Luiselli, Juliette

Master Thesis M2 IMaLis

04/01/2021 — 25/06/2021

–

2020 — 2021

École Normale Supérieure, Biology Department

PSL Université Paris

NeGA – Influence of effective population size on genome architecture in eukaryotes

Beagle Team – INRIA, France, Lyon

Supervisor: Beslon, Guillaume, Professor

Inria

1794

ENS | PSL

INSA INSTITUT NATIONAL
DES SCIENCES
APPLIQUÉES
LYON

Abstract

The genome structures of eukaryotes differ radically from prokaryotes in length, content, density and genes structures. Theses differences and their causes are not yet well understood, and a comparative study is practically impossible, because theses realms have diverged too much and because they both comprise a wide diversity. Models are therefore helpful to understand the determinant factors in genome architecture, as they reduce the number of factors taken into account and can determine the relative influence of each factor independently. In this study, I implemented a eukaryotic version of the originally prokaryote AEVOL framework, which is characterized by a diploid genome, sexual reproduction and recombination. This new version of the software allowed us to investigate the effect of both sexual reproduction and recombinations on the genome architecture and the evolution of eukaryotes, as well as their interaction with other determinant evolving factors such as the effective population size (Ne) and the mutation rate (μ). This study is part of the "NeGA" research project, which aims at understanding the influence of effective population size on animal genome architecture. We conclude that recombination has a heavy mutational load, requiring genomes to evolve protections, and entail neutral forces influencing the genome size.

Acknowledgment

I would like to heartfully thank Guillaume BESLON for this opportunity to do my internship under his superivision in the BEAGLE Team and for his flawless support throughout the months despite the hard conditions and the distance. I also want to thank Jonathan ROUZAUD-CORNABAS for his precious advice on the AEVOL code and the usage of the computational resources of INRIA, as well as Théotime GROHENS and Paul BANSE for their helpful advice on the manuscript. I also thank Marco, Laurent, and the whole AEVOL team for their support, advice, and help on the interpretion of my data. Finally, this internship would not have been the same without the material support of the ENS and its numerous coffee machines.

Summary

1 Introduction

Understanding the genomic structure of living organisms has been a key question in biology since the discovery of DNA itself. Indeed, it has long been thought that the amount of DNA would correlate with the organisms complexity, but this later proved to be wrong (C-value paradox) [Thomas, 1971].

Numerous hypotheses have been proposed to draw links between the characteristics and complexity of organisms, and their genomic structure. One of the key immediate descriptor of a genome is its size, which is known in eukaryotes to strongly correlate with the proportion of non-coding DNA and repeated sequences, but only weakly with the number of genes [Charlesworth and Barton, 2004; Lynch and Conery, 2003]. Lynch and Conery observed that genome size is also positively correlated with the level of neutral polymorphism, linked to the product of effective population size Ne, and the mutation rate μ [Lynch and Conery, 2003]. They concluded that selfish genetic elements and non-coding DNA are responsible for slightly deleterious mutations that could not be effectively counter selected, and lead to an increase in genome size. As a matter of fact, a low Ne or μ implies a low underlying diversity and a low selection strength, resulting in a predominance of neutral forces which are supposed to increase the genome size [Lefébure et al., 2017]. Other characteristics are nevertheless thought to influence or be influenced by the genome size, such as the doubling time for a cell, cell volume or the amount and behavior of parasite genomic elements.

Yet, with rare exceptions [Lefébure et al., 2017], the relationship between Ne and genome size has not been demonstrated in comparable taxa, but only on distantly related organisms, with a wide range of possible confounding factors [Charlesworth and Barton, 2004]. This theory therefore remains to be confirmed empirically.

“NeGA” is a research project, coordinated by Tristan Lefébure, which aims to evaluate the influence of effective population size (Ne) on animal Genome Architectures (GA). The project involves teams from experimental and theoretical backgrounds, including the Beagle Team (that develops the Aevol framework), in order to provide data to either support or challenge Lynch’s hypothesis. We propose here to model eukaryotes organisms in Aevol.

Aevol is an individual-based artificial life software that emulates the evolution of bacteria and enables repeated evolution experiments with known and fixed parameters [Knibbe, 2006]. It is an ideal tool to test hypothetical links between genome size and either population size or mutation rate. In particular, the population size can be set by the user, whereas it is difficult to recover and estimate from empirical data [Brevet and Lartillot, 2019]. The main interest of Aevol in the context of NeGA is that it includes an explicit genomic model, each organisms having a full genome that evolves and undergoes transcription and translation. Moreover, the genome size in itself, or non-coding DNA in general, have no direct effect on the fitness of the individuals in the model. We can thus observe the effect of genome size independently of the direct influence of selection for fitness, removing a number of possible confounding factors, and reducing the range of possible causes for the observed relationships between the conditions and the genomic structures. However, a central difficulty is that Aevol is based on a genomic model inspired by prokaryotes, with a unique circular chromosome and clonal reproduction.

A key question for this approach is therefore to highlight the main characteristics of eukaryotes in comparison to prokaryotes, in order to limit the amount of functionalities to implement while still having a model representative of eukaryotic organisms. First of all, sexual reproduction is a determinant key, which probably originated with the eukaryotic taxa [Goodenough and Heitman, 2014; Hofstatter et al., 2020; Michod and Levin, 1988], and is known to have a great influence of genome architecture of organisms [Misevic et al., 2006]. Sexual reproduction is tightly linked to recombination, which is mandatory in meiosis and strongly regulated [Baudat et al., 2013; de Massy, 2013; Haenel et al., 2018; Heyer et al., 2010; Rossignol, 1990]. While the recombination rates vary greatly across taxa, it is strongly regulated: the shortest chromosome always performs one cross-over per generation [Fernandes et al., 2018]. As the homology between sequences is strictly controlled during recombination, investigating eukaryotes in the context of Aevol requires the implementation of a complex recombination machinery, based on homologous sequence recognition, combined with sexual reproduction, meaning the fusion of two gametes from different parents.

A link between population size and genome size has already been observed in prokaryotic models in Aevol (unpublished data): the larger the population is, the shorter the genome. I will here investigate whether this can be generalized to eukaryotic organisms, as they are the subject of the NeGA project. In a first part of the project, this required to adapt Aevol to inlcude eukaryotic characteristics, then to investigate the influence of the population size and mutation rate on the genome size to search for possible links between these data and the mechanisms explaining it.

2 Materials & Methods

2.1 The Aevol framework

Initially developed by Guillaume Beslon and Carole Knibbe [Knibbe, 2006; Parsons et al., 2010], Aevol is an individual-based artificial life software. It emulates a population which is composed of a fixed number of individuals (Fig. 1). Each individual owns a double-stranded genomic sequence, composed of 0s and 1s. In order to compute the phenotype, sequences on the genome are recognized as promoters and mark the beginning of transcription, which stops when a palindromic hairpin structure in encountered. RNAs are then translated into proteic sequences an artificial genetic code with 3-base codons. This determines a mathematical function which is the phenotype of the individual. The distance between this function and a target function, which represents the ideal phenotype in the environment, gives the fitness of the individual.

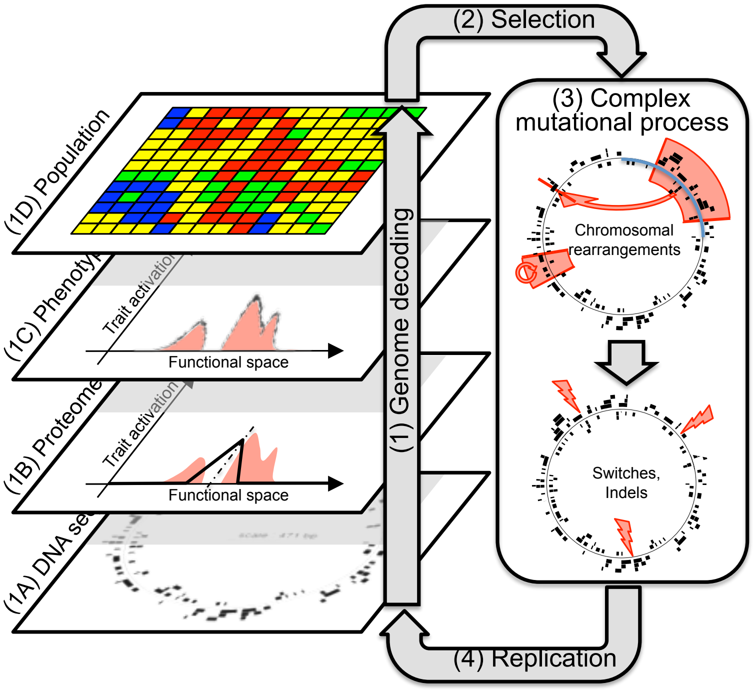

Figure 1: **The Aevol model.** A generation is composed of four steps: (1) Fitness computation from the genome; (2) Competition between neighbors and selection; (3) Occurrence of various mutations in the genome and (4) Replication.

Following a Wright-Fisher model, all individuals are replaced at each generation. The number of descendants of each individual depends on its relative fitness to its neighbors. At each reproduction event, point mutations or genomic rearrangements can occur, which creates diversity in the phenotypes and genomes, and allows the genome size to change due to both coding and non-coding sequences. This combination of variation and selection enables the evolution of these artificial organisms and the emergence of complex phenomena.

Aevol is based on prokaryotes: as illustrated in Fig. 1, genomes in Aevol are composed of a single circular chromosome, and they undergo clonal replication, without exchange of genetic material.

2.2 Construction and implementation of the eukaryotic model

In order to turn Aevol into a model suitable for eukaryotic organisms, I made several profound changes to the software.

As a first step, I introduced the handling of a second chromosome per individual, making the model *de facto* diploid. In this case, all genes are computed in the exact same way, regardless of the chromosome on which they are located. This does not necessarily imply that there are actually genes on both chromosomes, nor that each chromosome is more than 1 base long. Hence, organisms can lose diploidy by removing all genetic material from one of the two chromosomes.

As a next step, I implemented a model of sexual reproduction, in which each organism inherits one random chromosome from each of its parents. No autogamy is allowed in this model, and the parents are selected based on their fitness relative to their neighbors.

Finally, I added a new form of mutation: chromosomal recombinations. One recombination per pair of chromosome per generation is mandatory, and the choice of the recombination point is based on sequence homology: either we find a pair of points above a certain homology score (allelic recombination), or we take the best homology found after a certain number of tries, depending on the total genome size (illegitimate recombination). The recombination therefore depends on two parameters, the target recombination score and search density. Both factors tightly interact: if the density is sufficient with regard to the recombination score, the limit number of tries is rarely reached. In order to calibrate the recombination score, a range of 3 simulations for each of 8 different recombination scores were run during 10 000 generations. The score we chose is a compromise between computation speed (the lower the score, the less time is needed to find a recombination point) and biological accuracy: an individual starting to recombine should not lose fitness, as recombinations are not usually considered deleterious.

Once the score was chosen, we could investigate the impact of both sex and recombination by activating one, the other, or both in our simulations. A fifth option is to have neither sex nor recombination but to have one chromosome that is always a copy of the other, as a reference without possible overshooting.

The code is available on the INRIA gitlab (`https://gitlab.inria.fr/jluisell/aevol`). This version can run up to 8 000 generations per day with 1 024 individuals, while the haploid version of Aevol can run up to 1 million generations per day with 1 024 individuals. This time limitation is due to the homology search in the recombination algorithm and the total absence of clonal organisms, and requires to be parsimonious in the quantity of experiments. The simulations were run on the INRIA Grenoble cluster.

2.3 Generation of a wild type individual

As a starting point for the experiments, we use a “wild type” (WT) organism: the best individual of a population evolved for 10 million generations as a prokaryote in order to be well adapted to its environment. We then copy paste its chromosome to obtain a diploid organism and change the environment accordingly to account for dosage effect, creating what we call the WT1. After that, we let it evolve for 100 000 generations with sex and recombination, so that it is adapted to this new way of life (WT2). The WT1 also underwent a full genome duplication and the corresponding change of the environment, but conserved a single chromosome, remaining prokaryote. To have comparable simulation, this individual also evolved an extra 100 000 generations in this new environment (WT3).

2.4 Experimental design

Starting from WT1, we ran 100 000 generations with (♡) or without (♠) sex and with (✓) or without (✠) recombination. We also run 100 000 generations with neither activated and with one chromosome forced to be a copy of the other, to serve as a comparison.

Starting from a WT2 and WT3, we created 3 types of conditions: a constant population size, a population size divided by 4 or multiplied by 4 (256, 64 and 1 024) and we launched 10 repetitions of each experimental condition. Experiments were run in both the prokaryotic and eukaryotic models, to understand which mechanisms are typical of eukaryotes.

As the selection strength is characterized by both Ne and μ, the same experiment is repeated 10 times with a population of 256 but a point mutation rate 100 times higher than the initial point mutation rate, for 100 000 generations.

For each simulation, data are recorded at each generation on both the best individual and the whole population for a wide range of parameters such as the fitness, the total genome size, the coding and non-coding size of the genome, the number of genes, etc. Note that the first generations for experiments with 1 024 eukaryotes individuals were lost due to a shutdown of the cluster.

2.5 Robustness study

As fitness only considers the coding genome, other metrics are necessary to study the evolution of genome architecture, and especially the evolution of non-coding genome. We use robustness, which measures the probability that the descendants of an individual retain the same fitness as their parent.

In order to estimate the robustness of an individual, we compute randomly 10 000 of its descendants (in case of sexual reproduction, we consider it undergoes self-fecundation) and compare their fitness to the fitness of the focal individual. We define robustness as the percentage of descendants with the same fitness.

2.6 Recombination study

To study recombination further, we draw maps of the recombination events. To draw this map, one organism undergoes 10 000 random recombinations, each one being independent of the others, and record their position as well as the coding state of theses positions. Recombination maps and the distribution of recombinations along the chromosomes can be retrieved from this data. Dotplots for the alignment of one chromosome against another were generated with the BLAST tool of the NCBI (`https://blast.ncbi.nlm.nih.gov/Blast.cgi#`).

3 Results

3.1 Testing the model

Before investigating in details the interaction between population size and genome architecture, the new eukaryotic model has to be thoroughly tested. This also allows some questioning on the functioning of sexual reproduction and recombination in themselves.

3.1.1 The effect of sex

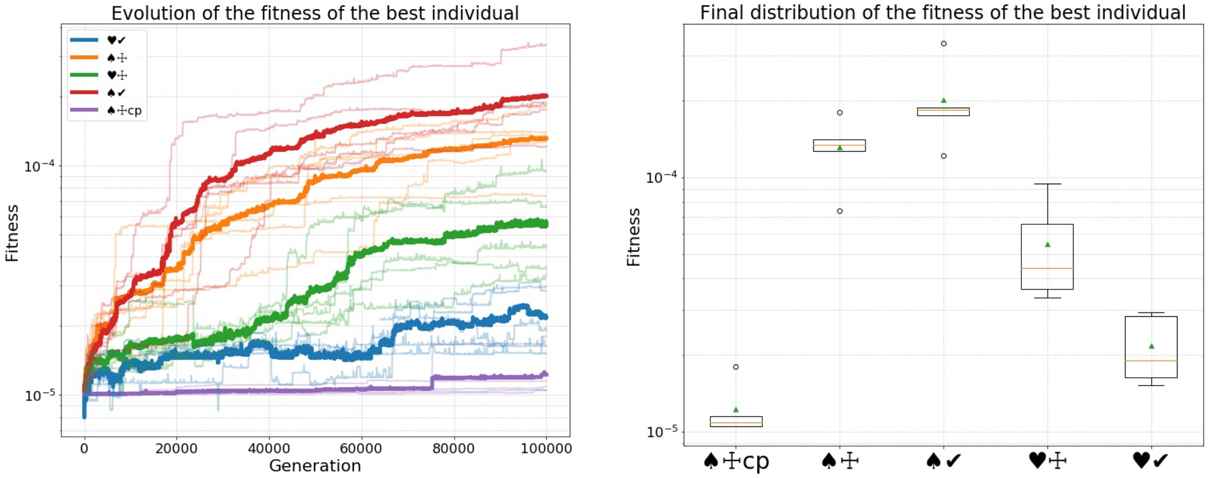

Figure 2: **Overview of the fitness for the 25 experiments testing the model.** Evolution of the fitness of the best individual for 100 000 generations (left), distribution at the end of the simulations (right) for 5 replicates of 5 scenarios: with (♡) or without (♠) sex and with (✓) or without (✠) recombination, and neither with one chromosome being a copy of the other (cp). Values are averaged over 50 generations to reduce noise. In bold is the mean value at each generation.

In the constant environment used for the simulations, sex limits the fitness gains in comparison to the experiments without sex (blue and green, against red and orange, Fig. 2). There is no apparent direct effect of sex on genome size (see Fig. 3).

Looking at the genome structure of the organisms, we observe that organisms that evolved without sex did not conserve their diploid state, so no chromosomes alignments could be undergone. Organisms that evolved with sex did conserve two chromosomes with coding genes, although some temporary losses were observed. Alignments are shown in Appendix A.2.2. We thus deduce that sex is a necessary factor for the maintenance of diploidy.

The difference of fitness between the simulations with and without sex is greater when considering the whole population (Appendix A.2.3) instead of only the best individual (Fig. 2).

This hints towards a lower robustness: the better individuals do not produce enough descendants without loss of fitness. This is confirmed by the robustness tests (see Fig. 4): sex strongly decrease the robustness of the individuals. One possible explanation is overshooting: the best individual of the population, the robustness of which is tested, is better than the others because it owns beneficial heterozygous mutations, but all its homozygous descendants are less fit than itself.

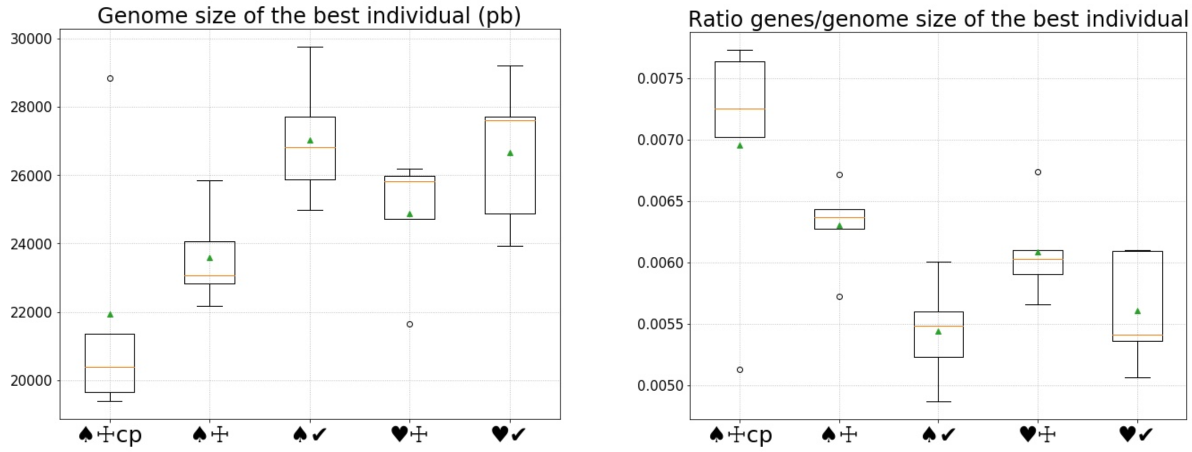

Figure 3: **Overview of the genome content for the 25 experiments testing the model.** Distribution at the end of the simulations for genome size (left) and number of genes over genome size ratio (right) for 5 replicates of 5 scenarios: with (♡) or without (♠) sex and with (✓) or without (✠) recombination and neither with one chromosome being a copy of the other (cp). Values are averaged over 50 generations to reduce noise.

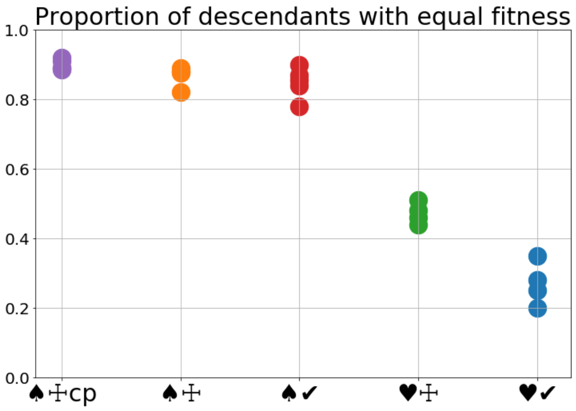

Figure 4: **Measured robustness after 100 000 generations** for 5 scenarios: with (♡) or without (♠) sex and with (✓) or without (✠) recombination and with one chromosome being a copy (cp).

3.1.2 The effect of recombination

In the model, recombination increases the genome size of the individuals: the ♠✓ and ♡✓ simulations have a greater genome size than the others (Fig. 3 left). This concerns mainly the non-coding genome, as the ratio of the number of genes over the total genome size is lower in these simulations (Fig. 3 right).

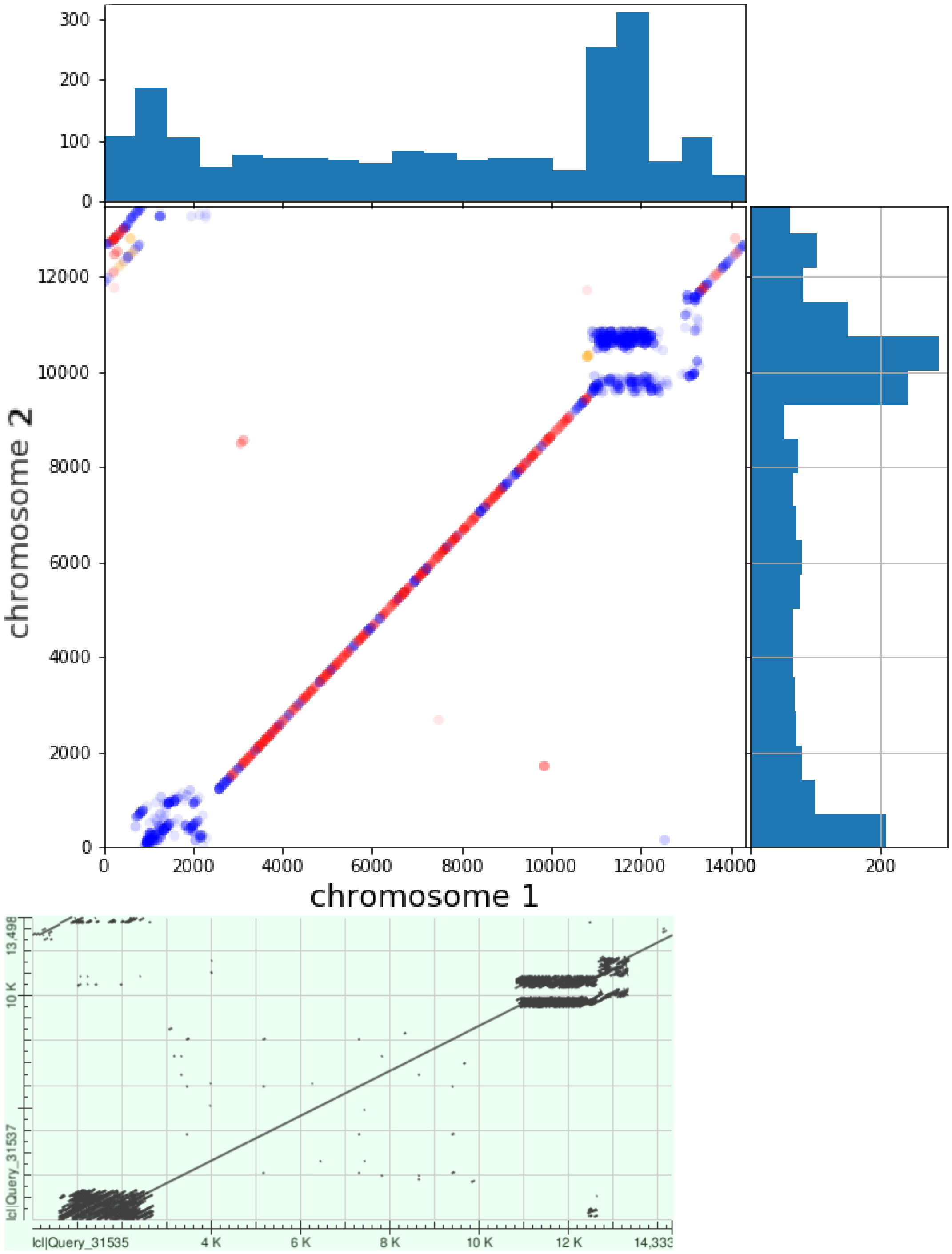

Figure 5: **Example of chromosomic structure.** Top: Localization of 10 000 unique recombination on the best individual of one simulation. Top and right are the density plots of the recombination points. This is a stochastic realization of the alignments found by the algorithm. Points color depends on whether the aligned sequences are coding: either both are non-coding (blue) or one is coding (yelow) or both are coding (red). Bottom is the dotplot associated to this individual computed with BLAST, revealing the existing alignments.

In the absence of recombination, chromosomes poorly align. When both sex and recombination are present, both chromosomes remain balanced, and the alignment is greater (see Appendix A.2.2). Some individuals may temporarily lose this property due to inequal recombination or inter-chromosomic translocation, but our experiments show that they can recover it quickly. It is also expected that unbalanced individuals are counter-selected at the next generation and produce virtually sterile descendants, as their fitness is too low for them to be selected.

When combined to sex, as is the case in real life, we observe that recombinations are deleterious to both fitness and robustness (see Fig. 2 and 4). In such a context, we would expect protection mechanisms to evolve in our experiment to limit these deleterious effects. Strikingly, we observe 1 or 2 hot-spots of recombination in each genome (see Fig. 5) even though the search for homology points is uniformly random in the algorithm. The dotplots show that these hot-spots are composed of short repeated sequences, generally in the non-coding parts of the genomes.

The recombination hot-spots limit the risk of illegitimate recombination, which could lead to unbalanced chromosomes and increase the deleterious effect of sex. Their presence in non-coding part of the genome can be explained by two factors. First, their repetitive structure is a strong constraint on the sequence, preventing their use as genes, but increases the chance that the hot-spot is selected by the algorithm. Moreover, slightly unequal recombinations regularly occur, changing the number of repeated sequences. This would have a high fitness cost within genes.

3.2 Relationship between population size and genome size

3.2.1 Prokaryotes

In this subsection, we use the version of the code without diploidy, sex or recombination to study the effect of a change in population size on genome size.

The populations of 1 024 individuals gain a lot more fitness than populations of 256 individuals, and experiments with 64 individuals globally maintain their fitness (see Fig. 6). Large population experiments have a reduced genome size, and small population experiments an increased genome size. Small population experiments also lose genes, leading to a global decrease in the proportion of coding genome (see Fig. 7). A bigger population induces a stronger selection, which leads in the case of prokaryotes to a reduced genome size. A weaker selection leads to an increased genome size and a decrease in the proportion of coding DNA: the individuals lose proportionally more coding sequences than non-coding sequences.

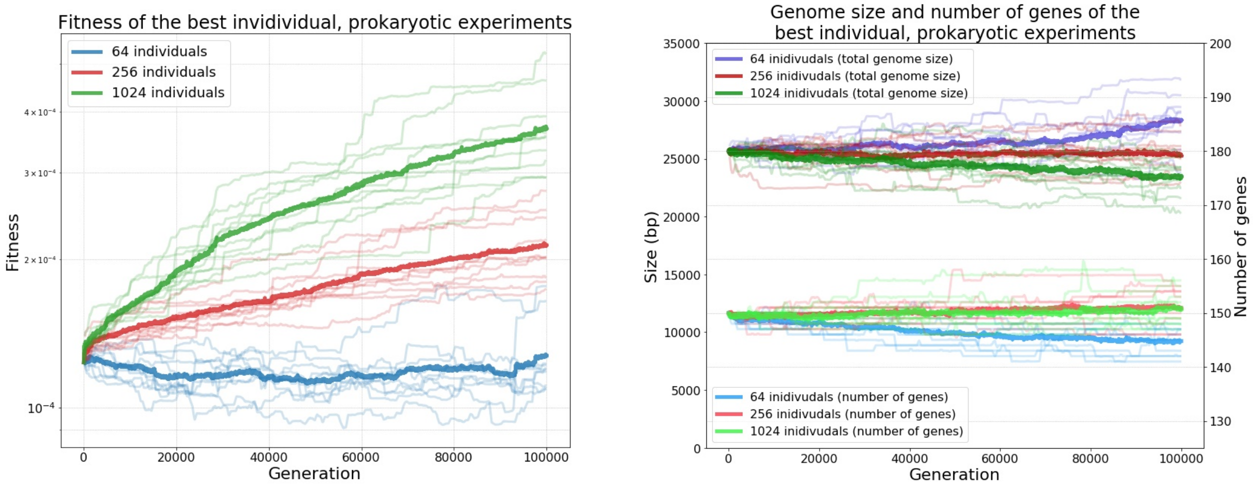

Figure 6: **Overview of the 30 experiments testing the different population sizes.** Fitness (left) and genome size (right) of the best individual for simulations with a population of size 64 (blue), 256 (red) or 1 024 (green), in the case of prokaryotes. Values are averaged over 50 generations to reduce noise. In bold is the mean value at each generation.

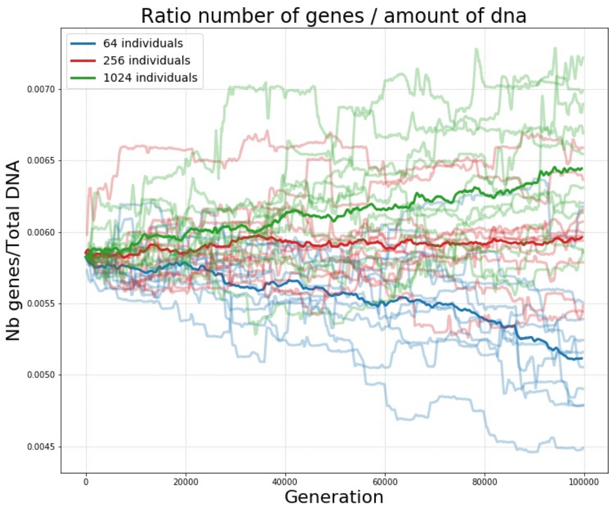

Figure 7: **Ratio of the number of functional genes over total genome size** for simulations with a population size of 64 (blue), 256 (red) or 1 024 (green), in the case of prokaryotes. Values are averaged over 500 generations for better readability. The number of genes is a proxy of the number of coding bases, as the mean length of genes does not vary significantly. Bold trait is the mean value at each generation.

This can be explained by a neutral bias in Aevol toward an increase in genome size. Duplications and deletions occur at the same rate, yet neutral duplications can be larger than neutral deletions (see Appendix A.3) and this lead to a bias for neutral mutations to expand the non-coding genome. This bias is counter-balanced by a selection for robustness, since larger genomes have a lower robustness because of a higher mutation risk (unpublished data). When the population size is lower, the selection pressure toward shorter genomes is released and the non-coding genome size increases. On the opposite, multiplying the population size increases the selection pressure and thus reduces the non-coding portion of the genome.

3.2.2 Eukaryotes

In this subsection, we use the version of the code with both sex and recombination to study the effect of a change in population size on genome size.

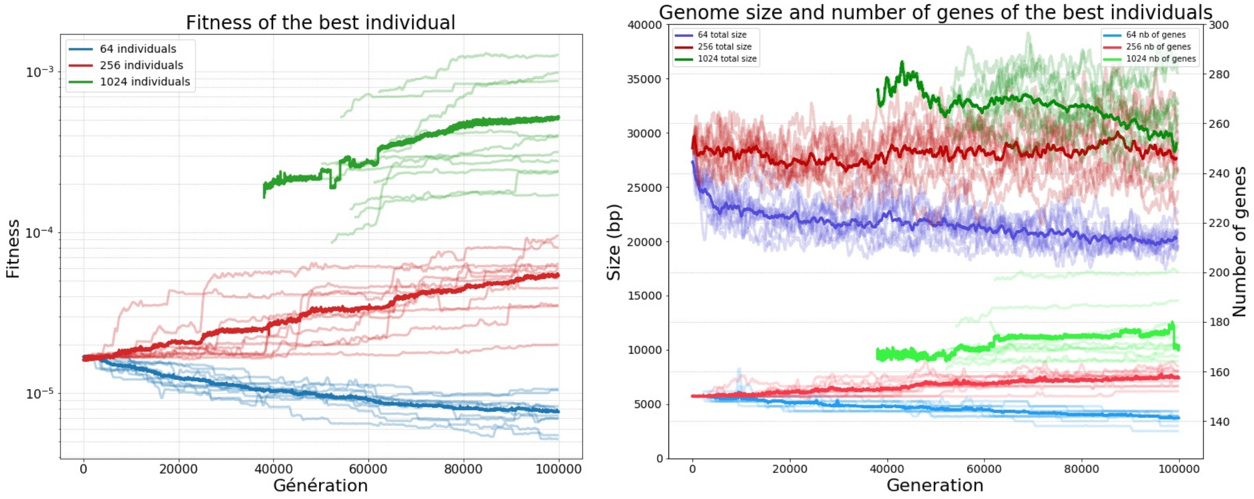

Figure 8: **Overview of the 30 experiments testing the different population sizes.** Fitness (left) and genome size (right) averaged over 500 generations of the best individual for simulations with a population size of 64 (blue), 256 (red) or 1 024 (green), in the case of eukaryotes. Some of the data for the 1 024 populations were lost.

The larger population gains a lot more fitness and has an increase in genome size and number of genes (see Fig. 8), while the smaller population loses fitness and undergoes a decrease in both genome size and number of genes. Comparing the ratio of coding genome size over total genome size, we see that the large population gains both coding and non-coding genome. By contrast, small population experiments lose more non-coding than coding genome compared to the reference experiments (see Fig. 9): the proportion of coding genome rises.

A bigger population induces a stronger selection, which leads in the case of eukaryotes to an increased genome size, contrary to the case of prokaryotes. A weaker selection leads to a decrease in genome size and an increase in the proportion of coding DNA: the individuals lose proportionally more non-coding than coding sequences, which is also contrary to what happens in the prokaryotic case.

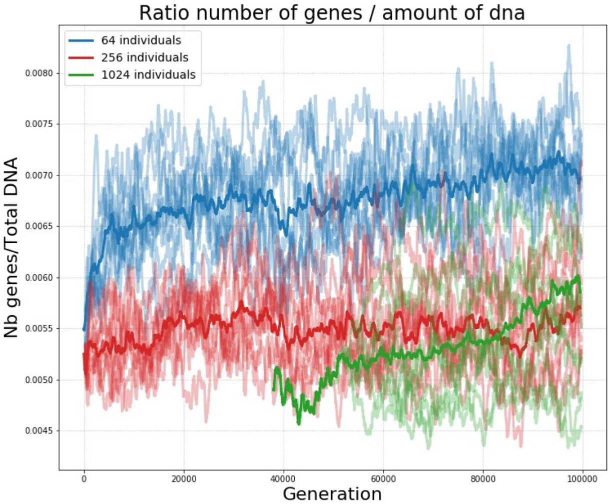

Figure 9: **Ratio of the number of functional genes over total genome size** for simulations with a population size of 64 (blue), 256 (red) or 1 024 (green), in the case of eukaryotes. Values are averaged over 500 generations for better readability. The number of genes is a proxy of the number of coding bases, as the mean length of genes does not vary significantly. Some of the data for the 1 024 populations were lost.

These results are surprising and opposite to what is observed in prokaryotes simulations. Our hypothesis is that the counter-balancing selection for robustness leading to a decrease in genome size has not the same strength in eukaryotes as in prokaryotes. While chromosomal mutations in Aevol have a per base rate, meaning that their number depends on the genome length, this is not the case for recombinations, as there is one, and only one, recombination per generation per reproduction in the model. Having a larger non-coding genome thus protects from a deleterious recombination, as it reduces the probability that coding regions are selected at random and destroyed, without increasing the recombination rate.

3.3 Population size and robustness

According to Fig. 10, individuals have typically a robustness around 0.3 to 0.4, as was observed in Fig. 4 when sex is present. Nevertheless, some outliers are observed in the smaller populations, with a robustness close to 0.9, which is the usual value without sex (see Fig. 4). In this particular case, the absence of effect of sex on robustness implies that both chromosomes are strictly equivalent and can be chosen interchangeably during the sexual reproduction event. The remaining 10 percent of fitness loss are explained by the other regular Aevol mutations and possible illegitimate recombinations.

This also means that in the other cases, the two chromosomes differ: the best individual carries at least one mutation on one of its chromosomes that is beneficial in the heterozygous state (AA' is better than AA), but deleterious, or not as beneficial, in the homozygous state (A'A' is not as good as AA'). As a matter of fact, mutations always appear in the heterozygous state and selection can lead to an increase in their frequency until they have a high probability of being in a homozygous individual, but there is no reason for this homozygous state to be better than the heterozygous one, as the mutation has not been selected on this basis. This phenomenon is called an overshooting mutation.

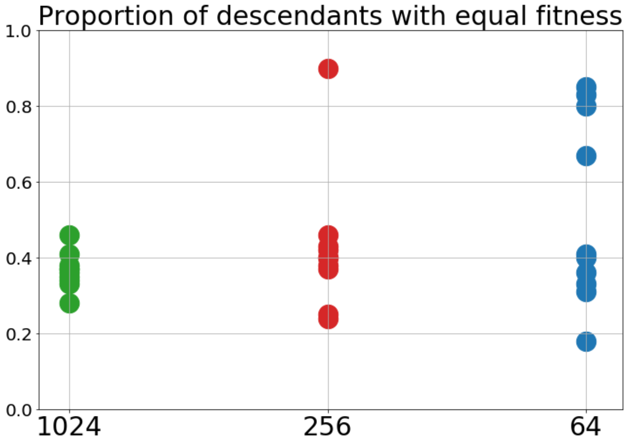

Figure 10: **Robustness for the different population sizes.** Proportion of neutral descendants after 1 generation of selfing for eukaryotic populations of 1 024 (green), 256 (red) or 64 (blue) individuals.

Robustness tests have been run on the best individual of each population. It is not surprising that in populations with overall mutation rate, the best individual is heterozygous as the occurrence of beneficial mutation is more frequent and their selection more efficient. In the smaller population, the best individual often has the same fitness than the rest of the population, and is totally homozygous.

3.4 The impact of the hypermutant phenotype on fitness

Selection pressure depends on both the effective population size Ne and the mutation rate ţ. We expect an increase in ţ to have similar effects than a decrease in population size [Lynch and Conery, 2003].

Fig. 11 shows that a large increase in the point mutation rate reduces the fitness of individuals, in both the prokaryotic and the eukaryotic cases. This loss is smaller in the eukaryotic case, but the initial fitness is not recovered, whereas the prokaryotes lose initially more fitness but recover it quickly.

This shows a certain resistance of eukaryotes to brutal changes, but they are less adaptable in the long term when the new envrionment is constant. Prokaryotes, on the other side, adapt their genome and recover a higher fitness [Rutten et al., 2019].

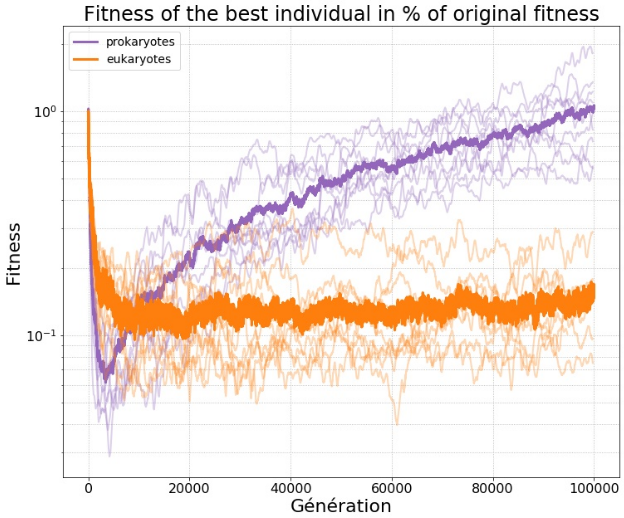

Figure 11: **Fitness of the best individual for the 20 experiments with a 100 times higher point mutation rate**, for eukaryotic and prokaryotic simulations, as a percentage of initial fitness.

4 Discussion

4.1 Sex and recombination as determinant processes of eukaryotic life

In our simulations, diploidy cannot be maintained without both sex and recombination. As a matter of fact, there is no selection pressure to maintain balanced chromosomes, and neutral drift simply deletes one of the two chromosomes. This is consistent with biological results: very complex mechanisms ensure the homology of the sequences at the recombination point [Sun et al., 2020], and sexual reproduction cannot occur if this sequence recognition fails. There is indeed a strong selection pressure on diploidy, which would otherwise be lost.

When diploidy is maintained, we observe that the best individual of the population is most often heterozygous, as deduced from Fig. 10. This is in accordance with the literature on the hybrid vigor, explained by a better action of the combination of two alleles than each of these allele when homozygous [Birchler et al., 2006]. A new mutation is first selected in its heterozygous state, as it appears for the first time on a single chromosome. When the mutation is beneficial, it can be selected and its frequency rises, until it can commonly be homozygous. However, nothing assures that the allele is as beneficial, or beneficial at all, when homozygous (overshooting).

4.2 Recombination hotspots

In the light of the complex biological mechanisms required to control the homology of sequences during recombination and correct possible errors [Sun et al., 2020], it is not surprising that such mechanisms also emerge in Aevol to protect against the potential deleterious effect of recombination. Indeed, this process appears to be tightly controlled both in biological life and in Aevol.

In Aevol, recombination locations concentrate in the short repeated sequences. Indeed, a point drawn in this area can match several points on the homologous chromosome, increasing the chances of finding a good alignment in this area of the chromosome, as shown in Fig 5.

Recombination in itself is not deleterious when happening on perfectly homologous sequences. Nevertheless, mismatchs or shifts can happen, having a potential deleterious effect when they are in coding sequences. Observing hotspots preferably in non-coding sequences in Aevol is therefore not surprising. Yet, hotspots are inequally distributed on the chromosome among species: in non-coding subtelomeric regions for mice and humans, but mostly in genes for maize [Lichten and Goldman, 1995; Mackiewicz et al., 2013]. The location on the chromosome cannot be studied in Aevol because the chromosomes are circular, but the fact that hotspots emerge in non-coding regions questions the reasons explaining the presence of recombination hotspots in genes.

The mechanism observed in Aevol most likely protects the organism from deleterious recombination. By comparison, it is likely that the strict control of recombination during meiosis also protects the organisms from deleterious recombination, an hypothesis worth studying in living organisms. Recombinations are necessary for the maintenance of diploidy, but they are highly deleterious and highly controlled, meaning that they are dangerous. This raises a recombination paradox, which can be merge with the sex paradox: why is diploidy so widespread when it is so costly to maintain ? Under what conditions do the benefits outweigh the costs ?

4.3 Genome size variations

In the original prokaryotic version of Aevol, non-coding sequences are, to some extent, selected against due to selection for robustness [Fischer et al., 2014; Knibbe et al., 2007]. Even if, in the model, a non-coding sequence does not cost in itself anything to the organism, a larger genome undergoes more rearrangements, as rearrangements occur on a per base rate: natural selection does not only select on fitness, but also on robustness [Liard et al., 2020; Singhal et al., 2019; Wilke et al., 2001]. Hence, selection tend to reduce the size of the genome. Drift also plays a role, as neutral mutations tend to increase genome size (unpublished Aevol data). An equilibrium point exists between both these selective forces and drift, stabilizing genome size. This is why, in our prokaryotic experiments, increasing selection reduces the genome and decreasing selection increases the genome size through the proliferation of non-coding sequences. This is in accordance with the literature on the subject [Lynch, 2002; Lynch and Conery, 2003; Yi and Streelman, 2005].

In the eukaryotic version of Aevol developed in this study, the number of recombinations is fixed to one per genome per generation. Increasing genome size still increases the frequency of the other rearrangements, but not the frequency of recombinations. Thus, it is not surprising that the genome size equilibrium is higher when recombination is present: the selection for non-coding sequences to protect against mutation is proportionally stronger in a longer genome, while the selection against non-coding sequences remains the same. When selection is relaxed (64 individuals), non-coding sequences are lost and the proportion of coding genome rises, whereas when selection is strong the proportion of coding sequences decreases. This behavior is opposite to what is observed in the literature, hinting that some determinant mechanisms underlying the regulation of genome size are missing in our model.

4.4 Muller's ratchet hypothesis

Muller's ratchet postulates that in a finite population, without sex and recombination, deleterious mutations will accumulate through drift and will eventually reduce the fitness of the population, until its extinction [Muller, 1950]. This process is enhanced by a small population size [Kimura et al., 1963]. Sex and recombination should be able to avoid Muller's ratchet because they allow the population to purge deleterious mutations [Barton and Charlesworth, 1998].

Yet, we observe here that experiments with 64 individuals lose fitness, in what seems to be a realization of Muller's ratchet. In our experiments, sex and recombination are not able to stop Muller's ratchet, probably because they carry their own mutational load: they are deleterious in themselves and cannot necessarily counter-balance this load by purging other mutations. Sex and recombination even amplify Muller's ratchet in our simulations, as the prokaryotic simulations with 64 individuals do not lose fitness significantly, meaning that the mutational load of sex and recombination is heavy with our model. This observation is enhanced by the fact that we compare here a similar number of individuals, instead of a similar number of chromosomes (128 chromosomes in the eukaryotic case). However, the amount of DNA in each individual is of the same order of magnitude in both experiments. One hypothesis is that sex and recombination are in fact a protection against Muller's ratchet, but some additional conditions have to be fulfilled and are not taken into account in our model.

The question of Muller's ratchet is also relevant in the case of hypermutants: the purge of deleterious mutations should preserve the fitness of the individuals in the case of sexual reproduction, while asexual reproduction is thought to be more penalized. This is not really shown in our data, suggesting that some more parameters have to be considered to explain Muller's ratchet and what protects from it. Prokaryotes better adapt themselves in the long term when the environment remains constant, probably by reorganizing their genome to adapt to this hypermutant phenotype, as observed by Rutten et al. [2019].

4.5 Circular or linear chromosomes

All our experiments, for eukaryotes or prokaryotes, were performed on circular chromosomes, as in the classical Aevol version. This modeling choice avoids many biases for most mutations, and changing it would have probably required many more weeks or months spent on the code.

Yet, this choice is questionable regarding recombination: recombining circular chromosomes requires the selection of two break points instead of just one. It also has an impact on the sense (direct or indirect) of alignment, as both alignments must have the same direction[1]. More specifically, locations of the emerging recombination hotspots on the chromosomes are meaningless as there is no begin or middle of the chromosome, while they would be more interesting in the case of linear chromosomes.

However, we argue that our results on robustness, overshooting, genome size, or even the emergence of recombination hotspots in themselves are quite independent from the shape of chromosomes. Circular chromosomes assure minimal biases in mutations and avoid problematic edge effects, without creating obvious obstacles to the processes under study.

5 Perspectives

A direct application of this research is to integrate it into the heart of the project NeGA ("Influence of effective population size (Ne) on animal Genome Architectures"), a collaboration of 4 research teams including both artificial and real organisms. The eukaryotic Aevol model allows us to compare populations of different sizes, in a perfectly controlled manner, contrary to *in vivo* experiments, and thus to study the impact of genetic drift on the evolution of genome architecture. Here, we investigated the model in itself and ensured its proper functioning, but a detailed comparison with empirical data still has to be done.

Sex slowing down the fitness gain in a constant environment was expected, as it highlights the well studied sex paradox: how could sexual reproduction evolve and maintain itself, as it is so costly ? Sex is expected to be beneficial in a varying environment [Misevic et al., 2010], or under specific conditions [Otto and Lenormand, 2002], but not especially in a

[1] This specific problem was not taken into account in the simulations, but showed little impact on the results.

constant environment. Yet, this implementation of an eukaryotic model in Aevol open perspectives for the study of sexual reproduction in itself in various environment and timescales. It could allow us to study the conditions under which the benefits of sexual reproduction outweigh its costs, or which prevent Muller's ratchet in small populations or with a high mutation rate.

Finally, the study of recombination could be pursued with a more complex model, this work being a first step. An obvious improvement would be to introduce a per base recombination rate (with a basal value of 1 per pair of chromosome), and introduce a varying number of pairs of chromosomes. This could change the bias toward the increase of non-coding sequences as a longer genome would undergo more recombination events, and reflect the literature results more. Yet, a key point to have a more realistic model of recombination would be to linearize the chromosomes to allow for the selection of a single recombination point. However, this would be a huge investment in development time, as all mutations in Aevol would have to be rethought and recoded to avoid biases.

6 Conclusion

The internship allowed me to develop of a fully functional eukaryotic model in the Aevol framework: a diploid organism undergoing recombination and performing sexual reproduction. Using this new framework, I have been able to investigate the impact of both sex and recombination on fitness and robustness, highlighting the effect of Muller's ratchet in sexual populations and the presence of overshooting mutations. I also shed light on some of the mechanisms linking the effective population size and the genome size of organisms, such as the mutational bias induced by the mandatory character of recombination and its associated mutational load.

The framework I developed can be refined to investigate more complex characteristics of eukaryotes, such as the theoretical consequences of the linear shape of chromosomes. In the context of the "NeGA" project, this framework can also be used to study in detail the relationship between effective population size and genome size in a wide variety of conditions, and with different constraints on evolution.

References

N. H. Barton and B. Charlesworth. Why sex and recombination? *Science*, 281(5385): 1986–1990, 1998.

F. Baudat, Y. Imai, and B. de Massy. Meiotic recombination in mammals: localization and regulation. *Nature Reviews. Genetics*, 14(11):794–806, 2013. ISSN 1471-0064. doi: 10.1038/nrg3573.

J. A. Birchler, H. Yao, and S. Chudalayandi. Unraveling the genetic basis of hybrid vigor. *Proceedings of the National Academy of Sciences*, 103(35):12957–12958, 2006. ISSN 0027-8424. doi: 10.1073/pnas.0605627103.

M. Brevet and N. Lartillot. Reconstructing the history of variation in effective population size along phylogenies. *bioRxiv*, page 793059, 2019. doi: 10.1101/793059.

B. Charlesworth and N. Barton. Genome size: Does bigger mean worse? *Current Biology*, 14(6):R233–R235, 2004. ISSN 0960-9822. doi: https://doi.org/10.1016/j.cub.2004.02.054.

B. de Massy. Initiation of meiotic recombination: How and where? conservation and specificities among eukaryotes. *Annual Review of Genetics*, 47(1):563–599, 2013. ISSN 0066-4197. doi: 10.1146/annurev-genet-110711-155423.

J. B. Fernandes, M. Séguéla-Arnaud, C. Larchevêque, A. H. Lloyd, and R. Mercier. Unleashing meiotic crossovers in hybrid plants. *Proceedings of the National Academy of Sciences*, 115(10):2431–2436, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1713078114.

S. Fischer, S. Bernard, G. Beslon, and C. Knibbe. A Model for Genome Size Evolution. *Bulletin of Mathematical Biology*, 76(9):2249 – 2291, 2014. doi: 10.1007/s11538-014-9997-8.

U. Goodenough and J. Heitman. Origins of eukaryotic sexual reproduction. *Cold Spring Harbor Perspectives in Biology*, 6(3):a016154, 2014. ISSN , 1943-0264. doi: 10.1101/cshperspect.a016154.

Q. Haenel, T. G. Laurentino, M. Roesti, and D. Berner. Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics. *Molecular Ecology*, 27(11):2477–2497, 2018. ISSN 1365-294X. doi: https://doi.org/10.1111/mec.14699.

W.-D. Heyer, K. T. Ehmsen, and J. Liu. Regulation of homologous recombination in eukaryotes. *Annual Review of Genetics*, 44(1):113–139, 2010. doi: 10.1146/annurev-genet-051710-150955. PMID: 20690856.

P. G. Hofstatter, G. M. Ribeiro, A. L. Porfírio-Sousa, and D. J. G. Lahr. The sexual ancestor of all eukaryotes: A defense of the "meiosis toolkit". *BioEssays*, 42(9):2000037, 2020. ISSN 1521-1878. doi: https://doi.org/10.1002/bies.202000037.

M. Kimura, T. Maruyama, and J. F. Crow. The mutation load in small populations. *Genetics*, 48(10):1303, 1963.

C. Knibbe. *Structuration des génomes par sélection indirecte de la variabilité mutationnelle: une approche de modélisation et de simulation.* PhD thesis, INSA de Lyon, 2006.

C. Knibbe, A. Coulon, O. Mazet, J.-M. Fayard, and G. Beslon. A long-term evolutionary pressure on the amount of noncoding dna. *Molecular biology and evolution*, 24(10):2344–2353, 2007.

T. Lefébure, C. Morvan, F. Malard, C. François, L. Konecny-Dupré, L. Guéguen, M. Weiss-Gayet, A. Seguin-Orlando, L. Ermini, C. D. Sarkissian, N. P. Charrier, D. Eme, F. Mermillod-Blondin, L. Duret, C. Vieira, L. Orlando, and C. J. Douady. Less effective selection leads to larger genomes. *Genome Research*, 27(6):1016–1028, 2017. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.212589.116.

V. Liard, D. P. Parsons, J. Rouzaud-Cornabas, and G. Beslon. The complexity ratchet: Stronger than selection, stronger than evolvability, weaker than robustness. *Artificial life*, 26(1):38–57, 2020.

M. Lichten and A. S. Goldman. Meiotic recombination hotspots. *Annual review of genetics*, 29(1):423–444, 1995.

M. Lynch. Intron evolution as a population-genetic process. *Proceedings of the National Academy of Sciences*, 99(9):6118–6123, 2002. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.092595699.

M. Lynch and J. S. Conery. The origins of genome complexity. *Science*, 302(5649):1401–1404, 2003. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1089370.

D. Mackiewicz, P. M. C. d. Oliveira, S. M. d. Oliveira, and S. Cebrat. Distribution of recombination hotspots in the human genome – a comparison of computer simulations with real data. *PLOS ONE*, 8(6):e65272, 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0065272.

R. Michod and B. Levin. The evolution of sex. *Letters to the Editor*, 147:151, 1988.

D. Misevic, C. Ofria, and R. E. Lenski. Sexual reproduction reshapes the genetic architecture of digital organisms. *Proceedings of the Royal Society B: Biological Sciences*, 273(1585):457–464, 2006. doi: 10.1098/rspb.2005.3338.

D. Misevic, C. Ofria, and R. E. Lenski. Experiments with digital organisms on the origin and maintenance of sex in changing environments. *Journal of Heredity*, 101:S46–S54, 2010. ISSN 0022-1503. doi: 10.1093/jhered/esq017.

H. J. Muller. Our load of mutations. *American journal of human genetics*, 2(2):111, 1950.

S. P. Otto and T. Lenormand. Resolving the paradox of sex and recombination. *Nature Reviews Genetics*, 3(4):252–261, 2002. ISSN 1471-0064. doi: 10.1038/nrg761.

D. Parsons, C. Knibbe, and G. Beslon. Aevol: un modèle individu-centré pour l'étude de la structuration des génomes. *MajecSTIC*, 10 2010.

J.-L. Rossignol. La recombinaison homologue : mecanismes et consequences. *Société Française de Génétique*, 6:6, 1990.

J. P. Rutten, P. Hogeweg, and G. Beslon. Adapting the engine to the fuel: mutator populations can reduce the mutational load by reorganizing their genome structure. *BMC evolutionary biology*, 19(1):1–17, 2019.

S. Singhal, S. M. Gomez, and C. L. Burch. Recombination drives the evolution of mutational robustness. *Current Opinion in Systems Biology*, 13:142–149, 2019. ISSN 2452-3100. doi: 10.1016/j.coisb.2018.12.003.

Y. Sun, T. J. McCorvie, L. A. Yates, and X. Zhang. Structural basis of homologous recombination. *Cellular and Molecular Life Sciences*, 77(1):3–18, 2020.

C. A. Thomas. The genetic organization of chromosomes. *Annual Review of Genetics*, 5(1): 237–256, 1971. doi: 10.1146/annurev.ge.05.120171.001321.

C. O. Wilke, J. L. Wang, C. Ofria, R. E. Lenski, and C. Adami. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844):331–333, 2001.

S. Yi and J. T. Streelman. Genome size is negatively correlated with effective population size in ray-finned fish. *Trends in Genetics*, 21(12):643–646, 2005. ISSN 0168-9525. doi: https://doi.org/10.1016/j.tig.2005.09.003.

Appendix

A.1 List of abbreviations

- GA = Genome Architecture
- Ne = Effective population size
- WT = Wild Type. An individual evolved for several millions of generations in a constant environment, and thus well adapted to it.

A.2 Supplementary data on the test of the model

A.2.1 Parameterization of the model

According to the test simulations, a recombination score of 60 is the lowest for which the best individual does not lose fitness when recombination is introduced. As a matter of fact, there are fitness loss around 30 and 50 generations for a score of 55, and none for 60 (see Fig. 12).

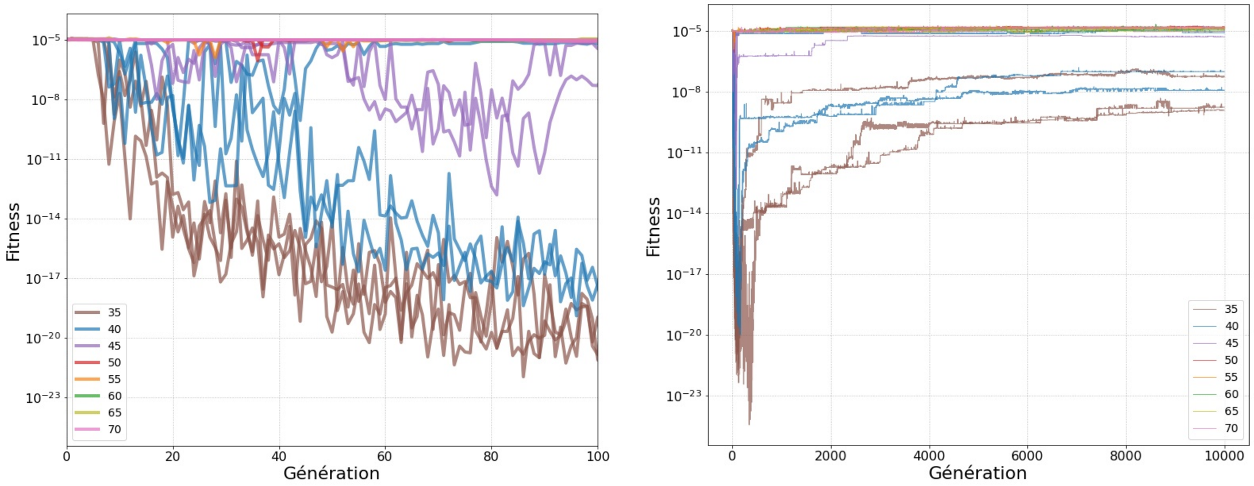

Figure 12: Fitness of the best individual across time for different recombination scores. Full run of 10 000 generations (right), and a zoom on the first 100 generation for more details (left).

A.2.2 Chromosome alignments with and without recombination

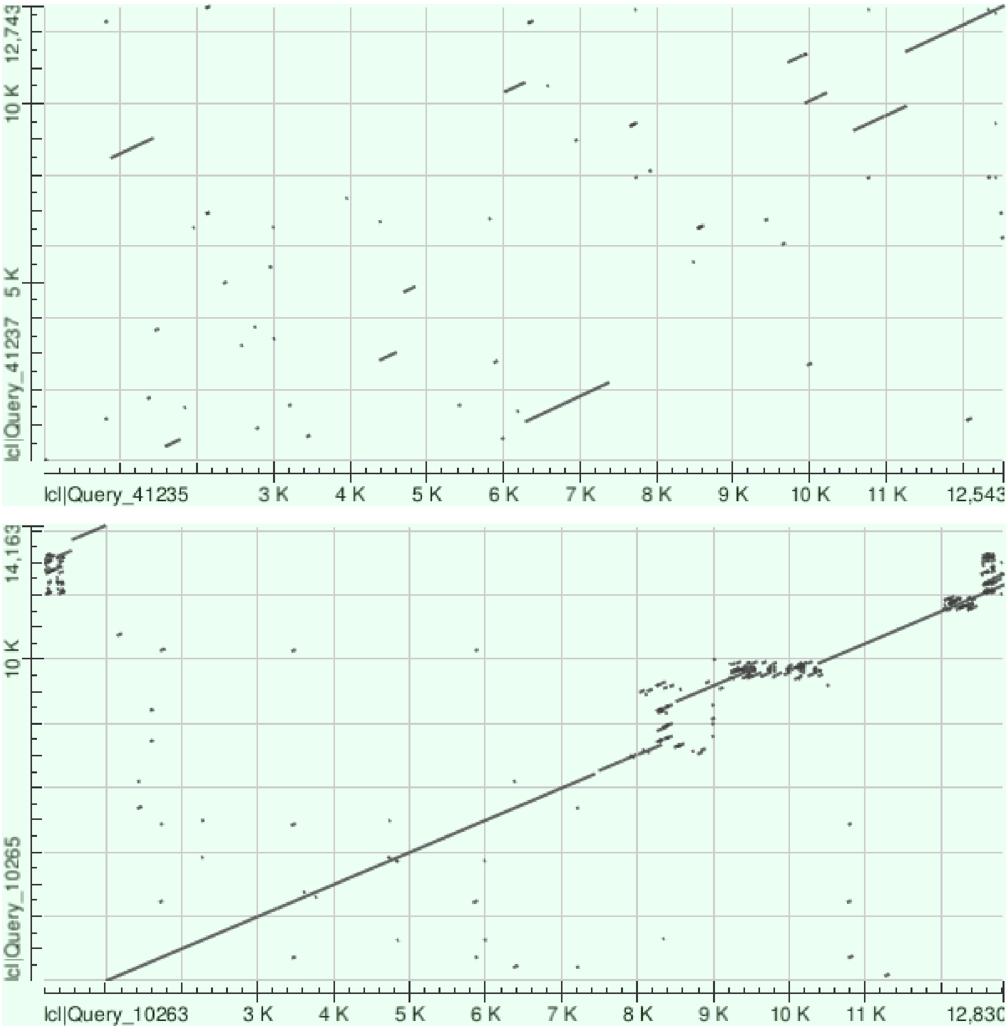

Figure 13: Example of a chromosome 1 VS chromosome 2 dotplot of the best individual for a simulation with sex but no recombination (top) and 1 simulation with both sex and recombination (bottom). Balanced chromosomes recovered from temporarily highly unbalanced individuals.

A.2.3 Global fitness

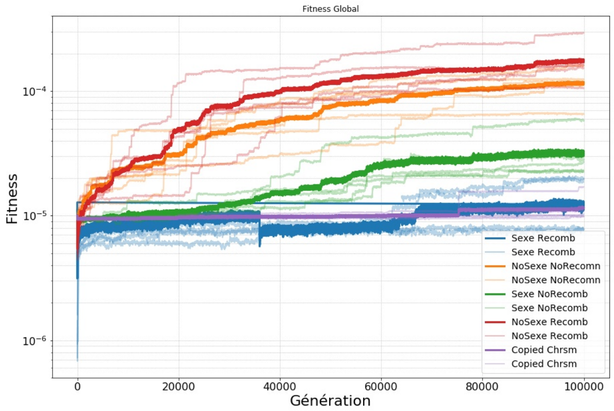

Figure 14: Mean fitness of the population across time for 5 replicates of 5 scenarios: sex and/or recombination and neither with one chromosome being a copy of the other

A.3 Mutational bias toward the increase in genome size

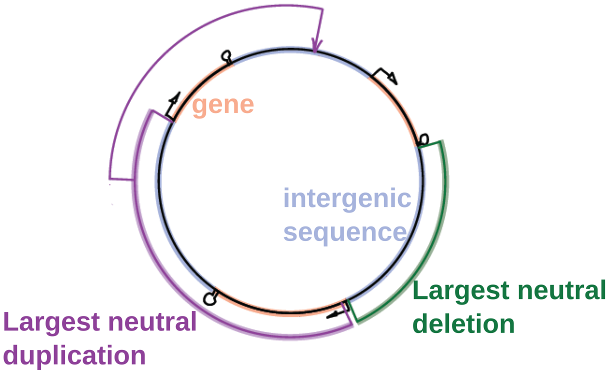

Figure 15: **Largest neutral deletion and largest neutral duplications in a model genome**. The difference of sizes between neutral deletions and neutral duplications explains the presence of a bias toward an increase in genome size. Figure from Marco Foley.