c,z_{nc},k)\right)} \\ &\underset{\mu\to 0}{\sim} \frac{-2\mu\Delta(g,z_c,z_{nc},k)}{1-\exp\left(2N\mu\Delta(g,z_c,z_{nc},k)\right)} \end{aligned}$$

$\mathbb{P}_{fix}$ appears to be a function of μ, $N\times\mu$ and other parameters. Thus, both δ_{dupl} and δ_{del} can also be written as μ times a function of $N\times\mu$. Since $B = \frac{\delta_{del}}{\delta_{dupl}}$, the μs cancel out and N and μ always appear in the form of a product in B. Given a fixed coding size z_c and a number of coding segments g (*i.e.* a fixed coding architecture), N and μ have therefore a similar impact on the equilibrium non-coding size. This can be illustrated by a numerical exploration of the relative effects of N and μ (Supplementary Materials Section C.3).

An alternative way to see this result is by noting that $\mu\Delta(g,z_c,z_{nc},k)$ is the selection coefficient associated with the effective fitness: $s = \frac{f_e(z_{nc}+k)-f_e(z_{nc})}{f_e(z_{nc})}$. Indeed, in the limit $\mu\to 0$ and $\frac{f_e(z_{nc}+k)}{f_e(z_{nc})}\to 1$, and

thus $s \to 0$, we have $s \sim \ln \frac{f_e(z_{nc}+k)}{f_e(z_{nc})} \sim \mu\Delta(g, z_c, z_{nc}, k)$. The mutation-selection-drift equilibrium, as a general rule, depends only on relative, not absolute, mutation rates (thus here, on the mutational bias). In addition, it depends on the various selective effects implicated in it only through their scaled selection coefficients. Here, $Ns \sim N\mu\Delta(g, z_c, z_{nc}, k)$, and thus, in the end, the mutation-selection-drift equilibrium depends on N and μ only through their product.

All these results show that an equilibrium non-coding genome size exists and depends on the coding genome architecture (g and z_c) and on the product $N \times \mu$.

Necessary condition for the existence of the equilibrium

So far, we only considered one type of mutation: structural variations. Other types of mutation change genome size and one could ask whether they would lead to a similar equilibrium. In particular, short indels (< 50 bp) can also add or remove bases to the genome and contribute to genome size changes, although less abruptly than structural variants. Most interestingly, if we replicate our model with only short indels (see Supplementary Materials Section C.1 and Section C.2), we don't observe an equilibrium genome size, except under very specific conditions (for extremely large population size and starting from a small enough genome). In all other cases, $\delta_{\text{indel}^+} > \delta_{\text{indel}^-}$ and so, in the absence of a sufficiently strong mutational bias in favor of deletions, indels induce an infinite growth of the non-coding size (Section C.4), confirming previous observations (Banse et al., 2024b). Indeed, except in very specific ranges of parameters (see Supplementary Material Section C.4), indels do not create a selection for shorter genomes on their own: although they are more numerous as genome size increases, they are also more often neutral due to their size being limited, and so their effect is more likely to be limited to non-coding parts of the genome. On the opposite, structural variations, being driven by double-strand breaks, conserve their mutational liability when non-coding genome size increases. This makes them a necessary component to observe a pervasive genome size equilibrium.

4.2.2 *Expanded model of non-coding genome size evolution*

Although the existence of an equilibrium specifically requires the presence of structural mutations, other types of mutations, with possibly different mutation rates, could contribute to genome size — directly (by changing the amount of non-coding sequences) or indirectly (due to their intrinsic mutational liability). To account for this, we added four types of mutations to our mathematical model: point mutations (pm), inversions (inv), small insertions (indel$^+$), small deletions (indel$^-$). Each mutation type i has its own mutation rate $\lambda_i\mu$. The

probability of being neutral for all these mutations, and the average contribution to changes in genome size for indels, is presented in the Supplementary Materials (Section C.1 and Section C.2).

Naturally, these mutations and biases displace the equilibrium value of our model, as they change both the robustness of the genomes and the probability of removing or adding new bases. However, Figure 4.5 shows that genome size (or equivalently the non-coding fraction) is always a decreasing function of $N\mu$, whatever the underlying mutational bias. Notably, when there is a deletion bias, the non-coding fraction remains bounded for all values of $N\mu$. The upper bound is the asymptotic value reached when $N\mu \to 0$, and it is smaller for more pronounced deletion biases. Thus, in this regime, not all coding fractions can be achieved by varying $N\mu$. On the other hand, when there is no bias or an insertion bias, the non-coding fraction diverges in the limit $N\mu \to 0$, such that arbitrarily large non-coding fractions can be achieved with sufficiently small $N\mu$. However, there is always an equilibrium value for $N\mu > 0$.

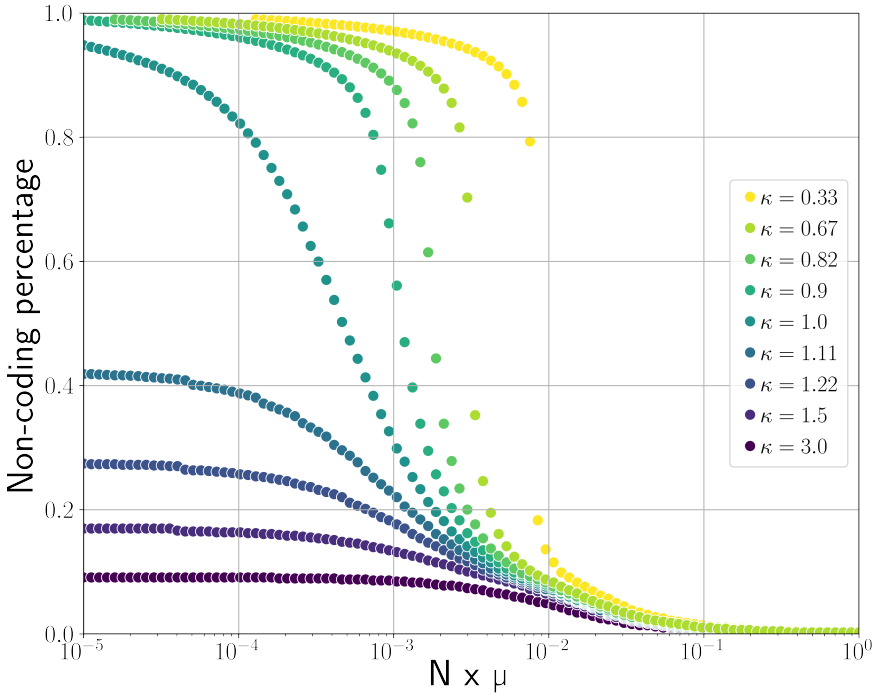

Figure 4.5: Predicted non-coding fractions for different values of $N \times \mu$ using the expanded version of the model with six types of mutations. Two sets of equilibrium percentages were run: with $\mu = 10^{-9}$ and N varying from 10^4 to 10^9, and with $N = 10^8$ and μ varying from 10^{-13} to 10^{-8}. We note the deletion bias $\kappa = \frac{\lambda_{\text{del}}}{\lambda_{\text{dupl}}} = \frac{\lambda_{\text{indel}^-}}{\lambda_{\text{indel}^+}}$. Note that we fix $\lambda_{\text{del}} = \lambda_{\text{indel}^-}$, $\lambda_{\text{dupl}} = \lambda_{\text{indel}^+}$, and these four λ_i sum to 4. Other parameters are fixed at $g = 2,000$, $z_c = 1,000,000$, $l_m = 50$ and $\lambda_{\text{pm}} = \lambda_{\text{inv}} = 1$.

Notably, the addition of new mutations and the variations in the mutational bias do not suppress the existence of the equilibrium,

as they do not fundamentally change the mechanisms at stake. The equilibrium is still determined by the product $N \times \mu$ and the coding genome architecture (z_c and g), and the variations are always in the same direction: a higher population size or a higher mutation rate is associated with a lower non-coding fraction.

4.2.3 *Insights from biological data*

Our model predicts that the *fraction* of the non-coding genome depends only on the compound parameter $N_e \times \mu$, with μ the structural mutation rate, as depicted by Figure 4.5, as well as on the relative insertion versus deletion rates. The non-coding absolute size has more complicated dependencies, as it also depends on the coding architecture (see discussion). These predictions could in principle be tested against empirical data. However, spontaneous structural mutation rates are unknown, as most structural mutations are strongly deleterious, hence frequently purged by selection and notoriously difficult to observe and quantify (Ho et al., 2020)). Notwithstanding, a tentative comparison with empirical data is shown in Figure 4.6, relying on nucleotide diversity to estimate $N_e\mu$ and assuming a 1:1 ratio for structural versus point mutation rates (ratios of either 1:10 or 10:1 would shift empirical points to the left or the right, respectively, compared to the theoretical curves).

With a structural mutation rate of this order of magnitude, our model globally predicts the overall trend of the distribution of non-coding fractions observed across species, as a function of nucleotide diversity. Thus, species with a high $N_e \times \mu$ present a lower coding fraction than species with a low $N_e \times \mu$. More precisely, eukaryotes are mostly located on the left of the figure and have both a higher non-coding fraction and a lower $N_e \times \mu$, with a tendency to follow that relationship within them, while prokaryotes are on the right of the figure and present both a lower non-coding fraction and a higher $N_e \times \mu$. Notably, the actual non-coding percentage of the prokaryotes shown here is higher than the one predicted by the model, but this is expected as our model assumes that the non-coding is purely non-functional, while the non-coding genome actually comprises regulatory RNAs and other functional sequences. This could indicate that most of the "non-coding" base pairs of prokaryotes have a phenotypic effect.

Altogether, and even if it is still far from a formal test, this comparison with empirical data gives an idea of the structural mutation rates for which second-order selection on genome rearrangements represents a key force preventing an infinite genome size growth. It also reproduces the global non-coding genome fraction variation patterns across cellular life. It shows that the potential role of structural mutations in non-coding genome size evolution should not be underestimated and deserves further investigation.

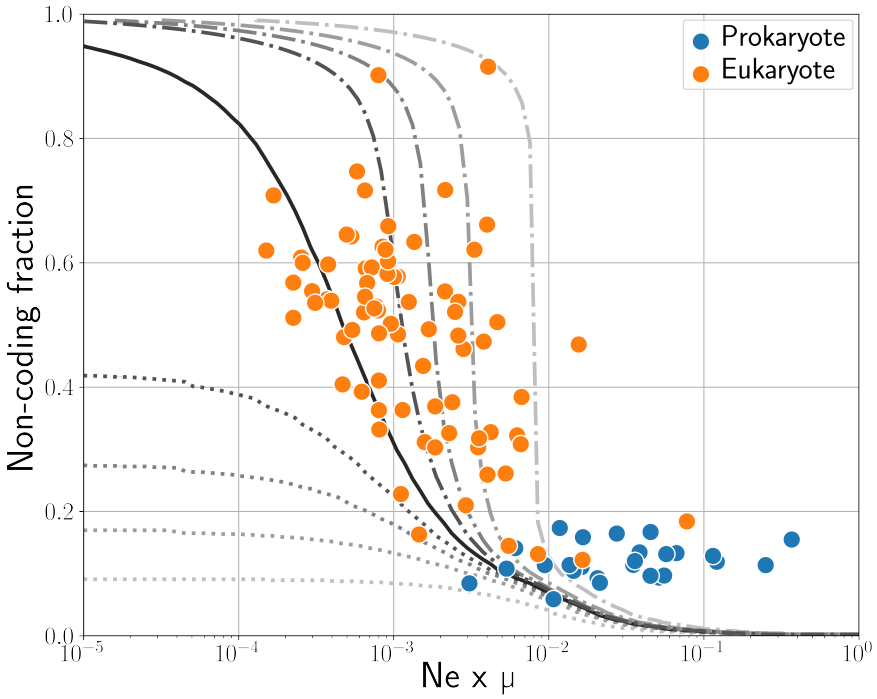

Figure 4.6: Non-coding fraction plotted against $N_e \times \mu$ for 129 species from Lynch et al., 2023. The mutation rate μ used here is the per generation per base substitution rate, which we assume to be correlated with the overall mutation rate. The gray lines show the equilibrium non-coding percentages predicted by our model for the same range of $N \times \mu$ and different mutational biases κ (see Figure 4.5).

4.3 DISCUSSION

Our model reveals simple yet important evolutionary dynamics on genome size due to two opposite effects. On the one hand, duplications and insertions are more often neutral than deletions, implying a neutral bias towards genome size increase, a mechanism akin to the border selection effect proposed by (He et al., 2019; Loewenthal et al., 2022). As an intuitive example, in the extreme case of a fully coding genome, it is still possible to neutrally insert a base between two genes, while no base can be removed. On the other hand, at constant phenotypical adaptation bigger genomes are counter-selected as they are more susceptible to structural mutations: lineages in which genomes get bigger are less likely to survive in the long term. This second-order selection for shorter genomes is imposed by the mere existence of structural mutations, which decrease the robustness of genomes when non-coding genome size increases — as previously conjectured (Knibbe et al., 2007a). With these two effects, and knowing only the genome's current coding architecture (size and number of segments), the effective population size of the species, and the rates for the different mutation types, we can determine an equilibrium

non-coding genome size towards which the species should be tending. Notably, this does not apply when structural mutations are absent from the model: with only indels, our model predicts that genomes are likely to grow indefinitely.

Naturally, parameters not considered here could impact this equilibrium quantitatively. Most importantly, the presence of transposable elements, non-coding but functional DNA, horizontal transfers, and other mutational processes, such as recombination, would displace the equilibrium by affecting both the neutral mutational bias and the robustness of genomes. Yet, they would not suppress either of the two above-elaborated effects and so the existence of the equilibrium remains — as well as the direction of change in non-coding proportion caused by changes in N or μ. Similarly, we have assumed a uniform distribution for the sizes of mutations. Relaxing this assumption would also displace the equilibrium, although robustness selection would still operate, as long as the size of the structural mutations increases with genome size — which is intuitive as some species' structural variants are longer than other species' genomes (Wellenreuther and Bernatchez, 2018). Therefore, we expect the two effects we present here to be pervasive. In particular, while the hypotheses of our model are closer to a prokaryote-like genome (a single haploid circular chromosome), there is no reason for the general mechanism to not be true in the case of eukaryotes, and we can use it to compute the predicted non-coding percentage around eukaryote-like values of N_e and μ (as shown by Figure 4.6).

As a first empirical confrontation of our theory, our comparison of the predictions of our model with empirical data (Figure 4.6) results in a globally coherent and insightful picture for both eukaryotes and prokaryotes, as predictions could align with biological observations of non-coding fractions. In particular, our model predicts more variability of non-coding fractions in the eukaryotic parameter range, which lies around the steepest part of the curves (see Figure 4.6). Conversely, prokaryotes, having a much larger $N_e \times \mu$ product, are predicted to be much more stable around lower non-coding fractions. In that case, although a deletion bias in prokaryotes could exist, the $N_e \times \mu$ values are in a range where the predicted non-coding percentage is only loosely affected by mutational biases. Mathematically, the high variation of non-coding percentages in eukaryotes could be explained by observing that the function B is flatter for lower values of $N_e \times \mu$, typical of eukaryotes (see Figure 4.4). In these ranges, it is harder to reach the equilibrium non-coding value, as the bias towards losing or gaining bases at each generation is very low, hence allowing more variability of non-coding sizes at constant $N_e \times \mu$ and constant mutational bias. Eukaryotes are supposed to be subjected to biases towards insertions (Ratcliff, 2024), making it more complicated to compare data to our model. Finally, the results presented also

depend on the relative rates of different types of mutations, which are largely unknown for structural mutations. Indeed, although they are frequently observed in all domains of life (Raeside et al., 2014; Fang and Edwards, 2024), their spontaneous rate is very difficult to estimate due to their strong deleterious effect. As such, interpretations should be taken with caution. In short, our model proposes an explanation for the very different non-coding percentages observed in eukaryotes and prokaryotes, without postulating a difference in nature between these two types of organisms but only relying on the existence of structural mutations and the different values of N_e and μ for eukaryotes and prokaryotes.

Surprisingly, our results point out that non-coding genome *fractions* evolution is determined by the product $N_e \times \mu$ and the mutational bias, which, to our knowledge, is a new prediction. Although these factors have already been pointed out as the potential determinant of (non-coding) genome *size* (Petrov, 2002; Lynch and Conery, 2003; Yi and Streelman, 2005; Kelkar and Ochman, 2012), we show that size also depends strongly on the coding architecture of a genome, with the latter probably highly driven by adaptation. This, as well as the fact that N_e and μ act jointly and should always be accounted for in the form of their product, could explain why some data are not aligned with the MHH (Ai et al., 2012; Sloan et al., 2012; Mohlhenrich and Mueller, 2016; Marino et al., 2024). Consequently, further research should focus on non-coding fractions, or account for differences in coding architectures when comparing non-coding genome size and N_e and μ between species.

Finally, although we have assumed fixed mutation rates, in reality, mutation rates are themselves susceptible to evolve. According to the drift barrier model (Sung et al., 2012), mutation rates are under directional selection limited by random drift, such that they reach an evolutionary equilibrium that depends on N_e. The selective force involved in this higher-order evolutionary process stems from the deleterious effects of new mutations on the offspring of the current generation (Sung et al., 2012, 2016; Lynch et al., 2016). Interestingly, in the case of structural mutations, this force is the same as the second-order selective force behind the mutational hazard (see Equation 4.2). Thus, any mutational hazard also represents a selective force acting on modifiers of the corresponding mutation rate. This raises interesting perspectives on the joint evolutionary dynamics of non-coding genome size and structural mutation rates, which would certainly deserve further theoretical exploration.

To conclude, our results show that indirect selection against mutational hazards, *i.e.* robustness selection, and differences in the neutrality of mutations increasing or decreasing genome size are sufficient to explain the existence of an equilibrium in non-coding genome size. Structural mutations are sufficient to fulfill both conditions. The

non-coding genome is constantly under indirect selection due to its mutagenic nature and the structural mutations it can initiate following double-strand breaks. As a consequence, a major determinant of the non-coding genome fraction is the product $N_e \times \mu$, which affects both the efficacy of selection and the robustness cost of each additional base pair. More research should be conducted into that area to reach quantitative results and to understand more precisely how each determinant of non-coding genome size (N, μ, mutation bias, types of mutations, number and length of genes) affects the equilibrium non-coding size of a species, and whether the species are at that equilibrium or tending towards it. Finally, the interaction of indirect selection on the non-coding genome and direct selection on the coding genome should be studied further by relaxing the hypotheses of a fixed coding architecture and a binary fitness.

4.4 MATERIALS AND METHODS

4.4.1 *Logic behind the model*

We consider a model with a simple genome architecture: organisms own a single circular chromosome. Mutations happen at random on the chromosome, each position for a mutation being drawn from a uniform distribution along the genome. While very simplistic, this approach carries an essential property of structural mutations: their size grows with total genome size (Hirabayashi and Owens, 2023). This is easily demonstrated by biological data, as some observed inversions of several Mb (Wellenreuther and Bernatchez, 2018) are bigger than the genomes of other organisms. The fact that bigger genomes are more susceptible to mutating and that structural mutations are increasingly dangerous implies a selection for shorter genomes on the lineage level, which we quantify mathematically.

The logic behind the model is the following (detailed computations are provided in the Supplementary Materials): the computation of each ν_i is made assuming a uniform draw of all positions needed for the mutation (two for deletions and three for duplications). Each mutation that has any phenotypic effect (deleting part of a gene, duplicating inside a gene, or duplicating a promoter) is lethal. To compute the effective fitness, we consider the probability for each base pair to initiate each type of mutation, and for the initiated mutation to be neutral. This corresponds to a binomial law where success is defined as either no mutation occurs or a neutral mutation occurs, the number of draws is the genome length multiplied by the number of mutation types. A non-lethal reproduction means there are only successes. The fixation probability is taken from Sella and Hirsh 2005 (Hirsh, 2005). The average contribution to genome size change of each mutation type is computed as an average of the mutation size weighted by the

probability of the mutation to be of this size and by the probability of a neutral mutation of this size to be fixed.

4.4.2 *Numerical resolution*

All the code for the numerical resolution of the mathematical model is available online in GitLab: `https://gitlab.inria.fr/jluisell/structural-mutations-set-an-equilibrium-non-coding-fraction`.

Notably, the code is written in `Python` using `Decimal` to increase float precision. This is necessary due to the very large values taken by our parameters (especially genome size and population size) and the several exponents in the computation.

To predict an equilibrium non-coding genome size, we fix the number of genes g, the coding genome size z_c, the effective population size N_e, the per-base per mutation type mutation rate μ, and the mutational bias κ. We then compute the resulting bias towards adding or losing bases for several possible non-coding sizes. Since the bias function is monotonous with respect to the non-coding genome size, we find the actual equilibrium with a bisect method.

4.4.3 *Biological data*

To compare our model to actual biological data, we gathered mutation rates and effective population sizes from Lynch et al., 2023. For each available species, we downloaded the annotated genome from NCBI and isolated the biggest chromosome. We counted any base pais annotated as part of a protein-coding gene as a "coding" base pair (z_c), all the other base pairs as "non-coding" (z_{nc}), and we counted the number of continuous non-coding segments (g). While more precise annotations could be used (*e.g.* regulatory RNAs should be counted as coding in our model, since they are functional), this approach reduced the annotation bias between well-studied model species and less annotated pioneer species.

4.5 ACKNOWLEDGMENTS

The authors thank Laurent Duret, Julien Joseph, Ivan Junier, and Flora Gaudillère for fruitful discussions on the model and the biological data. Funding: Agence Nationale de la Recherche (ANR-20-CE02-0008-01 NeGA).

CONCLUSION FOR PART I

The first part of this PhD thesis showed that chromosomal rearrangements create an evolutionary force that constrains genome architecture evolution: they pose a mutational hazard from which genomes have to protect themselves by staying small enough. Thus, there is an indirect selection for robustness that prevents genome growth. In a population, individuals with the larger genome size are less susceptible to reaching fixation as their offspring — and the offspring of their offspring — have a higher risk of harboring deleterious mutations than individuals with smaller genomes.

In Chapter 2, experiments with Aevol enabled us to show that chromosomal rearrangements are key to long-term genome evolution, both for fitness and genome architecture. Indeed, rearrangements open new possibilities for innovation by considerably expanding the neighborhood of genotypes, and they constrain genome growth as they represent a serious mutational hazard. The risk posed by chromosomal rearrangements is modulated by the mutation rate, while the efficacy of the selection for robustness against random drift is modulated by the population size. As such, the evolutionary force contributed by chromosomal rearrangements could be a key factor explaining how genome size changes in response to changes in population size and/or mutation rate. This is demonstrated in Chapter 3: an increase in population size increases the efficacy of selection for robustness to chromosomal rearrangements — thus reducing genome size —, and an increase in mutation rate drastically increases the mutational hazard posed by chromosomal rearrangements — thus forcing a stronger selection against them. This is an instance of the Mutational Hazard Hypothesis (MHH), proposed by M. Lynch and widely discussed in the literature (Lynch and Conery, 2003; Lynch, 2006b): the non-coding genome is deleterious *per se* and is filtered out by population genetics mechanisms. The novelty of Chapter 3 is to point out chromosomal rearrangements as major actors of this dynamics. This result was generalized beyond Aevol in Chapter 4, highlighting how very simple hypotheses suffice to explain the presence of an equilibrium non-coding genome size, determined by the effective population size, the mutation rate, and the coding structure — hence the history of the species.

Both of these results, the experiments in Aevol and the mathematical model, were based on prokaryote-like genomes with a single circular chromosome, no sexual reproduction, and no meiotic recombination. This raises the question of the generality of our results and their appli-

cability to eukaryotes. In Chapter 4, we assumed that there is no reason for the mechanisms we highlight not to act on eukaryotes, and we applied the model to both prokaryotes and eukaryotes, using estimates of effective population sizes and mutation rates reported in the literature for both groups. While this may be reasonable as a first approximation and a test of the theory, eukaryotes display unique characteristics in their life cycle and mutational processes, in particular because of sexual reproduction and meiosis. First, eukaryotes generally provoke double-strand breaks to perform an obligate meiotic recombination event. Those induced double-strand breaks represent an additional source of potential rearrangements caused by non-legitimate recombination. Conversely, the reshuffling of genetic variation implemented by meiotic recombination increases the selection efficacy. Finally, eukaryotes generally have linear chromosomes. This contrasts with the circular chromosomes of prokaryotes. This could have an impact on the combinatorial structure of the mutational landscape of rearrangements. All this could lead to different quantitative behaviors between eukaryotes and prokaryotes. As such, it is required to specifically model the eukaryote-like mechanisms and measure the impact of chromosomal rearrangements on genome architecture evolution within this framework. To this end, a new eukaryote-like version of Aevol has been developed specifically for this thesis.

Part II

EUKARYOTES

"The amount of DNA associated with just 30 human genes is equivalent to the entire genome size of an average prokaryote." (Lynch and Conery, 2003)

5 EUKARYOTE AEVOL

The first part of the thesis focused on prokaryote-like models and gave insight into the impact of chromosomal rearrangements on the evolution of their genome structure. Arguably, as briefly mentioned at the end of Part i, eukaryotes could answer differently to the same input because of some profound differences between eukaryotes and prokaryotes.

To test whether the results on the impact of chromosomal rearrangements on the evolution of genome structure hold true for eukaryotes, I developed a eukaryote-like version of Aevol. This chapter is dedicated to the presentation of the model, while the next two chapters present experimental results using it, testing the impact of changes in effective population size (Chapter 6) and of changes in reproduction mode (Chapter 7) on genome architecture evolution.

As this chapter is based on a comparison between the prokaryote version of Aevol and the new eukaryote implementation of the model, it is recommended first to read the methods section of Chapter 2 and Chapter 3 to understand the original functioning of Aevol. More details are also provided in Appendix A. Additionally, Aevol-9 — the version of Aevol that integrates the eukaryotic model —, is presented in a method paper entitled "Aevol-9: A simulation platform to decipher the evolution of genome architecture" and authored by Juliette Luiselli and David P. Parsons (co-first authors), Romain Gallé, Paul Banse, Jonathan Rouzaud-Cornabas, and Guillaume Beslon (Luiselli et al., 2025b). The full documentation, including the specifics of the eukaryote version of the model, is available on `www.aevol.fr`.

5.1 OVERVIEW OF THE MODEL

The idea behind the eukaryote version of Aevol is to retain the key blocks of Aevol and make the least amount of changes — although they are substantial in the end — to design the eukaryote model. This aims at being able to compare results between the prokaryote and eukaryote frameworks and making use of the knowledge gathered on the prokaryote version of Aevol to analyse the new experiments. This choice also limits development time, a crucial parameter in the context of a PhD thesis.

Briefly, we retain the key concepts that allow Aevol to be used for studying genome architecture evolution: a population of organisms compete with each other to populate the next generation. They each have their own genome, on which RNAs and genes are identified and

decoded to compute the phenotype of the organisms. Each phenotype is then compared to the optimum phenotype in this environment to retrieve the fitness, *i.e.* the level of phenotypical adaptation of each individual.

The main genomic difference with the prokaryote model is that the individuals have two homologous chromosomes (they are diploid), which are linear instead of circular. Then, there is sexual reproduction, each parent giving one chromosome to the offspring. Finally, upon reproduction, parents undergo a mandatory meiotic recombination, thus the chromosome they transmit to their offspring is recombined. In the rest of the chapter, I will focus on these changes introduced to transition from the historic prokaryote version of Aevol to the new eukaryote version (Section 5.2 to Section 5.4). Then, I present the methodology to run simulations with this new model (Section 5.5) and analyze them (Section 5.6).

5.2 DIPLOID ORGANISMS WITH LINEAR CHROMOSOMES

In the prokaryote version of Aevol, each organism owns a single circular chromosome (see Figure 2.1). To get closer to a eukaryote model, I started by going from a circular to a linear chromosome, and then added a second chromosome. This necessitated changes in the decoding of genomes and the mutational processes, entailing several modelling choices.

5.2.1 *Changes in genome decoding: incomplete RNAs*

In Aevol, RNAs start with a promoter — defined by a consensus sequence —, and end with a terminator — a hairpin-like pattern. If there is no terminator after the promoter, then the RNA is considered unstable and the genes it might bear will not de translated. The biological rational behind this choice is that the RNA would be recognized as abnormal and degraded by the cell. Consequently, any gene located after a promoter but not followed by a terminator is not expressed. This rarely happens with a circular chromosome, as there must be just one terminator in the whole genome for all promoters to be followed by a terminator. However, introducing linear chromosomes changes that: there can often be promoters near the edge of the chromosome and no terminator after them (on the leading strand, or before them for the lagging strand). Thus, it is common to have incomplete RNAs.

These incomplete RNAs had to be correctly identified: contrary to what occurs on a circular chromosome, a terminator located at position 10 can no longer be used to complete an RNA started by a promoter at position $10,000$ (on the leading strand). Similarly, for a genome of size $L = 10,000$ bp, there could be promoter or terminator overlapping the position 0 in the prokaryote model (*e.g.* starting at position $9,990$

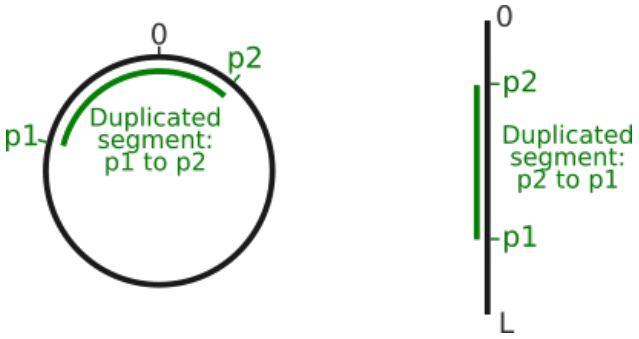

Figure 5.1: Mutation on a circular VS linear chromosome. Difference in segment choice for chromosomal rearrangement (here a duplication) between a circular chromosome (left) and a linear chromosome (right). Note that in this example, we have $0 < p_2 < p_1 < L$.

and ending at position 11), while this is no longer recognized as a promoter or a terminator in the eukaryote version as there is no continuity from positions $10,000$ to 0. As such, the correct handling of chromosome edges for promoters and terminators detection, as well as RNA delimitation has been a point of attention during development.

5.2.2 *Changes in the mutational process*

Chromosomal rearrangements

There are 4 types of chromosomal rearrangements in Aevol: deletion, duplications, inversions, and translocations. To perform chromosomal rearrangements, two points (p_1 and p_2) are drawn at random on the chromosome to delimit the segment to be deleted, duplicated, inverted or translocated. On a circular chromosome, there is no location bias for the segment and its size is distributed uniformly on $[1 : L]$, with L the size of the chromosome, as we take the two points in the order they are drawn. On a linear chromosome however, a segment cannot overlap from position L to position 1. If $p_2 < p_1$, we will have to take the segment from p_2 to p_1 instead of the segment from p_1 to p_2 (see Figure 5.1).

This introduces biases in our mutations: mutated segments are more often around the middle of the chromosome, and less often near its borders. Additionally, mutations are on average smaller: the expectation of segment size is now $L/3$ instead of $L/2$. While it is important to be aware of these new biases, they are inevitable on linear chromosomes, and we choose to maintain the null hypothesis of drawing positions at random rather than risking introducing uncontrolled biases in the model.

Insertion points

Another major change concerning mutations on a linear chromosome is that there are now $L + 1$ insertion points for small insertions, duplications, and translocations, instead of L. While inserting before the "first" base or after the "last" base is equivalent on a circular chromosome, these are two different positions on a linear chromosome. Following previous modelling choices of Aevol, the insertion point is drawn uniformly on the possible positions.

Length of small deletions

InDels differ from chromosomal rearrangements in Aevol in that only one position is drawn, before choosing randomly the (small) length of the mutational event (instead of a second position). While the length of small insertions is not affected by having linear chromosomes, as one can add an arbitrary number of bases at any insertion point, the length of small deletions has to be limited when occurring near the border of a linear chromosome. We choose to always draw the deletion size from the same distribution of possible length, and then perform the maximal possible deletion, rather than preemptively limiting the size of the event. For instance, if there are only 4 bases after the mutation position, and we are supposed to delete 6 bases, we simply delete the 4 remaining bases.

5.2.3 *Diploidization of the organisms*

Having two chromosomes per organism instead of one entailed further modeling choices, especially regarding how to compute the phenotype and the possibility of DNA transfer from one chromosome to the other.

Phenotype computation

In Aevol, proteins are triangles defined by 4 parameters: their height h, width w, position on the trait space m, and expression level e. To compute the phenotype, proteins are summed together. We retain this behavior in the eukaryote version, and proteins from both chromosomes will be summed together: there are no recessive or dominant alleles. Now, the order in which proteins are summed mattered, as summing the same numbers in a different order can change the final result due to the accumulation of floating rounding errors by the computer. To ensure that two organisms that have the exact same proteins and promoters have the exact same phenotype — thus the exact same fitness —, proteins are therefore sorted based on these 4 parameters before being added.

To keep this property in the eukaryote version, we first merge all proteins from both chromosomes into a single list, that is sorted before

summing the proteins. As a result, the chromosome on which a protein is located does not impact the final phenotype, which is the expected behavior since being chromosome "A" or "B" is arbitrary and does not reflect any biological reality.

Cross-chromosome mutations

Duplications and translocations copy or cute a segment of DNA and insert it somewhere in the genome. In theory, these mutations could be cross-chromosome, and a segment from chromosome "A" could be inserted into chromosome "B". However, to limit the complexity of software development, it has been decided that duplicated or translocated segments could only be inserted in the chromosome from which they originate. Since recombination already provides a mean of exchanging genetic material between chromosomes (see Section 5.4), it would have been useless to add further complexity in these mutational processes. As such, there are no cross-chromosome mutations other than recombinations.

5.3 SEXUAL REPRODUCTION

The second major step for transitioning from a prokaryote-based model to a eukaryote-based model was to introduce sexual reproduction. In the eukaryote version of Aevol, to populate a grid cell, we select two parents — instead of one in the prokaryote version. Each parent gives one chromosome to the offspring.

While the most usual form a selection in the prokaryote version is the local competition, in which neighboring individuals from a 3×3 square compete with one another to replicate into each grid cell, the default usage of the eukaryote model is global competition and mating. This allows for not being too restrictive in the choice of the two parents. Note that to avoid any unforeseen bias in chromosome order, especially because of recombination (see Section 5.4), the order of the chromosomes ("A" or "B") is randomized at each generation.

If the same parent is chosen twice, then there is a case of auto-fecundation. As this could introduce uncontrolled biases, we choose to control the proportion of auto-fecundation (or "selfing" events). It can be either set to a fixed proportion (from 0% to 100%), in which case we first choose between a selfing or non-selfing event before drawing the parent(s) accordingly, or be unspecified, in which case it just depends on how often the same parent is chosen twice. In the latter case, the selfing proportion thus depends on the distribution of fitness values in the population and can vary across generations.

5.4 MEIOTIC RECOMBINATION

The last step of the development of the eukaryote model was to introduce meiotic recombination. Instead of just giving one of its chromosomes to its offspring, a parent first performs a recombination between its two chromosomes and then gives one of the recombined chromosomes to the offspring. Notably, the recombined chromosomes are only temporarily stored and do not replace the parents' chromosomes. If the same individual is chosen several times as a parent, then it performs a new recombination each time, always starting from its initial pair of chromosomes. As such, if an individual has more than one offspring, the chromosomes inherited by the different offsprings will be different. This can be seen as indirectly implementing the life-cycle of a multi-cellular organism, in which gamete formation by meiosis occurs in the germline, creating a pool of gametes that are all genetically distinct, and this, without changing the genotype of the individual itself.

To perform a recombination, we look for a homologous sequence between the two chromosomes. To this end, several parameters are needed: the distribution from which we draw possible recombination points, the algorithm to measure the local alignment score, the target alignment score, and the maximal number of trials before giving up if the required alignment score cannot be reached. Due to the number of possible parameters and their combinations, most parameters have been chosen quite arbitrarily, and the following sections explain the rationale behind the choices.

5.4.1 *Distribution of possible recombination points*

The biological rational behind the algorithm for choosing the possible recombination breakpoints is that chromosomes have a general physical alignment in the cell, hence potential pairs of points should be, on average, on approximately the same position on the chromosome — although exceptions are possible. Therefore, we consider that chromosomes are more or less physically aligned: given a first point on the first chromosome, the position of the second point is not drawn uniformly on the second chromosome but with a biased probability towards the same position on the chromosome. Then, we look for an alignment on the sequence level: the breakpoints for the recombination are chosen based on sequence homology. This prevents breaking genes at each generation, while still allowing for illegitimate recombinations, given a local homology in the sequences.

To explain more precisely how the model works, let us call the two chromosomes A and B. They are of length L_A and L_B. A position p_A is drawn uniformly on the first chromosome (chromosome A). The second position of the pair of points is not taken uniformly on the

second chromosome, but in the "physical" proximity of p_A. To define this physical proximity, we take the position "in front of" p_A on the chromosome B, by scaling the with the length of both chromosomes ($p_{A'} = \frac{p_A * L_B}{L_A}$). Then, we draw the position p_B from a normal law around this position, with $\sigma = 0.05 \times L_B$ (see Figure 5.2). Thus, we are more likely to find a homology quickly (as both chromosomes stay roughly homologous throughout the simulations), but it is still possible to sometimes perform illegitimate recombination: the normal law maintains a probability to test pairs of points that are far away. Scaling σ with the chromosome size allows to still test positions that are very far away in larger chromosomes.

Once a pair of points p_A and p_B has been selected, the homology score at these positions is computed with the following algorithm:

Algorithm 1 Alignment score computation

```
s ← 0                                          ▷ current score
d ← ±1                                         ▷ direction of the alignment
for x = 0 ; x < m ; x++ do                     ▷ Where m is the maximal length
    if p_A + x × d = p_B + x × d then
        s ← s + 1
        if s > target then return s
        end if
    else
        s ← s − 2
        if s ≤ 0 then return s
        end if
    end if
end for
return s
```

In short, we first choose the direction in which to test the alignment and compare the two selected bases at positions p_A and p_B. If they match, we increase the alignment score by 1 and compare the following bases. On the other side, each mismatch will decrease the alignment score by 2. This is done until the score goes below 0, or until the maximal length of the alignment, defined 1.5 times the required alignment score, is reached.

While the computed score is below a required alignment score, and while the maximal number of trials is not reached either, we re-draw a pair of points. For my experiments, I took the maximal number of tries to be twice the genome length of the individual. It depends on the chromosome lengths, as bigger chromosomes are expected to have more points of physical interactions than smaller chromosomes during meiosis. As soon as the minimal alignment score is reached, the search stops and the recombination is performed between p_A and p_B: the two new chromosomes are composed of the segments from

0 to p_A and $p_B + 1$ to L_B on the one side, and 0 to p_B and $p_A + 1$ to L_A on the other side. Figure 5.3 depicts the formation of an offspring from two distinct parents that each perform a meiotic recombination event.

If the minimal alignment score is not reached within the maximal number of tries, the recombination is performed on the best-matching pair of points. For my experiments, I took a score of 50, which was both high enough for chromosome homology to be maintained in the long run and low enough for the search to be relatively quick and for the minimum score to be almost always reached within the maximal number of trials.

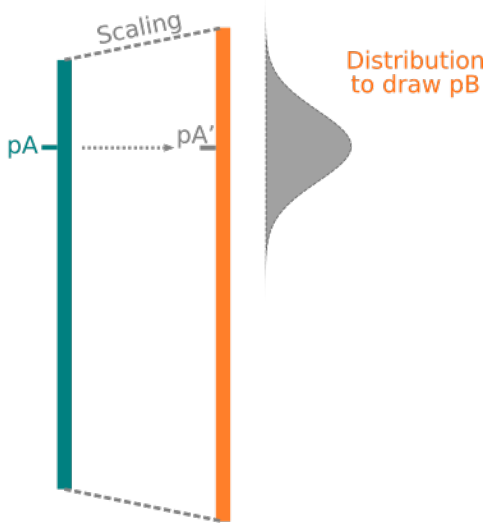

Figure 5.2: Distribution of potential recombination point on the second chromosome. The potential recombination point on chromosome A is drawn uniformly on chromosome A. Then, its equivalent $p_{A'}$ is computed on chromosome B, and a normal distribution around this equivalent position is used to draw the potential recombination point on chromosome B.

5.5 RUNNING SIMULATIONS

In the prokaryote version of Aevol, one can start a simulation either from a pre-evolved individual, given its DNA sequence, or from a randomly generated naive individual with one beneficial gene. This second option is not yet available in the eukaryote version of Aevol, due to a founding effect that prevents individuals from actually being diploid. Indeed, in the early stages of Aevol, it is very beneficial to quickly accumulate diverging gene copies. The fast selective sweeps cause the two chromosomes to diverge quickly, and one chromosome is selected above the other, resulting in one degenerated chromosome

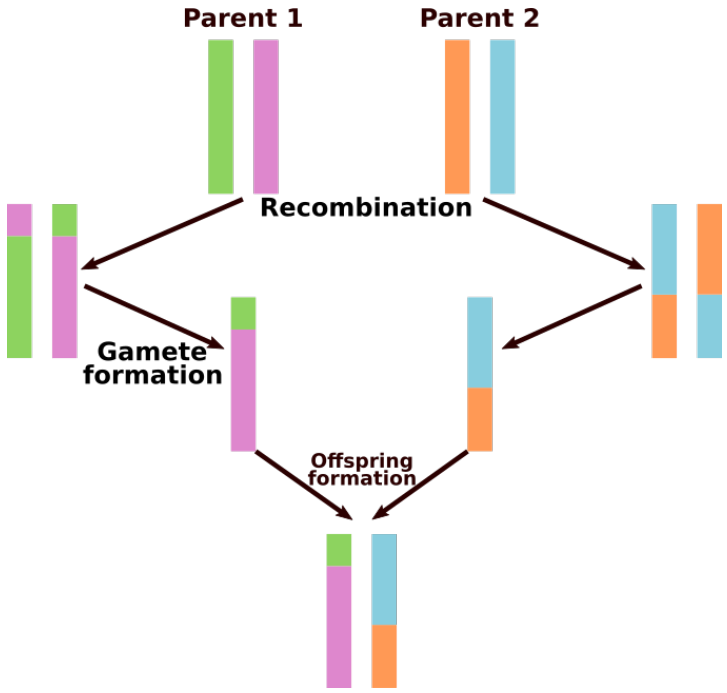

Figure 5.3: Reproduction event in the Eukaryotic version of Aevol. Two parents each undergo a meiotic recombination between their two chromosomes, and the offspring inherit one recombined chromosome from each parent.

and one chromosome carrying all the genes. Once this situation occurs, it is irreversible: having two copies of the chromosome that bears genes is equivalent to a whole genome duplication. In Aevol, that brutally multiplies by two the activation level of each trait, largely overshooting the phenotypic target. As such, it is a highly deleterious event that never goes to fixation.

Consequently, the only way to start a eukaryote simulation is to pre-evolve prokaryotic individuals with a halved phenotypic target, *i.e.* having divided by two the target activation level of all traits. Once a stable genome structure is reached in these conditions (typically after $10,000,000$ generations), we manually perform a whole genome duplication by importing a prokaryotic individual into a eukaryotic setting with twice the same chromosome and a restored phenotypic target. While not fully satisfying, this process ensures that we obtain stable diploid organisms that evolve with sexual reproduction and meiotic recombination. Before confronting these newly diploid organisms to different evolutionary conditions that we want to compare, it is recommended to let them evolve within the eukaryote framework for a while. This ensures that they adapt to these specific conditions and reach a stable state, thus that future differences we may observe are due to the later changes in evolutionary conditions and not to the adaptation to the eukaryote conditions.

Future extensions of the model could solve this problem by allowing whole genome duplication that would not be as deleterious. With this requirement, the first selected chromosome could be duplicated and replace the second degenerated in viable and diploid offspring. One way to achieve this would be to change the way the phenotype is computed in Aevol. Currently, each gene is translated and decoded into a triangle function with a given position on the trait axis, width, and height, and the phenotype is the sum of all these triangles. As such, a whole genome duplication doubles the phenotypes' function height and largely overshoots the target. To overcome this issue, a possibility could be to always normalize the phenotype such that the area under the curves remains constant. Thus, a whole genome duplication would not change the phenotype. While this would be another approximation, the idea is that the relative concentration of the different proteins is more important than their absolute concentration ("gene dosage effect"). However, this remains to be implemented and tested in Aevol.

5.6 POST TREATMENTS

5.6.1 *Impossible lineage study*

The most standard way to analyze a simulation in prokaryote Aevol is to retrieve the ancestral lineage of the final population and study its statistics. This allows following a single individual across time that harbors the mutations that went to fixation in the population. Due to sexual reproduction and recombination, this is no longer possible in the eukaryote version. While there can be coalescence when looking backward in time for any single gene of the extant population, the whole genome does not coalesce, and it is not possible to follow a single individual across time and gather lineage information.

The question of the study of the genetic ancestry of eukaryote-like populations with sexual reproduction and meiotic recombination has been explored during the PhD thanks to a collaboration with Manuel Lafond, Associate Professor at the University of Sherbrooke (Canada). While we did not find a satisfying way of studying a "lineage" in our Aevol eukaryote-like simulations, we found some interesting results on how different parameters such as the population size and the chromosome length affect how ancestral information is structured within populations. This work in progress in presented in the Appendix D.

5.6.2 *Population analyses*

As we could not restrict analyses to a single individual per generation, we resorted to record statistics on the whole population. Indeed, a standard Aevol simulation evolves populations under different con-

ditions and records at runtime their *fitness* and genome structure, *i.e.* total genome length, number of genes and RNAs, and coding and non-coding size. That alone suffices to observe the effect of certain parameters (such as population size, mutation rates, or selfing proportion) on genome architecture evolution.

Several post-processing options are also available in Aevol to better understand the processes and mechanisms explaining changes in genome structure at the population level. They allow us to study the difficulty of recombination (number of trials to find a recombination point and homology score at the recombination point), the mutational robustness of a population, or the replicative robustness of a population, *i.e.* the average loss in fitness after a replication event. While only statistics for the best individual and averages for the whole population are recorded at run time, it is also possible to extract detailed statistics for each individual at a given time step. This allows us to study the variance of our observed variables within populations.

Finally, additional post-treatment specific to a set of experiments are often developed, as is the case in Chapter 7, where we measure the replicative robustness while changing the reproduction mode (*i.e.* forcing selfing or non-selfing, regardless of the parameter used to simulate the population initially). These post-treatments will be presented in the chapter where they are used.

5.7 SOFTWARE AVAILABILITY STATEMENT

My work consisted of the development of a eukaryote prototype, which I used for the experiments presented in Chapter 6 and Chapter 7. It is available on my personal fork of Aevol on GitLab. Once validated, the prototype has been merged into the main repository for Aevol by David P. Parsons, research engineer in the Inria Beagle Team. Consequently, the eukaryote model is now also available *via* the main repository of Aevol and is presented in Luiselli et al. (2025b).

6

EUKARYOTE GENOME STREAMLINING ? EFFECT OF MUTATION RATE AND POPULATION SIZE ON EUKARYOTE GENOME SIZE REDUCTION

In Chapter 3, we observed that an increase in population size or mutation rate could provoke a genome size reduction in our prokaryote-like model organisms due to the selection for robustness to chromosomal rearrangements. However, these same parameters could have a different impact on eukaryote-like organisms, as it has been observed that population size reduction triggers different responses in prokaryotes and eukaryotes, which could be explained by different mutational biases: an insertion bias for eukaryotes and a deletion bias for prokaryotes (Ratcliff, 2024). While we showed in Chapter 3 that our results are robust to potential mutational biases, other differences between eukaryotes and prokaryotes could explain their different behavior.

To test whether the eukaryote-like characteristics introduced by the new eukaryote version of Aevol modify the reaction of genomes to changes in population size or mutation rate, we perform experiments similar to those of Chapter 3, within the new eukaryote framework. This serves both as a validation and test of the model and as a scientific exploration of the question of genome size evolution in a eukaryote context.

The results presented here were produced using Aevol-9, the version of the software that integrated the eukaryote prototype into the software's stable code base.

6.1 MATERIALS AND METHODS

6.1.1 *Model*

The eukaryote version of Aevol is described in Chapter 5.

6.1.2 *Experimental protocol*

Wild-Typing

Wild-Type generation

Starting from random genomes with one good gene, 10 prokaryote populations have been evolved for $10,100,000$ generations in a stable environment, allowing them to reach a stable genome structure. We then extract the common ancestor of the final population at generation $10,000,000$ and duplicate its chromosome to form a diploid organism.

We let the 10 diploid organisms evolve $1,000,000$ more generations in the eukaryotic setting with an adapted environment so that they adapt to diploidy, sexual reproduction, and meiotic recombination. There are 5 repetitions of each of these adaptation phases, totaling 50 simulations. Note that selfing is not allowed, and we transition from a local selection on a 3×3 neighborhood — the standard prokaryote selection condition — to a global selection.

In both the prokaryote and eukaryote parts of the experiments, the population size is kept constant at $N_0 = 1,000$, and the mutation rate for 6 types of mutations (substitution, small insertion, small deletion, duplication, large deletion and inversion) is constant at $\mu_0 = 10^{-6}$ per base pair. There are no translocations in these experiments.

During this wild-typing phase, the average fitness and genome structure (genome size, coding fraction) of the lineage — for the prokaryote phase — or of the populations — for the eukaryote phase — is recorded. This allows us to test whether the transition from prokaryote to eukaryote conditions yields unexpected changes in the evolution.

Wild-Type selection

To limit the number of experiments, only 5 of the 10 Wild-Types were selected for the second part of the experiment. As the aim is to observe changes in genome structure, we select the 5 Wild-Types that had, on average, the lowest variance in total genome size along the $500,000$ last generations of the experiments. There are two main reasons for this. First, the genome structure is expected to change in the early phase of the experiments as the organisms adapt to the new conditions, which is why the first half of the experiments are excluded from the variance computation. Second, it is known in Aevol that some Wild-Types are more stable than others, and we could observe that some repeats displayed huge variability in genome size despite the experimental conditions being kept constant, which could add noise to our observations. Since the Wild-Types selected have a very stable genome structure, any variation we detect will most probably be due to the changes in the experimental conditions.

Consequently, the second part of the experiments will be run from one random repeat extracted from Wild-Types 0, 3, 5, 6, and 7.

Effect of population size and mutation rate

Each of the 5 selected WTs is confronted with a change, either in population size or in mutation rates (both being multiplied by 4 or divided by 4), for 5 repetitions and $1,000,000$ generations. A control experiment is also run for the same duration, without changes in population size or mutation rate, totalling 125 simulations.

During the experiments, the average fitness and genome structure (genome size, coding fraction) for the population are recorded, allow-

ing us to compare genome architecture evolution under the different tested conditions.

6.1.3 *Data availability statement*

Simulations have been run with Aevol-9.1 (https://gitlab.inria.fr/aevol/aevol/-/tags/v9.1).

6.2 RESULTS

6.2.1 *Adaptation to the eukaryotic framework*

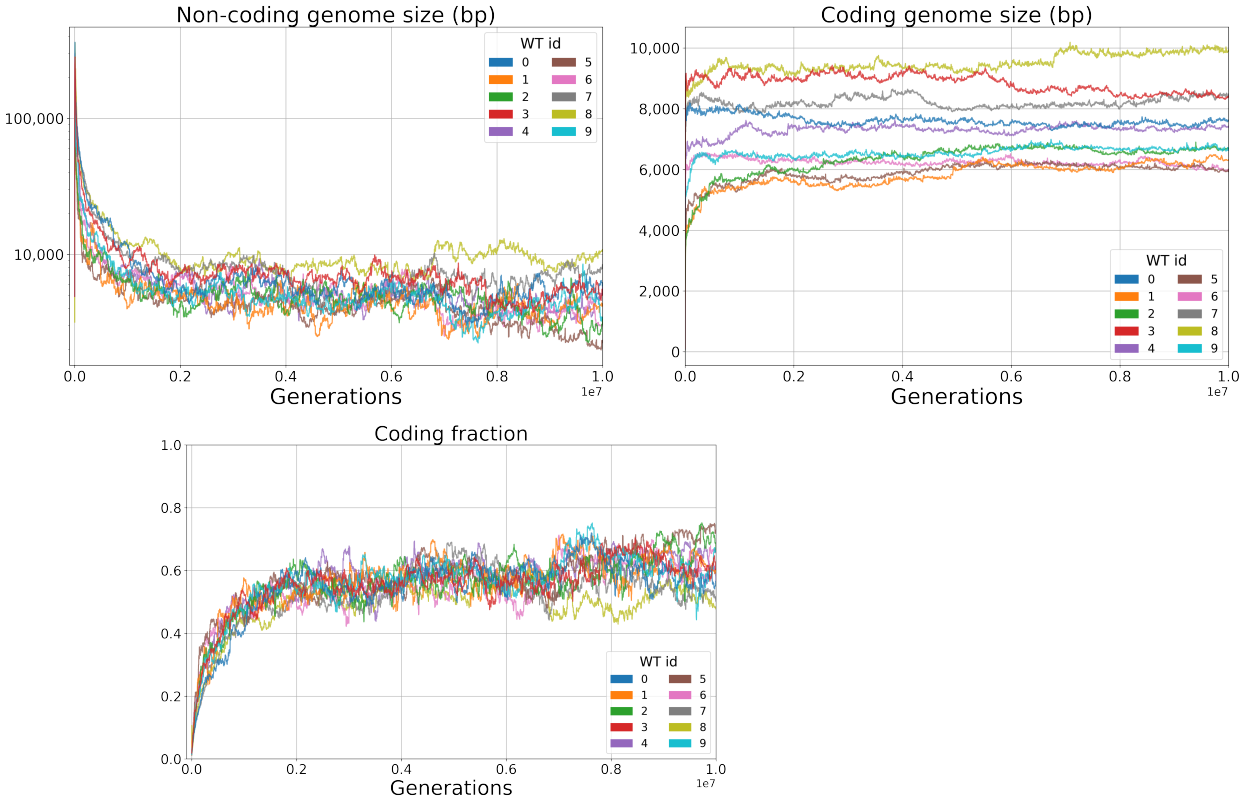

Figure 6.1: Non-coding (left) and coding (right) genome sizes, and coding fraction (bottom) of the prokaryotes Wild-Types, along 10 million generations. For each of the 10 WTs, the value displayed is the ancestor of the final population (at generation 10,100,000. The mutation rate is constant at $\mu_0 = 10^{-6}$ per base pair for each type of mutation, and the population size is constant at $N_0 = 1,000$.

First, we can note that the prokaryotes Wild-Types all behave very similarly (see Figure 6.1). They each undergo a substantial increase in non-coding genome size at the very start of the simulations, which will decrease slowly along generations to reach an equilibrium. The

coding size, on the other hand, is reached very rapidly and displays almost no variations after the first 1,000,000 generations.

After 10,000,000 generations, the prokaryote Wild-Types display a very stable genome structure, with a mean genome size of 12,621 bp across the simulations and a mean coding fraction of 60%.

Directly after the diploidization, genomes tend to contract slightly, losing mostly non-coding base pairs (see Figure 6.2). However, contrary to what happened in prokaryotes, this tendency is not retained, and the genome structure of many replicates is very unstable. The variation in coding fraction illustrates this well: while the prokaryote WTs vary between 50% and 70% of coding fraction, the eukaryotes vary between 40% and 80%. It is also notable that many individual simulations occasionally gain a lot of non-coding base pairs and sometimes also coding base pairs. These bursts of variability are always observed in the upward direction: no large genome contraction happens. Overall, there is no major change in genome structure introduced by the transition from the prokaryote to the eukaryote model.

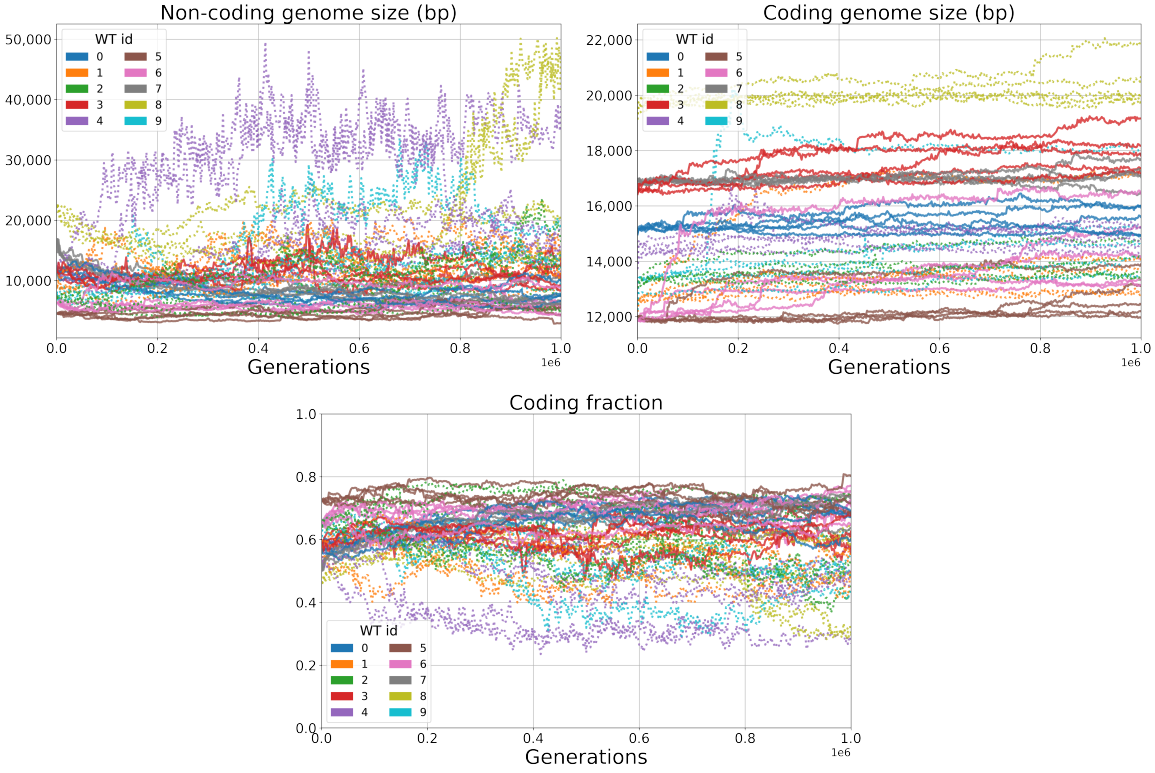

Figure 6.2: Non-coding (left) and coding (right) genome sizes, and coding fraction (bottom) of the eukaryote Wild-Types, along 1 million generations. For each of the 10 WTs, the value displayed is the mean of the population every 1,000 generations. The mutation rate is constant at $\mu_0 = 10^{-6}$ per base pair for each type of mutation, and the population size is constant at $N_0 = 1,000$. The 5 median Wild-Types, which are selected for the rest of the experiments, are in bold lines while the others are dotted.

To perform the experiments with varying population sizes and mutation rates, we extract the 5 Wild-Types with the lowest genome size variance, as described in Section 6.1.2.

6.2.2 *Change in population size*

Similarly to the experiments of Chapter 3, an increase in population size is associated with a decrease in the total genome size, which is driven by a decrease in the non-coding genome size (see Figure 6.3). Conversely, a decrease in population size is associated with an increase in the total genome size, which is driven by an increase in the non-coding genome size. The coding part of the genome also varies, but more slightly and in the opposite direction. Consequently, the coding fraction of the genomes is positively correlated with the population size, as was the case in the prokaryote experiments.

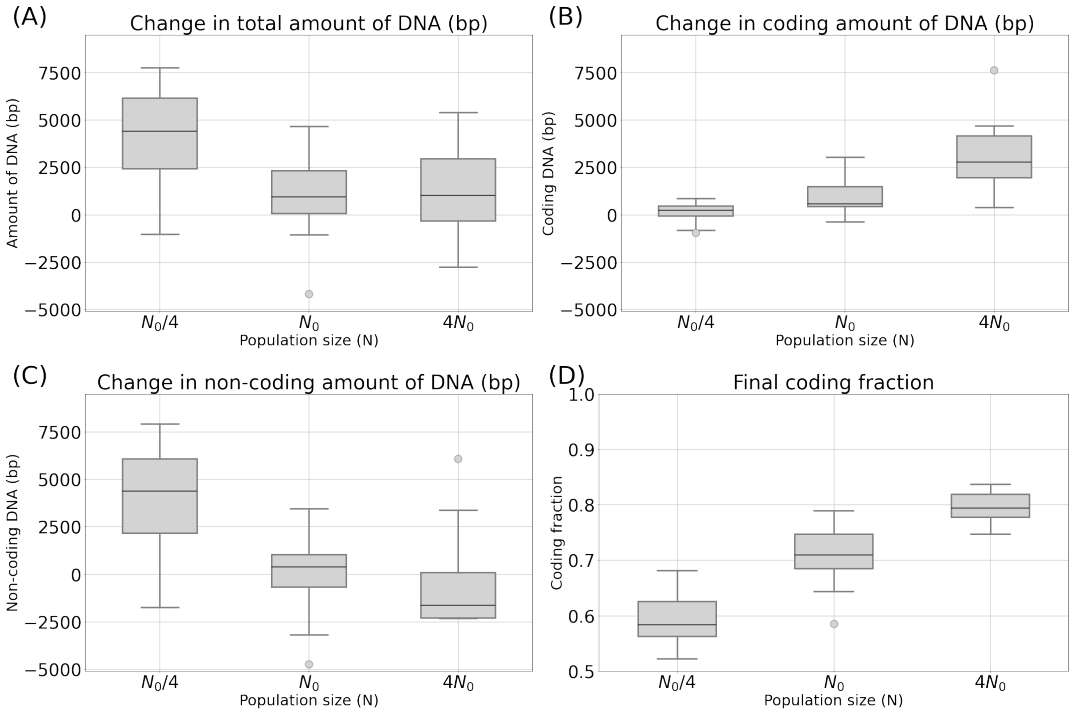

Figure 6.3: Total (A), coding (B) and non-coding (C) genome size variations, and final coding fraction (D), after 2 million generations. For each of the 5 WTs, 10 replicas were performed under a constant mutation rate ($\mu_0 = 10^{-6}$ per base pair for each type of mutation) with 3 different population sizes ($N_0 = 1,000$ being the control population size).

6.2.3 *Change in mutation rate*

Similarly to the experiments of Chapter 3, an increase in mutation rate is associated with a decrease in the total genome size, which is

driven mainly by a decrease in the non-coding genome — although the coding genome also decreases (see Figure 6.3). Conversely, a decrease in mutation rate is associated with an increase in the total genome size, but due to the low amount of mutations, this change happens slowly, and genomes are still far from their equilibrium sizes after 1,000,000 generations. Consequently, the coding fraction of the genomes is positively correlated with the mutation rate, as was the case in the prokaryote experiments.

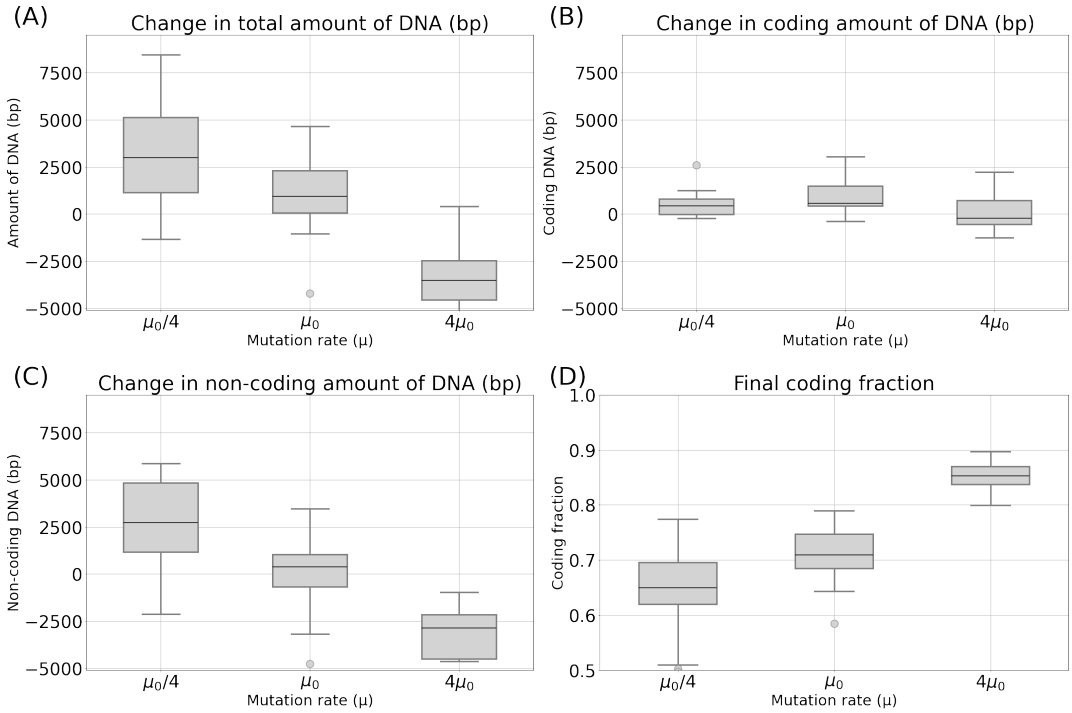

Figure 6.4: Total (A), coding (B) and non-coding (C) genome size variations, and final coding fraction (D), after 2 million generations. For each of the 5 WTs, 10 replicas were performed under a constant population size ($N_0 = 1,000$) with 3 different mutation rates ($\mu_0 = 10^{-6}$ per base pair for each type of mutation being the control mutation rate).

6.3 DISCUSSION

The results show that there is no difference in how genome architecture reacts to changes in either population size or mutation rate in the eukaryote-like model, compared to the prokaryote-like model. This tends to show that there is no difference in nature in how the genome architecture of both eukaryotes and prokaryotes is shaped. In particular, neither the transition to diploidy, to obligatory sexual reproduction, nor the introduction of meiotic recombination seems to provoke fundamental differences in how genomes react to changes in either N or μ. Consequently, the reactions to N and μ observed here are probably due to the same mechanisms as in prokaryotes,

i.e. the selection for robustness to chromosomal rearrangements. As such, differences in genome architecture between prokaryotes and eukaryotes could be solely — or mostly — explained by differences in population size and mutation rate, which lead to differences in the strength and importance of drift and robustness selection.

However, the simulation results could be further analyzed to measure the robustness of the organisms to a replication event and to the different types of mutations. Contrary to the prokaryote experiments, these analyses cannot be performed on the unique common ancestor of the population: because of sexual reproduction and recombination, we cannot isolate a single individual that would give information on the whole population. Therefore, new post-treatments have to be developed to analyze the population level instead of a lineage, and should be added to Aevol-9. This would allow reproducing the full set of experiments and measures performed in Chapter 3 to confirm that there is no difference in the selection of robustness to chromosomal rearrangements in both cases.

Despite eukaryotes and prokaryotes reacting the same way to N and μ within our framework, it is important to note that eukaryote genomes are much more unstable, as shown in Section 6.2.1. Indeed, some simulations undergo huge changes in genome size, which is overall much more variable in the eukaryote framework than in the prokaryote one. Several hypotheses could explain this: the eukaryote Wild-Typing phase lasted for only $1,000,000$ generations, while the prokaryote phase lasted $10,000,000$. While we assumed that a stable prokaryote would rapidly yield a stable eukaryote, the adaptation to the new experimental conditions could actually take longer than that, and it seems necessary to extend the experiments to $10,000,000$ generations to rigorously compare the variances in genome size for both types of organisms.

Another hypothesis is that the meiotic recombination, a newly introduced eukaryote characteristic that acts directly on the genome at each generation, could lead to genome instability due to the possible rapid accumulation of tandem duplications: once a segment is duplicated, it can recombine with its previous copy and create more copies at each generation. To better understand that, or to identify which other part of the eukaryote package leads to the genomic instability, it would be very interesting to run experiments with a subset of the eukaryote features: with and without recombination, with and without sexual reproduction, etc.

Finally, we know that in Aevol, the history of genomes carries a heavy weight: the genome instability could be the result of our prokaryote initiation. It is known in Aevol that experiments starting at a normal mutation rate that transition to a high mutation rate do not reach genome sizes as low as experiments immediately starting at high mutation rates. The early phase of the experiments influences the

range of possible genomes later on. As a consequence, it is possible that a native eukaryote would harbor a different — and potentially more stable — genome structure. To test this, profound changes must be brought to Aevol and the phenotype computation (see previous chapter, Section 5.5), but it would allow exploiting the full potential of this new version of Aevol.

Despite this current limitation, the eukaryote framework introduced already raises many interesting questions. While it has not altered the way genome architecture responds to changes in population size or mutation rate, it has added new parameters that could impact genome size evolution. Indeed, the reproductive mode (sexual *vs* asexual, proportion of auto- *vs* allo-fecundation) or the meiotic recombination parameters (minimal alignment score, number of recombination per generations, duration of homology search, etc.), could have a considerable impact on the evolution of genome architecture. The following chapter explores this by presenting experiments with different imposed selfing rates and studying the associated changes in genome architecture.

7

GENOME SIZE AND STRUCTURE: A DIRECT CONSEQUENCE OF REPRODUCTIVE MODE

FOREWORD

The following work is an ongoing collaboration with Diala Abu Awad, Associate Professor at the Paris-Saclay University.

Chapter 6 showed that eukaryotes' genomes react in the same way as prokaryotes' genomes when confronted with changes in population size or mutation rate. As such, the selection for robustness to chromosomal rearrangements seems similar in both realms, and the differences in genome architecture between them could be mainly due to their differences in effective population size and mutation rates, as shown in Chapter 4. The eukaryote version of Aevol, however, raises further questions on the specificities of eukaryotic genome architecture evolution. Indeed, it allows the exploration of new parameters, such as the proportion of auto- or allo-fecundation in a population.

In this work, we focus on the impact of the reproductive mode on genome architecture evolution by comparing populations with forced outcrossing, partial selfing, or almost mandatory selfing. We show that while outcrossing populations have a higher effective population size than selfing populations, they have larger genomes with more non-coding content. This contradicts the effect of effective population size documented in Chapter 3, Chapter 4, and Chapter 6 and shows that more complex phenomena can arise when introducing sexual reproduction.

The results presented here were produced using the eukaryote prototype I developed. They will be reproduced with Aevol-9, the stable version of the software that integrated my prototype, in the near future. Supplementary Materials for this chapter have been added as the Appendix F.

7.1 INTRODUCTION

Genome size is not only subject to rapid change on the evolutionary timescale but presents considerable variation within populations (Jeffery et al., 2016; González et al., 2022; Franco et al., 2024; Cang et al., 2023). Verbal and mathematical models predict that selection could act to reduce the total genome size to avoid unnecessary metabolic costs while still being able to carry out important functions (Lynch and Conery, 2003; Krakauer and Plotkin, 2002; Elena et al., 2007). Generally, the role of effective population size (N_e) is considered central to understanding the dynamics of genome size. Larger N_e should yield a better selection for streamlined genomes, whereas small N_e is expected to result in larger genome sizes. But to what extent genome size results from selection or genetic drift is an ongoing debate (Blommaert, 2020).

A proxy of N_e is the genetic diversity, which seems to be correlated with life-history strategies (Romiguier et al., 2014; Chen et al., 2017). There also are consistent findings of correlations between life-history traits and genome size (in eukaryotes Beaulieu et al., 2007; Knight and Beaulieu, 2008; Cutter, 2019; Bureš et al., 2024, and in prokaryotes Beier et al., 2022), pointing to the potential role of life-history in genome size evolution. In angiosperms, more specifically, it has been suggested that this correlation is due to the "large genome constraint" hypothesis (LGCH), wherein larger genome sizes negatively impact plant physiology and are thus selected against (Knight et al., 2005; Bureš et al., 2024). However, there is no direct evidence to support this hypothesis. Another angle would be to consider more mechanistic processes. Indeed, going beyond metabolic selection, observed tendencies may reflect a coevolution between life-history traits and mutation and recombination rates, processes that show a correlation with genome size (mutation: Sniegowski et al., 2000; Marais et al., 2008; recombination: Stapley et al., 2017, also see Lynch, 2007b) and are directly involved in reproduction and the transmission of genetic material.

As genome size evolution appears tightly linked to the way genetic material is transmitted, the reproductive mode is also an important parameter of the story. Self-fertilization is a widespread reproductive strategy in hermaphroditic animals and plants (Barrett, 2002; Jarne and Auld, 2006). It has been shown that the reproductive mode not only changes genome composition (*Arabidopsis thaliana*: Hu et al., 2011, *Caenorhabditis*: Fierst et al., 2015), but that there is a tendency for self-fertilizing species to evolve smaller genome sizes (Wright et al., 2008; Whitney et al., 2010). This trend is considered one of the markers of the "selfing syndrome" (Cutter, 2019). Smaller genome sizes in self-fertilizing populations are in disagreement with expectations on the role of N_e in genome size evolution, as selfing in associated with a lower N_e (Charlesworth and Charlesworth, 1987; Nordborg,

2000; Wang et al., 2016a). Indeed, controlling for phylogenetic non-independence, large-scale phylogenetic approaches yield no correlation between N_e and genome size in angiosperms (Whitney et al., 2010).

Transposable Elements (TEs) have been shown to be non-negligible contributors to genome size (Ågren and Wright, 2011; Hu et al., 2011; Kapusta et al., 2017). The duplication and spread of these elements are hypothesized to be better kept "under control" in self-fertilizing genome (Roze, 2023), potentially contributing to lower genome size in self-fertilizing species. However, while data on genome composition points to a correlation between reproductive mode and other elements of genome composition — such as the number and proportion of coding and non-coding DNA (Hu et al., 2011) —, there does not seem to be a correlation between reproductive mode and TE numbers, in some species at least (Tam et al., 2007). Therefore, self-fertilization may have an effect that goes beyond its interaction with the dynamics of TEs.

While solid groundwork has been laid in the development of theoretical models to study how self-fertilization affects recombination and mutation rates (Roze and Lenormand, 2005; Gervais and Roze, 2017; Stetsenko and Roze, 2022), it is not clear how the consequences of self-fertilization on drift and selection, congruently with these changes in rates, would influence genome size. Modeling the dynamics of genome composition and size is an arduous and simulation-heavy task. Aevol is an *in-silico* experimental evolution tool developed specifically to study genome structure (Knibbe et al., 2007a; Banse et al., 2024b). Initially based on microbial (bacterial) organisms, a recent extension to account for sex and recombination has been integrated (Luiselli et al., 2025b). We use this extension to study whether and how self-fertilization influences genome size within this framework. We found that populations that self-fertilize tend to have smaller genome sizes despite a smaller N_e.This difference in genome size is mostly explained by a reduction in the non-coding genome size in populations that self-fertilize, *i.e.* these populations present a more streamlined genome.

7.2 RESULTS

Due to model constraints, the initial stages of the simulations are run within a prokaryote framework. Individual then undergo a diploidization stage and are let to adapt to the eukaryote conditions for $1,000,000$ generations before being confronted with changes in their reproduction mode. Details are provided in Section 7.4

7.2.1 *Relationship between genome structure and fitness*

Right after the diploidization stage, simulations are run for a further one million generations to allow adaptation to the new genomic configuration (see Materials and methods, Section 7.4). During this time, both genome size and population fitness continue to evolve. In Figure 7.1, we show four of the ten simulations run to illustrate possible outcomes.

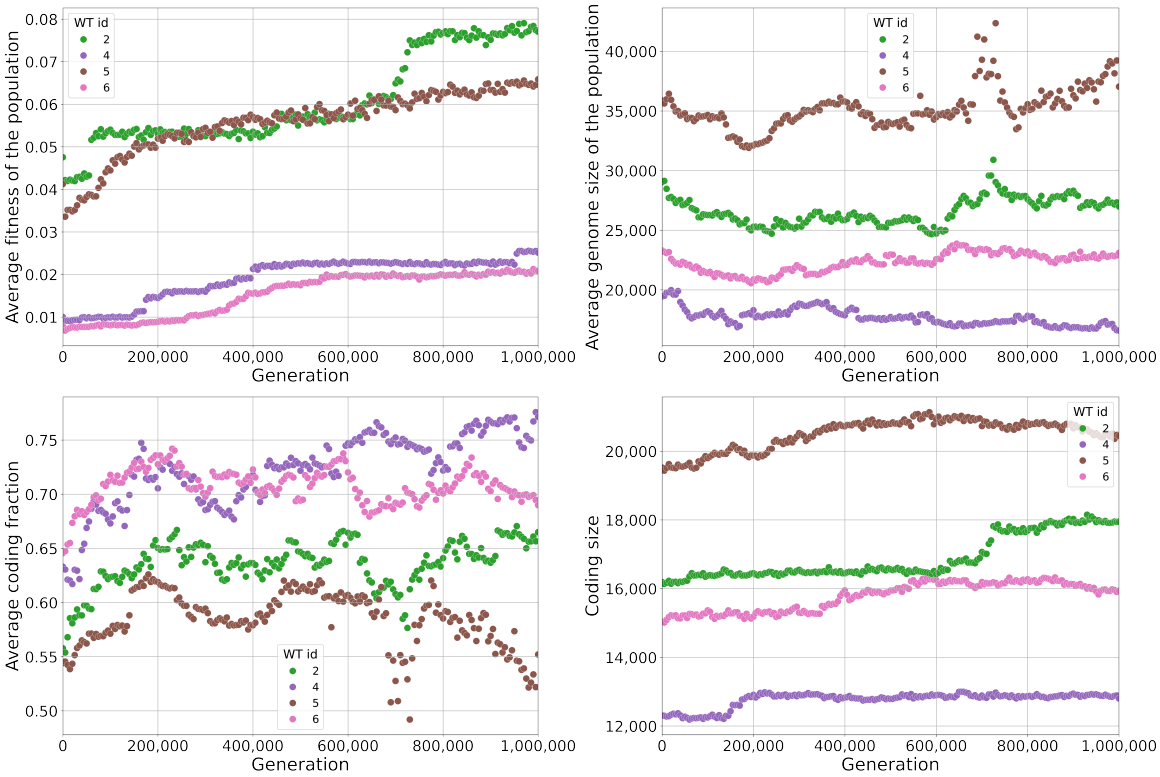

Figure 7.1: Average fitness (A), genome size (B), coding fraction (C), and coding size (D) for 4 different populations during 1,000,000 generations after their diploidization. The full data for the 10 populations are shown in Supp. Figure F.1

Changes in total genome size do not necessarily reflect a change in fitness (compare Figure 7.1A and Figure 7.1B, Table 7.1). Although, on average, the fitness and coding genome size of the populations seem positively correlated, this is not significant when using only the ten WT populations. This was probably due to low statistical power (Table 7.1) and to one of the WT simulations, WT3, presenting extreme values (see Supp. Figure F.2 and Figure F.4). Analyzing the same metrics using the results from 50 simulations run a further 500,000 generations (five replicates for each WT that remained outcrossing (self-fertilization rate of 0, see methods), we see the same tendencies in correlation coefficients as for the WT populations, but it is significant between fitness and coding, and fitness and total DNA (Table 7.1,

	Correlation coefficient	p-value
Coding DNA	0.576 (0.647)	0.082 (6.701×10^{-7})
Non-coding DNA	0.236 (0.205)	0.511 (0.162)
Total DNA	0.333 (0.372)	0.347 (0.009)
Proportion of coding DNA	-0.152 (-0.070)	0.676 (0.638)

Table 7.1: Spearman correlation coefficients and p-values for the relationship between fitness and coding, non-coding, and total DNA 500,000 generations after the beginning of the simulation for nine of the ten initial wild-types (WT3 was excluded, see text). Between brackets, the same values were calculated after a further 500,000 generations, but using five replicates of each wild-type population in which selfing was maintained at 0.

values between brackets). On the other hand, the absolute quantity of non-coding and the proportion of coding DNA are not significantly correlated with fitness (also see Figure 7.1a and Figure 7.1c).

The positive correlation between total genome size and fitness is mostly driven by the increase in coding DNA. This is in agreement with the underlying assumptions of the Aevol framework, as a larger coding genome allows for a better fine-tuning of the phenotype. Yet, the amount of coding DNA does not predict fitness, as populations with the same amount of coding DNA can have quite different fitnesses, and *vice versa*. Indeed, contingency plays a very important role in how genome structure (coding *vs* non-coding DNA) evolves with time.

Because of the extreme variance observed in population WT3's trajectory (see Supp. Figure F.2), it was excluded from the analyses that follow. Results including WT3 are in the Supp. Mat. and are referred to throughout the text. The tendencies in the relationship between self-fertilization and genome structure, though no longer significant, are unchanged when including this run.

7.2.2 *Effect of self-fertilization on genome size and fitness*

Once the WT lineages were deemed to be at equilibrium, self-fertilization was introduced. Each WT population was used as a starting point for five replicates per selfing rate tested (0%, 50%, and 95%). After 500,000 generations, the trajectories were not yet at a stable equilibrium, but we could make out some general tendencies. Though the changes were quite small, at 500,000 generations, we found significant negative correlations between the selfing rate and changes in total genome size, coding-DNA, and non-coding DNA (Table 7.2).

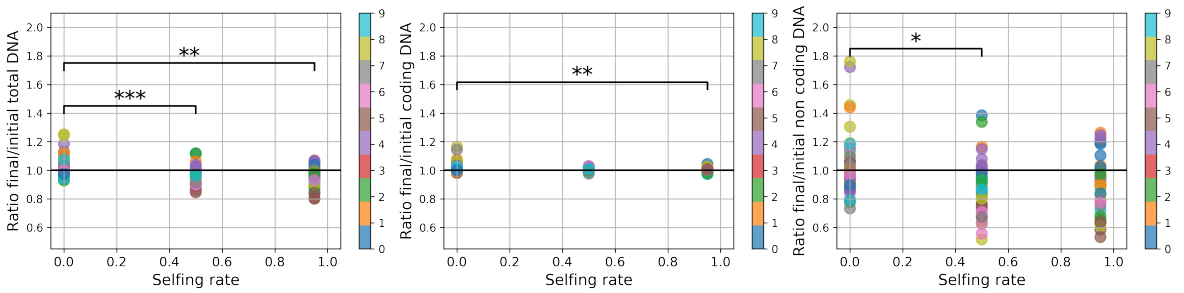

Figure 7.2: Changes in total, coding and non-coding DNA for all simulations after 500,000 generation, color-coded for WT lineage. Stars indicate significance threshold, after Bonferroni correction: $* : p < 0.05$, $** : p < 0.005$, $*** : p < 0.0005$ (Supp. Table F.1 for full results). Due to its atypical behavior, WT3 has been excluded from the analyses, but the full data are available in the Supp. Figure F.5

	Correlation coefficient	p-value
Fitness ratio	-0.225	0.007
Coding DNA ratio	-0.346	2.01×10^{-5}
Non-coding DNA ratio	-0.194	0.019
Total DNA ratio	-0.237	0.004

Table 7.2: Spearman correlation coefficients and p-values for the relationship between the selfing rate and ratios of fitness, coding, non-coding, and total DNA (value after 500,000 generations over value at the start of the simulation). For each selfing condition, there are five replicates of each of the nine wild-types (excluding WT3).

The differences in genome size between the different selfing rates is driven by both an increase in coding and non-coding DNA in non-selfing populations and a decrease in non-coding DNA in selfing populations (see Figure 7.2). The degree of self-fertilization (50% or 95%) is of little importance, the main differences being between non-selfing and selfing populations (significance indicators on Figure 7.2). Over time, the mean total genome size tends to decrease for higher self-fertilization rates and increase for non-selfing populations (Supp. Figure F.3). A closer look at the coding structure revealed that, while WT lines have an average of 15,667 coding base pairs and 248.8 genes, after 500,000 generations, outcrossing populations have an average of 16,012 coding base pairs and 251.5 genes and self-fertilizing populations (at a self-fertilization rate of 95%) have 15,678 coding base pairs and 247.2 genes. So, there is both an increase of base pairs and genes in the coding DNA of outcrossing populations, whereas self-fertilizing populations show no change in the number of base pairs but have a reduced number of genes.

The negative correlation in total genome size is mainly driven by the change in non-coding DNA, as attested by the strong similarity

in trajectories of the change in the total amount of DNA and in the amount of non-coding DNA (Figure F.3b and d) and the much higher relative change in the amount of non-coding DNA at 500,000 generations (Figure 7.2c). More precisely, WT lines have on average 7,258 non-coding base pairs, while after 500,000 generations outcrossers have 7,334 base pairs and selfers have 6,070: the difference is greater than in the coding size of the genomes. In the Aevol framework, a decrease in the amount of non-coding DNA is observed in populations with a large effective population size N_e or a higher mutation rate μ (Knibbe et al., 2007a; Luiselli et al., 2024). As we did not change the mutation rate, we estimated N_e in our simulations to quantify its effect on genome structure in the following section.

7.2.3 *Estimating the effective population size (N_e)*

Higher self-fertilization rates should decrease N_e even in the absence of selection, as the rate of coalescence is automatically increased (*i.e.* the homologous copies of a given gene in an individual can coalesce in a single generation Nordborg, 2000). As mentioned previously, within the Aevol framework, a decrease in the proportion of non-coding DNA occurs in populations with high N_e, as selection is more efficient and favors more streamlined genomes (Luiselli et al., 2024). In Table 7.3, we calculate the effective population size using equations based on the effective number of reproducing individuals and the variance in reproductive success. We find that the effective population size behaved as expected (Wang et al., 2016a), with higher self-fertilization rates resulting in smaller N_e (see Table 7.3).

This shows that the effective population sizes vary as expected, and thus a higher N_e is not the explanation of the streamlined genomes in the selfing populations. In Aevol, genome size is known to covary with the robustness of genomes: more compact genomes imply a better replicative robustness and potentially a lower mutational robustness (Luiselli et al., 2024, 2025a). We therefore examine below how self-fertilization affects these variables and processes and how they may contribute to the observed patterns in genome structure.

Selfing rate	Avg nb of reproducers	Ne [1]	Ne [2]
0	777	860	859
0.5	714	658	716
0.95	603	455	609

Table 7.3: Estimates of effective population sizes for different selfing rates. [1]: measure of N_e based on the variance in the number of offspring per reproductive individual: $N_e = \frac{4N}{2(1-\frac{\alpha}{2-\alpha})+V(1-\frac{\alpha}{2-\alpha})}$ with α the selfing rate and V the measured variance in the number of effectively reproducing individuals (Caballero and Hill, 1992). [2]: $N_e = \frac{4N-2}{2+V}$ from (Wright, 1938).

7.2.4 *Mutational robustness*

Mutational robustness is defined as the capacity of a genome to withstand a new mutation and maintain its fitness. To obtain the distributions of mutational fitness effects (DFEs) for each selfing rate, we randomly introduce mutations of each category (deletions, point mutations, duplications, etc.) into random individuals of our populations. Figure 7.3 shows that self-fertilization rates have little effect on the DFEs. Mutations are slightly more likely to be beneficial in outcrossing populations, but the difference is minimal, and the proportion of neutral or deleterious mutations remains approximately constant.

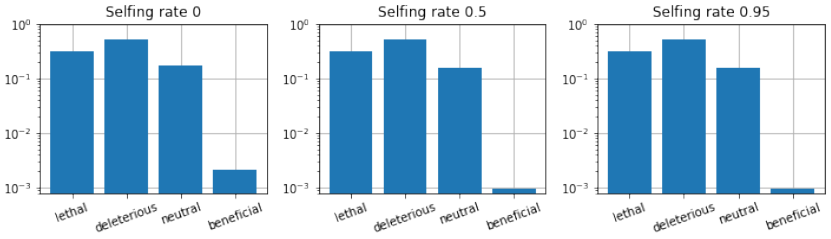

Figure 7.3: Measured mutational robustness to any mutation. For each of the 40 simulations, 10,000 mutations of each type are performed on random individuals, and we compare the fitness before/after the mutation. Plotted values are the proportion (on a log-scale) of mutation landing in each of the 4 categories based on their selective coefficient s: lethal ($s \leqslant -0.999$), deleterious ($-0.999 < s \leqslant -0.001$), neutral ($-0.001 < s \leqslant 0.001$), or beneficial ($s > 0.001$). These data exclude WT3, but the full data, as well as the DFEs per mutation type, are presented in the Supplementary Materials Section F.6.

7.2.5 *Recombination efficiency*

Another event able to change the genetic information is the recombination: misalignment or a lack of suitably similar genomic regions can be mutagenic. Recombination efficiency can be approximated in our simulations by measuring the number of trials to find homologous regions between the two chromosomes and the homology score at the recombination breakpoints. It gives information on the danger of the recombination events. We find that not only was recombination more efficient in self-fertilizing populations (less time was needed to find a suitable recombination site than in outcrossing populations, Figure 7.4A) but that misalignment is more frequent in outcrossing populations (Figure 7.4B).

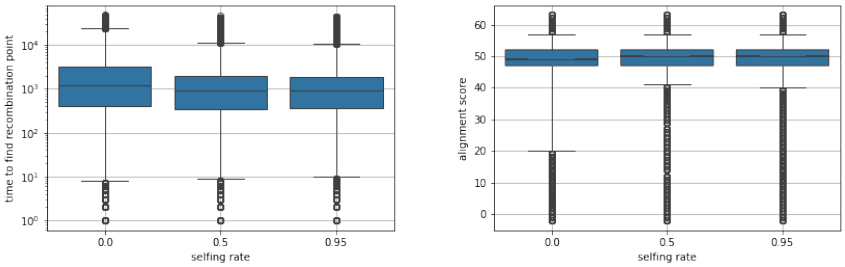

Figure 7.4: (A) Number of tries to find a successful recombination and (B) Alignment score at the recombination points for the different selfing rates. Whiskers of the box plots mark the 1^{st} and 99^{th} percentiles, highlighting that more recombination events have a lower score and that finding a suitable site takes more time in outcrossing populations. The detailed distributions are available in Section F.7.

7.2.6 *Replicative robustness*

We have shown that mutations and recombinations are both affected by the selfing rate, although rather slightly. These differences, in combination, could have had consequences on shaping selection on genome structure. To account for them simultaneously, we examined the replicative robustness of populations (*i.e.* the heritability of fitness). Replicative robustness is simply the similarity in mean parental and offspring fitness. High replicative robustness indicates offspring are, on average, close to the parental fitness. The more robust the genotype is to mutations and the more efficient recombination, the higher the replicative robustness. Genome size also impacts replicative robustness since mutations happen on a per-base basis and hence are, on average, more frequent in larger genomes.

To measure the replicative robustness in our populations, we randomly draw (with replacement) 10,000 individuals from each population and make them perform a reproduction event, with selfing or outcrossing. Then, we compare the fitness of the parent(s) to the fitness of the offspring.

Self-fertilizing populations are found to be more robust. A higher proportion of replication events were perceived as "neutral", *i.e.* offspring had the same fitness as the mid-parent, whether offspring were produced through outcrossing or self-fertilisation. On the other hand, outcrossing populations had a higher proportion of offspring with either a higher or a lower fitness than the mid-parental fitness: they have a lower robustness but also more potential for improvement. This could be due to comparing the offspring with the average fitness of the parents, but the statement holds when the comparison is with the best parent (see Supp. Figure F.16).

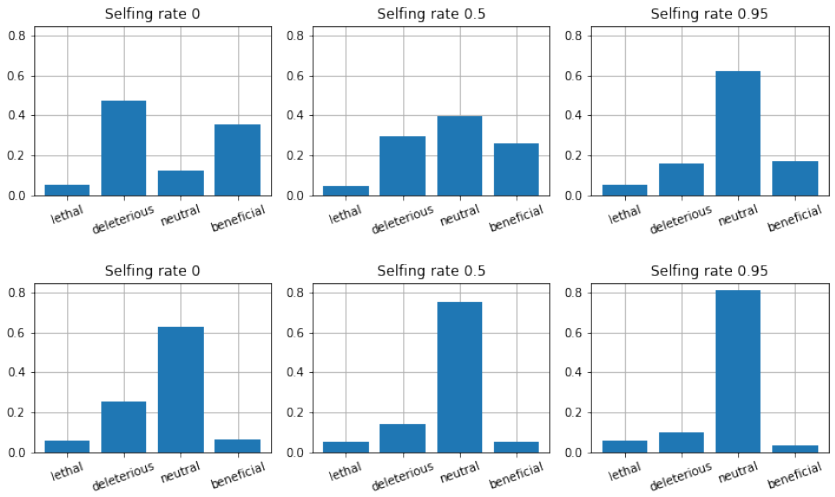

Figure 7.5: Replicative robustness (ratio of offspring fitness to mid-parental fitness) in offspring produced via outcrossing (top — first generation of outcrossing, compared to mid-parental fitness) and self-fertilization (bottom — comparison of F2 offspring with F1 offspring), independently of the source population's selfing rate. The selective coefficient of the replication event is the ratio of the fitness after and before the replication event minus 1; see Figure 7.3 for the definitions of the four categories. Comparison with the best parent instead of the mid-parental fitness is presented in Supp. Section F.8.

7.3 DISCUSSION

The Aevol environment provides a complete framework to examine how genome size and structure evolve. Previous works have explored genome size evolution in asexual prokaryote populations using Aevol (Knibbe et al., 2007a; Luiselli et al., 2024), and have explored the effects of different parameters (such as population size and mutation rates) on these dynamics. Here, we used a new version of Aevol that models eukaryote individuals, capable of sexual and non-random reproduction (Luiselli et al., 2025b). Our results agree with observations from natural populations, with increased self-fertilization decreasing genome size (Whitney et al., 2010), despite a reduced effective population size. It has been argued that the main difference between outcrossing and self-fertilizing populations is the number of Transposable Elements (TEs) (Wright et al., 2008). We found that although this hypothesis receives support in the literature (Ågren and Wright, 2011; Hu et al., 2011; Kapusta et al., 2017; Roze, 2023), it is not necessary to explain modifications of coding to non-coding DNA ratios. Even in the absence of TE dynamics, as is the case in our experiments, self-fertilization still favors a decrease in genome size. In our experiments, genome size differences between outcrossing and self-fertilizing populations were driven by an increase in both coding and non-coding DNA in the

former and a decrease in non-coding DNA in the latter. We discuss the implications of our results below.

7.3.1 *Larger genomes, better fitness*

In our simulations, the increase in fitness is directly correlated with an increase in coding DNA (Table 7.1). In Aevol, populations are continuously adapting to the optimum through small increments, and the longer the simulation runs, the more fitness is expected to increase. As outcrossing favors the maintenance and generation of phenotypic and genotypic variance (see Supp. Figure F.4 to see variance in replicate populations), this results in a faster increase in population fitness — and in coding DNA. Indeed, higher variability in offspring increases the probability of introducing new phenotypes that are better adapted (Clo et al., 2020).

The greater the genotypic variance of offspring, the greater the capacity for populations to branch out to non-local optima. This is often accompanied by an increase in the absolute amount of coding DNA, potentially allowing the introduction of more functions through an increase in the number of genes or fine-tuning the effects of existing genes through an increase in their length. Outcrossing populations show a clear increase in the number of coding genes, whereas self-fertilizing populations show a decrease in the number of genes. This result in is in agreement with what has been observed in *Arabidopsis* (Hu et al., 2011). More coding genes imply that outcrossing populations can have more complex phenotypes, and can more easily evolve higher fitnesses. These dynamics explain the mechanics behind the higher fitness in outcrossing populations.

Self-fertilizing populations, on the other hand, are more locally adapted, with less genetic variance. Though the absolute number of genes is smaller in self-fertilizing populations, there is a potential increase in the fine tuning of the functions these genes code for, attested to by a greater number of base-pairs per gene. Because of this, they have fewer opportunities to "jump" to other functional, and perhaps better, fitness optima.

Another consequence of this is that there are fewer potentially beneficial mutations in self-fertilizing populations (Figure 7.3), as any new mutation would take them away from their local optimum, but not far enough to approach a new optimum. Outcrossing populations are more spread out around the optimum due to their higher variance. Therefore, they are continuously adapting to it, which (slightly) increases the proportion of beneficial mutations. Additionally, Figure 7.5 compares the fitness of the offspring to the mean fitness of the parents. As parents are more likely to have different phenotypes in outcrossing populations, if an offspring has a phenotype that is closer to the fitter parent, this would seem to increase the replicative

robustness through outcrossing, despite not bringing true innovation (compare Figure 7.5 to Supp. Figure F.16 for a measure of replicative robustness by comparing offspring to the best of their parent instead of the average).

As outcrossing populations have more genetic variance and hence a higher effective population size N_e (Table 7.3), it would have been expected that they would also have less non-coding DNA (previous Aevol results). This was, however, not the case. The effect of the rate of self-fertilization on genome size reduction is thus a consequence of other processes.

7.3.2 *A streamlined and efficient genome*

A streamlined genome, *i.e.* with less non-coding DNA, is selected with a higher N_e or mutation rate (Luiselli et al., 2024, 2025a). The expectation is that a higher effective population size (N_e) would favor genomes with higher proportions of coding DNA, as "useless" non-coding DNA would be less likely to accumulate through drift (Lynch, 2007b). However, self-fertilization is known to decrease N_e (Charlesworth and Charlesworth, 1987; Nordborg, 2000), which also happens in our experiments (Table 7.3). As such, our results pointing to a reduction in non-coding DNA in selfing populations go against existing theoretical expectations.

Theoretical works have pointed out that self-fertilization favors a reduction in the mutation rate (Gervais and Roze, 2017). While in this paper this could only occur through a reduction of the per-base mutation rate, there are other ways of achieving a lower overall mutation rate: the reduction of illegitimate recombinations, or the reduction of the genome size. The latter hypothesis requires the mutation rate to be per base, which is supported by data, with smaller genomes having lower genome-wide mutation rates (Sniegowski et al., 2000; Lynch, 2007b).

First, illegitimate recombination, *i.e.* recombination between two non-homologous sites, can be mutagenic and have important consequences on fitness. In our simulations, self-fertilizing populations present a higher recombination efficiency (Figure 7.4). This means that there were few, if any, non-homologous recombination events compared to outcrossing populations. This could be part of a feedback loop favoring genomes that are not too structurally different within the population, or a direct side-effect of having fewer reproductive individuals and lower genetic diversity in the population.

Most importantly, our per base-pair mutation rate is fixed, but we observe a reduction in genome size in self-fertilizing populations, thus reducing the genome-wide mutation rate, in agreement with theoretical expectations (Gervais and Roze, 2017). Indeed, non-coding base pairs represent a mutational hazard in themselves due to the risk

of deleterious mutations in these regions (Lynch, 2007b), *e.g.* following double strand breaks in the case of structural mutations (Luiselli et al., 2025a). Reducing non-coding DNA is easier than reducing coding DNA because it has fewer direct consequences on population fitness and is less likely to be deleterious. Thus, the genome size reduction we observe is mainly in the non-coding part of the genome (see Figure 7.2).

This genome size reduction in self-fertilizing populations is consistent despite a lower effective population size. There can be several explanations for this: either the strength of selection is enhanced by self-fertilization, or the self-fertilization increases the selective cost of having more non-coding bases. The first hypothesis is supported by the literature, as non-random mating is known to change the efficiency of selection (Glémin, 2003; Roze, 2015), thus potentially facilitating the purge of non-coding DNA that harbors a mutational hazard. Another possible explanation could be that having more non-coding DNA contributes to decreasing the negative effect of mutations on fitness, notably by acting as a buffer, and thus be selected at higher N_e. This does not seem to be the case in our simulations, as the DFE was not greatly affected with regard to deleterious mutations (see previous section and Figure 7.3).

A last hypothesis on why smaller genomes could be favored in self-fertilizing populations is that in our simulations, a smaller genome implicitly increased the per base-pair recombination rate. The genome-wide recombination rate *per se* remained constant in our simulations, as it was fixed at one recombination event per reproductive event. So, following a logic similar to that concerning the mutation rate, there may be selection to modify the recombination rate. Theoretical works have indeed suggested that self-fertilization should increase the recombination rate, allowing for more efficient selection by decreasing linkage disequilibria between loci (Roze and Lenormand, 2005; Stetsenko and Roze, 2022).

7.3.3 *Conclusion*

Using individual-based simulations in which genome size and structure are allowed to evolve, we highlighted the possible impact of self-fertilization on the evolution of genome architecture. Our results reflected what has been observed in natural populations, namely that self-fertilizing populations evolve smaller genomes and do so consistently, despite a lower effective population size. We also confirmed previous findings that this change in genome size was due to a reduction in non-coding DNA. We hypothesized that this could be a consequence of selection for more streamlined genomes that were less likely to accumulate deleterious mutations due to a reduced genome-

wide mutation rate and an increased genome-wide recombination rate.

7.4 MATERIALS AND METHODS

7.4.1 *Model description*

Aevol is a software designed to study genome structure (Knibbe et al., 2007a; Banse et al., 2024b; Luiselli et al., 2025b). It presents a simple population structure: a fixed number of individuals, replaced at each generation, with a reproductive success biased by their level of phenotypical adaptation. The phenotype of an individual is encoded in its genome. Aevol focuses on the realism of the genome structure, with coding and non-coding regions that can evolve freely in size and content. In the sequence, consensus patterns are identified to mark promoters and Shine-Dalgarno-like sequences (start of RNAs and proteins), and hairpin structures are recognized to mark RNA terminators. Consensus sequences, terminators, and gene sequences are considered coding, and the rest of the genome is non-coding as it can be removed without causing any change to the phenotype. Any part of the genome can switch from coding to non-coding and *vice versa* through mutation events. Mutations happen directly on the sequence, independently of their effect on fitness.

The decoding from protein sequences to a phenotype and a fitness value relies on a simplified mathematical vision: from the primary sequence of each protein is computed a triangle function with a specific width (w), height (h) and position in the trait space (m), and the sum of all triangles represents the phenotype of the individual. That phenotype is compared to a target function, representing the ideal phenotype in a given environment, and their difference gives the individual's fitness.

At each reproduction, mutations can happen at random on the sequence, without an *a priori* fitness effect. There are different types of mutations: local mutations, which would follow a polymerase slippage (substitutions, short indels), and chromosomal rearrangements, which would follow double-strand breaks (inversions, deletions, duplications).

For this project, we used the eukaryote version of Aevol (Luiselli et al., 2025b): each individual owns two linear homologous chromosomes, and they reproduce sexually with a mandatory meiotic recombination event driven by sequence homology. More details on the model are available on the website (`aevol.fr`).

7.4.2 *Experimental protocol*

We started from 10 pre-evolved haploid populations, with a population size $N = 1,000$. These populations adapted to their environment for $10,000,000$ generations. One individual for each population is extracted and undergoes an artificial diploidization to populate 10 new populations ($N = 1,000$). These populations evolve during $1,000,000$ more generations, with sexual reproduction and meiotic recombination but without selfing, to adapt to these new genomic conditions.

Then, we take the best individual from each of the 10 populations and create 15 clonal populations of size $N = 1,000$ from it: five for each selfing condition (0% of selfing, 50% of selfing or 95% of selfing). This results in 150 simulations, for which we record the average fitness, coding, and non-coding genome sizes along $500,000$ generations.

7.4.3 *Post-evolution analyses*

Once our populations evolved for $500,000$ generations under the different selfing conditions, we measure the final fitness and genome architecture for all individuals of all populations, which allows measuring the variance of characters within and between populations. A final generation is also run to record which individuals would reproduce in the current generation.

We also measure the robustness to replications or mutation events: for each population, we draw $10,000$ individuals (with replacement), and we record their fitness and genome structure. To measure mutational robustness, we apply a mutation and record the new fitness and genome structure, enabling us to measure the average fitness loss for any mutation kind. To measure the robustness to a replication event, we perform a full replication (selection of second parent, recombination, and mutations) and compare the fitness of the offspring to the fitness of the parents. We measure this with either 100% or 0% of selfing, and in the first case, we perform two replications to compare F1 and F2 individuals. This accounts for the heterozygosity of the population and enables us to highlight how self-fertilizing populations and non-self-fertilizing populations have developed different genome structures that change their robustness. It guarantees that the differences we observe in the robustness measures are not solely due to the type of reproduction at a given time.

Finally, we also measure the average number of trials to find a sufficiently homologous pair of recombination points. To measure this, we draw $10,000$ individuals (with replacement) from the final population and apply the process to find the meiotic recombination breakpoints. We record the number of trials before finding a pair of sufficiently homologous points, and the homology score at the chosen positions.

7.5 DATA AVAILABILITY STATEMENT

The software code is available on GitLab: https://gitlab.inria.fr/jluisell/aevol-eukaryotes. More details on Aevol can be found on the website www.aevol.fr.

8 CONCLUSION

8.1 HOW CHROMOSOMAL REARRANGEMENTS SHAPE GENOMES

In this thesis, we explored the effect of chromosomal rearrangements on genome architecture evolution. We have shown that chromosomal rearrangements represent a mutational input that has important consequences for genome evolution. They provide unique paths of evolutionary innovations and induce a second-order selection that limits genome size (Chapter 2). This selection for robustness to chromosomal rearrangements is modulated by the mutation rate and biases and the population size (Chapter 3, Chapter 4, Chapter 6), and maybe also by the reproductive mode (Chapter 7). Rearrangements also contribute a mutational bias towards genome size growth, due to the asymetry of their probabilities of being neutral or lethal (Chapter 4). As a result, chromosomal rearrangements allow for an equilibrium non-coding genome fraction.

More specifically, we have shown that the effect of rearrangements is tightly linked to the selection for robustness, *i.e.* the selection for the capacity of genomes to withstand perturbations. Notice that robustness is difficult to characterize in itself, as it can refer to perturbations such as mutations (mutational robustness), a replication event and the associated probability to mutate (replicative robustness, which depends partially on the mutational robustness, as showed in Chapter 4), or even an environmental change or an outside perturbation (developmental or phenotypic robustness). This last type of robustness can also be correlated with mutational robustness (Masel and Siegal, 2009; Kaneko, 2009): if there is a regulatory buffer ensuring the cell ends up in the same state despite different inputs from the environment (*i.e.* through different regulatory paths, or on-off switches that remain in the same state for a range of environmental conditions), mutating a part of the regulatory network might be invisible in a majority of environments. As such, being robust to environmental change can be correlated with being robust to mutations under certain conditions, and the different types of robustness are not independent from one another. Additionally, and similarly to the selection for phenotypical adaptation, there can be several forms of selection for robustness: the positive selection for robustness would increase the resistance of genomes to the perturbation at stake, while the purifying selection for robustness would prevent any loss in robustness. These different modes of selection for the different types of robustness are likely to ex-

ert multiple — and potentially opposing — pressures on the evolution of genome architecture.

Indeed, there is a wide variety of possible interactions between these different forms of robustness and how they are selected: for example, genomes that have more non-coding DNA than coding DNA have a better mutational robustness since a given mutation has fewer risks to affect a gene and deteriorate the fitness of the individual. However, individuals that have a smaller total genome size have a better replicative robustness, as they undergo on average fewer mutations. As such, starting from a given genome architecture, removing non-coding base pairs increases the replicative robustness of the genome but decreases its mutational robustness. This means that the lineage would undergo fewer mutations, but these mutations would be, on average, more deleterious than in a non-reduced genome. Importantly, these two effects perfectly compensate each other in the case of substitutions, but not in the case of chromosomal rearrangements. Indeed, the mutational robustness does not increase as much as the replicative robustness decreases upon the addition of more non-coding base pairs, resulting in a selection for more reduced genomes. Note that in the case of chromosomal rearrangements, we consider the phenotype to not be robust to mutations affecting the coding part of the genome due to their large-scale effect. However, selection for replicative robustness would be affected by actually how robust the phenotype is to mutational events. As such, the selection for replicative robustness induced by InDels or other local mutation events might be more intertwined with selection for mutational robustness in more complex ways.

Chromosomal rearrangements not only induce a second-order selection for robustness, but they also entail an intrinsic bias in their neutrality that contributes to a neutral genome size growth, as demonstrated in Chapter 4. This can be associated to a border-induced phenomenon (Loewenthal et al., 2022) and is not specific to chromosomal rearrangements, as InDels also contribute to it. It must be accounted for when studying genome architecture: the insertion of some base pairs is more likely to be neutral — in terms of level of phenotypical adaptation – than the deletion of some. As a result, chromosomal rearrangements contribute in opposite ways to the evolution of genome architecture.

These opposing effects of chromosomal rearrangements on genome architecture give rise to complex dynamics, and one could be tempted to study them exhaustively to better understand under which conditions which genome architecture ends up evolving. However, rearrangements have a huge combinatorics (Banse et al., 2024a) that prevents extensive studies, even in models. Additionally, they are highly deleterious and hence rarely present in lineage and population studies from genomic data, such that an extensive exploration of

their potential effects is not possible *in vivo* or *in vitro* either. Having access to only a scarce part of a very broad range of mutational events complicates theoretical analysis around chromosomal rearrangements. Models can help grasp part of their complexity, but more theory is still needed, such that the impact of each type of chromosomal rearrangement could be studied individually, as well as their interactions with one another. This thesis opens new ways of thinking about and studying chromosomal rearrangements and raises new questions about genome architecture evolution.

8.2 PERSPECTIVES

In this thesis, we studied the impact of chromosomal rearrangements on the evolution of genome architecture through the prism of the robustness selection they induce, the variability in mutations they bring, and their neutral mutational bias. While we demonstrated that these dynamics depend notably on the mutation rate, the true mutation rate of chromosomal rearrangements is still debated. Indeed, since most chromosomal rearrangements are likely to be lethal, it is impossible to observe and count them. As a consequence, initial studies proposed quite low mutation rates (of the order of 10^{-14} per base in human, according to Shaffer and Lupski (2000)[1]). According to recent studies, however, the rate of rearrangements could be quite high: of the order of 10^{-11} per base pair for *E. coli* (Raeside et al., 2014)[2], or even from around 10 times less to the same order of magnitude than the substitution rates (Lipinski et al., 2011; Molari et al., 2025; Wei et al., 2018; Saxena and Baer, 2025)[3]). More precise estimates of these mutation rates will be needed to verify the importance of the second-order selection that chromosomal rearrangements can induce, as discussed in Chapter 4. Moreover, this rate can also evolve, thus changing the per-genome mutation rate at constant genome size, which is not possible in our experiments. A key perspective of this thesis would thus be to study the co-evolution of mutation rate and genome size induced by the second-order selection for robustness due to chromosomal rearrangements.

Other mutational events probably also contribute to the mutational bias towards genome growth and/or to the induction of a selection for robustness. Indeed, not all double-strand breaks result in chro-

1 To achieve this number, the authors approximate the spontaneous mutation rate with the observe ferquency of rearrangements in populations, which ignore any lethal rearrangemets. They give the rate per individual, which, divided by the genome size, gives a rate per base pair.

2 110 rearrangements over 40,000 generation and 12 lineages, in genomes of approximately 5Mb. This also ignores lethal and non-fixed rearrangements.

3 Note that Lipinski et al. (2011) compare per base substitution rates and per gene rearrangements rate, but also that they only consider duplications and not other rearrangements.

mosomal rearrangements: some are repaired without mutations, or induce a local mutation on the breakpoints, while some others are not repaired, probably causing cellular death or more complications later in the cellular cycle. Moreover, the double-strand breaks initiating chromosomal rearrangements in the models used in the thesis were assumed to be uniformly distributed along the genomes, but other distributions of the breakpoints could impact the evolution of genome architecture by changing the probability for rearrangements to be deleterious or the average size of mutational events.

Transposable Elements (TEs) are also important contributors to genome size in eukaryotes (Marino et al., 2024), and potentially interact with both chromosomal rearrangements and recombination if they happen on homologies since TEs spread similar sequences throughout the genome. As such, they would be a valuable addition to our models, both in the prokaryote (Insertion Sequences) and in the eukaryote frameworks. TEs would both contribute a mutational bias towards genome growth — due to them spreading in genomes, potentially in an exponential way — and create an additional pressure for robustness selection — due to the danger they pose in themselves and the additional ectopic recombination they could provoke. Each of these opposing dynamics depends on the number of TEs and their transposition rate, thus probably yielding complex dynamics that cannot necessarily be predicted beforehand.

Several behaviors could be observed depending on the strength of selection for robustness and the activity of the TEs. First, TEs could have only a limited impact on the equilibrium non-coding percentage but simply occupy the non-coding "space" enabled by the $N_e \times \mu$ conditions. Indeed, we have shown in Chapter 4 that, depending on the conditions, introducing an additional mutational bias towards insertion does not necessarily displace the equilibrium, contrary to what could be expected, probably because the selection for robustness is too strong and any new transposition would be quickly removed. Second, TEs could be strongly counter-selected and eliminated from genomes — in which case, they probably do not displace the equilibrium either. This probably happens under very high $N_e \times \mu$ conditions, in which the selection for streamlined genome is very strong (Lynch, 2006b). Finally, TEs could exert an important pressure on genome size increase and significantly displace the equilibrium non-coding percentage, as observed with mutational biases in Chapter 3 (in the Aevol framework) and in parts of the parameter space tested in Chapter 4 (in the mathematical model).

To understand these different regimes and the conditions of their emergence, we would probably need to first better characterize the different forms of selection for the various types of robustness. This would allow deriving which conditions influence which evolution-

ary force. One way of achieving this could be to implement more different measures of robustness in Aevol and conduct large-scale experiments to measure how they vary under different conditions before introducing Transposable Elements or other elements of complexity.

Such an extensive study could also focus more in depth on the discrete differences between prokaryotes and eukaryotes to understand how they may affect genome architecture. Indeed, the new eukaryote flavor of Aevol could allow studying separately the impact of being diploid, of performing sexual reproduction, and of having a mandatory meiotic recombination on genome architecture evolution. The influence of meiotic recombination could also be refined: for now, there is a single and mandatory recombination event in Aevol, but this number could be variable and depend on genome size or on other parameters, or be set to higher values to test how this would affect the different forms of robustness and the resulting genome architecture. However, an in-depth study of the discrete prokaryote/eukaryote differences would require the ability to evolve eukaryotes from scratch in Aevol, which is still an ongoing work. This would necessitate changing the way phenotypes and fitness are modelled and computed, which could have unpredicted impacts on genome evolution, as explained in Chapter 5.

Another potentiality of the new model to better understand the differences and similarities between eukaryotes and prokaryotes is to reproduce the main experiments that were carried out on the prokaryote version of Aevol and compare the results. This concerns for example the evolution of complexity (Liard et al., 2020), of mutator alleles (Rutten et al., 2019), or the impact of population structure (Misevic et al., 2015).

In addition to studying eukaryotes relative to prokaryotes, the new eukaryote version of Aevol developed during this PhD thesis also opens up broad perspectives for the study of eukaryotes in themselves. It can be use to study how ectopic recombination events influence genome architecture: the probability of ectopic recombinations can be modulated by the homology score required to perform a recombination event, as well as by the distribution from which the pairs of points to test are drawn. This could help unravel the potential adaptive role of ectopic recombinations, as well as their robustness cost and the strength of selection for robustness they induce.

Finally, this thesis focused, through the prism of chromosomal rearrangements, on selection for robustness and largely ignored selection for phenotypical adaptation. Indeed, while there is selection for phenotypic adaptation always ongoing in Aevol, and we supposed a strong purifying selection in the mathematical model, it was not the focus of our experiments and results. Despite that, we observed some interesting behaviors that would gain in being explored further: in Chapter 3, some lineages lose phenotypical adaptation and robustness upon their

exposure to new evolutionary conditions and have to first regain some robustness before resuming their phenotypical adaptation.

It goes without saying that organisms must first be able to survive and reproduce before selection on other, more complex features can emerge, but even then, a feature that cannot be faithfully inherited cannot be selected. Hence, both the selection for phenotypical adaptation and for robustness are tightly intertwined. While it was important to first isolate the effect of selection for robustness to chromosomal rearrangements to understand it and how it affects genome architecture, an important perspective of the presented work is to study its interactions with the selection for phenotypical adaptation — often simply called *fitness* selection. Large-scale experiments and more theory crafting could allow distinguishing conditions in which selection for one or the other is dominant, whether concerning positive or purifying selection. Studying interactions between fitness selection and robustness selection would also raise the question of their interaction with selection for evolvability, which can be defined as the ability of a population to generate adaptive genetic variation (Wagner and Altenberg, 1996; Pigliucci, 2008).

While the fitness value is already present in the experiments with Aevol, it is not easy to study its interaction with the robustness selection. A solution would be to compare the actual temporal trajectory of both fitness and robustness instead of comparing the end values. This requires a more complex analysis framework for the simulations. Instead, a first approach to study the interactions between fitness and robustness could be through the mathematical model presented in Chapter 4: for now, in our mathematical framework, the fitness is strictly binary (an organism is either alive with a perfect fitness, or dead), but it would be interesting to introduce a more complex fitness function, that could for example depend on the number of genes, and see how it changes the equilibrium of the resulting genome architecture. This would give a first insight into how fitness selection can interact with robustness selection to select for a specific genome architecture.

8.3 CONCLUSION

In summary, this thesis provides key insights into genome evolution and pinpoints chromosomal rearrangements and selection for robustness as major contributors to genome architecture. It raises new questions and opens perspectives for the study of chromosomal rearrangements and how they interact with other evolutionary factors.

Chromosomal rearrangements are complex mutations with huge combinatorics and many intricate effects on genome architecture evolution, pushing both towards a genome expansion and a genome reduction through different forces. They open the way for numerous

evolutionary paths and a wide diversity of possible outcomes. As such, the mere existence of chromosomal rearrangements and the complexity of their effect on genome evolution could partly explain the huge diversity of genome architectures observed throughout the Tree of Life, as we have shown here for the non-coding fraction.

In different chapters of the thesis, we changed the selection/drift equilibrium by changing the census population size and showed how the respective strengths of selection and drift change genome architecture. However, many other factors can change the strength of selection and the selection/drift equilibrium: the mutation rates and biases, the environment and its potential changes, population structure, the reproduction mode, etc. To summarize these factors, the selection/drift equilibrium is often quantified using N_e, the effective population size. Our experiments in general, and more particularly the mathematical model presented in Chapter 4, demonstrate that N_e greatly influences genome architecture evolution. Since virtually every parameter influences N_e, this means that virtually every parameter influences genome architecture evolution.

Moreover, genome architecture evolution is so complex that different factors influencing N_e in the same direction might actually have different effects on genome architecture evolution. Several factors could act on the strength of the positive or purifying selection, for robustness or for phenotypical adaptation, and result in similar changes in the measures of N_e, while impacting differently the genome evolution. For example, it can be noted that a diminution in the effective population size due to a reduction in the census population size leads to expanded genomes with more non-coding bases in Chapter 3 and Chapter 6, while the introduction of selfing in Chapter 7 both leads to a reduced effective population size and a reduced genome. As such, N_e might be a useful concept to quantify drift and compare different conditions, but it may hide too much of the complexity of the evolutionary forces at stake — thus preventing a better understanding of genome evolution. It is already well documented that the different proxies used to measure N_e give different information (Waples, 2022), and that there are short-term and long-term values of N_e that also bear different information (Brevet and Lartillot, 2021). Even for a given timescale, N_e is a measure that tries to condensate a lot of information into one value and thus could be misleading on the evolutionary forces acting on genome architecture evolution

Finally, fitness has been referred to throughout the manuscript and generally refers here to the level of phenotypical adaptation. This is the definition used within the Aevol framework, and also applies to the mathematical model of Chapter 4 — although, in that case, it is limited to "alive" or "dead". Outside Aevol however, several definitions coexist for the concept of fitness: the ability to produce

offspring in a given environment, or the capacity to survive and transmit its genes — *i.e.* the average contribution of an individual to the gene pool of the next generation. The thesis shows that the concept of robustness actually questions these definitions of fitness: Why limit the definition to the contribution to the following generation only? Is there a sense in quantifying the contribution to the next generation if the contributed offspring are not themselves able to contribute to the following generation? If robustness is the ability to faithfully transmit genotypic information to the next generations, is it not just a multi-generational fitness? In that sense, the opposition between fitness selection and robustness selection may just be an opposition of the immediate *vs* long-term time scale and not an opposition between two different features.

In this thesis, for the sake of simplicity, robustness was measured over a single generation, and hence it was closely intricated with fitness — the first condition for being able to faithfully transmit its genetic information is to have a chance of transmitting any information, *i.e.* to reproduce. To some extent, this definition of robustness could entail the fitness, hence the use of the level of phenotypical adaptation instead of fitness in most of the manuscript to avoid the ambiguity. But in the future, the pertinence of separating the two concepts, or at least the ways of measuring one independently of the other, should be discussed. Integrating these two concepts could be key to unraveling more complex dynamics of genome evolution.

This leads to a new paradox: we have suggested that N_e is a unique parameter that probably condenses too much information and hence loses an important part of said information, which makes it harder to understand genome evolution in the light of this value only. As a consequence, N_e should probably be split into several concepts and measures, better able to explain genome architecture evolution. On the other hand, fitness and robustness are two concepts often considered separately, while this separation could be quite artificial, and they could, in fact, be two sides of the same coin. Hence, an integrated view that considers both of them together instead of opposing them or studying them separately could yield a conceptual breakthrough in our comprehension of genome evolution. This shows how defining unifying parameters both helps us understand complex phenomena, but also limits our capacity to broaden our understanding.

Similarly, models such as Aevol provide a simplifying view of genome evolution, which helps us understand complex dynamics and isolate the impact of some parameters on genome architecture. To provide useful insights, some models need to get more complex to allow for the study of more phenomena — the eukaryote version of Aevol is more complex but allows for new experiments, to raise and answer new questions —, but models also need to be simplified to highlight the generality of what they are demonstrating — as is done

by our mathematical model of non-coding genome size evolution. In any case, intertwining simple and complex models helps to define and describe useful concepts while erasing part of the complexity they aim to describe. If "All models are wrong, but some are useful" (George Box), we may add that two models that are differently wrong are even more useful.

Part III

APPENDIX

A

SUPPLEMENTARY MATERIALS FOR *FORWARD-IN-TIME SIMULATION OF CHROMOSOMAL REARRANGEMENTS: THE INVISIBLE BACKBONE THAT SUSTAINS LONG-TERM ADAPTATION*

A.1 AEVOL: A FORWARD-IN-TIME EVOLUTIONARY SIMULATOR WITH COMPLEX MUTATIONS

Aevol (https://www.aevol.fr) is a forward-in-time evolutionary simulator that simulates the evolution of a population of haploid organisms through a process of variation and selection (Knibbe et al., 2007a; Beslon et al., 2010; Parsons et al., 2010; Frenoy et al., 2013; Batut et al., 2013). The design of the model focuses on the realism of the genome structure and of the mutational process. Aevol can therefore be used to decipher the effect of chromosomal rearrangements on genome evolution, including their interactions with other types of mutational events.

In short, Aevol is made of three components (Fig. 1A from the main text):

- A mapping that decodes the genomic sequence of an individual into a phenotype and computes the corresponding fitness value.
- A population of organisms, each owning a genome, hence its own phenotype and fitness. At each generation, the organisms compete to populate the next generation.
- A genome replication process during which genomes can undergo several kinds of mutational events, including chromosomal rearrangements and local mutations. The seven modelled types of mutation are depicted on Fig. 1B (main text) and entail three local mutations: substitutions, small insertion, small deletion, two balanced rearrangements (which conserve the genome size): inversions and translocations, and two unbalanced rearrangements: duplications and deletions. This allows the user to study the effect of chromosomal rearrangements and their interaction with other kinds of events such as substitutions and InDels.

A.1.1 *The Genotype-to-Phenotype-to-Fitness map*

Genome representation. Each artificial organism, similarly to prokaryotes, is asexual, haploid, and owns a single circular chromosome. The

genome is encoded as a double-strand binary string containing a variable number of genes separated by non-coding sequences (Figure A.1). Genes are delimited by predefined signaling sequences indicating transcription and translation. The number of proteins an organism owns thus depends on its signaling sequences, and can evolve through mutational events.

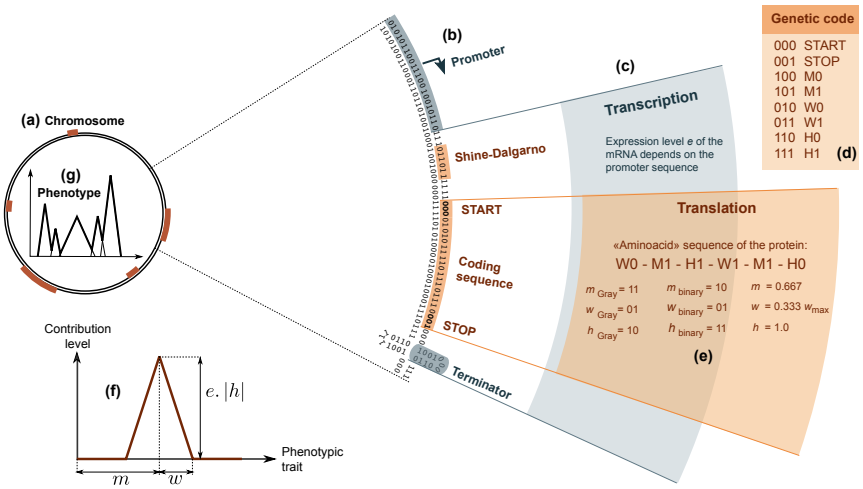

Figure A.1: The Aevol model. In the model, each organism owns a circular double-strand binary chromosome **(a)** along which genes are delimited by predefined signaling sequences **(b)**. Promoters and terminators mark the boundaries of RNAs **(c)**, within which coding sequences can in turn be identified between a Shine-Dalgarno-START signal and a STOP codon. Each coding sequence is then translated into the primary sequence of a protein, using a predefined genetic code **(d)**. This primary sequence is decoded into three real parameters called m, w and h **(e)**. Proteins, phenotypes, and environments are represented similarly through mathematical functions that associate a level in $[0,1]$ to each abstract phenotypic trait in $[0,1]$. For simplicity reasons, a protein's contribution is a piecewise-linear function with a triangular shape: the m, w and h parameters correspond respectively to the position, half-width and height of the triangle **(f)**. All proteins encoded in the chromosome are then summed to compute the phenotype **(g)** that, once compared to the environmental target, can be used to compute the fitness of the individual.

Transcription starts at promoters, which are defined in the model as sequences that are close enough to an arbitrarily chosen consensus sequence (0101011001110010010110 in all simulations presented here, with at most $d_{max} = 4$ mismatches). The expression level e of an mRNA is determined by the similarity between the actual promoter and the consensus sequence: $e = 1 - \frac{d}{d_{max}+1}$ with d the number of mismatches ($d \leq d_{max}$). This models the interaction of the RNA polymerase with the promoter, without additional regulation.

When a promoter is found, transcription proceeds until a terminator is reached. Terminators are defined as sequences that would form a

stem-loop structure, as the ρ-independent bacterial terminators do. The stem size is here set to 4 and the loop size to 3.

The translation initiation signal is $011011****000$, corresponding to a Shine-Dalgarno-like sequence followed by a START codon 000. When this signal is found on an mRNA, the downstream Open-Reading-Frame (ORF) is read until the termination signal (the STOP codon 001), is found. Each codon lying between the initiation and termination signals is translated into an abstract "amino-acid" using an artificial genetic code, thus giving rise to the sequence of the protein (Figure A.1). Transcribed sequences (mRNAs) can contain an arbitrary number of ORF, with some mRNAs possibly containing no ORF at all (non-coding mRNAs) and others possibly containing several ORFs (polycistronic mRNAs). Importantly, the relative fractions of non-coding, monocistronic and polycistronic mRNAs are not predefined but result from the evolutionary dynamics and are likely to be influenced by the evolutionary conditions (Parsons et al., 2010).

Protein function and phenotype computation. We define an abstract continuous one-dimensional space $\Omega = [0, 1]$ of phenotypic traits. Each protein is modeled as a mathematical function that associates a contribution level between -1.0 and 1.0 to a subset of phenotypic traits (negative contribution correspponding to inhibiting the trait). The range of phenotypic traits to which a single protein can contribute is limited by $2 \times W_{max}$, where W_{max} defines the maximum pleiotropy degree. Hence, increasing W_{max} indirectly reduces the total number of proteins required to cover the whole phenotypic space. Similarly to the K parameter of the classical NK-fitness landscape (Kauffman and Levin, 1987), increasing W_{max} increases the level of pleiotropy and hence the ruggedness of the fitness landscape.

For simplicity, we use piecewise-linear functions with a symmetric, triangular shape to model protein effect (Figure A.1). This way, only three parameters are needed to characterize the contribution of a given protein: the position $m \in \Omega$ of the triangle on the axis, its half-width w ($w \leq W_{max}$) and its height $h \in [-1, 1]$. This means that this protein contributes to the phenotypic traits in $[m - w, m + w]$, with a maximal contribution h for the traits closest to m. Thus, various types of proteins can co-exist, from highly efficient (high h) to poorly efficient (low h) and even inhibiting (negative h) and from highly specialized (low w) to versatile (high w).

In this framework, the primary sequence of a protein is interpreted in terms of three interlaced binary subsequences that will in turn be decoded as the values for the m, w and h parameters (Figure A.1). For instance, the codon 010 (resp. 011) is translated into the single amino acid $W0$ (resp. $W1$), which means that it concatenates a bit 0 (resp. 1) to the code of w. Mutations in the coding sequences, including of course local mutations but also chromosomal rearrangements, can

change these values and hence change the protein's contribution to the phenotype.

The contribution of all the proteins encoded in the genotype of an organism are combined to get the final level for each phenotypic trait. This is done by first scaling all protein contributions by the transcription rate e of the corresponding mRNA (see above), then by summing the mathematical functions of all the proteins, with bounds in 0 and 1. The resulting piecewise-linear function $f_P : \Omega \to [0,1]$ is called the phenotype of the organism.

Fitness computation. In the model, fitness depends only on the difference between the levels of the phenotypic traits and target traits levels, which are defined by a user-defined mathematical function $f_E : \Omega \to [0,1]$. This target function indicates the optimal level of each phenotypic trait in Ω and is called the environmental target. In usual Aevol experiments, f_E is the sum of several Gaussian lobes with different standard deviation, maximal height and centers. It can be stable over evolutionary time, or change stochastically.

The difference between f_P and f_E is defined as $\Delta := \int_\Omega |f_E(x) - f_P(x)|dx$, $\forall x \in \Omega$ and is called the "metabolic error". It is used to measure adaptation penalizing both the under-realization and the over-realization of phenotypic traits. Given the metabolic error of an individual, its fitness f is given by $f := \exp(-k\Delta)$ with k a fixed parameter regulating the selection strength (the higher k, the larger the effect of metabolic error variations on the fitness values).

To illustrate this computation, we can look at the organisms in Fig 2 of the main text: at generation 0 (Fig 2a), there is a single gene, thus a single protein in the proteome. Since in our experiment $k = 1,000$ (Figure A.3), we exponentiate $-1000\times$ the difference between the environmental target (the sum of Gaussian lobes in grey) and the phenotype (the single black triangle), and we obtain a fitness $f = 1.1 \times 10^{-66}$. To compare, our Wild Type (Fig. 2b) has 58 genes, so its phenotype is the sum of the 58 triangles depicted in the proteome and its fitness is $f = 0.0517$: its phenotype is much closer to the environmental target than at generation 0.

A.1.2 *Population model and selection process*

The population is modelled as a toroidal grid with one individual per grid cell. At each generation, the fitness of each individual is computed, and the individuals compete to populate each cell of the grid at the next generation. This competition can be fully local (the 9 individuals in the neighborhood of a given cell competing to populate it at the next generation, Figure 2.1A) or encompass a larger subpopulation. If the selection scope encompasses the whole population, all individuals compete for all grid cells. Importantly, the more local the selection scope, the more the population model diverges from the

panmictic Wright-Fisher model as local selection increases the effective population size N_e for a given census population size (Waples, 2010).

Given a selection scope, the individuals in the neighborhood $\mathcal{N}$ of a given grid-cell compete through a "fitness-proportionate" selection scheme: the probability p_j, for an individual j with fitness f_j to populate the focal grid-cell at the next generation is given by $p_j = f_j / \sum_{i \in \mathcal{N}} f_i$.

A.1.3 *Genetic operators*

During their replication, genomes can undergo sequence variations (Figure 2.1). An important feature of the model is that, given the Genotype-to-Phenotype map (Section A.1.1), any genome sequence can be decoded into a phenotype (although possibly with no trait activated if there is no ORF on the sequence). This allows to implement – and test – any kind of mutational process. In the classical usage of the simulator, seven different kinds of mutations are modelled (depicted on Figure 2.1B). Three mutations are local (substitutions and small insertions or deletions), and four are chromosomal rearrangements, either balanced (with no change in genome size): translocations and inversions, or unbalanced: duplications and deletions.

Local mutations happen at a position uniformly drawn on the genome. Substitutions change a single nucleotide. InDels insert (or delete) a small sequence of random length – and random composition for insertions. The length of the sequence is drawn uniformly between 1 and a maximum value (6 by default). Notably, InDels occurring within an ORF can shift the reading frame or simply add/remove codons, resulting in very different evolutionary outcomes.

Chromosomal rearrangement breakpoints are uniformly drawn on the chromosome, the number of breakpoints depending on the type of rearrangement (Figure 2.1B). Hence, chromosomal rearrangements can be of any size between 1 and the total genome size, allowing to investigate the effect of small structural variants that are indeed observed *in vivo* (Musumeci et al., 2000; Audrézet et al., 2004; Blakely et al., 2006; Xue et al., 2023).

The rates μ_t at which each type t of genetic mutation occur are defined as a per-base, per-replication probability. This means that the number of spontaneous events is linearly dependent on the length of the genome. However, its fixation probability depends on its phenotypic effect (for instance, a mutation affecting exclusively an untranscribed region is likely to be neutral). Hence, the distribution of fitness effects (DFE) of any kind of mutation is not predefined but depends on the intertwining of its effect on the sequence, and of the genome structure. For example, the fraction of coding sequences or the spatial distribution of the genes along the chromosome change the probability of a given mutation to alter the phenotype, and the fitness, an effect

that is especially important for chromosomal rearrangements. Having an emergent DFE instead of a predefined one enables investigating the complex direct and indirect effects of chromosomal rearrangements on the evolutionary dynamics.

A.2 SOFTWARE USAGE

Aevol is based on running and analyzing forward-in-time simulations. More specifically, any experiment with Aevol is divided into four main steps. The first step consists in preparing a simulation with the `aevol_create` command. This reads the parameter file (Figure A.3 and Table A.1) and creates a population of organisms at generation zero according to the specified values. `aevol_run` then simulates the evolution starting from the initial population or a from backed up population for a given number of generations.

`aevol_run` outputs several data files: summary statistics regarding the best individual at each generation (fitness, genome size, gene number...), backup files (to resume a simulation) and tree files. Tree files store the "replication reports" that log all replication and mutational events. Hence, by analyzing trees, one can precisely reconstruct the events that went to fixation along the line of descent of the final population. To that end, `aevol_post_lineage`, starts from the final population, reads the tree files backward-in-time to reconstruct the line of descent and outputs the corresponding replication reports. Finally, the fourth step, `aevol_post_ancestor_stats`, uses these replication reports to compute the statistics of the ancestral lineage and the list of mutational events that went to fixation along this lineage.

Users might be tempted to stop the experiments after the `aevol_run` step. However, the statistics of the best individuals along generations, although representative of the global trend of simulation, must not be confused with the statistics of the ancestral lineage as mutational events carried by the best individual may not get fixed on the long term.

A.2.1 *Basic usage: Starting from a naive individual*

Aevol allows to analyze the effect of various evolutionary parameters (typically mutation rates, mutational biases, population size...) on genomes by comparing simulations under various scenarios (see Table A.1 for a list of the main testable parameters). Once the parameter values have been chosen, the basic usage of Aevol consist in testing the effect of these parameters directly, starting from "naive" individuals.

In this case, `aevol_create` generates random sequences of a predefined length (typically 5,000 bp) until it finds a genome that has a better fitness than that of a gene-less genome. This approach enables to study evolution when starting far from the fitness optimum. How-

ever, in that case the evolutionary dynamics is strongly dominated by genes recruitment, with massive genome size variation as shown *e.g.* by figure 3 (main text), hence putting the emphasis on a very specific evolutionary dynamics. If one wishes to study more subtle effects, this basic usage is not appropriate and one can turn to a more advanced experimental design based on "Wild-Typing".

A.2.2 *Advanced usage: Wild-Typing*

Once populations have evolved for a sufficiently long time (from a few hundred thousand generations up to millions of generations depending on the parameters, see `https://www.aevol.fr/doc/user-doc/` for more details) under stable evolutionary conditions, individuals own a stable set of genes and are well adapted to their environments. "Wild-Typing" then consists in extracting one or more individuals in the coalescent lineage of the final population, and use these individuals as "Wild-Types" to initiate new evolution experiments, where one can change one or more of the parameters.

Wild-Typing allows studying the response of a well-adapted organism to different types of perturbations, and thus to analyze evolutionary trajectories of more biologically realistic scenarios (Batut et al., 2013).

A.2.3 *Post-evolution analyzes*

Once the simulations are complete, the general characteristics of the ancestors are available (genome size, gene number, coding proportion, etc.), as well as the list of all fixed mutations with their types, loci, and effects on fitness. Now, the ultimate objective is to decipher the relative role of the different evolutionary forces (direct and indirect selection, drift, and the different mutational events – local events, balanced and unbalanced chromosomal rearrangements) on the observed evolutionary dynamics.

Aevol provides several tools to help the user analyze the individuals along the line of descent by estimating their robustness, evolvability and distribution of the fitness effect (DFE) for all types of mutation. To this end, it generates large numbers of independent offspring and, by analyzing the fitness of this offspring, computes the robustness and the evolvability of the ancestors. Similarly, Aevol can generate and analyze single-mutant offspring to estimate the DFE and the mutational robustness for any type of mutation.

To illustrate this, Figure A.2 depicts the distribution of selection coefficients for a large number of mutations on the WT (median CRLM run after the initial evolution) used in the paper.

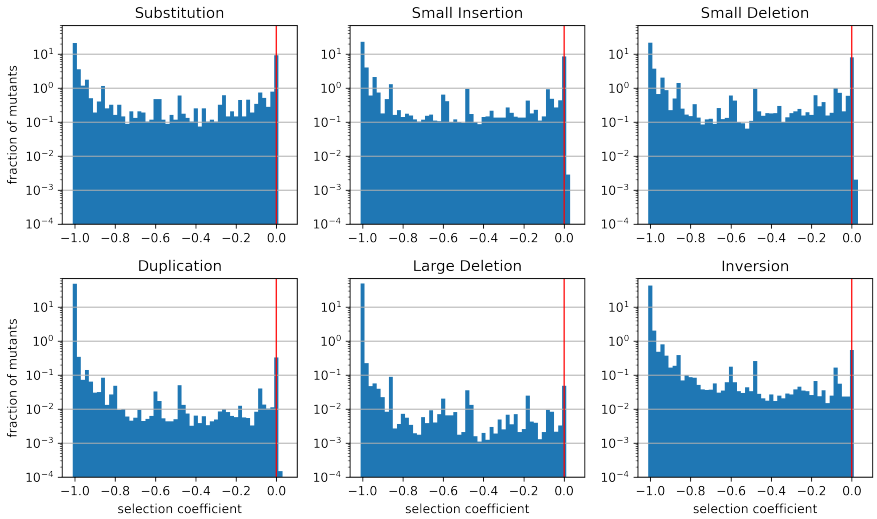

Figure A.2: Distribution of selection coefficients of the different mutation type, on the median individual of our CRLM experiment, after $1,000,000$ generation when starting from a naive individual. For each mutation type, $1,000,000$ mutants were generated, except for the substitution, which were exhaustively tested. The selection coefficient is computed as $s = \frac{f_{mutant}}{f_{parent}} - 1$. The vertical red line indicates neutrality.

A.3 SOFTWARE PARAMETERS

Section	Parameter	Usual range	Description
Genotype to Phenotype to Fitness map	Maximal pleiotropy (W_{max}) `MAX_TRIANGLE_WIDTH`	0.01 – 1 (default: 0.033333)	Largest range of phenotypic values a single protein can impact. Regulates the mean pleiotropy degree and impacts the maximal phenotypic contribution of a single gene (Knibbe et al., 2007a)
	Target function `ENV_ADD_GAUSSIAN`	Sum of 1 to 3 Gaussian functions	The target function is a linear combination of n Gaussian function G_i, each with a weight $\mathcal{H}_i$, a mean μ_i and a standard deviation σ_i: $Target = \sum_{i<n} \frac{\mathcal{H}_i}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2\right)$
	Length of a randomly generated genome `CHROMOSOME_INITIAL_LENGTH`	5000	Initial size of the chromosome when starting from a naive individual (see Section A.2.1)
Population / Selection	population size (N) `INIT_POP_SIZE`	256 – 4096	Census population size. Correlated with the effective population size N_e hence influencing the efficiency of the selection
	Grid size `WORLD_SIZE`	16x16 – 64x64	Shape of the grid. The grid shape influences the speed at which an individual can invade the population (Misevic et al., 2015)
	Selection neighborhood `SELECTION_SCOPE`	Local 3x3 – Global	Type of selection (local or global), and, in the local case, shape of the window used for competition. Local selection slows down the spreading of favorable mutants and increases the effective population size (Zhang, 2003)
	Intensity of the selection (k) `SELECTION_SCHEME`	fitness proportionate 250 – 2500	The selection strength influences the genome size of individuals by increasing/decreasing the indirect selection for robustness (Batut et al., 2013). Note that $k = 0$ suppresses the selection
Replication process	`POINT_MUTATION_RATE`	10^{-4} – 10^{-7}	Per base mutation rates for each kind of mutation. Changes in mutation rates have been shown to impact both the genome length and the genome structure (Knibbe et al., 2007a; Rutten et al., 2019)
	`SMALL_INSERTION_RATE`		
	`SMALL_DELETION_RATE`		
	`DUPLICATION_RATE`		
	`DELETION_RATE`		
	`INVERSION_RATE`		
	`TRANSLOCATION_RATE`		
	`MAX_INDEL_SIZE`	6	Maximal size of small insertion or deletion

Table A.1: Main parameters of the Aevol model.

```
##################################
#         AEVOL PARAMETERS       #
##################################

##### 0. Initial setup ############
STRAIN_NAME             Mol_Ecol_WT4_CRLM
SEED                    4575654216
INIT_METHOD             ONE_GOOD_GENE CLONE
CHROMOSOME_INITIAL_LENGTH   5000

### 1. Genotype-to-Fitness map ####
# Target function (H, mu, sigma)
ENV_ADD_GAUSSIAN        1.2   0.52  0.12
ENV_ADD_GAUSSIAN       -1.4   0.5   0.07
ENV_ADD_GAUSSIAN        0.3   0.8   0.03
# W_Max
MAX_TRIANGLE_WIDTH      0.033333333

### 2. Population and selection ###
INIT_POP_SIZE           1024
WORLD_SIZE              32 32
SELECTION_SCOPE         local 3 3
SELECTION_SCHEME        fitness_proportionate   1000

##### 3. Mutation rates ###########
# Local events
POINT_MUTATION_RATE     5e-6
SMALL_INSERTION_RATE    5e-6
SMALL_DELETION_RATE     5e-6
# Balanced chromosomal rearrangements
INVERSION_RATE          5e-6
TRANSLOCATION_RATE      0
# Unbalanced chromosomal rearrangements
DUPLICATION_RATE        5e-6
DELETION_RATE           5e-6

#### 4. Recording #################
BACKUP_STEP             100000
RECORD_TREE             true
TREE_STEP               1000
```

Figure A.3: Parameter file used for an example simulation (CRLM scenario).

A.4 ADDITIONAL RESULTS

We also tested a scenario "CRLMx2", where all individual mutation rates are doubled with respect to the CRLM scenario (1×10^{-5} per bp), and thus the individual rates are the same as in the LM and CR scenarios (see Table 1, main text). Notably, increasing the mutation rates only further favors the CRLM case, with CRLMx2 being the scenarios with the quickest fitness improvement, and the smallest genome.

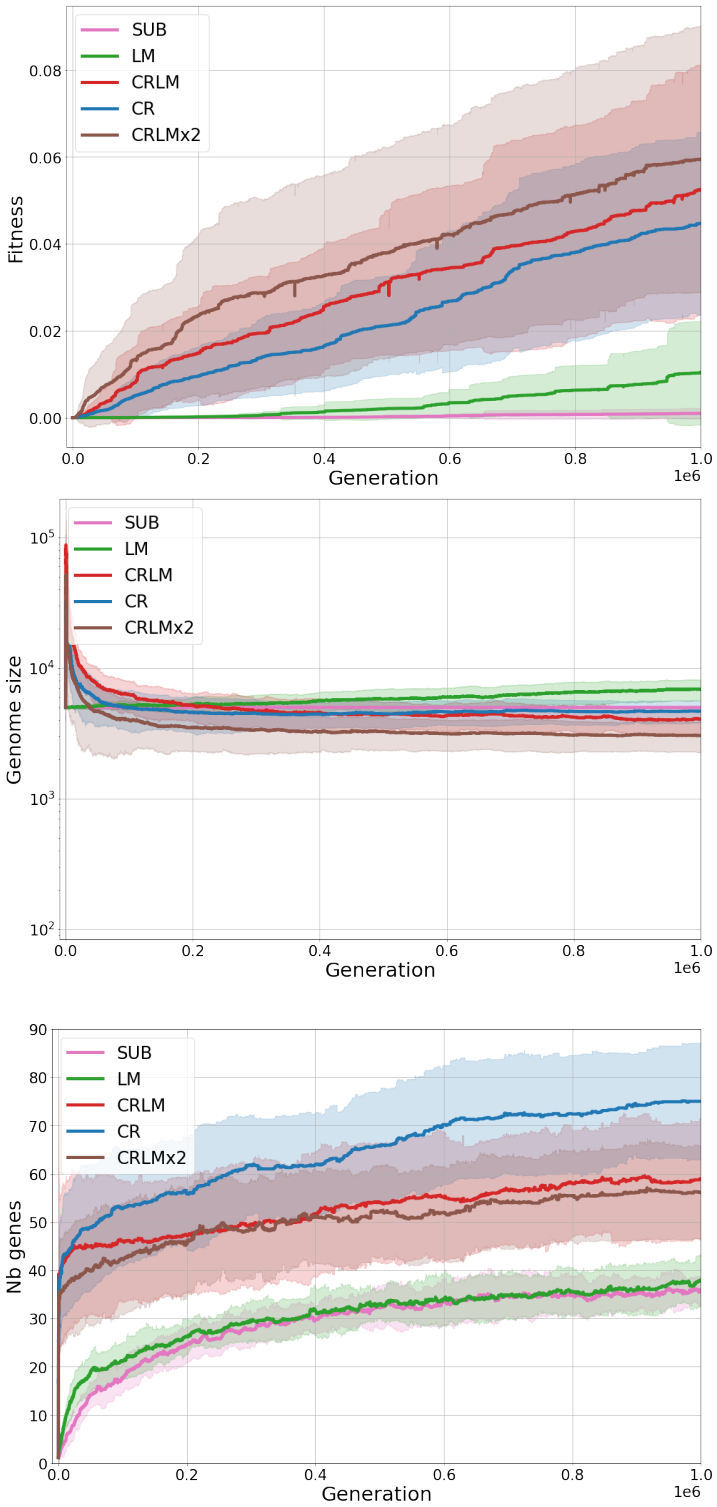

Figure A.4: Variation of fitness, genome size and gene number on the line of descent of the final population, starting from a naive individual for the four mutational scenarios, as well as an additional scenario with doubled mutation rates, and all mutation types present. The shaded areas indicate the variability across the 11 repetitions (standard deviation).

B

SUPPLEMENTARY MATERIALS FOR *GENOME STREAMLINING: EFFECT OF MUTATION RATE AND POPULATION SIZE ON GENOME SIZE REDUCTION*

B.1 EFFECTIVE POPULATION SIZE IN A MODEL WITH LOCAL COMPETITION.

Based on the work of Zhang et al. 2014, we computed the theoretical effective population size in Aevol, when the reproduction is limited to the direct neighborhood of the organism (9 possibilities). It appears that, approximately, $Ne \propto N \log(N)$. The difference is however relatively small for low N values.

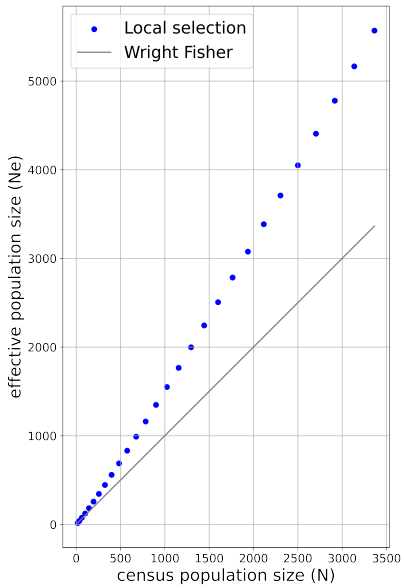

Figure B.1: N_e as a function of N for a Wright-Fischer model and for a model with local reproduction. This supposes the absence of selective sweep.

Note that, to divide the effective population size by 16 when starting from $N = 1024$, one should use $N = 81$ instead of $N = 64$. Some simulations tested with $N = 81$ and $\mu = 1.6 \times 10^{-6}$ do show that in these conditions the coding percentage goes closer to the initial value of 0.68% (*data not shown*).

Reference

Zhang, Y., Tan, Z. and Krishnamachari, B. (2014). On the Meeting Time for Two Random Walks on a Regular Graph. *arXiv:1408.2005*.

B.2 RESULTS WITH A MUTATIONAL BIAS IN THE INDELS.

To test the interaction between mutational biases and changes in population size or mutation rate, we evolved 5 Wild-Types with an insertion bias in the InDels distribution (twice more insertions than deletions), or a deletion bias (twice more deletions than insertions), keeping the total mutation rate constant. Similarly to what is described in the main text for the bias in the rearrangement rates (duplications and large deletions), the genome sizes of our Wild-Types are different from the one without a mutational bias: 28,941 and a coding fraction of 0.37 with the insertion bias *VS* 8,925 and 0.98 with a deletion bias.

When submitted to a change in population size or mutation rate, the median WT of both these experiments react similarly to what is predicted by our model (see Figure B.2) — although the loss of non-coding bases is limited in the deletion-bias experiment since the coding proportion is already close to 1.

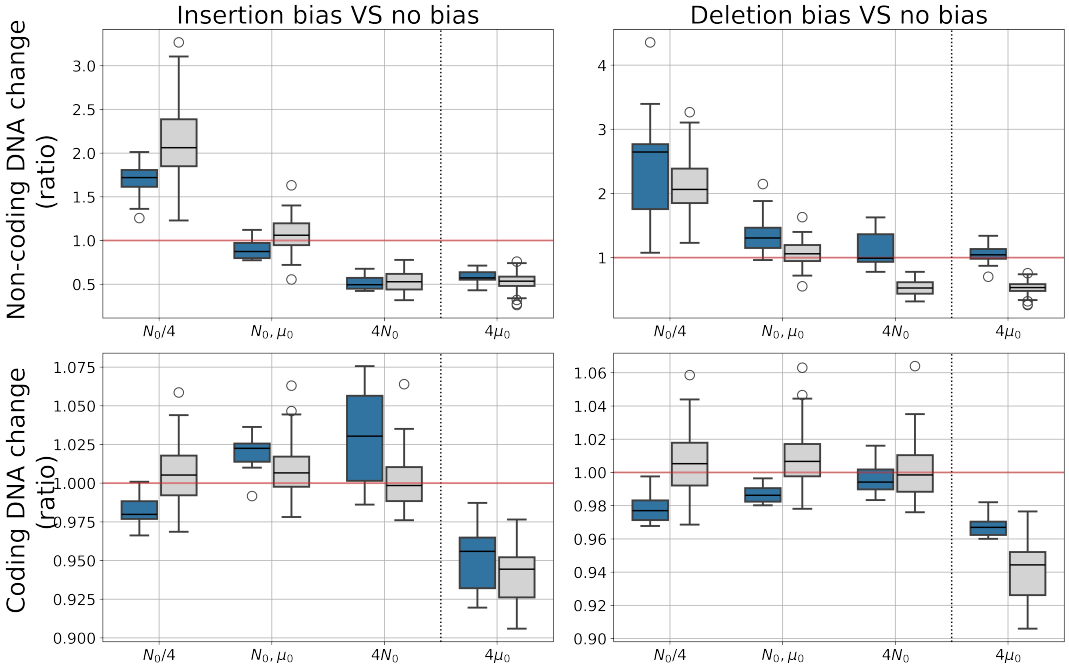

Figure B.2: Change in coding and non-coding genome sizes in reaction to changes in N or μ for the different mutational biases. Blue boxes show the results with a mutational bias (left: insertion bias in InDels, right: deletion bias in InDels), and gray boxes show the results without mutational bias. Depicted values are the ratio of the coding/non-coding size at the final generation over the value at generation 0.

B.3 TEMPORAL DATA FOR ALL TESTED CONDITIONS.

For each of the conditions tested (see Table 1 of the main document, section 4.2.2), we provide the temporal data of fitness, total amount of DNA, coding and non-coding sizes as well as the coding fraction for all 50 repetitions. Curves are colored by WT used to start the simulation (blue: WT1, orange: WT2, green: WT3, purple: WT4, brown: WT5). Individual simulations are depicted as shallow lines, and the thick curves show the average value for the Wild-Type.

B.3.1 *Control:* $\mu = 10^{-6} = \mu_0, \quad N = 1024 = N_0, \quad N \times \mu = N_0 \times \mu_0$

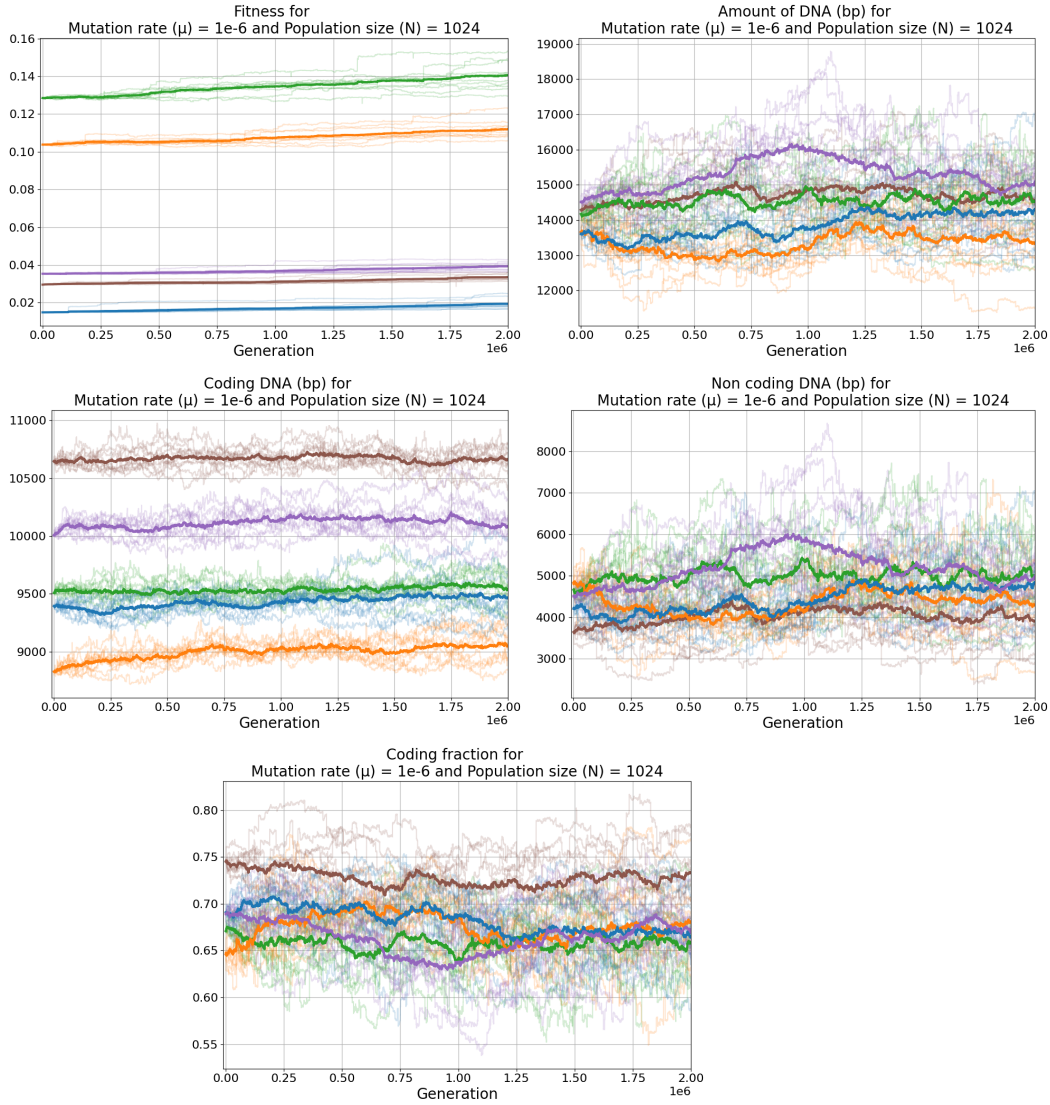

Figure B.3: Temporal data for $\mu = 10^{-6}$, $N = 1024$. From left to right, top to bottom : Fitness, total amount of DNA, coding genome size, non-coding genome size and coding fraction.

B.3.2 $\mu = 10^{-6} = \mu_0, \quad N = 64 = N_0/16, \quad N \times \mu = 1/16 N_0 \times \mu_0$

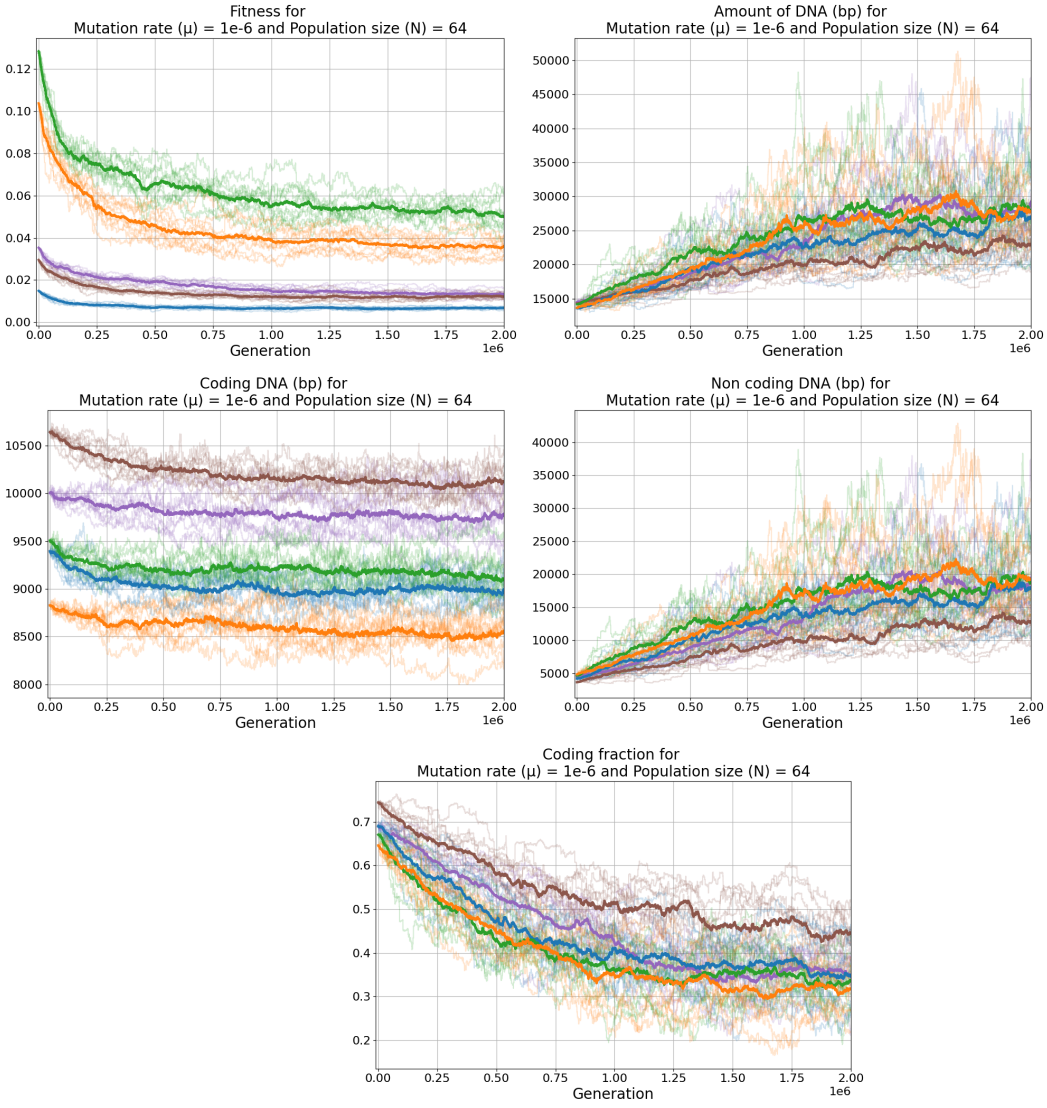

Figure B.4: Temporal data for $\mu = 10^{-6}$, $N = 64$. From left to right, top to bottom : Fitness, total amount of DNA, coding genome size, non-coding genome size and coding fraction.

B.3.3 $\mu = \mu_0 = 10^{-6}$, $N = N_0/4 = 256$, $N \times \mu = 1/4 N_0 \times \mu_0$

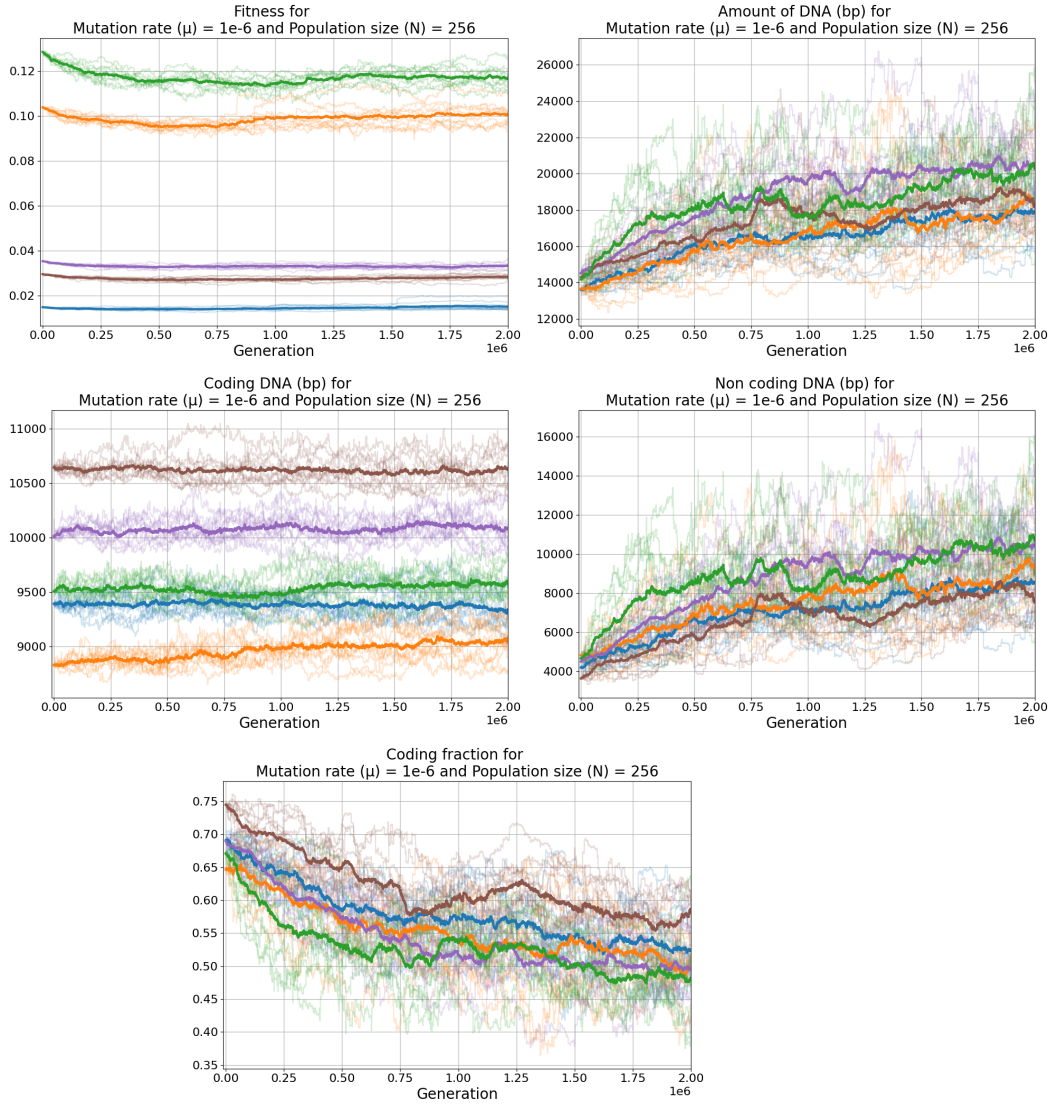

Figure B.5: Temporal data for $\mu = 10^{-6}$, $N = 256$. From left to right, top to bottom : Fitness, total amount of DNA, coding genome size, non-coding genome size and coding fraction.

B.3.4 $\mu = 10^{-6} = \mu_0, \quad N = 4096 = 4N_0, \quad N \times \mu = 4N_0 \times \mu_0$

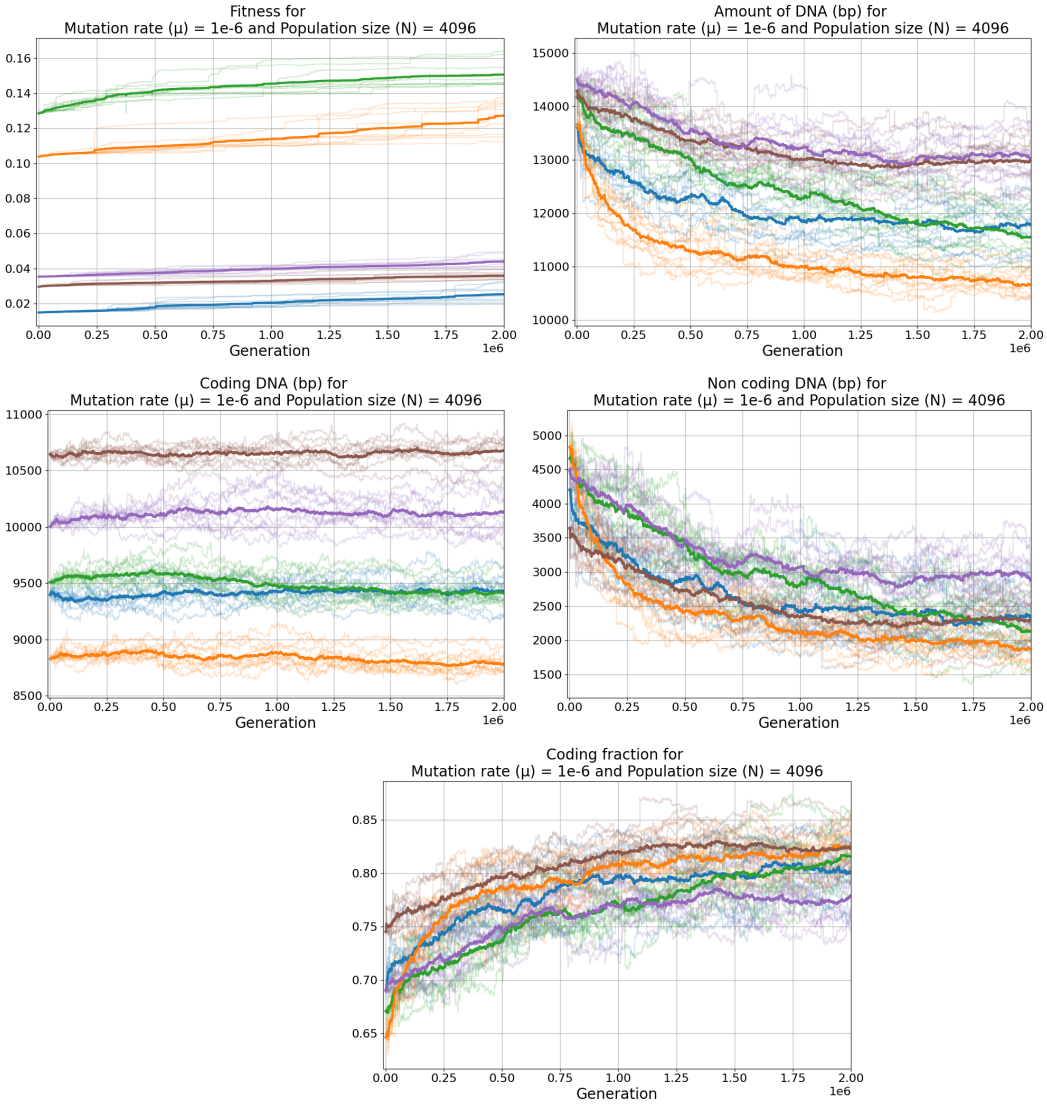

Figure B.6: Temporal data for $\mu = 10^{-6}$, $N = 4096$. From left to right, top to bottom : Fitness, total amount of DNA, coding genome size, non-coding genome size and coding fraction.

B.3.5 $\mu = \mu_0 = 10^{-6}, \quad N = 16N_0 = 16384, \quad N \times \mu = 16N_0 \times \mu_0$

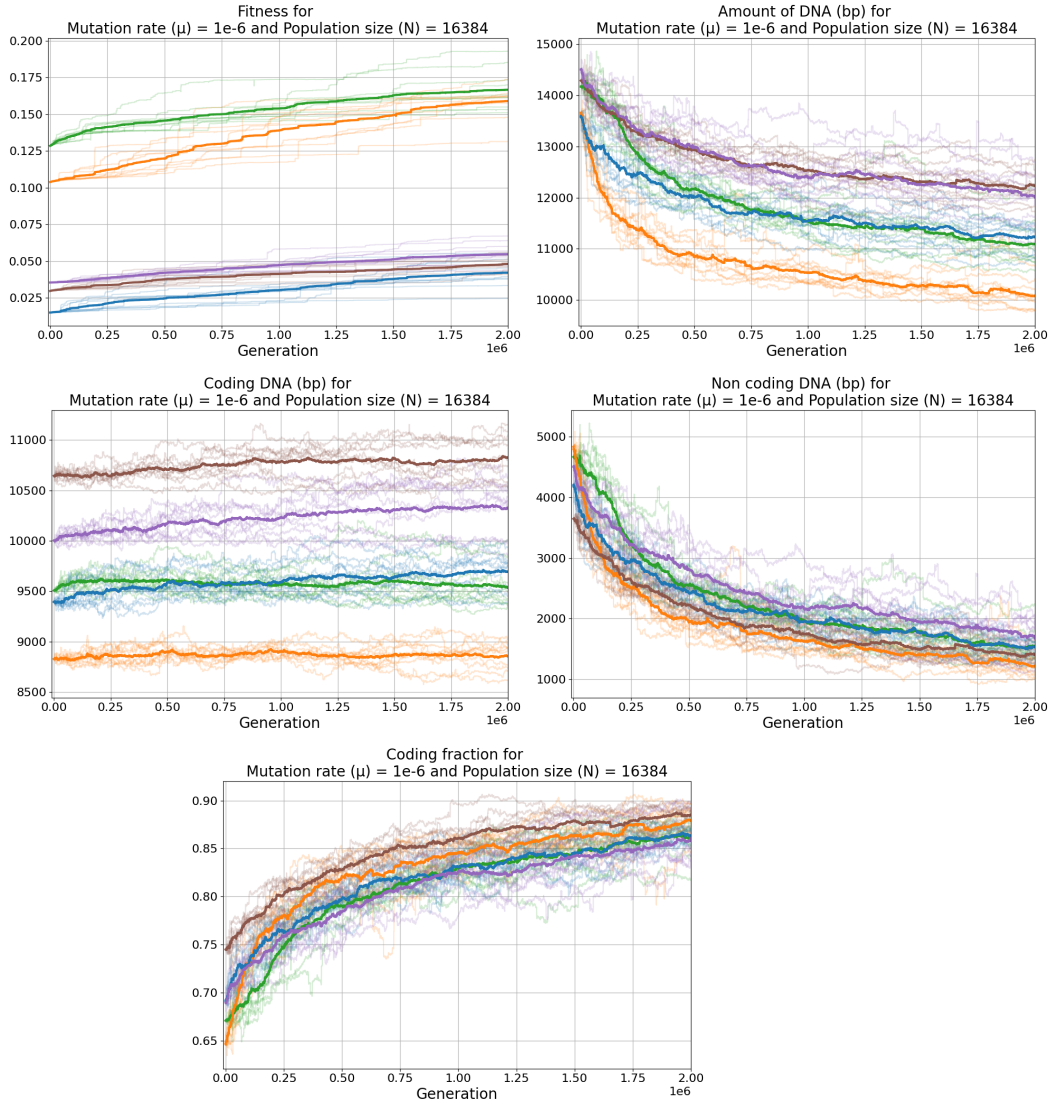

Figure B.7: Temporal data for $\mu = 10^{-6}$, $N = 16384$. From left to right, top to bottom : Fitness, total amount of DNA, coding genome size, non-coding genome size and coding fraction.

B.3.6 $\mu = 2 \times 10^{-6} = 2\mu_0, \quad N = 529 \approx N_0/2, \quad N \times \mu \approx N_0 \times \mu_0$

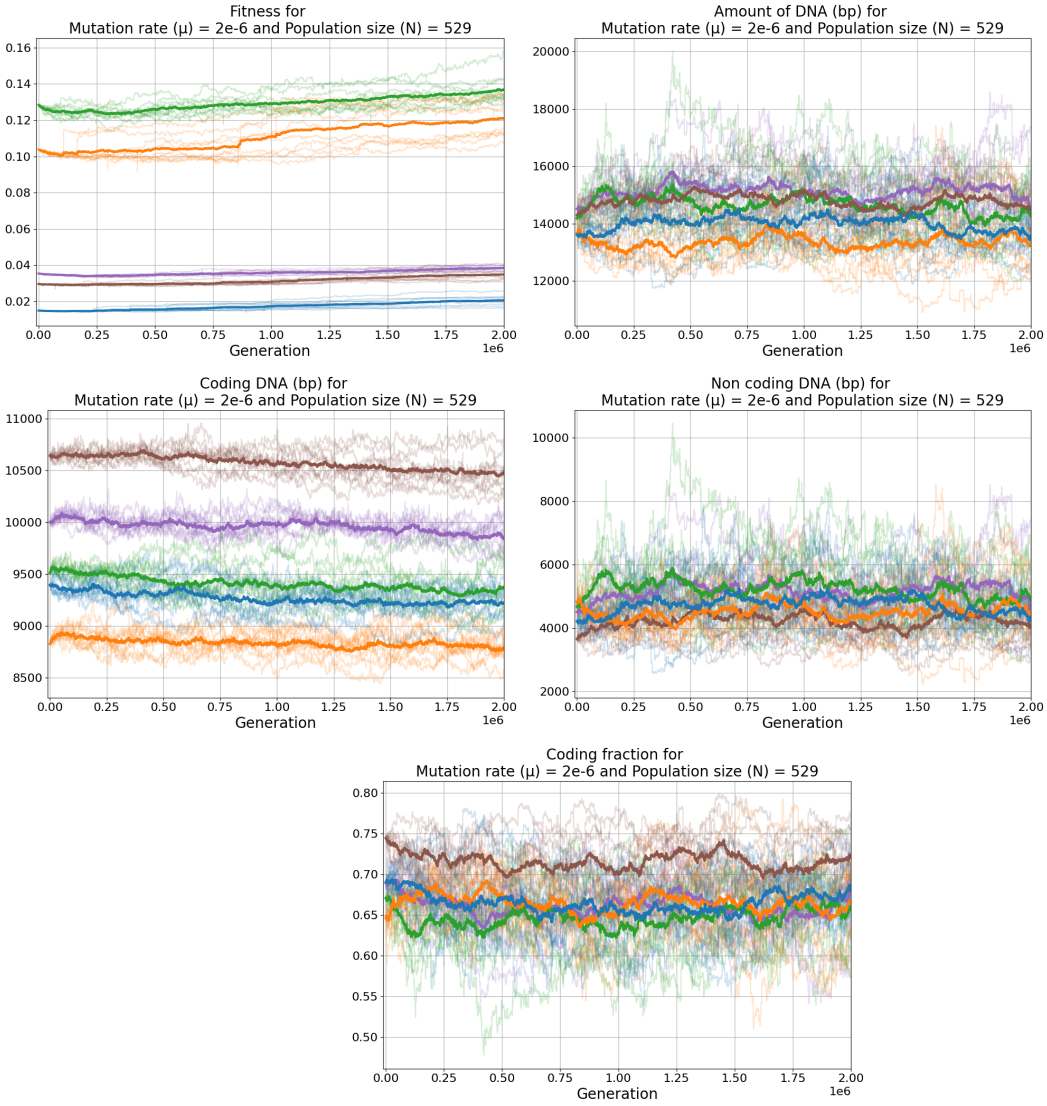

Figure B.8: Temporal data for $\mu = 2 \times 10^{-6}$, $N = 529$. From left to right, top to bottom : Fitness, total amount of DNA, coding genome size, non-coding genome size and coding fraction.

B.3.7 $\mu = 2 \times 10^{-6} = 2\mu_0, \quad N = 2025 \approx 2N_0, \quad N \times \mu \approx 4N_0 \times \mu_0$

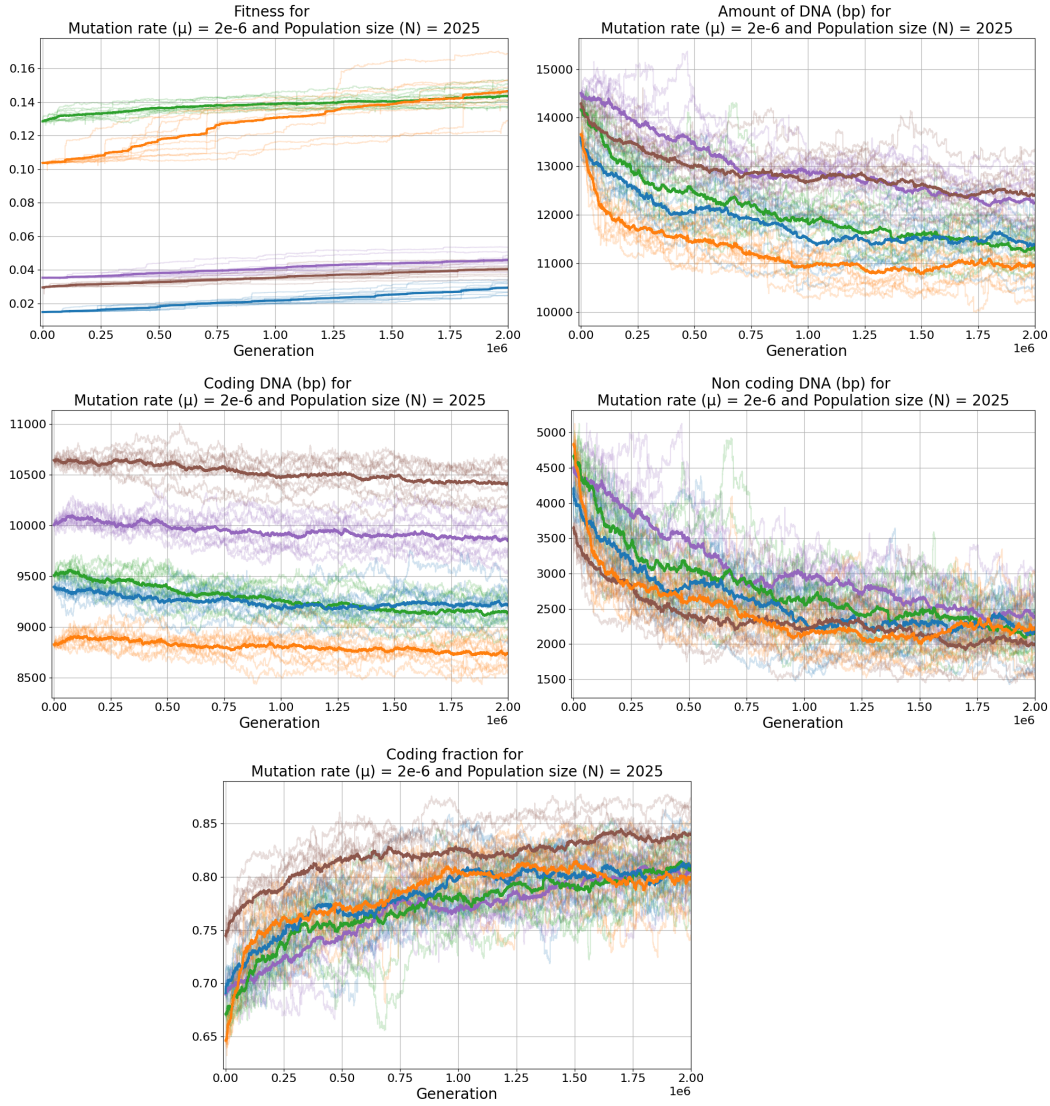

Figure B.9: Temporal data for $\mu = 2 \times 10^{-6}$, $N = 2025$. From left to right, top to bottom : Fitness, total amount of DNA, coding genome size, non-coding genome size and coding fraction.

B.3.8 $\mu = 4 \times 10^{-6} = 4\mu_0, \quad N = 256 = N_0/4, \quad N \times \mu = N_0 \times \mu_0$

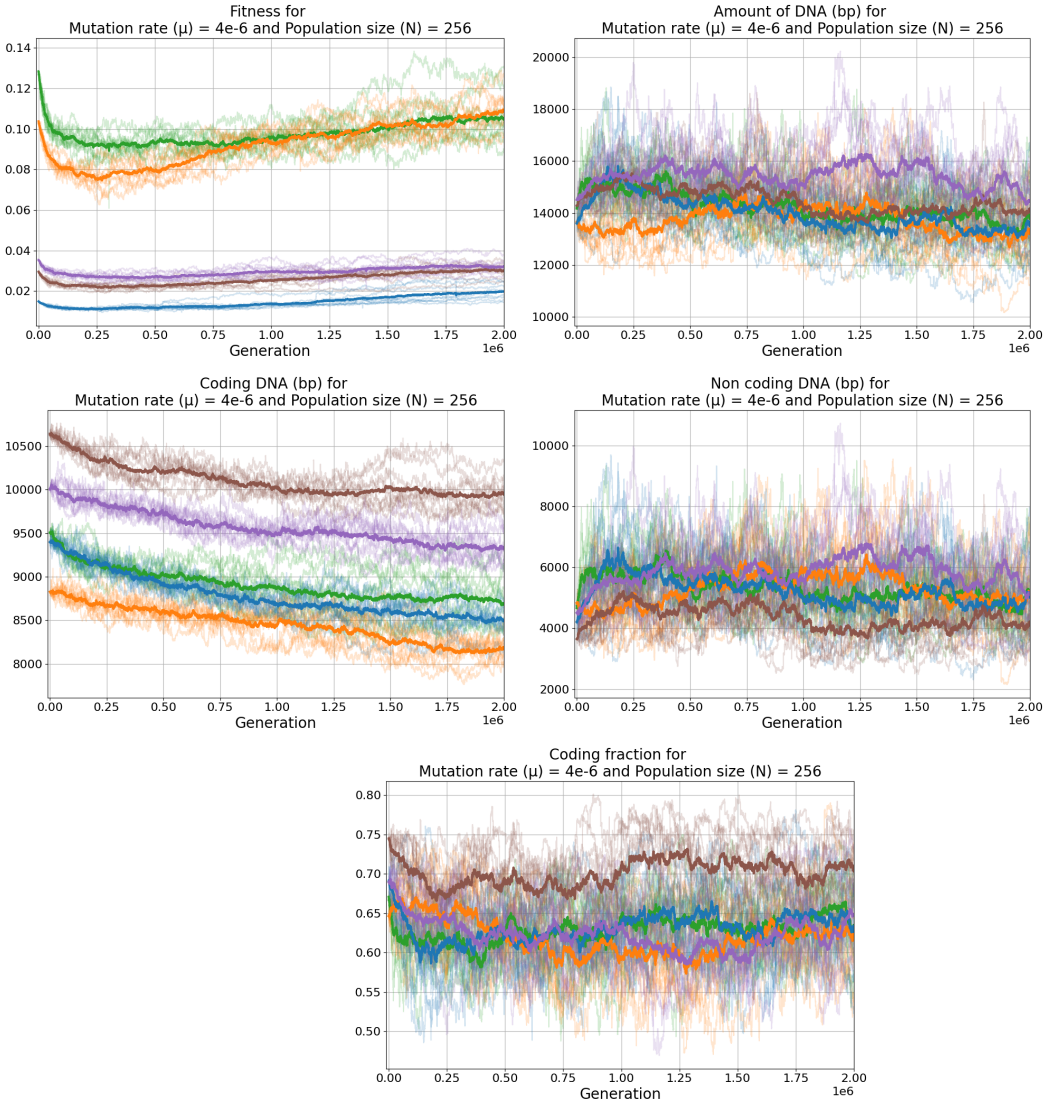

Figure B.10: Temporal data for $\mu = 4 \times 10^{-6}$, $N = 256$. From left to right, top to bottom : Fitness, total amount of DNA, coding genome size, non-coding genome size and coding fraction.

B.3.9 $\mu = 4 \times 10^{-6} = 4\mu_0, \quad N = 1024 = N_0, \quad N \times \mu = 4N_0 \times \mu_0$

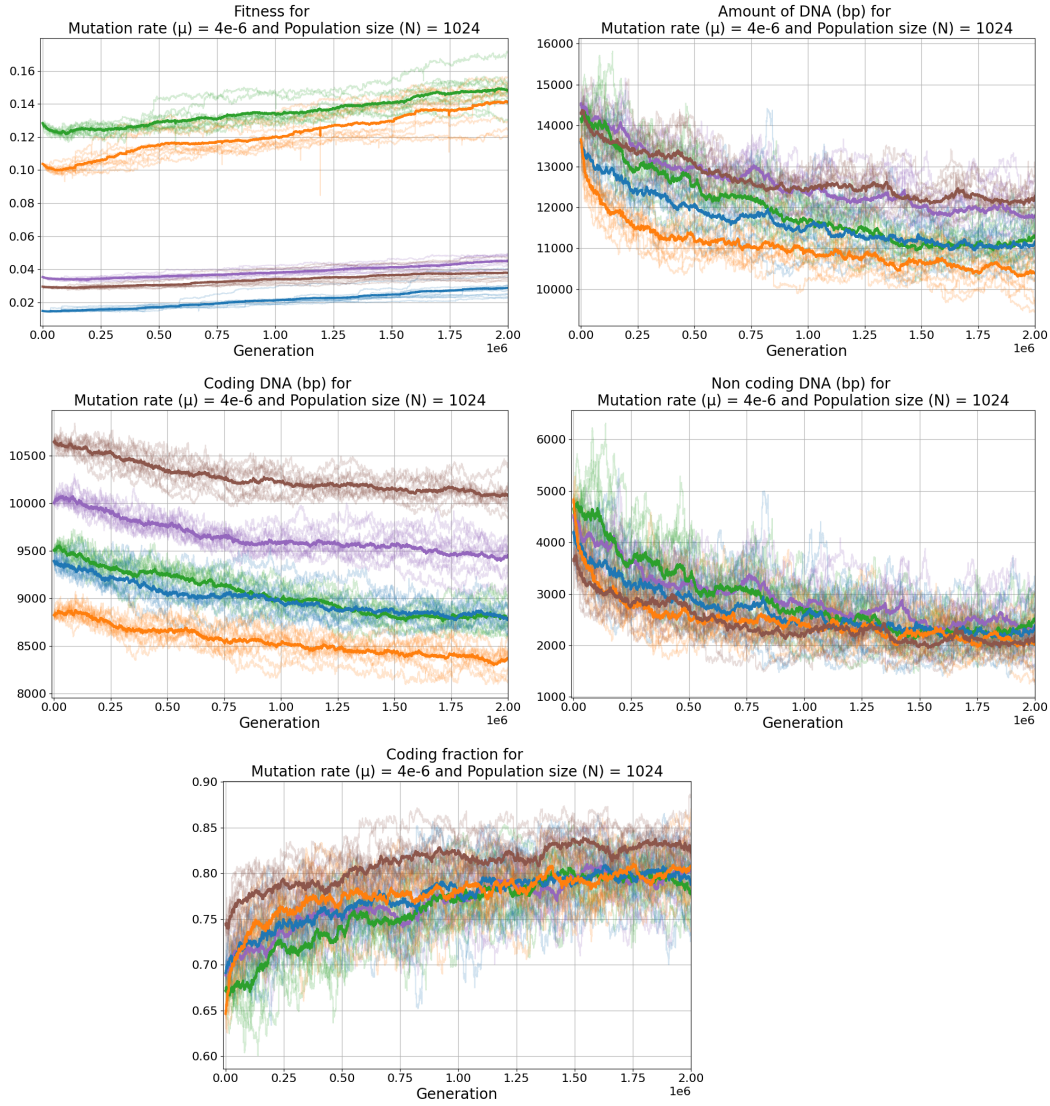

Figure B.11: Temporal data for $\mu = 4 \times 10^{-6}$, $N = 1024$. From left to right, top to bottom : Fitness, total amount of DNA, coding genome size, non-coding genome size and coding fraction.

B.3.10 $\mu = 4 \times 10^{-6} = 4\mu_0, \quad N = 4096 = 4N_0, \quad N \times \mu = 16N_0 \times \mu_0$

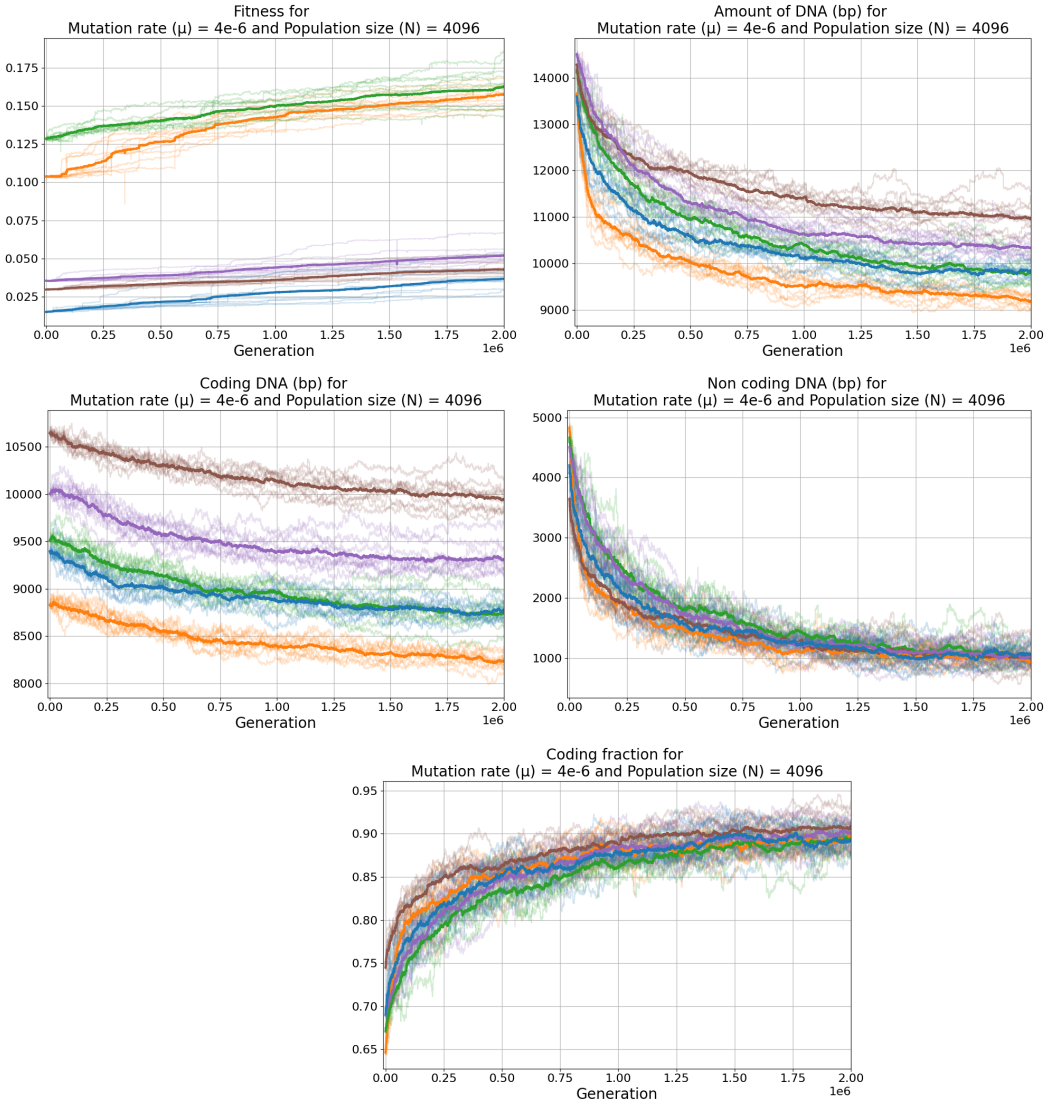

Figure B.12: Temporal data for $\mu = 4 \times 10^{-6}$, $N = 4096$. From left to right, top to bottom : Fitness, total amount of DNA, coding genome size, non-coding genome size and coding fraction.

B.3.11 $\mu = 1.6 \times 10^{-5} = 16\mu_0, \quad N = 64 = N_0/16, \quad N \times \mu = N_0 \times \mu_0$

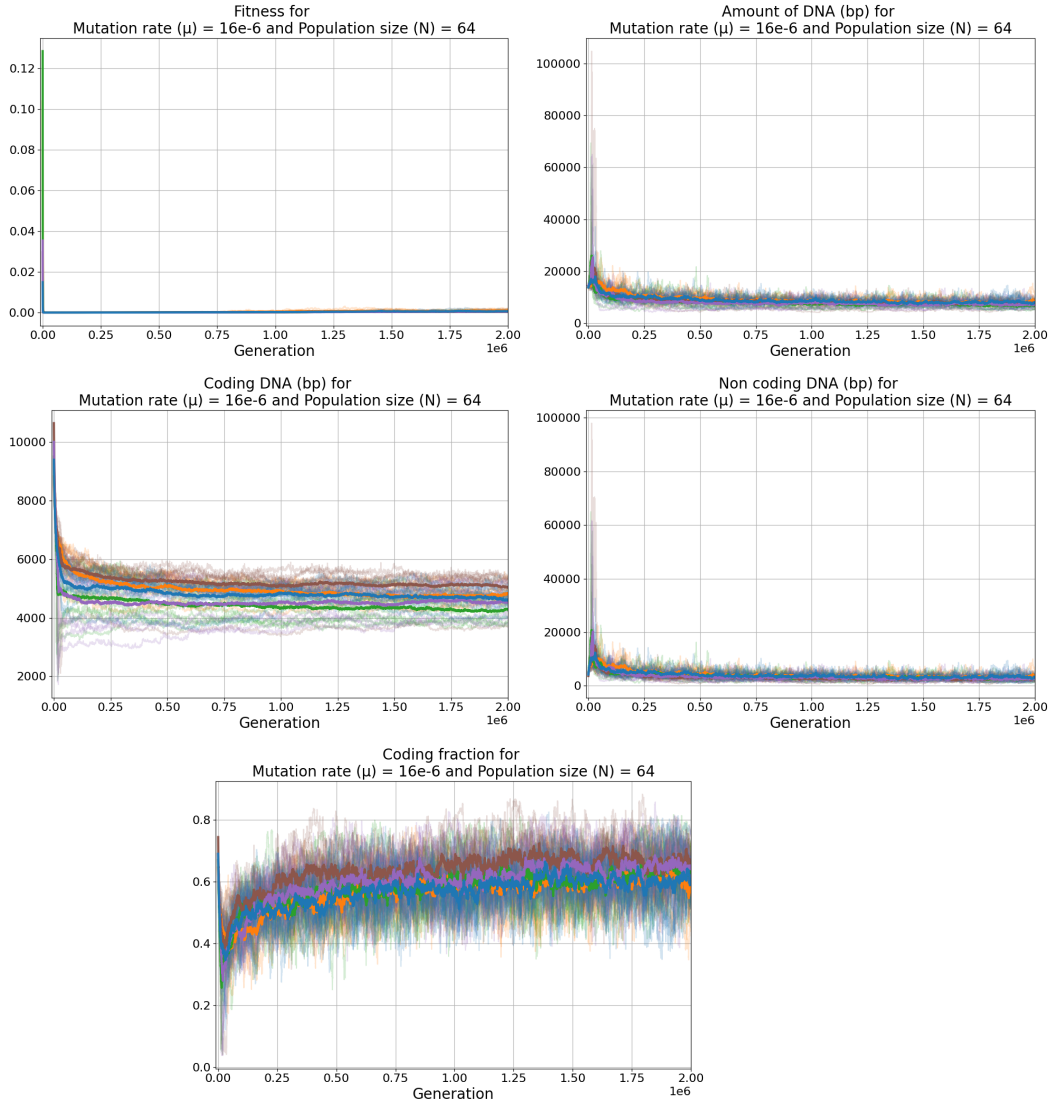

Figure B.13: Temporal data for $\mu = 1.6 \times 10^{-5}$, $N = 64$. From left to right, top to bottom : Fitness, total amount of DNA, coding genome size, non-coding genome size and coding fraction.

B.3.12 $\mu = 1.6 \times 10^{-5} = 16\mu_0, \quad N = 1024 = N_0, \quad N \times \mu = 16N_0 \times \mu_0$

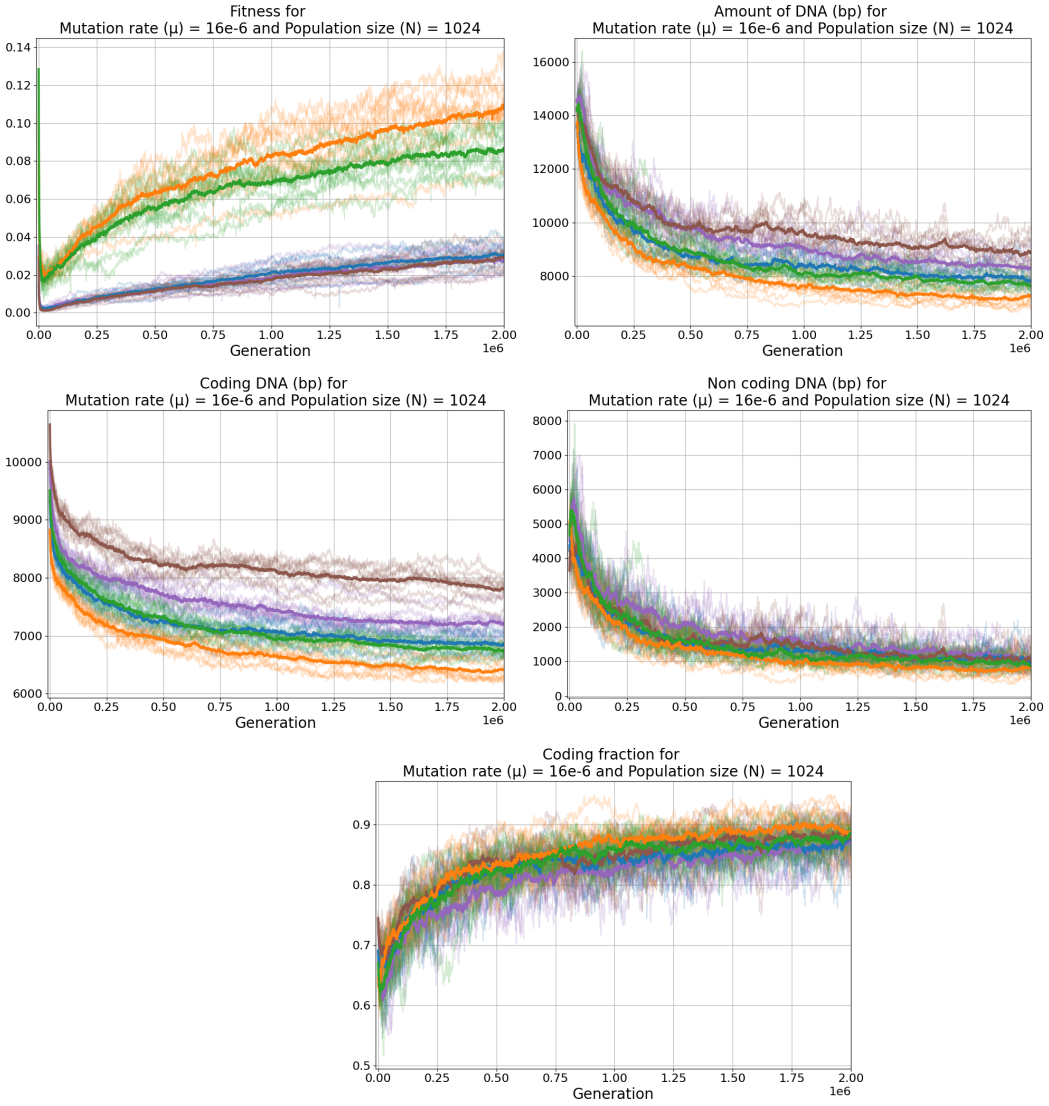

Figure B.14: Temporal data for $\mu = 1.6 \times 10^{-5}$, $N = 1024$. From left to right, top to bottom : Fitness, total amount of DNA, coding genome size, non-coding genome size and coding fraction.

B.3.13 $\mu = 1.6 \times 10^{-5} = 16\mu_0$, $N = 16,384 = 16N_0$, $N \times \mu = 256N_0 \times \mu_0$

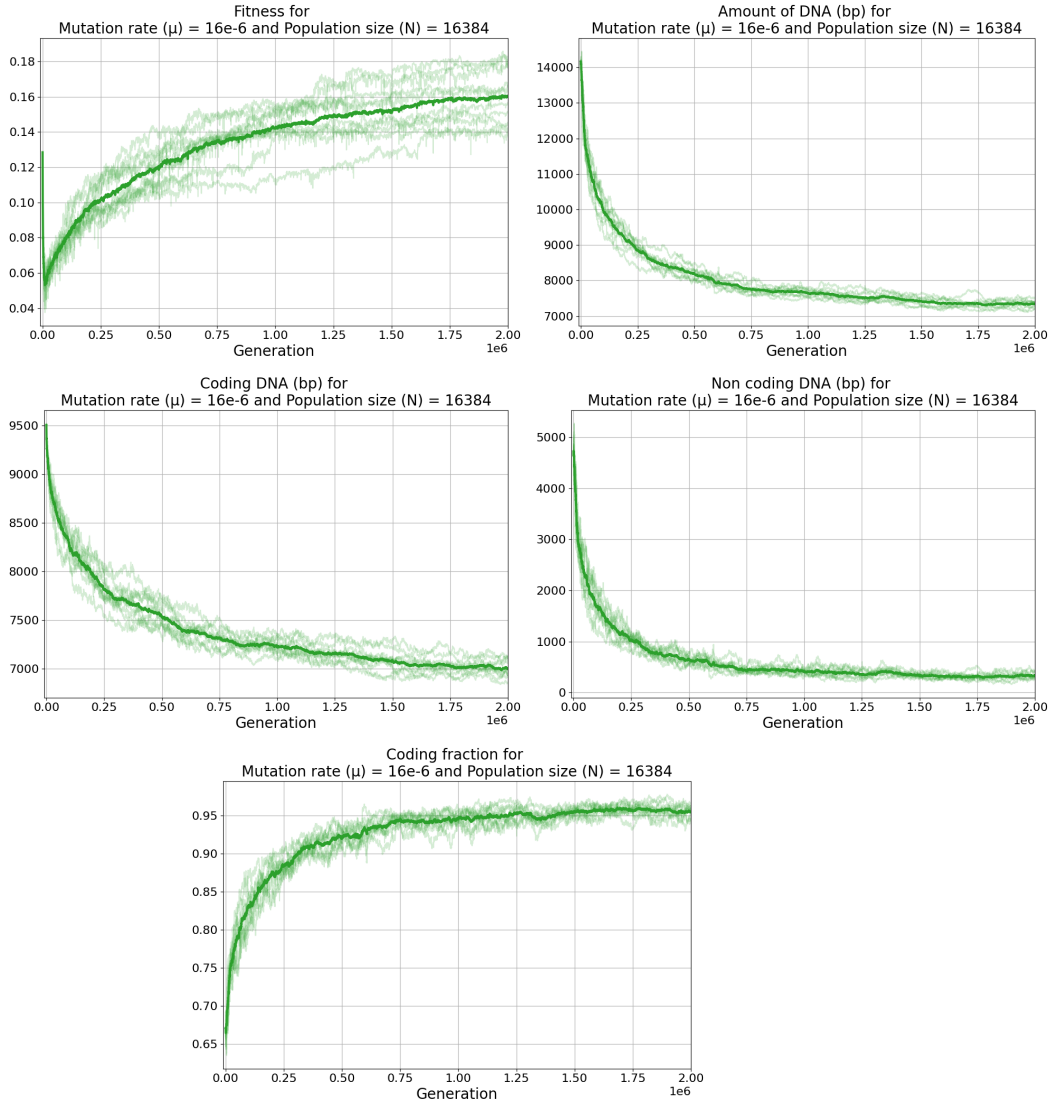

Figure B.15: Temporal data for $\mu = 1.6 \times 10^{-5}$, $N = 16,384$. From left to right, top to bottom : Fitness, total amount of DNA, coding genome size, non-coding genome size and coding fraction. To limit the computational load, only wild type 3 (WT3) was tested for this parameter set.

C

SUPPLEMENTARY MATERIAL FOR *STRUCTURAL MUTATIONS SET AN EQUILIBRIUM NON-CODING GENOME FRACTION*

C.1 PROBABILITY FOR A MUTATION TO BE NEUTRAL

For each type of mutation, we note ν_{mutation} its probability to be perfectly neutral regarding the viability of the individual. We note p_1 the first position uniformly drawn on the genome, and p_2 and p_3 the second and third when needed. As such, each base has a probability $\frac{1}{L}$ to be drawn. There are z_{nc} non-coding bases, distributed along g non-coding segments. The computations are provided for 6 types of mutations: deletions and duplications as well as point mutations, small insertions, small deletions, and inversions.

C.1.1 *Probability for deletions to be neutral*

A deletion is neutral if, and only if, the bases deleted are within one of the g non-coding segments. This means that if the first deleted base is at a position i, i must be in the non-coding part of the genome, and the second must be at a position j in the same non-coding region.

$$\begin{aligned}\nu_{\text{del}}(g, z_{\text{c}}, z_{\text{nc}}) &= g \sum_{i=1}^{z_{\text{nc}}/g} \left(\frac{1}{L} \sum_{j=i}^{z_{\text{nc}}/g} \frac{1}{L} \right) \\ &= \frac{g}{2L^2} \sum_{i=1}^{z_{\text{nc}}/g} \left(\frac{z_{\text{nc}}}{g} - i + 1 \right) \\ &= \frac{z_{\text{nc}} \left(\frac{z_{\text{nc}}}{g} + 1 \right)}{2L^2}\end{aligned}$$

C.1.2 *Probability for duplications to be neutral*

A duplication is neutral if, and only if, it duplicates a sequence without a promoter basis and copies it at any position in the non-coding regions. The sum over i starts at position 2 to avoid the first base (promoter), and then all duplications are valid as long as they do not encompass the next promoter. This probability is then multiplied by the probability for the insertion point to be in a non-coding region. Note that there are $z_{\text{nc}}/g + 1$ insertion points in a non-coding sequence of size z_{nc}/g as we can insert just before and just after the sequence.

$$\begin{aligned}
\nu_{\text{dupl}}(g, z_c, z_{nc}) &= g \sum_{i=2}^{L/g} \left(\frac{1}{L} \sum_{j=i}^{L/g} \frac{1}{L} \right) \left(g \sum_{k=0}^{z_{nc}/g} \frac{1}{L} \right) \\
&= \frac{g^2}{L^3} \sum_{i=2}^{L/g} \sum_{j=i}^{L/g} \left(\frac{z_{nc}}{g} + 1 \right) \\
&= \frac{g(z_{nc}+g)}{L^3} \sum_{i=2}^{L/g} \left(\frac{L}{g} - i + 1 \right) \\
&= \frac{g(z_{nc}+g)}{L^3} \sum_{i=1}^{L/g-1} \left(\frac{L}{g} - i \right) \\
&= \frac{g(z_{nc}+g)\left(\frac{L}{g}-1\right)\left(\frac{L}{g}\right)}{2L^3} \\
&= \frac{(z_{nc}+g)\left(\frac{L}{g}-1\right)}{2L^2}
\end{aligned}$$

C.1.3 *Probability for point mutations to be neutral*

Point mutations are neutral when they affect a non-coding base, and deleterious when they affect a coding base. The probability to affect a non-coding base is $\frac{z_{nc}}{L}$:

$$\begin{aligned}
\nu_{\text{pm}}(g, z_c, z_{nc}) &= g \sum_{i=1}^{z_{nc}/g} \frac{1}{L} \\
&= \frac{z_{nc}}{L}
\end{aligned}$$

C.1.4 *Probability for small insertions to be neutral*

Regardless of their size, small insertions are neutral when outside a coding segment, and deleterious when within a coding segment. Note however that there are $z_{nc}/g + 1$ insertion points in a non-coding sequence of size z_{nc}/g as we can insert just before and just after the sequence.

$$\begin{aligned}
\nu_{\text{indel}^+}(g, z_c, z_{nc}) &= g \sum_{i=0}^{z_{nc}/g} \frac{1}{L} \\
&= \frac{(z_{nc}+g)}{L}
\end{aligned}$$

C.1.5 *Probability for small deletions to be neutral*

The maximum size l_m of indels events is a parameter of the model. Here, we assume that $z_{nc}/g \geq l_m$. The rationale here is to calculate the probability of a neutral deletion by separating all non-coding sequences into the $z_{nc}/g - (l_m - 1)$ first bases that can witness deletions of size up to l_m, and the $l_m - 1$ other bases for which only a subset of the possible deletions are neutral. Since the length of the deletion is uniformly chosen between 1 and l_m, if the deletion starts from a basis i close to the end z_{nc}/g of the non coding zone, the probability that it is neutral is $\sum_{k=i}^{z_{nc}/g} \frac{1}{l_m}$ when $z_{nc}/g - i < l_m$

$$\begin{aligned}
\nu_{\text{indel}^-}(g, z_c, z_{nc}) &= g\left(\sum_{i=1}^{z_{nc}/g-(l_m-1)} \frac{1}{L} + \sum_{i=z_{nc}/g-(l_m-2)}^{z_{nc}/g} \frac{1}{L} \sum_{k=i}^{z_{nc}/g} \frac{1}{l_m}\right) \\
&= \frac{g}{L}\left(\frac{z_{nc}}{g} - (l_m - 1) + \frac{1}{l_m} \sum_{z_{nc}/g-(l_m-2)}^{z_{nc}/g} \frac{z_{nc}}{g} - i + 1\right) \\
&= \frac{1}{L}\left(z_{nc} - g(l_m - 1) + \frac{g}{l_m} \sum_{i=1}^{l_m-1} i\right) \\
&= \frac{1}{L}\left(z_{nc} - g\frac{l_m - 1}{2}\right)
\end{aligned}$$

C.1.6 *Probability for inversions to be neutral*

An inversion is neutral if the two breakpoints are outside coding regions. Note that the second breakpoint must be different from the first for an inversion to occur. The probability of the inversion to be neutral is thus the product of the two probabilities.

$$\begin{aligned}
\nu_{\text{inv}}(g, z_c, z_{nc}) &= \frac{(z_{nc} + g)}{L} \times \frac{(z_{nc} + g) - 1}{L - 1} \\
&= \frac{(z_{nc} + g)(z_{nc} + g - 1)}{L(L - 1)}
\end{aligned}$$

C.2 EXPECTED CONTRIBUTION TO GENOME SIZE CHANGE ALONG EVOLUTION

For each mutation changing the genome size, we can compute its expected contribution per generation to the average genome size change for a species in terms of base pairs.

C.2.1 *Contribution of deletions to genome size change*

This corresponds to the average size of a deletion weighted by the probability of neutrality times the probability of fixation.

$$\begin{aligned}
\delta_{\text{del}}(\mu, N, g, z_c, z_{nc}) &= g \sum_{i=1}^{z_{nc}/g} \left(\frac{1}{L} \sum_{j=i}^{z_{nc}/g} \frac{1}{L} (j-i+1) \mathbb{P}_{fix}(-(j-i+1)) \right) \\
&= \frac{g}{L^2} \sum_{i=1}^{z_{nc}/g} \sum_{j=i}^{z_{nc}/g} (j-i+1) \mathbb{P}_{fix}(-(j-i+1)) \\
&= \frac{g}{L^2} \sum_{i=1}^{z_{nc}/g} \sum_{k=1}^{z_{nc}/g-i+1} k \mathbb{P}_{fix}(-k) \\
&= \frac{g}{L^2} \sum_{k=1}^{z_{nc}/g} \sum_{i=1}^{z_{nc}/g-k+1} k \mathbb{P}_{fix}(-k) \\
&= \frac{g}{L^2} \sum_{k=1}^{z_{nc}/g} \left(\frac{z_{nc}}{g} - k + 1 \right) k \mathbb{P}_{fix}(-k) \\
&= \frac{1}{L^2} \sum_{k=1}^{z_{nc}/g} (z_{nc} - g(k+1)) k \mathbb{P}_{fix}(-k)
\end{aligned}$$

C.2.2 *Contribution of duplications to genome size change*

Similarly, this corresponds to the average size of a duplication weighted by the probability of neutrality times the probability of fixation.

$$\begin{aligned}
\delta_{\text{dupl}}(\mu, N, g, z_c, z_{nc}) &= g \sum_{i=2}^{L/g} \left(\frac{1}{L} \sum_{j=i}^{L/g} \frac{1}{L} \left(g \sum_{k=0}^{z_{nc}/g} \frac{1}{L} (j-i+1) \mathbb{P}_{fix}(j-i+1) \right) \right) \\
&= \frac{g(z_{nc}+g)}{L^3} \sum_{i=2}^{L/g} \sum_{j=i}^{L/g} (j-i+1) \mathbb{P}_{fix}(j-i+1) \\
&= \frac{g(z_{nc}+g)}{L^3} \sum_{i=2}^{L/g} \sum_{j=1}^{L/g-i+1} j \mathbb{P}_{fix}(j) \\
&= \frac{g(z_{nc}+g)}{L^3} \sum_{j=1}^{L/g-1} \sum_{i=2}^{L/g-j+1} j \mathbb{P}_{fix}(j) \\
&= \frac{g(z_{nc}+g)}{L^3} \sum_{j=1}^{L/g-1} \left(\frac{L}{g} - j \right) j \mathbb{P}_{fix}(j)
\end{aligned}$$

C.2.3 *Contribution of small insertions (InDel$^+$) to genome size change*

This corresponds to the average size of a small insertion weighted by its probability of neutrality times its probability of fixation.

$$\begin{aligned}\delta_{\text{indel}^+}(\mu, N, g, z_c, z_{nc}) &= g \sum_{i=0}^{z_{nc}/g} \frac{1}{L} \sum_{k=1}^{l_m} \frac{k\mathbb{P}_{fix}(k)}{l_m} \\ &= \frac{(z_{nc}+g)}{L\, l_m} \sum_{k=1}^{l_m} k\mathbb{P}_{fix}(k)\end{aligned}$$

C.2.4 *Contribution of small deletions (InDel$^-$) to genome size change*

This corresponds to the average size of a small deletion weighted by its probability of neutrality times its probability of fixation.

$$\begin{aligned}&\delta_{\text{indel}^-}(\mu, N, g, z_c, z_{nc}) \\ &= g \left(\sum_{i=1}^{z_{nc}/g-(l_m-1)} \left(\frac{1}{L} \sum_{k=1}^{l_m} \frac{k\mathbb{P}_{fix}(-k)}{l_m} \right) + \sum_{i=z_{nc}/g-(l_m-2)}^{z_{nc}/g} \frac{1}{L} \sum_{k=i}^{z_{nc}/g} \frac{(k-i+1)\mathbb{P}_{fix}(-(k-i+1))}{l_m} \right) \\ &= \frac{g}{L\, l_m} \left(\left(\frac{z_{nc}}{g} - (l_m-1) \right) \sum_{k=1}^{l_m} k\mathbb{P}_{fix}(-k) + \sum_{i=z_{nc}/g-(l_m-2)}^{z_{nc}/g} \sum_{j=1}^{z_{nc}/g-i+1} j\mathbb{P}_{fix}(-j) \right) \\ &= \frac{1}{L\, l_m} \left((z_{nc} - g(l_m-1)) \sum_{k=1}^{l_m} k\mathbb{P}_{fix}(-k) + \sum_{s=1}^{l_m-1} \sum_{j=1}^{s} j\mathbb{P}_{fix}(-j) \right)\end{aligned}$$

C.3 JOINED IMPACT OF N AND μ ON NON-CODING GENOME FRACTION AT EQUILIBRIUM

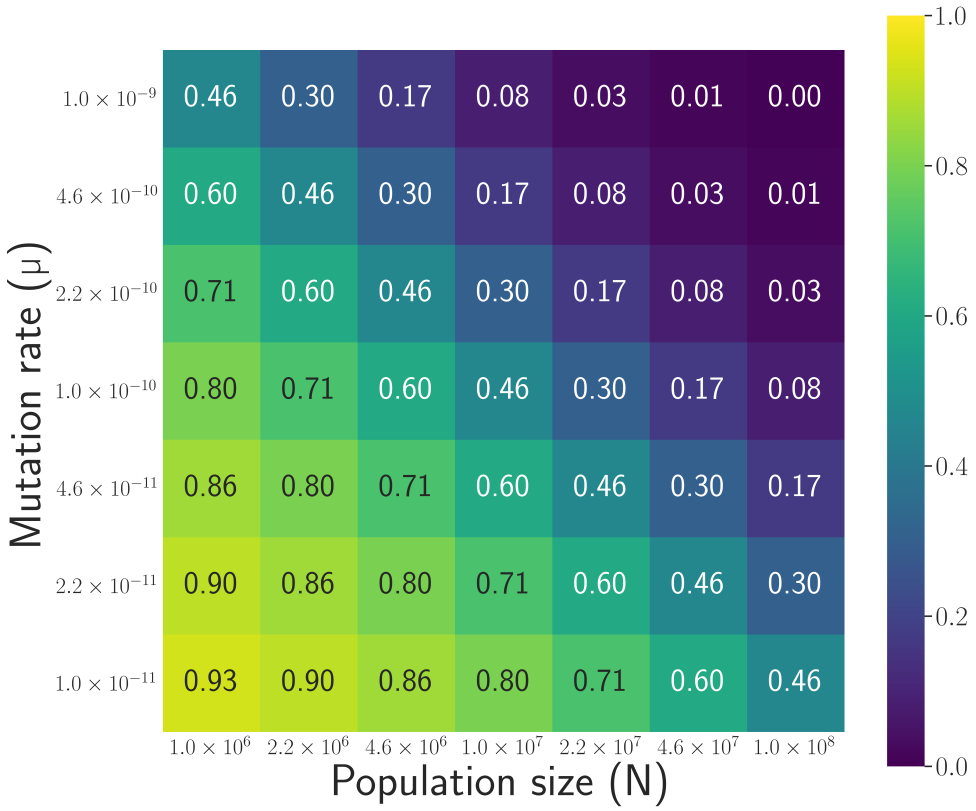

Figure C.1: Predicted non-coding fraction at equilibrium for different values of N and μ. The genome architecture is fixed at $z_c = 1,000,000$ and $g = 2,000$, and we have $\lambda_{\text{dupl}} = \lambda_{\text{del}} = 1$.

A change in N or in μ by the same factor results in the same non-coding fraction.

C.4 SIMPLIFIED MODEL WITH ONLY INDELS (NO STRUCTURAL MUTATIONS)

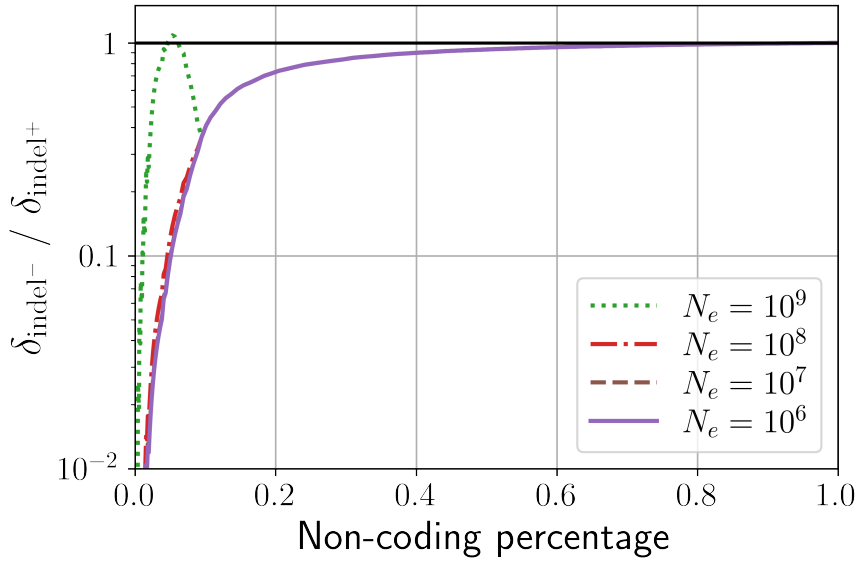

Figure C.2: Measured bias for different non-coding proportions. Genome architecture is fixed at $z_c = 1,000,000$ and $g = 2,000$, the mutation rate is fixed at $\mu = 1 \times 10^{-10}$ and $\lambda_{\text{indel}^-} = \lambda_{\text{indel}^+} = 1$. The maximum size of indels (l_m) is 50. z_{nc} varies in a logspace from 10^3 to 10^9, and four different values of N are depicted. The black horizontal line shows the potential equilibrium at $B = 1$, but is only crossed for $N_e = 10^9$.

Except for the highest value of N_e, the bias converges towards 1 as the non-coding proportion increases but is always below 1: when only indels are modeled, genome size would grow indefinitely in most cases in our model. Yet, when non-coding segments of the genome are small compared to the size of indels ($z_{nc}/g \ll l_m$), small deletions create a selection for robustness similar to the one of structural mutations: indels are more numerous as genome size increases but only marginally less deleterious and so genome growth could counter-selected. As a result, if the selection for robustness is very strong (very large N_e), it can counterbalance the difference in neutrality between small insertions and small deletions and thus there can be an equilibrium genome size, at a size where the intergenic segments are lower than the maximal size of indels.

C.5 EQUATIONS WITH THE FULL SET OF MUTATIONS

We note M the set of six mutations: duplications, deletions, inversions, small insertions, small deletions, and point mutations.

C.5.1 *Effective fitness*

$$f_e(\mu, g, z_c, z_{nc}) = \prod_{i \in M} (1 - \mu + \mu\, \nu_i(g, z_c, z_{nc}))^L \tag{C.1}$$

C.5.2 *Overall bias*

$$\begin{aligned} B(\mu, N, g, z_c, z_{nc}) &= \frac{\mu\ L\ N\ \delta_{del}(\mu, N, g, z_c, z_{nc}) + \mu\ L\ N\ \delta_{indel^-}(\mu, N, g, z_c, z_{nc})}{\mu\ L\ N\ \delta_{dupl}(\mu, N, g, z_c, z_{nc}) + \mu\ L\ N\ \delta_{indel^+}(\mu, N, g, z_c, z_{nc})} \\ &= \frac{\delta_{del}(\mu, N, g, z_c, z_{nc}) + \delta_{indel^-}(\mu, N, g, z_c, z_{nc})}{\delta_{dupl}(\mu, N, g, z_c, z_{nc}) + \delta_{indel^+}(\mu, N, g, z_c, z_{nc})} \end{aligned} \tag{C.2}$$

C.6 AVERAGE SIZE OF NEURTAL MUTATIONS

We want to compute the spontaneous contribution of the different mutations to genome size changes. As non-neutral mutations are lethal, their size is counted as 0: they cannot change the genome size. Four types of mutations can change the genome size: deletions and duplications, as well as small deletions and small insertions.

C.6.1 *Genome size change due to a neutral deletion*

$$\begin{aligned} \eta_{del} &= g \sum_{i=1}^{z_{nc}/g} \left(\frac{1}{L} \sum_{j=i}^{z_{nc}/g} \frac{1}{L} (j - i + 1) \right) \\ &= \frac{g}{L^2} \sum_{i=1}^{z_{nc}/g} \sum_{j=i}^{z_{nc}/g} (j - i + 1) \\ &= \frac{g}{L^2} \sum_{i=1}^{z_{nc}/g} \sum_{j=1}^{z_{nc}/g - i + 1} j \\ &= \frac{g}{2L^2} \sum_{i=1}^{z_{nc}/g} (z_{nc}/g - i + 1)(z_{nc}/g - i + 2) \\ &= \frac{z_{nc}(\frac{z_{nc}}{g} + 1)(\frac{z_{nc}}{g} + 2)}{6L^2} \end{aligned}$$

C.6.2 *Genome size change due to a neutral duplication*

$$
\begin{aligned}
\eta_{\text{dupl}}(g, z_c, z_{nc}) &= g \sum_{i=2}^{L/g} \left(\frac{1}{L} \sum_{j=i}^{L/g} \frac{1}{L} \left(g \sum_{k=0}^{z_{nc}/g} \frac{1}{L}(j-i+1) \right) \right) \\
&= \frac{g(z_{nc}+g)}{L^3} \sum_{i=2}^{L/g} \sum_{j=i}^{L/g} (j-i+1) \\
&= \frac{g(z_{nc}+g)}{L^3} \sum_{i=2}^{L/g} \sum_{j=1}^{L/g-i+1} j \\
&= \frac{g(z_{nc}+g)}{2L^3} \sum_{i=2}^{L/g} \left(\frac{L}{g} - i + 1 \right) \left(\frac{L}{g} - i + 2 \right) \\
&= \frac{g(z_{nc}+g)}{2L^3} \sum_{i=1}^{L/g-i} (i)(i+1) \\
&= \frac{g(z_{nc}+g) \left(\frac{L}{g} - 1 \right) \left(\frac{L}{g} \right) \left(\frac{L}{g} + 1 \right)}{6L^3} \\
&= \frac{(z_{nc}+g) \left(\frac{L}{g} - 1 \right) \left(\frac{L}{g} + 1 \right)}{6L^2}
\end{aligned}
$$

C.6.3 *Genome size change due to a neutral Indel$^-$*

A small deletion has size $k \leq l_m$ with probability $\frac{1}{l_m}$ so the mean size is :

$$
\begin{aligned}
&\eta_{\text{indel}^-}(g, z_c, z_{nc}) \\
&= g \left(\sum_{i=1}^{z_{nc}/g-(l_m-1)} \left(\frac{1}{L} \sum_{k=1}^{l_m} \frac{k}{l_m} \right) + \sum_{i=z_{nc}/g-(l_m-2)}^{z_{nc}/g} \frac{1}{L} \sum_{k=i}^{z_{nc}/g} \frac{(k-i+1)}{l_m} \right) \\
&= \frac{g}{L\, l_m} \left(\sum_{i=1}^{z_{nc}/g-(l_m-1)} \frac{l_m(l_m+1)}{2} + \sum_{i=z_{nc}/g-(l_m-2)}^{z_{nc}/g} \sum_{j=1}^{z_{nc}/g-i+1} j \right) \\
&= \frac{g}{L\, l_m} \left(\frac{l_m(l_m+1)}{2} \left(\frac{z_{nc}}{g} - (l_m-1) \right) + \sum_{i=z_{nc}/g-(l_m-2)}^{z_{nc}/g} \frac{\left(\frac{z_{nc}}{g}-i+1\right)\left(\frac{z_{nc}}{g}-i+2\right)}{2} \right) \\
&= \frac{1}{2L} \left((l_m+1)(z_{nc} - g(l_m-1)) + \sum_{j=1}^{l_m-1} j(j+1) \right) \\
&= \frac{1}{2L} \left((l_m+1)(z_{nc} - g(l_m-1)) + \frac{l_m(l_m-1)(2l_m-1)}{6} + \frac{l_m(l_m-1)}{2} \right) \\
&= \frac{1}{2L} \left(z_{nc}(l_m+1) - g({l_m}^2-1) + \frac{2{l_m}^2 - 3l_m + 1}{6} + \frac{l_m-1}{2} \right) \\
&= \frac{1}{L} \left(z_{nc} \frac{l_m+1}{2} + g \frac{1-{l_m}^2}{3} \right)
\end{aligned}
$$

C.6.4 *Genome size change due to a neutral Indel$^+$*

$$
\begin{aligned}
\eta_{\text{indel}^+}(g, z_c, z_{nc}) &= g \sum_{i=0}^{z_{nc}/g} \frac{1}{L} \sum_{k=1}^{l_m} \frac{k}{l_m} \\
&= \frac{g}{L\, l_m} \left(\sum_{i=0}^{z_{nc}/g} 1 \right) \left(\sum_{k=1}^{l_m} k \right) \\
&= \frac{(z_{nc}+g)(l_m+1)}{2L}
\end{aligned}
$$

After calculation, we have $\eta_{\text{indel}^+} > \eta_{\text{indel}^-}$, and since $z_c/g \geq 1$ unless all coding sections are only composed of promoter sequences, we also have $\eta_{\text{dupl}} > \eta_{\text{del}}$. Thus, there exists a neutral bias towards non coding genome size increase. If phenotypical adaptation was constant, genomes would tend to gain more new non-coding bases through duplications than what they lose through deletions. However, we do not observe an infinite growth of genome sizes, and that is due to a variation in the probability of fixation of our neutral mutations.

D

EUKARYOTIC ANCESTRY IN A FINITE WORLD

FOREWORD

The following work is an ongoing collaboration with Manuel Lafond, Associate Professor in computer sciences at the University of Sherbrooke. The project started during an international mobility to Sherbrooke from April to June 2024 and initially focused on how a eukaryote population across time, *i.e.* finding whether there are more appropriate measures than the average of the population to describe the evolutionary dynamics. As we read the literature on eukaryote ancestry studies, we found no framework that takes into account the whole complexity of possible mutations as well as recombination events while reconstructing the ancestry of a population. Here, we propose to study such a population with only recombination but no approximation on population size or chromosome length — contrary to many mathematical approaches — to understand how genetic information is structured backward in time.

D.1 INTRODUCTION

Understanding genealogical and genetic ancestry of populations is central to the coalescent theory, a widely applied model in population genetics to infer demographic histories (Sigwart, 2009). Several mathematical predictions can be derived from this theory, providing insights into the evolution of lineages of various populations. For instance, it is well-known that the genealogical ancestry of asexually reproducing organisms eventually coalesces into a single individual, with the time of convergence depending on the population size and structure (Hein et al., 2004). Sexual reproduction and diploidy complicate the picture, but several predictions on genealogical ancestry are still possible (Derrida et al., 2000; Brunet and Derrida, 2013). Genetic ancestry requires establishing which ancestors (or even chromosomes) have left genetic material in the extant population and is often more challenging to understand. While a single non-recombining genetic segment across a population coalesces to a single ancestral genome, mimicking haploid asexual dynamics (Hartfield et al., 2016), recombinations fragment the chromosome into several ancestral segments that are dispersed throughout the genealogical ancestors. Modeling the ancestry of a chromosome in eukaryotic recombining populations is therefore a challenge, and many questions are unanswered. For example, what proportion of the ancestral genealogical ancestors is also a genetic an-

cestor to one extant chromosome or the whole population? Is there an equilibrium regime regarding the number of segments and ancestors, and how many generations does it take to reach it?

These questions present overwhelming challenges in both theory and practice, which often need to be circumvented through simplifying assumptions or approximations. In particular, several mathematical predictions assume that variables such as population size, genome length, or evolutionary time tend to infinity. Notable examples include a prediction of an equilibrium state in which about a proportion of about 0.7968 of the population in each generation has extant descendants (Derrida et al., 2000; Brunet and Derrida, 2013), a closed-form formula for the expected number of ancestral segments of an extant segment or interest (Wiuf and Hein, 1997), or the distribution of the surviving segments of an ancestral genome (Baird et al., 2003). More recently, Gravel and Steel (2015) derived that the proportion of *superghosts*, which are genealogical ancestors of the whole population but are not genetic ancestors of anyone, tends to 0.7968, *i.e.* the proportion of genealogical ancestors, if the population size and the time tend to infinity. In all these works, studying the limits greatly simplifies calculations that would otherwise be impossible, although it is sometimes unclear whether these results connect to our finite reality.

Another simplification consists of viewing time as continuous (Hudson, 1983), as opposed to the classical Wright–Fisher model, which models time as a discrete sequence of generations. Several results were derived in the continuous approximation (Wiuf and Hein, 1997; Schweinsberg, 2001; Sagitov, 2003), but Davies et al. (2007) have argued that this can lead to inaccurate predictions of non-local quantities such as the equilibrium number of ancestors, or of the dynamics to reach that equilibrium. In terms of results in a finite reality, Chapman and Thompson (2003), Agranat-Tamir et al. (2024), and Derrida and Jung-Muller (1999) provided the exact expected number of ancestral segments or genetic ancestors, but only up to a few generations in the past, or in the case of a small chromosome length. Notably, the questions of the expected number of ancestors of a segment and the expected length of its ancestral segments in the equilibrium state were raised more than two decades ago by Derrida and Jung-Muller (1999), but remain completely unanswered, even when allowing the above simplifications.

In practice, simulations are commonly performed to gain insights into these difficult questions. However, simulation software for large diploid populations undergoing recombination also face limitations, as they are constrained by the significant computational resources required to maintain the state of individual and segment lineages. Forward simulators, which follow the evolution of populations from past to present, provide a realistic representation of genetic processes Yuan et al., 2012, but are more time- and memory-intensive, limiting their

scalability. As a result, several works have focused on the more efficient backward simulators. Hudson's classical *ms* (Hudson, 2002) is often regarded as a gold standard approach for backwards simulations, as it simulates the whole ancestral recombination graph exactly. Due to the limited population size and segment lengths it can handle, dozens of simulators were subsequently developed to achieve better scalability through various approximations (see Hoban et al., 2012 for a survey). One important category of approximate models consists of spatial algorithms, which simulate local trees on the sites of the segments from left to right, starting with an initial tree at the first site and then spawning new lineages on sites affected by recombination (e.g., MaCS (Chen et al., 2009), SMC (McVean and Cardin, 2005), SMC' (Marjoram and Wall, 2006)). These approaches achieve great scalability, but as argued in (Wang et al., 2014) the effects of these simplifications are not fully understood. Other approaches, for instance, msms (Ewing and Hermisson, 2010), fastsimcoal (Excoffier and Foll, 2011), or msprime (Kelleher et al., 2016), rely on sampling a portion of the population to achieve efficiency. *forqs* is one of the few software that can simulate whole populations exactly Kessner and Novembre, 2014, but only for a few dozen generations before running out of memory. Recent efforts have then focused on adding features and realism to the simulators, for instance, by allowing different recombination hotspots and migration Shlyakhter et al., 2014, admixing populations Agranat-Tamir et al., 2024, and others Laval and Excoffier, 2004; Virgoulay et al., 2021. While it is certainly desirable to incorporate biological realism into coalescent models, it is still unclear how well the aforementioned mathematical predictions hold up in a finite universe, and we found no implementation able to compute the proportion of superghosts at equilibrium for modest effective population sizes.

In this work, we study those mathematical questions in a finite universe using exact back-in-time simulations of whole diploid populations experiencing recombination, tracking genealogical and genetic ancestry without approximations or sampling. Using a combination of compressed data structures, algorithmic optimizations, and parallel processing, our simulator tracks populations as large as one million individuals, each with multiple chromosomes of hundreds of thousands of sites in length. Our simulation is not limited by the number of desired generations, allowing us to observe populations until they reach stable, equilibrium states. This allows us to verify which results from coalescent theory hold under realistic, finite conditions. We focus on three aspects: how much time is required to reach an equilibrium state; how are genetic segments distributed in genetic ancestors; and what is the proportion of superghosts?

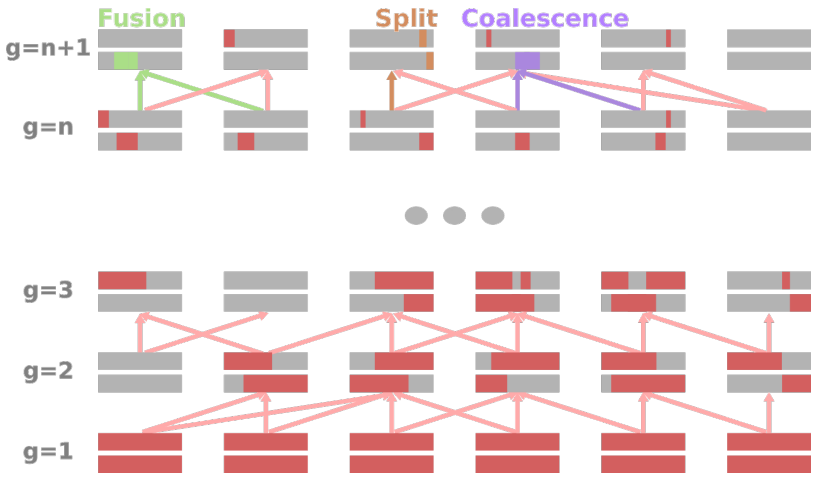

Figure D.1: Schematic representation of the model. Individuals have a single pair of chromosomes. Each individual of generation g chooses a parent for each of its chromosomes at generation $g+1$. Marked segments (in red) are followed in the previous generation. They can be split due to recombination events or fuse or coalesce into a single one.

D.2 MATERIAL AND METHODS

We first describe our evolutionary model in detail, along with the relevant implementation details of our simulator, and then present our experimental results in the next section.

D.2.1 *Model description*

We start our simulation from a population of size N, in which each individual owns c pairs of chromosomes (for a total of $2c$ chromosomes). We assume that each chromosome is of the same length L_c.Note that the length here is the number of possible recombination breakpoints along the chromosome and can have several interpretations: the number of base pairs if we consider that recombination can happen between any two base pairs; the number of genes if we consider that the only relevant information is which genes are on which side of the breakpoint; or any "block" which would represent the space between recombination hotspots. For simplicity, we will call each position of a chromosome a *base pair* in the rest of the paper.

For the genealogical ancestry graph, we use a standard discrete framework in which we simulate one generation at a time from present to past, with each generation containing N individuals. The first generation consists of the extant population, and the count goes backward in time (so higher generation numbers refer to populations from a more distant past). To obtain generation $g+1$ from generation g, each of the N individuals from generation g chooses two parents uniformly at random, among the N individuals from generation $g+1$. The choices are

made with replacement so that selfing is possible, although this has little bearing on the results according to our experiments. Note that our simulations follow a Wright-Fisher model (Wright, 1931; Fisher, 1923): all individuals are replaced at each generation, and we assume equal fitness and panmixia. As such, by definition, our census population size equals the effective population size.

For genetic ancestry, we "mark" each base pair of each individual in the extant population and follow them backward in time, considering recombinations (described below). The base pair at position i of an individual in a chromosome x is acquired from one of the parents, from the same base pair position i in either the same chromosome x or its homologous copy, depending on how the parents recombined. When viewed backward, this means that each base pair has exactly one parent among all base pairs present in the parent generation. Rearrangements that could alter the relative positions of base pairs are not modeled. The ancestors of the base pairs of the extant population are called *ancestral base pairs*. The set of base pairs on the same position and chromosome across the extant population can be seen as following a coalescent process, and eventually, they will share a single common ancestor. This also means that eventually, an ancestral population will possess exactly $c \cdot L_c$ ancestral base pairs. Note that the model also allows to "mark" initially only a sample of the initial population. This allows studying how the extent of the initial sampling of a population influences the information accessible about the ancestors from that population.

For our purposes, it is sufficient to track contiguous ancestral segments instead of individual base pairs. Initially, the base pairs of an individual are split into exactly $2c$ contiguous segments, as there are c pairs of chromosomes per individual, for a total of $2Nc$ segments to track when considering the whole population. As we go back in time, a segment can be split into two or more segments due to recombination events. More precisely, after an offspring has chosen its two parents (see previous paragraph), for each chromosome $x \in \{1, 2, \ldots, 2c\}$ a number k of recombinations is drawn according to a rate r of events per base pair per generation. The k recombination positions are then drawn uniformly at random along the length L_c. One of the two homologous copies of the chromosome x is chosen with equal probability for each parent, and each recombination position alternates the parental chromosome from which a segment is inherited. This allows us to determine how the segments currently tracked in the copies of x are partitioned among the chosen parental chromosome copies (for example, an ancestral segment $[i, j]$ could be split into $[i, l], [l + 1, j]$ if a recombination occurs at position l). After each individual is handled, each chromosome in the parental generation has a list of segments to track. Individuals with multiple children may contain overlapping segments, for example, it may need to track

$[i, j]$ and $[i', j']$ with $i < i' < j$, in which case the two segments are *fused* into $[i, j']$. In this manner, we only track *maximal* segments, i.e., segments that cannot be extended into a longer contiguous segment (note that a pair of segments $[i, j], [j + 1, j']$ will also be fused).

An ancestral individual is a *genetic ancestor* if it contains at least one tracked segment, that is, if it has a base pair ancestral to some base pair from an extant individual. We may also refer to a specific chromosome copy as a genetic ancestor if it contains a tracked segment.

D.2.2 *Experimental design*

To test the impact of variation in each of the relevant parameters (population size N, chromosome length L_c, number of pairs of chromosomes c, and per-base recombination rate r), we take a reference value for each of them and vary them separately. The reference values for our experiments are:

- $N = 20,000$
- $L_c = 10,000$
- $c = 36$
- $r = 1/L_c$

The reference population size is based on a usual estimate for human effective population size (Lynch et al., 2023). Having 36 chromosomes that undergo on average one recombination per generation also mimics the human genome, as its length is 36 Morgan, and is similar to what has been done by Gravel and Steel (2015).

Each combination of parameters is run with 3 different pseudo-random seeds to avoid degenerated cases and ensure the robustness of our results. To test the impact of L_c, we let it vary from $5,000$ to $500,000$, keeping the other parameters constant. Note that this does not reflect the physical length (in terms of base pairs) of the human chromosomes, as we test here the number of possible recombination breakpoints, more than the length in base pairs. To test the impact of N, we let it vary from 20 to $200,000$, $20,000$ being a standard approximation of human effective population size. Finally, to test the impact of genome structure, we change the number of pairs of chromosomes from 1 to 36, keeping the total genome length and recombination rate constant.

D.2.3 *Simulating until the equilibrium state*

Since several mathematical results assume that time tends to infinity, we aim to perform simulations until an equilibrium state is reached. It isn't easy to define such a state formally. Still, it can loosely be described as a state in which simulating further generations would provide no additional information on our variables of interest because their values do not change or only vary around a stable average.

In our experiments, we saw that among our variables of interest, the number of tracked bases (or more precisely, the sum of lengths of the tracked segments) usually took the longest time to converge. That is, the minimum number of tracked bases is $L_c \times c$, as no base can coalesce with another once it has no "twin" base at the same locus in any other individual. In other words, a base at position i in an extant chromosome has at least one ancestral base in the same position i in the same chromosome pair in any generation. This minimum will eventually be achieved once all bases at a given position in all individuals have coalesced.

We therefore define the *equilibrium* as the number of generations required to have exactly $L_c \times c$ tracked bases across the whole population. To estimate the time to reach this equilibrium, notice that for any position i, there are initially $2N$ distinct tracked base pairs at that position. Two of those tracked bases fuse when they choose the same parental chromosome, and so a fusion should occur with probability around $1/(2N)$. This mimics a standard coalescent process, in which case the waiting time to reach a single individual is linear in N. Hence, the expected time to reach equilibrium should be proportional to N. Do note that equilibrium requires coalescence of *all* positions, and the coalescence events cannot be treated as independent, so obtaining an exact formula for the expected time to equilibrium is left as an open problem. In any case, we choose to run simulations for $200,000$ generations, which is ten times the default population size and should therefore enable most parameters to converge.

D.2.4 *Technical aspects*

Our C++ simulator maintains the list of genealogical ancestors and the list of segments in memory only for the current and previous generations, which limits memory requirements to a fixed amount (with an exception for superghosts, see below) [1]. We face three major bottlenecks: computing random numbers, sorting segments, and computing the number of superghosts. Recall that each individual and chromosome chooses a random parent, along with a random number and location of breakpoints. This may require hundreds of millions of random integers per generation, which is too slow using default libraries. Instead, at the start of a generation, the number of necessary random integers is calculated in advance, and all random numbers are computed in large blocks in parallel using the recent P2RNG library (`https://github.com/arminms/p2rng`). From the breakpoints, we infer the segments in the next generation, with possible overlaps. By sorting these segments, we can determine in linear time which ones need to be fused, and we used the pattern-defeating quicksort

1 We do maintain statistics at each generation for the program output, but its memory is linear in the number of generations and is negligible)

implementation from Peters (2021), which sped up our simulations significantly.

Counting the number of superghosts was challenging for large populations. Recall that a superghost is a genealogical ancestor of the whole population, but is not a genetic ancestor, *i.e.*, it has no tracked segment. For each individual, we must check whether the whole population descends from it, which requires storing a set of size up to N representing its descendants (this set is the union of descendants of its children). This requires $O(N^2)$ space, which is prohibitively large when $N \geq 100,000$, even using compressed data structures. Instead, we store the ancestry graph, in which the vertices are the individuals from all generations, who have an edge towards their children from the previous generation. One can check whether a given individual is a superghost by checking which extant individuals it reaches in this graph. This approach is too slow, however. Instead, we split the extant population into blocks of size B, a parameter, and ask: which individuals are ancestors of this whole block? This step can be parallelized over the blocks, and the ancestors of all the population are in the intersection of the ancestors of all blocks, making their computation viable even with $N = 1,000,000$ (we used $B = 5,000$). The astute reader will notice that storing the ancestry graph takes $O(N)$ space per generation and therefore imposes memory limitations. However, it is known that after $O(\log N)$ generations, individuals are either ancestors of all or none of the extant population, at which point we do not need the graph. It was therefore sufficient to store the graph up to 100 generations, which is fine in terms of memory.

To give an idea of the scalability, we could simulate $N = 200,000$, $L_c = 500,000$, $c = 36$, $r = 1/L_c$ for 200,000 generations in about half a day on a laptop with 16Gb RAM. The code for the simulator is made accessible at `https://github.com/jluiselli/euktree-simulation`.

D.3 RESULTS

D.3.1 *Time to reach equilibrium*

As justified in the previous section, the time at which the equilibrium is reached, noted T_{eq}, is defined here as the time at which the number of tracked bases is equal to $L_c \times c$, the chromosome length times the number of pairs of chromosomes. Indeed, this is the minimal number of bases followed, and it is an absorbing state since no more bases can coalesce once this is reached. We compare T_{eq} for different population sizes, as it is the only parameter that significantly impacts the time to reach the equilibrium in our experiments.

As shown in Figure D.2 (left), the number of base pairs decreases very fast initially but starts to decrease more slowly as it gets closer to the minimum $L_c \times c$. As such, the equilibrium is not reached within

the 200,000 generations of the simulation for $N \geqslant 20,000$, despite it being seemingly very close for $N = 20,000$. Noting that T_{eq} seems to depend linearly on N for $N < 20,000$, we use the measures of T_{eq} for $N < 20,000$ to fit a regression of T_{eq} as a function of N (see Figure D.2 right). For $N = 20,000$, we predict $T_{eq} \simeq 570,000$, for $N = 100,000$, $T_{eq} \simeq 2,800,000$ and for $N = 200,000$, $T_{eq} \simeq 5,700,000$. Although our simulator could reach these numbers of generations, we believe that going so far in the past is not relevant to biological data, as species evolve and undergo major changes within these time frames. This also suggests that in some cases, mathematical results that require the time T to tend to infinity may sometimes require T to be extremely large to be applicable.

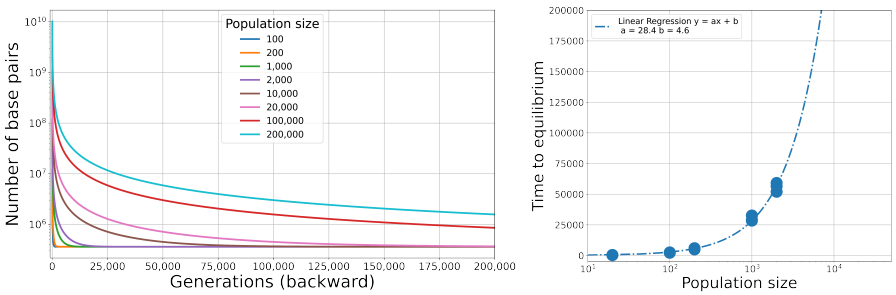

Figure D.2: (left) Number of ancestral bases followed across time. Note that the equilibrium is not reached for the three larger population sizes of $N \geqslant 20,000$. (right) Time at which the equilibrium is reached for different population sizes and the associated linear regression.

Since for the reference population size $N = 20,000$, the equilibrium is reached at the end of our simulations ($T = 200,000$), we will compare data at $T = 200,000$ for the rest of the manuscript. Detailed temporal data provided in the Supplementary Materials (Section E.1) support that the variables of interest are stable around this time. Additionally, differences in the measured variable of interest appear very early in the simulations, showing that the tendencies we describe are already relevant a few generations in the past, thus at relevant biological time scales.

D.3.2 *Segments lengths and distribution*

We now turn our attention to how the genetic information from the extant generation is distributed among the ancestors. This information can be more or less fragmented as it is spread across more or less segments, *i.e.*, contiguous portions of bases that are ancestral to extant individuals, which are themselves distributed among the ancestors. The question of tracking the genetic ancestry of a genetic segment of interest was initiated by Wiuf and Hein (1997). The authors focus on the history of a single chromosome from a single individual and discuss the fact that, at equilibrium, the rate at which segments get

separated by recombinations should roughly match the rate at which they coalesce (which occurs when segments spanning two adjacent loci choose the same parent). Therefore, although the number of segments can oscillate, the mean number of segments across the population should converge to a well-defined value going far enough back in time. Similar reasoning applies to the mean length of the segments and to the number of ancestral chromosomes or individuals that possess these segments.

Wiuf and Hein (1997) assume that the population size N and the chromosome length L_c tend to infinity. The recombination rate is also assumed to tend to 0 as its growth should be inversely proportional to L_c. In the following, we will fix that $r = 1/L_c$. Under all these assumptions, the main predictions of interest for our purposes are that, at equilibrium: (1) the mean number of segments across the population is proportional to N; (2) the mean number of ancestral chromosomes is proportional to $N/\log(N)$[2]. Let us also mention that Derrida and Jung-Muller (1999) comes to similar conclusions, albeit with a different approach based on spin models in physics. They also propose approximations for the mean number of segments and their length in the case of finite populations, but, to our knowledge, the question of obtaining exact and efficiently computable means for given — and finite — N, r, and L_c remains open.

We consider here the number and average length of ancestral segments after 200,000 generations back in time. At equilibrium, the total length of the ancestral segments converged towards $2cL_c$ base pairs. Note that unlike Wiuf and Hein (1997), we track the whole population instead of a single individual, but since we track the same number of bases at equilibrium, comparing their predictions with our empirical values is meaningful.

NUMBER OF SEGMENTS AND ANCESTRAL CHROMOSOMES. Figure D.3 compares the predicted mean number of segments across the population, at equilibrium, with those obtained in our simulations. On the left, we see that the total number of segments does grow as the population increases. Indeed, as the population size increases, the probability for two ancestral segments to coalesce decreases. On the other hand, the probability for a segment to split solely depends on chromosome length and recombination rate and is thus constant, resulting in more but shorter ancestral segments. As N gets larger, the number of segments appears to grow more slowly and diverges from the prediction. A possible intuitive explanation is that L_c is fixed in our analysis, and so the maximal number of segments is fixed. For a very large N, every segment would be of size 1, and increasing N

2 Let us note that Wiuf and Hein give the values in terms of R, the expected number of recombinations per N_e generations, the effective population size. This value tends to rL_cN_e, and since $rL_c = 1$ we estimate this as N.

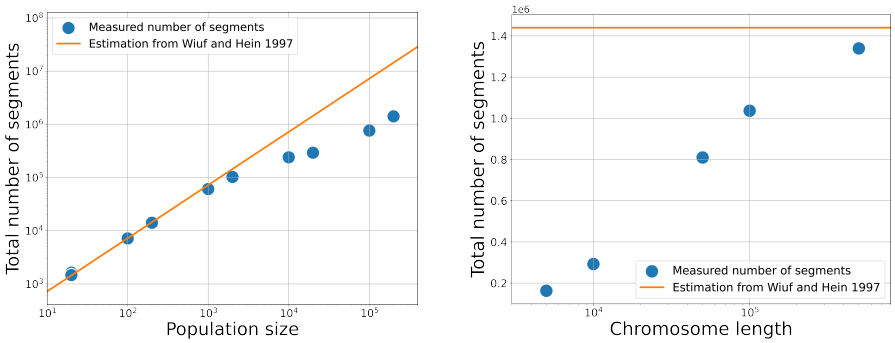

Figure D.3: Average number of segments at equilibrium, with respect to the population size (left) and chromosome length (right), with $r = 1/L_c$. Note that the computations of Wiuf and Hein (1997) are valid for one chromosome per individual. Since we have 36 pairs of chromosomes, we multiplied the prediction by 72. Temporal data for the number of segments are provided in the Supp. Figure E.1.

further does not increase the number of segments as they cannot split. Here, we are probably approaching this limit, and some segments are too small to split — hence, the split rate is not constant but decreases with the number of segments. This suggests that N and L_c must grow together for the prediction of Wiuf and Hein (1997) to hold.

On the right of Figure D.3, we recall that the prediction of the total number of segments does not depend on the chromosome length in Wiuf and Hein (1997) (assuming $r = 1/L_c$). The plot suggests that our simulations could reach the prediction once chromosomes are large enough ($L_c > 10^6$ or above), which is a very high number of possible breakpoints along a chromosome. This reiterates the need to be careful when using such predictions with finite parameters.

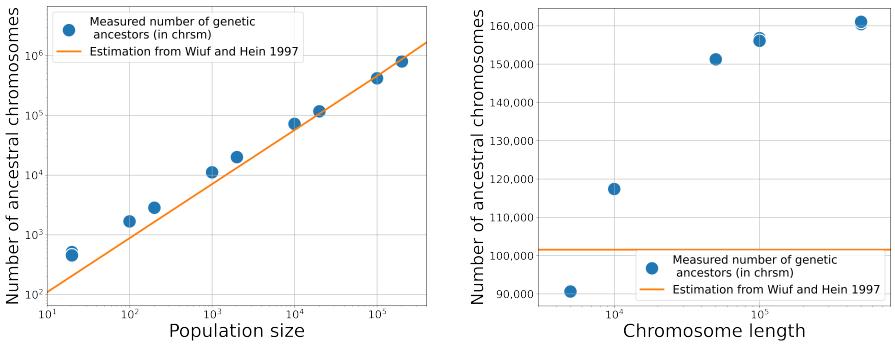

Figure D.4: Average number of ancestral chromosomes that possess extant genetic material, with respect to the population size (left) and chromosome length (right), still with $r = 1/L_c$. Note that the computations of Wiuf and Hein (1997) are valid for one chromosome per individual. Since we have 36 pairs of chromosomes, we multiplied the prediction by 72. Temporal data of the number of ancestral chromosomes in our simulations are provided in the Supp. Figure E.2.

Figure D.4 on the left shows the comparison between the mean number of ancestral chromosomes (i.e., that possess at least one segment) in our simulation and the prediction of Wiuf and Hein (1997). We find that the prediction is quite accurate, even for small population sizes. There is probably a limit to this accuracy: when L_c is fixed, the maximum number of possible ancestors is also fixed, and so the latter cannot keep increasing with N. Nevertheless, the plot suggests that this phenomenon occurs only when N gets very large, and in this case, the prediction appears usable on finite populations.

According to Wiuf and Hein (1997), the predicted number of ancestors does not depend on L_c when $r = 1/L_c$, whereas we observe that this value increases with chromosome size (Figure D.4, right). It is plausible that if we considered even larger L_c values, the number of ancestors would converge to a constant. Moreover, the plot suggests that this value of convergence could be close to the prediction, i.e., within a small constant factor. The discrepancy could be due to the fact that here, N is fixed. It is possible that a larger N would get us closer to the prediction.

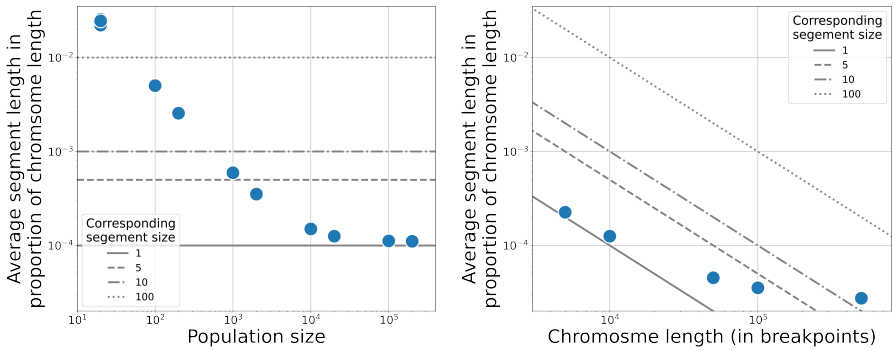

Figure D.5: Average segment length (in proportion of chromosome length) with respect to population size (left) and chromosome length (right). The segment sizes are divided by L_c to provide comparable measurements. For ease of reading, gray lines illustrate the correspondence in absolute segment size. Temporal data are provided in the Supp. Figure E.3.

SEGMENT LENGTHS. We now turn to the average length of segments at equilibrium, as seen in Figure D.5. When L_c remains fixed and the population grows, predictions state that the average segment length should tend to 1, as the probability of two segments coalescing into the same parent becomes much smaller than the probability of two consecutive bases being separated by a recombination. This trend is confirmed in Figure D.5 on the left, where early on a linear decrease in segment length is observed until it stabilizes close to 1, *i.e.* 1×10^{-4} of chromosome length.

On the right, we exhibit the relationship between segment length and chromosome length. Here, we measure segment lengths in the percentage of chromosome size, as our different sizes could represent

the same physical chromosome length but with different distributions of potential recombination breakpoints. That is, recall that we assume a constant recombination rate of $1/L_c$ and thus chromosomes of a size of 1 Morgan regardless of L_c. This implies that, for example, a segment of size s is much more likely to be broken on a chromosome of size $10,000$ than on a chromosome of size $100,000$, which makes the length proportion more meaningful than the absolute values.

We could expect the average segment length to be a constant proportion of the chromosome length, as the probability to coalesce depends solely on N, while the probability to split depends on segment length (s) and chromosome length (L_c) and should be $\frac{s}{L_c}$. However, this second statement does not hold, as a segment of size $s = 1$ cannot be split. As a result, there is a limit-induced effect when the average segment length is small in absolute value: as when there is a significant proportion of segments of size 1, their average split rate is lower while the coalescence rate remains constant. This results in larger segments (in proportion on chromosome length) for shorter chromosomes, as demonstrated in Figure D.5 right. As chromosome length increases, the intensity of this border-induced effect decreases, and the average segment should converge to a constant proportion of chromosome length. Indeed, we can see that this value is very close for $L_c = 100,000$ and $L_c = 500,000$.

This shows that the number of possible breakpoints along a chromosome (or the chromosome length) makes a difference in our variables of interest, and so the measure of chromosome length in terms of Morgan is not enough to determine the behavior of the genetic ancestry of a population along a chromosome.

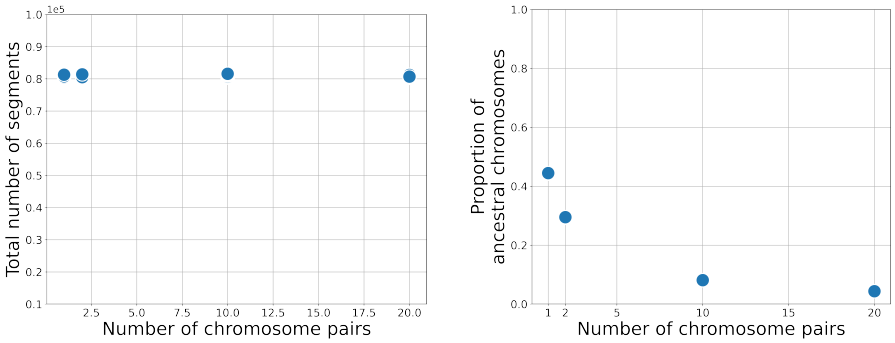

Figure D.6: Number of ancestral segments (left) and proportion of chromosomes that are genetic ancestors (right) for different genome structure. Other parameters are fixed at $c \times L_c = 200,000$, $r = 1/L_c$ and $N = 20,000$. The temporal data are presented in Supp Figure E.4.

IMPACT OF THE NUMBER OF CHROMOSOMES. Finally, the genome structure could change the distribution of ancestral segments in the population, as breaking the genome into chromosomes effectively adds obligatory recombination points. To test this, we compare simulations

with different numbers of chromosomes but a constant total genome size and average number of recombination per generation. To our knowledge, this question has not been studied in the literature, either in theory or in practice.

Interestingly, the additional breakpoints (between the chromosomes) have little to no effect on the number of segments (see Figure D.6, left), probably because their number is negligible compared to the total number of possible breakpoints. Yet, genome structure does impact the distribution of ancestral genetic material as it changes non-linearly the probability of a chromosome to be a genetic ancestor: having 1 chromosome instead of 2 does not double its probability to be a genetic ancestor (see Figure D.6, right), contrary to what would be expected with a constant number of segments uniformly distributed within the chromosomes. This makes the distribution of ancestral segments very difficult to study analytically, hence the need for simulations to understand it and the impact of complex parameters such as the genome structure.

D.3.3 *Ghosts and superghosts*

We now turn to ghosts and super-ghosts, as introduced in (Gravel and Steel, 2015). A *ghost* is an individual from a past generation that is a genealogical ancestor of at least one individual from the extant population, but that is *not* the genetic ancestor of any individual, *i.e.* that does not possess any ancestral segment. A *super-ghost* is a ghost that is a genealogical ancestor of *every* individual of the extant population. That is, super-ghosts are common ancestors of everyone, but they leave genetic material to no one.

IMPACT OF THE CHROMOSOME LENGTH AND POPULATION SIZE. In (Gravel and Steel, 2015), Proposition 2.2 states that

$$\lim_{N\to\infty}\lim_{T\to\infty} q(N,T) \simeq 0.7968,$$

where $q(N,T)$ is the probability that a randomly chosen individual in a population of size N at time T is a super-ghost, assuming that the chromosome size L_c is an arbitrary constant. Recall that 0.7968 is the equilibrium proportion of genealogical ancestors. In essence, this is saying that if we look far enough in the past and populations are large enough, virtually all ancestors that have descendants leave no genetic material in those descendants.

Importantly, the proof requires L_c to be fixed, the argument being that at equilibrium, a total of only L_c bases eventually remain in circulation, making the number of possible genetic ancestors constant, whereas N grows to infinity. Figure D.7 (top) confirms that if L_c is small compared to N, then the proportion of super-ghost does approach 0.796 after enough time (this is mostly visible for $L_c = 10$,

with $N = 20,000$). On the other hand, it is clear that this value is much harder, and perhaps impossible, to reach as L_c grows and N remains fixed. The most extreme $L_c = 500,000$ does not reach a proportion of super-ghosts beyond 0.05. In a finite universe, N/L_c does not tend to infinity, so it may be relevant to study the proportion of super-ghosts with respect to this ratio.

The authors also ask whether their result could hold when L_c is not fixed — specifically, the question is whether the same result holds in the continuous limit where genomes are represented as the set of real numbers $[0, 1]$, while still letting N tend to infinity (and maintaining the proportions of the recombination rate to one recombination per chromosome per generation).

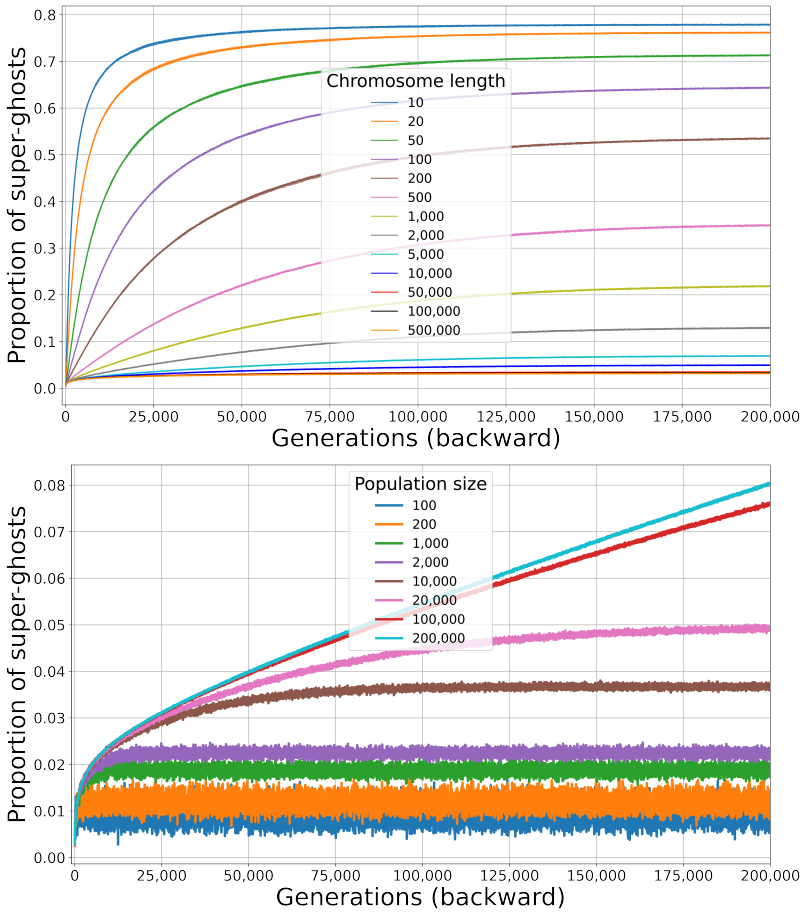

Figure D.7: Proportion of super-ghosts across time for different chromosome sizes (top) and population sizes (bottom). Note that N is fixed at $20,000$ on the left, and L_c at $20,000$ on the right. The shorter the chromosomes, the higher the percentage of super-ghosts, and he greater the population size, the higher the percentage of super-ghosts. Note that due to the large differences in values, the scale differs between the two plots.

To gain more insight on this question, Figure D.7 (bottom) shows the evolution in time of the proportion of super-ghosts as the population increases. Note that this analysis uses $L_c = 10,000$ and $c = 36$,

so the number of bases in the equilibrium is 360,000, a relatively large number that allows studying the behavior of super-ghosts as genomes allow numerous breakpoints. We see that the proportion of super-ghosts never goes above 0.1, well below 0.769. We reached an equilibrium for populations up to $N = 20,000$, but larger populations require much longer to attain equilibrium (see Section D.3.1). This analysis on large L_c suggests that there are two possibilities in the continuous limit of genome sizes: either the limit of $q(N, T)$ is strictly smaller than 0.769; or this proportion can be approached arbitrarily closely, but very slowly, that is, with extremely large populations and after waiting an extended amount of time.

Either way, we believe that more work is needed to predict the number of super-ghosts in realistic population sizes. It would be expected that the relevant parameter for biological populations would be the effective population size N_e. Our population follows a Wright-Fisher model, hence, the population size and the effective population size are the same. While $N_e = 20,000$ is a standard approximation of the effective population size of humans, the effective population size of some unicellular eukaryotes could be as large as tens or hundreds of millions (Lynch et al., 2023).

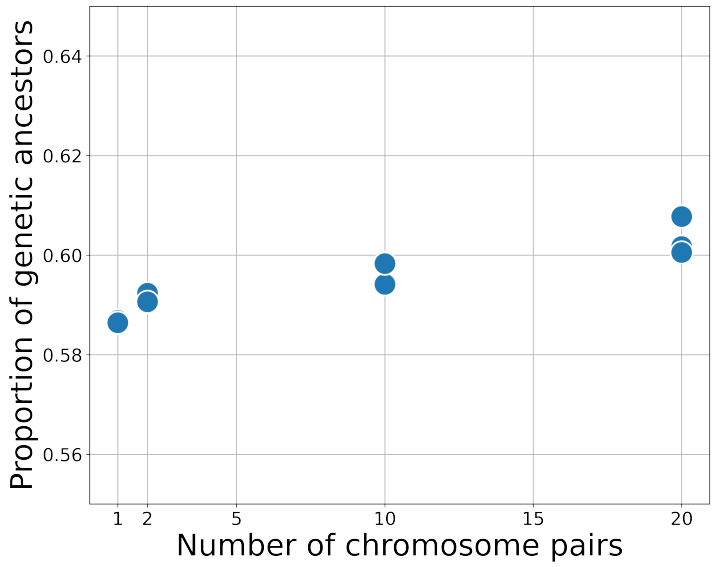

Figure D.8: Proportion of individuals that are genetic ancestors for different genome structures. Other parameters are fixed at $r = 0.0001$, $c \times L_c = 200,000$ and $N = 20,000$. Note that the difference in percentage is low but consistent: the more fragmented the genome is, the higher is the proportion of genetic ancestors.

IMPACT OF THE NUMBER OF CHROMOSOMES. In (Gravel and Steel, 2015), the model genome mimics a human genome with 36 pairs of chromosomes of size 1 Morgan to represent the 23 pairs of chromosomes of different sizes that undergo in total on average

36 recombination events per generation. Up to now, we have used the same approach for our simulation. Yet, the way chromosomes are partitioned could change the results, as shown in the previous section (Figure D.6). Figure D.8 shows that genome structure also changes the equilibrium fraction of super-ghosts due to a change in the number of individuals that are genetic ancestors. The more fragmented the genome is, the more genetic ancestors we have, and hence, the fewer super-ghosts. This cannot be explained solely by the additional breakpoints created by the presence of chromosomes, as the number of segments is sensibly the same with the different number of chromosomes (Figure D.6).

An explanation for our observations is that the ancestral segments are not distributed uniformly along the genome. To demonstrate that, let us assume the distribution is uniform. If we have one chromosome per individual, four ancestral segments, and a population size of 10, each individual has a probability $\frac{4}{10}$ to be a genetic ancestor. If we now have 2 chromosomes per individual but the same number of segments, each chromosome has a probability $\frac{4}{20}$ to be a genetic ancestor, hence each individual still has a probability $\frac{4}{10}$ to be a genetic ancestor. As this does not fit with our observations, the distribution of segments must not be uniform, which is expected since the probability to recombine between two segments and break their linkage depends on their physical distance.

To conclude, the approximation of using 36 same-sized chromosomes instead of 23 to model the human genome is questionable if the aim is to study the distribution of ancestral genetic material. This subtlety should be taken into account by future models.

D.4 DISCUSSION

Our work showed that many common approximations in the study of eukaryotic ancestry can have unexpected and unpredicted impacts on commonly studied variables and should thus be taken more into consideration when studying populations. Indeed, having a finite rather than an infinite population size (N) considerably changes the probability of ancestral segments to coalesce, hence changing their equilibrium number, size, and distribution. Similarly, the chromosome length (L_c), *i.e.* the number of possible breakpoints when studying recombination events, changes the probability to recombine between any two loci in the genome, which has a similar effect. Despite this, in our results, chromosome length has a significantly narrower impact on the ancestral segments distribution than population size. Therefore, approximating chromosomes with a continuous space (*i.e.* having infinite sized chromosomes) yields less questionable approximations than assuming an infinite population size.

Changes in the ancestral segment distribution, whether provoked by L_c or N, also change the proportion of genetic ancestors of the population (or the proportion of super-ghosts, as defined by Gravel and Steel (2015)). As such, it impacts the amount of information about past generations that is attainable by sampling and sequencing the extant population. Some invisible alleles probably transitively impacted selection and species adaptation to a given environment, advantaging some of the genealogical ancestors of the population, and yet were never transmitted to the extant population. A perspective of our work in this direction would be to carry out similar experiments but with an initial sample of the population instead of the whole population. This would allow retrieving the minimal proportion of individuals to sample to have the maximal amount of information on the ancestors at equilibrium.

Finally, this work opens the way to other interesting perspectives. The simulator could be extended to allow for the superimposition of neutral mutations on the ancestral graph, thus enabling the computation of polymorphism data. Indeed, polymorphism data are widely used to reconstruct species histories (Muller et al., 2006; Leaché and Oaks, 2017), and recent simulators allow for explicit sequence simulation and recombinations (Haller and Messer, 2023). However, more theoretical work is still needed to understand the impact of population size, of chromosome length, and of their interaction with recombination on polymorphism data. Similarly, many more extensions of the model could be proposed to perform large-scale simulations of ancestral graphs with recombination, including *e.g.* a population structure, a more detailed genomic structure (with chromosomes of different sizes and/or sex chromosomes, etc.), or the addition of a fitness function and non-neutral mutations.

Overall, we believe that more extensive studies on the reconstruction of eukaryotic ancestries are necessary to understand the classical approximations used in most studies. This would allow us to consciously, and on a case-by-case basis, choose which approximations are reasonable for ease of computing and which would have a too large impact on the results and should be avoided.

E

SUPPLEMENTARY MATERIALS FOR *EUKARYOTIC ANCESTRY IN A FINITE WORLD*

E.1 TEMPORAL DATA

E.1.1 *Number of ancestral segments across time*

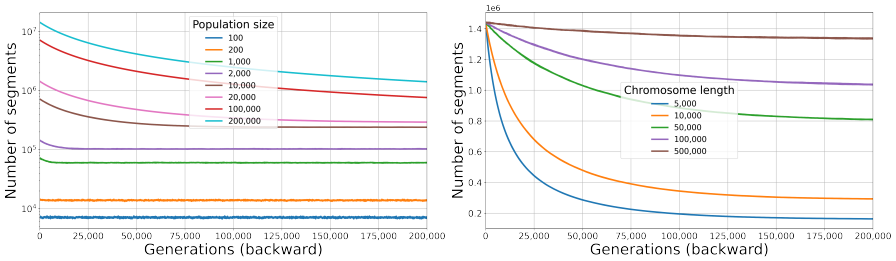

Figure E.1: Number of segments across time for different population sizes (left) and chromosome lengths (right). Note that due to large variability for the different population sizes, the scale on the left plot is logarithmic.

E.1.2 *Number of chromosomes that are genetic ancestors across time*

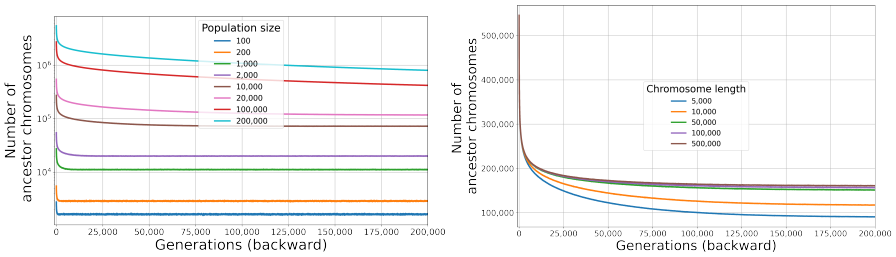

Figure E.2: Number of chromosomes that are genetic ancestors across time for different population sizes (left) and chromosome lengths (right). Note that due to large variability for the different population sizes, the scale on the left plot is logarithmic.

E.1.3 *Average segment length across time*

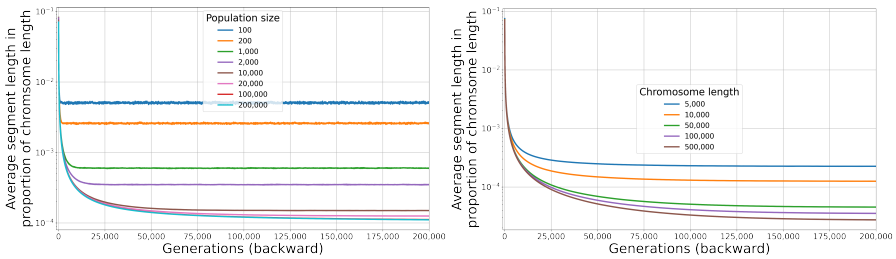

Figure E.3: Average length of ancestral segments across time for different population sizes (left) and chromosome lengths (right), in proportion of chromosome length. Note that the scales are logarithmic.

E.1.4 *Impact of the number of chromosome*

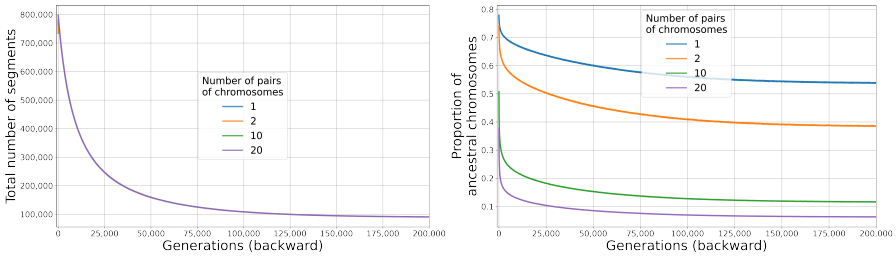

Figure E.4: Number of ancestral segments (left) and proportion of chromosomes that are genetic ancestors (right) across time, for different genome structures. The number of pairs of chromosomes varies, while the total genome size is fixed at $c \times L_c = 200000$, and the total per genome recombination rate at 20 per generation ($r = 1/L_c$).

F

SUPPLEMENTARY MATERIALS FOR *GENOME SIZE AND STRUCTURE: A DIRECT CONSEQUENCE OF REPRODUCTIVE MODE*

F.1 FULL WILD-TYPES DATA

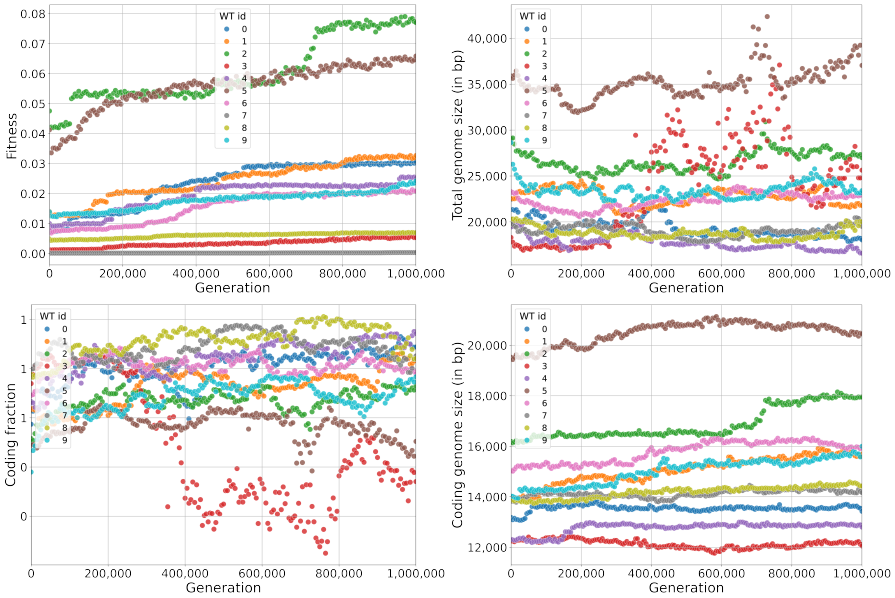

Figure F.1: Averages fitness (A), genome size (B), coding fraction (C), and coding size (D) for all 10 populations during $1,000,000$ generations after their diploidization.

We expect Wild-Types to have a stabilized genome structure. Each of them is relatively stable, except WT3, which undergoes wide variations in genome size and structure (B and C). We quantify this unusual behavior by measuring the variance in genome size along the experiment:

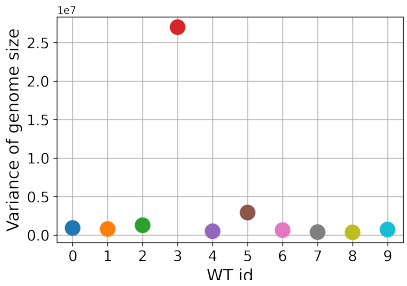

Figure F.2: Variance in genome size along 1,000,000 **generations for the** 10 **different populations**

Since WT3 has a very high variance in genome size and structure while the others are stable, we exclude it from further analyses. Full results with WT3 included will however always be presented in the supplementary material.

F.2 TRAJECTORIES OF FITNESS AND GENOMIC COMPONENTS AFTER THE INTRODUCTION OF SELF-FERTILIZATION

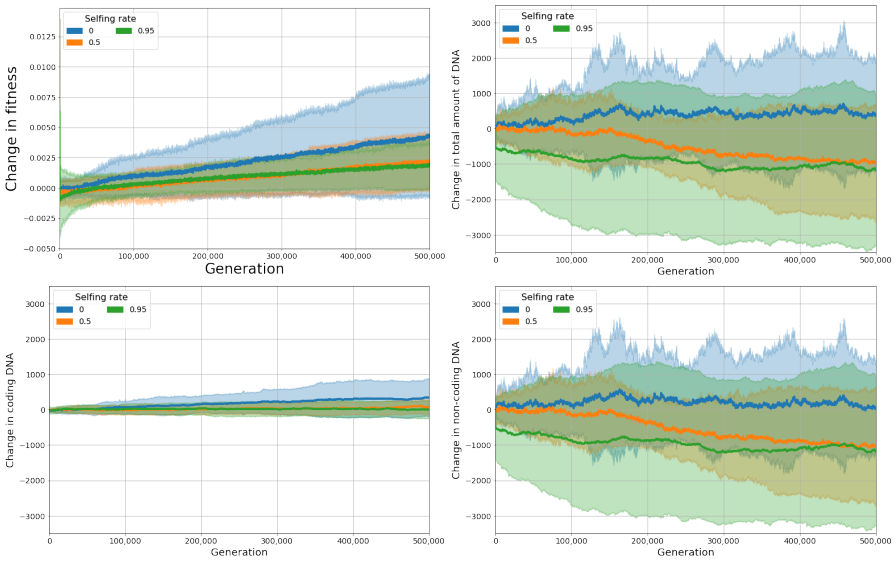

Figure F.3: Averages changes in fitness (A), genome size (B), coding size (C), and non-coding size (D) for the 3 different selfing rates (5 repetitions for each of the 9 wildtypes — WT3 excluded), measured for 500,000 generations. The colored area shows the standard deviation from the mean value.

F.3 VARIANCE WITHIN THE POPULATIONS

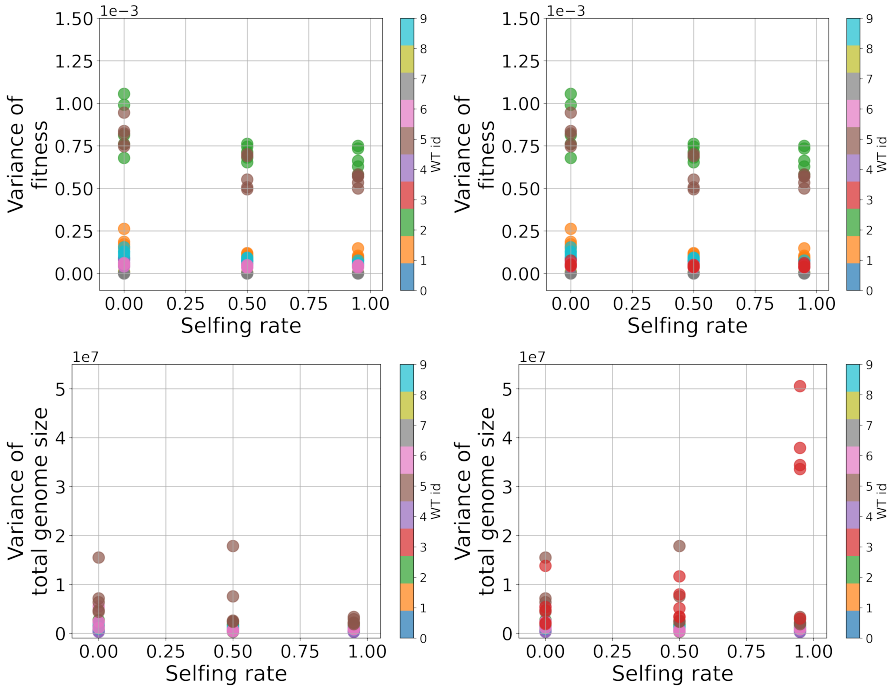

Figure F.4: Variance of fitness (top) and total genome size (bottom) within the populations, without (left) or with (right) the degenerated WT3, at generation 500,000.

F.4 VARIATIONS IN TOTAL, CODING AND NON-CODING DNA, WITH WT3 INCLUDED

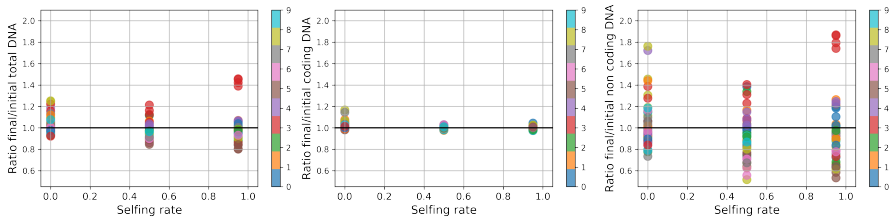

Figure F.5: Changes in total, coding and non-coding DNA for all simulations run after 500,000 generation, colour-coded for WT lineage. These data include WT3, which presents an atypical behaviour.

F.5 MANN-WHITNEY-U TESTS FOR PAIRWISE DIFFERENCES BETWEEN SELFING RATES

	0 to 0.5 selfing rates	0.5 to 0.95 selfing rates	0 to 0.95 selfing rates
Coding DNA ratio	0.0072	0.2110	**0.0003**
Non-coding DNA ratio	**0.0009**	0.6925	0.0067
Total DNA ratio	$\mathbf{2.6 \times 10^{-5}}$	0.6311	**0.0005**

Table F.1: P-values for Mann-Whitney-U tests between the different selfing rates, for data at generation 500,000 and 5 replicates for each of the 9 wild-types (excluding WT3). Differences that are significant after a Bonferroni correction are in bold.

F.6 MUTATIONAL ROBUSTNESS

Measured mutational robustness, *i.e.* ratio of fitness before and after a mutation. For each mutation type, 10,000 mutations were performed on random individuals for each of the 50 simulations. WT3 is included in these measures, which does not impact the results.

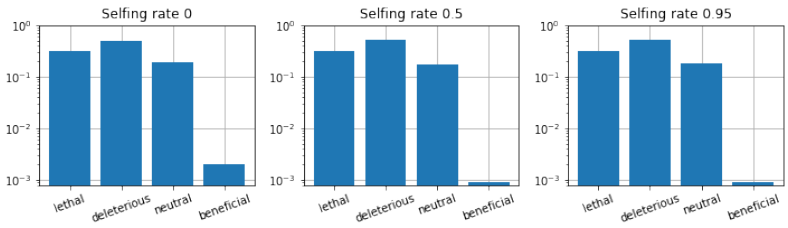

Figure F.6: Measured mutational robustness to any mutation. For each of the 50 simulations, 10,000 mutations of each type are performed on random individuals, and we compare the fitness before/after the mutation. Plotted values are the proportion of mutation landing in each of the 4 categories based on their selective coefficient s: lethal ($s \leqslant -0.999$), deleterious ($-0.999 < s \leqslant -0.001$), neutral ($-0.001 < s \leqslant 0.001$), or beneficial ($s > 0.001$).

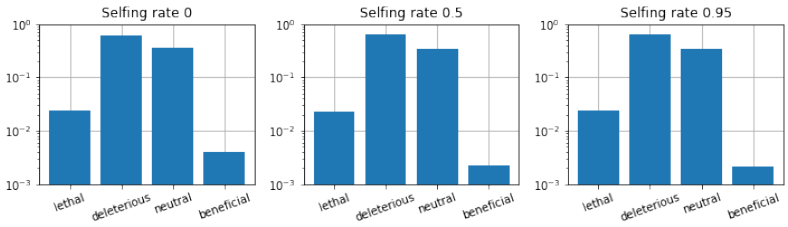

Figure F.7: Measured mutational robustness to a switch. For each of the 50 simulations, 10,000 mutations are performed on random individuals, and we compare the fitness before/after the mutation.

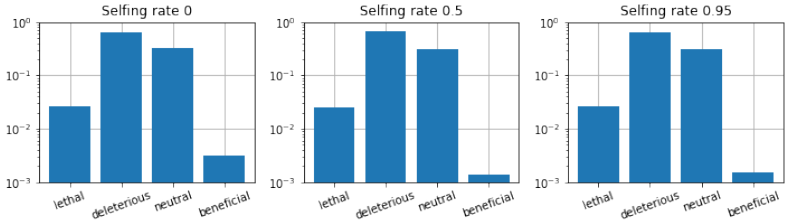

Figure F.8: Measured mutational robustness to a small insertion. For each of the 50 simulations, 10,000 mutations are performed on random individuals, and we compare the fitness before/after the mutation.

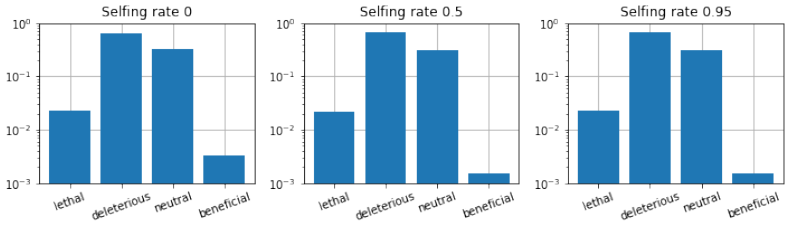

Figure F.9: Measured mutational robustness to a small deletion. For each of the 50 simulations, 10,000 mutations are performed on random individuals, and we compare the fitness before/after the mutation.

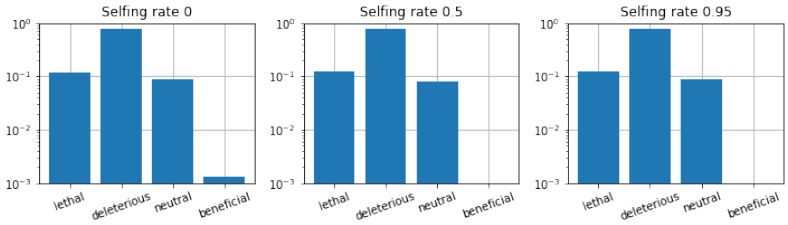

Figure F.10: Measured mutational robustness to an inversion. For each of the 50 simulations, 10,000 mutations are performed on random individuals, and we compare the fitness before/after the mutation.

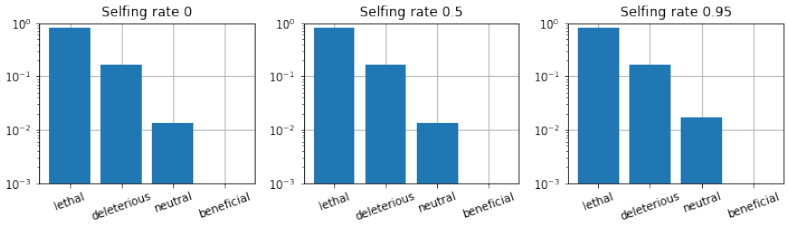

Figure F.11: Measured mutational robustness to a duplication. For each of the 50 simulations, 10,000 mutations are performed on random individuals, and we compare the fitness before/after the mutation.

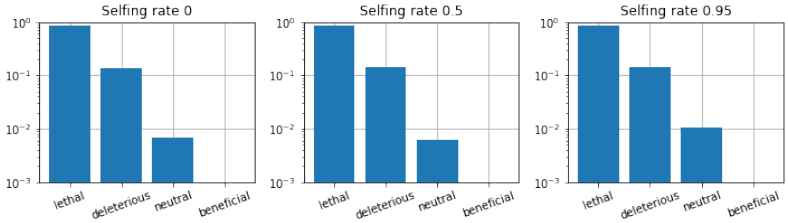

Figure F.12: Measured mutational robustness to a large deletion. For each of the 50 simulations, 10,000 mutations are performed on random individuals, and we compare the fitness before/after the mutation.

F.7 RECOMBINATION EFFICIENCIES

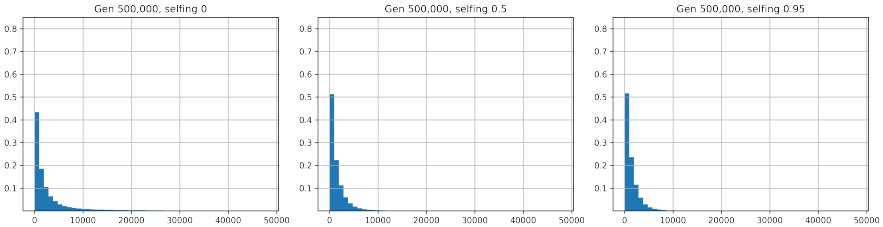

Figure F.13: Distribution of the number of tries before finding appropriate recombination points for the different selfing rates.

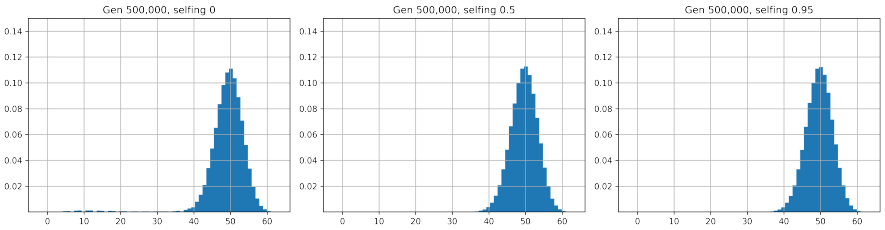

Figure F.14: Distribution of the alignment scores at the recombination points for the different selfing rates.

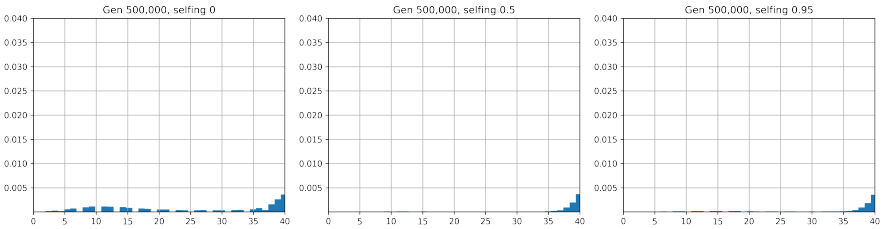

Figure F.15: Distribution of the alignment scores at the recombination points for the different selfing rates, zoomed on the lowest scores.

F.8 REPLICATIVE ROBUSTNESS

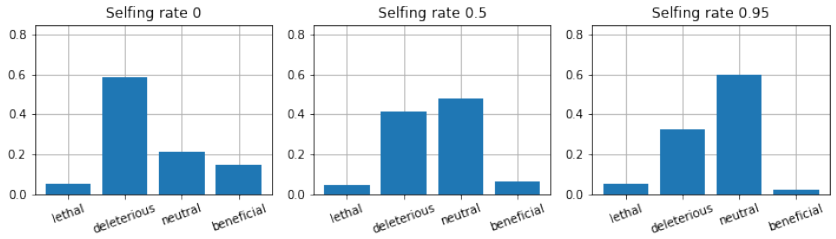

Figure F.16: Measured consequence of a replication event when comparing the fitness of the offspring to the fitness of its parents, in case of forced outcrossing (comparison with the best of both parents). The comparison is done similarly to mutational robustness: the selective coefficient of the replication event is the ratio of the fitness after and before the replication event minus 1. Plotted values are the proportion of replication landing in each of the 4 categories based on their selective coefficient s: lethal ($s \leqslant -0.999$), deleterious ($-0.999 < s \leqslant -0.001$), neutral ($-0.001 < s \leqslant 0.001$), or beneficial ($s > 0.001$).

BIBLIOGRAPHY

Adami, Christoph (2002). "What is complexity?" In: *BioEssays* 24.12, pp. 1085–1094. DOI: 10.1002/bies.10192. 2

– (2006). "Digital genetics: unravelling the genetic basis of evolution." In: *Nat Rev Genet* 7.2, pp. 109–118. DOI: 10.1038/nrg1771. 13, 46

Agranat-Tamir, Lily, Jazlyn A Mooney, and Noah A Rosenberg (Jan. 2024). "Counting the genetic ancestors from source populations in members of an admixed population." In: *Genetics* 226.4, iyae011. DOI: 10.1093/genetics/iyae011. 178, 179

Ahnert, Sebastian E., Thomas M. A. Fink, and Andrei Zinovyev (2008). "How much non-coding DNA do eukaryotes require?" In: *Journal of Theoretical Biology* 252.4, pp. 587–592. DOI: 10.1016/j.jtbi.2008.02.005. 7, 68

Ahsan, Mian Umair, Qian Liu, Jonathan Elliot Perdomo, Li Fang, and Kai Wang (2023). "A survey of algorithms for the detection of genomic structural variants from long-read sequencing data." In: *Nature methods* 20.8, pp. 1143–1158. DOI: 10.1038/s41592-023-01932-w. 12

Ai, Bin, Zhao-Shan Wang, and Song Ge (2012). "Genome size is not correlated with effective population size in the *Oryza* species." In: *Evolution* 66.10, pp. 3302–3310. DOI: 10.1111/j.1558-5646.2012.01674.x. 11, 69, 83

Alkan, Can, Bradley P. Coe, and Evan E. Eichler (2011). "Genome structural variation discovery and genotyping." In: *Nature Reviews Genetics* 12.5, pp. 363–376. DOI: 10.1038/nrg2958. 20

Almpanis, Apostolos, Martin Swain, Derek Gatherer, and Neil McEwan (2018). "Correlation between bacterial G+C content, genome size and the G+C content of associated plasmids and bacteriophages." In: *Microbial Genomics* 4.4, e000168. DOI: 10.1099/mgen.0.000168. 44

Andersson, A. I. Nilssonand S. Koskiniemiand S. Erikssonand E. Kugelbergand J. C. D. Hintonand D. I. (2005). "Bacterial genome size reduction by experimental evolution." In: *Proceedings of the National Academy of Sciences* 102.34, pp. 12112–12116. DOI: 10.1073/pnas.0503654102. 45

Arbeithuber, Barbara, Andrea J Betancourt, Thomas Ebner, and Irene Tiemann-Boege (2015). "Crossovers are associated with mutation and biased gene conversion at recombination hotspots." In: *Proceedings of the National Academy of Sciences* 112.7, pp. 2109–2114. DOI: 10.1073/pnas.141662211. 9

5 Ardlie, Kristin G, Leonid Kruglyak, and Mark Seielstad (2002). "Patterns of linkage disequilibrium in the human genome." In: *Nature Reviews Genetics* 3.4, pp. 299–309. DOI: 10.1038/nrg777.

20, 143 Audrézet, Marie-Pierre, Jian-Min Chen, Odile Raguénes, Nadia Chuzhanova, Karine Giteau, Cédric Le Maréchal, Isabelle Quéré, David N Cooper, and Claude Férec (2004). "Genomic rearrangements in the CFTR gene: extensive allelic heterogeneity and diverse mutational mechanisms." In: *Human mutation* 23.4, pp. 343–357. DOI: 10.1002/humu.20009.

12 Augustijnen, Hannah, Cristina Arias-Sardá, Marta Farré, and Kay Lucek (2024). "A Genomic Update on the Evolutionary Impact of Chromosomal Rearrangements." In: *Molecular Ecology* 33.24, e17602. DOI: 10.1111/mec.17602.

44 Bachmann, Konrad (1972). "Genome size in mammals." In: *Chromosoma* 37.1, pp. 85–93. DOI: 10.1007/BF00329560.

178 Baird, S.J.E., N.H. Barton, and A.M. Etheridge (2003). "The distribution of surviving blocks of an ancestral genome." In: *Theoretical Population Biology* 64.4, pp. 451–471. DOI: 0.1016/S0040-5809(03)00098-4.

68 Bales, Alex L. and Erika I. Hersch-Green (2019). "Effects of soil nitrogen on diploid advantage in fireweed, Chamerion angustifolium (Onagraceae)." In: *Ecology and Evolution* 9.3, pp. 1095–1109. DOI: 10.1002/ece3.4797.

35 Bank, Claudia, Ryan T Hietpas, Jeffrey D Jensen, and Daniel NA Bolon (2015). "A systematic survey of an intragenic epistatic landscape." In: *Molecular biology and evolution* 32.1, pp. 229–238. DOI: 10.1093/molbev/msu301.

14 Banse, Paul (2023). "Evolution beyond substitutions: Computational modeling of the impact of chromosomal rearrangements on evolutionary dynamics." PhD thesis. INSA de Lyon.

1, 128 Banse, Paul, Santiago F Elena, and Guillaume Beslon (2024a). "Innovation in viruses: fitness valley crossing, neutral landscapes, or just duplications?" In: *Virus Evolution* 10.1, veae078. DOI: 10.1093/ve/veae078.

19, 46, 47, 53, 60, 69, 78, 113, 124 Banse, Paul, Juliette Luiselli, David P. Parsons, Théotime Grohens, Marco Foley, Leonardo Trujillo, Jonathan Rouzaud-Cornabas, Carole Knibbe, and Guillaume Beslon (2024b). "Forward-in-time simulation of chromosomal rearrangements: The invisible backbone that sustains long-term adaptation." In: *Molecular Ecology* 33.24, e17234. DOI: 10.1111/mec.17234.

44 Barow, Martin and Armin Meister (2002). "Lack of correlation between AT frequency and genome size in higher plants and the effect of nonrandomness of base sequences on dye binding." In: *Cytometry* 47.1, pp. 1–7. DOI: 10.1002/cyto.10030.

112 Barrett, Spencer CH (2002). "The evolution of plant sexual diversity." In: *Nature reviews genetics* 3.4, pp. 274–284. DOI: 10.1038/nrg776.

Bartenhagen, Christoph and Martin Dugas (2013). "RSVSim: an R/Bioconductor package for the simulation of structural variations." In: *Bioinformatics* 29.13, pp. 1679–1681. DOI: 10.1093/bioinformatics/btt198. 13

Batut, Bérénice, Carole Knibbe, Gabriel Marais, and Vincent Daubin (2014). "Reductive genome evolution at both ends of the bacterial population size spectrum." In: *Nat Rev Microbiol* 12.12, pp. 841–850. DOI: 10.1038/nrmicro3331. 44, 45, 56

Batut, Bérénice, David P. Parsons, Stephan Fischer, Guillaume Beslon, and Carole Knibbe (2013). "In silico experimental evolution: a tool to test evolutionary scenarios." In: *BMC Bioinformatics* 14.15, S11. DOI: 10.1186/1471-2105-14-S15-S11. 22, 23, 46, 139, 145, 147

Beaulieu, Jeremy M, Angela T Moles, Ilia J Leitch, Michael D Bennett, John B Dickie, and Charles A Knight (2007). "Correlated evolution of genome size and seed mass." In: *New Phytologist* 173.2, p. 422. DOI: 10.1111/j.1469-8137.2006.01919.x. 112

Beier, Sara, Johannes Werner, Thierry Bouvier, Nicolas Mouquet, and Cyrille Violle (2022). "Trait-trait relationships and tradeoffs vary with genome size in prokaryotes." In: *Frontiers in Microbiology* 13, p. 985216. DOI: 10.3389/fmicb.2022.985216. 112

Berdan, Emma L, Alexandre Blanckaert, Roger K Butlin, and Claudia Bank (2021a). "Deleterious mutation accumulation and the long-term fate of chromosomal inversions." In: *PLoS genetics* 17.3, e1009411. DOI: 10.1371/journal.pgen.1009411. 39, 40

Berdan, Emma L, Alexandre Blanckaert, Tanja Slotte, Alexander Suh, Anja M Westram, and Inês Fragata (2021b). "Unboxing mutations: Connecting mutation types with evolutionary consequences." In: *Molecular Ecology* 30.12, pp. 2710–2723. DOI: 10.1111/mec.15936. 35, 37

Beslon, G., D.P. Parsons, Y. Sanchez-Dehesa, J.-M. Peña, and C. Knibbe (2010). "Scaling laws in bacterial genomes: A side-effect of selection of mutational robustness?" In: *Biosystems* 102.1, pp. 32–40. DOI: 10.1016/j.biosystems.2010.07.009. 23, 139

Bhatia, Sangeeta, Pedro Feijão, and Andrew R Francis (2018). "Position and content paradigms in genome rearrangements: the wild and crazy world of permutations in genomics." In: *Bulletin of Mathematical Biology* 80, pp. 3227–3246. DOI: 10.1007/s11538-018-0514-3. 20, 21

Bickmore, Wendy A and Bas Van Steensel (2013). "Genome architecture: domain organization of interphase chromosomes." In: *Cell* 152.6, pp. 1270–1284. DOI: 10.1016/j.cell.2013.02.001. 7

Blakely, Emma L. et al. (2006). "Sporadic Intragenic Inversion of the Mitochondrial DNA MTND1 Gene Causing Fatal Infantile Lactic Acidosis." In: *Pediatric Research* 59.3, pp. 440–444. DOI: 10.1203/01.pdr.0000198771.78290.c4. 143

Blommaert, Julie (2020). "Genome size evolution: towards new model systems for old questions." In: *Proceedings of the Royal Society B:* 68, 112

Biological Sciences 287.1933, p. 20201441. DOI: 10.1098/rspb.2020.1441.

20, 35 Blount, Zachary D., Jeffrey E. Barrick, Carla J. Davidson, and Richard E. Lenski (2012). "Genomic analysis of a key innovation in an experimental Escherichia coli population." In: *Nature* 489.7417, pp. 513–518. DOI: 10.1038/nature11514.

11 Bobay, Louis-Marie and Howard Ochman (2017). "The evolution of bacterial genome architecture." In: *Frontiers in genetics* 8, p. 72. DOI: 10.3389/fgene.2017.00072.

59 Bock, Ralph (2017). "Witnessing Genome Evolution: Experimental Reconstruction of Endosymbiotic and Horizontal Gene Transfer." In: *Annual Review of Genetics* 51, pp. 1–22. DOI: 10.1146/annurev-genet-120215-035329.

57 Boer, Folkert K. de and Paulien Hogeweg (2010). "Eco-evolutionary dynamics, coding structure and the information threshold." In: *BMC Evolutionary Biology* 10.1, p. 361. DOI: 10.1186/1471-2148-10-361.

45 Bourguignon, Thomas, Yukihiro Kinjo, Paula Villa-Martin, Nicholas V Coleman, Qian Tang, Daej A Arab, Zongqing Wang, Gaku Tokuda, Yuichi Hongoh, Moriya Ohkuma, et al. (2020). "Increased mutation rate is linked to genome reduction in prokaryotes." In: *Current Biology* 30.19, pp. 3848–3855. DOI: 10.1016/j.cub.2020.07.034.

4, 45, 133 Brevet, Mathieu and Nicolas Lartillot (2021). "Reconstructing the History of Variation in Effective Population Size along Phylogenies." In: *Genome Biology and Evolution* 13.8. DOI: 10.1093/gbe/evab150.

5 Bridges, Calvin B (1916). "Non-disjunction as proof of the chromosome theory of heredity (concluded)." In: *Genetics* 1.2, p. 107. DOI: 10.1093/genetics/1.2.107.

5 – (1921). "Genetical and cytological proof of non-disjunction of the fourth chromosome of Drosophila melanogaster." In: *Proceedings of the National Academy of Sciences* 7.7, pp. 186–192. DOI: 10.1073/pnas.7.7.186.

177, 178 Brunet, Éric and Bernard Derrida (2013). "Genealogies in simple models of evolution." In: *Journal of Statistical Mechanics: Theory and Experiment* 2013.01, P01006. DOI: 10.1088/1742-5468/2013/01/P01006.

2 Buffalo, Vince (2021). "Quantifying the relationship between genetic diversity and population size suggests natural selection cannot explain Lewontin's Paradox." In: *Elife* 10, e67509. DOI: 10.7554/eLife.67509.

112 Bureš, Petr et al. (2024). "The global distribution of angiosperm genome size is shaped by climate." In: *New Phytologist* 242.2, pp. 744–759. DOI: 10.1111/nph.19544.

117 Caballero, A and W G Hill (1992). "Effective size of nonrandom mating populations." In: *Genetics* 130.4, pp. 909–916. DOI: 10.1093/genetics/130.4.909.

20 Campo, N., M. J. Dias, M. L. Daveran-Mingot, P. Ritzenthaler, and P. Le Bourgeois (2004). "Chromosomal constraints in Gram-positive bacte-

ria revealed by artificial inversions: Experimental genome inversions in Streptococcaceae." In: *Molecular Microbiology* 51.2, pp. 511–522. DOI: 10.1046/j.1365-2958.2003.03847.x.

Canapa, Adriana, Marco Barucca, Maria A. Biscotti, Mariko Forconi, and Ettore Olmo (2015). "Transposons, Genome Size, and Evolutionary Insights in Animals." In: *Cytogenetic and Genome Research* 147.4, pp. 217–239. DOI: 10.1159/000444429. 69

Cang, F. Alice, Shana R. Welles, Jenny Wong, Maia Ziaee, and Katrina M. Dlugosch (2023). "Genome size variation and evolution during invasive range expansion in an introduced plant." In: *Evolutionary Applications* 17.1, e13624. DOI: 10.1111/eva.13624. 112

Cao, Liang, Eric Alani, and Nancy Kleckner (1990). "A pathway for generation and processing of double-strand breaks during meiotic recombination in S. cerevisiae." In: *Cell* 61.6, pp. 1089–1101. DOI: 10.1016/0092-8674(90)90072-M. 9

Cao, Sha, Gerrit Brandis, Douglas L. Huseby, and Diarmaid Hughes (2022). "Positive Selection during Niche Adaptation Results in Large-Scale and Irreversible Rearrangement of Chromosomal Gene Order in Bacteria." In: *Molecular Biology and Evolution* 39.4. DOI: 10.1093/molbev/msac069. 20

Chapman, N.H and E.A Thompson (2003). "A model for the length of tracts of identity by descent in finite random mating populations." In: *Theoretical Population Biology* 64.2, pp. 141–150. DOI: 10.1016/S0040-5809(03)00071-6. 178

Charlesworth, Brian and Deborah Charlesworth (2017). "Population genetics from 1966 to 2016." In: *Heredity* 118.1, pp. 2–9. DOI: 10.1038/hdy.2016.55. 1

Charlesworth, Brian and Jeffrey D Jensen (2022). "How Can We Resolve Lewontin's Paradox?" In: *Genome Biology and Evolution* 14.7, evac096. DOI: 10.1093/gbe/evac096. 2

Charlesworth, D. and B. Charlesworth (1987). "Inbreeding depression and its evolutionary consequences." In: *Annual Review of Ecology and Systematics* 18, 237–268. DOI: 10.1146/annurev.ecolsys.18.1.237. 112, 122

Chen, Gary K, Paul Marjoram, and Jeffrey D Wall (2009). "Fast and flexible simulation of DNA sequence data." In: *Genome research* 19.1, pp. 136–142. DOI: 10.1101/gr.083634.108. 179

Chen, Jun, Sylvain Glémin, and Martin Lascoux (2017). "Genetic Diversity and the Efficacy of Purifying Selection across Plant and Animal Species." In: *Molecular Biology and Evolution* 34.6, pp. 1417–1428. DOI: 10.1093/molbev/msx088. 112

Chen, Xingyu, Siyu Wei, Chen Sun, Zelin Yi, Zihan Wang, Yingyi Wu, Jing Xu, Junxian Tao, Haiyan Chen, Mingming Zhang, et al. (2025). "Computational Tools for Studying Genome Structural Variation." In: *OMICS: A Journal of Integrative Biology*. DOI: 10.1089/omi.2024.0200. 12

Chen, Zhuoyu, Xiaojun Wang, Yu Song, Qinglu Zeng, Yao Zhang, and Haiwei Luo (2022). "Prochlorococcus have low global mutation rate 45, 56

and small effective population size." In: *Nat Ecol Evol* 6.2, pp. 183–194. DOI: 10.1038/s41559-021-01591-0.

9 Chiyomaru, Katsumi and Kazuhiro Takemoto (2020). "Revisiting the hypothesis of an energetic barrier to genome complexity between eukaryotes and prokaryotes." In: *Royal Society Open Science* 7.2, p. 191859. DOI: 10.1098/rsos.191859.

44 Choi, Ik-Young, Eun-Chae Kwon, and Nam-Soo Kim (2020). "The C- and G-value paradox with polyploidy, repeatomes, introns, phenomes and cell economy." In: *Genes & Genomics* 42.7, pp. 699–714. DOI: 10.1007/s13258-020-00941-9.

44 Chong, Rebecca A, Hyunjin Park, and Nancy A Moran (2019). "Genome Evolution of the Obligate Endosymbiont Buchnera aphidicola." In: *Molecular Biology and Evolution* 36.7, pp. 1481–1489. DOI: 10.1093/molbev/msz082.

121 Clo, Josselin, Joëlle Ronfort, and Diala Abu Awad (2020). "Hidden genetic variance contributes to increase the short-term adaptive potential of selfing populations." In: *Journal of Evolutionary Biology* 33.9, pp. 1203–1215. DOI: 10.1111/jeb.13660.

6 Coissac, Eric, Tiayyba Riaz, and Nicolas Puillandre (2012). "Bioinformatic challenges for DNA metabarcoding of plants and animals." In: *Molecular ecology* 21.8, pp. 1834–1847. DOI: 10.1111/j.1365-294X.2012.05550.x.

20 Connallon, Tim and Colin Olito (2022). "Natural selection and the distribution of chromosomal inversion lengths." In: *Molecular Ecology* 31.13, pp. 3627–3641. DOI: 10.1111/mec.16091.

1 Cook, Laurence M, Bruce S Grant, Ilik J Saccheri, and Jim Mallet (2012). "Selective bird predation on the peppered moth: the last experiment of Michael Majerus." In: *Biology letters* 8.4, pp. 609–612. DOI: 10.1098/rsbl.2011.1136.

8, 9 Craig, Jack M., Sudhir Kumar, and S. Blair Hedges (2023). "The origin of eukaryotes and rise in complexity were synchronous with the rise in oxygen." In: *Frontiers in Bioinformatics* 3, p. 1233281. DOI: 10.3389/fbinf.2023.1233281.

112 Cutter, Asher D. (2019). "Reproductive transitions in plants and animals: selfing syndrome, sexual selection and speciation." In: *New Phytologist* 224.3, pp. 1080–1094. DOI: 10.1111/nph.16075.

20, 21, 39 Darling, Aaron E., István Miklós, and Mark A. Ragan (2008). "Dynamics of Genome Rearrangement in Bacterial Populations." In: *PLOS Genetics* 4.7, e1000128. DOI: 10.1371/journal.pgen.1000128.

178 Davies, Joanna L., František Simančík, Rune Lyngsø, Thomas Mailund, and Jotun Hein (Dec. 2007). "On Recombination-Induced Multiple and Simultaneous Coalescent Events." In: *Genetics* 177.4, pp. 2151–2160. DOI: 10.1534/genetics.107.071126. (Visited on 06/11/2024).

6 Deiner, Kristy, Holly M Bik, Elvira Mächler, Mathew Seymour, Anaïs Lacoursière-Roussel, Florian Altermatt, Simon Creer, Iliana Bista, David M Lodge, Natasha De Vere, et al. (2017). "Environmental

DNA metabarcoding: Transforming how we survey animal and plant communities." In: *Molecular ecology* 26.21, pp. 5872–5895. DOI: 10.1111/mec.14350.

Derrida, B. and B. Jung-Muller (Feb. 1999). "The Genealogical Tree of a Chromosome." In: *Journal of Statistical Physics* 94.3, pp. 277–298. DOI: 10.1023/A:1004579800589. 178, 186

Derrida, Bernard, Susanna C. Manrubia, and Damián H. Zanette (2000). "On the genealogy of a population of biparental individuals." In: *Journal of theoretical biology* 203 3, pp. 303–15. DOI: 10.1006/jtbi.2000.1095. 177, 178

Dijk, Bram van, Frederic Bertels, Lianne Stolk, Nobuto Takeuchi, and Paul B Rainey (2022). "Transposable elements promote the evolution of genome streamlining." In: *Philosophical Transactions of the Royal Society B* 377.1842, p. 20200477. DOI: 10.1098/rstb.2020.0477. 59

Diss, Guillaume and Ben Lehner (2018). "The genetic landscape of a physical interaction." In: *Elife* 7, e32472. DOI: 10.7554/eLife.32472. 35

Dobzhansky, Th and Alfred H Sturtevant (1938). "Inversions in the chromosomes of Drosophila pseudoobscura." In: *Genetics* 23.1, p. 28. DOI: 10.1093/genetics/23.1.28. 5

Dobzhansky, Theodosius (1930). "Translocations involving the third and the fourth chromosomes of Drosophila melanogaster." In: *Genetics* 15.4, p. 347. DOI: 10.1093/genetics/15.4.347. 5

Doolittle, W Ford (2013). "Is junk DNA bunk? A critique of ENCODE." In: *Proceedings of the National Academy of Sciences* 110.14, pp. 5294–5300. DOI: 10.1073/pnas.1221376110. 7, 68

Drake, John W (1991). "A constant rate of spontaneous mutation in DNA-based microbes." In: *Proceedings of the National Academy of Sciences* 88.16, pp. 7160–7164. DOI: 10.1073/pnas.88.16.7160. 38

Dufresne, Alexis, Laurence Garczarek, and Frédéric Partensky (2005). "Accelerated evolution associated with genome reduction in a free-living prokaryote." In: *Genome Biology* 6.2, R14. DOI: 10.1186/gb-2005-6-2-r14. 44, 56

Eigen, Manfred (1971). "Selforganization of matter and the evolution of biological macromolecules." In: *Naturwissenschaften* 58.10, pp. 465–523. DOI: 10.1007/BF00623322. 57

Eisen, Jonathan A, John F Heidelberg, Owen White, and Steven L Salzberg (2000). "Evidence for symmetric chromosomal inversions around the replication origin in bacteria." In: *Genome biology* 1, pp. 1–9. DOI: 10.1186/gb-2000-1-6-research0011. 12

El Houdaigui, Bilal, Raphaël Forquet, Thomas Hindré, Dominique Schneider, William Nasser, Sylvie Reverchon, and Sam Meyer (2019). "Bacterial genome architecture shapes global transcriptional regulation by DNA supercoiling." In: *Nucleic acids research* 47.11, pp. 5648–5657. DOI: 10.1093/nar/gkz300. 68

Elena, Santiago F., Claus O. Wilke, Charles Ofria, and Richard E. Lenski (2007). "Effects of population size and mutation rate on the 112

evolution of mutational robustness." In: *Evolution* 61.3, pp. 666–674. DOI: 10.1111/j.1558-5646.2007.00064.x.

7 Elliott, Tyler A and T Ryan Gregory (2015). "What's in a genome? The C-value enigma and the evolution of eukaryotic genome content." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1678, p. 20140331. DOI: 10.1098/rstb.2014.0331.

179 Ewing, Gregory and Joachim Hermisson (June 2010). "MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus." In: *Bioinformatics* 26.16, pp. 2064–2065. DOI: 10.1093/bioinformatics/btq322.

179 Excoffier, Laurent and Matthieu Foll (Mar. 2011). "fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios." In: *Bioinformatics* 27.9, pp. 1332–1334. DOI: 10.1093/bioinformatics/btr124.

7, 68 Fagundes, Nelson JR, Rafael Bisso-Machado, Pedro ICC Figueiredo, Maikel Varal, and André LS Zani (2022). "What we talk about when we talk about "junk DNA"." In: *Genome Biology and Evolution* 14.5, evac055. DOI: 10.1093/gbe/evac055.

83 Fang, Bohao and Scott V Edwards (2024). "Fitness consequences of structural variation inferred from a House Finch pangenome." In: *Proceedings of the National Academy of Sciences* 121.47, e2409943121. DOI: 10.1073/pnas.2409943121.

112 Fierst, Janna L., John H. Willis, Cristel G. Thomas, Wei Wang, Rose M. Reynolds, Timothy E. Ahearne, Asher D. Cutter, and Patrick C. Phillips (2015). "Reproductive Mode and the Evolution of Genome Size and Structure in Caenorhabditis Nematodes." In: *PLOS Genetics* 11.6. Ed. by Mark Blaxter, e1005323. DOI: 10.1371/journal.pgen.1005323.

46 Figuet, Emeric, Benoît Nabholz, Manon Bonneau, Eduard Mas Carrio, Krystyna Nadachowska-Brzyska, Hans Ellegren, and Nicolas Galtier (2016). "Life History Traits, Protein Evolution, and the Nearly Neutral Theory in Amniotes." In: *Molecular Biology and Evolution* 33.6, pp. 1517–1527. DOI: 10.1093/molbev/msw033.

45, 56 Filatov, Dmitry A. and Mark Kirkpatrick (2024). "How does evolution work in superabundant microbes?" In: *Trends in Microbiology* 32.9, pp. 836–846. DOI: 10.1016/j.tim.2024.01.009.

4, 75, 181 Fisher, R. A. (1923). "XXI.—On the Dominance Ratio." In: *Proceedings of the Royal Society of Edinburgh* 42, 321–341. DOI: 10.1017/S0370164600023993.

112 Franco, Ana Luiza, Ana Luisa Sousa Azevedo, Aryane Campos Reis, Elyabe Monteiro Matos, Marina Arantes Fonseca, Antônio Vander Pereira, Ilia J. Leitch, Andrew R. Leitch, and Lyderson Facio Viccini (2024). "Uncovering the genomic diversity of the wild forage crop Setaria sphacelata." In: *Crop Science* 64.6, pp. 3381–3398. DOI: 10.1002/csc2.21344.

Freeling, Michael, Jie Xu, Margaret Woodhouse, and Damon Lisch (2015). "A solution to the C-value paradox and the function of junk DNA: the genome balance hypothesis." In: *Molecular Plant* 8.6, pp. 899–910. DOI: 10.1016/j.molp.2015.02.009. 68

Frenoy, Antoine, FranÃ§ois Taddei, and Dusan Misevic (2013). "Genetic Architecture Promotes the Evolution and Maintenance of Cooperation." In: *PLoS Computational Biology* 9.11, e1003339. DOI: 10.1371/journal.pcbi.1003339. 23, 139

Gabzi, Tzahi, Yitzhak Pilpel, and Tamar Friedlander (2022). "Fitness Landscape Analysis of a tRNA Gene Reveals that the Wild Type Allele is Sub-optimal, Yet Mutationally Robust." In: *Molecular Biology and Evolution* 39.9. DOI: 10.1093/molbev/msac178. 56

Galtier, Nicolas (2024). "Half a Century of Controversy: The Neutralist/Selectionist Debate in Molecular Evolution." In: *Genome Biology and Evolution* 16.2, evae003. DOI: 10.1093/gbe/evae003. 10

Gao, Yuxia, Huayao Zhao, Yin Jin, Xiaoyu Xu, and Guan-Zhu Han (2017). "Extent and evolution of gene duplication in DNA viruses." In: *Virus research* 240, pp. 161–165. DOI: 10.1016/j.virusres.2017.08.005. 20

Gerlee, P. and T. Lundh (2008). "The Emergence of Overlapping Scale-free Genetic Architecture in Digital Organisms." In: *Artificial Life* 14.3, pp. 265–275. DOI: 10.1162/artl.2008.14.3.14303. 13

Gervais, Camille and Denis Roze (2017). "Mutation Rate Evolution in Partially Selfing and Partially Asexual Organisms." In: *Genetics* 207.4, pp. 1561–1575. DOI: 10.1534/genetics.117.300346. 113, 122

Gil, Rosario and Amparo Latorre (2012). "Factors Behind Junk DNA in Bacteria." In: *Genes* 3.4, pp. 634–650. DOI: 10.3390/genes3040634. 7, 68

Giovannoni, Stephen J., J. Cameron Thrash, and Ben Temperton (2014). "Implications of streamlining theory for microbial ecology." In: *ISME J* 8.8, pp. 1553–1565. DOI: 10.1038/ismej.2014.60. 45, 56

Glémin, S. (Dec. 2003). "How are deleterious mutations purged? Drift versus nonrandom mating." In: *Evolution* 57, pp. 2678–2687. DOI: 10.1554/03-406. 123

González, Raquel, Joan Vallès, and Teresa Garnatje (2022). "Genome Size Variation Assessment in Vitis vinifera L. Landraces in Ibiza and Formentera (Balearic Islands)." In: *Plants* 11.14, p. 1892. DOI: 10.3390/plants11141892. 112

Gravel, Simon and Mike Steel (2015). "The existence and abundance of ghost ancestors in biparental populations." In: *Theoretical Population Biology* 101, pp. 47–53. DOI: 0.1016/j.tpb.2015.02.002. 178, 182, 190, 192, 194

Guan, Peiyong and Wing-Kin Sung (2016). "Structural variation detection using next-generation sequencing data: A comparative technical review." In: *Methods* 102. Pan-omics analysis of biological data, pp. 36–49. DOI: 10.1016/j.ymeth.2016.01.020. 12

Gunbin, Konstantin Popadinand Leonard V. Polishchukand Leila Mamirovaand Dmitry Knorreand Konstantin (2007). "Accumulation 46

of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals." In: *Proceedings of the National Academy of Sciences* 104.33, pp. 13390–13395. DOI: 10.1073/pnas.0701256104.

13 Gupta, Aditi, Thomas LaBar, Miriam Miyagi, and Christoph Adami (2016). "Evolution of genome size in asexual digital organisms." In: *Scientific reports* 6.1, p. 25786. DOI: 10.1038/srep25786.

2 Hahn, Matthew W. and Gregory A. Wray (2002). "The g-value paradox." In: *Evolution & Development* 4.2, pp. 73–75. DOI: 10.1046/j.1525-142X.2002.01069.x.

5 Haldane, John Burdon Sanderson (1936). "A provisional map of a human chromosome." In: *Nature* 137.3462, pp. 398–400. DOI: 10.1038/137398b0.

21 Haller, Benjamin C and Philipp W Messer (2017). "SLiM 2: flexible, interactive forward genetic simulations." In: *Molecular biology and evolution* 34.1, pp. 230–240. DOI: 10.1093/molbev/msw211.

13, 194 Haller, Benjamin C and Philipp W Messer (2023). "SLiM 4: multispecies eco-evolutionary modeling." In: *The American Naturalist* 201.5, E127–E139. DOI: 10.1073/pnas.74.12.5463.

14 Haller, Benjamin C. and Philipp W. Messer (2024). *SLiM: An Evolutionary Simulation Framework*. URL: http://benhaller.com/slim/SLiM_Manual.pdf.

45 Ham, R. van et al. (2003). "Reductive genome evolution in Buchnera aphidicola." In: *Proceedings of the National Academy of Sciences* 100.2, pp. 581–586. DOI: 10.1073/pnas.0235981100.

20 Hanlon, Vincent C. T., Peter M. Lansdorp, and Victor Guryev (2022). "A survey of current methods to detect and genotype inversions." In: *Human Mutation* 43.11, pp. 1576–1589. DOI: 10.1002/humu.24458.

177 Hartfield, Matthew, Stephen I. Wright, and Aneil F. Agrawal (Jan. 2016). "Coalescent Times and Patterns of Genetic Diversity in Species with Facultative Sex: Effects of Gene Conversion, Population Structure, and Heterogeneity." In: *Genetics* 202.1, pp. 297–312. DOI: 10.1534/genetics.115.178004.

10, 81 He, Yang, Suyan Tian, and Pu Tian (2019). "Fundamental asymmetry of insertions and deletions in genomes size evolution." In: *Journal of Theoretical Biology* 482, p. 109983. DOI: 10.1016/j.jtbi.2019.08.014.

45 Heddi, Abdelaziz, Hubert Charles, Chaqué Khatchadourian, Guy Bonnot, and Paul Nardon (1998). "Molecular Characterization of the Principal Symbiotic Bacteria of the Weevil Sitophilus oryzae: A Peculiar G + C Content of an Endocytobiotic DNA." In: *J Mol Evol* 47.1, pp. 52–61. DOI: 10.1007/PL00006362.

177 Hein, Jotun, Mikkel Schierup, and Carsten Wiuf (2004). *Gene genealogies, variation and evolution: a primer in coalescent theory*. Oxford University Press, USA.

46 Hindré, Thomas, Carole Knibbe, Guillaume Beslon, and Dominique Schneider (2012). "New insights into bacterial adaptation through

in vivo and in silico experimental evolution." In: *Nature Reviews Microbiology* 10.5, pp. 352–365. DOI: 10.1038/nrmicro2750.

Hirabayashi, Kaede and Gregory L Owens (2023). "The rate of chromosomal inversion fixation in plant genomes is highly variable." In: *Evolution* 77.4, pp. 1117–1130. DOI: 10.1093/evolut/qpad027. 84

Hirsh, Guy Sellaand Aaron E. (2005). "The application of statistical physics to evolutionary biology." In: *Proceedings of the National Academy of Sciences* 102.27, pp. 9541–9546. DOI: 10.1073/pnas.0501865102. 75, 84

Ho, Steve S, Alexander E Urban, and Ryan E Mills (2020). "Structural variation in the sequencing era." In: *Nature Reviews Genetics* 21.3, pp. 171–189. DOI: 10.1038/s41576-019-0180-9. 12, 20, 80

Hoban, Sean, Giorgio Bertorelle, and Oscar E Gaggiotti (2012). "Computer simulations: tools for population and evolutionary genetics." In: *Nature Reviews Genetics* 13.2, pp. 110–122. DOI: 10.1038/nrg3130. 179

Hoffmann, Ary A., Carla M. Sgrò, and Andrews R. Weeks (2004). "Chromosomal inversion polymorphisms and adaptation." In: *Trends in Ecology & Evolution* 19.9, pp. 482–488. DOI: 10.1016/j.tree.2004.06.013. 20

Holder, Isabelle T and Jörg S Hartig (2014). "A matter of location: influence of G-quadruplexes on Escherichia coli gene expression." In: *Chemistry & biology* 21.11, pp. 1511–1521. DOI: 10.1016/j.chembiol.2014.09.014. 5

Hu, Jinghua and Jeffrey L. Blanchard (2008). "Environmental Sequence Data from the Sargasso Sea Reveal That the Characteristics of Genome Reduction in Prochlorococcus Are Not a Harbinger for an Escalation in Genetic Drift." In: *Molecular Biology and Evolution* 26.1, pp. 5–13. DOI: 10.1093/molbev/msn217. 56

Hu, Tina T. et al. (2011). "The Arabidopsis lyrata genome sequence and the basis of rapid genome size change." In: *Nature Genetics* 43.5, pp. 476–481. DOI: 10.1038/ng.807. 112, 113, 120, 121

Hudson, Richard R (1983). "Properties of a neutral allele model with intragenic recombination." In: *Theoretical population biology* 23.2, pp. 183–201. DOI: 10.1016/0040-5809(83)90013-8. 178

– (2002). "Generating samples under a Wright–Fisher neutral model of genetic variation." In: *Bioinformatics* 18.2, pp. 337–338. DOI: 10.1093/bioinformatics/18.2.337. 179

IHGSC, International Human Genome Sequencing Consortium (2001). "Initial sequencing and analysis of the human genome." In: *nature* 409.6822, pp. 860–921. DOI: 10.1038/35057062. 6, 7

Jarne, Philippe and Josh R Auld (2006). "Animals mix it up too: the distribution of self-fertilization among hermaphroditic animals." In: *Evolution* 60.9, pp. 1816–1824. DOI: 10.1111/j.0014-3820.2006.tb00525.x. 112

Jeffery, Nicholas W., Kristin Hultgren, Solomon Tin Chi Chak, T. Ryan Gregory, and Dustin R. Rubenstein (2016). "Patterns of genome size 112

variation in snapping shrimp." In: *Genome* 59.6, pp. 393–402. DOI: 10.1139/gen-2015-0206.

21 Jensen-Seaman, Michael I., Terrence S. Furey, Bret A. Payseur, Yontao Lu, Krishna M. Roskin, Chin-Fu Chen, Michael A. Thomas, David Haussler, and Howard J. Jacob (2004). "Comparative Recombination Rates in the Rat, Mouse, and Human Genomes." In: *Genome Research* 14.4, pp. 528–538. DOI: 10.1101/gr.1970304.

21 Kaback, David B., Vincent Guacci, Dianna Barber, and James W. Mahon (1992). "Chromosome Size-Dependent Control of Meiotic Recombination." In: *Science* 256.5054, pp. 228–232. DOI: 10.1126/science.1566070.

35, 36, 40 Kalhor, Reza, Guillaume Beslon, Manuel Lafond, and Celine Scornavacca (2023). "Classifying the Post-duplication Fate of Paralogous Genes." In: *Comparative Genomics*. Ed. by Katharina Jahn and Tomáš Vinař. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, pp. 1–18. DOI: 10.1007/978-3-031-36911-7_1.

47 – (2024). "A Rigorous Framework to Classify the Postduplication Fate of Paralogous Genes." In: *Journal of Computational Biology* 31.9, pp. 815–833. DOI: 10.1089/cmb.2023.0331.

127 Kaneko, Kunihiko (2009). "Relationship among phenotypic plasticity, phenotypic fluctuations, robustness, and evolvability; Waddington's legacy revisited under the spirit of Einstein." In: *Journal of biosciences* 34.4, pp. 529–542. DOI: 10.1007/s12038-009-0072-9.

10, 68 Kang, Ming, Jing Wang, and Hongwen Huang (2015). "Nitrogen limitation as a driver of genome size evolution in a group of karst plants." In: *Scientific Reports* 5.1, p. 11636. DOI: 10.1038/srep11636.

113, 120 Kapusta, Aurélie, Alexander Suh, and Cédric Feschotte (2017). "Dynamics of genome size evolution in birds and mammals." In: *Proceedings of the National Academy of Sciences* 114.8. Publisher: Proceedings of the National Academy of Sciences, E1460–E1469. DOI: 10.1073/pnas.1616702114.

20 Kara, Eleanna et al. (2014). "A 6.4 Mb Duplication of the α-Synuclein Locus Causing Frontotemporal Dementia and Parkinsonism: Phenotype-Genotype Correlations." In: *JAMA Neurology* 71.9, pp. 1162–1171. DOI: 10.1001/jamaneurol.2014.994.

20 Katju, Vaishali and Ulfar Bergthorsson (2013). "Copy-Number Changes in Evolution: Rates, Fitness Effects and Adaptive Significance." In: *Frontiers in Genetics* 4. DOI: 10.3389/fgene.2013.00273.

141 Kauffman, Stuart and Simon Levin (1987). "Towards a general theory of adaptive walks on rugged landscapes." In: *Journal of theoretical Biology* 128.1, pp. 11–45. DOI: 10.1016/S0022-5193(87)80029-2.

8 Kejnovsky, E, R Hobza, T Cermak, Z Kubat, and B Vyskot (2009). "The role of repetitive DNA in structure and evolution of sex chromosomes in plants." In: *Heredity* 102.6, pp. 533–541. DOI: 10.1038/hdy.2009.17.

Kelkar, Yogeshwar D. and Howard Ochman (2012). "Causes and Consequences of Genome Expansion in Fungi." In: *Genome Biology and Evolution* 4.1, pp. 13–23. DOI: 10.1093/gbe/evr124. 11, 69, 83

Kelleher, Jerome, Alison M Etheridge, and Gilean McVean (May 2016). "Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes." In: *PLOS Computational Biology* 12.5, pp. 1–22. DOI: 10.1371/journal.pcbi.1004842. 179

Kent, Tyler V, Jasmina Uzunović, and Stephen I Wright (2017). "Coevolution between transposable elements and recombination." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 372.1736, p. 20160458. DOI: 10.1098/rstb.2016.0458. 8

Kessner, Darren and John Novembre (2014). "forqs: forward-in-time simulation of recombination, quantitative traits and selection." In: *Bioinformatics* 30.4, pp. 576–577. DOI: 10.1093/bioinformatics/btt712. 179

Kidwell, Margaret G. (2002). "Transposable elements and the evolution of genome size in eukaryotes." In: *Genetica* 115.1, pp. 49–63. DOI: 10.1023/A:1016072014259. 59

Kirkpatrick, Mark (2010). "How and why chromosome inversions evolve." In: *PLoS biology* 8.9, e1000501. DOI: 10.1371/journal.pbio.1000501. 21

Knibbe, Carole, Antoine Coulon, Olivier Mazet, Jean-Michel Fayard, and Guillaume Beslon (2007a). "A Long-Term Evolutionary Pressure on the Amount of Noncoding DNA." In: *Molecular Biology and Evolution* 24.10, pp. 2344–2353. DOI: 10.1093/molbev/msm165. 14, 22, 23, 36, 38, 39, 46, 47, 53, 56, 60, 69, 81, 113, 117, 120, 124, 139, 147

Knibbe, Carole, Olivier Mazet, Fabien Chaudier, Jean-Michel Fayard, and Guillaume Beslon (2007b). "Evolutionary coupling between the deleteriousness of gene mutations and the amount of non-coding sequences." In: *Journal of Theoretical Biology* 244.4, pp. 621–630. DOI: 10.1016/j.jtbi.2006.09.005. 14

Knight, Charles A. and Jeremy M. Beaulieu (2008). "Genome Size Scaling through Phenotype Space." In: *Annals of Botany* 101.6, pp. 759–766. DOI: 10.1093/aob/mcm321. 10, 112

Knight, Charles A., Nicole A. Molinari, and Dmitri A. Petrov (2005). "The Large Genome Constraint Hypothesis: Evolution, Ecology and Phenotype." In: *Annals of Botany* 95.1, pp. 177–190. DOI: 10.1093/aob/mci011. 112

Knoll, Andrew H. (2015). *Life on a Young Planet: The First Three Billion Years of Evolution on Earth - Updated Edition*. Princeton University Press, p. 19. ISBN: 9781400866045. DOI: 10.1515/9781400866045. 17

Konstantinidis, Konstantinos T. and James M. Tiedje (2004). "Trends between gene content and genome size in prokaryotic species with larger genomes." In: *Proceedings of the National Academy of Sciences* 101.9, pp. 3160–3165. DOI: 10.1073/pnas.0308653100. 44

Koonin, Eugene V. (2009). "Evolution of genome architecture." In: *The International Journal of Biochemistry & Cell Biology*. Molecular and 5, 8, 9, 11, 12

Cellular Evolution: A Celebration of the 200th Anniversary of the Birth of Charles Darwin 41.2, pp. 298–306. DOI: 10.1016/j.biocel.2008.09.015.

45 Koskiniemi, Sanna, Song Sun, Otto G. Berg, and Dan I. Andersson (2012). "Selection-Driven Gene Loss in Bacteria." In: *PLOS Genetics* 8.6, e1002787. DOI: 10.1371/journal.pgen.1002787.

112 Krakauer, David C. and Joshua B. Plotkin (2002). "Redundancy, antiredundancy, and the robustness of genomes." In: *Proceedings of the National Academy of Sciences* 99.3, pp. 1405–1409. DOI: 10.1073/pnas.032668599.

6 Kreitman, Martin (1983). "Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster." In: *Nature* 304.5925, pp. 412–417. DOI: 10.1038/304412a0.

10, 11, 45 Kuo, Chih-Horng, Nancy A. Moran, and Howard Ochman (2009). "The consequences of genetic drift for bacterial genome complexity." In: *Genome Res.* 19.8, pp. 1450–1454. DOI: 10.1101/gr.091785.109.

10, 11, 51 Kuo, Chih-Horng and Howard Ochman (2009). "Deletional Bias across the Three Domains of Life." In: *Genome Biology and Evolution* 1, pp. 145–152. DOI: 10.1093/gbe/evp016.

56 LaBar, Thomas and Christoph Adami (2020). "Genome Size and the Extinction of Small Populations." In: *Evolution in Action: Past, Present and Future: A Festschrift in Honor of Erik D. Goodman*. Cham: Springer International Publishing, pp. 167–183. DOI: 10.1007/978-3-030-39831-6_14.

9 Lane, Nick and William Martin (2010). "The energetics of genome complexity." In: *Nature* 467.7318, pp. 929–934. DOI: 10.1038/nature09486.

7 Langmüller, Anna M, Viola Nolte, Marlies Dolezal, and Christian Schlötterer (2023). "The genomic distribution of transposable elements is driven by spatially variable purifying selection." In: *Nucleic Acids Research* 51.17, pp. 9203–9213. DOI: 10.1093/nar/gkad635.

179 Laval, Guillaume and Laurent Excoffier (2004). "SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history." In: *Bioinformatics* 20.15, pp. 2485–2487. DOI: 10.1093/bioinformatics/bth264.

194 Leaché, Adam D and Jamie R Oaks (2017). "The utility of single nucleotide polymorphism (SNP) data in phylogenetics." In: *Annual review of ecology, evolution, and systematics* 48.1, pp. 69–84. DOI: 10.1146/annurev-ecolsys-110316-022645.

45 Lefebure, Tristan et al. (2017). "Less effective selection leads to larger genomes." In: *Genome Research* 27.6, pp. 1016–1028. DOI: 10.1101/gr.212589.116.

13 Lei, Yu et al. (2022). "Overview of structural variation calling: Simulation, identification, and visualization." In: *Computers in Biology and Medicine* 145, p. 105534. DOI: 10.1016/j.compbiomed.2022.105534.

Leth Bak, A., Finn T. Black, C. Christiansen, and E. A. Freundt (1969). "Genome Size of Mycoplasmal DNA." In: *Nature* 224.5225, pp. 1209–1210. DOI: 10.1038/2241209a0. 44

Leushkin, Evgeny V, Georgii A Bazykin, and Alexey S Kondrashov (2012). "Insertions and deletions trigger adaptive walks in Drosophila proteins." In: *Proceedings of the Royal Society B: Biological Sciences* 279.1740, pp. 3075–3082. DOI: 10.1098/rspb.2011.2571. 37

Lewontin, Richard C (1974). *The genetic basis of evolutionary change*. Columbia University Press. 2

Liao, Xingyu, Wufei Zhu, Juexiao Zhou, Haoyang Li, Xiaopeng Xu, Bin Zhang, and Xin Gao (2023). "Repetitive DNA sequence detection and its role in the human genome." In: *Communications Biology* 6.1, p. 954. DOI: 10.1038/s42003-023-05322-y. 7

Liard, Vincent, David P. Parsons, Jonathan Rouzaud-Cornabas, and Guillaume Beslon (2020). "The Complexity Ratchet: Stronger than Selection, Stronger than Evolvability, Weaker than Robustness." In: *Artificial Life* 26.1, pp. 38–57. DOI: 10.1162/artl_a_00312. 1, 14, 22, 131

Lin, Bo, Jianan Hui, and Hongju Mao (2021). "Nanopore Technology and Its Applications in Gene Sequencing." In: *Biosensors* 11.7. DOI: 10.3390/bios11070214. 7

Lipinski, Kendra J, James C Farslow, Kelly A Fitzpatrick, Michael Lynch, Vaishali Katju, and Ulfar Bergthorsson (2011). "High spontaneous rate of gene duplication in Caenorhabditis elegans." In: *Current Biology* 21.4, pp. 306–310. DOI: 10.1016/j.cub.2011.01.026. 129

Liu, Gangiang, John Mattick, and Ryan J Taft (2013). "A meta-analysis of the genomic and transcriptomic composition of complex life." In: *Cell cycle* 12.13, pp. 2061–2072. DOI: 10.4161/cc.25134. 68

Loewenthal, Gil, Elya Wygoda, Natan Nagar, Lior Glick, Itay Mayrose, and Tal Pupko (2022). "The evolutionary dynamics that retain long neutral genomic sequences in face of indel deletion bias: a model and its application to human introns." In: *Open Biology* 12.12, p. 220223. DOI: 10.1098/rsob.220223. 10, 36, 81, 128

Luiselli, Juliette, Paul Banse, Olivier Mazet, Nicolas Lartillot, and Guillaume Beslon (2025a). "Structural mutations set an equilibrium non-coding genome fraction." In: *bioRxiv*. DOI: 10.1101/2025.02.03.636187. 67, 117, 122, 123

Luiselli, Juliette, David P Parsons, Romain Gallé, Paul Banse, Jonathan Rouzaud-Cornabas, and Guillaume Beslon (2025b). "Aevol-9: simulating the evolution of genome architecture." In: *BioRxiv*. DOI: 10.1101/2025.04.10.648095. 91, 101, 113, 120, 124

Luiselli, Juliette, Jonathan Rouzaud-Cornabas, Nicolas Lartillot, and Guillaume Beslon (Nov. 2024). "Genome streamlining: effect of mutation rate and population size on genome size reduction." In: *Genome Biology and Evolution*, evae250. ISSN: 1759-6653. DOI: 10.1093/gbe/evae250. 43, 69, 76, 117, 120, 122

44 Lynch, Michael (2006a). "Streamlining and simplification of microbial genome architecture." In: *Annu. Rev. Microbiol.* 60, pp. 327–349. DOI: 10.1146/annurev.micro.60.080805.142300.

10, 87, 130 – (2006b). "The origins of eukaryotic gene structure." In: *Molecular biology and evolution* 23.2, pp. 450–468. DOI: 10.1093/molbev/msj050.

10, 44 – (2007a). "The frailty of adaptive hypotheses for the origins of organismal complexity." In: *Proceedings of the National Academy of Sciences* 104.suppl_1, pp. 8597–8604. DOI: 10.1073/pnas.0702207104.

10, 11, 57, 69, 112, 122, 123 – (2007b). *The origins of genome architecture*. Vol. 98. Sinauer associates Sunderland, MA.

38 – (2010). "Evolution of the mutation rate." In: *Trends in Genetics* 26.8, pp. 345–352. DOI: 10.1016/j.tig.2010.05.003.

83 Lynch, Michael, Matthew S. Ackerman, Jean-Francois Gout, Hongan Long, Way Sung, W. Kelley Thomas, and Patricia L. Foster (2016). "Genetic drift, selection and the evolution of the mutation rate." In: *Nature Reviews Genetics* 17.11, pp. 704–714. DOI: 10.1038/nrg.2016.104.

58, 81, 85, 182, 192 Lynch, Michael, Farhan Ali, Tongtong Lin, Yaohai Wang, Jiahao Ni, and Hongan Long (2023). "The divergence of mutation rates and spectra across the Tree of Life." In: *EMBO reports* 24.10, e57561. DOI: 10.15252/embr.202357561.

10, 11, 44, 45, 56, 59, 68, 69, 83, 87, 89, 112 Lynch, Michael and John S. Conery (2003). "The Origins of Genome Complexity." In: *Science* 302, pp. 1401–1404. DOI: 10.1126/science.1089370.

9 Lynch, Michael and Georgi K Marinov (2017). "Membranes, energetics, and evolution across the prokaryote-eukaryote divide." In: *eLife* 6. Ed. by Paul G Falkowski. Publisher: eLife Sciences Publications, Ltd, e20437. DOI: 10.7554/eLife.20437.

20 Lynch, Michael et al. (2008). "A genome-wide view of the spectrum of spontaneous mutations in yeast." In: *Proceedings of the National Academy of Sciences* 105.27, pp. 9272–9277. DOI: 10.1073/pnas.0803466105.

10, 68 Malerba, Martino E., Giulia Ghedini, and Dustin J. Marshall (2020). "Genome Size Affects Fitness in the Eukaryotic Alga Dunaliella tertiolecta." In: *Current Biology* 30.17, 3450–3456.e3. DOI: 10.1016/j.cub.2020.06.033.

45, 56, 112 Marais, Gabriel A. B., Alexandra Calteau, and Olivier Tenaillon (2008). "Mutation rate and genome reduction in endosymbiotic and free-living bacteria." In: *Genetica* 134.2, pp. 205–210. DOI: 10.1007/s10709-007-9226-6.

7, 11, 59, 69, 83, 130 Marino, Alba, Gautier Debaecker, Anna-Sophie Fiston-Lavier, Annabelle Haudry, and Benoit Nabholz (2024). "Effective population size does not explain long-term variation in genome size and transposable element content in animals." In: DOI: 10.7554/elife.100574.1.

179 Marjoram, Paul and Jeff Wall (2006). "Fast "coalescent" simulation." In: *BMC Genetics* 7, pp. 16 –16. DOI: 10.1186/1471-2156-7-16.

Martinez-Gutierrez, Carolina A. and Frank O. Aylward (2022). "Genome size distributions in bacteria and archaea are strongly linked to evolutionary history at broad phylogenetic scales." In: *PLOS Genetics* 18.5. Publisher: Public Library of Science, e1010220. DOI: 10.1371/journal.pgen.1010220. 44

Martiny, Pedro Flombaumand José L. Gallegosand Rodolfo A. Gordilloand José Rincónand Lina L. Zabalaand Nianzhi Jiaoand David M. Karland William K. W. Liand Michael W. Lomasand Daniele Venezianoand Carolina S. Veraand Jasper A. Vrugtand Adam C. (2013). "Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*." In: *Proceedings of the National Academy of Sciences* 110.24, pp. 9824–9829. DOI: 10.1073/pnas.1307701110. 45

Martínez-Cano, David J., Mariana Reyes-Prieto, Esperanza Martínez-Romero, Laila P. Partida-Martínez, Amparo Latorre, Andrés Moya, and Luis Delaye (2015). "Evolution of small prokaryotic genomes." In: *Frontiers in Microbiology* 5. DOI: 10.3389/fmicb.2014.00742. 44

Masel, Joanna and Mark L Siegal (2009). "Robustness: mechanisms and consequences." In: *Trends in genetics* 25.9, pp. 395–403. DOI: 10.1016/j.tig.2009.07.005. 127

Mathur, Stephen J. Giovannoniand H. James Trippand Scott Givanand Mircea Podarand Kevin L. Verginand Damon Baptistaand Lisa Bibbsand Jonathan Eadsand Toby H. Richardsonand Michiel Noordewierand Michael S. Rappéand Jay M. Shortand James C. Carringtonand Eric J. (2005). "Genome Streamlining in a Cosmopolitan Oceanic Bacterium." In: *Science* 309.5738, pp. 1242–1245. DOI: 10.1126/science.1114057. 44, 45

McVean, Gil and Niall J. Cardin (2005). "Approximating the coalescent with recombination." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 360, pp. 1387 –1393. DOI: 10.1098/rstb.2005.1673. 179

Mérot, Claire, Rebekah A Oomen, Anna Tigano, and Maren Wellenreuther (2020). "A roadmap for understanding the evolutionary significance of structural genomic variation." In: *Trends in Ecology & Evolution* 35.7, pp. 561–572. DOI: 10.1016/j.tree.2020.03.002. 12, 20, 21, 35

Mira, A. and N.A. Moran (2002). "Estimating Population Size and Transmission Bottlenecks in Maternally Transmitted Endosymbiotic Bacteria." In: *Microbial Ecology* 44.2, pp. 137–143. DOI: 10.1007/s00248-002-0012-9. 44

Mira, Alex, Howard Ochman, and Nancy A. Moran (2001). "Deletional bias and the evolution of bacterial genomes." In: *Trends in Genetics* 17.10, pp. 589–596. DOI: 10.1016/S0168-9525(01)02447-7. 57

Misevic, Dusan, Antoine Frenoy, Ariel B Lindner, and François Taddei (2015). "Shape matters: Lifecycle of cooperative patches promotes cooperation in bulky populations." In: *Evolution* 69.3, pp. 788–802. DOI: 10.1111/evo.12616. 131, 147

11, 69, 83 Mohlhenrich, Erik Roger and Rachel Lockridge Mueller (2016). "Genetic drift and mutational hazard in the evolution of salamander genomic gigantism." In: *Evolution* 70.12, pp. 2865–2878. DOI: 10.1111/evo.13084.

12, 129 Molari, Marco, Liam P Shaw, and Richard A Neher (2025). "Quantifying the Evolutionary Dynamics of Structure and Content in Closely Related E. coli Genomes." In: *Molecular Biology and Evolution* 42.1, msae272. DOI: 10.1093/molbev/msae272.

44 Moran, N A (1996). "Accelerated evolution and Muller's rachet in endosymbiotic bacteria." In: *Proceedings of the National Academy of Sciences* 93.7, pp. 2873–2878. DOI: 10.1073/pnas.93.7.2873.

45 Moran, Nancy A. (2002). "Microbial Minimalism: Genome Reduction in Bacterial Pathogens." In: *Cell* 108.5, pp. 583–586. DOI: 10.1016/S0092-8674(02)00665-7.

58, 59 Moran, Nancy A (2003). "Tracing the evolution of gene loss in obligate bacterial symbionts." In: *Current Opinion in Microbiology* 6.5, pp. 512–518. DOI: 10.1016/j.mib.2003.08.001.

44, 45, 56 Moran, Nancy A. and Alex Mira (2001). "The process of genome shrinkage in the obligate symbiont Buchnera aphidicola." In: *Genome Biol* 2.12, pp. 1–12. DOI: 10.1186/gb-2001-2-12-research0054.

5 Morgan, LV (1938). "Effects of a compound duplication of the X chromosome of Drosophila melanogaster." In: *Genetics* 23.5, p. 423. DOI: 10.1093/genetics/23.5.423.

45 Morris, J. Jeffrey, Richard E. Lenski, and Erik R. Zinser (2012). "The Black Queen Hypothesis: Evolution of Dependencies through Adaptive Gene Loss." In: *mBio* 3.2, e00036–12. DOI: 10.1128/mBio.00036-12.

13 Mu, John C, Marghoob Mohiyuddin, Jian Li, Narges Bani Asadi, Mark B Gerstein, Alexej Abyzov, Wing H Wong, and Hugo YK Lam (2015). "VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications." In: *Bioinformatics* 31.9, pp. 1469–1471. DOI: 10.1093/bioinformatics/btu828.

69 Mueller, Rachel Lockridge and Elizabeth L. Jockusch (2018). "Jumping genomic gigantism." In: *Nature Ecology & Evolution* 2.11, pp. 1687–1688. DOI: 10.1038/s41559-018-0703-3.

194 Muller, M.-H., C. Poncet, J. M. Properi, S. Santoni, and J. Ronfort (2006). "Domestication history in the Medicago sativa species complex: inferences from nuclear sequence polymorphism." In: *Molecular Ecology* 15.6, pp. 1589–1602. DOI: 10.1111/j.1365-294X.2006.02851.x.

143 Musumeci, Olimpia, Antoni L Andreu, Sara Shanske, Nereo Bresolin, Giacomo P Comi, Rodney Rothstein, Eric A Schon, and Salvatore DiMauro (2000). "Intragenic inversion of mtDNA: a new type of pathogenic mutation in a patient with mitochondrial myopathy."

In: *The American Journal of Human Genetics* 66.6, pp. 1900–1904. DOI: 10.1086/302927.

Müller, Rebekka, Ingemar Kaj, and Carina F. Mugal (2022). "A Nearly Neutral Model of Molecular Signatures of Natural Selection after Change in Population Size." In: *Genome Biology and Evolution* 14.5, evac058. DOI: 10.1093/gbe/evac058. 45

Nattestad, Maria et al. (2018). "Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line." In: *Genome Research* 28.8, pp. 1126–1135. DOI: 10.1101/gr.231100.117. 20

Ngugi, David K., Silvia G. Acinas, Pablo Sánchez, Josep M. Gasol, Susana Agusti, David M. Karl, and Carlos M. Duarte (2023). "Abiotic selection of microbial genome size in the global ocean." In: *Nature Communications* 14.1, p. 1384. DOI: 10.1038/s41467-023-36988-x. 45, 59

Noor, Mohamed A. F., Katherine L. Grams, Lisa A. Bertucci, and Jane Reiland (2001). "Chromosomal inversions and the reproductive isolation of species." In: *Proceedings of the National Academy of Sciences* 98.21, pp. 12084–12088. DOI: 10.1073/pnas.221274498. 20

Nordborg, Magnus (2000). "Linkage Disequilibrium, Gene Trees and Selfing: An Ancestral Recombination Graph With Partial Self-Fertilization." In: *Genetics* 154.2, pp. 923–929. DOI: 10.1093/genetics/154.2.923. 112, 117, 122

O'Neill, Bill (2003). "Digital Evolution." In: *PLOS Biology* 1.1, e18. DOI: 10.1371/journal.pbio.0000018. 19

Oggenfuss, Ursula et al. (2021). "A population-level invasion by transposable elements triggers genome expansion in a fungal pathogen." In: *eLife* 10. Ed. by Detlef Weigel, Marie Mirouze, Zoé Joly-Lopez, and Leandro Quadrana, e69249. DOI: 10.7554/eLife.69249. 59

Ohno, Susumu (1972). "So much" junk" DNA in our genome. In" Evolution of Genetic Systems"." In: *Brookhaven Symposium in Biology*. Vol. 23, pp. 366–370. 7, 68

Ohtomo, Yoko, Takeshi Kakegawa, Akizumi Ishida, Toshiro Nagase, and Minik T Rosing (2014). "Evidence for biogenic graphite in early Archaean Isua metasedimentary rocks." In: *Nature Geoscience* 7.1, pp. 25–28. DOI: 10.1038/ngeo2025. 8

Olson, C Anders, Nicholas C Wu, and Ren Sun (2014). "A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain." In: *Current biology* 24.22, pp. 2643–2651. DOI: 10.1016/j.cub.2014.09.072. 35

Palazzo, Alexander F and T Ryan Gregory (2014). "The case for junk DNA." In: *PLoS genetics* 10.5, e1004351. DOI: 10.1371/journal.pgen.1004351. 7, 68

Parée, Tom, Luke Noble, João Ferreira Gonçalves, and Henrique Teotónio (2024). "Rec-1 loss of function increases recombination in the central gene clusters at the expense of autosomal pairing centers." In: *Genetics* 226.3, iyad205. DOI: 10.1093/genetics/iyad205. 7

7 Parée, Tom, Luke Noble, Denis Roze, and Henrique Teotónio (2025). "Selection can favor a recombination landscape that limits polygenic adaptation." In: *Molecular Biology and Evolution* 42.1, msae273. DOI: 10.1093/molbev/msae273.

23, 139, 141 Parsons, David P., Carole Knibbe, and Guillaume Beslon (2010). "Importance of the rearrangement rates on the organization of transcription." In: *Proceedings of Artificial Life XII*, pp. 479–486. ISBN: 978-0-262-29075-3.

22 Parsons, David (2011). "Indirect Selection in Darwinian Evolution: Mechanisms and Implications." PhD thesis. INSA Lyon.

44, 68 Pellicier, Jaume, Michael F. Fay, and Ilia J. Leitch (2010). "The largest eukaryotic genome of them all?" In: *Botanical Journal of the Linnean Society* 164.1, pp. 10–15. DOI: 10.1111/j.1095-8339.2010.01072.x.

184 Peters, Orson RL (2021). "Pattern-defeating quicksort." In: *arXiv*. DOI: 10.48550/arXiv.2106.05123.

10, 44 Petrov, Dmitri A. (2001). "Evolution of genome size: new approaches to an old problem." In: *Trends in Genetics* 17.1, pp. 23–28. DOI: 10.1016/S0168-9525(00)02157-0.

45, 64, 68, 83 – (2002). "Mutational Equilibrium Model of Genome Size Evolution." In: *Theoretical Population Biology* 61.4, pp. 531–544. DOI: 10.1006/tpbi.2002.1605.

132 Pigliucci, Massimo (2008). "Is evolvability evolvable?" In: *Nature Reviews Genetics* 9.1, pp. 75–82. DOI: 10.1038/nrg2278.

7 Pombo, Ana and Niall Dillon (2015). "Three-dimensional genome architecture: players and mechanisms." In: *Nature reviews Molecular cell biology* 16.4, pp. 245–257. DOI: 10.1038/nrm3965.

13 Qin, Maochun, Biao Liu, Jeffrey M Conroy, Carl D Morrison, Qiang Hu, Yubo Cheng, Mitsuko Murakami, Adekunle O Odunsi, Candace S Johnson, Lei Wei, et al. (2015). "SCNVSim: somatic copy number variation and structure variation simulator." In: *BMC bioinformatics* 16, pp. 1–6. DOI: 10.1186/s12859-015-0502-7.

20 Quandt, Erik M, Jimmy Gollihar, Zachary D Blount, Andrew D Ellington, George Georgiou, and Jeffrey E Barrick (2015). "Fine-tuning citrate synthase flux potentiates and refines metabolic innovation in the Lenski evolution experiment." In: *Elife* 4, e09696. DOI: 10.7554/eLife.09696.

7 Quesneville, Hadi (2020). "Twenty years of transposable element analysis in the Arabidopsis thaliana genome." In: *Mobile DNA* 11.1, p. 28. DOI: 10.1093/nar/gkad635.

20, 21, 39, 83, 129 Raeside, Colin, Joël Gaffé, Daniel E Deatherage, Olivier Tenaillon, Adam M Briska, Ryan N Ptashkin, Stéphane Cruveiller, Claudine Médigue, Richard E Lenski, Jeffrey E Barrick, et al. (2014). "Large chromosomal rearrangements during a long-term evolution experiment with Escherichia coli." In: *MBio* 5.5, e01377–14. DOI: 10.1128/mbio.01377-14.

Ratcliff, Emma P. Binghamand William C. (2024). "A nonadaptive explanation for macroevolutionary patterns in the evolution of complex multicellularity." In: *Proceedings of the National Academy of Sciences* 121.7, e2319840121. DOI: 10.1073/pnas.2319840121. 10, 45, 82, 103

Ravetch, Jeffrey V and BICE Perussia (1989). "Alternative membrane forms of Fc gamma RIII (CD16) on human natural killer cells and neutrophils. Cell type-specific expression of two genes that differ in single nucleotide substitutions." In: *The Journal of experimental medicine* 170.2, pp. 481–497. DOI: 10.1084/jem.170.2.481. 6

Rick, Charles M (1940). "On the nature of X-ray induced deletions in Tradescantia chromosomes." In: *Genetics* 25.5, p. 466. DOI: 10.1093/genetics/25.5.466. 5

Riley, Alex B., Dohyup Kim, and Allison K. Hansen (2017). "Genome Sequence of Candidatus Carsonella ruddii Strain BC, a Nutritional Endosymbiont of Bactericera cockerelli." In: *Genome Announcements* 5.17. DOI: 10.1128/genomea.00236-17. 68

Rinn, John L. and Howard Y. Chang (2012). "Genome Regulation by Long Noncoding RNAs." In: *Annual Review of Biochemistry* 81.1, pp. 145–166. DOI: 10.1146/annurev-biochem-051410-092902. 68

Rocha, Eduardo P. C. (2006). "Inference and Analysis of the Relative Stability of Bacterial Chromosomes." In: *Molecular Biology and Evolution* 23.3, pp. 513–522. DOI: 10.1093/molbev/msj052. 20, 39

Romiguier, J., V. Ranwez, E.J.P. Douzery, and N. Galtier (2012). "Genomic Evidence for Large, Long-Lived Ancestors to Placental Mammals." In: *Molecular Biology and Evolution* 30.1, pp. 5–13. DOI: 10.1093/molbev/mss211. 46

Romiguier, J. et al. (Nov. 2014). "Comparative population genomics in animals uncovers the determinants of genetic diversity." In: *Nature* 515.7526, pp. 261–263. DOI: 10.1038/nature13685. 112

Roze, D. (Oct. 2015). "Effects of Interference Between Selected Loci on the Mutation Load, Inbreeding Depression, and Heterosis." In: *Genetics* 201.2, pp. 745+. DOI: 10.1534/genetics.115.178533. 123

Roze, Denis (2023). "Causes and consequences of linkage disequilibrium among transposable elements within eukaryotic genomes." In: *Genetics* 224.2, iyad058. DOI: 10.1093/genetics/iyad058. 113, 120

Roze, Denis and Thomas Lenormand (2005). "Self-Fertilization and the Evolution of Recombination." In: *Genetics* 170.2, pp. 841–857. DOI: 10.1534/genetics.104.036384. 113, 123

Rutten, Jacob Pieter, Paulien Hogeweg, and Guillaume Beslon (2019). "Adapting the engine to the fuel: mutator populations can reduce the mutational load by reorganizing their genome structure." In: *BMC Evolutionary Biology* 19.1, p. 191. DOI: 10.1186/s12862-019-1507-z. 22, 131, 147

Sagitov, Serik (2003). "Convergence to the coalescent with simultaneous multiple mergers." In: *Journal of Applied Probability* 40.4, pp. 839–854. DOI: 10.1239/jap/1067436085. 178

6 Sanger, Frederick, Steven Nicklen, and Alan R Coulson (1977). "DNA sequencing with chain-terminating inhibitors." In: *Proceedings of the national academy of sciences* 74.12, pp. 5463–5467. DOI: 10.1073/pnas.74.12.5463.

44 Sanjuán, Selma Gagoand Santiago F. Elenaand Ricardo Floresand Rafael (2009). "Extremely High Mutation Rate of a Hammerhead Viroid." In: *Science* 323.5919, pp. 1308–1308. DOI: 10.1126/science.1169202.

12, 129 Saxena, Ayush Shekhar and Charles F. Baer (2025). "High rate of mutation and efficient removal by selection of structural variants from natural populations of Caenorhabditis elegans." In: *bioRxiv*. DOI: 10.1101/2025.03.22.644739.

11, 58 Schaack, Michael Lynchand Britt Koskellaand Sarah (2006). "Mutation Pressure and the Evolution of Organelle Genomic Architecture." In: *Science* 311.5768, pp. 1727–1730. DOI: 10.1126/science.1118884.

6 Schadt, Eric E, Steve Turner, and Andrew Kasarskis (2010). "A window into third-generation sequencing." In: *Human molecular genetics* 19.R2, R227–R240. DOI: 10.1093/hmg/ddq416.

44 Schneiker, Susanne et al. (2007). "Complete genome sequence of the myxobacterium Sorangium cellulosum." In: *Nature Biotechnology* 25.11, pp. 1281–1289. DOI: 10.1038/nbt1354.

8 Schrader, Lukas and Jürgen Schmitz (2019). "The impact of transposable elements in adaptive evolution." In: *Molecular Ecology* 28.6, pp. 1537–1549. DOI: 10.1111/mec.14794.

20 Schrider, Daniel R., David Houle, Michael Lynch, and Matthew W. Hahn (2013). "Rates and Genomic Consequences of Spontaneous Mutational Events in *Drosophila melanogaster*." In: *Genetics* 194.4, pp. 937–954. DOI: 10.1534/genetics.113.151670.

6 Schuster, Stephan C (2008). "Next-generation sequencing transforms today's biology." In: *Nature methods* 5.1, pp. 16–18. DOI: 10.1038/nmeth1156.

178 Schweinsberg, Jason Ross (2001). *Coalescents with simultaneous multiple collisions*. University of California, Berkeley.

129 Shaffer, Lisa G and James R Lupski (2000). "Molecular mechanisms for constitutional chromosomal rearrangements in humans." In: *Annual review of genetics* 34.1, pp. 297–329. DOI: 10.1146/annurev.genet.34.1.297.

179 Shlyakhter, Ilya, Pardis C Sabeti, and Stephen F Schaffner (2014). "Cosi2: an efficient simulator of exact and approximate coalescent with selection." In: *Bioinformatics* 30.23, pp. 3427–3429. DOI: 10.1093/bioinformatics/btu562.

177 Sigwart, Julia (Mar. 2009). "Coalescent Theory: An Introduction." In: *Systematic Biology* 58.1, pp. 162–165. DOI: 10.1093/schbul/syp004.

69, 83 Sloan, Daniel B., Andrew J. Alverson, John P. Chuckalovcak, Martin Wu, David E. McCauley, Jeffrey D. Palmer, and Douglas R. Taylor (2012). "Rapid Evolution of Enormous, Multichromosomal Genomes

in Flowering Plant Mitochondria with Exceptionally High Mutation Rates." In: *PLOS Biology* 10.1, e1001241. DOI: 10.1371/journal.pbio.1001241.

Smith, David Roy, Takashi Hamaji, Bradley J.S.C. Olson, Pierre M. Durand, Patrick Ferris, Richard E. Michod, Jonathan Featherston, Hisayoshi Nozaki, and Patrick J. Keeling (2013). "Organelle Genome Complexity Scales Positively with Organism Size in Volvocine Green Algae." In: *Molecular Biology and Evolution* 30.4, pp. 793–797. DOI: 10.1093/molbev/mst002. 11, 69

Smolke, Christina D and Jay D Keasling (2002). "Effect of gene location, mRNA secondary structures, and RNase sites on expression of two genes in an engineered operon." In: *Biotechnology and bioengineering* 80.7, pp. 762–776. DOI: 10.1002/bit.10434. 5

Sniegowski, Paul D., Philip J. Gerrish, Toby Johnson, and Aaron Shaver (2000). "The evolution of mutation rates: separating causes from consequences." In: *BioEssays* 22.12, pp. 1057–1066. DOI: 10.1002/1521-1878(200012)22:12<1057::AID-BIES3>3.0.CO;2-W. 112, 122

Stapley, Jessica, Philine G. D. Feulner, Susan E. Johnston, Anna W. Santure, and Carole M. Smadja (2017). "Variation in recombination frequency and distribution across eukaryotes: patterns and processes." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 372.1736, p. 20160455. DOI: 10.1098/rstb.2016.0455. 112

Starr, Tyler N and Joseph W Thornton (2016). "Epistasis in protein evolution." In: *Protein science* 25.7, pp. 1204–1218. DOI: 10.1002/pro.2897. 35

Stetsenko, Roman and Denis Roze (2022). "The evolution of recombination in self-fertilizing organisms." In: *Genetics* 222.1, iyac114. DOI: 10.1093/genetics/iyac114. 113, 123

Sultana, Tania, Alessia Zamborlini, Gael Cristofari, and Pascale Lesage (2017). "Integration site selection by retroviruses and transposable elements in eukaryotes." In: *Nature Reviews Genetics* 18.5, pp. 292–308. DOI: 10.1038/nrg.2017.7. 7

Sung, Way, Matthew S Ackerman, Marcus M Dillon, Thomas G Platt, Clay Fuqua, Vaughn S Cooper, and Michael Lynch (2016). "Evolution of the Insertion-Deletion Mutation Rate Across the Tree of Life." In: *G3 Genes Genomes Genetics* 6.8, pp. 2583–2591. DOI: 10.1534/g3.116.030890. 69, 83

Sung, Way, Matthew S Ackerman, Samuel F Miller, Thomas G Doak, and Michael Lynch (2012). "Drift-barrier hypothesis and mutation-rate evolution." In: *Proceedings of the National Academy of Sciences* 109.45, pp. 18488–18492. DOI: 10.1073/pnas.1216223109. 57, 83

Takeuchi, Nobuto and Paulien Hogeweg (2007). "Error-threshold exists in fitness landscapes with lethal mutants." In: *BMC Evolutionary Biology* 7.1, p. 15. DOI: 10.1186/1471-2148-7-15. 57

Tam, S. M., M. Causse, C. Garchery, H. Burck, C. Mhiri, and M.-A. Grandbastien (May 2007). "The distribution of copia-type retrotrans- 113

posons and the evolutionary history of tomato and related wild species." In: *Journal of Evolutionary Biology* 20.3, pp. 1056–1072. DOI: 10.1111/j.1420-9101.2007.01293.x.

2 Thomas CA, Jr (1971). "The genetic organization of chromosomes." In: *Annual Review of Genetics* 5, pp. 237–256. DOI: 10.1146/annurev.ge.05.120171.001321.

6 Treangen, Todd J and Steven L Salzberg (2012). "Repetitive DNA and next-generation sequencing: computational challenges and solutions." In: *Nature Reviews Genetics* 13.1, pp. 36–46. DOI: 10.1038/nrg3117.

12, 13, 37 Trujillo, Leonardo, Paul Banse, and Guillaume Beslon (2022). "Getting higher on rugged landscapes: Inversion mutations open access to fitter adaptive peaks in NK fitness landscapes." In: *PLoS Computational Biology* 18.10, e1010647. DOI: 10.1371/journal.pcbi.1010647.

37 Vakhrusheva, Anna A, Marat D Kazanov, Andrey A Mironov, and Georgii A Bazykin (2011). "Evolution of prokaryotic genes by shift of stop codons." In: *Journal of molecular evolution* 72, pp. 138–146. DOI: 10.1007/s00239-010-9408-1.

179 Virgoulay, Thimothée, François Rousset, and Raphaël Leblois (2021). "GSpace: an exact coalescence simulator of recombining genomes under isolation by distance." In: *Bioinformatics* 37.20, pp. 3673–3675. DOI: 10.1093/bioinformatics/btab261.

1 Wagner, Andreas (2008). "Robustness and evolvability: a paradox resolved." In: *Proceedings of the Royal Society B: Biological Sciences* 275.1630, pp. 91–100. DOI: 10.1098/rspb.2007.1137.

132 Wagner, Günter P and Lee Altenberg (1996). "Perspective: complex adaptations and the evolution of evolvability." In: *Evolution* 50.3, pp. 967–976. DOI: 10.1111/j.1558-5646.1996.tb02339.x.

20 Wala, Jeremiah A, Pratiti Bandopadhayay, Noah F Greenwald, Ryan O'Rourke, Ted Sharpe, Chip Stewart, Steve Schumacher, Yilong Li, Joachim Weischenfeldt, Xiaotong Yao, et al. (2018). "SvABA: genome-wide detection of structural variants and indels by local assembly." In: *Genome research* 28.4, pp. 581–591. DOI: 10.1101/gr.221028.117.

113, 117 Wang, J, E Santiago, and Armando Caballero (2016a). "Prediction and estimation of effective population size." In: *Heredity* 117.4, pp. 193–206. DOI: 10.1038/hdy.2016.43.

36 Wang, Jianbin, H Christina Fan, Barry Behr, and Stephen R Quake (2012). "Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm." In: *Cell* 150.2, pp. 402–412. DOI: 10.1016/j.cell.2012.06.030.

4 Wang, Jinliang (2005). "Estimation of effective population sizes from data on genetic markers." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 360.1459, pp. 1395–1409. DOI: 10.1098/rstb.2005.1682.

179 Wang, Ying, Ying Zhou, Linfeng Li, Xian Chen, Yuting Liu, Zhi-Ming Ma, and Shuhua Xu (2014). "A new method for modeling coalescent

processes with recombination." In: *BMC bioinformatics* 15, pp. 1–12. DOI: 10.1186/1471-2105-15-273.

Wang, Yinhua, Carolina Diaz Arenas, Daniel M Stoebel, Kenneth Flynn, Ethan Knapp, Marcus M Dillon, Andrea Wünsche, Philip J Hatcher, Francisco B-G Moore, Vaughn S Cooper, et al. (2016b). "Benefit of transferred mutations is better predicted by the fitness of recipients than by their ecological or genetic relatedness." In: *Proceedings of the National Academy of Sciences* 113.18, pp. 5047–5052. DOI: 10.1073/pnas.1524988113. 37

Waples, Robin S (2010). "Spatial-temporal stratifications in natural populations and how they affect understanding and estimation of effective population size." In: *Molecular Ecology Resources* 10.5, pp. 785–796. DOI: 10.1111/j.1755-0998.2010.02876.x. 4, 143

– (2022). "What Is Ne, Anyway?" In: *Journal of Heredity* 113.4, pp. 371–379. DOI: 10.1093/jhered/esac023. 2, 4, 133

Wei, Wen, Lifeng Xiong, Yuan-Nong Ye, Meng-Ze Du, Yi-Zhou Gao, Kai-Yue Zhang, Yan-Ting Jin, Zujun Yang, Po-Chun Wong, Susanna KP Lau, et al. (2018). "Mutation landscape of base substitutions, duplications, and deletions in the representative current cholera pandemic strain." In: *Genome Biology and Evolution* 10.8, pp. 2072–2085. DOI: 10.1093/gbe/evy151. 12, 129

Wei, Xinzhu and Jianzhi Zhang (2019). "Patterns and mechanisms of diminishing returns from beneficial mutations." In: *Molecular biology and evolution* 36.5, pp. 1008–1021. DOI: 10.1093/molbev/msz035. 37

Weissensteiner, Matthias H, Ignas Bunikis, Ana Catalán, Kees-Jan Francoijs, Ulrich Knief, Wieland Heim, Valentina Peona, Saurabh D Pophaly, Fritz J Sedlazeck, Alexander Suh, et al. (2020). "Discovery and population genomics of structural variation in a songbird genus." In: *Nature communications* 11.1, p. 3403. DOI: 10.1038/s41467-020-17195-4. 12

Weissman, Daniel B, Marcus W Feldman, and Daniel S Fisher (2010). "The rate of fitness-valley crossing in sexual populations." In: *Genetics* 186.4, pp. 1389–1410. DOI: 10.1534/genetics.110.123240. 20, 21

Wellenreuther, Maren and Louis Bernatchez (2018). "Eco-Evolutionary Genomics of Chromosomal Inversions." In: *Trends in Ecology & Evolution* 33.6, pp. 427–440. DOI: 10.1016/j.tree.2018.04.002. 21, 82, 84

Wellenreuther, Maren, Claire Mérot, Emma Berdan, and Louis Bernatchez (2019). "Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification." In: *Molecular ecology* 28.6, pp. 1203–1209. DOI: 10.1111/mec.15066. 20

Wernegreen, Jennifer J. (2002). "Genome evolution in bacterial endosymbionts of insects." In: *Nat Rev Genet* 3.11, pp. 850–861. DOI: 10.1038/nrg931. 45

– (2015). "Endosymbiont evolution: predictions from theory and surprises from genomes." In: *Annals of the New York Academy of Sciences* 1360.1, pp. 16–35. DOI: 10.1111/nyas.12740. 44, 58

7 Wessler, Susan R (2006). "Transposable elements and the evolution of eukaryotic genomes." In: *Proceedings of the National Academy of Sciences* 103.47, pp. 17600–17601. DOI: 10.1073/pnas.0607612103.

44 Westoby, Mark, Daniel Aagren Nielsen, Michael R. Gillings, Elena Litchman, Joshua S. Madin, Ian T. Paulsen, and Sasha G. Tetu (2021). "Cell size, genome size, and maximum growth rate are near-independent dimensions of ecological variation across bacteria and archaea." In: *Ecology and Evolution* 11.9, pp. 3956–3976. DOI: 10.1002/ece3.7290.

112, 113, 120 Whitney, Kenneth D. et al. (2010). "A role for nonadaptive processes in plant genome size evolution?" In: *Evolution* 64.7, pp. 2097–2109. DOI: 10.1111/j.1558-5646.2010.00967.x.

53 Wilke, Claus O. and Christoph Adami (2003). "Evolution of mutational robustness." In: *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 522.1, pp. 3–11. DOI: 10.1016/S0027-5107(02)00307-X.

1, 56 Wilke, Claus O., Jia Lan Wang, Charles Ofria, Richard E. Lenski, and Christoph Adami (2001). "Evolution of digital organisms at high mutation rates leads to survival of the flattest." In: *Nature* 412.6844, pp. 331–333. DOI: 10.1038/35085569.

22, 27, 32, 37 Wiser, Michael J, Noah Ribeck, and Richard E Lenski (2013). "Long-term dynamics of adaptation in asexual populations." In: *Science* 342.6164, pp. 1364–1367. DOI: 10.1126/science.1243357.

178, 185–188 Wiuf, Carsten and Jotun Hein (Nov. 1997). "On the Number of Ancestors to a DNA Sequence." In: *Genetics* 147.3, pp. 1459–1468. DOI: 10.1093/genetics/147.3.1459.

45 Wolf, Yuri I. and Eugene V. Koonin (2013). "Genome reduction as the dominant mode of evolution." In: *BioEssays* 35.9, pp. 829–837. DOI: 10.1002/bies.201300037.

4, 75, 181 Wright, Sewall (1931). "Evolution in Mendelian populations." In: *Genetics* 16.2, p. 97. DOI: 10.1093/genetics/16.2.97.

117 – (1938). "Size of population and breeding structure in relation to evolution." In: *Science* 87, pp. 430–431. ISSN: 0036-8075.

112, 120 Wright, Stephen I., Rob W. Ness, John Paul Foxe, and Spencer C. H. Barrett (2008). "Genomic Consequences of Outcrossing and Selfing in Plants." In: *International Journal of Plant Sciences* 169.1, pp. 105–118. DOI: 10.1086/523366.

13 Xia, Yuchao, Yun Liu, Minghua Deng, and Ruibin Xi (2017). "Pysim-sv: a package for simulating structural variation data with GC-biases." In: *BMC bioinformatics* 18, pp. 23–30. DOI: 10.1186/s12859-017-1464-8.

143 Xue, James R. et al. (2023). "The functional and evolutionary impacts of human-specific deletions in conserved elements." In: *Science* 380.6643, eabn2253. DOI: 10.1126/science.abn2253.

20, 21 Yancopoulos, Sophia, Oliver Attie, and Richard Friedberg (2005). "Efficient sorting of genomic permutations by translocation, inversion

and block interchange." In: *Bioinformatics* 21.16, pp. 3340–3346. DOI: 10.1093/bioinformatics/bti535.

Yi, Soojin and J. Todd Streelman (2005). "Genome size is negatively correlated with effective population size in ray-finned fish." In: *Trends in Genetics* 21.12. Publisher: Elsevier. DOI: 10.1016/j.tig.2005.09.003. 11, 69, 83

Yuan, Xiguo, David J Miller, Junying Zhang, David Herrington, and Yue Wang (2012). "An overview of population genetic data simulation." In: *Journal of Computational Biology* 19.1, pp. 42–54. DOI: 10.1089/cmb.2010.0188. 178

Zelkowski, Mateusz, Mischa A Olson, Minghui Wang, and Wojtek Pawlowski (2019). "Diversity and determinants of meiotic recombination landscapes." In: *Trends in Genetics* 35.5, pp. 359–370. DOI: 10.1016/j.tig.2019.02.002. 7

Zhang, Jianzhi (2003). "Evolution by gene duplication: an update." In: *Trends in ecology & evolution* 18.6, pp. 292–298. DOI: 10.1016/S0169-5347(03)00033-8. 35, 36, 39, 147

Ågren, J. Arvid and Stephen I. Wright (2011). "Co-evolution between transposable elements and their hosts: a major factor in genome size evolution?" In: *Chromosome Research* 19.6, pp. 777–786. DOI: 10.1007/s10577-011-9229-0. 113, 120

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede and Ivo Pletikosić. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*".

https://bitbucket.org/amiede/classicthesis/