Genome streamlining: effect of mutation rate and population size on genome size reduction

Abstract

Genome streamlining, *i.e.* genome size reduction, is observed in bacteria with very different life traits, including endosymbiotic bacteria and several marine bacteria, raising the question of its evolutionary origin. None of the hypotheses proposed in the literature is firmly established, mainly due to the many confounding factors related to the diverse habitats of species with streamlined genomes. Computational models may help overcome these difficulties and rigorously test hypotheses. In this work, we used Aevol, a platform designed to study the evolution of genome architecture, to test two main hypotheses: that an increase in population size (N) or mutation rate (μ) could cause genome reduction. Pre-evolved individuals were transferred into new conditions, characterized by an increase in population size or mutation rate. In our experiments, both conditions lead to streamlining. However, they lead to very different genome structures. Under increased population size, genomes loose a significant fraction of non-coding sequences, but maintain their coding size, resulting in densely packed genomes (akin to streamlined marine bacteria genomes). By contrast, under increased mutation rate, genomes loose coding and non-coding sequences (akin to endosymbiotic bacteria genomes). Hence, both factors lead to an overall reduction in genome size, but the coding density of the genome appears to be determined by $N \times \mu$. Thus, a broad range of genome size and density can be achieved by different combinations of N and μ. Our results suggest that genome size and coding density are determined by the interplay between selection for phenotypic adaptation and selection for robustness.

Keywords : Genome Architecture, Genome Evolution, Genome Streamlining, Mutation rate, Modeling, Population size

Genome streamlining: effect of mutation rate and population size on genome size reduction

Juliette Luiselli*[,1,2], Jonathan Rouzaud-Cornabas[1,2], Nicolas Lartillot[3,4] and Guillaume Beslon[1,2]

February 2024

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 (Gago et al., 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 largely reduced 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 in fact 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 some 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, 2007). In particular, it has been suggested that population genetics mechanisms could explain the reductive evolution observed in several bacterial strains (Lynch, 2006). 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 some free-living marine bacteria such as *Prochlorococcus marinus* (Dufresne et al., 2005) or *Pelagibacter ubique* (Giovannoni et al., 2005). Strikingly, both types of bacteria lies 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 constant 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 less coding genes than average bacteria (Heddi et al., 1998). Endosymbionts have typically lost both coding and non-coding genomic content (Moran and Mira, 2001; Wernegreen, 2002), leading to a coding fraction of the order of 85% (van 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 (Giovannoni et al., 2005; Batut et al., 2014), but are believed to have very large effective population sizes, undisputedly larger than endosymbionts (Marais et al., 2008; Flombaum et al., 2013; Giovannoni et al., 2014), although to what extent remains questioned (Chen et al., 2022). Noticeably, in their case, genome size reduction is primary contributed by the loss of non-coding sequences rather than coding sequences (Giovannoni et al., 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 a rise in mutation rate (Bourguignon et al., 2020) and either an increase or a decrease in effective population size N_e. While some observations link the decrease in genome size to the increase in random drift (Moran, 2002; Kuo et al., 2009), this is not consensual among the scientific community, as 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 quasi-neutral duplications, which would be more frequent than quasi-neutral deletions (Lynch and Conery, 2003; Lefebure et al., 2017).

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 (Popadin et al., 2007; Romiguier et al., 2012; Figuet et al., 2016), however, the precise quantitative relation between these proxies and effective population size is unknown. Moreover, the very different living conditions of the bacterial species that have undergone genome reduction introduce many

confounding factors. To avoid these pitfalls, we choose to turn to simulation to study the different phenomena leading to genome size reduction, and more specifically the influence of population size and mutation rate over the coding and non-coding genome sizes.

In silico experimental evolution provides tools to study in detail the genomic architecture (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. Then, running simulations in a perfectly controlled environment covering a broad range of populations 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 which 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., 2007; Banse et al., 2023). 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 conditions lead to genome size reduction. However, they lead to very different genome structure, as a high μ reduces both coding and non-coding compartments, while a high N reduces only the non-coding compartment, resulting in densely packed genomes. Surprisingly, they nevertheless 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 genome structure. Finally, by measuring the robustness of the lineages evolved in different conditions, we show that the observed variation in genome size and structure are due to the interaction between selection for phenotypical adaptation to the environment and selection for robustness.

2 Results

We perform our experiments using Aevol, a forward-in-time evolutionary simulator (Knibbe et al., 2007; Banse et al., 2023). 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 from 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 maps 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 locus, length or the intergenic distances, hence the total amount of non-coding DNA. While genes can be duplicated and diverge (Kalhor et al., 2023), they can also be deleted, and some emerge *de novo*. Hence, the impact of a given mutation highly depends on the pre-existing genome structure, which can in turn be indirectly selected (Knibbe et al., 2007). 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 millions 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. The five WTs display stable genome structures (with small random variations, as exemplified by cases N_0 and μ_0 on Figs. 1 and 2) although they still slowly gain fitness by fixing rare favorable mutations (see case N_0 on Fig. 4A). Their fitness and genomic characteristics are displayed in section 4.2, Table 1. In our experiments, these WTs are used as founders of new populations which are then confronted to new evolutionary conditions for 2 millions generations. In parallel, these same WTs were evolved in constant conditions, providing perfect control experiments. We then compare the fitness, genome size and genome structure of populations evolved in new conditions with those of the control populations.

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

2.1.1 Change in population size

Increasing the population size by a factor 4 or 16 results in a reduction in the total genome size (see Fig. 1A), yet this change does not impact proportionally the coding and non-coding parts of the genome: while the size of the coding compartment is barely affected (see Fig. 1B), the non-coding genome size is greatly reduced (see Fig. 1C). As a result, the proportion of coding genome increases (see Fig. 1D). Conversely, reducing the population size by a factor 4 or 16 increases the total genome size (Fig. 1A), by increasing greatly the non-coding genome size (Fig. 1B) and reducing slightly the coding genome size (in the extreme condition $N_0/16$, Fig. 1C), thus drastically reducing the coding fraction of the genome (Fig. 1D).

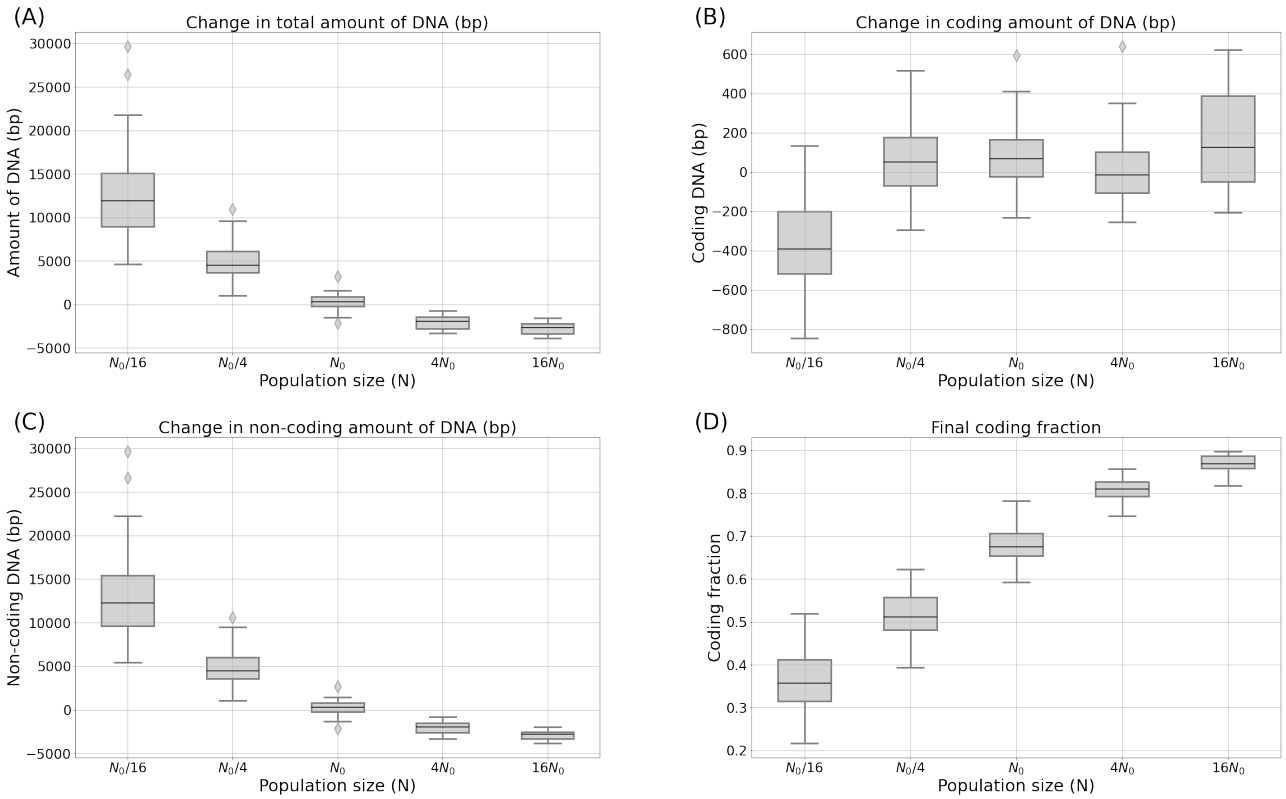

Figure 1: Total (A), coding (B) and non-coding (C) genome size variation, and final coding fraction (D), after 2 millions generations. For each of the 5 WTs, 10 repetitions were performed under a constant mutation rate ($\mu_0 = 10^{-6}$ per base pair for each type of mutation) with 5 different population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

2.1.2 Change in mutation rate

Increasing the mutation rate drastically reduces the total genome size (see Fig. 2A). Thus, at first sight, population size and mutation rate seem to have a similar impact on genome evolution. However, in the details, the impact 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 Fig.2B and C). Both compartments are nevertheless not proportionally impacted by the mutation rate μ: the decrease affects more strongly the non-coding part of the genome, such that the final coding fraction of the genome increases with μ (see Fig. 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.

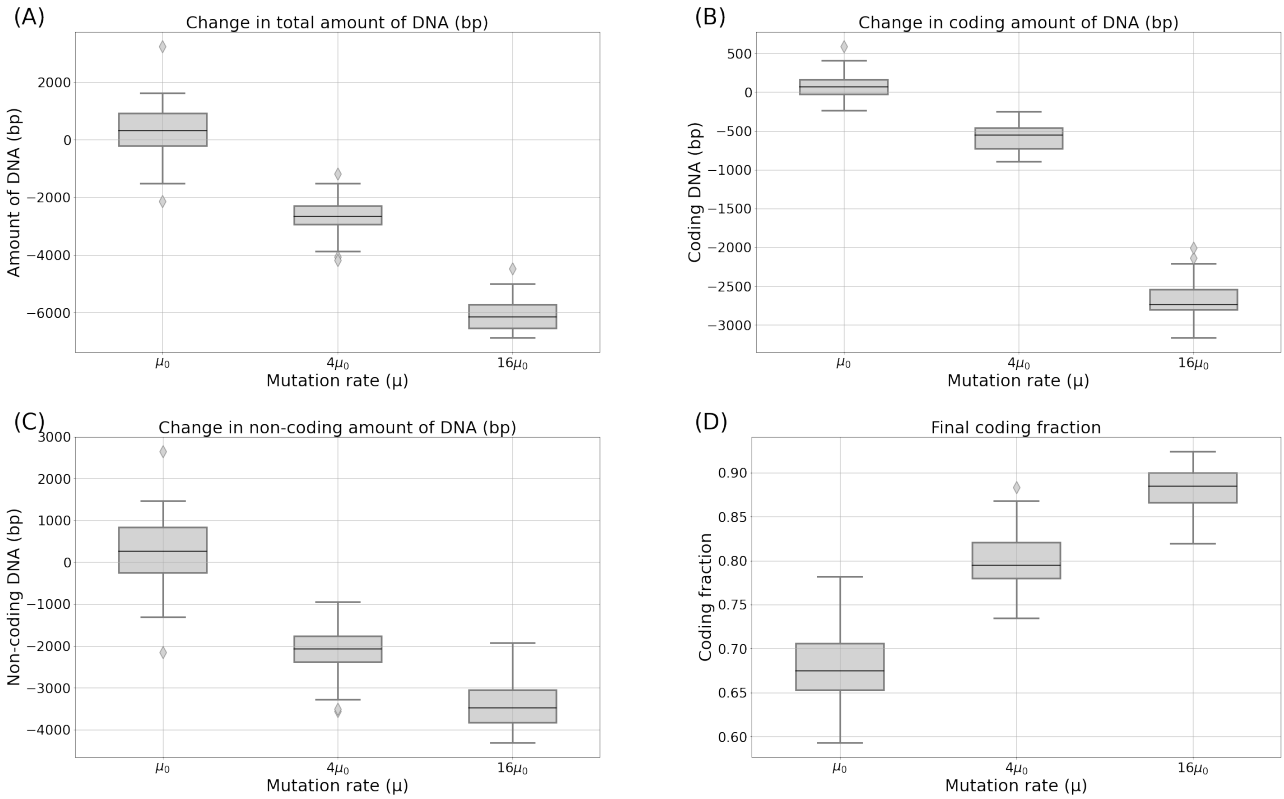

Figure 2: Total (A), coding (B) and non-coding (C) genome size variation, and final coding fraction (D), after 2 millions generations. For each of the 5 WTs, 10 repetitions were performed under a constant population size ($N_0 = 1024$ individuals) with 3 different mutation rates: the control $\mu_0 = 10^{-6}$ mutations per base pair for each type of mutation, $4 \times \mu_0$ and $16 \times \mu_0$.

Notably, and in spite of 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.

2.1.3 Linked effect of population sizes and mutation rates

Figure 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 all subfigures, the bottom line and the central column correspond to the values presented in Figs. 1 and 2).

Overall, as N increases, the total amount of DNA decreases, and this, whatever the value of mu (see Fig. 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 Fig. 3B). This is counterbalanced by the change in the non-coding size of the genomes (see Fig. 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 Fig. 3D). Indeed, we observe that, under constant $N \times \mu$, and despite the fact that these two factors taken individually have changed in various 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 Fig. 3D). Although the coding fraction varies slightly (from 68% to 63%) for the most extreme change tested ($N_0/16$ and $16\mu_0$), the diagonals of constant $N_0 \times \mu_0$ also display an almost constant coding fraction (Fig. 3D).

However, strikingly, the total genome size as well as the coding and non-coding genome sizes vary greatly, even for similar coding densities (Fig. 3B, C and D). For densities of 63% and 65%, we can halve the total amount of DNA (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 combined variation in both the population size N and the mutation rate μ.

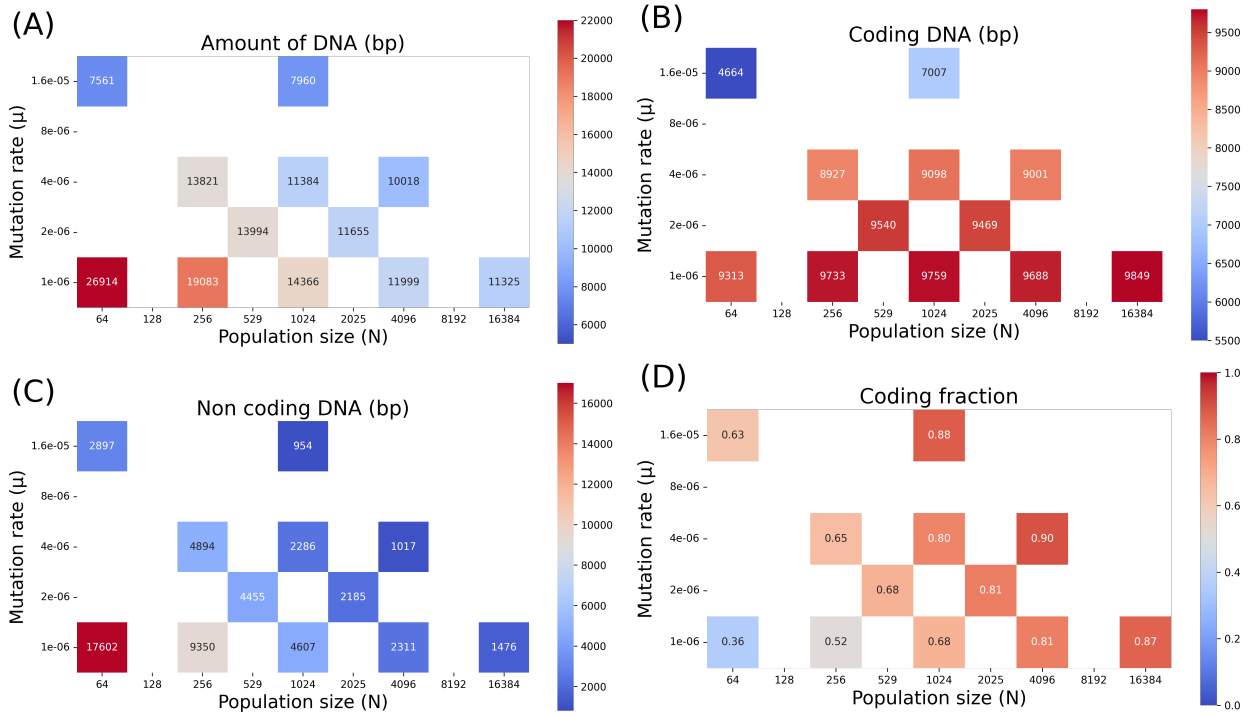

Figure 3: Amount of DNA (A), coding size (B), non-coding size (C) and coding fraction (D) for the different combinations of μ and N tested, after 2 millions generations. For each of the 5 WTs, 10 repetitions were performed for each condition. Control conditions ($N = 1024$ and $\mu = 1.10^{-6}$) are outlined in black. Due to computational limitations, not all combinations of parameters could be tested.

2.2 Robustness selection as the explanatory mechanism

We found 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.

We propose that these observations can be explained by an interplay between selection for phenotypic adaptation to the environment (hereafter *direct selection*), and selection for replicative robustness (hereafter *indirect selection*). We define the replicative robustness of an individual as the proportion of its offspring that did not acquire new deleterious mutations. In our case, wild-type organisms are very well adapted to their environment, thus most mutations will be strongly deleterious if they affect the coding part of the genome. This is particularly true for chromosomal rearrangements that can affect large genomic segments, hence imposing a heavy burden on genes. Conversely, beneficial mutations are extremely rare. We therefore approximate the robustness of our organisms by measuring the proportion of their offsprings 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 retain 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. Importantly, replicative robustness depends both on the probability for a given mutation to be neutral (hence on the fraction on non-coding sequences in the genome) and on the mean number of mutations underwent by the genome at each generation (hence on the total amount of DNA). 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, as some deleterious mutations that were quasi-neutral for low N can become effectively counter-selected for high N, changing the balance of beneficial *vs* deleterious fixed mutations. Figure 4A and 4B show the consequences of variations of N on the fitness and replicative robustness of the individuals respectively. They show that the increase in selection efficacy 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. However, a more efficient indirect selection is expected to decrease the overall genome size. The combination of both effects leads to a decrease in the non-coding genome size, and a maintenance of the coding genome size as exemplified by Fig. 1B and C. Conversely, when the population size is reduced,

the increased drift leads to the loss of coding sequences and an inflation of the non-coding compartment (Fig. 1B and C). This reorganization of the genome structure is associated with a loss in robustness (Fig. 4B).

In Aevol, genomes undergo different types of mutations that can be roughly grouped into local mutations (substitutions, InDels) and chromosomal rearrangements (duplications, deletions, inversions and translocations). Both kinds of events don't have the same impact on robustness. Figure 4C shows the change in robustness induced by the different types of events. It shows that the loss and gain in robustness are indeed driven by chromosomal rearrangements. On the opposite, local mutations (substitutions and InDels) do not have a significant impact on robustness.

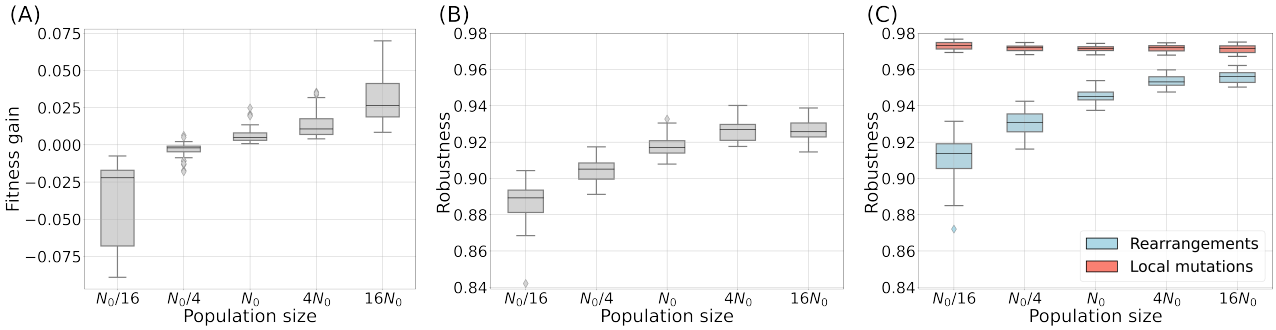

Figure 4: **Fitness gain (A) and Robustness (B: overall and C: by mutation type) at the end of the simulations, for different population size N.** Robustness is defined as the proportion of neutral offspring. The mutation rate is fixed to 10^{-6} per base pair for each type of mutation.

The case of an increased mutation rate is very different. Indeed, a sudden increase in μ results in an immediate drop of robustness at the start of the experiments (Fig. 5A). As the proportion of offspring that bears mutations rises with μ, we go from an initial robustness of 92% at the start of the simulations at μ_0, to 71% at $4\mu_0$, and only 26% at $16\mu_0$. 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 faithfully inherited cannot be selected. This indirect selection for robustness leads to the fixation of mutations decreasing drastically the genome size, even at the cost of the fitness of the individuals (see Fig. 5B): the only lineages that survive in the long term are the ones managing to mitigate the impact of the per-base mutation rate by decreasing their genome size, thus regaining some robustness (see Fig. 5C). Once the robustness has increased sufficiently, direct selection for phenotypical adaptation can resume and the fitness starts to increase again (see Fig. 5B). Interestingly, organisms manage here to continue to lose some coding base pairs while increasing their fitness, probably thanks to global genome restructurings allowing for a more compact encoding of the phenotype, for example through overlapping genes. This dynamics is very different to what happens when N is increased (and so the initial robustness is unaffected), as shown by Figure 5D, E and F.

Notably, robustness does not reach values as high as before the increase in mutation rate but 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 selection for robustness as soon as some organisms are able to pass on their genomic information.

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 focused on the non-coding compartment when caused by an increased population size.

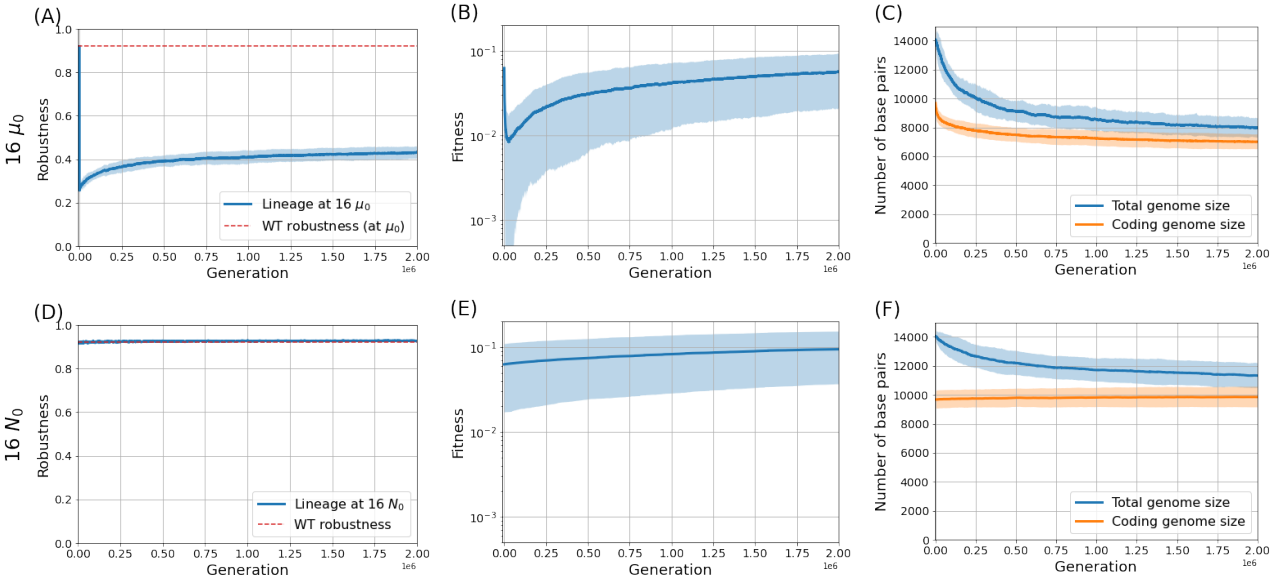

Figure 5: **Robustness, fitness, and genome architecture** across generations for $\mu = 1.6 \times 10^{-5}$ (16 μ_0) per base pair for each mutation type and $N = 1,024$ (N_0) (top row, panels A, B and C) and $N = 16,384$ (16 N_0) and $\mu = 1 \times 10^{-6}$ (μ_0) per base pair for each mutation type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

3 Discussion

We found that, in our experiments, genome size reduction can be caused by an increase in population size, in mutation rate, or both. These two factors can however be distinguished, as they have different effects on the coding and non-coding sequences of the genome. Their combination in various proportions opens a wide range of possibilities in the intensity of genome size reduction, and attainable coding density . Our results can therefore account for both endosymbiont-like and marine cyanobacteria-like genomes. 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 this has been recently questioned (Chen et al., 2022). 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 highlights that indirect selection for shorter genomes through robustness selection could be a key factor of genome evolution (Wilke et al., 2001; Gabzi et al., 2022), and especially of the evolution of genome size and structure.

Our observations confirm those made by Lynch and Conery (2003) that an increased genetic drift, here associated with a decreased population size, increases the genome size. Our result 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., 2007; 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). To explain this stabilization of genome size, we do not need to postulate a bias toward deletions in the mutational process (Mira et al., 2001), but simply an effective selection for robustness due to the mere existence of chromosomal rearrangements. 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. This mechanism requires no additional hypotheses and is very general. It should therefore be pervasive in the living world. We do nevertheless expect some other determinants of genome size, such as insertion sequences or effects of the non-coding genome over gene expression, to have an effect on genome size variation. These could be added to the model in order to study their impact on top of the mechanisms presented here. However, we argue that they would not fundamentally compromise the dynamics observed here and only displace the final equilibrium.

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 (Fig. 5 and S2 to S13). In particular, we observe that, when the mutation rate increases, the fitness initially drops drastically (Fig. 5B). This can be related to an error-threshold crossing mechanism (Eigen, 1971; Takeuchi and Hogeweg, 2007; de 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 (Fig. 5).

In our experiments, $N \times \mu$ stands out as a determining factor of genome structure, as isoclines of identical $N \times \mu$ values display similar coding densities, even in case of reduced genomes. Understanding this invariant is one of the most exciting perspectives opened by our work. Notably, there is a small variation in coding fraction along our 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 Supplementary Data S1). Further versions of the model could rely on various measures of the effective population size so as 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 Supplementary Data S1).

Our study also raises the question of the generalization of the relationships observed here between N, μ and the genome structure to eukaryotes. Indeed, Aevol is fundamentally a model of the evolution of prokaryotic genomes. The introduction of sexual reproduction and meiotic recombination to our model could have unexpected consequences, and may thus lead to qualitatively different results. The robustness of our observations to such big changes is yet to be assessed, and it would be interesting to study the interplay between selection for phenotypical adaptation and selection for robustness in the presence of sex and meiotic recombination.

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 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). 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.

4 Materials and Methods

4.1 The Aevol framework

Aevol (Knibbe et al., 2007; Banse et al., 2023) 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 which is composed of a fixed number of individuals on a grid (Fig. 6A). Each individual owns a double-stranded circular genomic sequence, composed of 0s and 1s. In order 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 (Fig. 6B). The distance between this function and a target function, which represents the ideal phenotype in the specified environment, gives the fitness of the individual. A detailed explanation can be found on the dedicated website www.aevol.fr.

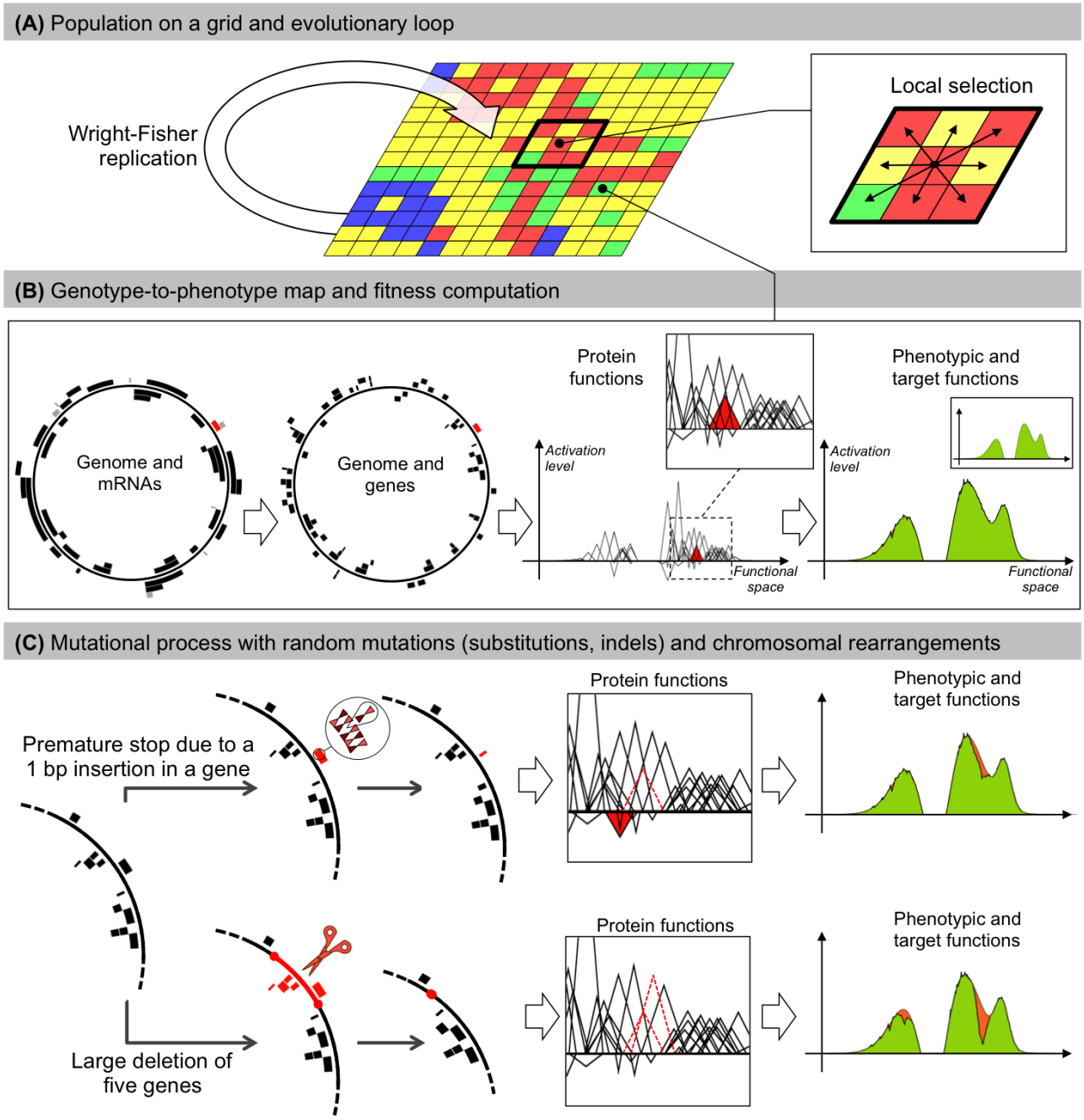

Figure 6: **The Aevol model.** **(A)** Individuals are distributed on a grid. A each generation, the whole population replicates according to a Wright-Fisher replication model in which selection operates locally within a 3×3 neighborhood. **(B)** Each grid-cell contains a single organism described by its genome. Genomes are decoded through a genotype-to-phenotype map with four main steps (transcription, translation, computation of protein functions, 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).

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 (Fig. 6C). They create diversity in the genomes, hence in the phenotypes, and allow the genome size and structure to change. Moreover, 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 offsprings as their genomes will be decoded thereafter, thus the model do not impose an *a priori* genome structure, which allows us to study the evolution of genome architecture in various experimental conditions.

4.2 Experimental design

4.2.1 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 a 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 millions 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 all types of mutations (point mutations, small insertions, small deletions, inversions, duplications, large deletions and translocations). Their fitnesses and genome structures are listed in Tab. 1.

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 1: **Wild-Types characteristics** at the start of our experiments.

4.2.2 Experimental conditions

A range of population sizes increases or decreases and mutations rates increases, as well as some combinations of both, are tested. All conditions are listed in Table 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.

Populations sizes and mutations rate are not equally distributed around the default parameters in order to better study the discussed cases of endosymbionts and marine bacteria, which both appear to have an increase in mutation rate but very different population sizes.

Population size	Mutation rate (per base pair, per mutation type)	$N \times \mu$ product
64 ($N_0/16$)	10^{-6} (μ_0)	$1/16N_0 \times \mu_0$
256 ($N_0/4$)	10^{-6} (μ_0)	$1/4N_0 \times \mu_0$
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} ($16 \times \mu_0$)	$N_0 \times \mu_0$
2025 ($\approx 2 \times N_0$)	2×10^{-6} ($2 \times \mu_0$)	$\approx 4N_0 \times \mu_0$
4096 ($4 \times N_0$)	10^{-6} (μ_0)	$4N_0 \times \mu_0$
1024 (N_0)	4×10^{-6} ($4 \times \mu_0$)	$4N_0 \times \mu_0$
4096 ($4 \times N_0$)	4×10^{-6} ($4 \times \mu_0$)	$16N_0 \times \mu_0$
16384 ($16 \times N_0$)	10^{-6} (μ_0)	$16N_0 \times \mu_0$
1024 (N_0)	16×10^{-6} ($16 \times \mu_0$)	$16N_0 \times \mu_0$

Table 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 mulpiplied by 2.

4.2.3 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, starting from a random individual of the final population, we identify all its ancestors. 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 which 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 repetitions of each experimental condition is provided in the supplementary material S2 (Figures S2 to S13).

4.2.4 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://zenodo.org/doi/10.5281/zenodo.10669479`.

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 for fruitful comments on the manuscript.

References

Adami, C. (2006). Digital genetics: unravelling the genetic basis of evolution. *Nat Rev Genet*, 7(2):109–118.

Almpanis, A., Swain, M., Gatherer, D., and McEwan, N. (2018). Correlation between bacterial G+C content, genome size and the G+C content of associated plasmids and bacteriophages. *Microbial Genomics*, 4(4):e000168.

Bachmann, K. (1972). Genome size in mammals. *Chromosoma*, 37(1):85–93.

Banse, P., Luiselli, J., Parsons, D. P., Grohens, T., Foley, M., Trujillo, L., Rouzaud-Cornabas, J., Knibbe, C., and Beslon, G. (2023). Forward-in-time simulation of chromosomal rearrangements: The invisible backbone that sustains long-term adaptation. *Molecular Ecology*.

Barow, M. and Meister, A. (2002). Lack of correlation between AT frequency and genome size in higher plants and the effect of nonrandomness of base sequences on dye binding. *Cytometry*, 47(1):1–7.

Batut, B., Knibbe, C., Marais, G., and Daubin, V. (2014). Reductive genome evolution at both ends of the bacterial population size spectrum. *Nat Rev Microbiol*, 12(12):841–850.

Batut, B., Parsons, D. P., Fischer, S., Beslon, G., and Knibbe, C. (2013). In silico experimental evolution: a tool to test evolutionary scenarios. *BMC Bioinformatics*, 14(15):S11.

Bourguignon, T., Kinjo, Y., Villa-Martin, P., Coleman, N. V., Tang, Q., Arab, D. A., Wang, Z., Tokuda, G., Hongoh, Y., Ohkuma, M., et al. (2020). Increased mutation rate is linked to genome reduction in prokaryotes. *Current Biology*, 30(19):3848–3855.

Brevet, M. and Lartillot, N. (2021). Reconstructing the history of variation in effective population size along phylogenies. *Genome Biology and Evolution*, 13(8).

Chen, Z., Wang, X., Song, Y., Zeng, Q., Zhang, Y., and Luo, H. (2022). Prochlorococcus have low global mutation rate and small effective population size. *Nat Ecol Evol*, 6(2):183–194.

Choi, I.-Y., Kwon, E.-C., and Kim, N.-S. (2020). The C- and G-value paradox with polyploidy, repeatomes, introns, phenomes and cell economy. *Genes & Genomics*, 42(7):699–714.

Chong, R. A., Park, H., and Moran, N. A. (2019). Genome Evolution of the Obligate Endosymbiont Buchnera aphidicola. *Molecular Biology and Evolution*, 36(7):1481–1489.

de Boer, F. K. and Hogeweg, P. (2010). Eco-evolutionary dynamics, coding structure and the information threshold. *BMC Evolutionary Biology*, 10(1):361.

Dufresne, A., Garczarek, L., and Partensky, F. (2005). Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biology*, 6(2):R14.

Eigen, M. (1971). Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58(10):465–523.

Figuet, E., Nabholz, B., Bonneau, M., Mas Carrio, E., Nadachowska-Brzyska, K., Ellegren, H., and Galtier, N. (2016). Life History Traits, Protein Evolution, and the Nearly Neutral Theory in Amniotes. *Molecular Biology and Evolution*, 33(6):1517–1527.

Flombaum, P., Gallegos, J. L., Gordillo, R. A., Rincón, J., Zabala, L. L., Jiao, N., Karl, D. M., Li, W. K. W., Lomas, M. W., Veneziano, D., Vera, C. S., Vrugt, J. A., and Martiny, A. C. (2013). Present and future global distributions of the marine cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proceedings of the National Academy of Sciences*, 110(24):9824–9829.

Gabzi, T., Pilpel, Y., and Friedlander, T. (2022). Fitness landscape analysis of a trna gene reveals that the wild type allele is sub-optimal, yet mutationally robust. *Molecular Biology and Evolution*, 39(9).

Gago, S., Elena, S. F., Flores, R., and Sanjuán, R. (2009). Extremely high mutation rate of a hammerhead viroid. *Science*, 323(5919):1308–1308.

Giovannoni, S. J., Cameron Thrash, J., and Temperton, B. (2014). Implications of streamlining theory for microbial ecology. *ISME J*, 8(8):1553–1565.

Giovannoni, S. J., Tripp, H. J., Givan, S., Podar, M., Vergin, K. L., Baptista, D., Bibbs, L., Eads, J., Richardson, T. H., Noordewier, M., Rappé, M. S., Short, J. M., Carrington, J. C., and Mathur, E. J. (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science*, 309(5738):1242–1245.

Heddi, A., Charles, H., Khatchadourian, C., Bonnot, G., and Nardon, P. (1998). Molecular characterization of the principal symbiotic bacteria of the weevil sitophilus oryzae: A peculiar g + c content of an endocytobiotic DNA. *J Mol Evol*, 47(1):52–61.

Hindré, T., Knibbe, C., Beslon, G., and Schneider, D. (2012). New insights into bacterial adaptation through in vivo and in silico experimental evolution. *Nature Reviews Microbiology*, 10(5):352–365.

Hu, J. and Blanchard, J. L. (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. *Molecular Biology and Evolution*, 26(1):5–13.

Kalhor, R., Beslon, G., Lafond, M., and Scornavacca, C. (2023). Classifying the post-duplication fate of paralogous genes. In Jahn, K. and Vinař, T., editors, *Comparative Genomics*, pages 1–18, Cham. Springer Nature Switzerland.

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

Konstantinidis, K. T. and Tiedje, J. M. (2004). Trends between gene content and genome size in prokaryotic species with larger genomes. *Proceedings of the National Academy of Sciences*, 101(9):3160–3165.

Koskiniemi, S., Sun, S., Berg, O. G., and Andersson, D. I. (2012). Selection-driven gene loss in bacteria. *PLOS Genetics*, 8(6):e1002787.

Kuo, C.-H., Moran, N. A., and Ochman, H. (2009). The consequences of genetic drift for bacterial genome complexity. *Genome Res.*, 19(8):1450–1454.

LaBar, T. and Adami, C. (2020). *Genome Size and the Extinction of Small Populations*, pages 167–183. Springer International Publishing, Cham.

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

Leth Bak, A., Black, F. T., Christiansen, C., and Freundt, E. A. (1969). Genome size of mycoplasmal DNA. *Nature*, 224(5225):1209–1210.

Lynch, M. (2006). Streamlining and simplification of microbial genome architecture. *Annu. Rev. Microbiol.*, 60:327–349.

Lynch, M. (2007). The frailty of adaptive hypotheses for the origins of organismal complexity. *Proceedings of the National Academy of Sciences*, 104(suppl_1):8597–8604.

Lynch, M. and Conery, J. S. (2003). The origins of genome complexity. *Science*, 302:1401–1404.

Marais, G. A. B., Calteau, A., and Tenaillon, O. (2008). Mutation rate and genome reduction in endosymbiotic and free-living bacteria. *Genetica*, 134(2):205–210.

Martinez-Gutierrez, C. A. and Aylward, F. O. (2022). Genome size distributions in bacteria and archaea are strongly linked to evolutionary history at broad phylogenetic scales. *PLOS Genetics*, 18(5):e1010220. Publisher: Public Library of Science.

Martínez-Cano, D. J., Reyes-Prieto, M., Martínez-Romero, E., Partida-Martínez, L. P., Latorre, A., Moya, A., and Delaye, L. (2015). Evolution of small prokaryotic genomes. *Frontiers in Microbiology*, 5.

Mira, A. and Moran, N. (2002). Estimating population size and transmission bottlenecks in maternally transmitted endosymbiotic bacteria. *Microbial Ecology*, 44(2):137–143.

Mira, A., Ochman, H., and Moran, N. A. (2001). Deletional bias and the evolution of bacterial genomes. *Trends in Genetics*, 17(10):589–596.

Moran, N. A. (1996). Accelerated evolution and muller's rachet in endosymbiotic bacteria. *Proceedings of the National Academy of Sciences*, 93(7):2873–2878.

Moran, N. A. (2002). Microbial Minimalism: Genome Reduction in Bacterial Pathogens. *Cell*, 108(5):583–586.

Moran, N. A. (2003). Tracing the evolution of gene loss in obligate bacterial symbionts. *Current Opinion in Microbiology*, 6(5):512–518.

Moran, N. A. and Mira, A. (2001). The process of genome shrinkage in the obligate symbiont buchnera aphidicola. *Genome Biol*, 2(12):1–12.

Morris, J. J., Lenski, R. E., and Zinser, E. R. (2012). The black queen hypothesis: Evolution of dependencies through adaptive gene loss. *mBio*, 3(2):e00036–12.

Müller, R., Kaj, I., and Mugal, C. F. (2022). A Nearly Neutral Model of Molecular Signatures of Natural Selection after Change in Population Size. *Genome Biology and Evolution*, 14(5):evac058.

Ngugi, D. K., Acinas, S. G., Sánchez, P., Gasol, J. M., Agusti, S., Karl, D. M., and Duarte, C. M. (2023). Abiotic selection of microbial genome size in the global ocean. *Nature Communications*, 14(1):1384.

Pellicier, J., Fay, M. F., and Leitch, I. J. (2010). The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society*, 164(1):10–15.

Petrov, D. A. (2001). Evolution of genome size: new approaches to an old problem. *Trends in Genetics*, 17(1):23–28.

Popadin, K., Polishchuk, L. V., Mamirova, L., Knorre, D., and Gunbin, K. (2007). Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *Proceedings of the National Academy of Sciences*, 104(33):13390–13395.

Romiguier, J., Ranwez, V., Douzery, E., and Galtier, N. (2012). Genomic Evidence for Large, Long-Lived Ancestors to Placental Mammals. *Molecular Biology and Evolution*, 30(1):5–13.

Schneiker, S., Perlova, O., Kaiser, O., Gerth, K., Alici, A., Altmeyer, M. O., Bartels, D., Bekel, T., Beyer, S., Bode, E., Bode, H. B., Bolten, C. J., Choudhuri, J. V., Doss, S., Elnakady, Y. A., Frank, B., Gaigalat, L., Goesmann, A., Groeger, C., Gross, F., Jelsbak, L., Jelsbak, L., Kalinowski, J., Kegler, C., Knauber, T., Konietzny, S., Kopp, M., Krause, L., Krug, D., Linke, B., Mahmud, T., Martinez-Arias, R., McHardy, A. C., Merai, M., Meyer, F., Mormann, S., Muñoz-Dorado, J., Perez, J., Pradella, S., Rachid, S., Raddatz, G., Rosenau, F., Rückert, C., Sasse, F., Scharfe, M., Schuster, S. C., Suen, G., Treuner-Lange, A., Velicer, G. J., Vorhölter, F.-J., Weissman, K. J., Welch, R. D., Wenzel, S. C., Whitworth, D. E., Wilhelm, S., Wittmann, C., Blöcker, H., Pühler, A., and Müller, R. (2007). Complete genome sequence of the myxobacterium Sorangium cellulosum. *Nature Biotechnology*, 25(11):1281–1289.

Takeuchi, N. and Hogeweg, P. (2007). Error-threshold exists in fitness landscapes with lethal mutants. *BMC Evolutionary Biology*, 7(1):15.

van Ham, R., Kamerbeek, J., Palacios, C., Rausell, C., Abascal, F., Bastolla, U., Fernández, J., Jiménez, L., Postigo, M., Silva, F., Tamames, J., Viguera, E., Latorre, A., Valencia, A., Morán, F., and Moya, A. (2003). Reductive genome evolution in buchnera aphidicola. *Proceedings of the National Academy of Sciences*, 100(2):581–586.

Wernegreen, J. J. (2002). Genome evolution in bacterial endosymbionts of insects. *Nat Rev Genet*, 3(11):850–861.

Wernegreen, J. J. (2015). Endosymbiont evolution: predictions from theory and surprises from genomes. *Annals of the New York Academy of Sciences*, 1360(1):16–35.

Westoby, M., Nielsen, D. A., Gillings, M. R., Litchman, E., Madin, J. S., Paulsen, I. T., and Tetu, S. G. (2021). Cell size, genome size, and maximum growth rate are near-independent dimensions of ecological variation across bacteria and archaea. *Ecology and Evolution*, 11(9):3956–3976.

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

Wolf, Y. I. and Koonin, E. V. (2013). Genome reduction as the dominant mode of evolution. *BioEssays*, 35(9):829–837.