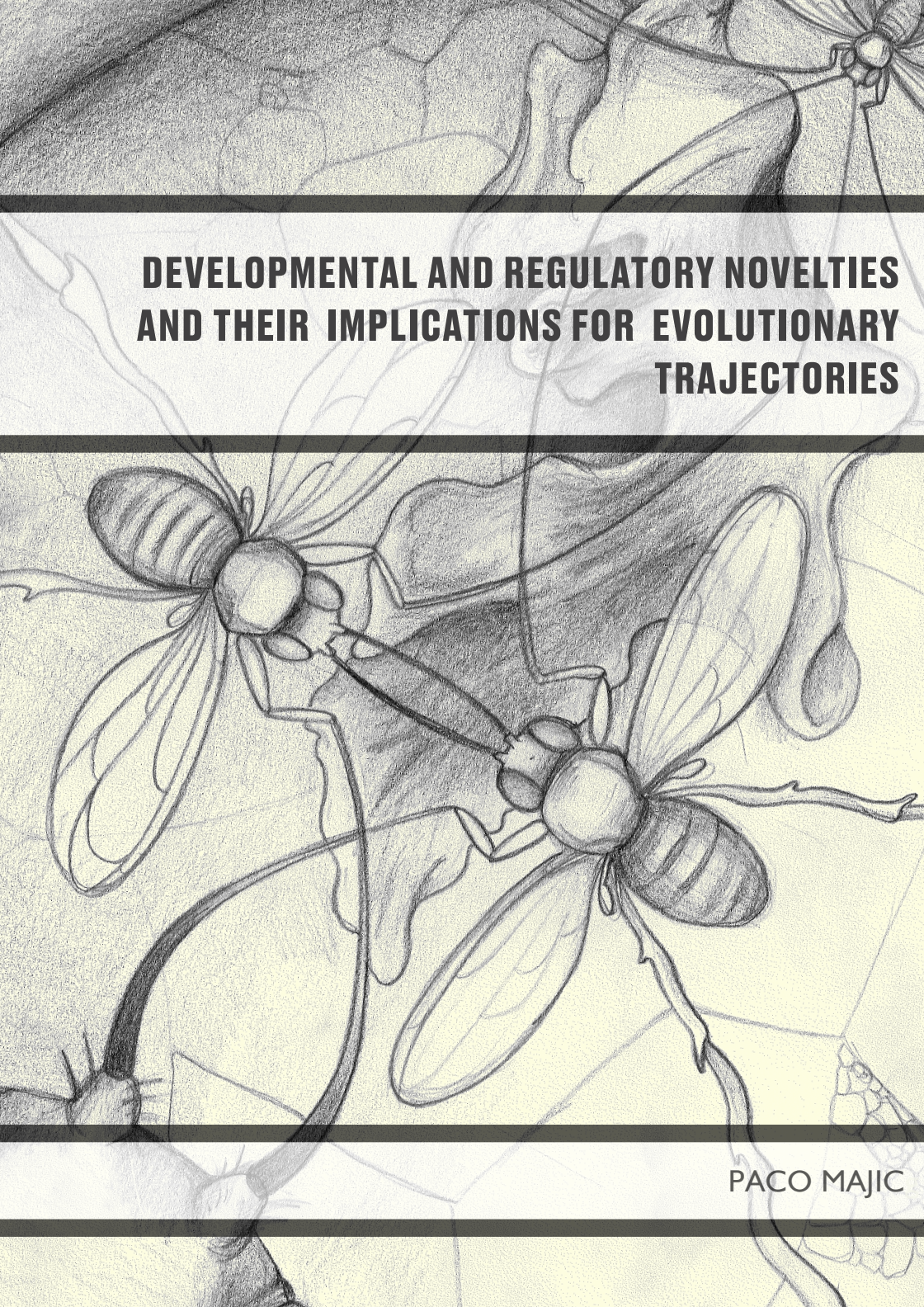
Regulatory and developmental novelties and their implications for evolutionary trajectories

Author(s):
Majic Bergara, Paco M.

Publication date:
2022

Permanent link:
https://doi.org/10.3929/ethz-b-000583476

DEVELOPMENTAL AND REGULATORY NOVELTIES AND THEIR IMPLICATIONS FOR EVOLUTIONARY TRAJECTORIES



PACO MAJIC

Diss. ETH No. 28551

Regulatory and developmental novelties and their implications for evolutionary trajectories

A dissertation submitted to attain the degree of

Doctor of Sciences of ETH Zurich
(Dr. sc. ETH Zurich)

presented by

Paco Matheus Majic Bergara
MSc, The University of Tokyo, Tokyo

born on 27 January 1991
citizen of Uruguay and Croatia

accepted on the recommendation of

Prof. Dr. Joshua L. Payne, ETH Zürich, examiner
Prof. Dr. Erich Bornberg-Bauer, University of Münster, co-examiner
Dr. Justin Crocker, EMBL Heidelberg, co-examiner
Prof. Dr. Patrick Tschopp, Universität Basel, co-examiner

2022

En la memoria de los ancestros
que me encausaron hasta acá.
Especialmente a mis abuelos,
Amilcar "Cacho" Bergara y María
Keco.

I think of man as an amoeba who sticks out pseudopods to catch and envelop his food. There are long and short pseudopods, movements, turnings. One day this all becomes fixed (what we call maturity, a full-grown man). On the one side he can go farther, on the other he can't see a lamp two steps away. In this way a guy goes on living fairly well convinced that nothing interesting will escape him, until an instantaneous landslide shows him for a second, unfortunately without giving him time to know what, shows him his divided being, his irregular pseudopods, the suspicion that farther on, where now I can see clear air, or in this indecision, at the crossroads of choice, I myself, in the rest of reality that I don't know, I'm waiting uselessly for myself.

(Rayuela, Julio Cortázar)

Table of Contents

Summary

One way to visualize the evolutionary process by which species can change and adapt is by thinking of organisms as explorers of the space of all possible genetic combinations. Mutations are the fuel that propagate populations within this space, and natural selection can be thought of as the guiding principle that dictates the viability of each step. What determines whether an organism is selected or not is its phenotype. Therefore, in order to figure out the actual evolutionary potential of organisms it is essential to understand how phenotypes are constructed and how those constructions can evolve. During the history of life on Earth some lineages have evolved key novelties that drastically impacted how phenotypes can develop, and, therefore, how those lineages evolve. One of the most drastic novelties in organismal organisation was the evolution of multicellularity in the stem-lineage to animals. This novelty was followed by an impressive exploration of shapes and habits that make up some of the most exquisite and intriguing diversity populating our planet. In this thesis, I will particularly focus on two features of animal organisation that represented novelties that accompanied the evolution of multicellularity.

First, I will explore the influence of the evolution of novel distal regulatory elements called enhancers on the evolution of gene regulatory networks. Regulatory networks control the spatiotemporal gene expression patterns that give rise to and define the individual cell types of multicellular organisms. In eumetazoa, enhancers play a key role in determining the structure of such networks, particularly the wiring diagram of "who regulates whom". Mutations that affect enhancer activity can therefore rewire regulatory networks, potentially causing adaptive changes in gene expression. Here, I will present results from analyzing whole-tissue and single-cell transcriptomic and chromatin accessibility data from mouse to show that enhancers play

an additional role in the evolution of regulatory networks: They facilitate network growth by creating transcriptionally active regions of open chromatin that are conducive to *de novo* gene evolution. I also show that open reading frames gradually acquire interactions with enhancers over macroevolutionary timescales, helping integrate genes—those that have arisen *de novo* or by other means—into existing regulatory networks. Taken together, these results highlight a dual role of enhancers in expanding and rewiring gene regulatory networks.

Secondly, I will present how the separation of reproductive and somatic cells during the development of animals can influence evolutionary trajectories. The evolution of a germline-soma separation in animals has the implication that much of the genetic variation that arises in the body of an individual cannot be inherited. Because of that, the adaptive potential of non-heritable somatic mutations has received limited attention in traditional evolutionary theory. I will here show how the ability of a germline genotype to express a novel phenotype via non-heritable somatic mutations can be selectively advantageous, and that this advantage can channel evolving populations toward germline genotypes that constitutively express the phenotype. I will present the results of simulations of evolving populations of developing organisms with an impermeable germline-soma separation navigating a minimal fitness landscape. Those simulations revealed the conditions under which non-heritable somatic mutations promote adaptation. Specifically, this can occur when the somatic mutation supply is high, when few cells with the advantageous somatic mutation are required to increase organismal fitness, and when the somatic mutation also confers a selective advantage at the cellular level. These results therefore provide proof-of-principle that non-heritable somatic mutations can promote adaptive evolution.

These discoveries have important implications for how we understand the evolutionary potential of organisms, showing how the evolution of complex molecular mechanisms can facilitate the evolution of completely novel biochemical components, and how the ontological configuration of an organism can guide the evolutionary trajectories of its lineage. I conclude by emphasizing the importance of developmental and historical processes for the understanding of biological evolution.

Résumé

Une façon de visualiser le processus évolutif par lequel les espèces peuvent changer et s'adapter est de penser aux organismes comme des explorateurs de l'espace de toutes les combinaisons génétiques possibles. Les mutations sont le carburant qui propage les populations dans cet espace, et la sélection naturelle peut être considérée comme le principe directeur qui dicte si chacune des étapes franchies est un faux pas ou non. Ce qui détermine si un organisme est sélectionné ou non est son phénotype. Par conséquent, afin de déterminer le potentiel évolutif réel des organismes, il est essentiel de comprendre comment les phénotypes sont construits et comment ces constructions peuvent évoluer. Au cours de l'histoire de la vie sur Terre, certaines lignées ont évolué vers des nouveautés clés qui ont eu un impact profond sur la façon dont les phénotypes peuvent se développer et, par conséquent, sur l'évolution de ces lignées. L'une des nouveautés les plus radicales dans l'organisation des organismes a été l'évolution de la multicellularité dans la lignée souche des animaux. Cette nouveauté a été suivie d'une exploration impressionnante de formes et d'habitudes, qui constituent une des parties les plus exquises et intrigantes de la diversité qui peuple notre planète. Dans cette thèse, je me concentrerai particulièrement sur deux caractéristiques de l'organisation animale, qui représentent des nouveautés qui ont accompagné l'évolution de la multicellularité.

Tout d'abord, j'explorerai l'influence de l'évolution de nouveaux éléments régulateurs, appelés amplificateurs , sur l'évolution des réseaux de régulation des gènes. Les réseaux de régulation contrôlent les schémas spatio-temporels d'expression des gènes, qui donnent naissance aux différents types de cellules des organismes multicellulaires et les définissent. Chez les eumétazoaires, les amplificateurs jouent un rôle clé dans la détermination de la structure de ces réseaux, en particulier le schéma de câblage de "qui régule qui". Les mutations qui affectent

l'activité des amplificateurs peuvent donc modifier les réseaux de régulation, ce qui peut entraîner des changements adaptatifs dans l'expression des gènes. Je présenterai ici les résultats d'analyses de données transcriptomiques et d'accessibilité de la chromatine sur des tissus entiers et des cellules uniques de souris, qui montrent que les amplificateurs jouent un rôle supplémentaire dans l'évolution des réseaux de régulation : Ils facilitent la croissance du réseau en créant des régions de chromatine ouverte, actives sur le plan transcriptionnel, qui sont propices à l'évolution *de novo* des gènes. Je montre également que les cadres de lecture ouverts acquièrent progressivement des interactions avec les amplificateurs sur des échelles de temps macro-évolutives, aidant à intégrer les gènes, ceux qui sont apparus *de novo* ou par d'autres moyens, dans les réseaux de régulation existants. Ces résultats mettent donc en évidence un double rôle des amplificateurs dans l'expansion et le recâblage des réseaux de régulation des gènes.

Dans un deuxième temps, je présenterai comment la séparation des cellules reproductives et somatiques au cours du développement des animaux peut influencer les trajectoires évolutives. L'évolution de la séparation entre la lignée germinale et somatique chez les animaux implique qu'une grande partie de la variation génétique qui se produit dans le corps d'un individu ne peut être héritée. Pour cette raison, le potentiel adaptatif des mutations somatiques non héritables a reçu une attention limitée dans la théorie traditionnelle de l'évolution. Je montrerai ici comment la capacité d'un génotype germinal à exprimer un nouveau phénotype par le biais de mutations somatiques non héritables peut être sélectivement avantageuse, et que cet avantage peut orienter l'évolution des populations vers des génotypes germinaux qui expriment le phénotype de manière constitutive. Je présenterai les résultats de simulations d'évolution de populations composées d'organismes en développement avec une séparation imperméable entre la lignée germinale et le soma naviguant dans un paysage adaptatif. Ces simulations ont révélé les conditions dans lesquelles les mutations somatiques non-héritables favorisent l'adaptation. Plus précisément, cela peut se produire lorsque le taux de mutations somatiques est élevée, lorsque peu de cellules présentant la mutation somatique avantageuse sont nécessaires pour augmenter la valeur adaptative de l'organisme, et lorsque la mutation

somatique confère également un avantage sélectif au niveau cellulaire. Ces résultats apportent donc la démonstration de principe que les mutations somatiques non héritables peuvent favoriser l'évolution adaptative.

Ces découvertes ont des implications importantes sur la façon dont nous comprenons le potentiel évolutif des organismes, en montrant comment l'évolution de mécanismes moléculaires complexes peut faciliter l'évolution de composants biochimiques totalement nouveaux, et comment la configuration ontologique d'un organisme peut guider les trajectoires évolutives de sa lignée. Je conclus en soulignant l'importance des processus développementaux et historiques pour la compréhension de l'évolution biologique.

1 Introduction: navigations of the finite fly

"*After we know the nature of a body, we cannot feign an infinite fly*"

Benedictus Spinoza

1.1 Transformation and inheritance

Organisms thrive whenever their biological capabilities are in tune with their environmental conditions. Species are said to undergo adaptation if they achieve a better suitability to those conditions following a shift in their structure, physiology or behaviour. These changes eventually may lead to the transformation of a species into another, thus creating the biodiversity that we observe today.

Although the idea that species are not fixed and that they can change over generations is currently well accepted, this was not the case until relatively recently. In the early 19^{th} century the fixity of species was the object of debate between Jean-Baptiste Lamarck and Georges Cuvier, two of the most prominent naturalists of their time. From his fruitless attempts to discretely categorize species based on morphology, Lamarck reached the conclusion that species could transform after noting the prevalence of intermediate morphs. Such gradient lead him to the conclusion that there must be a gradual transformation of one species into another. Lamarck's proposal met strong criticism from Cuvier, who advocated for the existence of fixed archetypes, arguing that species were so well-assembled in the act of creation, that the tweaking of any part of an organism would collapse the whole system. The debate between Lamarck and Cuvier involved what some consider the first empirical test of evolution (Curtis, Millar,

and Lambert 2018). After Napoleon's campaign in Egypt between 1798 and 1801, among the many archaeological artifacts his army took to Paris were vases containing mummified specimens of the African sacred ibis, *Threskiornis aethiopicus*. After studying the specimens, Cuvier remarked the similarities shared between the mummified specimens and the ibises that still lived in Egypt (Fig. 1.1). He used this argument to attack Lamarckian transformism, since there had been no change in the species over the three thousand years that separated the time of pharaohs from Napoleonic France (Cuvier 1804). Although Lamarck admitted the lack of change, he defended his transformist argument claiming that the time that had passed since the mummification of the specimens had not been sufficient for the environment to change to an extent such that it would make the ibis change its morphology (Lamarck 1809). While Cuvier's vision prevailed at the time of this debate (Curtis, Millar, and Lambert 2018), it would not take long until the idea that species can indeed change over generations became the ruling view of the origins and history of biodiversity. An acceptance that is owed to Darwin's thorough theory of evolution by natural selection published in 1859, thirty years after Lamarck's death.

Even when the idea that species could change over time managed to eventually reach a wider acceptance, the question still lingered about the causes and direction of that change. As it is well known, Lamarck's idea of how variation arose was one in which traits were altered by their use and disuse in response to environmental conditions, and he also defended that traits thus modified could then be passed on to the progeny. Although this inheritance of acquired traits was not the flagship idea of Lamarck's writings, it became a synonym of his name. Like Lamarck, Darwin also recognised the importance of variation and its inheritance, but he added that change emerges blindly relative to its adaptive value. Adaptation does not happen as a response to changes in the environment, but rather, change happens first and then that variation is filtered by natural selection. This would mean that the direction of evolutionary change would be given by natural selection, rather than by the sources of variation. Darwin was not sure about those sources of variation, nor of the way of its inheritance, and he himself seems to have eventually come to consider the possibility of a Lamarckian mode of inheritance (Morange 2016). To deal with inheritance, Darwin pondered a view in which the information

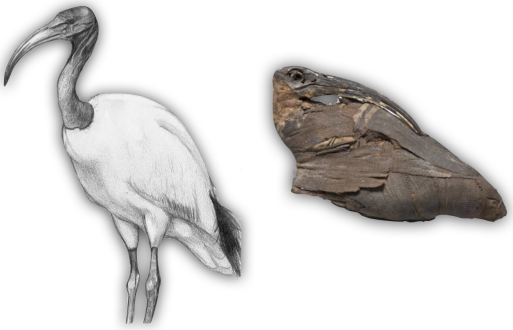


Figure 1.1: The invariance of the sacred ibis. Depiction of a modern exemplar of the sacred ibis *Threskiornis aethiopicus* and a photograph of a mummy of an individual of the same species. Egyptians honoured the sacred ibis as a symbol of Thoth, a god of the moon, of reckoning, of learning, and of writing. The countless mummies of this bird that have been preserved for thousands of years helped in early conceptualizations of the transformability of species. The mummified exemplar belongs to the collection of the Medelhavstmuseet in Stockholm, Sweden. Credit for the photograph of the mummified exemplar goes to that museum.

of the different tissues and organs would be transported to the gonads for it to then be passed on to the progeny - a mechanism called pangenesis (Darwin 1868). This hypothesis came to be disproven by the discovery of the germline-soma separation. August Weismann proposed that during embryogenesis there are cells that follow a pathway that will lead them to perform bodily functions in the soma, and cells that retain the potential to become reproductive cells, the germline. Such framework reinforced the idea that adaptive traits arise blindly relative to their adaptive advantage, since variation arising in the soma would not be inherited via cells developing from the germline (Weismann 1892). This findings prepared the ground for what came to be known as Neo-Darwinism (Mayr 1985; Morange 2016), a paradigm defending that variation emerges randomly and selection non-randomly defines the directions of change.

The ignorance that clouded the understanding of heredity and the origins of biological variation further dissipated with the development of genetics. At the turn of the century, there was a rediscovery of the experiments of Mendel, which lead to the conceptualization of three key concepts in evolutionary biology, the gene, the genotype and the phenotype (Johannsen

1911). Genes were the units of variation, genotypes the total collection of those genes in an organism, and phenotypes can encompass a diverse array of properties, which could be for example, macroscopic morphological traits, physiological functions, behavioural patterns, or microscopic traits, such as basic cellular process or the abundance of a gene product. Chromosomes composed of nucleic acids were discovered to be the physical substrate for the abstract genes, and it was only with the description of the structure of DNA by Watson and Crick (1953) that the underlying mechanism for the maintenance, transmission and mutability of genetic information began to be understood. Simply speaking, Watson and Crick discovered that the heritable information was encoded on sequences of nucleotides of four types - adenine [A], cytosine [C], guanine [G] and thymine [T], and the sources of variation could be said to originate from mutations altering those sequences. The description of the structure of DNA opened the gates for the development of the molecular biology revolution, which has helped us to understand more deeply the origins and the directions of heritable change, and to uncover how genetic information influences the development of phenotypes on which natural selection can act.

1.2 Evolutionary trajectories in the space of possibilities

1.2.1 Navigating the universal genotype space

The genomes of nearly all DNA-based life are contained in the space of genotypes built on all possible combination of the four base pairs A, C, T and G (Dennett 1995), and the evolutionary history of life on Earth could be said to have involved a journey within this space. Note that the size of the human genome is estimated to be composed of roughly 3 billion base pairs (Nurk et al. 2022), while the smallest known functional genome capable of sustaining cellular life corresponds to a synthetic construct of 531 kilobase pairs (Hutchison III et al. 2016). Therefore, for the smallest known functional genome size, there is a gargantuan number of 4^{531000} possible genotypic combinations. This implies the space of genomic possibilities is enormous and each organism is but a point in this universal genetic vastness. Much of this space will not produce

any viable functional biological properties, but as was pointed out by John Maynard Smith (1970) in regard to a similar problem with the space of all possible proteins, navigating the space of genotypes by small mutational steps allows for an exploration of the space in which the non-functional variants are evaded and in which evolution can find adaptive nucleotide arrangements. That is to say that a population can explore the genotypic space by means of mutational steps that can disperse its individuals in different directions of that space assessing which combinations are functional and which ones are not. Imagine we have a fragment of a genome of a sequence ATT (Fig. 1.2A). Single-step mutations could grant a population the potential to explore neighbouring sequences, ACT, TTT, ATG, etc., if any of those neighbours is viable, then the population can take an evolutionary step in that direction. By repeating this process, life can diversify within a mutational network of viable genotypes.

The exploration of a mutational network of genotypes was at the basis of the adaptive landscape metaphor that Sewall Wright used to abridge the mathematical insights of population genetic models to biological audiences (1932). Wright proposed that each genotype would represent a given selective advantage to individuals, thus constructing a landscape of fitness values that would be grounded by the mutational network of genetic combinationsn (Fig. 1.2B). Under a simple model, individuals could explore this landscape via mutational steps and natural selection would direct the population upward in the gradient of fitness elevation, eventually leading that population towards the summit of an adaptive peak. The complexity of the genotype space, meaning its high-dimensionality[1], implied that different allelic combinations could have epistatic interactions that would lead to the formation of valleys, making the landscape rugged. Populations could be trapped in local optima, depending on the starting point of the population and the evolutionary trajectory it followed. Another important aspect is that since the environment is always shifting, a genotype that is favourable in one environment, might not be favourable in another, which results in a shifting peak dynamic model that has

1. The high dimensionality Wright used as an example was one of 2^{1000}, based on estimations of the time suggesting that *Drosophila* had a thousand genes, for each of which Wright modelled two "allelomorphs". The current estimates of the genome size of *Drosophila* is of around 140 Mb. That would mean that, given that there are four possible alleles per base, the actual genotypic possibilities for *Drosophila* would be $4^{140000000}$. Therefore, Wright was correct in his assessment that "the population is confined to an infinitesimal portion of the field of possible gene combinations".

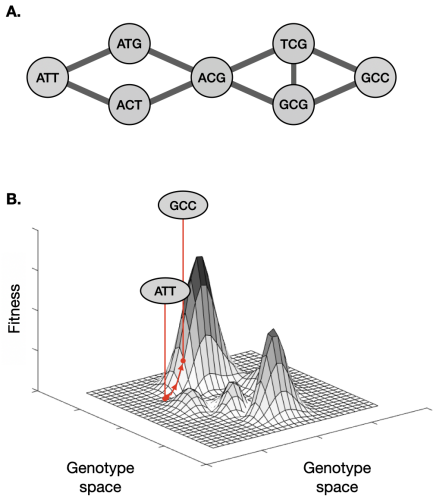


Figure 1.2: Genotype networks and adaptive landscapes. (A) A genotype network represents the single-step mutational connections between genotypes. The particular network depicted here shows a subset of possible mutational paths connecting genotype ATT and genotype GCC. (B) Representation of an adaptive landscape. The vertical axis shows the fitness values for each genotype of a two-dimensional plane of the space of genotypes. In the metaphor of Sewall Wright, the evolutionary process can be seen as a population exploring this landscape via mutations that displace populations on the genotype space. In red, the mutational route from ATT to GCC is depicted as an upward route in direction of the higher peak in this particular landscape. Note the presence of local optima that could trap populations if selection is the only guiding principle in this exploration.

been described as a seascape (Mustonen and Lässig 2009). Wright's model of the adaptive landscape is an effective illustration of how evolution is a local search in the space of possibilities (Jacob 1977; Kauffman 1993), and it epitomises the adaptationist notion that natural selection is the compass of evolutionary change. That said, moving towards higher degrees of biological realism, it becomes inescapably evident that there are other principles that complement the action of natural selection in the directions of adaptive dynamics and morphological diversification (Gavrilets 2010).

A layer of realism that can be added concerns the mutational process. The basic assumption Wright made about mutations was that they were "fortuitous in origin, infrequent in oc-

currence and deleterious when not negligible in effect" (Wright 1932). These assumptions are in line with the neo-Darwinian principles according to which mutations are chance events and evolutionary change is gradual. However, the fact that the *occurrence* of a mutation is fortuitous does not imply that the *directions* of those mutations are as well. In fact, the mutational process itself is strongly biased. Mutations can be of various different kinds, point mutations, insertions, deletions, duplications, recombination, rearrangements, transposition, etc. The rates of all these different mutation types differ. For example, point mutations are considerably more common than insertion-deletion mutations across different regions of the genome of different organisms (Chen et al. 2009). Furthermore the effects of these different kinds of mutations involve different manners in which individuals traverse the space of genotypes. Whereas point mutations might represent a step towards neighbouring genotypes differing in the identity of a single base as mentioned above, an insertion or a deletion represent a change in dimensionality (Martin and Ahnert 2021), transposition represent a leap to a different mutational neighbourhood (McFadden and Knowles 1997) and recombination can create novel genotypes by shuffling pre-existing parental genetic diversity (Klug, Park, and Krug 2019). Even within the same types of mutations there are biases that can favour that a population takes a mutational step in one direction over another, as it happens with transition-transversion biases of point mutations (Stoltzfus and Norris 2016; Cano and Payne 2020). Additionally, mutation rates and biases can be specific to the genomic location (Xie et al. 2019b), the organism (Cagan et al. 2022), the cell type (Milholland et al. 2017) or the environmental conditions (Liu and Zhang 2019). Thus, it may well be the case that mutational event itself is a fortuitous roll of a die, but the die is loaded, and so are the directions evolution will follow as a consequence.

Therefore, the study of the mutational process conveys the idea that the directions of evolutionary change are not defined by natural selection alone. If we continue to approach biological reality in our understanding of the evolutionary trajectories to adaptation, the next step to consider is the interface between the heritable genotype and its fitness. That is, we need to also account for the actual organism and its selected traits. By doing so, adaptive landscapes cease

to be a bodyless metaphor for abstract adaptive processes, and instead they can can start to approximate the actual diversification of biological forms and functions.

1.2.2 The influence of genotype-phenotype maps

Genotype-phenotype maps represent the way in which the heritable genetic information contained in chromosomes directs, coordinates or influences the development of phenotypes, which are the features of organisms on which natural selection can act. The fitness values that define the surface of adaptive landscapes are essentially a one-dimensional reduction of the countless dimensions of an organism's phenotypic complexity. Therefore, the topology of any fitness landscape will be intrinsically defined by how the underlying genotype space tethers to the intermediate phenotypic space. Consequently, the mapping of genotypes to phenotypes can have important evolutionary consequences when it comes to the potential to evolve new forms and the mutational trajectories populations might take as they explore the space of genotypes.

An important aspect of genotype-phenotype maps in relation to evolutionary change is that they define the likelihood that a genotypic change leads to a novel phenotype. To start illustrating this point, consider the case of the genetic code, which is a set of rules for the translation of proteins on the basis of the nucleotide sequence of the genetic material. Under the standard genetic code, any sequence of three nucleotide bases, a codon, maps to either one of twenty amino acids or to a STOP signal that halts protein elongation. For example, during protein assembly, a GCA codon recruits an alanine, while CCG recruits a proline. Because the genetic code maps $4^3 = 64$ combinations of nucleotides to 21 options, a property of this map is that many triplets code for the same amino acid. Alanine, for example, is not only encoded by GCA, but also by GCC, GCT and GCG. This has the fundamental implication that not all mutations have the same potential to cause phenotypic alterations. Whereas a change in the first or second position of GCA can change which amino acid is mapped to this sequence, changes in the third position would always still code for alanine. Therefore, many mutations on protein-coding sequences produce no phenotypic change, and they can be said to be neutral with

respect to amino acid sequence.

In 1968, Motoo Kimura published the results of his comparison of the rates of amino acid changes of different proteins over evolutionary time. The abstract of that publication was a single sentence stating that "(c)alculating the rate of evolution in terms of nucleotide substitutions seems to give a value so high that many of the mutations involved must be neutral ones" (Kimura 1968). Kimura thus founded the neutral theory of molecular evolution, which defended that most of the genetic variation fixing in a population would have no adaptive value. In the context of evolutionary trajectories, what the neutralist view suggests is that much of the exploration of the space of genotypes happens without any evident change in morphology or functions of organisms. As a result, the spread of populations over this space would mostly occur via non-selective processes such as genetic drift or the accumulation of mutations. Such a spread would be possible along a sub-set of connections of the total mutational network that map to the same phenotype, which can be referred to as "neutral networks" (Schuster et al. 1994; Van Nimwegen, Crutchfield, and Huynen 1999)[2].

The architecture of neutral networks (and therefore of the genotype-phenotype maps) has important consequences for the biological possibilities of attaining new forms over evolutionary time. Specifically, it can impact phenotypic robustness and evolvability. Phenotypic robustness refers to the invariance of a phenotype when an organism is exposed to environmental or genetic perturbations (De Visser et al. 2003), while evolvability can be considered to be "the ability of a biological system to produce phenotypic variation that is both heritable and adaptive" (Payne and Wagner 2019). The particular case of robustness to mutations implies that a population can mutate without changing phenotypes, and that the more robust phenotypes are those that map to broader neutral networks with a high number of internal connections (Ciliberti, Martin, and Wagner 2007). Although intuitively there is a conflict between the robustness of phenotypes and evolvability, there has recently been a number of studies indicating that this is not always the case, and that robustness can in fact promote phenotypic diversification under certain conditions (Ciliberti, Martin, and Wagner 2007; Wagner 2008; Draghi et al. 2010; Payne and Wagner 2014).

2. Sometimes also referred to as "genotype networks" e.g. (Payne and Wagner 2014)

One way in which robustness can increase evolvability is by allowing for the accumulation of "cryptic genetic variation" (Rutherford and Lindquist 1998; Bergman and Siegal 2003; McGuigan and Sgro 2009; Zheng, Payne, and Wagner 2019). As a population explores a neutral network, individuals can spread over the surface of the network thus accumulating high levels of standing genetic variation. Upon a change in conditions, such as a stressful trigger or a novel environment, it is possible that the mapping of genotypes shifts, thus mapping the accumulated genetic variation to novel phenotypes. Another way in which robustness can promote the evolution of novel phenotypes is when the architecture of neutral networks facilitates connectivity between the neutral networks of different phenotypes (Ciliberti, Martin, and Wagner 2007; Wagner 2008; Draghi et al. 2010; Payne and Wagner 2014). Although phenotypes do not change during the spread of a population over a neutral network, what might actually change is the mutational routes a population has access to. As populations spread over neutral networks, some individuals might reach that network's edge[3], making a novel phenotype mutationally accessible. The idea of how neutral networks can promote evolvability draws a picture of how gradual genotypic change can result in major phenotypic leaps. What this means is that the consideration of the genotype-phenotype map in the evolutionary process reconciles the idea of punctuated equilibrium at a phenotypic level with a Neo-Darwinian gradual change at the genotypic level (Pigliucci 2010).

Another important point about the mapping of genotypes to phenotypes is that it may lead to what has been called a "developmental bias" (Maynard-Smith et al. 1985; Arthur 2004; Uller et al. 2018). A developmental bias simply refers to the idea that evolutionary processes can produce some phenotypic variants more readily than others. Typically, developmental biases are taken from the perspective of how the origin of a shape might be constrained (Maynard-Smith et al. 1985), for example, as the result of a trade off, the covariance of traits with common genetic roots, or a physical impossibility[4]. Another source of developmental bias depends

3. The equivalent of what Oster and Alberch (1982) called a "bifurcation boundary" for morphologies: "The different stability domains in the developmental program correspond to the major morphological themes. These domains are bounded, and mutations and selection serve to disperse the phenotypes over the region, with no qualitative changes in morphology. However, when enough mutations (or a large one, depending on the nature of the genetic control system involved) accumulate to bring the developmental system near to a bifurcation boundary, then one can expect large changes in phenotype to accompany relatively minor genetic alterations."

4. Like the square circle or the infinite fly that Spinoza invites us to imagine.

on how easy it is to reach a novel phenotype given a starting point in the space of genotypes (Maynard-Smith et al. 1985). A certain phenotype might be more likely to evolve depending on the distribution of the genotypes that map to it in a mutational network (Schaerli et al. 2018). For example, given a neutral network that is explored by a population, that population will be more likely to reach a novel phenotype whose neutral network shares more mutational edges to the original network. Furthermore, phenotypes that are more commonly mapped within the genotype space will be more evolvable than rare phenotypes, in what has been termed "the ascent of the abundant" (Cowperthwaite et al. 2008) or "the arrival of the frequent" (Schaper and Louis 2014). A consequence of this bias is that evolutionary trajectories might often lead to a representation of sub-optimal phenotypes that is greater than what would be expected in an adaptationist view.

Additionally, genotype-phenotype maps are often non-linear (Alberch 1991) and single genotypes can guide the development of more than one phenotype. There is phenotypic plasticity whenever the change in phenotypic output of a genotype results from an evolved response to an environmental cue. Examples of this include the temperature-dependent sex determination of reptiles (Janzen and Phillips 2006), the phenotypic distinction between different castes of eusocial insects (Miura 2005), or the development of defensive structures in water fleas when exposed to predators (Laforsch and Tollrian 2004). Besides phenotype plasticity, a single genotype can also output different phenotypes as a consequence of internal stochastic events in cellular or developmental mechanisms. For example, the folding of proteins can sometimes be erroneous (Drummond and Wilke 2009), some proteins and RNAs have multistable native structures (Ancel and Fontana 2000; Drummond and Wilke 2009), gene expression can occasionally be noisy (Raser and O'shea 2005) and transcriptional, translational and somatic mutations are rampant and more common than mutations on the genomes of germline cells (Pouplana et al. 2014; Gout et al. 2013; Gout et al. 2017; Milholland et al. 2017). During the evolutionary trajectory of a population, a phenotype that originally was produced only under specific environmental circumstances or as the result of developmental stochasticity can eventually become canalized or assimilated in a way that it becomes the output of a default

developmental program (Crispo 2007; Pigliucci, Murren, and Schlichting 2006; Waddington 1942, 1953; Whitehead et al. 2008) - in other words, the population can get displaced more deeply into the neutral network of the alternative phenotype, away from the network of the original phenotype. This is especially facilitated when the area of a genotype network that represents phenotypic ambiguity overlaps the area bordering alternative neutral networks, which has been explored in the case for bistable protein structures (Sikosek, Chan, and Bornberg-Bauer 2012) and gene regulatory networks (Espinosa-Soto, Martin, and Wagner 2011). As a consequence to all this, the effects of genetic mutations on the development of a phenotype, even for mutations within a neutral network, might be variable.

The evolutionary trajectories life follows are therefore not only delimited by chance mutations and natural selection. There is an essential component which is the decoding of those new variables into a phenotype that define the fate of different mutations, and that can have major repercussions regarding the solutions life encounters as it evolves. Therefore, as anticipated by Oster and Alberch (1982), "the dynamics inherent in the process of development itself imposes constraints and biases on morphological evolution that cannot be comprehended from a genetic or population perspective alone". To figure out the actual evolutionary potential of organisms it is therefore fundamental to understand how phenotypes are constructed and how those constructions can evolve.

1.3 The body, its construction and its transformations

1.3.1 Basic principles of phenotypic development

The original conception of phenotypes going back to Johannsen (1911) describes how "(a)ll typical phenomena in the organic world are *eo ipso* phenotypical". By typical phenomena Johannsen meant all traits that permit discerning individuals in types for classification, which at the time were mostly based on morphology. Currently, we can do typological descriptions across several layers of organismal organization, and describe phenotypes that go from the structure of a protein to the length of a giraffe's ossicones and to the cooperative breeding

behaviour of guira cuckoos. At higher levels of organization, phenotypes are, in the most part, compositions. In fact, an individual organism itself could be considered to be a composite phenotype of several nested layers of other phenotypes. Ultimately, the development that maps a genotype to a phenotype can be regarded as the synergystic integration of different phenotypic layers.

The phenotypic potential of organisms relies strongly on the expression of the information encoded in the genome of each of their cells. While a genome can contain tens of thousands of protein-coding genes (Lander et al. 2001), within a cell, only a subset of those genes are active simultaneously. This offers an enormous combinatorial potential of gene activity, not only in the identity of genes that are active, but also in the levels of their expression. The regulation of gene expression can be highly complex, involving many diverse molecules interacting in a concerted way that allows the right set of genes to be expressed at the correct level under different circumstances. The activation and inactivation of specific genes allows cells to respond to their environment (Jacob and Monod 1961; Granados et al. 2018) and to control their own internal physiological activities and life cycle (Rowicka et al. 2007; Gestel, Ackermann, and Wagner 2019).

Gene regulation is a multilayered process that takes place across any of the different steps of gene expression, including the activation or inhibition of transcription, the stabilisation and transportation of active RNAs within the cell, and, in the case of protein-coding genes, the translation of the information contained in mRNAs into chains of amino acids and the subsequent modification, transportation or degradation of those chains. Transcriptional regulation, on which I will focus here, has been studied in depth since the discovery of the lactose operon in bacteria (Jacob and Monod 1961), and many of its molecular mechanisms are now well characterised. A major step in transcriptional regulation is the recruitment of the RNA-polymerase protein complex to the transcription start site of a gene. Across the tree of life, the recruitment of this complex is facilitated by genomic regions called promoters, which are sequences located immediately upstream of the transcription start sites of genes on which the polymerase complex docks to start the production of RNAs. The activity of promoters is based on their

interactions with transcription factors, which are proteins with DNA-binding domains that recognise short DNA sequence motifs that can be found on promoters. Across branches of the tree of life, transcriptional regulation has different layers of complexity. For example, in Eukaryotes, DNA forms a chromatin complex with proteins that allows for its packaging within the nucleus. Nucleosomes are the sub-unit of chromatin and they consist of an octamer formed by histones around which DNA is wrapped. The density of the nucleosome packaging, which can be regulated by epigenetic modification on specific amino acids of individual histones, can make different specific genomic regions accessible or not for the transcriptional machinery. Another layer of complexity is the presence of other cis-regulatory elements complementing the activity of promoters, either via the negative or positive control of gene expression.

In complex multicellular organisms such as animals, plants and fungi, gene regulation not only allows individual cells to handle environmental conditions or their own life cycle, but it is also at the basis of the coordination of morphogenetic processes (Carroll 1995; Peter and Davidson 2011). There is an enormous diversity of molecular process and signaling pathways that help coordinate these cellular activities along the trajectory that goes from a single-celled zygote to the fully developed adult organism. These processes regulate the behaviour of individual cells or group of cells by orchestrating their proliferation, migration, growth, death, and, ultimately, their differentiation. The differentiation of cells from an original totipotent state results from shifts in gene expression leading to a stable state representing a cell type with a specialized function, as could be a neuron, a cardiomyocyte, or an hepatocyte. This is a process that can be thought of as an epigenetic landscape or a dynamical system by which sinks of attraction pull towards stable patterns of gene expression (Waddington 2014; Kauffman 1993; Furusawa and Kaneko 2012). In molecular terms, these dynamics result from lineage regulators, such as some transcription factors, that enable major chromatin rearrangements that gradually commit cell activities towards a specialized use of their genetic repertoires (Stergachis et al. 2013).

1.3.2 The evolution of the molecular wirings within

A principle of gene expression is that it is controlled by genes themselves. The developmental process of complex phenotypes can therefore be pictured as a cybernetic network of interactions between molecules that assures the functional coherence of the different parts of organisms (Monod 1970). In 1969, Britten and Davidson presented a very rough model of genomic regulation in which they classified genes as having different roles within a functional network. The model of Britten and Davidson was very speculative as they themselves acknowledged. However, the growing evidence in the fields of molecular, cellular and developmental biology gradually gathered the pieces to reconstruct models of cellular functioning with resemblances to the one they envisioned, as well as other theoretical models of metabolic networks that were developed around the same time (Kauffman 1969). Gene regulatory networks are one of such models. In these models, some genes have a regulatory function and others have a structural function, a distinction that goes back to the very origins of the study of gene regulation (Jacob and Monod 1961). Transcription factors play a central role as genes of the first kind, and by binding cis-regulatory elements they can target the expression of either other regulatory genes or structural genes (Fig. 1.3). As such, transcription factors and the genomic regions they bind are the primary piece in the assembly of the wiring diagram of "who regulates whom" of gene regulatory networks that define what is the combination of genes that is expressed at a given time in a cell.

Gene regulatory networks have different properties that make them particularly relevant for phenotypic evolution (Gerhart and Kirschner 2007). One such property is that they offer modularity to genomic functions. We could imagine that a total network maps all the possible interactions between regulatory genes and regions of the genome that can affect the expression of target genes. But at any given time, a cell is only making use of some modules from within that total network. The specific utilisation of those modules defines the repertoire of genes that is expressed and, ultimately, the cell's state and phenotype. The modular nature of networks allows phenotypes to also evolve in a modular fashion, thus permitting adaptive changes without corrupting the whole and releasing from pleiotropic constraints (Wagner

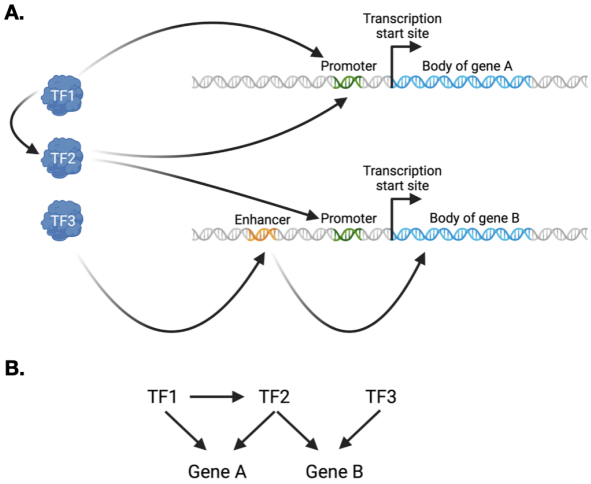


Figure 1.3: The structure of gene regulatory networks. (A) Illustration of how transcription factors (TF) interact with binding sites on regulatory elements such as promoters and enhancers to guide the transcription of hypothetical genes A and B. (B) Schematic representation of the regulatory circuit involving TF1, 2 and 3 and genes A and B.

1996; Solé et al. 2002). A modular architecture can also occasionally allow for the redeployment of entire subcircuits for them to be co-opted into novel functions, as has recently been shown with a regulatory circuit involved in the patterning of legs and antennae that was co-opted to develop eyespots on wings (Murugesan et al. 2022).

As a corollary of gene regulatory network architecture, mutations on different genes can differently impact the phenotypic output depending on that gene's position in a network (Stern and Orgogozo 2009; Erwin and Davidson 2009; Yang and Wittkopp 2017). For example, alterations of centrally located regulators that control the activity of many genes would have broader repercussions than alterations of peripheral genes. Furthermore, being regulated by many transcription factors might make genes more susceptible to mutations, since many more mutations can possibly affect at least one of those transcription factors, although genes that have a higher connectivity from incoming regulatory edges may also tend to have a gene ex-

pression that is more robust to genetic variation due to compensatory activity of other edges (Yang and Wittkopp 2017). Consider the case of one of the better characterised examples of a developmental network, the one modelling molecular interactions during the larval development of the purple sea urchin *Strongylocentrotus purpuratus* (Oliveri and Davidson 2004; Peter and Davidson 2009, 2010; Sharma and Ettensohn 2010). Studying the impact of naturally occurring genetic variation in populations of *S. purpuratus* in relation to the architecture of the regulatory network, Garfield et al (2013) found that the expression of genes that are expressed early in development and that show a higher interconnectivity is less affected by the genetic variation than the expression of genes expressed at final stages of development.

As in other genotype-phenotype maps, gene regulatory networks also show different degrees of robustness and evolvability (Aldana et al. 2007; Ciliberti, Martin, and Wagner 2007; Crombach and Hogeweg 2008; Payne, Moore, and Wagner 2014; Jiménez et al. 2015). Many different arrangements of the interactions of genes can lead to similar attractor states, either in the patterns of expression of genes, or in the encoding of a specific phenotypic feature. Populations can thus neutrally explore alternative regulatory circuits via "developmental system drift" (Dalal and Johnson 2017; Crombach et al. 2016; Halfon 2017; Haag and True 2021). Early evidence of this comes from a group of genes that was regulated by an activator in an ancestral fungal ancestor, which is regulated by a repressor in modern bakers' yeast while still maintaining the same logical output of the ancestral regulatory circuit (Tsong et al. 2006). As another example, the comparison of the expression of genes mediating the development of wings on different ant castes showed how different species used different elements of a regulatory module that drives the development of wings in *Drosophila* (Shbailat and Abouheif 2013). That is to say, at least in these ants, the regulatory circuit has drifted without altering the phenotypic output. Like in the case of the exploration of neutral networks, developmental systems drift over different topologies of gene regulatory networks can occasionally lead to the development of novel adaptive phenotypes. For example, the evolution of biofilm formation in the fungus *Candida albicans* was enabled by the rewiring of regulatory circuits involving the transcriptional regulator Ndt80 (Nocedal, Mancera, and Johnson 2017).

Gene regulatory networks are constructed basically on two components, the total number of genes in a genome and the regulatory sequences that enable the functional connections between regulatory genes and those sequences. The evolution of gene regulatory networks can therefore result from either a change in the repertoire of genes a genome contains, or in the connections between regulatory genes and their targets. I will now discuss how each of these can evolve.

The birth and death of genes

The biological capabilities of organisms are defined to a great extent by their genetic repertoires. Genes can arise or vanish in a lineage-specific manner, which has resulted in a genetic turnover that helped shape the genomic diversity existing today across the tree of life. For example, the repertoire of genes has expanded substantially in the lineage that lead to animals after the split with the common ancestor they share with choanoflagellates, while it has considerably shrunk before the radiation of deuterostomes and ecdysozoans (Fernández and Gabaldón 2020). At smaller evolutionary timescales, the gain and loss of genes has also diversified the genetic repertoire within primates (Hahn, Demuth, and Han 2007) and it has differentiated the genome of humans and chimpanzees (Wang, Grus, and Zhang 2006). The loss or gain of genes can aid adaptations. For example, the gain of a pancreatic ribonuclease has helped red-shanked doucs (monkeys from South East Asia) with their leaf-eating habits (Zhang, Zhang, and Rosenberg 2002), while the loss of a gene involved in iron retention likely contributed to the adaptation of vampire bats to their blood-based diet, which is high in iron (Blumer et al. 2022),

A turnover of the genomic repertoire does not necessarily imply that the total number of genes changes (Fernández and Gabaldón 2020), but it does have the potential to rearrange gene regulatory networks in such a way that the adaptability of an organism is enhanced or reduced. Experiments in yeast, for example, have shown how knockouts of different genes in different positions of gene regulatory networks can improve the evolvability of populations under changing environments (Helsen et al. 2020). On the other hand, the gain of genes can create

new edges in networks of interactions (Capra, Pollard, and Singh 2010; Zhang et al. 2015), thus promoting the exploration of novel network structures and functional innovations. Particularly, during the evolution of multicellular organisms like plants and animals there has been a step-wise increase in the repertoire of transcription factors encoded in the genome (Mendoza et al. 2013; Schmitz, Zimmer, and Bornberg-Bauer 2016), as well as a recurrent birth of new transcription factors in different animal lineages that with the help of transposable elements has rewired gene regulatory networks across several lineages (Cosby et al. 2021).

Distinct molecular evolution processes facilitate the gain or loss of genes. The mechanisms for gene loss are more straightforward than those for gene gain, since the loss of function of a gene tends to be more mutationally accessible than a gain of function. Gene loss can take place for example, by the neutral decay of a coding sequence as it happened for example in genes that are essential for eye development in fish and rodents that dwell in darkness (Policarpo et al. 2021). When it comes to the mechanisms of gene birth, they can be classified as gene birth working on the substrate of pre-existing protein-coding genes and *de novo* origination of genes. Most of the better characterised mechanisms of gene birth are of the first kind, and they include gene duplication, fusion, retrotransposition, the domestication of genomic parasites, and horizontal gene transfer (reviewed in Kaessmann, 2010). The birth of new genes by the employment of pre-existing coding sequences epitomises François Jacob's conceptualization of evolution acting as a tinkerer. The basic molecular machinery, Jacob (1977) defended, evolved at the origin of life, and thereafter the main way by which novelty could arise was by restructuring that machinery and creating new tools anew from materials that were at hand. When presenting this idea, Jacob emphasized the importance of gene duplication in evolution and casted doubt on the possibility of gene birth *de novo*, claiming that the probability of that happening was practically zero.

If protein-coding genes could only arise from other pre-existing genes as Jacob suggested, one would not expect to find genes without any degree of homology when comparing the genetic repertoire of closely related organisms. However, many protein-coding genes do appear to be restricted to some lineages, showing no degree of homology with the genes of other

groups of organisms (Levine et al. 2006; Begun et al. 2007; Cai et al. 2008; Toll-Riera et al. 2009; Carvunis et al. 2012; Schmitz, Ullrich, and Bornberg-Bauer 2018). This makes these so-called "orphan" genes candidates to have arisen *de novo* from non-coding genomic regions (Tautz and Domazet-Lošo 2011; McLysaght and Hurst 2016; Van Oss and Carvunis 2019). In the case of the more recently evolved orphan sequences, even if they did evolve an open reading frame *de novo*, that does not necessarily mean that the sequence encodes for a protein that has a cellular function, or that it is evolutionarily relevant. However, empirical evidence that has been gathered over the past decade and a half that strongly support the idea of de novo gene birth, with examples being identified across several organisms and many different characterizations of their functionality (Baalsrud et al. 2018; Zhang et al. 2019; Xie et al. 2019a; Vakirlis et al. 2020). Furthermore, some cells of *Escherichia coli* transformed with plasmids holding randomized open reading frames evolved antimicrobial resistance (Knopp et al. 2019), strongly supporting the view that peptides evolved from scratch can be functional and adaptive.

Regulatory evolution

For complex phenotypic development, it can be argued that the functional wiring connecting genes in a genome is as important as the genetic repertoire. But for evolutionary change, it may even be more important. In 1975, King and Wilson presented their results of the comparison of the amino acid sequences of homologous proteins from mouse and chimpanzee. The differences they found were so small that they suggested that most of the phenotypic differences that exist between both species are probably not due to changes in the proteins, but in how they are regulated. The idea that most evolutionary variation would come from regulatory changes had been anticipated by Britten and Davidson (1969) when presenting their theory for the gene regulation of complex cells, and it was also later eloquently illustrated by Jacob (1977) to argue in favour of the idea of evolution acting as a tinkerer. Jacob wrote:

> "It seems likely that divergence and specialization of mammals, for instance, resulted from mutations altering regulatory circuits rather than chemical structures. Small changes modifying the distribution in time and space of the same structures

> are sufficient to affect deeply the form, the functioning, and the behavior of the final product – the adult animal. It is always a matter of tinkering."

The logic behind the argument that regulatory changes would be more likely to generate adaptive evolutionary variation than structural variation is that regulatory functions offer a great combinatorial potential (Carroll 2001) and they can minimise fitness penalties (Prud'homme, Gompel, and Carroll 2007). This last point refers to the idea that regulatory change can be modular, affecting expression in developmental contexts were changes might be adaptive without affecting it where they might be disruptive. A beautiful illustration of this is the mutational story of how snakes lost their limbs. A key determinant of the development of limbs in tetrapods is the signaling protein Sonic hedgehog encoded by the *Shh* gene. In the lineage of snakes, mutations on transcription factor binding sites regulating this gene lead to a local loss of expression that prevents the development of the limb bud during embryogenesis (Kvon et al. 2016). Another way in which the reduction of limbs in snakes could have evolved is by mutations on the protein or perhaps the promoter of the gene. But *Shh* is essential for the development of other body parts, including the central nervous system, and changes in the structural protein or its core regulators would have system-wide deleterious effects.

The case for the prevalence of regulatory change in the evolution of novel traits gained support with the advent of fine-grained molecular techniques, especially with the dawn of the genomic era. When the human genome was published, researchers wondered to what extent our greater organismal complexity relative to yeast could be explained by the 2-3 fold increase in the number of genes that they estimated, suggesting that it was regulatory complexity that was behind organismal complexity (Lander et al. 2001). Since, many other comparative genomic studies have evidenced how regulatory changes affect the different morphologies of various animal clades, including the convergent evolution of the flightlessness of paleognathous birds (Sackton et al. 2019), the origin of feathers (Lowe et al. 2015), the atrophy of eyes in subterranean mammals (Partha et al. 2017), insect pigmentation and patterns (Prud'homme et al. 2006), and human facial features (Prescott et al. 2015).

In gene regulatory network evolution, there are several ways in which mutations can af-

fect the wiring of regulatory circuits. These include changes in connectivity due to mutations on transcription factor binding sites, alterations to the transcription factor itself, changes in the concentration of a transcription factor, or changes in co-accessibility of transcription factors and binding sites (Johnson 2017). The evolutionary likelihood of each of these events is strongly reliant on properties of the transcription factors, the sequences they bind, the nature of the interaction and the regulatory context. For example, Igler et al (2018) studied the rewiring of two repression proteins Lamba C1 and P22 C2 and evaluated the robustness, evolvability and tunability of each of these transcription factors in the face of mutations in their target binding sites. They found that whereas Lamba C1 is more robust and more evolvable, P22 C2 is more tunable, owing to the different mechanisms that these proteins use to bind DNA. Given the enormous diversity of transcription factors that exist in genomes (Mendoza et al. 2013; Schmitz, Zimmer, and Bornberg-Bauer 2016), it is expected that different regulatory wirings will represent several different degrees of robustness and evolvability (Payne and Wagner 2014).

In sum, the development of complex phenotypes can be modelled as the product of regulatory interactions between genes, and the structure of this network will determine the phenotypic potential of mutations. Thus, the conceptualization of gene regulatory networks as genotype-phenotype maps illustrates how development will bias the evolutionary trajectories of populations across the morphospace of complex traits.

1.4 Novelties and evolvability

I have so far discussed how the evolutionary process can be conceived as an exploration in the space of genotypes, and how the potential to attain new forms is strongly determined by the developmental process defining how genotypes translate into phenotypes. But there is another aspect of this exploration that is worth emphasising, and that is that evolution is a historical process. In the vastness of the space of genotypes, what evolutionary step follows is a process that is strongly contingent on its previous steps (Harms and Thornton 2014; Xie et al. 2021). One can imagine that neutral networks are vast spaces, and that the potential of discovering a

new phenotype that is adaptive will strongly rely on each step that the population takes in that space. An element that can influence this reliance is, for example, population size. This parameter will determine how many individuals can spread over that network. If populations are not infinite (which they are not), then the history of the spread of a population over a neutral network will define the potential to encounter new phenotypes. Furthermore, as a population explores the genotype space it might occasionally evolve a phenotype that radically impacts the biology of the organism in a way that the entire set of rules for the exploration of genotype spaces is altered.

The evolution of novelties can result in major organismal rearrangements that can strongly change the developmental wirings of organisms to an extent that they open up a complete new world in what relates to the construction of phenotypes, thus posing new evolutionary problems and solutions (Maynard Smith and Szathmary 1997; Erwin 2015). As a simple example, consider how evolving pectoral fins in the ancestor to all gnathostomes (Zhu et al. 2012) allowed branches of this lineage to explore different shapes of limbs. This diversification has lead to the evolution of a wide array of functions such as maintaining buoyancy, balancing swim, gliding, walking, courting, flying, grasping, etc. Novelties are typically thought of as the evolution of a morphological character, but there are also novelties at the molecular scale. One of the clearest examples being the origin of the V(D)J recombination system of the adaptive immune system of jawed vertebrates, which has allowed for a monumental exploration of phenotypes consisting of antibody repertoires. Novelties epitomise the idea that evolvability, or the capacity to encounter novel adaptive phenotypes, can evolve (Pigliucci 2008; Payne and Wagner 2019). As noted by Erwin (2017), “spaces evolve, not just the entities that occupy them”.

One of the most drastic novelties in organismal organisation was the evolution of multicellularity in the stem-lineage to animals. This novelty was followed by an impressive exploration of shapes and habits that make up some of the most exquisite and intriguing diversity populating our planet. In this thesis, I will particularly focus on two features of animal organisation that represented novelties that accompanied the evolution of multicellularity:

1) Enhancers as a regulatory novelty: Promoters are a common feature of gene regulatory activity in the three domains of life (Palmer and Daniels 1995; Cases and Lorenzo 1998; Fickett and Hatzigeorgiou 1997; Westmann et al. 2018), but in Eukaryotes, and in particular in animals, their activity is considerably more intricate than in Prokaryotes (Bylino, Ibragimov, and Shidlovskii 2020). In animals, promoter functions can be influenced by regulatory regions of the genome called enhancers, which are a metazoan innovation (Sebé-Pedrós et al. 2016). Like promoters, enhancers wire gene regulatory networks by binding transcription factors and helping recruit cofactors to initiate the transcription of target genes. But in contrast to promoters, enhancers tend to be found thousands of bases away from the transcription start sites of the genes they regulate, with some reports describing enhancers being at a distance of over a million base pairs from their target genes (Lettice et al. 2003). Enhancers interact with promoters in the tridimensional space of the nucleus, which is enabled by the formation of loops of accessible chromatin around which enhancers help generate transcriptionally active micro-environments (Calhoun and Levine 2003; De Laat and Duboule 2013; Levine, Cattoglio, and Tjian 2014). Enhancer-promoter interactions are highly dynamic throughout development and strongly depend on factors such as the presence of so-called pioneer transcription factors that can make the DNA strand encoding the enhancer accessible to binding proteins by locally depleting nucleosomes across cellular and developmental contexts (Zaret and Carroll 2011). The dynamics of enhancer activation and inactivation permits the differential deployment of distinct regulatory sub-networks in different cells, which helps define cell-type–specific spatiotemporal gene expression patterns (Davidson and Levine 2008; Spitz and Furlong 2012).

2) Germline-soma separation: The differentiation of gametes is key in organisms with sexual reproduction. All living animals have the potential to produce sperm and eggs, even those animals that can reproduce by fission, such as sponges, anthozoans or planarians. Nevertheless, there is a diversity of ways in which the germline is determined during animal development. In some lineages, which cells become the germline is defined early on during embryogenesis, where the differentiation of germ cells is determined by the inheritance of maternal effects,

while in other species the differentiation of the germ line is induced later in development by signals from surrounding tissues (Extavour and Akam 2003). It has been suggested that the origination of a germline would have permitted the stabilization of the cooperative multicellular system by silencing the potential to evolve cheaters (Reeve and Keller 1999). The need to protect the heritable material has had consequences such as the evolution of mechanisms for the maintenance of telomere length (Shore 1997), as well as a decreased mutation rate in the germline relative to somatic cells (Milholland et al. 2017). This has implications such as how the soma may decay more promptly as suggested by the disposable soma theory (Kirkwood 2017), but it also has the implication that genotype spaces will be explored widely and non-heritably in somatic cells.

1.5 Thesis outline

In this thesis, I will explore how the evolution of enhancers and the germline-soma separation affected some evolutionary processes in animals.

In the research presented in chapter 2, I studied the regulatory maturation of genes over evolutionary timescales. To do this, I integrated several publicly available datasets of different kinds (genomic, transcriptomic, epigenomic, etc.), and I studied the regulatory complexity of protein-coding sequences that originated at different timepoints during the evolutionary history of the lineage leading to mouse. I measured regulatory complexity as the number of enhancer interactions per gene. The insights I gained in this study support the notion that genes that have been conserved for longer periods of time, spanning from a few thousands of years ago all the way to the origin of cellular life, tend to evolve a higher regulatory complexity that helps them sustain functional robustness.

In chapter 3, I will present the results of studying how enhancers can offer a substrate for the early functionalization of budding genes. I will show how many open reading frames that are candidate to be *de novo* genes overlap genomic regions that are putative enhancers and how such open reading frames have higher evidence of functionalization than their counterparts arising far from regulatory elements. Furthermore, I will also show strong evidence suggesting

that at least a few protein-coding genes in the mouse genome stem from genomic regions that are homologous to enhancers in other species. These results suggest that enhancers might facilitate the evolution of novel functional peptides by enabling their early transcription. Thus, enhancers would play a double role in the evolution of gene regulatory networks, by helping in structure its wiring and its changing, as well as helping new genes integrate into them.

In chapter 4, I will present how the separation of reproductive and somatic cells during the development of animals can influence evolutionary trajectories. I will explore a hypothesis by which the enormous genetic variation existing in cells with no heritable potential can promote adaptation by revealing that a population is mutationally close to regions of the genotype space that map to adaptive phenotypes. I will present the results of simulations of evolving populations of developing organisms with an impermeable germline-soma separation navigating a minimal fitness landscape. Those simulations revealed the conditions under which non-heritable somatic mutations promote adaptation. Specifically, this can occur when the somatic mutation supply is high, when few cells with the advantageous somatic mutation are required to increase organismal fitness, and when the somatic mutation also confers a selective advantage at the cellular level. These results therefore provide proof-of-principle that non-heritable somatic mutations can promote adaptive evolution.

I will finish by discussing some perspectives on these points and emphasising how the evolution of these novelties, the complex regulation involving enhancer and a germline-soma separation, might have defined how animal populations evolve. These discoveries have important implications for how we understand the evolutionary potential of organisms, showing how the evolution of complex molecular mechanisms can facilitate the evolution of completely novel biochemical components, and how the ontological configuration of an organism can guide the evolutionary trajectories of its lineage. I conclude by emphasizing the importance of developmental and historical processes for the understanding of biological evolution.

2 Enhancer interactions and the regulatory maturation of genes

2.1 Introduction

One of the keys to Metazoan complexity is its diversity of cell types (Carroll 2001; Sebé-Pedrós et al. 2018b; Ros-Rocher et al. 2021). The identity of each cell type is specified by a particular combination of active genes that define what are the specialized functions a cell can perform. The different combinations of genes that are active in a cell are a consequence of well-integrated regulatory programs that coordinate cellular and tissue activities across the body of an animal, a fundamental component of which are enhancers. Enhancers are distally acting cis-regulatory elements that complement the activity of promoters by binding transcription factors and by helping recruit the transcriptional machinery to localized regions of the nucleus. Enhancers thus help direct specific gene expression patterns. Having originated in stem animals (Sebé-Pedrós et al. 2016; Sebé-Pedrós et al. 2018b), enhancers have evolved to become essential for the gene regulatory functions of organisms of this clade and they have strongly influenced the evolution of genomic architecture of metazoans (Nelson, Hersh, and Carroll 2004).

Over evolutionary time, enhancers have facilitated the search for adaptive phenotypes by remodelling regulatory networks. What enables this search is that mutations in enhancer sequences can create or ablate interactions with regulatory proteins, thus enabling modifications in gene use without affecting gene product or its expression in undesired developmental contexts (Prud'homme, Gompel, and Carroll 2007; Carroll 2008). Such remodelling of gene reg-

ulatory networks can cause changes in gene expression patterns that embody or lead to evolutionary adaptations or innovations (Peter and Davidson 2011). Examples include the archetypical pentadactyl limb anatomy of extant tetrapods (Kherdjemil et al. 2016), the ocular regression in subterranean rodents (Partha et al. 2017; Roscito et al. 2018), the loss of limbs in snakes (Kvon et al. 2016; Roscito et al. 2018), the convergent pigmentation patterns in East African cichlids (Kratochwil et al. 2018), the diversity of butterfly wing patterns (Wallbank et al. 2016), and the mammalian neocortex (Emera et al. 2016).

Each gene in a genome can be regulated by many enhancers during the lifetime of an animal (Levine and Tjian 2003; Cannavò et al. 2016). This is because the chromatin architecture is dynamic, which permits genes to be regulated by different repertoires of enhancers in time (during development) and space (across different cell types and tissues) (Gao et al. 2018; Cusanovich et al. 2018b). Furthermore, a gene can have its expression modulated by more than one enhancer simultaneously in a single cell (Hong, Hendrix, and Levine 2008; Fulco et al. 2016; Tsai, Alves, and Crocker 2019). Such seemingly redundant interactions can help ensure patterns of expression of genes, fine-tuning promoter activity (Perry, Boettiger, and Levine 2011) and providing organisms with phenotypic robustness against environmental (Tsai, Alves, and Crocker 2019) or genetic (Cannavò et al. 2016; Osterwalder et al. 2018) perturbations. The number of enhancers targeting a gene is not only dynamic during development, but it can also dramatically vary over evolutionary time as the result of the gain and loss of enhancer interactions. Enhancers can be lost, for example, by the accumulation of mutations that degenerate transcription factor binding sites (Kvon et al. 2016; Kvon et al. 2020; Fuqua et al. 2020), while they can be gained by means such as transposition (Lynch et al. 2015), gene duplication (Dorshorst et al. 2015) and the gradual emergence of transcription factor biding sites (Emera et al. 2016; Fong and Capra 2021). This gain and loss of enhancers can occur rapidly in evolutionary time when compared to the evolution rates of promoters or protein-coding genes, which likely underlies the high turnover of enhancers that has been observed in relatively recent phylogenetic branchings (Cotney et al. 2013; Vierstra et al. 2014; Villar et al. 2015; Prescott et al. 2015)

Because of the difference in numbers of enhancers with which a gene interacts, genes can be said to have different degrees of regulatory complexity in terms of the input that is involved in influencing their activity (Berthelot et al. 2018). This complexity can considerably impact the evolvability of the expression of a gene and the phenotypes that could be affected by changes in that expression. For example, a higher regulatory complexity in terms of the number of enhancer interactions can lead to a higher stability of gene expression over evolutionary time (Berthelot et al. 2018), it can entail modularity in trait evolution (Kvon et al. 2016), and it can offer phenotypic robustness (Osterwalder et al. 2018; Tsai, Alves, and Crocker 2019). On top of that, the gain of enhancer interactions has been hypothesised to help newly arising genes integrate into regulatory networks and to find functional niches that might help them stabilise in genomes (Tautz and Domazet-Lošo 2011). To understand how the regulatory complexity of genes has evolved and how it helps integrate genes into regulatory networks, we here study the number of enhancer interactions of thousands of open reading frames transcribed in mouse tissues.

2.2 Results

2.2.1 The regulatory complexity of thousands of transcribed open reading frames

To study the evolution of gene regulatory complexity in terms of enhancer interactions, we considered the regulatory background of 30,585 open reading frames (ORFs) encoded in the mouse genome that were reported as transcribed in liver, kidney and/or brain (Schmitz, Ullrich, and Bornberg-Bauer 2018). One way in which the interaction between enhancers and their target genes can be predicted is based on the detection of proximal genomic regions of open chromatin that are co-accessible within a single cell (Pliner et al. 2018). Using this principle, Cusanovich et al (2018a) reported thousands of potential interactions deduced from patterns of chromatin accessibility obtained by ATAC-sequencing in more than 100,000 cells corresponding to roughly 50 cell types from across 13 mouse tissues. We considered the subset of these predicted interactions that involved the first exon of each ORF, as well as regions

of open chromatin inferred to be enhancers. We consider a region of open chromatin to be candidate to be an enhancer when it overlapped epigenetic marks indicative of enhancer activity. Specifically, for the evaluation of the epigenetic status of each region of accessible chromatin, we considered chromatin immunoprecipitation sequencing (ChIP-seq) data produced with antibodies against the histone modifications H3K27ac (acetylation of a lysine residue in the 27^{th} N-terminal position of histone H3), H3K4me1 (monoamethylation of histone H3 on lysine 4) and H3K4me3 (trimethylation of histone H3 on lysine 4) (Davis et al. 2018). The function of these modifications is still under investigation, but there are experimental indications that they are involved in stabilizing nuclear micro-environments by helping localize transcription factors in the case of H3K4me1 (Atlasi and Stunnenberg 2017; Gandara et al. 2022), and H3K27ac is typically taken as separating active enhancers from poised enhancers evidenced by H3K4me1 activity (Creyghton et al. 2010). H3K4me3 on the other hand is involved in promoter function (Guenther et al. 2007) and therefore we did not consider any region of accessible chromatin overlapping this mark as an enhancer. In short, we considered a gene to be interacting with an enhancer candidate if i) their chromatin was co-accessible in the same cell types (as done by Pliner et al (2018) and Cusanovich et al (2018a), ii) if they overlapped either epigenetic mark that is indicative of enhancer function (H3K27ac and H3K4me1), and iii) if they did not overlap an epigenetic mark indicative of promoter activity (H3K4me3).

We characterised the regulatory complexity of each of the 30,585 ORFs as the total number of unique interactions predicted with the aforementioned criteria across all 50 cell types identified by Cusanovich et al (2018a) based on patterns of chromatin accessibility. As an illustrative example of this characterisation, consider the case of the *5-HT_{2A}* gene. This gene codes for a receptor that is active in the nervous system of mammals and is involved in important coordinating functions of brain activity, which can be affected by the binding of psychedelic substances such as psilocybin and lysergic acid diethylamide (Glennon, Titeler, and McKenney 1984; Béïque et al. 2007), and that is also involved in renal functions (Kaur and Krishan 2020). The first exon of this gene is in a genomic region that is accessible across 7 cell types according to the ATAC-seq data produced by Cusanovich et al (2018a), which, expectedly, given

the known physiological functions, include inhibitory neurons and kidney cells of the loop of Henle and the collective duct (data not shown). Across these cell-types, the first exon of this gene is co-accessible with 10 other genomic regions that are candidate enhancers, based on the fact that they are overlapping genomic regions marked by H3K27ac and/or H3K4me1, but not H3K4me3 (Fig. 2.1A).

5-HT$_{2A}$ has a below average regulatory complexity among the total set of ORFs we considered. The distribution of the regulatory complexity per gene across all ORFs goes from 0 to 117 enhancer interactions per gene, with a median of 17 and a mean of 19.62 enhancer interactions (Fig. 2.1B). Many of the considered ORFs comprise coding-sequences of a recent evolutionary origin that do not necessarily have a cellular function. Considering only a subset of 12,734 annotated *bona fide* genes with a relatively deep conservation of their protein sequence (see Materials and Methods), sharply increased the median number of enhancer interactions to 23 and the mean to 24.04. The difference in the distributions of the number of enhancer interactions between the total set of ORFs and the subset of annotated genes is suggestive of a major difference between genes that are well established in gene regulatory networks and genes having a recent evolutionary origin. To further explore this point, we next studied more in detail the change in regulatory complexity as a function of gene age across different evolutionary scales.

2.2.2 The regulatory complexity of evolutionarily young open reading frames

To study how regulatory complexity evolved over the lifetime of genes, we compared the number of enhancers targeting ORFs of different phylogenetic ages. The age of each of the ORFs we considered was estimated by Schmitz et al (2018) using phylostratigraphy, a method based on the detection of homologous protein-coding genes in the genome of other organisms based on sequence similarity (Domazet-Lošo, Brajković, and Tautz 2007). We considered four of the phylostrata studied by Schmitz et al (2018), separating the total set of ORFs into those that emerged before the split of placentary mammals and marsupials, those that emerged before the common ancestor shared between primates and rodents, those that emerged before the

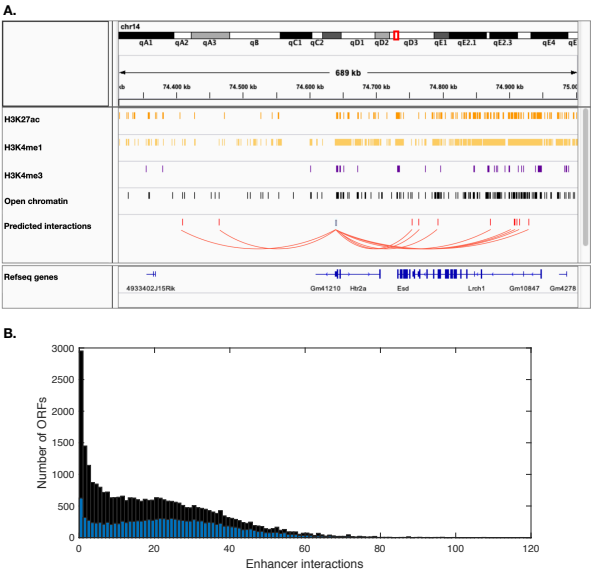


Figure 2.1: Regulatory complexity measured as the number of enhancer interactions. (A) Illustrative example of the regulatory profile of the gene *5-HT*$_{2A}$ based on the predicted interactions between the region coding for its first exon and enhancer candidates. We considered interactions to be putatively targeting a gene only when the first exon of that gene was detected as being in open chromatin in the same cell type as candidate enhancers (see Cusanovich et al 2018). We only considered regions of open chromatin interacting with the first exon of a gene to be an enhancer, if we identified it as being marked by the histone modifications H3K27ac and/or H3K4me1, which are indicative of enhancer activity, but not H3K4me3, which is indicative of promoter activity. The 10 enhancer interactions predicted for *5-HT*$_{2A}$ are shown in red in the lower track. The accessible chromatin for the first exon of *5-HT*$_{2A}$ is shown in blue, also in the lower track. (B) Distributions of the regulatory complexity of each ORF measured as the predicted number of enhancer interactions. In black is the total set of 30,585 ORFs identified as transcribed in the brain, liver and/or kidney of mouse (Schmitz et al, 2018). In blue is the subset of 12,734 ORFs that correspond to annotated *bona fide* genes conserved over macroevolutionary timescales.

common ancestor shared by mouse and rat, and those that emerged after that split (Fig. 2.2A). Nearly half (17725) of the total set of considered ORFs originated before the split between placental mammals and marsupials, while 795 corresponded to ORFs shared with human, 643 are shared with rat, and 11,422 were mouse specific. Among the mouse specific ORFs, Schmitz et al (2018) differentiated between those that are overlapping other pre-existing genes, which I

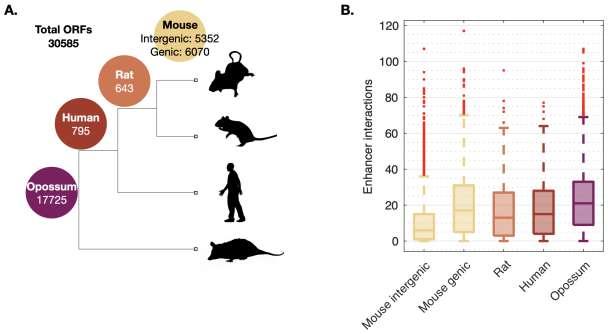


Figure 2.2: Enhancer interactions increase with gene age. (A) Phylogeny showing the four age classes of the 30,585 ORFs. The numbers on the branches indicate the number of ORFs that are either mouse-specific or shared with rat, human, and opossum. Mouse-specific ORFs are further classified as intergenic or genic. (B) Number of enhancer interactions per ORF as a function of the age category of each ORF.

will here call "genic", and those that are far from other genes, which I here refer to as "intergenic". From the total 11,422 mouse-specific ORFs, 5352 arose in intergenic genomic regions and 6070 arose in genic regions.

We uncovered a positive correlation between the age of an ORF and its number of enhancer interactions (Spearman's correlation coefficient $\rho = 0.23$, $P < 0.001$; Fig. 2.2B), with the number of enhancer interactions gradually increasing from a median of 10 for mouse-specific ORFs to a median of 13, 15, and 21 for ORFs that are shared with rat, human, and opossum, respectively. Among mouse-specific ORFs, intergenic ORFs had a median of 6 enhancer interactions, whereas genic ORFs had a median of 17, which makes them more similar to non-mouse-specific ORFs in their number of enhancer interactions. This suggests that many of the mouse-specific ORFs of genic origin may be co-opting the regulatory interactions of their host gene, or of nearby genes. Overall, younger genes tend to be regulated by fewer enhancer interactions than older genes, which is suggestive of a gradual increase in the regulatory complexity of genes and of their deeper embedding into regulatory networks as they age.

2.2.3 Enhancer interactions of loci with novel transcription

Our observation that enhancers are gradually acquired across roughy 160 My of mammalian evolution is consistent with the hypothesis that enhancers help integrate genes into regulatory networks as they are born (Tautz and Domazet-Lošo 2011). Ideally, we would have high-coverage transcriptomic data across a shallower phylogeny of mouse taxa, which would provide more convincing support for this hypothesis by facilitating the estimation of ORF age for the mouse-specific intergenic ORFs. To our knowledge, no such data exist. However, transcriptomic data from brain are available for mouse taxa spanning around 10 My of evolution (Neme and Tautz 2016). With these data, we were able to estimate when loci that encode for recently evolved intergenic mouse-specific open reading frames started to be transcribed in the evolution of our focal mouse species, *Mus musculus domesticus*.

The available transcriptomes were obtained from brain tissue from 10 different individuals belonging to different populations, subspecies, species and genera that diverged after the mouse-rat split. Specifically, the individuals come from 3 populations of *Mus musculus domesticus*, 2 populations of *M. m. musculus*, and 1 from *M. m. castaneus*, *M. spicilegus*, *M. spretus*, *M. mattheyi*, and *Apodemus uralensis*. We separated all mouse-specific intergeneic ORFs into 4 age categories (Fig. 2.3A), depending on whether expression could be detected in i) at least one of the three *M. m. domesticus* samples, ii) in *M. m. domesticus*, and also in at least one of the other *Mus musculus* subspecies, iii) in *M. m. domesticus*, at least one other subspecies of *M. musculus*, and at least one other *Mus* species, but not in the *Apodemus uralensis* sample, or iv) in *M. m. domesticus* and in *Apodemus uralensis*. We considered two different thresholds to consider a gene as being expressed in the lineage of *M. m. domesticus*, one corresponding to a minimum expression of at least one read, and the other to a more restrictive threshold of at least 20 reads. We assigned a total of 3,980 ORFs to each of the transcriptional age categories when using the low expression criterion, and 2,855 ORFs the more restrictive criterion.

In agreement with the trend we found studying ORFs of different age within the clade of mammals, we uncovered that mouse-specific intergenic ORFs whose expression can be detected in more modern branches of the recent mouse phylogeny have fewer enhancer interac-

tions than ORFs whose expression can be detected at more basal branches, considering both permissive (Fig. 2.3B) and more stringent (Fig. 2.3C) thresholds for the number of reads assigned to each ORF. This shows that ORFs that only recently acquired expression are less likely to be regulated by many interactions, and that these interactions may gradually be acquired as the transcription of a certain locus is stabilized over evolutionary timescales.

2.2.4 Trends in the regulatory complexity of genes with deep phylogenetic origins

To study the evolutionary patterns of regulatory complexity over macroevolutionary timescales, we shifted our focus to the subset of 12,734 ORFs that correspond to *bona fide* annotated genes (Fig. 2.1B, blue), all of which belong to the age class of opossum-shared ORFs. We separated this subset of ORFs into 15 new age classes sorted using phylostratigraphy by Neme & Tautz (2013) (Fig. 2.4A). These new age classes date back to the origin of cellular life, with phylostrata corresponding to major evolutionary events like the origin of eukaryotes (age class 14), animals (age class 11) and vertebrates (age class 5). With these age estimations, we again found a significant correlation between the age of a gene and its number of enhancer interactions (Spearman's correlation coefficient $\rho = 0.09$, $P < 0.001$; Fig. 2.4B).

Interestingly, the increase in the number of enhancer interactions over macroevolutionary time is not gradual and steady, but rather there is a saltational pattern. Based on the trend in Fig. 2.4B, we noted that groups of neighbouring age classes seem to have similar distributions of the number of enhancer interactions of their ORFs (Fig. 2.2B). To corroborate this, we used a two-sample Kolmogorov-Smirov to assess the probability that different number of enhancer interactions per ORF in each age class are drawn from the same distribution as any other age class. The results of these comparisons allowed us to group age classes based on whether the regulatory profile of the ORFs they contained is significantly different from the regulatory profile of their neighbouring age classes (Fig. 2.4C). The sole age class that was significantly different from all other age classes and could not be clustered with neighbouring age classes was age class 14 (Fig. 2.4B,C), which corresponds to ORFs shared between all eukaryotes, but not with bacteria or archaea. Overlooking this exception, we identified four groups of age classes

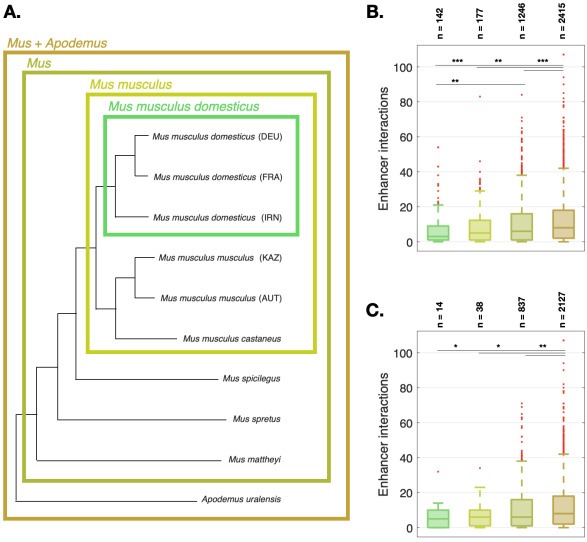


Figure 2.3: Mouse-specific intergenic ORFs with evidence of transcription at more basal branches of the mouse radiation tend to have more enhancer interactions. (A) Phylogeny adapted from Neme & Tautz (2016) showing ten mouse taxa that diverged after the mouse-rat split, for which brain transcriptomic data are available. Colored boxes indicate four estimates of ORF age, based on where in the mouse phylogeny an ORF from *Mus musculus domesticus* shows evidence of transcription. (B, C) Number of enhancer interactions of ORFs with evidence of transcription in only *M. m. domesticus* (green); in *M. m. domesticus* and at least one other *M. musculus subpsecies* (light olive green); in *M. m. domesticus*, at least one other *M. musculus* subspecies, and at least one other *Mus* species (dark olive green); and in *M. m. domesticus* and *Apodemus uralensis* (light brown). In B) we consider a permissive threshold of at least one read associated to an ORF in at least one of the *M. m. domesticus* samples, while in C) we consider a more stringent threshold of at least twenty reads associated to an ORF across the three *M. m. domesticus* samples. The number of ORFs assigned to each category are shown above the panels. Significant comparison based on Wilcoxon's signed rank test are indicated by asterisks, where * indicates $p < 0.05$, ** $p < 0.001$, and *** $p < 0.0001$.

(Fig. 2.4B,C) corresponding to ORFs shared between all amniotes (age classes 1 and 2), ORFs shared between vertebrates and urochordates (age classes 3 to 6), ORFs that emerged after the origin of animals (age classes 7 to 10), and ORFs that predate the common ancestor shared between sponges and all other animals (age classes 11 to 15). Reclassifying ORFs on the ba-

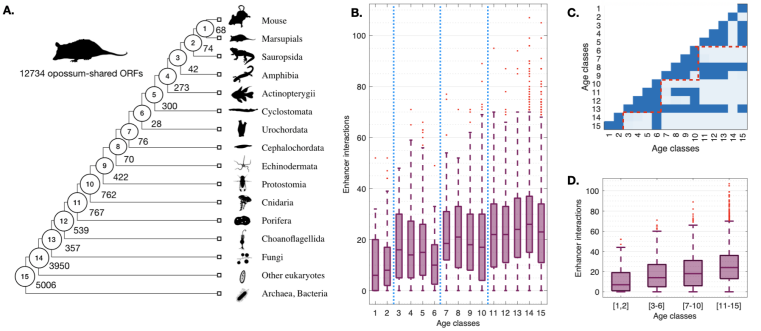


Figure 2.4: Enhancer interactions of genes with a deep phylogenetic origin increase saltationally with gene age. (A) Phylogeny adapted from Neme and Tautz (2013). The numbered circles indicate lineages representative of the age classes to which we assigned 12,734 opossum-shared ORFs, each of which corresponds to an annotated gene. The numbers on each branch represent the total number of annotated genes assigned to each age class. (B) Number of enhancer interactions per gene, shown in relation to the age classes depicted in (A). Blue dotted lines indicate the grouping of age classes based on the similarity of the distribution of the number of enhancer interactions per ORF (see main text). (C) Significance of the two-sample Kolmogorov-Smirnov test performed in each pair of age classes. Dark blue corresponds to non-significant comparisons ($p > 0.05$) and light blue corresponds to comparisons for which the test supports that the distribution of enhancer interactions per ORF are different between each pair of age classes. Red dashed lines indicate the grouping of age classes based on the similarity of the distribution of the number of enhancer interactions per ORF. (D) Number of enhancer interactions per gene, shown in relation to the groupings of age classes based on based on the similarity of the distribution of the number of enhancer interactions per ORF.

sis of these groupings shows an even stronger correlation between ORF age and the number of enhancer interactions per ORF (Spearman's correlation coefficient $\rho = 0.16$, $P < 0.001$; Fig. 2.4D). Therefore, although there is a trend in the acquisition of regulatory interactions as genes age over deep macroevolutionary timescales, there seem to be key evolutionary breakpoints during the millions of years of evolution separating different age classes that ultimately define what is going to be the regulatory complexity of genes that arise thereafter.

2.2.5 Expression breadth and homogeneity is influenced by the number of regulatory interactions

The results shown above allowed us to uncovered a positive correlation between the age of an ORF and its number of enhancer interactions across three evolutionary timescales, a shallow phylogeny spanning ∼10 My of murine evolution, a phylogeny spanning ∼160 My of mammalian evolution, and a deep phylogeny dating back to the origin of cellular life. To understand the functional implications of the increase in regulatory interactions over macroevolutionary time, we studied the expression breadth of opossum-shared annotated genes using single-cell transcriptomic data from 68 cell types of ten murine tissues (Consortium 2018), for which single-cell chromatin accessibility data were also available (Materials and Methods). We found that expression breadth increases with the number of enhancer interactions (Spearman's correlation coefficient $\rho = 0.49$, $p < 0.001$; Fig. 2.6D) and with gene age (Spearman's correlation coefficient $\rho = 0.13$, $p < 0.001$). The latter observation corroborates previous findings based on transcriptomic data from whole tissues (Kryuchkova-Mostacci and Robinson-Rechavi 2015). We next quantified the expression homogeneity of each gene across all cell types where expression was measurable. We calculated expression homogeneity as the entropy of the expression of genes measured across cell types where expression was detectable (Materials and Methods), uncovering a positive correlation between expression homogeneity and the number of enhancer interactions (Spearman's correlation coefficient $\rho = 0.45$, $p < 0.001$; Fig. 2.6E), as well as gene age (Spearman's correlation coefficient $\rho = 0.11$, $p < 0.001$). From these results, we can infer that as a consequence of a gain of enhancer interactions, a gene can expand the breadth of its expression and how homogeneous its activity is across the body.

2.3 Discussion

To understand how genes integrate into regulatory networks as they mature, we here compared the number of enhancer interactions affecting genomic regions encoding for thousands

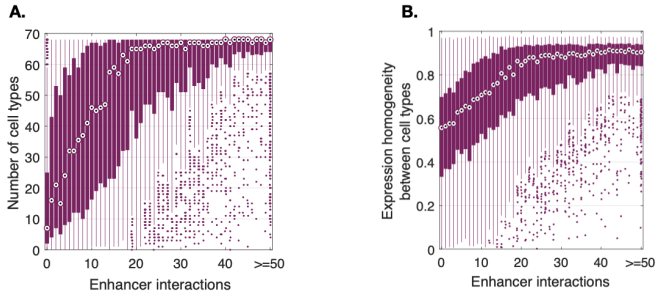


Figure 2.5: Impact of the number of enhancer interactions on the breadth and homogeneity of gene expression. (A) Expression breadth and (B) homogeneity of opossum-shared annotated genes as a function of the number of enhancer interactions.

of ORFs transcribed in mouse tissues, showing that the number of enhancers targeting an ORF tends to increase with its phylogenetic age. We detected this correlation at three evolutionary timescales, for loci that began being transcribed in recent branchings of the mouse phylogeny, for young open reading frames that are exclusive to different lineages of placentary mammals, and for genes that originated at some phylogenetic time-point between the origin of cellular life and the origin of mammals.

Our results have important implications for the understanding of how gene regulatory networks evolve in animals. In the context of gene regulatory network models, enhancers can be seen as mediators of the interactions between transcription factors and their target genes. Within this framework, an increase in enhancer connections also implies an increase in the number of edges connecting target genes to the transcription factors that bind those enhancers. The gain of enhancer connections can therefore either build new connections between a gene and a transcription factor, or, if such connections already exist, they can reinforce them. We can then say that the accumulation of enhancer interactions over evolutionary time causes genes to become more and more entrenched into regulatory networks as they age. The deeper integration of genes into regulatory networks as genes age has also been studied in yeast from the point of view of the gain of connections with transcription factors (Carvunis et al. 2012;

Abrusán 2013). Yeast are unicellular eukaryotes which lack enhancers and rely only on promoters and so-called upstream regulator sequences to congregate transcription factors for the regulation of gene expression (Phillips and Hoopes 2008). The fact that both animals and yeast show a similar trend in the integration of genes into regulatory networks in spite of their major differences in the mechanisms of gene regulation, suggests that the increase in connectivity as genes age is a common evolutionary trend in genome functionality. Indeed, other mechanisms of gene regulation also show a stronger connectivity in genes as they age. For example, older genes of the human genome tend to be more regulated at the post-transcriptional level by harbouring more miRNA targets and being more likely to be affected by nonsense-mediated decay and RNA editing (Warnefors and Eyre-Walker 2011), and the gene products of yeast and animals have been found to be more deeply interconnected in protein-protein interaction networks (Capra, Pollard, and Singh 2010; Zhang et al. 2015).

The deeper embedding of genes into regulatory networks as a consequence of the gain of enhancer interactions can have major repercussions for the functionality and the evolutionary potential of genes. As mentioned above, having a higher number of enhancer interactions can offer robustness to the expression of genes against environmental and genetic perturbations (Cannavò et al. 2016; Osterwalder et al. 2018; Tsai, Alves, and Crocker 2019; Kvon et al. 2021). Consequently, genes that are surrounded by more enhancers tend to have levels of gene expression that are more stable over evolutionary time, as has been shown for genes expressed in the liver of mammals (Berthelot et al. 2018). Since genes gain enhancer interactions as they age, they can thus evolve a higher expression stability, which could result in an entrenchment of genes into cellular networks to the point that they become essential for important biological functions.

Our study also supports the idea that an increase in the number of enhancer interactions can increase the functional diversification of genes. Having more enhancer interactions correlates with an expansion of the breadth and homogeneity of expression of genes across different cellular environments, which would allow genes to explore different organismal contexts in which a gene could impact phenotypes in an adaptive manner. The fact that the gain of en-

hancers can increase the spatial breadth of a gene's activity was already well known (Cotney et al. 2013; Prescott et al. 2015) and it has been hypothesised that the gain of enhancers could be behind the expansion of the expression of new genes, which tend to originate in localised developmental environment (Tautz and Domazet-Lošo 2011). By showing that enhancer interactions accumulate as genes grow older and that this can result in an expansion of the number of cell types where genes are expressed, we offer empirical support for this hypothesis, thus illuminating mechanisms for the functional and evolutionary maturation of genes.

We hypothesize that the increase in expression stability and the broadening of the organismal contexts where genes are expressed that might result from increases in enhancer interactions can explain the trends we observed when studying genes of deep phylogenetic origins. In that analysis we detected four profiles of enhancer interactions that correspond to sets of genes that arose before or after major morphological and molecular innovations in the history of animal evolution. The oldest of those genes predate the origins of metazoans. The ancestor to animals was likely a unicellular organism with the potential to occasionally aggregate, as it happens in choanoflagellates (Fairclough, Dayel, and King 2010). Therefore most of the oldest genes in animal genomes probably arose in the genome of a unicellular organism, and the functions it evolved to perform were functions that might impact the basic physiology of cells. Thus, it is reasonable to infer that genes with such fundamental cellular functions would evolve a higher degree of functional robustness, which can be attained by being targeted by a greater number of regulatory elements. An alternative, but not mutually exclusive interpretation, is that the evolution of a higher regulatory complexity might be related to the evolution of multicellularity in animals. It would be important for organismal stability of multicellualr organisms to contain gene functions that evolved originally in a single-celled organisms, given that functions that are beneficial for individual cells might jeopardize the livelihood of a society of cells. Indeed, mutations leading to the disregulation of genes shared between animals and unicellular organisms that are involved in single-cell processes can lead to excessive proliferation resulting in cancers (Trigos et al. 2019). The mutational containment of potential cheating cells via a fine-grained regulation could therefore help explain the more complex regulation of

the oldest genes predating the origin of metazoans.

Animal organismal biology is characterised by a multicellular organization in which different cell types are defined by different patterns of gene expression (Sebé-Pedrós et al. 2016; Sebé-Pedrós et al. 2018a; Sebé-Pedrós et al. 2018b). As a consequence, whenever new beneficial genes arise in the animal lineage, they tend to first be expressed in a tissue-specific manner (Levine et al. 2006; Toll-Riera et al. 2009; Tautz and Domazet-Lošo 2011). A new gene that can functionally integrate in the cellular physiology of, say, a neuron, might be deleterious if it is expressed in a different cell type. This means that selection can act against the ectopic expression of genes and that any mutation that might promote it, as has been suggested to explain evolutionary constraints on highly pleiotropic enhancers (Fuqua et al. 2020). This could help explain why, although there is a general trend of increasing the number of enhancer interactions as genes age, this increase seems to saturate across millions of years of evolution separating neighbouring deep phylogenetic age classes, since further gaining enhancer interactions could expand the breadth of expression of genes into tissues where such genes might be disruptive. Therefore, the first breakpoint in the trend of regulatory complexity of genes that emerged after the origins of metazoans could be explained by the evolution of division of labour among cell types. The second breakpoint occurs after the branching of the lineage leading to vertebrates and urochordates from the lineage leading to cephalochordates. Sometime after this branching there was a whole-genome duplication event that lead to major reorganizations of the genomic architecture in the lineage leading to vertebrates (Marlétaz et al. 2018). Enhancers played a key role in the events that followed this duplication, enabling major rewirings of regulatory networks that allowed for an even higher specialisation of gene expression across cell types (Marlétaz et al. 2018) and that helped greatly increase the interconnectivity between signaling pathways, which facilitated the formation of new cell types and morphological novelties (Gil-Gálvez et al. 2022). Thus, genes that emerged anew after this phyletic timepoint were probably even more constrained than older genes in the breadth of expression they could attain. The last breakpoint corresponding to the origin of amniotes is not characterised, to our knowledge, by any major shift in regulatory functions. But it is, how-

ever, marked by a major innovation in the lifecycle, with innovations like the amniotic egg and the elimination of a feeding larval stage. We presume that such a drastic change in basic embryology in this lineage, which involved innovations at the level of cell types and tissues such as the amnion, might be implicated also in a constraint in the breadth of expression for newer genes. Thus, we suggest that the trend in the gain of enhancer interactions over evolutionary time, might be saturated for newly born genes as organisms increase their biological complexity in terms of specialized cell types and morphological innovations over millions of years of evolution.

Another consequence of increasing enhancer interactions is how a higher regulatory complexity might facilitate the evolvability of regulatory sequences. By offering mutational robustness to the regulation of genes (Cannavò et al. 2016; Osterwalder et al. 2018; Wang and Goldstein 2020), a higher number of enhancer interactions targeting a gene can lead to the accumulation of genetic diversity in regulatory sequences. This reasoning finds support from results obtained by Danko et al (2018), who, by studying the number of enhancer interactions affecting genes expressed in CD4+ T cells of different primate species, found that the more enhancers targeting a gene, the lower the conservation at putative binding sites in enhancers that act in redundancy. The genetic diversity that can be accumulated as a result of the robustness offered by an increased number of interactions can promote the neutral exploration of the space of genotypes (Wagner 2008). This might eventually bring these regulatory sequences mutationally closer to novel binding sites, which might help build new connections in regulatory networks without affecting the pre-existing function of the regulatory element in question (Ciliberti, Martin, and Wagner 2007). Thus, by building new transcription factor binding sites a regulatory sequence could produce new phenotypes (Noon, Davis, and Stern 2016). Therefore, the higher mutational robustness that might be offered by an increase in the number of enhancers over evolutionary time, might also increase the evolvability of the regulation of the target gene.

Overall, we provide novel insights about the regulatory maturation of genes, showing how genes, as they age, can become more deeply integrated into regulatory networks by acquiring

enhancer interactions. This deeper integration can have evolutionary conseuqences such as the stabilization of the expression of genes, but also the expansion of the breadth of expression across different tissues and, potentially, it can increase the evolvability of gene regulation. Our results offer a perspective of how genes can gradually evolve a higher regulatory complexity after they are born. However, an open question still remains and that is what is the original regulatory background of genes in the moment they are born. In the next chapter we explore the idea that enhancers not only help newly born genes integrate into regulatory networks, but they can also help in the birth process itself.

2.4 Methods

ORF Age and Classification. Schmitz et al. (2018) identified a set of 58,864 ORFs from the transcriptomes of three murine tissues: liver, brain, and testis. Blasting against the transcriptomes of four other mammalian species (rat, human, kangaroo rat, and opossum), they estimated the age of each ORF by phylostratigraphic methods (Domazet-Lošo et al. 2007; Schmitz et al. 2018). Because of the small number of ORFs shared with the kangaroo rat (49 ORFs), we merged these ORFs together with those from the rat age class. We used the genomic coordinates of the first exon of each ORF in the mm10 mouse genome reference to study the regulatory properties of ORFs of different ages, for example, to study their distance to the nearest enhancer. We only considered ORFs that were transcribed from nuclear chromosomes and whose first exon was longer than 30 base pairs. If first exons were shared between more than one ORF, we only retained the oldest of the ORFs. Our filtered data set contained 56,262 ORFs before selecting those that had their first exon overlapping regions of open chromatin (see below).

Schmitz et al. (2018) annotated each ORF as belonging to one of eight different categories: "intergenic," "close to promoter same strand," "close to promoter opposite strand," "overlapping same strand," "overlapping opposite strand," "overlapping coding sequence same strand," "overlapping coding sequence opposite strand," and "overlapping annotated gene in frame." We considered all categories except "intergenic" to be "genic" in order to separate ORFs that

were born within or near existing genes from those that were not. This resulted in five classes: mouse-specific intergenic ORFs, mouse-specific genic ORFs, rat-shared ORFs, human-shared ORFs, and opossum-shared ORFs.

Chromatin Accessibility. We used single-cell ATAC-seq data from 13 different mouse tissues (bone marrow, cerebellum, large intestine, heart, small intestine, kidney, liver, lung, cortex, spleen, testes, thymus, and whole brain). We obtained the data from the Mouse ATAC atlas (Cusanovich et al. 2018), which composed 436,206 peaks of open chromatin. We used liftOver to convert the genome coordinates from mm9 to mm10. A total of 29 peaks could not be converted. Using the "closest" function of bedtools with the "-t first" option activated, we calculated the distance between ORFs and regions of open chromatin. We annotated regions of open chromatin as enhancers if they overlapped H3K27ac and/or H3K4me1 peaks but not H3K4me3 peaks, or as promoters if they overlapped H3K4me3 peaks. To do so, we used the bedtools intersect function with the -u option activated.

Cusanovich et al. (2018) used these single-cell ATAC-seq data to identify clusters of cells with similar patterns of chromatin accessibility. They assigned the clusters to 38 distinct cell types based on the chromatin accessibility of marker genes indicative of each cell type. We used these data to identify tissues and cell types where ORFs are in accessible chromatin. We considered an ORF-containing region of the genome to be in open chromatin in a certain cell type if it was accessible in at least 1% of the cells that made up at least one of the clusters of that cell type (Cusanovich et al. 2018).

Enhancer Interactions. Cusanovich et al. (2018) used single-cell ATAC-seq data to predict physical interactions between regions of open chromatin (Pliner et al. 2018), thus creating an atlas of enhancer interactions in single murine cells. We downloaded these data from the Mouse ATAC atlas (Cusanovich et al. 2018), which includes the cell clusters where the interactions occur, as well as the coaccessibility scores of pairs of regions of open chromatin—a measure of interaction strength. We disregarded cell clusters classified as "unknown" or "collisions," as well as interactions with a coaccessibility score < 0.25, following Pliner et al. (2018).

We also filtered out interactions with regions of open chromatin that overlapped ChIP-seq peaks for H3K4me3 marks or no enhancer marks, in order to focus solely on interactions with enhancers. An interaction was assigned to an ORF if the ORF's first exon was included in the interaction.

Expression within the Mouse Lineage. We considered the transcriptomes of brain from ten different mouse taxa that diverged after the mouse-rat split (three populations of *Mus musculus domesticus*, two populations of *M. m. musculus*, and one from *M. m. castaneus*, *M. spicilegus*, *M. spretus*, *M. mattheyi*, and *Apodemus uralensis*) (Neme and Tautz 2016). The data consisted of read counts from the transcriptomes of each taxon mapped to 200-bp windows of the mm10 mouse reference genome. We assigned each ORF to one of the 200-bp windows if the middle point of the ORF's first exon mapped to that window. For this analysis, we only considered mouse-specific intergenic ORFs that overlapped regions of open chromatin. We considered two different thresholds to evidence transcription of an ORF. Using the first, more permissive threshold, we only considered ORFs that had at least one read mapping to its 200-bp window in at least one of the three samples of *M. m. domesticus*. This resulted in 4,104 ORFs. Using the second, more stringent threshold, we only considered ORFs that had at least 20 reads from across the three samples of *M. m. domesticus* mapping to the ORF's 200-bp window. This resulted in 2,864 ORFs. We separated these ORFs into four age categories (Fig. 2.3), depending on whether expression could be detected using a highly conservative threshold of just a single read in 1) at least one of the three *M. m. domesticus* samples, 2) in *M. m. domesticus*, and also in at least one of the other *Mus musculus* subspecies, 3) in M. m. domesticus, at least one other subspecies of *M. musculus*, and at least one other *Mus* species, but not in the *A.uralensis* sample, or 4) in *M. m. domesticus* and in *A.uralensis*. We assigned a total of 3,980 ORFs to each of these categories when considering ORFs with at least 1 read detectable in the *M. m. domesticus* clade (97%), and 2,855 ORFs when considering ORFs with at least 20 reads detectable in the *M. m. domesticus* clade (99.7%).

Because of the low coverage of the transcriptomic data (1 sequencing depth), there is increased uncertainty in our estimation of ORF ages relative to the other phylogenies consid-

ered in this study. This is especially true of ORFs that are expressed at low levels, which are less likely to be detected across the phylogeny and are therefore more susceptible to the underestimation of their ages. We were therefore concerned by the observed positive correlation between the expression level of an ORF and its estimated age (Spearman's correlation coefficient $\rho = 0.20$, $P < 0.001$). To ameliorate this concern, we determined the probability of underestimating the age of an ORF in the *M. m. domesticus* clade under our most stringent detection limit of 20 reads per 200-bp window. Specifically, for each ORF assigned to the *M. m. domesticus* clade, we used the binomial formula to calculate the probability of underestimating the ORF's age due to lack of detection in the other seven clades, under the assumption that the ORF actually emerged at the base of the phylogeny and is expressed at the same low level across the phylogeny. This probability is low: given 27.76×10^8 trials (the total number of reads from the seven samples not in the *M. m. domesticus* clade) and a probability of success of $20/10.75 \times 10^8$ (the minimum fraction of reads from the three samples in the *M. m. domesticus* clade mapping to the ORF's 200-bp window), the probability of observing zero reads mapping to the ORF's 200-bp window in all of the seven samples from the outgroup is 1.4×10^{-13}, after Bonferroni correction for 14 tests (the number of ORFs assigned to *M. m. domesticus*).

Age of Annotated Genes. To study how genes acquire enhancer interactions over macroevolutionary timescales, we considered the subset of ORFs that belong to the opossum age class in Schmitz et al. (2018) and that are annotated as genes in the latest version of Ensembl (release 95) (Cunningham et al. 2019). We matched these genes to age estimates reported by Neme and Tautz (2013), based on a phylostratigraphic analysis of 20 lineages spanning 4 Gy from the last universal common ancestor to the common ancestor of mouse and rat. We further filtered the data set to only include ORFs that emerged in the first 15 of the 20 phylostrata, in order to focus on ORFs that are considered to have emerged before the split between the common ancestor of placental mammals and marsupials by both Schmitz et al. (2018) and Neme and Tautz (2013). This left us with 16,000 ORFs corresponding to 12,734 unique annotated genes that emerged prior to the origin of placental mammals.

Expression Breadth and Homogeneity of Annotated Genes. To study the transcription of annotated genes, we used the expression data reported by the Tabula Muris Consortium (2018) for the single-cell RNA sequencing performed with FACS-based cell capture in plates, for 20 different mouse tissues. The data include the log-normalization of 1+counts per million for each of the annotated genes in each of the sequenced cells. We considered ten tissues that were also used for the construction of the Mouse ATAC Atlas (Cusanovich et al. 2018). We measured the expression breadth of each ORF corresponding to an annotated gene as the number of cell types in which expression could be detected in at least 1% of the cells assigned to a cell type. The homogeneity of expression of a gene k across cell types was calculated as:

$$H_k = -\sum_{(i=1)}^{n} \frac{FPKM_i}{\sum_{j=1}^{n} FPKM_j} log_n(\frac{FPKM_i}{\sum_{j=1}^{n} FPKM_j}) \tag{2.1}$$

where n is the number of cell types where expression was detectable.

3 The transmutation of enhancers into protein-coding genes

3.1 Introduction

In his *Natural Selection and the Concept of a Protein Space* (1970), John Maynard Smith clarified how in spite of the vastness of a protein space, evolution by natural selection could find novel functional proteins by modifying pre-existing proteins through small mutational steps. Under this scenario, proteins only needed to arise once in the history of life and then they could diversify by exploring their neighbourhoods of functional variants. At the end of that piece, Maynard Smith asked whether all existing proteins were "part of the same continuous network, and if so, have they all been reached from a single starting point?" François Jacob was of the opinion that the answer to that question would be affirmative deeming the probability of a new protein arising *de novo* to be null (Jacob 1977). But as I teased in the introduction, it is becoming increasingly appreciated that new genes can arise *de novo* from non-coding regions of the genome, and that the essential prerequisites for protein-coding genes to arise *de novo*, which are the formation of an open reading frame (ORF), and the transcription and translation of that ORF, are often met (Carvunis et al. 2012; Li et al. 2014; McLysaght and Hurst 2016; Van Oss and Carvunis 2019; Willemsen, Félez-Sánchez, and Bravo 2019).

Some of the first reports of *de novo* genes come from the comparison of transcripts from the male reproductive system of different *Drosophila* species (Levine et al. 2006; Begun et al. 2007). In those studies, a handful of protein-coding sequences were detected as transcribed in some drosophilids that had no homologs even in closely related species. Subsequently, numerous

other *de novo* gene candidates were identified in yeast (Cai et al. 2008; Carvunis et al. 2012), in humans (Toll-Riera et al. 2009; Knowles and McLysaght 2009), and across several other branches of the tree of life (Schmitz, Ullrich, and Bornberg-Bauer 2018; Baalsrud et al. 2018; Xie et al. 2019a; Zhuang et al. 2019). The support for *de novo* gene origination does not only come from the finding of homologous sequences. There is also evidence of the translation of peptides encoded by candidate *de novo* genes (Zhang et al. 2019), as well as functional characterizations (Xie et al. 2019a), and reconstructions of the evolutionary steps that lead from non-coding intergenic sequences to the evolution of particular protein-coding genes (Zhuang et al. 2019). Because much of the genome is transcribed (Kapranov, Willingham, and Gingeras 2007; Neme and Tautz 2016) and many lineage-specific transcripts containing ORFs show evidence of translation (Wilson and Masel 2011; Ruiz-Orera et al. 2014; Ruiz-Orera et al. 2018; Prabh and Rödelsperger 2016; Schmitz, Ullrich, and Bornberg-Bauer 2018; Ruiz-Orera and Albà 2019; Zhang et al. 2019), the *de novo* evolution of new protein-coding genes offers a great potential for the growth of genetic repertoires.

An important question concerning new genes—those that have arisen *de novo* or by other means—is how they integrate into existing regulatory networks. As shown in the previous chapter, enhancer acquisition may allow new genes to expand their breadth of expression, providing opportunities to acquire new functions in different cellular contexts. Enhancers may therefore help new genes integrate into existing regulatory networks via edge formation and rewiring. Less appreciated is the role enhancers may play in the origination of *de novo* genes (Wu and Sharp 2013), and thus in the growth of those gene regulatory networks. The physical proximity between active enhancers and their target genes (Levine, Cattoglio, and Tjian 2014)—facilitated by DNA looping—creates a transcriptionally permissive environment that is engaged with RNA polymerase II, which may lead to the transcription of DNA near the enhancer, or to the transcription of the enhancer itself, producing so-called enhancer RNA (De Santa et al. 2010; Kim et al. 2010; Li, Notani, and Rosenfeld 2016; Haberle and Stark 2018). If the transcript contains an ORF, then such increased transcription will increase the likelihood of interaction with ribosomes, and because enhancers are typically active in a small number

of cell types (He et al. 2014), interactions with ribosomes will occur in a limited diversity of cellular contexts. This may help purge toxic peptides and enrich for benign peptides, a process that has been hypothesized to increase the likelihood of *de novo* gene birth (Wilson and Masel 2011). Moreover, similarities in the architectures of enhancers and promoters may facilitate the regulatory repurposing of the former into the latter (Carelli et al. 2018), reinforcing the transcription of new ORFs that emerge near enhancers. Thus, enhancers may play a dual role in the evolution of *de novo* genes, and consequently in the evolution of gene regulatory networks. By creating a transcriptionally permissive environment, enhancers may facilitate the origin of *de novo* genes; by physically interacting with gene promoters, enhancers may facilitate the integration of new genes—those emerging *de novo* or by other means—into existing regulatory networks.

The first evidence that enhancers can facilitate *de novo* gene birth was recently provided using whole-animal transcriptomic and epigenetic data from the nematode *Pristionchus pacificus* (Werner et al. 2018). Specifically, the transcription start sites of expressed genes that were in open chromatin and private to *P. pacificus* were found to be in closer proximity to histone modifications indicative of enhancers than the transcription start sites of expressed genes that were in open chromatin and shared with other nematode species. Although this evidence is compelling, additional systematic analyses are required to draw firm conclusions and to address remaining open questions. For example, we do not yet know about the generality of this mechanism, specifically whether it applies to other clades of eumetazoa. Furthermore, information on the stability of the transcribed ORFs or their potential for translation is still lacking. We also do not know about the cell-type specificity of the enhancers that facilitate *de novo* gene birth (because the data used to study *P. pacificus* were derived from the whole animal) or how the facilitating role of enhancers in *de novo* gene birth differs from that of other means of pervasive transcription (Neme and Tautz 2016).

Here, we take an integrative approach to address these open questions and to study the potential dual role of enhancers in the evolution of gene regulatory networks. We leverage whole-tissue and single-cell transcriptomic and functional genomics data from mouse that describe

gene expression levels, chromatin accessibility, and chemical modifications to histones, as well as phylostratigraphic estimates of the ages of transcribed ORFs. We find young ORFs are preferentially located near enhancers, whereas older ORFs are not. Some of these young ORFs likely are enhancers, as evidenced by their balanced bidirectional transcription—a hallmark of enhancer activity. Mouse-specific intergenic ORFs that are proximal to enhancers are more highly and stably transcribed than mouse-specific intergenic ORFs that are not proximal to enhancers or promoters, and they are transcribed in more cellular contexts, thus highlighting fundamental differences between the facilitating role of enhancers versus other forms of pervasive transcription in *de novo* gene birth. We find the transcripts of enhancer-proximal ORFs often associate with ribosomes, and we uncover several instances of mouse-specific intergenic ORFs that are proximal to promoters that are likely repurposed enhancers. Overall, our results indicate how the molecular properties of enhancers can facilitate the transmutation of regulatory sequences into effective protein-coding ones, as a case of what Siepel (2009) called "Darwinian alchemy".

3.2 Results

3.2.1 Mouse-Specific Intergenic ORFs Are Often Proximal to Enhancers

We considered a set of 56,262 ORFs from transcripts expressed in the liver, brain, and testis of mouse. Previous work assigned phylogenetic ages to these ORFs (Schmitz, Ullrich, and Bornberg-Bauer 2018), based on the presence of homologous sequences in the transcriptomes of other mammalian species, including rat, human, and opossum (Fig. 3.1A). We further classified the mouse-specific ORFs as genic or intergenic, based on whether or not they are proximal to older, annotated genes (Methods). We use the term proximal to mean within 500bp (in the the annex Figs. 3.8-11, we show our findings are qualitatively insensitive to changing this definition to 250 and 1,000bp), and we use an ORF's first exon to calculate its distance from other genomic features. To characterize the regulatory background of an ORF, we considered data describing histone modifications that are indicative of promoters and enhancers (Heintzman et

al. 2007). Specifically, we merged chromatin immunoprecipitation followed by DNA sequencing (ChIP-seq) data for H3K27ac, H3K4me1, and H3K4me3 obtained from 23 mouse tissues and cell types (ENCODE 2012). We considered enhancers to be those genomic regions where H3K27ac and/or H3K4me1 peaks do not overlap H3K4me3 peaks in any tissue, and promoters to be those genomic regions with H3K4me3 peaks (Creyghton et al. 2010; Berthelot et al. 2018) (Methods).

The majority of ORFs in each age class are proximal to a promoter or an enhancer (Fig. 3.1B). Remarkably, mouse-specific intergenic ORFs are the only class of ORFs that are more likely to be proximal to enhancers than to promoters. Although the first exons of nearly 45% (7,128) of mouse-specific intergenic ORFs are proximal to enhancers, fewer than 25% of rat, human, and opossum-shared ORFs are proximal to enhancers. Similar trends are observed when we restrict our attention to ORFs that are within, or proximal to, genomic regions of open chromatin in at least one of 13 mouse tissues (Fig. 3.1C; Methods). Specifically, $\sim$47% (3,513 out of 7,484) of mouse-specific intergenic ORFs are proximal to regions of open chromatin that harbor histone modifications indicative of enhancers, but not promoters, whereas fewer than 21% of rat, human, and opossum-shared ORFs are proximal to such regions. Similar trends are also observed when we consider histone modification data from individual tissues, as opposed to merging data across cell and tissue types. Specifically, 25% (281 ORFs), $\sim$36% (897 ORFs), and $\sim$20% (537 ORFs) of intergenic mouse-specific ORFs that are in open chromatin and expressed in liver, brain, and testis, respectively, are proximal to an enhancer in that tissue, as compared with fewer than 10% of genic and older ORFs, which are instead preferentially proximal to promoters (Fig. 3.1E). Finally, mouse-specific intergenic ORFs are more likely to show evidence of balanced bidirectional transcription—a hallmark of enhancer activity (Andersson et al. 2014)—than any other class of ORFs (Fig. 3.1D), with 12% of the ORFs overlapping a bidirectional capped analysis of gene expression (CAGE) peak and nearly 20% (1,429) of the ORFs proximal to a bidirectional CAGE peak (Fig. 3.7, shows that these trends are not driven by exon length). Taken together, these results support a model in which enhancers facilitate the expression of young ORFs (Wu and Sharp 2013; Werner et al. 2018).

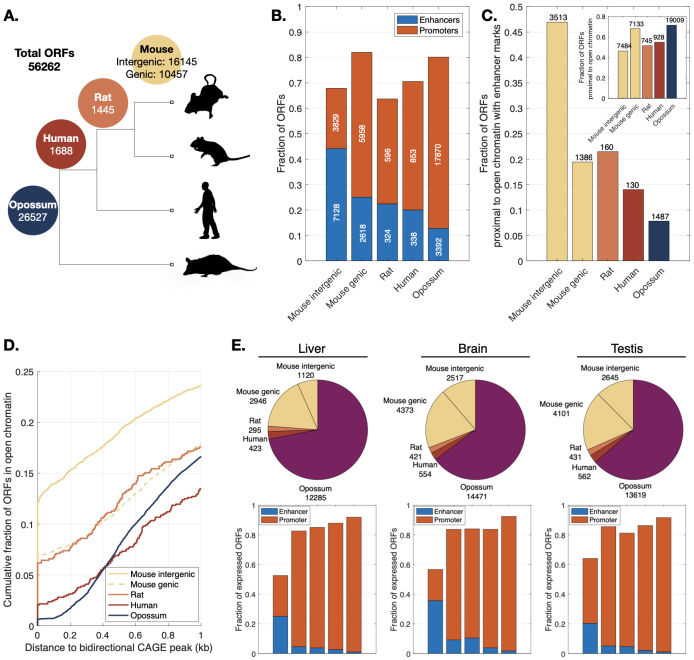


Figure 3.1: Enhancer interactions increase with gene age. (A) Phylogeny showing the four age classes of the 56,262 ORFs. The numbers on the branches indicate the number of ORFs that are either mouse-specific or shared with rat, human, and opossum. Mouse-specific ORFs are further classified as intergenic or genic. (B) Fraction of ORFs that are proximal to ChIP-seq peaks indicative of enhancers (H3K27ac and/or H3K4me1 without overlapping H3K4me3) or promoters (H3K4me3), shown in relation to ORF class. (C) Fraction of ORFs that are proximal to regions of open chromatin that contain enhancers, but not promoters, shown in relation to ORF class. The inset shows the fraction of ORFs that are proximal to regions of open chromatin, regardless of whether those regions contain promoters or enhancers. (D) Cumulative fraction of ORFs that are proximal to regions of open chromatin, shown in relation to their distance to the closest bidirectional CAGE peak. (E) Number of ORFs of each class that are in regions of open chromatin and that are expressed (FPKM>0) in liver, brain, and testis (upper row). Fraction of ORFs from each class (as presented in B) that are proximal to promoters or enhancers in liver, brain, and testis (lower row).

3.2.2 Levels and stability of expression of enhancer-associated intergenic ORFs

We next asked what differentiates the facilitating role of enhancers in *de novo* gene birth from other forms of pervasive transcription taking place away from promoters and enhancers. We

hypothesized that because enhancers are regularly engaged with the transcriptional machinery, they may confer higher levels of expression and greater expression stability. To test this hypothesis, we compared the expression levels and stabilities of intergenic mouse-specific ORFs that are proximal to enhancers with those of intergenic mouse-specific ORFs that are not proximal to enhancers or promoters, using transcriptomic, histone modification, and chromatin accessibility data from liver, brain, and testis (Methods).

In all three tissues, we observed that mouse-specific intergenic ORFs that are proximal to enhancers have a higher median expression level than mouse-specific intergenic ORFs that are not proximal to enhancers or promoters (Fig. 3.2A; Wilcoxon signed-rank test, $P < 0.001$ in liver, $P = 0.003$ in brain, and $P = 0.02$ in testis). To measure expression stability, we calculated the entropy of expression across biological replicates (Methods). When this measure equals its minimum of 0, the ORF is expressed in only one of the replicates; when it equals its maximum of 1, the ORF is expressed at equal levels across replicates. In all three tissues, we observed that expression stability is higher for mouse-specific intergenic ORFs that are proximal to enhancers than for mouse-specific intergenic ORFs that are not proximal to enhancers or promoters (Fig. 3.2B; Wilcoxon's signed-rank test, $P < 0.001$). These observations support the hypothesis that enhancers confer higher expression levels and greater expression stability to proximal ORFs than do other forms of pervasive transcription away from promoters and enhancers.

In liver and testis, we observed that mouse-specific intergenic ORFs that are proximal to enhancers have lower median expression levels and stabilities than mouse-specific intergenic ORFs that are proximal to promoters (Fig. 3.2A and B; Wilcoxon signed-rank test, $P = 0.03$ and $P < 0.001$ for expression level, and $P = 0.02$ and $P < 0.001$ for expression stability, in liver and testis, respectively). This observation is consistent with previous analyses of transcription emerging from enhancers and promoters, which showed that enhancers drive lower and less stable expression than promoters, despite the architectural similarities of these regulatory elements (Core et al. 2014). The increased expression levels and stabilities of promoter-associated transcription may derive from the sequence features of the corresponding tran-

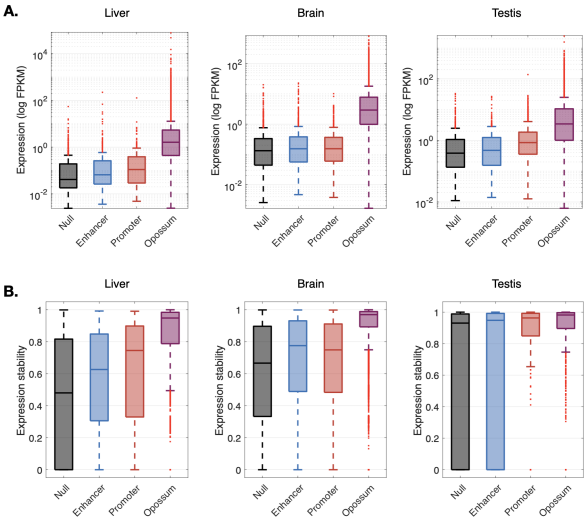


Figure 3.2: Mouse-specific intergenic ORFs that are proximal to enhancers are more highly expressed and have greater expression stability than mouse-specific intergenic ORFs that are not proximal to enhancers or promoters. (A) Expression level of mouse-specific intergenic ORFs proximal to enhancers, promoters, or neither ("Null") in liver, brain, and testis. (B) Expression stability of mouse-specific intergenic ORFs proximal to enhancers, promoters, or neither ("Null") in liver (eight replicates), brain (eight replicates), and testis (two replicates). The expression levels and stabilities of opossum-shared ORFs are shown as a point of comparison.

scripts, including the presence or absence of early polyadenylation sites and splicing signals, which are conducive to transcriptional elongation and may contribute to a positive feedback loop wherein elongation promotes subsequent rounds of initiation (Core et al. 2014).

3.2.3 Ribosomal association of ORFs transcribed from enhancers

Many noncoding transcripts associate with ribosomes (Wilson and Masel 2011; Ingolia et al. 2014; Ruiz-Orera et al. 2014; Zhang et al. 2015). It has been suggested that this may enrich the pool of transcribed ORFs for benign peptides, thus increasing the likelihood of *de novo* gene birth (Wilson and Masel 2011). We hypothesized that because of their increased expression levels and

stabilities, mouse-specific intergenic ORFs that are proximal to enhancers will be more likely to associate with ribosomes than mouse-specific intergenic ORFs that are not proximal to enhancers or promoters. To test this hypothesis, we considered liver, brain, and testis data from a ribosomal profiling assay called ribo-seq, which describes the transcriptome-wide binding patterns of ribosomes to RNA molecules (Ruiz-Orera et al. 2018; Ruiz-Orera and Albà 2019) (Methods).

Following Schmitz et al. (2018), we first consider a permissive definition of ribosomal association: at least one read mapping to the first exon of an ORF. We found that mouse-specific intergenic ORFs that are proximal to enhancers are ∼10% more likely to associate with ribosomes than mouse-specific intergenic ORFs that are not proximal to enhancers or promoters, and ∼10% less likely to associate with ribosomes than mouse-specific intergenic ORFs that are proximal to promoters (Fig. 3.3A). When we apply more conservative thresholds for ribosomal association, mouse-specific intergenic ORFs that are proximal to enhancers remain more likely to associate with ribosomes than mouse-specific intergenic ORFs that are not proximal to enhancers or promoters, and less likely than mouse-specific intergenic ORFs that are proximal to promoters, although the differences in ribosomal association between these classes decreases as the threshold for ribosomal association increases, both when evaluating reads per kilobase mapped to the first exon (Fig. 3.3A), or simply total number of reads mapped to the first exon (Fig. 3.3B). These trends remain when considering tissue-specific transcriptomic, histone modification, and ribosomal association data for liver, brain, and testis (Fig. 3.3C).

3.2.4 Intergenic ORFs That Are Proximal to Enhancers Are Expressed in More Cellular Contexts than Intergenic ORFs That Are Not Proximal to Enhancers or Promoters

In the model of enhancer-facilitated *de novo* gene birth studied here, ORFs emerging near enhancers are likely to have their expression restricted to cells where those enhancers are active. Enhancers are often specific to a small number of cell types (He et al. 2014), which may reduce the potential for enhancer-proximal ORFs to have deleterious pleiotropic effects, while

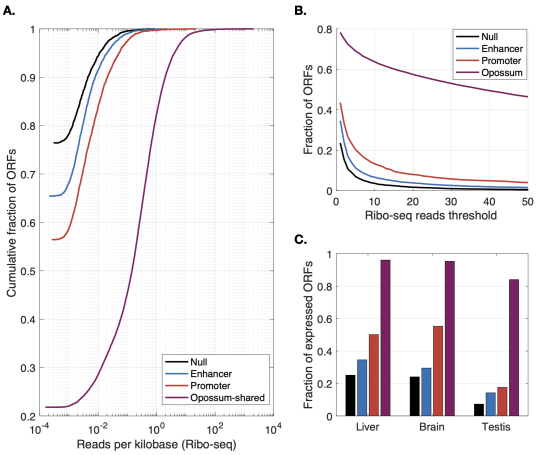


Figure 3.3: Mouse-specific intergenic ORFs that are proximal to enhancers are more likely to associate with ribosomes than mouse-specific intergenic ORFs that are not proximal to enhancers or promoters. (A) Cumulative fractions of mouse-specific intergenic ORFs that are proximal to an enhancer, a promoter, or neither ("Null"), or to opossum-shared ORFs, shown in relation to the number of ribo-seq reads mapped per kilobase to the first exon of each ORF. (B) Fraction of ORFs with ribosomal association, shown in relation to the minimum threshold for the number of reads mapped. (C) Fraction of ORFs expressed in liver, brain, and testis for which at least one tissue-specific ribo-seq read could be mapped to their first exon. The color scheme is the same as in (A).

simultaneously exposing the ORFs to a range of cellular contexts in which they may confer a selective advantage. To study the breadth of expression of ORFs, we considered two sources of data: whole tissue measurements of total RNA across 10 tissues and single-cell measurements of open chromatin across 38 cell types (Methods).

We found that mouse-specific intergenic ORFs that are proximal to enhancers are expressed in more tissues (Wilcoxon's signed-rank test, $P < 0.001$; Fig. 3.4A) and are in open chromatin in more cell types (Wilcoxon's signed-rank test, $P < 0.001$; Fig. 3.4B) than mouse-specific intergenic ORFs that are not proximal to enhancers or promoters. However, these ORFs are expressed in fewer tissues (Wilcoxon's signed-rank test, $P < 0.001$; Fig. 3.4A) and are in open chromatin in fewer cell types (Wilcoxon's signed-rank test, $P < 0.001$; Fig. 3.4B) than mouse-specific intergenic ORFs that are proximal to promoters. This result is expected, because en-

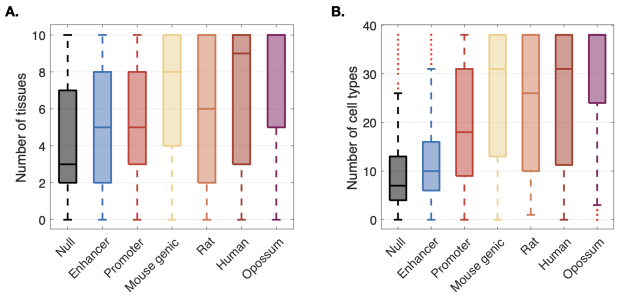


Figure 3.4: Mouse-specific intergenic ORFs that are proximal to enhancers are expressed in a limited diversity of cellular contexts. (A) Number of tissues in which ORFs have an average FPKM > 0 across replicates. (B) Number of cell types in which ORFs are in regions of open chromatin. In both panels, the "Null," "Enhancer," and "Promoter" categories correspond to mouse-specific intergenic ORFs.

hancers tend to be active in fewer tissues than promoters (Colbran, Chen, and Capra 2019). ORFs emerging near enhancers are therefore transcribed in more cellular contexts than ORFs emerging away from promoters and enhancers, but in fewer cellular contexts than ORFs associated with promoters. This may help balance the reward of sampling a diversity of cellular environments with the risk of the pleiotropic effects of broad expression.

3.2.5 Some Intergenic ORFs Are Proximal to Promoters That Show Evidence of Being Repurposed Enhancers

Similarities in the architectures of enhancers and promoters can facilitate the regulatory repurposing of enhancers into promoters (Wu and Sharp 2013; Carelli et al. 2018), which could reinforce the transcription of ORFs emerging near enhancers. We next assessed whether the mouse-specific intergenic ORFs that are proximal to promoters are cases of ORFs transcribed from enhancers that were repurposed into promoters. To do so, we considered 422 mouse-specific intergenic ORFs that are expressed and proximal to an active promoter in mouse liver (Methods). Subsequently, we assessed the chromatin modification status in the rat liver of those mapped genomic regions, using ChIP-seq data for H3K27ac and H3K4me3, marking enhancers and promoters, respectively.

Of the regions mapped to the rat genome, 335 are proximal to H3K27ac peaks and 245 are proximal to H3K4me3 peaks identified from rat liver samples. The majority (∼72%) of the regions that are proximal to H3K27ac peaks are also proximal to H3K4me3 peaks (Fig. 3.5A and B), implying they act as promoters in the liver of both mouse and rat. However, some mapped genomic regions are at such distances from H3K4me3 peaks that they could well be enhancers in rat and may therefore have been repurposed into promoters on the lineage to mouse (Fig. 3.5B). Considering those mapped genomic regions with H3K27ac peaks that are separated from an H3K4me3 peak by a conservative threshold of at least 5kb, we found 42 candidates for the repurposing of rat enhancers to mouse promoters (10% of the 422 ORFs; Fig. 3.5B). The ORFs corresponding to these mapped genomic regions show evidence of stable transcription, both in terms of expression stability across biological replicates (Fig. 3.5C) and in terms of their proximity to CAGE peaks (Fig. 3.5D), which provides evidence that the transcripts are 5'-capped. Of note, many of these CAGE peaks are bidirectional, despite their proximity to H3K4me3 (promoter) peaks, which further supports the hypothesis of an enhancer origin (Fig. 3.5D). Finally, ∼81% of the 42 ORFs show evidence of association with ribosomes in mouse (using our most permissive criterion), which is more than we would expect by randomly sampling mouse-specific intergenic ORFs that are proximal to promoters and expressed in liver (Fig. 3.3C; binomial test $P < 0.001$). Together, these observations give further support to a model in which enhancers provide fertile ground for *de novo* gene birth.

An alternative interpretation of these data is that promoters were repurposed as enhancers on the rat lineage, rather than enhancers being repurposed as promoters on the mouse lineage. To study the directionality of the repurposing, we considered ChIP-seq data for H3K27ac and H3K4me3 from the liver of rabbit, which served as an outgroup (Villar et al. 2015). Of the 42 candidate genomic regions, 11 could be mapped to the rabbit genome using liftOver and were proximal to an H3K27ac peak in rabbit liver. Of these, ten were proximal to an H3K27ac peak that was separated from an H3K4me3 peak by at least 5kb (see e.g., Fig. 3.5E). This provides further support for the hypothesis of an ancestral enhancer state, at least for these ten ORFs. Of note, all of the ORFs corresponding to these mapped genomic regions are surrounded by

other enhancer marks in the mouse liver (Fig. 3.6), which hints that enhancer redundancy may help prevent conflicts that arise in the repurposing of enhancers into promoters, a possibility we revisit in the discussion.

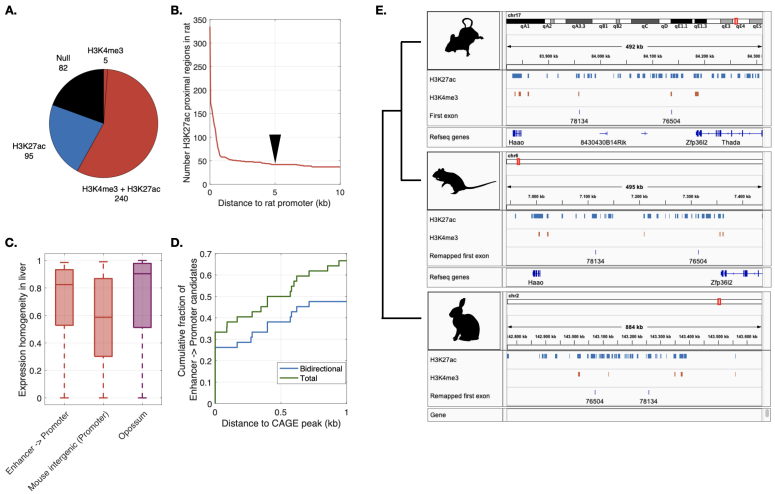


Figure 3.5: Some mouse-specific intergenic ORFs that are proximal to promoters show evidence of being repurposed enhancers. (A) Distribution of histone modification marks among genomic regions in rat. These regions are orthologous to genomic regions in mouse that harbor ORFs that are expressed in liver and are proximal to promoters. (B) Number of genomic regions in rat liver with H3K27ac peaks, shown in relation to their distance to the closest H3K4me3 peak. We use a conservative threshold of 5kb (black arrow) between a promoter and an enhancer mark to determine that an enhancer is not a promoter. This results in 42 candidate promoters that were potentially repurposed from enhancers. (C) Expression homogeneity in liver of these 42 ORFs ("Enhancer → Promoter"), mouse-specific intergenic ORFs that are expressed and proximal to promoters in liver, and opossum-shared ORFs. (D) Cumulative fraction of the 42 ORFs shown in relation to their distance to the nearest CAGE peak ("Total") or the nearest bidirectional CAGE peak ("Bidirectional"). (E) Example repurposed enhancers. Orthologous genomic regions in mouse, rat, and rabbit that in mouse include the first exon of an intergenic mouse-specific ORF. The blue tracks represent H3K27ac peaks (enhancers) and the red tracks represent H3K4me3 peaks (promoters), both measured in liver samples from each organism. Annotated refseq genes are also indicated.

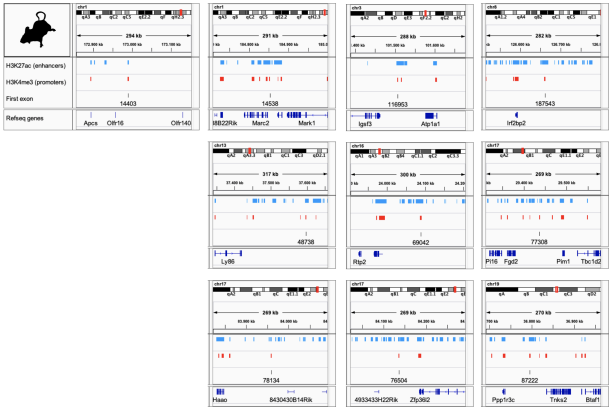


Figure 3.6: Enhancer redundancy may help prevent conflicts in the repurposing of enhancers to promoters. Visualization of 10 murine genomic regions that contain 10 mouse-specific intergenic ORFs that are proximal to promoters and show evidence of being repurposed enhancers, because they show indications of being enhancers in rat and rabbit, but indications of being promoters in mouse (Fig. 3.5).

3.3 Discussion

Our study provides empirical support for the hypothesis that enhancers facilitate *de novo* gene evolution, which to our knowledge was first proposed upon the discovery of enhancer RNA (Kim et al. 2010) and later expanded upon in a perspective piece by Wu and Sharp (2013). Our findings complement recent work on the regulatory architecture of the nematode *P. pacificus*, which showed that young genes — those private to *P. pacificus* — are in closer proximity to enhancers than genes with one-to-one orthologs in other nematode species (Werner et al. 2018). The observation that many young ORFs are proximal to enhancers in both nematodes and mammals suggests that this mode of gene evolution dates back to at least the common ancestor of Bilateria, and possibly even earlier, since cnidarians, ctenophores, and sponges also employ distal regulatory elements (Schwaiger et al. 2014; Gaiti et al. 2017; Sebé-Pedrós et al. 2018b; Sebé-Pedrós et al. 2018a). As the complexity of gene regulation increased during the evolution of some lineages, such as the lineage to vertebrates (Marlétaz et al. 2018), we speculate that

enhancer-facilitated *de novo* gene birth may have played an increasingly prominent role in the expansion of gene regulatory networks.

The facilitating role of enhancers in *de novo* gene birth is conceptually similar to the facilitating role of the permissive chromatin state of meiotic spermatocytes and postmeiotic round spermatids that underlies the "out-of-testis hypothesis", which proposes the testis as a primary tissue for the origination of new genes (Kaessmann 2010; Witt et al. 2019). Both scenarios envision regions of open chromatin that are exposed to the transcriptional machinery, and thus produce a transcriptionally active environment that is conducive to the evolution of new genes. The two scenarios differ, however, in at least three ways. First, genes that emerge from or near enhancers may rapidly acquire their own promoters, due to the similar architectural and functional features of enhancers and promoters, a similarity that facilitates the repurposing of the former to the latter (Wu and Sharp 2013). Indeed, we report several mouse-specific intergenic ORFs that are proximal to promoters that show evidence of being repurposed enhancers, complementing recent analyses of enhancer repurposing in primates and rodents (Carelli et al. 2018). Second, enhancers are often deployed in multiple cell types or developmental stages (Kvon et al. 2014), exposing enhancer-proximal young ORFs to selection in a limited diversity of cellular contexts. This may help to purge toxic peptides (Wilson and Masel 2011) and balance the benefit of expression in distinct cellular environments with the cost of pleiotropic effects. Third, because enhancers are often active in somatic cell types, *de novo* genes emerging near enhancers are more likely to be involved in physiological or morphological traits than *de novo* genes emerging from testis, which are more likely to be involved in reproductive traits. We emphasize that the enhancer-facilitated and out-of-testis scenarios are not mutually exclusive; in fact, they may be complementary or even interactive. Indeed, we found many young transcribed ORFs that associate with ribosomes in testis that are also proximal to enhancers (Figs. 3.8 and 9).

The three points that differentiate enhancer-facilitated *de novo* gene birth from the out-of-testis scenario also differentiate enhancer-facilitated *de novo* gene birth from pervasive transcription (Clark et al. 2011) taking place away from promoters or enhancers. An additional

difference is the relatively high and stable expression levels of enhancers, which increases the chances of ORF-bearing transcripts that stem from or near enhancers to associate with ribosomes. This is indeed what we observe when comparing ribosomal association among mouse-specific intergenic ORFs that are proximal to enhancers with mouse-specific intergenic ORFs that are not proximal to enhancers or promoters. However, we note that this observation may be a technological artifact. If the likelihood of the ribo-seq assay to detect ribosomal association increases with the level or stability of expression, then we would expect to see increased ribosomal association for mouse-specific intergenic ORFs that are proximal to enhancers, relative to mouse-specific intergenic ORFs that are not proximal to enhancers or promoters, even if these two classes of ORFs tend to associate with ribosomes to the same extent. Thus, the reason why enhancer proximity increases the likelihood of ribosomal association is the same reason why we cannot rule out the possibility that we observe this association due to a technological artifact.

An additional facet to enhancer-facilitated *de novo* gene birth is conflict between the enhancer and the emerging gene. If the enhancer is repurposed as a promoter to enforce directional transcription, then the ancestral function of the enhancer may be compromised. There are at least two ways to resolve this conflict. One is to maintain enhancer function; indeed, many promoters also act as enhancers (Medina-Rivera et al. 2018). Another is enhancer redundancy. Genes are often targeted by multiple enhancers, and in many of these cases, only a subset of the enhancers are necessary to drive correct expression under normal growth conditions (Osterwalder et al. 2018). Thus, we hypothesize that redundant enhancers are less likely to face conflict in facilitating *de novo* gene birth. Although our observation that repurposed enhancers tend to be surrounded by other enhancers provides anecdotal support for this hypothesis (Fig. 3.6), more systematic analyses are warranted.

Within the framework of the study of *de novo* gene origination there are two models that explain the emergence and maturation of these genes. One of the models is the continuum model, which proposes that there are intermediate steps of functionalization between non-genes and genes, the so-called “proto-genes” (Carvunis et al. 2012). The preadaptation model,

on the other hand, claims that candidate *de novo* genes represent an abrupt functional innovation, but that they must have first undergone a screening by which potentially toxic peptides were purged (Wilson et al. 2017). Our model of how enhancers can impact *de novo* gene origination would in principle be compatible with both models. For the continuum model, the expression of proto-genes facilitated by enhancer transcription can expose different stages of functionalization to natural selection, thus allowing for a gradual hill climbing towards a transition into a fully-fledged gene. On the other hand, the same expression could help purge toxic variants and promote the apparent sudden emergence of non-toxic *de novo* genes, while offering a preadapted transcriptional background to those genes.

Our results suggest that the power of enhancers in creating evolutionary novelties lies not only in their well-recognised ability to rewire gene regulatory networks but also in their ability to expand them, by providing fertile ground for *de novo* gene birth.

3.4 Methods

ORF Age and Classification. Schmitz et al. (2018) identified a set of 58,864 ORFs from the transcriptomes of three murine tissues: liver, brain, and testis. Blasting against the transcriptomes of four other mammalian species (rat, human, kangaroo rat, and opossum), they estimated the age of each ORF by phylostratigraphic methods (Domazet-Lošo, Brajković, and Tautz 2007; Schmitz, Ullrich, and Bornberg-Bauer 2018). Because of the small number of ORFs shared with the kangaroo rat (49 ORFs), we merged these ORFs together with those from the rat age class. We used the genomic coordinates of the first exon of each ORF in the mm10 mouse genome reference to study the regulatory properties of ORFs of different ages, for example, to study their distance to the nearest enhancer. We only considered ORFs that were transcribed from nuclear chromosomes and whose first exon was longer than 30 base pairs. If first exons were shared between more than one ORF, we only retained the oldest of the ORFs. Our filtered data set contained 56,262 ORFs.

Schmitz et al. (2018) annotated each ORF as belonging to one of eight different categories: “intergenic,” “close to promoter same strand,” “close to promoter opposite strand,” “overlap-

ping same strand," "overlapping opposite strand," "overlapping coding sequence same strand," "overlapping coding sequence opposite strand," and "overlapping annotated gene in frame." We considered all categories except "intergenic" to be "genic" in order to separate ORFs that were born within or near existing genes from those that were not. This resulted in five classes: mouse-specific intergenic ORFs, mouse-specific genic ORFs, rat-shared ORFs, human-shared ORFs, and opossum-shared ORFs.

Proximity to Enhancers and Promoters. We obtained ChIP-seq data for H3K27ac, H3K4me1, and H3K4me3 modifications from 23 different tissues and cell types from the ENCODE project (bone marrow, cerebellum, cortex, heart, kidney, liver, lung, olfactory bulb, placenta, spleen, small intestine, testis, thymus, embryonic whole brain, embryonic liver, embryonic limb, brown adipose tissue, macrophages, MEL, MEF, mESC, CH12 cell line, and E14 embryonic mouse) (ENCODE, 2012). We used liftOver (Kent et al. 2002) to convert the genomic coordinates of the peaks from mm9 to mm10. We used the "merge" function of bedtools (Quinlan and Hall 2010) with default parameters to collate the peaks for all tissues and cell types, considering any overlapping H3K27ac and H3K4me1 peak as part of the same enhancer. We used the "intersect" function of bedtools with default parameters to separate H3K27ac and H3K4me1 peaks that overlapped any length of H3K4me3 peaks from those that did not. This resulted in 172,930 H3K27ac and 277,187 H3K4me1 peaks that did not overlap H3K4me3 peaks. We considered genomic regions with H3K4me3 peaks to be promoters, and those exclusively with H3K27ac and/or H3K4me1 peaks to be enhancers (Berthelot et al. 2018). We measured the distance in base pairs between the first exon of an ORF to an enhancer or promoter using the "closest" function of bedtools with the "-t first" option activated. We considered an ORF to be proximal to an enhancer if the distance to the first exon was shorter than 500bp and there was no promoter within that distance. When controlling for the length of the first exon, we considered the distance to windows of 750-bp up- and down-stream of the central nucleotide of the first exon, rather than to the first exon itself.

We followed the same procedures when measuring the distance of ORFs to enhancers and promoters in liver, brain, and testis tissues separately. For brain tissue, we merged ChIP-seq

data from embryonic whole brain and cortex. The ORFs we considered as expressed in each tissue (Fig. 3.8A–C) were those with a mean fragments per kilobase per million mapped reads (FPKM)>0 across replicates of total RNA transcriptomic data (eight replicates for liver and brain and two replicates for testis) (Li et al. 2017).

5'-capping. We used CAGE data from the FANTOM5 consortium from 1,016 mouse samples including cell lines, primary cells, and tissues (Lizio et al. 2015; Noguchi et al. 2017). This method is based on the capture of 5'-capped ends of mRNA, which allows the mapping of regions of transcription initiation genome wide (Shiraki et al. 2003). Using the "closest" function from bedtools with the "-t first" option activated (Quinlan and Hall 2010), we measured the distance between an ORF's first exon and its closest CAGE peak. In the same manner, we also considered a subset of CAGE peaks which were annotated as bidirectional and transcribed from enhancers (Andersson et al. 2014; Dalby, Rennie, and Andersson 2018).

Expression Level and Stability. We measured the expression levels and stabilities of ORFs. To do so, we aligned paired reads produced by RNAseq from total RNA from ten tissues (liver, testis, brain, muscle, bone, small intestine, thymus, heart, lung and spleen) (Li et al. 2017) using STAR 2.5.3a (Dobin et al. 2013) to the mm10 build of the mouse genome. We chose these tissues because ChIP-seq data for histone modifications were also available. For each ORF, we calculated FPKM as the number of reads mapped to the first exon divided by a millionth of the number of reads sequenced in each sample and then by the length of the exon in kilobases. We considered an ORF to be expressed if it had an average FPKM>0 across replicates.

We calculated the expression stability of ORF k as

$$H_k = -\sum_{(i=1)}^{n} \frac{FPKM_i}{\sum_{j=1}^{n} FPKM_j} log_n(\frac{FPKM_i}{\sum_{j=1}^{n} FPKM_j}) \tag{3.1}$$

where n is the number of replicates for a given tissue (8 for liver and brain and 2 for testis). We refer to this measure as expression homogeneity when calculated across tissues or cell types, rather than across replicates for the same tissue or cell type.

Ribosome Association. We used ribosome profiling (ribo-seq) data from mouse liver, brain, and testis (Ingolia et al. 2014). We obtained the coordinates of mRNA segments detected by ribo-seq from GWIPS-viz (Michel et al. 2014), a database that includes such data from different studies. From this source, we considered samples from liver (three samples from three studies), brain (five samples from two studies), and testis (one sample). We combined the data sets for each tissue and merged the provided genomic coordinates using the bedtools merge function; we did so with the options "-c" and "-o absmax" activated. Following Ruiz-Orera et al (2018), we removed all merged coordinates shorter than 26bp, because these could be anomalous reads. We subsequently mapped these merged coordinates on the first exon of our set of ORFs using the bedtools "map" function and we summed the number of reads from each of the mapped merged coordinates. In this way, we were able to assign a number of ribo-seq reads to each ORF, which allowed us to estimate ribosomal association and thus potential for translation.

Enhancer Repurposing. We considered the set of 544 mouse-specific intergenic ORFs that were transcribed (average FPKM>0) and proximal to an H3K4me3 peak in mouse liver. We filtered this set to the 456 ORFs that were proximal to what Villar et al. (2015) considered to be replicated H3K4me3 peaks in mouse, in order to facilitate comparison with the histone methylation data from rat and rabbit that were generated for the same study. We used liftOver to map the genomic coordinates of these ORFs to the rat and rabbit genomes (builds r5 and oryCun2), requiring a minimum fraction of remapped bases of 0.6 and 0.4, respectively (Carelli et al. 2018). This resulted in 422 and 152 presumably orthologous genomic regions in rat and rabbit, respectively. Considering H3K27ac and H3K4me3 ChIP-seq peaks in the livers of mouse, rat, and rabbit (Villar et al. 2015), we then calculated the distance between these mapped regions and H3K4me3 and H3K27ac peaks using the bedtools "closest" function. We considered the promoter of an ORF in mouse to show evidence of being a repurposed enhancer if its mapped genomic region in rat or rabbit was proximal to an H3K27ac peak, yet more than 5kb from an H3K4me3 peak, in rat or rabbit liver.

3.5 Supplements

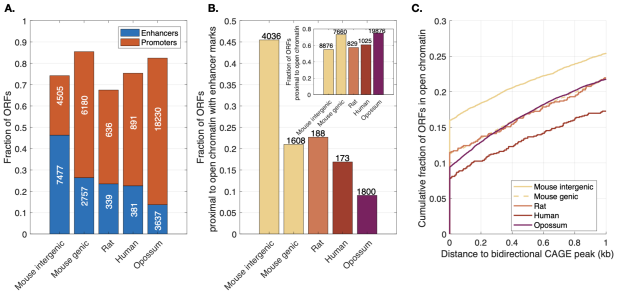


Figure 3.7: Exon length is not responsible for the proximity of young ORFs to enhancers. Repeating the analyses shown in Fig. 3.1B-D using windows of 1.5 kb around the central nucleotide of the first exon of each ORF, rather than the first nucleotide of the first exon, results in the same qualitative trend that mouse-specific intergenic ORFs are more likely to be proximal to enhancers than the other classes of ORFs studied here.

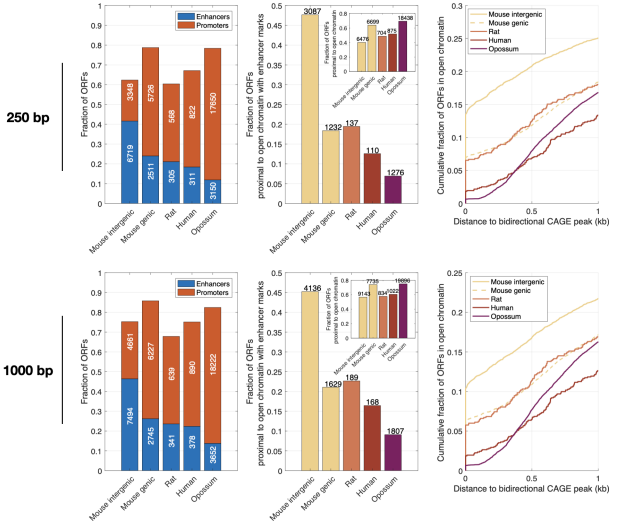


Figure 3.8: Changing our definition of proximal from within 500bp to within (A-C) 250bp or (D-F) 1000bp does not qualitatively alter the trends shown in Fig. 3.7.

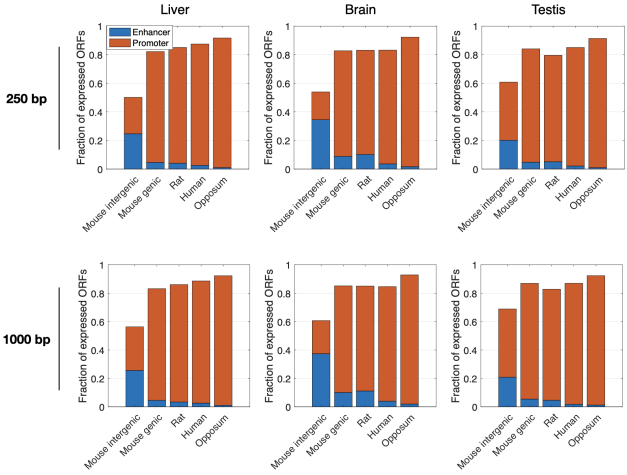


Figure 3.9: Changing our definition of proximal from within 500bp to within (A) 250bp or (B) 1000bp does not qualitatively alter the trends shown in Fig. 3.1E.

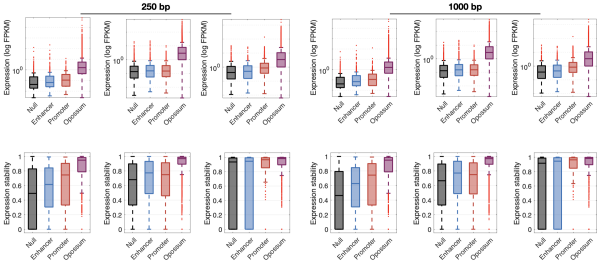


Figure 3.10: Changing our definition of proximal from within 500bp to within (A-C) 250bp or (B-D) 1000bp does not qualitatively alter the trends shown in Fig. 3.2.

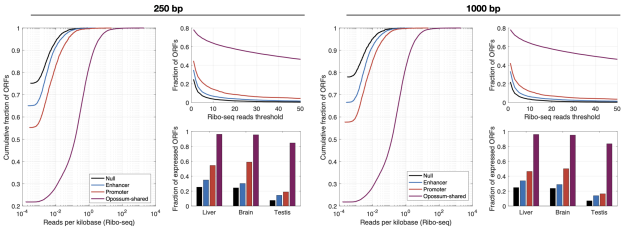


Figure 3.11: Changing our definition of proximal from within 500bp to within (A-C) 250bp or (D-F) 1000bp does not qualitatively alter the trends shown in Fig. 3.3.

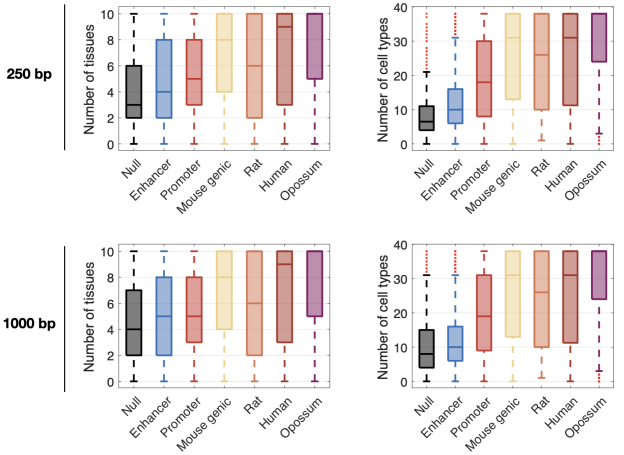


Figure 3.12: Changing our definition of proximal from within 500bp to within (A-C) 250bp or (B-D) 1000bp does not qualitatively alter the trends shown in Fig. 3.4.

4 The adaptive potential of non-heritable somatic mutations

4.1 Introduction

During the development of most animals, an early distinction occurs between the germline – the population of cells that are fated to differentiate into gametes – and the soma – the cells composing the rest of the body. August Weismann noted that any variation arising in the soma during the lifetime of an organism would be temporary and non-heritable, because it would not be present in the reproductive cells (Weismann 1892). The non-heritability of somatic variation weakened Lamarckian arguments concerning the role of acquired variation in adaptation and set the stage for a neo-Darwinian take on evolution (Mayr 1985; Morange 2016). Within this paradigm, the somatic organism came to be viewed as a mere "excrescence" (Bergson 1907) or a "dead-end replicator" (Dawkins 1982), and the non-heritable genetic variation arising in it as an evolutionary cul-de-sac (Buss 1983a; Dawkins 1982; Jablonka and Lamb 2005; Otto and Hastings 1998). Consequently, studies of somatic mutation in animal evolution mainly focused on their deleterious consequences at the level of the organism in which they arise, such as in cancer and senescence (Cairns 1975; Erten and Kokko 2020; Kennedy, Loeb, and Herr 2012; Kirkwood 1977; Kirkwood and Rose 1991; Medawar 1957), and on the evolutionary dynamics of tumors (Greaves and Maley 2012).

Somatic mutations are ubiquitous, thus effectively making multicellular organisms a genetic mosaic (Reusch, Baums, and Werner 2021), and, more often than not, these mutations are harmless (De 2011; Martincorena and Campbell 2015; Wijewardhane, Dressler, and Ciccarelli

2021; Yizhak et al. 2019). Somatic mutations are detected at different frequencies within the soma, partly depending on whether they arise early or late during development (Behjati et al. 2014; Ju et al. 2017; Lee-Six et al. 2018; Osorio et al. 2018), and partly because of the selective competitiveness of mutant cells (Lawson et al. 2020; Martincorena et al. 2015; Martincorena et al. 2018), meaning that somatic mutations can increase in frequency within the body when they confer a higher proliferative potential or lower mortality to the cells carrying them (Hanahan and Weinberg 2011).

Although the clonal expansion resulting from this somatic selective process is one of the characteristics of the evolutionary dynamics of cancers (Greaves and Maley 2012), positive selection of cells with somatic mutations can also occur without causing any apparent disease phenotypes in the tissues containing the mutant cells (Colom et al. 2020; Lawson et al. 2020; Martincorena et al. 2015; Martincorena et al. 2018; Martincorena and Campbell 2015; Yizhak et al. 2019). In some cases, somatic mutations are beneficial not only for individual cells, but also for the entire multicellular organism. A classic example is the adaptive immune system of jawed vertebrates, in which somatic mutants are selected within the body based on their affinity to the pathogens they help deter (Zhu et al. 2019). They can also ameliorate the consequences of deleterious mutations that cause hematopoietic diseases (Revy, Kannengiesser, and Fischer 2019) or serious developmental syndromes (Bar et al. 2017). In such cases, somatic mutations may facilitate the persistence of deleterious germline mutations by buffering their negative effects on organismal fitness until a compensatory or reversion mutation arises in the germline (De 2011).

Research on the evolutionary consequences of somatic variation that is beneficial above the cellular level has mainly focused on plants (Antolin and Strobeck 1985; Cruzan, Streisfeld, and Schwoch 2020; Gill et al. 1995; Schoen and Schultz 2019; Whitham and Slobodchikoff 1981), as well as other modular organisms such as corals (Van Oppen et al. 2011) and red algae (Monro and Poore 2009). However, somatic variation in these taxa may be partly heritable, either because there is a blurry germline-soma distinction or because they can reproduce clonally (Buss 1983b; Cruzan, Streisfeld, and Schwoch 2020; Leria et al. 2019; Reusch, Baums, and Werner

2021; Yu et al. 2020). Here we argue that there is also evolutionary potential in somatic variation that is strictly non-heritable. Indeed, if non-heritable somatic mutations can confer a fitness advantage to the organism carrying them, selection can act on the potential to acquire such mutations, as evidenced by the evolution of mechanisms that direct or intensify the production of somatic genetic variation, such as the molecular processes driving the recombination and mutagenesis of genes involved in the adaptive immune system (Müller et al. 2018; Odegard and Schatz 2006) and the somatic activation of transposable elements (McClintock 1950; Singer et al. 2010). These mechanisms tend to target specific genomic regions and function in specialized cell types, and what ultimately gets selected is the mechanisms producing the somatic mosaicism rather than the somatic mutant genotypes themselves (Caporale 2000; Jablonka and Lamb 2005; Müller et al. 2018; Singer et al. 2010; Whitham and Slobodchikoff 1981).

Here, we envision a complementary, general model in which adaptation is facilitated by selection acting on genotypes with a potential to acquire non-heritable somatic mutations that are beneficial to the organism, even in the absence of a mechanism to intensify somatic diversity. Given the sheer number of cells in the soma and their increased mutation rates relative to the germline (Lynch 2010; Milholland et al. 2017; Moore et al. 2021; Murphey et al. 2013), we reason that beneficial mutations often first arise in the soma. Similar to so-called phenotypic mutations (Whitehead et al. 2008), which arise due to errors in transcription or translation, non-heritable genotypes that are similar in sequence to a heritable beneficial genotype may occasionally confer the fitness benefit of the heritable beneficial genotype to an organism via somatic mutation. Placing this model in the context of an adaptive landscape (Wright 1932), the germline genotype can be one or more mutations away from an adaptive peak, and somatic mutations can confer a fitness benefit to an organism by producing non-heritable beneficial genotypes that are closer to or atop the adaptive peak. This can cause a smoothing of the fitness landscape (Frank 2011; Van Egeren, Madsen, and Michor 2018), which may promote adaptation towards an adaptive peak by increasing the probability that the beneficial mutation arises as a germline variant. This is because precursor genotypes that are near the adaptive peak are

more likely to be selected, as they exhibit a positive epistatic interaction with somatic mutations, thus increasing their frequency in the population relative to genotypes that are farther away from the peak. We thus hypothesize that non-heritable somatic mutations causing a fitness advantage may channel evolving populations towards adaptive peaks, thus promoting adaptation.

4.2 Results

4.2.1 Model overview

To study the potential of non-heritable somatic mutations to promote adaptation, we modelled an evolving population of N multicellular organisms with an impermeable germline-soma separation navigating a minimal fitness landscape. We used a haploid two-locus, two-allele model with alleles a and A for the first locus and b and B for the second locus, to represent a landscape with a single adaptive peak at genotype AB, which confers a selective advantage s_{organism} to the organism relative to the other genotypes in the landscape (Fig. 4.1A, Methods). We simulated evolution with non-overlapping generations, starting from an initial population composed exclusively of organisms with the ab genotype. Each generation consisted of a developmental phase followed by a reproductive phase (Fig. 4.1B, Methods). In the developmental phase, the soma of each individual developed from a single cell with a given zygotic genotype in D developmental cycles until reaching the final somatic size of 2^D cells. For each cell division, somatic mutations occurred at rate μ_{soma} per locus per daughter cell. Organismal fitness was defined by the proportion of somatic cells carrying the AB genotype at the end of development. As such, we allowed somatic mutations only during development and did not model the mature lifespan of organisms. In the reproductive phase, organismal reproductive success was proportional to fitness, and germline mutations occurred at rate μ_{germline} per locus.

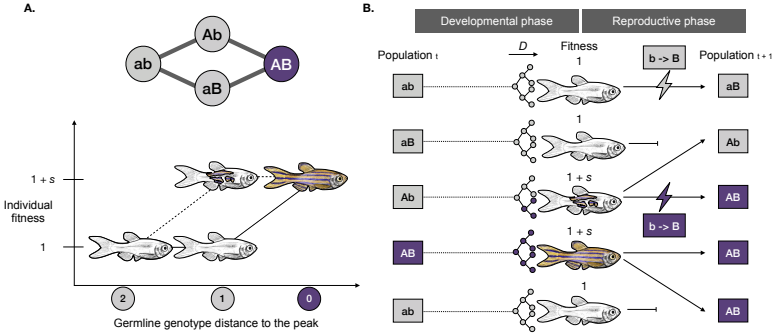


Figure 4.1: Baseline model. A) Fitness landscape represented as a two-locus two-allele model. Each line connecting genotypes in silver and purple nodes corresponds to a single mutational step. Genotype AB (purple, peak genotype) confers a higher fitness to its carrier, whereas all other genotypes confer no selective advantage in the absence of somatic mutations. The Cartesian axes represent the individual fitness value as a function of the distance of the heritable germline genotype to the peak. Solid lines represent evolutionary trajectories towards the peak from ab, in which somatic mutations are either not present or produce no selective advantage, such as in our control simulations. Dashed lines show how beneficial somatic mutations can smoothen the fitness landscape by increasing the organismal fitness of individuals with somatic mutations to the peak genotype. B) Representation of a single generation in our simulations. At generation t, individuals in a population of size N enter a developmental phase. During this phase, starting from a single cell with each individual's germline genotype, D developmental cycles occur until the final somatic size $2D$ is reached. At each developmental cycle, somatic mutations occurring at rate μ_{soma} can modify the distance to the peak of each somatic cell. Based on a fitness function, at the end of the developmental phase, the final genotypic composition of the soma defines the fitness of each individual. During the reproductive phase, the population is sampled based on the individual fitness values to create a new population for the next generation. Before entering the developmental phase of generation $t+1$, germline mutations may occur at rate μ_{germline}, represented by purple and silver lightning bolts in our diagram.

4.2.2 Non-heritable somatic mutations can promote adaptation

We ran two versions of our model across a range of somatic and germline mutation rates. In the first version, somatic mutations did not confer an organismal selective advantage. This served as a control and as a point of comparison with traditional population genetic models that disregard somatic development. We implemented this version of our model by simulating each generation using only the reproductive phase, thus ignoring somatic mutations that could arise during the developmental phase. Fitness was therefore defined by the germline genotype alone. In the second version, somatic mutations conferred an organismal selective

advantage. Such an advantage could arise via somatic mutations influencing cell signaling or spatial patterning during development, as we later discuss. We implemented this version of our model using a fitness function in which an individual attained the full selective advantage s_{organism} if at least one somatic cell had the peak genotype AB at the end of the developmental phase – an assumption we later relax. Fitness was therefore defined by the somatic composition of the organism.

Fig. 4.2A-D shows the evolutionary outcomes of these simulations for a range of germline and somatic mutation rates. In the control simulations, the mean fitness of the population increased for germline mutation rates beyond 5×10^{-7} mutations per locus per generation (Fig. 4.2A; Annex I). Under these high germline mutation rates, populations converged on the peak genotype via stochastic tunneling, which occurs when the double mutant AB arises in a neutral or deleterious aB or Ab background before the latter goes to extinction (Ashcroft, Michor, and Galla 2015; Iwasa, Michor, and Nowak 2004) (Fig. 4.2C). When somatic mutations conferred a selective advantage to the organism, the mean fitness of the population increased with both the germline and somatic mutation rates, thus expanding the parameter space in which higher fitness evolved, relative to the control simulations (compare Figs. 5.2A and 5.2B). For low somatic mutation rates, this expansion manifested via a decreased threshold on the germline mutation rate past which populations converged on the peak genotype (Fig. 4.2D). In contrast, for higher somatic mutation rates, fitness increases were not due to convergence on the peak genotype in the germline. Rather, germline genotypes remained one or two mutations away from the adaptive peak, but the peak fitness was nevertheless obtained due to somatic mutations (Fig. 4.2D). Additionally, the presence of somatic mutations increased the rate of convergence to the peak genotype in parameter regions where the peak was reached in both versions of our model (Fig. 4.2E).

We note that empirical mutation rates from mouse and human cells ($\mu_{\text{soma}}\ 5\times10^{-9}$, $\mu_{\text{germline}} \sim 1 \times 10^{-8}$; (Milholland et al. 2017)) fall within the range of mutations rates where somatic mutations can promote adaptation via convergence on the peak germline genotype (Fig. 4.2A-D, circle). For these mutation rates, an initial stage of drift with low frequencies of the Ab and

aB genotypes was often followed by two consecutive selective sweeps towards the AB genotype (Fig. 4.2F). In contrast, in control simulations, genetic drift dominated the evolutionary dynamics, such that populations remained two mutations from the peak (Fig. 4.2G). Taken together, these results provide proof-of-principle that non-heritable somatic mutations can promote adaptation, even under biologically realistic germline and somatic mutation rates.

4.2.3 Somatic mutation supply determines evolutionary outcomes

The results presented above revealed three distinct evolutionary outcomes depending upon the somatic and germline mutation rates (Fig. 4.3A): (i) populations did not increase in fitness, and remained two mutations away from the peak; (ii) populations increased in fitness by converging on the peak; and (iii) populations increased in fitness, but the germline genotype either remained one mutation away from the peak as Ab or aB, or two mutations away from the peak as ab. The mutation rate thresholds for each of the three outcomes can be estimated probabilistically (Annex I). At low somatic mutation rates ($\mu_{\text{soma}} = 1 \times 10^{-10}$, $D = 25$, Fig. 4.6), the probability of acquiring the peak AB genotype via somatic mutation from an intermediate aB or Ab germline genotype is essentially zero, so there is no selective advantage of genotypes Ab or aB over ab. At intermediate somatic mutation rates ($1 \times 10^{-10} < \mu_{\text{soma}} < 1 \times 10^{-7}$, $D = 25$, Fig. 4.6), somatic mutations to the peak AB genotype from intermediate aB and Ab germline genotypes occur with sufficient frequency to give the intermediate germline genotypes an average selective advantage over the germline ab genotype. However, these somatic mutation rates are not high enough to guarantee the somatic evolution of the AB genotype, rendering the AB germline genotype more fit than the intermediate aB and Ab genotypes. In contrast, at high somatic mutation rates, somatic evolution of the AB genotype is essentially guaranteed from the ab germline genotype ($\mu_{\text{soma}} > 1 \times 10^{-4}$, $D = 25$, Fig. 4.6) or from the intermediate aB and Ab germline genotypes ($\mu_{\text{soma}} > 1 \times 10^{-7}$, , $D = 25$, Fig. 4.6), rendering all germline genotypes selectively equivalent. The latter scenario is exemplified by the adaptive immune system of jawed vertebrates. Given the hypervariability induced by VDJ recombination and elevated rates of somatic mutation during affinity maturation, there is reduced

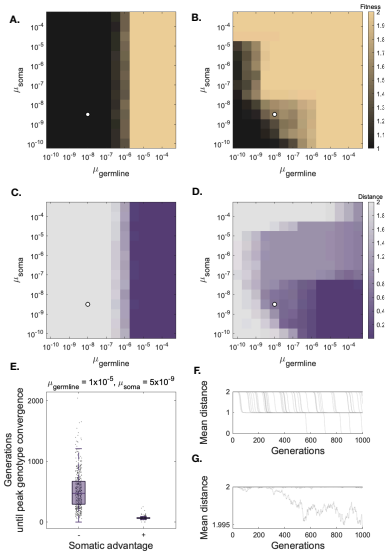


Figure 4.2: Non-heritable somatic mutations can promote adaptation. (A,B) Mean fitness of populations and (C,D) mean mutational distance to the peak after 5000 generations when somatic peak genotypes (A,C) did not provide an organismal selective advantage and (B,D) when they did, for a range of somatic mutation rates (mutations per locus per cell division) and germline mutation rates (mutations per locus per generation). The rows and columns that are not marked by a tick correspond to values of with a mantissa of 5. The values shown are the mean across 500 replicates for each parameter combination. White circles indicate the combination of $\mu_{\text{soma}} = 5 \times 10^{-9}$ and $\mu_{\text{germline}} = 1 \times 10^{-8}$, which approximates empirically estimated mutation rates from mouse and human cells (Milholland et al. 2017). (E) Distribution of the number of generations required to converge on the peak genotype in 500 simulations for $\mu_{\text{soma}} = 5 \times 10^{-9}$ and $\mu_{\text{germline}} = 5 \times 10-5$, when somatic peak genotypes provided a selective advantage (+) and when they did not (-). (F,G) Evolutionary trajectories over the first 1000 generations for 100 randomly chosen replicates when somatic peak genotypes (F) provided a selective advantage and (G) when they did not, under the mutation rates indicated by the asterisks in A-D. Notice that the limits on the y-axes differ in panels (F) and (G). We ran all simulations with a population size N = 100000, a selective advantage s_{organism} = 1 and a number of developmental cycles D = 25.

selective pressure to evolve an exact germline memory of past encounters with pathogens.

Somatic mutation supply can thus determine which evolutionary outcome emerges. Another factor that influences somatic mutation supply beside somatic mutation rate is the number of cells in the soma, which in our model is determined by the number of developmental

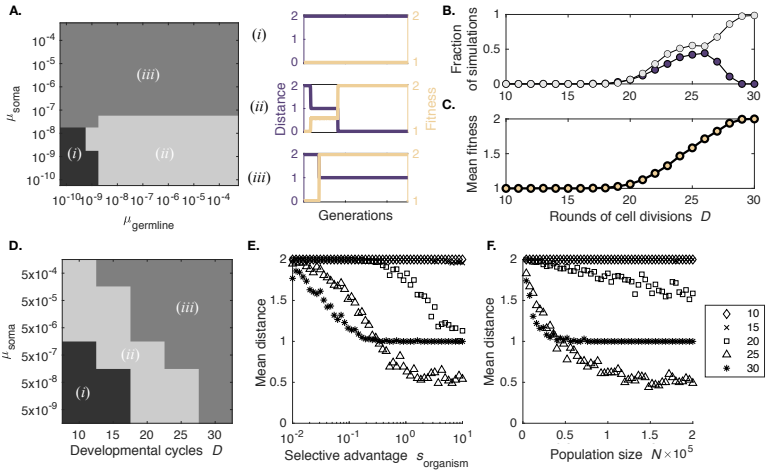


Figure 4.3: Evolutionary outcomes in relation to model parameters. (A) Three distinct evolutionary outcomes emerge in our model: (i) populations do not evolve the maximum fitness of 1 + s_{organism} after 5000 generations, (ii) populations evolve the maximum fitness of 1 + s_{organism} after 5000 generations and converge on the peak genotype, and (iii) populations evolve the maximum fitness of 1 + s_{organism} after 5000 generations, but do not converge on the peak genotype. Regions shaded according to the outcome realized by the simulations for each parameter combination. The rows and columns that are not marked by a tick correspond to values of with a mantissa of 5. Sub-panels to the right show the trajectories for mean distance to the peak (purple) and mean fitness (beige) in representative simulations, for each of the three evolutionary outcomes. (B) The fraction of simulations in which the population reached the peak genotype (purple) or remained one mutation away from the peak genotype (silver) after 5000 generations is shown in relation to D. (C) The mean population fitness after 5000 generations of the same simulations as in (B). (D) Evolutionary outcomes i, ii or iii for different combinations of somatic mutation rates and developmental cycles. (E,F) Mean distance of the germline genotype to the peak after 5000 generations across a range of values for (E) selective advantage in a log-scale and (F) population size. The different symbols correspond to different number of developmental cycles D, as shown in the legend in panel (F). For (D-F), we performed 100 simulations for each combination of parameters. The baseline parameters were N = 100000, s_{organism} = 1, $\mu_{\text{soma}} = 5 \times 10^{-9}$ and $\mu_{\text{germline}} = 1 \times 10^{-8}$.

cycles D. To assess how somatic mutations influence adaptation for organisms of different size, we modified our baseline model to include a range from $D = 10$ to $D = 30$, which produces final somatic cell counts between $2^{10} = 1024$ and $2^{30} = 1.07 \times 10^9$, respectively — values that approximate the number of cells in tissues, organs, and entire animals (Table 5.1). Modifying the mutation supply in this way resulted in similar evolutionary outcomes as when

Table 4.1: Representative biological examples. We study a range of developmental cycles D, which yield somas spanning the size of an adult worm to a newborn rat. These examples provide biological intuition for our model parameter D, though we emphasize that the functional effects of somatic mutations will often be restricted to the tissue in which they arise.

Developmental cycles (D)	**Somatic size (2^D cells)**	**Representative biological examples**
10	1024	Somatic cells of adult *Caenorhabditis elegans* (Kenyon 1988)
15	32768	Adult tardigrade (Seki and Toyoshima 1998), wing disc of fruit fly at metamorphosis (Day and Lawrence 2000)
20	1.05×10^6	Mouse pituitary gland (Gleiberman et al. 2008)
25	3.36×10^7	Adult mouse cerebellum (Surchev et al. 2007)
30	1.07×10^9	Newborn rat (Cairns 1975)

varying somatic mutation rates (Fig. 4.3B,C). Specifically, after 5000 generations, no populations increased in fitness when $D < 19$, populations tended to increase in fitness by evolving the peak genotype when $19 \leq D \leq 28$, and populations increased to a maximum fitness of 1 + s_{organism} without reaching the peak germline genotype when $D > 28$. These outcomes can again be described probabilistically, since, by extending the developmental phase of the simulation, the opportunities for the adaptive mutations to occur somatically also increase, making them likely for intermediately sized somas and guaranteed for larger somas (Annex I, Fig. 4.6). As expected under this logic, by varying the somatic mutation rate together with the number of developmental cycles, we observed that the evolutionary outcome of our model depended on the interaction between these two components of somatic mutation supply (Fig. 4.3D). For example, high somatic mutation rates facilitated convergence on the peak genotype even for populations of organisms with the smallest soma considered ($D = 10$, which approximates the somatic size of an adult *Caenorhabditis elegans*, Table 5.1).

To explore the importance of somatic mutation supply relative to other factors affecting evolutionary dynamics in our model, we additionally varied the population size N and the selection coefficient s_{organism} of the AB genotype (Methods). The number of populations converging on the peak genotype increased as either of these parameters increased, but only for some in-

termediately sized somas (Fig. 4.3E,F). Notably, even though our baseline values for s_{organism} were relatively high, reflecting scenarios in which a single mutation more than doubles fitness components like survival or reproductive maturation (e.g., (Karageorgi et al. 2019; Lanno et al. 2019)), we also observed somatic mutations promoting adaptation under more conservative selection coefficients between 0.01 and 0.1, and slightly higher empirical measurements of selection coefficients, such as those measured for missense mutations enhancing insecticide resistance in mosquitoes (s_{organism} = 0.16, (Lynd et al. 2010)) and improving camouflage in wild mice (s_{organism} = 0.32, (Barrett et al. 2019)) (Fig. 4.3E). Overall, although population size N and selection coefficient s_{organism} can influence the probability with which a population converges on the peak germline genotype, it is ultimately the somatic mutation supply, defined by the somatic mutation rate and the size of the organism that primarily influence which of the three evolutionary outcomes emerge.

Another factor that can influence the adaptive potential of non-heritable somatic mutations is the ruggedness of the fitness landscape. Thus far, we studied a fitness landscape with three genotypes of equal fitness (ab, Ab, and aB) and a fourth genotype (AB) with a selective advantage s_{organism} (Fig. 4.1A). Considering instead a rugged fitness landscape with two adaptive peaks separated by an adaptive valley (Ab and aB; Methods; Annex II, Fig. 4.7A), we found that somatic mutations could promote adaptation by facilitating valley crossing (Fig. 4.7B,C). Which of the three evolutionary outcomes emerged depended on the mutation supply, the depth of the valley, and the selective advantage of the peak genotype AB (Fig. 4.7D). These results suggest that somatic mutations not only promote adaptation on neutral networks, but also aid in the exploration of rugged fitness landscapes, which can otherwise hinder adaptive evolution as populations become trapped on local adaptive peaks (Gokhale et al. 2009).

4.2.4 Alternative fitness functions restrict the adaptive potential of somatic mutations

So far, we have used a fitness function in which a single somatic cell with the peak genotype is sufficient to confer the full selective advantage s_{organism}. Relaxing this assumption to account

for more realistic biological scenarios, we considered alternative functions in which the fitness of a given organism is proportional to the fraction of somatic cells with the peak genotype, σ_{peak}, so that the fraction F_i of s_{organism} attained by an individual i is:

$$F_i(\sigma_{\text{peak}}) = (\sigma_{\text{peak}})^f \tag{4.1}$$

where f is a constant defining the shape of the function. The function is concave when $f > 1$, convex when $0 < f < 1$, and linear when $f = 1$ (Fig. 4.4A). Concave functions require relatively few somatic cells with the peak genotype to confer the full selective advantage s_{organism}, whereas convex functions require a large number of somatic cells with the peak genotype to confer the full selective advantage. Figure 4B shows the probability of converging on the adaptive peak in relation to the selection coefficient s_{organism} for three different somatic mutation rates using seven parameterizations of the fitness functions given in Eq. 4.1. In the range of parameters explored, concave functions tended to promote adaptation across different values of μ_{soma} and s_{organism}, whereas linear or convex functions did not (Fig. 4.4B). Exploring further the evolutionary outcomes under different kinds of concave functions, we also considered a series of diminishing returns fitness functions with different thresholds for the number of cells needed to increase fitness, which was defined by a constant g so that:

$$F_i(\sigma_{\text{peak}}) = \frac{(\sigma_{\text{peak}} + 1)^{1-g} - 1}{2^{1-g} - 1} \tag{4.2}$$

We studied six of these fitness functions with different values for g (Fig. 4.4D; Methods). Under our baseline somatic mutation rate, beneficial somatic mutations only promoted adaptation at relatively high values of s_{organism} for the two fitness functions that required the fewest somatic cells with the peak genotype to confer the full selective advantage (Fig. 4.4E). With these fitness functions, in a soma of 225 cells, 30 and 300 cells with the peak genotype are needed to confer 10% of the selection coefficient s_{organism}, respectively, and at least 3000 and 30000 cells (representing less than 0.01% of the total somatic cells) with the peak genotype are needed to confer the full selection coefficient s_{organism}, respectively. For the same fitness functions,

increasing the somatic mutation rate by one or two orders of magnitude (Fig. 4.4E, second and third panels) increased the probability of converging on the adaptive peak for all values of s_{organism}. Moreover, this increase in the somatic mutation rate expanded the set of fitness functions under which somatic mutations promoted convergence to the peak genotype. Thus, even under high somatic mutation rates and with high selection coefficients, somatic mutations are unlikely to promote adaptation if more than a small fraction of somatic cells with the peak genotype are required to confer the full selective advantage s_{organism}.

4.2.5 The adaptive potential of non-heritable somatic mutations under multi-level selection

We have so far assumed that somatic mutations confer a selective advantage to the organism, but not to the cell. Yet, cell-lineage selection is common in development and it biases mosaic and chimeric cellular compositions (Buss 1982; Extavour and García-Bellido 2001; Morata and Ripoll 1975; Otto and Hastings 1998; Otto and Orive 1995; Schoen and Schultz 2019; Schwarz and Cadavid 2007; Simpson 1979; Yu et al. 2020), helping to explain why some somatic mutations are recurrently detected across different individuals (Melton et al. 2015). Cells can attain increased fitness if they better respond to signals in their developmental environment, make better use of resources, induce apoptosis of neighboring cells, proliferate more, or die less (Kim and Jain 2020; Moreno, Basler, and Morata 2002). We therefore included cell-lineage selection in our model, reasoning that it might expand the set of fitness functions under which non-heritable somatic mutations promote adaptation by increasing the number of cells with the peak genotype in the soma. We assumed somatic cells with peak genotypes had a cellular fitness advantage s_{cell} (Methods), but kept the final size of the soma at 2^D cells, inspired by systems with determinate growth (Hariharan 2015). In other words, we assumed that cells with a fitness advantage increased in frequency without affecting the final size of the organism. After applying these modifications to our model, we ran simulations using the same fitness functions as described above (Fig. 4.4A,D).

Doubling the cellular proliferative advantage of somatic peak genotypes ($s_{\text{cell}} = 2$) expanded

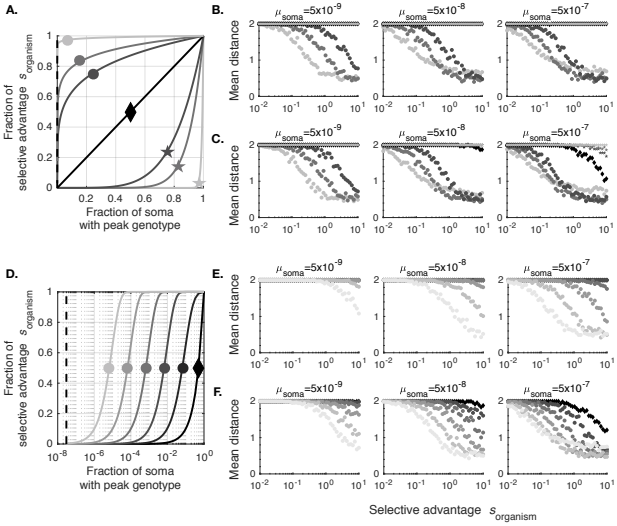


Figure 4.4: Influence of alternative fitness functions and cell-lineage selection on the adaptive potential of somatic mutations. (A,D) Fraction of the selective advantage conferred to an individual as a function of the fraction of cells in the soma with the peak genotype. In (A) there are three concave functions (circles), three convex functions (stars) and a linear function (diamond), generated with Eq. 4.1 for different values of f (Methods). In (D) we show different concave diminishing returns fitness functions of different shape, generated with Eq. 4.2 using different values of g (Methods). The shades of gray from light to dark indicate the distance to the linear fitness function, which is black. Note that the linear function is the same in both (A) and (D) and that (D) is presented in a log-scale on the x-axis. The black dashed line indicates the value of $\frac{1}{2^{25}}$, which is the minimum fraction of somatic cells with peak genotypes that are needed to confer the full selective advantage in the baseline model with D = 25 developmental cycles. (B,E) Mean distance to the peak in populations after 5000 generations across a range of selective advantages under each of the different fitness functions from (A) and (D) in (B) and (E), respectively. Each point represents the mean across 50 replicate simulations for each parameter combination. The shade and symbols of each point refer to the fitness functions presented in (A) and (D). For (C) and (F), we modified the simulations such that cells with peak genotypes had a fitness advantage s_{cell} =2 over cells with other genotypes. We performed the simulations in (B,C,E,F) with the somatic mutation rates indicated above the panels, while the remaining parameters were N = 100000, D = 25 and $\mu_{germline} = 1 \times 10^{-8}$.

the conditions under which non-heritable somatic mutations promoted adaptation, even when the fitness function was linear or convex (Fig. 4.4C, second and third panel). In the case of the concave diminishing returns functions, increases in the proliferation of somatic mutants

with the peak genotype expanded the set of functions in which somatic mutations promoted adaptation, even for the lowest somatic mutation rate considered (Fig. 4.4F, first panel), and increasing the somatic mutation rate by one or two orders of magnitude expanded the set even further, to the point that, across all considered functions some populations converged on the peak germline genotype (Fig. 4.4F, second and third panels). Thus, when a somatic mutation simultaneously benefits the organism and the somatic mutant cell within the context of development, non-heritable somatic mutations can promote adaptation across a broader range of conditions.

4.3 Discussion

Somatic mutations are abundant and sometimes increase organismal fitness (Bar et al. 2017; Revy, Kannengiesser, and Fischer 2019; Zhu et al. 2019). Nonetheless, because of their non-heritability, they are typically neglected as a source of adaptation in traditional evolutionary theory (Buss 1983a, 1983b). Here we provide proof-of-principle that non-heritable somatic mutations can promote adaptation and help traverse fitness valleys. They do so by exposing adaptive genotypes to selection ahead of their emergence in the germline, thus channeling populations toward peaks in adaptive landscapes, a process we refer to as somatic genotypic exploration. Below, we discuss the biological plausibility of somatic genotypic exploration given the simplifications in our model and the restrictions it uncovered, as well as the implications of this process for adaptive evolution.

4.3.1 Biological plausibility of somatic genotypic exploration

Our model makes assumptions that simplify many intricacies of organismal biology. We modelled organisms as haploid individuals with asexual reproduction, whose somas develop through consecutive symmetric and synchronized cell divisions. Complexifying the model could make somatic genotypic exploration more or less likely, depending on the circumstances. For example, if the organism was diploid, the chances of acquiring somatic peak genotypes would be

doubled, because there would be two copies of each allele per cell, but if the peak allele was recessive, somatic mutations would be less effective in revealing adaptive phenotypes. Moreover, the fate of somatic mutants can be affected by tissue architecture and growth dynamics (Frank and Nowak 2004; West et al. 2021). More realistic developmental models accounting for differentiation, asymmetrical divisions and stem cells, will likely affect how somatic genotypic exploration influences adaptation. For example, somatic mutations could have greater influence if they arise in stem cells that contribute substantially to the composition of a tissue, but if somatic mutations are beneficial only if they arise in cells with specialized functions, then acquiring the somatic mutations in specific developmental contexts where they are adaptive would be less likely.

We also made the simplifying assumption that fitness depends on two biallelic loci. One consequence of this assumption is that valley crossing requires the traversal of just a single deleterious intermediate. This assumption underlies most models of valley crossing, including those addressing the roles of mutation rates (Iwasa, Michor, and Nowak 2004; Komarova, Sengupta, and Nowak 2003), population sizes (Weinreich and Chao 2005), recombination (Altland et al. 2011; Christiansen et al. 1998; Michalakis and Slatkin 1996; Weissman, Feldman, and Fisher 2010), population structure (Bitbol and Schwab 2014), cooperation (Obolski et al. 2017), environmental fluctuations (Hadany 2003), and non-genetic phenotypic variation (Van Egeren, Madsen, and Michor 2018). We anticipate that somatic genotypic exploration is less likely to facilitate the crossing of valleys that comprise multiple deleterious intermediates, as has been observed in other modeling frameworks (Komarova 2014; Ram and Hadany 2014; Weissman et al. 2009). The reason is the "foresight" afforded by somatic genotypic exploration is limited to a small mutational radius around the germline genotype, which may not be sufficiently large to explore genotypes on the other side of wide valleys. A direction for future research is therefore to incorporate more than two loci in our model, which would also help us understand how somatic genotypic exploration influences population diffusions on neutral networks (Maynard Smith 1970; Schuster et al. 1994), with implications for the evolution of mutational robustness (Van Nimwegen, Crutchfield, and Huynen 1999) and genetic diversity (Wagner, Pavlicev, and

Cheverud 2007).

Despite these simplifications, our model is suggestive of the biological conditions under which somatic genotypic exploration is expected to influence adaptive evolution in nature. For example, our model suggests that non-heritable somatic mutations can promote adaptation when the somatic mutation supply is high, which can occur by an increased number of cellular divisions in development and by an increased somatic mutation rate. One could object that an increased somatic mutation rate would cause deleterious mutations elsewhere in the genome, thus offsetting any beneficial effects of somatic mutations. However, mutation rates are highly heterogeneous across the genome, sometimes varying by several orders of magnitude even amongst neighboring loci (Hodgkinson and Eyre-Walker 2011; Makova and Hardison 2015), which produces mutational hotspots that can be sources of evolutionary adaptations (Galen et al. 2015; Xie et al. 2019b). Similarly, these hotspots may confer the elevated somatic mutation rates suggested by our model to promote adaptation via non-heritable somatic mutation, without increasing the mutation load elsewhere in the genome. In our model, we studied beneficial non-heritable somatic mutations, although deleterious somatic mutations could also influence the evolutionary trajectories of populations. Extending our model to include deleterious mutations elsewhere in the genome will illuminate how somatic genotypic exploration steers a population through an adaptive landscape, not only toward adaptive peaks, but also away from adaptive valleys via the purging of deleterious genetic variation, akin to what has been shown for other types of non-heritable variation (Bratulic, Toll-Riera, and Wagner 2017; Kosinski and Masel 2020; Zheng, Guo, and Wagner 2021).

Our model also suggests that non-heritable somatic mutations can promote adaptation when a small number of cells with the adaptive somatic mutation are required to confer a selective advantage to the organism. Cell signaling offers a promising example of a biological process where this may occur, because a small number of cells with a somatic peak genotype influencing signal emission could orchestrate the behavior of many more cells with alternative genotypes. For example, somatic mutations in organs with endocrinological functions can drastically alter individual physiology and development (Kim and Kim 2019; Richter-Unruh

et al. 2002), even when those mutations do not cause enhanced cell proliferation and are in normal non-tumoral tissue (Azizan et al. 2013; Nishimoto et al. 2015). In a developmental context, small disturbances from signals could trigger the formation of new patterns in embryos (Schweisguth and Corson 2019), disturbances which could come about from somatic mutations affecting paracrine signaling in few cells amongst a population of cells that do not have the somatic mutation.

As an illustrative example of a system in which a few mutant cells can have major phenotypic consequences, consider the development of the striped pattern of zebrafish. This pattern arises from the precise arrangement of different pigmented cell types that dynamically interact with each other and with themselves to coordinate their positioning and proliferation (Owen, Kelsh, and Yates 2020; Patterson and Parichy 2019). Mutations to specific genes are capable of altering interactions between cell types or preventing their differentiation entirely, which result in major differences in the pattern phenotypes (Podobnik et al. 2020; Singh and Nüssslein-Volhard 2015). We presumed that if these mutations could occur or be reversed somatically during development, they could influence pattern formation. We tested this using a computational model of the cellular interactions between pigment cells that accurately predicts the pattern of wild-type fish and that of different mutants (Owen, Kelsh, and Yates 2020). Adding a single somatic mutant cell during the development of a fish that could not produce stripes recreated completely or partially the pattern of the striped wild-type zebrafish, depending on the timing of the mutation during development (Fig. 4.5). Supporting these outcomes, Maderspacher and Nüsslein-Volhard (2003) reported the rescue of a wild-type striped phenotype in mutant zebrafish incapable of producing stripes, enabled by a small number of wild type cells derived from grafting their progenitors into the mutant embryo. This is evidence supporting the potential of a small number of somatic mutant cells to substantially affect the development of a selectable phenotype.

Our model shows that non-heritable somatic mutations are more likely to promote adaptation when they confer a selective advantage not only to the organism, but also to the somatic cells in their developmental context. In other words, somatic genotypic exploration is facili-

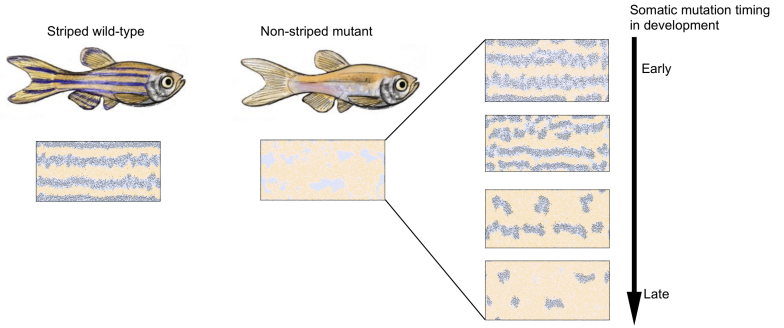


Figure 4.5: Non-heritable somatic mutations influence pattern formation in a model of zebrafish development. We modified a model of developmental patterning in zebrafish (Owen, Kelsh, and Yates 2020) to include somatic mutations (Annex III). Developing from a non-striped germline nacre mutant, the emergence of a single somatic mutation can cause stripe formation. The extent of stripe formation depends on when during development the somatic mutation arises, such that early-arising somatic mutations replicate the wild-type stripe pattern, whereas late-arising somatic mutations cause a more diffuse stripe pattern. This example highlights a biological scenario in which only few somatic mutations are required to cause drastic changes in a selectable phenotype.

tated if somatic mutations that are beneficial to the organism also confer a competitive advantage to the cells in which they arise, allowing these cells to increase in frequency within the organism. The further study of this angle of somatic genotypic exploration will greatly benefit from current approaches to the study of cancer in which the development of tumors based on clonal expansions is thought of as an ecological process in the tissue and cellular context where somatic variants arise (Lloyd et al. 2016; Neinavaie et al. 2022; Somarelli 2021) and from the general study of cell competition during development (Extavour and García-Bellido 2001; Morata and Ripoll 1975; Moreno, Basler, and Morata 2002; Otto and Hastings 1998). Phenotypes that could be affected by somatic genotypic exploration facilitated by proliferative somatic mutant cells include the many morphological innovations in animal evolution that resulted from changes in cell proliferation during development and the parameters controlling its onset and cessation (Alberch et al. 1979; Conlon and Raff 1999). For example, differences in cell proliferation in the facial development of some species of phyllostomid bats help explain the different lengths of their snouts in relation to the shape of the flowers they feed on (Cama-

cho et al. 2020), and genes involved in cell proliferation show indications of positive selection in animals of relatively large size such as capybaras (Herrera-Álvarez et al. 2021). Conceivably, different degrees of cell proliferation revealing beneficial phenotypes like in these examples might have arisen first from prolific somatic variants, and selection acting on those variants eventually promoted their emergence in the germline.

Ultimately, the biological plausibility of somatic genotypic exploration is an empirical question, which future work could address with a natural case study or using laboratory experiments. Identifying a natural case study may prove challenging, because of the need to pinpoint somatic variants that cause beneficial phenotypes. However, this may eventually be possible using single-cell genomics technologies, which are advancing rapidly (Dou et al. 2018). Alternatively, laboratory experiments with developing animals may provide a path forward. For example, gene editing technologies can be used to identify somatic mutations that induce measurable changes in organismal phenotypes, such as in the aforementioned example of stripe formation in zebrafish, which is a well characterized developmental system for which precise gene editing techniques are already available (Rosello et al. 2021). Additionally, laboratory evolution experiments could be carried out using developing model organisms that have short generation times, such as nematodes or flies, in which the somatic functionalization of a reporter gene could be selected.

4.3.2 Evolutionary implications of somatic genotypic exploration

Somatic genotypic exploration can impact evolution in at least four ways. First, somatic genotypic exploration can make the vast supply of non-heritable genetic diversity adaptively relevant. As was pointed out by Frank (2010), the cell lineage history of the development of a single human individual is tremendous, exceeding the lineage history of hominids. Given empirical rates of somatic mutations, a single body can thus harbor immense genetic diversity, which is only now starting to be explored in depth with sequencing technologies (Abascal et al. 2021; Blokzijl et al. 2016; García-Nieto, Morrison, and Fraser 2019; Lee-Six et al. 2018; Martincorena et al. 2015; Martincorena et al. 2018; Moore et al. 2021; Zhu et al. 2019). Such diversity cannot

directly enter the germline, but by fueling somatic genotypic exploration it could still influence evolutionary trajectories. Indeed, as our model shows, non-heritable somatic mutations can steer evolving populations toward adaptive peaks, as well as increase the rate of adaptation to these peaks. Although our study has focused on organisms with an impermeable germline-soma separation, we speculate that somatic genotypic exploration can also accelerate adaptation in organisms that do not have such a strict barrier between the germline and soma, because even in these organisms, most somatic mutations are not inherited. It would be interesting to adapt our model to explore how somatic genotypic exploration could influence adaptation of populations in which somatic variants are occasionally inherited, thus expanding on what is known about the influence of somatic mutations in the evolution of such organisms (Reusch, Baums, and Werner 2021).

Second, somatic genotypic exploration allows selection to act on the potential of genotypes to produce non-heritable adaptive phenotypes, facilitating the eventual fixation of those phenotypes via germline mutations. This makes somatic genotypic exploration akin to the genetic assimilation of plastic phenotypes triggered by environmental conditions(Crispo 2007; Pigliucci, Murren, and Schlichting 2006; Waddington 1942, 1953), by the stochasticity or "noise" of cellular processes (Kaern et al. 2005; Payne and Wagner 2019; Schmutzer and Wagner 2020; Whitehead et al. 2008), or by epigenetic modifications (Klironomos, Berg, and Collins 2013). Within this context, the so-called "look-ahead effect" (Whitehead et al. 2008) is particularly relevant; in this model, phenotypic mutations caused by transcription or translation errors create potentially adaptive protein variants, offering a mechanism for channeling populations towards adaptive genotypes, as in our model. However, because somatic genotypic exploration relies on the exploratory potential of somatic mutations that arise during development, it can act on substantially different phenotypes via substantially different biological processes. For instance, the effect of cell-lineage selection would be irrelevant in the absence of some degree of intra-organismal inheritance, which is provided by somatic mutations, but would be mostly absent in the transcriptional and translational errors enabling the look-ahead effect. Third, somatic genotypic exploration can influence the mosaic evolution of mutation rates across the

genome. Although high genome-wide mutation rates can be deleterious, individual loci can evolve higher mutation rates if selection favors their diversification (Sniegowski et al. 2000). The locus-specific rate can result from localized structural and functional properties, such as the fragility of segments of DNA strands with specific nucleotide sequences (Xie et al. 2019b), how often the locus is transcribed (Chen et al. 2017), the influence of chromatin organization (Schuster-Böckler and Lehner 2012), nucleotide composition and mutation biases (Cano et al. 2022; Fryxell and Moon 2005; Stoltzfus and McCandlish 2017), or specific targeting by biomolecular mechanisms (Odegard and Schatz 2006). Selection may favor elevated mutation rates in genomic regions where adaptive phenotypes can be revealed by somatic mutation, without influencing mutation rates elsewhere. The resulting mosaic of mutation rates across loci implies that different parts of the genome could be subject to the different evolutionary regimes we uncovered.

Fourth, somatic genotypic exploration can cause developmental bias. Developmental bias exists when certain phenotypes are produced more readily than others, thus influencing evolutionary trajectories and outcomes (Maynard Smith 1970; Uller et al. 2018). They can arise either from developmental constraints impeding the emergence of certain phenotypes (Zalts and Yanai 2017) or through developmental drive, which accounts for the increased likelihood of some phenotypes (Arthur 2001). Some causes of developmental drive are high mutation rates in genomic regions affecting an evolving trait (Galen et al. 2015; Xie et al. 2019b), the genetic architecture of the trait (Besnard et al. 2020; Stern and Orgogozo 2008), and the number of genotypes mapping to a phenotype (Dingle et al. 2022). Somatic genotypic exploration is a form of developmental drive in the latter sense, because it causes genotypes to intermittently express beneficial phenotypes that they would not otherwise express in the absence of somatic mutation, thus altering the genotype-phenotype map.

Overall, our study offers a theoretical grounding for the further analysis of non-heritable somatic mutations as a source of adaptation. Future empirical studies can help evaluate the plausibility of somatic genotypic exploration, through analyses of traits affected by genomic regions with high somatic mutation rates, phenotypes that can be altered by relatively few

or clonally expanding mutant cells, and phenotypic innovations derived from changes in the proliferation or mortality of cells during development. For these analyses, we need to better understand the dynamics of somatic mutant cells within the organism and how somatic genetic diversity affects phenotypes beyond cancer and senescence. By studying somatic genotypic exploration as a potential adaptive mechanism, we can elucidate whether and how the immense genetic diversity of the soma directs evolutionary trajectories toward adaptation. If that proves to be the case, the soma and by extension the organism as a whole, is not only the instantiation of the founding genotype present in the zygote, but also an important source of adaptive potential.

4.4 Methods

Baseline model. We modelled a population of multicellular organisms with an impermeable germline-soma division navigating a fitness landscape. We used a haploid two-locus two-allele model in which one of the four possible allele combinations represented the peak conferring a selective advantage s_{organism} over the other three genotypes (Fig. 4.1A). We used Wright-Fisher simulations with a population of N haploid individuals, where we represented each individual by the mutational distance of its germline genotype to the peak genotype. We initialized monomorphic populations at the maximum distance from the peak (i.e., genotype ab, which is two mutations from the peak). We ran each simulation for 5000 generations, each of which consisted of a developmental phase and a reproductive phase (Fig. 4.1B). In the developmental phase, we modelled the somatic growth of each individual in the population as a branching process with D developmental cycles consisting of synchronized rounds of cell divisions starting from a single cell, until the individual reached a reproductive somatic size 2D. The starting cell contained the germline genotype and at each cell division, somatic mutations occurred at rate μ_{soma} without altering the germline genotype. To implement this, at each round of cell division, we sampled the number of mutated cells from a binomial distribution B(n, μ_{soma}), where N was the number of somatic cells with each genotype (ab, aB, Ab or AB). The genotypic composition of the soma at the end of development defined organismal fitness. In the

case of our baseline model, having at least one somatic cell with the peak genotype provided the full selective advantage s_{organism}, producing an organismal fitness of 1 + s_{organism}; otherwise the fitness was 1.

In the reproductive phase, germline genotypes were selected with replacement from the population with a probability proportional to organismal fitness at the end of development. At this step, the germline genotypes of offspring were mutated at a rate μ_{germline} per locus. These selected and possibly mutated germline genotypes produced the population of the next generation.

As a control, we also considered a version of our model where somatic mutations did not affect fitness. To do so, we only included the reproductive phase in each generation. Organismal fitness was therefore defined exclusively by germline genotype. The default parameters for our baseline model were D = 25, N = 100000, s_{organism} = 1, $\mu_{\text{soma}} = 5 \times 10^{-9}$ and $\mu_{\text{germline}} = 1 \times 10^{-8}$. However, we also explored the parameter space by including ranges from D = 10 to D = 30, from N = 1000 to N = 200000, from s_{organism} = 0 to s_{organism} = 10, from $\mu_{\text{soma}} = 1 \times 10^{-10}$ to $\mu_{\text{soma}} = 5 \times 10^{-4}$ and from $\mu_{\text{germline}} = 1 \times 10^{-10}$ to $\mu_{\text{germline}} = 5 \times 10^{-4}$.

Fitness functions. We ran simulations in which organismal fitness was a function of σ_{peak}, which is the fraction of the developed organism's somatic cells with the peak genotype. To do so, we defined the fitness of each individual i as 1 + $s_{\text{organism}} F_i(\sigma_{\text{peak}})$, where $F_i(\sigma_{\text{peak}})$ was a monotonic function of σ_{peak} that yielded values between 0 and 1. $F_i(\sigma_{\text{peak}})$ thus determined the selective advantage s_{organism} conferred to an individual, according to its fraction of somatic cells with the peak genotype. We defined $F_i(\sigma_{\text{peak}})$ as indicated in equations 4.1 and 4.2. In Eq. 4.1 f was a constant defining the shape of the function. For the seven functions used the values were $f = 5$, 10 and 100 for concave functions (Fig. 4.4A, circles), $f = \frac{1}{100}$, $\frac{1}{10}$ and $\frac{1}{5}$ for convex functions (Fig. 4.4A, stars) and $f = 1$ for the linear function (Fig. 4.4A, diamond). For the fitness function defined by Eq. 4.2 we chose six values of g in order to study a range of fitness functions that required different numbers of somatic cells with the peak genotype to confer the full selective advantage s_{organism}. These values were $g = 0$ for the linear function (Fig. 4.4D, diamond), and $g = 12$, 107, 1055, 10538, and 10000 for the remaining diminishing

returns functions (Fig. 4.4D, circles). The set of baseline parameters we used in combination with these fitness functions was D = 25, N = 100000, a range of s_{organism} from s_{organism} = 0.01 to s_{organism} = 10, μ_{soma} = 5×10^{-9}, 1×10^{-8}, or 5×10^{-8}, and $\mu_{\text{germline}} = 1 \times 10^{-8}$.

Cell fitness. We ran simulations in which somatic mutations to the AB genotype conferred a selective advantage not only to the organism, but also to the somatic cell. To do so, we added an extra stage to the developmental phase, in which somatic cells with AB genotypes had a proliferative advantage over somatic cells with other genotypes. Specifically, they divided at a rate s_{cell} times that of somatic cells with other genotypes. Somatic mutations occurred at the same rate μ_{soma} in these cell divisions as in other cell divisions. Although under natural conditions the cellular fitness effect of mutations will depend on when in development and where in the genome the mutations occur (see Cannataro et al. (2018) for a comparison of the selective effect sizes of nucleotide variants in cancerous cells), we used a representative single fixed value of $s_{\text{cell}} = 2$, which doubles the reproductive capacity of cells with the peak genotype relative to other cells. This value approximates estimations for differential proliferative capacities among somatic variants (Morata and Ripoll 1975; Zhu et al. 2019) and populations of cells in stages of development that are comparable to each other across different species (Camacho et al. 2020).

Fitness valleys. We ran simulations in which the intermediate germline genotypes aB and Ab conferred a fitness disadvantage to the organism, relative to the genotypes ab and AB. Specifically, we modified our baseline model such that individuals with the ab germline genotype had a fitness of 1, AB germline genotype had a fitness of $1 + s_1$, and the intermediate germline genotypes had a basal fitness of $\frac{1}{1+s_2}$, which could be increased via somatic mutation to AB. In these simulations, we explored values for$s_1$and s_2 ranging from 0 to 10 and values for D ranging from 10 to 30.

4.5 Supplements

4.5.1 Annex I: Probabilistic analysis

To aid in the interpretation of Fig. 4.2, we present a probabilistic analysis of our baseline model. In our control simulations, in which somatic mutations do not influence organismal fitness, the probability of fixation for the peak germline genotype AB depends on μ_{germline}. Because intermediate Ab and aB germline genotypes are selectively neutral with respect to the initial ab germline genotype, their dynamics are governed by genetic drift. As these intermediate genotypes fluctuate in frequency within the population, they provide an opportunity for the origin and fixation of the peak germline genotype AB via stochastic tunneling (Iwasa, Michor, and Nowak 2004). At the lowest values of μ_{germline} considered, the mutation supply is too low to facilitate stochastic tunneling. For example, when $\mu_{\text{germline}} = 10^{-9}$ and $N = 10^5$, only a single germline mutation is expected in either locus within 5,000 generations. For the same population size, we begin to regularly observe fixation of the peak genotype as μ_{germline} exceeds 5×10^{-6} (Fig. 4.2A,B), when a single germline mutation is expected per locus every 10 generations.

When somatic mutations influence organismal fitness, the probability of fixation for the peak germline genotype AB depends on both μ_{germline} and μ_{soma}. In our baseline model, an individual will have fitness $1 + s_{\text{organism}}$ if at least one somatic cell carries the peak genotype AB. Starting from a zygotic cell with germline genotype ab, this requires that during D developmental cycles, a somatic mutation first produces a lineage of cells with an intermediate Ab or aB genotype, and a subsequent somatic mutation in a cell of that lineage produces the peak genotype AB. In contrast, starting from a zygotic cell with germline genotype Ab or aB, a single somatic mutation can produce the peak genotype AB.

To understand the trends reported in Fig. 4.2B,D, we need to calculate the probability of observing at least one somatic AB genotype in D developmental cycles. Because we model development as a multitype branching process, we can calculate this probability using probability generating functions (Agresti 1974). We begin by writing the generating function for the

two daughters of a cell with genotype ab as

$$f_{\text{ab}}(x, y, z) = ((1 - \mu_{\text{soma}})^2 x + 2\mu_{\text{soma}}(1 - \mu_{\text{soma}})y + ((\mu_{\text{soma}}{}^2)z)^2, \tag{4.3}$$

where $(1 - \mu_{\text{soma}})^2$ is the probability of ab not mutating into any other genotype (i.e., no mutations in either locus), $2\mu_{\text{soma}}(1 - \mu_{\text{soma}})$ is the probability of mutating into Ab or aB (i.e., a single mutation in one of the loci), and $\mu_{\text{soma}}{}^2$ is the probability of ab mutating into AB (i.e., mutation in both loci). The arguments x, y, and z, are used to extract the probabilities of genotypes ab, Ab or aB, and AB, respectively. For instance, the probability of having two daughters with an Ab or aB genotype is $f_{\text{ab}}(0, 1, 0) = (2\mu_{\text{soma}}(1 - \mu_{\text{soma}}))^2$, while the probability that neither of the daughter cells has the genotype AB is $f_{\text{ab}}(1, 1, 0) = ((1 - \mu_{\text{soma}})^2 + 2\mu_{\text{soma}}(1 - \mu_{\text{soma}}))^2$.

We can similarly write the generating functions starting from genotypes Ab/aB and AB as

$$f_{(\text{Ab/aB})}(x, y, z) = (\mu_{\text{soma}}(1-\mu_{\text{soma}})x+((1-\mu_{\text{soma}})^2+\mu_{\text{soma}}{}^2)y+\mu_{\text{soma}}(1-\mu_{\text{soma}})z)^2 \tag{4.4}$$

and

$$f_{\text{AB}}(x, y, z) = ((1 - \mu_{\text{soma}})^2 x + 2\mu_{\text{soma}}(1 - \mu_{\text{soma}})y + (\mu_{\text{soma}}{}^2)z)^2 \tag{4.5}$$

Starting from a single genotype, we need all three generating functions to calculate the probability of a given genotype mutating into the other genotypes through the division cycles of development. Therefore, we vectorize the generating functions given in Eqs. 5.3-5 as

$$F(x, y, z) = (f_{\text{ab}}(x, y, z), f_{(\text{Ab/aB})}(x, y, z), f_{\text{AB}}(x, y, z)) \tag{4.6}$$

We can then use Eq. 4.6 to capture the nested probabilities of observing different genotypes after a given number of developmental cycles: starting from an ab genotype, the probability of not observing a cell with genotype AB after two developmental cycles ($D = 2$) is $f_{\text{ab}}(F(1, 1, 0))$, after three developmental cycles ($D = 3$) is $f_{\text{ab}}(F(F(1, 1, 0)))$, and after D developmental cycles is $f_{\text{ab}}(F^{(D-1)}(1, 1, 0))$. The complementary probability $1-f_{\text{ab}}(F^{(D-1)}(1, 1, 0))$

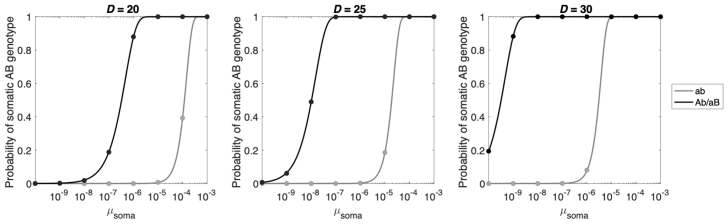


Figure 4.6: Probability generating functions. Probability of observing at least one somatic AB genotype during D = 20, 25, or 30 developmental cycles, shown in relation to the somatic mutation rate μ_{soma} for ab (gray) or Ab/aB (black) germline genotypes. The lines show numerical evaluation obtained using generating functions for 1000 logarithmically spaced values of the somatic mutation rates between $\mu_{\text{soma}} = 1 \times 10^{-10}$ and $\mu_{\text{soma}} = 1 \times 10^{-3}$, and symbols show simulation results, specifically the fraction of N = 10000 individuals in a single generation with at least one AB genotype in their soma at the end of the developmental phase, starting from a zygote with an ab (black circles) or Ab (gray circles) germline genotype.

then gives the probability of observing at least one cell with the genotype AB, having started the developmental process with one cell with the genotype ab. The same can be calculated for an individual that develops from an Ab or aB zygote as $1 - f_{(\text{Ab/aB})}(F^{(D-1)}(1,1,0))$. Fig. 4.6 shows these probabilities as a function of μ_{soma} for three values of D, which agree perfectly with data from our simulations.

These calculations help to explain the evolutionary outcomes of our baseline simulations (Fig. 4.2C,D). This is because the probability of an organism to produce offspring will not only be given by the selection coefficient s_{organism}, but also by the probability of mutating into the AB genotype somatically. This smoothens the adaptive landscape (Frank 2011) by producing different effective selection coefficients for each of the germline genotypes, so that the selection coefficient for the ab genotype is $s_{\text{ab}} = s(1 - f_{\text{ab}}(F^{(D-1)}(1,1,0)))$ and the selection coefficient for the intermediate genotypes is $s_{Ab/aB} = s(1 - f_{(\text{Ab/aB})}(F^{(D-1)}(1,1,0)))$. When μ_{soma} is sufficiently high, a somatic AB genotype is guaranteed to emerge from an ab germline genotype or from an intermediate (Ab or aB) germline genotype ($5 \times 10^{-5} < \mu_{\text{soma}}$, Fig. 4.6, $D = 25$), rendering the germline ab genotype selectively equivalent to the intermediate and peak germline genotypes. The evolutionary dynamics of the germline genotypes are therefore governed by genetic drift. As such, the population tends to remain at the ab germline geno-

type, with intermediate and peak germline genotypes only going to fixation at high μ_{germline} (notice the shading for the highest μ_{germline} values in the rows for $\mu_{\text{soma}} = 1 \times 10^{-4}$ and 5×10^{-4} in Fig. 4.2D). In contrast, for intermediate values of μ_{soma}, a somatic AB genotype is not guaranteed to emerge from an ab germline genotype, but it is guaranteed to emerge from an intermediate germline genotype ($1 \times 10^{-7} < \mu_{\text{soma}} < 5 \times 10^{-5}$, Fig. 4.6, D = 25). The ab germline genotype therefore has lower fitness than the intermediate and peak germline genotypes, which are selectively equivalent. As such, the intermediate germline genotypes are driven to fixation by selection. Once this occurs, the evolutionary dynamics of the intermediate and peak germline genotypes are governed by genetic drift, and the population tends to remain at the intermediate germline genotype. For lower values of μ_{soma}, where a somatic AB mutation may, but is not guaranteed, to emerge from an intermediate germline genotype ($1 \times 10^{-10} < \mu_{\text{soma}} < 1 \times 10^{-7}$), there is selective pressure on both the intermediate and peak germline genotypes, such that the population tends to evolve to the peak germline genotype.

4.5.2 Annex II: Fitness valleys

In our baseline model we studied a fitness landscape with three genotypes of equal fitness (ab, Ab, and aB) and a fourth genotype (AB) with a selective advantage s_{organism} (Fig. 4.1A). We now study a rugged fitness landscape with two adaptive peaks separated by an adaptive valley. Specifically, we assigned a fitness of 1 to genotype ab, a fitness of $1 + s_1$ to genotype AB, and a basal fitness of $1/(1 + s_2)$ to the intermediate genotypes aB and Ab, which could be increased via somatic mutation to AB (Fig. 4.7A). Such landscape ruggedness can hinder evolution, because populations can become trapped in local adaptive peaks (Gokhale et al. 2009). However, we reasoned that non-heritable somatic mutations might permit valley crossing, because the fitness disadvantage of the Ab or aB germline genotypes could be offset by the fitness advantage of somatic mutations to the AB genotype. Using our baseline model with a fitness valley of s_2 = 0.5, the same three qualitative evolutionary outcomes emerged as without the valley, but the parameter combinations yielding each outcome changed quantitatively (Fig. 4.7B). We

further explored how different combinations of selective values for peaks s_1 and valleys s_2 affected the likelihood of converging on the peak under different somatic mutation supplies, by varying the number of developmental cycles (Fig. 4.7C). For a low somatic mutation supply (D = 20), populations never converged on the peak if there was a valley, in contrast to our baseline model without a valley (s_2 = 0, Fig. 4.7D). For an intermediate somatic mutation supply (D = 25), populations converged on the peak so long as s_1 was not too low. For a high somatic mutation supply (D = 30), populations remained one mutation away from the peak regardless of the depth of the valley, so long as $s_1 > 0$. Thus, a high somatic mutation supply can mask the deleterious effects of the intermediate germline genotypes, such that evolving populations converge on genotypes that would be otherwise maladaptive in the absence of somatic mutations. In sum, this analysis shows that non-heritable somatic mutations can facilitate valley crossing, but whether or not they do depends on the depth of the valley and the somatic mutation supply.

4.5.3 Annex III: Zebrafish pattern development

The pigment patterns of wild type zebrafish consist of interspersed blue and golden stripes (Fig. 4.5) that arise from interactions between three different cell types: xanthophores containing yellow and orange pigments, melanophores with black pigments, and iridescent iridophores containing reflective crystals (Hirata et al. 2003; Patterson and Parichy 2019). Owen et al (2020) used a computational model of the development of zebrafish that included these interactions to replicate the patterns of wild-type and mutant zebrafish. The model is based on three overlaying and interacting lattices, each corresponding to a layer in the skin of zebrafish, where each of the different cell types are located.

We used this model to test how somatic mutations that arise during development in the gene nacre influence pattern formation, beginning with a germline nacre genotype that has a non-stripe phenotype. The nacre gene encodes a transcription factor that activates the differentiation of the cellular precursors of melanophores (Lister et al. 1999), and mutations disrupting its function result in patterns that are dominated by xanthophores and iridophores exclusively

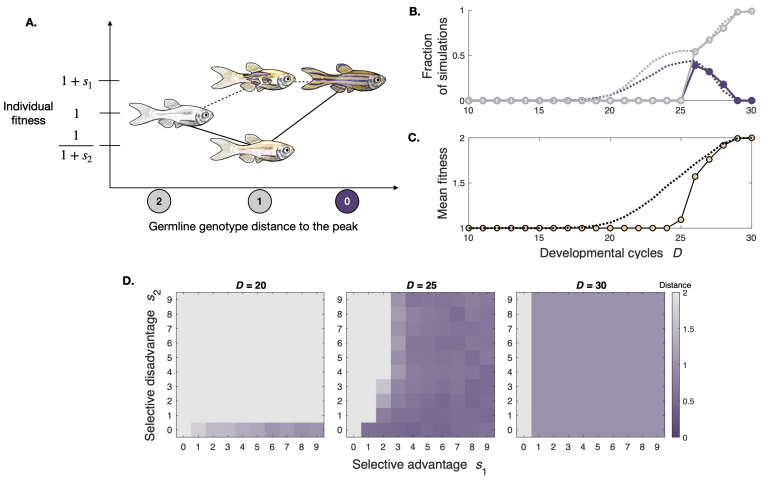


Figure 4.7: Somatic mutations can facilitate valley crossing. (A) We modified our baseline fitness landscape (Fig. 4.1A) to include a fitness valley at intermediate germline genotypes (i.e., at genotypes aB and Ab). (B) The fraction of simulations in which the population reached the peak genotype (purple) or remained one mutation away from the peak genotype (silver) after 5000 generations is shown in relation to D, for $s_2 = 0.5$. (C) Mean population fitness after 5000 generations for the same simulations as in (B). The dotted lines show the data from Fig. 4.3B and C, for reference. (D) Heatmaps of the mean distance to the peak after 5000 generations. Rows correspond to different values of the selective disadvantage s_2 of the genotypes aB and Ab, while columns correspond to different values of the selective advantage s_1 of genotype AB. The labels (i), (ii) and (iii) refer to the different evolutionary outcomes described in Fig. 4.3A. In all panels, $N = 100000$, $\mu_{\text{soma}} = 5 \times 10^{-9}$ and $\mu_{\text{germline}} = 1 \times 10^{-8}$.

(Fig. 4.5). We modified the model by adding a step that incorporates a single cell in a random position of the melanophore lattice, which is empty in the simulations for the nacre mutant. We ran simulations by fixing this step at different timepoints of development. The outcomes of representative simulations fixing the mutation step after 10000, 100000, 1000000 or 2000000 developmental steps are presented in Fig. 4.5 and discussed in the main text.

The script from Owen et al. (2020) is public and can be accessed at:

https://github.com/elifesciences-publications/Zebrafish-stripe-model

We made two modifications to the script main.m. These are clearly indicated in our modified version of this script, which is available as online supplementary material.

5 Synthesis and outlook

> "Welcome, O life! I go to encounter for the millionth time the reality of experience and to forge in the smithy of my soul the uncreated conscience of my race."
>
> James Joyce, A Portrait of the Artist as a Young Man.

Evolutionary trajectories are a historical product, and the paths on which life advances are paths that are built on the memory of past adaptive innovations. Some of these innovations can have a drastic impact on the way evolution unfolds, because they can have repercussions on the interaction between basic principles like mutation, population dynamics, selection, reproduction, morphogenesis, etc. This is what happens, for example, in the case of major evolutionary transitions, such as the evolution of multicellularity (Maynard Smith and Szathmary 1997). In this thesis I have studied three evolutionary implications of organizational innovations that arose in stem animals. In chapters 2 and 3, I focused on two evolutionary consequences of the origination of enhancers, distal cis-regulatory elements in the genome that represent an increase in the complexity of gene regulatory programs. In chapter 4, I explored a hypothesis by which the immense genetic diversity that is available in the non-reproductive cells of animals might fuel and guide their adaptation. In this chapter I will contextualize and discuss some of the implications of these findings.

5.1 Enhancers of evolution

An evolutionary novelty does not only affect the immediate physiology and structure of an organism in which it arises, but it also impacts the evolutionary future and the capacity to diversify that organism's lineage (Erwin 2015). If we consider the specific case of enhancers, one can only speculate about what were the original adaptive reasons leading to their arrival and fixation in the ancestor of animals. But what one can study is how the evolution of these novel regulatory features fundamentally impacted the developmental and evolutionary capabilities of metazoans.

Enhancers can tune the rate of transcription of target genes, as well as define their spatiotemporal patterns of expression. As mentioned in the first chapter of this thesis, one of the most prominent features of gene regulation mediated by enhancers is the regulatory and evolutionary modularity they confer to developmental systems. The fact that enhancers are distally located with respect to promoters, and that the genomic regions where they are located can become accessible or inaccessible under different circumstances, offers a degree of flexibility to the regulatory networks they help wire. This allows cells to fine tune the expression of cassettes of genes in specific contexts, and to avoid activating those genes when they are not needed. Such regulatory flexibility facilitates the division of labor among cells of multicellular organisms, by allowing for the differential deployment of distinct sub-circuits of a gene regulatory network across different cells. If we look at unicellular organisms lacking enhancers, we can also identify different cell states akin to different cell types, such as in cyanobacteria (Flores 2012) or in yeast (Galgoczy et al. 2004). But the diversity of cell states these organisms can achieve is meager relative to the diversity that can be attained in animals, or for that matter, plants, who also seem to employ distal-acting regulatory elements (Clark et al. 2006). Supporting this view, Sebé-Pedrós et al. (2018b) did a transcriptomic profiling of single cells of three basal clades in the metazoan radiation, finding that cell type diversity is richer in ctenophores than in sponges or placozoans, and that this may be due to the expanded use of distal regulatory elements in ctenophores. Therefore, upon their evolution, enhancers enriched the genotype-phenotype map for regulatory networks by increasing the number of

phenotypes that could be realized by any one genotype.

The same properties of enhancers allowing them to offer plasticity to regulatory networks during development also allows them to promote network rearrangements at evolutionary timescales. That is because the localized nature of their activity allows mutations occurring on enhancers to specifically modify or co-opt morphological modules, without disrupting other phenotypes (Prud'homme, Gompel, and Carroll 2007). By tweaking regulatory sub-circuits, the evolution of enhancers facilitate the exploration of alternative configurations of gene regulatory networks, promoting evolvability. However, phenotypes whose development is mediated by enhancer activity have different degrees of mutational robustness and evolvability depending on a number of factors. One of those has to do with the fact that enhancer function is derived from hosting transcription factors binding sites, which have different degrees of robustness and evolvability in their own right (Payne and Wagner 2014). Another factor is the regulatory syntax the enhancer utilizes. An enhancer's syntax consists of the identity and combination of transcription factors that bind to it, as well as the spacing, affinity, order and orientation of the binding (Long, Prescott, and Wysocka 2016). For example, enhancers regulating the expression of the *shavenbaby* gene in *Drosophila* harbour clusters of low-affinity binding sites for the transcription factor Ultrabithorax bound to a cofactor, which offers a high robustness to genetic alterations and opportunities for interspecific genetic diversification (Crocker et al. 2015). Another example is the case of a recently identified enhancer from the sponge *Amphimedon queenslandica* (Wong et al. 2020). This enhancer is not conserved at the level of the nucleotide sequence, but it has in fact conserved its transcription factor binding syntax across lineages that diverged 700 million years ago, before the Cambrian explosion. Strikingly, upon the transformation of this sponge enhancer into zebrafish embryos, the resulting pattern of gene expression was the same as when the zebrafish were transformed with the mouse and human homolog enhancer. Additionally, mutagenesis analysis of 23 enhancers that are highly conserved in different mammals have shown how these enhancers were functionally resilient to mutations (Snetkova et al. 2021). This shows how, by retaining transcription factor binding syntax, enhancers can explore broad neutral networks in the space of genotypes. And, as was

speculated by Wong et al (2020), the organization of enhancers "promotes robustness and may provide a foundation for further lineage-specific elaboration via the integration of additional transcription factor binding motifs and the dissociation of others". That said, however, even if an enhancer is highly evolvable when it comes to the formation of new regulatory connections in this way, deleterious pleiotropic effects might still constrain their evolution (Sabarís et al. 2019).

Another relevant factor for the evolution of an enhancer sequence is its context. As mentioned above, genes can be targeted by several enhancers, and whether an enhancer is acting alone or in the company of shadow enhancers will influence the robustness of the phenotypes that that enhancer helps develop (Osterwalder et al. 2018; Berthelot et al. 2018; Tsai, Alves, and Crocker 2019; Wang and Goldstein 2020; Kvon et al. 2021). Berthelot et al (2018) studied the regulatory landscape across 20 mammalian species and found that, in spite of the turnover of enhancers that had previously been described by Villar et al (2015), genes tended to maintain stable levels of expression if the number of enhancers in their vicinity was also stable. Therefore it would appear that in certain contexts, the identity of an enhancer is less important than the number of enhancers regulating a gene's function. Given this higher robustness resulting from more complex regulatory landscapes, it is expected that enhancers involved in complex regulatory hubs will be more likely to accumulate neutral mutations. This can have as a consequence that whenever an enhancer interaction is gained by a gene for its regulation, the mutational robustness of other regulatory elements involved in the interaction may also increase (Fig. 5.1). Enhancers acting in redundancy do indeed tend to accumulate higher levels of genetic diversity (Danko et al. 2018). Naturally, this might facilitate the exploration of the genotype space and thus also increase the opportunity to gain new regulatory interactions as previously described. But another way in which more enhancer interactions can impact evolvability is when they can act as evolutionary capacitors (Rutherford and Lindquist 1998) by enabling the accumulation of cryptic genetic diversity across regulatory elements. Enhancers show a high turnover both at evolutionary scales (Villar et al. 2015) as well as at developmental scales (Gao et al. 2018), which means that if an enhancer that is acting as a capacitor is eas-

ily lost, then the accumulated diversity can also be easily expressed. The fact that enhancers can increase the robustness of other regulatory elements might also have broadly affected the architecture and evolution of genomes when enhancers first evolved. That would be because upon their emergence they could have released pre-existing regulatory elements, such as promoters, from long-standing functional constraints. Relating to this point, Sebé-Pedrós et al. (2018b) have uncovered how the evolution of a broader usage of distal-acting regulatory elements is correlated with a lower specificity in promoter sequence. This implies that, upon the evolution of enhancers, part of the regulatory information that used to be encoded in promoters became distributed across different genomic regions. Thus, when enhancers are acting in redundancy they can have increased levels of mutational robustness, but they can also more broadly increase the robustness and evolvability of their co-regulators.

The mutational robustness showcased by some enhancer sequences offers them a high navigability across genotype networks. This has the potential of increasing the genetic diversity of transcription factor binding sites within enhancers, thus promoting the evolvability of more optimal binding patterns or the creation of new regulatory connections with new transcription factors (Payne, Moore, and Wagner 2014; Wong et al. 2020). A higher robustness within enhancer sequences can not only promote evolvability when it comes to the wiring of gene regulatory networks, but also when it comes to their expansion. Linking back to our findings shown in chapter 3, as enhancers neutrally explore the space of genotypes, they will likely encounter regions of that space coding for open reading frames (Fig. 5.1). Because enhancers are transcriptionally permissive, as they drift through a neutral network they can readily transcribe novel sequences, some of which might prove beneficial. This can be true even for protein-coding sequences. As we have shown, enhancers increase the opportunities for novel open reading frames to be expressed and translated by 1) granting them a higher and more stable transcription, 2) facilitating their association with ribosomes, 3) permitting a more stable transcription across different cell types, and 4) offering a facilitated path to the evolution of a promoter, because of the readiness with which enhancers can acquire promoter functionality (Wu and Sharp 2013; Carelli et al. 2018). All this implies that enhancers grant a degree

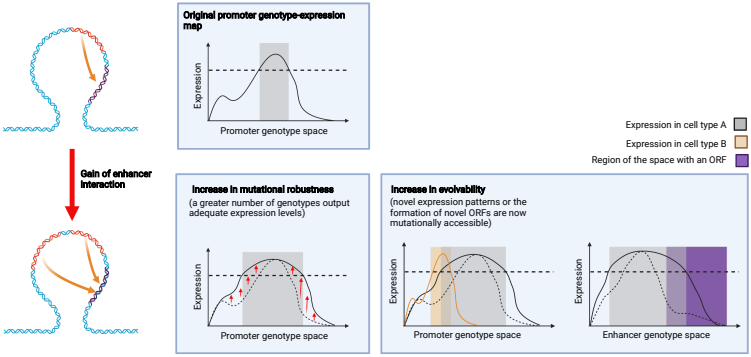


Figure 5.1: A gain of enhancer interactions might increase the robustness and evolvability of regulatory elements. An illustrative loop of DNA showing an enhancer-promoter interaction (orange arrow) and the profile of gene expression given the genotype on the promoter. The dashed line indicates a threshold above which the expression of the gene in question is adequate for a given biological function. The gray zone delineates the region of the genotype space that is viable given the resulting level of expression. Upon the gain of an additional enhancer interaction, the range of genotypes that outputs adequate levels of expression expands (red arrows). This can be regarded as an increase in the mutational robustness of the promoter. The panels on the right show two examples of how the changes in the range of genotypes outputting an adequate gene expression can influence the evolvability of the original promoter, as well as that of a redundant enhancer, by promoting the evolvability of new expression patters (yellow) or the evolution of new functional structures like an open reading frame (purple).

of pre-functionalizaton to novel open reading frames. Furthermore, another angle of how an enhancer might prime a *de novo* gene candidate for functionalization is that those open reading frames arising on enhancers will encounter a genomic environment enriched in specific transcription factor binding motifs corresponding to connections of a pre-existing regulatory network in which the enhancer was involved. Not to mention that they might arise in the proximity of the target gene of the enhancer on which they arose. These last two points have as a consequence that a newly evolved *de novo* gene can become readily integrated into a specific pre-existing regulatory circuit (Grandchamp et al. 2022). But as we have shown in chapter 2, new genes can gradually gain their own enhancer interactions and shuffle their regulatory connections as they mature, thus expanding their breadth of expression or stabilising their function.

In sum, enhancers offered a resourceful toolkit for animal innovation that greatly enhanced the evolvability of this clade. Firstly, the evolution of enhancers has allowed for a complexification of the possible phenotypes a single genome can map to, by greatly increasing the potential for the combinatorial and modular deployment of genes. Secondly, the molecular details of their action allows enhancers to be robust to mutations and to also increase the robustness of other regulatory elements, which not only helps stabilize adaptive phenotypes, but it can also help find novel configurations of regulatory networks that might also prove adaptive. Thirdly, the fact that enhancers are transcritpionally active and can readily evolve promoter activity makes them fertile ground for the evolution of novel genes. Overall, it is clear that enhancers were pivotal for the evolution of animal complexity in the form of an increase in the number of cell types (Carroll 2001; Sebé-Pedrós et al. 2016; Sebé-Pedrós et al. 2018b; Ros-Rocher et al. 2021), but, in addition to this, the novelty of enhancer regulation, with its evolvable and yet robust architecture, offered the animal lineage an enormous evolutionary potential to explore the space of genotypes and different configurations of their gene regulatory networks.

5.2 The adaptive guidance of somatic genetic variation

Only when a mutation is heritable can a population take a step in the space of possible genotypes as it explores a genotype space. Even in the instances in which a mutation has an adaptive outcome, the step will not be taken if it is not a trait that can be passed on to the progeny. This has the implication that, when stem animals evolved the germline-soma separation, much of the genetic variation arising as somatic mutations did not represent evolutionary steps that the whole population could follow. However, under the hypothesis presented in chapter 4, non-heritable somatic mutations still have the potential to guide evolutionary trajectories on the space of genotypes. We proposed that the non-heritable development of beneficial phenotypes via somatic mutations can act as a lighthouse, helping an evolving population to navigate towards the region of the genotype space that is adaptive. The traditional Neo-Darwinian paradigm is one that has equated the evolutionary process to a blind watchmaker that is blind "because it does not see ahead, does not plan consequences, has no purpose in

view" (Dawkins 1986). But under the model of somatic genotypic exploration, somatic mutations have the potential of "scanning the surroundings", of acting as a white cane to that blind watchmaker.

As pointed out by Mayr (1972), the inheritance of an adaptive trait is the *sine qua non* of the evolutionary process. Although somatic genotypic exploration is based on the development of non-heritable phenotypes, inheritance is still at the root of this process. Nevertheless, what we suggest is that selection not only can favour genotypes that produce an adaptive phenotype itself, but it can also favour genotypes with the heritable *potential* of evolving an adaptive phenotype. With each mutational step an individual takes on a space of genotypes, it moves closer or farther away from a region of that space that may be encoding an adaptive phenotype. Whenever the genotype resulting from that step is inherited, the relative distance to the adaptive region of the space is also inherited. For example, when a parent passes a genotype, let us say ACC, to its offspring, what the offspring is inheriting is not only the ACC genotype, but also the potential to mutate via a single point mutation to the genotype TCC, AGC, and all other mutational neighbours. Not only that, but it will also likely inherit any propensities or biases towards which that inherited genotype can mutate - for example, if it is a sequence that is more or less prone to mutate in a given direction. Therefore, by effectively taking those steps, even if non-heritably, selection will be able to favour the potential of evolving the adaptive trait. Evolution by natural selection can thus follow tendencies or propensities, since the selection on a phenotypic potential can eventually channel populations towards adaptation.

The idea of a potential that is inherited was intuited by Henri Bergson. When discussing the dogmatic negation of the importance of acquired traits that followed Weismann's exposition about the germ plasma, he wrote:

> "The acquired characters we are speaking of are generally habits or the effects of habit, and at the root of most habits there is a natural disposition. So that one can always ask whether it is really the habit acquired by the soma of the individual that is transmitted, or whether it is not rather a natural aptitude, which existed prior to the habit. This aptitude would have remained inherent in the germ-plasm which

> the individual bears within him, as it was in the individual himself and consequently in the germ whence he sprang" (Bergson 1907).

The idea that selection can act on the potential to evolve adaptive traits also has precedent. Baldwin (1896) argued that individuals show a degree of phenotypic plasticity that allows them to adapt within a generation, and that selection can act in directions of that plasticity. Eventually, genetic variation could arise in a way that traits that are originally triggered by the environment can become inherited.

A key difference between the selection on plasticity described by Baldwin and somatic genotypic exploration is that the plasticity Baldwin wrote about is an environmentally triggered plastic response, as is the case in more modern conceptualizations of phenotypic plasticity and genetic assimilation (Waddington 1961; West-Eberhard 2003; Pigliucci, Murren, and Schlichting 2006). On the other hand, in somatic genotypic exploration the source of the formation of the adaptive phenotype is the same genetic step that will be required for the evolution of a heritable version of the change. In the case of somatic genotypic exploration what is inherited is not only the potential to develop an adaptive phenotype, but also simultaneously the genotypic potential to mutate into it, to assimilate it genetically. In this last regard, somatic genotypic exploration shares some of its foundations with the look-ahead effect proposed by Whitehead et al (2008) for transcritptional and translational mutations.

On top of the potential of offering foresight of adaptive traits and channeling evolving populations, in the discussion of chapter four we have briefly mentioned a few other evolutionary implications of somatic genotypic exploration. Namely, we argued that, 1) the enormous genetic variation that is expected to exist in the thousands, millions or billions of somatic cells is not necessarily evolutionarily irrelevant, 2) it can lead to genetic assimilation of non-heritable phenotypes, 3) it can impact the evolution of localised mutation rates, and that 4) it can cause a developmental bias. The last point is particularly important for the question of how somatic genotypic exploration can influence evolutionary trajectories. As mentioned above, developmental bias can exist when the evolutionary paths leading to certain phenotypes are more accessible than the paths leading to other phenotypes (Maynard-Smith et al. 1985; Uller et

al. 2018), which can result, for example, from the relative abundances of genotypes mapping to each of the phenotypes (Cowperthwaite et al. 2008; Schaper and Louis 2014). Somatic genotypic exploration can result in a case of developmental bias in which there is a "pull" in the evolutionary direction of one phenotype over another whenever that phenotype is capable of being expressed via non-heritable somatic mutations (Fig. 5.2.A).

Additionally, somatic genotypic exploration can also impact evolutionary trajectories by biasing which paths are taken towards an adaptive peak (Fig. 5.2B). Consider a case in which the a/A locus corresponds to a receptor and the b/B locus to that receptor's ligand, and that the product of allele A bound to allele B induces an increase in the proliferation of the cells with the receptor A (thus replicating something like the scenario we described in chapter 4, section 2.5), but has no effect on cell proliferation when B binds the receptor a. This will have as a consequence that whenever an organism develops from a zygote with the aB genotype, if a somatic mutation leads to the peak genotype AB, then the mutant cells will be able to respond to the signal from B. Such cells will therefore proliferate and increase their proportion in the soma. This will not happen if the somatic mutation leading to AB arises in the context of an Ab developing soma. Consequently, in cases when many cells are needed for expressing the adaptive phenotype, aB will tend to have a higher fitness than Ab, and this path will be more likely to be taken on the way to the peak. Modifying the model used in chapter 4 to account for this situation indeed shows that this bias can occur (Fig. 5.2B, right panel).

The big question then is, what is the actual potential of real life phenotypes to be revealed by non-heritable mutations frequently enough so that they can push towards the genetic assimilation of the non-heritable mutation? Clearly not all phenotypes can be showcased by non-heritable mutations. Which phenotypes are prone to be evidenced in this way will mainly reside on two factors:

(i) The mutational neighbourhood: The development of the adaptive phenotype must be directed by a genotype that is mutationally close to the genotype in the genome of the zygote of a developing organism. That is because only when a genotype can repeatedly arise as the result of non-heritable mutations can selection efficiently act on the potential of evolving that

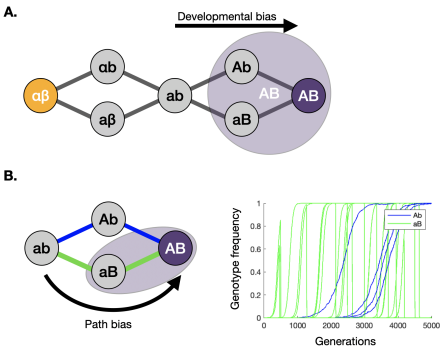


Figure 5.2: Somatic genotypic exploration as a source of evolutionary bias. (A) **Developmental bias.** Mutational network for 2 loci and 3 alleles (for simplicity, we do not consider allelic combinations of Greek and capital letters). AB and $\alpha\beta$ represent two genotypes that map to two adaptive phenotypes, purple and orange, respectively. In populations starting from genotype ab, both peak genotypes are mutationally equidistant. However, populations will more likely follow the evolutionary trajectories leading to the genotype that is capable of expressing an adaptive phenotype via somatic mutations, which in this case maps to the purple phenotype. (B) **Path bias.** When the genotype-phenotype map is structured so that one genetic background is more likely to express an adaptive phenotype than another, somatic genotypic exploration can bias the evolutionary path that is taken towards adaptation. In the mutational network shown, that would be the case of the green path. The panel on the right shows the result of adapting our model from chapter 4 to explore this scenario. I modified the model so that the B allele induces the cell proliferation of cells with the A allele (this could mimic the co-evolution of a ligand and its receptor controlling cellular division). In this case the developmental bias is towards the aB genotype that has a higher potential of expressing the adaptive phenotype non-heritably. The simulations consisted of 100 replicates, using the baseline parameters from the model in chapter 4 and using the least restrictive fitness function shown in Fig. 4.4.D (light gray).

genotype. This means that in order for somatic genotypic exploration to be efficient, the population needs to be at a "bifurcation boundary" (Oster and Alberch 1982) between the neutral networks of two different phenotypes. Note that the more mutational paths towards the neutral network of the adaptive phenotype, the more efficient somatic genotypic exploration will be, since many different non-heritable mutations can stabilize the population at the boundary.

A biological example in which these conditions could have been met is the case of the *de novo* origin of antifreeze peptides in gadid fish living in the cold waters in the Arctic Sea (Baalsrud et al. 2018; Zhuang et al. 2019). The mutational steps that lead to the functionalization of this gene from a non-coding sequence are well characterised (Fig. 5.3), and they involve a

series of intermediate genotypes with a high likelihood of expressing the antifreeze peptide as a result of somatic or phenotypic mutations. Note, how the last step for the functionalization of the peptide simply involved a frameshift on a sequence of repeats on a microsatellite-like sequence. In theory, there are many possible mutational paths that could then lead to the production of the antifreeze peptide at that last step. If we also take into account how the mutation rates on microsatellites are very high, and also that an adult gadid fish can have several million somatic cells with the potential of producing non-heritable variation, then the probability that a biologically relevant amount of antifreeze peptide could be produced by intermediate genotypes could be high. If that is the case, then it is likely that the intermediate genotypes would have been selected by expressing an adaptive trait by means of non-heritable mutations, which might have lead to the eventual genetic assimilation of a sequence coding for the *de novo* gene.

(ii) The developmental map: The second condition is that the developmental map of the ancestral phenotype needs to be sufficiently non-linear for the non-heritable mutations to be able to modify it. In other words, the ancestral phenotype needs to be susceptible to whatever change the new non-heritable genotypes impose. If the developmental map of the ancestral phenotype is excessively robust, then the novel phenotype will not be able to be expressed. An example of a developmental map that is susceptible to somatic mutations is the case of the pattern formation of zebrafish that I mentioned in chapter 4. In the following section I will refer more broadly to another developmental map that might facilitate the action of somatic genotypic exploration, the case in which multi-level selection facilitates the expression of adaptive phenotypes.

5.2.1 The ecology of the embryo, or reintroducing the struggle of the parts

The results from chapter 4 show how somatic genotypic exploration is especially likely when there is multi-level selection. That is, when the mutation that confers a benefit to the organism is aligned with a positive effect on the mutant cell's fitness relative to its sisters. Whenever a mutation confers a proliferative advantage to a cell, that cell can increase in frequency within

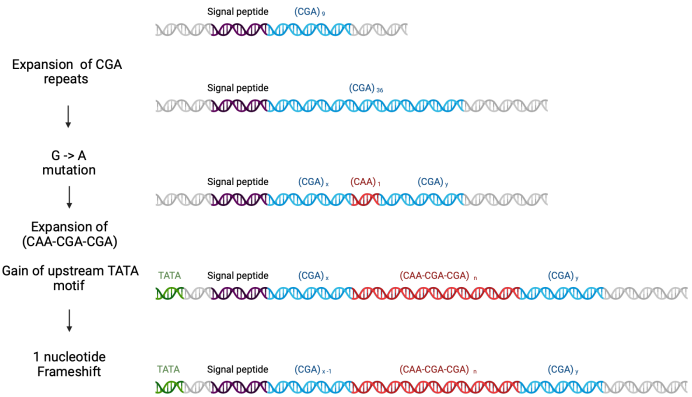


Figure 5.3: Evolutionary steps in the evolution of the antifreeze glycoprotein *de novo* in gadid fish Mutational steps that transformed a non-coding intergenic region of the genome of the ancestral gadid into a sequence coding for the antifreeze glycoporotein, as described by Zhuang et al (2019). The steps proposed are the following: (1) Ancestrally, there was a latent signal-peptide-coding exons associated to a Kozak motif (a signal for enhanced translation) a GCA repeat of 27 nucleotides. (2) The GCA repeat expanded resulting in four tandem copies of this 27 nucleotiede unit (108 nucleotides in total) repeats expanded. (3) A change from G to A formed the 9-nucleotides long coding coding sequence for the basic Thr-Ala-Ala. (4) Acquisition of a TATA-motif upstream of the whole sequence accompanied by a microsatellite expansions of the repeat unit coding for Thr-Ala-Ala. 5) A 1-nucleotide frameshift mutation between the signal peptide and the downstream sequence of repeats brought the signal peptide in frame with the (Thr-Ala-Ala) coding sequence.

the body, which increases the likelihood that the organism develops the adaptive phenotype enabled by the mutation in question. In such cases, somatic genotypic exploration can promote adaptation even when the number of cells needed to express the adaptive phenotype is large (Fig. 4.4). For this reason, the evolution of phenotypes that derive from increases in cell proliferation will be especially likely to be facilitated by somatic genotypic exploration.

The idea that competition between cells might promote evolution was proposed by the embryologist Wilhelm Roux just two decades after Darwin had published *On the Origin of Species*. In 1881, Roux published *The Struggle of the Parts in the Organism* (*Der Kampf der Theile im Organismus*). In this book he proposed that “the struggle of the molecules and the struggle of the cells produced a series of qualities which, in consequence of their general character, are

also extremely useful to the individual in his struggle for existence"[1]. In the case of cells, such a struggle, he claimed, was a struggle for space within a developing organism. Roux figured that the source of the struggle in the case of cells, was due to a differential efficacy in their uptake and consumption of resources for their nourishment that would affect their growth. He also noted that the developmental timepoint in which the struggle took place could differentially impact the process in the life of the individual[2]. Although Darwin himself judged Roux's book to be "the more important book on evolution, which has appeared for some time"[3], Roux's theory of the struggle of the parts did not have an impactful legacy. However, the recent advances in our capacity to understand the genetic and cellular dynamics during development and within adult organisms opens up the door to re-evaluate Roux's ideas and to ponder its connections with somatic genotypic exploration.

Modern sequencing technologies are uncovering an enormous genetic diversity existing within the soma of single individuals (see chapter 4), which supports Roux's notion that it is unlikely that the laws of inheritance would have defined every single detail of the behaviour of every cell of an individual beforehand[4]. This diversity makes the soma a genetic mosaic within which evolutionary processes can take place, such as the clonal expansions of certain genetic profiles. The advantage of a somatic mutant cell over its clonal sisters is particularly well studied in the context of cancers, where mutations in key genes can lead to a disruption of the control of cell proliferation, thus increasing the relative fitness of mutant cells. Another way by which a certain lineage of mutant cells can expand within an individual is via a more aggressive elimination or displacement of cells with alternative genotypes, in a process that has been termed cell competition[5]. Under this process a newly arriving mutant can clear its

1. "Wir sahen danach, dass der Kampf der Molekel und der Kampf der Zellen eine Reihe von Qualitätetn züchtete, welche in Folge ihres allgemeinenen Charakters auch dem Individuum in seinem Kampfe ums Dasein höchst nützlich sind." Roux (1881), page

2. "Die Wirkungsgrösse des Kampfes der Zellen ist bedingt durch die Zahl von Zellgenerationen, in welchen er zur Wirkung gelangt, und dies ist natürlich abhängig von dem Zeitpunkte des Auftrtetens der neuen Eigenschaft im Leben des Individuums." Ibid. Page 95

3. As cited by Heams (2012).

4. "Durch die Ungleichartigkeiten, welche durch den Wechsel der Bedingungen fortwährend nicht blos an den Ganzen, sondern auch an den Theilen hervorgebracht werden, war es von vornherein unmöglich, dass Vererbungsgesetzte such ausbilden konnten, welche das Einzelgeschehen bis in die letzte Zelle und das letzte Molekel von vornherein normirten." Ibid. Page 71

5. Although in the literature "cell competition" is often used to refer to particular cases in which there are active mechanisms to inhibit the proliferation of cells with alternative genotypes, I will here use this phrase to

surrounding of wild type cells making space for its own expansion, or alternatively wild type cells can eliminate arising mutant neighbours (Morata and Ripoll 1975; Simpson 1979; Kim and Jain 2020; Tseng et al. 2022).

Whether a somatic mutation increases a cell's fitness will depend on where in the genome the mutation happens, but also on the *embryological ecological context* of that mutated cell. The embryo can be considered to be an ecosystem in which cells in a population are consuming resources for their maintenance, growth and division. In well integrated developmental systems such as multicelluar animals, signaling mechanisms coordinate the behaviour of cells to orchestrate a coherent morphogenesis and physiology. The local environment of each cell in the system will include its cellular neighbours and different concentrations of signals and other molecules that will influence that cell's behaviour, physiology and molecular activity. However, a cell can better evolve the capacity to respond to a signal or to uptake a resource (or to not do either) within its embryological or tissue context. If it does then that cell can increase (or decrease) its fitness and expand clonally. This would not necessarily mean that it would lead to a malignant tumour, given that another signal for which those clonally expanding cells are not mutated could put a stop to that lineage's expansion and the body can regain control of its proliferation.

An evolutionary consequence of competition between cells during development is that it can bias inheritance. Mathematical models have shown how selection at the level of cells during organismal development can bias against the arrival of deleterious variants at the level of populations because deficient cells would not be significantly represented in the germline (Hastings 1989). In support of this idea, it has been show empirically that, in *Drosophila*, competition at the cellular level can indeed bias the allelic representation in gametes and in the progeny (Extavour and García-Bellido 2001; Tseng et al. 2022). If this is a prevalent mechanism, then it is likely that the germline-soma separation plays an important role in the existing genetic diversity within a population, and that it might promote the evolution of genomes that are fitter relative to other genomes. I believe this process may underlie some mutational bi-

more generally also include the more classical ecological sense of "competition" referring to the direct or indirect competition for resources.

ases that have recently been presented. Monroe et al (2022) found an under-representation of mutations on different genes with different degrees of "essentiality" in mutation accumulation lines of *Arabidopsis thaliana*. The authors claimed this result was evidence for a directed effect of mutations, in which they do not occur at random, but that they are biased towards avoiding deleterious consequences. They suggested that this would be the result of a decreased mutation rate consequence of a presumed repair mechanism directed by epigenetic marks indicative of gene functionality. However, I argue that the same patterns could arise from the more parsimonious explanation of intra-organismal cell competition, because the cell whose essential genes mutate in a developmental context in the organism is unlikely to proliferate and make it to the gametes of the organism. It would be interesting to reconsider these data exploring the effect of cell competition in development. Selection at the level of cells will not only purge deleterious variants, but it can also lead to an over-representation of variants that increase cell fitness in the germline, thus positively biasing inheritance. This was also shown mathematically by Otto and Hastings (1998), who noted that cell-level selection can be aligned with individual-level selection, and that "when selection at the cell and individual levels act in a cooperative manner, increased rather than decreased opportunity for germline selection will be favored by evolution".

Under somatic genotypic exploration, competition at the cell level can also promote the evolution of adaptive traits when there is a congruence in selection at the cell and individual levels. The evolution of many adaptive morphological traits are the result of general or local changes in growth relative to an ancestral state. Think of extreme examples such as the neck of a giraffe, or of a sauropod, or the disproportionate digits of the aye-aye, pterosaurs and bats, or the elongated tusks of narwhals, elephants and walruses. Traits such as those that represent changes in the shape and size of certain body parts are at the heart of classical morphological and palaeontological studies. These morphogeometric changes are usually the result of localized changes in developmental parameters. Alberch et al (1979) produced a model to describe the evolution of size and shape in terms of a dynamical function for development defined by parameters such as growth rate (which includes cell proliferation), and the onset and offset

signals for growth during an ontogenic trajectory of a structure or organ. Changes in these parameters can result in different heterochronic processes during development. For example, an increase in the offset of the signal would result in hypermorphosis, since the structure in question would have a longer time to grow, while increases in cell proliferation can lead to the acceleration of a developmental process. If a somatic mutation would lead to variation in one of these parameters, it could therefore affect the final form and impact the fitness of the individual. As was pointed out by D'Arcy Thompson (1942):

> "An organism is so complex a thing, and growth so complex a phenomenon, that for growth to be so uniform and constant in all the parts as to keep the whole shape unchanged would indeed be an unlikely and an unusual circumstance. Rates vary, proportions change, and the whole configuration alters accordingly."

And if rates change within the organism because of the non-heritable fluctuation of cell activities, the whole configuration might also change accordingly.

To illustrate the basic idea of how somatic mutation could work by means of cell competition and multilevel selection, consider an adaptive trait that results from an increased proliferation of cells. Among Neotropical leaf-nosed bats (Phyllostomidae) there is a great diversity of specialized diets. Some of these bats specialized in consuming insects, while others specialized in feeding on fruits, blood or nectar (Potter et al. 2021). Depending on the diet, different groups within this family evolved an array of adaptations, at the molecular (Potter et al. 2021) and morphological (Camacho et al. 2020) levels. Among such adaptations is skull morphology. Nectar feeding bats, such as Pallas's long-tongued bat (*Glossophaga soricina*) have a characteristic long snout that allows them to lick the nectar off the flowers, while the Jamaican fruit bat (*Artibeus jamaicensis*) has a short and wide face (Fig. 5.4.A). Camacho et al (2020) studied the ontogenic causes for the difference in cranial morphology of these species and they uncovered how those differences are the result of heterochronic changes during development. Even though snouts are shorter in fruit-eating bats, the shorter face is a consequence of an increase in cell proliferation within a cluster of cells that corresponds to a module responsible for the development of the adult midface. The reason why an increase in cell divisions re-

sults in a shorter face is because it causes terminal cell divisions to occur earlier, thus limiting the time for a greater proliferation of these cells. Linking back to the question of how somatic mutations can promote adaptation, it is conceivable that, somatic mutations in these specific developmental modules could lead to the development of intermediate phenotypes of snout length whenever they these mutations lead to genotypes that that increase the proliferation of cells, as it happened on the lineage of the Jamaican fruit bat (Fig. 5.4.B). If such somatic mutations are frequent enough, there would be a distribution of snout lengths within populations that would enable natural selection to favour genotypes that are mutationally closer to the short face-producing genotypes, thus pushing morphological evolution towards shorter faces as bats evolve further into their dietary specializations.

Note that the case of the length of bat snouts is an example of a continuous phenotype (a few somatic mutant cells could result in a snout that is a bit shorter, but more cells could result in a snout that is a bit larger), but additional examples of how somatic mutations increasing cell proliferation can affect phenotypic development could come also from discrete phenotypic states. For example, when sequential developmental steps, if the timely development of a structure precludes the subsequent development of other structures. Alberch et al (1979) pointed out how salamanders that achieve sexual maturity before the complete sequential development of cranial structures would lack the maxillary bones of the skull, and they also speculated about how continuous changes in the proliferation rate of cells of the apical ectodermal ridge of bird limbs could lead to a limbless state, a normal state or polydactylous state (Fig. 5.4.C). Therefore, somatic mutations modifying rates of cell proliferation could have a cascading effect that leads to major phenotypic changes, which could include the complete absence of a morphological structure.

Under this account, those mutations that cause increases in fitness to the cells will cause their own potential to arise again in the germline if they have an impact on the development of the organism that is adaptive. I have here focused on the case in which somatic expansions can impact phenotypic development, but other developmental processes could also reveal adaptive phenotypes via non-heritable mutations. One example is the case of the zebrafish pattern

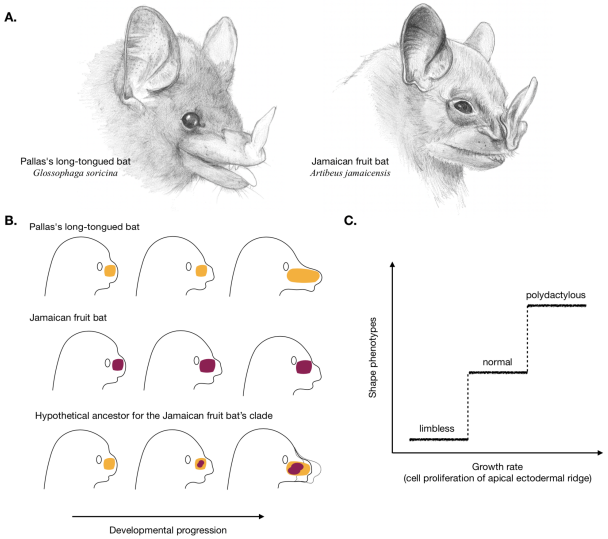


Figure 5.4: Potential phenotypic impact of somatic mutations that are beneficial at the cell level. (A) Illustration of the two species of bats referred to in the main text. Notice the differences in facial proportions. (B) Differential growth in different developmental stages can lead to changes in morphological proportions. The illustrations represent a side view of the head of the embryos of the two bats shown in (A). An early increase in cell proliferation in cells with a purple genotype as in the Jamaican fruit bat leads to a shorter snout. A slower cell proliferation in cells with an orange genotype in the middle stage of development allows for later terminal divisions, producing longer snouts as in the case of Pallas's long-tongued bat. Somatic mutations from orange to purple genotypes would increase the local rates of proliferation, possibly affecting the final facial morphology and leading to an intermediate phenotype between both bats. (C) Hypothetical relationship between the development of different wing shapes and the cell proliferation at the apical ectodermal ridge of a developing bird. Adapted from Alberch et al (1979).

discussed in chapter 4, another could be a scenario in which a somatic cell could evolve the capacity to respond to an attracting or repelling signal which can lead the cell to migrate towards a new embryonic niche where it can ground the development of a new adaptive structure or the expansion of an old one (a beneficial metastasis, we could say).

Overall, it is therefore possible that populations can explore a spectrum of phenotypic possibilities based on non-heritable mutations, and that multi-level selection might favour the development and evolution of adaptive traits. This opens up the intriguing question about

how much of morphological changes in evolutionary time can be the result of evolutionary processes within the ecological background of developmental systems. This is a view of development in which development is not simply the unfolding of a pre-existing plan, it is a vision of development in which each developmental step matters to accomplish the final form. At each step a novelty, genetic or otherwise, can push the evolution of shapes in one or another direction. The developmental process of each individual thus has the potential to influence the evolutionary fate of populations.

5.3 Concluding remarks, or the creative paths of memory

Chance and necessity have been two pillars of evolutionary thought since Darwin suggested that whenever variation arises, it is blind to its adaptive value, and that such value is only proven a posteriori, when selection filters out the favourable traits from the rest, thus guiding the evolution of forms. Having a reductionist view of biological evolution driven by chance and necessity has been extremely useful to conceptualise fundamental evolutionary principles. However, in order to understand the reality of material biology and the directions of its evolution it is fundamental to also understand the context in which chance and necessity act. This context is nothing other than the evolutionary history of organisms and their individual development in their specific time and place.

Even when the occurrence of a mutation might indeed be a chance event, it is an event that occurs in a pre-existing substrate, which is the genomic memory of a lineage. An organism's genome is the result of a long evolutionary process that has imprinted on that genome rules for the development of phenotypes. Sometimes these rules are quite precise, and not all mutations will be compatible with them in a way that they can be integrated functionally into the genome. The developmental processes orchestrated by the genome can evolve a certain robustness to impede chance mutations from disrupting its results. At the same time, the pre-existing developmental process has the potential to facilitate the propagation of mutations that can actually integrate with the functions that were in place, thus offering more flexibility to the development of some features over others. Throughout the history of life, features have

evolved that have affected how these developmental process work, with the potential consequence of redefining which are the mutations that will have a phenotypic impact and which ones will not. The origins of enhancers and of the germline-soma separation that I have discussed in this thesis are examples of this.

The evolutionary history of a lineage that has put in place the guidelines for development encoded in each genome thus defines what are the potential evolutionary trajectories ahead. This echoes a poetic idea of Henri Bergson by which an internal force pushes forward creatively in each lineage as it evolves. Bergson had to explain his ideas from a speculative and somewhat metaphysical angle given the tools of his time, but our contemporary understanding of development and molecular biology can now inform us as to how lineages can follow their internal tendencies. But these tendencies, which are incarnated in the genome of each organism, only offers guidelines, they are not a deterministic force. Another important aspect to consider is how each individual within a lineage manages to make use of the genomic memory it has inherited for its own development, and what part such usage has to play in defining a lineage's evolutionary trajectory.

Leowntin envisioned the organism as an object and subject of its own evolution from the perspective of how organisms can influence their own development, and also because they can define their own niche and their own ecological necessities. He writes:

> "The organism, irrespective of the internal and external forces that influence it, enters directly into the determination of its own future. The view of development that sees genes as determinative or even a view that admits interactions between gene and environment as determining the organism, places the organism as the end point, the object, of forces. The arrows of causation point from gene and environment to organism. In fact, however, the organism participates in its own development because the outcome of each developmental step is a precondition to the next. But the organism also participates in its own development because (...) it is the determinant of its own milieu." (Lewontin 1983)

The more complex the genotype-phenotype map, the more degrees of freedom an organ-

ism has for exploring the potentials of their own development. The non-linearity of complex developmental systems (Alberch 1991), such as those involving distal-acting regulatory elements and coordinated morphogenetic processes, further encouraged the lineage of animals to participate in their own development and to explore the most precious diversity of forms.

In sum, life pushes forward in time, and as it does it evolves its own rules, its own constraints and its own potentials. The living forms that exist today are not exclusively the product of chance and necessity, but over evolutionary time they have acquired the capacity to tame that chance and to define their own necessities, and, in so doing, it has permitted each lineage to define its own trajectories.

6 Acknowledgements

In 1981, the magazine *Cabildo* published a letter entitled “Borges does not exist”. In that letter it was claimed that Jorge Luis Borges was a literary ghost created by a pool of writers, and that the stuttering blind author was embodied by a second-rank Uruguayan author by the name of Aquiles Rosendo Scatamacchia. In an interview Borges offered to the Spanish newspaper *El País* from his flat facing the San Martín Square in Buenos Aires, he was asked about this letter. Borges, with comical insight, replied that it may indeed be true that he does not exist: “I am not sure that I exist, actually. I am, after all, all the authors I have read, all the people I have met, all the women I loved. All the cities I have visited, all my ancestors... But you can, actually, say that I have told you that I am not Uruguayan, nor an author, even if I am not sure I exist”.

Unlike Borges, I am Uruguayan, but like him, I admit that the stream of my life is, in great measure, a unique riverine confluence of many other life streams. In the uncertainty of my own existence, I also sometimes think of myself as a mosaic of the people I met and the experiences I lived. My probable existence, which now writes these closing words for this dissertation, calls me to name a few of the fundamental shards of my mosaic. I will start with Josh. Although it was sometimes challenging to find common ground in our interests due to my bias for zoological curiosities and his bias for abstract landscapes, when our interests did agree beautiful projects came to fruition. Our relationship went much farther than that of a simple mentorship and it developed into a proper friendship. This is exactly the kind of student-supervisor relationship that I would take as an inspiration if I ever come to become a supervisor myself.

I also want to mention members of the family I developed here in Switzerland. Alejandro offered me the best conversations I had and helped me mould a more joyful version of myself. Cauã’s wit and charisma continues to offer an endless source of laughter. The spontaneously

forming bond that I have with Epi still glitters brightly in spite of our countless quarrels. While Ana has dazzled me with an original and stoic nature that never cease to surprise and inspire me. Artemis has set the foundations for this place to feel like a home, and, because I know she likes appraisals, I will just say that the goddess should be humbled by this friend of mine carrying her name. Beeeeeerit is probably one of the people who best understood the winds of my moods and was always there to help me raise my mainsail (I am happy she didn't get scared by the brutal wildlife videos I used to send her). Fokko is a fellow air-sign clown laughing at the absurdity of this circus. Judith still holds the title of my favourite person (nobody ever has sacrificed so much for my dental hygiene). Quokka has been a curious man that made me realise what truly makes friends invaluable while being an icon of authenticity. Yagmur was my popcorn-loving flatmate, friend, colleague and Swisster. Of course, the list goes on with the likes of Fabienne, Louis, Cheyenne, Jonas, Cecilia, Inez, Damián, Alberto, Giaccomo, Fidel, Bharat, Alessia, Mark, Julian, Elena, Nico, Giulia, Jana and other people that I surely love and that I hope they never read this far to realize I forgot about them when writing this. There is also a special place for people that were fundamental in the workplace, such as Dave, Pete, Marco, Sonja, Francesca, Jessica, María, and, of course, Rita. Among the group-family that Josh managed to put together are Magda, Malvika and Hana, who are amazing, kind, funny and insanely smart, and Alex, who got here later but early enough to see me despair on the day of the deadline for this thesis.

I would also want to pay homage to Daniel Constanda and Keijão Kunigami, who in our endless chats in Tokyo, Rio, Uppsala or Taipei helped me steer my intellectual development to where they are now. But it would not be fair not to name also other shards from the Japanese face of my mosaic. Dieguito, Christine, Michael, Guillaume, Hasegawa, Aiko, Nazim, Mehdi, Marina, Akiko and the fundamental Kaida-san. I want to especially mention the vital meetings with the Kolabolabos, Svetlin, Erica and Edison, not only in Tokyo but also during the pandemic, and, of course, the Japoguayas del Bar Michigan, Ana-san and Tyana-san.

I'll finish this with a linguistic somersault. Ya acabadas las menciones de algunos miembros de la familia extendida que fui tejiendo en mi Tokio y en mi Zurich, vuelvo al sur a reconocer a

la familia nuclear. Porque, ¿qué sería yo sin la semilla de la locura que plantaron mis padres y mi hermanita, sin esa anarquía estética, sin ese cariño incondicional, sin esa irreverente originalidad, sin las asperezas de esos algodones? ¿Qué sería yo sin las milanesas de Mary, sin las maquinitas del Cacho, sin los cigarros de Nair o los timbrazos de Vinko? ¿Qué seria yo sin las fiestas en el Cerro, sin las cataratas, los glaciares y los cruceros? Y en verdad, ¿qué seria yo sin las historias de María, sin las clases de Hachiuma, sin los ecos de voces y risas rebotando en el mármol y granito de las paredes del Seminario? Y ya que estamos de nuevo en el Río de la Plata, volviendo a Borges, ¿qué seria yo sin la Plaza San Martín? En frente a esa Plaza, intentaba culminar el periodista su entrevista volviéndole a preguntar a Borges si existía, pero en tono de cierre Borges lo interrumpía con un "nada, nada, amigo mío; lo que le he dicho: no estoy seguro de nada, no sé nada. Imagínese que ni siquiera sé la fecha de mi muerte...".

Bibliography

Abascal, Federico, Luke MR Harvey, Emily Mitchell, Andrew RJ Lawson, Stefanie V Lensing, Peter Ellis, Andrew JC Russell, Raul E Alcantara, Adrian Baez-Ortega, Yichen Wang, et al. 2021. "Somatic mutation landscapes at single-molecule resolution." *Nature* 593 (7859): 405–410.

Abrusán, György. 2013. "Integration of new genes into cellular networks, and their structural maturation." *Genetics* 195 (4): 1407–1417.

Agresti, Alan. 1974. "Bounds on the extinction time distribution of a branching process." *Advances in Applied Probability* 6 (2): 322–335.

Alberch, Pere. 1991. "From genes to phenotype: dynamical systems and evolvability." *Genetica* 84 (1): 5–11.

Alberch, Pere, Stephen Jay Gould, George F Oster, and David B Wake. 1979. "Size and shape in ontogeny and phylogeny." *Paleobiology* 5 (3): 296–317.

Aldana, Maximino, Enrique Balleza, Stuart Kauffman, and Osbaldo Resendiz. 2007. "Robustness and evolvability in genetic regulatory networks." *Journal of theoretical biology* 245 (3): 433–448.

Altland, Alexander, Andrej Fischer, Joachim Krug, and Ivan G Szendro. 2011. "Rare events in population genetics: stochastic tunneling in a two-locus model with recombination." *Physical review letters* 106 (8): 088101.

Ancel, Lauren W, and Walter Fontana. 2000. "Plasticity, evolvability, and modularity in RNA." *Journal of Experimental Zoology* 288 (3): 242–283.

Andersson, Robin, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, Xiaobei Zhao, Christian Schmidl, Takahiro Suzuki, et al. 2014. "An atlas of active enhancers across human cell types and tissues." *Nature* 507 (7493): 455–461.

Antolin, Michael F, and Curtis Strobeck. 1985. "The population genetics of somatic mutation in plants." *The American Naturalist* 126 (1): 52–62.

Arthur, Wallace. 2001. "Developmental drive: an important determinant of the direction of phenotypic evolution." *Evolution & development* 3 (4): 271–278.

———. 2004. "The effect of development on the direction of evolution: Toward a twenty-first century consensus." *Evolution & development* 6 (4): 282–288.

Ashcroft, Peter, Franziska Michor, and Tobias Galla. 2015. "Stochastic tunneling and metastable states during the somatic evolution of cancer." *Genetics* 199 (4): 1213–1228.

Atlasi, Yaser, and Hendrik G Stunnenberg. 2017. "The interplay of epigenetic marks during stem cell differentiation and development." *Nature Reviews Genetics* 18 (11): 643–658.

Azizan, Elena AB, Hanne Poulsen, Petronel Tuluc, Junhua Zhou, Michael V Clausen, Andreas Lieb, Carmela Maniero, Sumedha Garg, Elena G Bochukova, Wanfeng Zhao, et al. 2013. "Somatic mutations in ATP1A1 and CACNA1D underlie a common subtype of adrenal hypertension." *Nature genetics* 45 (9): 1055–1060.

Baalsrud, Helle Tessand, Ole Kristian Tørresen, Monica Hongrø Solbakken, Walter Salzburger, Reinhold Hanel, Kjetill S Jakobsen, and Sissel Jentoft. 2018. "De novo gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data." *Molecular biology and evolution* 35 (3): 593–606.

Baldwin, J Mark. 1896. "A new factor in evolution (continued)." *The American Naturalist* 30 (355): 536–553.

Bar, Daniel Z, Martin F Arlt, Joan F Brazier, Wendy E Norris, Susan E Campbell, Peter Chines, Delphine Larrieu, Stephen P Jackson, Francis S Collins, Thomas W Glover, et al. 2017. "A novel somatic mutation achieves partial rescue in a child with Hutchinson-Gilford progeria syndrome." *Journal of medical genetics* 54 (3): 212–216.

Barrett, Rowan DH, Stefan Laurent, Ricardo Mallarino, Susanne P Pfeifer, Charles CY Xu, Matthieu Foll, Kazumasa Wakamatsu, Jonathan S Duke-Cohan, Jeffrey D Jensen, and Hopi E Hoekstra. 2019. "Linking a mutation to survival in wild mice." *Science* 363 (6426): 499–504.

Begun, David J, Heather A Lindfors, Andrew D Kern, and Corbin D Jones. 2007. "Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade." *Genetics* 176 (2): 1131–1137.

Behjati, Sam, Meritxell Huch, Ruben van Boxtel, Wouter Karthaus, David C Wedge, Asif U Tamuri, Iñigo Martincorena, Mia Petljak, Ludmil B Alexandrov, Gunes Gundem, et al. 2014. "Genome sequencing of normal cells reveals developmental lineages and mutational processes." *Nature* 513 (7518): 422–425.

Béïque, Jean-Claude, Mays Imad, Ljiljana Mladenovic, Jay A Gingrich, and Rodrigo Andrade. 2007. "Mechanism of the 5-hydroxytryptamine 2A receptor-mediated facilitation of synaptic activity in prefrontal cortex." *Proceedings of the National Academy of Sciences* 104 (23): 9870–9875.

Bergman, Aviv, and Mark L Siegal. 2003. "Evolutionary capacitance as a general feature of complex gene networks." *Nature* 424 (6948): 549–552.

Bergson, Henri. 1907. *L'évolution créatrice*. Bibliothèque de philosophie contemporaine F. Alcan, Paris.

Berthelot, Camille, Diego Villar, Julie E Horvath, Duncan T Odom, and Paul Flicek. 2018. "Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression." *Nature ecology & evolution* 2 (1): 152–163.

Besnard, Fabrice, Joao Picao-Osorio, Clément Dubois, and Marie-Anne Félix. 2020. "A broad mutational target explains a fast rate of phenotypic evolution." *Elife* 9:e54928.

Bitbol, Anne-Florence, and David J Schwab. 2014. "Quantifying the role of population subdivision in evolution on rugged fitness landscapes." *PLoS computational biology* 10 (8): e1003778.

Blokzijl, Francis, Joep De Ligt, Myrthe Jager, Valentina Sasselli, Sophie Roerink, Nobuo Sasaki, Meritxell Huch, Sander Boymans, Ewart Kuijk, Pjotr Prins, et al. 2016. "Tissue-specific mutation accumulation in human adult stem cells during life." *Nature* 538 (7624): 260–264.

Blumer, Moritz, Tom Brown, Mariella Bontempo Freitas, Ana Luiza Destro, Juraci A. Oliveira, Ariadna E. Morales, Tilman Schell, et al. 2022. “Gene losses in the common vampire bat illuminate molecular adaptations to blood feeding.” *Science Advances* 8 (12).

Bratulic, Sinisa, Macarena Toll-Riera, and Andreas Wagner. 2017. “Mistranslation can enhance fitness through purging of deleterious mutations.” *Nature communications* 8 (1): 1–9.

Britten, Roy J, and Eric H Davidson. 1969. “Gene Regulation for Higher Cells: A Theory: New facts regarding the organization of the genome provide clues to the nature of gene regulation.” *Science* 165 (3891): 349–357.

Buss, Leo W. 1983a. “Evolution, development, and the units of selection.” *Proceedings of the National Academy of Sciences* 80 (5): 1387–1391.

———. 1982. “Somatic cell parasitism and the evolution of somatic tissue compatibility.” *Proceedings of the National Academy of Sciences* 79 (17): 5337–5341.

———. 1983b. “Somatic variation and evolution.” *Paleobiology* 9 (1): 12–16.

Bylino, Oleg V, Airat N Ibragimov, and Yulii V Shidlovskii. 2020. “Evolution of regulated transcription.” *Cells* 9 (7): 1675.

Cagan, Alex, Adrian Baez-Ortega, Natalia Brzozowska, Federico Abascal, Tim HH Coorens, Mathijs A Sanders, Andrew RJ Lawson, Luke MR Harvey, Shriram Bhosle, David Jones, et al. 2022. “Somatic mutation rates scale with lifespan across mammals.” *Nature:* 1–8.

Cai, Jing, Ruoping Zhao, Huifeng Jiang, and Wen Wang. 2008. “De novo origination of a new protein-coding gene in Saccharomyces cerevisiae.” *Genetics* 179 (1): 487–496.

Cairns, John. 1975. “Mutation selection and the natural history of cancer.” *Nature* 255 (5505): 197–200.

Calhoun, Vincent C, and Michael Levine. 2003. “Long-range enhancer–promoter interactions in the Scr-Antp interval of the Drosophila Antennapedia complex.” *Proceedings of the National Academy of Sciences* 100 (17): 9878–9883.

Camacho, Jasmin, Rachel Moon, Samantha K Smith, Jacky D Lin, Charles Randolph, John J Rasweiler, Richard R Behringer, and Arhat Abzhanov. 2020. “Differential cellular proliferation underlies heterochronic generation of cranial diversity in phyllostomid bats.” *EvoDevo* 11 (1): 1–17.

Cannataro, Vincent L, Stephen G Gaffney, and Jeffrey P Townsend. 2018. “Effect sizes of somatic mutations in cancer.” *JNCI: Journal of the National Cancer Institute* 110 (11): 1171–1177.

Cannavò, Enrico, Pierre Khoueiry, David A Garfield, Paul Geeleher, Thomas Zichner, E Hilary Gustafson, Lucia Ciglar, Jan O Korbel, and Eileen EM Furlong. 2016. “Shadow enhancers are pervasive features of developmental regulatory networks.” *Current Biology* 26 (1): 38–51.

Cano, Alejandro V, and Joshua L Payne. 2020. “Mutation bias interacts with composition bias to influence adaptive evolution.” *PLoS computational biology* 16 (9): e1008296.

Cano, Alejandro V, Hana Rozhoňová, Arlin Stoltzfus, David M McCandlish, and Joshua L Payne. 2022. “Mutation bias shapes the spectrum of adaptive substitutions.” *Proceedings of the National Academy of Sciences* 119 (7): e2119720119.

Caporale, Lynn Helena. 2000. “Mutation is modulated: implications for evolution.” *Bioessays* 22 (4): 388–395.

Capra, John A, Katherine S Pollard, and Mona Singh. 2010. “Novel genes exhibit distinct patterns of function acquisition and network integration.” *Genome biology* 11 (12): 1–16.

Carelli, Francesco N, Angélica Liechti, Jean Halbert, Maria Warnefors, and Henrik Kaessmann. 2018. “Repurposing of promoters and enhancers during mammalian evolution.” *Nature communications* 9 (1): 1–11.

Carroll, Sean B. 2001. “Chance and necessity: the evolution of morphological complexity and diversity.” *Nature* 409 (6823): 1102–1109.

———. 2008. “Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution.” *Cell* 134 (1): 25–36.

———. 1995. “Homeotic genes and the evolution of arthropods and chordates.” *Nature* 376 (6540): 479–485.

Carvunis, Anne-Ruxandra, Thomas Rolland, Ilan Wapinski, Michael A Calderwood, Muhammed A Yildirim, Nicolas Simonis, Benoit Charloteaux, César A Hidalgo, Justin Barbette, Balaji Santhanam, et al. 2012. “Proto-genes and de novo gene birth.” *Nature* 487 (7407): 370–374.

Cases, Ildefonso, and Victor de Lorenzo. 1998. “Expression systems and physiological control of promoter activity in bacteria.” *Current opinion in microbiology* 1 (3): 303–310.

Chen, Jian-Qun, Ying Wu, Haiwang Yang, Joy Bergelson, Martin Kreitman, and Dacheng Tian. 2009. “Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria.” *Molecular biology and evolution* 26 (7): 1523–1531.

Christiansen, Freddy B, Sarah P Otto, Aviv Bergman, and Marcus W Feldman. 1998. “Waiting with and without recombination: the time to production of a double mutant.” *Theoretical population biology* 53 (3): 199–215.

Ciliberti, Stefano, Olivier C Martin, and Andreas Wagner. 2007. “Innovation and robustness in complex regulatory gene networks.” *Proceedings of the National Academy of Sciences* 104 (34): 13591–13596.

Clark, Michael B, Paulo P Amaral, Felix J Schlesinger, Marcel E Dinger, Ryan J Taft, John L Rinn, Chris P Ponting, Peter F Stadler, Kevin V Morris, Antonin Morillon, et al. 2011. “The reality of pervasive transcription.” *PLoS biology* 9 (7): e1000625.

Clark, Richard M, Tina Nussbaum Wagler, Pablo Quijada, and John Doebley. 2006. “A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture.” *Nature genetics* 38 (5): 594–597.

Colbran, Laura L, Ling Chen, and John A Capra. 2019. “Sequence characteristics distinguish transcribed enhancers from promoters and predict their breadth of activity.” *Genetics* 211 (4): 1205–1217.

Colom, Bartomeu, Maria P Alcolea, Gabriel Piedrafita, Michael WJ Hall, Agnieszka Wabik, Stefan C Dentro, Joanna C Fowler, Albert Herms, Charlotte King, Swee Hoe Ong, et al. 2020. “Spatial competition shapes the dynamic mutational landscape of normal esophageal epithelium.” *Nature genetics* 52 (6): 604–614.

Conlon, Ian, and Martin Raff. 1999. “Size control in animal development.” *Cell* 96 (2): 235–244.

Consortium, Tabula Muris. 2018. “Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris.” *Nature* 562 (7727): 367–372.

Core, Leighton J, André L Martins, Charles G Danko, Colin T Waters, Adam Siepel, and John T Lis. 2014. “Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers.” *Nature genetics* 46 (12): 1311–1320.

Cosby, Rachel L, Julius Judd, Ruiling Zhang, Alan Zhong, Nathaniel Garry, Ellen J Pritham, and Cédric Feschotte. 2021. “Recurrent evolution of vertebrate transcription factors by transposase capture.” *Science* 371 (6531): eabc6405.

Cotney, Justin, Jing Leng, Jun Yin, Steven K Reilly, Laura E DeMare, Deena Emera, Albert E Ayoub, Pasko Rakic, and James P Noonan. 2013. “The evolution of lineage-specific regulatory activities in the human embryonic limb.” *Cell* 154 (1): 185–196.

Cowperthwaite, Matthew C, Evan P Economo, William R Harcombe, Eric L Miller, and Lauren Ancel Meyers. 2008. “The ascent of the abundant: how mutational networks constrain evolution.” *PLoS computational biology* 4 (7): e1000110.

Creyghton, Menno P, Albert W Cheng, G Grant Welstead, Tristan Kooistra, Bryce W Carey, Eveline J Steine, Jacob Hanna, Michael A Lodato, Garrett M Frampton, Phillip A Sharp, et al. 2010. “Histone H3K27ac separates active from poised enhancers and predicts developmental state.” *Proceedings of the National Academy of Sciences* 107 (50): 21931–21936.

Crick, Francis, and James Watson. 1953. “A structure for deoxyribose nucleic acid.” *Nature* 171 (737-738): 3.

Crispo, Erika. 2007. “The Baldwin effect and genetic assimilation: revisiting two mechanisms of evolutionary change mediated by phenotypic plasticity.” *Evolution: International Journal of Organic Evolution* 61 (11): 2469–2479.

Crocker, Justin, Namiko Abe, Lucrezia Rinaldi, Alistair P McGregor, Nicolás Frankel, Shu Wang, Ahmad Alsawadi, Philippe Valenti, Serge Plaza, François Payre, et al. 2015. “Low affinity binding site clusters confer hox specificity and regulatory robustness.” *Cell* 160 (1-2): 191–203.

Crombach, Anton, and Paulien Hogeweg. 2008. “Evolution of evolvability in gene regulatory networks.” *PLoS computational biology* 4 (7): e1000112.

Crombach, Anton, Karl R Wotton, Eva Jiménez-Guri, and Johannes Jaeger. 2016. “Gap gene regulatory dynamics evolve along a genotype network.” *Molecular biology and evolution* 33 (5): 1293–1307.

Cruzan, Mitchell B, Matthew A Streisfeld, and Jaime A Schwoch. 2020. “Fitness effects of somatic mutations accumulating during vegetative growth.” *BioRxiv:* 392175.

Curtis, Caitlin, Craig D Millar, and David M Lambert. 2018. “The Sacred Ibis debate: The first test of evolution.” *PLoS biology* 16 (9): e2005558.

Cusanovich, Darren A, Andrew J Hill, Delasa Aghamirzaie, Riza M Daza, Hannah A Pliner, Joel B Berletch, Galina N Filippova, Xingfan Huang, Lena Christiansen, William S DeWitt, et al. 2018a. “A single-cell atlas of in vivo mammalian chromatin accessibility.” *Cell* 174 (5): 1309–1324.

Cusanovich, Darren A, James P Reddington, David A Garfield, Riza M Daza, Delasa Aghamirzaie, Raquel Marco-Ferreres, Hannah A Pliner, Lena Christiansen, Xiaojie Qiu, Frank J Steemers, et al. 2018b. "The cis-regulatory dynamics of embryonic development at single-cell resolution." *Nature* 555 (7697): 538–542.

Cuvier, Georges. 1804. *Mémoire sur l'ibis des anciens Egyptiens.* Frères Lavrault.

Dalal, Chiraj K, and Alexander D Johnson. 2017. "How transcription circuits explore alternative architectures while maintaining overall circuit output." *Genes & Development* 31 (14): 1397–1405.

Dalby, M, S Rennie, and R Andersson. 2018. "FANTOM5 transcribed enhancers in mm10." *Zenodo. doi* 10.

Danko, Charles G, Lauren A Choate, Brooke A Marks, Edward J Rice, Zhong Wang, Tinyi Chu, Andre L Martins, Noah Dukler, Scott A Coonrod, Elia D Tait Wojno, et al. 2018. "Dynamic evolution of regulatory element ensembles in primate CD4+ T cells." *Nature Ecology & Evolution* 2 (3): 537–548.

Darwin, Charles. 1868. *The variation of animals and plants under domestication*. Vol. 2. J. murray.

Davidson, Eric H, and Michael S Levine. 2008. "Properties of developmental gene regulatory networks." *Proceedings of the National Academy of Sciences* 105 (51): 20063–20066.

Davis, Carrie A, Benjamin C Hitz, Cricket A Sloan, Esther T Chan, Jean M Davidson, Idan Gabdank, Jason A Hilton, Kriti Jain, Ulugbek K Baymuradov, Aditi K Narayanan, et al. 2018. "The Encyclopedia of DNA elements (ENCODE): data portal update." *Nucleic acids research* 46 (D1): D794–D801.

Dawkins, Richard. 1986. *The Blind Watchmaker*. W. Norton & Co., New York.

———. 1982. *The extended phenotype: The gene as the unit of selection*. Oxford, W.H. Freeman.

De, Subhajyoti. 2011. "Somatic mosaicism in healthy human tissues." *Trends in Genetics* 27 (6): 217–223.

De Laat, Wouter, and Denis Duboule. 2013. "Topology of mammalian developmental enhancers and their regulatory landscapes." *Nature* 502 (7472): 499–506.

De Santa, Francesca, Iros Barozzi, Flore Mietton, Serena Ghisletti, Sara Polletti, Betsabeh Khoramian Tusi, Heiko Muller, Jiannis Ragoussis, Chia-Lin Wei, and Gioacchino Natoli. 2010. "A large fraction of extragenic RNA pol II transcription sites overlap enhancers." *PLoS biology* 8 (5): e1000384.

De Visser, J Arjan GM, Joachim Hermisson, Günter P Wagner, Lauren Ancel Meyers, Homayoun Bagheri-Chaichian, Jeffrey L Blanchard, Lin Chao, James M Cheverud, Santiago F Elena, Walter Fontana, et al. 2003. "Perspective: evolution and detection of genetic robustness." *Evolution* 57 (9): 1959–1972.

Dennett, Daniel. 1995. *Darwin's Dangerous Idea*. Simon & Schuster. New York: Touchstone.

Dingle, Kamaludin, Fatme Ghaddar, Petr Šulc, and Ard A Louis. 2022. "Phenotype bias determines how natural RNA structures occupy the morphospace of all possible shapes." *Molecular biology and evolution* 39 (1): msab280.

Dobin, Alexander, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. 2013. "STAR: ultrafast universal RNA-seq aligner." *Bioinformatics* 29 (1): 15–21.

Domazet-Lošo, Tomislav, Josip Brajković, and Diethard Tautz. 2007. "A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages." *Trends in Genetics* 23 (11): 533–539.

Dorshorst, Ben, Mohammad Harun-Or-Rashid, Alireza Jian Bagherpoor, Carl-Johan Rubin, Chris Ashwell, David Gourichon, Michèle Tixier-Boichard, Finn Hallböök, and Leif Andersson. 2015. "A genomic duplication is associated with ectopic eomesodermin expression in the embryonic chicken comb and two duplex-comb phenotypes." *PLoS genetics* 11 (3): e1004947.

Dou, Yanmei, Heather D Gold, Lovelace J Luquette, and Peter J Park. 2018. "Detecting somatic mutations in normal cells." *Trends in Genetics* 34 (7): 545–557.

Draghi, Jeremy A, Todd L Parsons, Günter P Wagner, and Joshua B Plotkin. 2010. "Mutational robustness can facilitate adaptation." *Nature* 463 (7279): 353–355.

Drummond, D Allan, and Claus O Wilke. 2009. "The evolutionary consequences of erroneous protein synthesis." *Nature Reviews Genetics* 10 (10): 715–724.

Emera, Deena, Jun Yin, Steven K Reilly, Jake Gockley, and James P Noonan. 2016. "Origin and evolution of developmental enhancers in the mammalian neocortex." *Proceedings of the National Academy of Sciences* 113 (19): E2617–E2626.

ENCODE. 2012. "An integrated encyclopedia of DNA elements in the human genome." *Nature* 489 (7414): 57.

Erten, E Yagmur, and Hanna Kokko. 2020. "From zygote to a multicellular soma: Body size affects optimal growth strategies under cancer risk." *Evolutionary applications* 13 (7): 1593–1604.

Erwin, Douglas H. 2015. "Novelty and innovation in the history of life." *Current Biology* 25 (19): R930–R940.

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

Erwin, Douglas H, and Eric H Davidson. 2009. "The evolution of hierarchical gene regulatory networks." *Nature Reviews Genetics* 10 (2): 141–148.

Espinosa-Soto, Carlos, Olivier C Martin, and Andreas Wagner. 2011. "Phenotypic plasticity can facilitate adaptive evolution in gene regulatory circuits." *BMC evolutionary biology* 11 (1): 1–14.

Extavour, Cassandra, and Antonio García-Bellido. 2001. "Germ cell selection in genetic mosaics in Drosophila melanogaster." *Proceedings of the National Academy of Sciences* 98 (20): 11341–11346.

Extavour, Cassandra G, and Michael Akam. 2003. "Mechanisms of germ cell specification across the metazoans: epigenesis and preformation."

Fairclough, Stephen R, Mark J Dayel, and Nicole King. 2010. "Multicellular development in a choanoflagellate." *Current Biology* 20 (20): R875–R876.

Fernández, Rosa, and Toni Gabaldón. 2020. "Gene gain and loss across the metazoan tree of life." *Nature ecology & evolution* 4 (4): 524–533.

Fickett, James W, and Artemis G Hatzigeorgiou. 1997. "Eukaryotic promoter recognition." *Genome research* 7 (9): 861–878.

Flores, Enrique. 2012. "Restricted cellular differentiation in cyanobacterial filaments." *Proceedings of the National Academy of Sciences* 109 (38): 15080–15081.

Fong, Sarah L, and John A Capra. 2021. "Modeling the Evolutionary Architectures of Transcribed Human Enhancer Sequences Reveals Distinct Origins, Functions, and Associations with Human Trait Variation." *Molecular biology and evolution* 38 (9): 3681–3696.

Frank, Steven A. 2011. "Natural selection. II. Developmental variability and evolutionary rate." *Journal of evolutionary biology* 24 (11): 2310–2320.

———. 2010. "Somatic evolutionary genomics: mutations during development cause highly variable genetic mosaicism with risk of cancer and neurodegeneration." *Proceedings of the National Academy of Sciences* 107 (suppl 1): 1725–1730.

Frank, Steven A, and Martin A Nowak. 2004. "Problems of somatic mutation and cancer." *Bioessays* 26 (3): 291–299.

Fryxell, Karl J, and Won-Jong Moon. 2005. "CpG mutation rates in the human genome are highly dependent on local GC content." *Molecular Biology and Evolution* 22 (3): 650–658.

Fulco, Charles P, Mathias Munschauer, Rockwell Anyoha, Glen Munson, Sharon R Grossman, Elizabeth M Perez, Michael Kane, Brian Cleary, Eric S Lander, and Jesse M Engreitz. 2016. "Systematic mapping of functional enhancer–promoter connections with CRISPR interference." *Science* 354 (6313): 769–773.

Fuqua, Timothy, Jeff Jordan, Maria Elize van Breugel, Aliaksandr Halavatyi, Christian Tischer, Peter Polidoro, Namiko Abe, Albert Tsai, Richard S Mann, David L Stern, et al. 2020. "Dense and pleiotropic regulatory information in a developmental enhancer." *Nature* 587 (7833): 235–239.

Furusawa, Chikara, and Kunihiko Kaneko. 2012. "A dynamical-systems view of stem cell biology." *Science* 338 (6104): 215–217.

Gaiti, Federico, Katia Jindrich, Selene L Fernandez-Valverde, Kathrein E Roper, Bernard M Degnan, and Miloš Tanurdžić. 2017. "Landscape of histone modifications in a sponge reveals the origin of animal cis-regulatory complexity." *Elife* 6:e22194.

Galen, Spencer C, Chandrasekhar Natarajan, Hideaki Moriyama, Roy E Weber, Angela Fago, Phred M Benham, Andrea N Chavez, Zachary A Cheviron, Jay F Storz, and Christopher C Witt. 2015. "Contribution of a mutational hot spot to hemoglobin adaptation in high-altitude Andean house wrens." *Proceedings of the National Academy of Sciences* 112 (45): 13958–13963.

Galgoczy, David J, Ann Cassidy-Stone, Manuel Llinás, Sean M O'Rourke, Ira Herskowitz, Joseph L DeRisi, and Alexander D Johnson. 2004. "Genomic dissection of the cell-type-specification circuit in Saccharomyces cerevisiae." *Proceedings of the National Academy of Sciences* 101 (52): 18069–18074.

Gandara, Lautaro, Albert Tsai, Mans Ekelöf, Rafael Galupa, Ella Preger-Ben Noon, Theodore Alexandrov, and Justin Crocker. 2022. "Developmental phenomics suggests that H3K4 monomethylation catalyzed by Trr functions as a phenotypic capacitor." *bioRxiv*.

Gao, Lei, Keliang Wu, Zhenbo Liu, Xuelong Yao, Shenli Yuan, Wenrong Tao, Lizhi Yi, Guanling Yu, Zhenzhen Hou, Dongdong Fan, et al. 2018. "Chromatin accessibility landscape in human early embryos and its association with evolution." *Cell* 173 (1): 248–259.

García-Nieto, Pablo E, Ashby J Morrison, and Hunter B Fraser. 2019. "The somatic mutation landscape of the human body." *Genome biology* 20 (1): 1–20.

Garfield, David A, Daniel E Runcie, Courtney C Babbitt, Ralph Haygood, William J Nielsen, and Gregory A Wray. 2013. "The impact of gene expression variation on the robustness and evolvability of a developmental gene regulatory network." *PLoS biology* 11 (10): e1001696.

Gavrilets, Sergey. 2010. "High-dimensional fitness landscapes and speciation." *Evolution: the extended synthesis:* 45–79.

Gerhart, John, and Marc Kirschner. 2007. "The theory of facilitated variation." *Proceedings of the National Academy of Sciences* 104 (suppl_1): 8582–8589.

Gestel, Jordi van, Martin Ackermann, and Andreas Wagner. 2019. "Microbial life cycles link global modularity in regulation to mosaic evolution." *Nature Ecology & Evolution* 3 (8): 1184–1196.

Gil-Gálvez, Alejandro, Sandra Jiménez-Gancedo, Alberto Pérez-Posada, Martin Franke, Rafael D Acemel, Che-Yi Lin, Cindy Chou, Yi-Hsien Su, Jr-Kai Yu, Stephanie Bertrand, et al. 2022. "Gain of gene regulatory network interconnectivity at the origin of vertebrates." *Proceedings of the National Academy of Sciences* 119 (11): e2114802119.

Gill, Douglas E, Lin Chao, Susan L Perkins, and Jason B Wolf. 1995. "Genetic mosaicism in plants and clonal animals." *Annual review of Ecology and Systematics* 26 (1): 423–444.

Glennon, Richard A, Milt Titeler, and JD McKenney. 1984. "Evidence for 5-HT2 involvement in the mechanism of action of hallucinogenic agents." *Life sciences* 35 (25): 2505–2511.

Gokhale, Chaitanya S, Yoh Iwasa, Martin A Nowak, and Arne Traulsen. 2009. "The pace of evolution across fitness valleys." *Journal of theoretical biology* 259 (3): 613–620.

Gout, Jean-François, W Kelley Thomas, Zachary Smith, Kazufusa Okamoto, and Michael Lynch. 2013. "Large-scale detection of in vivo transcription errors." *Proceedings of the National Academy of Sciences* 110 (46): 18584–18589.

Gout, Jean-Francois, Weiyi Li, Clark Fritsch, Annie Li, Suraiya Haroon, Larry Singh, Ding Hua, Hossein Fazelinia, Zach Smith, Steven Seeholzer, et al. 2017. "The landscape of transcription errors in eukaryotic cells." *Science advances* 3 (10): e1701484.

Granados, Alejandro A, Julian MJ Pietsch, Sarah A Cepeda-Humerez, Iseabail L Farquhar, Gašper Tkačik, and Peter S Swain. 2018. "Distributed and dynamic intracellular organization of extracellular information." *Proceedings of the National Academy of Sciences* 115 (23): 6088–6093.

Grandchamp, Anna, Katrin Berk, Elias Dohmen, and Erich Bornberg-Bauer. 2022. "New genomic signals underlying the emergence of human proto-genes." *bioRxiv*.

Greaves, Mel, and Carlo C Maley. 2012. "Clonal evolution in cancer." *Nature* 481 (7381): 306–313.

Guenther, Matthew G, Stuart S Levine, Laurie A Boyer, Rudolf Jaenisch, and Richard A Young. 2007. "A chromatin landmark and transcription initiation at most promoters in human cells." *Cell* 130 (1): 77–88.

Haag, Eric S, and John R True. 2021. "Developmental system drift." *Evolutionary developmental biology: a reference guide:* 99–110.

Haberle, Vanja, and Alexander Stark. 2018. "Eukaryotic core promoters and the functional basis of transcription initiation." *Nature reviews Molecular cell biology* 19 (10): 621–637.

Hadany, Lilach. 2003. "Adaptive peak shifts in a heterogenous environment." *Theoretical population biology* 63 (1): 41–51.

Hahn, Matthew W, Jeffery P Demuth, and Sang-Gook Han. 2007. "Accelerated rate of gene gain and loss in primates." *Genetics* 177 (3): 1941–1949.

Halfon, Marc S. 2017. "Perspectives on gene regulatory network evolution." *Trends in Genetics* 33 (7): 436–447.

Hanahan, Douglas, and Robert A Weinberg. 2011. "Hallmarks of cancer: the next generation." *cell* 144 (5): 646–674.

Hariharan, Iswar K. 2015. "Organ size control: lessons from Drosophila." *Developmental cell* 34 (3): 255–265.

Harms, Michael J, and Joseph W Thornton. 2014. "Historical contingency and its biophysical basis in glucocorticoid receptor evolution." *Nature* 512 (7513): 203–207.

Hastings, IM. 1989. "Potential germline competition in animals and its evolutionary implications." *Genetics* 123 (1): 191–197.

He, Bing, Changya Chen, Li Teng, and Kai Tan. 2014. "Global view of enhancer–promoter interactome in human cells." *Proceedings of the National Academy of Sciences* 111 (21): E2191–E2199.

Heams, Thomas. 2012. "Selection within organisms in the nineteenth century: Wilhelm Roux's complex legacy." *Progress in biophysics and molecular biology* 110 (1): 24–33.

Heintzman, Nathaniel D, Rhona K Stuart, Gary Hon, Yutao Fu, Christina W Ching, R David Hawkins, Leah O Barrera, Sara Van Calcar, Chunxu Qu, Keith A Ching, et al. 2007. "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome." *Nature genetics* 39 (3): 311–318.

Helsen, Jana, Karin Voordeckers, Laura Vanderwaeren, Toon Santermans, Maria Tsontaki, Kevin J Verstrepen, and Rob Jelier. 2020. "Gene loss predictably drives evolutionary adaptation." *Molecular biology and evolution* 37 (10): 2989–3002.

Herrera-Álvarez, Santiago, Elinor Karlsson, Oliver A Ryder, Kerstin Lindblad-Toh, and Andrew J Crawford. 2021. "How to make a rodent giant: genomic basis and tradeoffs of gigantism in the capybara, the world's largest rodent." *Molecular biology and evolution* 38 (5): 1715–1730.

Hirata, Masashi, Kei-ichiro Nakamura, Takaaki Kanemaru, Yosaburo Shibata, and Shigeru Kondo. 2003. "Pigment cell organization in the hypodermis of zebrafish." *Developmental dynamics: an official publication of the American Association of Anatomists* 227 (4): 497–503.

Hodgkinson, Alan, and Adam Eyre-Walker. 2011. "Variation in the mutation rate across mammalian genomes." *Nature reviews genetics* 12 (11): 756–766.

Hong, Joung-Woo, David A Hendrix, and Michael S Levine. 2008. "Shadow enhancers as a source of evolutionary novelty." *Science* 321 (5894): 1314–1314.

Hutchison III, Clyde A, Ray-Yuan Chuang, Vladimir N Noskov, Nacyra Assad-Garcia, Thomas J Deerinck, Mark H Ellisman, John Gill, Krishna Kannan, Bogumil J Karas, Li Ma, et al. 2016. "Design and synthesis of a minimal bacterial genome." *Science* 351 (6280): aad6253.

Igler, Claudia, Mato Lagator, Gašper Tkačik, Jonathan P Bollback, and Călin C Guet. 2018. "Evolutionary potential of transcription factors for gene regulatory rewiring." *Nature Ecology & Evolution* 2 (10): 1633–1643.

Ingolia, Nicholas T, Gloria A Brar, Noam Stern-Ginossar, Michael S Harris, Gaëlle JS Talhouarne, Sarah E Jackson, Mark R Wills, and Jonathan S Weissman. 2014. "Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes." *Cell reports* 8 (5): 1365–1379.

Iwasa, Yoh, Franziska Michor, and Martin A Nowak. 2004. "Stochastic tunnels in evolutionary dynamics." *Genetics* 166 (3): 1571–1579.

Jablonka, Eva, and Marion J. Lamb. 2005. *Evolution in four dimensions : genetic, epigenetic, behavioral, and symbolic variation in the history of life: Life and mind*. MIT Press, Cambridge, Massachusetts.

Jacob, François. 1977. "Evolution and tinkering." *Science* 196 (4295): 1161–1166.

Jacob, François, and Jacques Monod. 1961. "Genetic regulatory mechanisms in the synthesis of proteins." *Journal of molecular biology* 3 (3): 318–356.

Janzen, Fredric J, and Patrick C Phillips. 2006. "Exploring the evolution of environmental sex determination, especially in reptiles." *Journal of evolutionary biology* 19 (6): 1775–1784.

Jiménez, Alba, James Cotterell, Andreea Munteanu, and James Sharpe. 2015. "Dynamics of gene circuits shapes evolvability." *Proceedings of the National Academy of Sciences* 112 (7): 2103–2108.

Johannsen, Wilhelm. 1911. "The genotype conception of heredity." *The American Naturalist* 45 (531): 129–159.

Johnson, Alexander D. 2017. "The rewiring of transcription circuits in evolution." *Current opinion in genetics & development* 47:121–127.

Ju, Young Seok, Inigo Martincorena, Moritz Gerstung, Mia Petljak, Ludmil B Alexandrov, Raheleh Rahbari, David C Wedge, Helen R Davies, Manasa Ramakrishna, Anthony Fullam, et al. 2017. "Somatic mutations reveal asymmetric cellular dynamics in the early human embryo." *Nature* 543 (7647): 714–718.

Kaern, Mads, Timothy C Elston, William J Blake, and James J Collins. 2005. "Stochasticity in gene expression: from theories to phenotypes." *Nature Reviews Genetics* 6 (6): 451–464.

Kaessmann, Henrik. 2010. "Origins, evolution, and phenotypic impact of new genes." *Genome research* 20 (10): 1313–1326.

Kapranov, Philipp, Aarron T Willingham, and Thomas R Gingeras. 2007. "Genome-wide transcription and the implications for genomic organization." *Nature Reviews Genetics* 8 (6): 413–423.

Karageorgi, Marianthi, Simon C Groen, Fidan Sumbul, Julianne N Pelaez, Kirsten I Verster, Jessica M Aguilar, Amy P Hastings, Susan L Bernstein, Teruyuki Matsunaga, Michael Astourian, et al. 2019. "Genome editing retraces the evolution of toxin resistance in the monarch butterfly." *Nature* 574 (7778): 409–412.

Kauffman, Stuart A. 1969. "Metabolic stability and epigenesis in randomly constructed genetic nets." *Journal of theoretical biology* 22 (3): 437–467.

Kauffman, Stuart A, et al. 1993. *The origins of order: Self-organization and selection in evolution*. Oxford University Press, USA.

Kaur, Gagandeep, and Pawan Krishan. 2020. "Understanding Serotonin 5-HT2A Receptors-regulated cellular and molecular Mechanisms of Chronic Kidney Diseases." *Renal Replacement Therapy* 6 (1): 1–11.

Kennedy, Scott R, Lawrence A Loeb, and Alan J Herr. 2012. "Somatic mutations in aging, cancer and neurodegeneration." *Mechanisms of ageing and development* 133 (4): 118–126.

Kent, W James, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. 2002. "The human genome browser at UCSC." *Genome research* 12 (6): 996–1006.

Kherdjemil, Yacine, Robert L Lalonde, Rushikesh Sheth, Annie Dumouchel, Gemma de Martino, Kyriel M Pineault, Deneen M Wellik, H Scott Stadler, Marie-Andrée Akimenko, and Marie Kmita. 2016. "Evolution of Hoxa11 regulation in vertebrates is linked to the pentadactyl state." *Nature* 539 (7627): 89–92.

Kim, Jinuk, and Young Zoon Kim. 2019. "Analysis of Somatic Variants in Growth Hormone Secreting Pituitary Adenomas by Whole Exome Sequencing." *American Journal of Biomedical Science Research* 4:445–454.

Kim, Tae-Kyung, Martin Hemberg, Jesse M Gray, Allen M Costa, Daniel M Bear, Jing Wu, David A Harmin, Mike Laptewicz, Kellie Barbara-Haley, Scott Kuersten, et al. 2010. "Widespread transcription at neuronal activity-regulated enhancers." *Nature* 465 (7295): 182–187.

Kim, Wonho, and Rajan Jain. 2020. "Picking winners and losers: cell competition in tissue development and homeostasis." *Trends in Genetics* 36 (7): 490–498.

Kimura, Motoo, et al. 1968. "Evolutionary rate at the molecular level." *Nature* 217 (5129): 624–626.

King, Mary-Claire, and Allan C Wilson. 1975. "Evolution at Two Levels in Humans and Chimpanzees: Their macromolecules are so alike that regulatory mutations may account for their biological differences." *Science* 188 (4184): 107–116.

Kirkwood, TB. 2017. "The disposable soma theory." *The evolution of senescence in the tree of life:* 23–39.

Kirkwood, Thomas BL. 1977. "Evolution of ageing." *Nature* 270 (5635): 301–304.

Kirkwood, Thomas BL, and Michael R Rose. 1991. "Evolution of senescence: late survival sacrificed for reproduction." *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 332 (1262): 15–24.

Klironomos, Filippos D, Johannes Berg, and Sinéad Collins. 2013. "How epigenetic mutations can affect genetic evolution: model and mechanism." *BioEssays* 35 (6): 571–578.

Klug, Alexander, Su-Chan Park, and Joachim Krug. 2019. "Recombination and mutational robustness in neutral fitness landscapes." *PLoS computational biology* 15 (8): e1006884.

Knopp, Michael, Jonina S Gudmundsdottir, Tobias Nilsson, Finja König, Omar Warsi, Fredrika Rajer, Pia Ädelroth, and Dan I Andersson. 2019. "De novo emergence of peptides that confer antibiotic resistance." *MBio* 10 (3): e00837–19.

Knowles, David G, and Aoife McLysaght. 2009. "Recent de novo origin of human protein-coding genes." *Genome research* 19 (10): 1752–1759.

Komarova, Natalia L. 2014. "Spatial interactions and cooperation can change the speed of evolution of complex phenotypes." *Proceedings of the National Academy of Sciences* 111 (Supplement 3): 10789–10795.

Komarova, Natalia L, Anirvan Sengupta, and Martin A Nowak. 2003. "Mutation–selection networks of cancer initiation: tumor suppressor genes and chromosomal instability." *Journal of theoretical biology* 223 (4): 433–450.

Kosinski, Luke J, and Joanna Masel. 2020. "Readthrough errors purge deleterious cryptic sequences, facilitating the birth of coding sequences." *Molecular biology and evolution* 37 (6): 1761–1774.

Kratochwil, Claudius F, Yipeng Liang, Jan Gerwin, Joost M Woltering, Sabine Urban, Frederico Henning, Gonzalo Machado-Schiaffino, C Darrin Hulsey, and Axel Meyer. 2018. "Agouti-related peptide 2 facilitates convergent evolution of stripe patterns across cichlid fish radiations." *Science* 362 (6413): 457–460.

Kryuchkova-Mostacci, Nadezda, and Marc Robinson-Rechavi. 2015. "Tissue-specific evolution of protein coding genes in human and mouse." *PLoS One* 10 (6): e0131673.

Kvon, Evgeny Z, Yiwen Zhu, Guy Kelman, Catherine S Novak, Ingrid Plajzer-Frick, Momoe Kato, Tyler H Garvin, Quan Pham, Anne N Harrington, Riana D Hunter, et al. 2020. "Comprehensive in vivo interrogation reveals phenotypic impact of human enhancer variants." *Cell* 180 (6): 1262–1271.

Kvon, Evgeny Z, Rachel Waymack, Mario Gad, and Zeba Wunderlich. 2021. "Enhancer redundancy in development and disease." *Nature Reviews Genetics* 22 (5): 324–336.

Kvon, Evgeny Z, Tomas Kazmar, Gerald Stampfel, J Omar Yáñez-Cuna, Michaela Pagani, Katharina Schernhuber, Barry J Dickson, and Alexander Stark. 2014. "Genome-scale functional characterization of Drosophila developmental enhancers in vivo." *Nature* 512 (7512): 91–95.

Kvon, Evgeny Z, Olga K Kamneva, Uirá S Melo, Iros Barozzi, Marco Osterwalder, Brandon J Mannion, Virginie Tissières, Catherine S Pickle, Ingrid Plajzer-Frick, Elizabeth A Lee, et al. 2016. "Progressive loss of function in a limb enhancer during snake evolution." *Cell* 167 (3): 633–642.

Laforsch, Christian, and Ralph Tollrian. 2004. "Inducible defenses in multipredator environments: cyclomorphosis in Daphnia cucullata." *Ecology* 85 (8): 2302–2311.

Lamarck, Jean-Baptiste de Monet de. 1809. *Philosophie zoologique ou Exposition des considérations relatives à l'histoire naturelle des animaux. Tome premier.* A Paris: chez Dentu: Imprimerie de Duminil-Lesueur.

Lander, Eric S, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. 2001. "Initial sequencing and analysis of the human genome."

Lanno, Stephen M, Serena J Shimshak, Rubye D Peyser, Samuel C Linde, and Joseph D Coolon. 2019. "Investigating the role of Osiris genes in Drosophila sechellia larval resistance to a host plant toxin." *Ecology and evolution* 9 (4): 1922–1933.

Lawson, Andrew RJ, Federico Abascal, Tim HH Coorens, Yvette Hooks, Laura O'Neill, Calli Latimer, Keiran Raine, Mathijs A Sanders, Anne Y Warren, Krishnaa TA Mahbubani, et al. 2020. "Extensive heterogeneity in somatic mutation and selection in the human bladder." *Science* 370 (6512): 75–82.

Lee-Six, Henry, Nina Friesgaard Øbro, Mairi S Shepherd, Sebastian Grossmann, Kevin Dawson, Miriam Belmonte, Robert J Osborne, Brian JP Huntly, Inigo Martincorena, Elizabeth Anderson, et al. 2018. "Population dynamics of normal human blood inferred from somatic mutations." *Nature* 561 (7724): 473–478.

Leria, Laia, Miquel Vila-Farré, Eduard Solà, and Marta Riutort. 2019. "Outstanding intraindividual genetic diversity in fissiparous planarians (Dugesia, Platyhelminthes) with facultative sex." *BMC evolutionary biology* 19 (1): 1–19.

Lettice, Laura A, Simon JH Heaney, Lorna A Purdie, Li Li, Philippe de Beer, Ben A Oostra, Debbie Goode, Greg Elgar, Robert E Hill, and Esther de Graaff. 2003. "A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly." *Human molecular genetics* 12 (14): 1725–1735.

Levine, Mia T, Corbin D Jones, Andrew D Kern, Heather A Lindfors, and David J Begun. 2006. "Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression." *Proceedings of the National Academy of Sciences* 103 (26): 9935–9939.

Levine, Michael, Claudia Cattoglio, and Robert Tjian. 2014. "Looping back to leap forward: transcription enters a new era." *Cell* 157 (1): 13–25.

Levine, Michael, and Robert Tjian. 2003. "Transcription regulation and animal diversity." *Nature* 424 (6945): 147–151.

Lewontin, Richard C. 1983. "The organism as the subject and object of evolution." *Scientia* 77 (18).

Li, Bin, Tao Qing, Jinhang Zhu, Zhuo Wen, Ying Yu, Ryutaro Fukumura, Yuanting Zheng, Yoichi Gondo, and Leming Shi. 2017. "A comprehensive mouse transcriptomic BodyMap across 17 tissues by RNA-seq." *Scientific reports* 7 (1): 1–10.

Li, Dan, Zhihui Yan, Lina Lu, Huifeng Jiang, and Wen Wang. 2014. "Pleiotropy of the de novo-originated gene MDF1." *Scientific reports* 4 (1): 1–4.

Li, Wenbo, Dimple Notani, and Michael G Rosenfeld. 2016. "Enhancers as non-coding RNA transcription units: recent insights and future perspectives." *Nature Reviews Genetics* 17 (4): 207–223.

Lister, James A, Christie P Robertson, Thierry Lepage, Stephen L Johnson, and David W Raible. 1999. "Nacre encodes a zebrafish microphthalmia-related protein that regulates neural-crest-derived pigment cell fate." *Development* 126 (17): 3757–3767.

Liu, Haoxuan, and Jianzhi Zhang. 2019. "Yeast spontaneous mutation rate and spectrum vary with environment." *Current Biology* 29 (10): 1584–1591.

Lizio, Marina, Jayson Harshbarger, Hisashi Shimoji, Jessica Severin, Takeya Kasukawa, Serkan Sahin, Imad Abugessaisa, Shiro Fukuda, Fumi Hori, Sachi Ishikawa-Kato, et al. 2015. "Gateways to the FANTOM5 promoter level mammalian expression atlas." *Genome biology* 16 (1): 1–14.

Lloyd, Mark C, Jessica J Cunningham, Marilyn M Bui, Robert J Gillies, Joel S Brown, and Robert A Gatenby. 2016. "Darwinian dynamics of intratumoral heterogeneity: not solely random mutations but also variable environmental selection forces." *Cancer research* 76 (11): 3136–3144.

Long, Hannah K, Sara L Prescott, and Joanna Wysocka. 2016. "Ever-changing landscapes: transcriptional enhancers in development and evolution." *Cell* 167 (5): 1170–1187.

Lowe, Craig B, Julia A Clarke, Allan J Baker, David Haussler, and Scott V Edwards. 2015. "Feather development genes and associated regulatory innovation predate the origin of Dinosauria." *Molecular biology and evolution* 32 (1): 23–28.

Lynch, Michael. 2010. "Evolution of the mutation rate." *TRENDS in Genetics* 26 (8): 345–352.

Lynch, Vincent J, Mauris C Nnamani, Aurélie Kapusta, Kathryn Brayer, Silvia L Plaza, Erik C Mazur, Deena Emera, Shehzad Z Sheikh, Frank Grützner, Stefan Bauersachs, et al. 2015. "Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy." *Cell reports* 10 (4): 551–561.

Lynd, Amy, David Weetman, Susana Barbosa, Alexander Egyir Yawson, Sara Mitchell, Joao Pinto, Ian Hastings, and Martin J Donnelly. 2010. "Field, genetic, and modeling approaches show strong positive selection acting upon an insecticide resistance mutation in Anopheles gambiae ss." *Molecular Biology and Evolution* 27 (5): 1117–1125.

Maderspacher, Florian, and Christiane Nüsslein-Volhard. 2003. "Formation of the adult pigment pattern in zebrafish requires leopard and obelix dependent cell interactions."

Makova, Kateryna D, and Ross C Hardison. 2015. "The effects of chromatin organization on variation in mutation rates in the genome." *Nature Reviews Genetics* 16 (4): 213–223.

Marlétaz, Ferdinand, Panos N Firbas, Ignacio Maeso, Juan J Tena, Ozren Bogdanovic, Malcolm Perry, Christopher DR Wyatt, Elisa de la Calle-Mustienes, Stephanie Bertrand, Demian Burguera, et al. 2018. "Amphioxus functional genomics and the origins of vertebrate gene regulation." *Nature* 564 (7734): 64–70.

Martin, Nora S, and Sebastian E Ahnert. 2021. "Insertions and deletions in the RNA sequence–structure map." *Journal of the Royal Society Interface* 18 (183): 20210380.

Martincorena, Iñigo, and Peter J Campbell. 2015. "Somatic mutation in cancer and normal cells." *Science* 349 (6255): 1483–1489.

Martincorena, Iñigo, Amit Roshan, Moritz Gerstung, Peter Ellis, Peter Van Loo, Stuart McLaren, David C Wedge, Anthony Fullam, Ludmil B Alexandrov, Jose M Tubio, et al. 2015. "High burden and pervasive positive selection of somatic mutations in normal human skin." *Science* 348 (6237): 880–886.

Martincorena, Iñigo, Joanna C Fowler, Agnieszka Wabik, Andrew RJ Lawson, Federico Abascal, Michael WJ Hall, Alex Cagan, Kasumi Murai, Krishnaa Mahbubani, Michael R Stratton, et al. 2018. "Somatic mutant clones colonize the human esophagus with age." *Science* 362 (6417): 911–917.

Maynard-Smith, J, Richard Burian, Stuart Kauffman, Pere Alberch, John Campbell, Brian Goodwin, Russell Lande, David Raup, and Lewis Wolpert. 1985. "Developmental constraints and evolution: a perspective from the Mountain Lake conference on development and evolution." *The Quarterly Review of Biology* 60 (3): 265–287.

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

Maynard Smith, John, and Eors Szathmary. 1997. *The major transitions in evolution.* OUP Oxford.

Mayr, Ernst. 1972. "Lamarck revisited." *Journal of the History of Biology:* 55–94.

———. 1985. "Weismann and evolution." *Journal of the History of Biology:* 295–329.

McClintock, Barbara. 1950. "The origin and behavior of mutable loci in maize." *Proceedings of the National Academy of Sciences* 36 (6): 344–355.

McFadden, Johnjoe, and Greg Knowles. 1997. "Escape from evolutionary stasis by transposon-mediated deleterious mutations." *Journal of theoretical biology* 186 (4): 441–447.

McGuigan, Katrina, and Carla M Sgro. 2009. "Evolutionary consequences of cryptic genetic variation." *Trends in ecology & evolution* 24 (6): 305–311.

McLysaght, Aoife, and Laurence D Hurst. 2016. "Open questions in the study of de novo genes: what, how and why." *Nature Reviews Genetics* 17 (9): 567–578.

Medawar, PB. 1957. *The uniqueness of the individual. 36 Essex St.*

Medina-Rivera, Alejandra, David Santiago-Algarra, Denis Puthier, and Salvatore Spicuglia. 2018. "Widespread enhancer activity from core promoters." *Trends in biochemical sciences* 43 (6): 452–468.

Melton, Collin, Jason A Reuter, Damek V Spacek, and Michael Snyder. 2015. "Recurrent somatic mutations in regulatory regions of human cancer genomes." *Nature genetics* 47 (7): 710–716.

Mendoza, Alex de, Arnau Sebé-Pedrós, Martin Sebastijan Šestak, Marija Matejčić, Guifré Torruella, Tomislav Domazet-Lošo, and Iñaki Ruiz-Trillo. 2013. "Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages." *Proceedings of the National Academy of Sciences* 110 (50): E4858–E4866.

Michalakis, Yannis, and Montgomery Slatkin. 1996. "Interaction of selection and recombination in the fixation of negative-epistatic genes." *Genetics Research* 67 (3): 257–269.

Michel, Audrey M, Gearoid Fox, Anmol M. Kiran, Christof De Bo, Patrick BF O'Connor, Stephen M Heaphy, James PA Mullan, Claire A Donohue, Desmond G Higgins, and Pavel V Baranov. 2014. "GWIPS-viz: development of a ribo-seq genome browser." *Nucleic acids research* 42 (D1): D859–D864.

Milholland, Brandon, Xiao Dong, Lei Zhang, Xiaoxiao Hao, Yousin Suh, and Jan Vijg. 2017. "Differences between germline and somatic mutation rates in humans and mice." *Nature communications* 8 (1): 1–8.

Miura, Toru. 2005. "Developmental regulation of caste-specific characters in social-insect polyphenism." *Evolution & development* 7 (2): 122–129.

Monod, Jacques. 1970. *Le hasard et la nécessite, essai sur la philosophie naturelle de la biologie moderne.* Du Seuil. Paris.

Monro, Keyne, and Alistair GB Poore. 2009. "Performance benefits of growth-form plasticity in a clonal red seaweed." *Biological Journal of the Linnean Society* 97 (1): 80–89.

Monroe, J, Thanvi Srikant, Pablo Carbonell-Bejerano, Claude Becker, Mariele Lensink, Moises Exposito-Alonso, Marie Klein, Julia Hildebrandt, Manuela Neumann, Daniel Kliebenstein, et al. 2022. "Mutation bias reflects natural selection in Arabidopsis thaliana." *Nature:* 1–5.

Moore, Luiza, Alex Cagan, Tim HH Coorens, Matthew DC Neville, Rashesh Sanghvi, Mathijs A Sanders, Thomas RW Oliver, Daniel Leongamornlert, Peter Ellis, Ayesha Noorani, et al. 2021. "The mutational landscape of human somatic and germline cells." *Nature* 597 (7876): 381–386.

Morange, Michel. 2016. *Une histoire de la biologie (inédit).* Média Diffusion.

Morata, Ginés, and Pedro Ripoll. 1975. "Minutes: mutants of Drosophila autonomously affecting cell division rate." *Developmental biology* 42 (2): 211–221.

Moreno, Eduardo, Konrad Basler, and Ginés Morata. 2002. "Cells compete for decapentaplegic survival factor to prevent apoptosis in Drosophila wing development." *Nature* 416 (6882): 755–759.

Müller, Viktor, Rob J De Boer, Sebastian Bonhoeffer, and Eörs Szathmáry. 2018. "An evolutionary perspective on the systems of adaptive immunity." *Biological Reviews* 93 (1): 505–528.

Murphey, Patricia, Derek J McLean, C Alex McMahan, Christi A Walter, and John R McCarrey. 2013. "Enhanced genetic integrity in mouse germ cells." *Biology of reproduction* 88 (1): 6–1.

Murugesan, Suriya Narayanan, Heidi Connahs, Yuji Matsuoka, Mainak Das Gupta, Galen JL Tiong, Manizah Huq, V Gowri, Sarah Monroe, Kevin D Deem, Thomas Werner, et al. 2022. "Butterfly eyespots evolved via cooption of an ancestral gene-regulatory network that also patterns antennae, legs, and wings." *Proceedings of the National Academy of Sciences* 119 (8): e2108661119.

Mustonen, Ville, and Michael Lässig. 2009. "From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation." *Trends in genetics* 25 (3): 111–119.

Neinavaie, Fargam, Arig Ibrahim-Hashim, Andrew M Kramer, Joel S Brown, and Christina L Richards. 2022. "The genomic processes of biological invasions: From invasive species to cancer metastases and back again." *From Ecology to Cancer Biology and Back Again.*

Nelson, Craig E, Bradley M Hersh, and Sean B Carroll. 2004. "The regulatory content of intergenic DNA shapes genome architecture." *Genome biology* 5 (4): 1–15.

Neme, Rafik, and Diethard Tautz. 2016. "Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence." *elife* 5:e09977.

———. 2013. "Phylogenetic patterns of emergence of new genes support a model of frequent de novoevolution." *BMC genomics* 14 (1): 1–13.

Nishimoto, Koshiro, Scott A Tomlins, Rork Kuick, Andi K Cani, Thomas J Giordano, Daniel H Hovelson, Chia-Jen Liu, Aalok R Sanjanwala, Michael A Edwards, Celso E Gomez-Sanchez, et al. 2015. "Aldosterone-stimulating somatic gene mutations are common in normal adrenal glands." *Proceedings of the National Academy of Sciences* 112 (33): E4591–E4599.

Nocedal, Isabel, Eugenio Mancera, and Alexander D Johnson. 2017. "Gene regulatory network plasticity predates a switch in function of a conserved transcription regulator." *Elife* 6:e23250.

Noguchi, Shuhei, Takahiro Arakawa, Shiro Fukuda, Masaaki Furuno, Akira Hasegawa, Fumi Hori, Sachi Ishikawa-Kato, Kaoru Kaida, Ai Kaiho, Mutsumi Kanamori-Katayama, et al. 2017. "FANTOM5 CAGE profiles of human and mouse samples." *Scientific data* 4 (1): 1–10.

Noon, Ella Preger-Ben, Fred P Davis, and David L Stern. 2016. "Evolved repression overcomes enhancer robustness." *Developmental cell* 39 (5): 572–584.

Nurk, Sergey, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V Bzikadze, Alla Mikheenko, Mitchell R Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, et al. 2022. "The complete sequence of a human genome." *Science* 376 (6588): 44–53.

Obolski, Uri, Ohad Lewin-Epstein, Eran Even-Tov, Yoav Ram, and Lilach Hadany. 2017. "With a little help from my friends: cooperation can accelerate the rate of adaptive valley crossing." *BMC evolutionary biology* 17 (1): 1–10.

Odegard, Valerie H, and David G Schatz. 2006. "Targeting of somatic hypermutation." *Nature Reviews Immunology* 6 (8): 573–583.

Oliveri, Paola, and Eric H Davidson. 2004. "Gene regulatory network controlling embryonic specification in the sea urchin." *Current opinion in genetics & development* 14 (4): 351–360.

Osorio, Fernando G, Axel Rosendahl Huber, Rurika Oka, Mark Verheul, Sachin H Patel, Karlijn Hasaart, Lisanne de la Fonteijne, Ignacio Varela, Fernando D Camargo, and Ruben van Boxtel. 2018. "Somatic mutations reveal lineage relationships and age-related mutagenesis in human hematopoiesis." *Cell reports* 25 (9): 2308–2316.

Oster, George, and Pere Alberch. 1982. "Evolution and bifurcation of developmental programs." *Evolution:* 444–459.

Osterwalder, Marco, Iros Barozzi, Virginie Tissières, Yoko Fukuda-Yuzawa, Brandon J Mannion, Sarah Y Afzal, Elizabeth A Lee, Yiwen Zhu, Ingrid Plajzer-Frick, Catherine S Pickle, et al. 2018. "Enhancer redundancy provides phenotypic robustness in mammalian development." *Nature* 554 (7691): 239–243.

Otto, Sarah P, and Ian M Hastings. 1998. "Mutation and selection within the individual." *Genetica* 102:507–524.

Otto, Sarah P, and Maria E Orive. 1995. “Evolutionary consequences of mutation and selection within an individual.” *Genetics* 141 (3): 1173–1187.

Owen, Jennifer P, Robert N Kelsh, and Christian A Yates. 2020. “A quantitative modelling approach to zebrafish pigment pattern formation.” *Elife* 9:e52998.

Palmer, John R, and Charles J Daniels. 1995. “In vivo definition of an archaeal promoter.” *Journal of bacteriology* 177 (7): 1844–1849.

Partha, Raghavendran, Bharesh K Chauhan, Zelia Ferreira, Joseph D Robinson, Kira Lathrop, Ken K Nischal, Maria Chikina, and Nathan L Clark. 2017. “Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling.” *Elife* 6:e25884.

Patterson, Larissa B, and David M Parichy. 2019. “Zebrafish pigment pattern formation: insights into the development and evolution of adult form.” *Annual Review of Genetics* 53:505–530.

Payne, Joshua L, Jason H Moore, and Andreas Wagner. 2014. “Robustness, evolvability, and the logic of genetic regulation.” *Artificial life* 20 (1): 111–126.

Payne, Joshua L, and Andreas Wagner. 2019. “The causes of evolvability and their evolution.” *Nature Reviews Genetics* 20 (1): 24–38.

———. 2014. “The robustness and evolvability of transcription factor binding sites.” *Science* 343 (6173): 875–877.

Perry, Michael W, Alistair N Boettiger, and Michael Levine. 2011. “Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo.” *Proceedings of the National Academy of Sciences* 108 (33): 13570–13575.

Peter, Isabelle S, and Eric H Davidson. 2011. “Evolution of gene regulatory networks controlling body plan development.” *Cell* 144 (6): 970–985.

———. 2009. “Modularity and design principles in the sea urchin embryo gene regulatory network.” *FEBS letters* 583 (24): 3948–3958.

———. 2010. “The endoderm gene regulatory network in sea urchin embryos up to mid-blastula stage.” *Developmental biology* 340 (2): 188–199.

Phillips, T, L Hoopes, et al. 2008. “Transcription factors and transcriptional control in eukaryotic cells.” *Nature Education* 1 (1): 119.

Pigliucci, Massimo. 2010. “Genotype–phenotype mapping and the end of the ‘genes as blueprint’metaphor.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 365 (1540): 557–566.

———. 2008. “Is evolvability evolvable?” *Nature Reviews Genetics* 9 (1): 75–82.

Pigliucci, Massimo, Courtney J Murren, and Carl D Schlichting. 2006. “Phenotypic plasticity and evolution by genetic assimilation.” *Journal of Experimental Biology* 209 (12): 2362–2367.

Pliner, Hannah A, Jonathan S Packer, José L McFaline-Figueroa, Darren A Cusanovich, Riza M Daza, Delasa Aghamirzaie, Sanjay Srivatsan, Xiaojie Qiu, Dana Jackson, Anna Minkina, et al. 2018. “Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data.” *Molecular cell* 71 (5): 858–871.

Podobnik, Marco, Hans Georg Frohnhöfer, Christopher M Dooley, Anastasia Eskova, Christiane Nüsslein-Volhard, and Uwe Irion. 2020. "Evolution of the potassium channel gene Kcnj13 underlies colour pattern diversification in Danio fish." *Nature communications* 11 (1): 1–13.

Policarpo, Maxime, Julien Fumey, Philippe Lafargeas, Delphine Naquin, Claude Thermes, Magali Naville, Corentin Dechaud, Jean-Nicolas Volff, Cedric Cabau, Christophe Klopp, et al. 2021. "Contrasting gene decay in subterranean vertebrates: insights from cavefishes and fossorial mammals." *Molecular Biology and Evolution* 38 (2): 589–605.

Potter, Joshua HT, Kalina TJ Davies, Laurel R Yohe, Miluska KR Sanchez, Edgardo M Rengifo, Monika Struebig, Kim Warren, Georgia Tsagkogeorga, Burton K Lim, Mario Dos Reis, et al. 2021. "Dietary diversification and specialization in neotropical bats facilitated by early molecular evolution." *Molecular biology and evolution* 38 (9): 3864–3883.

Pouplana, Liuís Ribas de, Manuel AS Santos, Jun-Hao Zhu, Philip J Farabaugh, and Babak Javid. 2014. "Protein mistranslation: friend or foe?" *Trends in biochemical sciences* 39 (8): 355–362.

Prabh, Neel, and Christian Rödelsperger. 2016. "Are orphan genes protein-coding, prediction artifacts, or non-coding RNAs?" *BMC bioinformatics* 17 (1): 1–13.

Prescott, Sara L, Rajini Srinivasan, Maria Carolina Marchetto, Irina Grishina, Iñigo Narvaiza, Licia Selleri, Fred H Gage, Tomek Swigut, and Joanna Wysocka. 2015. "Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest." *Cell* 163 (1): 68–83.

Prud'homme, Benjamin, Nicolas Gompel, and Sean B Carroll. 2007. "Emerging principles of regulatory evolution." *Proceedings of the National Academy of Sciences* 104 (suppl 1): 8605–8612.

Prud'homme, Benjamin, Nicolas Gompel, Antonis Rokas, Victoria A Kassner, Thomas M Williams, Shu-Dan Yeh, John R True, and Sean B Carroll. 2006. "Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene." *Nature* 440 (7087): 1050–1053.

Quinlan, Aaron R, and Ira M Hall. 2010. "BEDTools: a flexible suite of utilities for comparing genomic features." *Bioinformatics* 26 (6): 841–842.

Ram, Yoav, and Lilach Hadany. 2014. "Stress-induced mutagenesis and complex adaptation." *Proceedings of the Royal Society B: Biological Sciences* 281 (1792): 20141025.

Raser, Jonathan M, and Erin K O'shea. 2005. "Noise in gene expression: origins, consequences, and control." *Science* 309 (5743): 2010–2013.

Reeve, H Kern, and Laurent Keller. 1999. "Levels of selection: Burying the units-of-selection debate and unearthing the crucial new issues." *Levels of selection in Evolution*. 3–14.

Reusch, Thorsten BH, Iliana B Baums, and Benjamin Werner. 2021. "Evolution via somatic genetic variation in modular species." *Trends in Ecology & Evolution* 36 (12): 1083–1092.

Revy, Patrick, Caroline Kannengiesser, and Alain Fischer. 2019. "Somatic genetic rescue in Mendelian haematopoietic diseases." *Nature Reviews Genetics* 20 (10): 582–598.

Richter-Unruh, A, HT Wessels, U Menken, M Bergmann, K Schmittmann-Ohters, J Schaper, S Tappeser, and BP Hauffa. 2002. "Male LH-independent sexual precocity in a 3.5-year-old boy caused by a somatic activating mutation of the LH receptor in a Leydig cell tumor." *The Journal of Clinical Endocrinology & Metabolism* 87 (3): 1052–1056.

Ros-Rocher, Núria, Alberto Pérez-Posada, Michelle M Leger, and Iñaki Ruiz-Trillo. 2021. "The origin of animals: an ancestral reconstruction of the unicellular-to-multicellular transition." *Open Biology* 11 (2): 200359.

Roscito, Juliana G, Katrin Sameith, Genis Parra, Bjoern E Langer, Andreas Petzold, Claudia Moebius, Marc Bickle, Miguel Trefaut Rodrigues, and Michael Hiller. 2018. "Phenotype loss is associated with widespread divergence of the gene regulatory landscape in evolution." *Nature communications* 9 (1): 1–15.

Rosello, Marion, Juliette Vougny, François Czarny, Marina C Mione, Jean-Paul Concordet, Shahad Albadri, and Filippo Del Bene. 2021. "Precise base editing for the in vivo study of developmental signaling and human pathologies in zebrafish." *Elife* 10:e65552.

Roux, Wilhelm. 1881. *Der kampf der theile im organismus*. W. Engelmann.

Rowicka, Maga, Andrzej Kudlicki, Benjamin P Tu, and Zbyszek Otwinowski. 2007. "High-resolution timing of cell cycle-regulated gene expression." *Proceedings of the National Academy of Sciences* 104 (43): 16892–16897.

Ruiz-Orera, Jorge, and M Mar Albà. 2019. "Translation of small open reading frames: roles in regulation and evolutionary innovation." *Trends in Genetics* 35 (3): 186–198.

Ruiz-Orera, Jorge, Xavier Messeguer, Juan Antonio Subirana, and M Mar Alba. 2014. "Long non-coding RNAs as a source of new peptides." *elife* 3:e03523.

Ruiz-Orera, Jorge, Pol Verdaguer-Grau, José Luis Villanueva-Cañas, Xavier Messeguer, and Mar Albà. 2018. "Translation of neutrally evolving peptides provides a basis for de novo gene evolution." *Nature ecology & evolution* 2 (5): 890–896.

Rutherford, Suzanne L, and Susan Lindquist. 1998. "Hsp90 as a capacitor for morphological evolution." *Nature* 396 (6709): 336–342.

Sabarís, Gonzalo, Ian Laiker, Ella Preger-Ben Noon, and Nicolás Frankel. 2019. "Actors with multiple roles: pleiotropic enhancers and the paradigm of enhancer modularity." *Trends in Genetics* 35 (6): 423–433.

Sackton, Timothy B, Phil Grayson, Alison Cloutier, Zhirui Hu, Jun S Liu, Nicole E Wheeler, Paul P Gardner, Julia A Clarke, Allan J Baker, Michele Clamp, et al. 2019. "Convergent regulatory evolution and loss of flight in paleognathous birds." *Science* 364 (6435): 74–78.

Schaerli, Yolanda, Alba Jiménez, José M Duarte, Ljiljana Mihajlovic, Julien Renggli, Mark Isalan, James Sharpe, and Andreas Wagner. 2018. "Synthetic circuits reveal how mechanisms of gene regulatory networks constrain evolution." *Molecular systems biology* 14 (9): e8102.

Schaper, Steffen, and Ard A Louis. 2014. "The arrival of the frequent: how bias in genotype-phenotype maps can steer populations to local optima." *PloS one* 9 (2): e86635.

Schmitz, Jonathan F, Kristian K Ullrich, and Erich Bornberg-Bauer. 2018. "Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover." *Nature ecology & evolution* 2 (10): 1626–1632.

Schmitz, Jonathan F, Fabian Zimmer, and Erich Bornberg-Bauer. 2016. "Mechanisms of transcription factor evolution in Metazoa." *Nucleic acids research* 44 (13): 6287–6297.

Schmutzer, Michael, and Andreas Wagner. 2020. "Gene expression noise can promote the fixation of beneficial mutations in fluctuating environments." *PLoS computational biology* 16 (10): e1007727.

Schoen, Daniel J, and Stewart T Schultz. 2019. "Somatic mutation and evolution in plants." *Annual Review of Ecology, Evolution, and Systematics* 50:49–73.

Schuster, Peter, Walter Fontana, Peter F Stadler, and Ivo L Hofacker. 1994. "From sequences to shapes and back: a case study in RNA secondary structures." *Proceedings of the Royal Society of London. Series B: Biological Sciences* 255 (1344): 279–284.

Schuster-Böckler, Benjamin, and Ben Lehner. 2012. "Chromatin organization is a major influence on regional mutation rates in human cancer cells." *nature* 488 (7412): 504–507.

Schwaiger, Michaela, Anna Schönauer, André F Rendeiro, Carina Pribitzer, Alexandra Schauer, Anna F Gilles, Johannes B Schinko, Eduard Renfer, David Fredman, and Ulrich Technau. 2014. "Evolutionary conservation of the eumetazoan gene regulatory landscape." *Genome research* 24 (4): 639–650.

Schwarz, Ryan S, and Luis F Cadavid. 2007. "Dynamics of Somatic Cell Lineage Competition in Chimeras of Hydractinia symbiolongicarpus (CNIDARIA: HYDROZOA)." *Acta Biológica Colombiana* 12:13–26.

Schweisguth, François, and Francis Corson. 2019. "Self-organization in pattern formation." *Developmental cell* 49 (5): 659–677.

Sebé-Pedrós, Arnau, Baptiste Saudemont, Elad Chomsky, Flora Plessier, Marie-Pierre Mailhé, Justine Renno, Yann Loe-Mie, Aviezer Lifshitz, Zohar Mukamel, Sandrine Schmutz, et al. 2018a. "Cnidarian cell type diversity and regulation revealed by whole-organism single-cell RNA-Seq." *Cell* 173 (6): 1520–1534.

Sebé-Pedrós, Arnau, Elad Chomsky, Kevin Pang, David Lara-Astiaso, Federico Gaiti, Zohar Mukamel, Ido Amit, Andreas Hejnol, Bernard M Degnan, and Amos Tanay. 2018b. "Early metazoan cell type diversity and the evolution of multicellular gene regulation." *Nature ecology & evolution* 2 (7): 1176–1188.

Sebé-Pedrós, Arnau, Cecilia Ballaré, Helena Parra-Acero, Cristina Chiva, Juan J Tena, Eduard Sabidó, José Luis Gómez-Skarmeta, Luciano Di Croce, and Inaki Ruiz-Trillo. 2016. "The dynamic regulatory genome of Capsaspora and the origin of animal multicellularity." *Cell* 165 (5): 1224–1237.

Sharma, Tara, and Charles A Ettensohn. 2010. "Activation of the skeletogenic gene regulatory network in the early sea urchin embryo." *Development* 137 (7): 1149–1157.

Shbailat, Seba Jamal, and Ehab Abouheif. 2013. "The wing-patterning network in the wingless castes of Myrmicine and Formicine ant species is a mix of evolutionarily labile and non-labile genes." *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 320 (2): 74–83.

Shiraki, Toshiyuki, Shinji Kondo, Shintaro Katayama, Kazunori Waki, Takeya Kasukawa, Hideya Kawaji, Rimantas Kodzius, Akira Watahiki, Mari Nakamura, Takahiro Arakawa, et al. 2003. "Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage." *Proceedings of the National Academy of Sciences* 100 (26): 15776–15781.

Shore, David. 1997. "Telomere length regulation: getting the measure of chromosome ends." *Biological chemistry* 378 (7): 591–597.

Siepel, Adam. 2009. "Darwinian alchemy: Human genes from noncoding DNA." *Genome research* 19 (10): 1693–1695.

Sikosek, Tobias, Hue Sun Chan, and Erich Bornberg-Bauer. 2012. "Escape from adaptive conflict follows from weak functional trade-offs and mutational robustness." *Proceedings of the National Academy of Sciences* 109 (37): 14888–14893.

Simpson, Pat. 1979. "Parameters of cell competition in the compartments of the wing disc of Drosophila." *Developmental biology* 69 (1): 182–193.

Singer, Tatjana, Michael J McConnell, Maria CN Marchetto, Nicole G Coufal, and Fred H Gage. 2010. "LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes?" *Trends in neurosciences* 33 (8): 345–354.

Singh, Ajeet Pratap, and Christiane Nüsslein-Volhard. 2015. "Zebrafish stripes as a model for vertebrate colour pattern formation." *Current Biology* 25 (2): R81–R92.

Snetkova, Valentina, Athena R Ypsilanti, Jennifer A Akiyama, Brandon J Mannion, Ingrid Plajzer-Frick, Catherine S Novak, Anne N Harrington, Quan T Pham, Momoe Kato, Yiwen Zhu, et al. 2021. "Ultraconserved enhancer function does not require perfect sequence conservation." *Nature genetics* 53 (4): 521–528.

Sniegowski, Paul D, Philip J Gerrish, Toby Johnson, and Aaron Shaver. 2000. "The evolution of mutation rates: separating causes from consequences." *Bioessays* 22 (12): 1057–1066.

Solé, Ricard V, Ramon Ferrer-Cancho, Jose M Montoya, and Sergi Valverde. 2002. "Selection, tinkering, and emergence in complex networks." *Complexity* 8 (1): 20–33.

Somarelli, Jason A. 2021. "The hallmarks of cancer as ecologically driven phenotypes." *Frontiers in ecology and evolution* 9.

Spitz, François, and Eileen EM Furlong. 2012. "Transcription factors: from enhancer binding to developmental control." *Nature reviews genetics* 13 (9): 613–626.

Stergachis, Andrew B, Shane Neph, Alex Reynolds, Richard Humbert, Brady Miller, Sharon L Paige, Benjamin Vernot, Jeffrey B Cheng, Robert E Thurman, Richard Sandstrom, et al. 2013. "Developmental fate and cellular maturity encoded in human regulatory DNA landscapes." *Cell* 154 (4): 888–903.

Stern, David L, and Virginie Orgogozo. 2009. "Is genetic evolution predictable?" *Science* 323 (5915): 746–751.

———. 2008. "The loci of evolution: how predictable is genetic evolution?" *Evolution: International Journal of Organic Evolution* 62 (9): 2155–2177.

Stoltzfus, Arlin, and David M McCandlish. 2017. "Mutational biases influence parallel adaptation." *Molecular biology and evolution* 34 (9): 2163–2172.

Stoltzfus, Arlin, and Ryan W Norris. 2016. "On the causes of evolutionary transition: transversion bias." *Molecular biology and evolution* 33 (3): 595–602.

Tautz, Diethard, and Tomislav Domazet-Lošo. 2011. "The evolutionary origin of orphan genes." *Nature Reviews Genetics* 12 (10): 692–702.

Thompson, Darcy Wentworth, and D'Arcy W Thompson. 1942. *On growth and form*. Vol. 2. Cambridge university press Cambridge.

Toll-Riera, Macarena, Nina Bosch, Nicolas Bellora, Robert Castelo, Lluis Armengol, Xavier Estivill, and M Mar Alba. 2009. "Origin of primate orphan genes: a comparative genomics approach." *Molecular biology and evolution* 26 (3): 603–612.

Trigos, Anna S, Richard B Pearson, Anthony T Papenfuss, and David L Goode. 2019. "Somatic mutations in early metazoan genes disrupt regulatory links between unicellular and multicellular genes in cancer." *Elife* 8:e40947.

Tsai, Albert, Mariana RP Alves, and Justin Crocker. 2019. "Multi-enhancer transcriptional hubs confer phenotypic robustness." *Elife* 8:e45325.

Tseng, Chen-Yuan, Michael Burel, Michael Cammer, Sneh Harsh, Maria Sol Flaherty, Stefan Baumgartner, and Erika A Bach. 2022. "chinmo-mutant spermatogonial stem cells cause mitotic drive by evicting non-mutant neighbors from the niche." *Developmental Cell* 57 (1): 80–94.

Tsong, Annie E, Brian B Tuch, Hao Li, and Alexander D Johnson. 2006. "Evolution of alternative transcriptional circuits with identical logic." *Nature* 443 (7110): 415–420.

Uller, Tobias, Armin P Moczek, Richard A Watson, Paul M Brakefield, and Kevin N Laland. 2018. "Developmental bias and evolution: A regulatory network perspective." *Genetics* 209 (4): 949–966.

Vakirlis, Nikolaos, Omer Acar, Brian Hsu, Nelson Castilho Coelho, S Branden Van Oss, Aaron Wacholder, Kate Medetgul-Ernar, Ray W Bowman, Cameron P Hines, John Iannotta, et al. 2020. "De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences." *Nature communications* 11 (1): 1–18.

Van Egeren, Debra, Thomas Madsen, and Franziska Michor. 2018. "Fitness variation in isogenic populations leads to a novel evolutionary mechanism for crossing fitness valleys." *Communications biology* 1 (1): 1–9.

Van Nimwegen, Erik, James P Crutchfield, and Martijn Huynen. 1999. "Neutral evolution of mutational robustness." *Proceedings of the National Academy of Sciences* 96 (17): 9716–9720.

Van Oppen, Madeleine JH, Petra Souter, Emily J Howells, Andrew Heyward, and Ray Berkelmans. 2011. "Novel genetic diversity through somatic mutations: fuel for adaptation of reef corals?" *Diversity* 3 (3): 405–423.

Van Oss, Stephen Branden, and Anne-Ruxandra Carvunis. 2019. "De novo gene birth." *PLoS genetics* 15 (5): e1008160.

Vierstra, Jeff, Eric Rynes, Richard Sandstrom, Miaohua Zhang, Theresa Canfield, R Scott Hansen, Sandra Stehling-Sun, Peter J Sabo, Rachel Byron, Richard Humbert, et al. 2014. "Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution." *Science* 346 (6212): 1007–1012.

Villar, Diego, Camille Berthelot, Sarah Aldridge, Tim F Rayner, Margus Lukk, Miguel Pignatelli, Thomas J Park, Robert Deaville, Jonathan T Erichsen, Anna J Jasinska, et al. 2015. "Enhancer evolution across 20 mammalian species." *Cell* 160 (3): 554–566.

Waddington, Conrad H. 1942. "Canalization of development and the inheritance of acquired characters." *Nature* 150 (3811): 563–565.

———. 1953. "Genetic assimilation of an acquired character." *Evolution:* 118–126.

Waddington, Conrad Hal. 1961. "Genetic assimilation." *Advances in genetics* 10:257–293.

———. 2014. *The strategy of the genes*. Routledge.

Wagner, Andreas. 2008. "Robustness and evolvability: a paradox resolved." *Proceedings of the Royal Society B: Biological Sciences* 275 (1630): 91–100.

Wagner, Günter P. 1996. "Homologues, natural kinds and the evolution of modularity." *American Zoologist* 36 (1): 36–43.

Wagner, Günter P, Mihaela Pavlicev, and James M Cheverud. 2007. "The road to modularity." *Nature Reviews Genetics* 8 (12): 921–931.

Wallbank, Richard WR, Simon W Baxter, Carolina Pardo-Diaz, Joseph J Hanly, Simon H Martin, James Mallet, Kanchon K Dasmahapatra, Camilo Salazar, Mathieu Joron, Nicola Nadeau, et al. 2016. "Evolutionary novelty in a butterfly wing pattern through enhancer shuffling." *PLoS biology* 14 (1): e1002353.

Wang, Xiaoxia, Wendy E Grus, and Jianzhi Zhang. 2006. "Gene losses during human origins." *PLoS biology* 4 (3): e52.

Wang, Xinchen, and David B Goldstein. 2020. "Enhancer domains predict gene pathogenicity and inform gene discovery in complex disease." *The American Journal of Human Genetics* 106 (2): 215–233.

Warnefors, Maria, and Adam Eyre-Walker. 2011. "The accumulation of gene regulation through time." *Genome Biology and Evolution* 3:667–673.

Weinreich, Daniel M, and Lin Chao. 2005. "Rapid evolutionary escape by large populations from local fitness peaks is likely in nature." *Evolution* 59 (6): 1175–1182.

Weismann, August. 1892. *Das Keimplasma: eine theorie der Vererbung*. G. Fischer.

Weissman, Daniel B, Marcus W Feldman, and Daniel S Fisher. 2010. "The rate of fitness-valley crossing in sexual populations." *Genetics* 186 (4): 1389–1410.

Weissman, Daniel B, Michael M Desai, Daniel S Fisher, and Marcus W Feldman. 2009. "The rate at which asexual populations cross fitness valleys." *Theoretical population biology* 75 (4): 286–300.

Werner, Michael S, Bogdan Sieriebriennikov, Neel Prabh, Tobias Loschko, Christa Lanz, and Ralf J Sommer. 2018. "Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation." *Genome research* 28 (11): 1675–1687.

West, Jeffrey, Ryan O Schenck, Chandler Gatenbee, Mark Robertson-Tessi, and Alexander RA Anderson. 2021. "Normal tissue architecture determines the evolutionary course of cancer." *Nature communications* 12 (1): 1–9.

West-Eberhard, Mary Jane. 2003. *Developmental plasticity and evolution*. Oxford University Press.

Westmann, Cauã A, Luana de Fatima Alves, Rafael Silva-Rocha, and María-Eugenia Guazzaroni. 2018. "Mining novel constitutive promoter elements in soil metagenomic libraries in Escherichia coli." *Frontiers in microbiology* 9:1344.

Whitehead, Dion J, Claus O Wilke, David Vernazobres, and Erich Bornberg-Bauer. 2008. "The look-ahead effect of phenotypic mutations." *Biology Direct* 3 (1): 1–15.

Whitham, Thomas G, and CN Slobodchikoff. 1981. "Evolution by individuals, plant-herbivore interactions, and mosaics of genetic variability: the adaptive significance of somatic mutations in plants." *Oecologia* 49 (3): 287–292.

Wijewardhane, Neshika, Lisa Dressler, and Francesca D Ciccarelli. 2021. "Normal somatic mutations in cancer transformation." *Cancer Cell* 39 (2): 125–129.

Willemsen, Anouk, Marta Félez-Sánchez, and Ignacio G Bravo. 2019. "Genome plasticity in papillomaviruses and de novo emergence of E5 oncogenes." *Genome biology and evolution* 11 (6): 1602–1617.

Wilson, Benjamin A, and Joanna Masel. 2011. "Putatively noncoding transcripts show extensive association with ribosomes." *Genome biology and evolution* 3:1245–1252.

Wilson, Benjamin A, Scott G Foy, Rafik Neme, and Joanna Masel. 2017. "Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth." *Nature ecology & evolution* 1 (6): 1–6.

Witt, Evan, Sigi Benjamin, Nicolas Svetec, and Li Zhao. 2019. "Testis single-cell RNA-seq reveals the dynamics of de novo gene transcription and germline mutational bias in Drosophila." *Elife* 8:e47138.

Wong, Emily S, Dawei Zheng, Siew Z Tan, Neil I Bower, Victoria Garside, Gilles Vanwalleghem, Federico Gaiti, Ethan Scott, Benjamin M Hogan, Kazu Kikuchi, et al. 2020. "Deep conservation of the enhancer regulatory code in animals." *Science* 370 (6517): eaax8137.

Wright, Sewall. 1932. "The roles of mutation, inbreeding, crossbreeding, and selection in evolution."

Wu, Xuebing, and Phillip A Sharp. 2013. "Divergent transcription: a driving force for new gene origination?" *Cell* 155 (5): 990–996.

Xie, Chen, Cemalettin Bekpen, Sven Künzel, Maryam Keshavarz, Rebecca Krebs-Wheaton, Neva Skrabar, Kristian Karsten Ullrich, and Diethard Tautz. 2019a. "A de novo evolved gene in the house mouse regulates female pregnancy cycles." *Elife* 8:e44392.

Xie, Kathleen T, Guliang Wang, Abbey C Thompson, Julia I Wucherpfennig, Thomas E Reimchen, Andrew DC MacColl, Dolph Schluter, Michael A Bell, Karen M Vasquez, and David M Kingsley. 2019b. "DNA fragility in the parallel evolution of pelvic reduction in stickleback fish." *Science* 363 (6422): 81–84.

Xie, Victoria Cochran, Jinyue Pu, Brian PH Metzger, Joseph W Thornton, and Bryan C Dickinson. 2021. "Contingency and chance erase necessity in the experimental evolution of ancestral proteins." *Elife* 10:e67336.

Yang, Bing, and Patricia J Wittkopp. 2017. "Structure of the transcriptional regulatory network correlates with regulatory divergence in Drosophila." *Molecular biology and evolution* 34 (6): 1352–1362.

Yizhak, Keren, François Aguet, Jaegil Kim, Julian M Hess, Kirsten Kübler, Jonna Grimsby, Ruslana Frazer, Hailei Zhang, Nicholas J Haradhvala, Daniel Rosebrock, et al. 2019. "RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues." *Science* 364 (6444): eaaw0726.

Yu, Lei, Christoffer Boström, Sören Franzenburg, Till Bayer, Tal Dagan, and Thorsten BH Reusch. 2020. "Somatic genetic drift and multilevel selection in a clonal seagrass." *Nature Ecology & Evolution* 4 (7): 952–962.

Zalts, Harel, and Itai Yanai. 2017. "Developmental constraints shape the evolution of the nematode mid-developmental transition." *Nature Ecology & Evolution* 1 (5): 1–7.

Zaret, Kenneth S, and Jason S Carroll. 2011. "Pioneer transcription factors: establishing competence for gene expression." *Genes & development* 25 (21): 2227–2241.

Zhang, Jianzhi, Ya-ping Zhang, and Helene F Rosenberg. 2002. "Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey." *Nature genetics* 30 (4): 411–415.

Zhang, Li, Yan Ren, Tao Yang, Guangwei Li, Jianhai Chen, Andrea R Gschwend, Yeisoo Yu, Guixue Hou, Jin Zi, Ruo Zhou, et al. 2019. "Rapid evolution of protein diversity by de novo origination in Oryza." *Nature ecology & evolution* 3 (4): 679–690.

Zhang, Wenyu, Patrick Landback, Andrea R Gschwend, Bairong Shen, and Manyuan Long. 2015. "New genes drive the evolution of gene interaction networks in the human and mouse genomes." *Genome biology* 16 (1): 1–14.

Zheng, Jia, Ning Guo, and Andreas Wagner. 2021. "Mistranslation Reduces Mutation Load in Evolving Proteins through Negative Epistasis with DNA Mutations." *Molecular Biology and Evolution* 38 (11): 4792–4804.

Zheng, Jia, Joshua L Payne, and Andreas Wagner. 2019. "Cryptic genetic variation accelerates evolution by opening access to diverse adaptive peaks." *Science* 365 (6451): 347–353.

Zhu, Min, Xiaobo Yu, Brian Choo, Junqing Wang, and Liantao Jia. 2012. "An antiarch placoderm shows that pelvic girdles arose at the root of jawed vertebrates." *Biology Letters* 8 (3): 453–456.

Zhu, Min, Tianshi Lu, Yuemeng Jia, Xin Luo, Purva Gopal, Lin Li, Mobolaji Odewole, Veronica Renteria, Amit G Singal, Younghoon Jang, et al. 2019. "Somatic mutations increase hepatic clonal fitness and regeneration in chronic liver disease." *Cell* 177 (3): 608–621.

Zhuang, Xuan, Chun Yang, Katherine R Murphy, and C-H Christina Cheng. 2019. "Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids." *Proceedings of the National Academy of Sciences* 116 (10): 4400–4405.

Curriculum vitae

Paco Matheus Majic Bergara

Date of birth: January 27th, 1991
Nationality: Uruguayan / Croatian
Home address: Berninastrasse 11, 8057, Zürich, Switzerland
Work address: CHN H 74, Universitätstrasse 16, 8092, Zürich, Switzerland
Cell phone number: (+41) 78 218 18 77
Work e-mail address: paco.majic@env.ethz.ch
Permanent e-mail address: paco.majic@outlook.com
OrcID: 0000-0003-0670-7413

Education

Doctor in Science Sep.2017 – Jun.2022

Thesis title: "Developmental and regulatory novelties and their implications for evolutionary trajectories"

Advisor: Prof. Dr. Joshua L. Payne

Institute of Integrative Biology, ETH Zürich, Switzerland

Master in Science Apr.2014 – Mar.2016

Thesis title: "Molecular assessment of the cellular processes involved in arm regeneration of the feather star *Oxycomanthus japonicus* (Echinodermata : Crinoidea)"

Advisor: Prof. Dr. Mariko Kondo

Department of Biological Sciences, The University of Tokyo, Japan

Licensed in Biological Sciences (BSc), Genetics and Evolution Mar.2009 – Jan.2013

Thesis title: "Evolution of Cytochrome b in the Abrotrichini tribe (Rodentia: Sigmodontinae): fossoriality and altitude"

Advisor: Prof. Dr. Enrique Lessa

Faculty of Sciences, Universidad de la República, Uruguay

Research experience

PhD student researcher Sep.2017 – Aug.2022

Advisor: Prof. Dr. Joshua L. Payne

Institute of Integrative Biology, ETH Zürich, Switzerland

Researcher Dec.2016 – Mar.2017

Transcriptomic analysis of killifish embryos during diapause

Advisor: Dr. Matias Feijoo

Faculty of Sciences, Universidad de la República, Uruguay

Field Assistant Aug.2016 – Nov.2016

Behavioural ecology of the Damaraland mole rat

Advisor: Dr. Markus Zöttl

The Kuruman River Reserve, South Africa

Masters student researcher Apr.2014 – Mar.2016

Molecular assessment of the cellular processes involved in arm regeneration of the feather star *Oxycomanthus japonicus*

Advisor: Prof. Dr. Mariko Kondo

Misaki Marine Biology Station, The University of Tokyo, Japan

Researcher Oct.2013 – Mar.2014

Analysis of the hox and parahox clusters of crinoids

Advisor: Prof. Dr. Mariko Kondo

Misaki Marine Biology Station, The University of Tokyo, Japan

Research intern Jan.2013 – Aug.2013

Comparative study of the C-peptide sequence of thee evolutionarily divergent insulin of caviomporh rodents

Advisor: Prof. Dr. Enrique Lessa

Faculty of Sciences, Universidad de la República, Uruguay

BSc Student researcher Apr.2012 – Jan.2013

Footprints of natural selection on the Cytochrome b of rodents of the tribe Abrotrichini

Advisor: Prof. Dr. Enrique Lessa

Faculty of Sciences, Universidad de la República, Uruguay

Institutional responsibilities

- Representative in the PhD committee of the Institute of Integrative Biology of ETH Zurich

Approved research projects

- "Genetic anticipation in adaptation to changing environments"

Project approved as part of the Young Researchers Exchange Program between Japan and Switzerland 2021 (cancelled due to the COVID pandemic)

- "Evolution of brood parasitism in cuckoos from the perspective of maternal effects and familial conflict"

Project approved and grant awarded to carry out my doctoral studies in the University of Groningen (refused to carry out my doctoral studies in ETH Zurich)

Supervision of junior researchers/students

- Sun Yiqiao Oct.2021– May.2022

Supervision of term paper "The evolution of gene regulation and the stabilization of a multicellular lifestyle"

- Fabian Schaich Oct.2020–May.2021

Supervision of term paper "On the possibility of Amphibian Antimicrobial Peptides to genetically emerge *de novo*"

- Jasmine Gamblin Nov.2019–Dec.2019

Supervision of laboratory rotation project "Comparative study of human and mouse regulons in kidney using the SCENIC method"

Teaching activities

- Teaching at international intensive course Mar.2016

"Biology of Marine Animals – Fertilization, Development and Regeneration"
The University of Tokyo

Scientific Reviewing

- Elife (2021), Genome Biology and Evolution (2022)

Membership in scientific societies

- Society for Molecular Biology and Evolution (SMBE),
- Society for the Study of Evolution (SSE)

Organization of conferences

- Zurich Interaction Seminar (ZIS) Feb.2018–Dec.2018

Prizes, awards, fellowships

- FY2021 JSPS fellowship for research in Japan (Strategic Program). 2022 (discontinued)

Fellowship consisting of a monthly allowance of 220000 Japanese Yen per month for the duration of five months – discontinued due to COVID regulations

- Japanese Government (Monbukagakusho) Scholarship for Students through embassy recommendation as Graduate Research Student Apr.2014–Mar.2016

Fellowship awarding a monthly stipend of 110000 Japanese Yen per month to carry out research for graduate studies

- Japanese Government (Monbukagakusho) Scholarship for Students through embassy recommendation as Research Student Oct.2013–Mar.2014

Fellowship awarding a monthly stipend of 110000 Japanese Yen per month to carry out research

Skills

Languages

Spanish (Mother tongue) – English (Fluent) – German (Intermediate) – French (Beginner) – Japanese (Beginner)

Programming languages

MATLAB – Python – Bash

Laboratory practices

PCR – Cloning – *in situ* hybridization – Microtome sectioning – Histological staining – Sequencing – RNA and DNA extractions – Light and Fluorescence Microscopy – Handling of small mammals and marine invertebrates

Bioinformatic tools and experience with datasets

bedtools – vcftools – MEGA – PAML – PhyloP – Bulk and single-cell genomic and transcriptomics – ChIP-seq – ATAC-seq

Peer-reviewed publications

- Majic, P., & Payne, J. L. (2020). Enhancers facilitate the birth of de novo genes and gene integration into regulatory networks. *Molecular biology and evolution*, *37*(4), 1165-1178. https://doi.org/10.1093/molbev/msz300

- Majic, P., Erten, Y. E., & Payne, J. L. (2022). The adaptive potential of non-heritable somatic mutations. *The American Naturalist*. https://doi.org/10.1086/721766.
- Majic. P. (2022) The molecular scaffolds of the élan vital. *Parrhesia: a journal of analytical philosophy.* (Accepted)

Unpublished work

- Majic. P. & Payne, J. L. (In preparation) Developmental selection and the perception of mutation bias.

Contributions to conferences

- Euro Evo Devo, Naples, Italy, May, Jun. 2022. – Poster
- EMBO Workshop: The evolution of animal genomes, Granada (virtual), Spain, Sep.2021. – Poster
- EMBO Workshop: Predicting Evolution, EMBL Heidelberg (online), Germany, Jun.2021. – Poster
- Bergson and Vitalism(s): an online workshop, Ghent (virtual), Belgium. Apr.2021. – Panellist and talk
- FEBS/EMBO Venice Summer School 2019: Mechanism in development and evolution, Venice, Italy, Aug.2019. – Flash talk
- Systems Genetics: From Genomes to Complex Traits, Heidelberg, Germany, Sep.2019. – Poster
- Society for Molecular Biology and Evolution, Manchester, UK, Jul.2019. –Poster
- 68th Meeting of the Zoological Society of Japan. Yokohama, Japan. Mar.2016. –Poster

Outreach activities

- Participation in the Swedish podcast Dataspaning where I explained principles of gene regulation and evolution
  Available at: https://podcasts.apple.com/gb/podcast/dataspaning/id1428304138