Evolution beyond substitutions : Computational modeling of the impact of chromosomal rearrangements on evolutionary dynamics

Paul Banse

n° d'ordre : 2023ISAL0129

THESE de DOCTORAT DE L'INSA LYON, membre de l'Université de Lyon

Ecole Doctorale N° 512
InfoMaths

Spécialité/ discipline de doctorat :
Informatique

Soutenue publiquement le 18/12/2023, par :
Paul Banse

Evolution beyond substitutions :
Computational modeling of the impact of chromosomal rearrangements on evolutionary dynamics

---

Devant le jury composé de :

Achaz, Guillaume	Professeur des Universités	Université Paris-Cité	Rapporteur
Verel, Sébastien	Professeur des Universités	Université du Littoral Côte d'Opale	Rapporteur
Carbone, Alessandra	Professeure des Universités	Sorbonne Universités	Examinatrice
Varoquaux, Nelle	Chargée de recherche	CNRS	Examinatrice
Rajon, Etienne	Maître de conférences	Université Claude Bernard Lyon 1	Examinateur
Beslon Guillaume	Professeur des Universités	INSA Lyon	Directeur de thèse

INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Département de la Formation par la Recherche et des Études Doctorales (**FEDORA**)

Bâtiment INSA direction, 1er étage
37, av. J. Capelle
69621 Villeurbanne Cédex
fedora@insa-lyon.fr

Référence : TH1053_BANSE Paul

L'INSA Lyon a mis en place une procédure de contrôle systématique via un outil de détection de similitudes (logiciel Compilatio). Après le dépôt du manuscrit de thèse, celui-ci est analysé par l'outil. Pour tout taux de similarité supérieur à 10%, le manuscrit est vérifié par l'équipe de FEDORA. Il s'agit notamment d'exclure les auto-citations, à condition qu'elles soient correctement référencées avec citation expresse dans le manuscrit.

Par ce document, il est attesté que ce manuscrit, dans la forme communiquée par la personne doctorante à l'INSA Lyon, satisfait aux exigences de l'Etablissement concernant le taux maximal de similitude admissible.

INSA LYON
Campus LyonTech La Doua
20, avenue Albert Einstein - 69621 Villeurbanne cedex - France
Tél. +33 (0)4 72 43 83 83 - Fax +33 (0)4 72 43 85 00
www.insa-lyon.fr

membre de UNIVERSITÉ DE LYON

Département FEDORA – INSA Lyon - Ecoles Doctorales

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
ED 206 CHIMIE	**CHIMIE DE LYON** https://www.edchimie-lyon.fr Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage secretariat@edchimie-lyon.fr	**M. Stéphane DANIELE** C2P2-CPE LYON-UMR 5265 Bâtiment F308, BP 2077 43 Boulevard du 11 novembre 1918 69616 Villeurbanne directeur@edchimie-lyon.fr
ED 341 E2M2	**ÉVOLUTION, ÉCOSYSTÈME, MICROBIOLOGIE, MODÉLISATION** http://e2m2.universite-lyon.fr Sec. : Bénédicte LANZA Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.e2m2@univ-lyon1.fr	**Mme Sandrine CHARLES** Université Claude Bernard Lyon 1 UFR Biosciences Bâtiment Mendel 43, boulevard du 11 Novembre 1918 69622 Villeurbanne CEDEX e2m2.codir@listes.univ-lyon1.fr
ED 205 EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTÉ** http://ediss.universite-lyon.fr Sec. : Bénédicte LANZA Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.ediss@univ-lyon1.fr	**Mme Sylvie RICARD-BLUM** Laboratoire ICBMS - UMR 5246 CNRS - Université Lyon 1 Bâtiment Raulin - 2ème étage Nord 43 Boulevard du 11 novembre 1918 69622 Villeurbanne Cedex Tél : +33(0)4 72 44 82 32 sylvie.ricard-blum@univ-lyon1.fr
ED 34 EDML	**MATÉRIAUX DE LYON** http://ed34.universite-lyon.fr Sec. : Yann DE ORDENANA Tél : 04.72.18.62.44 yann.de-ordenana@ec-lyon.fr	**M. Stéphane BENAYOUN** Ecole Centrale de Lyon Laboratoire LTDS 36 avenue Guy de Collongue 69134 Ecully CEDEX Tél : 04.72.18.64.37 stephane.benayoun@ec-lyon.fr
ED 160 EEA	**ÉLECTRONIQUE, ÉLECTROTECHNIQUE, AUTOMATIQUE** https://edeea.universite-lyon.fr Sec. : Philomène TRECOURT Bâtiment Direction INSA Lyon Tél : 04.72.43.71.70 secretariat.edeea@insa-lyon.fr	**M. Philippe DELACHARTRE** INSA LYON Laboratoire CREATIS Bâtiment Blaise Pascal, 7 avenue Jean Capelle 69621 Villeurbanne CEDEX Tél : 04.72.43.88.63 philippe.delachartre@insa-lyon.fr
ED 512 INFOMATHS	**INFORMATIQUE ET MATHÉMATIQUES** http://edinfomaths.universite-lyon.fr Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage Tél : 04.72.43.80.46 infomaths@univ-lyon1.fr	**M. Hamamache KHEDDOUCI** Université Claude Bernard Lyon 1 Bât. Nautibus 43, Boulevard du 11 novembre 1918 69 622 Villeurbanne Cedex France Tél : 04.72.44.83.69 direction.infomaths@listes.univ-lyon1.fr
ED 162 MEGA	**MÉCANIQUE, ÉNERGÉTIQUE, GÉNIE CIVIL, ACOUSTIQUE** http://edmega.universite-lyon.fr Sec. : Philomène TRECOURT Tél : 04.72.43.71.70 Bâtiment Direction INSA Lyon mega@insa-lyon.fr	**M. Jocelyn BONJOUR** INSA Lyon Laboratoire CETHIL Bâtiment Sadi-Carnot 9, rue de la Physique 69621 Villeurbanne CEDEX jocelyn.bonjour@insa-lyon.fr
ED 483 ScSo	**ScSo[1]** https://edsciencessociales.universite-lyon.fr Sec. : Mélina FAVETON Tél : 04.78.69.77.79 melina.faveton@univ-lyon2.fr	**M. Bruno MILLY** (INSA : J.Y. TOUSSAINT) Univ. Lyon 2 Campus Berges du Rhône 18, quai Claude Bernard 69365 LYON CEDEX 07 Bureau BEL 319 bruno.milly@univ-lyon2.fr

[1] ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

Contents

List of Figures

List of Tables

Remerciements

Mes premiers remerciements vont évidemment à Guillaume Beslon pour son encadrement, sa patience, sa gentillesse, son intelligence et son respect en toute situation.

Je voudrais ensuite remercier les rapporteurs de ma thèse, Guillaume Achaz et Sébastien Verel pour avoir accepté la charge de me relire et pour leurs retours bienveillants et encourageants. Je voudrais aussi remercier les membres du jury: Alessandra Carbone, Nelle Varoquaux et Etienne Rajon pour avoir accepté de m'évaluer.

Ensuite viennent les remerciements plus personnels :

À Juliette Luiselli pour avoir partagé tant de choses depuis des années.

À Lisa Chabrier pour avoir souris quand je me sentais seul.

À Laurent Turpin pour nos débats enflammés et nos éclats de rires.

À Nathan Quiblier pour m'avoir fait voyager du haut des pistes à son école primaire.

À Marco Foley pour avoir gardé son calme et fait preuve de sagesse à chaque moment.

À Julie Etienne pour son soutient sans faille et ses qualité organisationnelles.

À Théotime Grohens pour avoir ouvert la voie et montré le chemin.

À Lisa Blum pour m'avoir montré qu'on peut penser autrement.

À Arnaud Hubert pour son amitié, son accueil et son affection.

À Leonardo Trujillo pour son temps, son énergie et son expérience.

À Claire Sauer pour avoir été le roc au milieu de la tempête qui soutient tout le monde.

À Audrey Denizot, parce que certains de ses conseils ont fait beaucoup de chemin.

À Elise Philippine Lengrand pour avoir été un soutien de chaque jour.

D'autres personnes méritent d'être remerciées:

David Pichardie pour m'avoir permis de trouver le stage qui a mené à cette thèse.

Mes autres collaborateurs et collaboratrice: Jonathan Rouzaud Cornabas, David Parsons, Yvan Junier, Olivier Mazet, Nicolas Latrillot et Carole Knibbe.

Au sein du centre Inria Lyon, le personnel de soutien à la recherche : Laetitia Lécot Gauthé et Anne-Laure Fogliani.

De nombreuses autres personnes m'ont accompagné et soutenu : Arsène Marzorati, Charlotte Camus, Hana Sébia, Astrée Faraguet, Aurélien Oddou, Salim El Houat, Lou Charbrier, Elie Dumont, Florian Dupeuble, Jean De Piero, Nicolas Simon, Romain Gallé, Christophe Rigotti, Anton Crombach, Cordula Reisch, Valentin Melot, Flore Vandier, Sarah Léon, Moccha Melot, Victor Ferry, Vincent Liard, Axelle Gaigé, Ella Baumann, Chloé Bancelle, Andréa Ducos.

J'ai aussi eu l'honneur de participer à l'encadrement de stagiaires : Yannis Sindt-Baret, Aoife Orla Igoe, Félicie Chaudron.

Ensuite mes colocataires: Romain Banse, Antoine Amaté, Carole Doucerain.

Ma famille : mes parents, mes frères, ma soeur et le reste de la famille

Pour la relecture : Estelle Giner, Marthe Banse, Nadira Gayan, Basile Clément et Quentin Chevalier.

Enfin, un grand merci à tout les gens avec qui j'ai joué à des jeux de société, discuté via internet, ou croisé en soirée pour avoir participé à mon ouverture au monde.

Part A

Why Model Chromosomal Rearrangements?

Chapter I

General Introduction

"Choisir, c'est renoncer."

X.B. librement adapté d'André Gides

1 A question of choices

Model programming is at its core a question of choices. In fact, when modeling any complex phenomenon, the complexity prevents taking into account all details of the context (parameters, events, etc.). Hence, modelers in all area of knowledge have to decide on what is important for their model and what is not. This choice can be motivated by prior theories or by empirical data, but is ultimately a subjective choice. Indeed, it is impossible to investigate the effect of a parameter that was not included in the model, and it is just as impossible to include all parameters in the model. Only a constant questioning of modeling choices and a frequent comparison to the ground truth can lead to models that can provide insight on the real mechanism at stake.

In the particular case of models of evolution, an extensive amount of simplification is needed. Life has proven to be extremely diverse and the product of a long adaptation with multiple exogenous factors. The concept of natural selection is a typical example, as it is by definition context-dependent. The theory of adaptation through natural selection by Darwin is revolutionary because it conciliates both simple mechanical assumptions (namely heredity of traits with incidence on progeny, and random emergence of such traits) and a conceptual data-driven model, that if valid is sufficient to generate the observed patterns. This conceptual model of biological evolution was inherently faulty: observations were limited by the technology available, and it is an over-simplification of the real phenomenon. However, following George Box (Box, 1976) "All models are wrong, but some are useful", and as a matter of fact, this conceptual model was wrong, but it was immensely useful in understanding evolution and biology.

Following the success of Darwin's theory of adaptation through natural selection, numerous models have been proposed: conceptual models such as Darwin's, experimental models such as the chemostat (Novick et Szilard, 1950), formal ones such as fitness landscapes (Wright, 1932), Fisher's geometric model (Fisher, 1930) and many others.

As evolution is a complex process, a conventional simplification is to divide it in three parts: (i) the process of variation of information for an individual, (ii) conservation of

this variation through reproduction in an ecosystem and (iii) selection of this variation by competition and environmental constraints. Each part encompasses a wide diversity of situations, and they are closely intertwined. Indeed, mutations influence the population size, populations can change the environment, and the environment can have an impact on the mutation rate. However, knowledge in evolution can only be gained by studying examples and not all ecosystems together at all evolutionary scales. Hence, there is a need to divide the process into different parts, and also to focus on so-called model experiments and model species. None of these models perfectly recreate the situations in the wild, yet they remain useful as they ease the conceptual understanding of the real phenomenon at stake.

Reproductive competition and environmental influence is the core of Darwin's initial theory. It has started with empirical observations of environmental adaptation, followed by conceptual models and theories (Grinnell, 1924), experimental models to analyse and reproduce partially these phenomenons, and ultimately formal (mathematical) models (Goel *et al.*, 1971) that can predict evolution of populations. Indeed, the influence of environment and ecology on evolution is still a very active research topic.

Similarly, populations with different structures experiencing different modes of selection are a well studied field of research that has followed a similar trajectory. For example, it has been shown early that empirical observations on diversity of related species cannot be due to sole Darwinian selection, demanding new concepts (Gulick, 1887). These concepts allowed building experimental models of evolution, and multiple testing of experimental models is needed to eventually lead to formal models of diversity (Kimura et Crow, 1963; Wright, 1938).

Because mutations occur at molecular scale, the study of mutations is intrinsically more difficult. Indeed, the first direct observations of the mechanisms of genetic mutations date from the early twentieth century, while mechanisms of natural selections could be observed visually by Darwin in the 1830s. Sturtevant (1921) recorded the first case of chromosomal inversion and proposed a mechanistic explanation of this phenomenon. Interestingly, this discovery occurred long before the discovery of DNA structure (Franklin et Gosling, 1953; Watson et Crick, 1953). Due to the lack of knowledge about the inner mechanism of mutations and DNA structure, the conceptual frameworks were based on alleles and loci (Dobzhansky, 1950; Haldane, 1937)[1], these frameworks included both alleles modifications and chromosome rearrangements.

After the discovery of DNA structure by Franklin et Gosling (1953); Watson et Crick (1953), and the following development of sequencing technologies, point mutations were shown to be an important source of genetic variation (Edwards *et al.*, 2007). Short read sequencing outputs short sequences and thus cannot detect whether long sequences were rearranged. Understandably, conceptual work has been focused on the substitution and other mutations commonly witnessed in short reads (the local mutations: substitutions and Indels[2]). For example, Freese (1959) introduces the terms "transition" and "transversions", creating the precise vocabulary needed for theoretical work. Maynard Smith (1970) conceptualized mutational networks on substitution and proposed the basis of mathematical reasoning on such networks. Due to this background conceptual work on substitutions,

[1]Haldane's article is also interesting in a historical perspective, as it was written while the author was defending Madrid against Francoists during the Spanish Civil War.

[2]Indels are mutations that insert (In) or delete (del) few base pairs.

new ideas and models (computational and mathematical) arose based on substitutions, for example on the Avida platform, the focus is made on substitutions and Indels as the cause of genetic variation (Lenski *et al.*, 2003; Adami, 2006; Bray Speth *et al.*, 2009).

Due to short read sequencing, chromosomal rearrangements were rarely witnessed, and since models need simplicity, in most cases modeling and theoretical work ignored rearrangements (Wellenreuther *et al.*, 2019). Recently, improvements in sequencing techniques shed a new light on this theoretical blind spot(Hanlon *et al.*, 2022; Ho *et al.*, 2020). Indeed, it seems that chromosomal rearrangements, while rare, have an unforeseen importance in adaptation. For example, in the Long Term E.coli Experiment, rearrangements have been shown to be a decisive part of bacterial evolution (Raeside *et al.*, 2014; Blount *et al.*, 2012a)). The goal of the present thesis is to contribute to filling this theoretical blind spot. To introduce the topic, we will briefly present first the molecular mechanisms responsible for local mutations and chromosomal rearrangements and then provide an overview of the computational and theoretical models related to rearrangements.

2 Nature of a chromosomal rearrangement

Genetic information is transmitted to the descendants during reproduction, by definition any variation in this information is a mutation. Mutations are needed for evolution, as they provide the necessary variations for selection to work on. Despite that, since they are a loss of information in the genome, it is common to consider that mutations are "errors" in the replication. Among the mutations, it is usual to discriminate between the most common small sized mutations (local mutations), and the rarer ones affecting significant parts of the genome (chromosomal rearrangements also known as structural variations). While all domains of life experience these mutations, in the context of this thesis, we will focus on haploid (typically microbial) genomes.

Semantically, local mutations affect only a limited portion of the chromosome, while chromosomal rearrangements affect a long sequence. In experimental data, the distinction is usually made on the size of the variation introduced : when more than 50 base pairs are affected, the mutation can be called a structural variation, i.e., a rearrangement of a part of the chromosome (Mérot *et al.*, 2020). However, there is no consensus on this definition. Another approach, more systemic, is to focus on the mechanisms that can create large scale changes. In this case, all mutations generated by such mechanisms, independently of the size of the changes introduced, can be called a rearrangement (Darling *et al.*, 2008).

DNA as a polymer is composed of two long chains of nucleotides, it can be subject to random deterioration. Damage can come from a multitude of factors, such as environmental mutagens, endogenous mutagens such as reactive oxygen species (Sakai *et al.*, 2006), or spontaneous reactions within the polymer (isomerization, hydrolysis, ...) (Loeb, 1989). DNA replication also creates a fragility, as the two strands are separated to allow replication. When the damage is limited in size, it can turn into a local mutation, while if the damage concerns the whole sequence, such as a double-strand break, the resulting mutation is a chromosomal rearrangement. The main chromosomal rearrangements are deletions, duplications, inversions, and translocations and they can affect sequences within genes as well as outside of genes (Periwal et Scaria, 2015). A deletion consists of a deletion of a sequence. A duplication consists of the copy of a sequence into another

part of the genome. An inversion is a mutation that exchanges both strands of the DNA and its orientation, and a translocation consists of deletion of a sequence and its copy at another location of the genome. It has been shown by Hastings *et al.* (2009) that one of the main events leading to a chromosomal rearrangement is when the DNA replication fork encounter a nick in a template strand, this breaks one arm of the fork and frees the newly formed sequence. The broken arm can then invade other parts of the DNA molecule, when this happens, new replication forks can emerge and copy part of the other molecule into the broken arm strand. Other mechanisms exist, in general they result from damage on the DNA molecule followed by a faulty repair (Burssed *et al.*, 2022).

Since there are multiple types of genome rearrangements, and since they can occur on sequences of small sizes up to millions of base pairs in length, these mutations are complex to study theoretically. Indeed, while chromosomal rearrangement is a focus for numerous experimental articles (Wellenreuther et Bernatchez, 2018; Willemsen *et al.*, 2016), the need for a "common framework for studying all structural variants" (Mérot *et al.*, 2020) has been identified in the community, calling for more theorization of the problem. Among all different types of rearrangements, those linked to sexual reproduction are probably those that have gathered most of the attention from theoretical studies. As an example of what could be done theoretically, we hereby present succinctly sex viewed under the prism of theoretical rearrangements.

In terms of its impact on genomic information, sex could indeed be considered as a sort of chromosomal rearrangement, as it consists in a rearrangement of genetic sequences between the maternal and paternal chromosomes. However, due to its visibility in karyotypes and to a relative ease of experimentation, it has been better studied, experimented on and theorized than any other type of rearrangements. In this introduction, we will use it as an example to illustrate the kinds of theoretical work that could (and should) be performed on chromosomal rearrangements.

The genetic process of sexual reproduction consists in having two parents whose genome contains a specific number of chromosomal pairs. During gamete production, each chromosome is paired and then split such that a gamete contains only one chromosome of each pair. The offspring's genome results in the fusion of two gamete genomes, one for each parent. During the pairing, homologous sequences are likely to intertwine, this leads to double strand breaks, that can be repaired by a recombination. Since recombination is based on homologies, it is often neutral in a lot of cases. Note that similar mechanisms of homologous recombination exist in bacteria, for damage repair.

After sexual reproduction, the chromosomes of the parents are shuffled by pair and sampled randomly. It allows genes in different chromosomes to be separately filtered by selection, instead of the whole chromosome being selected or counter selected. In a more complex manner, homologous recombination between paired chromosomes during meiosis partially allows the differentiated selection of genes within the same chromosome. Interestingly, sex, and to a lesser extent homologous recombination, received a lot more attention from the theoretical evolutive community than any other chromosomal rearrangements. This may be due to the observation that most multicellular life forms can reproduce sexually, rising the question of the interest of such reproduction.

A good introduction to theoretical analysis on sex and homologous recombination can be found in (Barton et Charlesworth, 1998). Is sexual reproduction a selective trait? What are its consequences on evolution? Can they explain empirical observations such as

the overwhelming presence of sex in multicellular life forms? Similarly to all mutations, sex and homologous recombination are mostly deleterious in a well adapted population. In particular, sexual reproduction requires two partners to produce one offspring, this is the so-called "cost of males" (Smith, 1978). It may also create an epistatic effect between two distant genomes, up to infertile offsprings. Thus, for sexual reproduction to have such a prevalence in evolution, there must be a great and general evolutive advantage. Hartfield et Keightley (2012) analyses all hypotheses to explain the presence of sex, thus investigating its evolutionary advantages. While the question is not totally resolved, there are two favored hypotheses. The first hypothesis is that homologous recombination allows selection to act more precisely by separating genes within the chromosomes, by lowering genetic linkage. Selection then acts at the allelic scale: if two beneficial mutations on different genes arise in different individuals, there are still chances that both can be fixed at the same time, this limits the effect of clonal interference and increasing the parallelism of selection. The second hypothesis is the Red Queen hypothesis: sex and homologous recombination can be valuable to maintain a diversity in the population and to create novel phenotypes. In the case of an evolutionary arm race, for example against parasites, diversity is advantageous both for resilience and for adaptation to a changing environment. The predominance of a hypothesis over the other can be context specific, and a complete theory is still under investigation (Tarkhnishvili *et al.*, 2023). The conceptual models also make some simplifications, such as non-disjunctive meiosis and ectopic recombination that lead respectively to monosomy or trisomy and non-allelic mutations, but the theoretical work allows a general understanding of the mechanisms and its impact on evolution. From a conceptual model, it is possible to generate computational models to investigate specific evolutionary impacts, for example, to investigate why recombination seems more favorable in small populations (Otto et Barton, 2001). Indeed, by simulating the effect of negative epistatic interaction and drift on genomes, it is possible to show that drift plays the major part in selection for recombination in small to medium populations (Otto et Barton, 2001).

While some theoretical work has been done on specific instances of rearrangements, for the moment there lacks a framework to study them in the general case (Mérot *et al.*, 2020). Indeed, there is a large gap between the theoretical and computational work made on sexual reproduction and recombination compared to more basic rearrangement events. The object of this thesis is to take a part in bridging this gap.

3 Models of rearrangements

As said in the first part of this introduction, chronologically, chromosomal rearrangements have been observed long before local mutations. As rearrangements were the only experimentally visible mutation type, it naturally caused reflections on their impact and evolutionary consequences. However, sequencing and the improvement of detection of local mutations overshadowed modeling work in this topic, up to a point that in textbooks such as *Theoretical evolutionary genetics* (Felsenstein, 2005), there are no theoretical results on rearrangements, while mutations are the topic of a whole chapter of the book.

As theories date from before the broadening of sequencing, these models are mostly conceptual and use a simplified representation of the genome. Their representation of

the genome could be called a "string of pearl" representation: without sequencing, the only visible effect of chromosomal rearrangement is the modification in gene copy number or in gene order. Hence, the genome can be represented as a string of a defined length with different pearls/genes on it and rearrangements only change the ordering of the pearls. In these basic representations, rearrangements do not affect genes, but just their ordering (synteny) and the distance between them. We will present here some theoretical models that leveraged this string of pearl approach to understand the evolutive impact of chromosomal rearrangements.

The first example is the hypothesis that chromosomal rearrangement such as inversions and translocations enable the control of homologous recombination. Inversions and translocation are neutral in the string of pearl metaphor, but they affect the pearl order, which in turn reduces recombination. In (Noor *et al.*, 2001), inversions are shown to prevent recombination, retrogression, and hybridization events, keeping alleles with a positive epistatic interaction together (Kirkpatrick, 2010; Hoffmann *et al.*, 2004). Similarly, it has been proposed that translocations reduce the recombination rate, keeping apart alleles that could be lethal together (de Waal Malefijt et Charlesworth, 1979).

Another well-supported theory concerns the origin of genes. It has been long shown that gene duplication is the primary source of new genes (Ohno, 1970a). With the string of pearl representation, a duplication of a coding part of the genome always corresponds to a full duplication of one or more coding genes. In this case, the two genes can either be conserved, diverge, specialize, adapt to each other, or disappear (Walsh, 2003). Interestingly, a shift towards more formal definitions of the outcomes of gene duplication can be observed in a recent publication (Kalhor *et al.*, 2023). In the conclusion, the authors acknowledge that the definition of gene duplication and their fate is more complex outside the string of pearl model.

There is an intrinsic difficulty in the theoretical modeling of chromosomal rearrangements: they are known to be large scale events (as witnessed by the string of pearl metaphor), but numerous examples show that their local effect on the gene sequence is not as negligible as the string of pearl metaphor suggests (Blount *et al.*, 2012a; Schobel *et al.*, 2016). This multiscale complexity, coupled by the diversity of life and of chromosomal rearrangements, limits theoretical development.

Computational modeling provides a valuable way to investigate the effects of chromosomal rearrangements. Indeed, computer simulations provide a perfectly controlled environment where any experiment is possible, as long as there is enough computational power. Computational modeling can have different goals. One can be to create a computational simulation of the world, in order to be as precise as possible with the computational constraints and possibly simulate impossible experiments. But models can also be created to support theoretical works, in order to test and illustrate quantitative hypotheses by computer simulations. However, the scientific value of these two types of computational models is really different. Indeed, on the one hand, the goal is to partially recreate the world to allow studying it more easily, and the other is mostly a computational verification of a mathematical theory. While there is a gray area between the two approaches, since we covered in the last paragraphs the theories, we will try to focus now on the computational models providing tools for simulations.

Among the simulators of evolution, very few are designed to model chromosomal rearrangements. For example, INDELible (Fletcher et Yang, 2009) has been made for

benchmarking purposes in phylogeny, to include Indels[1]. This benchmarking tool does not include chromosomal rearrangement and is still largely used today (Olabode *et al.*, 2023).

Some computational models are indeed used for phylogenic benchmarking or sequencing inference, as for example, ALF (Dalquen *et al.*, 2012) and EvolSimulator (Beiko et Charlebois, 2007). Since their primary goal is to study DNA sequence evolution, they simulate one genome per population and thus ignore all population effects. While these tools could prove interesting, their main focus is on sequence simulation and they both have significant drawbacks that prevent rearrangement studies. For example, in ALF all mutations are assumed to be either neutral or lethal, and in EvolSimulator the duplication probability is a gene-specific pre-assigned value (Knibbe et Parsons, 2014). Furthermore, the lack of population prevents the study of mutation fixation and epistatic effects that are likely to be particularly strong for chromosomal rearrangements.

There are few long-standing models that focus on studying evolution, namely, Aevol, Avida and SLiM. The former is used in this thesis and will be described in the next chapter. Among the two other, SLiM is more focused on analyzing population effects, and the implementation of the mutations is not adapted for rearrangement. Indeed, in SLiM, the effect of a mutation is coded to be located in a given position in the genome, including rearrangements would require a lot of changes. Berdan *et al.* (2021a) studied the fate of inversions in SLiM, but it models the inversion by manually adding the inversion and taking an allelic view of the genome, not by considering inversions as a part of the evolutive process. On Avida however, some rearrangements have been modelled. In Avida, the genome is represented by a series of instructions that are executed to compute the fitness of an organism. Duplications and deletions can occur from a "divide" mutation and they appear as "pivotal" mutation to acquire a specific function (Lenski *et al.*, 2003). Duplications are then introduced more formally as results of a slippage in genome reproduction by Lalejini *et al.* (2017). They show that duplications that conserve the gene content of the duplicated section are more efficient. Indeed, they study the impact both in fitness and evolvability of duplications compared to mutations that copy part of the genome but with a worse conservation of information. As explained in the conclusion of the article, conservation of information is decisive in the evolutionary influence of duplication. However, it could be argued that the amount of information conserved when breaking a sequence depends on the way this information is coded. Because genomes in Avida are represented as a series of computational instructions with no structure such as open reading frames, it can be argued that it is not an optimal model to study rearrangements.

Surprisingly, many results in theorizing mutations come from pure computer science. For example, genetic algorithms are bio-inspired optimization algorithms that rely on evolutionary processes for optimization. As early as 1992, Holland (1992) states that "Like the rules of a well-constructed game [...], genetic operators are simply defined but subtle in their consequences". Recombinations (generally called "cross-overs" in genetic algorithms) are shown to provide a great level of parallelism. While being interested in the genetic operators, the author states that for the field of genetic algorithms to move forward, it is important to move past the biological constraints. Vie *et al.* (2020) reviews multiple genetically inspired types of mutation, such as duplications, deletions,

[1]This program was made in London in 2009, not to be mistaken by InDelible, developped in Manchester (Gardner *et al.*, 2021).

translocation, inversions. There is also a "shuffle" mutation that randomly reorders genes, similar to a rare rearrangement called chromothripsis, and a "hyper mutation" which entails modifying a portion of the genome with random entries. Other mechanisms that are not biologically inspired are described in (Mirjalili *et al.*, 2020), for example, (Mauldin, 1984) focused on a new reproductive mutational operator that guarantees the diversity of the population.

A specific kind of evolutionary algorithm is genetic programming, which involves applying genetic algorithms to search for a solution to a problem in a space of programs. One of the important questions when thinking about programming is the programming language and the structure of the programs it involves. In the particular case of genetic programming, the language also determines the possible mutational operators and their potential effects (Whigham *et al.*, 1995). The information structure is often to be associated with the specific problem the genetic program is supposed to solve, from trees, to grammars, to graphs,... In (Ahvanooey *et al.*, 2019) a list of different types of genetic programming is proposed depending on the type of data structure (and thus the available genetic operators). Information structure also influences the resulting program in terms of complexity. Indeed, unnecessary complexity in the resulting program is a common problem called "bloating" in this scientific community, and methods such as increasing the mutation rate have been proposed to counter it (ONeill et Brabazon, 2019). This suggests that in order to simulate and study chromosomal rearrangements in biology, the structure of the genome in the simulation must be as close as possible to a biological genome.

In the next chapter, we will describe the Aevol model, a computational model that has been precisely designed to mimic the structure of information on biological genes. While Aevol is not used in all the chapters of this thesis, it was used as a stepping stone for both the computational models and the theoretical work presented here. The next chapter is thus an introductory chapter presenting Aevol. It also shows an example of how it can be used to study the impact of chromosomal rearrangements, and it raises the questions that we will try to answer in the rest of the thesis.

Chapter II

Forward-in-time Simulation of Chromosomal Rearrangements: The Invisible Backbone that Sustains Long-term Adaptation

Paul Banse*,[1], Juliette Luiselli*,[1], David P. Parsons[1],Théotime Grohens[2], Marco Foley[1], Leonardo Trujillo[1], Jonathan Rouzaud-Cornabas[1], Carole Knibbe[3], Guillaume Beslon[1]

*Paul Banse and Juliette Luiselli are joint first author
[1]Université de Lyon, INSA-Lyon, Inria, CNRS, Université Claude Bernard Lyon 1, ECL, Université Lumière Lyon 2, LIRIS UMR5205, Lyon, F-69621, France
[2]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain
[3]Université de Lyon, CarMeN Laboratory, INSERM, INRAE, INSA Lyon, Université Claude Bernard Lyon1, Pierre-Bénite, France

This article is accepted for publication in *Molecular Ecology* for the special issue "A genomic update on the evolutionary impact of chromosomal rearrangements".

Abstract

While chromosomal rearrangements are ubiquitous in all domains of life, very little is known about their evolutionary significance, mostly because, apart from a few specifically studied and well-documented mechanisms (interaction with recombination, gene duplication, etc.), very few models take them into account. As a consequence, we lack a general theory to account for their direct and indirect contributions to evolution. Here, we propose Aevol, a forward-in-time simulation platform specifically dedicated to unraveling the evolutionary significance of chromosomal rearrangements (CR) compared to local mutations (LM). Using the platform, we evolve populations of organisms in four conditions characterized by an increasing diversity of mutational operators from substitutions alone to a mix of substitutions, InDels and CR but with a constant global mutational rate. Despite being almost invisible in the phylogeny owing to the scarcity of their fixation in the lineages, we show that CR make a decisive contribution to the evolutionary dynamics by comparing the outcome in these four conditions. As expected, chromosomal rearrangements allow fast expansion of the gene repertoire through gene duplication, but they also reduce the effect of diminishing-returns epistasis, hence sustaining adaptation on the long-run. Last, we show that chromosomal rearrangements tightly regulate the size of the genome through indirect selection for reproductive robustness. Overall, these results confirm the need to improve our theoretical understanding of the contribution of chromosomal rearrangements to evolution and show that dedicated platforms like Aevol can efficiently contribute to this agenda.

Keywords : Chromosomal Rearrangements, Evolution, InDels, Modelling, Simulation.

1 Introduction

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

While new sequencing techniques and discoveries have shed a new light on chromosomal rearrangements (Quandt *et al.*, 2015; Ho *et al.*, 2020), theoretical frameworks have been slow to adapt. Indeed, the effect of chromosomal rearrangements is generally not addressed in theoretical articles and textbooks. In most models of evolution, substitutions are still the sole source of variation, with recombination merely expected to shuffle these variations among individuals (Weissman *et al.*, 2010). In the rare cases where ectopic recombination is considered in evolutionary models, its effect is generally limited to gene permutations or variation of copy number, excluding *a priori* any effect on gene sequence itself (Bhatia *et al.*, 2018; Yancopoulos *et al.*, 2005). Similarly, inversions are often viewed as just an evolutionary pathway that prevents recombination, hybridization, and introgression (Noor *et al.*, 2001), thus keeping specific alleles together (Kirkpatrick, 2010; Hoffmann *et al.*, 2004). Nevertheless, the ubiquity of these rearrangements calls for more in-depth studies of their potential other effects (Wellenreuther et Bernatchez, 2018).

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

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

The variety and complexity of chromosomal rearrangements makes it challenging to build a theoretical understanding of their effect on evolution. In this context, forward-in-time simulations are a promising tool to observe the effect of rearrangements and unravel their importance in adaptation to new environments (Mérot *et al.*, 2020). However, in forward-in-time models – like the well-known SLiM (Haller et Messer, 2017) –, the effect of mutations is often either an allelic change, drawn from a predefined DFE, or a positional change of the gene. This prevents these models from considering any combination of small- and large-scale effects, and makes it difficult to account for non-independent events (where some kinds of events modify the DFEs of others). To overcome these difficulties, a model designed to study rearrangements should not rely on explicit *a priori* DFEs. On the opposite, DFEs must emerge from how the mutations affect the genomic sequence, which depends on the characteristics of mutations themselves but also on the – evolving – genomic structure (gene number and positions, size of intergenic sequences, etc.). Hence, a model designed to study chromosomal rearrangements should provide an explicit genome with both coding and non-coding regions, in which rearrangements can happen blindly and have both direct (when altering coding regions) and indirect (when modifying the DFE of the different mutational operators – including rearrangements themselves) effects on fitness.

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

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

2 Material and Methods

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

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

In short, Aevol is made of three components (Fig. II.1A):

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

A detailed presentation of the model is available in the Supplementary Materials (Fig. S1, S2 and S3).

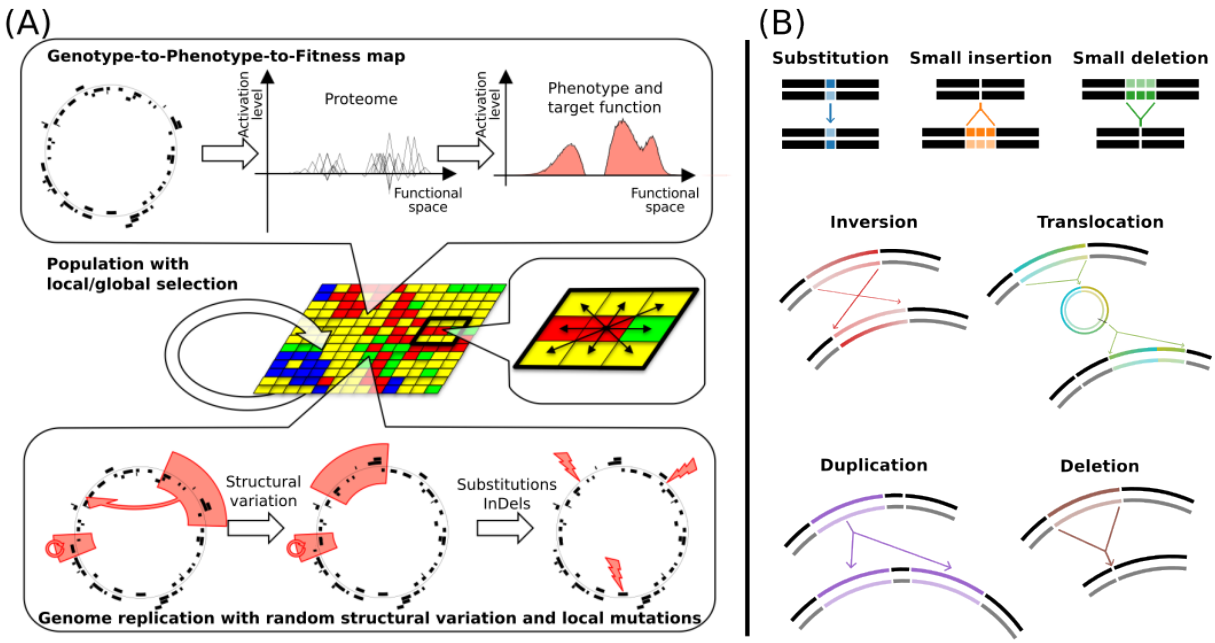


Figure II.1: **The Aevol model.** The left panel shows all steps of a generation in Aevol. (top) Overview of the genotype-to-phenotype map. Note that the organism shown here is a real organism evolved within Aevol for $1,000,000$ generations with a typical target. Hence, it contains many Open-Reading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

2.2 Experimental setup: Evolution with limited mutations

2.2.1 Experiment starting from naive individuals

We run 11 replicate simulations for four types of conditions: substitutions only (SUB), local mutations only (LM – substitutions and InDels), chromosomal rearrangements only (CR – duplications, deletions and inversions), and both chromosomal rearrangements and local mutations (CRLM). Note that translocations, although possible in Aevol, are excluded here as to have as many local mutations as chromosomal rearrangements. The median (in terms of final fitness) CRLM run will be used to start the second set of simulations. The simulations begin with naive individuals owning a single gene, and run for $1,100,000$ generations, which is enough to reach a stable genome.

All replicates share the same population size ($1,024$ individuals on a 32×32 square grid), the same environment (a sum of three Gaussian lobes, see Fig. II.2 and Supplementary Material, Fig. S5) and the same selection mode (local competition against the direct neighbors). The only difference lays in the mutation rates, as shown in Table II.1. Importantly, for each condition, mutation rates are equally balanced between all mutation types and adjusted such that the overall mutation probability per locus is constant throughout all experiments. An example parameter file for the CRLM setup is provided in the Supplementary Material (Fig. 4).

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

Table II.1: Mutation rates per base pair per generation for the four mutational scenarios: SUB, LM, CR and CRLM. Note that the total per base per generation mutation rate is constant across experiments.

For every simulation, we start from a single individual of the final population and reconstruct its lineage by tracking its ancestor at every generation. We then compute fitness, genome size, coding and non-coding sizes, and number of genes of the individuals in this lineage. We also extract all fixed mutations, and record their type and effect on fitness. To ensure that the lineage we study is composed of the ancestors of the whole final population, and that the observed mutations went to fixation, we remove the data between generations $1,000,000$ and $1,100,000$.

Finally, we also study the distribution of fitness effect for each type of mutation on the median (in terms of final fitness) CRLM individual. This allows to better understand the differences between local mutations and chromosomal rearrangements in terms of impact on the fitness and chances of fixation.

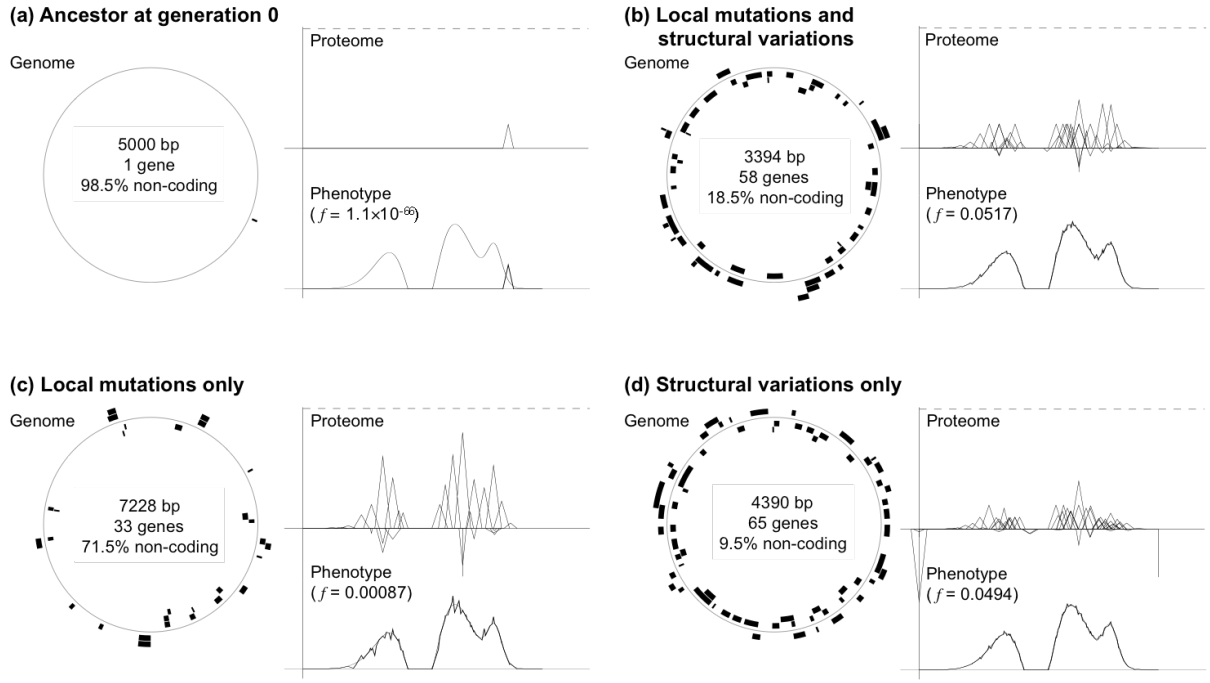


Figure II.2: Initial ancestor (a) and examples of evolved organisms in the CRLM (b), LM (c) and CR (d) conditions after 1,000,000 generations. The organism presented in (b) corresponds to the Wild-Type used for the second step of the experiments. For each organism, there is on the left a visualisation of its genes localised on the genome. On the right, the proteome shows all the single proteins, and the phenotype (black curve) is their sum. The grey curve plotted in addititon to the phenotype is the environmental target, a sum of 3 Gaussian lobes (2 positives and 1 negative) – see Supplementary Material, Fig. S5.

2.2.2 Evolution from Wild-Types

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

2.2.3 Fitting fitness trajetories

In order to estimate diminishing-returns epistasis, *i.e.* how fast the advantage provided by each new beneficial mutation reduces over time, we fit the fitness values along the line of descent with power laws of type $f = (bt + 1)^a$ where f is the fitness, t is the time in generations (Wiser *et al.*, 2013). a and b are the parameters to be fitted with a

corresponding to the diminishing-returns epistasis when $0 \leq a \leq 1$ ($a = 1$ corresponding to linear fitness growth without diminishing-returns epistasis) and b corresponding to an initial fitness growth parameter.

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

3 Results

To investigate the contribution of chromosomal rearrangements to evolutionary innovation, we compare the evolutionary dynamics of four sets of runs: SUB, with only substitutions; LM, with only local mutations; CRLM, with both local mutations and chromosomal rearrangements; and CR, with only chromosomal rearrangements. As we suspect that the relative contribution of chromosomal rearrangements versus local mutations depends on the distance to the fitness optimum, we repeated these experiments in two conditions: starting with naive individuals (see Section 3.1) or with pre-evolved ones (WT – see Section 3.2).

3.1 Local mutations are dispensable when far from the optimum

As shown in Figure II.2 and Figure II.3, the evolutionary trajectories in terms of fitness, genome size and number of genes without local mutations (CR) are similar to the evolutionary trajectories with both rearrangements and local mutations (CRLM), whereas the simulations without rearrangements (SUB and LM) produce significantly less adapted organisms, with fewer genes and a smaller coding genome size despite a greater total genome size.

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

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

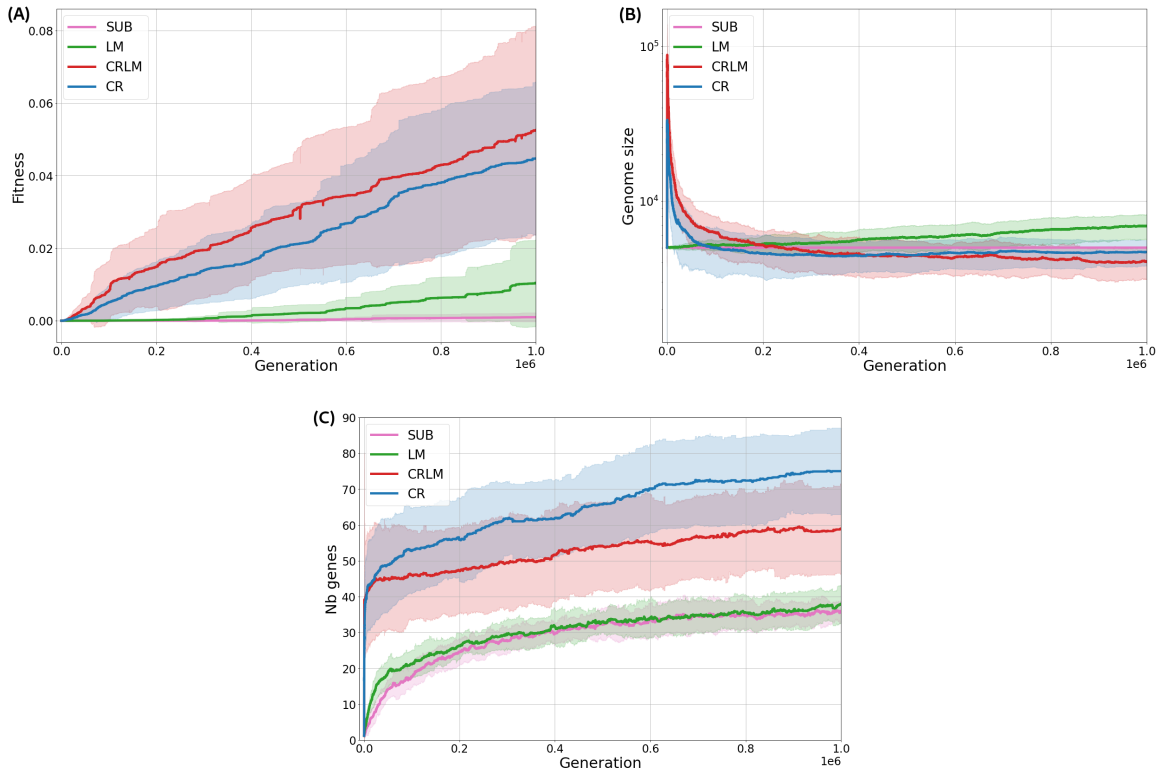


Figure II.3: Variation of fitness (A), genome size (B) and gene number (C) on the line of descent of the final population, starting from a naive individual for the four mutational scenarios. The shaded areas indicate the variability across the 11 repetitions (standard deviation).

pace. This is caused by the fixation of more small insertions than small deletions (see Fig. II.4B). Ultimately, genomes evolved under the LM setup are longer than genomes evolved under the CR and CRLM setups but they contain much fewer genes, resulting in a larger proportion of non-coding DNA (see Fig. II.2).

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

To better understand the origin of these differences, we first look at the contribution of each mutation type to the end fitness. We computed the total gain of fitness per mutation type along the ancestral lineage during the $1,000,000$ generations of each experiment (Fig. II.4A). Interestingly, although CRLM are much fitter than LM, it is still the local mutations that contribute the most to the overall fitness gain in CRLM. Local mutations are crucial to evolution, and it is not surprising that they are the most impactful. However, the difference between SUB, LM and CRLM shows that their potential is only fully unleashed when chromosomal rearrangements are also present and create a substrate that local mutations can then finely tune.

The number of non-neutral mutations fixed along the line of descent (Fig. II.4B) shows that rearrangements, although rarely fixed compared to local events and hence

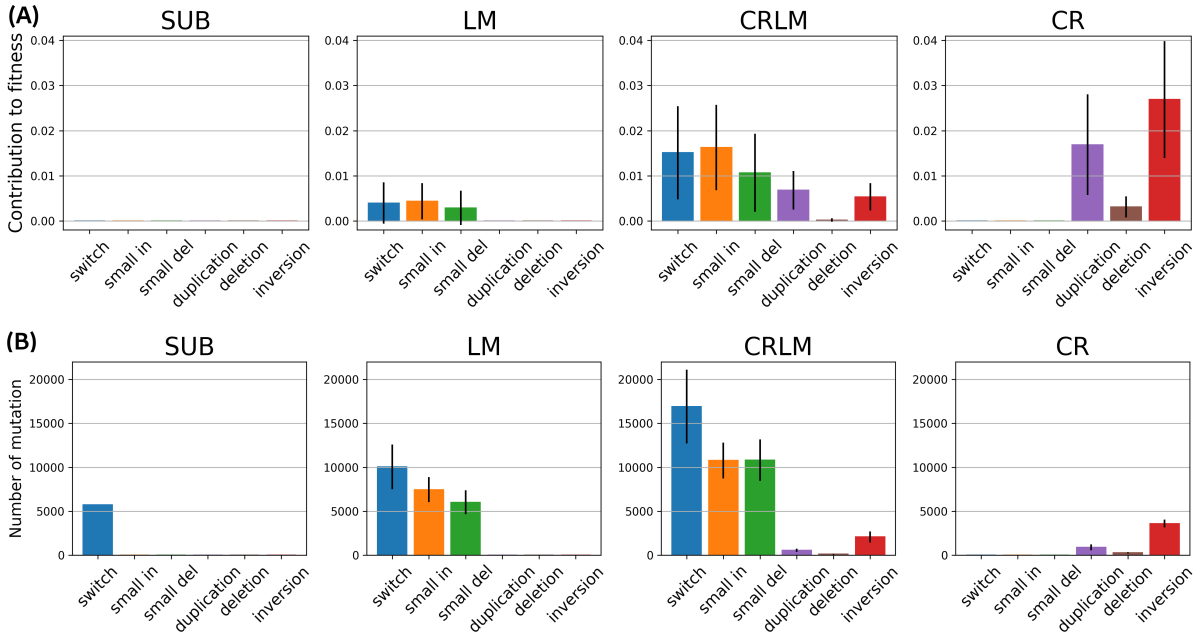


Figure II.4: **Fitness contribution and number of mutations fixed during the initial evolution from naive individuals** (A) Contribution of each type of mutation to the total fitness gains, measured as the sum of the change in fitness of each mutation on the line of descent of the final best individuals, starting from naive individuals. Histograms show the mean values across the 11 repetitions, and the bars show their standard deviation. Reverted mutations (mutations which effect on fitness was exactly compensated by the following one) were filtered out to reduce noise. Fitness increase in the SUB simulations are negligible at this scale. (B) Number of non-neutral and non-reverted mutations fixed for the different mutation types and for the four conditions, normalized by the number of mutations occurring ($L \times \mu$, with L the genome size), on the line of descent of the final best individuals, starting from naive individuals. Histograms show the mean values across the 11 repetitions, and the bars show their standard deviation.

almost invisible in the phylogeny, favor the fixation of beneficial local mutations. This is consistent with the dynamics of gene number shown on Fig.II.3C: by allowing for the recruitment of more genes, rearrangements increase the number of potential mutational targets on which local events can have an effect, hence favoring the fixation of more favorable local events.

The very rare fixation of rearrangements compared to the fixation rate of local mutations can be better understood by looking at the distribution of fitness effects (DFE) for each type of mutation (see Fig. II.5). Duplications and deletions have a very broad effect and can disturb, delete or imbalance essential genes: they are therefore very often lethal (in approximately 95% of cases here). Local mutations, on the other hand, have a smaller chance of disrupting an essential gene, as they affect a restricted section of the genome. They are more often neutral or "simply" deleterious, and lethal only in less than 40% of cases. Finally, inversions have two breakpoints while local mutations have only one, and are therefore more lethal than local mutations (80%), but, as inversions are balanced rearrangements, they are less likely to be deleterious than duplications or deletions.

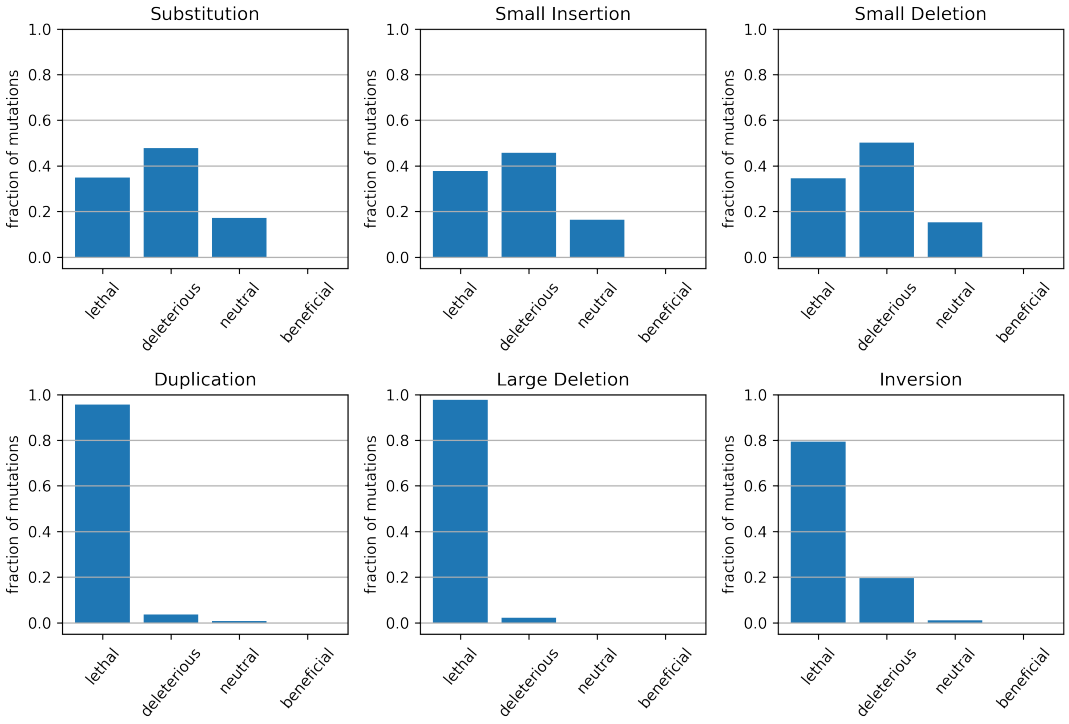


Figure II.5: **Distribution of fitness effect of the different types of mutation**, on the median individual of the CRLM experiment, after 1,000,000 generations when starting from a naive individual. For each mutation type, 1,000,000 mutants were generated, except for the substitution, which were exhaustively tested. The selection coefficient is computed as $s = \frac{f_{mutant}}{f_{parent}} - 1$. Lethality is defined as $s < -0.999$, and neutrality as $s \in [-0.001, 0.001]$. The detailed Distribution of Fitness Effect (DFE) is presented in Supplementary Material (Fig. S4).

3.2 Chromosomal rearrangements sustain long-term adaptation

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

Fig.II.6A shows that the four conditions result in very different dynamics of genome size. While the genome size of CR and CRLM experiments is quite stable, as observed at the end of the previous experiments, in LM conditions the genome size increases continuously during the 3 million generations of the experiment. At first sight, this result may seem contradictory, as the genome size is much more likely to vary in the presence of long segmental duplications/deletions than in the sole presence of small InDels. This shows the complex effect of chromosomal rearrangements in regulating genome size, and highlights the difference between InDels and rearrangements in doing so.

As expected, when looking at the fitness gain along the 3 million generations of the

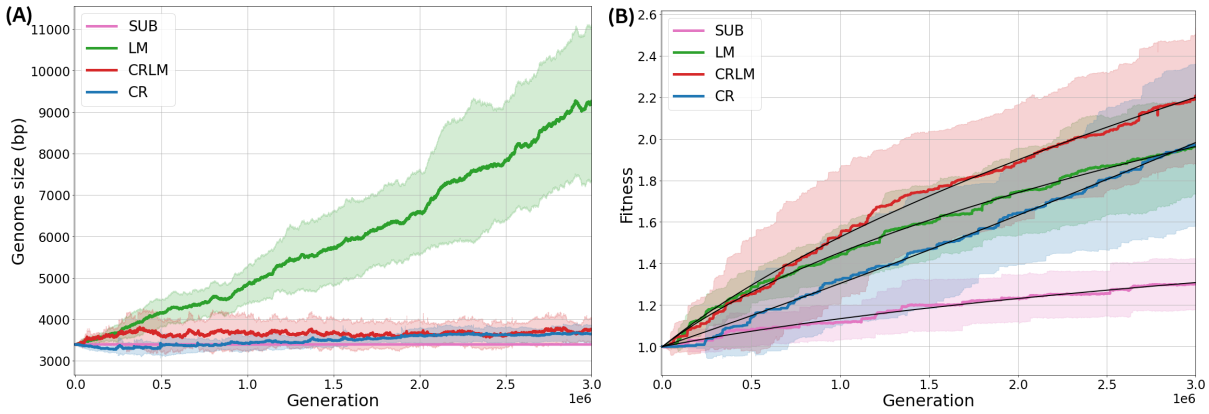


Figure II.6: **Temporal changes in genome size and fitness in evolution started from the WT.** (A) Mean change in genome size on the line of descent of the final populations, for the 11 repetitions and the 3 conditions. All simulations started from the same Wild Type with a genome length of 3394 bp (Fig. II.2.b) and evolved for 3,000,000 generations. The shaded areas indicate the variability across repetitions (standard deviation). (B) Relative fitness variation on the line of descent of the final population, starting from a Wild Type. The shaded areas indicate the variability across repetitions (standard deviation). Black curves show the fitted power laws for the mean fitness values of the four sets of simulations (see Methods, section 2.2.3). The fitted parameters are: $a_{SUB} = 0.2$, $b_{SUB} = 7.0 \times 10^{-7}$, $a_{LM} = 0.4$, $b_{LM} = 1.8 \times 10^{-6}$; $a_{CR} = 1.5$, $b_{CR} = 2.0 \times 10^{-7}$, $a_{CRLM} = 0.5$, $b_{CRLM} = 1.5 \times 10^{-6}$.

experiment (Fig. II.6B) the difference between the mutational scenarios is not as marked as what was observed when far from the optimum, at least for the LM, CR and CRLM scenarios. Yet, the SUB scenario still clearly lags behind in terms of fitness, showing again that substitutions alone are not sufficient in fine-tuning genes. In the four conditions, fitness improves all along the experiment, albeit with a clear diminishing-returns epistasis in the SUB, LM and CRLM conditions. Following Wiser *et al.* (2013), we used power-law curve fitting to estimate the amount of diminishing-returns epistasis in the four conditions (black lines on Fig. II.6B). Results show that diminishing-returns epistasis is higher in the SUB and LM conditions than in the CRLM conditions ($a_{SUB} = 0.2$; $a_{LM} = 0.4$; $a_{CRLM} = 0.5$ – see Methods, section 2.2.3) which, in the long run, advantages the CRLM over the other scenarios. Strikingly, when evolving only with chromosomal rearrangements (CR scenario), populations show no diminishing-returns epistasis throughout the duration of the experiment ($a_{CR} = 1.5 > 1$). This contrasts with the other conditions and allows the CR populations to catch up with the SUB and LM ones, despite an initial disadvantage.

As previously, we measured the total fitness effect and the number of non-neutral mutations fixed along the lineage for the different types of mutation and for the four mutational scenarios (Fig. II.7A and II.7B respectively). As already noticed when starting far from the optimum, this shows that chromosomal rearrangements, although very rarely fixed in the lineage, have a dual contribution to fitness. While, in the CRLM, fixed rearrangements have a small impact on fitness on their own (Fig. II.7A), they also contribute to increasing the number of favorable substitutions. Indeed, substitutions and InDels are more likely to be favorable and fixed in the CRLM populations than in the

LM populations and almost as likely – for the substitutions – as in the SUB ones (Fig. II.7B). This leads to a sustained evolutionary dynamics, despite rearrangements being almost invisible in the phylogeny owing to their very low fixation probability.

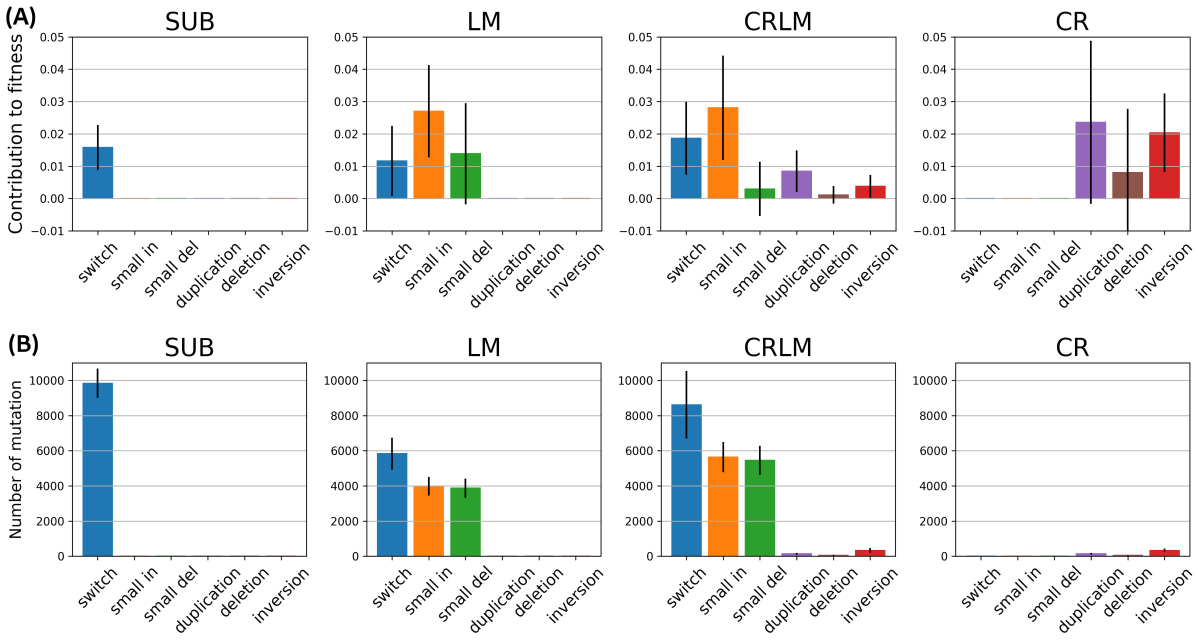


Figure II.7: **Fitness contribution and number of mutations during the evolution from WT individuals** (A) Contribution of each type of mutation to the total fitness gains, measured as the sum of the change in fitness of each mutation on the line of descent of the final best individuals, starting from WT individuals. Histograms show the mean values across the 11 repetitions, and the bars show their standard deviation. Reverted mutations (mutations which effect on fitness was exactly compensated by the following one) were filtered out to reduce noise. (B) Number of fixed non-neutral and non-reverted mutations per generation for the different mutation types per million generation, normalized by the number of mutations occurring ($L \times \mu$, with L the genome size), on the line of descent of the final best individuals, starting from WT individuals. Histograms show the mean values across the 11 repetitions, and the bars show their standard deviation.

4 Discussion

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

Diss et Lehner, 2018), the epistatic effect of rearrangements is, with very few exceptions (Blount *et al.*, 2012a), *terra incognita*.

Here we used Aevol to simulate genome evolution under several conditions characterized by an increased mutational diversity but a constant overall mutational rate (see Table II.1). We completed these experiments by testing evolution under the exclusive pressure of chromosomal rearrangements, in order to estimate their capacity to generate enough variation to allow sustained evolution. This enables an experimental (though simulated) exploration of the consequences of chromosomal rearrangements on the evolutionary dynamics. Specifically, we analyzed the results of the simulations with a focus on two levels: genome structure, which is likely to be largely impacted by rearrangements, and individuals' fitness.

Regarding the evolution of genome structure, our results show two clear differences when genomes evolve with (CRLM and CR simulations) or without (SUB and LM simulations) chromosomal rearrangements. First, they confirm the well-established theory of evolution by gene duplication (Zhang, 2003; Kalhor *et al.*, 2023): in our simulations, rearrangements are essential for the rapid acquisition of a large gene repertoire, and duplications are the main cause of increase in gene number (see Section 3.1). Indeed, gene number rapidly increases in the very first thousands of generations for CR and CRLM (Fig. II.3C), and this process of gene recruitment is maintained throughout the simulation, though at a lower pace. On the opposite, lineages evolving without rearrangements only acquire a limited gene repertoire (see Fig. II.3C).

In a less intuitive way, our simulations show an important contribution of chromosomal rearrangements to the stabilization of genome length during evolution. Indeed, figures II.3B and II.6A show that, after an initial burst of genome size at the very beginning of the evolution (corresponding to the phase of fast gene acquisition through duplications), CR and CRLM lineages quickly undergo a reduction of their genome size (while preserving their gene repertoire – see Fig. II.3B and C). Continuing the simulation for 3 million generations, we see that genome size varies very little thereafter (Fig. II.6A). This dynamic contrasts sharply with that of the LM lineages, which show a steady increase in genome size, both when starting far or close to the optimum. This sustained growth of genome size under the sole pressure of InDels advocates in favor of the mechanism of border-induced selection, which has been recently conceptualized by Loewenthal *et al.* (2022). Indeed, despite their spontaneous mutation rates being equal, the probability of fixation of neutral insertions is slightly higher than the probability of fixation of neutral deletions, due to interference with gene borders (Loewenthal *et al.*, 2022): a small insertion close to a gene is most often harmless, while a small deletion at the same point can impact a gene if the size of the deletion is larger than the distance to this gene. In the absence of other constraints on the genome size, this bias leads to an infinite growth in genome size, as we observe on Fig. II.3B and II.6A. Strikingly, in the presence of chromosomal rearrangements, this bias is not visible anymore, showing that rearrangements generate an evolutionary pressure that prevents genome growth. As already proposed by Knibbe *et al.* (2007b), deleterious chromosomal rearrangements lead to selection for robustness, favoring smaller genomes as these undergo fewer rearrangements than longer ones. This hypothesis is sustained by the low rate of fixation of chromosomal rearrangements (Fig. II.7B): they are largely filtered-out by purifying selection, suggesting that they have a strong robustness effect.

The low number of fixed rearrangements, due to their high lethality, (Fig. II.5) questions the concept of mutation rate. Indeed, by measuring mutation rates on a live population, a bias is introduced towards non-lethal mutations. This bias has been observed in the case of substitutions (Wang *et al.*, 2012) but we hypothesize that this could be even more important in the case of genome rearrangements, and models should take into account that spontaneous mutation rates could be very different from observed and fixed ones.

The influence of chromosomal rearrangements on fitness evolution is also very different depending on whether the simulations start far from the optimum (hence requiring them to acquire new genes) or close to the optimum (with a gene pool already acquired but still to be optimized). In the former situation, lineages evolving in the presence of chromosomal rearrangements have a much higher fitness than those evolving with only substitutions, or even with all local mutations (Fig. II.3A). This confirms that, in such a situation, gene duplication has a decisive contribution (Zhang, 2003; Kalhor *et al.*, 2023), enabling both the CR and CRLM lineages to largely overcome the LM and the SUB lineages. Strikingly, lineages evolving with chromosomal rearrangements only (CR) perform almost as well as those evolving with both chromosomal rearrangements and local mutations (CRLM). This illustrates the multiscale nature of chromosomal rearrangements that can both enlarge the gene repertoire through large duplications but also optimize gene sequences by reorganizing them through *e.g.* inversions, as shown by Trujillo *et al.* (2022). Interestingly, the fitness of the SUB lineages (that evolved under the sole pressure of substitutions) is much lower than the fitness of the LM lineages (that evolved through substitutions and InDels) despite a very similar dynamic of gene recruitment. This confirms that small insertion and small deletions are decisive operators when the evolution of protein sequence is concerned, as they can add/remove codons when substitutions can only mutate existing ones (Vakhrusheva *et al.*, 2011; Leushkin *et al.*, 2012).

When starting close to the fitness optimum, the differences between the experiments are more subtle, except when substitutions are the sole mutational operator (SUB curve on Fig. II.6B), in which case fitness gains are much lower than in the three other conditions, highlighting the importance of the diversity of mutational operators (Berdan *et al.*, 2021b). In all experiments, the dynamics of fitness is similar to what can be observed *in vitro*, for example in experimental evolution with bacteria (Wiser *et al.*, 2013; Wang *et al.*, 2016), or yeast strains (Wei et Zhang, 2019): simulations show a sustained fitness gain all along the experiment albeit with a more or less pronounced diminishing-returns epistasis. Inspired by Wiser *et al.* (2013), we estimated the diminishing-returns epistasis in these different conditions, and showed that, in the long run, chromosomal rearrangements reduce diminishing-returns epistasis, hence enabling sustained evolutionary dynamics. Moreover, as shown by Fig.II.7A, the effect of rearrangements is mainly indirect: they have a small effect by themselves but potentiate other factors. Indeed, in the CRLM lineage, substitutions have a larger impact than in the SUB and LM lineage. This suggests that rearranged sequences open new targets to substitutions, hence increasing the probability to fix beneficial local events (Fig.II.7B). Finally, as Fig.II.7B also shows, this effect is due to only a very low number of fixed rearrangements. Hence, while rearrangements sustain long-term adaptation by reducing the effect of diminishing-returns epistasis, they are almost invisible in the phylogeny.

When quantifying the diminishing return, a striking result was the apparent accelerat-

ing evolution in the CR populations ($a_{CR} > 1$). We hypothesize that this is due to the low fixation rate of chromosomal rearrangements (Fig. II.7B). As CR populations undergo only rearrangements, fitness comparatively evolve by bigger steps but with longer waiting times between mutations, and this creates an initial lag in the fitness gain (Fig. II.6B), hence the appearance of acceleration. Now, the number of possible rearrangements for a given genome is much larger than the number of possible local events (it is indeed mainly linked to the number of breakpoints to be chosen for a given type of event: one for local mutations, two for inversions and deletions, three for duplications – see Fig. II.1B). A direct consequence is that, contrary to substitutions and InDels, rearrangements neighborhood cannot be explored in a reasonable time, hence the lower diminishing-returns epistasis observed on the duration of our simulations when rearrangements are allowed. Further, exploring this question, *e.g.* by estimating the contribution of each type of rearrangement to the phenomenon, is a very promising research direction opened by our results.

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

All models rely on simplifying assumptions, and ours makes no exception. However, the interest of modelling is precisely to reduce the complexity of the system to be studied. Here, studying only a limited number of mutational operators has enabled us to identify effects that could have been blurred in a more complex setting. Indeed, our experimental strategy, which relies on a progressive complexification of the mutational repertoire, has enabled us to uncover profound differences between chromosomal rearrangements and small InDels, both in the evolution of genome size and in the adaptation of organisms. Both kinds of events may seem rather similar at first sight, but they differ on two important aspects: first, contrary to duplications that copy preexisting genomic sequences, small insertions add random sequences to the genome. Hence, they cannot duplicate genes, while this process is central in evolution (Zhang, 2003). Second, even though both types of mutation add/remove genomic segments to the chromosome, the distribution of the size of these segments is different: in the case of InDels, this distribution is fixed while in the case of rearrangements, the distribution depends on the size of the genome. A direct consequence of this property is that because a larger genome leads to more numerous

deleterious rearrangements, it also leads to a lower robustness (Knibbe *et al.*, 2007b). In our simulations, large duplications and deletions, far from randomly shuffling the genome size as could have been expected, impose a tight constraint on it.

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

Our conclusions are drawn from the comparison of the evolutionary trajectories of different experiments and open up several interesting perspectives. For example, Aevol also includes several analysis tools, such as the computation of the distribution of fitness effect for all mutation types and for all genomes along a lineage, as illustrated by Fig. II.5. Taking advantage of the perfect record of the mutational events, these measures help quantify the evolutionary forces at work, as well as the relative contribution of the different types of mutation to these forces. As exemplified on Fig. II.4A and II.7A, the impact of the different types of mutation on the fitness can easily be quantified, allowing to estimate the direct contribution of each type of mutation. Although it would be very computationally demanding, it could be interesting to also quantify the consequences of each mutation type on robustness and evolvability as this could allow to estimate their indirect effect and explain how the different types of mutations interact. Finally, as long as chromosomal rearrangements are concerned, an obvious prospect is to extend the model to diploid eukaryote-like genomes with recombination. This would enable exploring their interplay with recombination (Berdan *et al.*, 2021a).

The experiments we presented here only scratch the surface of what can be done with Aevol. Indeed, as Table S1 of the Supplementary Material shows, many other experiments can be done, including testing the effect of mutation rates, mutation biases or population size. Aevol is available to any team that would like to test hypotheses regarding the effect of these parameters on the evolutionary dynamics and on genome structure. Moreover, as the code is open and freely available, any team can modify it to test some specific mutation type that would not already be implemented (see Section 4).

Despite the highly artificial nature of our model, our simulations are consistent with

the classical view of evolution: among the variety of mutational operators, substitutions and small InDels are by far the most visible adaptive events both in terms of their number (Fig. II.4B and II.7B) and their contribution to the fitness (Fig. II.4A and II.7A). However, our simulations also show that the scarcity of rearrangements that we observe in the phylogenies masks an important contribution to adaptation. While the vast majority of models and simulators of molecular evolution still implements a solely allelic view of evolution, where rearrangements can modify gene organization but cannot create new gene sequences, our results suggest that the innovative potential of rearrangements is not marginal, and that it is essential to integrate them into population genetics models.

Acknowledgements

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

Data accessibility and Benefit-Sharing

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

Author Contributions

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

Supplemental Information for:

Forward-in-time simulation of chromosomal rearrangements: The invisible backbone that sustains long-term adaptation

Paul Banse*, Juliette Luiselli*, David P. Parsons,
Théotime Grohens, Marco Foley, Leonardo Trujillo,
Jonathan Rouzaud-Cornabas, Carole Knibbe, Guillaume Beslon

5 Supplementary Material

5.1 S1. Aevol: a forward-in-time evolutionary simulator with complex mutations

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

In short, Aevol is made of three components (Fig. II.8A):

- A mapping that decodes the genomic sequence of an individual into a phenotype and computes the corresponding fitness value.
- A population of organisms, each owning a genome, hence its own phenotype and fitness. At each generation, the organisms compete to populate the next generation.
- A genome replication process during which genomes can undergo several kinds of mutational events, including chromosomal rearrangements and local mutations. The seven modelled types of mutation are depicted on Fig. II.8B and entail three local mutations: substitutions, small insertion, small deletion, two balanced rearrangements (which conserve the genome size): inversions and translocations, and two unbalanced rearrangements: duplications and deletions. This allows the user to study the effect of chromosomal rearrangements and their interaction with other kinds of events such as substitutions and InDels.

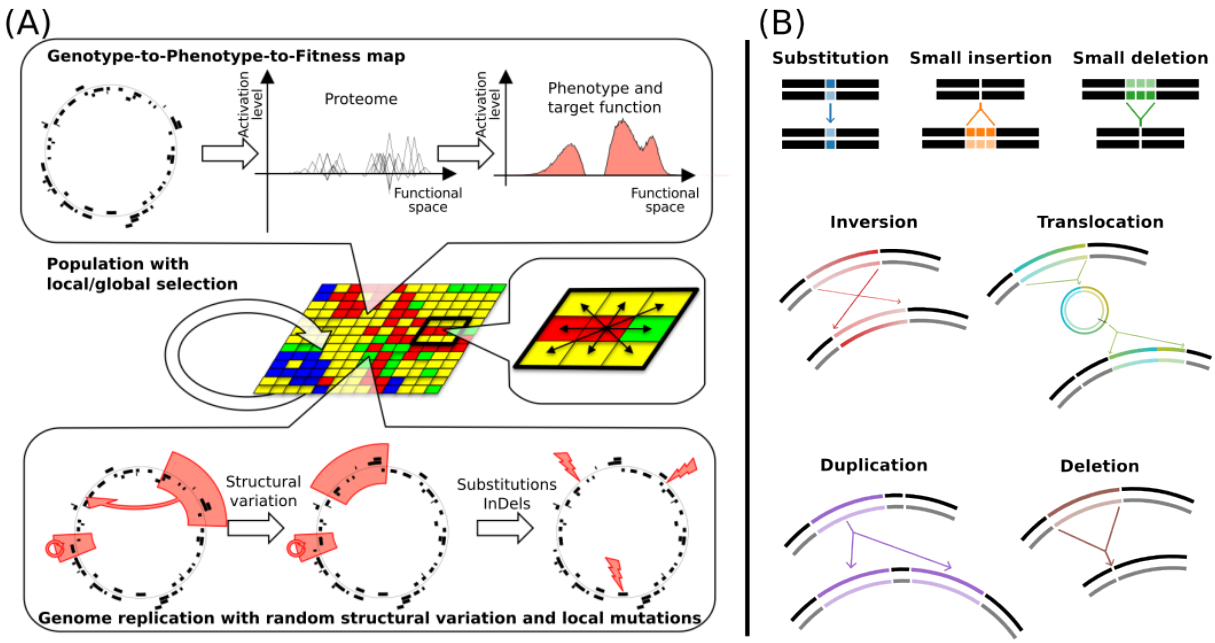


Figure II.8: **The Aevol model.** The left panel shows all steps of a generation in Aevol. (top) Overview of the genotype-to-phenotype map. Note that the organism shown here is a real organism evolved within Aevol for $1,000,000$ generations with a typical target. Hence, it contains many Open-Reading Frames (ORF) on both strands, many proteins and it is well adapted to its environment (*i.e.* its phenotypic function black curve is very close to the target function light red plain curve). (middle) Population on a grid is fully renewed every generation. Example of a local selection process occurring with a 3×3 neighborhood. (bottom) Mutation operators include chromosomal rearrangements (duplications, deletions, translocations and inversions – here a translocation and an inversion are shown) and local mutations (substitutions and InDels). These mutations are described more precisely in the right panel: (top) Local mutations: substitution (one base pair is mutated to another), small insertion and small deletion (a few base pairs are inserted or deleted). (middle) Balanced chromosomal rearrangements: inversion (two points are drawn and the segment in between is rotated) and translocation (a segment is excised, circularized, re-cut and inserted elsewhere in the genome). (bottom) Unbalanced chromosomal rearrangements: duplication (copy-paste of a segment in the genome) and deletion (suppression of a segment of the genome).

5.1.1 The Genotype-to-Phenotype-to-Fitness map

Genome representation. Each artificial organism, similarly to prokaryotes, is asexual, haploid, and owns a single circular chromosome. The genome is encoded as a double-strand binary string containing a variable number of genes separated by non-coding sequences (Fig. II.9). Genes are delimited by predefined signaling sequences indicating transcription and translation. The number of proteins an organism owns thus depends on its signaling sequences, and can evolve through mutational events.

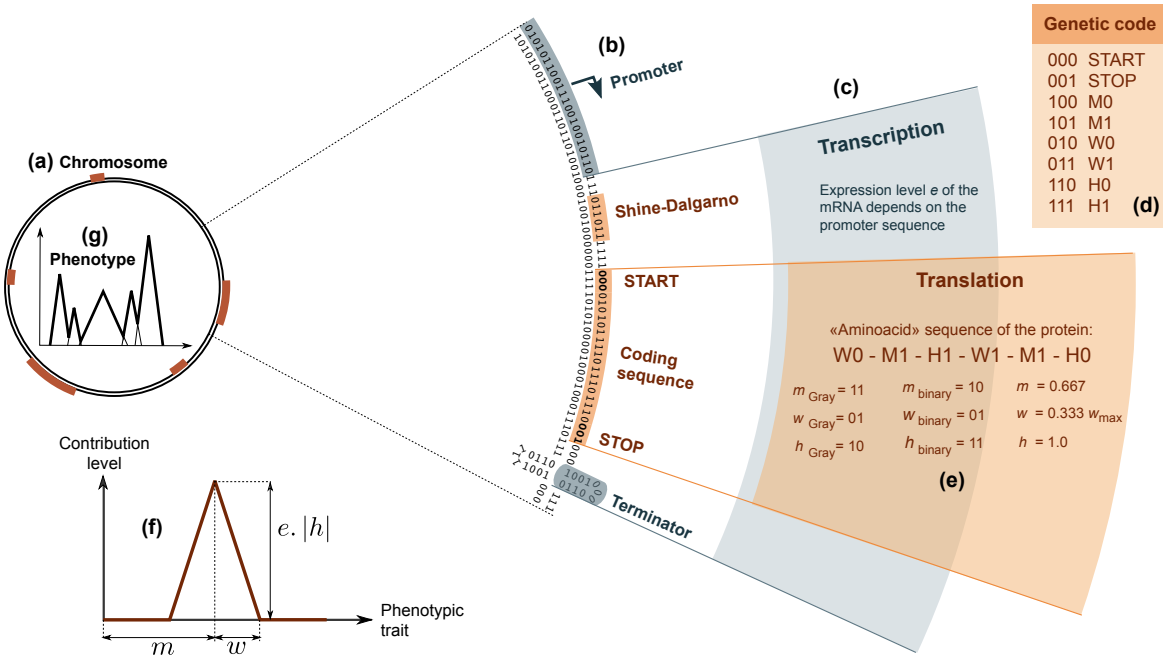


Figure II.9: In the model, each organism owns a circular double-strand binary chromosome **(a)** along which genes are delimited by predefined signaling sequences **(b)**. Promoters and terminators mark the boundaries of RNAs **(c)**, within which coding sequences can in turn be identified between a Shine-Dalgarno-START signal and a STOP codon. Each coding sequence is then translated into the primary sequence of a protein, using a predefined genetic code **(d)**. This primary sequence is decoded into three real parameters called m, w and h **(e)**. Proteins, phenotypes, and environments are represented similarly through mathematical functions that associate a level in $[0, 1]$ to each abstract phenotypic trait in $[0, 1]$. For simplicity reasons, a protein's contribution is a piecewise-linear function with a triangular shape: the m, w and h parameters correspond respectively to the position, half-width and height of the triangle **(f)**. All proteins encoded in the chromosome are then summed to compute the phenotype **(g)** that, once compared to the environmental target, can be used to compute the fitness of the individual.

Transcription starts at promoters, which are defined in the model as sequences that are close enough to an arbitrarily chosen consensus sequence (010101100111001001011 0 in all simulations presented here, with at most $d_{max} = 4$ mismatches). The expression level e of an mRNA is determined by the similarity between the actual promoter and the consensus sequence: $e = 1 - \frac{d}{d_{max}+1}$ with d the number of mismatches ($d \leq d_{max}$). This

models the interaction of the RNA polymerase with the promoter, without additional regulation.

When a promoter is found, transcription proceeds until a terminator is reached. Terminators are defined as sequences that would form a stem-loop structure, as the ρ-independent bacterial terminators do. The stem size is here set to 4 and the loop size to 3.

The translation initiation signal is $011011{*}{*}{*}{*}000$, corresponding to a Shine-Dalgarno-like sequence followed by a Start codon 000. When this signal is found on an mRNA, the downstream Open-Reading-Frame (ORF) is read until the termination signal (the Stop codon 001), is found. Each codon lying between the initiation and termination signals is translated into an abstract "amino-acid" using an artificial genetic code, thus giving rise to the sequence of the protein (Fig. II.9). Transcribed sequences (mRNAs) can contain an arbitrary number of ORF, with some mRNAs possibly containing no ORF at all (non-coding mRNAs) and others possibly containing several ORFs (polycistronic mRNAs). Importantly, the relative fractions of non-coding, monocistronic and polycistronic mRNAs are not predefined but result from the evolutionary dynamics and are likely to be influenced by the evolutionary conditions (Parsons *et al.*, 2010).

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

For simplicity, we use piecewise-linear functions with a symmetric, triangular shape to model protein effect (Fig. II.9). This way, only three parameters are needed to characterize the contribution of a given protein: the position $m \in \Omega$ of the triangle on the axis, its half-width w ($w \leq W_{max}$) and its height $h \in [-1, 1]$. This means that this protein contributes to the phenotypic traits in $[m - w, m + w]$, with a maximal contribution h for the traits closest to m. Thus, various types of proteins can co-exist, from highly efficient (high h) to poorly efficient (low h) and even inhibiting (negative h) and from highly specialized (low w) to versatile (high w).

In this framework, the primary sequence of a protein is interpreted in terms of three interlaced binary subsequences that will in turn be decoded as the values for the m, w and h parameters (Fig. II.9). For instance, the codon 010 (resp. 011) is translated into the single amino acid $W0$ (resp. $W1$), which means that it concatenates a bit 0 (resp. 1) to the code of w. Mutations in the coding sequences, including of course local mutations but also chromosomal rearrangements, can change these values and hence change the protein's contribution to the phenotype.

The contribution of all the proteins encoded in the genotype of an organism are combined to get the final level for each phenotypic trait. This is done by first scaling all protein contributions by the transcription rate e of the corresponding mRNA (see above), then by summing the mathematical functions of all the proteins, with bounds in 0 and

1. The resulting piecewise-linear function $f_P : \Omega \to [0, 1]$ is called the phenotype of the organism.

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

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

To illustrate this computation, we can look at the organisms in Fig II.10: at generation 0 (Fig II.10a), there is a single gene, thus a single protein in the proteome. Since in our experiment $k = 1,000$ (Fig. II.11), we exponentiate $-1000\times$ the difference between the environmental target (the sum of Gaussian lobes in grey) and the phenotype (the single black triangle), and we obtain a fitness $f = 1.1 \times 10^{-66}$. To compare, our Wild Type (Fig. II.10b) has 58 genes, so its phenotype is the sum of the 58 triangles depicted in the proteome and its fitness is $f = 0.0517$: its phenotype is much closer to the environmental target that at generation 0.

5.1.2 Population model and selection process

The population is modelled as a toroidal grid with one individual per grid cell. At each generation, the fitness of each individual is computed, and the individuals compete to populate each cell of the grid at the next generation. This competition can be fully local (the 9 individuals in the neighborhood of a given cell competing to populate it at the next generation, Fig. II.8A) or encompass a larger subpopulation. If the selection scope encompasses the whole population, all individuals compete for all grid cells. Importantly, the more local the selection scope, the more the population model diverges from the panmictic Wright-Fisher model as local selection increases the effective population size N_e for a given census population size (Waples, 2010).

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

5.1.3 Genetic operators

During their replication, genomes can undergo sequence variations (Fig. II.8). An important feature of the model is that, given the Genotype-to-Phenotype map (section 5.1.1), any genome sequence can be decoded into a phenotype (although possibly with no trait activated if there is no ORF on the sequence). This allows to implement – and test – any kind of mutational process. In the classical usage of the simulator, seven different kinds

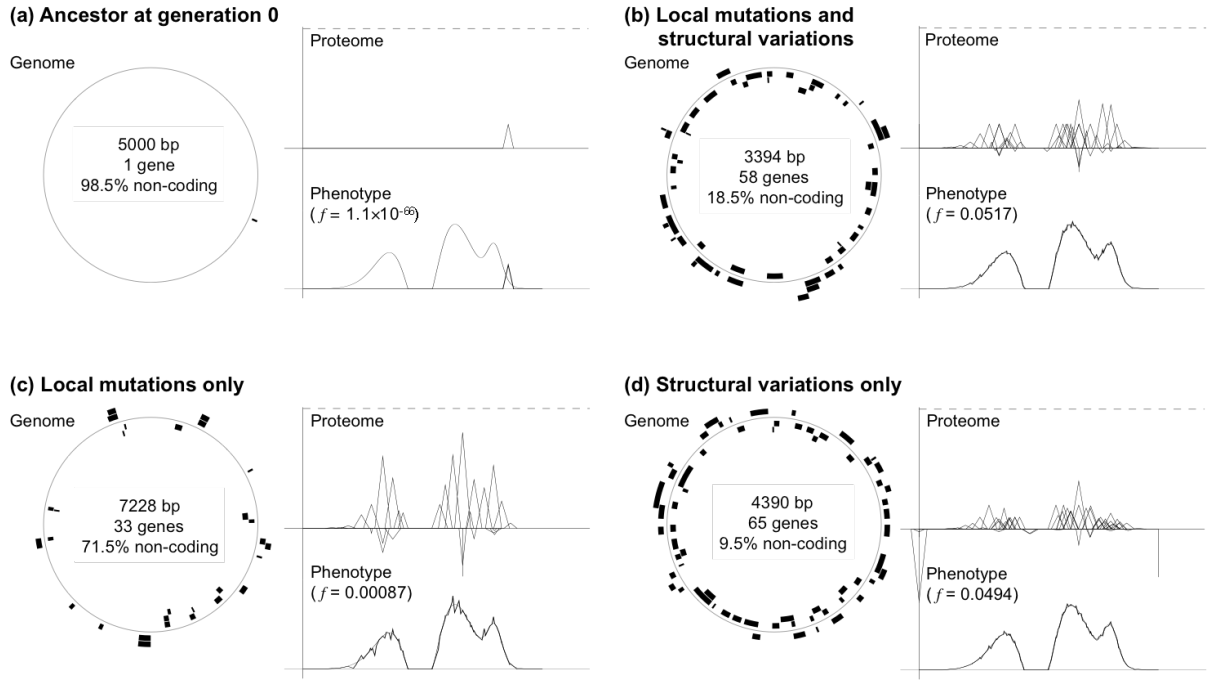


Figure II.10: Initial ancestor (a) and examples of evolved organisms in the CRLM (b), LM (c) and CR (d) conditions after 1,000,000 generations. The organism presented in (b) corresponds to the Wild-Type used for the second step of the experiments. For each organism, there is on the left a visualisation of its genes localised on the genome. On the right, the proteome shows all the single proteins, and the phenotype (black curve) is their sum. The grey curve plotted in addititon to the phenotype is the environmental target, a sum of 3 Gaussian lobes (2 positives and 1 negative – See Fig. II.11).

of mutations are modelled (depicted on Fig. II.8B). Three mutations are local (substitutions and small insertions or deletions), and four are chromosomal rearrangements, either balanced (with no change in genome size): translocations and inversions, or unbalanced: duplications and deletions.

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

Chromosomal rearrangement breakpoints are uniformly drawn on the chromosome, the number of breakpoints depending on the type of rearrangement (Fig. II.8B). Hence, chromosomal rearrangements can be of any size between 1 and the total genome size, allowing to investigate the effect of small structural variants that are indeed observed *in vivo* (Musumeci *et al.*, 2000; Audrézet *et al.*, 2004; Blakely *et al.*, 2006; Xue *et al.*, 2023).

The rates μ_t at which each type t of genetic mutation occur are defined as a per-base, per-replication probability. This means that the number of spontaneous events is linearly dependent on the length of the genome. However, its fixation probability depends on its

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

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

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

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

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

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

Figure II.11: Parameter file used for an example simulation (CRLM scenario).

phenotypic effect (for instance, a mutation affecting exclusively an untranscribed region is likely to be neutral). Hence, the distribution of fitness effects (DFE) of any kind of mutation is not predefined but depends on the intertwining of its effect on the sequence, and of the genome structure. For example, the fraction of coding sequences or the spatial distribution of the genes along the chromosome change the probability of a given mutation to alter the phenotype, and the fitness, an effect that is especially important for chromosomal rearrangements. Having an emergent DFE instead of a predefined one enables investigating the complex direct and indirect effects of chromosomal rearrangements on the evolutionary dynamics.

5.2 S2. Software usage

Aevol is based on running and analyzing forward-in-time simulations. More specifically, any experiment with Aevol is divided into four main steps. The first step consists in preparing a simulation with the `aevol_create` command. This reads the parameter file (Fig. II.11 and Table II.2) and creates a population of organisms at generation zero according to the specified values. `aevol_run` then simulates the evolution starting from the initial population or a from backed up population for a given number of generations.

Section	Parameter	Usual range	Description
Genotype to Phenotype to Fitness map	Maximal pleiotropy (W_{max}) `MAX_TRIANGLE_WIDTH`	0.01 – 1 (default: 0.033333)	Largest range of phenotypic values a single protein can impact. Regulates the mean pleiotropy degree and impacts the maximal phenotypic contribution of a single gene (Knibbe *et al.*, 2007b)
	Target function `ENV_ADD_GAUSSIAN`	Sum of 1 to 3 Gaussian functions	The target function is a linear combination of n Gaussian function G_i, each with a weight $\mathcal{H}_i$, a mean μ_i and a standard deviation σ_i: $Target = \sum_{i<n} \frac{\mathcal{H}_i}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2\right)$
	Length of a randomly generated genome `CHROMOSOME_INITIAL_LENGTH`	5000	Initial size of the chromosome when starting from a naive individual (see section 5.2.1)
Population / Selection	population size (N) `INIT_POP_SIZE`	256 – 4096	Census population size. Correlated with the effective population size N_e hence influencing the efficiency of the selection
	Grid size `WORLD_SIZE`	16x16 – 64x64	Shape of the grid. The grid shape influences the speed at which an individual can invade the population (Misevic *et al.*, 2015)
	Selection neighborhood `SELECTION_SCOPE`	Local 3x3 – Global	Type of selection (local or global), and, in the local case, shape of the window used for competition. Local selection slows down the spreading of favorable mutants and increases the effective population size (Zhang, 2003)
	Intensity of the selection (k) `SELECTION_SCHEME`	fitness proportionate 250 – 2500	The selection strength influences the genome size of individuals by increasing/decreasing the indirect selection for robustness (Batut *et al.*, 2013). Note that $k = 0$ suppresses the selection
Replication process	`POINT_MUTATION_RATE`	10^{-4} – 10^{-7}	Per base mutation rates for each kind of mutation. Changes in mutation rates have been shown to impact both the genome length and the genome structure (Knibbe *et al.*, 2007b; Rutten *et al.*, 2019)
	`SMALL_INSERTION_RATE`		
	`SMALL_DELETION_RATE`		
	`DUPLICATION_RATE`		
	`DELETION_RATE`		
	`INVERSION_RATE`		
	`TRANSLOCATION_RATE`		
	`MAX_INDEL_SIZE`	6	Maximal size of small insertion or deletion

Table II.2: Main parameters of the Aevol model.

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

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

5.2.1 Basic usage: Starting from a naive individual

Aevol allows to analyze the effect of various evolutionary parameters (typically mutation rates, mutational biases, population size) on genomes by comparing simulations under various scenarios (see Table II.2 for a list of the main testable parameters). Once the parameter values have been chosen, the basic usage of Aevol consist in testing the effect of these parameters directly, starting from "naive" individuals.

In this case, `aevol_create` generates random sequences of a predefined length (typically 5,000 bp) until it finds a genome that has a better fitness than that of a gene-less genome. This approach enables to study evolution when starting far from the fitness optimum. However, in that case the evolutionary dynamics is strongly dominated by genes recruitment, with massive genome size variation as shown *e.g.* by figure 3 (main text), hence putting the emphasis on a very specific evolutionary dynamics. If one wishes to study more subtle effects, this basic usage is not appropriate and one can turn to a more advanced experimental design based on "Wild-Typing".

5.2.2 Advanced usage: Wild-Typing

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

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

5.2.3 Post-evolution analyzes

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

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

To illustrate this, Fig II.12 depicts the distribution of selection coefficients for a large number of mutations on the WT (median CRLM run after the initial evolution) used in the paper.

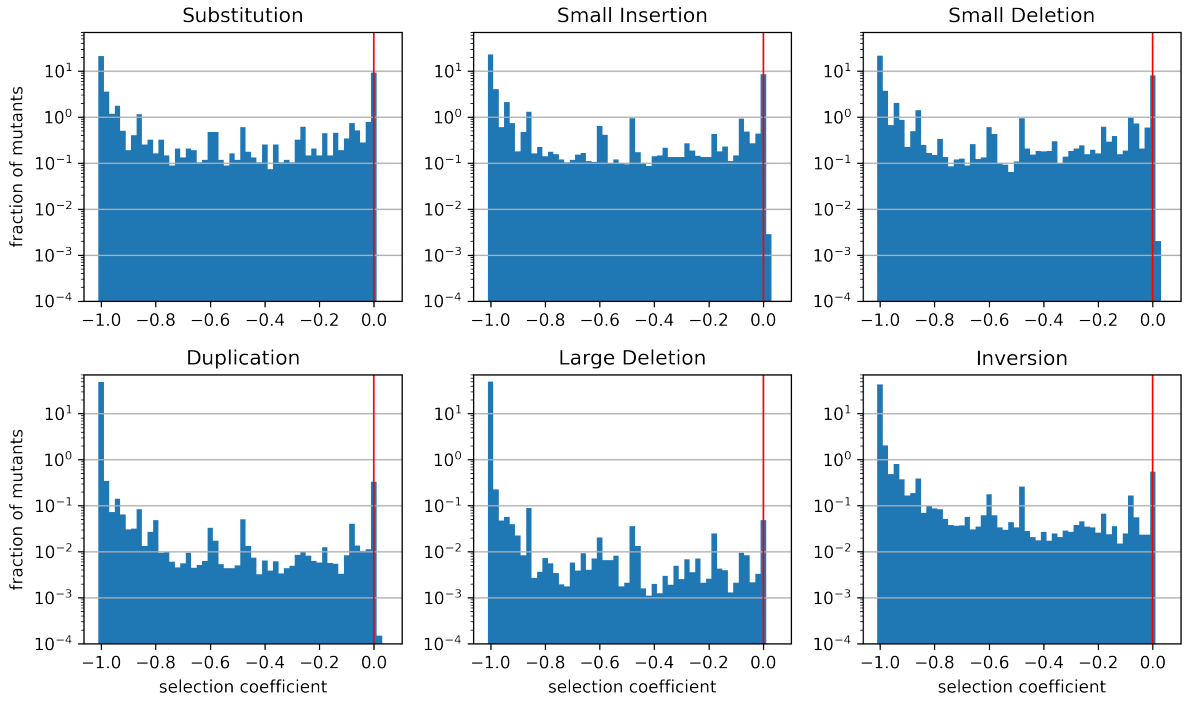


Figure II.12: **Distribution of selection coefficients** of the different mutation type, on the median individual of our CRLM experiment, after $1,000,000$ generation when starting from a naive individual. For each mutation type, $1,000,000$ mutants were generated, except for the substitution, which were exhaustively tested. The selection coefficient is computed as $s = \frac{f_{mutant}}{f_{parent}} - 1$. The vertical red line indicates neutrality.

Chapter III

Organization of the Manuscript

The previous chapter shows an example on how simulations with Aevol can shed a new light on the effect of genomic rearrangements. However, shedding light is not explaining and, as it is often the case in research, these results also raise many questions. The inclusion of chromosomal rearrangements in models of molecular evolution is not trivial. As we just saw, it leads to very different dynamics, both in terms of fitness and adaptation but also in terms of gene creation and accumulation of non-coding sequences. There is a dramatic change in complexity compared to models with solely point mutations, where the only adaptation can come from this small sized, well studied variations. Models such as Aevol introduce a twofold complexity cost: they include complex mutations such as sequence duplications and they create an interplay between these mutations and more classical operators like substitutions or Indels.

Following chapter II, many questions arise. How and why can rearrangements increase fitness more than local mutations? Which chromosomal rearrangements have the largest effect on fitness? How does the interplay between different types of mutations influence the evolutionary dynamics? What is the consequence of genome structure on the distribution of fitness effect of rearrangements, and inversely, what is the consequence of rearrangements on genome structure?

As stated in the general introduction, models need simplicity to be useful. As far as models of genome rearrangements are concerned, the standard Aevol seems already very complex, providing a wide diversity of results owing to the multiple evolutionary forces at stake. Simulations are highly stochastic due to drift, and also influenced by genome structure, founding effects and selection. While Aevol is a simplified model of biological evolution and a valuable tool to unravel the importance of chromosomal rearrangements, it is still, for the moment, too complex to allow understanding the impact of these rearrangements in details. In order to better understand the impact of chromosomal rearrangements in Aevol, we propose to develop "models of the model", following a "rematematization" approach (Varenne et Silberstein, 2013). In the second part of this thesis, we will present three articles (one published and two in preparation) in which we "rematematize" Aevol's results using different formalisms in order to decipher the dynamics of chromosomal rearrangements. Note that some of these articles may not even mention Aevol observations even though, in all cases, the scientific reasoning emerged from the study of Aevol simulations.

The chapter IV focuses on the characteristics of chromosomal rearrangements as an

exploration operator of the neighbourhood of accessible mutants. Indeed, as explained in the chapter II, the number and the variety of accessible mutants changes greatly with the type of mutations. For any given genome, the neighbourhood of accessible mutants defines the connectedness of the mutational network and influences greatly the ruggedness of the fitness landscape, hence the probability that a population get stuck on a local optimum. In order to study this property, we use a model of evolution with a simpler (compared to Aevol) genotype to fitness function: the NK-fitness landscape (Kauffman et Levin, 1987). While this model uses a fully coding genome and is classically used with substitutions only, we introduce an inversion operator and study its effect under a Strong Selection Weak Mutation (SSWM) regime, where population and robustness effects are absent. The population is represented by one individual and can only increase fitness until it reaches a local fitness maximum. In this article, we focus on how the general characteristics of the mutational network emerges from mutational operators and how they influence evolution. Indeed, the number of accessible mutants is not the sole determinant, as the final fitness value reached in the presence of inversions is different when the epistasis is local or global. This hints that the most important characteristics to determine the effect of a mutation over evolution are its "combinatorics" (the number of accessible mutants) and its DFE (Distribution of Fitness Effects). This article has been published in *PLoS Computational Biology* (Trujillo *et al.*, 2022).

While chapter IV is mainly about the final fitness reached. Another question raised in chapter II is how dynamics of adaptation changes in the presence of rearrangements. In chapter II, we observed that rearrangements sustain long term adaptation but the overall dynamics of rearrangements, substitutions and InDels has not been studied in details. In chapter V, we will focus on shorter and denser genomes (similar to viral genomes). Compared to the simulations presented in chapter II, this will reduce the "mutational noise" in the populations hence allowing to better distinguish the respective effects of rearrangements and local mutations. In this simpler context, we were able to identify periods of alternating mutational bursts and stasis in the simulations. We show that these bursts are similar to the ones we could observe after an environmental change, but that they are triggered stochastically without an exogenous event. By studying the key innovations that allowed some populations to escape a fitness peak, we show that duplications are the main cause of mutational bursts in our viral populations and show that the bursty dynamics is caused by the interaction of multiple mutation types. As a matter of fact, the number of possible mutations and the distribution of their fitness effect is very different between local mutations (such as substitutions) and rearrangements (such as duplications). Duplications have a larger impact on the genome, and so are significantly more often deleterious than substitution. Thus, for similar mutation rates, our simulations witness more frequent fixation of substitutions than duplications. However, from a given genome, there are much more possible duplications than substitutions in the mutational neighboorhood, and so the favorable substitutions are exhausted faster, leading to a stasis period (at least in short genomes). Eventually, a lucky duplication may appear, creating new opportunities and initiating a burst substitutions fixation. This shows that saltational evolution can emerge with very few assumptions, simply due to the interaction of different mutation types.

In chapters IV and V, we focused mainly on fitness effects, thus ignoring the dynamics of non-coding parts of the genome. However, as shown by the simulations of chapter II

and by many previous results with Aevol, non-coding DNA is regulated in the model (Knibbe *et al.*, 2007a; Batut *et al.*, 2013; Rutten *et al.*, 2019; Carde *et al.*, 2019). In chapter VI we will focus on the genome structure viewed as an succession of coding and non-coding sections on a circular genome. Chromosomal rearrangements obviously influence the genome structure but they have complex direct and undirect effects that make their contribution to genome structure difficult to predict. In chapter VI, we propose a mathematical model of chromosomal rearrangements. By computing the probability for a given type of mutation (being it a rearrangement or a local event) to modify a coding sequence, and under the approximation that any modification of a coding sequence is lethal, we show that, on the one hand, rearrangements contribute to eliminate non-coding DNA through a mechanism of selection for robustness, and that, on the other hand, they contribute to increasing non-coding DNA content through a bias in favor of duplication neutrality (with regards to deletions neutrality). We show that, qualitatively, both mechanisms lead to an equilibrium, hence to the maintenance and regulation of non-coding sequences in genomes. Quantitatively, the model predicts that in a given genome defined by the length and number of its coding sequences, the non-coding fraction is directly related to the product of the effective population size by the mutation rate. Although our mathematical model uses a lot of simplifying assumptions, it suggests that maintenance of a constant fraction of non-coding sequences in genomes is simply due to the presence of chromosomal rearrangements that result both in selection for robustness and in a mutational bias. It also suggests that the resulting amount of non-coding sequences is driven by the $Ne \times \mu$ factor, a very classical measure in population genetics.

Part B

Development

Chapter IV

Getting Higher on Rugged Landscapes: Inversion Mutations Open Access to Fitter Adaptive Peaks in NK Fitness Landscapes

Leonardo Trujillo, Paul Banse, Guillaume Beslon

Univ. de Lyon, INSA-Lyon, INRIA, CNRS, Univ. Claude Bernard Lyon 1, ECL, Univ. Lumière Lyon 2, LIRIS UMR5205, F-69622, Lyon, France

This article has been published in *PLOS Computational Biology* in October 31, 2022.

Abstract

Molecular evolution is often conceptualised as adaptive walks on rugged fitness landscapes, driven by mutations and constrained by incremental fitness selection. It is well known that epistasis shapes the ruggedness of the landscape's surface, outlining their topography (with high-fitness peaks separated by valleys of lower fitness genotypes). However, within the strong selection weak mutation (SSWM) limit, once an adaptive walk reaches a local peak, natural selection restricts passage through downstream paths and hampers any possibility of reaching higher fitness values. Here, in addition to the widely used point mutations, we introduce a minimal model of sequence inversions to simulate adaptive walks. We use the well known NK model to instantiate rugged landscapes. We show that adaptive walks can reach higher fitness values through inversion mutations, which, compared to point mutations, allows the evolutionary process to escape local fitness peaks. To elucidate the effects of this chromosomal rearrangement, we use a graph-theoretical representation of accessible mutants and show how new evolutionary paths are uncovered. The present model suggests a simple mechanistic rationale to analyse escapes from local fitness peaks in molecular evolution driven by (intragenic) structural inversions and reveals some consequences of the limits of point mutations for simulations of molecular evolution.

Author summary

Ninety years ago, Wright translated Darwin's core idea of survival of the fittest into rugged landscapes—a highly influential metaphor—with peaks representing high values of fitness separated by valleys of lower fitness. In this picture, once a population has reached a local peak, the adaptive dynamics may stall as further adaptation requires crossing a valley. At the DNA level, adaptation is often modelled as a space of genotypes that is explored through point mutations. Therefore, once a local peak is reached, any genotype fitter than that of the peak will be away from the neighbourhood of genotypes accessible through point mutations. Here we present a simple computational model for inversion mutations, one of the most frequent structural variations, and show that adaptive processes in rugged landscapes can escape from local peaks through intragenic inversion mutations. This new escape mechanism reveals the innovative role of inversions at the DNA level and provides a step towards more realistic models of adaptive dynamics, beyond the dominance of point mutations in theories of molecular evolution. Supplementary Material

1 Introduction

The fitness landscape is a very influential metaphor introduced by Wright (1932, Fig. 2) to describe evolution as explorations through a "field of possible genes combinations", where high values of fitness are represented as peaks separated by valleys of lower fitness. The topography of the fitness landscape has important evolutionary consequences, e.g. speciation via reproductive isolation (Gavrilets, 2004). Within this framework, the evolution of any population can be conceptualised as adaptive walks driven by successive mutations constrained by incremental or neutral fitness steps. Thus, in the absence of additional evolutionary forces such as drift or environmental variations, once a population reaches a local peak, natural selection hampers any further mutational paths that decrease fitness. However, there are empirical evidences showing that populations do not stop indefinitely at a local peak and can explore alternative trajectories on the landscape (Schrag *et al.*, 1997; Maisnier-Patin *et al.*, 2002; Salverda *et al.*, 2011; Cervera *et al.*, 2016; Bloom *et al.*, 2010), *ergo*, the following question arises: *How does evolution escape from a local peak to a fitter one?*

Since it has been formulated, considerable progress have been made on this question (Gillespie, 1983; Iwasa *et al.*, 2004; Weinreich et Chao, 2005; Jain et Krug, 2007; Serra et Haccou, 2007; Durrett et Schmidt, 2008; Weissman *et al.*, 2009; Altland *et al.*, 2011; de Lima Filho *et al.*, 2012; Grewal *et al.*, 2018; Belinky *et al.*, 2019; Guo *et al.*, 2019; Aguilar-Rodríguez *et al.*, 2018; Zheng *et al.*, 2019; Cano et Payne, 2020). However, conventional theoretical approaches still state that a genotype mutates into another through *point mutations* (e.g. single-nucleotide variations). If one takes a look at the molecular scale of DNA and the different mutation types, this may seem contradictory as it is well known that many other kinds of variation operators (including insertions, deletions, duplications, translocations and inversions) act on the genome. Hence, a fundamental aspect of this challenge is to understand—at the scale of molecular evolution—the roles played by these different mutation types. Indeed, there is a gap between the theoretical models, that account for a very limited set of mutations types—typically only point

mutations—and the reality of molecular evolution, where multiple variation operators act on the sequence.

As a contribution to bridge this gap, we present a minimal DNA-inspired mechanistic model for inversion mutations, and explore their relationship with the escape dynamics from local fitness peaks. Inversion mutations are one of the most frequent chromosomal rearrangements (Griffiths *et al.*, 2012, Ch. 17.2) with lengths covering a wide range of sizes. For example, in Long Term Evolution Experiments with *E. coli*, chromosomal rearrangements have been characterised by optical mapping (hence limiting the resolution to rearrangements larger than 5000 bp) (Raeside *et al.*, 2014). In this study, 75% of evolved populations showed inversion events—ranging in size from $\sim$ 164 Kb to $\sim$ 1.8 Mb (Raeside *et al.*, 2014) (for other examples of large inversions in different clades, see (Wellenreuther et Bernatchez, 2018, Table 1)). With the development of novel sequencing technologies (Wolfe et Li, 2003; Brockhurst *et al.*, 2011; Wellenreuther *et al.*, 2019), it has been possible to identify intragenic (submicroscopic) DNA inversions, for example, an inversion of seven nucleotides in mitochondrial DNA, resulting in the alteration of three amino acids and associated to an unusual mitochondrial disorder (Musumeci *et al.*, 2000, Fig. 1). Intragenic inversions have also been suggested to be an important mechanism implied in the evolution of eukaryotic cells (Korneev et O'Shea, 2002). Although chromosomal inversions are ubiquitous in many evolutionary processes (Raeside *et al.*, 2014; Merrikh et Merrikh, 2018; Wellenreuther et Bernatchez, 2018; Musumeci *et al.*, 2000; Korneev et O'Shea, 2002; Ranz *et al.*, 2007; Hoffmann et Rieseberg, 2008; Kirkpatrick, 2010; Faria *et al.*, 2019; Huang et Rieseberg, 2020; Wolfe et Li, 2003; Brockhurst *et al.*, 2011; Mérot *et al.*, 2020; Berdan *et al.*, 2021b; Griffiths *et al.*, 2012), very little is known about their theoretical description and computational simulation at the sequence level, as models generally focus on very large inversions (typically larger than a single gene), hence on their effect on synteny (or the deleterious effects at breakpoints), but neglect the possibility that small inversions occur inside coding sequences (Fertin *et al.*, 2009).

Here, we simulate a representation of molecular evolution of digital organisms (replicators), each of which contains a single piece of DNA. We engineer a computational method to cartoon the double-stranded structure of DNA, and simulate inversion-like mutations consisting of a permutation of a segment of the complementary strand, which is then exchanged with the main strand segment (see Methods, schema IV.6). For the sake of simplicity, we consider digital genotypes made up of binary nucleotides (i.e. a binary alphabet $\{0, 1\}$ instead of the four-nucleotides alphabet $\{\mathtt{A}, \mathtt{T}, \mathtt{C}, \mathtt{G}\}$). The sequences are arranged in circular strings with constant number of base-pairs. In an abstract sense, the model mimics the molecular evolution of some viruses (Solé et Elena, 2018) and (animal) mitochondrial DNA (Kolesnikov et Gerasimov, 2012) with compact genomes and closed double-stranded DNA circles (Solé et Elena, 2018; Tisza *et al.*, 2020; DiMauro, 2001). It is very important to emphasise that our computational model simulates intragenic-like mutations (Musumeci *et al.*, 2000; Korneev et O'Shea, 2002; Bank *et al.*, 2016). We are modelling asexual replication, therefore recombination is not considered. To build rugged fitness landscapes, we adopt the well known Kauffman NK model, where N denotes the length of the genome and K parameterizes the epistatic coupling between nucleotides (Kauffman et Levin, 1987; Kauffman et Weinberger, 1989; Weinberger, 1991; Kauffman, 1993). We do not include environmental changes, so the landscape remains constant through the simulations. Finally, it is worth mentioning that all our simu-

lations were conducted in the evolutionary regime of strong selection weak mutations (SSWM) (Gillespie, 1991, Ch. 5).

2 Results

2.1 Who is next to whom: the mutational network

We first study how inversion mutations can increase the number of accessible mutants. For this we translate the canonical notion of neighbour genotypes (see Methods, Eq. IV.1) into a graph theory approach, and analyse the simplest and most familiar geometric object in molecular evolutionary theory: the discrete space of binary sequences (unless explicitly stated we will henceforth consider only binary alphabets). Thinking topologically, all the sequences $\mathbf{x}$ with N binary-nucleotides $x_i \in \{0,1\}, \forall i = 1,\ldots,N$ and $N \in \mathbb{N}_{\geq 2}$, define the set $\mathcal{X} \in \{0,1\}^N$ of 2^N possible genotype combinations. A canonical measure to characterise the topology of the set $\mathcal{X}$ is the Hamming distance

$$d_H(\mathbf{x},\mathbf{x}') := \sum_{i=1}^{N}(x_i - x_i')^2.$$

A convenient way to organise such a set is by graphs connecting two sequences $\mathbf{x}$ and $\mathbf{x}'$ that differ by one point mutation (i.e. $d_H(\mathbf{x},\mathbf{x}') = 1$). This is the so-called Hamming graph $\mathcal{H}(N,2)$—a special case of the hypercube graph $\mathcal{Q}_N$ (the well-known graph representation of the genotype space). On the other hand, the Hamming distance for inversion mutations forms a set of integers satisfying

$$0 \leq d_H(\mathbf{x},\mathbf{x}') \leq N,$$

meaning that, contrary to point mutations, the Hamming distance of inversion mutations range from zero to N (see Methods). Note that if the inversion spans the entire chromosome, then $d_H(\mathbf{x},\mathbf{x}') = N$, all loci have changed, but it also implies that 5'— 3' becomes 3'—5' and vice versa and nothing has changed biologically.

We propose that, for a sequence $\mathbf{x} \in \mathcal{X}$, the mutation operation (i.e. the mechanistic representation of point or inversion mutations) build the set $\mathcal{N}_\nu(\mathbf{x})$ of accessible mutants $\mathbf{x}' \in \mathcal{N}_\nu(\mathbf{x}), \forall \mathbf{x}' \neq \mathbf{x}, \implies d_H(\mathbf{x},\mathbf{x}') \neq 0$ (the subindex ν denotes the type of mutation: P for point mutations and I for inversion mutations). Therefore, the number of neighbouring mutants can be reformulated as

$$D_\nu(\mathbf{x}) = |\mathcal{N}_\nu(\mathbf{x})|.$$

Inversion's combinatorics is not trivial since it involves the permutation of a subsequence and its flips between each strand (see Methods, schemata IV.2 and IV.6). Nevertheless, from the algorithmic point of view, for a given genotype $\mathbf{x}$ the mutational operations can be used to enumerate all the accessible mutants (see Methods, Algorithm 1: `Mutate`). Also, the combinatorics can be represented as a directed multigraph of mutations $\mathrm{m}(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}$ (see the mathematical definition in (Bollobás, 2013, p.8)), i.e. the ordered triple

$$\mathrm{m} = (V(\mathrm{m}), E(\mathrm{m}), I_{\mathrm{m}}),$$

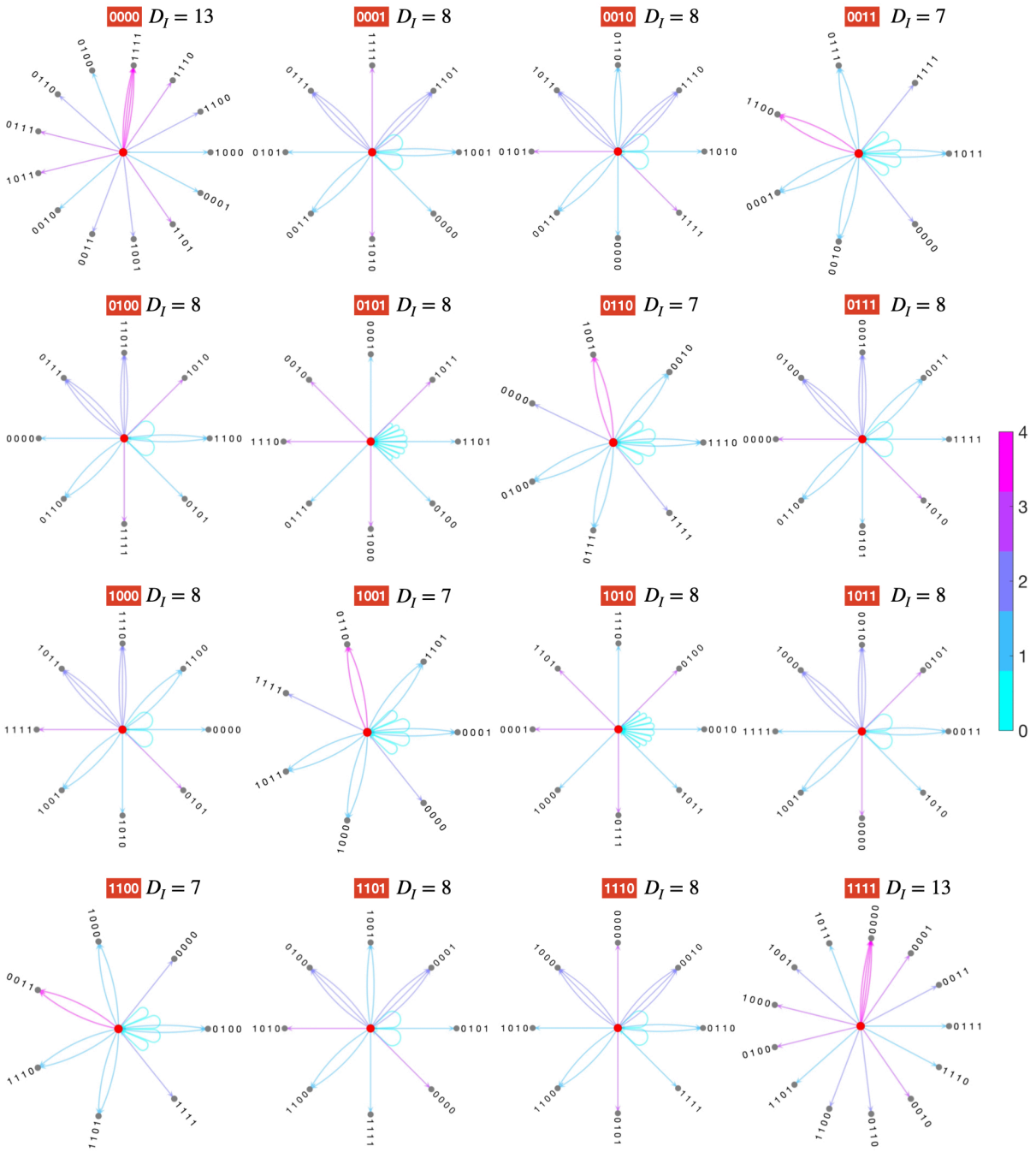


Figure IV.1: **Atlas of accessible-mutants.** Example of the total enumeration of inversion mutations, represented as *graphs of accessible mutants*, for each one of the 2^4 genotypes (central red nodes) with size $N = 4$. Edges colour quantifies the Hamming distance $d_H(\mathbf{x}, \mathbf{x}')$, between the central nodes $\mathbf{x}$ (wild-types) and their mutants $\mathbf{x}'$. Each wild type is labeled in red and its number of accessible mutants D_I is also displayed. Let us remark that this enumeration also depends on the fact that in our model the sequences are circular (i.e. periodic boundary conditions: $x_{N+i} = x_i, \forall i \in \{1, \ldots, N\}$).

Genome size N	Number of accessible-mutants via inversions D_I	Number of accessible-mutants via point mutations D_P
2	$2_2, 3_2$	2_4
3	$5_6, 7_2$	3_8
4	$7_4, 8_{10}, 13_2$ (See Fig. IV.1)	4_{16}
5	$13_{20}, 17_{10}, 21_2$	5_{32}
6	$16_{30}, 17_{18}, 18_2, 22_{12}, 31_2$	6_{64}
7	$25_{70}, 29_{42}, 37_{14}, 43_2$	7_{128}
8	$28_{16}, 29_{52}, 30_{112}, 32_2, 34_8, 36_{48}, 46_{16}, 57_2$	8_{256}
9	$39_6, 40_{18}, 41_{234}, 45_{162}, 52_{36}, 53_{36}, 64_{18}, 73_2$	9_{512}
10	$45_{100}, 46_{150}, 47_{420}, 50_2, 52_{40}, 53_{200}, 62_{40}, 63_{50}, 77_{20}, 91_2$	10_{1024}

Table IV.1: **Enumeration of accessible-mutants.** Number of neighboring mutants accessible via inversion mutations D_I and via point mutations D_P for all genomes of size $N \in [\![2, 10]\!]$ (subscripts numbers denote the number of occurrence of each value of D_I and D_P).

where $V(\mathrm{m}) = \mathbf{x} \cup \mathcal{N}_\nu(\mathbf{x})$ is the set of vertices (formed by a given genotype $\mathbf{x}$ and its mutated genotypes $\{\mathbf{x}'\} \in \mathcal{X}$), $E(\mathrm{m})$ is the set of directed edges (from genotype $\mathbf{x}$ to a mutated genotype $\mathbf{x}'$) and $I_\mathrm{m} : \mathcal{X} \longrightarrow \mathcal{X}$ is an incidence relation that associates to each element of $E(\mathrm{m})$ an ordered pair of $V(\mathrm{m})$. In fact, the incidence relation I_m corresponds to a mutation operation (see Methods, Eqs. (IV.3), (IV.4), and (IV.5)). As an example, in Fig. IV.1 we display the atlas of accessible-mutants for $N = 4$, constructed by calculating all inversion mutations for each one of the 2^4 (wild-type) sequences (central red dots). In this example (see also Table IV.1), it is verified that the number of accessible-mutants for inversion mutations $D_I(\mathbf{x}), \forall \mathbf{x} \in \{0, 1\}^4$ is in the set $\{7, 8, 13\}$ (unlike the case for point mutations, which must be a singleton, i.e. the single-value set: $D_P(\mathbf{x}) \in \{4\}, \forall \mathbf{x} \in \{0, 1\}^4$). We can also verify that $\min(D_I(\mathbf{x})) \geq \max(D_P(\mathbf{x}))$. In Table IV.1 we show the enumeration of accessible-mutants for inversions and point mutations for genome sizes ranging from $N = 2$ to 10. From Fig. IV.1 and Table IV.1, we can see that the combinatorics of the inversion mutations is not trivial. We can verify that the maximum number of accessible mutants is equal to $N^2 - N + 1$, which corresponds to the trivial cases of genotypes $\mathbf{x}$ with $x_i = 0, \forall i \in \{0, \dots N\}$ and $x_i = 1, \forall i \in \{0, \dots N\}$. Note that for a circular sequence of size N, the total number of inversion mutations is N^2, while for point mutations this number is equal to N. However, the number of mutants accessible by inversions is lower than the total number of inversions mutations ($D_I < N^2$). This is due to "degenerate" inversion mutations: several inversions—occurring between different loci and/or for different interval sizes—may mutate the initial sequence to the same accessible mutant (see the multiple edges in Fig. IV.1). In Fig. IV.1, we can also verify that there are loops (an "edge" joining a vertex to itself), that is, "invariant inversions" that preserves the nucleotide sequence after the inversion operation (i.e. $d_H(\mathbf{x}, \mathbf{x}') = 0$). It can easily be shown that the fraction of invariant inversions converges to $1/N$ (see S1 Text, section 1). A very important consequence of inversions is that mutated sequences can differ with the wild type by more than one nucleotide, i.e. $d_H(\mathbf{x}, \mathbf{x}') > 1$ (edges colours in Fig. IV.1 denote the values of the Hamming distance). This result allows us to gain a first insight

of how inversions can promote the escape from local fitness peaks: they can "connect", in a single mutational event, genotypes that are at two or more point-mutational steps away. It is pertinent to remark that the combinatorics of inversions for alphabets with size $|\mathcal{A}_\mathcal{L}| = 2n, \forall n \geq 2$ would imply even more connections.

It should be noted that the inversion combinatorics would be slightly different for linear sequences. In that case, the number of possible inversions is $N(N+1)/2$.

Up to this point, we have shown how inversion mutations can actually broaden the horizon of evolutionary exploration in the genotype space.

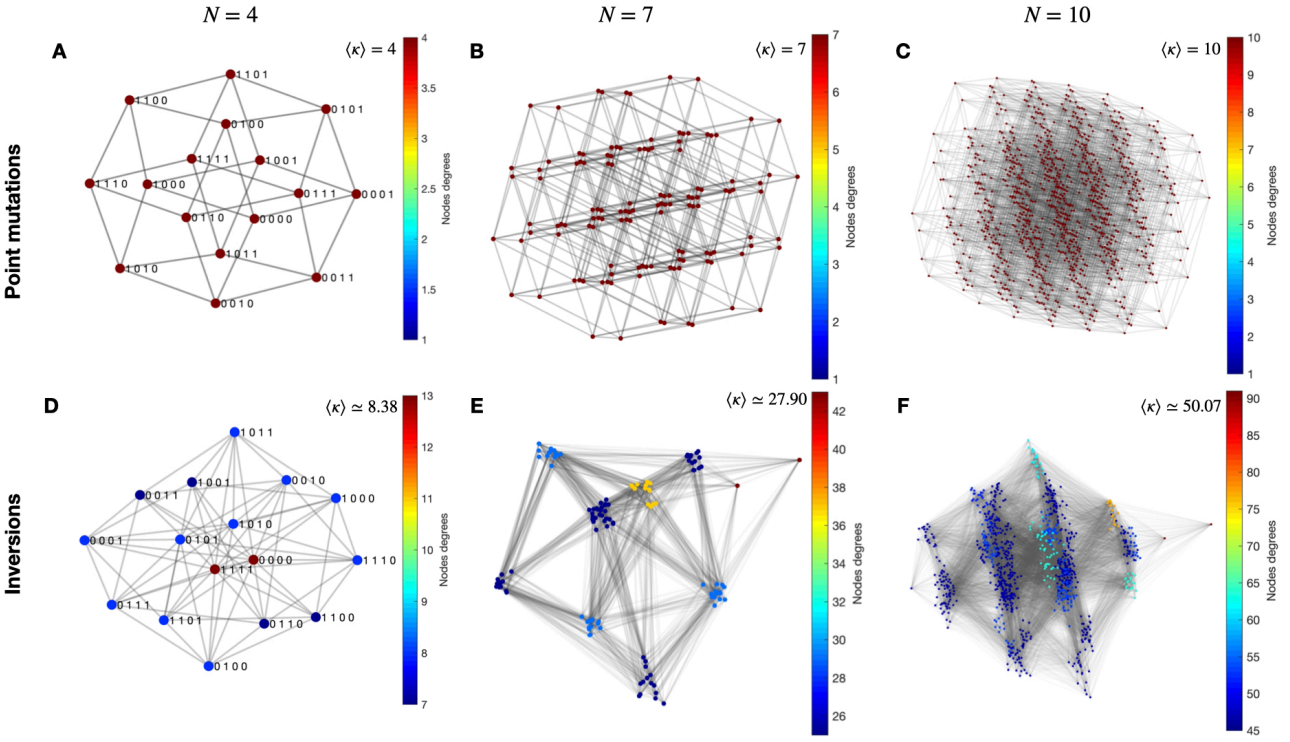


Figure IV.2: **Examples of mutational networks.** Representative examples for $N = 4, 7$ and 10. Colour indicates the node's degree κ. The reported values correspond to average node degrees $\langle\kappa\rangle$. The upper graphs show the point mutation case, verifying that the mutational networks are Hamming graphs and therefore isomorphic to their genotype spaces: (A) $\mathcal{H}(4,2)$; (B) $\mathcal{H}(7,2)$ and (C) $\mathcal{H}(10,2)$. The lower graphs (D), (E) and (F) correspond to the inversion mutations cases, where we can note that these mutational networks are not isomorphic to their genotype spaces.

Inversions rewire the adjacency of the genotype space Now, for each directed graph of mutations $\mathrm{m}(\mathbf{x})$, we can associate a graph $\mathrm{M}(\mathbf{x})$ on the same set of vertices $V(\mathrm{m})$. Corresponding to each directed edge of m, there is an edge of M with the same ends (loops being excluded). In this sense, the graph $\mathrm{M}(\mathbf{x})$ is the underlying (simple) graph of the directed graph $\mathrm{m}(\mathbf{x})$. That is, the graphs without self- and directed-edges so that when several edges (mutations) connect the genotype $\mathbf{x}$ with the same accessible-mutant $\mathbf{x}'$, only one (undirected) edge is kept. Thus, every directed graph $\mathrm{m}(\mathbf{x})$ defines a unique, up

to isomorphism, reduced graph M($\mathbf{x}$) (see the mathematical definition in (Bollobás, 2013, p.3)). Now it is natural to do the union of each graph M($\mathbf{x}$), to describe how genotypes can be reached from somewhere in the genotype space in one mutation operation. We call this object the mutational network. It is defined as:

$$\mathcal{M}(\mathcal{X}) := \bigcup_{\mathbf{x}\in\mathcal{X}} \mathrm{M}(\mathbf{x}).$$

For point mutations the mutational network is the Hamming graph $\mathcal{H}(N, 2)$ (Hwang *et al.*, 2018, p. 230) (as we can see in Figs. IV.2 A, B and C for $\mathcal{H}(4, 2)$, $\mathcal{H}(7, 2)$ and $\mathcal{H}(10, 2)$ respectively), which is isomorphic to the canonical genotype space $\mathcal{H}(N, 2) = \mathcal{Q}_N$. The notion of isomorphism means that the mutational network for point mutations preserves the adjacency of the edge structure of the genotype space. Historically, the canonical graph of the genotype space overshadowed the richness of the (full) mutation graph, since theoretically only point mutations are usually considered as generators of mutational networks. For inversions, the mutational network does not necessarily inherit the (local) topology of the genotype space. For example, Figs. IV.2 D, E and F outline the structure of the mutational networks for inversion mutations for $N = 4, 7$ and 10. From the point of view of graph theory, inversion mutations "rewire" the adjacency of the genotype space, i.e. they link genomes such that $d_H(\mathbf{x}, \mathbf{x}') \geq 1$. Also, in graph terms, the total number of accessible mutants per genotype corresponds to the node's degree $\kappa_{\mathbf{x}}$ (defined as the number of edges in the graph incident on $\mathbf{x}$ (Bollobás, 2013, p. 3)). Therefore, $\kappa_{\mathbf{x}} = D_\nu(\mathbf{x})$. On the other hand, the average node degree quantifies accessible-mutants (nodes) interconnections:

$$\langle\kappa\rangle := \frac{1}{2^N}\sum_{\mathbf{x}} \kappa_{\mathbf{x}}.$$

It can be verified that for point mutations $\langle\kappa\rangle = N$, while for inversions mutations $\langle\kappa\rangle > N$ (for $N \geq 2$) and therefore the genotypes are "more connected" to each other. Paraphrasing in terms of evolutionary biology, they are "more mutable". In this sense, $\langle\kappa\rangle$ defines a mean mutability, which quantifies the ability to reach a different genome when the sequence undergoes a mutation. This property also holds for linear chromosomes, although as mentioned above, the average of node degrees is smaller since the number of possible inversions is lower than for circular chromosomes.

2.2 Inversions can reveal new evolutionary paths

Even though the nature of the genotype-to-fitness function is still largely unknown, an easy way to introduce it into computational models is by assuming that for genotypes $\mathbf{x} \in \mathcal{X}$ there exists a map from the set $\mathcal{X}$ to the real numbers $f : \mathcal{X} \to \mathbb{R}$. In the graph-based representation, each node (genotype) then possesses a fitness value $f(\mathbf{x})$. This fitness landscape graph F is isomorphic to the hypercube graph $\mathcal{Q}_N$ (i.e. the genotype space) and therefore can also be represented as Hamming graphs, providing a fitness value per node. So, the fitness landscape graph is univocally defined as:

$$\mathrm{F}(\mathcal{X}, \mathcal{N}_P, f) := (\mathcal{H}(N, 2), f).$$

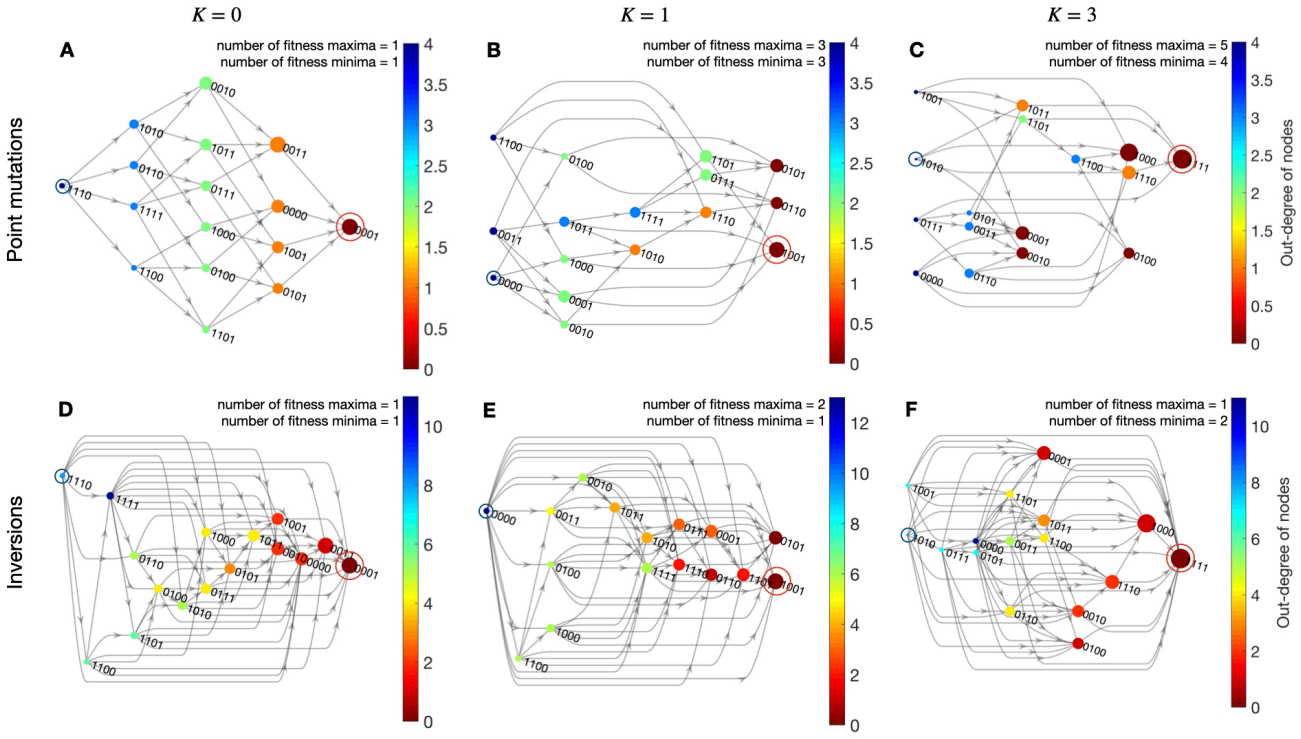


Figure IV.3: **NK fitness networks for epistatic interactions with random neighbouring.** Representative instances of the NK model for $N = 4$ and their fitness networks in layered representation. The layers are constructed such that each node is assigned to the first possible layer, with the constraint that all its predecessors must be in earlier layers. The colors of the nodes correspond to the values of the out-degrees, i.e. the number of edges going out of a node (note that color scales differ in range between panels). Therefore, nodes with node out-degrees equal to zero correspond to local fitness maxima (sink nodes). The landscapes' ruggedness are: single peaks $K = 0$, intermediate ruggedness $K = 1$ and full rugged case $K = 3$. Node sizes are scaled with fitness values (the best fitness, the largest, and vice versa). Global maximum of fitness are encircled in red. While the global minimum in blue. The total number of fitness maxima and minima are also reported. See Fig. IV.7 for epistatic interactions with adjacent neighbouring.

Likewise, as the mutational network we can also define the fitness network $\mathcal{F}$, but in this case the edges are directed from genotypes with lower fitness to genotypes with higher fitness:

$$\mathcal{F}(\mathcal{X}, \mathcal{N}_\nu, f) := (\mathcal{X},\ \{(\mathbf{x}, \mathbf{x}') \in \mathcal{X}^2 \mid \mathbf{x}' \in \mathcal{N}_\nu(\mathbf{x}) \text{ and } f(\mathbf{x}') > f(\mathbf{x})\}, f),$$

which also depends on the neighbouring $\mathcal{N}_\nu$, with ν denoting the type of mutation (P and I for point and inversion mutations respectively). Therefore, the fitness network is the anisotropic version of the mutational network, the direction of the evolutionary paths being fitness-dependent. Precisely what mutational and fitness networks reveal to us is the ensemble of possible evolutionary paths. But in this case, fitness networks are diagrams showing the paths upward and their "altitudes" (fitness values).

To illustrate the definition of fitness network used in this work, we need to build fitness landscapes instances. For that, we use the well-known NK-model (Kauffman, 1993; Kauffman et Levin, 1987; Kauffman et Weinberger, 1989), recalling that for a genome of size N, the parameter $K \in \{0, \ldots, N-1\}$ corresponds to the epistatic coupling between loci and thus tunes the ruggedness of the landscape (for a brief introduction see Methods). Here, we use two neighbourhood coupling models between loci: i) the adjacent model in which the K loci are those closest to a focal locus x_i; ii) the random neighbourhood model, where the K loci are chosen randomly among the $N-1$ loci other than x_i (an illustration of epistatic interactions is sketched in Fig. IV.6 in Methods). In Fig. IV.3 we show representative examples of fitness networks engineered for $N = 4$ with K epistatic random neighbours (see IV.7 for the fitness networks with K epistatic adjacent neighbours). The landscapes range from single peaked $K = 0$ (no epistatic interaction) to full rugged landscapes $K = 3$ (highly connected epistatic interactions). Global fitness maxima and minima are highlighted by encircled nodes. In Figs. IV.3 (A)-(C) we show an instance of fitness networks for genotypes connected through pathways with point mutation steps. They have the same topology as the genotype space $\mathcal{Q}_{N=4}$ and, therefore, are isomorphic to the graph representation of the fitness landscape $\mathcal{F} \cong \mathrm{F} = (\mathcal{H}(4,2), f)$. When $K = 0$ all the trajectories arrive to the single global maximum of fitness. When we construct the fitness network with inversion mutations, we can verify in Figs. IV.3 (D)-(F) that there are more paths between all the genotypes, and so it is easier for an evolutionary process to explore more domains of the fitness landscape compared to point mutations. Many of these paths connect genotypes such that $d_H(\mathbf{x}, \mathbf{x}') > 1$, and therefore, are like jumps between distant domains of the landscape. We can also verify that, for a given fitness function f, a node that is a local optimum on the fitness graph $\mathrm{F} = (\mathcal{H}(N,2), f)$, is not necessarily a local peak for inversion mutations in the fitness network. In most of the cases, the fitness landscape is "smoothed out" by inversion mutations, since the notion of local peak fades in the fitness network. However, the local peaks are not always smoothed out, as we can see in Fig. IV.3 (E) for $K = 1$, where genotype 0101 remains as a local peak. This is because 0101 cannot be mutated to genotype 1001 by any inversion mutation (see also the combinatorics of accessible mutants for 0101 in Fig. IV.1). Note that for inversion mutations, it is verified that in some cases the global maximum can be reached from the global minimum in a single evolutionary step with $d_H(\mathbf{x}, \mathbf{x}') \neq 1$, e.g. Figs. IV.3 (E) with $d_H = 2$.

Finally, contrary to point mutations, inversions are not commutative: in many cases, two overlapping inversions applied to a same initial sequence in direct or reversed order lead to different final sequences. This can easily be shown on an example:

$$\begin{array}{c} 0111100 \xrightarrow{inv(2,3)} 0100100 \xrightarrow{inv(2,4)} 0101100 \\ 1000011 \quad\circlearrowleft\quad 1011011 \quad\circlearrowleft\quad 1010011 \end{array}$$

$$\begin{array}{c} 0111100 \xrightarrow{inv(2,4)} 0100100 \xrightarrow{inv(2,3)} 0111100 \\ 1000011 \quad\circlearrowleft\quad 1011011 \quad\circlearrowleft\quad 1000011 \end{array}$$

where, starting from the same sequence, the two inversions ($inv(2,3)$ and $inv(2,4)$) give different outcomes depending on their order. Note that, given this property, the classical definition of mutational epistasis does not hold for inversion mutations.

2.3 Getting higher on rugged landscapes

Up to now, we have shown results on the combinatorial (topological) differences between point and inversion mutations. Inversions cannot be mapped to the classical "fitness landscape" metaphor—being better represented through mutational networks and their juxtaposition with fitness landscapes through fitness networks. This is because, for inversion mutations, there are shortcut routes connecting distant sequences (differing by more than one base) in the genotype space and consequently in the fitness landscape. Therefore, this can be interpreted as "escape routes" from local peaks. We want to verify if as a consequence of these escape routes, an evolutionary process will be able to reach higher peaks of fitness. For that, we performed computer simulations in the SSWM setting, where adaptation occurs by sequential fixing of novel beneficial mutations (see Adaptive walks in Methods).

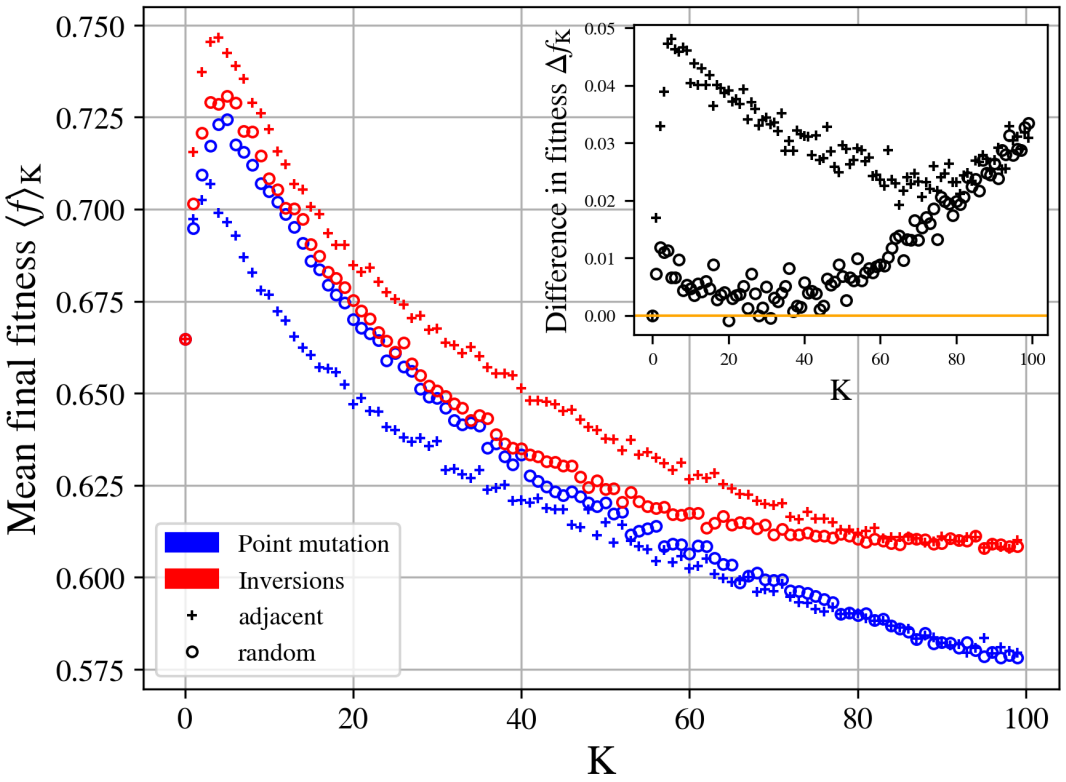


Figure IV.4: **The average value of local fitness maxima suggests the escape from local fitness peaks by inversion mutations.** Changes in mean final fitness for different epistatic parameter K for inversion (red) and point mutations (blue), averaged for 100 instances of adaptive walks simulations in NK landscapes. The circle (respectively cross) markers corresponds to random (respectively adjacent) neighbouring epistatic interactions. Inset: Difference Δf_K between the mean of local fitness maxima of inversions and point mutations, for random (circles) and adjacent (cross) neighbouring epistatic interactions.

We focus our study on a series of n-repetitions of adaptive walks, where the evolution-

ary process is driven by (random) mutational steps. For a given set of independent initial random genomes with size $N = 100$, $\{\mathbf{x}_0\} \in \{0,1\}^{100}$, we create two pools of $n = 100$ simulations for point mutations and inversions respectively. As before, we use the NK model to engineer rugged fitness landscapes. In each round, the landscape is the same for simulations with point mutations and inversions, respectively. For independent explorations over (sub)domains of the landscape, we monitor the time-evolution of the fitness values until a fitness optimum is reached. This is when it is verified, in the simulation, that a genotype $\mathbf{x}_\nu^{\text{loc}} \in \mathcal{X}$ satisfies

$$f(\mathbf{x}') < f(\mathbf{x}_\nu^{\text{loc}}),\ \forall \mathbf{x}' \in \mathcal{N}_\nu(\mathbf{x}_\nu^{\text{loc}}).$$

Subindex ν denotes the type of mutation (P for point mutations and I for inversion mutations). Then, we calculate the mean fitness value per K as:

$$\langle f_\nu \rangle_K := \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{x}_\nu^{\text{loc}}(i)) \Bigg|_K ,$$

where the notation $|_K$ means that the average is calculated for a fixed value of $K \in \{0, \ldots, N-1\}$.

In Fig. IV.4 we show the behaviour of the average fitness $\langle f_\nu \rangle_{\text{K}}$, calculated from $n = 100$ instances of adaptive walks simulations in NK landscapes, with $N = 100$ and values of K ranging from 0 to 99. The simulations correspond to the case of epistatic interactions with K closest adjacent loci ('+' marker), and with K randomly chosen loci ('o' marker). The markers of the simulations with point mutations and inversions are coloured with blue and red respectively.

For the simplest case $K = 0$, with no epistatic interaction between neighbouring loci, we can verify in Fig. IV.4, that the average fitness for point mutations and inversions are equal $\langle f_P \rangle_{K=0} = \langle f_I \rangle_{K=0} \simeq 0.667$. In this case, the landscape is smooth with only a single peak. Hence, mutations that increase fitness are not hard to find and $\langle f_\nu \rangle_{K=0}$ is independent of mutation types. This result also agrees with the Kauffman's (analytical) result $\langle f \rangle_{K=0} = \frac{2}{3}$, which was calculated using order statistics arguments (Kauffman, 1993, p. 55). Then, for $K > 0$, the average fitness $\langle f_\nu \rangle_K$ increases with K until reaching a maximum value of fitness. In relation to this maximum, in Fig. IV.4 we can identify the following four cases: i) point mutations with adjacent epistatic interactions, $\langle f_P \rangle_K = 0.707$ for $K = 2$; ii) point mutations with random epistatic interactions, $\langle f_P \rangle_K = 0.722$ for $K = 5$; iii) inversions with adjacent epistatic interactions, $\langle f_I \rangle_K = 0.747$ for $K = 4$; and iv) inversions with random epistatic interactions $\langle f_I \rangle_K = 0.732$ for $K = 4$. After these maximum, the fitness values decrease as K increases. This trend is also consistent with the seminal simulations carried out by Kauffman (1993); Kauffman et Weinberger (1989). For example, when $K \to N - 1$, the mean fitness converge to the same value, regardless of the type of epistatic interaction neighbourhood. For point mutations with random and adjacent epistatic interactions, we obtained $\langle f_P \rangle_{K=96} \simeq 0.580$ and this agrees very well with the Kauffman's numerical outcomes for $K = 96$, c.f. (Kauffman, 1993, Tables 2.1 and 2.2). It is worth mentioning that these trends—lower fitness being associated with increasing epistatic interactions—correspond to the well-known "complexity catastrophe" described by Kauffman (Kauffman, 1993, p. 52) (see also Refs.(Solow *et al.*, 1999a,b)).

These numerical outcomes confirm that our numerical set-up reproduces, with point mutations, what is known about $\langle f_P \rangle_K$ vs K in the NK model (Kauffman, 1993; Kauffman et Weinberger, 1989). Now, what's new is that for inversions, the average fitness values are higher than those for point mutations. Indeed, for $K \to N-1$, the average fitness trend is very different from that of point mutations. For example, the complexity catastrophe estimates that as K increases, the expected fitness of the local maximum (for point mutations) decreases toward 1/2 (Kauffman, 1993), which is indeed verified here. But for inversions, the evolutionary process reaches higher expected fitness values $\langle f_I \rangle_{\mathrm{K}=99} = 0.610 > \langle f_P \rangle_{\mathrm{K}=99} = 0.579$, for both random and adjacent epistatic interactions. To generalize these results, we reproduced this experiment with other, more restrictive, definition of inversion mutations. More specifically, we tested inversions on linear chromosomes (with boundary conditions) and circular inversions with upper size limit $s \leq N$, ranging from 1%—a single locus (i.e. inversions are like point mutations)—up to 100% of chromosome size. Simulations with linear chromosomes show no significant difference from our reference circular model (see supplementary S1 Text, section 2 for detailed results). Simulation with an upper size limit show that the final fitness values increases as the size limit increases. However, the gain is maximal for small s values (typically up to $s = 16$) showing that small and mid-sized inversions are sufficient to reach high fitness peaks.

Figure IV.4 also show that, in average and for almost all K, the adaptive walks reach higher fitness peaks through inversions than through point mutations. To better visualise this statement, in the inset of Fig. IV.4 we plot the following difference:

$$\Delta f_K := \langle f_I \rangle_K - \langle f_P \rangle_K,$$

for the two neighbourhoods. To specify which type of epistatic interaction we are referring to, in what follows, we will use the notation $(\Delta f_K)_{\mathrm{rnd}}$ for the case in which the fitness differences correspond to simulations with random neighbourhood, and $(\Delta f_K)_{\mathrm{adj}}$ for the adjacent ones (respectively the markers 'o' and '+' in Fig. IV.4). In the absence of epistatic interactions, i.e. $K = 0$, we can note that $(\Delta f_K)_{\mathrm{rnd}} = (\Delta f_K)_{\mathrm{adj}} = 0$. Then, in the presence of epistatic interactions, $(\Delta f_K)_{\mathrm{rnd}}$ is monotonically increasing between $0 < K \leq 2$. Then for $2 < \mathrm{K} \leq 31$, $(\Delta f_K)_{\mathrm{rnd}}$ is monotonically decreasing, and for $\mathrm{K} > 31$ it is again monotonically increasing. For random epistatic interactions between $2 < K \leq 50$, the fitness values for the case with inversion are not very different from those of point mutations (note in Fig. IV.4 that, in this interval, the red and blue curves with marker 'o' are very close to each other).

On the other hand, we can observe that for $5 < \mathrm{K} \leq 65$, $(\Delta f_K)_{\mathrm{adj}}$ is monotonically decreasing, and for $\mathrm{K} > 65$ it is again monotonically increasing. We can also observe that, between $K > 0$ and $K \doteq 80$, $(\Delta f_K)_{\mathrm{rnd}} < (\Delta f_K)_{\mathrm{adj}}$. Contrary to the case of random epistatic interactions, for adjacent interactions $(\Delta f_K)_{\mathrm{adj}}$ is higher since the fitness values reached by inversions are higher than those reached by point mutations (note in Fig. IV.4 the gap between the red and blue curves with the marker '+'). So, we infer that an inversion—modifying several loci—results in a mutually advantageous conjunction with local epistatic interactions, that allows explorations of more combinations that can be beneficial. Finally, for $K > 80$, $(\Delta f_K)_{\mathrm{rnd}} \approx (\Delta f_K)_{\mathrm{adj}}$ (still monotonically increasing), i.e. regardless of the epistatic interaction neighbourhood, inversions can reach higher fitness values and attenuate the complexity catastrophe by not decreasing towards 1/2 (compare

with (Kauffman, 1993, Tables 2.1 and 2.2) and also note in Fig. IV.4, the gap between the tails of the red and blue curves).

Therefore, our results show that in the presence of inversion it is possible to reach higher fitness when compared to adaptive walks with only point mutations.

A direct interpretation of this result is given by the properties of the inversion's mutational network as it has been described above (see e.g. Fig. IV.2). Indeed, as it is more densely connected than the point-mutation mutational network, it is likely to allow a larger exploration of the fitness landscape and thus reach higher peaks, as observed here. However, given that the ruggedness of a fitness landscape depends on the mutational operator at work, an alternative explanation is that inversion mutations result in a smoother fitness landscape than that of point mutations, hence facilitating the finding of trajectories leading to higher peaks.

To test this assumption, we generalized the roughness measure introduced by Aita *et al.* (2001). More precisely, we measured deviations from fitness additivity (in the language of the NK model, we say that a landscape is additive when it is non-epistatic, that is, $K = 0$). We here use the term roughness from (Aita *et al.*, 2001) to distinguish this measure, which is a local one, from the classical ruggedness of the NK-fitness landscape which is a global property of the landscape. See, for example, Refs. (Lobkovsky *et al.*, 2011) and (Szendro *et al.*, 2013), for other definitions of roughness and how they are calculated. Following the approach introduced in (Aita *et al.*, 2001), we computed the roughness of the fitness landscape as the root mean square fitness variation due to each possible mutation, for both point mutations and inversions (see Methods for a formal mathematical definition of this measure). The results are shown in Fig. IV.5. As expected, for point mutations the roughness of the fitness landscape is (almost) linearly proportional to the epistatic interaction parameter K, for both types of epistatic interaction neighborhoods (adjacent and random). In contrast, in the case of inversion mutations, the roughness is always greater than that of point mutations, this trend being particularly visible for random epistatic neighborhood when compared to adjacent neighborhood. Interestingly, even for $K = 0$, the roughness is already higher than for point mutations (both epistatic neighborhood being equivalent in that case) while for $K = N - 1$ the roughness converges approximately towards similar values both for inversions and point mutations whatever the epistatic neighborhood.

This result shows that inversion mutations actually don't smooth the fitness landscape. On the opposite, the average roughness increases much faster with K in the case of inversions than in the case of point mutations (the roughness of the inversion-based fitness landscape with $K = 1$ being similar to the one for point mutations with $K = 50$, Fig. IV.5). This result also suggests a new explanation for the advantage of inversion mutations over point mutations. While the high connectivity of the inversions mutational network enables a better exploration of the fitness landscape, this effect is hampered (and not facilitated) by the effect of inversions on the roughness and while the combination of both positive (connectivity) and negative (roughness) is favorable for all values of K in the case of adjacent neighborhood, it is only favorable for high values of K in the random epistatic neighborhood. Indeed, for epistatic interactions with a random neighborhood, there is no noticeable difference between the average fitness values up to $K \simeq 40$ (red and blue curves with markers 'o' in Fig. IV.4). This is likely to be due to the fact that inversions are segmental operators. When epistatic interactions are confined to a segment

close to the focal nucleotide (which is the case for the adjacent neighborhood but not the random one), both segments can largely overlap, hence limiting the effect of the inversion to a set of epistatically interacting genes. This reduces the average roughness (compared to random epistatic neighborhood), leading to a more efficient exploration of the fitness landscape. Although a full mathematical proof is out of the scope of this paper, we develop a representative mathematical analysis that illustrates the origin of this pattern in S1 Text (section 3).

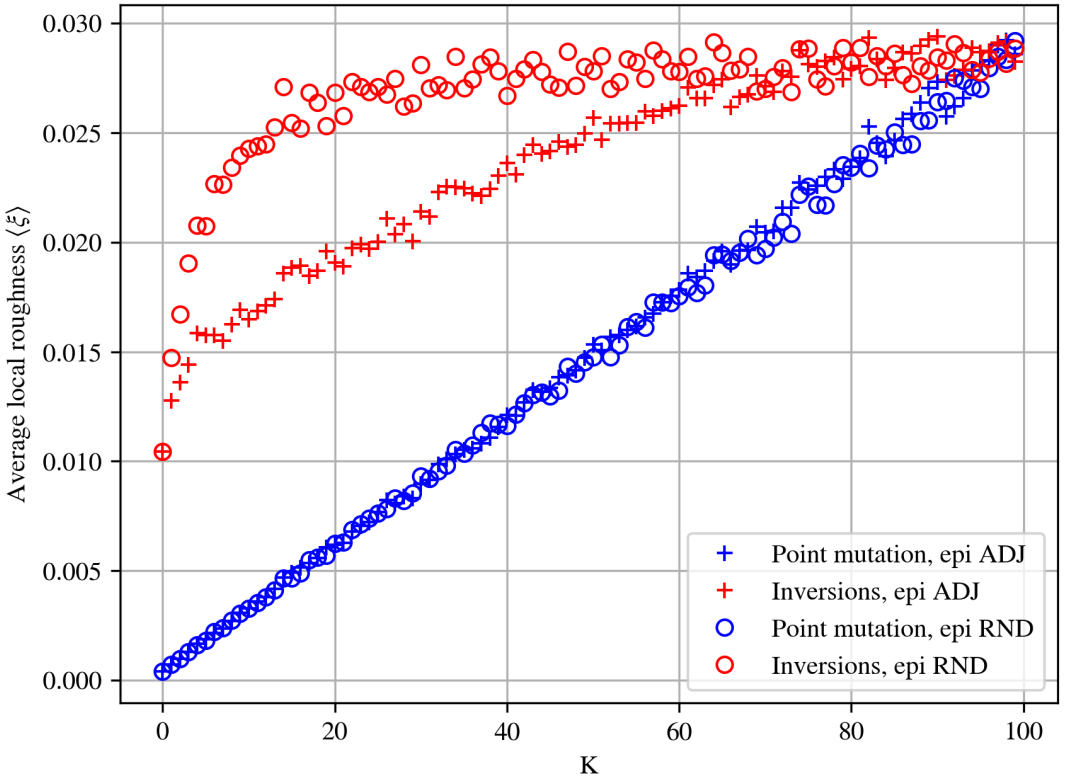


Figure IV.5: **Average value of the local measure of roughness.** Local roughness measured as the mean square differences between the fitness in a point of the landscape and its neighbours, for inversion (red) and point mutations (blue), for adjacent (crosses) and random (circles) epistatic interaction neighbourhoods, averaged for 100 instances of NK fitness landscapes.

3 Discussion

The results presented in this paper show that intragenic inversion mutations lead adaptive walks to reach higher fitness peaks on rugged landscapes. We have performed simulations in NK landscapes and have shown that the expected fitness values are higher for inversions than for point mutations. This holds for all degrees of ruggedness (epistasis), ranging from

single-peak (K = 0), moderately rugged ($1 \leq K < N - 1$), to fully rugged landscapes (K = $N - 1$). Simulations with point mutations agreed well with the already known characteristics of the NK model, and made it possible to establish a "control group" to ensure the reliability of the differences with inversion mutations. We also observed that for adjacent epistatic interactions, the differences of expected fitness values between inversions and point mutations are greater than in the case of random epistatic interactions. In this sense, we conjecture that this should be the consequence of a synergistic effect between inversions and adjacent epistatic interactions, epistatic adjacency enabling a set of interacting loci to be inverted at once without affecting other, non-interacting, loci (see S1 Text for a detailed discussion). We believe that the relationship between epistasis and structural inversions is an area that has not yet been deeply explored.

Our analysis consisted of adaptive processes driven by mutation-specific evolutionary steps, i.e. in addition to the widely used point mutations, we introduced a minimal model of inversion mutations in double-stranded (digital) genomes. This model has also revealed some consequences of the limits of simulations with point mutations. We showed that, in addition to ruggedness due to epistatic interactions, the escape process can also depend on the interrelationships between the genotype space and the fitness landscape that are mediated by the type of mutation. In particular, we showed that for inversion mutations, the graph-theoretical representation of accessible mutants displayed a complex topology, in comparison with the canonical genotype space constructed with point mutations. By definition, the node degree of the graph of mutations is the number of accessible mutants. In the case of inversions this number is no longer constant over the node set—as in the case of point mutations—but varies depending on the specific sequence composition. Therefore, although it is correct that we can generate genotypic space through point mutations, that does not (strictly) imply that evolutionary paths in the fitness landscape have to be solely through point mutations. In this sense, the inversion mutations allowed us to reveal new topological properties of the interconnection between genotypes through what we have defined as mutational networks. Indeed, it is this mutational network that mediates the interconnection between genotype space and the adaptive dynamics in the fitness landscape. The mutational network can be translated as a fitness network when the fitness values of each mutant genotype are included. Thus, revealing the directions of possible evolutionary paths. Let us remark that graph theory is a framework used in various models of evolution (see for example Refs. (Stadler, 1995; Stadler *et al.*, 2001; Stadler et Stadler, 2002; Beerenwinkel *et al.*, 2007b,a; Crona, 2014; Greene et Crona, 2014; Crona, 2018; Capitan *et al.*, 2015; Aguirre *et al.*, 2009, 2018)) and has been advantageous in analysis that use the Kauffman model, such as (Sarkar, 1990; Nowak et Krug, 2015; Hwang *et al.*, 2018; Kaznatcheev, 2019; Yubero *et al.*, 2017; Catalán *et al.*, 2017). However, most of these graph representations for mutations and fitness landscapes are isomorphic to the (hypercubic) genotype space, whereas our definition of mutational and fitness networks are not necessarily isomorphic to the genotype space. Moreover, we showed that for fitness networks generated by inversions, there are more mutational pathways between genotypes compared to point mutations. Therefore, for inversion mutations, at each step an evolutionary process can potentially explore more accessible mutants, than the "classical" estimation $(|\mathcal{A}_\mathcal{L}| - 1)N$, for any alphabet $\mathcal{A}_\mathcal{L}, \forall \mathcal{L} \in \{2n : n \in \mathbb{N}\}$, with size $|A_\mathcal{L}| = 2n$. In this sense, our work can straightforwardly complement the results reported in (Zagorski *et al.*, 2016), for $|\mathcal{A}_\mathcal{L}| > 2$ with point mutations. The main message

here is that in addition to the well-known utility of the fitness landscape metaphor, its topographic properties (due to epistasis) are not sufficient for modelling the escape from local peaks, and additional information should be included via the topology of mutational networks and fitness networks. This information can be useful to predict evolutionary trajectories in fitness landscapes (Lobkovsky *et al.*, 2011; Franke *et al.*, 2011; Koonin, 2011; Szendro *et al.*, 2013; De Visser et Krug, 2014; Bank *et al.*, 2016).

An important takeaway from this work is that an effort must be made to incorporate features of the structural variations of genomes, such as submicroscopic intragenic rearrangements. In this paper, we have taken a first step in this direction by modelling inversion mutations. A by-product of the construction of our model of inversion mutations revealed topological properties associated with the genotypic space and the accessible mutants for potential evolutionary paths. This is a consequence of the combinatorics of accessible mutants, which depends on structural aspects of this type of chromosomal rearrangement. On the other hand, our graph theoretical representations are consistent with the idea of adaptive walks in complex networks (Campos et Moreira, 2005; de Lima Filho *et al.*, 2012; Grewal *et al.*, 2018). The key difference is that in our model we do not have to postulate *a priori* a network that satisfies a certain topology (e.g. scale-free or random) as in (Campos et Moreira, 2005; de Lima Filho *et al.*, 2012; Grewal *et al.*, 2018). In our case, the topology arises as a consequence of the type of mutations. Following the interpretation provided in (Grewal *et al.*, 2018) about multiple mutations in a single evolutionary step, we suggest that another alternative to justify (or interpret) their *topological inspired walks* is via generic structural variations in concomitance with our model. In the case of the simulations of adaptive walks on complex networks reported in (Campos et Moreira, 2005), the authors state that "it seems more realistic to ponder sequence spaces where the node's connectivity is not the same for every node, as it is in hypercubes". We agree with the authors of (Campos et Moreira, 2005) and (Grewal *et al.*, 2018) that the degree of a genotype (in the mutational network for us) measures the availability of accessible mutants. This is what we have proposed to define as mutability, that is, the ability to change from one genotype to another under a mutation. We believe that it could be interesting to explore this notion of mutability and its relationship with the genetic potential of mutations that give rise to novel (beneficial) phenotypes(Ancel Meyers *et al.*, 2005).

In this work, we studied the effect of inversion mutations on the maximum fitness reached on rugged landscapes and compare it to the maximum fitness reached by point mutations. Importantly, both kind of mutations have been tested in independent simulations under the Strong Selection Weak Mutation regime. Hence, although we have shown that inversions reach higher fitness values in these conditions, we cannot compare the evolutionary dynamics between these two mutational setups. However, it is important to stress that, in a real population, both kind of mutations would not occur in isolation. Any evolving population undergoes all mutation types, including both point mutations and inversions. Hence, inversion mutations should not be considered in competition with point mutations, but rather as a synergistic interaction. Studying the consequences of this interaction on the evolutionary dynamics constitutes one of the most exciting perspective of this work.

In our model we did not include recombination (Klug *et al.*, 2019) and other chromosomal rearrangements, such as duplications, deletions, and translocations. However, our purpose with inversion mutations has been to exemplify a simple mechanistic model

of structural variation. Of course, our sequence model includes many simplifications. In particular, we use binary sequences and simulated a fully coding compact genome with a circular double-stranded DNA. Although most of our theoretical results hold in a more general case, the effect of inversions on more realistic coding sequences (with e.g. a 4 bases alphabet, multiple reading frames and ORF identified by start and stop codons and separated by non-coding sequences) could reveal other properties of interest. For instance, micro-inversions effect on reading frames is likely to be specific and very different from the effect of point mutations (that don't shift the reading frames) or InDels (that are likely to shift it). Indeed, inversions can alter a subsequence of an ORF without changing the reading frame of the start/stop codons. On the opposite, an inversion can easily remove (or create) a stop codon.

All throughout this study, we focused on binary sequences, simplifying the 4-bases nature of real genomes. Although the binary description is very common in theoretical and computational models, it is important to mention that the properties of the different mutational operators, hence of the generated mutational networks, may differ depending on the size of the alphabet (Zagorski *et al.*, 2016). In the case of inversions, although our main conclusions about the complex structure of the mutational network still hold, a 4-bases alphabet with two pairs of complementary bases (`A`-`T` and `C`-`G`) would introduce important properties compared to the binary case. Indeed, given the mathematical definition of inversions (Eq. IV.5), it is straightforward that the composition of the inverted segment will conserve the relative fraction of `AT` and `CG` pairs relatively to the original one (a specific situation being the inversion of a segment of size one that can only switch `A` and `T` or `C` and `G`). It immediately follows that inversions cannot change the `AT`/`CG` ratio of the sequence and that, for a sequence of length N, the mutational network generated by inversions contains at least $N+1$ disconnected sub-networks (which sizes will depend on the `AT`/`GC` ratio of the sequence, strongly biased sequences leading to smaller sub-networks). Hence, compared to the binary model, the 4-bases model increases the size and connectivity of the mutational network generated by point mutations (Zagorski *et al.*, 2016). On the opposite, in the inversion-generated mutational network, it isolates several sub-networks from each others. The effect on evolutionary dynamics, as we studied it here using the NK landscape, is still to be explored. Indeed, depending on the composition of the initial sequence, with a 4-bases alphabet some local/global optima may not be accessible. Consequently, the advantage of inversion mutations may be reduced, and even be cancelled if the number of local optimum is very low (i.e. for $K \ll N$). However, it is worth mentioning that real genomes undergo both kinds of mutational events and that point mutations connect the inversion-isolated sub-networks by changing the `AT`/`GC` ratio of the sequence. Exploring how both kinds of mutations (and others) interact is clearly beyond the scope of this manuscript, but studying the synergistic effect of inversions and point mutations in a more sequence-realistic model like the Aevol model (Rutten *et al.*, 2019; Beslon *et al.*, 2021) is clearly an appealing perspective.

Despite the simplifications used in this study, our results show that structural inversions could be considered not only as changes in the orientation of sequences that don't alter the genetic content, as classically supposed in the literature (Fertin *et al.*, 2009), but also as a source of intragenic variations. In this sense, our phenomenological model is supported by the empirical evidence of an intragenic inversion associated with the creation of new regulatory elements—required e.g. for the termination-activation of transcription in

the nitric oxide synthase gene in *Lymnaea stagnalis* (Korneev et O'Shea, 2002). As well, the pathogenic mutation due to an intragenic inversion of seven nucleotides in human mitochondrial DNA (Musumeci *et al.*, 2000). Furthermore, although first generation sequencing technologies were unable to identify submicroscopic rearrangements (Ho *et al.*, 2020), the development and availability of novel sequencing technologies (Ho *et al.*, 2020; Wellenreuther *et al.*, 2019), opens the possibility of characterising intragenic structural variations and may be of particular importance to unravel new aspects of mutations in molecular evolution. Therefore, we may soon require new theoretical and computational models to simulate the fullness of chromosomal rearrangements in evolutionary biology. We hope this work makes a first step in this direction.

4 Conclusion

The statements presented in this paper provided computational evidence that, for a very simple model of evolution in the strong selection weak mutation limit, an adaptive process in rugged landscapes driven by intragenic inversion mutations can reach higher fitness values (compared to a same process driven by point mutations). Therefore, this implies that intragenic inversion mutations can lead evolution to escape local fitness peaks in rugged landscapes. The way our model was conceived also proves that escape from a local peak of fitness can occur in constant environments without contingencies. Our model for inversion mutations not only elucidated an escape mechanism, but have also made it possible to uncover interesting aspects about the combinatorics of inversions and their relationship with mutated genotypes, genotype spaces and fitness landscapes in terms of graphs representations.

5 Methods

5.1 The model

Preamble: Limits in the single-nucleotide mutation scenario It is worthwhile to state the main issue when point mutations are the only source of genetic variations in evolutionary models. At the molecular level and besides the fitness values, the structure of DNA mutations constrains the way evolution can move through the genotype space. For example, a common reasoning in molecular evolution theory is: for any alphabet $\mathcal{A}_\mathcal{L}, \forall \mathcal{L} \in \{2n : n \in \mathbb{N}\}$, with size $|A_\mathcal{L}| = 2n$ (this total number of letters with even parity being due to the double-stranded structure), a sequence $\mathbf{x}$ of N nucleotides, would have

$$D(\mathbf{x}) = (|\mathcal{A}_\mathcal{L}| - 1)N \tag{IV.1}$$

mutant neighbours differentiated by a single point mutation (Smith, 1970; Gillespie, 1984, 1991; Kauffman et Levin, 1987). These D neighbouring genotypes are available for natural selection. Then, at the SSWM limit, only one of these neighbouring sequences can be fixed, chosen among those with fitness values higher than the wild-type fitness. Once a new mutant is fixed, a new set of D mutant neighbours is available for selection. Repeating this process unfolds an evolutionary path until it reaches a local (or global) fitness peak

in the rugged landscape. However, if a local peak is reached (i.e. the state when all D accessible genotypes have strictly lower fitness values), then the evolutionary process is "trapped" because any other fitter genotype is at two or more point-mutational steps away (i.e. only attainable by "descending" through a valley in the fitness landscape).

5.1.1 Digital sequence scheme

First, let us recall the very basic and well-known notion that DNA is a double strand molecule with two nucleotides chains, held together by complementary pairing of adenine (`A`) with thymine (`T`) and guanine (`G`) with cytosine (`C`). Given a DNA strand, as for example `ATCGATTGAGCTCTAGCG`, its complementary strand is `TAGCTAACTCGAGATCGC`, which in the IUPAC's notation is

$$\begin{matrix} 5' - \texttt{ATCGATTGAGCTCTAGCG} - 3' \\ 3' - \texttt{CGCTAGAGCTCAATCGAT} - 5' \end{matrix} ,$$

where the leading strand is on top and the DNA strand orientation is by convention $5' \to 3'$.

Throughout the presentation of our model, we adopt the alphabet $\mathcal{A}_2 = \{0, 1\}$, so the genotypes are binary sequences of (constant) length $N \in \mathbb{N}_{\geq 2}$. As a low-level structural representation consistent with the DNA molecular biology, these genotypes are double-stranded sequences

$$\mathbf{x} := x_1 x_2 \cdots x_{N-1} x_N,$$

with N digital nucleotides $x_i \in \mathcal{A}_2, \forall i \in \{1, \ldots, N\}$, where the complementary sequence $\bar{\mathbf{x}}$ is defined such that $\forall x_i \in \mathcal{A}_2$, $\bar{x}_i := 1 - x_i$, $\forall i \in \{1, \ldots, N\}$. All the sequences $\mathbf{x}$ of size N define the set $\mathcal{X} \in \{0, 1\}^N$ of possible genotypes.

In analogy with the example above, the representation of this double-stranded digital sequence is:

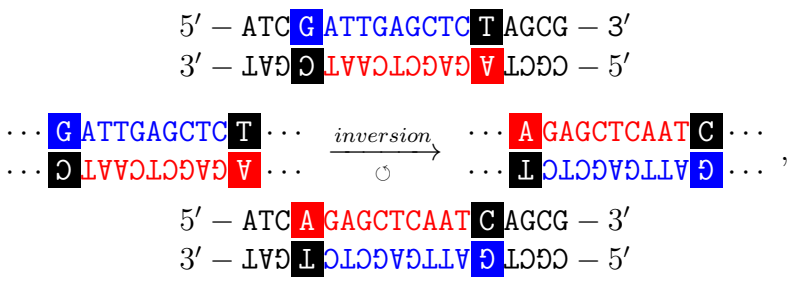
$$\begin{matrix} \texttt{011001100011110010} \\ \texttt{100110011100001101} \end{matrix} .$$

It is very important to clarify that our schematic representation should not be confused with the usual encoding $\mathcal{A}_{\texttt{DNA}} \to \mathcal{A}_2$, with the convention purines $\{\texttt{A}, \texttt{G}\} \to 0$ and pyrimidines $\{\texttt{T}, \texttt{C}\} \to 1$. The artificial genetics in our model only have two nucleotides, and they complement each other. We also adopt the following approximations:

- We assume that the sequences are circular, i.e. periodic boundaries: $x_{N+i} = x_i, \forall i \in \{1, \ldots, N\}$.
- We neglect the existence of coding sequences separated by non coding regions. Biologically, this corresponds to compact genomes with almost no non-coding sequences. In this sense, our model mimics the molecular evolution of some viruses (Solé et Elena, 2018) or (animal) mitochondrial DNA (DiMauro, 2001). Consequently, we are modelling intragenic mutations(Ho *et al.*, 2020; Bank *et al.*, 2016).
- We neglect geometrical aspects such as physical configurations like folding, twists, coiled structures, hairpin loops, etc.
- We do not consider transcription-translation processes.
- We do not include recombination, so we are modelling asexual replication.

5.1.2 Structural inversions model

To establish the main idea of DNA inversion mutations well, now let us illustrate this mechanism with the following representation:



(IV.2)

where in the middle is depicted the segment where the mutation occurs and their corresponding inversion (boxes and colours highlight the segment where the inversion occurs). A glance over schema (IV.2) shows why a double-stranded-like model is unavoidable to model intragenic inversions.

For computational purposes, an inversion mutation can be split in two operations:

- The conjugation operation $\hat{\mathrm{C}}$:

$$\hat{\mathrm{C}} : x_i, x_{i+1}, \ldots, x_{j-1}, x_j \longrightarrow \bar{x}_i, \bar{x}_{i+1}, \ldots, \bar{x}_{j-1}, \bar{x}_j, \tag{IV.3}$$

  for $i, j \in \{1, ..., N\}$.

- The permutation operation $\hat{\mathrm{P}}$:

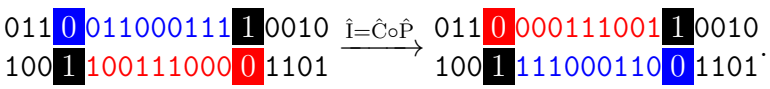
$$\hat{\mathrm{P}} : x_i, x_{i+1}, \ldots, x_{j-1}, x_j \longrightarrow x_j, x_{j-1}, \ldots, x_{i+1}, x_i, \tag{IV.4}$$

  for $i, j \in \{1, ..., N\}$.

Note that, as the genome is circular, there is no relation of order between i and j (i.e. $\hat{\mathrm{C}}$ and $\hat{\mathrm{P}}$ are well-defined even when $i > j$). Other sites $k \notin \{i, \ldots, j\}$, remain unchanged. Then, we have the (two-step) inversion operation:

$$\hat{\mathrm{I}} \equiv \hat{\mathrm{C}} \circ \hat{\mathrm{P}}. \tag{IV.5}$$

For example:



(IV.6)

Trivially, it can also be verified that $\hat{\mathrm{C}}$ and $\hat{\mathrm{P}}$ commutes:

$$\hat{\mathrm{C}} \circ \hat{\mathrm{P}} = \hat{\mathrm{P}} \circ \hat{\mathrm{C}}. \tag{IV.7}$$

Besides, Eq. IV.5 can also define a single-locus mutation: when $i = j$, then $\hat{\mathrm{I}}$ is a single bit-flip i.e. a point mutations.

Computationally, these operations can be easily implemented through Algorithm 1: `Mutate`. With this simple algorithm, it is possible to calculate the combinatorics of the inversion mutations. For example, the enumeration of accessible mutants for each genotype with $N = 4$ shown in Fig. IV.1.

Algorithm 1: `Mutate` $(\mathbf{x}, i, j, N)$

input: $\mathbf{x} \in \mathcal{X}$, $[i, j] \in \{1, \ldots, N\}$

$l \leftarrow i$
$u \leftarrow j$
$\mathbf{y} \leftarrow \mathbf{x}$
repeat
  $\mathbf{y}_l \leftarrow (1 - \mathbf{x}_u)$
  $l \leftarrow l + 1 \mod N$
  $u \leftarrow u - 1 \mod N$
until $l = j + 1$
return $\mathbf{y} \in \{0, 1\}^N$

5.2 NK model

A well known model of genetic epistatic interactions is the NK family of rugged multipeaked fitness landscapes (Kauffman et Levin, 1987; Kauffman et Weinberger, 1989; Weinberger, 1991; Kauffman, 1993). In this model, besides the genome length $N \in \mathbb{N}$, the integer $K \in \mathbb{Z}_{(0,N-1)}$, describes the epistatic interactions between loci in the genome and the contribution of each component to the total fitness, which depends on its own value as well as the values of K other loci. The fitness per locus is formally defined as:

$$f_i : \{0, 1\}^{K+1} \longrightarrow [0, 1), \quad \forall 1 \leq i \leq N.$$

Here $f_i(x_i, x_{i_1}, \ldots, x_{i_K})$ depends on the state of locus $x_i \in \{0, 1\}$ and K other loci $x_{i_K} \in \{0, 1\}$. The f_i's are given by $N \cdot 2^{K+1}$ independent and identically distributed random variables sampled from a given uniform probability distribution. See the example shown in (Weinberger, 1991, Table I), for a very illustrative description for the computing of the epistatic contribution per locus. The pattern into which the scheme of interaction between loci is connected is known as the epistatic neighbourhood (Kauffman et Weinberger, 1989; Weinberger, 1991). In our simulations we use two popular neighbourhood models:

- The adjacent neighbourhood model, where i and the K other sites are successively ordered, i.e. $i, i+1, \ldots, i+K$ (each variable modulo N when using periodic boundary conditions).
- The random neighbourhood model, where i and the K other loci are chosen at random according to a uniform distribution from $\{1, 2, \ldots, N\}$.

Examples for $N = 4$ are depicted in Fig. IV.6.

The total fitness $f \in [0, 1)$ for the genotype $\mathbf{x} \in \mathcal{X}$ is then defined as:

$$f(\mathbf{x}) := \frac{1}{N} \sum_{i=1}^{N} f_i(x_i, x_{i_1}, \ldots, x_{i_K}), \tag{IV.8}$$

where $\{i_1, \ldots, i_K\} \subset \{1, \ldots, i-1, i+1, \ldots, N\}$.

The most important feature of the NK model is that the parameter K tunes the landscape ruggedness, that is the distribution of fitness local maximums, ranging from

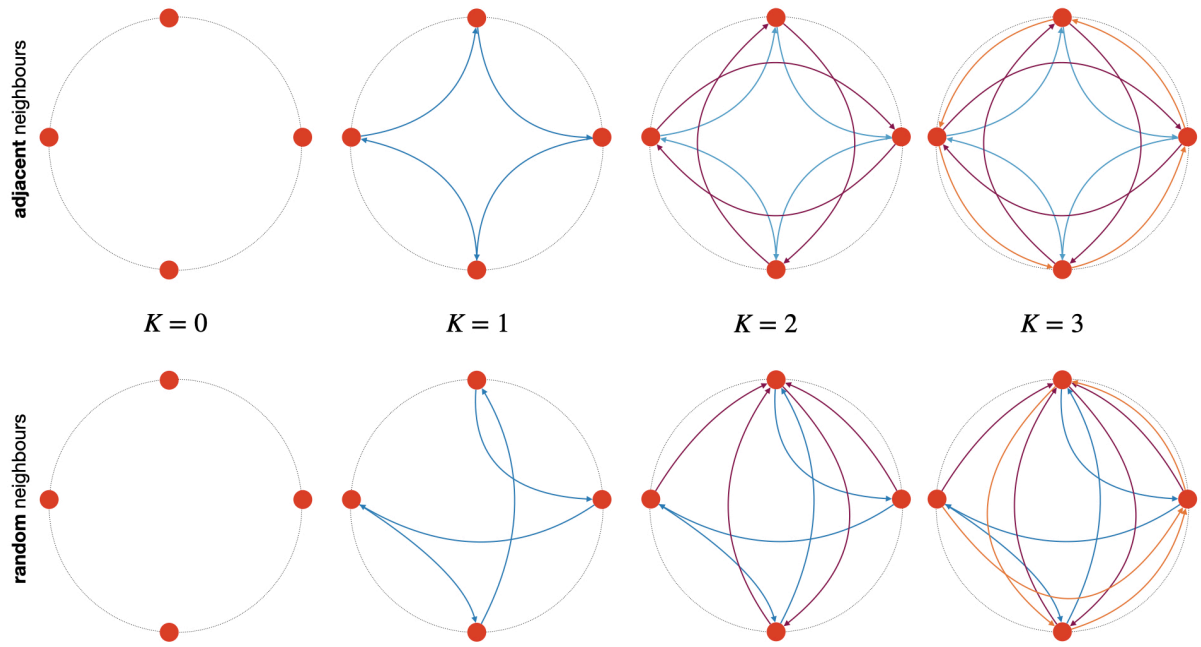


Figure IV.6: **Neighbourhood epistatic interactions of the NK model.** Schematic representation of a circular genome with $N = 4$ and epistatic interactions $K = \{0, 1, 2, 3\}$. Top: Example of adjacent neighbours. Bottom: Example of random neighbours (let us remark that in this case, for K $>$ 0 the in and out degrees per node do not have to be equal).

non-epistatic interactions when $K = 0$ (a Mount-Fuji-like landscape with a single peak), to the full rugged (or random) landscape when $K = N - 1$.

5.3 Adaptive walks

The zero-order approximation of our model to population genetics theory is on the limit of strong selection weak mutation (SSWM) (Gillespie, 1991, Ch. 5). In this limit, the adaptive walk model describes very well the molecular evolution of isogenic (monomorphic) populations, as the sequential fixing of novel beneficial mutations. Therefore, the simulation of evolutionary processes in our digital model can easily be translated within this framework. That is, instead of describing a population of organisms with a pool of genotypes, it is sufficient to simulate the evolutionary trajectory over the fitness landscape of a single initial genotype and its successive mutations.

The procedure goes as follows: a (randomly chosen) starting genotype $\mathbf{x} \in \mathcal{X}$ varies through successive mutations (calculated with Algorithm 1: `Mutate`) resulting in a mutated sequence $\mathbf{y} \in \mathcal{N}_\nu(\mathbf{x})$, where $\mathcal{N}_\nu(\mathbf{x})$ is the set of all accessible mutants of genotype $\mathbf{x}$. Then, the fitness $f(\mathbf{x})$ and $f(\mathbf{y})$ are calculated according to the NK model with Eq. (IV.8). If $f(\mathbf{y}) > f(\mathbf{x})$, the mutated genotype is selected, otherwise other mutations on $\mathbf{x}$ are tested until the fitness increases. In Algorithm 1: `Mutate`, the loci $[i, j] \in \{1, \ldots, N\}$ are

drawn from a pseudo random number generator function. With this recipe, the evolutionary dynamics is simulated up to a local fitness maximum is reached, i.e. $\mathbf{x} \in \mathcal{X}$ satisfies: $f(\mathbf{y}) < f(\mathbf{x})$, $\forall \mathbf{y} \in \mathcal{N}_\nu(\mathbf{x})$. In other words, we verify that all mutants for a given genotype do not have higher fitness values, if not, the simulation continues.

The main routine to simulate adaptive walks on the Kauffman's NK-fitness landscape model, with point mutations (as usual) and inversions (as new) is available at `https://gitlab.inria.fr/letrujil/getting-higher`. Our code is based on the one developed by Wim Hordijk (in its version of August 23, 2010 and which is available at `http://www.cs.unibo.it/~fioretti/CODE/NK/`), which uses some code from Terry Jones (`https://github.com/terrycojones/nk-landscapes`).

5.4 Roughness measure

The roughness to slope ratio proposed by Aita *et al.* (2001), can be re-interpreted in terms of the local measure of roughness of the surface of a solid material or an irregular interface, i.e. as the root mean square surface width in function of the height at a given place on the surface (see for example (Barabási *et al.*, 1995, p.22)). In our case if we assume that the "height" is equivalent to the value of the fitness $f(\mathbf{x})$ of genotype $\mathbf{x}$, and "the place on the surface" corresponds to a domain (on the surface) of the fitness landscape, then we can define the measure of local roughness given a genome $\mathbf{x} \in \{0,1\}^N$ as:

$$\xi_\nu(\mathbf{x}) := \left(\frac{1}{|E(m_\nu(\mathbf{x})|} \sum_{(\mathbf{x},\mathbf{y}) \in E(m_\nu(\mathbf{x}))} \left(f_\nu(\mathbf{x}) - f_\nu(\mathbf{y})\right)^2 \right)^{\frac{1}{2}}, \forall \mathbf{x} \in \{0,1\}^N, \tag{IV.9}$$

where the index ν denotes point mutations (P) or inversions (I), and $E(m_\nu(\mathbf{x}))$ is the set of all possible mutations $\mathbf{y}$ of a given genotype $\mathbf{x}$ (i.e. the edges of the directed multigraph of mutations $m_\nu(\mathbf{x})$).

Acknowledgments

We wish to acknowledge insightful conversations with members of the Beagle team, especially David P. Parsons and Aoife O. Igoe also for their suggestions on the manuscript. L.T. thanks the Institut National des Sciences Appliquées (INSA) as well as the Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS) for hospitality while part of this research was done and would like to thank Anton Crombach, Harold P. de Vladar and Ivan Junier for useful discussions and valuable support. P.B. is grateful to Laurent Turpin and Nathan Quiblier for stimulating discussions.

Data availability

The source code of all models and stimulation files used in the present paper can be found in `https://gitlab.inria.fr/letrujil/getting-higher`

Competing interests

The authors have declared that no competing interests exist.

Author contributions

Conceptualization: Leonardo Trujillo, Guillaume Beslon.

Data Curation: Leonardo Trujillo, Paul Banse.

Formal Analysis: Leonardo Trujillo, Paul Banse.

Investigation: Leonardo Trujillo, Paul Banse.

Methodology: Leonardo Trujillo, Paul Banse, Guillaume Beslon.

Project Administration: Leonardo Trujillo.

Software: Leonardo Trujillo, Paul Banse, Guillaume Beslon.

Supervision: Leonardo Trujillo, Guillaume Beslon.

Validation: Leonardo Trujillo, Paul Banse, Guillaume Beslon.

Visualization: Leonardo Trujillo, Paul Banse.

Writing Original Draft Preparation: Leonardo Trujillo.

Writing Review & Editing: Leonardo Trujillo, Paul Banse, Guillaume Beslon.

Open Researcher and Contributor ID (ORCID)

0000-0001-9995-4135 Leonardo Trujillo

0000-0003-2373-6785 Paul Banse

0000-0001-8562-0732 Guillaume Beslon

Funding

L.T. was partially supported by Institut National des Sciences Appliquées (INSA) Visiting Professor Fellowship. P.B. was supported by Ministère de l'Enseignement Supérieur et de la Recherche. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supporting Figure S1 Fig

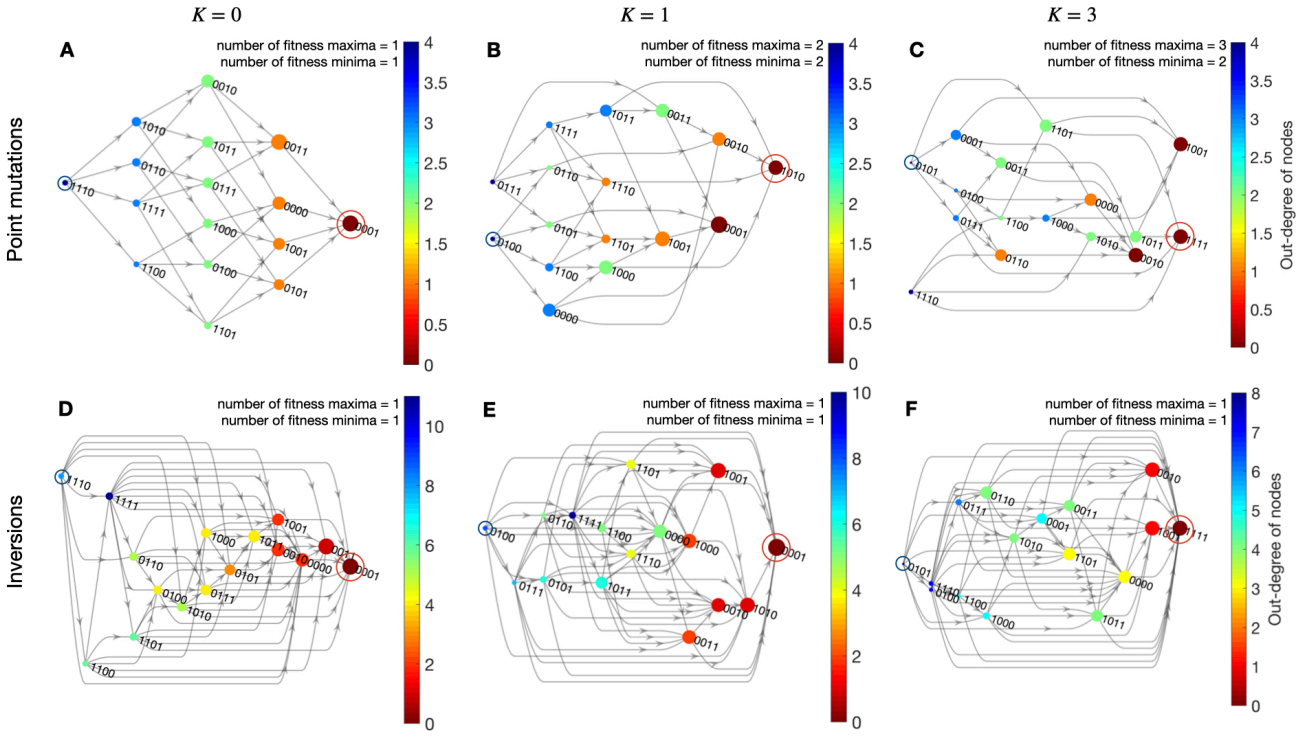


Figure IV.7: **NK fitness networks for epistatic interactions with adjacent neighbouring.** Representative instances of the NK model for $N = 4$ and their fitness networks in layered representation. The layers are constructed such that each node is assigned to the first possible layer, with the constraint that all its predecessors must be in earlier layers. The colors of the nodes correspond to the values of the out-degrees, i.e. the number of edges going out of a node (note that color scales differ in range between panels). Therefore, nodes with node out-degrees equal to zero correspond to local fitness maxima (sink nodes). The landscapes' ruggedness are: single peaks $K = 0$, intermediate ruggedness $K = 1$ and full rugged case $K = 3$, for adjacent neighbouring epistatic interactions. Node sizes are scaled with fitness values (best fitness, largest and vice versa). Global maximum of fitness are encircled in red. While the global minimum in blue. The total number of fitness maxima and minima are also reported. See IV.3 in main text for epistatic interactions with random neighbouring.

Supplemental Information for:

Getting higher on rugged landscapes: Inversion mutations open access to fitter adaptive peaks in NK fitness landscapes

Leonardo Trujillo, Paul Banse, Guillaume Beslon

6 Supplementary Text

6.1 Fraction of invariant inversions

To enumerate the average number of invariant inversions, let us characterise what an invariant inversion is.

An inversion of size j on a sequence $x_i...x_{i+j-1}$ is invariant if and only if:

$$x_i...x_{i+j-1} = \bar{x}_{i+j-1}...\bar{x}_i \tag{IV.10}$$

Thus, $\forall m \in [0, j], x_{i+m} = \bar{x}_{i+j-1-m}$ and if j is an odd number, then for $m = \frac{j-1}{2}$ we have $x_{i+m} = \bar{x}_{i+m}$. This is not possible, hence only even sized inversion can be invariant.

Then, given a random inversion of even size $j = 2 \times q$, what is the probability that this inversion is invariant?

For any sequence, $x_i...x_{i+q-1}$ there is only one sequence $x_{i+q}...x_{i+j}$ such that the property (IV.10) is respected. Thus, the probability for an inversion of even size $j = 2 \times q$ to be invariant in a binary alphabet is $\frac{1}{2^q}$.

Let us note the event: “a random inversion is invariant on a binary sequence" as R^2 and “a random inversion of size j is invariant on a binary sequence" as R^2_j. By cumulating all the possible inversion sizes for a genome of size N we get :

$$
\begin{aligned}
\mathcal{P}(R^2) &= \sum_{j=1}^{N} \mathcal{P}(\text{size} = j)\ \mathcal{P}(R_j^2|\ \text{size} = j) \\
&= \sum_{j=1}^{N} \frac{1}{N} \mathcal{P}(R_j^2|\ \text{size} = j) \\
&= \frac{1}{N}\left[\sum_{q=1}^{\lfloor \frac{N}{2} \rfloor} \frac{1}{2^q} + \sum_{q=0}^{\lfloor \frac{N+1}{2} \rfloor} 0\right] \\
&= \frac{1}{2N} \frac{1 - 1/2^{\lfloor \frac{N}{2} \rfloor}}{1 - 1/2} \\
&= \frac{1}{N}\left(1 - \frac{1}{2^{\lfloor \frac{N}{2} \rfloor}}\right) \\
&\underset{n \to +\infty}{=} \frac{1}{N}
\end{aligned}
$$

It is worth noticing that small inversions are the main drivers of invariability. Similarly, it can be proven that for a four-bases alphabet, $\mathcal{P}(R^4) \underset{n \to +\infty}{=} \frac{1}{3N}$.

6.2 Linear or limited sized inversions

To complement our analysis, here we present the results of adaptive walks simulations for inversion mutations with different constraints: bounded maximum size and linear chromosome. The simulation setup is the same as that used in the calculations of the mean fitness shown in Fig. 4 of the main manuscript.

6.3 Bounded sized inversion

We simulated inversion mutations with an upper limits $s \leq N$ that ranges from 1% of chromosome size —a single locus (i.e. inversions are like point mutations)—up to 100% of chromosome size. Figures IV.8 and IV.9 shows eight examples of maximum inversion mutations sizes $s = \{1, 2, 4, 8, 16, 32, 64, 100\}$, for a genome of size $N = 100$ and for adjacent and random epistatic neighbourhood respectively. The mean fitness is averaged over 100 instances of adaptive walks simulations, for all values of epistatic interaction, $K = \{0, 1, \ldots, N - 1 = 99\}$.

In agreement with Fig. 4, for $K = 0$ (no epistatic interaction between neighboring loci) the average fitness for all inversions sizes are equal, i.e. $\langle f \rangle_{K=0} \simeq 0.667$. This is also consistent with the analytical result, $\langle f \rangle_{K=0} = \frac{2}{3}$, for smooth landscapes with a single peak [(Kauffman, 1993), p. 55]. For $K > 0$, the average fitness $\langle f \rangle_K$ increases with K until reaching a maximum value of fitness, and then decreases towards a minimum at $K = N - 1$, with a really similar behaviour than in Fig. 4. Given that when the maximum inversion size is $s = 1$, the only possible mutation is a point mutation and when $s = N$ all the inversions are available, a smooth transition from the blue line to the red one is expected, and indeed observed. Also, consistently with what we observed on Fig. 4 for

the random epistatic neighbourhood, in this case no difference of fitness is observed for low K values, whatever s (IV.9).

However, importantly, the gain of fitness due to the increased s value is not linear, for example: for low values of K, in IV.8, the transition from $s = 1$ to $s = 2$ correspond to an average final fitness gain comparable to the transition from $s = 4$ to $s = 100$. This non-linearity on low values of K fades away when K reaches very high values (typically higher than $K = 80$). Indeed, for such high values of K the correlation between the fitness of a genome and its mutants is very low (See roughness in Fig. 5). In this case, the final fitness reached becomes mainly a combinatorial challenge, and the combinatorics of available mutations increases with s.

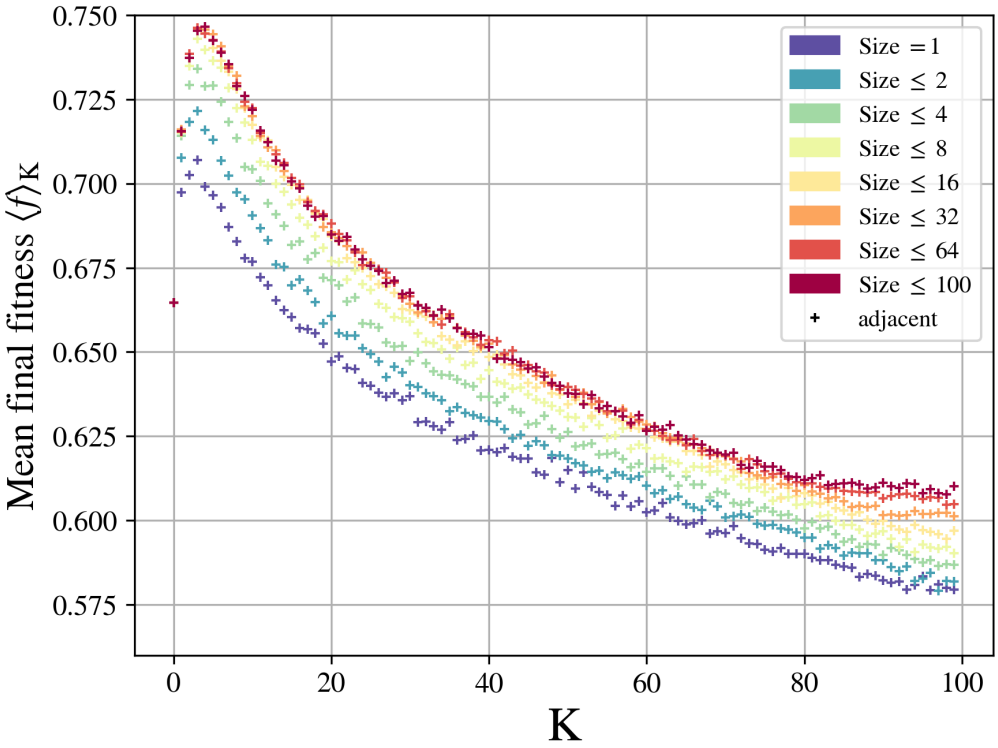


Figure IV.8: **Mean final fitness for inversion mutations restricted to a certain size range (epistatic interaction with adjacent neighborhood).** Changes in mean final fitness for different epistatic parameter K for point mutations (blue) and inversions restricted to a certain size range (sizes values and colors nomenclature are displayed in the figures), averaged for 100 instances of adaptive walks simulations in NK landscapes.

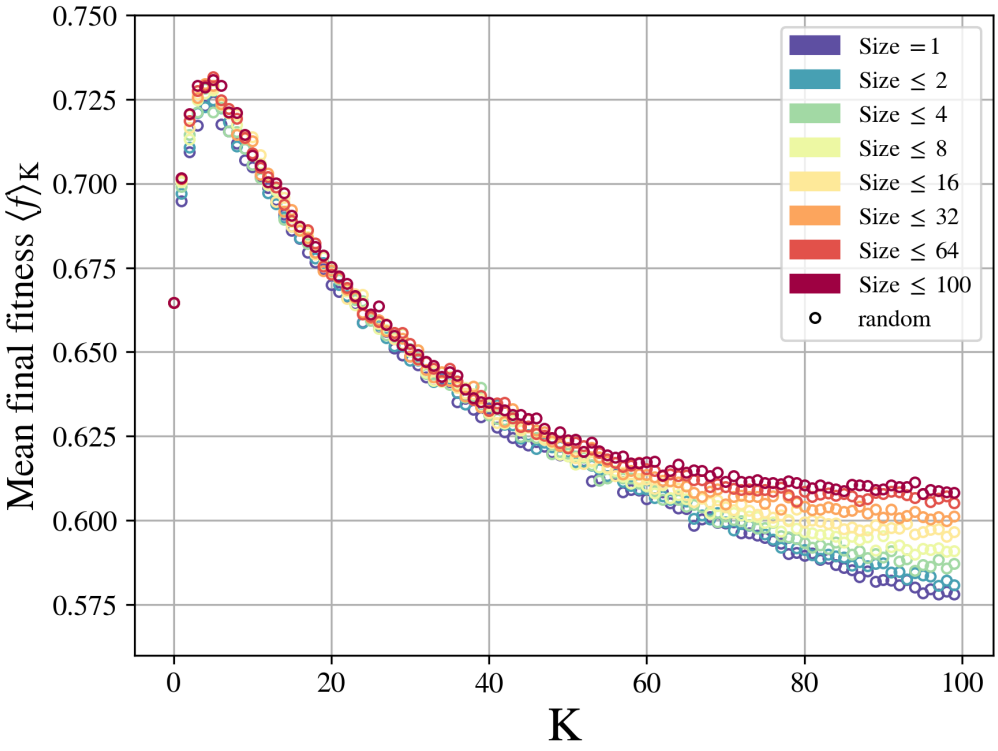


Figure IV.9: **Mean final fitness for inversion mutations restricted to a certain size range (epistatic interaction with random neighborhood).** Changes in mean final fitness for different epistatic parameter K for point mutations (blue) and inversions restricted to a certain size range (sizes values and colors nomenclature are displayed in the figures), averaged for 100 instances of adaptive walks simulations in NK landscapes.

6.4 Linear chromosomes

Linear chromosomes imply a limitation of the general inversion on circular genome. All inversions are allowed, except the ones overlapping with the boundaries. A direct consequence is that inversions of larger size are more likely to cross the boundaries and thus are more likely to be forbidden. The figure IV.10 shows the average final fitness when considering only linear inversions for all K values and for the two epistatic neighbourhood (green marks). To ease the comparison with the general case, the final fitness with inversions on a circular genome and with point mutations is presented in shaded red and blue respectively The values are relatively similar to inversions, the main difference being around $K = N - 1$. This behaviour is the same as the one observed previously with bounded inversions (with an upper size limit between $s = 32$ and $s = 64$). This is consistent with the idea that the presence of inversions in linear chromosomes mainly affect large sized inversions.

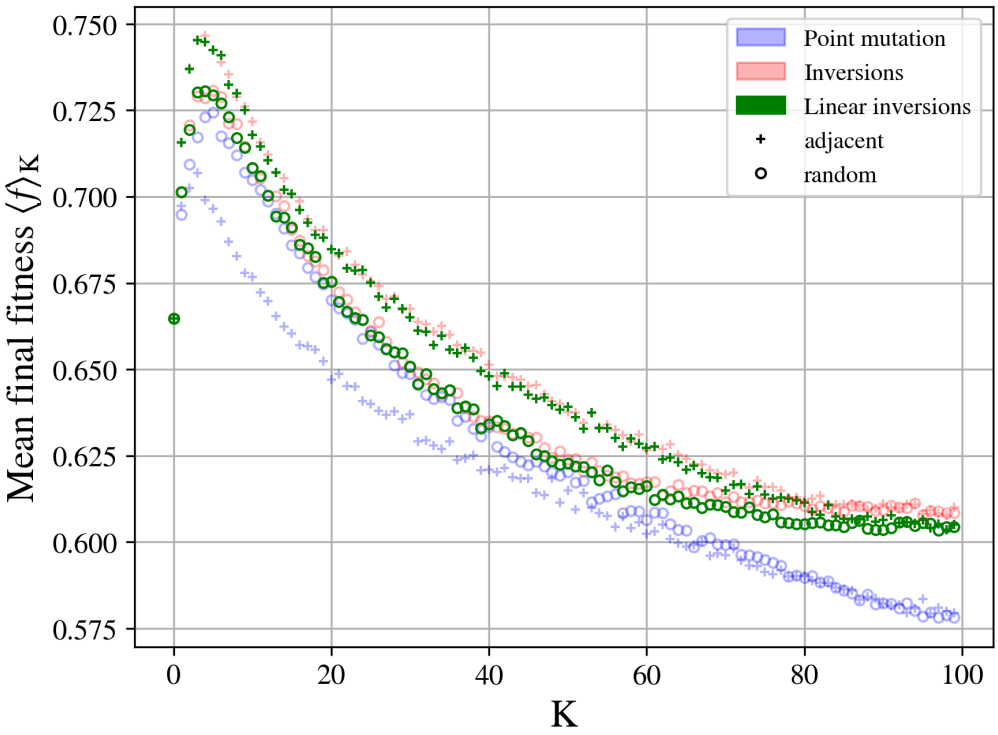


Figure IV.10: **Comparison of mean final fitness between circular and linear chromosomes.** Changes in mean final fitness for different epistatic parameter K for linear chromosomes (green markers), averaged for 100 instances of adaptive walks simulations in NK landscapes. For comparison with circular genomes, the initial values are shown dimly for inversions (shaded red) and point mutations (shaded blue). The circle (respectively cross) markers correspond to random (respectively adjacent) neighbouring epistatic interactions.

6.5 On the synergistic effect of inversions in the adjacent neighbourhood

Simulation of inversion mutations on the NK fitness landscape show that this operator enables reaching higher peaks than point mutations, especially for large values of K (typically $K > 40$, for $N = 100$). However, for lower values of K, this result only holds for the adjacent epistatic neighbourhood but not for the random epistatic neighbourhood (in that case, the fitness reached by inversions being similar to the fitness reached by point mutations). This is likely to be due to the fact that inversions effect is confined to a segment (we call inversions "segmental operators"). When epistatic interactions are also confined to a segment close to the focal nucleotide (which is the case for the adjacent neighborhood but not the random one), both segments can largely overlap, hence limiting the effect of the inversion to a set of epistatically interacting genes. Although a full mathematical proof is out of the scope of this paper, we develop here a representative mathematical analysis that illustrates this point.

To develop our argument, first let's recall Eq. 8 (main text). In an NK fitness land-

scape, the fitness $f(\mathbf{x})$ is given by:

$$f(\mathbf{x}) := \frac{1}{N}\sum_{i=1}^{N} f_i(x_i; x_{i_1}, \ldots, x_{i_K}). \tag{8}$$

Let us consider a genome $\mathbf{x} \in \{0,1\}^N$ with size $N = 12$ as an illustrative example that we can handle:

$$\mathbf{x} = x_1x_2x_3x_4x_5x_6x_7x_8x_9x_{10}x_{11}x_{12},\ \forall x_i \in \{0,1\}.$$

For simplicity, we also assume that the genome is circular (i.e. with periodic boundary conditions such that for example $x_{12+1} = x_1$, $x_{12+2} = x_2$ or $x_{1-1} = x_{12}$, $x_{1-2} = x_{11}$ and so on). Now suppose that an inversion is (randomly) located between loci x_3 and x_5, i.e. $s = 3$. Therefore $x_3x_4x_5 \to y_3y_4y_5$ and the mutated genome is (red color highlights the mutated loci):

$$\mathbf{y} = x_1x_2y_3y_4y_5x_6x_7x_8x_9x_{10}x_{11}x_{12},\ \forall x_i, y_i \in \{0,1\}.$$

For K = 0 (no epistatic interaction), it is trivial that only three fitness contributions per loci would change.

$$\begin{aligned} f_3(x_3) &\to f_3(y_3), \\ f_4(x_4) &\to f_4(y_4), \\ f_5(x_5) &\to f_5(y_5). \end{aligned}$$

Therefore, the fitness variation of the mutated genome $\mathbf{y}$ depends on the size of the mutation s.

Here are three examples of the parameter K for an adjacent neighborhood (let us remember that for $K \neq 0$, the fitness contribution at the ith locus depends upon itself and on K other loci, and that, for the adjacent neighborhood, these K loci are located just downstream of the focal locus):

1. For $K = 1$, only the following four local fitness contributions change:

$$\begin{aligned} f_2(x_2, x_3) &\to f_2(x_2, y_3), \\ f_3(x_3, x_4) &\to f_3(y_3, y_4), \\ f_4(x_4, x_5) &\to f_4(y_4, y_5), \\ f_5(x_5, x_6) &\to f_5(y_5, x_6), \end{aligned}$$

   Note that not only do loci within the inversion contribute, but also, as a consequence of the epistatic interaction pattern, two other loci, x_1 and x_2, outside the mutated segment of the genome also contribute to the fitness of the mutant genome $f(\mathbf{y})$ (x_1 because of epistatic effect of x_3 on this locus, x_2 because of its own epistatic effect on x_3 and x_4).

1. For $K = 2$, five local fitness contributions change

$$\begin{aligned} f_1(x_1, x_2, x_3) &\rightarrow f_1(x_1, x_2, y_3), \\ f_2(x_2, x_3, x_4) &\rightarrow f_2(x_2, y_3, y_4), \\ f_3(x_3, x_4, x_5) &\rightarrow f_3(y_3, y_4, y_5), \\ f_4(x_4, x_5, x_6) &\rightarrow f_4(y_4, y_5, x_6), \\ f_5(x_5, x_6, x_7) &\rightarrow f_5(y_5, x_6, x_7). \end{aligned}$$

2. For $K = 3$, we can verify that

$$\begin{aligned} f_{12}(x_{12}, x_1, x_2, x_3) &\rightarrow f_{12}(x_{12}, x_1, x_2, y_3), \\ f_1(x_1, x_2, x_3, x_4) &\rightarrow f_1(x_1, x_2, y_3, y_4), \\ f_2(x_2, x_3, x_4, x_5) &\rightarrow f_2(x_2, y_3, y_4, y_5), \\ f_3(x_3, x_4, x_5, x_6) &\rightarrow f_3(y_3, y_4, y_5, x_6), \\ f_4(x_4, x_5, x_6, x_7) &\rightarrow f_4(y_4, y_5, x_6, x_7), \\ f_5(x_5, x_6, x_7, x_8) &\rightarrow f_5(y_5, x_6, x_7, x_8). \end{aligned}$$

2. For $K = 4$, we can verify that

$$\begin{aligned} f_{11}(x_{11}, x_{12}, x_1, x_2, x_3) &\rightarrow f_{11}(x_{11}, x_{12}, x_1, x_2, y_3), \\ f_{12}(x_{12}, x_1, x_2, x_3, x_4) &\rightarrow f_{12}(x_{12}, x_1, x_2, y_3, y_4), \\ f_1(x_1, x_2, x_3, x_4, x_5) &\rightarrow f_1(x_1, x_2, y_3, y_4, y_5), \\ f_2(x_2, x_3, x_4, x_5, x_6) &\rightarrow f_2(x_2, y_3, y_4, y_5, x_6), \\ f_3(x_3, x_4, x_5, x_6, x_7) &\rightarrow f_3(y_3, y_4, y_5, x_6, x_7), \\ f_4(x_4, x_5, x_6, x_7, x_8) &\rightarrow f_4(y_4, y_5, x_6, x_7, x_8), \\ f_5(x_5, x_6, x_7, x_8, x_9) &\rightarrow f_5(y_5, x_6, x_7, x_8, x_9). \end{aligned}$$

   In this case the four loci x_{11}, x_{12}, x_1 and x_2 are outside the mutated segment, and the total number of fitness contribution per locus that were modified by the mutation is 7.

From the above examples we can infer that, with an adjacent epistatic neighbourhood, the number of loci contributing to the variation of fitness after an inversion of size s is equal to $K + s$. On the other hand, if the epistatic links are randomly assigned according to a uniformly distributed probability, the pattern of epistatic interaction becomes more complex. For example, the mutated base y_3 can change the fitness value of any f_i (the fitness contribution of locus i). Hence, for small values of K the number of fitness contributions that will be changed by an inversion is of the order of $K \times s$ which is much higher than for adjacent epistasis $(K + s)$.

Given that, when close to a peak, in average the more fitness values are changed, the less correlated the fitness of the genome after the mutation will be compared to the genome before the mutation, the worst the mutation (because the values of f_i are drawn in a random uniform distribution between 0 and 1 and the total fitness is greater than 0.5). This explains why, for low K values, the roughness (local standard deviation of the

fitness values) of the landscape if different for the two different epistatic neighbourhood (see Fig 5). Now, of course, the maximum number of locus contributing the fitness change after an inversion is N. Hence, for high K values, this number saturates and the difference between the two neighbourhood vanishes.

Chapter V

Innovation in viruses: fitness valley crossing, neutral landscapes, or just duplications?

Paul Banse[1], Santiago Elena[2] and Guillaume Beslon[1]

[1]Université de Lyon, INSA-Lyon, Inria, CNRS, Université Claude Bernard Lyon 1, ECL, Université Lumière Lyon 2, LIRIS UMR5205, Lyon, F-69621, France
[1]Consejo Superior de Investigaciones Cientificas (CSIC) - Universitat de València (UV)

This article is in preparation.

Abstract

Viruses are known to evolve by bursts, which are often triggered by exogenous factors such as environmental changes, antiviral therapies or spill-overs from reservoirs into novel host species. However, other types of events have been suggested to be able to trigger evolutionary burst: either fitness valley crossing or a neutral exploration of a fitness plateau until an escape mutant is found on a neutral ridge. In order to investigate the importance of these different causes of evolutionary burst, we use a simulation software to perform massive evolution experiments of viral-like genomes. We tested two conditions: after an "environmental" change or in constant conditions, this latter situation guaranteeing the absence of an exogenous triggering factor. As expected, an environmental change is almost systematically followed by an evolutionary burst. However, we show that bursts also occur, although much less frequently, in constant conditions. We analyze how many of these latter bursts are triggered by deleterious, neutral or beneficial mutations and we show that while bursts can occasionally be triggered by valley crossing or traveling along neutral ridge walking, many of them were triggered by chromosomal rearrangements, and in particular segmental duplications. Our results suggest that the difference in combinatorics between the different mutation types leads to punctuated evolutionary dynamics, with long periods of stasis occasionally interrupted by short periods of rapid evolution, akin to what is observed in virus evolution.

1 Introduction

Viral genomes are among the shortest of all genomes. They are dense and compact genomes with very high mutation rates (Belshaw *et al.*, 2008). They also undergo intense selection due to their parasitic nature and competition with the host's immune system. The evolution of viral populations has been shown to have complex dynamics, occurring in unpredictable bursts of mutation fixation (Bedford *et al.*, 2011). This dynamics, which can lead to major outbreaks, pose a threat to human health and jeopardize the sustainability of the food supply. It is therefore essential to understand these bursts in order to manage, prevent or predict them. However, there is no consensus on their origin, with numerous hypotheses having been proposed in the literature. These hypotheses can be classified into two distinct groups, depending on whether the bursts have an exogenous or endogenous origin.

Exogenous causes correspond to events that occur in the environment of the virus. In this view, the evolutionary bursts are triggered by changes that occur outside of the viral population like changes in the genetic composition of the host populations (Morley et Turner, 2017; Turner et Elena, 2000), hosts' immune response developed against previously common variants, (Elde *et al.*, 2012), population bottleneck due to random transmission events of very few particles (Escarmís *et al.*, 2006), or new niche colonization (Stapleford *et al.*, 2016). Other types of exogenous events can include interaction with other viruses, including the availability of new genetic material through genetic recombination. Indeed, this phenomenon has been observed empirically (Parra, 2019), and theorized as an efficient pathway to new phenotypes. (Wagner, 2011; Crona, 2018).

By opposition, endogenous causes do not invoke such events. In this view, bursts find their origin within the viral population and have an intrinsic evolutionary cause that can be explained using the *fitness landscape* and *valley crossing* metaphors. In the fitness landscape metaphor, initially proposed by Wright (1932), the population is viewed as evolving on a map, where horizontal coordinates represent the phenotypic traits and altitude represents fitness. On this map, the population evolves by fitness increases through mutations until a local optimum is reached. The population is then stuck on this optimum because most mutants are deleterious, hence filtered out by purifying selection, resulting in an evolutionary stasis. If there exists evolutionary mechanisms allowing the population to eventually leave this local optimum and cross a fitness valley, it can fix new mutations, comparatively triggering an evolutionary burst.

By definition, valley crossing theories are conceptualized on a stable fitness landscapes (Wright, 1932) in the absence of any exogenous event. Valley crossing has been a matter of debates in the evolutionary community for decades. Two main mechanisms leading to valley crossing have been proposed in the literature. However, there is no scientific consensus on which one is the most common (Østman et Adami, 2014).

The first mechanism is obviously to descend the slope: one or more deleterious mutations occur, which are not immediately filtered out by selection, leading part of the population to the depth of the valley where multiple alternative evolutionary paths are now available, including some giving access to a new peak of higher fitness. The likelihood of such valley crossing depends on many parameters, including the population size and structure, the mutation rate and the depth of the fitness valley (Kessinger et Cleve, 2018). Various mechanisms could ease the process by lowering the strength of selection

such as population structure (Wright, 1932), hitchhiking of deleterious mutations (Hill, 2020), phenotypic stochasticity (Van Egeren *et al.*, 2018) or stochastic tunneling (Iwasa *et al.*, 2004).

The second mechanism is the ridge line: several authors have pointed out that the importance given of fitness valleys is a direct consequence of the three-dimensional representation of adaptive landscapes (Gavrilets, 1997). Indeed, in a high-dimensional space, many neutral paths are likely to connect fitness "peaks" to each other. In this view, the population can spread and wander within a neutral plateau until a fitter genome is found (Wilke, 2001). Quasi-neutral landscape that allow cryptic deleterious mutations have also been proposed (Masel, 2006) and a general view of this phenomenon, called epochal evolution, was proposed by Crutchfield (2003).

Both exogenous and endogenous events are mentioned in the literature, although exogenous events such as environment change are generally given greater prominence (Ispolatov *et al.*, 2019). Indeed, exogenous causes are often witnessed and are well studied experimentally (Morley et Turner, 2017). Comparatively, endogenous events are difficult to identify *in vivo* and, by definition, cannot be experimentally triggered. This discretion generates a need for more experiments, for example by experimentally emulating specific mutations and analyzing the resulting landscape (Willemsen *et al.*, 2016), by simulating viral populations and comparing the outcome with biological data (Bedford *et al.*, 2011), or by pushing forward new models of evolution (Manrubia, 2012). Here we adopt a mix of the two latter approaches. As endogenous bursts are, by definition, stochastic events that cannot be triggered by an extrinsic change (environmental or ecological) we performed large scale *in silico* experiments — *i.e.*, numerous repetitions of long-lasted evolutionary simulations in a rigorously stable environment — in order to isolate the few populations that had undergone an endogenous evolutionary burst. To differentiate the origin of these bursts, we then compare them with those triggered by an environmental change.

As we aim at identifying the molecular origin of endogenous bursts, we used Aevol (Batut *et al.*, 2013; Liard *et al.*, 2020b), a simulator which accurately simulate DNA sequence and architecture as well as the various mutational events undergone by viral sequences. Indeed, viruses are known to evolve not only by point substitutions but also by more complex events such as insertions and deletions (collectively known as indels) or structural variation, including the generation of very short genomes that have lost all their coding capacity and that behave as parasites of the wild-type virus for their replication and encapsidation. (Bhange *et al.*, 2021; Rao *et al.*, 2021; Pita *et al.*, 2007) It has been shown that, in Aevol , a high mutation rate and a large population generate short and compact genomes (Knibbe *et al.*, 2007a). Thus, by setting population size to 4.096 individuals and mutation rates to 10^{-4} mutations per base-pair per generation (a rate that is relatively close to RNA virus mutation rates (Drake *et al.*, 1998a)), we were able to obtain virus-like genomes and to characterize their evolutionary dynamics.

In this paper, we use the Aevol framework, to perform two sets of 900 simulations each, starting from pre-evolved (hence well adapted) viral-like sequences. In the first set of simulations, we slightly modify the virus environment, triggering exogenous bursts. In the second set of simulations, we simulate the same viruses but in a strictly constant environment. As it could be expected, in the latter situation, most populations fix very few mutations with virtually no fitness gains at the end of the experiment. However, a few lineages substantially improve in the constant environment. We show that, in

both conditions, fitness improvements occur through short evolutionary bursts. We then identify the endogenous events that trigger the bursts in the second set of experiments and analyze the importance of evolutionary bursts depending on the kind of triggering events. Our results show that, in our simulations, duplication events are the most important source of evolutionary bursts, echoing empirical studies and questioning the limitations of studying evolution with models based solely on point mutations.

2 Methods

2.1 The Aevol platform

Aevol is a forward-in-time evolutionary simulation platform designed to study the evolution of genome structure (Knibbe *et al.*, 2007a; Batut *et al.*, 2013; Liard *et al.*, 2020b). It uses a low level representation of genetic information (sequence of nucleotides formalism (Hindré *et al.*, 2012)) in which most elements in the genome (coding and non-coding sequences, promoters, genes, operons, RNAs...) are being represented and evolve in number and position under the pressure of a wide variety of mutational operators, including substitutions, InDels and structural variants (Banse *et al.*, 2023). In the following paragraphs, we will give insights of the model that are relevant for this paper. A detailed description of the model can be found on http://www.aevol.fr.

Genome structure: In Aevol the genome is a binary circular double-stranded sequence containing scattered genes and intergenic non-coding sequences. This is similar to prokaryotic genomes and to many DNA viruses, and it allows to experimentally study the evolution of the genetic structure (typically the size of the genome, its coding proportion, the number, and position of the genes...) in different conditions (mutation rates, population size...).

Genotype-to-phenotype-to-fitness mapping: The genome sequence is decoded to compute the phenotype and the fitness of the individuals. The decoding is carried out in four steps directly inspired by the central dogma of molecular biology: transcription, translation, protein folding and protein-protein interactions.

Transcription: The genome is first read to search for promoters and terminators. In Aevol promoters are 22 bp-long consensus sequences while terminators are small sequence able to form hairpins akin ρ-independent terminators. The sequence between the promoter and the terminator is transcribed into an RNA, with a transcription rate depending on the number of differences between the promoter sequence and the consensus sequence.

Translation: Once all mRNAs are transcribed, Aevol searches for translation initiation signals. These are small 6bp-long consensus signals representing ribosome binding sites, followed by 4 bp downstream, by a Start codon. Once such signal is found, the corresponding open reading frame is translated into a protein sequence until a Stop codon is found on the same reading frame. Importantly, this process allows

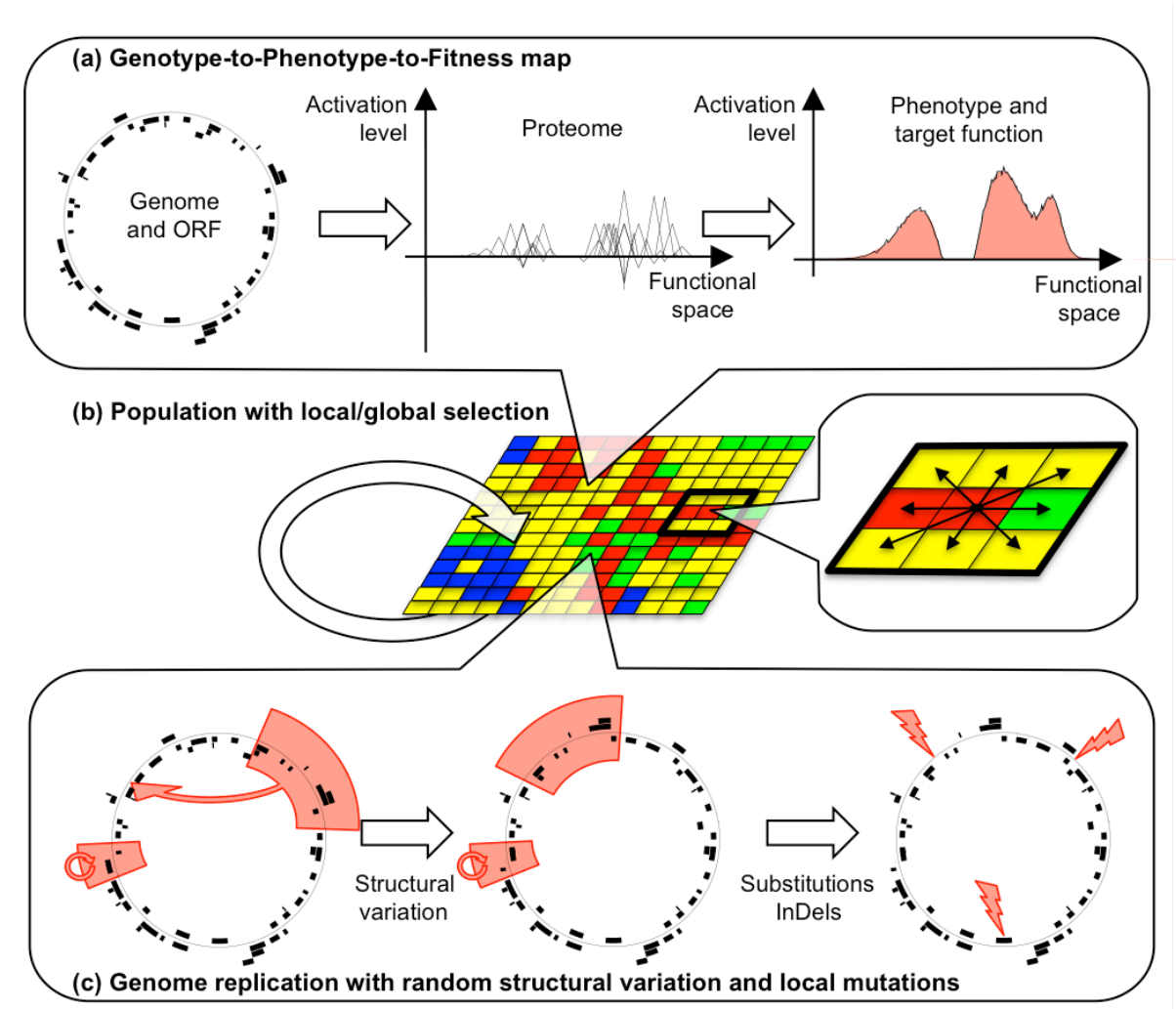


Figure V.1: Summary figure of the Aevol simulation platform. (A) The genome codes for proteins, each protein is represented as a triangle in the functional space, the ideal phenotype (in pink) is to be fitted as closely as possible by the combined effects of all proteins coded in a genome. (B) Population is represented on a grid and reproductive competition occurs locally. (C) Mutations can occur during reproduction, both chromosomal rearrangements and local mutations (substitutions and Indels).

for multiple different coding structure such as polycistronic sequences, overlapping genes, nested mRNA, that are often encountered in viral sequences (see Fig. V.2).

Folding: As the two previous points show it, the Aevol model describes quite accurately the various features of biological sequences carrying genetic information. Now, when it comes to model the functional levels, Aevol switches to an abstract, mathematical representation of biological functions, hence enabling fast computation of phenotypes and fitness for a given genomic sequence. In this representation, biological functions are represented by numerical values ϕ in the $[0, 1]$ interval. The degree to which these functions can be activated (or inhibited) can then be represented as $\mathcal{A}(\phi) \in [-1, 1]$ with $\mathcal{A}(\phi) = +1$ representing full activation and $\mathcal{A}(\phi) = -1$ full inhibition. In this formalism, a protein can be represented by three values, its main function m, the level of activation/inhibition of this function h and its pleiotropic activity w (the latter corresponding to functions close to the main one but lesser activated/inhibited, the activation/inhibition linearly decreasing with the distance up to full vanishing at distance w). In this view, once the primary sequence of the protein has been obtained through translation, it is "folded" to compute the protein's function, i.e. the three values m, w and h. Graphically, protein function can then be represented by triangular-shaped functions in a 2D space where the first axis represents the function ϕ and the second axis represents the activation level $\mathcal{A}(\phi)$ (Fig V.1.a).

Protein-Protein interactions: The whole set of proteins translated from the genome to a network with proteins sharing parts of their functions interacting together. For the sake of computational efficiency, this is modelled as a linear interaction: all protein functions (activating or inhibiting) are summed up, the resulting $[0, 1] \rightarrow [0, 1]$ function representing the phenotype of the organism (Fig. V.1.a). This phenotype is then compared with a reference function representing the ideal set of function an organism can perform in its environment. This reference function is generally defined as the weighted sum of an arbitrary number of Gaussian functions, and can be parameterized to represent more or less demanding environments (Liard *et al.*, 2020b). The reference function can also change at specific time-points to simulate exogenous events. Finally, the fitness of a given organism is computed as a function of the difference between its phenotype and the reference function, the lower the difference, the higher the fitness.

Population structure and selection: Aevol uses a spatialized Wright-Fisher reproduction model with selection and mutation (Fig. V.1.b). Individuals are distributed over a squared grid, and at each generation they compete according to their fitness values to populate neighbour grid cells at the next generation. Note that, consistently with Wright-Fisher model, Aevol does not consider any mutation as lethal. However, organisms bearing mutations inducing large phenotypic variations (e.g. loss of important genes) usually have a very low fitness and are virtually sterile.

Mutations: During replication, organisms may undergo different kinds of mutations. There are seven different kinds of mutations represented in Aevol (Fig. V.1.c). **Substitutions**, or point mutations, are the most commonly modelled mutation that changes

a nucleotide into another. Given the binary sequence, in Aevol , substitutions change a 0 into a 1 or the other way around. **Small insertions** are mutations that add a short random sequence to the genome. In all the experiments presented here, the inserted sequence has a size comprised between 1 and 6 nucleotides. **Small deletions** consist in the deletion of a small sequence of nucleotides (up to six bases). **Duplications** are genome rearrangements that copy-and-paste a part of the genome. The copied sequence and the insertion locus are chosen randomly. **Deletions** are genome rearrangements that delete a random sequence from the genome. **Translocations** are genome rearrangements that consist in the selection of a random sequence. This sequence is then extracted to form a circular plasmid that is then cut opened at a random position and inserted back in the genome at a random locus. **Inversions** can easily be modelled owing the double-stranded nature of genomes in Aevol: A random sequence is inverted, such that the first nucleotide of a strand becomes the last nucleotide of the complementary strand.

The mutation rate is defined as a probability for each nucleotide to mutate at each generation (usually between 10^{-4} and 10^{-6} mutations per nucleotide per generation). This means that the mean number of mutations undergone by an individual is directly correlated to its genome size. As a consequence, the genome size and structure are strongly influenced by mutation rates (Knibbe *et al.*, 2007a; Batut *et al.*, 2013). In particular, it has been shown that high mutation rates lead to short, dense genomes with a high proportion of overlapping genes and very few non-coding sequences, akin to viral double-stranded genomes.

2.2 Lineages tracking and analysis

Lineage tracking: During simulation, Aevol tracks the characteristics of the best individuals of the population at each generation. However, due to various classical processes in evolution (drift, clonal interference, selection for robustness...), the best organisms are not guaranteed to ultimately go to fixation. To track the mutational history of the populations, Aevol keeps a perfect record of all replication processes (including the mutations that underwent during these replications). In an asexual population and without recombination, this allows tracking the exact line of descent of the final population, to reconstruct all genomes and to extract all mutations on this line of descent. To ensure that we only consider fixed genomes and mutations (i.e. those descending directly from the coalescent), we systematically get rid of the last 5,000 generations of the lineage (see section Experimental Setup below).

As a result of the lineage tracking process, all the fitness trajectories and genome structures shown in this paper relate to individuals along the lineage of the final population. Importantly, these lineages may not be representative of the population at the given generation, but they are representative of the true evolutionary history of the final population. This implies that there is a forward dependency of our lineage analysis. For example, if a deleterious mutation is fixed in a lineage, it is very likely to be reversed or compensated in the future on the same lineage or to be hitchhiked from a favorable mutation (see eg. Fig. V.9 and related explications in the main text). Tracking such forward dependencies enables to unravel epistatic relationships through lineage analysis.

Lineage analysis and computation of evolvability: Using lineage data, we reconstructed all genomes along the line of descent of the final populations and estimated their fitness and evolvability. In the broad sense, evolvability is the ability of an individual to adapt and evolve on the long run (while fitness corresponds to instantaneous adaptation). Here we used an operational definition inspired by Woods *et al.* (2011). We measured evolvability as the expected degree to which a given genotype is likely to increase in fitness after a replication: For each genotype along the line of descent, we generated $10,000,000$ independent offsprings and evaluated their fitness in the same environment. We then computed evolvability as the fraction of beneficial offspring multiplied by their mean fitness improvement.

Along the line of descent, two successive genomes differ by a single mutation. Hence, from the fitness and evolvability of the genomes, we computed, for each type of mutation, the mean variations of fitness and evolvability it induces. For each type of mutation, we also computed the mean waiting times before and after a mutation of this type is fixed.

2.3 Fitting the fitness trajectories

To characterize the shape of fitness trajectories and classify them between stasis, gradual and saltational, we fitted them with mathematical laws, using a methodology inspired from (Wiser *et al.*, 2013, 2018). We used three different laws: *Constant (f_C)*, *Hyperbola (f_H)* and *Power law* (f_P), representing respectively stasis, bounded and open-ended fitness increase (Wiser *et al.*, 2013, 2018).

Formally, the three mathematical functions are:

$$\begin{aligned} f_C(t) &= f_{init} \\ f_H(t) &= f_{init} + \frac{a \times s(t, t_{start})}{s(t, t_{start}) + b} \\ f_P(t) &= f_{init} + (b^{-1} s(t, t_{start}) + 1)^a - 1 \end{aligned} \tag{V.1}$$

where f_{init} is the fitness of the ancestor of the lineage (i.e. $f_{init} = f(0)$), $s()$ is a delay function that enables fitting the start of a burst, its value is 0 until t is greater than t_{start} and then it increases linearly with time (i.e. $s(t, t_{start}) = \text{Max}(0, t - t_{start})$). $s()$ allows for a stasis period at the beginning the lineages and to distinguish gradual from bursty dynamics: gradual dynamics would lead to $t_{start} \approx 0$ while, in bursty evolution, the start parameter should be close to the burst start.

$f_C()$ has no parameter (f_{init} being equal to the fitness of the ancestor) while $f_H()$ and $f_P()$ have three parameters (a, b and t_{start}). Similarly to (Wiser *et al.*, 2018), we fit these curves with the least square method using the `lmfit` python package and use the BIC criterion to select the best function, the preferred model being the one with the lowest BIC value. The curve fitting is performed on the list of point where the lineage fitness curve changed (i.e. one generation before and one generation after each beneficial or deleterious mutation fixed in the lineage). In the case of $f_H()$ and $f_P()$, the initial values of the parameters have been set as followed. b, being homogeneous to a time, has been set to $1,000$. Given b, the a parameter is initialized such as the curve reaches the maximum fitness value reached by the lineage at $t = 25,000$ generations (the end of the simulation). Finally, t_{start} has been initialized at the generation of the first mutation fixed in the lineage.

In the test whether saltational evolution could fit our data, we tested a fourth function: a sum of hyperbola $f_{\Sigma H}()$. The rational is that open-ended evolution could correspond to power-law fitness curves as proposed by Wiser *et al.* (2013) but also to successive hyperbolas. The former corresponds to open-ended evolution with diminishing return epistasis, and the latter corresponds to open-ended evolution with saltational dynamics. Formally, we define $f_{\Sigma H}(t)$ as:

$$f_{\Sigma H}(t) = f_{init} + \sum_{i=0}^{i<n} \left(\frac{a_i \times s(t, t_{start_i})}{s(t, t_{start_i}) + b_i} \right) \tag{V.2}$$

$f_{\Sigma H}(t)$ has $3 \times n$ parameters. The curve fitting is performed iteratively: for all simulations, we fit the fitness curve with no hyperbola ($n = 1$) and computed the BIC value. We then fit it with a sum of two hyperbolas ($n = 2$), if it fits better than one hyperbola (i.e. the BIC value decreases), we continue with $n = 3$, and so on until the BIC value stops decreasing. The best fitting sum of hyperbola is the penultimate one. This allows avoiding the issue of multiple testing by only comparing one model to another. Note that the lineage fitness curve being a step-wise function, it could be perfectly fitted with a sum of N hyperbola (N being the number of fixed non-neutral mutations). In order to avoid this caveat, we added constraints to the hyperbolas parameters: the a_i parameter should be at least a fifth of the total fitness increase.

2.4 Identifying peak-shift and key innovations

In order to identify the precise events opening paths to a new fitness peak (i.e. peak-shift triggering events), we need a formal definition of fitness peaks that allows exact identification of the first mutant leaving the peak and whose offsprings will eventually invade the whole population. Indeed, the intuitive notion of mutational burst corresponds to the fixation of several beneficial mutations in a short period of time, which is not accurate enough for such identification as it would require an analysis of the density of mutation and thus cannot be precisely linked to a specific triggering event.

We classically define a fitness peak (or a fitness plateau) has a point (or a set of points) in the fitness landscape where no genotype with a higher fitness is accessible through a single point mutation (Wright, 1932).

For each genome along the line of descent of the final population, we performed all possible point-mutations and computed the fitness of the corresponding mutants. If none of these mutants have greater fitness, the focal genome is on a "local peak" of the fitness landscape (peak being here understood in a broad sense, meaning that a fitness plateau is a peak with neutral mutations available).

Having identified the genomes corresponding to fitness peaks, a **peak shift** can easily be defined as a sequence of mutations between two different peaks. We here consider that two peaks are different if their fitness values are different. This avoids mistaking reversed deleterious mutations for peak shifts. Following (Erwin, 2017; Hochberg *et al.*, 2017), the first mutation of a peak shift (ie. the mutation for which the non-mutate genome is on a peak and the mutant is shifting to a new peak) will be called a **key innovation**. Hence, a key innovation is an endogenous event that triggers a peak shift.

Using this method we identified all peak shifts in our simulations and isolated all key innovations. We then classified the peak shifts according to the characteristics of the key

innovations, deleterious mutations corresponding to fitness valley crossing and neutral mutations corresponding to travelling along fitness ridges.

It is important to note that our method does not rely on any arbitrary threshold, and is not dependent on mutation density or on fitness difference between two peaks. Hence, some peak-shifts can be very short and contain few mutations, and thus hardly corresponding to mutational bursts. To relate peak-shifts and mutational bursts, we computed the number of favorable mutations and the fitness difference between the pre- and post- peaks. We then observed which types of key innovations are the most likely to trigger large peak-shifts.

2.5 Experimental setup

Wild-Type evolution: We first used Aevol to evolve 30 populations of $4,096$ individuals with a mutation rate of 10^{-4}mutation.bp^{-1}.generation^{-1} and the target function containing two different subfunctions at positions $x_1 = 0.33$ and $x_2 = 0.66$ in the functional domain of the model (see figure V.1). These parameters were chosen to model the evolution of viral populations characterized by high mutation rates, high population size and limited biological functionality (Belshaw *et al.*, 2008; Frost *et al.*, 2001; García-Arenal *et al.*, 2003). Indeed, previous results with the model have shown that high mutation rates and large populations lead to compact, dense genomes with a limited gene repertoire and very efficient promoters (Knibbe *et al.*, 2007a; Batut *et al.*, 2013). The 30 populations evolved for 200,000 generations in a constant environment in order to obtain "Wild-Type" populations well adapted to their environment.

Evolution in a constant environment: We duplicated the Wild-Type populations to initiate 30 replicates per Wild-Type. The resulting 900 replicates evolved in the same conditions for 30,000 generations. We reconstructed the lineages of the 900 replicates, suppressing the last 5,000 generations in order to get rid of non-fixed events (see section *Lineages tracking and analysis* above). All genomes, fixed mutational events and peak-shifts along the 25,000 generations of the lineages were then analyzed to search for endogenous mutational bursts and key innovations.

Evolution after an environmental shift: The same 30 Wild-Type populations were duplicated 30 times each, and the resulting 900 populations evolved in a new environment. To this aim, we slightly shifted the position of the first subfunction: $x_1' = 0.3285$. The environmental change has been calibrated to induce a clear but limited drop of fitness (median fitness loss: 4.1×10^{-2}). These 30 populations evolved for 30,000 generations and the 25,000 first generations of the lineage were analyzed. Since these populations are not adapted to the new environment, they are expected to witness exogenous evolutionary bursts starting at generation $t_{start} = 0$.

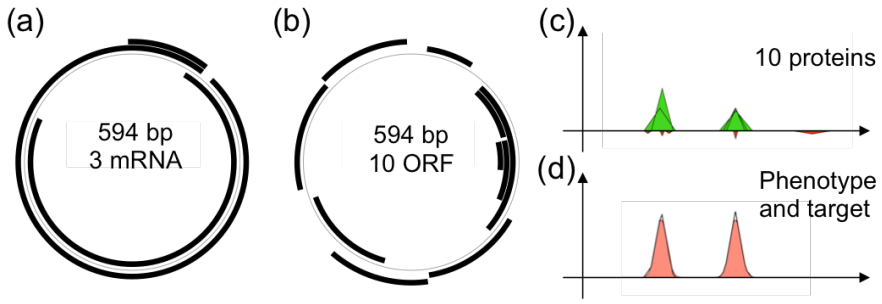


Figure V.2: Example of a Wild-Type master-sequence (Wild-Type 2, corresponding to the simulation presented on Fig. V.9). Left: genome (thin circle) and the 3 mRNAs (black segments). Center: genome and Open-Reading Frames (ORFs, black segments). Top-Right: activating (green triangle) and inhibiting (red triangle) proteins encoded by the 10 ORFs represented in the functional space of the model (see main text). Bottom-Right: Phenotype (black line) and target function (red filled area).

3 Results

3.1 Wild-Type populations:

Given the very high mutation rates (see Methods 2.5), Wild-Type populations are composed of a master sequence and a large cloud of mutants. After 200,000 generations of evolution, the master sequences are well adapted to their environment: Fitness values of the 30 master sequences range from 2.4×10^{-2} to 1.2×10^{-1} with a mean fitness of 5.6×10^{-2}.

As expected under such a mutation rate the genomes of the master sequences are short (max length: 760 bp, min: 405 bp, median: 558.5 bp) with dense sequences similar to viral sequences (Knibbe *et al.*, 2007a; Belshaw *et al.*, 2007). In average, the master sequences contained 11.13 genes with an average of 2.9 non coding base pairs. Interestingly, most mRNAs are polycistronic (mean number of gene per coding mRNA: 3.5). Figure V.2 shows an example of a wild-type genome and its phenotype.

3.2 Fitness gain in the replicates:

Starting from the wild-type populations, we ran 900 simulations in constant environment and 900 simulations after an environmental change (see Methods 2.5). Figure V.3 shows the cumulative histogram of the fitness gain during the 25,000 generations of the experiment for the two sets of simulations (Blue corresponding to the fitness gain in constant environment and orange to the fitness gain after an environmental change). All simulations evolving in the new environment recover from the initial fitness loss, at least partly (median fitness gain: 5.3×10^{-2}; Max: 2.1×10^{-1}; Min: 1.4×10^{-2}). Comparatively, the median fitness gain for the 900 simulations evolving in constant environment is 2.1×10^{-5}, most simulations showing no fitness gain as illustrated by the sharp blue peak in zero on Fig. V.3. This is coherent with Wild-Type populations being already well adapted to their environment and with the idea that environmental change is an exogenous triggering

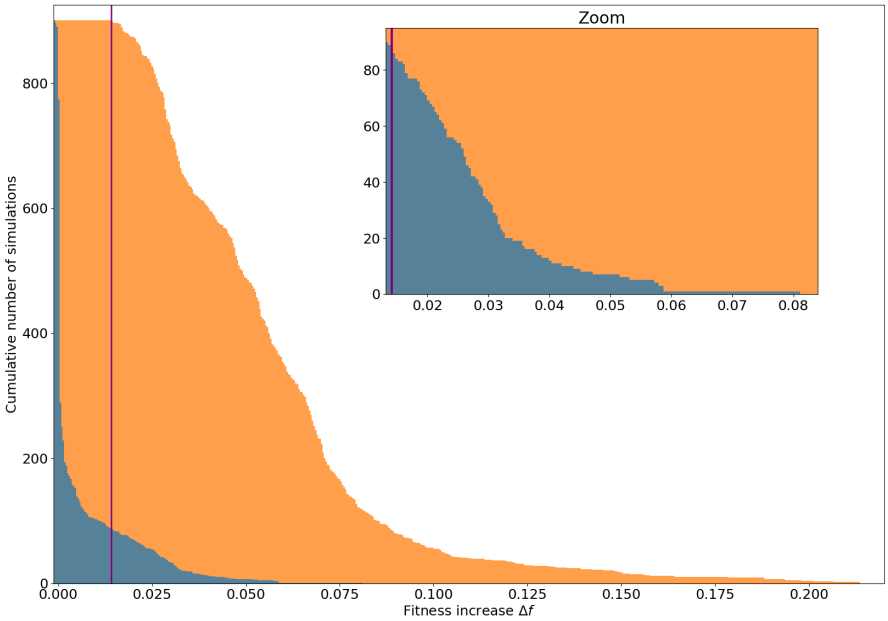


Figure V.3: Cumulative decreasing histogram of the fitness gains for the 900 simulations in a constant environment (blue) and the 900 simulations in a new environment (orange). The sharp blue peak around on zero corresponds to the majority of the simulations in constant environment showing no fitness gain ($\Delta f = 0$) or drifting around their initial fitness value through quasi-neutral mutations. On the opposite, after an environmental change, all simulations show a clear fitness gain with a minimal fitness gain after an environmental change of 1.4×10^{-2} (purple line). Interestingly, 87 simulations in constant environment (in blue) have a fitness gain greater than this limit, showing that some populations evolving in constant conditions can escape from their initial local optimum.

factor for evolutionary bursts. However, in constant environment, the distribution of the fitness gains also shows that several simulations witness fitness gains of the same order of magnitude as population adapting to a new environment (max fitness gain in constant environment: 8.1×10^{-2} (Fig. V.3). Indeed, among the 900 populations evolving in a constant environment, 87 show larger fitness gains than the worst population adapting to the new environment (purple line on Fig. V.3). this shows that some populations evolving in constant conditions can escape from their initial local optimum.

3.3 Evolutionary dynamics of the replicates:

Figures V.4 and V.5, show six examples of fitness trajectories after an environment change and in constant environment, respectively. In order to specifically observe those trajectories that experience fitness gains, we plotted the 1st, 4th, 16th and 64th bests in fitness increase which all show larger fitness gain than the worst population adapting to the new environment (from top left to bottom right). We also show an example of a random simulation (first repeat of the first wild-type) and then the worst simulation in terms of fitness gain. For all trajectories, we also show the density of mutation fixation along a sliding window of 5,000 generations. As expected, all populations adapting to a new

environment show a sharp increase in fitness and a high rate of mutation fixation at the beginning of the experiment (Fig.V.4). However, this initial exogenously triggered evolutionary burst quickly vanishes with mutation fixation rate going down to almost zero after a few thousands generations, although the best populations often show secondary bursts.

The situation is completely different in the populations evolving in a constant environment. As expected in such conditions, there is no initial evolutionary burst and most fitness trajectories are flat as illustrated by the random and worst replicates (Fig. V.5, bottom left and bottom right respectively). Nonetheless, bursts are clearly visible in the best populations, although these seem to start randomly (Fig. V.6) as exemplified by the four populations on top of Fig. V.5 that all experience sudden increases in fitness and a high rate of beneficial mutation fixation in a short time period. Remarkably, these bursts are very similar to the ones triggered by environmental change, although these are not triggered by an exogenous event.

3.4 Analysis of the evolution dynamics:

In order to quantify the visual intuition given by Fig. V.4 and V.5, we fit the fitness trajectories with either a flat function, a hyperbola or a power law (see Methods 2.3).

As expected, among the 900 simulations following an environmental change, none fit the flat function: 592 simulations are best fitted with a hyperbola, and 308 with a power law. On the opposite, in the case of a constant environment, the majority of simulations are best fitted by a flat function (505 simulations), 248 being best fitted by an hyperbola and 147 by a power law. Hence, as expected, most simulations on constant environment can be considered in evolutionary stasis.

Figure V.6 shows the distribution of the t_{start} parameter for all hyperbolas and power laws after environmental change (orange) and for a constant environment (blue). As expected, for almost all simulations starting in a new environment, $t_{start} \approx 0$ indicating that fitness starts increasing at the beginning of the simulation. This is consistent with the theory: the environmental change lowers the fitness and reorganizes the fitness landscape, giving populations access to new paths for fitness improvement, hence triggering evolutionary bursts. However, more tan half fitness trajectories (592/900) are best fitted by hyperbolas, showing the in most cases the exogenously triggered bursts quickly end, and the populations enter a new stasis phase.

This situation contrasts with the observed distribution of t_{start} for simulations evolving in a constant environment (excluding those 505 simulations best fitted by a flat function and for which t_{start} is not defined). In that case, the fitness increase starts randomly during the simulation. This suggests that these populations were experiencing evolutionary stasis at the beginning of the experiment but that an endogenous event triggered a change in dynamics, further illustrating the bursty nature of evolution in these populations.

3.5 Identifying peak-shifts:

The fitness trajectories clearly show that the evolutionary dynamics is dominated by alternating of long stasis periods and rare evolutionary bursts triggered either by exogenous

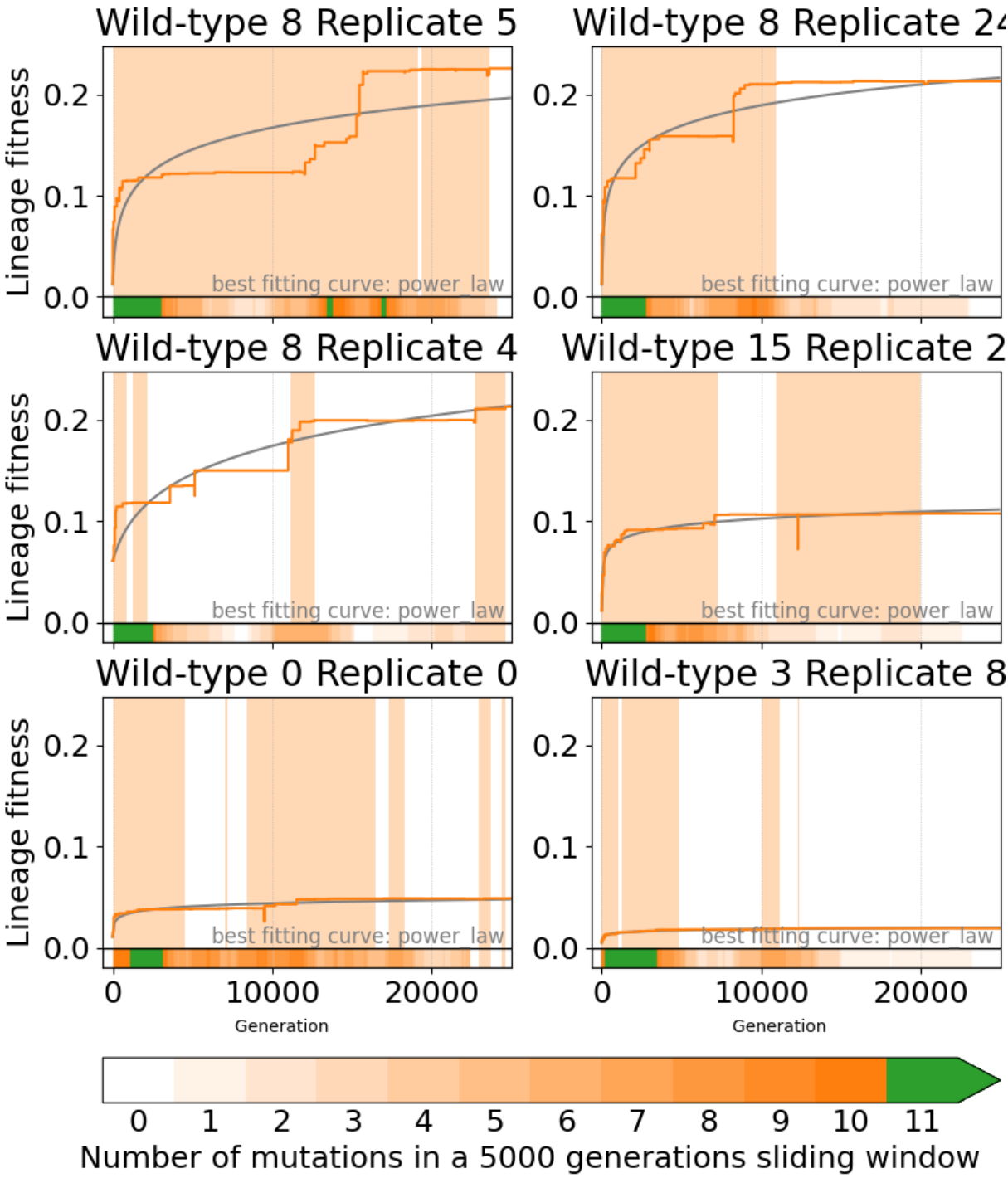


Figure V.4: Six examples of lineages from simulations after environmental change. Orange line: fitness curve of the simulation from generation 0 to generation 25 000. Grey line: the best fitting curve using the methodology described in 2. Shaded orange areas corresponds to peak-shift periods. Bottom of each graph: density of fixed beneficial mutation in a 5000 generation sliding window (see legend). The six examples are (from top left to bottom right) the best simulation (maximum fitness gain between generations 0 and 25,000), the 4th best, the 16th best, the 64th best, a random simulation (population 0) and the worst (900th best).

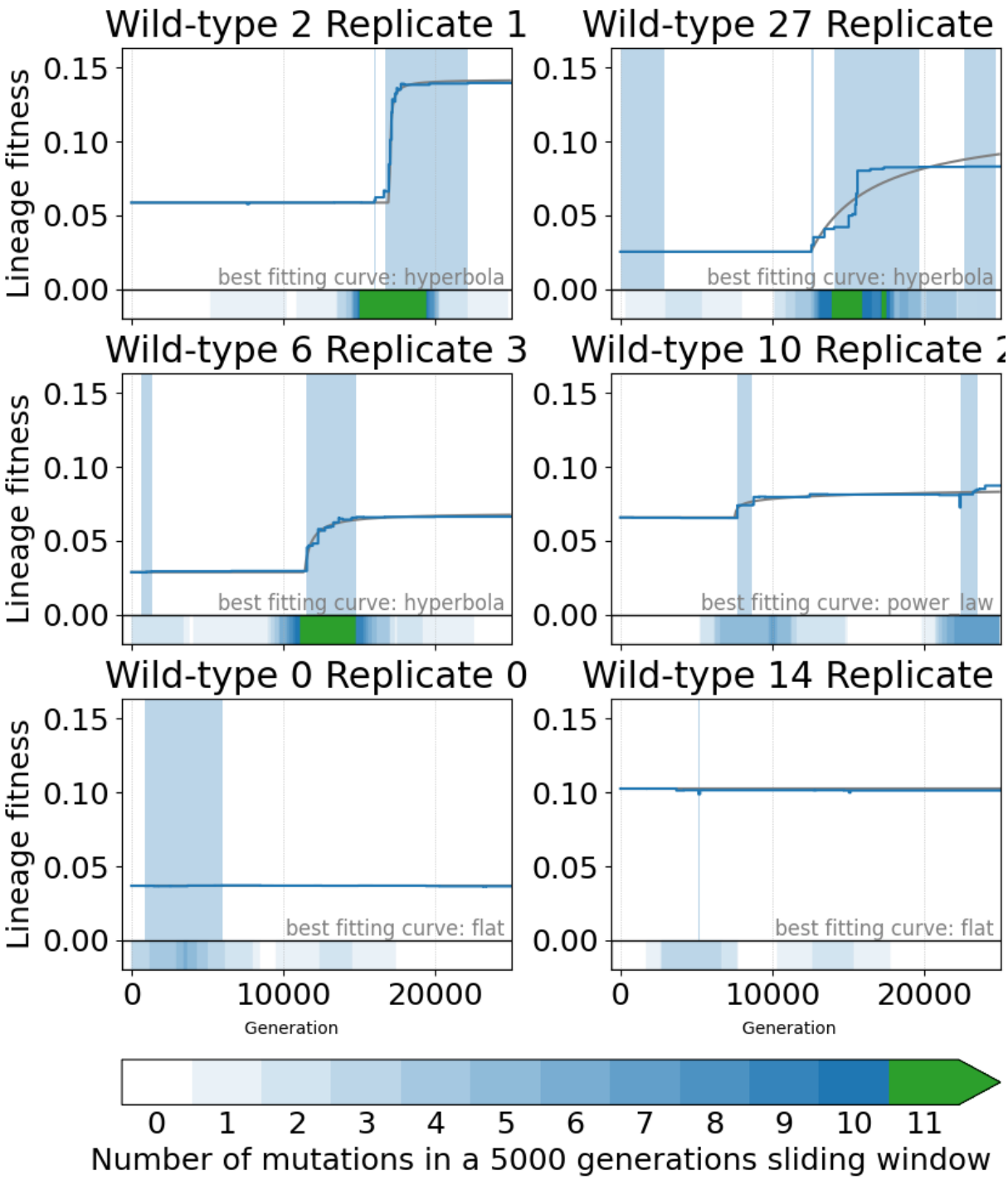


Figure V.5: Six examples of lineages from simulations in constant environment. Blue line: fitness curve of the simulation from generation 0 to generation 25 000. Grey line: the best fitting curve using the methodology described in 2. Shaded blue areas corresponds to peak-shift periods. Bottom of each graph: density of fixed beneficial mutation in a 5000 generation sliding window (see legend). The six examples are (from top left to bottom right) the best simulation (maximum fitness gain between generations 0 and 25,000), the 4th best, the 16th best, the 64th best, a random simulation (population 0) and the worst (900th best).

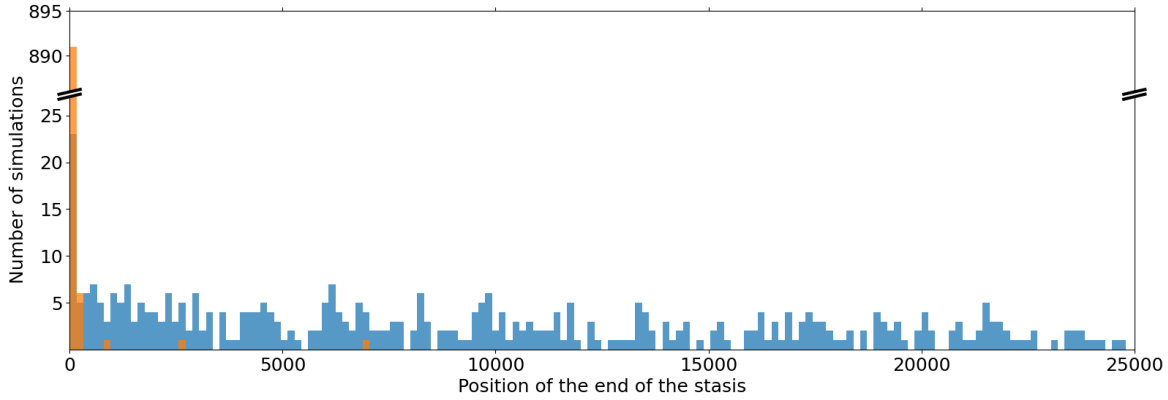


Figure V.6: Histogram of the position of the start for the simulations when fitting a hyperbola or a power law. In orange, the simulation with environmental variation, most starts are really close to the generation 0 (99.4% are within the first 1% generations of the simulation). In blue, simulation in the constant environment, the start positions are spread along all the 25 000 generations of the lineage. This distribution is neither uniform nor a power law because of edge effects.

(environment change) or endogenous events. However, so far, our results are either empirical, based on observation of the fitness trajectories and of the rate of mutation fixation (Fig. V.4 and V.5) or indirect, based on the distribution of t_{start} values in the fitness trajectories fits (Fig. 4). In order to understand what kind of event triggers the evolutionary bursts in our simulations, we need a more precise analysis of the fitness trajectories. To this aim, we linked the fitness trajectories with the structure of the fitness landscape: for all the different genomes encountered along the lineage, we tested all point mutations and measured their fitness effect. This allows us to precisely identify peak-shifts as periods of time during which at least one favorable substitution is immediately accessible from the extant genome (see Methods).

Using this approach, we identified a total of 3,631 peak-shifts in our simulations. Figures V.4 and V.5 show an example of peak-shift periods for six simulations in the two tested conditions. They show that although our method often detects "shallow peak-shifts" (i.e. peak-shifts that does not correspond to substantial fitness gains), it captures most evolutionary bursts, being they defined by the density of mutation fixation or by fitness gains. As expected, simulations which started with an environmental change experience more peak-shifts (mean: 3.16 peak-shifts per simulation with all simulations experiencing at least a peak-shift during the 25,000 generations of the experiment) than simulations in constant environment (mean: 0.87 peak-shifts per simulation). This is consistent with our previous analyses of the fitness trajectories. Indeed, most flat-fitted trajectories (373 over 505) had no peak-shift at all. On the opposite, all simulations starting with an environment variation experienced at least one peak-shift, 873 being actually already shifting to a new peak at generation 0, showing that environmental variation indeed triggers peak-shifts in a large majority of populations (by contrast, in a constant environment, only 161 populations were already shifting to a new peak at generation 0).

However, when comparing the peak shifts that occurred after an environmental change

and the ones occurring in a constant environment, both show a similar dynamic: the average peak-shift duration is 2,734 generations in the former experiments and 3,378 generation in the latter ones, both values being very short compared to the total duration of the experiment (25,000 generations) while peak-shifts concentrate respectively 80% and 96% of the total fitness gain in both kinds of experiments. Similarly, there is no difference in the number of beneficial mutations fixed during the peak-shift (on average 3.7 vs. 3.8) or in the total fitness gain between the pre- and post-peaks (on average 4.3×10^{-3} vs. 5.4×10^{-3}). These similar values back the idea that peak-shifts are similar in both kinds of experiments. They suggest that the differences lie more in the frequency of the peak-shifts than in their inner nature. Indeed, in simulations starting with an environmental variation the mean waiting time between two peak-shifts is much lower than that in a constant environment (5,177 vs. 25,358 generations respectively) as shown by figures V.4 and V.5.

3.6 Saltational dynamics:

Both our empirical observations of fitness trajectories and formal characterisation of peak-shifts point at a saltational dynamics. To further test this hypothesis, we fitted the fitness trajectories with more complex functions than the three previous ones, namely sums of n hyperbola (note that sums of n hyperbola encompass both the constant function – when $n = 0$ – and the hyperbola function – when $n = 1$ – see Methods). Compared to the power-law function that correspond to open-ended dynamics with diminishing return epistasis (Wiser *et al.*, 2013), a sum of n hyperbola would indeed correspond to saltational open-ended dynamics.

Figure V.7 shows the same fitness trajectories as figure V.4 though with the sum of n hyberbola fits. It shows that most trajectories that were originally best fitted with power-laws are now best fitted by a sum of n hyperbola. Indeed, Over the 1,800 simulations, still 504 are best fitted by a flat function (zero hyperbola) but 1043 are now best fitted by a sum of n hyperbola ($n \geq 1$) while only 253 are still best fitted by a power-law. This confirms that in our simulations the evolutionary dynamics is mostly saltational, with populations alternating between short evolutionary bursts and long periods of evolutionary stasis, the former being triggered either by exogenous events (here an environmental change) or by endogenous ones.

3.7 Triggering events:

Our formal characterization of peak-shifts allows to precisely identify the key innovations that triggered them, hence, the nature of the peak-shift. Indeed, depending on the characteristics of the triggering events, one could distinguish valley crossing (triggered by a deleterious mutation) from traveling along neutral ridges (triggered by a neutral mutation). To avoid direct and indirect effects of the initial environmental variation, we hereby focus on the 900 populations that evolved in a constant environment. Among the 787 peak-shifts observed in these experiments, we first excluded the 161 peak-shifts starting at generation zero and for which the triggering event is not formally identifiable as it has occurred in the lineage of the Wild-Type populations. Indeed, as the initial population contains a cloud of mutants, it is possible that some of these mutants already bear mu-

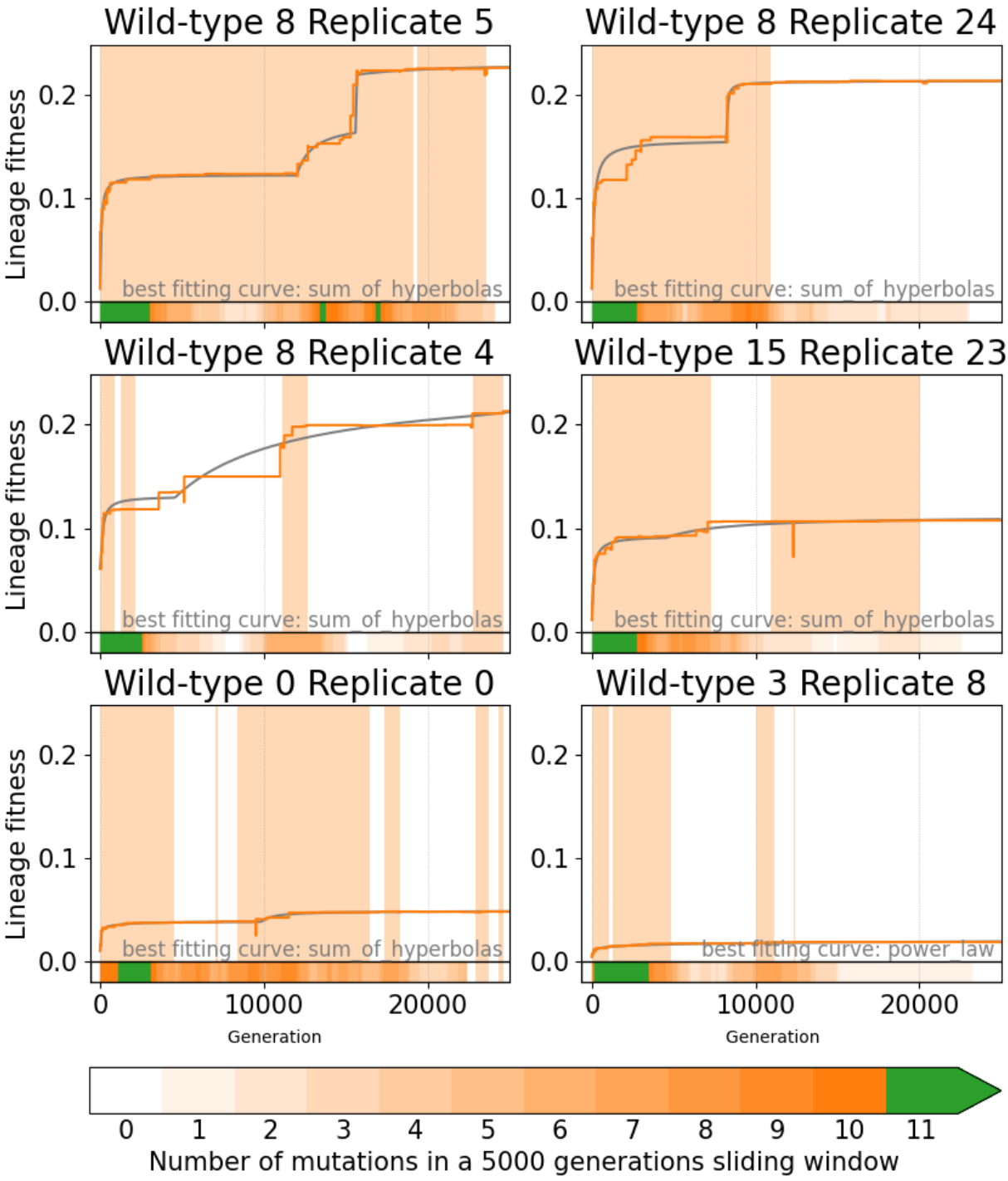


Figure V.7: Sum of n hyperbola fits on the six fitness trajectories presented on Fig. V.4 and corresponding to the best simulation, the 4th best, the 16th best, the 64th best, a random simulation (population 0) and the 900th best (all these simulations experienced an environment change at generation 0).

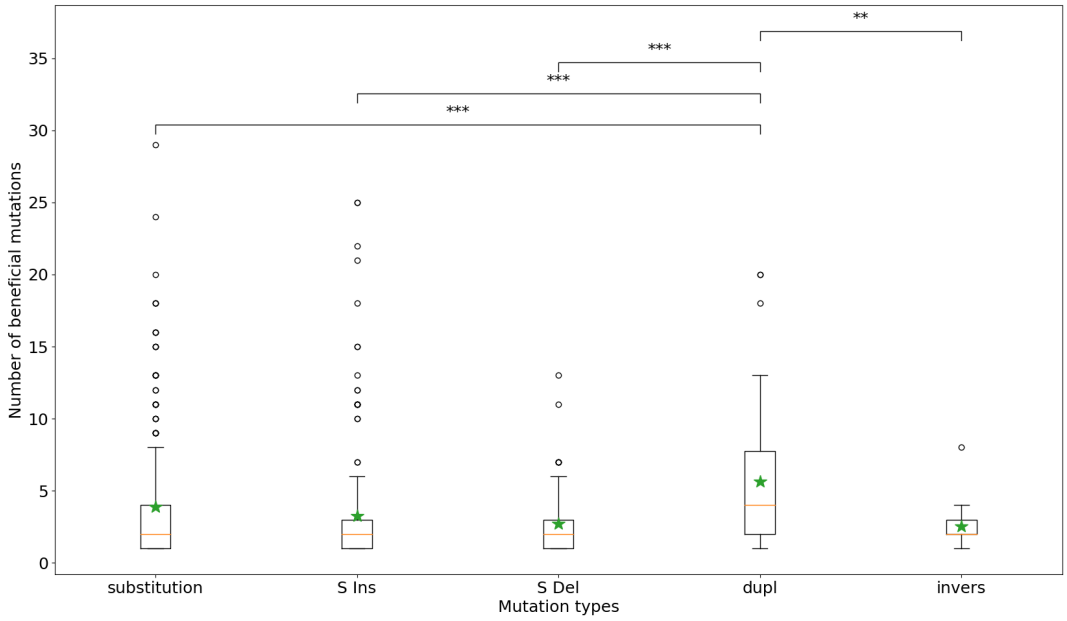
tations triggering a peak-shift. We also excluded the 7 peak-shifts starting with double mutations. Among the remaining 619 peak-shifts, 401 have been triggered by a deleterious event and 40 by a neutral event. However, surprisingly, there are 178 peak-shifts that were triggered neither by a deleterious nor by a neutral event. These were indeed triggered by beneficial events, more precisely by beneficial InDels (105 events) and beneficial rearrangements (73 events) – beneficial substitutions being impossible owning to the formal definition of peak-shifts (see Methods). Hence, our results show, first that valley crossing is much more probable than travelling along neutral ridge in our simulations, but also, and more surprisingly, that a substantial fraction of peak-shifts are triggered by beneficial events, providing these events are complex mutational events affecting more than one nucleotide at a time.

So far, we mainly analyzed peak-shifts on the basis of their temporal characteristics. However, as show by figures V.4 and V.5 show it, peak-shifts are highly variable in terms of "intensity", with "shallow peak-shifts" resulting in quasi-neutral fitness variations (e.g. fig. V.5, wild-type 0, replicate 0) to strong ones, resulting in large fitness gains and during which many mutations are fixed in a very short period of time (e.g. fig. V.5, wild-type 2, replicate 1). Indeed, not all peak-shifts correspond to mutational bursts. To further characterize key innovations, we computed, for each type of key innovation, the average fitness gain of the peak-shifts it triggered as well as the average number of favorable mutations fixed during the peak-shift. We first looked at the fitness gain depending on the sign of the mutation but find no major difference between deleterious, neutral or beneficial key innovations: the average fitness gain of a peak shift is indeed 3.74×10^{-3} when triggered by a deleterious mutation, 2.21×10^{-3} when triggered by a neutral mutation, and 5.88×10^{-3} when triggered by a beneficial mutation. However, when looking at the effect of the different types of mutation (Fig. V.8). It immediately appears that, among all peak-shifts, those triggered by segmental duplications result in fitness gains almost ten times higher than all the other ones and in the fixation of more beneficial mutations. These results show, first, that the *nature* of the peak-shift (valley-crossing or neutral landscape) is less important than the type of mutation that triggered the shift and, second, that, among the different types of mutations, segmental duplications are by far the most likely to trigger strong peak-shifts and mutational bursts.

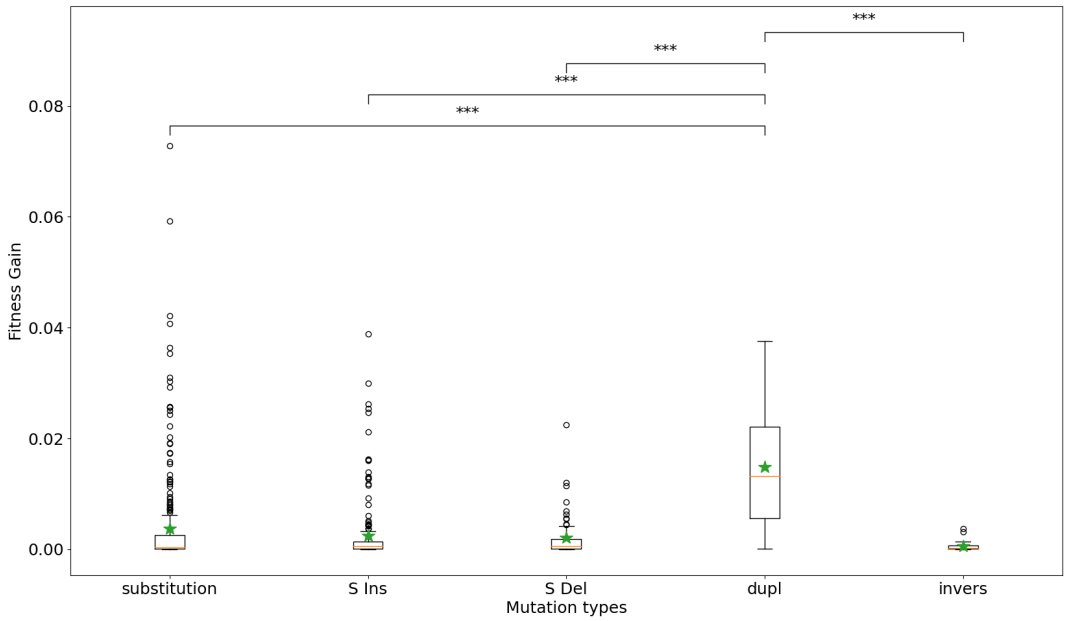
3.8 Analysis of the different types of mutation:

Previous results have shown first that the evolutionary dynamics is saltational, with most fitness gains concentrated in rare peak-shifts, and that the duplications are more likely to trigger strong peak-shifts than the other types of mutations. To better understand the contribution of the different types of mutations to the evolutionary dynamics, we analyzed the 33,598 mutations that went to fixation in the 900 simulation in constant environment. Then, for each type of mutation, we quantified the number and fraction of event that triggered a peak-shift (`Nb.trig.`), the mean fitness effect of a mutation (Δ`fitness`), the mean time since the previous fixed mutation (Δt_{pre}), the mean time before the next fixed mutation (Δt_{post}) and the mean contribution to evolvability (Δ`evolv.`, evolvability being defined as the fraction of beneficial offspring multiplied by their mean fitness improvement – see Methods).

Table V.1 summarizes the results for the seven different types of mutations. It first



(a) Number of mutations



(b) Fitness gains

Figure V.8: Comparison of the key innovations. Top: average number of favorable mutations fixed during a peak-shift for the different types of key innovations (substitutions, small insertions, small deletions, duplications and inversions) for the 619 peak-shifts occurring in a constant environment. Bottom: mean fitness gain of peak-shifts for the different types of key innovations. Large deletions and translocations have been excluded because they don't trigger enough peak-shifts (0 and 2 respectively). Mann-Whitney u-Test with Bonferoni corrections for multiple tests, $p_{corrected} < 2 \times 10^{-9}$. All other paired tests were non-significative ($p_{corrected} > 0.4$).

shows the huge difference in the number of fixed events. Given that the mutation rate – hence the number of spontaneous mutations – is the same for all types of events (10^{-4} mutation per base pair per generation – see Methods), this illustrates the difference in the Distribution of Fitness Effect for the different types of mutations. Indeed, the number substitutions (fixed point mutations), is one order of magnitude greater than the number of fixed small insertions or deletions (Indels). On that matter, chromosomal rearrangements shows a striking pattern with three types of rearrangements (duplications, deletions and translocations) being highly deleterious (hence hardly fixed) an the fourth one (inversions) which fixation rate is of the same order of magnitude than InDels. This is due first to unbalanced rearrangement (duplications and deletions) being significantly more deleterious than balanced ones (translocations and inversions) and second, to translocations being especially deleterious on dense genomes owing to their number of breakpoints. The situation is notably different for inversions but one has to note that small inversions can have no effect on the sequence, which increases their neutrality (Trujillo *et al.*, 2022). However, when looking at the fraction of fixed mutations that triggered a peak-shift, it immediately appears that, although rarely fixed in the lineage, duplications are much more likely to trigger a peak-shift. Indeed, 25% of the fixed duplications are key innovations, a much higher fraction than all other types of mutations. Moreover, $\Delta\texttt{Fitness}$ values show that fixed duplications are in average much more favorable than all other types of mutations, further suggesting that, despite their very low fixation rate, they make decisive direct and indirection contributions to adaptation.

The two columns $\Delta\mathtt{t}_{\mathtt{pre}}$ and $\Delta\mathtt{t}_{\mathtt{post}}$ respectively show the mean number of generations before and after a mutation of a given type until the next fixed mutation (whatever its type). Overall, the mean waiting time between two mutations is 643 generations and any large deviation of $\Delta\mathtt{t}_{\mathtt{post}}$ from this value indicates whether a specific type of mutation, when fixed, changes the evolutionary dynamics. On the opposite, any large deviation of $\Delta\mathtt{t}_{\mathtt{pre}}$ indicates that the corresponding mutation type is only fixed in a specific dynamics. Here the effect of duplications is particularly marked with $\Delta\mathtt{t}_{\mathtt{post}} = 215$ indicating a clear change of dynamic. Indeed, the fixation of a duplication triples the rate of fixation of mutations. The pattern in inverted for large deletions ($\Delta\mathtt{t}_{\mathtt{pre}} = 255$), indicating that large deletions are only likely to be fixed when occurring during an evolutionary burst. Note the very specific pattern of inversions, indicating that inversions are often fixed during long stasis periods, probably because, as explained above, they may have no effect at all on the sequence. Finally, when computing the mean effect of the mutation types on evolvability (column $\Delta\texttt{Evolv.}$ on table V.1), the effect of duplications appears even more marked. Indeed, duplications are the only type of mutational events that markedly increase evolvability. All other types of events either have a negligible contribution (substitutions, translocations, inversions) or reduce evolvability (InDels).

These results show that, although their rate of fixation is extremely low (261 duplications fixed for a total of 900 experiments lasting 25,000 generations each), segmental duplications, when fixed, creates new evolutionary opportunities and changes radically the dynamics of evolution. It is thus tempting to invoke gene duplication to explain these observations (Ohno, 1970b). We thus analyzed the genetic content of the genomes before and after each of the 261 duplications fixed in our experiments. Results show that, out of these 261 events, only 11 were gene duplications. However, 126 duplications resulted in the creation of a new functional gene (by duplicating segments of existing genes). Finally,

Mutation type	Nb.fixed	Nb.Trig.	ΔFitness	Δt_{pre}	Δt_{post}	ΔEvolv.
Point mutation	17387	274 (1.5%)	5.2×10^{-5}	642	646	-3.0×10^{-8}
Small insertion	4737	196 (4.1%)	1.0×10^{-4}	593	411	-4.0×10^{-7}
Small deletion	5117	58 (1.1%)	1.6×10^{-4}	404	590	-2.7×10^{-7}
Duplication	261	66 (25%)	3.4×10^{-3}	566	215	1.4×10^{-5}
Deletion	200	0 (0%)	9.1×10^{-5}	255	412	-7.4×10^{-8}
Translocation	130	2 (1.5%)	1.9×10^{-4}	773	632	-4.9×10^{-8}
Inversion	5766	23 (0.4%)	1.4×10^{-5}	914	900	2.0×10^{-8}

Table V.1: Properties of the 33,598 mutations fixed in the 900 simulations in a constant environment. For each type of mutations, columns show (from left to right) the number of occurrences in the 900 lineages, the number and fraction of mutations that triggered peak-shifts (i.e. key innovations), the average effect on fitness, the average number of generation since the last fixed mutation, the average number of generations until the next fixed mutation and the average contribution to evolvability (see Methods 2.2).

the remaining 124 events result in alteration of existing genes without adding a new ORF to the sequence.

3.9 Illustration: Wild-Type 2, experiment 1:

In order to illustrate the results presented above, Fig V.9 details the evolutionary dynamics of a specific experiment: the experiment 1 starting from the wild-type 2 and evolving in a constant environment (highest fitness gain among the 900 simulations in constant environment). Top panel shows the fitness trajectory as well as the local density of mutation fixation (both being identical to the top-left panel on figure V.5), middle and bottom panels respectively show the mutations fixed during the experiment (identified by their types and loci, the blue line corresponding to the size of the genome) and the variations of evolvability during the 25,000 generations of the experiment. Shaded areas on the three panels show the two peak-shift periods detected on this simulation.

As most simulations in a constant environment, this experiment starts in a stasis period. During this period, neutral mutations are occasionally fixed, most of them being short inversions (× grey symbols) or substitutions and InDels (+ grey symbols) at the rare neutral locus (especially at loci 5 and 7). Note the deleterious substitution at generation 7610 (blue + symbol) that is almost immediately reverted (red + symbol at generation 7731). The fixation of the deleterious mutation triggers a strong increase of evolvability as a favorable mutation is now available at short mutational distance. However, after the exact reversion of this mutation evolvability drops back to its initial value. Note that although meanwhile a favorable mutation is accessible, this episode is not considered as a fitness valley crossing because the deleterious mutation is exactly reverted.

The initial stasis period ends at generation 16,003 when the lineage enters a bursting period (as shown by the density of mutation fixation on the top panel). Shaded areas show the two peak-shifts periods. The first one is weak, stopping after 44 generations and only one favorable mutation. This first peak-shift is triggered by a short beneficial duplication of a segment of 46 bp (plain red segment), inserted close to the original

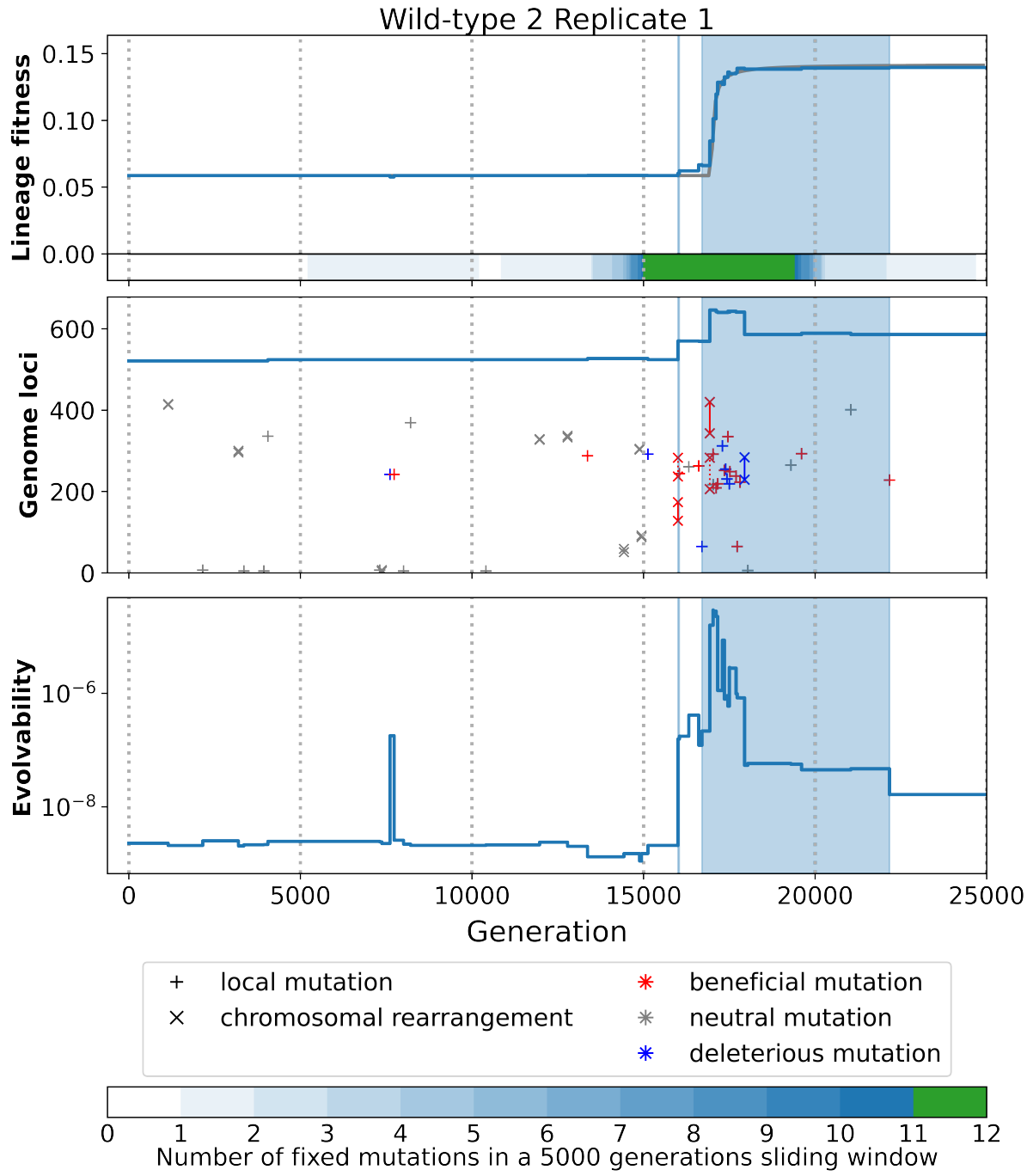


Figure V.9: Example on constant environment of analysis to find the key innovations. On **top** is the fitness of the lineage, as it was shown in Fig.V.5 with the density of mutations. In the **middle part**: the genome of each individual of the lineage, with the mutations fixed during evolution: plus signs correspond to local mutations: point mutations and indels, times signs correspond to chromosomal rearrangement: duplication, large deletions, inversions, and translocations. Note that for chromosomal rearrangement, multiples crosses indicate the position of the sequences of interest. Deep blue crosses: deleterious mutations, red crosses: beneficial mutations, gray crosses: neutral mutations. On the **lower part** evolvability computed as explained in 2.2. The shaded area in **both plots** correspond to the generations where a peak shift is identified with the method in 2.4. This area corresponds to the increase in mutation density, in evolvability and in fitness.

segment (dashed red segment). This duplication has a marginal effect on fitness, but it strongly increases the evolvability of the sequence (notice the log scale for evolvability) while initiating the peak-shift. This first peak-shift ends by the fixation of a beneficial substitution within the inserted sequence (red + symbol). A short pseudo-stasis period follows, during which two mutations are fixed within the inserted sequence: a neutral substitution (grey + symbol) and a favorable short deletion (red + symbol). Although this period is not detected as a peak-shift (because no favorable substitution is directly available), it is characterized by a high evolvability value, showing that the evolutionary potential created by the duplication is not yet exhausted. Indeed, a second peak shift immediately follows, which is more pronounced, starting at generation $16,702$, lasting $5,467$ generations and containing 11 favorable mutations and 6 deleterious ones. The triggering event of the second peak-shift is a deleterious substitution (blue + symbol) although the pattern is actually more complex. Indeed, this substitution has a marginal effect on fitness and on evolvability, and it is reverted at generation $17,729$. meanwhile, the initial duplication has been followed by a second one (which inserted a new segment nearby the first insertion), resulting in a second step of evolvability. Indeed, altogether, both duplications had increased evolvability by four orders of magnitude (from 2.1×10^{-9} to 2.2×10^{-5}). This creates a mutational hot-spot in the vicinity of the inserted segments in which several mutations go to fixation, either due to selection (favorable mutations) or to hitchhiking (deleterious ones, including one large deleterious deletion that strongly reduces the evolutionary potential – plain blue segment). Finally, two favorable mutations (one small insertion and one substitution, both in the inserted segment) end the sequence, further reducing evolvability and switch to a new stasis period lasting $2,831$ generations and during which not even a single mutation went to fixation.

Although mostly illustrative, this example clearly shows how duplications interact with mutations of different types to generate saltational dynamics. Indeed, rare fixation of duplications add new genetic material to the sequence. On this new genetic material, further events of various types (including substitutions, InDels but also other rearrangements) can play, resulting in a burst of mutation fixation driven by selection and hitch-hiking.

4 Discussion

In this study we used the Aevol simulation platform to study the evolutionary dynamics of viral-like genomes and to investigate whether the mutational bursts that are frequently observed in viral strains could be of endogenous origin or whether they are always triggered by exogenous causes.

Our results show that, although environmental variations indeed almost systematically trigger bursts, saltational dynamics is observed in the simulation even in the absence of such variation. Moreover, using a systematic exploration of the mutational neighborhood of the genomes along the line of descent, we have been able to relate mutational bursts to peak-shifts on the fitness landscape, even though many of these peak-shifts appear to be "shallow peak-shifts" during which both the fitness gain and the mutational fixation remain low. Having precisely identified the peak-shifts, we have been able to identify the type of mutational event that triggered them (the "key-innovations"). Surprisingly, only a small fraction of these key-innovations were neutral events, showing that, despite

the very large number of dimensions of the search space, moving along neutral ridge is not frequent and that fitness valley crossing is still the main peak-shift mechanism. However, a striking result is the peak-shift is frequently triggered by favorable events, mainly favorable insertions and favorable duplications. Finally, by analyzing the types of mutations triggering peak-shifts, we have been able to show that segmental duplication, although rarely fixed in the lineage, are triggering the most powerful peak-shifts, that is to say the peak-shifts corresponding to mutational bursts. This is confirmed by an exhaustive analysis of all mutational events fixed along the lineages, as the fixation of a duplication appears to increase the rate of fixation of all types of mutations and to increase the evolvability of the lineage.

Our results highlight the role of the complex interplay between different types of mutations. They show that even in conditions where the mutation rates are perfectly stable through time in the population, this interplay can result in a saltational dynamics in the rate of fixation of mutations. This dynamics is characterized by short periods of intense mutation fixation separated by long stasis periods. We propose that this dynamics is due to the difference in combinatorics of the various types of mutations. Combinatorics here refers to the number of different genomes that are accessible through a specific type of mutation. In other words, combinatorics quantifies the size of the mutational neighborhood of a genome for a given type of mutation. For instance, it is easy to show that, from an ancestral genome of length L, substitutions lead to a combinatorics of $L * 3$ (L possible locus times 3 possible substitutions per locus). To the best of our knowledge, combinatorics is not considered as a key parameter in evolutionary biology. However, its importance becomes evident when we compare this simple formula with the combinatorics of more complex mutations. Indeed, considering, for instance, a segmental deletion in which a random subsequence of the genome is suppressed. The combinatorics of this mutation is now $L * (L - 1)$: L possible locus for the beginning of the deleted segment times $L - 1$ possible locus for its end (for a circular genome). This shows that the mutational neighborhood of this event grows quadratically with the size of the genome, hence surpassing by orders of magnitude the neighborhood of substitutions, even for small genomes like viral ones. Table V.2 shows the combinatorics of the main types of mutations for a genome of size L as well as the rationales of the computation.

Importantly combinatorics relates to the number of different mutational events of a given type that are possible for a given genome. It is directly related to the size of the mutational neighborhood of a given genome although it can be larger as since different mutational events could possibly lead to the same genome. Given the large variations of combinatorics (e.g. from $3,000$ for substitutions to $999,900,000$ for segmental duplications for a genome of length $L = 1,000$ bp), it immediately follows that the time needed to explore a substantial fraction of the mutational neighborhood varies by orders of magnitudes depending on the mutation type. Hence, the overall evolutionary dynamics results from the juxtaposition of different mutational processes, each with its own timescale. However, the combinatorics alone cannot explain saltational dynamics, as it does not consider the interactions between the different types of mutations. Indeed, as our results show it, some types of mutational events increase the probability of fixation of the others through an increase in the overall evolvability. Hence, the saltational dynamics observed in our simulation results from the interplay of two mechanisms: combinatorics, that leads to the juxtaposition of very different evolutionary time scales (fast ones for substitutions

Point mutations	$3L$	L possible locus; 3 possible substitutions per locus.
Small insertion (max length l)	$L * \left(\sum_{1<i<l}(4^i)\right)$	L possible locus; 4^i possible random sequences of length i; $\sum_{1<i<l}(4^i)$ possible random sequences of length lower or equal to l.
Small deletion (max length l)	$L * l$	L possible locus; l possible deletion length.
Segmental duplication	$(L * (L-1)) * L$	Two different breakpoints; one insertion point.
Tandem duplication	$(L * (L-1)) * 2$	Two different breakpoints; two possible insertion locus.
Segmental deletion	$L * (L-1)$	Two different breakpoints.
Translocation	$((L * (L-1)) * L$	Two different breakpoints; one insertion point.
Inversion	$L * (L-1)$	Two different breakpoints.

Table V.2: Combinatorics of different types of mutations. Third column gives the rationals of the computation. Note that the formulas can change depending on specific mutational mechanics.

and short InDels and slow ones for chromosomal rearrangements), and evolvability that reignite the fixation of fast events after the fixation of slow ones. Ultimately, this interplay leads to the fixation of many fast events: in our simulations, the rate of fixation of substitutions is more than 50 times higher than the rate of fixation of duplications, despite their same spontaneous rate (table V.1). But these fast events are mostly fixed after the occurrence of a slow event that opened new favorable evolutionary pathways on which they immediately rush on.

In our simulation, duplications are by far the most beneficial type of chromosomal rearrangements when it comes to triggering evolutionary bursts (Fig. V.8). A natural hypothesis would be gene duplication-divergence (Ohno, 1970b; Zhang, 2003; Gao *et al.*, 2017). However, in-depth analysis of the duplication events show that only a small fraction of fixed events correspond to gene duplication. On the opposite, most fixed duplications actually add a small segment to extant genes. Table V.1 suggests that the advantage of duplications comes both from their direct effect on fitness (on average, fixed duplications are more beneficial than any other type of mutation), and from their indirect effect of

evolvability (as duplications are the sole type of mutation that substantially increases evolvability). This suggests that, apart from combinatorics, the mechanistic properties of the different types of chromosomal rearrangements play a role in their fixation and in the burst-triggering probabilities. Classically, rearrangements are classified as *unbalanced* or *balanced* depending on whether they change the genome length (duplications and deletions) or not (translocations and inversions (Mérot *et al.*, 2020)). Our results suggest that an important factor is whether rearrangements are *conservative* or not, conservative events being those keeping the rearranged segment intact (typically duplications) while unconservative ones modify it (typically deletions and translocations). We propose that the advantage of duplications in triggering evolutionary burst comes from them being conservative and unbalanced. Being conservative limits how deleterious the events are (duplications being only harmful at their insertion point) while being unbalanced creates a strong potential for other types of mutations to modify the inserted sequence, hence the gain of evolvability and the subsequent burst of mutation fixation (see Fig. V.9 and table V.1), including substitutions, InDels but also large segmental deletions that remove parts of the inserted sequence. Hence, our results support the genomic accordion model (Elde *et al.*, 2012; Filée, 2013) but show that the accordion can play a role even in a constant environment and that it is not restricted to gene copy-number variations. Indeed, although gene duplications are frequent and play a major role in double-stranded DNA viruses (Elde *et al.*, 2012; Gao *et al.*, 2017) and have been observed in the long RNA virus Citrus tristeza virus (Kang *et al.*, 2018), they are rare in other types of viruses. However, it is worth noting that small sized duplications, similar to those we observed in our simulations, have been shown to lead to significant increase in fitness and mutation rate in hepatitis C virus (Le Guillou-Guillemette *et al.*, 2017), or in respiratory syncytial virus with a 72 nucleotides duplication (Schobel *et al.*, 2016).

Given that in our simulations 25% of fixed duplications trigger peak-shifts, it would be tempting to conclude that segmental duplications predict mutational bursts. However, it is important to keep in mind that we analyzed fixed events, hence being dependent on the survivorship bias: the fixation probability of a key innovation depends on the fitness gain of the following peak-shift that cannot be predicted. This is especially true for segmental duplication that bear a deleterious effect owing to the increase of genome size (Willemsen *et al.*, 2016). This may explain why the fraction of fixed duplications triggering peak-shift is so high and why duplications trigger higher peak-shifts: only duplications followed by large mutational bursts come to fixations by hitch-hiking the burst they triggered.

Our results open intriguing questions on the interpretation of the fitness landscape metaphor and of the related question of fitness valley crossing. Several authors have argued that the classical 2D representation of fitness landscapes is misleading and that in large dimensions, fitness landscape are likely to be "holey-landscapes" in which many neutral ridges or neutral landscapes connect fitness peaks (Gavrilets, 1997; Wilke, 2001; Crutchfield, 2003; Fragata *et al.*, 2019). Our results first show that very few key innovations are neutral mutations. Although this can be due to the specifics of the Aevol genetic code in which there is no synonymous codons (hence less neutrality), this observation suggests that neutral landscape are not an efficient way to escape local fitness peaks. Hence, our results support the findings of Chatterjee *et al.* (2014) that showed that the time needed in the neutral basin grows exponentially with the size of the genome, mechanistically limiting the evolutionary potential of neutral landscapes. Interestingly, Chatterjee

et al. (2014) also suggest that chromosomal rearrangements would allow reducing the search process to a polynomial one in the size of the duplicated sequence.

Even more interestingly, our results suggest that the fitness landscape metaphor is not only biased by the classical 2D representation but also by its mere topological representation. Indeed, the "surface" of the landscape is composed of local mutations: substitutions, and sometimes Indels. But the metaphor does not allow representing – let alone reasoning – other kind of mutational events and, in particular chromosomal rearrangement. Indeed, those would look like "jumps" over the landscape. Moreover, as combinatorics shows it, these jumps are much more numerous that possible local moves. Metaphorically speaking, every point in the landscape is an airport, only weakly connected to its neighbors by a few roads (substitutions), but connected by direct flights (rearrangements) to a very large number of other airports all over the world. As most jumps are deleterious, the local connections are those that drive short term behavior. However, we conjecture that in such a "multi-mutational fitness landscape" the number of available jumps is such that there are virtually no fitness valleys that cannot be jumped over by a chromosomal rearrangement.

Interestingly, the dynamics we observe in our simulations are very similar to the ones conjectured by Kauffman et Levin (1987). When "correlated" mutations (with small fitness effect) and "uncorrelated" mutations (whose fitness is mostly uncorrelated to the fitness of their ancestors) coexist. According to Kauffman et Levin (1987), in such a case, evolution proceeds in three phases: initially, uncorrelated mutations have the best chances to increase fitness, then on the midterm, mutations that are correlated allow fine-tuning. Finally, in the long term, the population is stuck waiting for a beneficial uncorrelated mutation to make a "jump" through the fitness landscape. However, our simulations suggest that not all rearrangements are evenly "uncorrelated" and that duplications are the most likely to make favorable jumps.

In this study, we focused on very short and dense genomes and showed that, saltational evolution similar to the one observed in viruses can emerge in a stable environment with very few assumptions on the evolutionary process. This raises the question of the evolutionary dynamics of longer, possibly less dense, sequences (e.g. prokaryotic genomes) under the mixed effect of fast and slow mutations. Indeed, alternations of burst and stasis is mainly due to the difference in tempo between fast fixation of beneficial local mutations and long waiting times for key innovations. In this view, increasing the genome size will both increase the time needed to reach a peak and the time needed to explore the duplication neighborhood. However, combinatorics shows that the both will not grow at the same rates (table V.2). A direct consequence is that, under similar evolutionary conditions (mutation rates, population size...) the evolutionary dynamics is likely to be very different depending on genome size. Exploring the interplay of the different types of mutations for various genome structures constitutes a very promizing perspective of our work. In particular, we conjecture that the saltational dynamics that we observed in compact genomes will be less marked or even absent in longer sequences owing to the time needed to reach a marked stasis period. Yet, even if the dynamics are different, the potential of rearrangements as key innovations will be maintained. Indeed, duplication-triggered evolutionary bursts have been observed in bacteria. In the Long Term Evolution Experiment, a duplication has been shown to allow a bacterial strain to metabolize citrate (Blount *et al.*, 2012b). Interestingly, the duplication was not a gene duplication, but a partial promoter duplication. that was not in itself highly beneficial, but that triggered

an evolutive burst.

An intriguing question and exciting perspective of our work would be to link combinatorics of mutations with micro/macro evolution. Indeed, our results suggest that, depending on the timescale, evolution will not show the same kind of dynamics. The idea is that, on short time scales (micro-)evolution would be limited to the context of optimization of a new trait or to a new environment, mainly *via* fast mutations (substitutions and InDels). Now, on long time scales, (macro-)evolution would be mainly driven by rare bursts of innovations triggered by slow mutations (chromosomal rearrangements) that would maintain the microevolution active on the long run despite diminishing-returns epistasis (Banse *et al.*, 2023). This hypothesis to bridge the gap between long term and short term evolution reminds of proposals by Uyeda *et al.* (2011) where a corpus of biological and archaeological data is fitted with different models of long term evolution, suggesting multiple evolutionary bursts that the authors attribute to environmental changes. Our results suggest that these bursts could also have an endogenous origin, with slow mutations occasionally triggering innovations.

Acknowledgements

We would like to thank Juliette Luiselli for fruitful discussions. Part of this work has been conducted during the FP7 EvoEvo project.

Chapter VI

Predicting the Non-Coding Genome Size: a Robustness Equilibrium

Juliette Luiselli[*,1], Paul Banse[*,1], Olivier Mazet[2], Ivan Junier[3], Nicolas Lartillot[4], Guillaume Beslon[1]

[*]Paul Banse and Juliette Luiselli are joint first author
[1]Université de Lyon, INSA-Lyon, Inria, CNRS, Université Claude Bernard Lyon 1, ECL, Université Lumière Lyon 2, LIRIS UMR5205, Lyon, F-69621, France
[2] Institut National des Sciences Appliquées, Institut de Mathématiques de Toulouse, Université de Toulouse, Toulouse, France.
[3] Univ. Grenoble Alpes, CNRS, UMR 5525, VetAgro Sup, Grenoble INP, TIMC, 38000 Grenoble, France
[4] Laboratoire de Biométrie et Biologie Evolutive UMR 5558, CNRS, Université de Lyon, Villeurbanne, France

This article is in preparation.

1 Introduction

Genome size changes greatly throughout the tree of life: from 641 kbp in *buchnera Aphidicola* (Silva *et al.*, 2001), to 150 Gbp in *paris Japonica* (Leitch et Leitch, 2012). While it is arguable that coding sequences undergo adaptive changes, there are parts of the genome that seem devoid of phenotypic function. This "junk" non-coding DNA (Ohno, 1970a), sparks debates in the community about the reasons of its maintenance, both in prokaryotes (Gil et Latorre, 2012) and in eukaryotes (Doolittle, 2013).

Non-coding DNA (ncDNA) can refer to a wide range of different type of sequences, and its definition is still subject to debate (Fagundes *et al.*, 2022). In general, it refers to parts of the genome that seem to have no phenotypic effect other than marginally increasing DNA replication cost. This DNA is non-coding in the strict sense of the term, unlike, for example, non-coding RNAs, which can have regulatory impacts (Rinn et Chang, 2012). This purely non-coding DNA seems to be ubiquitous in all domains of life, and for all genome sizes (Ahnert *et al.*, 2008; Gil et Latorre, 2012; Doolittle, 2013). Even if there are debates on the precise proportion of non-coding sections, it challenges our understanding of evolution as an adaptive process, since the existence and maintenance of non-coding DNA seem impossible with the Darwinian conception of natural selection.

There are two main questions raised by the existence of non-coding DNA. How does it appear, and, if it can be easily generated, what are the constraints that prevent an infinite growth? Several hypotheses, adaptive or neutralist, have been proposed to address these issues (reviewed by Blommaert (2020)). We will try to give an overview of these theories, from the most adaptive to the most neutral.

In adaptive theories, the genome size has an intrinsic phenotypic impact, albeit small, which regulates its growth. Remarkably, non-coding genome size regulation can be performed by limitation in nutrient poor environments (Kang *et al.*, 2015; Bales et Hersch-Green, 2019). Furthermore, the positions of genes with respect to centromere or with respect to other genes influences their expression (El Houdaigui *et al.*, 2019), which may create a pressure for regulating non-coding DNA (Freeling *et al.*, 2015). However, the proportion of non-coding genome can vary by a factor of 6 within symbiotic bacteria (Giovannoni *et al.*, 2014), and the bias between fixed insertion and deletion can vary 100-fold in the tree of life (Petrov, 2002). According to Petrov (2002), these variations are too high to be the sole effect of stabilizing selection for non-coding genome size regulation. Indeed, indirect phenotypic effects of these noncoding sequences are not significant enough to explain all observed patterns, as they mostly apply to specific clades, and cannot account for the whole extent of non-coding DNA variations.

In neutral theories, the two main explanations are the mutational hazard hypothesis (MHH) proposed by Lynch et Conery (2003), and the mutational equilibrium hypothesis (MEH) proposed by Petrov (2002). In the MHH, Lynch et Conery (2003) postulate that the non-coding genome expands due to drift, for example by insertion of selfish elements, and imposes a slight selective cost on the genome due to the increased target for deleterious mutations. A mechanistic explanation of this phenomenon, based on chromosomal rearrangements, has been proposed via *in silico* experiments (Knibbe *et al.*, 2007a). Hence, in this view, there is a trend toward genome size growth in the absence of selection and the drift/selection equilibrium would determine the genome size. The MEH, on the

other side, proposes two different mutation rates biases of opposite direction (negative for InDels and positive for long insertion/deletion), which would mechanistically explain the equilibrium genome size, although the factors driving these biases are not detailed. Both theories receive some support from observations (Kelkar et Ochman, 2012; Sung *et al.*, 2016; Mueller et Jockusch, 2018; Canapa *et al.*, 2015; Smith *et al.*, 2013), but are also challenged by others (Mohlhenrich et Mueller, 2016; Sloan *et al.*, 2012). They do not offer a detailed mechanistic explanation as to why they would be valid, and are very difficult to test (Blommaert, 2020).

Other theories have been proposed to explain genome size evolution and regulation, based on unavoidable mechanistic constrains. For example, Fischer *et al.* (2014) propose that chromosomal rearrangements (such as large deletions and duplications) gain importance compared to InDels as genome size increases. Indeed, rearrangements are multiplicative operators: while a duplication at most doubles a genome size, a deletion can divide the genome size by much more than two. This simple principle imposes a mechanistic upper bound on the maximal genome size, even in the case of a reasonable bias in favor of duplications. More recently, He *et al.* (2019) a theory based on the fundamental asymmetry between insertion and deletion: in a given non-coding sequence of α bases, there are always $\alpha + 1$ positions of possible insertion of one base (following the classical asymmetry of picket-fence). Notably, for indels of greater size, this advantage is greater, as quantified by Loewenthal *et al.* (2022). Thus, insertions have a one base advantage, which, when counterbalances by the bias in mutation rates towards deletions observed in fixed mutations (Petrov, 2002; Kuo et Ochman, 2009), can lead to stable equilibrium.

In summary, although multiple hypotheses tackle the question of non-coding genome size (Blommaert, 2020), adaptive hypotheses struggle to address the wide variation observed in similar conditions, and the neutral hypotheses struggle to explain the presence and variation of equilibrium values. These models either make assumptions on a possible large phenotypic effect of ncDNA that need to be regulated, or are based on two assumptions: one for genome growth and one genome shrinkage.

Here, we show by mathematical evidence that the sole existence of chromosomal rearrangements without bias is sufficient to regulate the amount of ncDNA in a genome, regardless of any other assumption about putative deleterious/beneficial selective effect of these sequences or putative insertion/deletion bias in the occurrence of mutations. We propose a mathematical model of chromosomal rearrangements which explains two mechanisms for the growth and regulation of the non-coding genome size: a bias in the neutrality of mutations which favors the gain of new non-coding base pairs, and a selection for robustness which constrains the genome and prevents its infinite growth. Importantly, both effects are mechanistic consequences of the existence of unbalanced chromosomal rearrangements. We prove that the equilibrium point between these forces is determined by the coding genome size and its architecture – and thus dependent on the selective history of the species –, and by the mutation rate (μ) and the effective population size (N_e). This equilibrium exists even with strong variations in the mutation rates in a wide range of effective population size. We obtain trends that are coherent with literature and observations. The model allows predictions on the evolution of non-coding genome size with respect to changes in population sizes or mutation rates.

2 Model

To address the question of non coding genome size evolution, we propose to study a model of a circular genome and its possible mutations. Broadly, we look at a population of wild-type genomes and model their evolution: mutations that are neutral in terms of fitness (*i.e.* which do not affect the coding part of the genome) can go to fixation and possibly change the genome size. For the sake of simplicity, in a similar approach to the Holey fitness landscapes (Gavrilets, 1997) we will assume that all mutations that are not neutral are lethal. We are interested in studying the existence and determinants of an equilibrium in genome size, at constant fitness.

2.1 Description

2.1.1 Genome

We consider a circular genome of length L bp, composed of g coding segments (and thus g non coding segments given the circular nature of the sequence). Let's note the number of non coding bases z_{NC} and the number of coding bases z_C For simplicity, we will assume:

- There are no overlapping genes. All the genes are coding and of equal size: $\beta = (L - z_{NC})/g$,
- Genomes are considered "regular": the non-coding sections of the genome are equally distributed between the g genes, and are thus of size $\alpha = z_{NC}/g \geq 1$ each,
- Note that in this case we have $\alpha + \beta = \frac{L}{g}$

Keeping in mind that the genome is circular, without loss of generality we can assume that the first base (base of index 1) is the start of a gene.

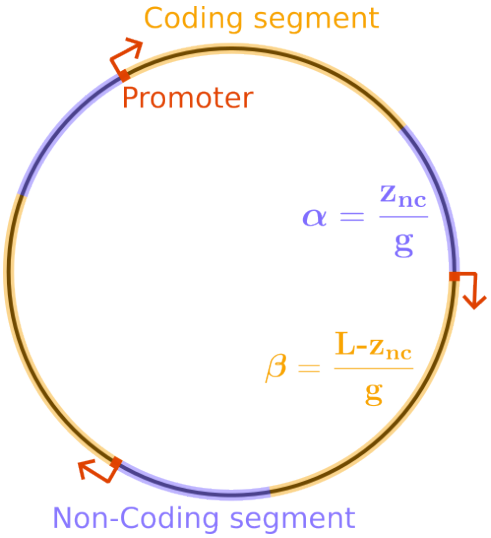


Figure VI.1: Schematic representation of a circular genome. Orange parts represent coding segments beginning with a promoter sequence (one base in the model), blue parts represent non-coding segments

2.1.2 Mutations

For simplification purposes, we admit that each mutation affecting at least one coding base is lethal (no homologous mutations or quasi-neutrality). Conversely, if a mutation only affects non-coding segments of the genome and does not create a new coding sequence, this mutation is assumed to be perfectly neutral. Finally, promoter sequences cannot be created at random, the beginning of each coding sequence is considered to be the promoter sequence, and only its duplication can create a new gene. Creating a new coding sequence is considered to be always deleterious. The maximum size of the Indel mutations is denoted l_m. For the numerical applications, we will use $l_m = 6$.

We will here consider six types of mutation:

- Point mutation (pm): change one base pair.
- Small insertion ($indel^+$): insert between one and l_m base pairs.
- Small deletion ($indel^-$): delete between one and l_m base pairs.
- Duplication ($dupl$): copy a random sequence of the genome and insert it elsewhere.
- Deletion (del): delete a random sequence of the genome.
- Inversion (inv): take a random sequence, and transpose it.

In order to assess whether a certain mutation can be fixed, we look at its neutrality. We assess, for each type of mutation i, its probability to be perfectly neutral (ν_i).

2.1.3 Probability for point mutations to be neutral

A point mutation is neutral when it affects a non-coding base, and deleterious when it affects a coding base. Since non-coding sequences are of length α, if we note 0 the index of the first base of the non-coding sequence, then the base of index α is the first base of the next coding sequence.

$$\begin{aligned}\nu_{pm} &= g \sum_{i=0}^{\alpha-1} \frac{1}{L} \\ &= \frac{g\alpha}{L}\end{aligned}$$

Probability for small insertions to be neutral

Regardless of their size, small insertion ($indel^+$) are neutral when inside a non-coding segment, and deleterious when within a coding segment. Note however that comparatively to substitutions, there are $\alpha + 1$ possible insertion points in a non-coding sequence of size α.

$$\begin{aligned}\nu_{indel^+} &= g\sum_{i=0}^{\alpha}\frac{1}{L}\\ &= \frac{g(\alpha+1)}{L}\end{aligned}$$

Probability for small deletions to be neutral

Small deletions ($indel^-$) are usually considered the opposite of small insertions, the maximum size l_m of InDels events depends on the model. Here, for simplicity, we assume that $\alpha \geq l_m$. For the numerical analysis and the simplicity of mathematical developments, we will here use $l_m = 6$. Deletion start at a random position and can delete between 1 and l_m forward. The rationale here is to calculate the probability of a neutral deletion by separating all non-coding sequences into the $\alpha - (l_m - 1)$ first bases that can witness deletions of size up to l_m, and the $l_m - 1$ other bases for which only a subset of the possible deletions are neutral.

$$\begin{aligned}\nu_{indel^-} &= g\left(\sum_{i=1}^{\alpha-(l_m-1)}\frac{1}{L} + \sum_{i=\alpha-(l_m-2)}^{\alpha}\frac{1}{L}\sum_{k=i}^{\alpha}\frac{1}{l_m}\right)\\ &= g\left(\sum_{i=1}^{\alpha-5}\frac{1}{L} + \sum_{i=\alpha-4}^{\alpha}\frac{1}{L}\sum_{k=i}^{\alpha}\frac{1}{6}\right)\\ &= \frac{g}{L}\left(\alpha - 5 + \sum_{i=\alpha-4}^{\alpha}\frac{\alpha-i+1}{6}\right)\\ &= \frac{g}{L}\left(\alpha - 5 + \sum_{i=1}^{5}\frac{1}{6}\right)\\ &= \frac{g}{L}\left(\alpha - 5 + \frac{15}{6}\right)\\ &= \frac{g}{L}\left(\alpha - \frac{5}{2}\right)\end{aligned}$$

Probability for inversions to be neutral

An inversion (inv) is neutral if the two breakpoints are outside coding regions. The two breakpoints have to be in different positions. The probability of the inversion to be neutral is thus the product of two probabilities.

$$\begin{aligned}\nu_{inv} &= \frac{g(\alpha+1)}{L} \times \frac{g(\alpha+1)-1}{L-1}\\ &= \frac{g(\alpha+1)(g\alpha+g-1)}{L(L-1)}\end{aligned}$$

Probability for segmental deletions to be neutral

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

$$\begin{aligned}\nu_{del} &= g\sum_{i=1}^{\alpha}\left(\frac{1}{L}\sum_{j=i}^{\alpha}\frac{1}{L}\right)\\ &= \frac{g}{2L^2}\sum_{i=1}^{\alpha}(\alpha-i+1)\\ &= \frac{g\alpha(\alpha+1)}{2L^2}\end{aligned}$$

Probability for duplications to be neutral

A duplication (*dupl*) is neutral if and only if it duplicates a sequence that does not include a promoter and copy it at any position in a non-coding region. The sum over s starts at position 2 to avoid the beginning of the first gene (promoter), and then all duplications are valid as long as they do not encompass the next promoter. This probability is then multiplied by the probability for the insertion point to be in a non-coding region, similarly to ν_{indel^+}

$$\begin{aligned}\nu_{dupl} &= g\sum_{i=2}^{\alpha+\beta}\left(\frac{1}{L}\sum_{j=i}^{\alpha+\beta}\frac{1}{L}\right)\left(g\sum_{k=0}^{\alpha}\frac{1}{L}\right)\\ &= \frac{g^2}{L^3}\sum_{i=2}^{\alpha+\beta}\sum_{j=i}^{\alpha+\beta}(\alpha+1)\\ &= \frac{g^2(\alpha+1)}{L^3}\sum_{i=2}^{\alpha+\beta}(\alpha+\beta-i+1)\\ &= \frac{g^2(\alpha+1)}{L^3}\sum_{i=1}^{\alpha+\beta-1}(\alpha+\beta-i)\\ &= \frac{g^2(\alpha+1)(\alpha+\beta-1)(\alpha+\beta)}{2L^3}\\ &= \frac{g(\alpha+1)(\alpha+\beta-1)}{2L^2}\end{aligned}$$

2.2 Fixation of changes in the non-coding segments

All neutral mutations have a chance to be fixed in the population, but this chance may be different depending on the change induced in genome architecture. To study the mutations that do go to fixation, we need to place our genomes in a populational context.

2.2.1 Mutational robustness

The mutational robustness is the probability for a genome to maintain its fitness despite the occurrence of a mutation. It is thus the proportion of neutral mutants. In this model, we assume that all mutations occur at the same rate, hence, the mutational robustness is just the average of the probability of neutrality for each mutation: $\overline{\nu} = \frac{1}{|M|} \sum_{i \in M} \nu_i$, where M is the set of all possible mutations.

2.2.2 Replicative robustness

The replicative robustness is the proportion of offsprings that have the same fitness as their progenitor. During a replication, the number of mutations follows a binomial law. Given that μ is the probability for each base to undergo a mutation and $\overline{\nu}$ the average probability for a mutation to be neutral, for an "average mutation", the probability for one base to be replicated without mutation is $1 - \mu$ and the probability for a base to mutate and still be neutral is on average $\mu\overline{\nu}$, hence the average probability for one base to be neutrally replicated is $1 - \mu(1 - \overline{\nu})$. If we assume that, in the same reproduction event, mutations don't affect the neutrality of following mutations, the number of neutral mutations follows a binomial law of parameter $\mathcal{B}(|M|L, 1 - \mu + \mu\overline{\nu})$

Hence, the probability for a given genome to undergo a neutral replication is the following

$$\mathcal{R}_{\text{replicative}} = (1 - \mu + \mu\overline{\nu})^{|M|L}$$

With $|M|$ the number of mutation types.

2.2.3 Effective fitness

In this model, all genomes have an intrinsic fitness which is either 1 (when the coding part of the genome is maintained) or 0 (if the coding genome has changed). To investigate the effect of replicative robustness in genome size regulation, we here define the **effective fitness** $\widetilde{f}$ of a genome as the proportion of its descendants which are viable : $\widetilde{f} \in [0; 1]$. According to the last subsection:

$$\begin{aligned} \widetilde{f} &= \sum_{a \in \text{offsprings}} f(a) \\ &= \mathcal{R}_{\text{replicative}} \\ &= (1 - \mu + \mu\overline{\nu})^{|M|L} \end{aligned}$$

Where $\mathcal{M}$ is the set of all possible mutations, μ is the mutation rate of all types of mutation and $\overline{\nu}$ is the average probability for a mutation to be neutral. This probability depends on the genome structure (z_{NC}, L, g) as shown by previous equations.

The effective fitness of each individual depends on its genome structure, and in particular z_{NC}, z_C and g. Since the coding part of the genome cannot change neutrally, the size and the number of coding segments is stable. In contrast, z_{NC} varies, by taking into account the probability of fixation based on effective fitness, we show that robustness can be selected.

2.2.4 Selection

We consider a mutant with an effective fitness $\widetilde{f}_m$ within an population of effective size Ne, homogeneously composed of wild-types with an effective fitness of $\widetilde{f}_{wt}$. Following Sella et Hirsh (2005), in a Wright-Fischer model with haploid individuals, the probability of fixation of the mutant in a population of wild-types is approximately:

$$\mathbb{P}_{fix} = \frac{1 - \left(\frac{\widetilde{f}_{wt}}{\widetilde{f}_m}\right)^2}{1 - \left(\frac{\widetilde{f}_{wt}}{\widetilde{f}_m}\right)^{2Ne}}$$

We can now mathematize the effect of selection for robustness. We assume that z_{NC} is mostly affected by the fixation of neutral deletions and duplications, and we study the expected change in genome size due to these two mutations. We consider that after each mutation, the genome is neutrally reshuffled and so non-coding segments are regular and of size $\alpha = \frac{z_{NC}}{g}$. Hence, considering that z_C and g fixed in the model, the only parameter that changes is z_{NC}. We can define the function P_{fix} where for all $k \in \mathbb{Z}^*, P_{fix}(k)$ is the probability of a mutant with $z_{NC} + k$ non-coding bases to be fixed in an ancestral population with z_{NC} non-coding bases (note that k can be negative). $P_{fix}(k)$ is thus $\mathbb{P}_{fix}$ with $\widetilde{f}_m = (1 - \mu + \mu\overline{\nu})^{|M|(L+k)}$, and $\widetilde{f}_{wt} = (1 - \mu + \mu\overline{\nu})^{|M|(L)}$.

2.2.5 Expected change in genome size

Mean size change due to neutral deletions This corresponds to the average size of a deletion weighted by its probability of neutrality times its probability of fixation (for a deletion of a segment starting at locus i and ending at locus j, this probability is $P_{fix}(-(j-i+1))$).

$$\begin{aligned}
\delta_{del} &= g \sum_{i=1}^{\alpha} \left(\frac{1}{L} \sum_{j=i}^{\alpha} \frac{1}{L} (j-i+1) P_{fix}(-(j-i+1)) \right) \\
&= \frac{g}{L^2} \sum_{i=1}^{\alpha} \sum_{j=i}^{\alpha} (j-i+1) P_{fix}(-(j-i+1)) \\
&= \frac{g}{L^2} \sum_{i=1}^{\alpha} \sum_{j=1}^{\alpha-i+1} j P_{fix}(-j)
\end{aligned}$$

Mean Size change due to neutral duplications Similarly, this corresponds to the average size of a duplication weighted by its probability of neutrality times its probability

of fixation (for a duplication of a segment starting at locus i and ending at locus j, this probability is $P_{fix}(j-i+1)$).

$$\begin{aligned}\delta_{dupl} &= g\sum_{i=2}^{\alpha+\beta}\left(\frac{1}{L}\sum_{j=i}^{\alpha+\beta}\frac{1}{L}\big(g\sum_{k=0}^{\alpha}\frac{1}{L}(j-i+1)P_{fix}(j-i+1)\big)\right)\\ &= \frac{g^2(\alpha+1)}{L^3}\sum_{i=2}^{\alpha+\beta}\sum_{j=i}^{\alpha+\beta}(j-i+1)P_{fix}(j-i+1)\\ &= \frac{g^2(\alpha+1)}{L^3}\sum_{i=2}^{\alpha+\beta}\sum_{j=1}^{\alpha+\beta-i+1}jP_{fix}(j)\end{aligned}$$

2.3 Neutral trend towards larger genomes

We want to highlight that while selection for robustness is limiting the growth of non-coding genome, there is a trend towards genome increase due to an imbalance between duplications and deletion. P_{fix}, the probability of fixation takes into account the robustness, while the rest of the summation takes into account the neutrality probability of the mutation. As a thought experiment, it is possible to imagine what would happen in a situation without selection for robustness. In order to avoid mistakes, we will use η which is δ without the probability of fixation :

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

$$
\begin{aligned}
\eta_{dupl} &= g\sum_{i=2}^{\alpha+\beta}\left(\frac{1}{L}\sum_{j=i}^{\alpha+\beta}\frac{1}{L}\big(g\sum_{k=0}^{\alpha}\frac{1}{L}(j-i+1)\big)\right)\\
&= \frac{g^2(\alpha+1)}{L^3}\sum_{i=2}^{\alpha+\beta}\sum_{j=i}^{\alpha+\beta}(j-i+1)\\
&= \frac{g^2(\alpha+1)}{L^3}\sum_{i=2}^{\alpha+\beta}\sum_{j=1}^{\alpha+\beta-i+1} j\\
&= \frac{g^2(\alpha+1)}{2L^3}\sum_{i=2}^{\alpha+\beta}(\alpha+\beta-i+1)(\alpha+\beta-i+2)\\
&= \frac{g^2(\alpha+1)}{2L^3}\sum_{i=1}^{\alpha+\beta-i}(i)(i+1)\\
&= \frac{g^2(\alpha+1)(\alpha+\beta-1)(\alpha+\beta)(\alpha+\beta+1)}{6L^3}\\
&= \frac{g(\alpha+1)(\alpha+\beta-1)(\alpha+\beta+1)}{6L^2}
\end{aligned}
$$

Then we have :

$$
\begin{aligned}
\eta_{dupl}-\eta_{del} &= \frac{g(\alpha+1)(\alpha+\beta-1)(\alpha+\beta+1)}{6L^2} - \frac{g\alpha(\alpha+1)(\alpha+2)}{6L^2}\\
&= \frac{g(\alpha+1)[(\alpha+\beta-1)(\alpha+\beta+1)-\alpha(\alpha+2)]}{6L^2}
\end{aligned}
$$

We have $\beta \geq 1$, so, unless all coding sections are only composed of promoter sequences, $\eta_{dupl} > \eta_{del}$. There exists a neutral bias towards non coding genome size increase. If phenotypical adaptation was constant, genomes would tend to grow infinitely as they gain more new non-coding bases through duplications than what they lose through deletions. However, this bias is not neutral in terms of effective fitness, which prevents an infinite growth of genome size.

2.3.1 General equilibrium

Assuming that chromosomal rearrangements are the main contributors of genome size change. Taking into account the probability of fixation, the ratio of non-coding genome size change, determined by the changes due to deletions and duplications, can be depicted as

$$
B = \frac{\delta_{del}}{\delta_{dupl}} = \frac{\frac{g}{L^2}\sum_{i=1}^{\alpha}\sum_{j=1}^{\alpha-i+1} jP_{fix}(-j)}{\frac{g^2(\alpha+1)}{L^3}\sum_{i=2}^{\alpha+\beta}\sum_{j=1}^{\alpha+\beta-i+1} jP_{fix}(j)}
$$

Although we did not find an analytical value for this, we can numerically compute this

bias, given the following parameters :

- Non-coding genome size z_{NC} (within α)
- Coding genome size z_C (within β)
- Number of coding segments g
- Spontaneous mutation rate μ (within P_{fix})
- Effective population size N_e (within P_{fix})

Notably, when an organism is evolving neutrally (well adapted, no beneficial mutations), we can assume that the genome is at its equilibrium size and $B = 1$. This enables us to retrieve numerically one of the input parameters, typically the non-coding size Z_{NC} or the fraction of non-coding DNA ($\frac{z_{NC}}{z_{NC}+z_C}$), if we know the others (Z_C,g,Ne,μ).

3 Numerical analysis

3.1 Influence of parameters and stability of the model

In this section, we start from values inspired from *E. coli* ($Ne = 1.8 \times 10^8$, $\mu = 5.4 \times 10^{-10}$ per bp, $Z_C = 4.6 \times 10^6$ and $g = 4536$, see Table. VI.1), and test the robustness of the predicted non-coding percentages (0.4%, see section 3.2.2) to changes in various parameters. Among the parameters, Ne and μ can vary due to external changes (environmental change, exposure to mutagenesis agents, etc.), and g can vary as some mutational events can separate or fuse multiple coding sequences.

Notably, we suppose that one parameter varies at a time, while all others (in particular the coding size of the genome) are kept constant. In real populations, the situation can be more complex, a change in mutation rate or population size could affect the selection and thus the coding part of the genome.

3.1.1 Variation in population size

When increasing the population size, the fraction of non-coding DNA reduces (see Fig. VI.2), as we keep the coding size constant but remove non-coding bases.

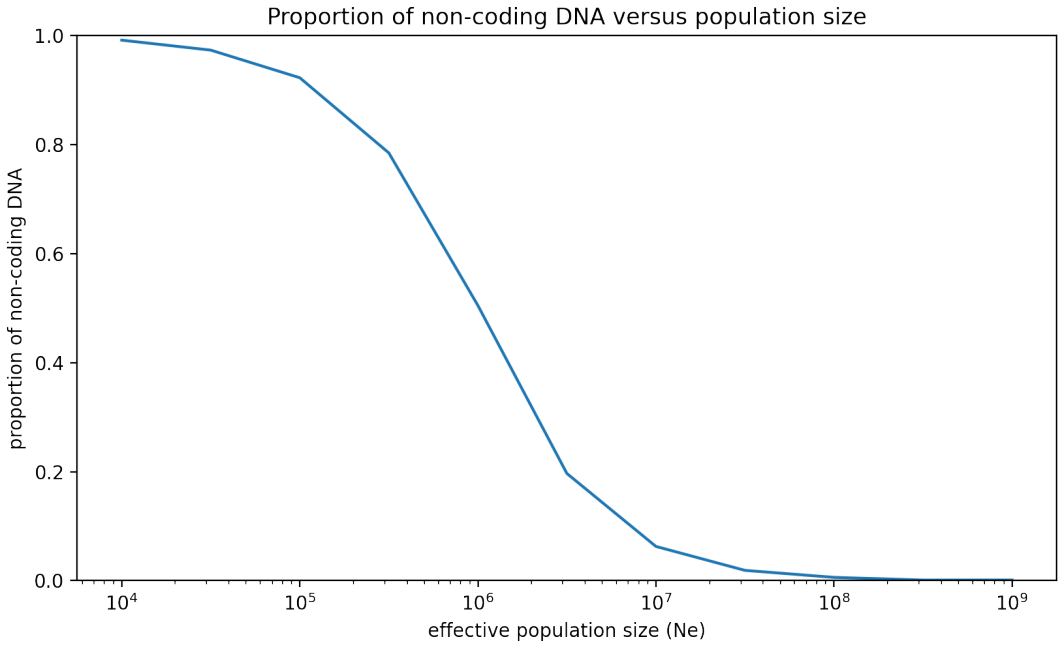


Figure VI.2: Proportion of non-coding genome at equilibrium for various population sizes. All other parameters are fixed, and taken from *E. Coli* (see Table VI.1): $\mu = 5.4 \times 10^{-10}$, $L = 4.6 \times 10^6$, $g = 4536$.

Indeed, a larger population increases the probability of fixation of beneficial mutations, which corresponds in this model to mutations increasing effective fitness, i.e. the mutations that reduce the genome size. Conversely, under a small population size genetic drift allows the fixation of more mutations that increases the non-coding quantity of the genome, enabling to reach more than 90% of non-coding bases in the genome.

3.1.2 Variation in mutation rate

When increasing the mutation rate, the percentage of non-coding genomes also reduces (see Fig. VI.3), as we keep the coding size constant but remove non-coding bases.

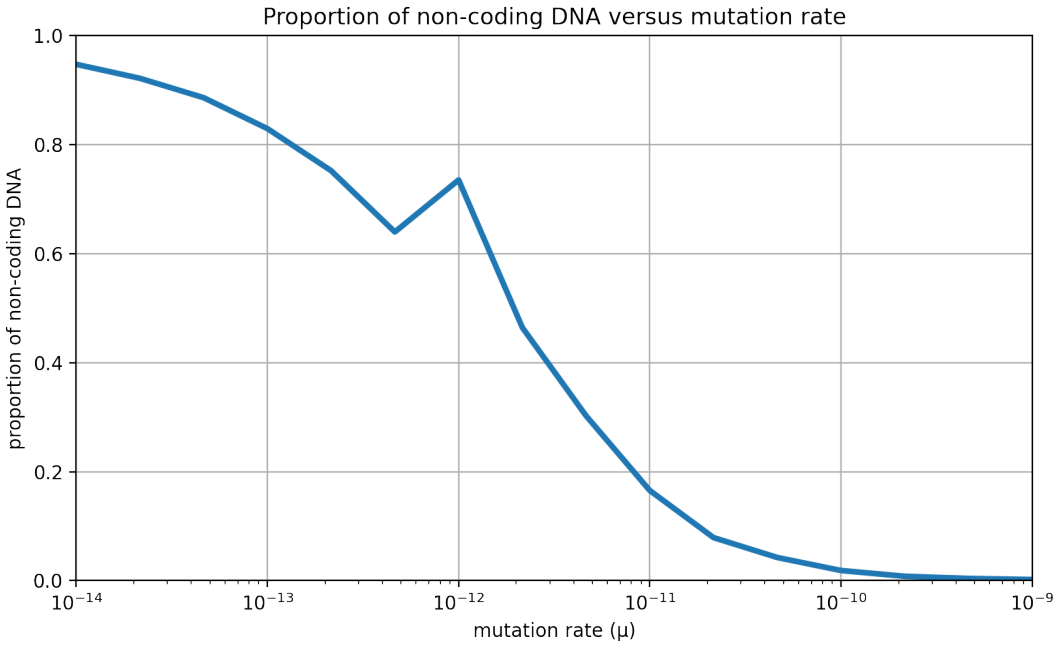


Figure VI.3: Proportion of non-coding genome at equilibrium for various mutation rates. All other parameters are fixed, and taken from *E. Coli* (see Table VI.1): $Ne = 1.8 \times 10^8$, $L = 4.6 \times 10^6$, $g = 4536$. The spike at $\mu = 10^{-12}$ is due to numerical inconsistencies at the inflection point

Indeed, increasing the mutation rate increases the number of mutant offsprings, creating more pressure for robustness, thus lowering the probability of fixation of neutral duplications and increasing the probability of neutral deletions. It is also worth noting that the lower the mutation rate, the slower the transition towards the equilibrium.

3.1.3 Variation in genome structure

As we increase the genome fragmentation (through an increase of the number of non-coding segments, g) the predicted percentage of non-coding bases increase (see Fig. VI.4).

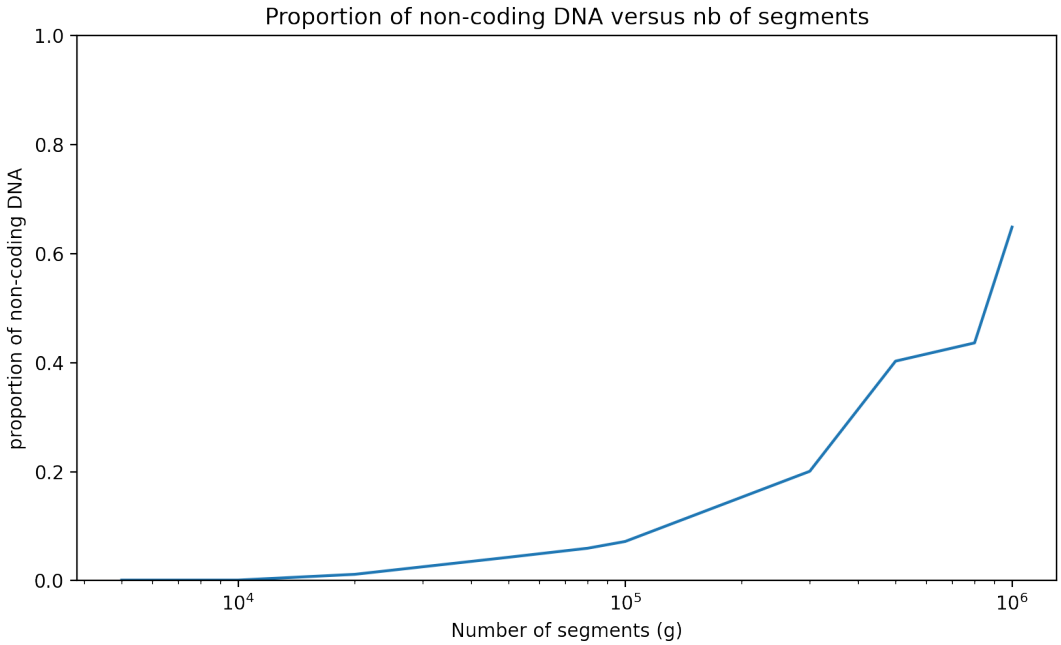


Figure VI.4: Proportion of non-coding genome at equilibrium for various number of non-coding segments g. All other parameters are fixed, and taken from *E. Coli* (see Table VI.1): $Ne = 1.8 \times 10^8$, $L = 4.6 \times 10^6$, $\mu = 5.4 \times 10^{-10}$

Indeed, the number of coding segments directly influences the insertion bias: for every segment, there is more neutral insertion position than neutral deletion.

This shows that the structure of the genome influences the quantity of non-coding bases organisms will have to deal with. Thus, a genome with a lot of genes overlap could have its coding size concentrated in only a few coding chunks, and consequently would maintain less non-coding bases. Nevertheless, there are strong constrained from the phenotypical adaptation on the number of coding segments, as there cannot be more coding segments than transcripts. Notably, the variation in non-coding percentage observed here for 3 order of magnitude of coding segments is quite low. The factors having the most influence on the non-coding size in the genomes are therefore the effective population size Ne and the mutation rate μ.

3.2 Comparison to biological data

3.2.1 Methods

To validate our model, we compare its predictions with data gathered for three organisms: *Escherichia Coli*, *Mycoplasma Pneumoniae* and *Pseudomonas Aeruginosa*. We retrieved data for N_e and μ from the literature (see Table VI.1). We also need the coding size and the number of coding (or non-coding) sequences. Non-coding sequences are defined in our model as all sequences which have absolutely no phenotypic effect. To identify what could biologically correspond to a non-coding sequences in a biological genome, we use transcriptomics and annotation data.

Estimating the non-coding sequences in a genome is far from trivial. Non-coding sequences correspond to regions that are outside genes and never transcribed, as they can be broken, extended, or reduced without any consequences on the phenotypical adaptation. On the other hand, some groups of genes need to stay together: some bases between them would be untranscribed, but we consider them coding as they cannot be separated without consequences. Additionally, non-coding RNA are known to play a role in regulation (Yoon *et al.*, 2013), and we therefore included all transcribed regions in our "coding" sections. Thus, we consider that a base is non-coding if it is never transcribed in a transcriptomic dataset personally communicated by Ivan Junier, and does not belong to any known gene. This also gives us the total number of coding and non-coding bases for the studied genomes within our framework.

3.2.2 Results

For all three species, and all mutation types, we use the estimate $\mu = 5.4 \times 10^{-10}$ (Drake *et al.*, 1998b), as there are few estimates of spontaneous mutation rates for different bacterial species. The number of non-coding segments and the total coding size are measured from transcriptomics data (Ivan Junier personal communication).

Species	Mutation rate	Effective population size (Ne)	Genome size	Non-coding segments	Non-coding percentage	**Predicted non-coding percentage**
E. Coli	5.4×10^{-10}	5×10^7 (a) 1.8×10^8 (b) 4.2×10^8 (c)	4.6×10^6	4536	2.3% (d)	1.6% 0.4% 0.1%
M. Pneumoniae	5.4×10^{-10}	3.8×10^6 (c)	8.2×10^5	371	3.4% (d)	10.4%
P. Aeruginosa	5.4×10^{-10}	4.0×10^8 (c) 2.1×10^8 (b)	6.3×10^6	6042	0.7% (d)	0.2% 0.3%

Table VI.1: Used parameters, predicted and expected non-coding percentages. Data are from : (a): Charlesworth et Eyre-Walker (2006), (b): (Lynch *et al.*, 2016), (c): (Bobay et Ochman, 2018), (d): transcriptomic analysis (Ivan Junier personal communication).

The non-coding percentages predicted by our for these three bacterial species, although not absolutely correct, is close to what was expected with respect to the uncertainty of the data. We are able to retrieve the correct order of magnitude, supporting that the mechanisms highlighted in our model could partly explain the amount of purely non-coding DNA in living organisms. Notably, *P. Aeruginosa* is the most coding species, and *M. Pneumoniae* the least coding species, in both data and prediction.

4 Model analysis

Using numerical computation based on the parameter values of *E. Coli*, we can show that, strikingly, the percentage of non-coding genome is constant when the factor $N_e \times \mu$ remains constant.

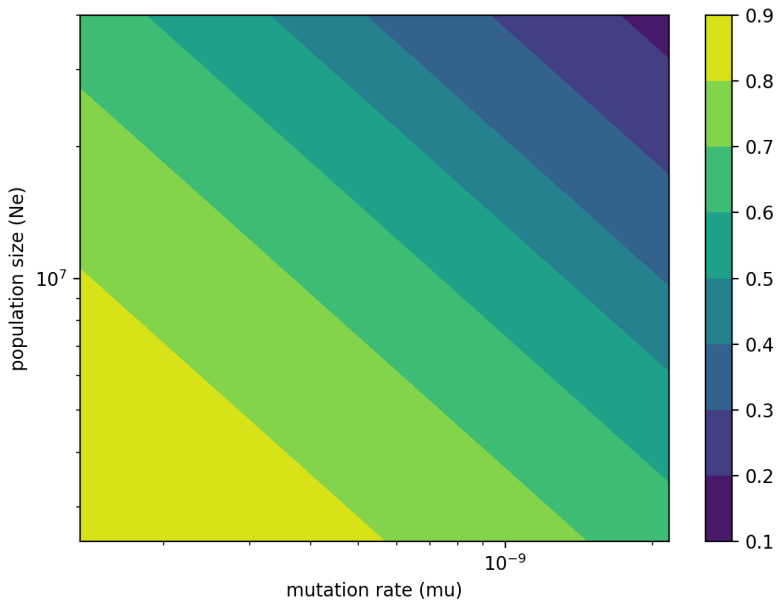


Figure VI.5: Non-coding percentage at equilibrium for 25 different population sizes and mutation rates.

Indeed, in most biologically relevant cases (*i.e.* μ small and N_e large), N_e and μ have a symmetric function at equilibrium in our model. Using our mathematical model, we can investigate for an analytic proof of the symmetrical roles of μ and Ne on the equilibrium.

Ne and μ only appear in the $\mathbb{P}_{fix}$ part of the equation. Let us start with $\mathcal{R}_{\text{replicative}}$. In the common case, the mutation rate μ is negligible compared to 1, we can thus use the first order of the Taylor expansion to simplify the equation:

$$\begin{aligned}\mathcal{R}_{\text{replicative}} &= (1-\mu(1-\overline{\nu}))^{|M|L} \\ &= \exp(|M|L\ln(1-\mu(1-\overline{\nu}))) \\ &\underset{\mu\to 0}{=} \exp(-|M|L\mu(1-\overline{\nu}))\end{aligned}$$

Then the ratio of effective fitness values between a rearranged mutant undergoing a neutral duplication (or deletion) of size k (with $k<0$ in the case of a deletion) and the ancestral population of non-mutated individuals is:

$$\begin{aligned}\frac{\mathcal{R}_{\text{replicative}}(L)}{\mathcal{R}_{\text{replicative}}(L+k)} &= \frac{\exp(-|M|L\mu(1-\overline{\nu}))}{\exp(-|M|(L+k)\mu(1-\overline{\nu}))} \\ &= \exp\left(|M|k\mu(1-\overline{\nu})\right))\end{aligned}$$

Then the probability of fixation of a duplication of size $k > 0$ is :

$$\begin{aligned}\mathbb{P}_{fix}(L,k) &= \frac{1-\left(\frac{\mathcal{R}_{\text{replicative}}(L)}{\mathcal{R}_{\text{replicative}}(L+k)}\right)^2}{1-\left(\frac{\mathcal{R}_{\text{replicative}}(L)}{\mathcal{R}_{\text{replicative}}(L+k)}\right)^{2N_e}} \\ &= \frac{1-\exp\left(2|M|k\mu(1-\overline{\nu})\right)}{1-\exp\left(2N_e|M|k\mu(1-\overline{\nu})\right)} \\ &\underset{\mu\to 0}{=} \frac{-2|M|k\mu(1-\overline{\nu})}{1-\exp\left(2N_e|M|k\mu(1-\overline{\nu})\right)} \\ &= \mu \times \mathscr{P}(N_e\mu, k)\end{aligned}$$

Where $\mathscr{P}$ is a function of $N_e\mu$ and other parameters, but N_e and μ always appear together. Thus, δ_{dupl} can also be written as μ times a function $\mathscr{P}$ of $N_e\mu$.

Similarly for a deletion with $k < 0$ we have $\mathbb{P}_{fix}(L,k) \underset{\mu\to 0}{=} \mu \times \mathscr{P}(N_e\mu, k)$.

Since δ_{del} and δ_{dupl} are sums of products of probabilities of fixation multiplied by factors independent of μ and Ne, there exist two functions d_{del} and d_{dupl} such that $\delta_{del} = \mu \times d_{del}(Ne\mu)$ and $\delta_{dupl} = \mu \times d_{dupl}(Ne\mu)$. Then since $B = \frac{\delta_{del}}{\delta_{dupl}} = \frac{\mu \times d_{del}(Ne\mu)}{\mu \times d_{dupl}(Ne\mu)}$, the μ cancels out and thus B is only a function of $N_e\mu$.

Intuitively, the larger Ne, the more efficient the selection for robustness, and the larger μ, the more robustness is needed to survive. This symmetry is not in the design of the model but emerges after the numerical analysis. It allows us to simplify our understanding of non-coding genome evolution: our two parameters Ne and μ correspond to only one degree of freedom in the model. Notably, this property holds for the equilibrium value, but not for the dynamic of convergence towards the equilibrium. For the same value of $Ne\mu$, convergence is probably much slower in the case of small μ and large Ne than the reverse.

5 Discussion

We have shown that an equilibrium between a trend towards higher genomes due to duplication/deletion imbalance, and the selection for replicative robustness can explain the maintenance of non-coding DNA in genomes. We could partially predict the non-coding equilibrium size in three species, independently from any phenotypical adaptation mechanism. This confirms that robustness is a key factor of evolution (De Visser *et al.*, 2003), not only in phenotypic adaptation(Wilke *et al.*, 2001), but also in neutral or quasi-neutral dynamics of genome size regulation. This equilibrium between a mutational bias and selection for robustness is determined by few parameters: on one hand, the population size, which drives the efficacy of the selection, and on the other hand, the mutation rate and the genome structure (size and segmentation of the coding genome), which both influence the robustness of the genomes. Ne and μ stands out in our result as having the most effect on the amount of non-coding sequences in a genome, as had been suggested previously by the Mutational Hazard Hypothesis (Lynch et Conery, 2003). However, the mechanism presented here is different, and we show analytically why and how these two parameters have the same effect on genome structure.

Hence, we expect species with a larger population size or a higher mutation rate to have a shorter genome, as we observe here a reduction in the non-coding genome. As a matter of fact, species with a high mutation rate, or a high population size, are observed to have a lower genome size (Dufresne *et al.*, 2005; Giovannoni *et al.*, 2005). While these relationships are complicated by the impact of the genome structure (g and z_C) and the effect of selection for phenotypical adaptation, our model can reach a wide range of values by changing only Ne and μ. Interestingly, our model predicts that, if Ne and μ change but $Ne\mu$ stays constant, the fraction of non-coding DNA stays constant in the genome.

5.1 Limits

One of the limit of our model is that we assume non-coding and coding sequences are equally spread around the genome, and of constant length. According to (Biller *et al.*, 2016), due to the action of neutral inversions, non-coding sections of the genome are predicted to be distributed along the genome following a flat Dirichlet distribution. Moreover, in the model, the size of rearrangements spans uniformly the whole length of the genome. Accounting to Darling *et al.* (2008), their true distribution is more likely to be geometric. However, this would not change profoundly the dynamics, as long as parameters of the distribution of rearrangement sizes still scale linearly with the size of the genome. We conjecture that this could reduce the robustness load, as large rearrangements only contribute to robustness and are never fixed, but this would only marginally change the equilibrium value of non-coding percentage, and not the direction of variations in non-coding size nor the mechanisms behind the equilibrium.

A high precision mathematical model would also be irrelevant because there is a lack of biological data about the parameters of the model. For example, most measurements only take into account mutations that went to fixation, while our model distinguish a bias in possible neutral mutations and mutations that actually go to fixation which are influenced by selection for robustness. There are also large differences in the literature in measures of effective population size for the same specie.

5.2 Perspectives

Other mutations could be added to the model, such as translocation, tandem duplications or Insertion Sequences. We believe this could change the final equilibrium, as the relative strength of the two opposite forces would vary, but not the dynamic we point out. Indeed, increasing Ne or μ would still decrease the non-coding proportion, and conversely.

The limits of our assumption prevent the model from being applicable to all the genomes. Indeed, in the numerical simulations, we have assumed that non-coding sequences are of size 6 at least, and we cannot compute the bias if the non-coding percentage reaches 0. In very dense genomes, or when considering subparts of the genomes with very short non-coding segments, the bias in the neutrality of InDels could become stronger than the bias in larger duplications and deletion we describe, hence dominating the dynamic of genome size evolution. Such a mechanism has for example been described in the evolution of introns size (Loewenthal *et al.*, 2022). Conversely, for this model to be relevant, we need biological genomes to reach their equilibrium size quick enough for the environment and the size of the coding sections to be considered constant until equilibrium

is reached. This supposes a relatively fast evolving population with short generations and high mutation rate. Also, our model is designed for prokaryotes, and it would be difficult to adapt it to eukaryotes, due to a large difference in genome structure and mutations. Any extension of our results would also mean reconsidering, and thoroughly discussing our preliminary hypotheses.

5.3 Conclusion

The mutational bias we observe differs to what is generally reported in the literature: indeed, a bias toward deletion is often reported (Petrov, 2002; Kuo et Ochman, 2009). However, our model produces results similar to the Mutational Hazard Hypothesis (MHH), as the non-coding percentage of the genome decreases with the increase in population size and mutation rates (Lynch et Conery, 2003; Blommaert, 2020). This has also been observed multiple times in living species (Dufresne *et al.*, 2005; Giovannoni *et al.*, 2005), but not mechanistically explained so far. Our predictions show similar order of magnitude than data in the literature, although refining the model would require less incertitude on the measures for effective population size, spontaneous mutation rate, and the amount of "truly non-coding" base pairs. Our work opens new opportunities for theorists as well as experimentalists. The former because of the many simplifications and assumptions made in the model need to be questioned, and require deeper investigation. The latter because this model can be applied to a broad range of prokaryotes, for example, to predict the effect of increase in mutation rate over the non-coding genome size. This would also help to verify whether the trends of coding fraction are coherent with our predictions on a larger scale.

To conclude, our mathematical model of genome size evolution is coherent with previous theoretical (Lynch, 2010) and empirical literature (Giovannoni *et al.*, 2014). It is based solely on the existence of chromosomal rearrangements, and thus needs much less preliminary hypotheses than all previously proposed explanation of non coding DNA evolution (Blommaert, 2020). We have shown that both an intrinsic mutation neutrality bias towards genome growth, due to duplications being generally less deleterious than deletions, and the selection for replicative robustness to chromosomal rearrangements are driving genome size evolution in prokaryotes. We have shown that the equilibrium between these opposite forces determines the percentage of non-coding sequences a genome withholds. It depends on spontaneous mutation rates, effective population size and architecture of the coding sections of the genome. And finally, we have shown this equilibrium is predictable.

Part C

Conclusion

In this thesis, we intended to push forward the understanding of the impact of chromosomal rearrangements on evolution. Indeed, while they have been the first type of mutation discovered, they were overshadowed by the study of short read sequencing. For the sake of simplicity, most theoretical and computational models focused on local mutations that were more often observed in biological data. While this led to interesting conceptual developments, the scientific community needs a better general understanding of chromosomal rearrangements (Mérot *et al.*, 2020). Indeed, there are very few theoretical computational models of evolution that includes them, even if there are well known examples of rearrangements having a major role in evolution (Blount *et al.*, 2012a).

The first step of this thesis was to use Aevol to model chromosomal rearrangements and to witness how they have strong effects on adaptive behaviors. On one hand, adaptation of genomes evolved without chromosomal rearrangements slowed down significantly faster than genomes evolved with them. Chromosomal rearrangements limit the diminishing return on successive beneficial mutations because their number of accessible mutants, their combinatorics, is multiple orders of magnitude larger than local mutations. On the other hand, genome sizes seemed regulated by chromosomal rearrangements, while simulations with only local mutations witnessed a seemingly limitless increase in genome size. In fact, rearrangements originating from non-coding sections of the genome could impact coding sections of the genome, creating a selection for robustness. Although simulations show large differences in the evolutionary process due to chromosomal rearrangements, their "fossil record" stays discrete. Indeed, looking at the mutations fixed during simulation, an overwhelming majority of them are local mutations and few are chromosomal rearrangements: while chromosomal rearrangements seem to be the backbone of long term evolution, they are almost invisible in the phylogeny.

Starting from these empirical observations, we build a theoretical framework to study one of the key components of rearrangements: the number of accessible mutants. The number of accessible mutants for a given type of mutation is a combinatorial number that allows quantifying the potential to reach new genomes. We study evolution with either local mutations (substitutions) or chromosomal rearrangements (inversions) in the NK-fitness landscape and show that inversions, arguably one of the simplest chromosomal rearrangement, adds a lot of complexity in the understanding. Indeed, concepts like epistasis are not fit to describe inversions, and others like mutational distance and smoothness of the landscape have to be adapted to the new mutational operator. The number of accessible mutants by inversions is dependent on the genome, while being always greater than the one of substitution. The evolutive potential of inversions also depends on how the information is coded in the genome: inversions operate on subsequences of the genome, and their distribution of fitness effect is less deleterious on average (smoother) when the epistatic information is also concentrated on subsequences. However, in all experiments, as the graph of accessible mutants of inversions contains the graph of accessible mutants of substitutions, simulations with inversions always get higher in the fitness landscape comparatively to simulations with only substitutions.

Empirically, our first results have shown that, when rearrangements are present, greater fitness values can be reached in the long term. By simulating viral-like genomes in Aevol, we show that in such short, almost fully-coding genomes, the evolution is intrinsically bursty. We used a formal characterization of fitness peaks (or plateaus) to

precisely identify peak shifts periods, which concentrated most of the fitness gains in a short period of time. Peak shifts are initiated by key innovation, and we show that some types of mutations are more likely than others to be key innovations. Furthermore, not all peak shifts correspond to mutational bursts. We show that some key innovations are more likely to trigger bursts of beneficial mutations and that these key innovations are not defined by their own fitness gain but by their mutation type: they are segmental duplications. While long term adaptation may exhibit an open-ended gradual behavior (as exemplified by our first results), in the short term the fitness mainly increases by series of close-ended bursts of local mutations triggered by duplications. Moreover, duplications able to trigger bursts are exceptionally rare and thus long to find compared to the fast tempo of local mutations. This explains how innovations can burst out in viral genome, even after hundreds of thousands of generations of apparent stasis.

In our first simulations, rearrangements have been shown to regulate genome size and the amount of non-coding sequences. Using a mathematical model, we show that the mere existence of chromosomal rearrangements is enough to explain the existence and maintenance of a significant fraction of non-coding sequences in prokaryotes. The contribution of rearrangements to the dynamics of "junk DNA" is twofold. On one hand, assuming similar spontaneous mutation rates, there is a bias in the size of neutral rearrangements: in average neutral duplications are larger than the neutral deletions. This leads to a neutral increase of non-coding genome size. On the other hand, due to the long range effect of chromosomal rearrangement, non-coding bases can potentially initiate deleterious rearrangements and are then counter-selected. This "selection for robustness" can be taken into account by introducing an "effective fitness" value. Such an indirect effect is difficult to see, as it is caused by mutations that do not go to fixation. Our model shows how an equilibrium can emerge, just from the existence of long ranged mutations (duplications and segmental deletions). Furthermore, based on an estimation of the robustness and of the mutation's neutrality bias, we show that it is possible to predict non-coding genome size from a limited number of parameters.

While this work is still ongoing and showing new promising perspectives, this thesis needs to be concluded. The articles presented here open multiple research opportunities, both in terms of theoretical work, computational simulation and biological experiments. It is likely that not all these opportunities will be explored, but it can still be valuable to list some of them here.

The first opportunity is mathematical. Our model is built on multiple simplifying assumptions, it would be interesting to test which of these assumptions are really required. Also, an analytical result on the non-coding percentage equilibrium seems possible given the results of the numerical simulations, however, we did not achieve it yet.

A theoretical computer science discussion is also ongoing to try understanding mutational operators and how they influence the properties of the mutational network. It would also be interesting to include fitness in these discussions to see how the fitness of a mutant is more or less correlated to the fitness of its ancestor, depending on the types of mutations and on information coding.

Another question is the redundancy of mutations, given that there are only four base pairs, a short sized chromosomal rearrangement can easily be classified as two local mutations if most of the sequence is conserved. This leads to the question of how difficult it

is to estimate spontaneous mutation rates, as selection is constantly operating.

In terms of computational modeling, for simplicity and historical reasons the ideas presented here have focused on microbial prokaryotic genome, it would be interesting to expand them to eukaryotic genomes and to other type of mutation such as Horizontal Gene Transfer (HGT) or Transposable Elements (TE).

About virus evolution, it would be promising to wonder if a duplication event can predict a mutational burst: indeed after few generations, if the duplication is not filtered out by selection, it may be going through an evolutive burst. This idea of prediction can also be applied on real viral population, for example looking at viral strains, one could try to see how many of them underwent a duplication before a major evolutive event.

In the viral evolution article (chapter V), duplication, by being slightly less deleterious than other chromosomal rearrangements, was the main cause of evolutive bursts. It is important to mention that evolution is a race where only top competitors are broadcasted. If, in biological evolution, the relevance of the various theories of valley crossing is different, some mechanisms showed here may be overshadowed by others. Since our models of chromosomal rearrangement require few hypotheses and are backed by some biological data, we are still confident that they are relevant.

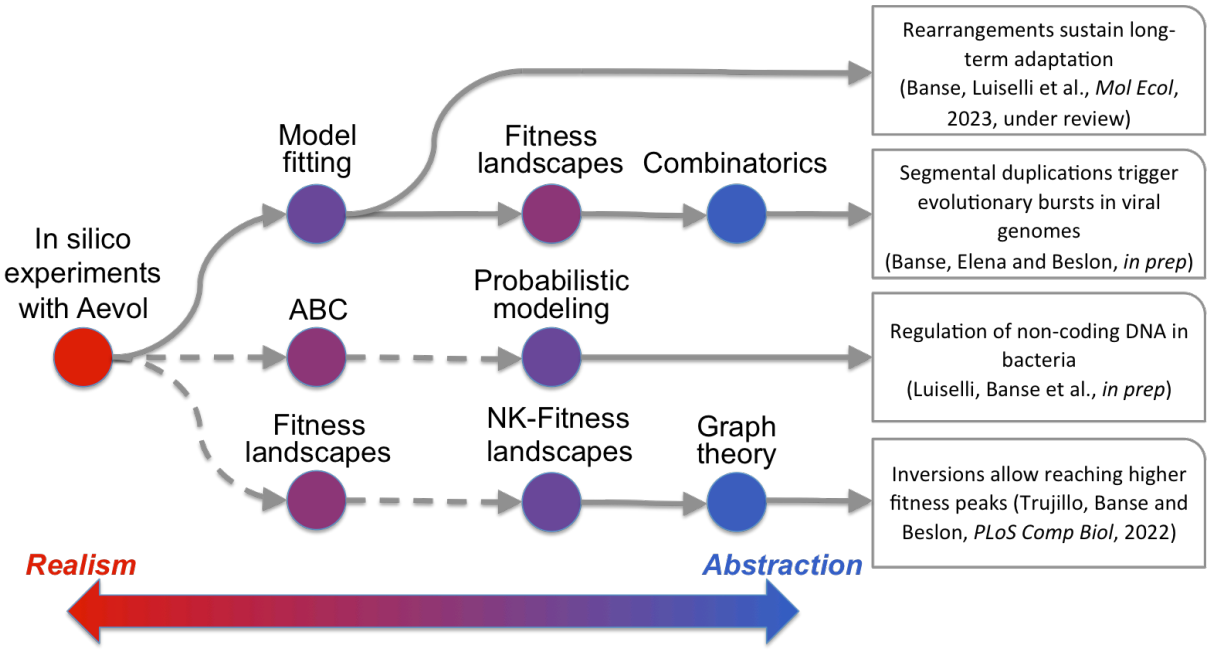


Figure VII.1: Overview of the models and computational concepts leveraged in this thesis. As exemplified in chapter II, ideas all come from empirical simulations with Aevol. We then progressively introduced abstracts models and concepts to propose new results published or in preparation for different journals. Plain lines represent connections that are acknowledged in the final articles. Dash lines represent models that have been of central importance in developing the concepts but are no longer needed afterward[1] and are thus absent from the final publications/chapters.

A good overview of this work is the number of new, adapted or refined computational tools and concepts that we had to use in order to understand and classify the effects of rearrangements (Figure VII.1). All these concepts were needed to design and explain both mathematical and computational models presented here. Indeed, ultimately, rearrangements are informational operators that strongly modify the genomic information. It is hardly surprising, then, that the importation of concepts from the information sciences (graphs, combinatorics, probabilistic modeling etc.) into the field of evolutionary biology can help shed further light on their properties. This is what we tried to do in this thesis, and we hope to have convinced the reader of the relevance of this approach.

[1] *The decisive thing with modelling is not the model per se, but what the model and working with the model does to our mind. It could even be argued that a criterion to determine good models is that they are no longer needed afterwards* (Grimm, 1999).

Bibliography

Adami, C. (2006). Digital genetics: unravelling the genetic basis of evolution. *Nature Reviews Genetics*, 7(2):109–118.

Aguilar-Rodríguez, J., Peel, L., Stella, M., Wagner, A. et Payne, J. L. (2018). The architecture of an empirical genotype-phenotype map. *Evolution*, 72(6):1242–1260.

Aguirre, J., Buldú, J. M. et Manrubia, S. C. (2009). Evolutionary dynamics on networks of selectively neutral genotypes: Effects of topology and sequence stability. *Physical Review E*, 80(6):066112.

Aguirre, J., Catalán, P., Cuesta, J. et Manrubia, S. (2018). On the networked architecture of genotype spaces and its critical effects on molecular evolution. *Open Biology*, 8(7):180069.

Ahnert, S. E., Fink, T. M. A. et Zinovyev, A. (2008). How much non-coding DNA do eukaryotes require? *Journal of Theoretical Biology*, 252(4):587–592.

Ahvanooey, M. T., Li, Q., Wu, M. et Wang, S. (2019). A survey of genetic programming and its applications. *KSII Trans. Internet Inf. Syst.*, 13(4):1765–1794.

Aita, T., Iwakura, M. et Husimi, Y. (2001). A cross-section of the fitness landscape of dihydrofolate reductase. *Protein engineering*, 14(9):633–638.

Alkan, C., Coe, B. P. et Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12(5):363–376. Number: 5 Publisher: Nature Publishing Group.

Altland, A., Fischer, A., Krug, J. et Szendro, I. G. (2011). Rare events in population genetics: stochastic tunneling in a two-locus model with recombination. *Physical Review Letters*, 106(8):088101.

Ancel Meyers, L., Ancel, F. D. et Lachmann, M. (2005). Evolution of genetic potential. *PLoS Computational Biology*, 1(3):e32.

Audrézet, M.-P., Chen, J.-M., Raguénes, O., Chuzhanova, N., Giteau, K., Maréchal, C. L., Quéré, I., Cooper, D. N. et Férec, C. (2004). Genomic rearrangements in the cftr gene: extensive allelic heterogeneity and diverse mutational mechanisms. *Human mutation*, 23(4):343–357.

Bales, A. L. et Hersch-Green, E. I. (2019). Effects of soil nitrogen on diploid advantage in fireweed, Chamerion angustifolium (Onagraceae). *Ecology and Evolution*, 9(3):1095–1109.

Bank, C., Hietpas, R. T., Jensen, J. D. et Bolon, D. N. (2015). A systematic survey of an intragenic epistatic landscape. *Molecular biology and evolution*, 32(1):229–238.

Bank, C., Matuszewski, S., Hietpas, R. T. et Jensen, J. D. (2016). On the (un) predictability of a large intragenic fitness landscape. *Proceedings of the National Academy of Sciences*, 113(49):14085–14090.

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

Barabási, A.-L., Stanley, H. E. *et al.* (1995). *Fractal concepts in surface growth*. Cambridge university press.

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

Batut, B., Parsons, D. P., Fischer, S., Beslon, G. et Knibbe, C. (2013). In silico experimental evolution: a tool to test evolutionary scenarios. *In BMC bioinformatics*, volume 14, page S11. BioMed Central.

Bedford, T., Cobey, S. et Pascual, M. (2011). Strength and tempo of selection revealed in viral gene genealogies. *BMC evolutionary biology*, 11(1):1–16.

Beerenwinkel, N., Pachter, L. et Sturmfels, B. (2007a). Epistasis and shapes of fitness landscapes. *Statistica Sinica*, pages 1317–1342.

Beerenwinkel, N., Pachter, L., Sturmfels, B., Elena, S. F. et Lenski, R. E. (2007b). Analysis of epistatic interactions and fitness landscapes using a new geometric approach. *BMC Evolutionary Biology*, 7(1):1–12.

Beiko, R. G. et Charlebois, R. L. (2007). A simulation test bed for hypotheses of genome evolution. *Bioinformatics*, 23(7):825–831.

Belinky, F., Sela, I., Rogozin, I. B. et Koonin, E. V. (2019). Crossing fitness valleys via double substitutions within codons. *BMC biology*, 17(1):1–15.

Belshaw, R., Gardner, A., Rambaut, A. et Pybus, O. G. (2008). Pacing a small cage: mutation and rna viruses. *Trends in ecology & evolution*, 23(4):188–193.

Belshaw, R., Pybus, O. G. et Rambaut, A. (2007). The evolution of genome compression and genomic novelty in rna viruses. *Genome research*, 17(10):1496–1504.

Berdan, E. L., Blanckaert, A., Butlin, R. K. et Bank, C. (2021a). Deleterious mutation accumulation and the long-term fate of chromosomal inversions. *PLoS genetics*, 17(3):e1009411.

Berdan, E. L., Blanckaert, A., Slotte, T., Suh, A., Westram, A. M. et Fragata, I. (2021b). Unboxing mutations: Connecting mutation types with evolutionary consequences. *Molecular Ecology*, 30(12):2710–2723.

Beslon, G., Liard, V., Parsons, D. P. et Rouzaud-Cornabas, J. (2021). Of evolution, systems and complexity. *In Evolutionary Systems Biology*, pages 1–18. Springer.

Beslon, G., Parsons, D., Sanchez-Dehesa, Y., Peña, J.-M. et Knibbe, C. (2010). Scaling laws in bacterial genomes: A side-effect of selection of mutational robustness? *Biosystems*, 102(1):32–40.

Bhange, D., Prasad, N., Singh, S., Prajapati, H. K., Maurya, S. P., Gopalan, B. P., Nadig, S., Chaturbhuj, D., Jayaseelan, B., Dinesha, T. R. *et al.* (2021). The evolution of regulatory elements in the emerging promoter-variant strains of hiv-1 subtype c. *Frontiers in Microbiology*, page 3442.

Bhatia, S., Feijão, P. et Francis, A. R. (2018). Position and content paradigms in genome rearrangements: the wild and crazy world of permutations in genomics. *Bulletin of Mathematical Biology*, 80:3227–3246.

Biller, P., Guéguen, L., Knibbe, C. et Tannier, E. (2016). Breaking good: accounting for fragility of genomic regions in rearrangement distance estimation. *Genome biology and evolution*, 8(5):1427–1439.

Blakely, E. L., Rennie, K. J., Jones, L., Elstner, M., Chrzanowska-Lightowlers, Z. M. A., White, C. B., Shield, J. P. H., Pilz, D. T., Turnbull, D. M., Poulton, J. et Taylor, R. W. (2006). Sporadic Intragenic Inversion of the Mitochondrial DNA MTND1 Gene Causing Fatal Infantile Lactic Acidosis. *Pediatric Research*, 59(3):440–444.

Blommaert, J. (2020). Genome size evolution: towards new model systems for old questions. *Proceedings of the Royal Society B: Biological Sciences*, 287(1933):20201441.

Bloom, J. D., Gong, L. I. et Baltimore, D. (2010). Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. *Science*, 328(5983):1272–1275.

Blount, Z. D., Barrick, J. E., Davidson, C. J. et Lenski, R. E. (2012a). Genomic analysis of a key innovation in an experimental Escherichia coli population. *Nature*, 489(7417):513–518. Number: 7417 Publisher: Nature Publishing Group.

Blount, Z. D., Barrick, J. E., Davidson, C. J. et Lenski, R. E. (2012b). Genomic analysis of a key innovation in an experimental escherichia coli population. *Nature*, 489(7417):513–518.

Bobay, L.-M. et Ochman, H. (2018). Factors driving effective population size and pan-genome evolution in bacteria. *BMC Evolutionary Biology*, 18(1):153.

Bollobás, B. (2013). *Modern graph theory*, volume 184. Springer Science & Business Media.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.

Bray Speth, E., Long, T. M., Pennock, R. T. et Ebert-May, D. (2009). Using avida-ed for teaching and learning about evolution in undergraduate introductory biology courses. *Evolution: Education and Outreach*, 2(3):415–428.

Brockhurst, M. A., Colegrave, N. et Rozen, D. E. (2011). Next-generation sequencing as a tool to study microbial evolution. *Molecular Ecology*, 20(5):972–980.

Burssed, B., Zamariolli, M., Bellucco, F. T. et Melaragno, M. I. (2022). Mechanisms of structural chromosomal rearrangement formation. *Molecular Cytogenetics*, 15(1):1–15.

Campo, N., Dias, M. J., Daveran-Mingot, M. L., Ritzenthaler, P. et Le Bourgeois, P. (2004). Chromosomal constraints in Gram-positive bacteria revealed by artificial inversions: Experimental genome inversions in Streptococcaceae. *Molecular Microbiology*, 51(2):511–522.

Campos, P. R. et Moreira, F. B. (2005). Adaptive walk on complex networks. *Physical Review E*, 71(6):061921.

Canapa, A., Barucca, M., Biscotti, M. A., Forconi, M. et Olmo, E. (2015). Transposons, Genome Size, and Evolutionary Insights in Animals. *Cytogenetic and Genome Research*, 147(4):217–239.

Cano, A. V. et Payne, J. L. (2020). Mutation bias interacts with composition bias to influence adaptive evolution. *PLoS Computational Biology*, 16(9):e1008296.

Cao, S., Brandis, G., Huseby, D. L. et Hughes, D. (2022). Positive Selection during Niche Adaptation Results in Large-Scale and Irreversible Rearrangement of Chromosomal Gene Order in Bacteria. *Molecular Biology and Evolution*, 39(4). msac069.

Capitan, J. A., Aguirre, J. et Manrubia, S. (2015). Dynamical community structure of populations evolving on genotype networks. *Chaos, Solitons & Fractals*, 72:99–106.

Carde, Q., Foley, M., Knibbe, C., Parsons, D. P., Rouzaud-Cornabas, J. et Beslon, G. (2019). How to reduce a genome? alife as a tool to teach the scientific method to school pupils. *In Artificial Life Conference Proceedings*, pages 497–504. MIT Press One Rogers Street, Cambridge, MA 02142-1209.

Catalán, P., Arias, C. F., Cuesta, J. A. et Manrubia, S. (2017). Adaptive multiscapes: an up-to-date metaphor to visualize molecular adaptation. *Biology Direct*, 12(1):1–15.

Cervera, H., Lalić, J. et Elena, S. F. (2016). Efficient escape from local optima in a highly rugged fitness landscape by evolving rna virus populations. *Proceedings of the Royal Society B: Biological Sciences*, 283(1836):20160984.

Charlesworth, J. et Eyre-Walker, A. (2006). The Rate of Adaptive Evolution in Enteric Bacteria. *Molecular Biology and Evolution*, 23(7):1348–1356.

Chatterjee, K., Pavlogiannis, A., Adlam, B. et Nowak, M. A. (2014). The time scale of evolutionary innovation. *PLoS computational biology*, 10(9):e1003818.

Connallon, T. et Olito, C. (2022). Natural selection and the distribution of chromosomal inversion lengths. *Molecular Ecology*, 31(13):3627–3641. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.16091.

Crona, K. (2014). Polytopes, graphs and fitness landscapes. *In Recent Advances in the Theory and Application of Fitness Landscapes*, pages 177–205. Springer.

Crona, K. (2018). Recombination and peak jumping. *Plos one*, 13(3):e0193123.

Crutchfield, J. P. (2003). When evolution is revolution - origins of innovation. *Evolutionary dynamics: exploring the interplay of selection, neutrality, accident, and function*, pages 101–134.

Dalquen, D. A., Anisimova, M., Gonnet, G. H. et Dessimoz, C. (2012). Alfa simulation framework for genome evolution. *Molecular biology and evolution*, 29(4):1115–1123.

Darling, A. E., Miklós, I. et Ragan, M. A. (2008). Dynamics of Genome Rearrangement in Bacterial Populations. *PLOS Genetics*, 4(7):e1000128. Publisher: Public Library of Science.

de Lima Filho, J., Moreira, F., Campos, P. et De Oliveira, V. M. (2012). Adaptive walks on correlated fitness landscapes with heterogeneous connectivities. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(02):P02014.

De Visser, J. A. G., Hermisson, J., Wagner, G. P., Meyers, L. A., Bagheri-Chaichian, H., Blanchard, J. L., Chao, L., Cheverud, J. M., Elena, S. F., Fontana, W. *et al.* (2003). Perspective: evolution and detection of genetic robustness. *Evolution*, 57(9):1959–1972.

De Visser, J. A. G. et Krug, J. (2014). Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics*, 15(7):480–490.

de Waal Malefijt, M. et Charlesworth, B. (1979). A model for the evolution of translocation heterozygosity. *Heredity*, 43(3):315–331.

DiMauro, S. (2001). Lessons from mitochondrial dna mutations. *Seminars in Cell & Developmental Biology*, 12(6):397–405.

Diss, G. et Lehner, B. (2018). The genetic landscape of a physical interaction. *Elife*, 7:e32472.

Dobzhansky, T. (1950). Mendelian populations and their evolution. *The American Naturalist*, 84(819):401–418.

Doolittle, W. F. (2013). Is junk dna bunk? a critique of encode. *Proceedings of the National Academy of Sciences*, 110(14):5294–5300.

Drake, J. W. (1991). A constant rate of spontaneous mutation in dna-based microbes. *Proceedings of the National Academy of Sciences*, 88(16):7160–7164.

Drake, J. W., Charlesworth, B., Charlesworth, D. et Crow, J. F. (1998a). Rates of spontaneous mutation. *Genetics*, 148(4):1667–1686.

Drake, J. W., Charlesworth, B., Charlesworth, D. et Crow, J. F. (1998b). Rates of Spontaneous Mutation. *Genetics*, 148(4):1667–1686.

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

Durrett, R. et Schmidt, D. (2008). Waiting for two mutations: with applications to regulatory sequence evolution and the limits of darwinian evolution. *Genetics*, 180(3): 1501–1509.

Edwards, D., Forster, J. W., Chagné, D. et Batley, J. (2007). What are snps? *In Association mapping in plants*, pages 41–52. Springer.

El Houdaigui, B., Forquet, R., Hindré, T., Schneider, D., Nasser, W., Reverchon, S. et Meyer, S. (2019). Bacterial genome architecture shapes global transcriptional regulation by dna supercoiling. *Nucleic acids research*, 47(11):5648–5657.

Elde, N. C., Child, S. J., Eickbush, M. T., Kitzman, J. O., Rogers, K. S., Shendure, J., Geballe, A. P. et Malik, H. S. (2012). Poxviruses deploy genomic accordions to adapt rapidly against host antiviral defenses. *Cell*, 150(4):831–841.

Erwin, D. H. (2017). The topology of evolutionary novelty and innovation in macroevolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1735): 20160422.

Escarmís, C., Lázaro, E. et Manrubia, S. C. (2006). *Population bottlenecks in quasispecies dynamics*, pages 141–170. Springer Berlin Heidelberg, Berlin, Heidelberg.

Fagundes, N. J., Bisso-Machado, R., Figueiredo, P. I., Varal, M. et Zani, A. L. (2022). What we talk about when we talk about junk dna. *Genome Biology and Evolution*, 14(5):evac055.

Faria, R., Johannesson, K., Butlin, R. K. et Westram, A. M. (2019). Evolving inversions. *Trends in Ecology & Evolution*, 34(3):239–248.

Felsenstein, J. (2005). Theoretical evolutionary genetics joseph felsenstein. *University of Washington, Seattle*.

Fertin, G., Labarre, A., Rusu, I., Vialette, S. et Tannier, E. (2009). *Combinatorics of genome rearrangements*. MIT press.

Filée, J. (2013). Route of ncldv evolution: the genomic accordion. *Current opinion in virology*, 3(5):595–599.

Fischer, S., Bernard, S., Beslon, G. et Knibbe, C. (2014). A model for genome size evolution. *Bulletin of mathematical biology*, 76:2249–2291.

Fisher, R. A. (1930). *The genetical theory of natural selection*. Oxford University Press.

Fletcher, W. et Yang, Z. (2009). Indelible: a flexible simulator of biological sequence evolution. *Molecular biology and evolution*, 26(8):1879–1888.

Fragata, I., Blanckaert, A., Louro, M. A. D., Liberles, D. A. et Bank, C. (2019). Evolution in the light of fitness landscape theory. *Trends in ecology & evolution*, 34(1): 69–82.

Franke, J., Klözer, A., de Visser, J. A. G. et Krug, J. (2011). Evolutionary accessibility of mutational pathways. *PLoS Computational Biology*, 7(8):e1002134.

Franklin, R. E. et Gosling, R. G. (1953). Molecular configuration in sodium thymonucleate. *Nature*, 171:740–741.

Freeling, M., Xu, J., Woodhouse, M. et Lisch, D. (2015). A solution to the c-value paradox and the function of junk dna: the genome balance hypothesis. *Molecular Plant*, 8(6):899–910.

Freese, E. (1959). The difference between spontaneous and base-analogue induced mutations of phage t4. *Proceedings of the National Academy of Sciences*, 45(4):622–633.

Frenoy, A., Taddei, F. et Misevic, D. (2013). Genetic architecture promotes the evolution and maintenance of cooperation. *PLoS Computational Biology*, 9(11):e1003339.

Frost, S. D., Dumaurier, M.-J., Wain-Hobson, S. et Brown, A. J. L. (2001). Genetic drift and within-host metapopulation dynamics of hiv-1 infection. *Proceedings of the National Academy of Sciences*, 98(12):6975–6980.

Gao, Y., Zhao, H., Jin, Y., Xu, X. et Han, G.-Z. (2017). Extent and evolution of gene duplication in dna viruses. *Virus research*, 240:161–165.

García-Arenal, F., Fraile, A. et Malpica, J. M. (2003). Variation and evolution of plant virus populations. *International Microbiology*, 6:225–232.

Gardner, E. J., Sifrim, A., Lindsay, S. J., Prigmore, E., Rajan, D., Danecek, P., Gallone, G., Eberhardt, R. Y., Martin, H. C., Wright, C. F., FitzPatrick, D. R., Firth, H. V. et Hurles, M. E. (2021). Detecting cryptic clinically relevant structural variation in exome-sequencing data increases diagnostic yield for developmental disorders. *The American Journal of Human Genetics*, 108:2186–2194.

Gavrilets, S. (1997). Evolution and speciation on holey adaptive landscapes. *Trends in ecology & evolution*, 12(8):307–312.

Gavrilets, S. (2004). *Fitness landscapes and the origin of species (MPB-41)*. Princeton University Press.

Gil, R. et Latorre, A. (2012). Factors behind junk dna in bacteria. *Genes*, 3(4):634–650.

Gillespie, J. H. (1983). A simple stochastic gene substitution model. *Theoretical Population Biology*, 23(2):202–215.

Gillespie, J. H. (1984). Molecular evolution over the mutational landscape. *Evolution*, 38(5):1116–1129.

Gillespie, J. H. (1991). *The causes of molecular evolution*. Oxford University Press.

Giovannoni, S. J., Cameron Thrash, J. et Temperton, B. (2014). Implications of streamlining theory for microbial ecology. *The ISME journal*, 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. et Mathur, E. J. (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science*, 309(5738):1242–1245.

Goel, N. S., Maitra, S. C. et Montroll, E. W. (1971). On the volterra and other nonlinear models of interacting populations. *Reviews of modern physics*, 43(2):231.

Greene, D. et Crona, K. (2014). The changing geometry of a fitness landscape along an adaptive walk. *PLoS Computational Biology*, 10(5):e1003520.

Grewal, R. K., Sinha, S. et Roy, S. (2018). Topologically inspired walks on randomly connected landscapes with correlated fitness. *Frontiers in Physics*, 6:138.

Griffiths, A. J. F., Wessler, S. R., Carroll, S. B. et Doebley, J. (2012). *Introduction to genetic analysis*. publisher: W. H. Freeman.

Grimm, V. (1999). Ten years of individual-based modelling in ecology: what have we learned and what could we learn in the future? *Ecological modelling*, 115(2-3):129–148.

Grinnell, J. (1924). Geography and evolution. *Ecology*, 5(3):225–229.

Gulick, J. T. (1887). *Divergent evolution through cumulative segregation*. Smithsonian Report.

Guo, Y., Vucelja, M. et Amir, A. (2019). Stochastic tunneling across fitness valleys can give rise to a logarithmic long-term fitness trajectory. *Science Advances*, 5(7):eaav3842.

Haldane, J. B. S. (1937). A dialectical account of evolution. *Science & Society*, pages 473–486.

Haller, B. C. et Messer, P. W. (2017). Slim 2: flexible, interactive forward genetic simulations. *Molecular biology and evolution*, 34(1):230–240.

Hanlon, V. C. T., Lansdorp, P. M. et Guryev, V. (2022). A survey of current methods to detect and genotype inversions. *Human Mutation*, 43(11):1576–1589.

Hartfield, M. et Keightley, P. D. (2012). Current hypotheses for the evolution of sex and recombination. *Integrative zoology*, 7(2):192–209.

Hastings, P., Ira, G. et Lupski, J. R. (2009). A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS genetics*, 5(1):e1000327.

He, Y., Tian, S. et Tian, P. (2019). Fundamental asymmetry of insertions and deletions in genomes size evolution. *Journal of Theoretical Biology*, 482:109983.

Hill, G. E. (2020). Genetic hitchhiking, mitonuclear coadaptation, and the origins of mt dna barcode gaps. *Ecology and evolution*, 10(17):9048–9059.

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

Ho, S. S., Urban, A. E. et Mills, R. E. (2020). Structural variation in the sequencing era. *Nature Reviews Genetics*, 21(3):171–189.

Hochberg, M. E., Marquet, P. A., Boyd, R. et Wagner, A. (2017). Innovation: an emerging focus from cells to societies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1735):20160414.

Hoffmann, A. A. et Rieseberg, L. H. (2008). Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation? *Annual Review of Ecology, Evolution, and Systematics*, 39:21–42.

Hoffmann, A. A., Sgrò, C. M. et Weeks, A. R. (2004). Chromosomal inversion polymorphisms and adaptation. *Trends in Ecology & Evolution*, 19(9):482–488.

Holland, J. H. (1992). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press.

Huang, K. et Rieseberg, L. H. (2020). Frequency, origins, and evolutionary role of chromosomal inversions in plants. *Frontiers in Plant Science*, 11:296.

Hwang, S., Schmiegelt, B., Ferretti, L. et Krug, J. (2018). Universality classes of interaction structures for nk fitness landscapes. *Journal of Statistical Physics*, 172(1): 226–278.

Ispolatov, I., Alekseeva, E. et Doebeli, M. (2019). Competition-driven evolution of organismal complexity. *PLoS computational biology*, 15(10):e1007388.

Iwasa, Y., Michor, F. et Nowak, M. A. (2004). Stochastic tunnels in evolutionary dynamics. *Genetics*, 166(3):1571–1579.

Jain, K. et Krug, J. (2007). Deterministic and stochastic regimes of asexual evolution on rugged fitness landscapes. *Genetics*, 175(3):1275–1288.

Jensen-Seaman, M. I., Furey, T. S., Payseur, B. A., Lu, Y., Roskin, K. M., Chen, C.-F., Thomas, M. A., Haussler, D. et Jacob, H. J. (2004). Comparative Recombination Rates in the Rat, Mouse, and Human Genomes. *Genome Research*, 14(4):528–538.

Kaback, D. B., Guacci, V., Barber, D. et Mahon, J. W. (1992). Chromosome Size-Dependent Control of Meiotic Recombination. *Science*, 256(5054):228–232. Publisher: American Association for the Advancement of Science.

Kalhor, R., Beslon, G., Lafond, M. et Scornavacca, C. (2023). Classifying the post-duplication fate of paralogous genes. *In RECOMB International Workshop on Comparative Genomics*, pages 1–18. Springer.

Kang, M., Wang, J. et Huang, H. (2015). Nitrogen limitation as a driver of genome size evolution in a group of karst plants. *Scientific Reports*, 5(1):11636.

Kang, S.-H., Atallah, O. O., Sun, Y.-D. et Folimonova, S. Y. (2018). Functional diversification upon leader protease domain duplication in the citrus tristeza virus genome: Role of rna sequences and the encoded proteins. *Virology*, 514:192–202.

Kara, E., Kiely, A. P., Proukakis, C., Giffin, N., Love, S., Hehir, J., Rantell, K., Pandraud, A., Hernandez, D. G., Nacheva, E., Pittman, A. M., Nalls, M. A., Singleton, A. B., Revesz, T., Bhatia, K. P., Quinn, N., Hardy, J., Holton, J. L. et Houlden, H. (2014). A 6.4 Mb Duplication of the α-Synuclein Locus Causing Frontotemporal Dementia and Parkinsonism: Phenotype-Genotype Correlations. *JAMA Neurology*, 71(9):1162–1171.

Katju, V. et Bergthorsson, U. (2013). Copy-number changes in evolution: rates, fitness effects and adaptive significance. *Frontiers in genetics*, 4:273.

Kauffman, S. et Levin, S. (1987). Towards a general theory of adaptive walks on rugged landscapes. *Journal of theoretical Biology*, 128(1):11–45.

Kauffman, S. A. (1993). *The origins of order: Self-organization and selection in evolution*. Oxford University Press, USA.

Kauffman, S. A. et Weinberger, E. D. (1989). The nk model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of Theoretical Biology*, 141(2):211–245.

Kaznatcheev, A. (2019). Computational complexity as an ultimate constraint on evolution. *Genetics*, 212(1):245–265.

Kelkar, Y. D. et Ochman, H. (2012). Causes and Consequences of Genome Expansion in Fungi. *Genome Biology and Evolution*, 4(1):13–23.

Kessinger, T. et Cleve, J. V. (2018). Genetic draft and valley crossing. *bioRxiv*.

Kimura, M. et Crow, J. F. (1963). The measurement of effective population number. *Evolution*, 17(3):279–288.

Kirkpatrick, M. (2010). How and why chromosome inversions evolve. *PLoS biology*, 8(9):e1000501.

Klug, A., Park, S.-C. et Krug, J. (2019). Recombination and mutational robustness in neutral fitness landscapes. *PLoS Computational Biology*, 15(8):e1006884.

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

Knibbe, C., Mazet, O., Chaudier, F., Fayard, J.-M. et Beslon, G. (2007b). Evolutionary coupling between the deleteriousness of gene mutations and the amount of non-coding sequences. *J. Theor. Biol.*, 244(4):621–630.

Knibbe, C. et Parsons, D. (2014). What happened to my genes? insights on gene family dynamics from digital genetics experiments. *In Artificial Life Conference Proceedings*, pages 33–40. MIT Press.

Kolesnikov, A. et Gerasimov, E. (2012). Diversity of mitochondrial genome organization. *Biochemistry (Moscow)*, 77(13):1424–1435.

Koonin, E. V. (2011). Are there laws of genome evolution? *PLoS Computational Biology*, 7(8):e1002173.

Korneev, S. et O'Shea, M. (2002). Evolution of nitric oxide synthase regulatory genes by dna inversion. *Molecular Biology and Evolution*, 19(8):1228–1233.

Kuo, C.-H. et Ochman, H. (2009). Deletional Bias across the Three Domains of Life. *Genome Biology and Evolution*, 1:145–152.

Lalejini, A., Wiser, M. J. et Ofria, C. (2017). Gene duplications drive the evolution of complex traits and regulation. *In Artificial Life Conference Proceedings*, pages 257–264. MIT Press.

Le Guillou-Guillemette, H., Pivert, A., Bouthry, E., Henquell, C., Petsaris, O., Ducancelle, A., Veillon, P., Vallet, S., Alain, S., Thibault, V. *et al.* (2017). Natural non-homologous recombination led to the emergence of a duplicated v3-ns5a region in hcv-1b strains associated with hepatocellular carcinoma. *Plos one*, 12(4):e0174651.

Leitch, I. J. et Leitch, A. R. (2012). Genome size diversity and evolution in land plants. *In Plant genome diversity Volume 2: Physical structure, behaviour and evolution of plant genomes*, pages 307–322. Springer.

Lenski, R. E., Ofria, C., Pennock, R. T. et Adami, C. (2003). The evolutionary origin of complex features. *Nature*, 423(6936):139–144.

Leushkin, E. V., Bazykin, G. A. et Kondrashov, A. S. (2012). Insertions and deletions trigger adaptive walks in drosophila proteins. *Proceedings of the Royal Society B: Biological Sciences*, 279(1740):3075–3082.

Liard, V., Parsons, D. P., Rouzaud-Cornabas, J. et Beslon, G. (2020a). The Complexity Ratchet: Stronger than Selection, Stronger than Evolvability, Weaker than Robustness. *Artificial Life*, 26(1):38–57.

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

Lobkovsky, A. E., Wolf, Y. I. et Koonin, E. V. (2011). Predictability of evolutionary trajectories in fitness landscapes. *PLoS Computational Biology*, 7(12):e1002302.

Loeb, L. A. (1989). Endogenous carcinogenesis: molecular oncology into the twenty-first centurypresidential address. *Cancer research*, 49(20):5489–5496.

Loewenthal, G., Wygoda, E., Nagar, N., Glick, L., Mayrose, I. et Pupko, T. (2022). The evolutionary dynamics that retain long neutral genomic sequences in face of indel deletion bias: a model and its application to human introns. *Open Biology*, 12(12):220223.

Lynch, M. (2010). Evolution of the mutation rate. *Trends in Genetics*, 26(8):345–352. Publisher: Elsevier.

Lynch, M., Ackerman, M. S., Gout, J.-F., Long, H., Sung, W., Thomas, W. K. et Foster, P. L. (2016). Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, 17(11):704–714.

Lynch, M. et Conery, J. S. (2003). The Origins of Genome Complexity. *Science*, 302(5649):1401–1404.

Lynch, M., Sung, W., Morris, K., Coffey, N., Landry, C. R., Dopman, E. B., Dickinson, W. J., Okamoto, K., Kulkarni, S., Hartl, D. L. et Thomas, W. K. (2008). A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proceedings of the National Academy of Sciences*, 105(27):9272–9277.

Maisnier-Patin, S., Berg, O. G., Liljas, L. et Andersson, D. I. (2002). Compensatory adaptation to the deleterious effect of antibiotic resistance in salmonella typhimurium. *Molecular microbiology*, 46(2):355–366.

Manrubia, S. C. (2012). Modelling viral evolution and adaptation: challenges and rewards. *Current opinion in virology*, 2(5):531–537.

Masel, J. (2006). Cryptic genetic variation is enriched for potential adaptations. *Genetics*, 172(3):1985–1991.

Mauldin, M. L. (1984). Maintaining diversity in genetic search. *In AAAI Conference on Artificial Intelligence*.

Maynard Smith, J. (1970). Natural selection and the concept of a protein space. *Nature*, 225(5232):563–564.

Mérot, C., Oomen, R. A., Tigano, A. et Wellenreuther, M. (2020). A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends in Ecology & Evolution*, 35(7):561–572.

Merrikh, C. N. et Merrikh, H. (2018). Gene inversion potentiates bacterial evolvability and virulence. *Nature Communications*, 9(1):1–10.

Mirjalili, S., Dong, J. S. et Lewis, A. (2020). Nature-inspired optimizers. *Studies in Computational Intelligence*, 811:7–20.

Misevic, D., Frenoy, A., Lindner, A. B. et Taddei, F. (2015). Shape matters: Lifecycle of cooperative patches promotes cooperation in bulky populations. *Evolution*, 69(3):788–802.

Mohlhenrich, E. R. et Mueller, R. L. (2016). Genetic drift and mutational hazard in the evolution of salamander genomic gigantism. *Evolution*, 70(12):2865–2878.

Morley, V. J. et Turner, P. E. (2017). Dynamics of molecular evolution in rna virus populations depend on sudden versus gradual environmental change. *Evolution*, 71(4): 872–883.

Mueller, R. L. et Jockusch, E. L. (2018). Jumping genomic gigantism. *Nature Ecology & Evolution*, 2(11):1687–1688.

Musumeci, O., Andreu, A. L., Shanske, S., Bresolin, N., Comi, G. P., Rothstein, R., Schon, E. A. et DiMauro, S. (2000). Intragenic inversion of mtdna: a new type of pathogenic mutation in a patient with mitochondrial myopathy. *The American Journal of Human Genetics*, 66(6):1900–1904.

Nattestad, M., Goodwin, S., Ng, K., Baslan, T., Sedlazeck, F. J., Rescheneder, P., Garvin, T., Fang, H., Gurtowski, J., Hutton, E., Tseng, E., Chin, C.-S., Beck, T., Sundaravadanam, Y., Kramer, M., Antoniou, E., McPherson, J. D., Hicks, J., McCombie, W. R. et Schatz, M. C. (2018). Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Research*, 28(8):1126–1135.

Noor, M. A. F., Grams, K. L., Bertucci, L. A. et Reiland, J. (2001). Chromosomal inversions and the reproductive isolation of species. *Proceedings of the National Academy of Sciences*, 98(21):12084–12088.

Novick, A. et Szilard, L. (1950). Experiments with the chemostat on spontaneous mutations of bacteria. *Proceedings of the National Academy of Sciences*, 36(12):708–719.

Nowak, S. et Krug, J. (2015). Analysis of adaptive walks on nk fitness landscapes with different interaction schemes. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(6):P06014.

Ohno, S. (1970a). The enormous diversity in genome sizes of fish as a reflection of nature's extensive experiments with gene duplication. *Transactions of the American Fisheries Society*, 99(1):120–130.

Ohno, S. (1970b). *Evolution by gene duplication*. Springer Science & Business Media.

Olabode, A. S., Mumby, M. J., Wild, T. A., Muñoz-Baena, L., Dikeakos, J. D. et Poon, A. F. (2023). Phylogenetic reconstruction and functional characterization of the ancestral nef protein of primate lentiviruses. *Molecular Biology and Evolution*, 40(8):msad164.

Olson, C. A., Wu, N. C. et Sun, R. (2014). A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Current biology*, 24(22):2643–2651.

Østman, B. et Adami, C. (2014). Predicting evolution and visualizing high-dimensional fitness landscapes. *In Recent advances in the theory and application of fitness landscapes*, pages 509–526. Springer.

Otto, S. P. et Barton, N. H. (2001). Selection for recombination in small populations. *Evolution*, 55(10):1921–1931.

ONeill, M. et Brabazon, A. (2019). Mutational robustness and structural complexity in grammatical evolution. *In 2019 IEEE Congress on Evolutionary Computation (CEC)*, pages 1338–1344. IEEE.

Parra, G. I. (2019). Emergence of norovirus strains: A tale of two genes. *Virus evolution*, 5(2):vez048.

Parsons, D. (2011). *Indirect Selection in Darwinian Evolution: Mechanisms and Implications*. Phd thesis, INSA Lyon.

Parsons, D. P., Knibbe, C. et Beslon, G. (2010). Importance of the rearrangement rates on the organization of transcription. *In Proceedings of Artificial Life XII*, pages 479–486.

Periwal, V. et Scaria, V. (2015). Insights into structural variations and genome rearrangements in prokaryotic genomes. *Bioinformatics*, 31(1):1–9.

Petrov, D. A. (2002). Mutational equilibrium model of genome size evolution. *Theoretical population biology*, 61(4):531–544.

Pita, J. S., De Miranda, J. R., Schneider, W. L. et Roossinck, M. J. (2007). Environment determines fidelity for an rna virus replicase. *Journal of virology*, 81(17): 9072–9077.

Quandt, E. M., Gollihar, J., Blount, Z. D., Ellington, A. D., Georgiou, G. et Barrick, J. E. (2015). Fine-tuning citrate synthase flux potentiates and refines metabolic innovation in the lenski evolution experiment. *Elife*, 4:e09696.

Raeside, C., Gaffé, J., Deatherage, D. E., Tenaillon, O., Briska, A. M., Ptashkin, R. N., Cruveiller, S., Médigue, C., Lenski, R. E., Barrick, J. E. *et al.* (2014). Large chromosomal rearrangements during a long-term evolution experiment with escherichia coli. *MBio*, 5(5):e01377–14.

Ranz, J. M., Maurin, D., Chan, Y. S., Von Grotthuss, M., Hillier, L. W., Roote, J., Ashburner, M. et Bergman, C. M. (2007). Principles of genome evolution in the drosophila melanogaster species group. *PLoS Biology*, 5(6):e152.

Rao, R. S. P., Ahsan, N., Xu, C., Su, L., Verburgt, J., Fornelli, L., Kihara, D. et Xu, D. (2021). Evolutionary dynamics of indels in sars-cov-2 spike glycoprotein. *Evolutionary Bioinformatics*, 17:11769343211064616.

Rinn, J. L. et Chang, H. Y. (2012). Genome Regulation by Long Noncoding RNAs. *Annual Review of Biochemistry*, 81(1):145–166.

Rocha, E. P. C. (2006). Inference and Analysis of the Relative Stability of Bacterial Chromosomes. *Molecular Biology and Evolution*, 23(3):513–522.

Rutten, J. P., Hogeweg, P. et Beslon, G. (2019). Adapting the engine to the fuel: mutator populations can reduce the mutational load by reorganizing their genome structure. *BMC Evolutionary Biology*, 19(1):191.

Sakai, A., Nakanishi, M., Yoshiyama, K. et Maki, H. (2006). Impact of reactive oxygen species on spontaneous mutagenesis in escherichia coli. *Genes to Cells*, 11(7): 767–778.

Salverda, M. L., Dellus, E., Gorter, F. A., Debets, A. J., Van Der Oost, J., Hoekstra, R. F., Tawfik, D. S. et de Visser, J. A. G. (2011). Initial mutations direct alternative pathways of protein evolution. *PLoS Genetics*, 7(3):e1001321.

Sarkar, S. (1990). On adaptation: a reduction of the kauffman-levin model to a problem in graph theory and its consequences. *Biology and Philosophy*, 5(2):127–148.

Schobel, S. A., Stucker, K. M., Moore, M. L., Anderson, L. J., Larkin, E. K., Shankar, J., Bera, J., Puri, V., Shilts, M. H., Rosas-Salazar, C. *et al.* (2016). Respiratory syncytial virus whole-genome sequencing identifies convergent evolution of sequence duplication in the c-terminus of the g gene. *Scientific reports*, 6(1):26311.

Schrag, S. J., Perrot, V. et Levin, B. R. (1997). Adaptation to the fitness costs of antibiotic resistance in escherichia coli. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 264(1386):1287–1291.

Schrider, D. R., Houle, D., Lynch, M. et Hahn, M. W. (2013). Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. *Genetics*, 194(4):937–954.

Sella, G. et Hirsh, A. E. (2005). The application of statistical physics to evolutionary biology. *Proceedings of the National Academy of Sciences*, 102(27):9541–9546.

Serra, M. C. et Haccou, P. (2007). Dynamics of escape mutants. *Theoretical Population Biology*, 72(1):167–178.

Silva, F. J., Latorre, A. et Moya, A. (2001). Genome size reduction through multiple events of gene disintegration in buchnera aps. *TRENDS in Genetics*, 17(11):615–618.

Sloan, D. B., Alverson, A. J., Chuckalovcak, J. P., Wu, M., McCauley, D. E., Palmer, J. D. et Taylor, D. R. (2012). Rapid Evolution of Enormous, Multichromosomal Genomes in Flowering Plant Mitochondria with Exceptionally High Mutation Rates. *PLOS Biology*, 10(1):e1001241.

Smith, D. R., Hamaji, T., Olson, B. J., Durand, P. M., Ferris, P., Michod, R. E., Featherston, J., Nozaki, H. et Keeling, P. J. (2013). Organelle Genome Complexity Scales Positively with Organism Size in Volvocine Green Algae. *Molecular Biology and Evolution*, 30(4):793–797.

Smith, J. M. (1970). Natural selection and the concept of a protein space. *Nature*, 225(5232):563–564.

Smith, J. M. (1978). *The evolution of sex*, volume 4. Cambridge University Press Cambridge.

Solé, R. et Elena, S. F. (2018). *Viruses as complex adaptive systems*, volume 15. Princeton University Press.

Solow, D., Burnetas, A., Roeder, T. et Greenspan, N. S. (1999a). Evolutionary consequences of selected locus-specific variations in epistasis and fitness contribution in kauffman's nk model. *Journal of Theoretical Biology*, 196(2):181–196.

Solow, D., Burnetas, A., Tsai, M.-C. et Greenspan, N. S. (1999b). Understanding and attenuating the complexity catastrophe in kauffman's nk model of genome evolution. *Complexity*, 5(1):53–66.

Stadler, B. M. et Stadler, P. F. (2002). Generalized topological spaces in evolutionary theory and combinatorial chemistry. *Journal of Chemical Information and Computer Sciences*, 42(3):577–585.

Stadler, B. M., Stadler, P. F., Wagner, G. P. et Fontana, W. (2001). The topology of the possible: Formal spaces underlying patterns of evolutionary change. *Journal of Theoretical Biology*, 213(2):241–274.

Stadler, P. F. (1995). Towards a theory of landscapes. *In* López-Peña, R., Waelbroeck, H., Capovilla, R., García-Pelayo, R. et Zertuche, F., éditeurs : *Complex systems and binary networks*, Lectures Notes in Physics, pages 78–163. Springer.

Stapleford, K. A., Moratorio, G., Henningsson, R., Chen, R., Matheus, S., Enfissi, A., Weissglas-Volkov, D., Isakov, O., Blanc, H., Mounce, B. C. *et al.* (2016). Whole-genome sequencing analysis from the chikungunya virus caribbean outbreak reveals novel evolutionary genomic elements. *PLoS neglected tropical diseases*, 10(1):e0004402.

Starr, T. N. et Thornton, J. W. (2016). Epistasis in protein evolution. *Protein science*, 25(7):1204–1218.

Sturtevant, A. H. (1921). A case of rearrangement of genes in drosophila. *Proceedings of the National Academy of Sciences*, 7(8):235–237.

Sung, W., Ackerman, M. S., Dillon, M. M., Platt, T. G., Fuqua, C., Cooper, V. S. et Lynch, M. (2016). Evolution of the Insertion-Deletion Mutation Rate Across the Tree of Life. *G3 Genes|Genomes|Genetics*, 6(8):2583–2591.

Szendro, I. G., Schenk, M. F., Franke, J., Krug, J. et de Visser, J. A. G. M. (2013). Quantitative analyses of empirical fitness landscapes. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(01):P01005.

Tarkhnishvili, D., Yanchukov, A. et Böhne, A. (2023). Advantages, limitations, and evolutionary constraints of asexual reproduction: An empirical approach. *Frontiers in Ecology and Evolution*, 11:1184306.

Tisza, M. J., Pastrana, D. V., Welch, N. L., Stewart, B., Peretti, A., Starrett, G. J., Pang, Y.-Y. S., Krishnamurthy, S. R., Pesavento, P. A., McDermott, D. H. *et al.* (2020). Discovery of several thousand highly diverse circular dna viruses. *Elife*, 9:e51971.

Trujillo, L., Banse, P. et Beslon, G. (2022). Getting higher on rugged landscapes: Inversion mutations open access to fitter adaptive peaks in nk fitness landscapes. *PLoS Computational Biology*, 18(10):e1010647.

Turner, P. E. et Elena, S. F. (2000). Cost of host radiation in an rna virus. *Genetics*, 156(4):1465–1470.

Uyeda, J. C., Hansen, T. F., Arnold, S. J. et Pienaar, J. (2011). The million-year wait for macroevolutionary bursts. *Proceedings of the National Academy of Sciences*, 108(38):15908–15913.

Vakhrusheva, A. A., Kazanov, M. D., Mironov, A. A. et Bazykin, G. A. (2011). Evolution of prokaryotic genes by shift of stop codons. *Journal of molecular evolution*, 72:138–146.

Van Egeren, D., Madsen, T. et Michor, F. (2018). Fitness variation in isogenic populations leads to a novel evolutionary mechanism for crossing fitness valleys. *Communications biology*, 1(1):1–9.

Varenne, F. et Silberstein, M. (2013). Modèles et simulations : pluriformaliser, simuler, remathématiser. *In* Franck Varenne, M. S., éditeur : *Modéliser & simuler. Epistémologies et pratiques de la modélisation et de la simulation*, volume 1, chapitre 10, pages 299–328. Editions Matériologiques, Oxford.

Vie, A., Kleinnijenhuis, A. M. et Farmer, D. J. (2020). Qualities, challenges and future of genetic algorithms: a literature review. *arXiv preprint arXiv:2011.05277*.

Wagner, A. (2011). The low cost of recombination in creating novel phenotypes: Recombination can create new phenotypes while disrupting well-adapted phenotypes much less than mutation. *BioEssays*, 33(8):636–646.

Wala, J. A., Bandopadhayay, P., Greenwald, N. F., O'Rourke, R., Sharpe, T., Stewart, C., Schumacher, S., Li, Y., Weischenfeldt, J., Yao, X. *et al.* (2018). Svaba: genome-wide detection of structural variants and indels by local assembly. *Genome research*, 28(4):581–591.

Walsh, B. (2003). *Population-genetic models of the fates of duplicate genes*, pages 279–294. Springer Netherlands, Dordrecht.

Wang, J., Fan, H. C., Behr, B. et Quake, S. R. (2012). Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell*, 150(2):402–412.

Wang, Y., Diaz Arenas, C., Stoebel, D. M., Flynn, K., Knapp, E., Dillon, M. M., Wünsche, A., Hatcher, P. J., Moore, F. B.-G., Cooper, V. S. *et al.* (2016). Benefit of transferred mutations is better predicted by the fitness of recipients than by their ecological or genetic relatedness. *Proceedings of the National Academy of Sciences*, 113(18):5047–5052.

Waples, R. S. (2010). Spatial-temporal stratifications in natural populations and how they affect understanding and estimation of effective population size. *Molecular Ecology Resources*, 10(5):785–796.

Watson, J. D. et Crick, F. H. (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

Wei, X. et Zhang, J. (2019). Patterns and mechanisms of diminishing returns from beneficial mutations. *Molecular biology and evolution*, 36(5):1008–1021.

Weinberger, E. D. (1991). Local properties of kauffmans n-k model: A tunably rugged energy landscape. *Physical Review A*, 44(10):6399.

Weinreich, D. M. et Chao, L. (2005). Rapid evolutionary escape by large populations from local fitness peaks is likely in nature. *Evolution*, 59(6):1175–1182.

Weissman, D. B., Desai, M. M., Fisher, D. S. et Feldman, M. W. (2009). The rate at which asexual populations cross fitness valleys. *Theoretical Population Biology*, 75(4):286–300.

Weissman, D. B., Feldman, M. W. et Fisher, D. S. (2010). The rate of fitness-valley crossing in sexual populations. *Genetics*, 186(4):1389–1410.

Wellenreuther, M. et Bernatchez, L. (2018). Eco-evolutionary genomics of chromosomal inversions. *Trends in ecology & evolution*, 33(6):427–440.

Wellenreuther, M., Mérot, C., Berdan, E. et Bernatchez, L. (2019). Going beyond snps: The role of structural genomic variants in adaptive evolution and species diversification. *Molecular ecology*, 28(6):1203–1209.

Whigham, P. A. *et al.* (1995). Grammatically-based genetic programming. *In Proceedings of the workshop on genetic programming: from theory to real-world applications*, volume 16, pages 33–41. Citeseer.

Wilke, C. O. (2001). Adaptive evolution on neutral networks. *Bulletin of mathematical biology*, 63(4):715–730.

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

Willemsen, A., Zwart, M. P., Tromas, N., Majer, E., Daròs, J.-A. et Elena, S. F. (2016). Multiple barriers to the evolution of alternative gene orders in a positive-strand rna virus. *Genetics*, 202(4):1503–1521.

Wiser, M. J., Dolson, E. L., Vostinar, A., Lenski, R. E. et Ofria, C. (2018). The boundedness illusion: Asymptotic projections from early evolution underestimate evolutionary potential. *PeerJ Preprints*, 6:e27246v2.

Wiser, M. J., Ribeck, N. et Lenski, R. E. (2013). Long-term dynamics of adaptation in asexual populations. *Science*, 342(6164):1364–1367.

Wolfe, K. H. et Li, W.-H. (2003). Molecular evolution meets the genomics revolution. *Nature Genetics*, 33(3):255–265.

Woods, R. J., Barrick, J. E., Cooper, T. F., Shrestha, U., Kauth, M. R. et Lenski, R. E. (2011). Second-order selection for evolvability in a large escherichia coli population. *Science*, 331(6023):1433–1436.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proceedings of the Sixth Annual Congress of Genetics*.

Wright, S. (1938). The distribution of gene frequencies under irreversible mutation. *Proceedings of the National Academy of Sciences*, 24(7):253–259.

Xue, J. R., Mackay-Smith, A., Mouri, K., Garcia, M. F., Dong, M. X., Akers, J. F., Noble, M., Li, X., Consortium, Z., Lindblad-Toh, K. *et al.* (2023). The functional and evolutionary impacts of human-specific deletions in conserved elements. *Science*, 380(6643):eabn2253.

Yancopoulos, S., Attie, O. et Friedberg, R. (2005). Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 21(16): 3340–3346.

Yoon, J.-H., Abdelmohsen, K. et Gorospe, M. (2013). Posttranscriptional gene regulation by long noncoding rna. *Journal of molecular biology*, 425(19):3723–3730.

Yubero, P., Manrubia, S. et Aguirre, J. (2017). The space of genotypes is a network of networks: implications for evolutionary and extinction dynamics. *Scientific Reports*, 7(1):1–12.

Zagorski, M., Burda, Z. et Waclaw, B. (2016). Beyond the hypercube: evolutionary accessibility of fitness landscapes with realistic mutational networks. *PLoS Computational Biology*, 12(12):e1005218.

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in ecology & evolution*, 18(6):292–298.

Zheng, J., Payne, J. L. et Wagner, A. (2019). Cryptic genetic variation accelerates evolution by opening access to diverse adaptive peaks. *Science*, 365(6451):347–353.

Abstract en français

L'évolution telle qu'elle a été décrite par Darwin est un processus simple qui aboutit à une extrême complexité. En effet, étudier l'évolution biologique aujourd'hui correspond à étudier un phénomène allant d'échelles nanométriques à des échelles planétaires. En plus de cela, le processus est aussi affecté par des biais dus à la méthode d'écriture et de conservation de l'information. Finalement, il faut rappeler que chaque changement évolutif a pour origine une mutation, qui est un évènement aléatoire, et que la survie des mutants est, elle aussi, un processus aléatoire. Face à une telle complexité, il est nécessaire de réduire le champ d'étude pour espérer aboutir à une compréhension. Que ce soit avec des modèles expérimentaux, comme les boites de pétri, avec des modèles formels, comme les équations différentielles, ou avec des modèles computationnels, par exemple des simulations, toutes les simplifications sont bonnes à prendre pour décortiquer l'évolution. Parmi ces simplifications, il est courant de ne considérer que certains types de mutation. En particulier, une simplification logique des modèles d'évolution est souvent d'ignorer les réarrangements chromosomiques (ces mutations qui réorganisent et réassemblent l'ADN et qui sont souvent létales pour l'organisme qui les porte). D'autant plus que jusqu'à récemment, les séquençages d'ADN réalisés n'étaient pas adaptés à les repérer. Dans cette thèse, nous allons montrer qu'en incluant les réarrangements, bien que les modèles obtenus soient plus complexes, il est possible d'en tirer une connaissance. Nous utiliserons des méthodes algorithmiques pour montrer que non seulement l'étude des réarrangements chromosomiques est non seulement utile, mais nécessaire sur de nombreuses questions liées à l'évolution. Par exemple, sur les ressorts de l'évolution à long terme, puisquils permettent une amélioration et de nouvelles opportunités d'évolution. Mais aussi, pour expliquer les dynamiques dévolution par à-coups, ainsi que la maintenance de segments non codants dans les génomes.

Abstract in English

Evolution as described by Darwin is a simple process that leads to extreme complexity. Indeed, the study of biological evolution today means studying a phenomenon that ranges from nanometric to planetary scales. On top of this, the process is also subject to biases due to the way in which information is written and preserved. Finally, it should be remembered that every evolutionary change has its origin in a random event, the mutation, and that the survival of mutants is also a random process. In the face of such complexity, it is necessary to narrow the field of study if we are to achieve any understanding. From experimental models such as petri dishes, formal models such as differential equations, or computational models such as simulations, all simplifications are welcomed to dissect evolution. Among these simplifications, it is common to consider only some types of mutation. In particular, a natural simplification of evolutionary models is often to ignore chromosomal rearrangements (mutations that reorganize and reassemble DNA, and which are often lethal to the organism carrying them). Especially since, until recently, DNA sequencing was not well suited to detecting them. In this thesis, we will show that by including rearrangements, although the models obtained are more complex, knowledge can be gained from them. We will use algorithmic methods to show that the study of chromosomal rearrangements is not only useful, but necessary for many questions related to evolution. For example, on the drivers of long-term evolution, since they allow for improvement and new evolutionary opportunities. But also, to explain the alternance of burst and stasis in evolution, as well as the maintenance of non-coding segments in genomes.

FOLIO ADMINISTRATIF

THESE DE L'INSA LYON, MEMBRE DE L'UNIVERSITE DE LYON

NOM : BANSE
(avec précision du nom de jeune fille, le cas échéant)

DATE de SOUTENANCE : 18/12/2023

Prénoms : Paul

TITRE : Evolution beyond substitutions: Computational modeling of the impact of chromosomal rearrangements on evolutionary dynamics

NATURE : Doctorat

Numéro d'ordre : 2023ISAL0129

École doctorale : InfoMaths

Spécialité : Informatique

RESUME :
L'évolution telle qu'elle a été décrite par Darwin est un processus simple qui aboutit à une extrême complexité. En effet, étudier l'évolution biologique aujourd'hui correspond à étudier un phénomène allant d'échelles nanométriques à des échelles planétaires. En plus de cela, le processus est aussi affecté par des biais dus à la méthode d'écriture et de conservation de l'information. Finalement, il faut rappeler que chaque changement évolutif a pour origine une mutation, qui est un évènement aléatoire, et que la survie des mutants est, elle aussi, un processus aléatoire. Face à une telle complexité, il est nécessaire de réduire le champ d'étude pour espérer aboutir à une compréhension. Que ce soit avec des modèles expérimentaux, comme les boites de pétri, avec des modèles formels, comme les équations différentielles, ou avec des modèles computationnels, par exemple des simulations, toutes les simplifications sont bonnes à prendre pour décortiquer l'évolution. Parmi ces simplifications, il est courant de ne considérer que certains types de mutation. En particulier, ignorer les réarrangements chromosomiques, ces mutations qui réorganisent et réassemblent l'ADN et qui est souvent létales pour l'organisme qui les porte, est souvent considéré comme une simplification logique des modèles d'évolution. D'autant plus que jusqu'à récemment, les séquençages d'ADN réalisés n'étaient pas adaptés à les repérer. Dans cette thèse, nous allons montrer qu'en incluant les réarrangements, bien que les modèles obtenus soient plus complexes, il est possible d'en tirer une connaissance. Nous utiliserons des méthodes algorithmiques pour montrer que non seulement l'étude des réarrangements chromosomiques est non seulement utile, mais nécessaire sur de nombreuses questions liées à l'évolution. Par exemple, sur les ressorts de l'évolution à long terme, puisqu'ils permettent une amélioration et de nouvelles opportunités d'évolution. Mais aussi, pour expliquer les dynamiques d'évolution par à-coups, ainsi que la maintenance de segments non codants dans les génomes.

MOTS-CLÉS : Évolution génomique, réarrangements chromosomiques, modélisation

Laboratoire (s) de recherche : LIRIS

Directeur de thèse: Guillaume Beslon

Président de jury :

Composition du jury :