Unravelling the determinants of the rate of adaptive evolution at the molecular level

Dissertation
in fulfilment of the requirements for the degree
Doctor rerum naturalium

of the Faculty of Mathematics and Natural Sciences
Christian Albrechts University of Kiel

Submitted by
Ana Filipa Moutinho

Max Planck Institute for Evolutionary Biology

Plön, March 2020

First examiner: Dr. Julien Y. Dutheil
Second examiner: Prof. Dr. Tal Dagan

Date of the oral examination: 28th of April 2020

Kurzfassung

Seit Darwin die natürliche Auslese als Motor der Evolution vorgestellt hat, sind Evolutionsbiologen bestrebt zu verstehen, wie vorteilhafte Mutationen die Anpassung der Arten an ihre Umwelt beeinflussen. Die Erforschung der Anpassung erfordert jedoch ein Verständnis der komplexen Dynamik zwischen Nukleotiden, Sequenzen, Proteinen, Organismen, Populationen und Arten. Mit anderen Worten, es erfordert die Bewertung des Zusammenspiels evolutionärer Prozesse über Systeme hinweg. Hier habe ich die Anpassung auf diese Weise untersucht, indem ich die Häufigkeit und Art der adaptiven Mutationen innerhalb der Gene, innerhalb der Genome und zwischen den Arten untersucht habe.

Auf intramolekularer Ebene zeigte dieses Projekt, dass die Zugänglichkeit des Rückstandes zu den Lösungsmitteln als primäre Determinante der Raten adaptiver Substitutionen sowohl bei Tieren als auch bei Pflanzen wirkt, wo adaptive Mutationen an der Proteinoberfläche häufiger vorkommen. Diese Analysen zeigten außerdem höhere Anpassungsraten für Gene, die für Proteine mit zentralen zellulären Funktionen kodieren, auf die Krankheitserreger bei einer Wirtsinfektion normalerweise abzielen. Diese Befunde legten daher nahe, dass die adaptive Evolution von Proteinen durch Interaktionen zwischen Molekülen abläuft, insbesondere auf der interspezifischen Ebene, wo die Wirt-Pathogen-Koevolution wahrscheinlich eine zentrale Rolle spielt.

Durch einen Schritt zurück und die Betrachtung der Anpassung auf verschiedenen Zeitskalen innerhalb des Genoms zeigte diese Arbeit die Rolle junger Gene in der adaptiven Evolution auf. Da diese Gene weiter von ihrem Fitness-Optimum entfernt sind, suggerieren diese Ergebnisse vor, dass sich die Proteine auf eine "adaptive Walk"-Art und Weise anpassen. Dieses Projekt hob ferner hervor, dass die Verteilung der adaptiven Mutationen über die Zeit einem Muster abnehmender Erträge folgt.

Wenn man eine noch breitere Skala betrachtet, indem man die Anpassung auf der Ebene der Spezies untersucht und den Effekt der intramolekularen Variation über mehrere Tierarten hinweg betrachtet, zeigte diese Arbeit eine negative Korrelation zwischen den Raten der adaptiven Substitutionen und der effektiven Populationsgröße (N_e). Trotz des relativ schwachen Signals widersprechen diese Ergebnisse der ursprünglichen Populationsgenetik-Theorie. Stattdessen scheinen sie mit den theoretischen Erwartungen an den phänotypischen Raum übereinzustimmen. Die Ergebnisse bezüglich der negativen Selektion wiederum bestätigen die N_e-Hypothese, wonach die Effizienz der Selektion bei großen N_e-Arten stärker ist. Dieser Effekt wurde gut in den Unterschieden der Verteilung der Fitnesseffekte zwischen vergrabenen und exponierten Rückständen dargestellt, wobei erstere vergleichsweise mildere Effektmutationen in niedrigen N_e-Spezies akkumulieren. Dieses Projekt erweiterte unsere Ergebnisse auf intramolekularer Ebene, indem es den starken Einfluss der makromolekularen Struktur des Proteins auf die Raten der molekularen Anpassung über mehrere Taxa hinweg aufzeigte.

Durch die Bewertung des Zusammenspiels adaptiver Mutationen über verschiedene Organisationsebenen hinweg lieferte diese Arbeit ein tieferes Verständnis der Raten der adaptiven Evolution auf molekularer Ebene und damit eine umfassende Sicht auf die molekulare Basis der Anpassung.

Abstract

Ever since Darwin presented natural selection as a driver of evolution, evolutionary biologists have thrived to understand how beneficial mutations shape species adaptation to their environment. Studying adaptation, however, requires an understanding of the complex dynamics between nucleotides, sequences, proteins, organisms, populations, and species. In other words, it requires assessing the interplay of evolutionary processes across systems. Here, I studied adaptation in such a way by exploring the frequency and nature of adaptive mutations within genes, within genomes, and between species.

At the intramolecular level, this project revealed that the residue's solvent accessibility acts as the primary determinant of rates of adaptive substitutions both in animals and in plants, where adaptive mutations are more frequent at the protein surface. These analyses further showed higher rates of adaptation for genes encoding proteins with central cellular functions, which are the ones usually targeted by pathogens during host infection. These findings, therefore, suggested that protein adaptive evolution proceeds through interactions between molecules, particularly at the interspecific level, where host-pathogen coevolution likely plays a central role.

By taking a step back and looking at adaptation at different time-scales within the genome, this thesis revealed the role of young genes in adaptive evolution. As these genes are further away from their fitness optimum, these findings suggested that proteins adapt in an "adaptive walk" manner. This project further highlighted that the distribution of adaptive mutations across time follows a pattern of diminishing returns.

Looking at an even broader scale by studying adaptation at the species level and considering the effect of intramolecular variation across several animal species, this thesis demonstrated a negative correlation between rates of adaptive substitutions and the effective population size (N_e). Despite the relatively weak signal, these findings contradict initial population genetics theory. Instead, they seem to agree with theoretical expectations at the phenotypic space. In turn, the results regarding negative selection confirm the N_e hypothesis, where the efficiency of selection is stronger in large-N_e species. This effect was well depicted in the differences of the distribution of fitness effects between buried and exposed residues, where the former accumulates comparatively more mild effect mutations in low-N_e species. This project further expanded our findings at the intramolecular level, by revealing the strong influence of the protein's macromolecular structure on rates of molecular adaptation across several taxa.

By assessing the interplay of adaptive mutations across distinct organizational levels, this thesis provided a more profound understanding of rates of adaptive evolution at the molecular level, thus delivering a comprehensive view of the molecular basis of adaptation.

Table of Contents

CHAPTER I

General Introduction

Understanding evolution requires one to account for the complex dynamics between molecules, cells, tissues, organisms, and populations. In other words, to study evolution, one needs to explore the remarkable interactions across systems, rather than focusing on particular elements. Evolution can be defined as the accumulation of changes in the elements that constitute a system over a specific time. Such changes can occur at different time-scales. Some occur rapidly, as evidenced in the evolution of antibiotic resistance in a microbial population (Laehnemann et al. 2014), or the genetic changes endured during host-pathogen interactions (Schulte et al. 2010; Alves et al. 2019). Many others, however, extend over thousands or millions of years, such as the evolution of new species. Understanding how such changes occur constitutes the sole basis of evolutionary thinking.

1.1 Towards an understanding of evolution: the first steps

More than 150 years ago, in the iconic book "The origin of species", Charles Darwin proposed that species evolve through natural selection by looking at the gradual changes of phenotypes (1859). Despite being unaware of the laws of inheritance, he argued that natural selection acts through a steady accumulation of differences rather than a burst of episodic events. As he mentioned, "[…] she [natural selection] can never take a leap, but must advance by the shortest and slowest steps". This theory provided the foundation for the rise of quantitative approaches to measure the impact of selection on phenotypic traits: the so-called biometric school of evolution pioneered by Weldon (1895) and Pearson (1898). This system allowed studying how traits are passed through generations and how evolution responds to selection in a continuous and gradual scale (Weldon 1895; Pearson 1898). Although influential, this "micromutational" view of evolution did not appeal to everyone's eyes. Galton (1894), Darwin's cousin, was the first to refute this theory. He believed that evolution proceeded by discontinuous steps with small bursts of selective events. This conflict continued to grow as the school of Mendelian genetics started to rise (Morgan 1903; Bateson 1913; Punnet 1915), leading to the introduction of concepts such as discrete inheritable units, later defined as "genes" (Johannsen 1911), and independent assortment. This debate was later reconciled, as Fisher (1918) demonstrated that the biometrical and the Mendelians' views were, in fact, compatible. In this classical paper, he developed the mathematical framework to understand how genes produce phenotypes. He described the infinitesimal model, which assumes that a phenotypic trait is affected by an infinite number of genes, all unlinked to each other, with no interactions, each having a small infinitesimal effect on the trait of interest. Based on the segregation analysis founded by Weinberg (1908) and Hardy (1908) - the Hardy-Weinberg law (Stern 1943), Fisher's astonishing work provided the avenue for a deeper understanding of how genes interact within a population.

1.2 The modern synthesis and the rise of population genetics

As Mendelian genetics became prominent, a new perspective started to rise: the notion of evolution as a random process. Hagedoorn and Hagedoorn (1921) were the first to point out that some genes may be lost simply by chance because the number of reproducing individuals is considerably smaller than of those that compose the species. Fisher (1922, 1930a, b) and Wright (1931) performed the mathematics of the so-called "Hagedoorn Effect" and reached the solution for the rate of decay in a population of finite size due to the random sampling of genetic variants every new generation, a process known as genetic drift. The pioneering work of Fisher and Wright was followed by Haldane (1939) and Malécot (1944), leading to a deeper understanding of how gene frequencies change over time, which culminated in the birth of the field of population genetics.

As Sewall Wright (1949, 1951) stated, the determinants of gene frequency variation can be seen as two sorts: systematic, such as selection, migration, and mutation, which tend to move the gene frequency towards an equilibrium; and dispersive, like the chance fluctuations in finite populations, which cause gene frequencies to spread. This "process of trial and error" as referred by Motoo Kimura (1955), which combines natural selection with population genetics, defines the evolutionary theory that is still considered today: the "Modern Synthesis" (Huxley 1942).

1.3 The nature of evolutionary changes: neutral evolution and natural selection

The 1950s were characterized by the "Watson-Crick bombshell", citing James F. Crow (2003). The discovery of the DNA molecule (Crick and Watson 1953) led to the rise of molecular genetics, allowing for a more thorough understanding of evolution and species differences. The analysis of molecular data between and within species (e.g., Freese 1962; Sueoka 1962; Zuckerkandl and Pauling 1962) identified two types of genetic variants: polymorphisms, corresponding to the variation within a population, and substitutions, consisting of the differences between species. Up until the 1960s, the nature of evolutionary changes was attributed to directional natural selection, and balancing selection was the fuel that maintained alleles at intermediate frequencies within a population (e.g., Dobhansky 1955; Ford 1964, 1975; Mayr 1965). In the late 1960s and 1970s, however, the discovery of large amounts of protein polymorphisms in natural populations raised the question of whether selection was the main force maintaining them (Shaw 1965; Harris 1966; Lewontin and Hubby 1966). The controversy started to grow: are genetic differences prompted by natural selection or by random genetic drift? This question has long been critical in the study of molecular evolution and established the long-standing debate between the so-called "selectionists" and "neutralists". These two fronts laid in the two most conspicuous theories of molecular evolution: the theory of evolution by natural selection, also known as Darwinian evolution, and the neutral theory, later proposed by Kimura (1968).

The selectionist and the study of adaptation

The selectionist front has been around since Darwin's evolution was presented. Fisher, although a pioneer in the stochastic population genetics theory, was well-known for being Darwin's advocate. In 1930, he presented the first model that allowed for different mutations to have different effects on the phenotype: the geometric model of adaptation (Fisher 1930a). With this model, Fisher wanted to answer a simple question: is adaptation made of phenotypically small or large mutations? By considering that an environmental change moves the population away from its optimal conditions, Fisher proposed that adaptation occurs through the accumulation of mutations with different size and random effects on the phenotype (either towards or away from the optimum), each having a pleiotropic effect on the trait (*i.e.*, the same mutation might have different effects on different traits). Fisher's model then inferred the fitness effect of a mutation according to its size and direction on the phenotypic space. From this, he estimated that the probability that a mutation is beneficial is 50%, although this falls rapidly with increasing mutational size. Fisher, therefore, concluded that adaptation proceeds through the acquisition of mutations with small effect size on fitness. Early after, Wright introduced the shifting balance theory of evolution and gave rise to the concept of fitness and adaptive landscape (Wright 1931). Wright's view of adaptation was different from Fisher's: he believed that adaptation could not be explained solely by natural selection. His model combined the effect of genetic drift, which shifts local populations to temporarily lower fitness, and natural selection, which brings the population back to higher fitness. Wright's landscapes were then characterized by its ruggedness, a concept later fully developed by Kauffman and Levin (1987), where adaptation results from the complex interaction between population structure, epistasis, drift, and migration. These landscapes typically represented the fitness values of a "field of [all] possible gene combinations," where the valleys represent the lowest fitness, and the "hills" illustrate the highest fitness (Figure 1).

Fisher and Wright's models of adaptation considered the Mendelian nature of mutations but were lacking the knowledge of the molecular basis of inheritance. John Maynard Smith was a key contributor in this sense by introducing one of the first sequence-based models of adaptation (Smith 1962). He presented the idea that adaptation occurs in the sequence space, which, unlike the phenotypic space, is discrete. Maynard Smith further suggested that adaptation consists of an "adaptive walk" throughout the space of all possible sequences, from one functional protein, or DNA sequence, to another (Smith 1970a). Maynard Smith's ideas, however, were ignored for almost two decades due to the rise of the Neutral Theory. It was only in the late 1980s that the theoretical study of adaptation returned, with John Gillespie playing a significant role (Gillespie 1983, 1984). Gillespie's work focused on understanding the distribution of fitness effects of beneficial mutations at the molecular level. Contrary to Fisher's geometric model of adaptation, where mutations have a direct phenotypic effect, molecular models of adaptation derive allele fitness's from a certain probability distribution (Gillespie 1983, 1984, 1991; Kimura 1983). Gillespie's key insight, in this sense, was the use of extreme value theory (EVT) to estimate the distribution of fitness effects of beneficial mutations. His use of EVT relied on the fact that the wild-type allele has high fitness. Hence, the fittest few alleles are always drawn from the right tail of the fitness distribution (Gillespie 1983, 1984). In other words, the probability distribution does not matter, as beneficial alleles will behave in a similar way. EVT then provides the differences in fitness of beneficial mutations, which is used to estimate the distribution of fitness effects. Gillespie used this theory to study adaptation over his "mutational landscape" by considering a strong selection-weak mutation model. The idea of weak mutations was similar to that proposed by Maynard Smith: the per-site mutation rate is low enough for one to overlook double mutants. The assumption of strong selection relied on the fact that mutations are either beneficial or deleterious, leaving no room for neutral mutations. One of the most important contributions of Gillespie's work was the estimation of the "move rule" in a mutational landscape under positive natural selection (Gillespie 1983, 1984, 1991). He proposed that the probability of fixation of a beneficial allele depends only on its selective advantage, large-effect beneficial mutations being, therefore, more likely to be fixed. Moreover, Gillespie (1991) suggested that, while neutral evolution leaves a signature of a simple molecular clock, natural selection is represented by a dispersed clock, due to the small bursts of substitutions. He argued that his adaptive view of evolution explained the data better than neutrality (Gillespie 1986, 1989, 1991).

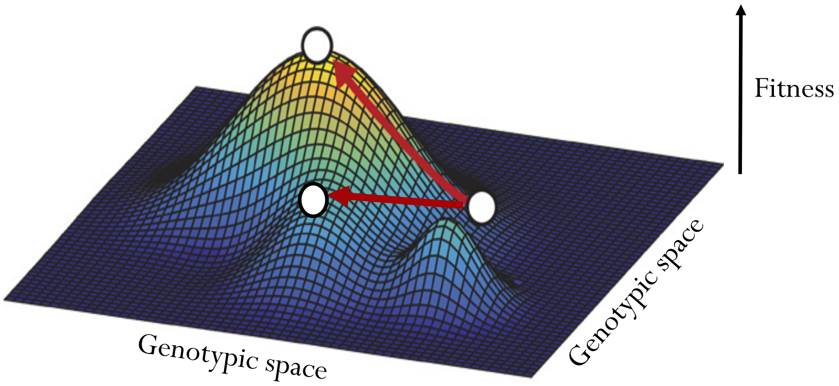

Figure 1. Representation of a fitness landscape. The genotypes are arranged in the x-y plane and fitness is depicted on the z axis. This landscape is rugged, having three adaptive peaks separated by fitness "valleys". Two alternatives evolutionary routes are represented in red. The white circles denote the different the different states a gene can take. Adapted from Steinberg and Ostermeier (2016).

Following studies of adaptation were based on Gillespie's model of molecular evolution, among which is the seminal work of Allen Orr. Orr supported Gillespie ideas of adaptation occurring through large jumps in fitness, and further showed that, in most cases, an increased in fitness was derived by a single substitution (Orr 2002). He, therefore, characterized adaptation with the "Pareto principle", where "the majority of an effect (increased fitness) is due to a minority of causes (one substitution)" (Orr 2005). Moreover, Orr assessed Fisher's geometrical model of phenotypic evolution (1930) and suggested that adaptation comprises not just mutations of small effects, but also a few mutations of relatively large effect on fitness (Orr 1998, 1999). His work further showed that the mean selection coefficient in the course of an "adaptive walk" decreases almost proportionally, roughly approaching a geometric sequence. This view of adaptation was characterised by a pattern of diminishing returns, where mutations of larger effect reach fixation earlier than small effect ones (Barton 1998; Orr 1998, 1999, 2005; Barton and Keightley 2002). This pattern agrees with the findings of previous studies suggesting that the distribution of selection coefficients of beneficial mutations should be exponentially distributed (Rozen et al. 2002; Orr 2003).

The neutralist

The controversy started when, in 1968, Kimura suggested that the bulk of segregating polymorphisms and substitutions do not alter protein function, being therefore neutral and subject only to random genetic drift. He presented the Neutral Theory of molecular evolution, which states that most of the new mutations are either deleterious, therefore unlikely to become fixed due to purifying selection, or neutral, where selection is so weak that these become fixed by genetic drift (Kimura 1968, 1983; King and Jukes 1969). Conversely, beneficial mutations are thought to be sufficiently rare to contribute much to the segregating variation, mainly because they reach fixation at a higher rate when compared to neutral mutations (Kimura 1968, 1983; King and Jukes 1969). Similar to what Morgan proposed (Morgan 1925, 1932), the Neutral Theory is based on the fact that evolution proceeds through mutation, and that the main role of natural selection is to remove variants that damage gene function. With this theory, Kimura solved several problems in theoretical population genetics, such as the probability of fixation of a new mutation as well as the time needed for fixation (Kimura 1968, 1983). The simplicity of this theory provided a remarkable explanation for the reasonably constant evolutionary rate across lineages in individual proteins, such as haemoglobins (Kimura 1969) and cytochrome C (King and Jukes 1969). Different patterns of protein polymorphism were then assessed by Kimura and Ohta (1971), leading to the conclusion that the neutral mutation-random drift hypothesis of molecular evolution can be used to explain such patterns.

Later on, Ohta (1973, 1976) extended this theory by proposing that there is a class of mutations that are affected both by drift and selection: the slightly deleterious mutations. In the so-called nearly neutral theory of molecular evolution, Ohta suggested that a considerable fraction of mutant substitutions in a population were produced by the random fixation of slightly deleterious mutations. She further demonstrated that in populations with larger effect sizes, the impact of selection was stronger, while in smaller populations, the effect of drift prevailed. This observation led to the conclusion that evolution is determined both by population size and mutation rate (Ohta 1992).

Today's views

The Neutral Theory revolutionised the way evolution at the molecular was perceived. In the 1980s, the data at the DNA level brought knowledge on the substantial variation on non-functional sequences, side-tracking the debate to the neutralist view (Li et al. 1981; Miyata and Yasunaga 1981; Kimura 1983; Nei 1987). Indeed, neutral evolution provided a simple and elegant way to explain levels of genetic variation at the divergence and polymorphism levels. Natural selection, however, is a complex process that can take a myriad of forms, making the development of mathematical methods an arduous task.

Hudson, Kreitman, and Aguadé (1987) motivated the first attempts of studying the impact of positive natural selection on molecular evolution. They introduced a statistical method that tests neutral evolution under the assumption that polymorphisms and substitutions are uniformly distributed under neutrality. To do so, the authors compared two types of loci: a non-coding region, which is assumed to evolve neutrally, and a protein-coding gene, which is assumed to be under selection. If the patterns of polymorphism and substitution in the coding locus differ from that in the neutral region, then it is assumed to be under selection. This statistical test, also known as the HKA test, provided the first evidence for the role of natural selection in maintaining polymorphic variants in *Drosophila*. This approach paved the way for the study of the impact of positive selection on segregating genetic variants (e.g., McDonald and Kreitman 1991; Eanes et al. 1993), leading to the question of whether adaptation plays a significant role in molecular evolution.

Today, the debate is still ongoing. Some authors argue that the Neutral Theory should be revisited (Hahn 2008; Kern and Hahn 2018), while others emphasize its undeniable role, even in light of the recent findings suggesting pervasive effects of positive selection along the genome (Graur et al. 2013; Jensen et al. 2019). With the thriving of genome-scale data, however, the role of adaptive mutations in molecular evolution can be addressed with a lot more accuracy. Studies assessing the genetic basis of phenotypic differences revealed several quantitative trait loci that may have experienced adaptive evolution (e.g., Sucena and Stern 2000; Colosimo et al. 2004). At the genome level, association studies also provide evidence for several loci linked to phenotypic traits under selection (e.g., Shapiro et al. 2006; Carneiro et al. 2014; Boyle et al. 2017; Alves et al. 2019; Liu et al. 2019). These studies provide the link between phenotypic and molecular evolution. However, they are limited in scope and cannot discern how much of the observed variation is actually adaptive. Some questions remain: "How much of the genetic variation can be explained by adaptive evolution? What is the frequency of adaptive mutations along the genome? Are there regions where adaptive mutations are more likely to occur?" These are some of the questions that can now be tackled by combining population genomics data and a new generation of methods for detecting and quantifying selection, thus providing a deeper understanding of the molecular rate of adaptation.

1.4 Measuring selection and adaptive evolution

This section provides a summary of the methods used to infer the rate of adaptive evolution from sequence data (for a more detailed description see (Moutinho et al. 2019a). Two main approaches are described: (1) phylogenetic methods, applied at the divergence between several species, and (2) population genetic methods, which contrast the within-species polymorphisms to the divergence with an outgroup species.

Phylogenetic methods

The strength and direction of selection in a given gene can be measured with the d_N/d_S (ω) ratio, which contrasts the rate of non-synonymous (d_N) and synonymous substitutions (d_S) (e.g., Miyata et al. 1979; Li et al. 1985; Yang and Nielsen 2002). Under the assumption that synonymous substitutions are effectively neutral, and that mutations rates at synonymous and non-synonymous sites are constant and equal, neutrally evolving genes are expected to have an ω ratio equal to 1. Genes evolving under positive selection at the protein level display an $\omega > 1$, while genes evolving under negative selection have $\omega < 1$. This is because non-synonymous substitutions are either favoured or discarded compared to neutral synonymous substitutions. However, as ω averages the substitution rate across multiple sites that experience both positive and negative selection, tests based on ω can only detect a strong signal of positive selection (e.g., Yang and Nielsen 2002). This is because the majority of non-synonymous substitutions are expected to be either neutral or deleterious, thus making the average d_N lower than d_S, leading to an ω generally lower than 1, even in the presence of positive selection (e.g., Yang and Nielsen 2002).

More complex phylogenetic models have been developed to account for variable selective pressure among sites (Nielsen and Yang 1998; Yang et al. 2000, 2005), branches (Yang and Nielsen 1998), and the combination of the two (Yang and Nielsen 2002; Zhang et al. 2005; Kosakovsky Pond et al. 2011), thus accounting for the great variation in selective constraints in space and/or in time. Even though these methods have the potential to detect adaptation at the site level, they tend to be more conservative in measuring selection throughout a specific region or lineage (Rodrigue and Lartillot 2017). This is due to the fact that adaptive events are often spread across several positions in the genome, rather than being concentrated on specific sites (Rodrigue and Lartillot 2017). Moreover, branch-site models underestimate the rate of adaptation in proteins that experience frequent adaptation over long evolutionary periods, as they assume that evolution is neutral on most branches and that adaptive processes are rare and usually isolated (Nielsen and Yang 1998; Yang et al. 2000, 2005; Rodrigue and Lartillot 2017). Finally, since these methods rely on multi-species alignments, they only account for more ancient genes that are shared by all species, being, therefore, more conserved. Fast-evolving genes are thus typically discarded from such analyses, as their alignment becomes less reliable with increasing divergence times between species.

Population genetics methods

Population genetics approaches require data from only two closely-related species: typically several individuals in the target species and one individual from an outgroup species (McDonald and Kreitman 1991). McDonald and Kreitman (1991) were the first to extend the HKA test (1987) to detect adaptive evolution in proteins. The MK test (1991) contrasts the number of polymorphisms and substitutions at two classes of sites: synonymous, which are assumed to evolve neutrally, and non-synonymous, which are potentially under selection. It is usually represented through the so-called MK table:

	Polymorphisms	*Substitutions*
Synonymous	P_s	D_s
Non-synonymous	P_n	D_n

The MK test is based on the fact that a beneficial mutation reaches fixation faster than neutral mutations, thus contributing comparatively more to divergence than to polymorphism levels. Hence, it can test three scenarios: (1) neutral evolution, where D_n/D_s is expected to be equal to P_n/P_s, (2) positive selection, in which D_n/D_s is higher than P_n/P_s, and (3) balancing selection, where D_n/D_s is lower than P_n/P_s.

Extensions of this method estimate the proportion of amino-acid substitutions driven by positive selection: α = 1- $(D_sP_n)/(D_nP_s)$ (Charlesworth 1994; Smith and Eyre-Walker 2002). As the numbers of polymorphic sites and non-synonymous substitutions are generally low, estimates of α for genes taken individually have usually large sampling variances. This prevents the use of this statistic for single genes and requires pooled data across multiple genes (Smith and Eyre-Walker 2002; Stoletzki and Eyre-Walker 2011). Such pooling can be done by summing numbers of polymorphisms and substitutions (Fay et al. 2001), or by taking the average across genes (Smith and Eyre-Walker 2002). However, a limitation of these approaches is that they do not consider the segregation of slightly deleterious mutations, which can bias estimates of α depending on the demographic history of the population (Eyre-Walker and Keightley 2009). On the one hand, α can be underestimated if the population size remained relatively constant or has undergone a decrease compared to the ancestral population. This is because slightly deleterious mutations may be observed as polymorphism while having a much lower chance of fixation when compared to neutral mutations. One way to mitigate this effect is to remove polymorphisms that are segregated at low frequencies (Charlesworth 1994; Smith and Eyre-Walker 2002). On the other hand, α may be overestimated if the population has gone through a demographic expansion: as polymorphism levels are very low, there is an apparent excess of substitutions (Eyre-Walker 2002). It is, therefore, crucial to account for the full range of fitness effects of mutations, as well as the demography of the population, to reach more precise estimates of α.

More recent methods specifically model the distributions of fitness effects (DFE) from the site frequency spectrum (SFS) of the derived allele in order to infer the molecular adaptive rate. These likelihood methods assume that the numbers of segregating mutations and substitutions are Poisson distributed and that polymorphism levels can be summarized by summing the categories of the unfolded (when the ancestral allele is known) or folded (counts of the minor allele frequency) SFS. The differences between methods rely on how demography is accounted for and the type of distribution of selection coefficients (N_es), *i.e.* DFE, used to infer the rate of adaptive evolution (Moutinho et al. 2019a). The first models only accounted for the DFE of deleterious and neutral mutations (Bierne and Eyre-Walker 2004; Welch 2006a; Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009). Further extensions also consider the distribution of positively selected mutations (Schneider et al. 2011; Galtier 2016; Tataru et al. 2017; Tataru and Bataillon

2019), where some are based on fitness landscape models (see Bataillon and Bailey 2014, for a detailed review), and others are driven by statistical convenience, where they fit the data with a flexible distribution (reviewed in Moutinho et al. 2019a).

Statistics used to estimate the rate of adaptive substitutions

From these methods, two major statistics are generally used to infer the rate of adaptive evolution: ω_a and α. ω_a is the rate of adaptive non-synonymous substitutions relative to the mutation rate and is given by ω_a = ω - ω_{na}, where ω_{na} denotes for the portion of the ω ratio contributed by neutral and deleterious mutations. α is the proportion of adaptive amino-acid substitutions and is estimated as ω_a/ω. Although all of these statistics provide an estimate for the molecular adaptive rate, they cannot be used in the same context. For instance, α is contingent on both ω_a and ω_{na}, making it unsuitable for distinguishing between the effects of positive and negative selection. In turn, ω_a cannot be used to evaluate the impact of mutation rate, as the mutation rate itself normalizes it (e.g., Castellano et al. 2016). It is, therefore, important to accurately assess the context of the question one aims to address, to choose the best measure of the rate of adaptation.

1.5 Variation of the adaptive substitution rate between species

Over the last few years, there has been an increased interest in understanding how the molecular adaptive rate varies between and within species (e.g., Gossmann et al. 2012; Galtier 2016; Zhen et al. 2018; Moutinho et al. 2019a). Previous studies have shown that the rate of adaptive evolution varies across species, where, for example, the fruit fly (e.g., Brookfield and Sharp 1994; Smith and Eyre-Walker 2002; Sella et al. 2009), the wild mouse (Halligan et al. 2010), and the European rabbit (Carneiro et al. 2012a) have a much higher rate of adaptation when compared to plant species (Gossmann et al. 2010) and primates (e.g., Boyko et al. 2008; Hvilsom et al. 2012; Galtier 2016). The determinants of such variability, however, remain unclear.

Multiples studies proposed that the cross-species variation of the adaptation rate is explained by differences in effective population size (N_e) (Eyre-Walker et al. 2006; Eyre-Walker and Keightley 2009; Gossmann et al. 2012). This hypothesis assumes that species with smaller N_e accumulate more slightly deleterious mutations and less advantageous mutations due to the effect of drift, therefore increasing ω_{na} while decreasing ω_a and, consequently, reducing estimates of α. In turn, in species with larger N_e, the impact of purifying and positive selection is more efficient, thus eliminating deleterious mutations from the allele pool at a faster rate while allowing for the fixation of advantageous mutations.

Another hypothesis relied on the so-called cost of complexity (Orr 2000), which is based on Fisher's geometric model of adaptation (Fisher 1930). According to this theory, more complex organisms, *i.e.* larger long-lived species, typically increase in fitness at a slower rate, thus theoretically needing more consecutive beneficial mutations to reach their fitness optimum. This assumption was previously proposed to explain the differences in the adaptation rates between humans and flies (Lourenço et al. 2013; Rousselle et al. 2018, 2019; Zhen et al. 2018). Despite these efforts, the underlying cause of the variation in the molecular adaptive rate between species is not fully resolved and more research is required to clarify this issue.

1.6 Variation of adaptive substitution rate within genomes

Since the time of the first sequence data, it was known that genetic diversity varies comparatively more between genes than between species (Kimura 1983; Ohta 1992). In the early 1970s, it was observed that rates of protein evolution were highly dependent on their function and structure (e.g., King and Jukes 1969; Dickerson 1971). Moreover, the effect of selection on closely linked sites was also shown to cause allele frequency changes throughout the genome (e.g., Maynard Smith and Haigh 1974; Charlesworth et al. 1993; Barton 1995; Andolfatto 2007). Linked selection can take the form of selective sweeps, in a process known as genetic draft (Gillespie 2000a) when genetic variation is reduced due the spread of beneficial mutations (Maynard Smith and Haigh 1974) (Figure 2a), and background selection, which causes the removal of neutral variants that are linked to deleterious mutations (Charlesworth et al. 1993; Charlesworth 2012) (Figure 2b). In turn, the effect of linkage is counteracted by recombination, which increases the levels of genetic variation (Begun and Aquadro 1992).

Molecular rates of adaptation seem to follow such pattern, where there is substantially more variation within genomes than between species (Moutinho et al. 2019a). At the genome level, recombination, mutation rate, and gene density were found to positively impact the rate of adaptive evolution (Marais and Charlesworth 2003; Campos et al. 2014; Castellano et al. 2016). As the recombination rate breaks down linkage disequilibrium, it is expected to favour the fixation of adaptive substitutions (Marais and Charlesworth 2003; Campos et al. 2014; Castellano et al. 2016). Genes in low recombining regions suffer from the Hill-Robertson interference (HRi; Hill and Robertson 1966): the interaction between favourable mutations occurring at linked sites, eventually leads to the fixation of only one of the mutations, unless a recombination event generates a haplotype carrying both of them (Figure 2c). Consequently, genes in low recombining regions tend to have lower rates of adaptive substitutions. Following the same reasoning, regions with high gene density might be subject to stronger HRi and lower molecular adaptive rates (Castellano et al. 2016). Conversely, genes with high mutation rates might adapt faster by increasing genetic diversity levels, which increases the chance for an adaptive process to occur.

At the gene level, previous studies have shown that protein function substantially impacts the rate of adaptive substitutions, where genes implicated in the immune response present the highest adaptation rates in several species (Nielsen et al. 2005; Sackton et al. 2007; Kosiol et al. 2008; Obbard et al. 2009; Slotte et al. 2011). Besides, studies focusing on sex-related genes also reported high rates of adaptive evolution across taxa (Pröschel et al. 2006; Haerty et al. 2007; Hvilsom et al. 2012; Gossmann et al. 2014; Crowson et al. 2017). At the intra-genic level, however, little is known about the factors influencing the frequency and nature of adaptive mutations.

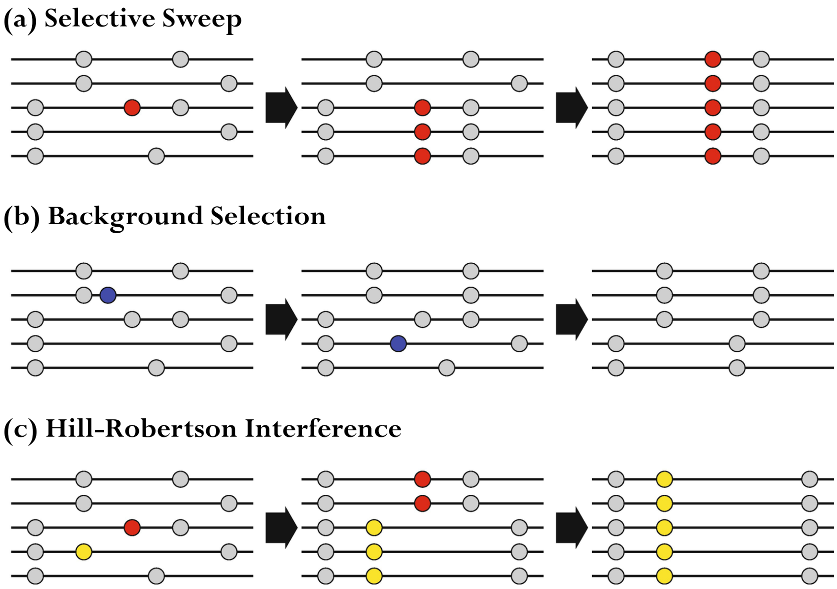

Figure 2. Impact of linked selection on genetic diversity. Black lines represent individual genomes. Filled circles denote SNP variants. Distinct variants at the same position are depicted with different colours: neutral variants in grey, positive variants in red or yellow, and negative variant in blue. **(a)** A positively selected new variant spreads in the population and removes genetic diversity at linked loci, generating a selective sweep**. (b)** Reduction of neutral diversity because of linkage to deleterious mutations (background selection). **(c)** Competitive segregation of positively selected variant at distinct loci, resulting in the loss of advantageous variants (Hill–Robertson interference). Figure and legend adapted from Gustavo et al. (2020).

1.7 Scope of the thesis

My thesis addresses patterns of selection at different organizational levels, particularly aiming to unravel the main determinants of adaptive evolution between-species, within-genomes, and within-genes. These three domains are individually addressed in the following chapters of my thesis in the form of three major questions:

(1) **Chapter II: Does protein architecture impact the rate of adaptive evolution?**

In the second chapter, I looked at molecular evolution on a fine-scale by studying the impact of protein architecture on the rate of adaptive evolution. By assessing the frequency and nature of adaptive mutations at the intramolecular level both in animals and in plants, I aimed to understand how protein biophysics influences fitness and adaptation.

(2) **Chapter III: How do rates of adaptation vary across time?**

In the third chapter, I took a step back to look at a broader scale of molecular evolution, aiming to understand how rates of adaptation vary in time. I studied genes with different evolutionary origins in animal and plant species to assess the dynamics of the distribution of beneficial mutations across the phylogeny of the species.

(3) **Chapter IV: What is the interplay between intramolecular variation and patterns of adaptation at the species level?**

In the fourth chapter, I looked even at a broader scale by studying patterns of adaptation at the species level. By analysing several animal species, I aimed to understand the interplay between patterns of intramolecular variation and the differences in the molecular adaptive rate between species.

CHAPTER II

Does Protein Architecture Impact the Rate of Adaptive Evolution?

2.1 Abstract

Adaptive mutations play an important role in molecular evolution. However, the frequency and nature of these mutations at the intra-molecular level is poorly understood. To address this, we analysed the impact of protein architecture on the rate of adaptive substitutions, aiming to understand how protein biophysics influences fitness and adaptation. Using *Drosophila melanogaster* and *Arabidopsis thaliana* population genomics data, we fitted models of distribution of fitness effects and estimated the rate of adaptive amino-acid substitutions both at the protein and amino-acid residue level. We performed a comprehensive analysis covering genome, gene and protein structure, by exploring a multitude of factors with a plausible impact on the rate of adaptive evolution, such as intron number, protein length, secondary structure, relative solvent accessibility, intrinsic protein disorder, chaperone affinity, gene expression, protein function and protein-protein interactions. We found that the relative solvent accessibility is a major determinant of adaptive evolution, with most adaptive mutations occurring at the surface of proteins. Moreover, we observe that the rate of adaptive substitutions differs between protein functional classes, with genes encoding for protein biosynthesis and degradation signalling exhibiting the fastest rates of protein adaptation. Overall, our results suggest that adaptive evolution in proteins is mainly driven by inter-molecular interactions, with host-pathogen coevolution likely playing a major role.

2.2 Introduction

A long-standing focus in the study of molecular evolution is the role of natural selection in protein evolution (Eyre-Walker 2006). One can measure the strength and direction of selection at the divergence level through the d_N/d_S ratio (ω). However, because ω represents a summary statistic across nucleotide sites, it can only provide the average trend, while proteins will typically undergo both negative and positive selection. Branch-site models address this issue by fitting phylogenetic models with heterogeneous d_N/d_S ratio among codons and branches, thus considering the great heterogeneity in selective constraints among sites, both in space and time (Nielsen and Yang 1998; Yang et al. 2005; Zhang et al. 2005). Although these methods potentially allow to study adaptation at the site level, they require large amounts of data across species and are therefore restricted to more conserved genes along the phylogeny. Conversely, the McDonald and Kreitman (MK) test (McDonald and Kreitman 1991) is applied at the population level and it only requires data from two closely-related species, usually several individuals from the study species and one individual from the other. Because adaptive mutations contribute relatively more to substitution than to polymorphism, the MK test disentangles positive and negative selection by contrasting the number of substitutions to the number of polymorphisms at synonymous and non-synonymous sites. Charlesworth (1994) extended this method to estimate the proportion of substitutions that are adaptive (α). Yet, one limitation of this approach was that it didn't account for the segregation of slightly deleterious mutations, which can either over- or underestimate measurements of α according to the demography of the population (Eyre-Walker 2002; Smith and Eyre-Walker 2002). Recent methods solved this issue by taking into consideration the distribution of fitness effects (DFE) of both slightly deleterious (Fay et al. 2001; Smith and Eyre-Walker 2002; Bierne and Eyre-Walker 2004; Eyre-Walker et al. 2006; Eyre-Walker and Keightley 2009; Stoletzki and Eyre-Walker 2011) and slightly beneficial mutations (Galtier 2016; Tataru et al. 2017). By allowing the estimation of the rate of non-adaptive ($\omega_{na} = \widehat{d_N^{na}}/d_S$) and adaptive ($\omega_a = \omega - \omega_{na}$) non-synonymous substitutions, in addition to measurements of α (ω_a/ω), these methods triggered new insights on the impact of both negative and positive selection on the rate of protein evolution.

Several studies have reported substantial levels of adaptive protein evolution in various animal species, including the fruit fly (Smith and Eyre-Walker 2002; Sawyer et al. 2003; Bierne and Eyre-Walker 2004; Haddrill et al. 2010), the wild mouse (Halligan et al. 2010) and the European rabbit (Carneiro et al. 2012b), but also in bacteria (Charlesworth and Eyre-Walker 2006) and in plants (Ingvarsson 2010; Slotte et al. 2010; Strasburg et al. 2011). Whereas for other taxa, such as primates (Boyko et al. 2008; Hvilsom et al. 2012; Galtier 2016), and many other plants (Gossmann et al. 2010), the rate of adaptive mutations was observed to be very low, wherein amino-acid substitutions are expected to be nearly neutral and fixed mainly through random genetic drift (Boyko et al. 2008). Several authors proposed that this across-species variation in the molecular adaptive rate is explained by an effective population size (N_e) effect, where higher rates of adaptive evolution are observed for species with larger N_e due to a lower impact of genetic drift (Eyre-Walker 2006; Eyre-Walker and Keightley 2009; Gossmann et al. 2012). Galtier (2016), however, reported that N_e had an impact on α and ω_{na} but not ω_a. Hence, he proposed that the relationship with N_e is mainly explained by deleterious effects, wherein slightly deleterious non-synonymous substitutions accumulate at lower rates in large-N_e species due to a higher efficiency of purifying selection, thus decreasing ω_{na} and consequently inflating α.

The rate of adaptive substitutions, however, was observed to vary extensively along the genome. On a genome-wide scale, it was reported that ω_a correlates positively with both the recombination and mutation rates, but negatively with gene density (Campos et al. 2014; Castellano et al. 2016). When looking at the gene level, previous studies have demonstrated the role of protein function in the rate of adaptive evolution, wherein genes involved in immune defence mechanisms appear with higher rates of adaptive mutations in Drosophila (Sackton et al. 2007; Obbard et al. 2009), humans and chimpanzees (Nielsen et al. 2005). In Drosophila, sex-related genes also display higher levels of adaptive evolution, being directly linked with species differentiation (Pröschel et al. 2006; Haerty et al. 2007). At the intra-genic level, however, the factors impacting the frequency and nature of adaptive mutations remain poorly understood.

There are several structural factors that have been reported to influence the rate of protein evolution but have not been investigated at the population level. Molecular evolution studies of protein families revealed that protein structure, for instance, significantly impacts the rate of amino-acid substitutions, with exposed residues evolving faster than buried ones (Liberles et al. 2012). As a stable conformation is often required to ensure proper protein function, mutations that impair the stability or the structural conformation of the folded protein are more likely to be counter-selected. Moreover, distinct sites in a protein sequence differ in the extent of conformational change they endure upon mutation, a pattern generally well predicted by the relative solvent accessibility of a residue (Goldman et al. 1998; Mirny and Shakhnovich 1999; Franzosa and Xia 2009). In this way, residues at the core of proteins evolve slower than the ones at the surface due to their role in maintaining a stable protein structure (Perutz et al. 1965; Overington et al. 1992; Goldman et al. 1998; Bustamante et al. 2000; Dean et al. 2002; Choi et al. 2006; Lin et al. 2007; Conant and Stadler 2009; Franzosa and Xia 2009; Ramsey et al. 2011). Inter-specific comparative sequence analyses also revealed that positively selected sites are often found at the surface of proteins (Proux et al. 2009; Adams et al. 2017). Hence, exploring the role that these structural elements play in shaping the rate of adaptive evolution is crucial in order to fully understand what are the main drivers of adaptation within proteomes.

Our study addresses protein adaptive evolution at a fine scale by analysing the impact of several functional variables among protein-coding regions at the population level. To further assess the potential generality of the inferred effects, we carried our comparison on two model species with distinct life-history traits: the dipter *Drosophila melanogaster* and the brassicaceae *Arabidopsis thaliana*. We fitted models of DFE and estimated the rate of adaptive substitutions, both at the protein and amino-acid residue scale, across several variables and found that solvent exposure is the most significant factor influencing protein adaptation, with exposed residues undergoing ten times faster ω_a than buried ones. Moreover, we observed that the functional class of proteins has also a strong impact on the rate of protein adaptation, with genes encoding for processes of protein regulation and signalling pathways exhibiting the highest ω_a values. We therefore hypothesized that inter-molecular interactions are the main drivers of adaptive substitutions in proteins. This hypothesis is consistent with the proposal that, at the inter-organism level, coevolution with pathogens constitute a so-far under-assessed component of protein evolution (Sackton et al. 2007; Obbard et al. 2009; Enard et al. 2016; Mauch-Mani et al. 2017).

2.3 Results and Discussion

In order to identify the genomic and structural variants driving protein adaptive evolution we looked at 10,318 protein-coding genes in 114 *Drosophila melanogaster* genomes, analysing polymorphism data from an admixed sub-Saharan population from Phase 2 of the *Drosophila* Population Genomics Project (DPGP2, Pool et al. 2012) and divergence out to *D. simulans*; and 18,669 protein-coding genes in 110 *Arabidopsis thaliana* genomes, with polymorphism data from a Spanish population (1001 Genomes Project, Weigel and Mott 2009) and divergence to *A. lyrata*. The rate of adaptive evolution was estimated with the Grapes program (Galtier 2016). The Grapes method extends the approach pioneered by the DoFE program (Fay et al. 2001; Smith and Eyre-Walker 2002; Bierne and Eyre-Walker 2004; Eyre-Walker et al. 2006; Eyre-Walker and Keightley 2009; Stoletzki and Eyre-Walker 2011), by explicitly accounting for mutations with slightly advantageous effects. Grapes estimates the rate of non-adaptive non-synonymous substitutions (ω_{na}), which is then used to estimate the rate of adaptive non-synonymous substitutions (ω_a) and the proportion of adaptive non-synonymous substitutions (α). A high α can be potentially explained both by a higher ω_a or a lower ω_{na}, and therefore does not allow to disentangle the two effects. Thus, we explored whether, and how, ω_a and ω_{na}, as well as the total ω, depend on the different functional variables analysed here.

Results from model comparison of DFE showed that the Gamma-Exponential model is the one that best fits our data according to Akaike's information criterion (Akaike 1973) (Table S1 in supplementary file S1, Appendix I). This model combines a Gamma distribution of deleterious mutations with an exponential distribution of beneficial mutations. In agreement with previous surveys within animal species, this model suggests the existence of slightly deleterious, as well as slightly beneficial segregating mutations in *D. melanogaster* and *A. thaliana* genomes (Galtier 2016). Genome-wide estimates of ω_a for *A. thaliana* and *D. melanogaster* are 0.05 and 0.09, respectively, and are in the range of previously reported estimates for these species (Bierne and Eyre-Walker 2004; Gossmann et al. 2012; Smith and Eyre-Walker 2002).

In order to investigate the main drivers of protein adaptive evolution, we divided the datasets into sets of genes and amino-acid residues according to the variables analysed, and fitted models of DFE in each subset independently. We distinguished two types of analyses: gene-based and site-based, where we looked into how the molecular adaptive rate varies across different categories of genes and amino-acid residues, respectively. Gene-based analyses allowed us to explore the impact of the background recombination rate, number of introns, mean expression levels and breadth of expression. At the protein level, we investigated the effect of binding affinity to the molecular chaperone *DnaK*, protein length, cellular localization of proteins, protein functional class and number of protein-protein interactions. Finally, site-based analyses enabled us to study the effect of the secondary structure of the protein, by comparing residues present in β-sheets, α-helices and loops; the tertiary structure, by considering the relative solvent accessibility of a residue (RSA) and the residue intrinsic disorder; and whether an amino-acid residue participated or not in an annotated active site.

The impact of gene and genome architecture on adaptive evolution

To study the impact of gene and genome architecture on the rate of adaptive evolution we looked at recombination rate and the number of introns. Recombination rate was previously reported to favour the fixation of adaptive mutations in Drosophila by breaking down linkage disequilibrium (Marais and Charlesworth 2003; Castellano et al. 2016). Our results are consistent with previous observations by showing a significant positive correlation in estimates of ω_a with increasing levels of recombination rate for *D. melanogaster* (Table 1, supplementary figure S1, and file S2 in Appendix I). This was also observed in *A. thaliana* (Table 1, supplementary figure S1, and file S2 in Appendix I), thus corroborating the effect of recombination in the rate of adaptive evolution.

Previous studies proposed that genes containing more introns are under stronger selective constraints due to the high cost of transcription, especially in highly expressed genes (Castillo-Davis et al. 2002). Hence, we would expect regions with more introns to be under stronger purifying selection. Conversely, by increasing the total gene length, introns might also effectively increase the intra-genic recombination rate, which could in turn increase the efficacy of positive selection and have a positive impact on ω_a. To disentangle the two effects, analyses were performed by comparing genes with different intron content. Results showed a significant negative correlation of ω_{na} with increasing number of introns in *D. melanogaster* (Table 1, supplementary figure S2, and file S2 in Appendix I). Conversely, the number of introns did not significantly correlate with ω_a (Table 1, supplementary figure S2, and file S2 in Appendix I). These findings suggest that the effect of the intron content on the rate of protein evolution is essentially due to stronger purifying selection, while having a negligible influence on the rate of adaptive substitutions.

The impact of protein structure on adaptive evolution

We further explored the impact of three different levels of protein structure (*i.e.*, primary, secondary and tertiary) on the rate of adaptive evolution. We first looked at the primary structure by categorizing proteins according to their length. Former studies correlating gene length and d_N/d_S have shown that smaller genes evolve more rapidly (Zhang 2000; Lipman et al. 2002; Liao et al. 2006). Here, we investigated whether this faster evolution is followed by a higher rate of adaptive substitutions. Results show significant negative correlations with protein length for values of ω and ω_{na} in both species (Table 1, supplementary figure S3, and file S2 in Appendix I). The same trend was observed for ω_a, although it was only significant in *D. melanogaster* (Table 1, supplementary figure S3, and file S2 in Appendix I). These findings suggest that smaller protein-coding regions are indeed under more relaxed purifying selection but might also evolve, in some cases, under a higher rate of adaptive substitutions.

The analysis at the secondary structural level showed significant differences in the evolutionary rate between the structural motifs, with loops demonstrating the highest values of ω, followed by α-helices and β-sheets (Table 2 and Figure 1). When considering adaptive and non-adaptive substitutions separately, β-sheets show significantly lower values of ω_{na} in *A. thaliana* and ω_a in both species, with marginally significant values observed for *D. melanogaster* (Table 2, Figure 1, and supplementary file S3 in Appendix I).

This implies that the structural motif has an impact on the selective constraints in *A. thaliana* and also contributes to the rate of adaptation in the two species. Previous studies investigating protein tolerance to amino-acid change have similarly shown that loops and turns are the most mutable, followed by α-helices and β-sheets (Goldman et al. 1998; Guo et al. 2004; Choi et al. 2006). Some authors posed this relationship as an outcome of residue exposure (Goldman et al. 1998; Guo et al. 2004), while others associate it to the degree of structural disorder, where ordered proteins are under stronger selective constraint (Choi et al. 2006). In order to clarify this, we further look into the impact of tertiary structure, by exploring the relationship between residue exposure to solvent and intrinsic protein disorder with the rate of adaptive evolution.

Considering the relative solvent accessibility, several studies previously demonstrated that residues at the surface of proteins evolve faster than the ones at the core (e.g. Goldman et al. 1998; Choi et al. 2007; Lin et al. 2007; Franzosa and Xia 2009). This higher substitution rate can be either due to a reduced selective constraint at exposed residues and/or to an increased rate of adaptive substitutions. To disentangle the two effects, we compared the site frequency spectra across several categories of RSA. Our results recapitulate those of previous studies on divergence and demonstrate a significant positive correlation with solvent exposure for values of ω (Table 1 and Figure 2a). Moreover, we demonstrate that both a relaxation of the selective constraints (ω_{na}) and a higher rate of adaptive non-synonymous substitutions (ω_a) explain the higher evolutionary rate at the surface of proteins (Table 1, Figure 2a, and supplementary file S2 in Appendix I).

Intrinsically disordered proteins are defined by lacking a well-defined three-dimensional fold (Dunker et al. 2002; Dyson and Wright 2005), more specifically, proteins that have a higher degree of loop dynamics ("hotloops") (Linding et al. 2003). As these structures are more flexible we expect them to be under less structural constraint and to accumulate more substitutions (Guo et al. 2004; Wilke et al. 2005; Choi et al. 2006; Afanasyeva et al. 2018), either deleterious and/or beneficial. To test this hypothesis, we asked two different questions: (1) Are intrinsically disordered protein regions more likely to respond to adaptation? (2) Are proteins with more disordered regions undergoing more adaptive substitutions? For the first question, we divided amino-acid residues based on their predicted value of intrinsic disorder. We report a significant positive correlation with ω, ω_a and ω_{na} with residue intrinsic disorder for both species (Table 1, Figure 2b, and supplementary file S2 in Appendix I). For the second question, proteins were categorized according to their proportion of disordered residues (see Material and Methods). Our results reveal a significant positive correlation of protein disorder with ω in both species, ω_{na} in *A. thaliana* and ω_a in *D. melanogaster* (Table 1, supplementary figure S4, and file S2 in Appendix I). These findings suggest that, at the residue level, intrinsically disordered regions are more likely to respond to adaptation and are also under less selective constraint in both species. However, when considering the whole protein, we observe that intrinsically disordered proteins have different effects between species. In particular, they contribute to the relaxation of purifying selection in *A. thaliana* and to a higher rate of adaptation in *D. melanogaster*. The reason for the difference between species is unclear and will require further analyses.

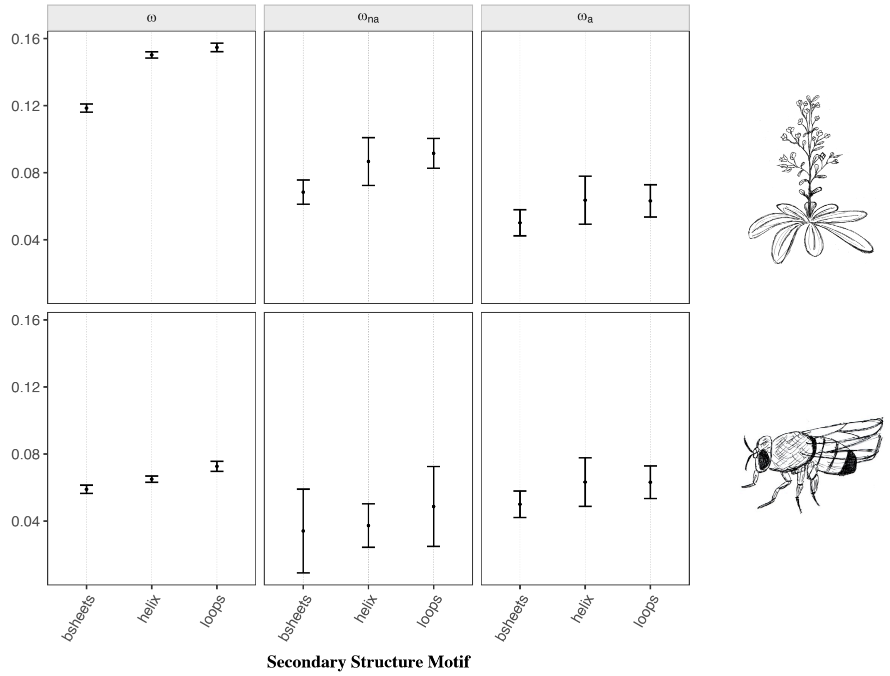

Figure 1. Estimates of the rate of protein evolution (ω), non-adaptive non-synonymous substitutions (ω_{na}) and adaptive non-synonymous substitutions (ω_a) for each of the secondary structural motif (β-sheets, α-helices and loops) in *A. thaliana* (top) and *D. melanogaster* (bottom). Mean values of ω, ω_{na} and ω_a for each motif are represented with the black points. Error bars denote for the 95% confidence interval for each category, computed over 100 bootstrap replicates. The hand-drawings of *A. thaliana* and *D. melanogaster* were made by AFM.

Finally, we tested whether the rate of adaptive substitutions is affected by the binding affinity of proteins to molecular chaperones. It has been suggested that binding to a chaperone leads to a higher evolutionary rate due to the buffering effect for slightly deleterious mutations (Bogumil and Dagan 2010; Kadibalban et al. 2016). Here, we investigate whether binding to the chaperone *DnaK* could also favour the fixation of adaptive mutations. In agreement with previous studies, we find a higher ω and ω_{na} in proteins binding to *DnaK* in *D. melanogaster* (Table 2; supplementary figure S5 in Appendix I), but no impact on ω_a (Table 2, supplementary figure S5, and file S3 in Appendix I), suggesting that the interaction with a molecular chaperone does not influence the fixation of beneficial mutations.

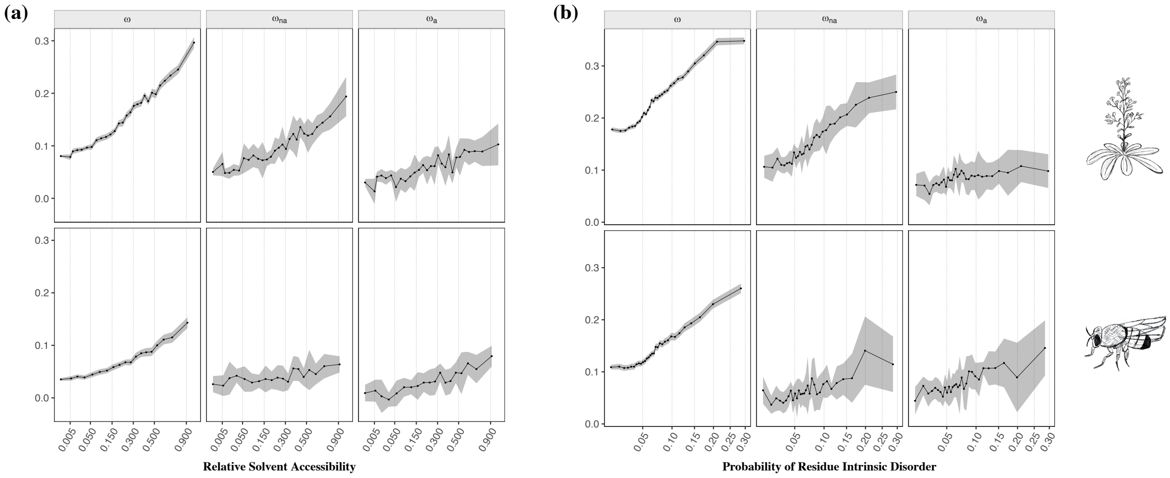

Figure 2. Relationship between ω, ω_{na} and ω_a with **(a)** the relative solvent accessibility (RSA) and **(b)** the probability of residue intrinsic disorder for *A. thaliana* (top) and *D. melanogaster* (bottom). The x axis is scaled using a squared root function. Mean values of each estimate for each category are represented with connected black dots. The shaded area represents the 95% confidence interval of each category, computed over 100 bootstrap replicates.

Table 1. Number of genes and categories analysed for each continuous variable.

	A. thaliana					*D. melanogaster*				
	Number of Categories	**Number of genes**	ω_a	ω_{na}	**ω**	**Number of Categories**	**Number of genes**	ω_a	ω_{na}	**ω**
Recombination rate	50	18668	0.2065 (*)	-0.2212 (*)	0.0857	30	8485	0.3839 (**)	-0.402 (**)	0.0759
Intron number	13	15347	-0.1538	-0.3590 (.)	-0.7949 (***)	10	10318	-0.3333	-0.866 (***)	-0.7333 (**)
Protein length	30	18669	-0.1310	-0.6735 (***)	-0.6782 (***)	50	10318	-0.4775 (***)	-0.6963 (***)	-0.7763 (***)
Relative Solvent Accessibility	28	9034	0.7513 (***)	0.8466 (***)	0.9841 (***)	19	4944	0.8129 (***)	0.5789 (***)	0.9766 (***)
Protein Intrinsic Disorder (Site)	30	18668	0.6000 (***)	0.9172 (***)	0.9770 (***)	30	8485	0.7057 (***)	0.6690 (***)	0.9540 (***)
Proportion of Disordered Residues (Gene)	30	18668	0.1908	0.7333 (***)	0.7517 (***)	20	8485	0.7263 (***)	0.0631	0.5684 (***)
Breadth of Expression	4	17999	-0.6667	-1.0000 (*)	-1.0000 (*)	6	4601	-0.7333 (*)	-0.4667	-0.7333 (*)
Mean Gene Expression	40	17999	-0.1385	-0.9154 (***)	-0.9282 (***)	15	6247	-0.5048 (**)	-0.6190 (**)	-0.7714 (***)
Protein-Protein Interactions	-	-	-	-	-	19	5628	-0.3099 (.)	-0.1111	-0.3684 (*)

Note. For each variable, the Kendall's τ is shown with the respective significance (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; "." $0.05 \leq P < 0.10$) for ω, ω_{na} and ω_a in *A. thaliana* and *D. melanogaster*.

Table 2. Number of genes and categories analysed for each discrete variable and the corresponding difference between the mean values of each category is reported for ω, ω_{na} and ω_a for *A. thaliana* and *D. melanogaster*.

		A. thaliana					***D. melanogaster***				
	Pairwise comparisons	**Number of Categories**	**Number of genes**	ω_a	ω_{na}	**ω**	**Number of Categories**	**Number of genes**	ω_a	ω_{na}	**ω**
Secondary structure	β-sheets - α-helices	3	9034	-0.01346 (*)	-0.0182 (.)	-0.0317 (*)	3	4944	-0.0132 (.)	-0.0033	-0.0060 (*)
	β-sheets - loops			-0.0130 (*)	-0.0231 (*)	-0.0361 (*)			-0.0131 (.)	-0.0146	-0.0137 (*)
	α-helices - loops			0.0004	-0.0049	-0.0045 (*)			0.00009	-0.0114	-0.0076 (*)
Affinity to Molecular Chaperone	Binder - Non-Binder	2	17775	0.0092	0.0260	0.0352 (*)	2	9420	0.00009	0.0606 (*)	0.0515 (*)
Protein Location [a]		7	18669				7	10318			
Protein Functional Class [a]		27	3780				23	2948			

Note. Significance levels as in Table 1.

[a] Due to the large amount of comparisons, the detailed pairwise comparisons and the corresponding p-values are detailed in supplementary Files S3 and S4 in Appendix I.

Protein function and adaptive evolution

We further explored the impact of protein function on sequence evolution. To do so, we analysed the effect of mean gene expression, breadth of expression, protein location and protein functional class on the rate of adaptive substitutions. Several studies on both Eukaryote (Pal et al. 2001; Subramanian and Kumar 2004; Wright et al. 2004; Lemos et al. 2005) and Prokaryote (Rocha and Danchin 2004) organisms have shown that highly expressed genes have lower rates of protein sequence evolution. Here we investigated if the lower evolutionary rate is followed by a reduced rate of adaptive substitutions. Our results support previous findings by displaying a significant negative correlation of mean gene expression with estimates of ω and ω_{na} in both species (Table 1, Figure 3, and supplementary file S2 in Appendix I). Besides, we find that mean gene expression is also significantly negatively correlated with ω_a in *D. melanogaster* (Table 1, Figure 3, and supplementary file S2 in Appendix I), suggesting that gene expression also constrains the rate of adaptation, in addition to the well-known effect on purifying selection. It has been hypothesized that the higher selective constraint in highly expressed genes could be driven by the reduced probability of protein misfolding, wherein selection acts by favouring protein sequences that accumulate less translational missense errors (Drummond et al. 2005). Hence, the higher selective pressure to increase stability in highly expressed proteins could also be hampering the fixation of adaptive mutations. Moreover, as mean gene expression is positively correlated with the breadth of expression (Kendall's $\tau = 0.3376$, $p < 2.2\text{e-}16$ in *A. thaliana*; Kendall's $\tau = 0.2170$, $p < 2.2\text{e-}16$ in *D. melanogaster*; supplementary figure S6 in Appendix I), and the latter is a good proxy for the pleiotropic effect of a gene, which is known to impose high selective constraints (*i.e.*, Salvador-Martínez et al. 2018), we also analysed the impact of the number of tissues where a gene is expressed on the rate of adaptive evolution. We report a significant negative correlation of the breadth of expression (number of tissues) with ω in both species (Table 1 and supplementary figure S7 in Appendix I), thus corroborating previous findings (Duret and Mouchiroud 2000; Slotte et al. 2011; Salvador-Martínez et al. 2018). When looking at adaptive and non-adaptive substitutions separately, we observe a significant negative impact on values of ω_a in *D. melanogaster* and ω_{na} in *A. thaliana* (Table 1, supplementary figure S7, and file S2 in Appendix I). This suggests that the breadth of expression is acting together with the mean expression levels, although with an apparently lower magnitude effect both in ω_{na} and ω_a.

In order to assess the impact of protein location we classified genes into the following cellular categories: cytoplasmic, endomembrane system, mitochondrial, nuclear, plasma membrane and secreted proteins (Tables S2 and S3 in supplementary file S1, Appendix I). Results show significantly higher rates of protein evolution in nuclear and secreted proteins, with the lowest values observed in the mitochondria, plasma membrane and endomembrane system (pairwise comparisons; $p = 0.0128$ in *A. thaliana*; $p = 0.0104$ in *D. melanogaster*; supplementary figure S8 in Appendix I). However, this result seems to be explained by a reduced purifying selection, with significantly higher values of ω_{na} observed in cytoplasmic, nuclear and secreted proteins (pairwise comparisons; $p = 0.0128$ in *A. thaliana*; $p > 0.0729$ in *D. melanogaster*; supplementary figure S8 in Appendix I), and not by a higher rate of adaptive substitutions, since no significant differences were found between the categories in the estimates of ω_a (supplementary figure S8 and file S3 in Appendix I).

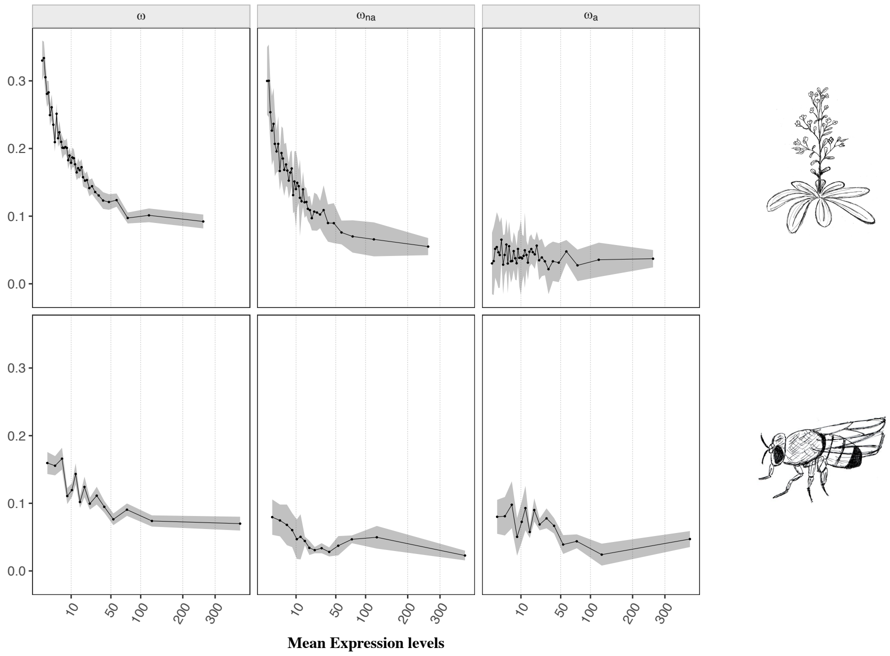

Figure 3. Estimates of ω, ω_{na} and ω_a for each category of genes with distinct mean gene expression levels for *A. thaliana* (top) and *D. melanogaster* (bottom). The x axis is scaled using a squared root function. Legend as in Figure 2.

By analysing the different categories of protein functional class (Tables S2 and S3 in supplementary file S1 in Appendix I), we observe that genes involved in protein biosynthesis (*i.e.*, mRNA and ribosome biogenesis and transcription machinery) and signalling for protein degradation (ubiquitin system) exhibit the highest rates of adaptive substitutions (Figure 4 and supplementary file S4 in Appendix I), functions coded mostly by nuclear and cytoplasmic proteins. Signal transduction pathways also appear to play a role in adaptation, since protein phosphatases also present high rates of adaptive mutations (Hunter 1995). Moreover, in *A. thaliana*, cytochrome P450 proteins are also in the top categories of ω_a (Figure 4 and supplementary file S4 in Appendix I). We fitted a linear model to the ω_a values of the shared categories (21 categories in total) to see if results were consistent between the two species and found a positive correlation (Kendall's τ = 0.257, p = 0.1101; supplementary figure S9a in Appendix I), which is stronger after discarding the two outliers, mRNA biogenesis and glycosyltransferases (Kendall's τ = 0.333, p = 0.0490; supplementary figure S9b in Appendix I). Our findings therefore suggest that adaptive mutations occur mainly through processes of protein regulation and signalling pathways.

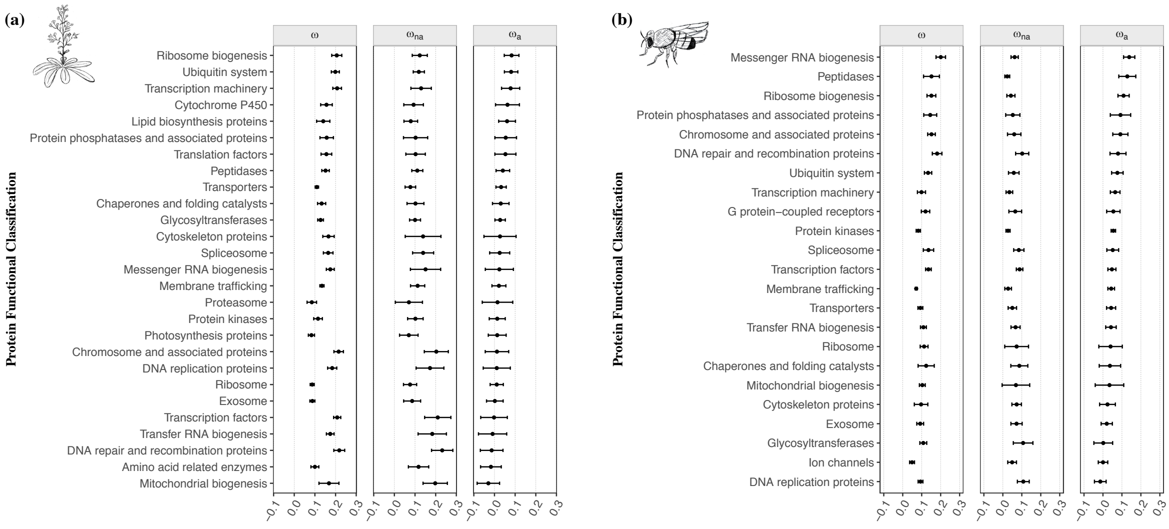

Figure 4. Estimates of ω, ω_{na} and ω_a for each category of protein functional class in **(a)** *A. thaliana* and **(b)** *D. melanogaster*. Categories are ordered according to the values of ω_a. Mean values of ω, ω_{na} and ω_a for each class are represented with the black points. Error bars denote the 95% confidence interval for each category, computed over 100 bootstrap replicates.

What are the major drivers of adaptive evolution along the genome?

Overall, we found multiple factors influencing protein adaptive evolution, specifically recombination rate (positive correlation), protein length (negative correlation), secondary structural motif (lower values observed for β-sheets), relative solvent accessibility (positive correlation), protein intrinsic disorder (positive correlation), gene expression levels (negative correlation) and protein functional class. Since some of these variables are intrinsically correlated we next asked whether some of the inferred effects are spurious. First of all, it is known that protein length and gene expression are negatively correlated, wherein highly expressed genes tend to be shorter, as previously reported for vertebrates (Subramanian and Kumar 2004), yeast (Coghlan and Wolfe 2000; Akashi 2003) and observed in this study (Kendall's τ = -0.015, *p* = 1.22e-02 in *A. thaliana*; τ = -0.093, *p* = 1.70e-28 in *D. melanogaster*; supplementary figure S10 in Appendix I). Since highly expressed genes have lower rates of adaptive substitutions and shorter genes have higher rates of adaptive evolution, we may conclude that these two variables independently impact the rate of adaptation in proteins. Protein length is also negatively correlated with the proportion of exposed residues (Kendall's τ = -0.310, *p* = 0.00 in *A. thaliana*; τ = -0.404, *p* = 1.03e-223 in *D. melanogaster*; supplementary figure S11 in Appendix I), as the surface / volume ratio of globular proteins decrease when protein length increases (Janin 1979). By estimating the rate of adaptive mutations of buried and exposed sites separately, we observe that the effect of protein length is no longer significant (Table 3, Figure 5a, and supplementary file S5 in Appendix I). This suggests that the effect of protein length on the rate of adaptive substitutions is a by-product of the effect of the residue's solvent exposure. Furthermore, mean gene expression is positively correlated with solvent exposure (Kendall's $\tau = 0.016$, $p = 0.1037$ in *A. thaliana*; $\tau = 0.327$, $p =$ 4.50e-45 in *D. melanogaster*; supplementary figure S12 in Appendix I), as expected since highly expressed genes are shorter and shorter genes have a greater proportion of exposed residues (supplementary figures S10 and S11 in Appendix I). These two variables, however, have opposite effects on ω_a, and we therefore conclude that gene expression is acting independently from solvent exposure on the rate of adaptive protein evolution.

We further note that the secondary structure motif is intrinsically correlated with the degree of intrinsic disorder, where loops and turns represent the most flexible motifs (supplementary figure S13 in Appendix I), consistent with previous studies (Choi et al. 2006). When analysing different degrees of protein disorder across the structural motifs, we observe that secondary structure has only an impact on estimates of ω, while intrinsic protein disorder is significantly positively correlated with ω within the three motifs in both species, and ω_a within β-sheets in *A. thaliana* and within α-helices in *D. melanogaster* (supplementary figure S14 and file S5 in Appendix I). Moreover, we report that the secondary structure motif is correlated with solvent exposure (supplementary figure S15 in Appendix I), β-sheets being mostly found at the core of proteins, while α-helices and loops have, on average, higher solvent exposure (Bowie et al. 1990; Guo et al. 2004). By estimating the rate of adaptive substitutions in buried and exposed residues across the three motifs, the impact of secondary structure is no longer noticeable on estimates of ω_a (Table 3, supplementary figure S16, and file S5 in Appendix I), thus suggesting that the effect of secondary structure motif is also a by-product of solvent exposure. When looking at the tertiary structure level, in agreement with Choi et al. (2006), we report that structures with more exposed residues tend to be more flexible (Kendall's $\tau = 0.001$, $p = 0.4726$ in *A. thaliana*; $\tau = 0.015$, $p = 0.0256$ in *D. melanogaster*; supplementary figure S17 in Appendix I). Estimation of the rate of adaptive mutations in buried and exposed sites across different levels of residue intrinsic disorder shows that solvent exposure plays the main role in protein adaptive evolution, with a significant positive impact of protein disorder only observed in values of ω in both species and ω_a in exposed residues for *D. melanogaster* (Table 3, Figure 5b, and supplementary file S5 in Appendix I). To further clarify the relative contribution of solvent exposure and protein disorder on the rate of adaptive evolution we performed an analysis of covariance (ANCOVA), using both measures and their interaction as explanatory variables. Results show that the RSA explains 95% ($p =$3.176e-14) and 99% ($p <$ 2.2e-16) of the variation in ω_a and ω_{na}, respectively, in *A. thaliana*; and 87% ($p =$ 1.011e-13) and 62% ($p = 0.00012$) in ω_a and ω_{na}, respectively, in *D. melanogaster*. These findings suggest that the level of exposure of a residue in the protein structure is the main driver of adaptive evolution, and that structural flexibility potentially constitutes a comparatively small, if any, effect to protein adaptation. By comparing the level of exposure of the residues across the different classes of protein function, no differences were observed (supplementary figure S18 in Appendix I), thus suggesting that these two variables independently affect the rate of protein adaptation.

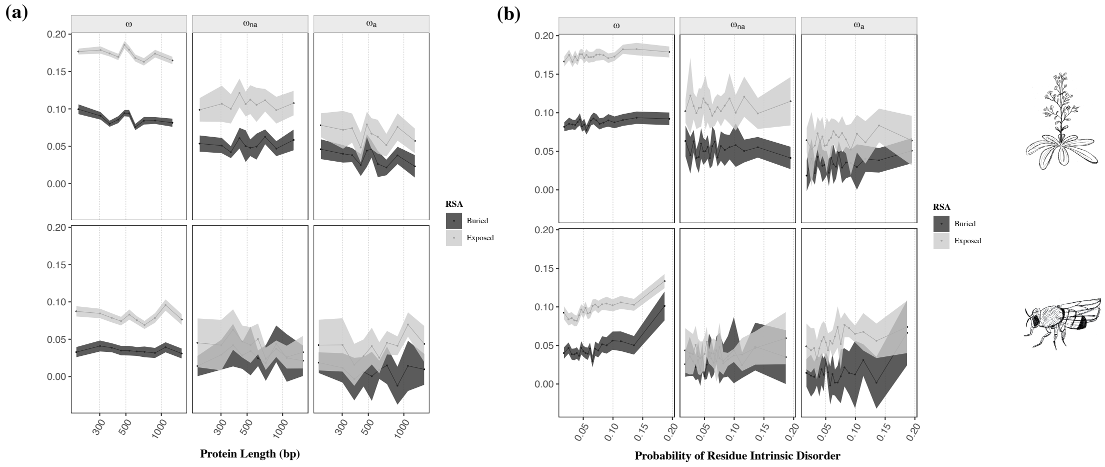

Figure 5. Estimates of ω, ω_{na} and ω_a plotted as a function of **(a)** the relative solvent accessibility and protein length and **(b)** the relative solvent accessibility and the probability of residue intrinsic disorder in *A. thaliana* (top) and *D. melanogaster* (bottom). The x axis is log scaled. Analyses were performed by comparing buried (RSA < 0.05) and exposed (RSA >= 0.05) residues across 10 categories of protein length in (a) and 20 categories of intrinsic disorder in (b) for both species. Legend as in Figure 2.

Summarizing, after accounting for potentially confounding effects, our results show that besides population genetic processes such as recombination and mutation rate (Hill and Robertson 1966; Marais and Charlesworth 2003; Castellano et al. 2016), three major protein features significantly impact the rate of protein adaptive evolution: gene expression, relative solvent accessibility and the protein functional class. When looking at the magnitude effect of each of these variables, we observe that exposed residues have a ten-fold higher rate of adaptive substitutions when compared to completely buried sites (Figure 2a and supplementary file S2 in Appendix I). The effect of gene expression seems to be of lower magnitude, wherein less expressed genes have a two-fold higher rate of adaptive substitutions with a significant negative correlation observed only in *D. melanogaster* (Figure 3 and supplementary file S2 in Appendix I). As a comparison, genes in highly recombining regions have up to a ten-fold higher rate of adaptive substitutions compared to genes within regions with the lowest recombination rates (supplementary figure S1 and file S2 in Appendix I), being therefore similar to that observed with solvent exposure. Previous studies reported that the type of amino-acid change also plays an important role in protein adaptive evolution, where more similar amino-acids present higher rates of adaptive substitutions (Grantham 1974; Miyata et al. 1979; Bergman and Eyre-Walker 2019). In order to evaluate a potential bias on the type of amino-acid at the surface and at the core of proteins, we computed the proportion of conservative and radical residue changes, according to volume and polarity indices, as defined by Grantham (Grantham 1974). We found similar frequencies of conserved and radical changes in buried and exposed residues, thus suggesting that our results at the structural level are not influenced by the type of amino-acid mutation (97% of conservative and 3%

changes on buried residues; 96% of conservative and 4% changes on exposed sites). Our findings therefore suggest that protein architecture strongly influences the rate of adaptive protein evolution, wherein selection acts by favouring a greater accumulation of adaptive mutations at the surface of proteins.

Why does adaptation occur mainly at the surface of proteins?

Our results show that solvent exposure is the protein feature with the strongest impact on the rate of adaptive substitutions at the intra-molecular level. To explain this effect, we discuss three hypotheses in which protein adaptive evolution occurs through (1) the acquisition of new biochemical activities at the surface of proteins, (2) the emergence of new functions via network rewiring at the level of protein-protein interactions, and (3) inter-molecular interactions between organisms, as a consequence of host-pathogen coevolution.

We first hypothesized that protein adaptation results from new catalytic activities, wherein adaptive mutations arise within active sites. Barlett et al (2002) reported that active sites are mostly present in more intrinsically disordered regions of the protein. Moreover, they proposed that apo-enzymes, which are not yet bound to the substrate or cofactor, present a greater residue flexibility and more exposed catalytic residues, which could favour a higher rate of adaptive substitutions. In order to test this, we estimated the rate of adaptive substitutions on active and non-active sites, controlling for solvent exposure, and observed only significant differences in ω within buried residues in *A. thaliana* (Table 3, supplementary figure S19, and file S5 in Appendix I), although with higher values observed for non-active sites. While the non-significant differences in the rate of adaptive mutations could result from incomplete annotations, which tend to be biased towards motifs highly conserved across species (De Castro et al. 2006), this suggests that being present in an active site does not influence the rate of adaptation. Active sites, however, are rather mobile, presenting different levels of solvent exposure and residue flexibility according to the stage of the enzymatic reaction (Bartlett et al. 2002). Therefore, it may be arbitrary to assign them a certain solvent exposure class based on the phase the enzymes were crystallized, limiting our capacity to test their role on adaptive evolution.

Several studies discussed the impact of protein-protein interactions (PPI) on the rate of protein evolution. Valdar and Thornton (2001) and Caffrey (2004) proposed that PPI may be acting as an inhibitor of protein evolution by enhancing the efficiency of purifying selection due to a higher degree of protein connectivity, typically associated with more complex functions. Mintseris and Weng (2005) supported this assumption but proposed that the proteins evolving slowly are the ones involved in obligate interactions, while proteins involved in transient interactions evolve at faster rates due to a higher interface plasticity. Here, we ask whether the higher rate of adaptive mutations at the surface of proteins could have arisen through inter-molecular interactions at the protein network level. We addressed this question by estimating the rate of adaptive mutations in genes with different degrees of PPI. This was only possible in *D. melanogaster* since there was limited data available for *A. thaliana*. We report a negative correlation between the number of PPI and ω, ω_{na} and ω_a, respectively, with only significant values observed for ω (Table 1, supplementary figure S20 and file S2 in Appendix I). These findings suggest that a higher degree of protein connectivity leads to lower rates of protein sequence evolution, but prevent us to assess with confidence whether this effect is due to a stronger purifying selection and/or a slower rate of adaptive substitutions. A potential limitation of this analysis is the low number of genes with PPI information available and the noise associated with the BioGRID annotations. As a physical interaction does not necessarily imply a functional link, we might lack statistical power to detect any putative effect of PPI on ω_a (Chatr-Aryamontri et al. 2017).

In support to our third hypothesis, several studies have described the role of the immune and defence responses in molecular evolution across taxa (Sackton et al. 2007; Obbard et al. 2009; Enard et al. 2016; Mauch-Mani et al. 2017). These studies suggest that pathogens could be key drivers of protein adaptation, by acting as a powerful selective pressure through the coevolutionary arms race between hosts and parasites. This could be driving the higher rate of adaptive mutations in protein biosynthesis enzymes (Figure 4), which are the ones typically hijacked by pathogens during host infection (Dangl and Jones 2001; Enard et al. 2016). Moreover, one of the fastest evolving protein class is the ubiquitin system (Figure 4), which is known to be involved in the defence mechanism, both by the host, through processes like the activation of innate immune responses and degradation signalling of pathogenic proteins; and by the pathogen, which inhibits and/or uses this system in order to modulate host responses (Loureiro and Ploegh 2006; Collins and Brown 2010; Dielen et al. 2010; Trujillo and Shirasu 2010; Hiroshi et al. 2014). Membrane trafficking proteins are also well-known for being involved in the immune response mechanisms, a functional class that also presents high values of ω_a, and "DNA replication" together with "mRNA biogenesis" and "transcription machinery" are typical signatures of viruses' activities (Figure 4). Likewise, in *A. thaliana*, cytochrome P450 proteins present a high rate of adaptive mutations (Figure 4), which have been reported to play a crucial role in the defence response in plants (Schuler and Werck-Reichhart 2003). Besides, the reduced selective pressure on nuclear and secreted proteins (supplementary figure S6 in Appendix I) may be also a consequence of their role in disease and pathogen immunity (*i.e.*, Motion et al. 2015; Mosmann et al. 2016), as observed in yeast (Julenius and Pedersen 2006), insects (Sackton et al. 2007; Obbard et al. 2009) and primates (Nielsen et al. 2005).

Our findings therefore support the hypothesis that coevolutionary arms race of the host-pathogen interactions, in particular intra-cellular pathogens such as viruses, are a major driver of adaptation in proteins. While we do not rule out that protein-protein interactions and the acquisition of new biochemical functions could also have an impact, more and better annotation data is required to further evaluate their role. In conclusion, our study reveals that, in addition to genome architecture, protein structure has a substantial impact on adaptive evolution consistent between *D. melanogaster* and *A. thaliana*, unravelling the potential generality of such effect. Our study further emphasizes that the rate of adaptation not only varies substantially between genes, but also at the intra-genic scale, and we posit that accounting for a fine-scale, intra-molecular evolution is necessary to fully understand the patterns of molecular adaptation at the species level.

2.4 Materials and Methods

Population Genomic Data and Data filtering

The *D. melanogaster* data set included alignments of 114 genomes for one chromosome arm of the two large autosomes (2L, 2R, 3L and 3R) and one sex chromosome (X) pooled from 22 sub-Saharan populations with negligible amount of population structure ($F_{ST} = 0.05$; DPGP2, Pool et al. 2012). Release 5 of the Berkeley Drosophila Genome Project (BDGP5, http://www.fruitfly.org/sequence/release5genomic.shtml, last updated June 2018) was used as the reference genome. Estimations of divergence were performed with *D. simulans*, for which genome alignments with the reference genome were available (http://www.johnpool.net/genomes.html). For *A. thaliana*, analyses were carried out with 110 genomes for the 5 chromosomes of the Spanish population from the 1001 Genomes Project (Weigel and Mott 2009), using the release 10 from The Arabidopsis Information Resource (TAIR10, ftp://ftp.ensemblgenomes.org/pub/plants/release-40/fasta/arabidopsis_thaliana/dna/) as reference genome. Divergence estimates were made with *A. lyrata* as an outgroup species, for which a pairwise alignment with the reference genome was available (ftp://ftp.ensemblgenomes.org/pub/plants/release-38/maf). Data processing was conducted with the help of GNU parallel (Tange 2011).

Estimation of the population genetic parameters and model selection

Coding DNA sequences (CDS) were extracted from the alignments with MafFilter (Dutheil et al. 2014) according to the General Feature Format (GFF) file of the reference genome of both species. First, a cleaning and filtering process was performed to keep only non-overlapping genes with the longest transcript, in cases of multiple transcripts per gene. At this stage, 12,801 and 27,072 genes, for *D. melanogaster* and *A. thaliana* respectively, were kept for further analysis. CDS sequences were then concatenated in order to obtain the full coding region per gene. For the analysis with *A. thaliana*, the alignment of *A. lyrata* with the reference sequence was re-aligned with each gene alignment of the ingroup using MAFFT v7.38 (Katoh and Standley 2013) with the options *add* and *keeplength* so that no gaps were included in the ingroup. CDS alignments with premature stop codons were excluded and alignment positions lacking a corresponding sequence in the outgroup were discarded. Final datasets included 10,318 genes for *D. melanogaster/D. simulans* and 18,669 genes for *A. thaliana/A. lyrata*. These datasets were then used to infer both the synonymous and non-synonymous unfolded and folded site frequency spectra (SFS), and synonymous and non-synonymous divergence based on the rate of synonymous and non-synonymous substitutions. Sites for which the outgroup allele was missing were considered as missing data. All calculations were performed using the BppPopStats program from the Bio++ Program Suite (Guéguen et al. 2013). The Grapes program was then used to compute a genome-wide estimate of the rate of non-adaptive (ω_{na}) and adaptive non-synonymous substitutions (ω_a) (Galtier 2016). This method assumes that all sites were sampled in the same number of chromosomes and since some sites were not successfully sampled in all individuals, the original dataset was reduced to 110 and 105 individuals for *D. melanogaster* and *A. thaliana* respectively, by randomly down-sampling polymorphic alleles at each site. The following models were fitted and compared using Akaike's information criterion: Neutral, Gamma, Gamma-Exponential, Displaced Gamma, Scaled Beta and Bessel K. A model selection procedure was conducted on the two datasets using the complete set of genes for comparison (see Table S1 in supplementary file S1, Appendix I). Following analyses consist in fitting the selected model on several subsets of the data according to the variables analysed, comprising sets of genes (see Tables S2 and S3 in supplementary file S1 for detailed information on the genes used for each variable as well as the population genetic parameters estimated per gene for *A. thaliana* and *D. melanogaster* respectively, Appendix I) and amino-acid residues (see Tables S4 and S5 in supplementary file S1 for detailed information on the amino-acid residues used for each category as well as the population genetic parameters estimated per site for *A. thaliana* and *D. melanogaster* respectively, Appendix I). We next described the different variables analysed.

Categorization of gene and genome architecture

Recombination rates were obtained with the R package "MareyMap" (Rezvoy et al. 2007), by using the cubic splines interpolation method. Hereafter we computed the mean recombination rate in cM/Mb units for each gene. Discretization of the observed distribution of recombination rate was performed in 50 and 30 categories with around 350 and 280 genes each for *A. thaliana* and *D. melanogaster* respectively. Intronic information was obtained using the GenomeTools from a GFF with exon annotation and the option *addintrons* (Gremme et al. 2013). Genes were discretized into 13 and 10 categories according to their intron content for *A. thaliana* and *D. melanogaster* respectively.

Categorization of protein structure

Genes were discretized according to the total size of the coding region, for which 30 and 50 categories with around 620 and 210 genes each were made for *A. thaliana* and *D. melanogaster* respectively.

In order to obtain structural information for each protein sequence, blastp (Schaffer 2001) was first used to assign each protein sequence to a PDB structure, and respective chain, by using the "pdbaa" library and an E-value threshold of 1e-10. When multiple matches occurred, for instance in cases of multimeric proteins, the match with the lowest E-value was kept. This resulted in 5,008 genes for which a PDB structure was available, making a total of 3,834 PDB structures for *D. melanogaster* and 9,121 genes with a total of 3,832 PDB structures for *A. thaliana*. The corresponding PDB structures were then downloaded and further processed to only keep the corresponding chain per polymer. PDB manipulation and analysis were carried on using the R package "bio3d" (Grant et al. 2006). Values for secondary structure (SS) and solvent accessibility (SA) per residue were obtained using the "dssp" program with default options, and were successfully retrieved for 3,613 PDB files corresponding to 4,944 genes for *D. melanogaster* and 3,806 PDB files for a total of 9,106 genes for *A. thaliana*. Subsequently, to map SS and SA values to each residue of the protein sequence a pairwise alignment between each protein and the respective PDB sequence was performed with MAFFT, allowing gaps in both sequences in order to increase the block size of sites aligned. The final data set comprised a total of 1,397,885 and 1,395,666 sites with SS and SA information, respectively, out of 4,821,113 total codon sites obtained with BppPopStats for the complete set of genes of

D. melanogaster; and 2,585,468 and 2,585,467 sites mapped with SS and SA information, respectively, out of 7,479,808 codon sites of *A. thaliana*. We computed the relative solvent accessibility (RSA) by dividing SA by the amino-acid's solvent accessible area (Tien et al. 2013).

Categorization of secondary structure was performed by comparing 460,702, 975,934 and 523,880 amino-acid residues in β-sheets, α-helices and loops respectively in *A. thaliana*, and 258,898, 516,356 and 282,588 sites in β-sheets, α-helices and loops respectively in *D. melanogaster*. RSA values were analysed with 28 categories with around 85,000 sites each, with the exception of the totally buried residues (RSA = 0) category containing 299,684 sites in *A. thaliana*; and 19 categories with approximately 69,000 residues each, except for 151,417 completely buried residues in *D. melanogaster*. For the analysis of correlation between variables two categories of RSA were considered, comparing buried (RSA < 0.05) and exposed (RSA >= 0.05) residues, following Miller et al (Miller et al. 1987).

Estimates of intrinsic protein disorder were acquired via the software DisEMBL (Linding et al. 2003), wherein intrinsic disorder was estimated per site and classified according to the degree of "hot loops", meaning loops with a high degree of mobility. This analysis was successfully achieved for a total of 7,479,807 out of 7,479,808 sites for *A. thaliana* and 3,952,602 out of 4,821,113 sites for *D. melanogaster*. Amino-acid residues were divided into 30 categories with an average of 249,000 and 131,000 sites in *A. thaliana* and *D. melanogaster* respectively. For the proportion of disordered regions per protein, we considered a residue "disordered" if it was in the top 25% of the measured probabilities of disorder across the proteomes of each species. Analyses were performed with 30 categories with around 620 and 420 genes for *A. thaliana* and *D. melanogaster* respectively.

Identification of proteins binding to a molecular chaperone

Prediction of the molecular chaperone *DnaK* binding sites in the protein sequence was estimated with the LIMBO software using the default option *Best overall prediction*. This setting implies 99% specificity and 77.2% sensitivity (Van Durme et al. 2009). Genes were categorized according to this prediction setting, which suggests that every peptide scoring above 11.08 is a predicted *DnaK* binder. Genes scoring below that value were not consider as possible binders.

Categorization of gene expression

Mean gene expression data was obtained from the database Expression Atlas (http://www.ebi.ac.uk/gxa; Petryszak et al. 2016), wherein one baseline experiment was used for each species (*D. melanogaster*, E-MTAB-4723; *A. thaliana*, E-GEOD-38612). In addition, for *D. melanogaster*, we obtained the breadth of expression data over the embryo anatomy from the BDGP database (Tomancak et al. 2007) and the data was processed and analysed as in Salvador-Martínez et al. (2018). Mean gene expression levels were obtained by averaging across samples and tissues for each gene, ending up with 40 and 15 categories with around 450 and 430 genes each for *A. thaliana* and *D. melanogaster* respectively. For the analysis on the breadth of expression, expression patterns in *A. thaliana* were analysed in four different tissues: roots, flowers, leaves and siliques; and for *D. melanogaster* we used the anatomical structures of the embryo development, analysing 18 structures (see Tomancak et al. 2007 and Salvador-Martínez et al. 2018). Analyses were carried with four and six categories in *A. thaliana* and *D. melanogaster* respectively, according to the number of tissues/organs a gene is expressed (see Tables S2 and S3 in supplementary file S1 for detailed information, Appendix I).

Protein cellular localization and protein functional class

Cellular localization of each protein sequence was predicted with the software ProtComp (from Softberry, http://www.softberry.com/) with the default options and genes were classified into the following cellular categories: cytoplasmic, endomembrane system, mitochondrial, nuclear, peroxisome, plasma membrane and secreted proteins. The category peroxisome was excluded from further analysis due to the small number of annotated genes (114 and 250 genes in *D. melanogaster* and *A. thaliana* respectively; detailed information in Tables S2 and S3 in supplementary file S1, Appendix I). Protein functional classes were obtained with the Bioconductor package for R "KEGGREST", using the KEGG BRITE database (Kanehisa et al. 2002). Analysis were carried out with 2,950 and 3,780 genes for *D. melanogaster* and *A. thaliana* respectively, discretized into the highest levels of each of the three top categories of protein classification: metabolism, genetic information processing and signalling and cellular processes (see Tables S2 and S3 in supplementary file S1, Appendix I).

Enzymatic active sites and protein-protein interactions

In order to check whether a residue was present in an active site, we used the ScanProsite software (De Castro et al. 2006). Datasets included 1,061,876 and 1,870,166 active sites for *D. melanogaster* and *A. thaliana* respectively. All sites that were not predicted by the program were considered as non-active (see Tables S4 and S5 in supplementary file S1, Appendix I). Data on the degree of protein-protein interactions was obtained with the BioGRID database (Chatr-Aryamontri et al. 2017). This was only possible for *D. melanogaster* since the data available for *A. thaliana* was very limited (only 878 annotated genes mapping to our dataset). Analyses were carried out with 5,628 genes divided into 19 categories, with 1,114 genes in the first category, and the others ranging from 700 to 130 according to the respective number of interactions (see Tables S2 and S3 in supplementary file S1, Appendix I).

Estimation of the adaptive and non-adaptive rate of non-synonymous substitutions

For all gene and amino-acid sets, 100 bootstrap replicates were generated by randomly sampling genes or sites in each category. The Grapes program was then run on each category and replicate with the Gamma-Exponential distribution of fitness effects (Galtier 2016). The first step included the removal of replicates for which the distribution of fitness effects parameters was not successfully fitted. For this purpose, we discarded 1% in the maximum and minimum values for the mean and shape parameters of the DFE (see supplementary files for detailed R scripts in Appendix I). Results for ω, ω_{na} and ω_a were plotted using the R package "ggplot2" (Wickham 2017) by taking the mean value and the 95% confidence interval of the 100 bootstrap replicates computed for each category (both for main and supplementary figures, for continuous and discrete variables, see supplementary files Appendix I).

Statistical analyses

Significance for all continuous variables, including protein length, number of introns, gene expression, intrinsic residue disorder, proportion of disordered regions, recombination rate, number of protein-protein interactions and RSA, was assessed through Kendall's correlation tests. Kendall's correlation test is non-parametric and does not make any assumption on the distribution of the input data. Furthermore, it can be applied to ordinal data, making it appropriate to analyse discretized continuous variables. To do so, the mean value of the 100 bootstrap replicates was taken for each category (see detailed script as well as all statistical results in supplementary file S2 in Appendix I). Significance values for discrete variables, comprising binding affinity to *DnaK*, protein location, protein functional class and secondary structure motif, were achieved by estimating the differences between each pair of the categories analysed, by randomly subtracting each bootstrap replicate. Following steps included counting the number of times the differences between categories were below and above 0, which by taking the minimum of those values gives us a statistic that we call k. The two-tailed p-value was then estimated by applying the following equation: $p = (2\mathrm{k} + 1)/(\mathrm{N} + 1)$, where N in the number of bootstrap replicates used. For variables comparing more than two categories we corrected the p-value for multiple testing using the FDR method (Benjamini and Hochberg 1995) as implemented in R (R Core Team 2015) (see detailed script and all statistical results in supplementary files S3 and S4 in Appendix I). Analyses on the correlations between variables are described in supplementary Files S5 and S6. The analysis of covariance (ANCOVA) was performed by applying a linear model to the values of ω_{na} and ω_a with the interaction between RSA and protein disorder following a control for the normality, homoscedasticity and independence of the corresponding error (supplementary file S5 in Appendix I).

CHAPTER III

How Do Rates of Adaptation Vary Across Time?

3.1 Abstract

Understanding the dynamics of species adaptation to their environments has long been a central focus in the study of molecular evolution. Early adaptive theories proposed that populations evolve by "walking" in an adaptive landscape. This "adaptive walk" is characterized by a pattern of diminishing returns, where populations further away from their fitness optimum take larger steps than the ones closer to their optimal conditions. This pattern seems to reflect the faster evolution of young genes: as these genes are theoretically further away from their fitness optimum, they need to take larger steps to reach their full potential. Testing the impact of gene age on molecular evolution, however, constitutes an arduous task. Young genes are small, have a higher degree of intrinsic disorder, are expressed at lower levels, and are involved in species-specific adaptations. These factors could, therefore, be mystifying the high rates of evolution of young genes. By controlling for multiple confounding factors, we provide the first attempt to test the effect of gene age on the molecular rate of adaptation both in plants and in animals. To estimate the rate of adaptive substitutions, we fitted models of the distribution of fitness effects both at the protein and amino-acid residue levels. Our findings suggest that the evolutionary origin of a gene acts as a primary determinant of the molecular adaptive rate at the gene level, thus supporting a model of adaptation in young genes in an "adaptive-walk" manner.

3.2 Introduction

How does adaptive evolution proceed in space and in time? This question has long intrigued evolutionary biologists as adaptive mutations are often too rare to study. At the phenotype level, Fisher (1930) proposed that adaptation relies on mutations with small effect sizes. He presented the geometric model of adaptation where phenotypic evolution occurs in a continuous and gradual scale towards some optimum fitness (Fisher 1930a). At the molecular level, Wright (1931, 1932) was the first to introduce the idea that populations evolve in the space of all possible gene combinations to acquire higher fitness. He characterized this model of evolution as a walk in an adaptive landscape. He proposed the shifting balance theory of adaptation, where drift moves the population away from its local peak, and natural selection directs the population to higher fitness, the so-called "global optimum" in a fitness landscape. With the rise of molecular genetics, John

Maynard Smith (Smith 1962, 1970a) extended this idea to a sequence-based model of adaptation. He introduced the concept of an "adaptive walk," where a protein "walks" in the space of all possible amino-acid sequences towards the ones with increasingly higher fitness values. Gillespie (1983, 1984, 1991) further developed Wright's model of adaptation and presented the "move rule" in an adaptive landscape. He suggested that adaptation proceeds in large steps, where mutations with higher effects on fitness are more likely to reach fixation. The ”adaptive walk” was later fully developed by Allen Orr (1998, 1999). Orr extended Fisher's geometric model of adaptation and demonstrated that, apart from small effect mutations, adaptation also relies on mutations of large fitness effects. He, therefore, characterized the adaptive walk with a pattern of diminishing returns. Under this model, a sequence that is further away from its local optimum will tend to accumulate large-effect mutations at the beginning of the "walk." Small-effect mutations will then only be fixed when the sequence is approaching its high fitness. Experimental studies tracing the evolution of microbial populations (e.g., Lenski et al. 1991; Cooper and Lenski 2000; Gerrish 2001; Imhof and Schlötterer 2001; Rozen et al. 2002) and fungi (Schoustra et al. 2009) provided evidence for this view of adaptation as a walk with diminishing returns. These studies, however, can only assess patterns of adaptation at relatively short time scales. The challenge lies in studying adaptation across time: how does the distribution of beneficial mutations vary along the phylogeny of the species?

One way to look at molecular evolution in time is to study genes with different evolutionary origins. Different genes within a genome not only differ in function, expression, or length but also age (e.g., Lynch 2002; Daubin and Ochman 2004; Tautz and Domazet-Lošo 2011; Neme and Tautz 2013). One can estimate the age of a gene by using sequence similarity searches (BLAST; Altschul et al. 1998) across the phylogeny of the species. A gene is considered "old" if a homolog is identified in several taxa over a deep evolutionary scale, or "young" or lineage-specific if the recognized homologs are only present in closely-related species. This approach is known as phylostratigraphy (Domazet-Lošo et al. 2007).

Multiple studies suggested that young or lineage-specific protein-coding genes evolve faster than old ones (Thornton and Long 2002; Domazet-Loso and Tautz 2003; Krylov et al. 2003; Daubin and Ochman 2004; Albà and Castresana 2005; Wang et al. 2005; Cai et al. 2006; Wolf et al. 2009; Cai and Petrov 2010; Zhang et al. 2010; Vishnoi et al. 2010; Tautz and Domazet-Lošo 2011; Cui et al. 2015). In humans, Albà and Castresana (2005) showed a negative correlation between d_N/d_S and gene age, where young genes present higher d_N/d_S. Cai and Petrov (2010) confirmed these findings also in chimpanzees. They further suggested that the faster evolution in young primate genes may be due to the lack of selective constraint posed by purifying selection and provided evidence that positive selection might be also at play. Similar patterns were observed in fungi (Cai et al. 2006), *Drosophila* (Domazet-Loso and Tautz 2003; Zhang et al. 2010; Domazet-Lošo et al. 2017), bacteria (Daubin and Ochman 2004), viruses (García-Vallvé et al. 2005), plants (Arendsee et al. 2014; Cui et al. 2015), and protozoan parasites (Kuo and Kissinger 2008).

Despite the observed consistency across taxa, the drivers of such an effect remain unclear. Besides, young and old genes differ in their structural properties, expression level, and protein function. Young genes tend to be smaller (Cai and Petrov 2010; Vishnoi et al. 2010; Neme and Tautz 2013), have a higher level of intrinsic disorder (Wilson et al. 2017), and are expressed at lower levels (Wolf et al. 2009; Cai and Petrov 2010; Vishnoi et al. 2010; Tautz and Domazet-Lošo 2011). Moreover, young genes tend to encode proteins involved in the development of species-specific characteristics (e.g., Hughes 1994; Lynch 2002; Zhang et al. 2002), as well as in the immune and stress responses (e.g., Hughes 1994; Lynch 2002; Zhang et al. 2002). As the macromolecular structure (Afanasyeva et al. 2018; Moutinho et al. 2019b), gene expression levels (e.g., Rocha and Danchin 2004; Subramanian and Kumar 2004; Moutinho et al. 2019b), and protein function (e.g., Stukenbrock et al. 2011; Enard et al. 2016; Moutinho et al. 2019b) are known determinants of the rate of protein adaptation; they could be mystifying the effect of gene age. Several studies reported the substantial impact of gene expression on the adaptive rate of proteins, where highly expressed proteins are significantly more constrained and have lower rates of adaptation (Pal et al. 2001; Rocha and Danchin 2004; Subramanian and Kumar 2004; Moutinho et al. 2019b). Moreover, studies have shown that the structure of a protein significantly impacts the molecular adaptive rate, where highly disordered (Afanasyeva et al. 2018; Moutinho et al. 2019b) and exposed residues (Moutinho et al. 2019b) present higher rates of adaptive evolution. Proteins involved in the immune and stress response were also reported with higher rates of molecular adaptation (e.g., Sackton et al. 2007; Obbard et al. 2009; Stukenbrock et al. 2011; Enard et al. 2016; Moutinho et al. 2019b). It is thus crucial to account for these confounding factors to better assess the impact of gene age on the molecular adaptive rate.

Here, we further investigate the impact of gene age on protein adaptive evolution to test whether adaptation along the phylogeny of a species follows an "adaptive walk" model. To assess the consistency of the inferred effects, we used two species with different life-history traits: the dipteran *Drosophila melanogaster* and the Brassicaceae *Arabidopsis thaliana*. To estimate the molecular rate of adaptation, we fitted models of the distribution of fitness effects (DFE) both at the protein and amino-acid residue levels across different age classes. Moreover, we assessed whether protein length, gene expression, relative solvent accessibility (RSA), intrinsic protein disorder, protein divergence, and protein function act as confounding factors of the effect of gene age. Our study aims to achieve a more comprehensive understanding of how the age of a gene impacts the rate of protein adaptation and gives light to the potential determinants of the higher evolutionary rate in young genes.

3.3 Results

We assessed the role of gene age on adaptive evolution using the divergence and polymorphism data published in Moutinho et al. (2019b). The data included 10,318 protein-coding genes in 114 *Drosophila melanogaster* individuals from an admixed sub-Saharan population from Phase 2 of the *Drosophila* Genomics Project (DPGP2, Pool et al. 2012) and divergence estimates from *D. simulans*; and 18,669 protein-coding genes in 110 *Arabidopsis thaliana* genomes comprising polymorphism data from a Spanish population (1001 Genomes Project, Weigel and Mott 2009) and divergence out to *A. lyrata*. The rate of adaptive evolution was estimated with the Grapes program (Galtier 2016). Grapes disentangles the effects of negative and positive selection on the d_N/d_S ratio (ω) by inferring the rate of non-adaptive (ω_{na}) and adaptive (ω_a)

non-synonymous substitutions, as well as the proportion of adaptive amino-acid substitutions (α). In our study, we focused on analysing the impact of the age of a protein on ω_{na} and ω_a, as well as the total ω.

The individual effect of gene age on adaptive evolution

The age of protein-coding genes was obtained from published data in *Drosophila melanogaster* (Domazet-Lošo et al. 2017) and *Arabidopsis thaliana* (Arendsee et al. 2014). The analyses of *D. melanogaster* were carried with 12 age categories corresponding to the phylogenetic branches defined in the work of Domazet-Lošo et al. (2017) (Figure 1a). We report a significant positive correlation between ω, ω_{na}, and ω_a with increasing phylostrata level for all chromosomes considered together (Table 1 and Figure 1b). As X-linked genes are known to evolve faster (Vicoso and Charlesworth 2006, 2009), we performed separate analyses for the X and the autosomal chromosomes to evaluate whether there were significant differences between them. While we observed a positive correlation between ω, ω_{na}, and ω_a with the phylostrata level, the correlations were weaker, as only marginally significant estimates were reported for ω_a in the X and ω_{na} for autosomal genes (Table 1 and Figure 1b). We performed an analysis of covariance (ANCOVA) to assess whether the chromosome had an impact on the effect of gene age, by comparing a model M1 that included the effects of chromosome, age, and their interaction, with a model M0 that included age only. We found low support for the effect of the chromosome ($p = 0.041$ for ω_{na} and $p = 0.094$ for ω_a) and, therefore, combined all chromosomes for subsequent analyses.

The analyses of *A. thaliana* were performed with 15 categories according to the clades defined in Arendsee et al. (2014) (Figure 1a). We reported a consistent pattern with that observed in *D. melanogaster*, where a significant positive correlation can be observed for estimates of ω, ω_{na}, and ω_a (Table 1 and Figure 1b).

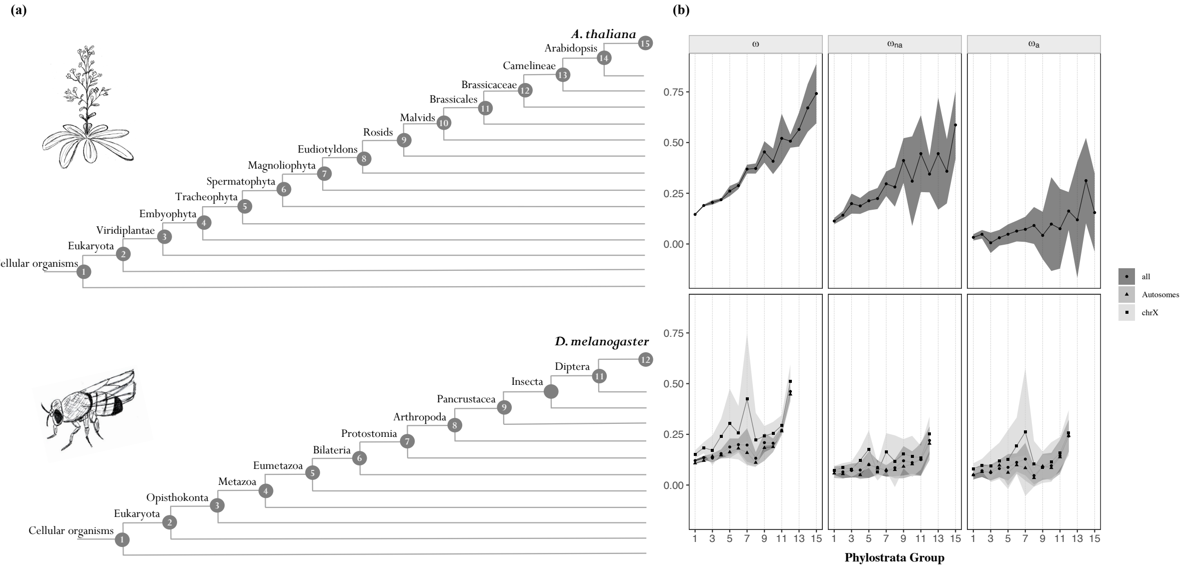

Figure 1. (a) Phylogenetic relationship between the clades analysed for *A. thaliana* (top) and *D. melanogaster* (bottom). **(b)** Relationship between the rate of protein evolution (ω), non-adaptive non-synonymous substitutions (ω_{na}) and adaptive non-synonymous substitutions (ω_a) with gene age in *A. thaliana* (top) and in *D. melanogaster* (bottom). Clades are ordered according to (a). In *D. melanogaster*, the results for X-linked, autosomal, and total genes are showed. Mean values of ω, ω_{na} and ω_a for each category are represented with the black points. Error bars denote for the 95% confidence interval for each category, computed over 100 bootstrap replicates.

Table 1. Statistical results for the analysis of the individual effect of gene age on ω, ω_{na}, and ω_a.

	A. thaliana			***D. melanogaster***		
	ω	$\boldsymbol{\omega_{na}}$	$\boldsymbol{\omega_a}$	**ω**	$\boldsymbol{\omega_{na}}$	$\boldsymbol{\omega_a}$
All chromosomes	0.962 (***)	0.848 (***)	0.733 (***)	0.727 (***)	0.697 (**)	0.636 (**)
X chromosome	-	-	-	0.576 (***)	0.636 (**)	0.485 (.)
Autosomes	-	-	-	0.756 (**)	0.424 (.)	0.424 (*)

Note. For each variable, the Kendall's τ of gene age is shown with the respective significance (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; "." $0.05 \leq P < 0.10$) for ω, ω_{na} and ω_a in *A. thaliana* and *D. melanogaster*.

Table 2. ANCOVA estimates for the contribution of gene age, protein length, gene expression, residue intrinsic disorder, RSA, and sequence similarity in ω_{na}, ω_a, and the interaction between the variables analysed.

	A. thaliana		***D. melanogaster***	
	$\boldsymbol{\omega_{na}}$	$\boldsymbol{\omega_a}$	$\boldsymbol{\omega_{na}}$	$\boldsymbol{\omega_a}$
Gene Age	96.76 (***)	25.90 (.)	77.21 (***)	88.52 (**)
Protein Length	0.70	5.19	12.83 (.)	6.41
Interaction	2.30 (*)	63.00 (*)	7.34	0.25
Gene Age	76.90 (***)	89.79 (**)	83.81 (***)	65.19 (**)
Gene Expression	21.78 (***)	3.01	0.75	10.83
Interaction	0.74	2.92	12.79 (.)	20.43 (.)
Gene Age	41.77 (***)	70.02 (***)	29.11 (**)	47.89 (***)
Relative Solvent Accessibility	46.88 (***)	27.78 (***)	62.48 (***)	49.87 (***)
Interaction	9.15 (.)	0.98	6.70 (.)	0.633
Gene Age	97.09 (***)	73.93 (**)	84.61 (***)	84.06 (**)
Exposed Residues/Gene	0.11	0.01	13.13 (*)	0.84
Interaction	0.31	18.92	0.07	7.87
Gene Age	87.55 (**)	93.02 (***)	67.80 (***)	83.81 (***)
Residue Intrinsic Disorder	11.94 (**)	1.22	25.75 (*)	10.73
Interaction	0.03	4.27	2.24	0.30

Note. For each variable, the proportion of explained variance is shown with the respective significance (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; "." $0.05 \leq P < 0.10$) for ω_{na} and ω_a in *A. thaliana* and *D. melanogaster*.

Is gene age the main determinant of the molecular adaptive rate?

We assessed whether gene age constitutes the main determinant of the rate of molecular adaptation by controlling for multiple confounding factors. As estimates of the rate of adaptive substitutions for single genes generate large sampling variances (Smith and Eyre-Walker 2002; Stoletzki and Eyre-Walker 2011), analyses were performed by pooling genes according to the categories analysed. Hence, each confounding factor was analysed individually and their magnitude effects were compared.

Previous studies reported that younger genes encode shorter proteins (Vishnoi et al. 2010; Ding et al. 2012; Neme and Tautz 2013) and are expressed at lower levels (Wolf et al. 2009; Cai and Petrov 2010; Vishnoi et al. 2010; Tautz and Domazet-Lošo 2011), a pattern that we also observe in our data set (gene age vs. protein length: τ = -0.485, p = 2.81e-02; τ = 0.752, p = 9.249e-05, Figure S1a; gene age vs. gene expression: τ = 0.636, p = 3.976e-03; τ = -0.880, p = 5.154e-06, Figure S1b in Appendix II; for *D. melanogaster* and *A. thaliana,* respectively). As protein length and gene expression are known to have an impact on the rate of protein evolution (Rocha and Danchin 2004; Liao et al. 2006; Moutinho et al. 2019b), we performed the analysis on gene age controlling for these two factors to assess whether the effect of gene age persisted. When looking at short and long genes separately (see Material and Methods), we observed that gene age is positively correlated with ω, ω_{na}, and ω_a in *D. melanogaster* (Figure 2a). In *A. thaliana*, we reported the same pattern, although with a comparatively weaker correlation observed for ω_a (Figure 2a). To further assess the relative contribution of protein length and gene age on ω_{na} and ω_a, we performed analyses of covariance (ANCOVA), using both factors and their interaction as explanatory variables. Our analyses showed that gene age is the largest contributor for the observed correlation with estimates of ω_{na} and ω_a in both species, although with only marginally significant estimates for ω_a in *A. thaliana*. Moreover, in *A. thaliana*, the interaction between protein length and gene age was also significant, suggesting that the two factors may be acting together (Table 2).

The analysis considering low and highly expressed genes individually reported a positive correlation for estimates of ω, ω_{na}, and ω_a in both species (Figure 2b). By examining the relative contribution of each of the variables, we showed that gene age is the main determinant of both ω_a and ω_{na} in both species, with gene expression only significantly contributing to ω_{na} in *A. thaliana*. Moreover, the interaction between the two variables also appears to slightly affect the rate of molecular adaptation in *D. melanogaster*, with marginally significant results observed for ω_a and ω_{na} (Table 2).

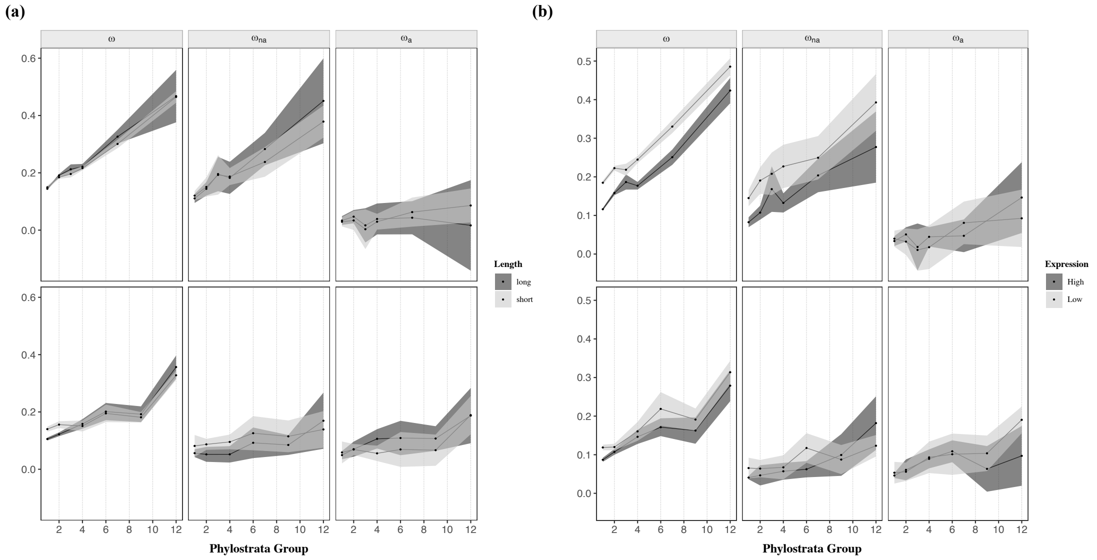

Figure 2. Estimates of ω, ω_{na} and ω_a plotted as a function of (a) protein length and (b) mean expression levels and gene age in *A. thaliana* (top) and *D. melanogaster* (bottom). Analyses were performed by comparing short and long (a), low and highly expressed (b) genes (see Methods) across 6 categories of gene age for both species. Legend as in Figure 1.

As proteins encoded by young genes are short, we expect them to have more exposed residues (Moutinho et al. 2019b), a pattern that we observed in our dataset (τ = 0.697, p = 0.0016; τ = 0.676, p = 0.0004, for *D. melanogaster* and *A. thaliana* respectively; Figure S2a in Appendix II). Moreover, as exposed residues are more flexible (Moutinho et al. 2019b), young genes tend to encode for proteins with a higher degree of intrinsic disorder, a pattern previously reported in mice (Wilson et al. 2017). We confirm this pattern in *D. melanogaster* (τ = 0.697, p = 0.002) and *A. thaliana* (τ = 0.505, p = 0.009; Figure S2b in Appendix II). The analysis of gene age on exposed and buried residues shows a positive correlation for estimates of ω, ω_{na}, and ω_a in both species (Figure 3a). By looking at the relative contribution of each of the variables, we observed that both RSA and gene age act as determinants of ω_{na} and ω_a in both species, with gene age contributing relatively more to ω_a in *A. thaliana*, and RSA to ω_{na} in *D. melanogaster* (Table 2). As RSA constitutes a main determinant of the rate of adaptive substitutions in these species (Moutinho et al. 2019b), we further assessed if the observed effect of gene age was driven by the variation of ω_a and ω_{na} within each category of RSA. We did so by reducing the dataset into two groups of sites with similar RSA levels (see Material and Methods) and re-analysed the effect of gene age on both. Our analyses showed that the effect of gene age persisted in estimates of ω, ω_{na}, and ω_a in *A. thaliana*. In *D. melanogaster*, however, only marginally significant results were observed for estimates of ω_a in all residues and for ω_{na} in exposed residues (Figure S3 and Table S1 in Appendix II).

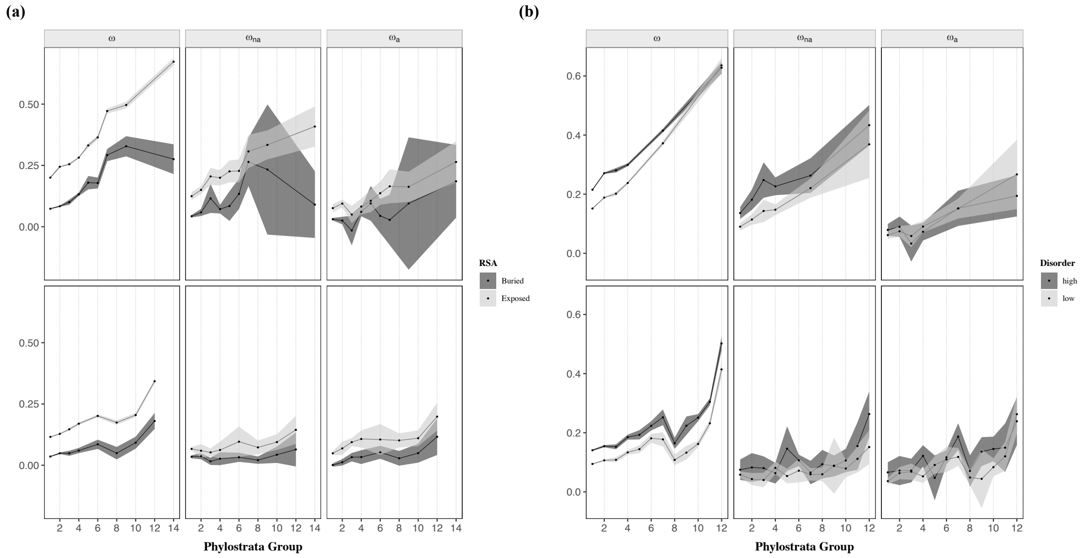

Figure 3. Estimates of ω, ω_{na} and ω_a plotted as a function of (a) relative solvent accessibility and (b) residue intrinsic disorder and gene age in *A. thaliana* (top) and *D. melanogaster* (bottom). Analyses were performed by comparing buried and exposed (a), low and highly disordered (b) residues (see Methods) across 12 and 6 categories of gene age in (a) for *D. melanogaster* and *A. thaliana* respectively, and with 6 and 4 categories in (b) for *D. melanogaster* and *A. thaliana* respectively. Legend as in Figure 1.

To further disentangle the effect of these two variables, we analysed the correlation between RSA and gene age at the gene level. We stratified the dataset into two groups of genes according to their proportion of exposed residues (see Material and Methods) and assessed the effect of gene age on both. We reported a positive correlation between gene age and estimates of ω, ω_{na}, and ω_a in both species (Figure S4 in Appendix II). ANCOVA analyses showed that, at the gene level, gene age constitutes the main determinant of ω_a and ω_{na} in both species, with the proportion of exposed residues only having a significant impact on ω_{na} in *D. melanogaster* (Table 2).

When performing the analysis of gene age on residues with high and low intrinsic disorder, we observed a positive correlation for estimates of ω, ω_{na}, and ω_a in both species (Figure 3b). ANCOVA analyses showed that gene age constitutes the main determinant of both ω_{na} and ω_a in both species, with residue intrinsic disorder only contributing for estimates of ω_{na} (Table 2).

In summary, the correlations performed with protein length, gene expression, RSA, and residue intrinsic disorder, show that gene age is the major factor determining the molecular adaptive rate at the gene level. When looking at the site the site level, however, our findings suggest that both RSA and gene age substantially impact the rate of adaptive evolution.

The effect of protein divergence on the relation between gene age and the molecular adaptive rate

While physlostratigraphy is the most-widely used approach to identify the emergence of new genes, some studies have pointed out its potential limitations (Elhaik et al. 2006; Albà and Castresana 2007; Moyers and Zhang 2015, 2016; Domazet-Lošo et al. 2017). The problem lies on the fast-evolving and short genes: as BLAST homology searches might fail to identify homologs in these genes, they could be mistakenly classified as young. To assess whether the correlation of gene age with the rate of adaptive evolution could be explained by BLAST's false negative rate, we analysed the effect of gene age by correcting for protein divergence. As expected, we observed that younger phylostrata groups present higher rates of protein divergence (τ = 0.757, p = 0.0006; τ = 0.886, p = 4.178e-06, for *D. melanogaster* and *A. thaliana* respectively; Figure S5a in Appendix II). To remove this effect, we randomly sampled a subset of genes (see Material and Methods) for which the positive correlation between gene age and protein divergence was no longer significant (τ = 0.156, p = 0.531; τ = 0.182, p = 0.411, for *D. melanogaster* and *A. thaliana* respectively; Figure S5b in Appendix II), and analysed the effect of gene age on the selected genes. We observed that the effect of gene age prevailed for estimates of ω and ω_a in *A. thaliana* (ω: τ = 0.697, p = 0.002; ω_{na}: τ = -0.424, p = 0.055; ω_a: τ = 0.515, p = 0.020; Figure S6). In *D. melanogaster*, however, we found no significant positive correlations (ω: τ = -0.652, p = 0.652; ω_{na}: τ = 0.333, p = 0.293; ω_a: τ = -0.333, p = 0.293; Figure S7 in Appendix II), suggesting a comparatively weaker effect of gene age in *Drosophila*.

The functions of lineage-specific genes

Lineage-specific genes are known to be involved in species-specific adaptive processes, such as the evolution of morphological diversity (Khalturin et al. 2009) and immune and stress responses (e.g., Kuo and Kissinger 2008; Khalturin et al. 2009; reviewed in Tautz and Domazet-Lošo 2011). As proteins encoding such functions tend to have higher molecular rates of adaptation (Sackton et al. 2007; Obbard et al. 2009; Slotte et al. 2011; Stukenbrock et al. 2011; Enard et al. 2016; Moutinho et al. 2019b), we further assessed whether these could be confounding the effect of gene age. We first examined which functions are encoded by young genes in these species. Our analyses showed that, in *Drosophila*, lineage-specific genes (Clades 11 and 12 in Figure 1a) encode mostly functions involved in response to stress, nervous system processes, enzyme regulators, and immune system mechanisms (Figure 4a). In *Arabidopsis*, young genes (Clades 14 and 15 in Figure 1a) seem to be involved in a large variety of cellular processes, but also in response to stress and external stimulus, protein binding, and signal transduction (Figure 4b). To note, however, that these functions represent general terms and not direct gene products due to the difficulty of annotating young genes.

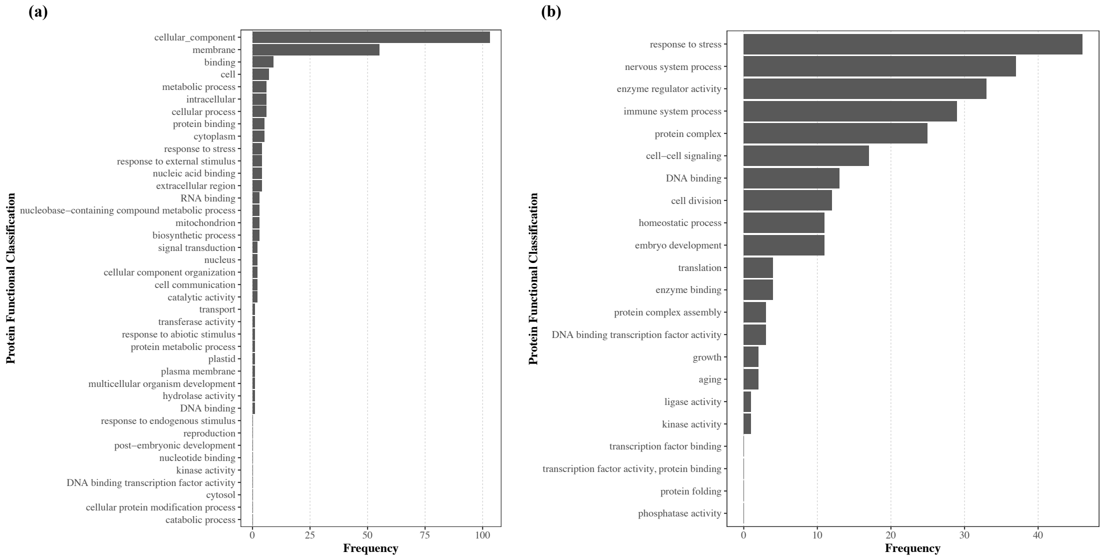

Figure 4. Frequency of young genes for the respective protein function in (a) *D. melanogaster* and (b) *A. thaliana.*

To further correct for the potential bias of gene function, we used the gene ontology (GO) terms (Gene Ontology Consortium 2004) holding the highest numbers of young proteins (above 10) that were also well distributed throughout the oldest clades. Due to the limited number of genes available for analyses, we could only compare two age classes, which we classified as "old" and "young". In *D. melanogaster*, proteins were considered "old" if they were in the root of the tree (clade 1 in Figure 1a) and "young" otherwise. In *A. thaliana*, genes belonging to clades 1 to 7 were considered "old", and other age classes as "young" (Figure 1a). In *D. melanogaster*, we observed a strong effect of gene age on ω_a for proteins involved in the homeostatic process, protein complex, and response to stress, with younger genes presenting higher molecular adaptive rates (Figure 5a). These are known functions involved in immune and stress responses, particularly in the co-evolutionary arms-race between the host and parasites (Obbard et al. 2009). Likewise, in *A. thaliana*, we found that the impact of gene age on ω_a is stronger in proteins implicated in stress response, extracellular regions, and cellular components (Figure 5b). Although the GO terms extracellular regions and cellular components represent broad annotations, they denote for the cellular compartments where processes such as signal transduction and membrane trafficking occur, which are essential for the maintenance of the cell homeostasis (Geldner and Robatzek 2008; Groen et al. 2008). Estimates of ω_{na} revealed a strong influence of gene age in all functions analysed in *A. thaliana*, where young genes present higher rates of non-adaptive substitutions (Figure 5b). In *D. melanogaster*, the same pattern is observed for proteins involved in the immune and stress response (Figure 5a). These results suggest that, by restricting the analysis to proteins involved in the immune and stress responses, which are known to adapt faster (e.g., Slotte et al. 2011; Stukenbrock et al. 2011; Enard et al. 2016), gene age still has an impact on the efficiency of selection acting upon a protein.

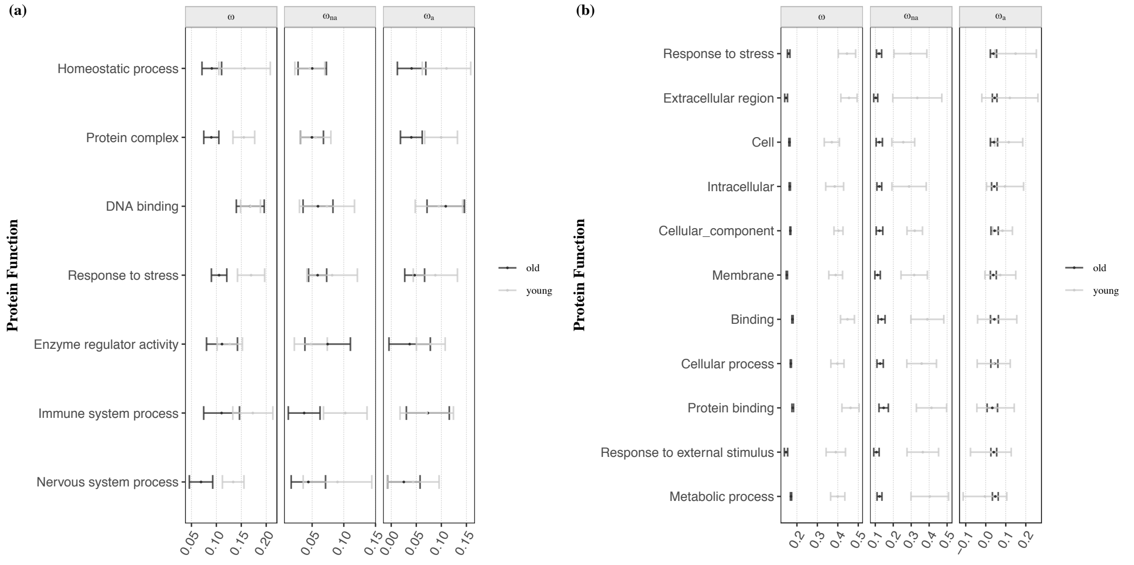

Figure 5. Estimates of ω, ω_{na} and ω_a plotted as a function of protein function and gene age in **(a)** *A. thaliana* and **(b)** *D. melanogaster*. Categories are ordered according to the values of ω_a. Mean values of ω, ω_{na} and ω_a for each class are represented with the black points. Error bars denote the 95% confidence interval for each category, computed over 100 bootstrap replicates.

3.4 Discussion

Overall, our findings suggest that gene age significantly impacts the rate of protein adaptive evolution, with young genes presenting higher rates of adaptive substitutions. The same pattern is observed when looking at the efficiency of purifying selection, where young genes accumulate comparatively more deleterious mutations (Figure 1b). By looking at the magnitude effect of gene age, we observed that young genes present a 25-fold higher adaptation rate in *D. melanogaster* and around 30-fold in *A. thaliana*, higher than that observed for recombination rate and solvent exposure in these species (Castellano et al. 2016; Moutinho et al. 2019b). Moreover, the analyses of the potential confounding effects of protein length, gene expression, RSA, protein disorder, and protein function revealed that the age of a protein is a key contributor to the molecular adaptive rate at the gene level (Table 2 and Figure 5). When looking at protein divergence, however, the effect of gene age only persisted in *Arabidopsis* (Figure S6), suggesting a comparatively weaker impact of gene age in *Drosophila*. We further discuss the inherent limitations of our study as well as the potential drivers of the higher adaptive substitution rates of young genes.

<u>Potential limitations of this study</u>

Even though our approach of protein divergence was extremely conservative, we cannot rule out the possibility that, in *Drosophila*, the lack of effect of gene age after correcting for protein divergence from the false negative's rates of phylostratigraphy. Multiple studies have discussed the potential limitations of this method, with authors questioning its accuracy (Elhaik et al. 2006; Moyers and Zhang 2015, 2016), and others defending it (Albà and Castresana 2007; Domazet-Lošo et al. 2007). The fast evolution of young proteins raised the question of whether old but fast-evolving genes could be misclassified as "young," as BLAST might fail to identify homologs in these proteins. Domazet-Lošo et al. (2017), however, provided evidence for the reliable identification of young genes even when considering a false negative rate of 11-15% in BLAST searches. We, therefore, propose three other scenarios that could explain the weak signal observed in *Drosophila*. First, this effect could result from the low number of genes analysed in each clade. While for *A. thaliana,* we managed to sample a total of 1,529 genes, for *D. melanogaster*, only a sample of 421 genes was possible. Second, in *Drosophila*, the observed effect of gene age on ω_a appears to be mostly driven by the two youngest clades (Figure 1b), whereas for the rest of the phylostrata, the correlation loses its power (ω: $\tau = 0.600$, $p = 0.016$; ω_{na}: $\tau = 0.556$, $p = 0.025$; ω_a: $\tau = 0.467$, $p = 0.060$). Hence, by removing the number of genes for analysis, we could be removing this effect. In contrast, in *Arabidopsis*, the effect of gene age still stands after removing the two youngest clades (ω: $\tau = 0.9487$, $p = 6.342\text{e-}06$; ω_{na}: $\tau = 0.872$, $p = 3.345\text{e-}05$; ω_a: $\tau = 0.692$, $p = 9.86\text{e-}04$). Last and somewhat related to the latter, the weaker effect of gene age in *D. melanogaster* could be derived from the fact that multiple adaptive peaks occurred along the phylogeny. Indeed, the shape of the correlation between ω_a and gene age in *Drosophila* is not gradually increasing, but instead has a peak in the adaptive substitution rate around the clades 6 and 7 (Figure 1b). Intriguingly, this pattern seems to follow the rate of emergence of young genes in this species (Tautz and Domazet-Lošo 2011). This adaptive peak appears before the major radiation of animal phyla, around the time when Earth was passing through glacial cycles (Hoffman et al. 1998). In turn, the burst of the emergence of new genes in *Arabidopsis* was reported to coincide with the plant-specific radiation, right before the emergence of Brassicaceae (Wang et al. 2009; Tautz and Domazet-Lošo 2011). This trend is consistent with our results in *A. thaliana*, where the more pronounced adaptive peaks occur in younger clades (after clades 11 and 12 in Figure 1b). These patterns were also observed in the analysis of gene age with the same number of genes in all clades (Figure S7).

Another challenge that we had to overcome was the lack of structural annotations for young genes. Even though we observed a relatively good correlation between the prediction method and the annotated PDB structures (see Material and Methods), it remains the possibility of potential artefacts from this analysis. Hence, more annotated PDB structures would be required to further confirm the effect of RSA and gene age. Besides, we have also to point out that our study could only assess the effect of gene age in proteins for which a homolog exists in the outgroup species, as estimates of divergence are needed to infer the molecular adaptive rate (Eyre-Walker et al. 2006; Eyre-Walker and Keightley 2009; Galtier 2016). We could, therefore, be underestimating the impact of gene age on the rate of adaptive substitutions by not accounting for the most recently emerged genes. Nonetheless, our study represents a first attempt of inferring the impact of gene age on the molecular adaptive rate and gives light to the potential strong influence of the evolutionary origin of a gene in species adaptation to their environments.

The adaptive interplay between gene age and protein function

Our findings suggested that the effect of gene age on the molecular adaptive rate differs between species, likely being correlated with major species diversifications. Indeed, studies across taxa have proposed that young genes are involved in lineage-specific adaptations to their environments, which could explain this pattern. In protozoa, Kuo and Kissinger (2008) found that many genus- or species-specific genes code for surface antigens that are involved in host-parasite interactions. In basal metazoans, such as *Nematostella* (Babonis et al. 2016) and *Hydra* (Khalturin et al. 2009), young genes were also found to be important in the defence and stress responses. In plants, the same pattern is observed, with lineage-specific genes playing a central role in species specification and defence response against pathogens (Cui et al. 2015). In *Drosophila*, Chen et al. (2010) reported that new genes are involved in essential functions for the viability of this organism. Our study supported these findings and further revealed that young genes also contribute to the higher rates of adaptive evolution in proteins involved in the defence mechanisms in *Drosophila* and *Arabidopsis*. Moreover, we showed that young genes are likely enhancers of the relaxation of purifying selection detected in these proteins (Figure 5). This study, therefore, highlights the strong relationship between the age of a protein and its function, suggesting that both factors may be contributing to the higher rates of adaptive evolution observed in young genes.

What drives the higher rates of adaptation in young genes?

Our study further emphasized that the faster evolution observed in young genes is driven both by a higher rate of adaptive and non-adaptive substitutions. These findings suggest that, after their emergence, young genes evolve through relaxed selection, as first proposed by Ohno (1970), but also by acquiring beneficial mutations, as described in the "adaptive-conflict" model (Piatigorsky and Wistow 1991; Hughes 1994). Ohno's idea of evolution was "non-Darwinian" in its nature, as he believed that "natural selection merely modified while redundancy created" (Ohno 1970). He proposed that new genes evolve through the accumulation of "forbidden" mutations, where they are only preserved if the development of a formerly non-existent function occurs, a process known as neo-functionalization. In this scenario, natural selection only acts at the stage of acquiring a new function. Further extensions of this theory suggested that the preservation of a new gene can also occur through sub-functionalization, where the accumulation of deleterious mutations leads to a complementary loss of function in both copies of the gene (Force et al. 1999; Prince and Pickett 2002). In contrast, the "adaptive-conflict" model assumed that the ancestral gene can carry more than two, although pleiotropically constrained functions (Piatigorsky and Wistow 1991; Hughes 1994). Once the duplication event occurs, each copy then becomes specialized in one of the ancestral functions. In this case, the split of the ancestral gene proceeded through positive Darwinian selection (Piatigorsky and Wistow 1991; Hughes 1994). These theories are based on the evolution of gene duplicates and are in line with the idea of evolution as a "tinkerer" proposed by Jacob (1977), where evolution adjusts the already existing elements. In *de novo* evolution, however, new genes emerge by acquiring new functions from the non-coding fragments of the genome (Cai et al. 2008; Heinen et al. 2009; Tautz and Domazet-Lošo 2011). This process is thought to proceed through a stochastic phase followed by the successive accumulation of beneficial mutations, ultimately leading to a new function with a species-specific selective advantage (Carvunis et al. 2012; Neme and Tautz 2014; Palmieri et al. 2014; Zhao et al. 2014).

If we look at the fundamental ideas behind these theories, we can draw one major feature that portraits the evolution of new genes: young genes are further away from their optimal conditions. Hence, we posit that adaptation in these genes agrees with an "adaptive walk" model (Wright 1932; Smith 1970b; Orr 2002). As their full potential has yet to be met, more consecutive beneficial mutations are theoretically needed to reach their fitness optimum, thus leading to the higher molecular adaptive rates observed in these genes. In turn, older genes are closer to their optimal features, thus only accumulating mutations with small effects on fitness, translating into the lower rates of adaptation observed in these proteins. Our study, therefore, highlights that the distribution of beneficial mutations across deep evolutionary time-scales follows a pattern of diminishing returns.

3.5 Material and Methods

The *D. melanogaster* and *A. thaliana* datasets were taken from Moutinho et al. (2019b) and included a total of 10,318 and 18,669 genes respectively, with data on protein length, gene expression. Gene age data was obtained from published data sets, wherein 9,004 Drosophila (Domazet-Lošo et al. 2017) and 15,935 Arabidopsis (Arendsee et al. 2014) genes were used. Analyses were performed dividing the genes into 12 and 15 phylostrata for *D. melanogaster* and *A. thaliana*, respectively, according to the branches annotated. The analyses on the X-linked and autosomal genes in *D. melanogaster* were performed with 1478 and 7526 genes respectively.

Categorization of protein length and gene expression with gene age

For the comparison between variables at the gene level we divided the dataset into two categories of protein length and gene expression, trying to keep similar number of genes between them. For the analysis of protein length, we used the full set of genes for which gene age data was available. Short proteins had length up to 366 and 389 amino-acids, and long proteins were the ones with length up to 4,674 and 5,098 amino-acids in *A. thaliana* and *D. melanogaster* respectively. Due to the low number of genes across clades, the stratification analyses were accomplished by combining genes across phylostrata. For *D. melanogaster,* gene age was categorized in 6 main clades by combining clades 3 and 4, 5 and 6, 7 to 10, and 11 and 12, keeping the others unchanged. In *A. thaliana*, the 15 clades were combined in 6 main clades by merging clades 5 to 8 and clades 9 to 15. For gene expression, a total of 16,117 and 6,247 genes were used for *A. thaliana* and *D. melanogaster* respectively. Genes were categorized as lowly expressed if the mean expression levels were up to 10.3 and 6.8, and highly expressed genes were the ones with expression up to 6,632.8 and 4,237.0 in *A. thaliana* and *D. melanogaster* respectively. For *D. melanogaster*, gene age was categorized in 6 categories by combing clades 3 to 5, 6 to 9, and 10 to 12.

Categorization of protein structure with gene age

Since most young genes lack a defined three-dimensional structure (Wilson et al. 2017), they do not have information on the residue's solvent accessibility. Hence, we used a deep learning approach, NetSurfP-2.0, that predicts the RSA of each residue from the amino-acid sequence (Klausen et al. 2019) by applying the HH-suite sequence alignment tool for protein similarity searches (Remmert et al. 2012). To assess whether this approach provided reliable results, we compared the RSA estimates of NetSurfP-2.0 with the ones obtained from the PDB structures (Moutinho et al. 2019b). We found a good correlation between the two approaches for both species ($\tau = 0.571$, $p < 2e\text{-}216$; $\tau = 0.462$, $p < 2e\text{-}216$, for *D. melanogaster* and *A. thaliana* respectively). RSA estimates were successfully obtained for a total of 4,238,686 and 7,479,807 amino-acid residues for *A. thaliana* and *D. melanogaster* respectively. The stratification analysis was performed by comparing buried (RSA < 0.05) and exposed (RSA >= 0.05) residues, according to Miller et al. (1987). The phylostrata groups were defined by combining clades 5-6, 7-8, 9-10, and 11-12 in *D. melanogaster*, and 8-11, and 12-15 in *A. thaliana*. To correct for the variation within each category of RSA we then took two subsets of sites with similar RSA estimates. For lower RSA estimates we took sites with values between 0.03 and 0.05 in *Drosophila,* making a total of 187,026 sites, and among 0.10 and 0.20 in *Arabidopsis*, for a total of 816,047 sites. For higher RSA estimates, we used sites with values between 0.55 and 0.60 in *Drosophila*, making a total of 386,586 sites, and among 0.60 and 0.65 in *Arabidopsis*, for a total of 444,995 sites. For this analysis, the phylostrata groups were defined by combining clades 7-9 and 11-12 in *D. melanogaster*, and 9-11 and 13-15 in *A. thaliana*. The stratification analysis of RSA per gene was performed for the total number of genes in both species by making two categories of genes according to their proportion of exposed residues (RSA > 0.05). Genes with lower proportions of exposed residues had values between 0.44 and 0.92 in *Drosophila*, and among 0.689 and 0.89 in *Arabidopsis*. Genes with higher proportion of exposed sites had values between 0.92 and 1 in *Drosophila*, and among 0.89 and 1.00 in *Arabidopsis*. The phylostrata groups were defined by combining clades 5-7, 8-10, and 11-12 in *D. melanogaster*, and 10-11, and 12-15 in *A. thaliana*.

The analysis of residue intrinsic disorder was successfully achieved for a total of 7,126,304 and 3,645,645 sites for *A. thaliana* and *D. melanogaster* respectively. Sites classified as having low intrinsic disorder were the ones with a value up to 0.066 and 0.068, and the ones with high intrinsic disorder had a value up to 0.586 and 0.590 for *A. thaliana* and *D. melanogaster* respectively. In *D. melanogaster*, analyses were accomplished with the 12 clades initially described. In *A. thaliana*, the 15 clades were combined in 6 main clades by merging clades 5 to 8 and clades 9 to 15.

Correcting for protein divergence

Protein divergence was obtained for each gene by computing the proportion of amino-acid differences. To remove the positive correlation between protein divergence and gene age we chose an arbitrary value (0.02 in both species) and randomly sampled around 100 and 150 genes, in *D. melanogaster* and *A. thaliana* respectively, that were at the maximum difference of 0.01 to that value. Due to the low number of genes available for this analysis, we combined clades 3-4, 5-6, and 7-10, each containing between 15 to 98 genes, making a total of 421 genes in *Drosophila*. In *Arabidopsis*, the phylostrata groups were defined by combining clades 11-12 and 13-15, each containing between 43 and 150 genes, making a total of 1,529 genes for analysis.

Categorization of protein function with gene age

Gene ontology terms were obtained from the "dmelanogaster_gene_ensembl" and the "athaliana_eg_gene" databases, for *D. melanogaster* and *A. thaliana* respectively, using the R package "biomaRt" (Durinck et al. 2005). These analyses were accomplished with a total of 2,710 and 15,604 genes for *D. melanogaster* and *A. thaliana*, respectively, for which annotations were available. The comparison between old and young genes was performed by considering the genes in the root of the tree (Clade 1 in Figure 1a) as "old" and the rest as "young" in *D. melanogaster*. In *A. thaliana*, genes belonging to clades 1 to 7 were considered old, and young otherwise.

Estimation of the adaptive and non-adaptive rate of non-synonymous substitutions

For all analysis, 100 bootstrap replicates were made by randomly sampling genes or sites in each category. The Grapes program was then run with the Gamma-Exponential distribution of fitness effects (Galtier 2016). Results for ω, ω_{na} and ω_a were plotted using the R package "ggplot2" (Wickham 2016) by taking the mean value and the 95% confidence interval of the 100 bootstrap replicates performed for each category. Statistical significance was assessed with Kendall's correlation tests. To do so, the mean value of the 100 bootstrap replicates was taken for each category. An analysis of covariance (ANCOVA) was performed using the estimates of ω_{na} and ω_a as response variables, and gene age as an explanatory variable, in combination with chromosome type (X or autosome), protein length, gene expression, residue intrinsic disorder, and RSA, and their respective interactions. Normality, homoscedasticity, and independence of the error terms of the model were assessed with the package "lmtest" (Zeileis and Hothorn 2002) in R (R Core Team 2017).

CHAPTER IV

What Is the Interplay Between Intramolecular Variation and Patterns of Adaptation at the Species Level?

4.1 Abstract

The frequency and nature of adaptive mutations are widely heterogeneous between species. For instance, fruit flies and wild mice exhibit higher adaptation rates than primates and plants. What determines this variation is, however, not fully understood. Over the years, several studies have proposed different hypotheses to explain such heterogeneity in rates of adaptation. Some rely on the stochastic population genetics theory at the molecular level, while others consider the phenotypic space, where each organism is represented as a number of dimensions climbing a fitness landscape. Molecular rates of adaptation, however, also vary between and within genes. Such variation can confound comparative analyses at the species level. Here, we try to understand the variability in adaptation rates between species by accounting for patterns of variation at the intramolecular level. We used a comparative population genomics approach across multiple animal species with distinct life-history traits. To estimate the rate of adaptive substitutions, we fitted models of distributions of fitness effects at the amino-acid residue level. We found a negative correlation between molecular rates of adaptation and the effective population size (N_e). Despite the relatively weak effect, our findings contradict the N_e hypothesis on positive selection. Instead, they are in line with the theoretical expectations at the phenotypic space. Conversely, when looking at the efficiency of negative selection, our findings support the N_e hypothesis. Moreover, we found that this effect reflects the differences in the distribution of fitness effects between buried and exposed residues. In lower-N_e species, exposed residues accumulate more mutations of mild effects due to weak selection. In turn, buried residues will only fix mutations of large effect due to stronger selective constraints. Our study, therefore, emphasizes the importance of assessing the interplay of selective patterns at different organizational levels to shed light on the molecular basis of adaptation.

4.2 Introduction

Molecular rates of adaptation are widely diverse among species (e.g., Gossmann et al. 2010; Halligan et al. 2010; Hvilsom et al. 2012; Galtier 2016; Moutinho et al. 2019a). For instance, fruit flies (e.g., Brookfield and Sharp 1994; Smith and Eyre-Walker 2002; Welch 2006; Sella et al. 2009), mice (Halligan et al. 2010), rabbits (Carneiro et al. 2012a), bacteria (Charlesworth and Eyre-Walker 2006), and some plant species (Ingvarsson 2010; Slotte et al. 2010) present higher proportions of adaptive substitutions (α) when compared to primates (e.g., Boyko et al. 2008; Eyre-Walker and Keightley 2009; Hvilsom et al. 2012; Castellano et al. 2019) and many other plants (Gossmann et al. 2010). Unravelling the determinants of such variation, however, is not an easy task.

At the molecular level, several studies proposed that the observed cross-species variation in rates of adaptation could be attributed to differences in effective population sizes (N_e) (Eyre-Walker et al. 2006; Eyre-Walker and Keightley 2009; Jensen and Bachtrog 2011; Gossmann et al. 2012). On the one hand, population genetics theory predicts that adaptation is limited by the population mutation rate (θ=4$N_e\mu$) (Charlesworth 2009; Karasov et al. 2010; Cutter et al. 2013). Under this assumption, populations with larger N_e, such as fruit flies and mice, adapt faster due to the higher availability of mutations (Gillespie 1999, 2001). On the other hand, the nearly neutral theory (Kimura and Ohta 1971; Ohta 1972, 1973, 1992) predicts that, as the effect of genetic drift is stronger in small-N_e species, the probability of fixation of an advantageous mutation decreases, while the accumulation of slightly deleterious mutations increases (Gillespie 1999; Lanfear et al. 2014). (Gillespie 1999; Lanfear et al. 2014). Conversely, large-N_e species are under stronger selection, thus removing deleterious mutations at a faster rate and increasing the probability of fixation of a beneficial mutation (Ohta 1972, 1973, 1992; Orr 1998; Gillespie 1999; Lanfear et al. 2014). By performing a comparative analysis across 44 different species, Galtier (2016) showed that N_e was positively correlated with α, which would corroborate the "adaptation limited by mutation" theory. Fluctuations in α, however, can be explained both by the effect of negative (ω_{na}) and positive (ω_a) selection, as $\alpha = \omega_a / (\omega_{na} + \omega_a)$. By looking at the individual effect in these two components, he showed that N_e was negatively correlated with the rate of non-adaptive substitutions (ω_{na}). However, he did not find any significant correlation between the rate of adaptive substitutions (ω_a) and N_e, suggesting that the effect on α was derived from the fraction of mutations that have deleterious effects.

In turn, long-term fluctuations in N_e can bias estimates of molecular adaptive rates (Eyre-Walker and Keightley 2009; Jensen and Bachtrog 2011; Rousselle et al. 2018). On the one hand, a decrease in population size may underestimate rates of adaptation because slightly deleterious mutations might be detected as polymorphism, while negligibly contributing to divergence. On the other hand, a demographic expansion may overestimate the molecular adaptive rate, as the low polymorphism levels mirror a pattern of an excess of substitutions (Eyre-Walker 2002). Rousselle et al. (2019) recently assessed the long-term and short-term effects of N_e by comparing results among groups of distantly-related and closely-related species, respectively. The authors found that, when comparing closely-related species, there was a positive correlation between ω_a and N_e. By contrasting groups of distantly-related species, however, they found a weak negative correlation between ω_a and N_e. Rousselle et al. (2019) suggested that the observed differences between time-frames reflect the hypothesis that the long-term N_e affects the distribution of fitness effects (DFE) and, consequently, the molecular rate of adaptation.

At the phenotypic level, these theoretical expectations take a turn. Populations are expected to suffer from the so-called "cost of complexity" (Orr 2000). This theory is based on the Fisher's geometric model of adaptation (Fisher 1930a). Fisher suggested that, in more complex species, *i.e.*, larger long-lived organisms, mutations are more likely to be detrimental than beneficial. As Orr (2000) mentioned: "Changing the length of an arbitrary mechanical part by one inch, for instance, is more likely to derail the function of a microscope than a hammer". This idea derives from the concept of high dimensionality: as a larger number of dimensions is available in more complex organisms, the adaptive walk takes more steps to reach their fitness peak. Consequently, these organisms, which typically have small-N_e, adapt slower than simple ones (Orr 2000; Welch and Waxman 2003). Intriguingly, a higher proportion of adaptive substitutions should be expected in such less efficient adaptive walks: as the number of traits (*i.e.*, dimensions) is larger, a higher number of adaptive changes are necessary to "climb" fitness peaks (Lourenço et al. 2013). This hypothesis was used by Rousselle et al. (2019) to explain the negative correlation observed between ω_a and long-term N_e.

These different findings suggest that the interaction between rates of molecular adaptation and measures of genetic diversity across species is a complex process. Moreover, it is known that molecular adaptive rates vary substantially within genomes. Linked selection, for instance, creates heterogeneous patterns of polymorphism along the genome (Maynard Smith and Haigh 1974; Charlesworth 1994; Gillespie 2000b). Besides, we have recently shown that the macromolecular structure of proteins acts as a major determinant of the molecular adaptive rate, where beneficial mutations accumulate at a faster rate on residues at the surface of proteins (Moutinho et al. 2019b). These factors could, therefore, be confounding comparative population genomic inferences of the relationship between ω_a and N_e. Huber et al. (2017) indeed suggested that models accounting for different biological factors, such as mutational robustness and organism complexity, lead to different predictions on how the DFE varies among species.

Here, we control for this potential bias by analysing patterns of intra-molecular variation between species. With this, we aim to understand the interplay between patterns of selection at different organizational levels. To do so, we analysed a wide range of species with different life-history traits from a previously published dataset (Galtier 2016). We fitted different DFE models across species to estimate the molecular rate of adaptation at the amino-acid residue level. By analysing how the effect of the relative solvent accessibility (RSA) varies across species, our study aims to deliver a better understanding of the variation in molecular rates of adaptation at larger taxonomic scales.

4.3 Results

We analysed patterns of intramolecular variation across 41 species of animals from a previously published dataset (Galtier 2016). The data included eleven mammals, ten arthropods, five sauropsids, four echinoderms, four molluscs, two tunicates, one annelid, one nematode, one ribbon worm, one cnidarian, and one teleost (Table S1 in Appendix III). The number of genes per species varied between 836 and 13,584 (Table S1 in Appendix III). The DFE model comparison showed that the ScaledBeta and GammaExponential models had the best fit for the majority (~64%) of the species (Table S1 in Appendix III), suggesting the prevalence of segregating beneficial mutations in these taxa, in agreement with what Galtier (2016) reported. To infer the effect of positive and purifying selection across species, we estimated rates of adaptive (ω_a) and non-adaptive (ω_{na}) amino-acid substitutions with the Grapes program (Galtier 2016).

<u>The impact of the macromolecular structure on rates of adaptive and non-adaptive substitutions between species</u>

To assess the effect of the macromolecular structure on the efficiency of selection between species, we analysed the relationship between the effect of the residue's RSA and species genetic diversity (π_S), here used as a proxy for the effective population size (N_e). The separate analysis of buried and exposed residues across species suggests a substantial variation on the magnitude effect of RSA both on ω_{na} (Figure S1 in Appendix III) and ω_a (Figure S2 in Appendix III). By looking at the singular correlation of buried and exposed residues with the log-transformed π_S, we observed a significant negative correlation for estimates ω and ω_{na} for both types of residues (Table 1 and Figure 1a). While we observed a negative trend for estimates of ω_a, the correlation was not significant (Table 1 and Figure 1a). When looking at the relationship between the differences in ω, ω_{na}, and ω_a between exposed and buried residues, we confirmed the negative correlation with the log-transformed π_S (Figure 1b). Moreover, by assessing the effect of RSA between species, we observed a much higher variability for estimates of ω_a than ω_{na}, particularly in lower π_S species (Figure 1b). For instance, primates and ants present a higher variation between residues than molluscs and butterflies (Figure 1b).

To further assess the interaction between RSA and π_S, we discretized RSA values in ten categories with similar numbers of sites for each species. Our results suggested that the correlations for ω_{na} and ω_a with π_S are stronger for lower values of RSA, suggesting a lower variation in estimates of ω_{na} and ω_a for buried residues (Table 2 and Figure 2). When looking at the slope of the linear regression, we observed a strong relationship between the log-transformed π_S and ω_{na}, which becomes steeper with higher RSA values (Table 2 and Figure 2). For ω_a, however, this relationship appears to be weaker, with higher values observed for intermediate values of RSA (Table 2 and Figure 2). By jointly analysing all species with an ANCOVA analysis, we observed that, for ω_{na}, there is a significant effect for the log-transformed π_S and a marginally significant interaction between RSA and the former, suggesting a stronger impact of π_S for higher RSA values (Table 3). For ω_a, our results suggested a significant effect of both RSA and log-transformed

π_S. These effects, however, seem to be purely additive, since the interaction between the two variables was not retained by the model selection procedure (Table 3). As we were dealing with a wide range of different taxa, we assessed whether the phylogenetic relationship between species was biasing our results (Felsenstein 1985). After correcting for the effect of the phylogeny, we observed that the significant negative correlation prevails for estimates of ω_{na} in all categories of RSA. For ω_a, however, no significant negative correlation was found (Table S2 in Appendix III). This pattern suggests a generally weaker effect of the log-transformed π_S on estimates of ω_a.

Table 1. Statistical results for the analysis of the effect of the log-transformed π_S on ω_{na} and ω_a in buried and exposed residues.

RSA	$\boldsymbol{\omega_{na}}$	$\boldsymbol{\omega_a}$
Buried	-0.652 (***)	-0.021
Exposed	-0.722 (***)	-0.211

Note. For each variable, the Pearson correlation coefficient (ρ) of the log-transformed π_S is shown with the respective significance (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; "." $0.05 \leq P < 0.10$) for ω_{na} and ω_a.

Table 2. Statistical results for the analysis of the effect of the log-transformed π_S on ω_{na} and ω_a in ten categories of RSA.

RSA	$\boldsymbol{\omega_{na}}$		$\boldsymbol{\omega_a}$	
	Correlation (ρ)	*Slope*	*Correlation (ρ)*	*Slope*
0.031	-0.729 (***)	-0.027 (***)	-0.011	-0.004
0.130	-0.723 (***)	-0.028 (***)	-0.178	-0.009
0.253	-0.707 (***)	-0.028 (***)	-0.223	-0.015
0.370	-0.646 (***)	-0.033 (***)	-0.324 (*)	-0.014
0.471	-0.706 (***)	-0.035 (***)	-0.198	-0.012
0.559	-0.616 (***)	-0.032 (***)	-0.281 (.)	-0.013
0.630	-0.613 (***)	-0.040 (***)	-0.130	-0.008
0.683	-0.494 (**)	-0.039 (***)	-0.085	-0.010
0.731	-0.555 (***)	-0.037 (***)	-0.098	-0.003
0.781	-0.453 (**)	-0.036 (***)	-0.040	-0.002

Note. For each variable, the Pearson correlation coefficient (ρ) and the coefficient of the linear regression of the log-transformed π_S is shown with the respective significance (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; "." $0.05 \leq P < 0.10$) for ω_{na} and ω_a.

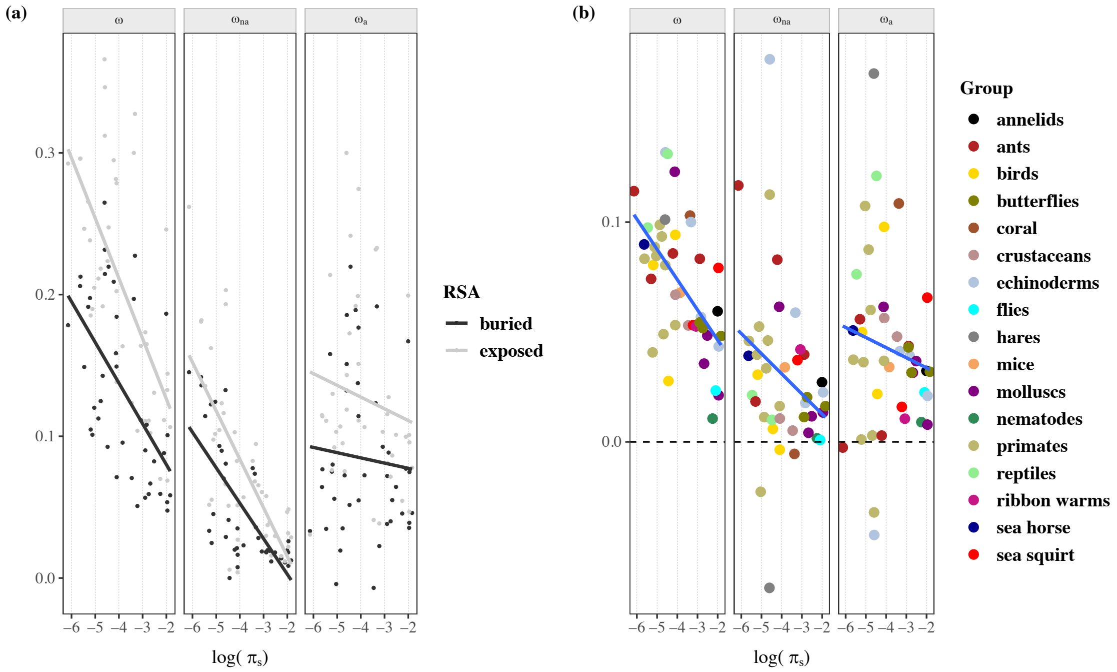

Figure 1. Relationship between the rate of protein evolution (ω), non-adaptive non-synonymous substitutions (ω_{na}), and adaptive non-synonymous substitutions (ω_a) with the log-transformed π_S for **(a)** the separate analysis of buried and exposed residues, and **(b)** the differences in ω, ω_{na}, and ω_a estimates between exposed and buried residues. **(a)** Each dot represents the mean values of the 100 bootstrap replicates performed for buried and exposed residues in each species. **(b)** Each dot represents the difference in ω, ω_{na}, and ω_a between exposed and buried residues for the respective species. Species are coloured according to the taxonomic group (see Table S1 in Appendix III). In both **(a)** and **(b)**, lines represent a linear model performed as a function of RSA with the log-transformed π_S.

Table 3. ANCOVA estimates for the proportion of contribution of the log-transformed π_S, RSA, and their interaction obtained with the best model procedure in ω_{na} and ω_a.

	$\boldsymbol{\omega_{na}}$	$\boldsymbol{\omega_a}$
log ($\boldsymbol{\pi_S}$)	86.67% (***)	22.25% (**)
RSA	11.62% (***)	74.68% (***)
Interaction	1.26% (.)	-

Note. For each variable, the proportion of explained variance is shown with the respective significance (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; "." $0.05 \le P < 0.10$) for ω_{na} and ω_a.

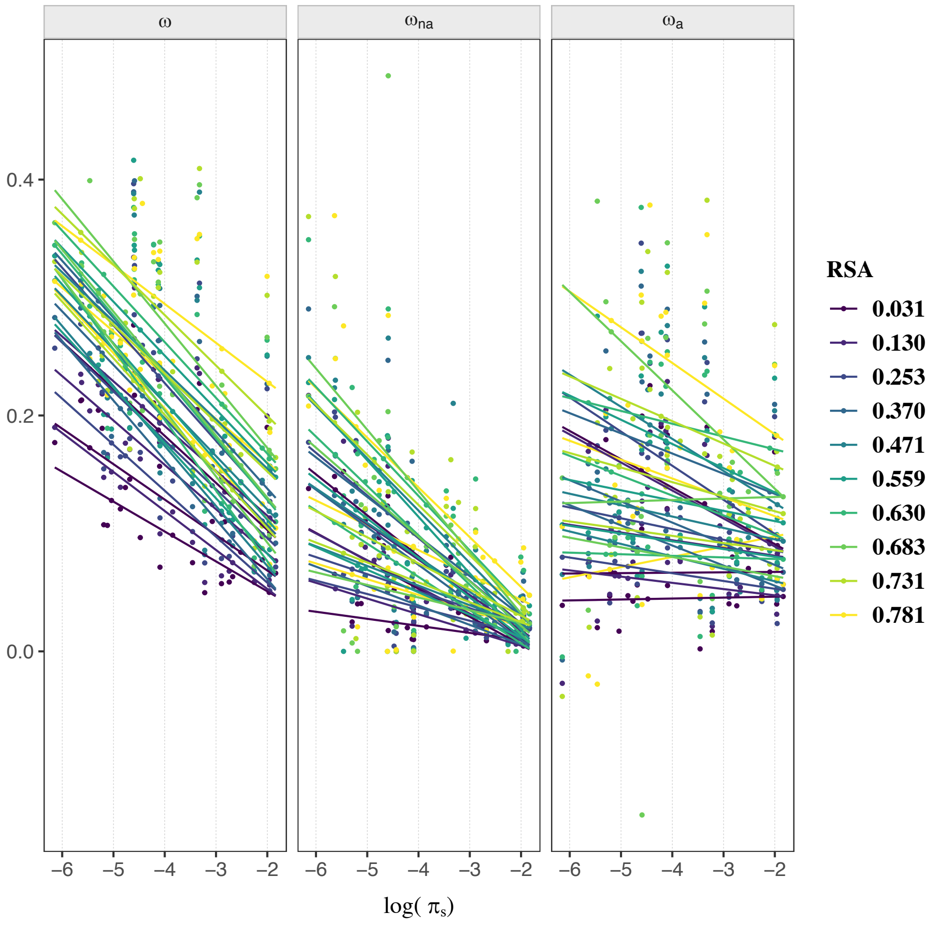

Figure 2. Relationship between ω, ω_{na}, and ω_a with the log-transformed π_S for 10 categories of RSA. Each dot represents the respective estimates for the different RSA values in each species. For each category of solvent exposure, a quantile regression was fitted to the data, which is represented with the respective colours of each RSA category.

4.4 Discussion

We first discuss some potential limitations associated with the study of adaptation between species with different life-history traits. We then provide an overview of the potential drivers of the variation in molecular adaptive rates at the species level by discussing the interplay of adaptive mutations at distinct levels of organization.

Potential limitations of the study of adaptation at the species level

The analyses of several different species carry potential limitations. On the one hand, the wide range of gene numbers across taxa could bias estimates of ω, ω_{na}, and ω_a due to the different number of sites available for analyses. To correct this issue, we randomly down-sampled the same number of sites in each species and re-analysed Grapes on that subset of the data. Our results showed a good correlation between the reduced and the full datasets for all estimates ($\rho > 0.484$, $p < 0.001$; Figure S3 in Appendix III), thus suggesting that the number of analysable sites does not bias our inferences of the rate of adaptive and non-adaptive substitutions.

Besides, the effect of linked selection could also be biasing our results, since this effect differs between species. As the frequency of selective sweeps is stronger in populations with larger N_e (Castellano et al. 2018; Chen et al. 2020), a higher proportion of neutral genetic variants at closely linked beneficial mutations (*i.e.*, genetic draft) will be removed (Gillespie 2000b, 2001; Castellano et al. 2018; Chen et al. 2020). This effect could, therefore, be confounding estimates of ω_{na} and ω_a in large-N_e species. With the data available, however, we could not test for this effect. One way to overcome this limitation would be to estimate the recombination map of each species, as this would provide a thorough overview of the patterns of linkage between alleles throughout the genome. More and high-quality genomic data would be required to perform such analyses.

Moreover, the use of π_S as a proxy for N_e constitutes another limitation, as π_S reflects both N_e and levels of mutation rate. Assessing the mutation rate landscape for each species would, therefore, contribute to better inferences of adaptation rates across species.

What is the interplay of adaptation at distinct organizational levels?

Overall, our results suggest a strong effect of solvent exposure both on the rate of adaptive and non-adaptive amino-acid substitutions across species, thus expanding our previous findings on *Drosophila* and *Arabidopsis* (Moutinho et al. 2019a). When looking at the magnitude of this effect between species, our analyses suggest a stronger negative correlation between the rate of non-adaptive substitutions and the log-transformed π_S, here used as a proxy for N_e. Intriguingly, we found that this pattern is amplified for higher values of RSA (Table 2 and Figure 2). These results suggest that, for low-N_e species, such as primates and ants, the differences between RSA classes is enlarged, where buried residues appear under stronger purifying selection than exposed ones. In contrast, in large-N_e species, like butterflies and flies, the differences between exposed and buried residues substantially decrease, suggesting a stronger effect of purifying selection both on residues at the surface and the core of the protein structure (Figure 2).

Analyses on the molecular adaptive rate, however, showed a weaker negative correlation with the log-transformed π_S. Despite the lack of significance, the interaction between N_e and RSA seems to follow the same trend as in ω_{na} (Table 1 and Figure 2). In this way, lower-N_e species would be more likely to fix advantageous mutations at the surface of the proteins when compared to species with larger effect sizes. However, as beneficial mutations are rare, we may be lacking power in the species comparisons performed in this study. By using a larger number of species and a deeper evolutionary scale, Rousselle et al. (2019) reported a significant negative correlation between ω_a and the long-term N_e. This signal could, therefore, be amplified if more species were included. In contrast to the weak effect of π_S on ω_a, our results suggest a strong impact of solvent exposure on the variation of ω_a within species, thus supporting our previous findings on the relevance of the macromolecular structure on the rates of protein adaptation (Moutinho et al. 2019a). This strong effect could also be influencing the weaker pattern observed at the comparison between species.

Our study further revealed that, by contrasting different structural classes of residues, we could detect a significant negative relationship between the molecular adaptive rate and the effective population size of the species (Table 2). These findings suggest that lower-N_e species have a higher chance of accumulating beneficial mutations. This pattern contradicts the initial prediction of the stochastic population genetics theory: that populations with larger π_S adapt at higher rates (e.g., Eyre-Walker 2006; Charlesworth 2009; Karasov et al. 2010; Gossmann et al. 2012). Instead, our findings seem to agree with the theoretical expectations at the phenotypic space. Under these assumptions, the rate of adaptive substitutions is expected to vary according to the rate of environmental change, which, in turn, is proportional to the generation time (Gillespie 2001; Lourenço et al. 2013). These predictions are directly linked with the notion of dimensionality and the Fisher's geometric model of adaptation (Fisher 1930a): more complex species, which usually have longer generation times, take more steps in an adaptive walk, thus accumulating comparatively more beneficial mutations (Orr 2000; Welch and Waxman 2003; Lourenço et al. 2013). Welch and Waxman (2003) indeed suggested that adaptation in the phenotypic space better resembles a rugged fitness landscape, comprising alternative phenotypic optima (e.g., Kauffman and Levin 1987). Under this model, species with more traits under selection potential acquire higher rates of adaptation due to the comparatively higher availability of multiple optima (Welch and Waxman 2003). Our results are, therefore, in line with this hypothesis. Besides, there are studies suggesting that smaller-N_e species might have a higher proportion of beneficial mutations by merely increasing the mutation load due to weak selection (Weissman and Barton 2012).

Our study confirmed the N_e hypothesis regarding the effect of purifying selection across species, an effect that is amplified at higher levels of solvent exposure. For positive selection, however, our findings contradict the initial assumptions of the stochastic population genetics theory. Instead, our results agree with the hypothesis at the phenotypic space, where species with more traits under selection tend to accumulate more beneficial mutations due to longer adaptive walks. These findings further emphasize the importance of integrating distinct levels of organization to better assess the fitness effects of mutations, thus providing a more profound understanding of the molecular basis of adaptation.

4.5 Material and Methods

Population Genomics Data and Data Filtering

We reanalysed a total of 41 species from a previously published dataset (Galtier 2016) (Table S1 in Appendix III). From this dataset, seven species pairs were "mirror species", as referred by Galtier (2016), where each served as the outgroup for the other (Table S1 in Appendix III). We started by filtering the data to keep only one sequence with the lowest amount of missing data missing data for each outgroup species. Gene alignments with premature stop codons were discarded. Final dataset sizes ranged from 836 to 13,584 genes per species (Table S1 in Appendix III). The synonymous and non-synonymous unfolded site frequency spectrum (SFS), the number of synonymous (Lps) and non-synonymous (Lpn) polymorphic sites, and the synonymous (d_S) non-synonymous (d_N) divergence were estimated using the BppPopStats program from the Bio++ Program Suite (Guéguen et al. 2013). As this dataset included genes with little polymorphism and a substantial amount of missing data, we first estimated the ts/tv ratio per gene with BppPopStats. We then used the estimated median value to correct the estimations of polymorphisms and substitutions counts in each species (Li et al. 1985; Yang and Bielawski 2000) (Table S1 in Appendix III). Moreover, because in most species a large amount of positions was missing genotype information in one or several individuals, we randomly down-sampled polymorphic alleles at each site by keeping 70% of the sample size of each species (see Table S1 in Appendix III). The Grapes program was then used to compute a genome-wide estimate of the rate of adaptive (ω_a) and non-adaptive (ω_{na}) non-synonymous substitutions. The best distribution of fitness effects (DFE) for each species was inferred by comparing six different models using Akaike's information criterion: Neutral, Gamma, Gamma-Exponential, Displaced Gamma, Scaled Beta, and Bessel K. This model comparison was performed on every dataset using the complete set of genes (see Table S1 in Appendix III). The selected model was then used to fit the different subsets of the data according to the macromolecular structure.

As our filtering method differed from the one used by Galtier (2016), we assessed whether estimates of π_N, π_S, π_N/π_S, d_N, d_S, ω, α, ω_{na}, and ω_a were well corroborated between approaches. Our analyses suggested a good correlation between all parameters ($\rho > 0.658$, $p < 1.195e\text{-}06$; Figure S4 in Appendix III). We further assessed the correlation between ω, ω_{na}, and ω_a and the effective population size (N_e) and found the same trend as Galtier (2016) reported: a significant negative correlation between ω ($\rho = -0.630$, $p = 7.529e\text{-}06$) and ω_{na} ($\rho = -0.659$, $p = 2.291e\text{-}06$) with the log-transformed π_S, but no significant correlation between ω_a and the log-transformed π_S ($\rho = -0.111$, $p = 0.474$) (Figure S5 in Appendix III).

Analysis of the Macromolecular Structure

To estimate the relative solvent accessibility (RSA) of each amino-acid residue, we used the program NetSurfP-2.0, which uses a deep learning approach to predict the RSA of each amino-acid from the protein sequence (Klausen et al. 2019). For this, we used the sequence of the focal species with less missing positions and applied the HH-suite sequence alignment tool for protein similarity searches (Remmert et al. 2012). To assess the effect of RSA on each species we divided the sites in buried (RSA < 0.05) and exposed (RSA > 0.05) residues according to Miller et al. (1987). For the continuous analysis of RSA, we discretized amino-acid residues in 10 categories of solvent exposure by keeping similar number of sites in each. Mean values of RSA ranged between 0.031 and 0.781 across categories.

Estimation of the adaptive and non-adaptive rate of non-synonymous substitutions

We performed 100 bootstrap replicates by randomly sampling sites in each category. The Grapes program was then run with the respective DFE for each category of RSA with the total number of sites in each species (Table S1 in Appendix III). Results for ω, ω_{na} and ω_a were plotted using the R package "ggplot2" (Wickham 2016) by taking the mean value and the 95% confidence interval of the 100 bootstrap replicates performed for each category. For the continuous analysis of RSA, results for ω, ω_{na} and ω_a were plotted by fitting a quantile regression to the data. Statistical significance of the correlations between ω, ω_{na} and ω_a and the log-transformed π_S for each RSA class were assessed with the Pearson correlation coefficient. An analysis of covariance (ANCOVA) was performed using the estimates of ω_{na} and ω_a as response variables, and RSA and the log-transformed π_S, as well as their respective interactions, as explanatory variables. A model selection procedure was conducted using the "step" function (Hastie and Pregibon 1992; Venables and Ripley 2002) in R, which sequentially removes single effects and selects the model with the lowest AIC. Normality, homoscedasticity, and independence of the error terms of the selected model were assessed with the package "lmtest" (Zeileis and Hothorn 2002) in R (R Core Team 2017). To analyse the potential effect of phylogeny, a phylogenetic tree was obtained from the NCBI (http://www.ncbi.nlm.nih.gov/) taxonomy using the R package "taxize" (Chamberlain and Szöcs 2013). A generalized least square (GLS) model was used, with Grafen's correlation structure as implemented in the R package "ape" (Paradis et al. 2004). The impact of the phylogeny was fitted using the parameter "rho", jointly estimated with the parameters of the linear model (Grafen 1989). A linear model in the form "response variable ~ log (π_S)", for both ω_{na} and ω_a, was fitted independently for each RSA class.

CHAPTER V

General Discussion

How does adaptation proceed? More than 150 years have passed since Charles Darwin published "The origin of species". With the almost unlimited amount of data and methods to study selection, we now have a deeper understanding of Darwin's evolution. We know that a mutation at the DNA level may lead to a change in the protein sequence, which can cause dramatic changes at a higher organizational level, such as the organism. Such interplay across systems defines the evolutionary path through distinct organizational levels: from the nucleotide to the DNA sequence, to the protein, to the organism, to the population, and, eventually, to the species. An adaptive event follows a similar route. When a new beneficial mutation arises within a population, selection and drift will determine its fate. If this mutation provides a fitness advantage to the organism, then selection will act by increasing its frequency. The spread of this beneficial mutation throughout the population occurs at the DNA level, through the process of inheritance. In turn, this process depends on the fitness effect of that mutation, which is determined at the residue level. These fitness effects may vary along the genome, being contingent on factors such as the functional or structural importance of that region, mutation, and recombination rates. The way selection and drift act at the population level, however, will depend on demography: in small populations, drift will dominate, whereas, in large populations, selection will be more efficient.

Understanding adaptation, therefore, requires a multilevel study of the patterns of selection: a study across systems. This thesis approached adaptation in such a form. By exploring the frequency and nature of adaptive mutations between species, within genomes, and within genes, this project delivered a comprehensive understanding of the molecular basis of adaptation.

5.1 Adaptation within genes

What was already known?

Before the rise of genomics, quantifying the frequency and nature of adaptive mutations within genes was challenging. Instead, most of the studies focused on the variation in rates of protein evolution. One of the most relevant factors under study was the macromolecular structure of a protein. As a stable conformation is usually required to assure proper protein function, mutations that impair this stability are more likely to be counter-selected. Hence, residues at the core of the protein, which are essential to sustain a stable structure, are expected to be more conserved. Several studies have indeed shown that exposed (e.g., Perutz et al. 1965; Choi et al. 2006; Liberles et al. 2012; Chi and Liberles 2016) and disordered residues (Guo et al. 2004; Wilke et al. 2005; Afanasyeva et al. 2018) evolve comparatively faster. However, a question remained: are these residues evolving faster due to less efficient purifying selection or due to the stronger effect of positive selection?

What is new?

This project showed that both a relaxation of purifying selection and a higher rate of adaptive substitutions explain the faster evolution observed in exposed and disordered residues. By analysing multiple confounding factors in animals and in plants, this study further revealed that the residue's solvent accessibility acts as the main determinant of the rate of adaptive evolution at the intramolecular level, being even higher than the effect of mean gene expression levels. Moreover, these analyses showed a higher number of beneficial mutations in genes encoding proteins with central functions in the cell, which are mostly conserved across species. Interestingly, such proteins are targeted by pathogens during host infection, notably viruses. These findings, therefore, suggest that adaptation in proteins is mainly driven by the interactions between molecules, particularly at the between-species level, with host-pathogen coevolution likely playing a major role.

5.2 Adaptation within genomes: a perspective in space and time

What was already known?

Stemming on many years of research, recombination (Hill and Robertson 1966; Marais and Charlesworth 2003; Campos et al. 2014; Castellano et al. 2016) and mutation rates (King and Jukes 1969; Kimura 1983; Ohta 1992; Castellano et al. 2016) are well-known determinants of rates of protein evolution and adaptation at the genome level. Highly recombining regions favour the fixation of adaptive substitutions by breaking down linkage disequilibrium. In turn, regions with high mutation rates adapt faster due to the higher levels of genetic diversity, which increases the chance for adaptation to occur. At the gene level, proteins involved in the immune and stress response (e.g., Nielsen et al. 2005; Sackton et al. 2007; Stukenbrock et al. 2011; Enard et al. 2016) and in sex-related functions (Pröschel et al. 2006; Crowson et al. 2017) were reported with higher rates of adaptive evolution in several species. These studies reflect the vast variability in the frequency of adaptive mutations in the genomic space. The dynamics of these mutations across time, however, remained unexplored.

What is new?

This thesis explored the dynamics of adaptation across time by analysing genes with different evolutionary origins. By accounting for multiple confounding factors, this study overcame the difficulty of assessing the impact of gene age on rates of adaptation. These analyses revealed that young genes adapt at higher rates when compared to more ancient ones. As these genes are theoretically further away from their fitness optimum, these findings suggest that adaptation in young proteins proceeds in an "adaptive walk" manner (e.g., Gillespie 1984; Orr 1998, 1999). This study, therefore, emphasized that the dynamics of beneficial mutations across deep evolutionary scales follow a pattern of diminishing returns.

5.3 Adaptation at the species level

What was already known?

Molecular rates of adaptation vary widely across species. For instance, primates and plants generally have lower rates of adaptive substitutions when compared to fruit flies and mice. Several studies hypothesized that such variation is due to the differences in effective population sizes (N_e), where species with higher N_e potentially adapt faster (Eyre-Walker et al. 2006; Eyre-Walker and Keightley 2009; Jensen and Bachtrog 2011; Gossmann et al. 2012). This rationale follows the effect of mutation rate along the genome, lying in the theory that adaptation is limited by mutation. In turn, studies at the phenotypic level suggest that adaptation mostly occurs in response to an environmental change. This hypothesis follows Fisher's geometric model of adaptation, which suggests that species with more traits under selection, such as primates, accumulate more mutations with beneficial mutations simply because the adaptive walk is much slower (Orr 2000; Welch and Waxman 2003; Lourenço et al. 2013). The observed controversy in these findings highlights the complex dynamics of rates of adaptive evolution across taxa.

What is new?

To shed light on the determinants of such cross-species variation in molecular adaptive rates, I assessed the effect of intramolecular variation across several animal species. This study showed that in species with higher N_e, the efficiency of purifying selection is much stronger both at buried and exposed residues, leading to generally lower evolutionary rates. Conversely, as the effect of selection is weaker in lower-N_e species, the variation at the intramolecular level becomes stronger, as only mutations with the strongest fitness effects are removed from the population. This leads to a higher accumulation of mutations in exposed residues when compared to buried ones. This project, therefore, supports the N_e hypothesis for the efficiency of negative selection.

Despite the generally weaker signal found for rates of adaptation, these analyses suggested that species with lower N_e tend to accumulate more beneficial mutations. These findings, therefore, contradict the expectations of the N_e hypothesis for rates of adaptive evolution. Instead, they seem to agree with the assumptions at the phenotypic level, where large, long-lived species (which typically have lower N_e) accumulate more adaptive mutations during the adaptive walk. Moreover, this study highlighted the strong impact of the macromolecular structure on rates of adaptive evolution across several taxa, as the distribution of fitness effects varies between buried and exposed residues.

5.4 Final remarks

What are the major determinants of the molecular rate of adaptation?

This thesis revealed the vast variability in rates of molecular adaptation at distinct scales of evolution. Intriguingly, such variation becomes more pronounced as we zoom in across organizational levels. By comparing the magnitude effect in rates of adaptive substitutions between species (Figure 1a), between genes (Figure 1b), and within genes (Figure 1c), one can observe that rates of adaptation vary comparatively more within genomes. This observation goes back to the initial assumptions of the neutral theory: that rates of evolution are relatively constant along the phylogeny while substantially varying among proteins (Kimura 1983; Ohta 1992). What causes such variation within genomes?

In summary, at the intramolecular level, solvent exposure acts as the primary determinant of the rate of adaptive evolution. Intriguingly, such effect seems to act independently of gene age, recombination rate, protein function, and gene expression. When looking at the gene level, however, adaptation seems to follow a more rugged path, where the interplay between gene age and protein function plays a significant role. At the species level, the effect of N_e interacts with the one of solvent exposure, where the distribution of fitness effects between buried and exposed residues varies according to the demography of the population. These findings emphasise the role of different factors in the rate of adaptive evolution across all organizational levels, thus highlighting the importance of a systematic study of adaptation.

By including patterns of intramolecular variation at the scale of systems evolution, this thesis brought the study of adaptation to its most elemental level. We are now one step closer to obtain a holistic comprehension of the molecular basis of adaptation.

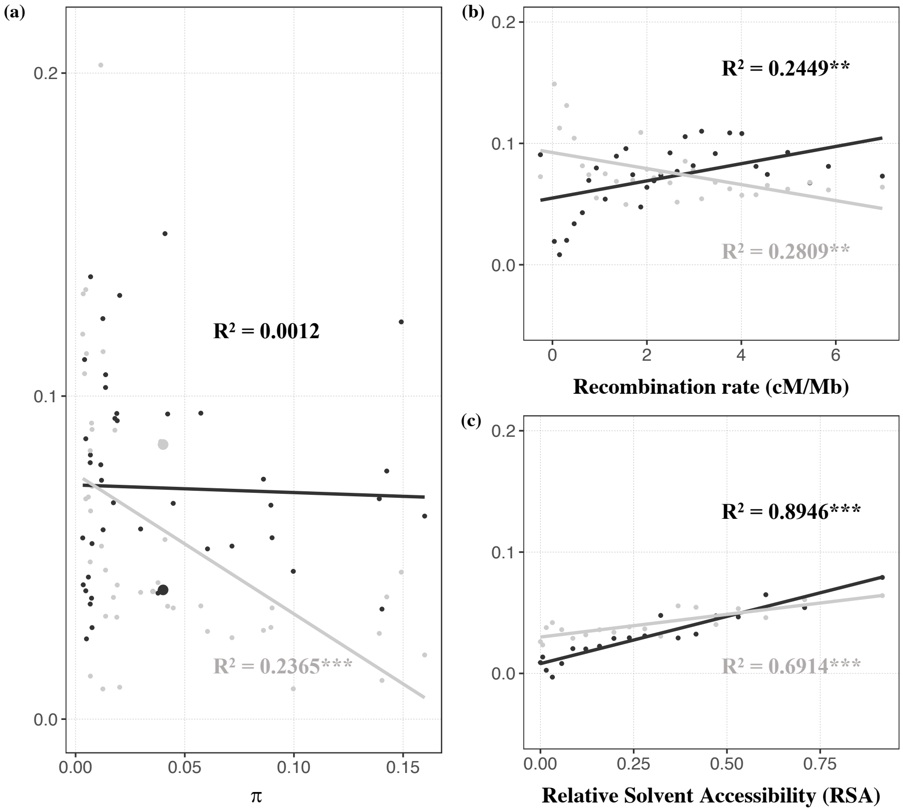

Figure 1. Variation of the rate of adaptive non-synonymous substitutions (ω_a; in black) and the rate of non-adaptive non-synonymous substitutions (ω_{na}; in grey) between species (a), within genomes (b), and within genes (c). The R^2 Pearson's correlation coefficient is given along with significance denoted by asterisks (** p-value < 0.01, *** p-value < 0.001). (a) Relationship between ω_a and ω_{na} with the level of species nucleotide diversity (π), used as a proxy for effective population size, obtained from Galtier (2016). Each sample point represents one species. Dots with bigger sizes correspond to *D. melanogaster* (data from Moutinho et al. 2019b), which is the focus species of plots (b) and (c). (b) Relationship between ω_a and ω_{na} with the recombination rate in cM/Mb, taken from Moutinho et al. (2019b). Each dot represents the mean value of ω_a or ω_{na} for each recombination rate class. (c) Relationship between ω_a and ω_{na} with the relative solvent accessibility (RSA), obtained from Moutinho et al. (2019b). Each dot represents the mean value of ω_a or ω_{na} for each RSA class. This figure and legend were taken from Moutinho et al. (2019a).

Bibliography

Adams J, Mansfield MJ, Richard DJ, Doxey AC. 2017. Lineage-specific mutational clustering in protein structures predicts evolutionary shifts in function. Bioinformatics. 33(9):1338–1345.

Afanasyeva A, Bockwoldt M, Cooney CR, Heiland I, Gossmann TI. 2018. Human long intrinsically disordered protein regions are frequent targets of positive selection. Genome Res. 28(7):975–982.

Akaike H. 1973. Maximum likelihood identification of Gaussian autoregressive moving average models. Biometrika 60(2):255-265.

Akaike H (1987) Factor analysis and AIC. In Selected papers of hirotugu akaike. Springer, New York, NY. (pp. 371-386).

Akashi H. 2003. Translational selection and yeast proteome evolution. Genetics. 164(4):1291–1303.

Albà MM, Castresana J (2005) Inverse relationship between evolutionary rate and age of mammalian genes. Mol Biol Evol 22:598–606.

Albà MM, Castresana J (2007) On homology searches by protein Blast and the characterization of the age of genes. BMC Evol Biol 7:1–8.

Altschul S, Madden T, Schaffer A, et al (1998) Gapped blast and psi-blast: a new generation of protein database search programs. FASEB J 12:3389–3402.

Alves JM, Alves JM, Carneiro M, et al (2019) Parallel adaptation of rabbit populations to myxoma virus. Science, 363(6433): 1319-1326.

Andolfatto P (2007) Hitchhiking effects of recurrent beneficial amino acid substitutions in the. Genome Res 17(12):1755-1762.

Arendsee ZW, Li L, Wurtele ES (2014) Coming of age: Orphan genes in plants. Trends Plant Sci 19:698–708.

Babonis LS, Martindale MQ, Ryan JF (2016) Do novel genes drive morphological novelty? An investigation of the nematosomes in the sea anemone Nematostella vectensis. BMC Evol Biol 16:1–22.

Barroso GV, Moutinho AF, Dutheil JY (2020) A Population Genomics Lexicon. In: Statistical Population Genomics. Methods in Molecular Biology, vol 2090. Humana, New York, NY.

Bartlett GJ, Porter CT, Borkakoti N, Thornton JM (2002) Analysis of catalytic residues in enzyme active sites. J Mol Biol 324:105–121.

Barton N (1998) The geometry of adaptation. Nature 395:751–752.

Barton NH (1995) Linkage and the limits to natural selection. Genetics 140:821–841

Barton NH, Keightley PD (2002) Understanding quantitative genetic variation. Nat Rev Genet 3:11–21.

Bataillon T, Bailey SF (2014) Effects of new mutations on fitness: Insights from models and data. Ann N Y Acad Sci 1320:76–92.

Bateson W (1913) Mendel's Principles of Genetcis. Cambridge Univ. Press

Begun DJ, Aquadro CF (1992) Levels of natrually occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature 356:519–520.

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. J R Stat Soc 57:289–300.

Bergman J, Eyre-Walker A (2019) Does Adaptive Protein Evolution Proceed by Large or Small Steps at the Amino Acid Level? Mol Biol Evol 2:1–9.

Bierne N, Eyre-Walker A (2004) The genomic rate of adaptive amino acid substitution in Drosophila. Mol Biol Evol 21:1350–1360.

Bogumil D, Dagan T (2010) Chaperonin-dependent accelerated substitution rates in prokaryotes. Genome Biol Evol 2:602–608.

Boyko AR, Williamson SH, Indap AR, et al (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet 4(5), e1000083.

Boyle EA, Li YI, Pritchard JK (2017) An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell 169:1177–1186.

Brookfield JF, Sharp PM (1994) Neutralism and selectionism face up to DNA data. Trends Genet 10:109–111.

Bustamante CD, Townsend JP, Hartl DL. 2000. Solvent accessibility and purifying selection within proteins of Escherichia coli and Salmonella enterica. Mol Biol Evol. 17(2):301-308.

Caffrey DR (2004) Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci 13:190–202.

Cai J, Zhao R, Jiang H, Wang W (2008) De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. Genetics 179:487–496.

Cai JJ, Petrov DA (2010) Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. Genome Biol Evol 2:393–409.

Cai JJ, Woo PCY, Lau SKP, et al (2006) Accelerated evolutionary rate may be responsible for the emergence of lineage-specific genes in Ascomycota. J Mol Evol 63:1–11.

Campos JL, Halligan DL, Haddrill PR, Charlesworth B (2014) The relation between recombination rate and patterns of molecular evolution and variation in drosophila melanogaster. Mol Biol Evol 31:1010–1028.

Carneiro M, Albert FW, Melo-Ferreira J, et al (2012a) Evidence for widespread positive and purifying selection across the european rabbit (oryctolagus cuniculus) genome. Mol Biol Evol 29:1837–1849.

Carneiro M, Albert FW, Melo-Ferreira J, et al (2012b) Evidence for widespread positive and purifying selection across the european rabbit (oryctolagus cuniculus) genome. Mol Biol Evol 29:1837–1849.

Carneiro M, Rubin C-J, Palma F Di, et al (2014) Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. Science 345(6200): 1074-1079.

Carvunis AR, Rolland T, Wapinski I, et al (2012) Proto-genes and de novo gene birth. Nature 487:370–374.

Castellano D, Coronado-Zamora M, Campos JL, et al (2016) Adaptive evolution is substantially impeded by hill-Robertson interference in drosophila. Mol Biol Evol 33:442–455.

Castellano D, James J, Eyre-Walker A (2018) Nearly Neutral Evolution across the Drosophila melanogaster Genome. Mol Biol Evol 35:2685–2694.

Castellano D, Macià MC, Tataru P, et al (2019) Comparison of the full distribution of fitness effects of new amino acid mutations across great apes. Genetics 213(3): 953-966.

Castillo-Davis CI, Mekhedov SL, Hartl DL, et al (2002) Selection for short introns in highly expressed genes. Nat Genet 31:415–418.

Chamberlain SA, Szöcs E (2013) taxize: taxonomic search and retrieval in R. F1000Research 2:191.

Charlesworth B (1994) The effect of background selection against deleterious mutations on weakly selected, linked variants. Genet Res 63:213–227.

Charlesworth B (2012) The effects of deleterious mutations on evolution at linked sites. Genetics 190:5–22.

Charlesworth B (2009) Effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10:195–205.

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289–303.

Charlesworth J, Eyre-Walker A (2006) The rate of adaptive evolution in enteric bacteria. Mol Biol Evol 23:1348–1356.

Chatr-Aryamontri A, Oughtred R, Boucher L, et al (2017) The BioGRID interaction database: 2017 update. Nucleic Acids Res 45:D369–D379.

Chen J, Glémin S, Lascoux M (2020) From drift to draft: How much do beneficial mutations actually contribute to predictions of Ohta's slightly deleterious model of molecular evolution? Genetics 1–39.

Chen S, Zhang YE, Long M (2010) New genes in Drosophila quickly become essential. Science 330:1682–1685.

Chi PB, Liberles DA (2016) Selection on protein structure, interaction, and sequence. Protein Sci 25:1168–1178.

Choi SC, Hobolth A, Robinson DM, et al (2007) Quantifying the impact of protein tertiary structure on molecular evolution. Mol Biol Evol 24:1769–1782.

Choi SS, Vallender EJ, Lahn BT (2006) Systematically assessing the influence of 3-dimensional structural context on the molecular evolution of mammalian proteomes. Mol Biol Evol 23:2131–2133.

Coghlan A, Wolfe KH (2000) Relationship of codon bias to mRNA and concentration protein length in Saccharomyces cerevisiae. Yeast 16:1131–1145.

Collins CA, Brown EJ (2010) Cytosol as battleground: Ubiquitin as a weapon for both host and pathogen. Trends Cell Biol 20:205–213.

Colosimo PF, Peichel CL, Nereng K, et al (2004) The genetic architecture of parallel armor plate reduction in threespine sticklebacks. PLoS Biol 2:635–641.

Conant GC, Stadler PF (2009) Solvent exposure imparts similar selective pressures across a range of yeast proteins. Mol Biol Evol 26:1155–1161.

Cooper VS, Lenski RE (2000) The population genetics of ecological specialization in evolving Escherichia coli populations. Nature 407:736–739.

Crick F, Watson J (1953) Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid.

Nature 171:737–738.

Crow JF (2003) Was there life before 1953? Nat Genet 33:449–450.

Crowson D, Barrett SCH, Wright SI (2017) Purifying and Positive Selection Influence Patterns of Gene Loss and Gene Expression in the Evolution of a Plant Sex Chromosome System. Mol Biol Evol 34:1140–1154.

Cui X, Lv Y, Chen M, et al (2015) Young genes out of the male: An insight from evolutionary age analysis of the pollen transcriptome. Mol Plant 8:935–945.

Cutter AD, Jovelin R, Dey A (2013) Molecular hyperdiversity and evolution in very large populations. Mol Ecol 22:2074–2095.

Dangl JL, Jones JD (2001) Plant pathogens and integrated defence responses to infection. Nature 411:826–833.

Darwin C (1859) The Origin of Species. J. Murray, London.

Daubin V, Ochman H (2004) Bacterial genomes as new gene homes: The genealogy of ORFans in E. coli. Genome Res 14:1036–1042.

De Castro E, Sigrist CJA, Gattiker A, et al (2006) ScanProsite: Detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res. 34(suppl_2): W362-W365.

Dean AM, Neuhauser C, Grenier E, Golding GB (2002) The pattern of amino acid replacements in α/β-barrels. Mol Biol Evol 19:1846–1864.

Dickerson RE (1971) The structure of cytochrome c and the rates of molecular evolution. J Mol Evol 1:26–45.

Dielen AS, Badaoui S, Candresse T, German-Retana S (2010) The ubiquitin/26S proteasome system in plant-pathogen interactions: A never-ending hide-and-seek game. Mol Plant Pathol 11:293–308.

Ding Y, Zhou Q, Wang W (2012) Origins of New Genes and Evolution of Their Novel Functions. Annu Rev Ecol Evol Syst 43:345–363.

Dobhansky T (1955) A review of some fundamental concepts and problems of population genetics. Cold Spring Harb Symp Quant Biol 20:1–15.

Domazet-Lošo T, Brajković J, Tautz D (2007) A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet 23:533–539.

Domazet-Lošo T, Carvunis AR, Albà MM, et al (2017) No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. Mol Biol Evol 34:843–856.

Domazet-Loso T, Tautz D (2003) An evolutionary analysis of orphan genes in Drosophila. Genome Res 13:2213–2219.

Drummond DA, Bloom JD, Adami C, et al (2005) Why highly expressed proteins evolve slowly. Proc Natl Acad Sci 102:14338–14343.

Dunker AK, Brown CJ, Lawson JD, et al (2002) Intrinsic disorder and protein function. Biochemistry 41:6573–6582.

Duret L, Mouchiroud D (2000) Determinants of Substitution Rates in Mammalian Genes: Expression Pattern Affects Selection Intensity but Not Mutation Rate. Mol Biol Evol 17:68–74.

Durinck S, Moreau Y, Kasprzyk A, et al (2005) BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. Bioinformatics 21:3439–3440.

Dutheil JY, Gaillard S, Stukenbrock EH. 2014. MafFilter: A highly flexible and extensible multiple genome alignment files processor. BMC Genomics. 15(1):53.

Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 6:197–208.

Eanes WF, Kirchner M, Yoon J (1993) Evidence for adaptive evolution of the G6pd gene in the Drosophila melanogaster and Drosophila simulans lineages. Proc Natl Acad Sci U S A 90:7475–7479.

Elhaik E, Sabath N, Graur D (2006) The "inverse relationship between evolutionary rate and age of mammalian genes" is an artifact of increased genetic distance with rate of evolution and time of divergence. Mol Biol Evol 23:1–3.

Enard D, Cai L, Gwennap C, Petrov DA (2016) Viruses are a dominant driver of protein adaptation in mammals. Elife 5:1–25.

Eyre-Walker A (2002) Changing effective population size and the McDonald-Kreitman test. Genetics 162:2017–2024

Eyre-Walker A (2006) The genomic rate of adaptive evolution. Trends Ecol Evol 21:569–75.

Eyre-Walker A, Keightley PD (2009) Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol 26:2097–2108.

Eyre-Walker A, Woolfit M, Phelps T (2006) The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics 173:891–900.

Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. Genetics 158:1227–34.

Felsenstein J (1985) Phylogenies and the Comparative Method. Am Nat 125:1–15.

Fisher R (1930a) The Genetical Theory of Natural Selection. Oxford Univ. Press, Oxford.

Fisher R (1930b) The distribution of gene ratios for rare mutations. Proc R Soc Edinburgh 205–220.

Fisher RA (1918) Reproduced by permission of the Royal Society of Edinburgh from Transactions of the Society, vol. 52: 399-433 (1918). Society 52:399–433.

Fisher RA (1922) On the Dominance Ratio. Proc R Soc Edinburgh 42:321–341.

Force A, Lynch M, Pickett FB, et al (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545.

Ford EB (1975) Ecological genetics. Chapman and Hall, London.

Ford EB (1964) Ecological genetics. John Wiley, New York.

Franzosa EA, Xia Y (2009) Structural determinants of protein evolution are context-sensitive at the residue level. Mol Biol Evol 26:2387–2395.

Freese E (1962) On the evolution of the base composition of DNA. J Theor Biol 3:82–101.

Galtier N. 2016. Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. PLoS

Genet. 12(1): e1005774.

Galton F (1894) Natural Inheritance. Macmillan and Company.

García-Vallvé S, Alonso Á, Bravo IG (2005) Papillomaviruses: Different genes have different histories. Trends Microbiol 13:514–521.

Geldner N, Robatzek S (2008) Plant receptors go endosomal: A moving view on signal transduction. Plant Physiol 147:1565–1574.

Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32:258D – 261.

Gerrish P (2001) The rhythm of microbial adaptation. Nature 413:299–302.

Gillespie JH (1991) The Causes of Molecular Evolution. Oxford University Press.

Gillespie JH (1984) Molecular evolution over the mutational landscape. Evolution (NY) 38:1116–1129.

Gillespie JH (1983) A Simple Stochastic Gene Substitution Model. Theor Popul Biol 23:202–215.

Gillespie JH (2000a) The neutral theory in an infinite population. Gene 261:11–18.

Gillespie JH (2001) Is the Population Size of a Species Relevant To Its Evolution? Evolution (NY) 55:2161.

Gillespie JH (1999) The role of population size in molecular evolution. Theor Popul Biol 55:145–156.

Gillespie JH (1986) Natural selection and the molecular clock. Mol Biol Evol 3:138–155.

Gillespie JH (1989) Lineage effects and the index of dispersion of molecular evolution. Mol Biol Evol 6:636–647.

Gillespie JH (2000b) Genetic Drift in Infinite Populations: The Pseudohitchhiking Model. Genet Soc Am 155:909–919.

Goldman N, Thorne JL, Jones DT (1998) Assessing the Impact of Secondary Structure and Solvent Accessibility on Protein Evolution. Genetics 149:445–458.

Gossmann TI, Keightley PD, Eyre-Walker A (2012) The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. Genome Biol Evol 4:658–667.

Gossmann TI, Schmid MW, Grossniklaus U, Schmid KJ (2014) Selection-driven evolution of sex-biased genes is consistent with sexual selection in arabidopsis thaliana. Mol Biol Evol 31:574–583.

Gossmann TI, Song BH, Windsor AJ, et al (2010) Genome wide analyses reveal little evidence for adaptive evolution in many plant species. Mol Biol Evol 27:1822–1832.

Grafen A (1989) The Phylogenetic Regression. Philos Trans R Soc B Biol Sci 326:119–157.

Grant BJ, Rodrigues APC, ElSawy KM, et al (2006) Bio3d: An R package for the comparative analysis of protein structures. Bioinformatics 22:2695–2696.

Grantham R (1974) Amino Acid Difference Formula to Help Explain Protein Evolution. Science 185:862–4.

Graur D, Zheng Y, Price N, et al (2013) On the immortality of television sets: “Function” in the human genome according to the evolution-free gospel of encode. Genome Biol Evol 5:578–590.

Gremme G, Steinbiss S, Kurtz S (2013) Genome tools: A comprehensive software library for efficient processing of structured genome annotations. IEEE/ACM Trans Comput Biol Bioinforma 10:645–656.

Groen AJ, De Vries SC, Lilley KS (2008) A proteomics approach to membrane trafficking. Plant Physiol 147:1584–1589.

Guéguen L, Gaillard S, Boussau B, et al (2013) Bio++: Efficient extensible libraries and tools for computational molecular evolution. Mol Biol Evol 30:1745–1750.

Guo HH, Choe J, Loeb LA (2004) Protein tolerance to random amino acid change. Proc Natl Acad Sci 101:9205–9210.

Haddrill PR, Loewe L, Charlesworth B (2010) Estimating the parameters of selection on nonsynonymous mutations in Drosophila pseudoobscura and D. miranda. Genetics 185:1381–1396.

Haerty W, Jagadeeshan S, Kulathinal RJ, et al (2007) Evolution in the fast lane: Rapidly evolving sex-related genes in Drosophila. Genetics 177:1321–1335.

Hagedoorn AL, Hagedoorn AC (1921) The Relative Value of the Processes causing Evolution. The Hague, Martinus Nijhoff.

Hahn MW (2008) Toward a Selection Theory of Molecular Evolution. Evolution (NY) 62:255–265.

Haldane J (1939) The Equilibrium Between Mutation and Random Extinction. Ann Eugen 9:400–405

Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD. 2010. Evidence for pervasive adaptive protein evolution in wild mice. PLoS Genet. 6(1): e1000825.

Hardy AGH (1908) Mendelian Proportions in a Mixed Population. Science. 28:49–50.

Harris H (1966) Enzyme polymorphisms in man. Genetics 54:298–310.

Hastie TJ, Pregibon D (1992) "Chapter 6." Statistical models in S. Generalized linear models. Wadsworth and Brooks/Cole.

Heinen TJAJ, Staubach F, Häming D, Tautz D (2009) Emergence of a New Gene from an Intergenic Region. Curr Biol 19:1527–1531.

Hill WG, Robertson A (1966) The effect of linkage on limits to artificial selection. Genet Res. 8:269–294.

Hiroshi A, Minsoo K, Chihiro S (2014) Exploitation of the host ubiquitin system by human bacterial pathogens. Nat Rev Microbiol 12:399–413.

Hoffman PF, Kaufman AJ, Halverson GP, Schrag DP (1998) A neoproterozoic snowball earth. Science. 281:1342–1346.

Huber CD, Kim BY, Marsden CD, Lohmueller KE (2017) Determining the factors driving selective effects of new nonsynonymous mutations. Proc Natl Acad Sci. 114:4465–4470.

Hudson RR, Kreitman M, Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116:153–159.

Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. Proc R Soc B 256:119–124.

Hunter T (1995) Protein Kinases and Phosphatases: The Yin and Yang of Protein Phosphorylation and Signaling. Cell 80:225–236.

Huxley J (1942) Evolution. The Modern Synthesis. London: George Alien & Unwin Ltd.

Hvilsom C, Qian Y, Bataillon T, et al (2012) Extensive X-linked adaptive evolution in central chimpanzees. Proc Natl Acad Sci 109:2054–2059.

Imhof M, Schlötterer C (2001) Fitness effects of advantageous mutations in evolving Escherichia coli populations. Proc Natl Acad Sci. 98:1113–1117.

Ingvarsson PK (2010) Natural Selection on Synonymous and Nonsynonymous Mutations Shapes Patterns of Polymorphism in Populus tremula. Mol Biol Evol 27:650–660.

Jacob F (1977) Evolution and Tinkering. Science. 196:1161–1166.

Janin J (1979) Surface and inside volumes in globular proteins. Nature 277(5696):491.

Jensen JD, Bachtrog D (2011) Characterizing the influence of effective population size on the rate of adaptation: Gillespie's Darwin domain. Genome Biol Evol 3:687–701.

Jensen JD, Payseur BA, Stephan W, et al (2019) The importance of the Neutral Theory in 1968 and 50 years on: A response to Kern and Hahn 2018. Evolution. 73:111–114.

Johannsen W (1911) The Genotype Conception of Heredity. Am Nat 45:129–159.

Julenius K, Pedersen AG (2006) Protein evolution is faster outside the cell. Mol Biol Evol 23:2039–2048.

Kadibalban AS, Bogumil D, Landan G, Dagan T (2016) DnaK-Dependent Accelerated Evolutionary Rate in Prokaryotes. Genome Biol Evol 8:1590–1599.

Kanehisa M, Goto S, Kawashima S, Nakaya A. 2002. The KEGG databases at GenomeNet. Nucleic Acids Res. 30(1):42–46.

Karasov T, Messer PW, Petrov DA (2010) Evidence that adaptation in Drosophila is not limited by mutation at single sites. PLoS Genet 6:1–10.

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol 30:772–780.

Kauffman S, Levin S (1987) Towards a General Theory of Adaptive Walks on Rugged Landscapes. J theor Biol 128:11–45.

Keightley PD, Eyre-Walker A (2007) Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177:2251–2261.

Kern AD, Hahn MW (2018) The neutral theory in light of natural selection. Mol Biol Evol 35:1366–1371.

Khalturin K, Hemmrich G, Fraune S, et al (2009) More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet 25:404–413.

Kimura M (1969) The rate of molecular evolution considered from the standpoint of population genetics. Proc Natl Acad Sci. 63:1181–1188.

Kimura M (1955) Random Genetic Drift in Multi-Allelic Locus. Evolution. 9:419–435

Kimura M (1983) The Neutral Theory of Molecular Evolution. Cambridge Univeristy Press.

Kimura M (1968) Evolutionary Rate at the Molecular Level. Nature 217:624–626.

Kimura M, Ohta T (1971) Protein Polymorphism as a Phase of Molecular Evolution. Nature 229:235–237

King JL, Jukes TH (1969) Non-Darwinian evolution. Science. 164:788–798.

Klausen MS, Jespersen MC, Nielsen H, et al (2019) NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. Proteins Struct Funct Bioinforma 87:520–527.

Kosakovsky Pond SL, Murrell B, Fourment M, et al (2011) A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol 28:3033–3043.

Kosiol C, Vinař T, Da Fonseca RR, et al (2008) Patterns of positive selection in six mammalian genomes. PLoS Genet 4(8): e1000144.

Krylov DM, Wolf YI, Rogozin IB, Koonin E V. (2003) Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res 13:2229–2235.

Kuo CH, Kissinger JC (2008) Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites Plasmodium and Theileria. BMC Evol Biol 8:1–16.

Laehnemann D, Peña-Miller R, Rosenstiel P, et al (2014) Genomics of rapid adaptation to antibiotics: Convergent evolution and scalable sequence amplification. Genome Biol Evol 6:1287–1301.

Lanfear R, Kokko H, Eyre-Walker A (2014) Population size and the rate of evolution. Trends Ecol Evol 29:33–41.

Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL (2005) Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Mol Biol Evol 22:1345–1354.

Lenski RE, Rose MR, Simpson SC, Tadler SC (1991) Long-Term Experimental Evolution in Escherichia coli. I. Adaptation and Divergence During 2,000 Generations. Am Nat 138:1315–1341.

Lewontin RC, Hubby JL (1966) A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. Genetics 54:595–609.

Li W-H, Gojobori T, Nei M (1981) Pseudogenes as a paradigm of neutral evolution. Nature 292:237–239.

Li W-H, Wu C-I, Luo C-C (1985) A New Method for Extimating Synonymous and Nonsynonymous Rates of Nucleotide Substitution Considering the Relative Likelihood of Nucleotide and Codon Usage. Mol Biol Evol 2:150–174.

Liao BY, Scott NM, Zhang J (2006) Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. Mol Biol Evol 23:2072–2080.

Liberles D a., Teichmann S a., Bahar I, et al (2012) The interface of protein structure, protein biophysics, and molecular evolution. Protein Sci 21:769–785.

Lin YS, Hsu WL, Hwang JK, Li WH (2007) Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. Mol Biol Evol 24:1005–1011.

Linding R, Jensen LJ, Diella F, et al (2003) Protein disorder prediction: Implications for structural proteomics. Structure. 11(11):1453–1459.

Lipman DJ, Souvorov A, Koonin E V., et al (2002) The relationship of protein conservation and sequence length. BMC Evol Biol 2:1–10.

Liu X, Li YI, Pritchard JK (2019) Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. Cell 177(4), 1022-1034.

Loureiro J, Ploegh HL (2006) Antigen Presentation and the Ubiquitin-Proteasome System in Host–Pathogen

Interactions. Adv Immunol 92:225–305.

Lourenço JM, Glémin S, Galtier N (2013) The rate of molecular adaptation in a changing environment. Mol Biol Evol 30:1292–1301.

Lynch M (2002) Genomics: Gene duplication and evolution. Science. 297:945–947.

Malécot G (1944) Sur un problème de probabilitiés en chaîne que pose la génétique. C R Acad Sci Paris 219:379–381.

Marais G, Charlesworth B (2003) Genome evolution: Recombination speeds up adaptive evolution. Curr Biol 13:68–70.

Mauch-Mani B, Baccelli I, Luna E, Flors V (2017) Defense Priming: An Adaptive Part of Induced Resistance. Annu Rev Plant Biol 68:485–512.

Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res Cambridge 23:23–25.

Mayr E (1965) Discussion. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins. Ev Academic Press, New York., pp 293–294.

McDonald JH, Kreitman M (1991) Adaptive protein evolution ate the Adh locus in Drosophile. 20; 351(6328):652–4.

Miller S, Lesk AM, Janin J, Chothia C (1987) The accessible surface area and stability of oligomeric proteins. Nature 329:855–857.

Mintseris J, Weng Z (2005) Structure, function, and evolution of transient and obligate protein-protein interactions. Proc Natl Acad Sci 102:10930–10935.

Mirny L a, Shakhnovich EI (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol 291:177–196.

Miyata T, Miyazawa S, Yasunaga T (1979) Two Types of Amino Acid Substitutions in Protein Evolution. J Mol Evol 12:219–236.

Miyata T, Yasunaga T (1981) Rapidly evolving mouse α-globin-related pseudo gene and its evolutionary history. Proc Natl Acad Sci U S A 78:450–453.

Morgan TH (1903) Evolution and Adaptation. Macmillan, New York.

Morgan TH (1925) Evolution and genetics. Princeton University Press, Princeton, N.J.

Morgan TH (1932) The scientific basis of evolution. WW Norton, New York.

Mosmann VR, Cherwinski H, Bond MW, et al (2016) Two types of murine helper T cell clone. I. Definition according to profiles of lymphokine activities and secreted proteins. J Immunol. 136(7): 2348-2357.

Motion GB, Amaro TMMM, Kulagina N, Huitema E (2015) Nuclear processes associated with plant immunity and pathogen susceptibility. Brief Funct Genomics 14:243–252.

Moutinho A., Bataillon T, Dutheil JY (2019a) Variation of the adaptive substitution rate between species and within genomes. Evol Ecol 1–40.

Moutinho AF, Trancoso FF, Dutheil JY (2019b) The impact of protein architecture on adaptive evolution. Mol Biol Evol. 36(9), 2013-2028.

Moyers BA, Zhang J (2015) Phylostratigraphic bias creates spurious patterns of genome evolution. Mol Biol

Evol 32:258–267.

Moyers BA, Zhang J (2016) Evaluating Phylostratigraphic Evidence for Widespread de Novo Gene Birth in Genome Evolution. Mol Biol Evol 33:1245–1256.

Nei M (1987) Molecular Evolutionary Genetics. Columbia University Press.

Neme R, Tautz D (2013) Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. BMC Genomics 14: 14(1): 117.

Neme R, Tautz D (2014) Evolution: Dynamics of de novo gene emergence. Curr Biol 24:R238–R240.

Nielsen R, Bustamante C, Clark AG, et al (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3:0976–0985.

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936.

Obbard DJ, Welch JJ, Kim KW, Jiggins FM (2009) Quantifying adaptive evolution in the Drosophila immune system. PLoS Genet 5(10): e1000698.

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

Ohta T (1976) Role of very slightly deleterious mutations in molecular evolution and polymorphism. Theor Popul Biol 10:254–275.

Ohta T (1973) Slightly deleterious mutant substitutions in evolution. Nature 246:96–98.

Ohta T (1992) The nearly neutral theory of molecular evolution. Annu Rev Ecol Syst 23:263–286.

Ohta T (1972) Population size and rate of evolution. J Mol Evol 1:305–314.

Orr AH (1999) The evolutionary genetics of adaptation: A simulation study. Genet Res 74:207–214.

Orr HA (2002) the Population Genetics of Adaptation: the Adaptation of Dna Sequences. Evolution. 56(7): 1317-1330.

Orr HA (2005) The genetic theory of adaptation: a brief history. Nat Rev Genet 6:119–127.

Orr HA (1998) The Population Genetics of Adaptation: The Distribution of Factors Fixed during Adaptive Evolution. 52(4): 935-949.

Orr HA (2003) The distribution of fitness effects among beneficial mutations. Genetics 163:1519–1526

Orr HA (2000) Adaptation and the cost of complexity. Evolution. 54:13–20.

Overington J, Donnelly D, Johnson MS, et al (1992) Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds. Protein Sci 1:216–226.

Pal C, Papp B, Hurst LD (2001) Highly Expressed Genes in Yeast Evolve Slowly. Genetics 158:927–931.

Palmieri N, Kosiol C, Schlötterer C (2014) The life cycle of Drosophila orphan genes. Elife 3:1–21.

Paradis E, Claude J, Strimmer K (2004) APE: Analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.

Pearson K (1898) Mathematical Contributions to the Theory ot Evolution. Proc R Soc L 62:386–412.

Perutz MF, Kendrew JC, Watson HC (1965) Structure and function of haemoglobin: II. Some relations between polypeptide chain configuration and amino acid sequence. J Mol Biol 13:669–678.

Petryszak R, Keays M, Tang YA, et al (2016) Expression Atlas update - An integrated database of gene and protein expression in humans, animals and plants. Nucleic Acids Res. 44(D1), D746-D752.

Piatigorsky J, Wistow G (1991) The recruitment of crystallins: new functions precede gene duplication. Science. 252:1078–1079.

Pool JE, Corbett-Detig RB, Sugino RP, et al (2012) Population Genomics of Sub-Saharan Drosophila melanogaster: African Diversity and Non-African Admixture. PLoS Genet. 8(12): e1003080.

Prince VE, Pickett FB (2002) Splitting pairs: The diverging fates of duplicated genes. Nat Rev Genet 3:827–837.

Pröschel M, Zhang Z, Parsch J (2006) Widespread adaptive evolution of Drosophila genes with sex-biased expression. Genetics 174:893–900.

Proux E, Studer RA, Moretti S, Robinson-Rechavi M (2009) Selectome: A database of positive selection. Nucleic Acids Res 37:404–407.

Punnet RC (1915) Mimicry in butterflies. Cambridge Univ. Press.

R Core Team (2017) A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ramsey DC, Scherrer MP, Zhou T, Wilke CO (2011) The relationship between relative solvent accessibility and evolutionary rate in protein evolution. Genetics 188:479–488.

Remmert M, Biegert A, Hauser A, Söding J (2012) HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 9:173–175.

Rezvoy C, Charif D, Guéguen L, Marais GAB (2007) MareyMap: An R-based tool with graphical interface for estimating recombination rates. Bioinformatics 23(16):2188-2189.

Rocha EPC, Danchin A (2004) An Analysis of Determinants of Amino Acids Substitution Rates in Bacterial Proteins. Mol Biol Evol 21:108–116.

Rodrigue N, Lartillot N (2017) Detecting adaptation in protein-coding genes using a Bayesian site-heterogeneous mutation-selection codon substitution model. Mol Biol Evol 34:204–214.

Rousselle M, Mollion M, Nabholz B, et al (2018) Overestimation of the adaptive substitution rate in fluctuating populations. Biol Lett 14: 20180055.

Rousselle M, Simion P, Tilak M-K, et al (2019) Is adaptation limited by mutation? A timescale dependent effect of genetic diversity on the adaptive substitution rate in animals. bioRxiv 643619. doi: 10.1101/643619

Rozen DE, De Visser JAGM, Gerrish PJ (2002) Fitness effects of fixed beneficial mutations in microbial populations. Curr Biol 12:1040–1045.

Sackton TB, Lazzaro BP, Schlenke TA, et al (2007) Dynamic evolution of the innate immune system in Drosophila. Nat Genet 39:1461–1468.

Salvador-Martínez I, Coronado-Zamora M, Castellano D, et al (2018) Mapping Selection within Drosophila melanogaster Embryo's Anatomy. Mol Biol Evol 35:66–79.

Sawyer SA, Kulathinal RJ, Bustamante CD, Hartl DL (2003) Bayesian Analysis Suggests that Most Amino Acid Replacements in Drosophila Are Driven by Positive Selection. J Mol Evol 57:154–164.

Schaffer AA (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 29:2994–3005.

Schneider A, Charlesworth B, Eyre-Walker A, Keightley PD (2011) A method for inferring the rate of occurrence and fitness effects of advantageous mutations. Genetics 189:1427–1437.

Schoustra SE, Bataillon T, Gifford DR, Kassen R (2009) The properties of adaptive walks in evolving populations of fungus. PLoS Biol 7(11): e1000250.

Schuler MA, Werck-Reichhart D (2003) Functional Genomics of P450S. Annu Rev Plant Biol 54:629–667.

Schulte RD, Makus C, Hasert B, et al (2010) Multiple reciprocal adaptations and rapid genetic change upon experimental coevolution of an animal host and its microbial parasite. Proc Natl Acad Sci. 107:7359–7364.

Sella G, Petrov DA, Przeworski M, Andolfatto P (2009) Pervasive natural selection in the Drosophila genome? PLoS Genet 5: e1000495.

Shapiro MD, Marks ME, Peichel CL, et al (2006) Erratum: Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks (Nature (2004) 428 (717-723)). Nature 439:1014.

Shaw CR (1965) Electrophoretic variation in enzymes. Science. 149:936–943.

Slotte T, Bataillon T, Hansen TT, et al (2011) Genomic determinants of protein evolution and polymorphism in arabidopsis. Genome Biol Evol 3:1210–1219.

Slotte T, Foxe JP, Hazzouri KM, Wright SI (2010) Genome-wide evidence for efficient positive and purifying selection in capsella grandiflora, a plant species with a large effective population size. Mol Biol Evol 27:1813–1821.

Smith JM (1962) An Anthology of Partly-Baked Ideas. In: Good IJ (ed) The Scientist Speculates. Basic Books, New York, pp 252–256.

Smith JM (1970a) Natural selection and the concept of a protein space. Nature 225:563–564.

Smith JM (1970b) Natural selection and the concept of a protein space. Nature 225:563–564.

Smith NGC, Eyre-Walker a (2002) Adaptive protein evolution in Drosophila . Nature 415:1022–1024.

Steinberg B, Ostermeier M (2016) Environmental changes bridge evolutionary valleys. Sci Adv 2: e1500921.

Stern C (1943) The Hardy-Weinberg law. Science. 97:137–138

Stoletzki N, Eyre-Walker A (2011) Estimation of the neutrality index. Mol Biol Evol 28:63–70.

Strasburg JL, Kane NC, Raduski AR, et al (2011) Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. Mol Biol Evol 28:1569–1580.

Stukenbrock EH, Bataillon T, Dutheil JY, et al (2011) The making of a new pathogen: Insights from comparative population genomics of the domesticated wheat pathogen Mycosphaerella graminicola and its wild sister species. Genome Res 21:2157–2166.

Subramanian S, Kumar S (2004) Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics 168:373–381.

Sucena É, Stern DL (2000) Divergence of larval morphology between Drosophila sechellia and its sibling species caused by cis-regulatory evolution of ovo/shaven-baby. Proc Natl Acad Sci. 97:4530–4534.

Sueoka N (1962) On the Genetic Basis of Variation and Heterogeneity of DNA Base Composition. Proc Natl Acad Sci 58:582–592.

Tataru P, Bataillon T (2019) polyDFEv2.0: testing for invariance of the distribution of fitness effects within and across species. Bioinformatics 1–2.

Tataru P, Mollion M, Glémin S, Bataillon T (2017) Inference of Distribution of Fitness Effects and. Genetics 207:1103–1119.

Tautz D, Domazet-Lošo T (2011) The evolutionary origin of orphan genes. Nat Rev Genet 12:692–702.

Thornton K, Long M (2002) Rapid divergence of gene duplicates on the Drosophila melanogaster X chromosome. Mol Biol Evol 19:918–925.

Tien MZ, Meyer AG, Sydykova DK, et al (2013) Maximum allowed solvent accessibilites of residues in proteins. PLoS One PLoS One. 8(11): e80635.

Tomancak P, Berman BP, Beaton A, et al (2007) Global analysis of patterns of gene expression during Drosophila embryogenesis. Genome Biol Genome Biol. 8(7):R145.

Trujillo M, Shirasu K (2010) Ubiquitination in plant immunity. Curr Opin Plant Biol 13:402–408.

Valdar WSJ, Thornton JM (2001) Protein-protein interfaces: Analysis of amino acid conservation in homodimers. Proteins Struct Funct Genet 42:108–124.

Van Durme J, Maurer-Stroh S, Gallardo R, et al (2009) Accurate prediction of DnaK-peptide binding via homology modelling and experimental data. PLoS Comput Biol PLoS Comput Biol. 5(8): e1000475.

Venables WN, Ripley BD (2002) Modern applied statistics with S. Springer.

Vicoso B, Charlesworth B (2006) Evolution on the X chromosome: Unusual patterns and processes. Nat Rev Genet 7:645–653.

Vicoso B, Charlesworth B (2009) Effective population size and the faster-X effect: An extended model. Evolution. 63:2413–2426.

Vishnoi A, Kryazhimskiy S, Bazykin GA, et al (2010) Young proteins experience more variable selection pressures than old proteins. Genome Res 20:1574–1581.

Wang H, Moore MJ, Soltis PS, et al (2009) Rosid radiation and the rapid rise of angiosperm-dominated forests. Proc Natl Acad Sci U S A 106:3853–3858.

Wang W, Zheng H, Yang S, et al (2005) Origin and evolution of new exons in rodents. Genome Res 15:1258–1264.

Weigel D, Mott R (2009) The 1001 Genomes Project for Arabidopsis thaliana. Genome Biol. 8(12): e1003080.

Weinberg W (1908) Uber den Nachweis der Vererbung beim Menshen. Jahresh Ver Vaterl Naturkd Wuerttemb 64:368–382.

Weissman DB, Barton NH (2012) Limits to the rate of adaptive substitution in sexual populations. PLoS Genet 8(6): e1002740.

Welch JJ (2006a) Estimating the genomewide rate of adaptive protein evolution in drosophila. Genetics 173:821–837.

Welch JJ (2006b) Estimating the genomewide rate of adaptive protein evolution in drosophila. Genetics 173:821–837.

Welch JJ, Waxman D (2003) Modularity and the cost of complexity. Evolution. 57:1723–1734.

Weldon WFR (1895) Attempt to measure the death-rate due to the selective destruction of Carcinus moenas with respect to a particular dimension. Proc R Soc L 58:360–379.

Wickham H (2017) ggplot2 - Elegant Graphics for Data Analysis (2nd Edition). J Stat Softw 77:2–5.

Wilke CO, Bloom JD, Drummond DA, Raval A (2005) Predicting the tolerance of proteins to random amino acid substitution. Biophys J 89:3714–20.

Wilson BA, Foy SG, Neme R, Masel J (2017) Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. Nat Ecol Evol 1: 0146.

Wolf YI, Novichkov PS, Karev GP, et al (2009) The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. Proc Natl Acad Sci 106:7273–7280.

Wright S (1951) Fisher and Ford on "The Sewall Wright effect." Am Sci 39:452–458.

Wright S (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. Sixth Int. Congr. Genet. 1:356–366.

Wright S (1931) Evolution in Mendelian Populations. Genetics 16:97–159.

Wright S (1949) Adaptation and Selection., Genetics,. Princeton University Press.

Wright SI, Yau CBK, Looseley M, Meyers BC (2004) Effects of gene expression on molecular evolution in Arabidopsis thaliana and Arabidopsis lyrata. Mol Biol Evol 21:1719–1726.

Yang Z, Bielawski JR (2000) Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15:496–503.

Yang Z, Nielsen R (2002) Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages. Mol Biol Evol 19:908–917.

Yang Z, Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol 46:409–418.

Yang Z, Nielsen R, Goldman N, Pedersen AK (2000) Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites.

Yang Z, Wong WSW, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol 22:1107–1118.

Zeileis A, Hothorn T (2002) Diagnostic Checking in Regression Relationshipslmtest citation info. R News 2:7–10.

Zhang J (2000) Protein-length distributions for the three domains of life. Trends Genet 16:107–109.

Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22:2472–2479.

Zhang J, Zhang Y ping, Rosenberg HF (2002) Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nat Genet 30:411–415.

Zhang YE, Vibranovski MD, Krinsky BH, Long M (2010) Age-dependent chromosomal distribution of male-biased genes in Drosophila. Genome Res 20:1526–1533.

Zhao L, Saelao P, Jones CD, Begun DJ (2014) Origin and spread of de novo genes in Drosophila melanogaster populations. Science. 343:769–772.

Zhen Y, Huber CD, Davies RW, Lohmueller KE (2018) Stronger and higher proportion of beneficial amino acid changing mutations in humans compared to mice and flies. bioRxiv.

Zuckerkandl E, Pauling L (1962) Molecular disease, evolution, and genetic heterogeneity. In: Kasha M, Pullman B (eds) Horizons in biochemistry. Academic Press, New York., pp 189–225.

Acknowledgments

First, I would like to thank Julien for the incredible guidance, discussions, and all the support throughout these years. I learned a lot since I came here and I also have to thank you for that. It was amazing working with you. Thank you very much for everything!

I would also like to thank my Thesis Advisory Committee, Tal Dagan and Chaitanya Gokhale, for all the fruitful discussions and all the suggestions for this project. And once more to Tal for agreeing to be the second examiner of my thesis.

I would also like to thank Adam Eyre-Walker, Guy Sella, Luis-Miguel Chevin, Nicolas Galtier, and Hinrich Schulenburg for all the fruitful discussions that gave some guidance to this thesis. Also, again to Hinrich for agreeing to be an examiner of this project, and Adam for accepting our invitation to come to Plön and attending my defence. To David Castellano for providing data, and Joost Schymkowitz and Floor Stam for helping with software. Tomislav Domazet-Lošo and Diethard Tautz for sharing the gene age data, with a special thanks to Diethard for the nice suggestions regarding this thesis. I also thank Thomas Bataillon for agreeing writing the review article with us and for the nice discussions. I also thank Toni Gossman for the writings suggestions and Nico for the help with the German abstract. Also Fernanda and Viplav for their internships, I also learned from you. And Matthias Leippe for agreeing being the chair in my defence.

This project would also not be possible without the invaluable help of all my friends here at the MPI. To Maria, we were together since the beginning. A lot of stories and a huge friendship that I really. To the Portuguese crowd, with a special thanks to Ana and Carolina for all the dinners, conversations, parties, and, in general, everything. This path was made much easier because of you girls. To Gillian, my mate, Juan and Loukas for all the fun that we had together. I really appreciate your friendship guys. To the people in the 'Bebe que la vida es breve', Karen, Tania, Federica and Ezgi, indeed life is short and I hope we can have more fun times together girls. To Gustavo, Bilal, and Natasha for all the fun times that we had in the MSE group, and also for all the suggestions that you gave to this project. To Alice, Dusica, Devika, Malavi, and Zahra for the nice talks and fun times. I'm sure I'm forgetting some people and I'm sorry for that, but I'm already really tired. So, here goes a huge thanks to all the family at the MPI, this was a better journey with you in it.

Aos meus amigos de Portugal, que já estão comigo há imenso tempo. À Sara, à Maria, à Carolina, ao Fábio, e á Kati. Apesar de não termos passado esta jornada juntos foi como se vocês tivessem sempre comigo.

Ao Joel, tudo isto não teria sido possível sem ti. A sério, a tua paciência, amor e amizade incondicional foram imprescindíveis nestes últimos anos. Sei que não foi nada fácil para ti e espero um dia poder compensar-te todo o apoio que me deste, em todos os sentidos. Um muito obrigado não chega de todo, mas e o que consigo fazer aqui. Obrigado!

E por último, à minha família: os meus pais nem o meu irmão. Nem sei por onde começar, este projeto não teria sido possível sem vocês. Todo o amor incondicional e paciência nesta minha jornada. Foi difícil, tive, e tenho, muitas saudades vossas ao longo deste caminho. Obrigado por tudo não inclui todo o amor e agradecimento que tenho por vocês, mas aqui vai: obrigado!

List of Manuscripts

Peer-reviewed Publications

Chapter 2:

Moutinho AF, Trancoso FF, Dutheil JY, The impact of protein architecture on adaptive evolution. *Molecular Biology and Evolution* 36 (9), pp. 2013 - 2028 (2019). https://doi.org/10.1093/molbev/msz134.

Review Article:

Moutinho AF, Bataillon T, Dutheil JY. Variation of the adaptive substitution rate between species and within genomes. *Evolutionary Ecology* (2019) https://doi.org/10.1007/s10682-019-10026-z.

Book Chapter

Barroso GV*, **Moutinho AF***, Dutheil JY (2020) A Population Genomics Lexicon. In: Dutheil J. (eds) Statistical Population Genomics. Methods in Molecular Biology, vol 2090. Humana, New York, NY

*contributed equally

Author Contributions

Chapter 2:

AFM and JYD designed the study. **AFM** performed the analyses and FFM helped filtering the *Arabidopsis* data. **AFM** and JYD interpreted the results and wrote the manuscript.

Chapter 3:

AFM and JYD designed the study. **AFM** performed the analyses. **AFM** and JYD interpreted the results. **AFM** wrote the thesis chapter.

Chapter 4:

AFM and JYD designed the study. **AFM** performed the analyses. **AFM** wrote the thesis chapter.

Chapters 1 and 5 are my own contribution.

Review Article:

AFM, JYD, and TB planned the structure of the review. **AFM** wrote the manuscript and JYD and TB contributed with suggestions.

Book Chapter:

AFM, GVB, and JYD designed the book chapter. **AFM** wrote the sections 3) Statistics on Nucleotide Diversity and 4) Selective Processes.

Authors given in alphabetic order:

AFM: **Ana Filipa Moutinho**; FFT: Fernanda Fontes Trancoso; GVB: Gustavo Valadares Barroso; JYD: Julien Yann Dutheil; TB: Thomas Bataillon.

Appendices

Appendix I

Appendix I is available at: https://doi.org/10.6084/m9.figshare.771874.

Appendix II

Table S1. Statistical results for the analysis of the effect of gene age as a function of RSA on ω, ω_{na}, and ω_{a}.

	A. thaliana			*D. melanogaster*		
RSA	**ω**	$\boldsymbol{\omega_{na}}$	$\boldsymbol{\omega_{a}}$	**ω**	$\boldsymbol{\omega_{na}}$	$\boldsymbol{\omega_{a}}$
Buried	0.854 (**)	0.818 (**)	0.709 (**)	0.667 (*)	0.333	0.667 (.)
Exposed	0.927 (***)	0.491 (*)	0.782 (***)	0.722 (**)	0.444 (.)	0.444 (.)

Note. For each variable, the Kendall's τ of gene age is shown with the respective significance (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; "." $0.05 \leq P < 0.10$) for ω, ω_{na} and ω_{a} in *A. thaliana* and *D. melanogaster*.

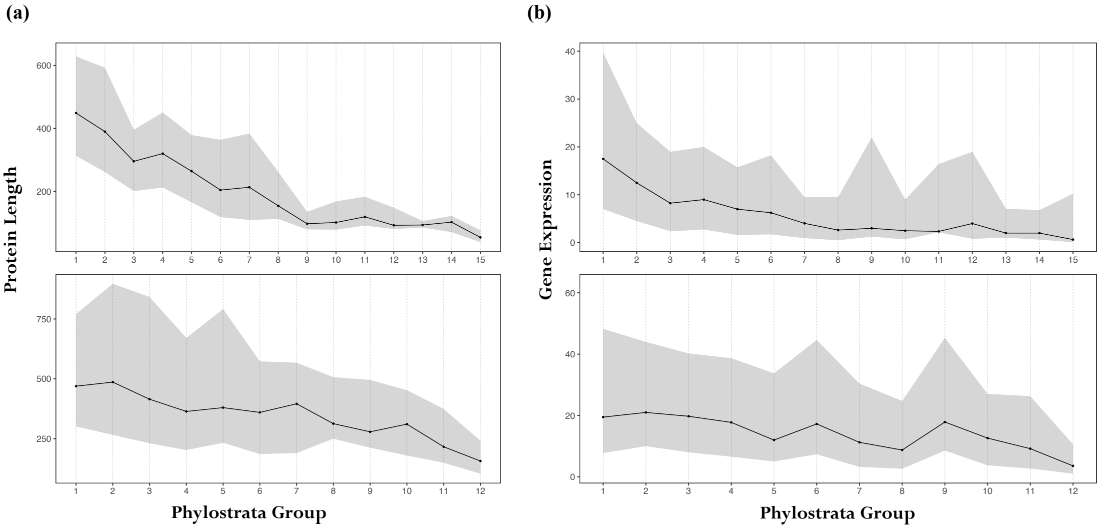

Figure S1. Relationship between gene age and gene length (a) and gene expression (b) for *A. thaliana* (top) and *D. melanogaster* (bottom). This analysis was performed by categorizing gene age according to the clades defined in Figure 1a. For each clade, the median value of gene length and gene expression is depicted with the black dot. The shaded area represents the values of gene length and mean expression levels within the 1st and 3rd quartile.

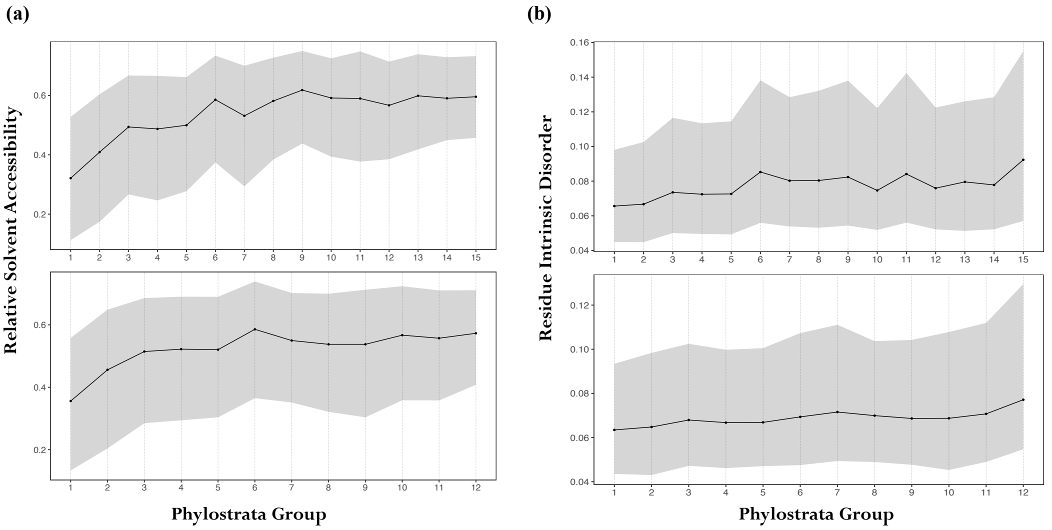

Figure S2. Relationship between gene age and RSA (a) and residue intrinsic disorder (b) for *A. thaliana* (top) and *D. melanogaster* (bottom). Legend as in Figure S1.

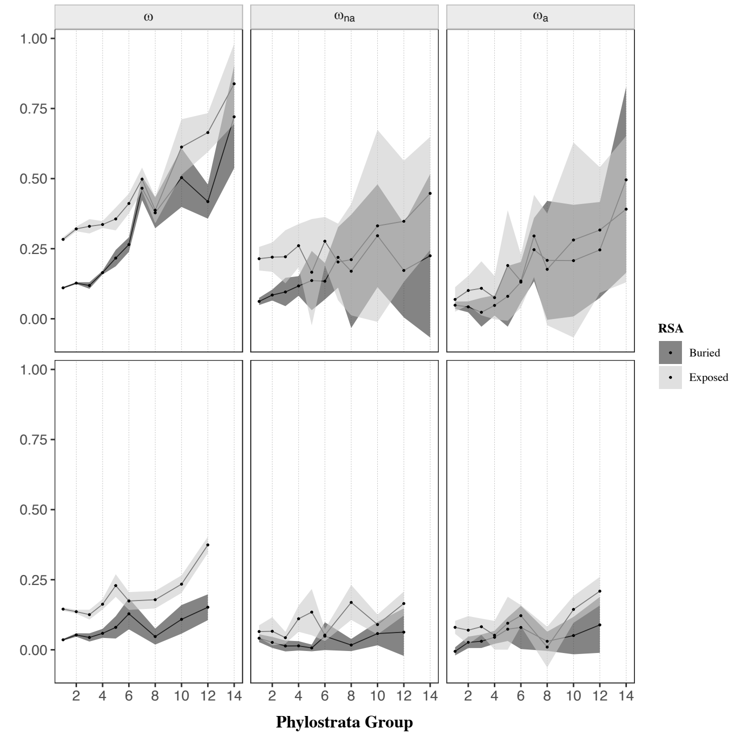

Figure S3. Estimates of ω, ω_{na} and ω_a plotted as a function of RSA and gene age in *A. thaliana* (top) and *D. melanogaster* (bottom). Analyses were performed by comparing a subset of buried and exposed residues (see Methods) across 11 and 9 categories of gene age in *A. thaliana* and *D. melanogaster*, respectively. Mean values of ω, ω_{na} and ω_a for each category are represented with the black points. Error bars denote for the 95% confidence interval for each category, computed over 100 bootstrap replicates.

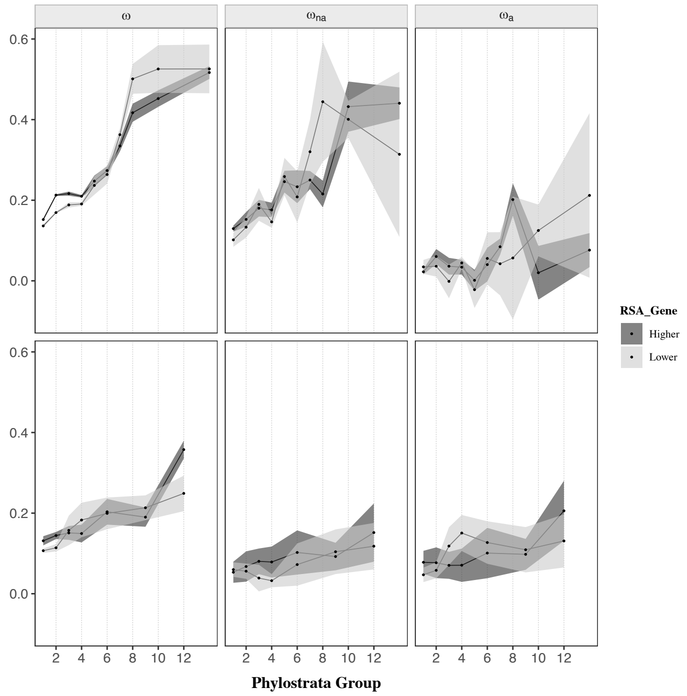

Figure S4. Estimates of ω, ω_{na} and ω_a plotted as a function of the proportion of exposed residues per protein and gene age in *A. thaliana* (top) and *D. melanogaster* (bottom). Analyses were performed by comparing proteins with higher and lower proportion of exposed residues (see Methods) across 10 and 7 categories of gene age in *A. thaliana* and *D. melanogaster*, respectively. Legend as in Figure S3.

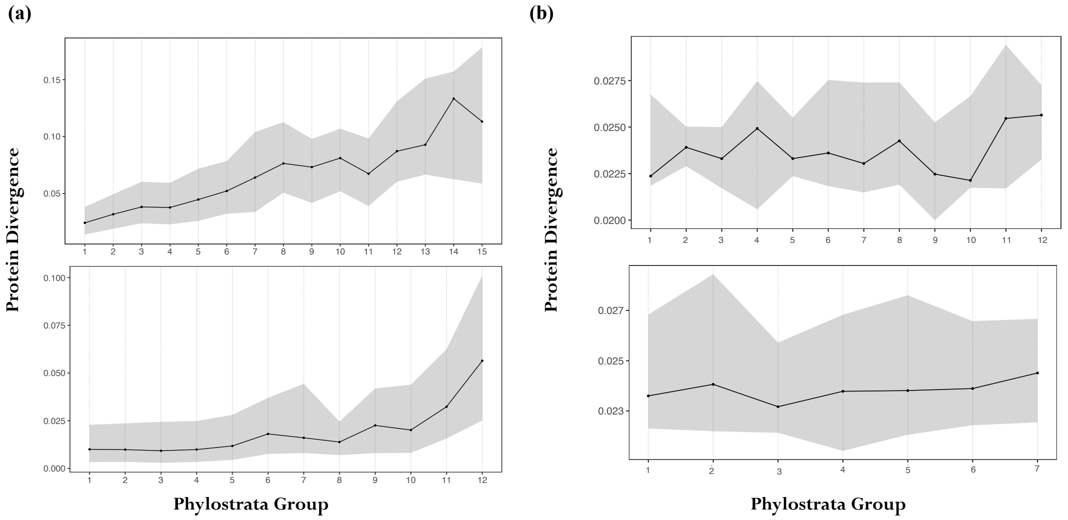

Figure S5. Relationship between gene age and protein divergence before (a) and after (b) correction for *A. thaliana* (top) and *D. melanogaster* (bottom). Legend as in Figure S1.

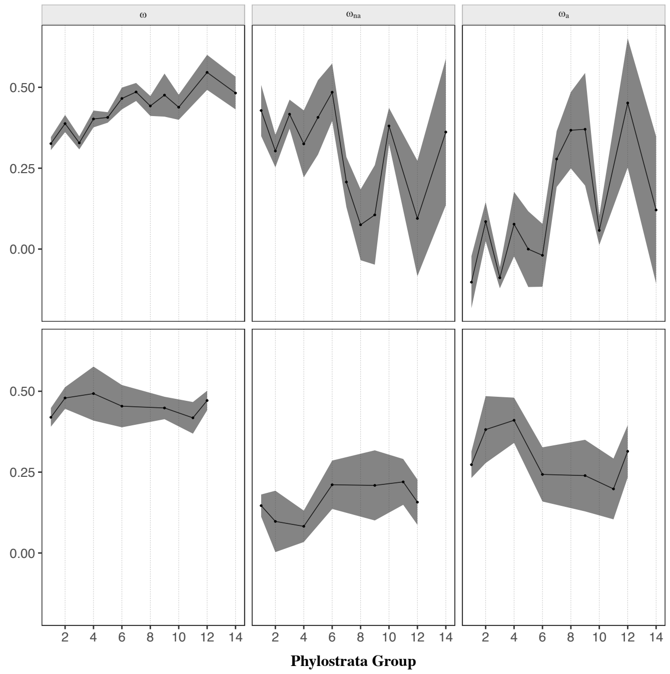

Figure S6. Estimates of ω, ω_{na} and ω_a plotted as a function of gene age by correcting for protein divergence in *A. thaliana* (top) and *D. melanogaster* (bottom). Legend as in Figure S3.

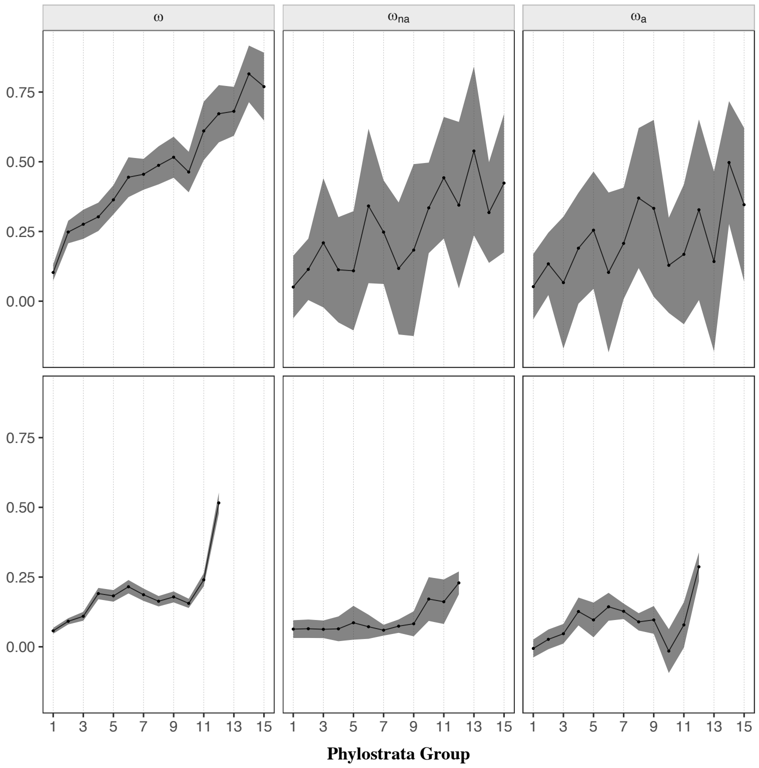

Figure S7. Relationship between ω, ω_{na} and ω_a and gene age with the same number of sites for each phylostrata in *A. thaliana* (top) and *D. melanogaster* (bottom). Legend as in Figure S3.

Appendix III

Table S1. Information on the number of genes, taxonomic group, the DFE model, and the transition to transversion ratio (ts/tv) for each species pair analysed in this study.

Focal Species	**Outgroup Species**	**Genes**	**Group**	**DFE Model**	**ts/tv**
Abatus_cordatus	Abatus_agassizi	2,144	echinoderms	ScaledBeta	1.438
Allolobophora_chlorotica_L2	Aporrectodea_icterica	9,751	annelids	ScaledBeta	2.421
Aptenodytes_patagonicus	Aptenodytes_forsteri	2,479	birds	GammaExpo	0.999
Armadillidium_vulgare	Armadillidium_nasatum	9,893	ants	GammaExpo	1.780
Artemia_franciscana	Artemia_sinica	7,464	crustaceans	GammaZero	2.664
Caenorhabditis_brenneri	Caenorhabditis_sp.10	836	nemtodes	GammaZero	2.502
Camponotus_ligniperdus	Camponotus_aethiops	7,588	ants	GammaZero	6.347
Chelonoidis_nigra	Chelonoidis_carbonaria	2,474	reptiles	ScaledBeta	1.403
Chlorocebus_aethiops	Macaca_mulatta	6,686	primates	GammaZero	1.105
Ciona_intestinalis_A	Ciona_intestinalis_B	3,750	sea_squirt	GammaZero	2.393
Ciona_intestinalis_B	Ciona_intestinalis_A	3,727	sea_squirt	GammaExpo	2.373
Crepidula_fornicata	Crepidula_plana	1,677	molluscs	ScaledBeta	2.113
Culex_pipiens	Culex_torrentium	3,704	flies	ScaledBeta	2.396
Cyanistes_caeruleus	Parus_major	1,433	birds	GammaExpo	1.939
Echinocardium_cordatum_B2	Echinocardium_mediterraneum	9,957	echinoderms	GammaExpo	2.396
Echinocardium_mediterraneum	Echinocardium_cordatum_B2	9,896	echinoderms	ScaledBeta	1.939
Emys_orbicularis	Trachemys_scripta	2,387	reptiles	GammaZero	1.980
Eudyptes_moseleyi	Pygoscelis_papua	2,453	birds	ScaledBeta	2.502
Eulemur_coronatus	Eulemur_mongoz	5,918	primates	GammaZero	1.037
Eulemur_mongoz	Eulemur_coronatus	5,857	primates	GammaZero	1.008
Eunicella_cavolinii	Eunicella_verrucosa	11,583	coral	ScaledBeta	2.097
Galago_senegalensis	Nycticebus_coucang	2,894	primates	GammaZero	2.448
Halictus_scabiosae	Halictus_simplex	3,495	ants	ScaledBeta	2.555
Hippocampus_guttulatus	Hippocampus_kuda	13,584	sea_horse	GammaZero	1.020
Homo_sapiens	Pan_troglodytes	6,102	primates	GammaZero	1.020
Lepus_granatensis	Lepus_americanus	1,137	hares	DisplGamma	3.706
Lineus_longissimus	Lineus_ruber	9,257	ribbon warms	GammaExpo	1.982
Macaca_mulatta	Chlorocebus_aethiops	6,686	primates	GammaZero	1.119
Melitaea_cinxia	Melitaea_didyma	5,155	butterflies	GammaZero	2.330
Messor_barbarus	Messor_structor	7,894	ants	GammaExpo	7.640
Microtus_arvalis	Microtus_glareolus	6,741	mice	ScaledBeta	3.918
Mytilus_galloprovincialis	Mytilus_californianus	10,844	molluscs	ScaledBeta	1.577
Necora_puber	Carcinus_aestuarii	6,551	crustaceans	ScaledBeta	2.436
Nycticebus_coucang	Galago_senegalensis	2,893	primates	GammaZero	2.406
Ophioderma_longicauda_L1	Ophioderma_longicauda_L3	6,461	echinoderms	GammaExpo	2.162
Ostrea_edulis	Ostrea_chilensis	2,823	molluscs	ScaledBeta	1.980
Pan_troglodytes	Homo_sapiens	6,097	primates	GammaExpo	1.167
Physa_acuta	Physa_gyrina	4,208	molluscs	ScaledBeta	1.681
Propithecus_coquereli	Varecia_variegata_variegata	5,825	primates	ScaledBeta	2.202
Thymelicus_lineola	Thymelicus_sylvestris	12,654	butterflies	GammaZero	2.538
Thymelicus_sylvestris	Thymelicus_lineola	12,649	butterflies	GammaExpo	2.431

Table S2. Statistical results for the analysis of the effect of the log-transformed π_S on ω_{na} and ω_a by accounting for the effect of the phylogeny.

RSA	$\boldsymbol{\omega_{na}}$	$\boldsymbol{\omega_a}$
0.031	1.687e-02 (***)	2.111e-02
0.130	6.9952-10 (***)	3.961 e-10
0.253	2.873e-01 (***)	8.495 e-02
0.370	1.010e-01 (***)	3.184 e-02
0.471	2.675e-01 (***)	5.884 e-03
0.559	1.241e-01 (***)	7.194 e-02
0.630	6.029e-02 (***)	8.504 e-03
0.683	2.233e-01 (***)	2.055 e-01
0.731	3.590e-02 (***)	6.184 e-02
0.781	3.203e-10 (***)	1.976 e-10

Note. For each variable, the rho coefficient of the phylogenetic regression (Grafen 1989) of the log-transformed π_S is shown with the respective significance (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; "." $0.05 \leq P < 0.10$) for ω_{na} and ω_a.

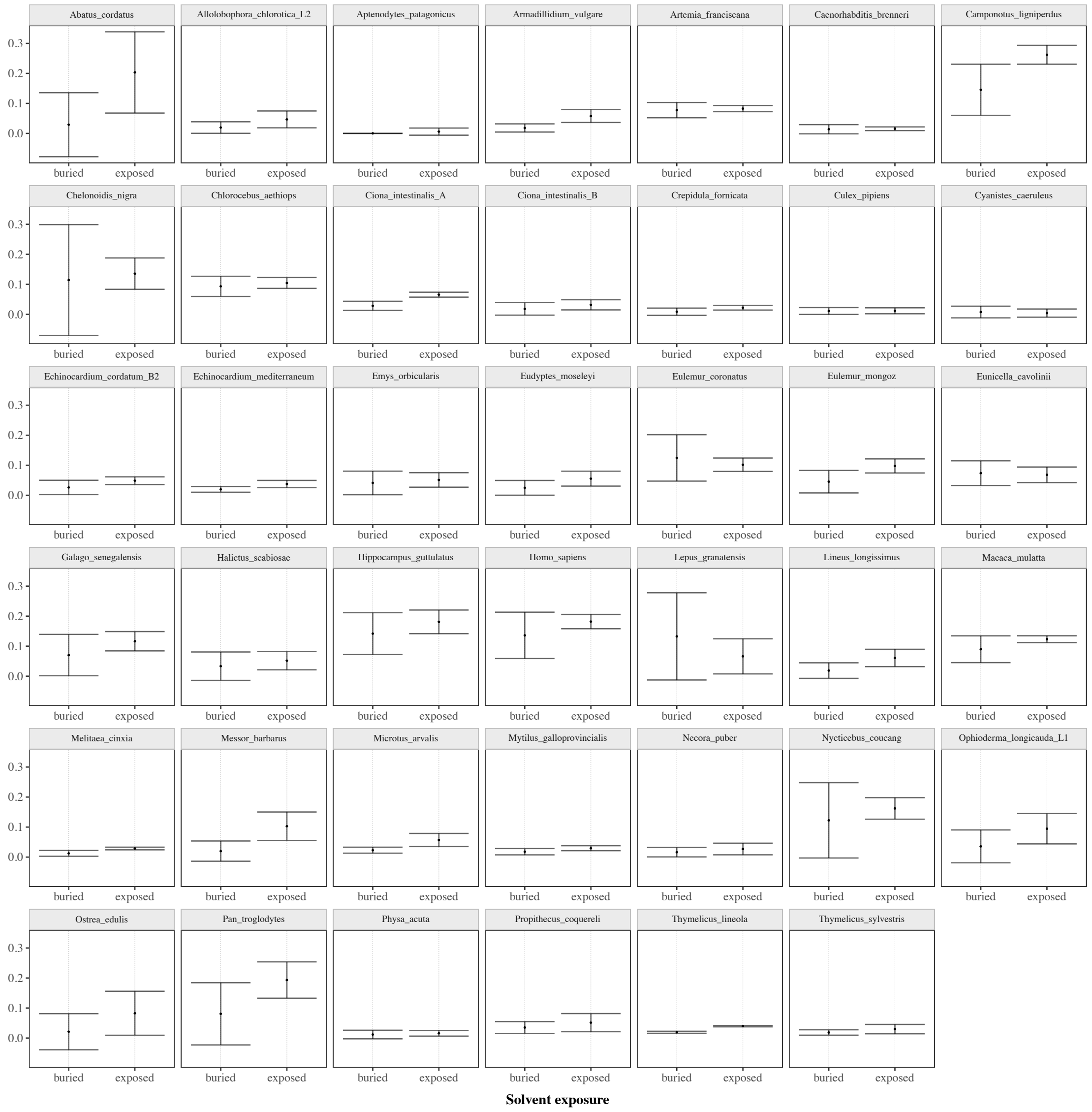

Figure S1. Estimates of ω_{na} for buried and exposed residues for each species analysed. Mean values of ω_{na} for each category of RSA is represented with the black points. Error bars denote for the 95% confidence interval for each category, computed over 100 bootstrap replicates.

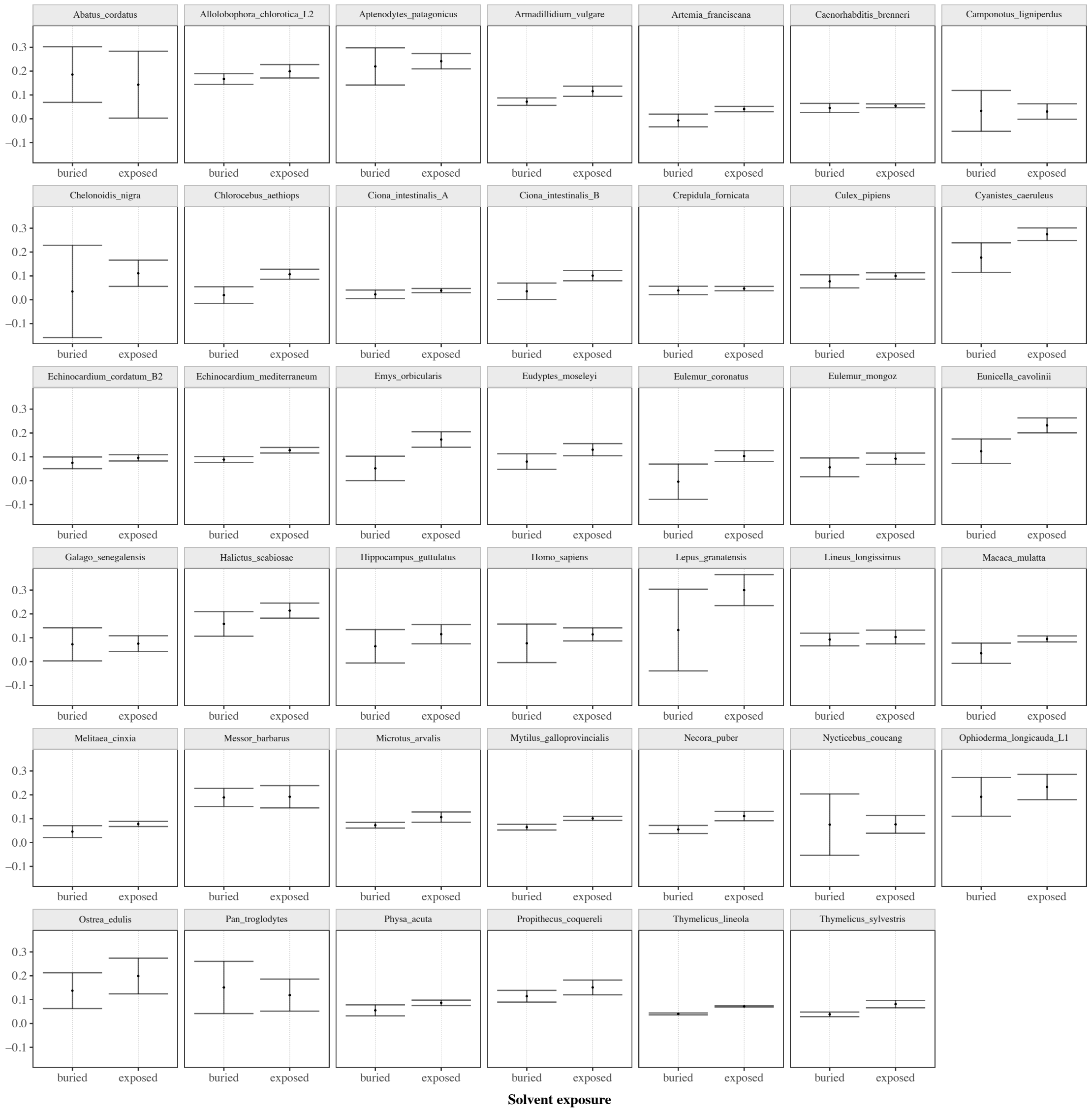

Figure S2. Estimates of ω_a for buried and exposed residues for each species analysed. Legend as in Figure S1.

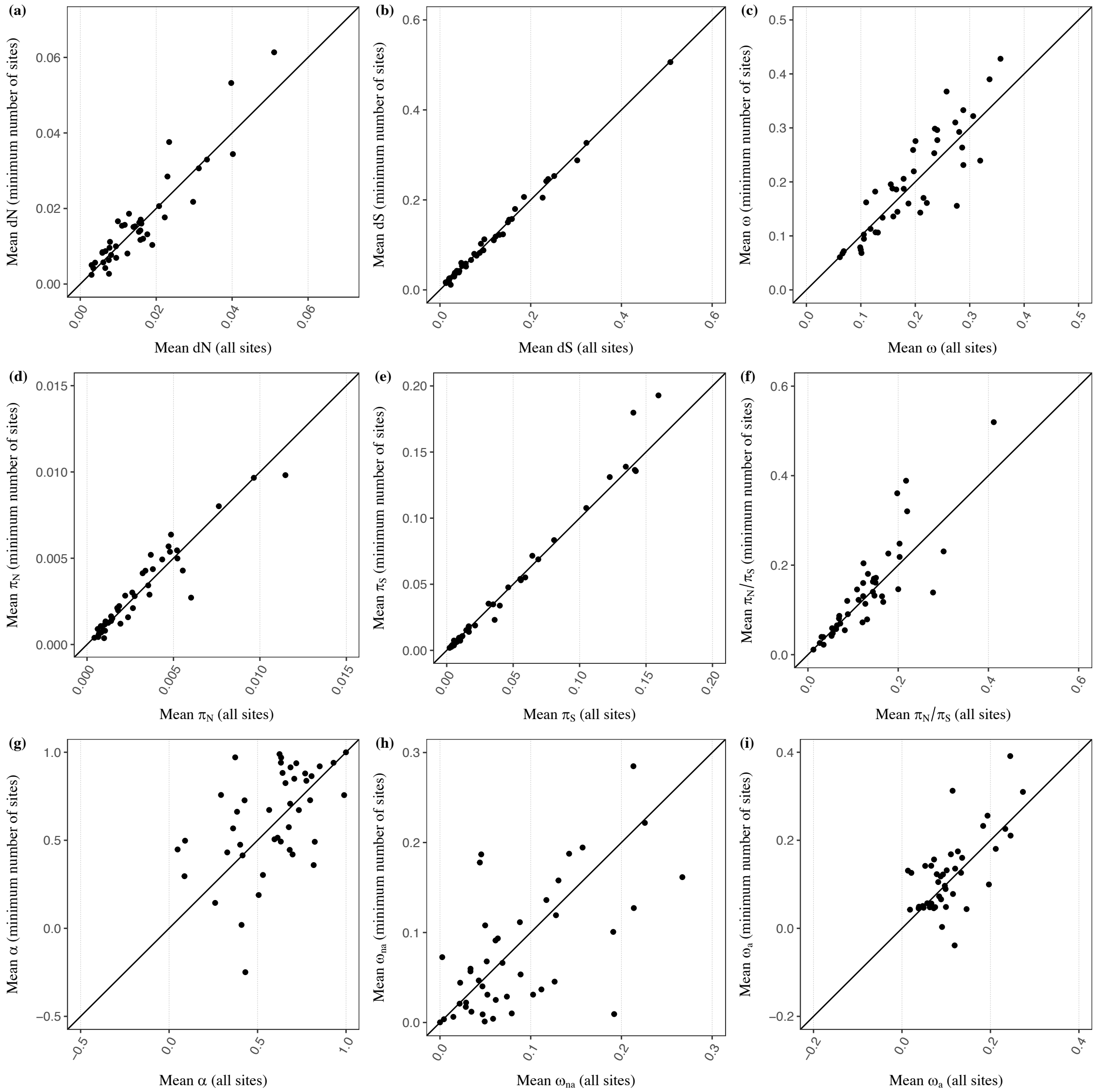

Figure S3. Correlation between estimates of **(a)** d_N, **(b)** d_S, (*c*) ω, **(d)** π_N, **(e)** π_S, **(f)** π_N/π_S, **(g)** α, **(h)** ω_{na}, and **(i)** ω_a, for all sites and the minimum number of sites in each species. Each dot represents one species.

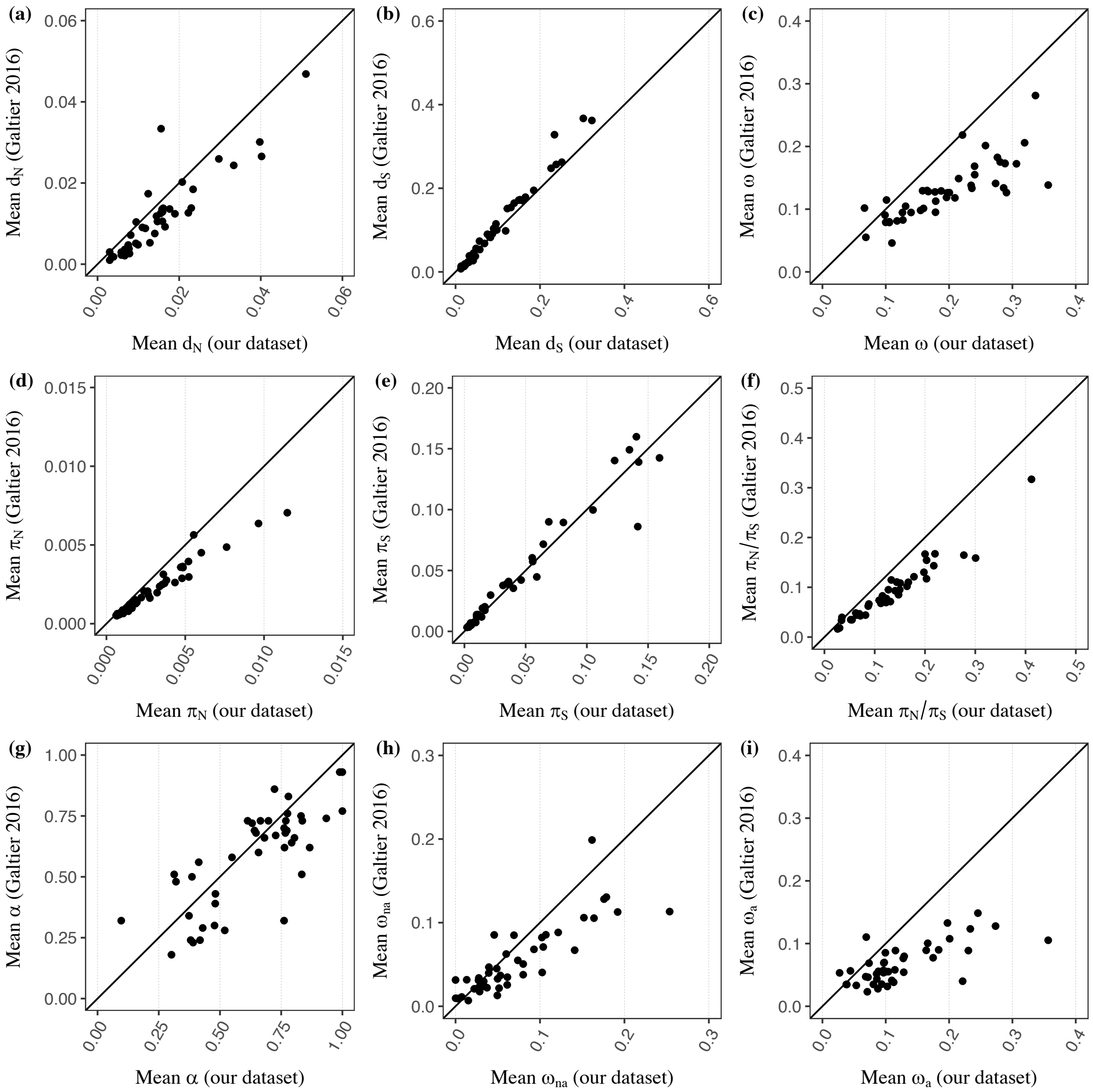

Figure S4. Correlation between our estimates of **(a)** d_N, **(b)** d_S, **(c)** ω, **(d)** π_N, **(e)** π_S, **(f)** π_N/π_S, **(g)** α, **(h)** ω_{na}, and **(i)** ω_a, with the ones obtained in Galtier (2016) for the full dataset. Each dot represents on species.

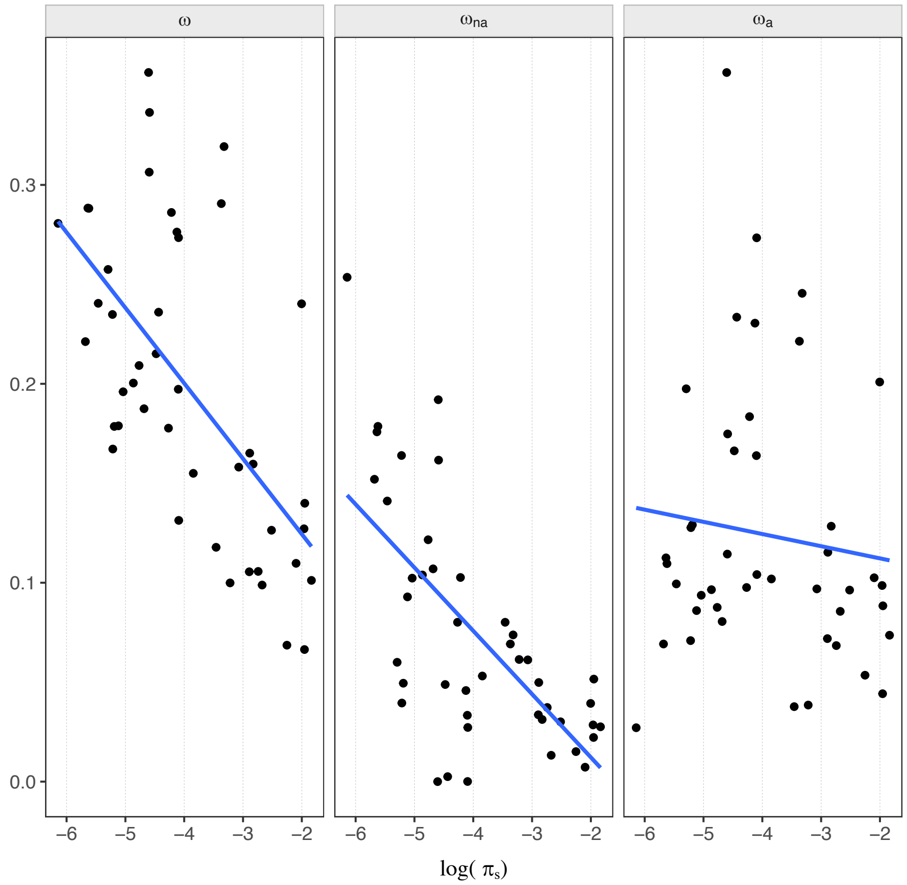

Figure S5. Relationship between the rate of protein evolution (ω), non-adaptive non-synonymous substitutions (ω_{na}) and adaptive non-synonymous substitutions (ω_a) with the log-transformed π_S. Each dot represents the mean values of the 100 bootstrap replicates performed for each species. The blue lines represent a linear model performed as a function of the log-transformed π_S.

Declaration

I hereby declare that:

i. Apart from my supervisor's guidance, the content and design of this thesis is the product of my own work. The co-authors' contributions are listed in the dedicated section;

ii. This thesis has not been already submitted either partially or wholly as part of a doctoral degree to another examination body, and no other materials are published or submitted for publication than indicated in the thesis;

iii. The preparation of the thesis has been subjected to the Rules of Good Scientific Practice of the German Research Foundation;

iv. Prior to this thesis, I have not attempted and failed to obtain a doctoral degree.

Plön, March 2020

Ana Filipa Moutinho

Published Manuscripts

Chapter 1

A Population Genomics Lexicon

Gustavo V. Barroso, Ana Filipa Moutinho, and Julien Y. Dutheil

Abstract

Population genomics is a growing field stemming from soon a 100 years of developments in population genetics. Here, we summarize the main concepts and terminology underlying both theoretical and empirical statistical population genomics studies. We provide the reader with pointers toward the original literature as well as methodological and historical reviews.

Key words Population genetics, Neutral theory, Coalescent theory, Mutation, Recombination, Selection, Lexicon

1 Genomic Variation

1.1 Loci, Alleles, and Polymorphism

Population genomics studies the evolution of genome variants in populations. A *locus (pl. loci)* refers to a given location in the genome. The particular sequence at a given locus may vary between individuals, each variant being termed an *allele*. We call loci with at least two alleles *polymorphic* and invariant loci *monomorphic*. The term *polymorphism* refers to the presence of multiple alleles but is commonly used as a countable noun as a substitute for "polymorphic locus" (*one polymorphism*, *several polymorphisms*).

Alleles may differ because of the nucleotide content, but also in length, as a result of nucleotide insertions or deletions (*a.k.a. indels*). Variable loci of length one can have up to four distinct alleles (A, C, G, or T) and are termed *single nucleotide polymorphisms (SNPs)*. SNPs constitute, so far, the majority of the data accounted for by population genetic models.

1.2 Mutations

Molecular events altering the genome are termed *mutations*. Mutations include substitution of a nucleotide into another one, removal or addition of one or several nucleotides, as well as multiplication of some part of the genome. Mutation is the process by which new

Authors Gustavo V. Barroso and Ana Filipa Moutinho contributed equally to this work.

Julien Y. Dutheil (ed.), *Statistical Population Genomics*, Methods in Molecular Biology, vol. 2090,
https://doi.org/10.1007/978-1-0716-0199-0_1, © The Author(s) 2020

alleles are formed. The *infinite site model* assumes that during the timeframe of evolution modeled, each locus have undergone at most one mutation [1–3]. This model also implies that each mutation creates a new allele in the population and that there is no "backward" or "reverse" mutation. The infinite site model is a generally reasonable assumption as the mutation rate is typically low and genomes are large. It might be locally invalidated, however, in case of mutation hotspots or when larger evolutionary timescales are considered. Under this premise, at most two alleles are expected per locus. Loci with two alleles are termed *diallelic* or *biallelic*, the first term having historical precedence and being more accurate [4], while the second is more commonly used since the 1990s. Furthermore, in a population genomic dataset, a sampled diallelic locus is called a *singleton* if one of the two alleles is present in only one haploid genome, and a *doubleton* if it is present in precisely two haploid genomes.

1.3 The Wright–Fisher Model

The simplest process of allele evolution within a single population is named the *Wright–Fisher model*. It describes the evolution of alleles in a population of fixed and constant size, where all alleles have the same fitness, and therefore the same chance to be transmitted to the next generation (*neutral evolution*). The population is assumed to be *panmictic*, that is, individuals are randomly mating. Time is discretized in *non-overlapping generations* so that the alleles in the current generation are a random sample of the alleles from the previous generation, without new alleles being generated by mutation. Under such conditions, allelic frequencies evolve only because of the stochasticity in the sampling of gametes that will contribute to the next generation, a process termed *genetic drift*. Because populations are of finite size, alleles will be sampled at their actual frequencies on average only and the ultimate fate of any allele is either to reach frequency zero in the population and be lost, when by chance no individual carrying this allele has any descendant in the next generation or to become fixed when all other alleles have been lost. The time until fixation depends on the population size: smaller populations will show a stronger sampling effect and shorter times to fixation. When genetic drift is the only force acting on a population, the number of alleles at a given locus is necessarily decreasing over time.

The *Wright–Fisher model with mutation* extends the Wright–Fisher model by introducing new alleles in the population, at a given rate. As the mutation rate is low, new mutations appear in a single copy, their initial frequency is then $1/2N$ in a diploid population. Mutation and drift act in opposite direction and a *mutation-drift equilibrium* is reached when the rate of allele creation by mutation equals the rate of allele loss by drift. The genetic diversity is then determined by the sole product of the population size N and the mutation rate u. Under the infinite site model, the expected heterozygosity at a locus in a population of diploid individuals is approximated by [1]

$$\hat{h} = \frac{4 \cdot N \cdot u}{4 \cdot N \cdot u + 1}$$

while the expected number of distinct alleles and their respective frequencies can be estimated using *Ewens's sampling formula* [5].

A *substitution* occurs when a new mutation has spread in the population, increasing from frequency $1/(2N)$ to 1 (*see* **Note 1**). Kimura showed that the average time to fixation of a new mutation is $4N$ in a population of diploid individuals [6]. Furthermore, as a neutral mutation has a probability of reaching fixation equal to $1/(2N)$ and given that there are $2N \cdot u$ new mutations per generation, in a purely neutrally evolving population, the expected number of substitutions per generation is equal to $2N \cdot u \cdot 1/(2N) = u$. The substitution rate is therefore independent of the population size and, assuming that the mutation rate is constant in time, the number of substitutions between two populations is a direct measure of the number of generations separating them, a phenomenon termed *molecular clock* [7].

1.4 The Backward Wright–Fisher Model: The Standard Coalescent

While the Wright–Fisher process naturally describes the evolution of sequences within populations one generation after the other, population genetic data typically represent individuals sampled at a given time point. For inference purposes, it is therefore convenient to model the history of the genetic material that gave rise to the sample. The modelization of the ancestry of a sample (also known as the *genealogy*) is typically done backward in time, as every locus find a common ancestor in the past, until the *most recent common ancestor (MRCA)* of the sample. The merging of two lineages in the past is called a *coalescence event*, and the set of mathematical tools describing this process under a variety of demographic models is referred to as the *coalescence theory.* Kingman [8] first described the *standard coalescent*, the genealogical model corresponding to the Wright–Fisher model (but *see* refs. 9 and 10 for a historical perspective). The standard coalescent is, therefore, also referred to as the *Kingman's coalescent.*

2 Beyond the Wright–Fisher Model

The Wright–Fisher model has been extended in several ways to include more realistic assumptions on the underlying evolutionary process. These extensions led to the concept of *Effective population size (Ne)*, originally defined as the number of individuals contributing to the gene pool. When a population deviates from the assumptions of the Wright–Fisher model, *Ne* is no longer equal to the census population size (N). Often (but not always) in such cases, *Ne* can be obtained by a linear scaling of *N* such that it reflects the number of individuals from an idealized Wright–Fisher population that would display the same genetic diversity as the actual population under study [11].

2.1 Demography

A possible deviation from the Wright–Fisher assumptions happens when the population size is not constant across generations. The term *demographic history* generally refers to the collection of demographic parameters (effective sizes, growth rates) that describes the history of the population until its most recent common ancestor [12]. When population size varies in a cyclic manner with relatively small period *n* generations, the resulting genealogies can be modeled by a Wright–Fisher process with a population size equal to the harmonic mean of the historical population sizes, so that

$$Ne = \frac{n}{\sum_i^n \frac{1}{N_i}},$$

where N_i refer to the *i*th population size [13]. More drastic demographic effects include *genetic bottlenecks*, corresponding to a sharp decrease (shrinkage) in population size.

2.2 Population Structure

In the absence of *panmixia*, genetic exchanges occur more often between certain individuals, resulting in *population structure* with several subpopulations. Population structure may occur for different reasons such as overlapping generations, assortative mating, or geographic isolation [12]. *Assortative mating* occurs when individuals choose their mates according to some similarity between their phenotypes. If the phenotype is genetically determined, assortative mating can influence the level of heterozygosity in the population [14].

Gene flow describes the migration of genetic variants between subpopulations under a scenario of population structure. It reduces genetic differentiation among subpopulations [15]. Ultimately, subpopulations can diverge and become genetically isolated, a process called *speciation*. The simplest speciation processes involve spontaneous isolation (*isolation model*) or spontaneous isolation followed by a period of gene flow (*isolation with migration model*) [16].

When speciation events occur in a short timeframe and ancestral population sizes are large, ancestral polymorphism may persist in the ancestral species, a phenomenon called *incomplete lineage sorting (ILS)* [17]. The expected amount of ILS depends on the number of generations between two isolation events (Δ_T) and the ancestral effective population size Ne_A [18]:

$$\Pr(ILS) = \frac{2}{3} e^{\left(-\frac{2 \cdot \Delta_T}{Ne_A}\right)}$$

The term *introgression* is used to depict the transfer of genetic material between diverged populations or species through secondary contact [19]. As a result, extant lineages share a common ancestor that predates the two isolation or speciation events. The resulting genealogy may, therefore, be incongruent with the phylogeny defined by the two splits, depending on the order of coalescence events between lineages [20].

3 Statistics on Nucleotide Diversity

Statistics are needed to infer population genetics parameters from polymorphism data. The *site frequency spectrum (SFS)* describes the empirical distribution of allele frequencies across segregating sites of a given (set of) loci in a population sample. For a sample of n sequences (in n haploid individuals or $n/2$ diploid individuals), the so-called unfolded SFS is the set of counts of derived alleles $X = (X_1, X_2, \ldots, X_{n-1})$, where sample configurations X_i denote the number of sites that have $n - i$ ancestral and i derived alleles. The ancestral state is usually estimated using an outgroup sequence. In cases where we cannot assess the ancestral allele, the folded site frequency spectrum, X', may be calculated instead. X' represents the distribution of the minor allele frequencies, such as $X'_i = X_i + X_{n-i}$ for $i < n/2$ and $X'_{n/2} = X_{n/2}$ [13, 21, 22]. The shape of the SFS is affected by underlying population genetic processes, such as demography and selection, and therefore serves as the input of many population genetics methods [23] (*see* Fig. 1).

Watterson's theta, here noted $\hat{\theta}_S$, is an estimator of the population mutation rate $\theta = 4Ne \cdot u$, where Ne is the (diploid) effective population size and u the mutation rate. It is derived from the number of segregating sites S_n of a sample of size n [25]. Assuming an infinite sites model, S_n is equal to the product of u and the expected time to coalescence, corrected by the sample size:

$$E[Sn] = u \cdot 4 \cdot Ne \sum_{i=1}^{n-1} i.$$

Since $4Ne \cdot u = \theta$ the equation may be written as $E[Sn] = \theta \cdot a_n$, where $a_n = \sum_{i=1}^{n-1} i$. The proposed estimator of θ for the sample is

$$\hat{\theta}_S = \frac{\hat{S}_n}{a_n} = \frac{\hat{S}_n}{\left(1 + \frac{1}{2} + \ldots + \frac{1}{n-1}\right)},$$

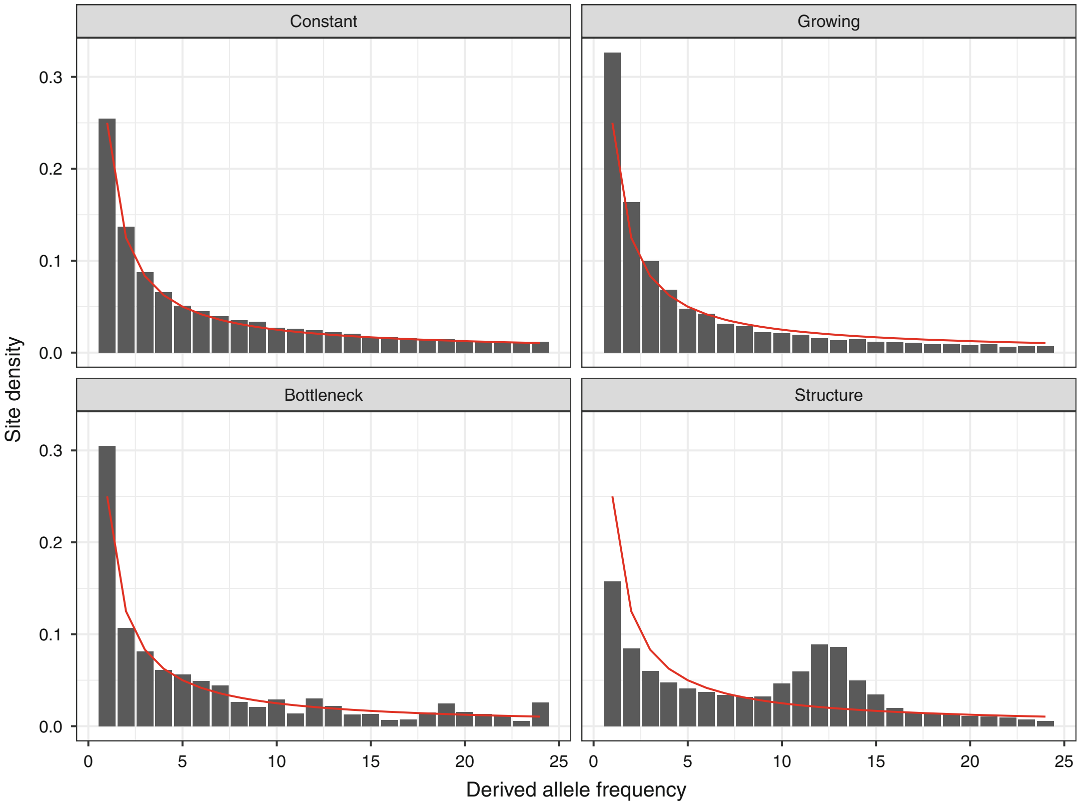

Fig. 1 Effect of demography on the shape of the site frequency spectrum (SFS). The figure depicts four scenarios: constant population size, exponential growth, genetic bottleneck, and population structure. The red curve shows the expectation under a constant population size. In the case of exponential growth or a genetic bottleneck, the SFS displays an excess of low-frequency variants. Population structure, here simulated as two subpopulations exchanging migrants at a low rate, results in an excess of intermediate frequency variant when we reconstruct a single SFS from the two subpopulations. Simulations were performed using the `msprime` software [24] (*see* also Chapter 9 and the online companion material)

where $\hat{S}_n$ is the observed number of segregating sites in the sample. In order to be comparable, values of θ are usually reported per site, and $\hat{\theta}_S$ is then further divided by the sequence length L. This estimator is unbiased when the data is generated from a Wright–Fisher process but is not robust to deviations from it, due to selection or demography [26].

Tajima's π, the *average pairwise heterozygosity* is a measure of nucleotide diversity defined as the number of pairwise differences between a set of sequences [27]. Under the infinite sites model, the number of mutations separating two orthologous chromosomes D_{ij} is equal to the number of nucleotide differences between sequences i and j. As the expectation of the average pairwise nucleotide differences between all pairs of sequences in a sample is equal to $\theta = 4Ne \cdot u$ [28], Tajima's estimator of θ is:

$$\hat{\theta}_{\pi} = \frac{2}{n(n-1) \cdot L} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} D_{ij},$$

where L is the total sequence length.

4 Selective Processes

4.1 Protein-Coding Genes

The coding region of a protein-coding gene, also known as *Coding DNA Sequence (CDS)* is the portion of DNA, or RNA, that encodes a protein. A start and stop codons limit the coding region at the five-prime and three-prime end, respectively. In mRNAs, the CDS is bounded by the five-prime untranslated region (5-UTR) and the three-prime untranslated region (3'-UTR), also included in the exons. Mutations within coding regions are expected to be of distinct types: *synonymous mutations* lead to no change of amino-acid at the protein level due to the redundancy of the genetic code, as opposed to *non-synonymous mutations*. Non-synonymous mutations can further be classified as *conservative* and *non-conservative (= radical)*, whether they replace an amino-acid by a biochemically similar one or not. Because of the structure of the genetic code, the four types of mutations at one site (toward A, C, G, or T) can be in principle both synonymous and non-synonymous. Sites where *n* out of four possible mutations are synonymous are called *n-fold degenerated*. *Four-fold degenerated* sites only undergo synonymous mutations, while a mutation at a so-called *zero-fold degenerated* site is necessarily non-synonymous. Most of second codon positions are zero-fold degenerated, while many of the third positions are four-fold degenerated.

4.2 Fitness Effect

The resulting change of fitness at the organism level characterizes the type of mutations: neutral mutations have no impact on the fitness, while harmful or deleterious mutations induce a lower fitness. Conversely, advantageous mutations increase the fitness of the organism compared to the wild-type genotype. There is, however, a wide range of selective effects, which extends the categorization of mutations from strongly deleterious, through weakly deleterious, neutral to mildly and highly adaptive mutations. The relative frequencies of these types of mutations represent the distribution of fitness effects [29, 30].

The *selection coefficient (s)* is a measure of differences in fitness, which determines the changes in genotype frequencies that occur due to selection. It is commonly expressed as a relative fitness. If one considers a single locus with two alleles A and a, a standard parametrization is to attribute a fitness of 1 to the homozygote AA and relative fitness of $1 + s$ for the homozygote aa. The heterozygote Aa is attributed a fitness of $1 + h \cdot s$, where h is the so-called *coefficient of dominance*. The s parameter varies between -1 and $+\infty$ (but *see* **Note 2**), wherein values comprised among -1 and 0 are indicative of negative selection, while positive values correspond to positive selection [13, 31]. The efficiency of selection, however, depends on both s and the effective population size, Ne, so that mutations with $Ne \cdot s \ll 1$ behave in effect like neutral mutations, whose fate is determined by genetic drift only [29].

4.3 Types of Selection

Positive selection acts on alleles that increase fitness, raising their frequency in the population over time, while *negative selection* (= *purifying selection)* decreases the frequency of alleles that impair fitness. Both positive and negative selection decrease genetic diversity. Conversely, *balancing selection* acts by maintaining multiple alleles in the gene pool of a population at frequencies higher than expected by drift alone. Three mechanisms are generally acknowledged: *heterozygous advantage*, where heterozygotes have a higher fitness than homozygotes and maintain genetic polymorphism; *frequency-dependent selection*, where the fitness of the genotype is inversely proportional to its frequency in the population; and *environment-dependent fitness* of genotypes (also known as *local adaptation*) [31, 32].

4.4 Inference of Selection in Protein-Coding Sequences

The strength and direction of selection acting on protein-coding regions may be assessed by contrasting the rate of non-synonymous (potentially under selection, dN) to synonymous (assumed to be neutral, dS, but see, for instance, [33]) substitutions between species. In a population of sequences evolving neutrally, all substitutions are neutral and the two rates are equal, leading to a dN/dS ratio equal to one on average. Assuming non-synonymous mutations are either neutral or deleterious while synonymous mutations are always neutral, the rate of non-synonymous substitutions will be lower than the rate of synonymous substitutions, and the dN/dS ratio will be lower than one. Conversely, if non-synonymous mutations are positively selected, their rate of fixation may exceed the rate of synonymous mutation, leading to a higher substitution rate and a dN/dS ratio higher than one.

At the population level, the ratio of non-synonymous (pN) and synonymous (pS) polymorphism is indicative of the strength of purifying selection acting on a protein. Because non-synonymous mutations are more likely to have a negative fitness effect and be counter-selected, they tend to be removed from the population by purifying selection or segregate at low-frequency. We can estimate the synonymous and non-synonymous genetic diversity by computing the average pairwise heterozygosity π separately for non-synonymous and synonymous mutations, noted π_N and π_S, respectively. The π_N/π_S ratio is therefore generally below one, the stronger the purifying selection, the closer the ratio is to zero.

Contrasting the dN/dS and pN/pS ratios allows to test the selection regime acting on the sequences [34]. If mutations are all neutral or deleterious, we expect the ratios dN/dS and pN/pS to be equal. Positively selected mutations will tend to quickly rise to fixation and will not be observed as polymorphism, leading to an increased dN/dS ratio higher than pN/pS. Conversely, balancing selection will lead to an excess of polymorphism detectable as $dN/dS < pN/pS$ [35]. A simple measure of the proportion of amino-acid substitutions resulting from positive selection (α) is given by $1 - (dS \cdot pN/dN \cdot pS)$ [36]. Using the complete synonymous and non-synonymous site frequency spectra, it is further possible to estimate the distribution of fitness effects and account for slightly deleterious and slightly advantageous mutations when estimating the rate of adaptive substitutions (*see* Chapter 5) [37].

5 Linkage and Recombination

5.1 The Coalescent with Recombination

In sexually reproducing species, *recombination* refers to both the shuffling of non-homologous chromosomes and the rearrangement of homologous chromosomes during meiosis. Such crossover events cause each chromosome to have two parent chromosomes in the previous generation, which are themselves the products of recombination events in the previous generations. Therefore, any chromosome in the current generation can be viewed as a mosaic of chromosomes that existed in the past (*see* Fig. 2) [38]. The collection of coalescence and recombination events that describes the history of sampled chromosomes until the most recent common ancestor of each non-recombining block is reached (*see* Fig. 2) is called the *ancestral recombination graph (ARG)* [39]. Compared to a tree-like genealogy of a sample without recombination, whose complexity depends only on the sample size, the complexity of the ARG grows with the sample size and the number of recombination events in the ancestry of the sample.

Backward-in-time, the *most recent common ancestor (MRCA)* denotes the first individual where the entire sample (population) coalesces for a particular non-recombining block. The *TMRCA* notes the timing of such event. DNA sequences provide no information beyond the MRCA in a sample of genomes since all individuals will share any mutation that happens further back in time [40]. In the presence of recombination, different parts of the genome will have different MRCAs. In this case, all ancestral material is eventually found as a contiguous sequence in the *grand most*

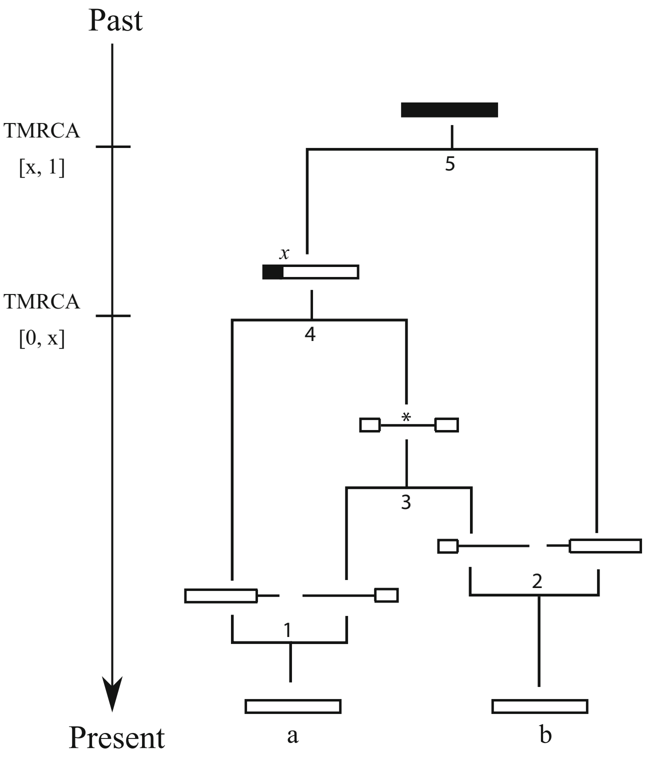

Fig. 2 An ancestral recombination graph. An ancestral recombination graph is a collection of recombination (1–2) and coalescence (3–5) events. In each depicted chromosome, white bars represent segregating ancestral material, black bars represent coalesced ancestral material, and thin lines represent non-ancestral material. The asterisk denotes trapped non-ancestral material. Note that "1" does not impact the sample because the resulting segments are joined back together in "4" before coalescing in "5." There are thus only two relevant TMRCAs in the ARG, separated at position *x*

recent common ancestor (GMRCA) of the sample (*see* Fig. 2). If the GMRCA is not an MRCA for any nucleotide, this individual does not have any significance for DNA sequences [39].

In the ARG, nucleotide segments that are found both in past chromosomes and in contemporary samples are termed *ancestral genetic material* (*see* Fig. 2). Conversely, *non-ancestral genetic material* refers to segments that are found in past chromosomes but not in contemporary samples. Furthermore, non-ancestral genetic material flanked on both sides by ancestral genetic material is referred to as *trapped genetic material*. In this setting, recombination events that happen in trapped genetic material can affect linkage disequilibrium between present-day nucleotides (*see* Fig. 2). Thus the existence of trapped genetic material introduces long-range correlations between genealogies rendering the coalescent with recombination a non-Markovian process along chromosomes

[41]. The *Sequentially Markov coalescent (SMC)* is an approximation to the coalescent with recombination whereby recombination events are assumed to happen only within ancestral material. This approximation allows the use of efficient algorithms in both simulation and data analysis [42, 43].

5.2 Impact of Linkage on Selection

An excess of linkage between loci compared to a random association is termed *linkage disequilibrium (LD)*. LD arises from genetic drift, population admixture, and selection, but is reduced by recombination each generation. It is, therefore, higher between close loci and decays with increasing physical distance [44].

Linked selection refers to the reduction of diversity at neutral sites that happens as a result of their physical linkage to variants under selection [45]. In the absence of recombination, all variants segregating in a chromosome would undergo the same shift in frequency as the selected variant. However, recombination creates new allelic combinations and reduces this correlation as the physical distance from the selected locus increases (*see* Fig. 3).

Background selection refers to a form of linked selection where the reduction of diversity at neutral loci results from linkage to a locus under purifying selection [46], and *genetic hitchhiking* is commonly used to depict linked selection due to linkage to a locus under positive selection [47], where a new beneficial mutation will rise in frequency in a population. As the new positively selected allele increases its frequency, nearby linked alleles on the chromosome will "hitchhike" along with it, also growing in frequency, thus producing a *selective sweep* of genetic diversity (*see* Fig. 3d). *Hard sweeps* occur when a new mutation is positively selected and is therefore exclusively associated with the genetic background where it arose. Conversely, *soft sweeps* occur when a mutation is already segregating in the population at the onset of selection. This mutation may exist in several genetic backgrounds and therefore does not prompt a complete loss of genetic variation after the selective sweep [47] (*see* Fig. 3a–c).

Linkage of two or more loci can also impair the efficacy of positive selection, a phenomenon termed *Hill–Robertson interference (HRI)* [48]. When two advantageous mutations at distinct loci in distinct individuals segregate in the population, one will be lost unless a recombination event brings them together. In the absence of recombination between the selected loci, only the unlikely event of recurrent mutations can generate the optimal haplotypic combination [49] (*see* Fig. 3e).

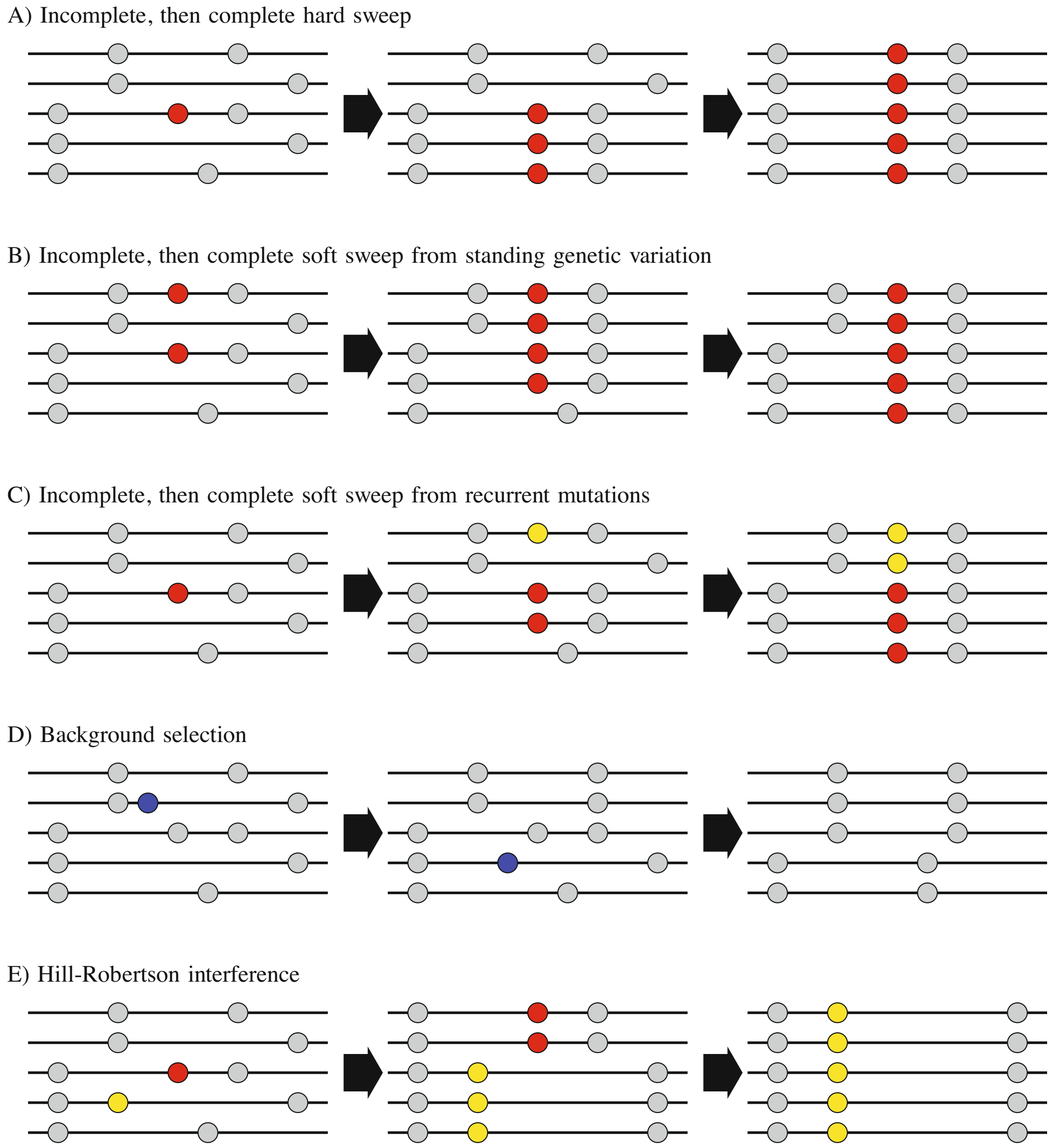
A) Incomplete, then complete hard sweep
B) Incomplete, then complete soft sweep from standing genetic variation
C) Incomplete, then complete soft sweep from recurrent mutations
D) Background selection
E) Hill-Robertson interference

Fig. 3 Impact of selection on genetic diversity. Black lines represent individual genomes. SNP variants are displayed by filled circles. Distinct variants at the same position are depicted with different colors: neutral variants in gray, positive variants in red or yellow, and negative variant in blue. (**a**) A positively selected new variant spreads in the population and removes genetic diversity at linked loci, generating a hard selective sweep. (**b** and **c**) Segregation of several positively selected variants in different genetic backgrounds, either from standing variation or recurrent mutations, resulting in a soft selective sweep. (**d**) Reduction of neutral diversity because of linkage to deleterious mutations (background selection). (**e**) Competitive segregation of positively selected variant at distinct loci, resulting in the loss of advantageous variants (Hill–Robertson interference)

6 Notes

1. The use of the term *substitution* differs in population genetics and molecular biology. In the latter case, it describes a particular type of mutation where a single nucleotide replaces a distinct one (as opposed to insertions/deletions, for instance).
2. In some instances, s is substituted by $-s$, so that the relative fitnesses become $\omega_{AA} = 1$, $\omega_{Aa} = 1 - h \cdot s$ and $\omega_{aa} = 1 - s$.

References

1. Kimura M, Crow JF (1964) The number of alleles that can be maintained in a finite population. Genetics 49:725–738
2. Kimura M (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61(4):893–903
3. Crow JF (1989) Twenty-five years ago in genetics: the infinite allele model. Genetics 121(4):631–634
4. Elston RC, Satagopan J, Sun S (2017) Statistical genetic terminology. Methods Mol Biol 1666:1–9. https://doi.org/10.1007/978-1-4939-7274-6_1
5. Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theor Popul Biol 3(1):87–112
6. Kimura M (1970) The length of time required for a selectively neutral mutant to reach fixation through random frequency drift in a finite population. Genet Res 15(1):131–133
7. Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge. http://ebooks.cambridge.org/ref/id/CBO9780511623486
8. Kingman JFC (1982) The coalescent. Stoch Process Appl 13(3):235–248. https://doi.org/10.1016/0304-4149(82)90011-4
9. Barton NH (2016) Richard Hudson and Norman Kaplan on the coalescent process. Genetics 202(3):865–866. https://doi.org/10.1534/genetics.116.187542
10. Kingman JFC (2000) Origins of the Coalescent: 1974–1982. Genetics 156(4):1461–1463. http://www.genetics.org/content/156/4/1461
11. Sjödin P, Kaj I, Krone S, Lascoux M, Nordborg M (2005) On the meaning and existence of an effective population size. Genetics 169(2):1061–1070. https://doi.org/10.1534/genetics.104.026799
12. Wakeley J (2008) Coalescent theory: an introduction, 1st edn. Roberts and Company Publishers, Reading
13. Wright S (1938) Size of population and breeding structure in relation to evolution. Science 87:430–431
14. Jiang Y, Bolnick DI, Kirkpatrick M (2013) Assortative mating in animals. Am Nat 181(6):E125–138. https://doi.org/10.1086/670160
15. Sousa V, Hey J (2013) Understanding the origin of species with genome-scale data: modelling gene flow. Nat Rev Genet 14(6):404–414. https://doi.org/10.1038/nrg3446
16. Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. Genetics 167(2):747–760. https://doi.org/10.1534/genetics.103.024182
17. Dutheil JY, Hobolth A (2012) Ancestral population genomics. Methods Mol Biol 856:293–313. https://doi.org/10.1007/978-1-61779-585-5_12
18. Hobolth A, Christensen OF, Mailund T, Schierup MH (2007) Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genet 3(2):e7. https://doi.org/10.1371/journal.pgen.0030007
19. Martin SH, Jiggins CD (2017) Interpreting the genomic landscape of introgression. Curr Opin Genet Dev 47:69–74. https://doi.org/10.1016/j.gde.2017.08.007
20. Mailund T, Munch K, Schierup MH (2014) Lineage sorting in Apes. Annu Rev Genet

https://doi.org/10.1146/annurev-genet-120213-092532

21. Bustamante CD, Wakeley J, Sawyer S, Hartl DL (2001) Directional selection and the site-frequency spectrum. Genetics 159 (4):1779–1788
22. Wright S (1968) Evolution and the genetics of populations, vol 2. The theory of gene frequencies. The University of Chicago Press, Chicago
23. Schraiber JG, Akey JM (2015) Methods and models for unravelling human evolutionary history. Nat Rev Genet 16(12):727–740. https://doi.org/10.1038/nrg4005
24. Kelleher J, Etheridge AM, McVean G (2016) Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS Comput Biol 12(5):e1004842. https://doi.org/10.1371/journal.pcbi.1004842
25. Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7(2):256–276
26. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3):585–595
27. Nei M, Tajima F (1981) Genetic drift and estimation of effective population size. Genetics 98(3):625–640. http://www.genetics.org/content/98/3/625
28. Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105(2):437–460
29. Eyre-Walker A, Keightley PD (2007) The distribution of fitness effects of new mutations. Nat Rev Genet 8(8):610–618. https://doi.org/10.1038/nrg2146
30. Orr HA (2009) Fitness and its role in evolutionary genetics. Nat Rev Genet 10(8):531–539. https://doi.org/10.1038/nrg2603
31. Gillespie JH (2004) Population genetics: a concise guide. JHU Press, Baltimore
32. Nielsen R (2005) Molecular signatures of natural selection. Annu Rev Genet 39:197–218. https://doi.org/10.1146/annurev.genet.39.073003.112420
33. Pouyet F, Bailly-Bechet M, Mouchiroud D, Guéguen L (2016) SENCA: a multilayered codon model to study the origins and dynamics of codon usage. Genome Biol Evol 8 (8):2427–2441. https://doi.org/10.1093/gbe/evw165
34. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351(6328):652–654. https://doi.org/10.1038/351652a0
35. Parsch J, Zhang Z, Baines JF (2009) The influence of demography and weak selection on the McDonald-Kreitman test: an empirical study in Drosophila. Mol Biol Evol 26(3):691–698. https://doi.org/10.1093/molbev/msn297
36. Smith NGC, Eyre-Walker A (2002) Adaptive protein evolution in Drosophila. Nature 415 (6875):1022–1024. https://doi.org/10.1038/4151022a
37. Keightley PD, Eyre-Walker A (2007) Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177 (4):2251–2261. https://doi.org/10.1534/genetics.107.080663
38. Stumpf MPH, McVean GAT (2003) Estimating recombination rates from population-genetic data. Nat Rev Genet 4(12):959–968. https://doi.org/10.1038/nrg1227
39. Hein J, Schierup MH, Wiuf C (2005) Gene genealogies, variation and evolution: a primer in coalescent theory. Oxford University Press, Oxford
40. Rosenberg NA, Nordborg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat Rev Genet 3 (5):380–390. https://doi.org/10.1038/nrg795
41. Rasmussen MD, Hubisz MJ, Gronau I, Siepel A (2014) Genome-wide inference of ancestral recombination graphs. PLoS Genet 10(5): e1004342. https://doi.org/10.1371/journal.pgen.1004342
42. McVean GAT, Cardin NJ (2005) Approximating the coalescent with recombination. Philos Trans R Soc Lond B Biol Sci 360 (1459):1387–1393. https://doi.org/10.1098/rstb.2005.1673
43. Marjoram P, Wall JD (2006) Fast "coalescent" simulation. BMC Genet 7:16. https://doi.org/10.1186/1471-2156-7-16
44. Slatkin M (2008) Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9 (6):477–485. https://doi.org/10.1038/nrg2361
45. Cutter AD, Payseur BA (2013) Genomic signatures of selection at linked sites: unifying the disparity among species. Nat Rev Genet 14

(4):262–274. https://doi.org/10.1038/nrg3425

46. Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134 (4):1289–1303
47. Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23(1):23–35
48. Hill WG, Robertson A (1966) The effect of linkage on limits to artificial selection. Genet Res **8**(3):269–294
49. Roze D, Barton NH (2006) The Hill-Robertson effect and the evolution of recombination. Genetics 173(3):1793–1811. https://doi.org/10.1534/genetics.106.058586

Evolutionary Ecology
https://doi.org/10.1007/s10682-019-10026-z

ORIGINAL PAPER

Variation of the adaptive substitution rate between species and within genomes

Ana Filipa Moutinho[1] · Thomas Bataillon[2] · Julien Y. Dutheil[1,3]

Received: 15 May 2019 / Accepted: 4 December 2019

Abstract

The importance of adaptive mutations in molecular evolution is extensively debated. Recent developments in population genomics allow inferring rates of adaptive mutations by fitting a distribution of fitness effects to the observed patterns of polymorphism and divergence at sites under selection and sites assumed to evolve neutrally. Here, we summarize the current state-of-the-art of these methods and review the factors that affect the molecular rate of adaptation. Several studies have reported extensive cross-species variation in the proportion of adaptive amino-acid substitutions (α) and predicted that species with larger effective population sizes undergo less genetic drift and higher rates of adaptation. Disentangling the rates of positive and negative selection, however, revealed that mutations with deleterious effects are the main driver of this population size effect and that adaptive substitution rates vary comparatively little across species. Conversely, rates of adaptive substitution have been documented to vary substantially within genomes. On a genome-wide scale, gene density, recombination and mutation rate were observed to play a role in shaping molecular rates of adaptation, as predicted under models of linked selection. At the gene level, it has been reported that the gene functional category and the macromolecular structure substantially impact the rate of adaptive mutations. Here, we deliver a comprehensive review of methods used to infer the molecular adaptive rate, the potential drivers of adaptive evolution and how positive selection shapes molecular evolution within genes, across genes within species and between species.

Keywords Adaptive evolution · Between-species · Within-genomes · Intra-molecular · Molecular evolution

✉ Ana Filipa Moutinho
moutinho@evolbio.mpg.de

[1] Research Group Molecular Systems Evolution, Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Biology, August-Thienemann-Str. 2, 24306 Plön, Germany

[2] Bioinformatics Research Center, Aarhus University, C.F. Møllers Allé 8, 8000 Aarhus C, Denmark

[3] UMR 5554, Institut des Sciences de l’Evolution, CNRS, IRD, EPHE, Université de Montpellier, Place E. Bataillon, 34095 Montpellier, France

Published online: 14 December 2019

Introduction

After Darwin proposed that natural selection acts as a main driver of evolution, a major goal of evolutionary biologists has been to understand how beneficial mutations shape species adaptation to their environment. Over the years, the number of approaches used to detect positive selection has increased substantially, making use of the increasing amount of genome data available. In particular, methods have been developed to pinpoint genes, or positions within these genes, that exhibit a pattern of genetic variation statistically incompatible with a pure nearly-neutral scenario (Ohta 1992), where mutations are considered to be neutral, nearly neutral or deleterious (*i.e.* Nielsen et al. 2005; Ometto et al. 2005; Kosiol et al. 2008). The ecological relevance of such candidate genes can be further tested using functional annotations, when available, or experimentally, for instance, by using reverse genetics and ancestral allele reconstruction (*i.e.* Hilson et al. 2004; Nielsen et al. 2005; Voight et al. 2006; Roux et al. 2014). This allowed to detect instances of adaptive evolution in many functional categories, such as immune genes in ants (Roux et al. 2014) and in hominids (Nielsen et al. 2005), virulence associated genes in pathogens (Stukenbrock et al. 2011; Dong et al. 2014), and coat-color related genes in hares (Jones et al. 2018) and mice (Hoekstra et al. 2006). While such methods allow a detailed understanding of case-studies, they do not enable one to assess the genome-wide distribution of the fitness effects of mutations.

By contrast, mutation accumulation (MA) experiments are specifically designed to estimate a genome-wide rate of mutation and distribution of effects of mutations on fitness (*i.e.* Shaw et al. 2002; Bataillon 2003; Rutter et al. 2012). With this approach, one can infer (1) the number of mutations that led to the divergence between MA lines, and (2) the fitness effects of these mutations on the (fitness-related) trait of interest (*i.e.* viability or lifetime reproductive success; see Glossary). Previous studies have inferred the presence of beneficial mutations in MA line experiments both in the field and in greenhouse studies of *A. thaliana* (Shaw et al. 2002; Rutter et al. 2012). Nonetheless, MA approaches can only give insight on recent adaptive events, and, therefore, provide little information regarding the proportion of adaptive genetic differences between species. Furthermore, MA experiments yield too few beneficial mutations to be able to test for the occurrence of genomic regions where adaptive mutations are more likely to occur. Conversely, population genomic approaches only offer indirect insights on mutation rates and fitness effects but can leverage patterns of sequence variation between and within species to infer rates of adaptive evolution, thus providing knowledge on the drivers of adaptation at deeper scales of evolution.

The role of positive (a.k.a. Darwinian) selection in molecular evolution is still widely debated (Hey 1999; Gillespie 2000; Kern and Hahn 2018; Jensen et al. 2019). The neutral theory of molecular evolution (Kimura 1968) states that the bulk of segregating polymorphisms is either neutral or deleterious and that the genetic differences between species are explained mainly by neutral substitutions (see Glossary), while beneficial mutations are considered to be too rare to contribute much to the observed polymorphism and divergence. With an increasing amount of data becoming available, however, the question of whether adaptive mutations play a role in molecular evolution can be investigated with a greater precision. “How much of the genetic variation can be explained by adaptive evolution? What is the frequency of adaptive mutations along the genome? Are there regions where adaptive mutations are more likely to occur?” are some of the questions that can now be addressed with population genomics data and statistical methods for the inference of selection.

Here, we present the current state-of-the-art methods used to model the distribution of fitness effects (DFE) and infer the frequency of adaptive mutations. We then review evidence for variation in the rate of adaptive evolution within genes, within genomes and between species.

Synthesis of methods

In the following section, we review the methods that can be used to estimate the rate of adaptive evolution from sequence data. We distinguish two main approaches: phylogenetic methods, based on the divergence between multiple species; and population genetics approaches, which contrast within-species polymorphism to the divergence with an outgroup species.

Glossary

Mutation accumulation (MA): experimental design where a single inbred line is used to create various sub-lines that are propagated under conditions minimizing the opportunity for selection. MA lines are allowed to diverge independently for several generations. The number of mutations that led to the divergence between MA lines and the fitness effects of these mutations on the trait of interest influence the empirical distribution of the mean phenotypic value of the trait. If the trait measured is fitness or a fitness component, this setting can be used to infer the genome-wide mutation rates and the underlying distribution of fitness effects (DFE, see below).

Synonymous mutation: a mutation, in a protein-coding region, that leaves the amino-acid residue unchanged.

Non-synonymous mutation: a mutation, in a protein-coding region, that leads to a change in the amino-acid residue.

Substitution: a fixed difference between species.

Polymorphism: a mutation segregating within a population (or a species).

Positive selection: selective process by which a beneficial mutation increases in frequency within a population.

Adaptive evolution: at the molecular level, it occurs in a certain genomic region through the successive fixation of advantageous mutations (Charlesworth and Charlesworth 2010).

Negative/Purifying selection: natural selection against a deleterious mutation.

Distribution of fitness effects (DFE) of mutations: represents the distribution of the relative frequencies of selection coefficients (s), extending from strongly and weakly deleterious, through neutral mutations to slightly and strongly advantageous.

d_N: number of non-synonymous substitutions per site.

d_S: number of synonymous substitutions per site.

D_n: number of non-synonymous substitutions per gene/region.

D_s: number of synonymous substitutions per gene/region.

P_n: number of non-synonymous polymorphisms per gene/region.

P_s: number of synonymous polymorphisms per gene/region.

α: proportion of amino-acid substitutions that are adaptive.

Genetic drift: random changes in allele frequencies produced by the sampling of the genetic variants that compose a population every new generation.

Genetic draft: a process that induces allele frequency changes through recurrent selective sweeps at linked positions.

Selective sweep: the process by which a beneficial substitution reduces genetic diversity at linked positions.

Background selection: the process by which negatively selected deleterious mutations reduce neutral genetic diversity at linked positions.

$\boldsymbol{\omega_a}$: rate of adaptive amino-acid non-synonymous substitutions relative to the mutation rate.

$\boldsymbol{K_{a+}}$: rate of adaptive amino-acid substitutions, denoted as: αK_a, where K_a represents an alternative notation of d_N.

Quantifying the proportion of adaptive substitutions

(1) Phylogenetic methods

The strength and direction of selection on the branch of a phylogenetic tree can be measured by contrasting the nonsynonymous (d_N) and synonymous divergence (d_S) in a given gene (*e.g.* Miyata et al. 1979; Li et al. 1985; Yang and Nielsen 2002; Eyre-Walker 2006). The d_N/d_S ratio, noted as ω, provides an estimate of the rate of nonsynonymous substitutions relative to the rate of synonymous substitutions. Assuming that mutation rates at synonymous and non-synonymous sites are constant and equal, and that synonymous substitutions are selectively neutral, genes with $\omega > 1$ are considered to be evolving under positive selection, while genes with $\omega < 1$ are evolving under negative selection. Because ω is based on averages of substitution rates across multiple nucleotide sites that undergo both positive and negative selection, this statistic can only detect strong positive selection (e.g. Yang and Nielsen 2002; Eyre-Walker 2006). As most nonsynonymous mutations are expected to be either neutral or deleterious, d_N will tend to be much lower than d_S, hence ω will tend to be globally lower than one (i.e. Yang and Nielsen 2002; Eyre-Walker 2006).

In order to consider variation in selective constraints in space and time, models have been developed to account for variable selective pressure among sites (Nielsen and Yang 1998; Yang et al. 2000, 2005), branches (Yang and Nielsen 1998), or both (so-called branch-site models; Yang and Nielsen 2002; Zhang et al. 2005; Kosakovsky Pond et al. 2011). In site-based models, the ω ratio varies across sites and positive selection is inferred at a specific site if the average d_N is higher than d_S over all lineages. In branch-based models, the ω ratio varies among lineages and positive selection is detected if the average d_N is higher than d_S across all sites in a certain branch or a series of branches defining a lineage in a phylogenetic tree. In turn, branch-site models allow the ω ratio to vary both across sites and lineages. Using this framework, distinct models can be compared to test for the occurrence of positive selection at particular sites or branches (*e.g.* Yang and Nielsen 2002; Zhang et al. 2005). Although these methods detect adaptation at the site level, it has been shown that they are conservative in measuring selection over a certain region and/or lineage (Rodrigue and Lartillot 2017). This higher conservatism could be due to adaptive processes not being concentrated on a small number of sites but rather scattered across a large number of positions in a certain genomic region (Rodrigue and Lartillot 2017). Moreover, branch-site models assume that evolution on the majority of branches is neutral and that adaptive processes are rare and usually isolated. Hence, events of frequent adaptation over long evolutionary periods would not be captured, leading to underestimates of the rate of adaptive evolution in the tested proteins (Nielsen and Yang 1998; Yang et al. 2000, 2005; Rodrigue and Lartillot 2017). Besides, as these approaches are based on multiple-species alignments, the analysis is focused on genes that are shared by all species, which are more ancient and typically more conserved. Rapidly evolving genes are typically discarded from such analysis since their alignment becomes less reliable as the divergence between species increases.

(2) Population genetics methods

a. The McDonald and Kreitman (MK) test

Population genetic methods pioneered by Hudson, Kreitman, and Aguadé (1987) test a neutral evolution scenario by comparing the number of polymorphic sites within a population with the number of substitutions with a distinct species (HKA test). Under a neutral scenario, the relative amount of polymorphism and divergence is constant between loci. The HKA test compares these values between at least two genomic regions to test this prediction (Hudson et al. 1987). McDonald and Kreitman (1991) first extended this approach to detect adaptive protein evolution (Fig. 1). The so-called MK test requires data from as little as two closely-related species, typically including several individuals in the study species and one individual from an outgroup species. It compares the number of polymorphisms to the number of substitutions for a locus in two classes of sites: synonymous, which are assumed to evolve neutrally, and non-synonymous, which are potentially under selection (McDonald and Kreitman 1991). The number of nonsynonymous substitutions is denoted as D_n, the number of synonymous substitutions as D_s, the number of nonsynonymous polymorphisms as P_n and the number of synonymous polymorphisms as P_s (see Glossary), leading to the so-called MK-table:

	Polymorphisms	Substitutions
Synonymous	P_s	D_s
Non-synonymous	P_n	D_n

Under a scenario where all mutations are either strongly deleterious or neutral, D_n/D_s is expected to be equal to P_n/P_s. Conversely, D_n/D_s higher than P_n/P_s is taken as a signature of positive selection, and D_n/D_s lower than P_n/P_s can be observed in case of balancing selection. As a beneficial mutation reaches fixation at a faster rate than a neutral mutation, it contributes comparatively more to divergence than to polymorphism levels (McDonald and Kreitman 1991; Eyre-Walker 2006).

b. Extensions of the MK-test: Estimation of the proportion of amino-acid substitutions (α)

By applying a derivative of the MK-table, Charlesworth (1994) estimated the proportion of amino-acid substitutions that are driven by positive selection, a measure referred to as α (Fig. 1; see Glossary) (Charlesworth 1994; Smith and Eyre-Walker 2002): $\alpha = 1 - (D_sP_n)/(D_nP_s)$. However, as the levels of nucleotide diversity and amino-acid divergence are generally low, the numbers of polymorphic sites and nonsynonymous substitutions are very

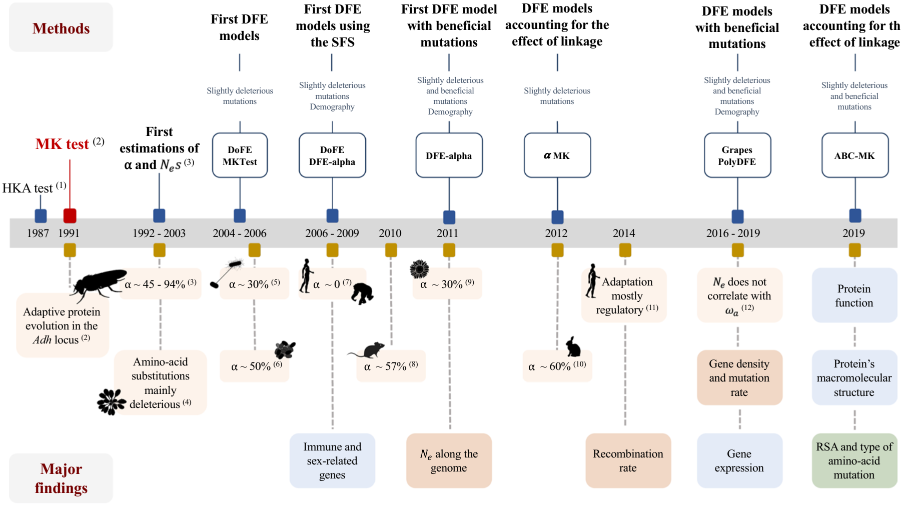
Methods
MK test (2)
HKA test (1)
First estimations of α and $N_e s$ (3)
First DFE models
Slightly deleterious mutations
DoFE MKTest
First DFE models using the SFS
Slightly deleterious mutations Demography
DoFE DFE-alpha
First DFE model with beneficial mutations
Slightly deleterious and beneficial mutations Demography
DFE-alpha
DFE models accounting for the effect of linkage
Slightly deleterious mutations
α MK
DFE models with beneficial mutations
Slightly deleterious and beneficial mutations Demography
Grapes PolyDFE
DFE models accounting for the effect of linkage
Slightly deleterious and beneficial mutations
ABC-MK
1987
1991
1992 - 2003
2004 - 2006
2006 - 2009
2010
2011
2012
2014
2016 - 2019
2019
Adaptive protein evolution in the Adh locus (2)
α ~ 45 - 94% (3)
Amino-acid substitutions mainly deleterious (4)
α ~ 30% (5)
α ~ 50% (6)
α ~ 0 (7)
Immune and sex-related genes
α ~ 57% (8)
α ~ 30% (9)
N_e along the genome
α ~ 60% (10)
Adaptation mostly regulatory (11)
Recombination rate
N_e does not correlate with ω_a (12)
Gene density and mutation rate
Gene expression
Protein function
Protein's macromolecular structure
RSA and type of amino-acid mutation
Major findings

Fig. 1 Timeline presenting the state-of-the-art population genetic methods to infer the rate of adaptive evolution (top) and the major findings on the factors impacting the variation of the molecular adaptive rate across species, along the genome, between and within genes (bottom). References for DFE methods can be found in Table 1. Light orange boxes correspond to the variation of the molecular adaptive rate between species; dark orange boxes represent the variation along the genome; blue boxes represent variation between protein-coding genes; and the green box correspond to the factors impacting the molecular adaptive rate at the intra-genic level. References for these studies can be found in the corresponding section in the main text. α: proportion of adaptive amino-acid substitutions; N_e: effective population size; s: selection coefficient; ω_a: rate of adaptive non-synonymous substitutions; RSA: relative solvent accessibility. References: (1) Hudson et al. 1987; (2) McDonald and Kreitman 1991; (3) Sawyer and Hartl 1992, Charlesworth 1994, Smith and Eyre-Walker 2002, Fay et al. 2001, Bustamante et al. 2002, Sawyer et al. 2003; (4) Bustamante et al. 2002; (5) Charlesworth and Eyre-Walker 2006; (6) Williamson 2003, Nielsen and Yang 2003; (7) i.e. Hvilsom et al. 2012, Eyre-Walker and Keightley 2009; (8) Halligan et al. 2010; (9) Gossmann et al. 2010, Strasburg et al. 2011; (10) Carneiro et al. 2012; (11) Enard et al. 2014; (12) Galtier 2016. Species figures were taken from PhyloPic (http://www.phylopic.org)

small for most genes taken individually. Hence, estimates of α for single genes have inherently large sampling variances, leading to the need for pooling data across many genes (Stoletzki and Eyre-Walker 2011). Such pooling is often done by summing counts of polymorphisms and divergence in each category (Fay et al. 2001) or by taking the average across genes (Smith and Eyre-Walker 2002). By using a different parametrization of the MK test, Sawyer and Hartl (1992) used a Poisson random field (PRF) model to derive expectations for the counts of D_n, D_s, P_n and P_s by considering the processes of mutation, selection, and genetic drift (see Glossary) acting independently and simultaneously at multiple sites (Sawyer and Hartl 1992). From the PRF model, one can relate the scaled selection coefficient ($\gamma = N_{e\ s}$, where N_e represents the effective population size and s the selection coefficient) and counts of polymorphism and divergence. Based on this approach, Bayesian models were developed where the posterior distribution of scaled selection coefficients for a given locus is inferred either by assuming a fixed-effects model, where γ is constant across sites (Bustamante et al. 2002); or a random-effects model, where γ of each new mutation is drawn from a single underlying normal distribution (Sawyer et al. 2003).

However, a limitation of these approaches is that they do not account for the segregation of slightly deleterious mutations, which can bias estimates of α in a demography-dependent manner (Eyre-Walker and Keightley 2009). On the one hand, α can be underestimated if the population size has been relatively constant or decreased since the divergence from the outgroup species, because slightly deleterious mutations may be observed as polymorphisms while having a much lower chance of fixation when compared to neutral mutations. This, however, can be controlled by removing polymorphisms segregating at low frequencies (Charlesworth 1994; Smith and Eyre-Walker 2002). On the other hand, α can be overestimated if the tested population experienced a demographic expansion: as the level of polymorphism is much lower, it leads to an apparent excess of substitutions (Eyre-Walker 2002). Modelling of the full range of the fitness effects of mutations and proper accounting of the underlying demography of the sample is, therefore, needed to achieve more accurate estimates of α.

Inferring α and the distribution of fitness effects (DFE) from the site frequency spectrum (SFS)

In the following, we briefly present methods that are specifically designed to infer the distribution of fitness effects from the frequency of the derived alleles across the genome in order to estimate the rate of adaptive evolution.

a. The folded/unfolded Site Frequency Spectrum (SFS)

The site frequency spectrum (SFS) is used to summarize the levels of polymorphisms in a sample of individuals. It represents the empirical distribution of the allelic frequencies for a given set of loci in the population. If the information on the ancestral allele at each variable position is available, the unfolded SFS can be computed, where the set of counts of the derived allele will be given. Conversely, if the ancestral allele cannot be inferred, the folded SFS may be calculated instead, representing the distribution of the minor allele frequencies. In these approaches, the SFS of potentially selected sites is compared to a neutral SFS. Most methods do so by comparing a non-synonymous to a synonymous SFS, however, this can also be done by contrasting genic with intergenic regions (Racimo and Schraiber 2014) or protein-binding with non-binding sites (Jenkins et al. 1995). The shape of both SFS provides crucial information on the underlying population genetic processes, such as demography and selection (Schraiber and Akey 2015; Barroso et al. 2019). For instance, slightly deleterious mutations segregate more often at low frequencies relative to neutral ones, while positively selected mutations are typically segregating at a higher frequency. But demography can also impact the SFS. For example, an expanding population has an excess of rare variants relative to what is expected in a stable population (Tajima 1989; Schraiber and Akey 2015; Barroso et al. 2019). The challenge is, therefore, to distinguish between the effect of selection and demography. This is done by assuming a neutral reference, for instance, the synonymous SFS, to which a demographic model is fitted. Selection is then inferred from the non-synonymous SFS. This assumption, together with the assumption of site independence is central to all methods inferring the distribution of fitness effects from the SFS.

b. The use of divergence data

The number of substitutions is usually computed at the codon level, distinguishing non-synonymous from synonymous substitutions, or an equivalent if non-coding DNA is used, by comparing the study species with at least one outgroup species. The outgroup sequences have to be selected with care. First, a closely-related outgroup species can potentially bias estimates of the rate of adaptive substitutions due to potentially shared polymorphisms. Second, a distantly-related outgroup species may lead to an underestimation of the divergence, and consequently of the rate of adaptive evolution, due to the possible presence of multiple "invisible" substitutions between the two species. One can potentially overcome this limitation by using multiple outgroup species, in order to span several levels of divergence and get more accurate estimates of the local substitution rate (Keightley and Jackson 2018). Moreover, if the divergence between the outgroup and the ingroup species is too high, we may suffer from the same bias as phylogenetic methods towards the more conserved genes, as fast evolving genes will not yield reliable sequence alignments. This would potentially underestimate the rate of adaptive substitutions by losing information on lineage-specific genes.

c. First likelihood models of DFE accounting for slightly deleterious mutations

The first likelihood model used to estimate the molecular rate of adaptive evolution was developed by Bierne and Eyre-Walker (2004) (Fig. 1). The authors developed an extension of the MK test allowing nonsynonymous mutations to be potentially strongly advantageous. This model assumes that, for a given gene, estimates of D_n, D_s, P_n and P_s are Poisson distributed and infers the number of adaptive amino-acid substitutions (η) and α by assuming that the selection parameters are either constant across all loci or that they follow a certain DFE, in this case, a Gamma or a Beta distribution (see Box 1). Welch (2006) extended the method developed by Bierne and Eyre-Walker (2004) by including models with a continuous distribution of selection coefficients and a two weighted spikes probability distribution of α, where α takes the value α_0 or α_1 with probabilities q and $1-q$ (Eqs. 4 and 8, respectively; Welch 2006). This likelihood framework has the advantage of enabling the comparison between nested models (Mangel and Hilborn 1996; Barton 2000): to test the occurrence of positive selection, we compare a model that potentially includes adaptive substitutions (η or $\alpha > 0$) with a neutral model (η or $\alpha = 0$) (Bierne and Eyre-Walker 2004; Welch 2006).

Further extensions of these methods model a deleterious DFE in the form of a Gamma distribution (Eyre-Walker et al. 2006; Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009). Each mutation arising at a site is ascribed a scaled selection coefficient, $4N_e s$, where the effective population size (N_e) is constant among loci, and s is drawn from an underlying DFE to be estimated from the data. Moreover, the SFS jointly estimates demographic parameters that allow for temporal changes in the effective population size (Eyre-Walker et al. 2006; Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009). These models come together in two of the most widely used inference methods: DoFE and dfe-alpha (Fig. 1, Table 1).

d. Extensions accounting for beneficial mutations

The fitness effect of new mutations is unlikely to be uniform within a given gene, but is rather expected to vary according to the sequence context and the nature of the functional changes that are incurred. It is, therefore, also important to consider the contribution of beneficial mutations to the SFS in addition to deleterious mutations. Some model-based inference methods account for mutations with positive effects in the DFE. Some of these

Table 1 Summary of population genetic methods that infer the rate of adaptive substitutions with sequence data

Reference(s)	Input Data	$N_e s$ distribution (DFE)	Model Inference	Method
Bierne and Eyre-Walker (2004)	polymorphism levels (P_n[a], P_s[b]); divergence data	Gamma; Beta	ML[c]	DoFE
Eyre-Walker et al. (2006) and Eyre-Walker and Keightley (2009)	folded SFS; divergence data	Gamma	ML[c]	
Stoletzki and Eyre-Walker (2011)				
Keightley and Eyre-Walker (2007) and Eyre-Walker and Keightley (2009)	folded SFS; divergence data	Gamma	ML[c]	DFE-alpha
Schneider et al. (2011)	unfolded SFS; divergence data	Gamma	ML[c]	
Welch (2006)	polymorphism levels (P_n[a], P_s[b]); divergence data	Continuous, two-spiked probability	ML[c]	MKTest
Galtier (2016)	unfolded/folded SFS; divergence data[d]	Gamma; GammaExponential; Displaced Gamma; FGMBesselK; SclaedBeta	ML[c]	Grapes
Tataru et al. (2017) and Tataru and Bataillon (2019)	unfolded SFS; divergence data (optional)	Gamma; Exponential; GammaExponential; Displaced Gamma; K bins	ML[c]	polyDFE
Messer and Petrov (2012)	unfolded SFS; divergence data	Exponential	ML[c]	αMK
Uricchio et al. (2019)	unfolded SFS; divergence data	Gamma; Continuous	ABC[e]	ABC-MK
Gronau et al. (2013)	unfolded SFS; divergence data	Categorical	ML[c]	INSIGHT

[a] P_n is the number of non-synonymous polymorphisms

[b] P_s is the number of synonymous polymorphisms

[c] ML corresponds to the maximum-likelihood approach

[d] Divergence data can be ignored if the unfolded SFS is used

[e] ABC corresponds to the approximate Bayesian computation

Box 1 The likelihood model of Bierne and Eyre-Walker (2004)

The method developed by Bierne and Eyre-Walker (2004) represents the first likelihood model that extends the MK test to estimate the rate of adaptive evolution. We further describe the parameters and the underlying assumptions of this model, which constitute the foundation for the methods developed hereafter

For a sample of n_i sequences at a locus i the expected numbers of synonymous polymorphisms $\left(\hat{P}_{si}\right)$ and substitutions $\left(\hat{D}_{si}\right)$ and numbers of nonsynonymous polymorphisms $\left(\hat{P}_{ni}\right)$ and substitutions $\left(\hat{D}_{ni}\right)$ are denoted as

	Polymorphisms	Substitutions
Synonymous	$\hat{P}_{si} = \Theta L_i$	$\hat{D}_{si} = \lambda_i L_i$
Non-synonymous	$\hat{P}_{ni} = \omega_i \Theta_i L_i$	$\hat{D}_{ni} = \omega_i \lambda_i + \eta_i L_i$
		$= \omega_i \lambda_i L_i / \left(1 - \alpha_i\right)^{(*)}$

By assuming that sites evolve independently (i.e. are in linkage equilibrium), this method uses a likelihood framework to model the data for n loci where observed data at each locus is summarized via the statistics $\left(\hat{P}_{si},\ \hat{P}_{ni},\ \hat{D}_{si}\right.$ and $\left.\hat{D}_{ni}\right)$ that are each Poisson distributed. This model has four parameters per locus and a maximum of 4n parameters. It is possible to reduce the number of parameters by assuming that, either some parameters are constant across loci, or selection parameters follow a certain probability density function, which constitutes the distribution of fitness effects. The authors evaluated different models where η and α are constant over all loci, or where η follows a gamma distribution and αis beta distributed

Θ_i = synonymous diversity (i.e. mean number of synonymous polymorphisms per codon); L_i = length of the sequence (i.e. number of codons); ω_i = nonsynonymous to synonymous diversity ratio $\left(= \hat{P}_{ni}/\hat{P}_{si}\right)$; $_i$ = synonymous substitution rate per codon; ω_{ii} = expected number of neutral nonsynonymous substitutions; η = expected number of adaptive nonsynonymous substitutions per codon; α = proportion of amino-acid substitutions that are adaptive; *Because α_i is denoted as $1 - \left(\hat{D}_{si}\hat{P}_{ni}/\hat{D}_{ni}\hat{P}_{si}\right)$ (Smith and Eyre-Walker 2002)

distributions are theoretically motivated by explicit fitness landscape models (see Bataillon and Bailey (2014) for a review of theoretically plausible distributions) while others are motivated by statistical convenience (to fit the data with a flexible distribution). An extension of the dfe-alpha method described above (Schneider et al. 2011) uses the unfolded SFS together with divergence data to model a Gamma DFE that also accounts for positively selected mutations (Table 1, Fig. 1). The Grapes method (Galtier 2016) can be used with both unfolded and folded SFS combined with divergence data (which is optional when the unfolded SFS is used) to model five different DFE, including the traditional Gamma distribution of deleterious mutations and four other models that account for mutations with beneficial effects (Table 1, Fig. 1). Galtier (2016) analyzed the performance of these models over 44 different datasets and observed that the GammaExponential model, which combines a Gamma distribution of deleterious mutations with an exponential distribution of beneficial mutations, and the ScaledBeta model, which uses a Beta-shaped distribution of slightly deleterious and advantageous mutations, were the ones with the best AIC scores, thus highlighting the important role of beneficial mutations in shaping the SFS. Using a similar framework, polyDFE (Tataru et al. 2017) infers the DFE from an unfolded SFS but does not require divergence data, thus allowing the estimation of the molecular adaptive rate on the branch of the study species. PolyDFE can model different DFE, including a model comprising a combination of gamma and exponential distributions to model mutations with negative and positive effects, respectively (Table 1, Fig. 1). At the level of noncoding DNA, INSIGHT (Gronau et al. 2013) contrasts the unfolded SFS and divergence in the non-coding elements of interest with those in flanking neutral sites. This method applies a generative probabilistic model by pooling data across non-coding elements considering the within-genome variation in mutation rates and coalescent times. INSIGHT models a categorical DFE, where each site is assumed to evolve under one of four different selective processes: neutral drift, strong negative selection, weak negative selection or positive selection (Table 1).

Despite their similarity, the methods above make slightly different assumptions when modeling polymorphism (SFS counts) and divergence (divergent sites relative to an outgroup). All methods assume a Poisson random field model and that the polymorphism data can be summarized by counts of the unfolded or folded SFS. Grapes, dfe-alpha and DoFE assume that the SFS is known without error, while polyDFE can model an independent rate of misorientation in the data, and INSIGHT uses a low dimensional projection of the SFS, by treating the ancestral allele as a hidden random variable in the model. Demography is either modeled via a set of nuisance parameters (Grapes, polyDFE) or assuming a fixed demographic model featuring a specific change of population size back in time that is also estimated (DFE-alpha, DoFE). Last but not least, most methods model a single SFS (synonymous versus non-synonymous) across genes, but a recent extension of polyDFE allows for fitting jointly several SFS datasets simultaneously (Tataru and Bataillon 2019). This can be used to determine whether distinct genomic regions and/or species share a common DFE, or provide evidence for differences in DFE among genomic regions/species.

e. aMK and ABC-MK models

The previously described methods assume that sites evolve independently. However, there has been growing evidence that selection at linked sites might be shaping genome-wide patterns of polymorphism (Barton 1995; Andolfatto 2007; Macpherson et al. 2007). Theoretical and empirical studies showed that, besides genetic drift and purifying selection, the frequency of a given allele can also be affected by recurrent selective sweeps at closely linked positions, a process known as genetic draft (see Glossary) (Gillespie 2000). Moreover, background selection (see Glossary) can also affect polymorphism levels at neutral sites if slightly deleterious mutations are segregating, creating interference at linked sites (Charlesworth et al. 1993; Bustamante et al. 2005; Keightley and Eyre-Walker 2007; Charlesworth 2012). Messer and Petrov (2012) developed an extension of the MK test that accounts for the effects of background selection and genetic draft on the levels of polymorphisms. They define $\alpha(x)$ as a function of the frequency of the derived mutation: $\alpha(x) = 1 - (d_0 \cdot p_{(x)})/d \cdot p_{0(x)}$, where $p_{(x)}$ and $p_{0(x)}$ represent the polymorphism levels at nonsynonymous and synonymous sites, for a specific derived allele frequency x. Here, any bias affecting the synonymous and nonsynonymous SFS, either demography or selection at linked sites, will be excluded, as $\alpha(x)$ only depends on the ratio $p_{(x)}/p_{0(x)}$. The asymptotic value of $\alpha(x)$ is then estimated in the limit $x \rightarrow 1$, where it should converge to the true value of α under the MK assumptions: in practice, this is done by fitting an exponential function to the data, given by: $\alpha(x) \approx \alpha + bexp(-cx)$. This function, however, assumes that all deleterious mutations have the same selection coefficient and that levels of nonsynonymous mutations decrease roughly exponentially with increasing frequency of neutral polymorphisms. Uricchio et al. (2019) extended this method by exploring the impact of background selection on the rate of adaptation using an approximate Bayesian computation (ABC) method, which the authors call ABC-MK (Table 1, Fig. 1). As in the αMK approach, this model is less sensitive to the demography of the population. Besides, it separately infers α for both weakly and strongly beneficial alleles, thus accounting for the strength of selection. To do so, ABC-MK assumes that deleterious mutations are gamma-distributed and allows α to follow a continuous distribution, from weakly to strongly beneficial mutations. As these models are less sensitive to the uncertainty associated with the demography of the population, they have the power to deliver more robust estimates of the molecular rate of adaptation on non-model organisms.

f. Statistics used to infer the rate of adaptive substitutions

From the above-described methods, three major statistics are often used to qualify the rate of adaptive non-synonymous substitutions: ω_a, α and K_{a+}. The rate of adaptive non-synonymous substitutions relative to the mutation rate, denoted as ω_a, is given by $\omega - \omega_{na}$, where ω_{na} represents the fraction of the ω ratio contributed by neutral and deleterious mutations. The proportion of positively selected amino-acid substitutions, α, is then estimated as ω_a/ω. Finally, K_{a+} represents the rate of adaptive amino-acid substitutions and is given by αK_a, where K_a is an alternative symbol of d_N, which is the number of non-synonymous substitutions per site. Each of these statistics has its limitations. For instance, α depends both on ω_a and ω_{na}, thus differences in α may be due to variations in any of the two rates or both, making it unsuitable for distinguishing the impact of negative and positive selection. On the other hand, ω_a is normalized by the mutation rate and, therefore, cannot be used to assess the impact of the mutation rate itself, which is an important varying factor along the genome. In this case, K_{a+} is more appropriate (Castellano et al. 2016).

Between-species variation in the molecular adaptive rate

Several studies investigated the prevalence of positive selection in the evolution of distinct species. Here, we provide a summary of their main conclusions.

a. Drosophila

Building on a long history of genetic studies, the Drosophila species complex was used in some of the pioneering research on adaptive evolution (Haudry et al. 2019). Brookfield and Sharp (1994) were the first to use the MK test to scan for signs of positive selection in Drosophila. They reported that three out of the seven genes analyzed had an excess of non-synonymous substitutions, thus suggesting that adaptive evolution was pervasive. By studying 35 genes, Smith and Eyre-Walker (2002) confirmed this hypothesis by reporting that ~45% of the amino-acid substitutions between *D. simulans* and *D. yakuba* were driven by positive selection. In the same year, Fay et al. (2002) estimated that ~70% of the amino-acid substitutions between *D. simulans* and *D. melanogaster* were adaptive. Further genome-wide studies also reported similar levels of adaptive evolution in the Drosophila genome (reviewed in Sella et al. 2009): 25 ± 20% (Bierne and Eyre-Walker 2004; Shapiro et al. 2007); 40 ± 10% (Welch 2006b); ~50% (Andolfatto 2007). Looking at the divergence between *D. pseudoobscura* and *D. affinis*, Haddrill et al. (2010) estimated even higher values of α, suggesting that 70–90% of the amino-acid substitutions differentiating the two species were driven by positive selection. By applying a Bayesian approach (Sawyer and Hartl 1992; Bustamante et al. 2001), Sawyer et al. (2003) estimated that ~94% of the substitutions were adaptive, although weakly selected ($N_e s \approx 5$, where s is the selection coefficient). It has been suggested, however, that these values of α could be overestimated if the current N_e is larger than the ancestral species (Eyre-Walker 2006; Rousselle et al. 2018). Nonetheless, analyses across the Drosophila genus led to similar estimates of α and, at least for *D. melanogaster*, the population size was inferred to have decreased (Akashi 1996; Haudry et al. 2019). Moreover, a recent study considering the past demography of the ancestral species found similar values of α to those previously reported in *D. melanogaster* (~49%, Zhen et al. 2018). These studies, therefore, provide evidence that positive selection may indeed be a prevalent mode of evolution in Drosophila genus.

b. Hominids

Alongside Drosophila, humans and apes have been focal species for studies of adaptive evolution. Fay et al. (2001) reported that ~35% of the fixed amino-acid differences between humans and old-world monkeys were positively selected. This study, however, had the shortcoming of using a very conserved set of polymorphisms, which can overestimate the rate of non-synonymous substitutions, and consequently α (Eyre-Walker 2006). Conversely, several studies proposed that the rate of adaptive evolution is almost zero in chimpanzees (Mikkelsen et al. 2005; Hvilsom et al. 2012; Castellano et al. 2019) and within hominids (Zhang and Li 2005; Boyko et al. 2008; Eyre-Walker and Keightley 2009), suggesting that only ~10% of the fixed differences between humans and chimpanzees are adaptive (Bustamante et al. 2005; Boyko et al. 2008). In turn, Enard et al. (2014) found genome-wide signals of positive selection in the human genome after correcting for the effects of background selection and suggested that adaptation in humans is mainly driven by regulatory rather than by coding differences. A recent study using an improved modeling of segregating weakly deleterious mutations and accounting for the demographic history of the ancestral species reported an α value around 20%, which is consistent when using the chimpanzee or the macaque as the outgroup species (Zhen et al. 2018). The authors argued that considering the same population size for the outgroup and ancestral species could bias estimations of α, especially in humans, where the human ancestral population is known to be much smaller than that of, for example, chimpanzees or macaques. We discuss in more detail these differences across studies in the last section of this topic (f).

c. Non-primate mammals

Halligan et al. (2010) reported that 57% of the amino-acid substitutions were adaptively driven in *Mus musculus castaneus*, a species of murid rodents. In two subspecies of the European rabbit, *Oryctolagus cuniculus algirus* and *O. c. cuniculus*, more than 60% of the amino-acid substitutions were found to be adaptive (Carneiro et al. 2012). Furthermore, a study performed on 44 non-model organisms, reported a mean value of α of around 50% in twelve mammal species (Galtier 2016).

d. Plants

Studies of plants led to a huge variation in the inferred rate of molecular adaptation across species. High rates of adaptive evolution have been measured for the grand shepherd's-purse (Slotte et al. 2010), the European aspen (Ingvarsson 2010) and species of sunflowers (Gossmann et al. 2010; Strasburg et al. 2011), where more than 30% of the amino-acid substitutions were estimated to be driven by positive selection. For the majority of plant species studied, though, α was observed to be close to zero (Gossmann et al. 2010). For example, in *Arabidopsis thaliana*, amino-acid substitutions are predominantly deleterious (Bustamante et al. 2002) with an average adaptive substitution rate very close to zero (Slotte et al. 2011). Authors proposed that this could be due to the *Arabidopsis* mating system, which by having a high frequency of inbreeding makes it harder to remove deleterious mutations (Bustamante et al. 2002). There are studies, however, reporting signs of adaptive evolution in the *Arabidopsis* genome. Barrier et al. (2003) found signs of positive selection in ~ 5% of the genes and Moutinho et al. (2019) showed that rates of adaptive evolution of sites at the surface of proteins are higher than the average across the genome, thus suggesting that some regions of the *Arabidopsis* genome are undergoing positive selection.

Slightly deleterious mutations were also observed to be prevalent in the genomes of *A. lyrata* (Barnaud et al. 2008; Foxe et al. 2008), *Sorghum bicolor* (Hamblin et al. 2006), and *Zea* species, (Bijlsma et al. 1986; Ross-Ibarra et al. 2009), thus suggesting very low rates of adaptive evolution also for these organisms. The reason behind such low rates of adaptive evolution in plant species is still unclear and further studies are needed to link plant adaptation at the ecological and molecular levels.

e. Other species

The rate of adaptive evolution was also studied in a wide range of other organisms. For yeast (Liti et al. 2009) and the giant Galapagos tortoise (Loire et al. 2013), α was observed to be close to zero. Conversely, studies on the sea squirt (Tsagkogeorga et al. 2012) and enterobacteria (Charlesworth and Eyre-Walker 2006) reported that ~ 50% of the amino-acid substitutions are adaptive. For viruses, a high rate of adaptive substitutions is also observed: Williamson (2003) suggested that ~ 50% of the substitutions in the env gene of HIV-1 were positively selected. By accounting for the distribution of d_N/d_S across codons, Nielsen and Yang (2003) inferred slightly higher rates of adaptive evolution (75%). Moreover, they reported an α of about 85% in the hemagglutinin gene of the human influenza virus.

f. What causes the across species variation of the rate of molecular adaptive evolution?

In the previous sections, we gave an overview of the wide range of data obtained across taxa, highlighting the great variation in the inferred rate of adaptive evolution across species (Fig. 2a). The factors determining this variability, however, remain unclear. Several studies have proposed that cross-species variation is explained by differences in effective population size (Eyre-Walker 2006; Eyre-Walker and Keightley 2009; Gossmann et al. 2012). According to this hypothesis, species with smaller N_e accumulate more weakly deleterious mutations simply by chance, thus increasing ω_{na} and consequently reducing estimates of α. Conversely, large-N_e species are under more efficient purifying selection, hence removing mutations with negative effects from the allele pool at a faster rate. By performing a study on 44 different species, Galtier (2016) confirmed this hypothesis by showing that N_e was positively correlated with α and ω_{na}, but not ω_a.

On the other hand, if the population size decreases, α can also be strongly underestimated due to segregating slightly deleterious mutations, which will remain within the population (Eyre-Walker and Keightley 2009; Zhen et al. 2018). Such a scenario was reported to be the cause of very low rates of adaptive evolution in the human genome (Zhen et al. 2018). By considering the demography of the ancestral population, Zhen et al. (2018)

revealed an α value of around 30%, higher than previous estimates for this species (Boyko et al. 2008; Eyre-Walker and Keightley 2009). Moreover, they found more strongly selected and/or more abundant advantageous mutations in humans when compared with mice and fruit flies. The authors proposed that these differences could reflect the number of traits under selection (Lourenço et al. 2013; Zhen et al. 2018). According to this hypothesis, larger long-lived organisms, such as humans, have less capacity to adapt to new environments, due to the greater number of traits under selection. Such organisms are theoretically expected to need more consecutive beneficial mutations to reach their fitness optimum, and thus a higher proportion of beneficial mutations should be accordingly detected in these

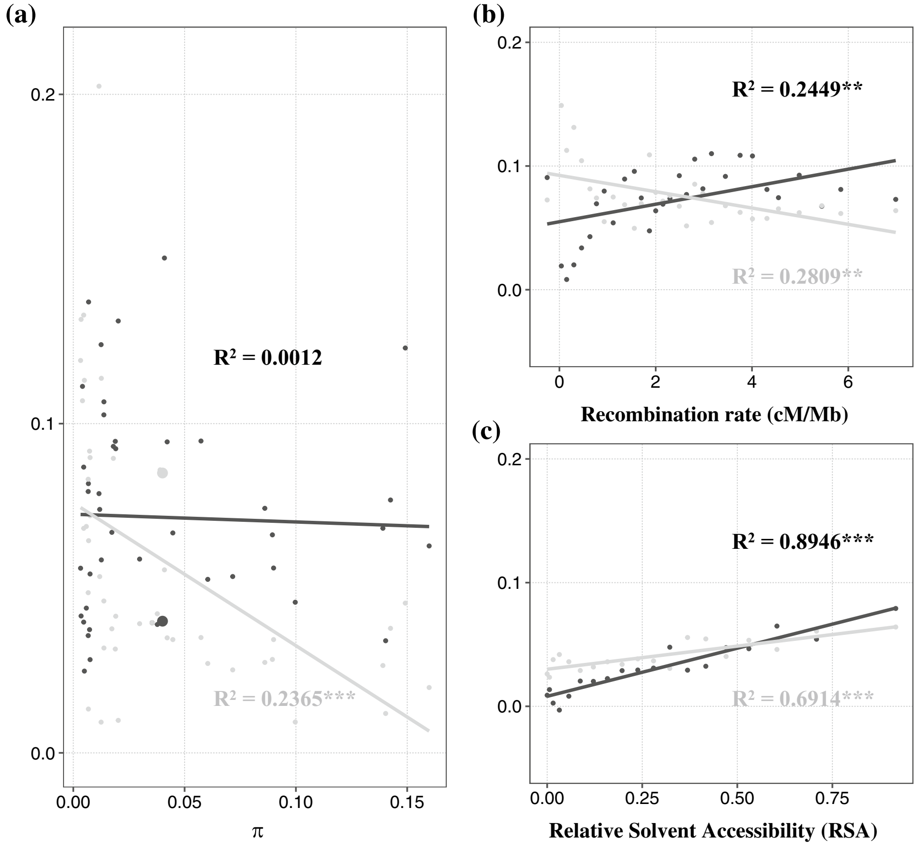

Fig. 2 Variation of the rate of adaptive non-synonymous substitutions (ω_a; in black) and the rate of non-adaptive non-synonymous substitutions (ω_{na}; in grey) between species (**a**), within genomes (**b**) and within genes (**c**). The R^2 Pearson's correlation coefficient is given along with significance denoted by asterisks (***P* value < 0.01, ****P*-value < 0.001). **a** Relationship between ω_a and ω_{na} with the level of species nucleotide diversity (π), used as a proxy for effective population size, obtained from Galtier (2016). Each sample point represents one species. Dots with bigger sizes correspond to *D. melanogaster* (data from Moutinho et al. 2019), which is the focus species of plots (**b**) and (**c**). **b** Relationship between ω_a and ω_{na} with the recombination rate in cM/Mb, taken from Moutinho et al. (2019). Each dot represents the mean value of ω_a or ω_{na} for each recombination rate class. **c** Relationship between ω_a and ω_{na} with the relative solvent accessibility (RSA), obtained from Moutinho et al. (2019). Each dot represents the mean value of ω_a or ω_{na} for each RSA class

species (Lourenço et al. 2013; Rousselle et al. 2018, 2019b). More studies are needed to clarify what is causing the observed differences between species.

Within-genome variation of the molecular rate adaptation

Several studies provided evidence for a substantial variation in the rate of adaptive substitutions along the genome. In this section, we summarize the factors that were found to influence the distribution of adaptive substitutions within species (Fig. 1).

a. Genome-wide variables

At the genome level, recombination, mutation and gene density are important determinants of the rate of adaptive substitutions (ω_a) (Marais and Charlesworth 2003; Campos et al. 2014; Castellano et al. 2016). Recombination rate is predicted to favor the fixation of adaptive substitutions (Fig. 2b) by breaking down linkage disequilibrium (Marais and Charlesworth 2003; Campos et al. 2014; Castellano et al. 2016). Advantageous mutations occurring at linked sites but in distinct individuals will interfere, so that only one will ultimately reach fixation unless a recombination event creates a haplotype carrying both of them (Hill-Robertson interference, HRi; Hill and Robertson 1966; Felsenstein 1974). As a result, genes in low recombining regions are expected to have overall lower rates of adaptive substitutions. Following a similar rationale, genes present in regions with high gene density may be subject to stronger HRi and slow rates of adaptive evolution (Castellano et al. 2016). In turn, genes with high mutation rates potentially adapt faster because they increase the levels of genetic diversity, which, consequently, increases the chance of selection operating such that adaptive processes may occur. Interestingly, Castellano et al. (2016) found that the positive correlation between mutation rate and the rate of adaptive substitutions no longer holds for genes located in regions with low recombination rate and high gene density, thus suggesting a strong effect of HRi in the presence of a large number of selected mutations with a small genetic distance between them. Similarly, Gossmann et al. (2011) observed that variations in N_e resulting from linked selection along the genome significantly impact the efficiency of natural selection in *C. grandiflora* and *A. thaliana*, where regions with larger N_e are subject to stronger purifying selection.

b. Protein-coding: gene-wide variables

On a gene-wide scale, it has been reported that protein function strongly influences the rate of adaptive evolution, with genes involved in the immune response presenting the highest rates of adaptation in *Drosophila* (Sackton et al. 2007; Obbard et al. 2009), *Arabidopsis* (Slotte et al. 2011), hominids (Nielsen et al. 2005; Kosiol et al. 2008) and other mammals (Kosiol et al. 2008). Sex-related genes were also reported to present higher rates of adaptive evolution in *Drosophila* (Pröschel et al. 2006; Haerty et al. 2007) chimpanzees (Hvilsom et al. 2012) and in plants (Gossmann et al. 2014; Crowson et al. 2017). Moreover, a recent study showed that genes involved in protein biosynthesis and signaling for protein degradation exhibit the highest rates of adaptive substitutions in *Drosophila* and *Arabidopsis* (Moutinho et al. 2019). Cytochrome P450 proteins, which are involved in defense response in plants, were also characterized by high rates of adaptation in *Arabidopsis* (Moutinho et al. 2019). Several studies have described that host–pathogen interactions act as key drivers of protein evolution in several taxa (Sackton et al. 2007; Obbard et al. 2009; Enard et al. 2016; Ebel et al. 2017; Mauch-Mani et al. 2017; Uricchio et al. 2019; Grandaubert et al. 2019), which could explain the observed high levels of adaptive evolution in the functions described above. Moreover, mean gene expression levels and the breadth of expression negatively impact the rate of adaptive evolution in *Drosophila*, where the two factors may be acting together (Duret and Mouchiroud 2000; Salvador-Martínez et al. 2018; Moutinho et al. 2019). This relationship with expression may be a consequence of stronger purifying selection in highly expressed genes, where selection acts by favoring proteins with the lowest probability of misfolding, which occurs if the protein sequence accumulates translational missense errors (Drummond et al. 2005). Additionally, the macromolecular structure of the protein was also observed to substantially impact the rate of protein adaptation in humans (Afanasyeva et al. 2018), *Drosophila* and *Arabidopsis* (Moutinho et al. 2019). In this case, proteins with a higher proportion of disordered regions (Afanasyeva et al. 2018; Moutinho et al. 2019) and/or exposed residues (Moutinho et al. 2019) are prone to accumulate more adaptive mutations, acting as important targets of positive selection.

c. Protein-coding: intra-molecular factors

There is growing evidence that adaptive substitution rates also vary significantly at the intra-genic level. Studies both at the population and divergence level, have shown that the relative solvent accessibility (RSA) significantly impacts the rate of amino-acids substitutions (Fig. 2c), with exposed residues accumulating more adaptive mutations than buried ones (Goldman et al. 1998; Mirny and Shakhnovich 1999; Franzosa and Xia 2009; Liberles et al. 2012; Moutinho et al. 2019). When contrasted with the effect of residue intrinsic disorder, RSA was observed to contribute with most of the variation in ω_a (95% and 87% of variance explained for *A. thaliana* and *D. melanogaster*, respectively; Moutinho et al. 2019). This suggests that solvent exposure is the main determinant of adaptive evolution at the level of protein structure, and that protein intrinsic disorder contributes with a mere additive small effect to the rate of protein adaptation (Moutinho et al. 2019). Furthermore, the type of amino-acid mutation was also reported to be an important factor affecting the rate of adaptive evolution, with more similar amino-acid changes presenting higher rates of adaptive substitutions (Grantham 1974; Miyata et al. 1979; Bergman and Eyre-Walker 2019).

d. Non-coding DNA

While much attention has been given to the study of the adaptive evolution of protein-coding genes, there is increasing evidence that the non-coding regions of the genome are also key targets of positive selection. By using an MK-like approach, contrasting numbers of polymorphisms and substitutions at protein-binding and non-binding sites, Jenkins et al. (1995) reported signatures of adaptive change in the control for gene expression in *D. melanogaster*. Kohn et al. (2004) estimated that ~ 50% of all substitutions in the 5′ region of eight *Drosophila* genes were adaptively driven. By extending these approaches, Andolfatto (2005) investigated patterns of molecular evolution in multiple classes of non-coding DNA in *D. melanogaster* and found that around 60% and 20% of the total nucleotide divergence with *D. simulans* were fixed by positive selection, in UTRs and intronic/intergenic regions respectively. These findings suggest that the noncoding regions of the *D. melanogaster* genome are key determinants of adaptive evolution. Likewise, Haddrill et al. (2008) found signs of adaptive evolution in the non-coding regions of the *D. simulans* genome. These patterns go beyond the *Drosophila* genus since there is evidence of widespread positive selection in noncoding conserved regions along the Brassicaceae phylogeny (Williamson et al. 2014). In hominids, however, the opposite pattern is observed. Keightley et al. (2005) analyzed the downstream and upstream regions of protein-coding genes using an MK approach and found no signs of adaptive evolution. This result might reflect the overall low levels of adaptive evolution in hominid genomes due to the lower effective population sizes. With the thrive of full genome sequence data, adaptive evolution can now be more extensively studied outside the coding regions (Gronau et al. 2013), which, until now, were the focus of most studies.

Current limitations and future perspectives

In the last two decades, numerous methods have been developed to detect and quantify adaptive evolution. This, together with the availability of datasets spanning many genes and species, increased our knowledge of the factors underlying the heterogeneity of rates of molecular adaptation within genomes and between species. However, existing methods rely on several assumptions that can create biases in the estimates of adaptive evolution when not met. For instance, the methods reviewed here assume that synonymous mutations are neutral, which may not always be a valid approximation, especially in species with large effective population sizes (Lawrie et al. 2013). Several studies have documented that selection for codon usage also affects the rate of synonymous substitutions in several species, including *Drosophila* (Akashi 1994; Comeron et al. 1999), the European aspen (Ingvarsson 2010) and non-model animals (Galtier et al. 2018), mammals and birds (Rousselle et al. 2019a). Finding a proper neutral reference remains a challenging goal. Yet, a similar approach to that used in codon models (Yang and Nielsen 2008; Spielman and Wilke 2016; Rodrigue and Lartillot 2017) could, in principle, be considered for methods inferring the rate of adaptive evolution by accounting for the evolution of synonymous sites. This would lead to a more realistic null model of neutral evolution and, consequently, less biased estimates of the molecular rate of adaptation (Rodrigue and Lartillot 2017).

Another challenge consists of better accounting for the confounding effects of demography. Some methods fit a simplified demographic model (DFE-alpha, DoFE) while others correct for demography by adding extra parameters, one per frequency category of the SFS (Grapes, polyDFE). The number of such parameters, therefore, increases with the sample size and can quickly lead to model overparameterization issues. Extending the methods to use a continuous SFS constitutes one perspective to accommodate increasingly larger datasets. Alternatively, the demography of the population could also be estimated from the currently available coalescent methods (i.e. the SMC++, Terhorst et al. 2017; or ∂a∂i, Gutenkunst et al. 2009).

Besides, current models often assume a constant DFE across the whole genome. This can lead to a bad model fit because selection varies within and between genomic regions. Such an assumption can be relaxed by allowing DFE parameters to vary along the genome. Moreover, the use of an outgroup species to infer the ancestral allele (polymorphism orientation) can lead to biases in the estimates of adaptive evolution, whether the outgroup is a very closely-related species or a very distantly-related one (Hernandez et al. 2007). This can be alleviated by using multiple outgroup species and probabilistic ancestral allele reconstructions (e.g. Keightley and Jackson 2018). Furthermore, by using only one outgroup sequence, these methods are estimating divergence on the total branch separating the focal and the outgroup species. Using a second outgroup species and a phylogenetic approach, however, would allow restricting the estimation of the divergence parameters to the branch of the study species.

Furthermore, these methods assume that all sites are equally sampled in all individuals and do not intrinsically account for the possibility of missing data. Pre-processing of the data is therefore required, which can introduce biases if too many sites have to be discarded. Finally, methods relying on patterns of polymorphism cannot track positively selected mutations of individual sites, limiting the power of these analyses in detecting positive selection at the site level. Combing such population genetics approaches with mutation accumulation experiments is a promising avenue to further understand the fitness effect of particular mutations. This, however, would have to be done across several generations so that enough mutations could be generated.

Conclusions

The development of statistical approaches based on the pioneering work of McDonald and Kreitman (1991), together with the increasing availability of genome sequences at the population level, paved the way for the qualitative and quantitative assessment of rates of adaptive evolution, both between species and within genomes. Growing evidence suggests a substantial variation of the molecular adaptive rate at distinct levels of molecular evolution, emphasizing the multitude of factors that can influence the rate of adaptation. These studies introduced a conceptual and theoretical framework that, we posit, will serve as a basis for increasingly realistic models that will strengthen our understanding of the fitness effect of new mutations and, therefore, the molecular basis of adaptation.

Acknowledgements Open access funding provided by Max Planck Society. The authors thank Adam Eyre-Walker for fruitful discussions. JYD acknowledges funding from the Max Planck Society.

Complaince with ethical standards

Conflict of interest No conflict of interest is to be declared.

References

Afanasyeva A, Bockwoldt M, Cooney CR et al (2018) Human long intrinsically disordered protein regions are frequent targets of positive selection. Genome Res 28(7):975–982

Akashi H (1994) Synonymous codon usage in drosophila melanogaster: natural selection and translational accuracy. Genet Soc Am 136:927–935

Akashi H (1996) Molecular evolution between drosophila melanogaster and D. sirnulam reduced codon bias, faster rates of amino acid substitution, and larger proteins in D. melarwgaster. DNA Seq 144:1297–1307

Andolfatto P (2005) Adaptive evolution of non-coding DNA in Drosophila. Nature 437(7062):1149–1152

Andolfatto P (2007) Hitchhiking effects of recurrent beneficial amino acid substitutions in the. Genome Res 17(12):1755–1762

Barnaud A, Trigueros G, McKey D, Joly HI (2008) High outcrossing rates in fields with mixed sorghum landraces: how are landraces maintained? Heredity 101(5):445–452

Barrier M, Bustamante CD, Yu J, Purugganan MD (2003) Selection on rapidly evolving proteins in the arabidopsis genome. Genetics 163(2):723–733

Barroso GV, Moutinho AF, Dutheil JY (2019) A population genetics lexicon. In: Dutheil JY (ed) Statistical population genomics. Springer, Berlin

Barton NH (1995) Linkage and the limits to natural selection. Genetics 140(2):821–841

Barton NH (2000) Estimating linkage disequilibria. Heredity 84(2):373–389

Bataillon T (2003) Shaking the “deleterious mutations” dogma? Trends Ecol Evol 18(7):315–317

Bataillon T, Bailey SF (2014) Effects of new mutations on fitness: insights from models and data. Ann N Y Acad Sci 1320(1):76–92

Bergman J, Eyre-Walker A (2019) Does adaptive protein evolution proceed by large or small steps at the amino acid level? Mol Biol Evol 36(5):990–998

Bierne N, Eyre-Walker A (2004) The genomic rate of adaptive amino acid substitution in Drosophila. Mol Biol Evol 21(7):1350–1360

Bijlsma R, Allard RW, Kahler AL (1986) Non random mating in an open-pollinated maize population. Genetics 112(3):669–680

Boyko AR, Williamson SH, Indap AR et al (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet 4(5):e1000083

Brookfield JF, Sharp PM (1994) Neutralism and selectionism face up to DNA data. Trends Genet 10(4):109–111

Bustamante CD, Wakeley J, Sawyer S, Hartl DL (2001) Directional selection and the site-frequency spectrum. Genetics 159(4):1779–1788

Bustamante CD, Nielsen R, Sawyer SA et al (2002) The cost of inbreeding in arabidopsis. Nature 416(6880):531–534

Bustamante CD, Fledel-Alon A, Williamson S et al (2005) Natural selection on protein-coding genes in the human genome. Nature 437(7062):1153–1157

Campos JL, Halligan DL, Haddrill PR, Charlesworth B (2014) The relation between recombination rate and patterns of molecular evolution and variation in drosophila melanogaster. Mol Biol Evol 31(4):1010–1028

Carneiro M, Albert FW, Melo-Ferreira J et al (2012) Evidence for widespread positive and purifying selection across the european rabbit (oryctolagus cuniculus) genome. Mol Biol Evol 29(7):1837–1849

Castellano D, Coronado-Zamora M, Campos JL et al (2016) Adaptive evolution is substantially impeded by hill-Robertson interference in drosophila. Mol Biol Evol 33(2):442–455

Castellano D, Macià MC, Tataru P et al (2019) Comparison of the full distribution of fitness effects of new amino acid mutations across great apes. Genetics 213:696971

Charlesworth B (1994) The effect of background selection against deleterious mutations on weakly selected, linked variants. Genet Res 63(3):213–227

Charlesworth B (2012) The effects of deleterious mutations on evolution at linked sites. Genetics 190(1):5–22

Charlesworth B, Charlesworth D (2010) Elements of evolutionary genetics. Roberts and Company Publishers, Englewood

Charlesworth J, Eyre-Walker A (2006) The rate of adaptive evolution in enteric bacteria. Mol Biol Evol 23(7):1348–1356

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134(4):289–303

Comeron JM, Kreitman M, Aguadé M (1999) Natural selection on synonymous sites is correlated with gene length and recombination in drosophila. Genetics 151(1):239–249

Crowson D, Barrett SCH, Wright SI (2017) Purifying and positive selection influence patterns of gene loss and gene expression in the evolution of a plant sex chromosome system. Mol Biol Evol 34(5):1140–1154

Dong S, Stam R, Cano LM et al (2014) Effector specialization in a lineage of the Irish potato famine pathogen. Science 343(6170):552–555

Drummond DA, Bloom JD, Adami C et al (2005) Why highly expressed proteins evolve slowly. Proc Natl Acad Sci 102(40):14338–14343

Duret L, Mouchiroud D (2000) Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol 17(1):68–74

Ebel ER, Telis N, Venkataram S et al (2017) High rate of adaptation of mammalian proteins that interact with Plasmodium and related parasites. PLoS Genet 13(9):e1007023

Enard D, Messer PW, Petrov DA (2014) Genome-wide signals of positive selection in human evolution. Genome Res 24(6):885–895

Enard D, Cai L, Gwennap C, Petrov DA (2016) Viruses are a dominant driver of protein adaptation in mammals. Elife. 5:e12469

Eyre-Walker A (2002) Changing effective population size and the McDonald-Kreitman test. Genetics 162(4):2017–2024

Eyre-Walker A (2006) The genomic rate of adaptive evolution. Trends Ecol Evol 21:569–575

Eyre-Walker A, Keightley PD (2009) Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol 26(9):2097–2108

Eyre-Walker A, Woolfit M, Phelps T (2006) The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics 173(2):891–900

Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. Genetics 158:1227–1234

Fay J, Wyckoff G, Wu C (2002) Testing the neutral theory of molecular evolution with genomic data from Drosophila. Nature 415(6875):1024–1026

Foxe JP, Dar VUN, Zheng H et al (2008) Selection on amino acid substitutions in Arabidopsis. Mol Biol Evol 25(7):1375–1383

Franzosa EA, Xia Y (2009) Structural determinants of protein evolution are context-sensitive at the residue level. Mol Biol Evol 26(10):2387–2395

Galtier N (2016) Adaptive protein evolution in animals and the effective population size hypothesis. PLoS Genet 12(1):e1005774

Galtier N, Roux C, Rousselle M et al (2018) Codon usage bias in animals: disentangling the effects of natural selection, effective population size, and GC-biased gene conversion. Mol Biol Evol 35(5):1092–1103

Gillespie JH (2000) Genetic drift in infinite populations: the pseudohitchhiking model. Genetics 155:909–919

Goldman N, Thorne JL, Jones DT (1998) Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics 149:445–458

Gossmann TI, Song BH, Windsor AJ et al (2010) Genome wide analyses reveal little evidence for adaptive evolution in many plant species. Mol Biol Evol 27(8):1822–1832

Gossmann TI, Woolfit M, Eyre-Walker A (2011) Quantifying the variation in the effective population size within a genome. Genetics 189(4):1389–1402

Gossmann TI, Keightley PD, Eyre-Walker A (2012) The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. Genome Biol Evol 4(5):658–667

Gossmann TI, Schmid MW, Grossniklaus U, Schmid KJ (2014) Selection-driven evolution of sex-biased genes is consistent with sexual selection in arabidopsis thaliana. Mol Biol Evol 31(3):574–583

Grandaubert J, Dutheil JY, Stukenbrock EH (2019) The genomic determinants of adaptive evolution in a fungal pathogen. Evolution Letters 3(3):299–312

Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185(4154):862–864

Gronau I, Arbiza L, Mohammed J, Siepel A (2013) Inference of natural selection from interspersed genomic elements based on polymorphism and divergence. Mol Biol Evol 30(5):1159–1171

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5:e1000695

Haddrill PR, Bachtrog D, Andolfatto P (2008) Positive and negative selection on noncoding DNA in Drosophila simulans. Mol Biol Evol 25(9):1825–1834

Haddrill PR, Loewe L, Charlesworth B (2010) Estimating the parameters of selection on nonsynonymous mutations in Drosophila pseudoobscura and D. miranda. Genetics 185(4):1381–1396

Haerty W, Jagadeeshan S, Kulathinal RJ et al (2007) Evolution in the fast lane: rapidly evolving sex-related genes in Drosophila. Genetics 177(3):1321–1335

Halligan DL, Oliver F, Eyre-Walker A et al (2010) Evidence for pervasive adaptive protein evolution in wild mice. PLoS Genet 6(1):e1000825

Hamblin MT, Casa AM, Sun H et al (2006) Challenges of detecting directional selection after a bottleneck: lessons from Sorghum bicolor. Genetics 173(2):953–964

Haudry A, Laurent S, Kapun M (2019) Statistical population genomics of fruit flies. In: Dutheil JY (ed) Statistical population genomics. Springer, Berlin

Hernandez RD, Williamson SH, Bustamante CD (2007) Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol Biol Evol 24(8):1792–1800

Hey J (1999) The neutralist, the fly and the selectionist. Trends Ecol Evol 14:35–38

Hill WG, Robertson A (1966) The effect of linkage on limits to artificial selection. Genet Res 8(3):269–294

Hilson P, Allemeersch J, Altmann T et al (2004) Versatile gene-specific sequence tags for Arabidopsis functional genomics: transcript profiling and reverse genetics applications. Genome Res 14:2176–2189

Hoekstra HE, Hirschmann RJ, Bundey RA et al (2006) A single amino acid mutation contributes to adaptive beach mouse color pattern. Science 313(5783):101–104

Hudson RR, Kreitman M, Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116:153–159

Hvilsom C, Qian Y, Bataillon T et al (2012) Extensive X-linked adaptive evolution in central chimpanzees. Proc Natl Acad Sci 109(6):2054–2059

Ingvarsson PK (2010) Natural selection on synonymous and nonsynonymous mutations shapes patterns of polymorphism in populus tremula. Mol Biol Evol 27(3):650–660

Jenkins DL, Ortori CA, Brookfield JFY (1995) A test for adaptive change in DNA sequences controlling transcription. Proc R Soc B Biol Sci 261(1361):203–207

Jensen JD, Payseur BA, Stephan W et al (2019) The importance of the neutral theory in 1968 and 50 years on: a response to Kern and Hahn 2018. Evolution 73(1):111–114

Jones MR, Mills LS, Alves PC et al (2018) Adaptive introgression underlies polymorphic seasonal camouflage in snowshoe hares. Science 360(6395):1355–1358

Keightley PD, Eyre-Walker A (2007) Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177(4):2251–2261

Keightley PD, Jackson BC (2018) Use of transgene-induced rnai to regulate endogenous gene expression. Genetics 209:897–906

Keightley PD, Lercher MJ, Eyre-Walker A (2005) Evidence for widespread degradation of gene control regions in hominid genomes. PLoS Biol 3(2):0282–0288

Kern AD, Hahn MW (2018) The neutral theory in light of natural selection. Mol Biol Evol 35:1366–1371

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626

Kohn MH, Fang S, Wu CI (2004) Inference of positive and negative selection on the 5′ regulatory regions of drosophila genes. Mol Biol Evol 21(2):374–383

Kosakovsky Pond SL, Murrell B, Fourment M et al (2011) A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol 28(11):3033–3043

Kosiol C, Vinař T, Da Fonseca RR et al (2008) Patterns of positive selection in six mammalian genomes. PLoS Genet 4(8):e1000144

Lawrie DS, Messer PW, Hershberg R, Petrov DA (2013) Strong purifying selection at synonymous sites in D. melanogaster. PLoS Genet 9:33–40

Li W-H, Wu C-I, Luo C-C (1985) A new method for extimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon usage. Mol Biol Evol 2(2):150–174

Liberles DA, Teichmann SA, Bahar I et al (2012) The interface of protein structure, protein biophysics, and molecular evolution. Protein Sci 21:769–785

Liti G, Carter DM, Moses AM et al (2009) Population genomics of domestic and wild yeasts. Nature 458(7236):337–341

Loire E, Chiari Y, Bernard A et al (2013) Population genomics of the endangered giant Galápagos tortoise. Genome Biol 14(12):R136

Lourenço JM, Glémin S, Galtier N (2013) The rate of molecular adaptation in a changing environment. Mol Biol Evol 30(6):1292–1301

Macpherson JM, Sella G, Davis JC, Petrov DA (2007) Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in Drosophila. Genetics 177(4):2083–2099

Marais G, Charlesworth B (2003) Genome evolution: recombination speeds up adaptive evolution. Curr Biol 13(2):68–70

Mauch-Mani B, Baccelli I, Luna E, Flors V (2017) Defense priming: an adaptive part of induced resistance. Annu Rev Plant Biol 68(1):485–512

McDonald J, Kreitman M (1991) Adaptive evolution at the Adh locus in Drosophila. Nature 351:652–654

Messer PW, Petrov DA (2012) Frequent adaptation and the McDonald-Kreitman Test. Proc Natl Acad Sci 110(21):8615–8620

Mikkelsen TS, Hillier LW, Eichler EE et al (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437(7055):69–87

Mirny LA, Shakhnovich EI (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol 291:177–196

Miyata T, Miyazawa S, Yasunaga T (1979) Two types of amino acid substitutions in protein evolution. J Mol Evol 12:219–236

Moutinho AF, Trancoso FF, Dutheil JY (2019) The impact of protein architecture on adaptive evolution. Mol Biol Evol 36(9):2013–2028

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936

Nielsen R, Yang Z (2003) Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. Mol Biol Evol 20(8):1231–1239

Nielsen R, Bustamante C, Clark AG et al (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3(6):0976–0985

Obbard DJ, Welch JJ, Kim KW, Jiggins FM (2009) Quantifying adaptive evolution in the Drosophila immune system. PLoS Genet 5(10):e1000698

Ohta T (1992) The nearly neutral theory of molecular evolution. Annu Rev Ecol Syst 23:263–286

Ometto L, Glinka S, De Lorenzo D, Stephan W (2005) Inferring the effects of demography and selection on Drosophila melanogaster populations from a chromosome-wide scan of DNA variation. Mol Biol Evol 22:2119–2130

Pröschel M, Zhang Z, Parsch J (2006) Widespread adaptive evolution of Drosophila genes with sex-biased expression. Genetics 174(2):893–900

Racimo F, Schraiber JG (2014) Approximation to the distribution of fitness effects across functional categories in human segregating polymorphisms. PLoS Genet 10(11):e1004697

Rodrigue N, Lartillot N (2017) Detecting adaptation in protein-coding genes using a Bayesian site-heterogeneous mutation-selection codon substitution model. Mol Biol Evol 34(1):204–214

Ross-Ibarra J, Tenaillon M, Gaut BS (2009) Historical divergence and gene flow in the genus Zea. Genetics 181(4):1399–1413

Rousselle M, Mollion M, Nabholz B et al (2018) Overestimation of the adaptive substitution rate in fluctuating populations. Biol Lett 14(5):20180055

Rousselle M, Laverré A, Figuet E et al (2019a) Influence of recombination and GC-biased gene conversion on the adaptive and nonadaptive substitution rate in mammals versus birds. Mol Biol Evol 36:458–471

Rousselle M, Simion P, Tilak M-K et al (2019b) Is adaptation limited by mutation? A timescale dependent effect of genetic diversity on the adaptive substitution rate in animals. bioRxiv 64:3619

Roux J, Privman E, Moretti S et al (2014) Patterns of positive selection in seven ant genomes. Mol Biol Evol 31:1661–1685

Rutter MT, Roles A, Conner JK et al (2012) Fitness of Arabidopsis thaliana mutation accumulation lines whose spontaneous mutations are known. Evolution 66(7):2335–2339

Sackton TB, Lazzaro BP, Schlenke TA et al (2007) Dynamic evolution of the innate immune system in Drosophila. Nat Genet 39(12):1461–1468

Salvador-Martínez I, Coronado-Zamora M, Castellano D et al (2018) Mapping selection within drosophila melanogaster Embryo's Anatomy. Mol Biol Evol 35(1):66–79

Sawyer S, Hartl D (1992) Population genetics of polymorphism and divergence. Genetics 132:1161–1176

Sawyer SA, Kulathinal RJ, Bustamante CD, Hartl DL (2003) Bayesian analysis suggests that most amino acid replacements in drosophila are driven by positive selection. J Mol Evol 57:154–164

Schneider A, Charlesworth B, Eyre-Walker A, Keightley PD (2011) A method for inferring the rate of occurrence and fitness effects of advantageous mutations. Genetics 189:1427–1437

Schraiber JG, Akey JM (2015) Methods and models for unravelling human evolutionary history. Nat Rev Genet 16(12):727–740

Sella G, Petrov DA, Przeworski M, Andolfatto P (2009) Pervasive natural selection in the Drosophila genome? PLoS Genet 5(6):e1000495

Shapiro JA, Huang W, Zhang C et al (2007) Adaptive genic evolution in the Drosophila genomes. Proc Natl Acad Sci 104(7):2271–2276

Shaw FH, Geyer CJ, Shaw RG (2002) A comprehensive model of mutations affecting fitness and inferences for Arabidopsis thaliana. Evolution 56:453–463

Slotte T, Foxe JP, Hazzouri KM, Wright SI (2010) Genome-wide evidence for efficient positive and purifying selection in capsella grandiflora, a plant species with a large effective population size. Mol Biol Evol 27(8):1813–1821

Slotte T, Bataillon T, Hansen TT et al (2011) Genomic determinants of protein evolution and polymorphism in arabidopsis. Genome Biol Evol 3:1210–1219

Smith NGC, Eyre-Walker A (2002) Adaptive protein evolution in Drosophila. Nature 415:1022–1024

Spielman SJ, Wilke CO (2016) Extensively parameterized mutation-selection models reliably capture site-specific selective constraint. Mol Biol Evol 33(11):2990–3001

Stoletzki N, Eyre-Walker A (2011) Estimation of the neutrality index. Mol Biol Evol 28:63–70

Strasburg JL, Kane NC, Raduski AR et al (2011) Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. Mol Biol Evol 28(5):1569–1580

Stukenbrock EH, Bataillon T, Dutheil JY et al (2011) The making of a new pathogen: insights from comparative population genomics of the domesticated wheat pathogen Mycosphaerella graminicola and its wild sister species. Genome Res 21(12):2157–2166

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Tataru P, Bataillon T (2019) PolyDFEv2.0: testing for invariance of the distribution of fitness effects within and across species. Bioinformatics 3:1–2

Tataru P, Mollion M, Glémin S, Bataillon T (2017) Inference of distribution of fitness effects and. Genetics 207:1103–1119

Terhorst J, Kamm JA, Song YS (2017) Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat Genet 49:303–309

Tsagkogeorga G, Cahais V, Galtier N (2012) The population genomics of a fast evolver: high levels of diversity, functional constraint, andmolecular adaptation in the tunicate Ciona intestinalis. Genome Biol Evol 4:740–749

Uricchio LH, Petrov DA, Enard D (2019) Exploiting selection at linked sites to infer the rate and strength of adaptation. Nat Ecol Evol. 3(6):977

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4:0446–0458

Welch JJ (2006) Estimating the genomewide rate of adaptive protein evolution in drosophila. Genetics 173(2):821–837

Williamson S (2003) Adaptation in the env gene of HIV-1 and evolutionary theories of disease progression. Mol Biol Evol 20(8):1318–1325

Williamson RJ, Josephs EB, Platts AE et al (2014) Evidence for widespread positive and negative selection in coding and conserved noncoding Regions of Capsella grandiflora. PLoS Genet 10(9):e1004622

Yang Z, Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol 46(4):409–418

Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19(2):908–917

Yang Z, Nielsen R (2008) Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol 25(3):568–579

Yang Z, Nielsen R, Goldman N, Pedersen AK (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155(1):431–449

Yang Z, Wong WSW, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol 22(4):1107–1118

Zhang L, Li WH (2005) Human SNPs reveal no evidence of frequent positive selection. Mol Biol Evol 22(12):2504–2507

Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22(12):2472–2479

Zhen Y, Huber CD, Davies RW, Lohmueller KE (2018) Stronger and higher proportion of beneficial amino acid changing mutations in humans compared to mice and flies. https://doi.org/10.1101/427583

The Impact of Protein Architecture on Adaptive Evolution

Ana Filipa Moutinho,*[,1] Fernanda Fontes Trancoso,[1] and Julien Yann Dutheil[1,2]

[1]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Biology, Plön, Germany

[2]Unité Mixte de Recherche 5554 Institut des Sciences de l'Evolution, CNRS, IRD, EPHE, Université de Montpellier, Montpellier, France

***Corresponding author:** E-mail: moutinho@evolbio.mpg.de.

Associate editor: Jianzhi Zhang

Abstract

Adaptive mutations play an important role in molecular evolution. However, the frequency and nature of these mutations at the intramolecular level are poorly understood. To address this, we analyzed the impact of protein architecture on the rate of adaptive substitutions, aiming to understand how protein biophysics influences fitness and adaptation. Using *Drosophila melanogaster* and *Arabidopsis thaliana* population genomics data, we fitted models of distribution of fitness effects and estimated the rate of adaptive amino-acid substitutions both at the protein and amino-acid residue level. We performed a comprehensive analysis covering genome, gene, and protein structure, by exploring a multitude of factors with a plausible impact on the rate of adaptive evolution, such as intron number, protein length, secondary structure, relative solvent accessibility, intrinsic protein disorder, chaperone affinity, gene expression, protein function, and protein–protein interactions. We found that the relative solvent accessibility is a major determinant of adaptive evolution, with most adaptive mutations occurring at the surface of proteins. Moreover, we observe that the rate of adaptive substitutions differs between protein functional classes, with genes encoding for protein biosynthesis and degradation signaling exhibiting the fastest rates of protein adaptation. Overall, our results suggest that adaptive evolution in proteins is mainly driven by intermolecular interactions, with host–pathogen coevolution likely playing a major role.

***Key words:* protein structure, protein function, adaptation, population genetics, *Drosophila melanogaster*, *Arabidopsis thaliana*.**

Introduction

A long-standing focus in the study of molecular evolution is the role of natural selection in protein evolution (Eyre-Walker 2006). One can measure the strength and direction of selection at the divergence level through the $d_{\mathrm{N}}/d_{\mathrm{S}}$ ratio (ω). However, because ω represents a summary statistic across nucleotide sites, it can only provide the average trend, while proteins will typically undergo both negative and positive selection. Branch-site models address this issue by fitting phylogenetic models with heterogeneous $d_{\mathrm{N}}/d_{\mathrm{S}}$ ratio among codons and branches, thus considering the great heterogeneity in selective constraints among sites, both in space and time (Nielsen and Yang 1998; Yang et al. 2005; Zhang et al. 2005). Although these methods potentially allow studying adaptation at the site level, they require large amounts of data across species and are therefore restricted to more conserved genes along the phylogeny. Conversely, the McDonald and Kreitman (MK) test (McDonald and Kreitman 1991) is applied at the population level and it only requires data from two closely related species, usually several individuals from the study species and one individual from the other. Because adaptive mutations contribute relatively more to substitution than to polymorphism, the MK test disentangles positive and negative selection by contrasting the number of substitutions to the number of polymorphisms at synonymous and nonsynonymous sites. Charlesworth (1994) extended this method to estimate the proportion of substitutions that is adaptive (α). Yet, one limitation of this approach was that it did not account for the segregation of slightly deleterious mutations, which can either over- or underestimate measurements of α according to the demography of the population (Eyre-Walker 2002; Smith and Eyre-Walker 2002). Recent methods solved this issue by taking into consideration the distribution of fitness effects (DFE) of both slightly deleterious (Fay et al. 2001; Smith and Eyre-Walker 2002; Bierne and Eyre-Walker 2004; Eyre-Walker et al. 2006; Eyre-Walker and Keightley 2009; Stoletzki and Eyre-Walker 2011) and slightly beneficial mutations (Galtier 2016; Tataru et al. 2017). By allowing the estimation of the rate of nonadaptive ($\omega_{\mathrm{na}} = d_{\mathrm{N}}^{\hat{na}}/d_{\mathrm{S}}$) and adaptive ($\omega_{\mathrm{a}} = \omega - \omega_{\mathrm{na}}$) nonsynonymous substitutions, in addition to measurements of α ($\omega_{\mathrm{a}}/\omega$), these methods triggered new insights on the impact of both negative and positive selection on the rate of protein evolution.

Several studies have reported substantial levels of adaptive protein evolution in various animal species, including the fruit fly (Smith and Eyre-Walker 2002; Sawyer et al. 2003; Bierne and Eyre-Walker 2004; Haddrill et al. 2010), the wild mouse (Halligan et al. 2010), and the European rabbit (Carneiro et al. 2012), but also in bacteria (Charlesworth and Eyre-Walker 2006) and in plants (Ingvarsson 2010; Slotte et al. 2010; Strasburg et al. 2011). Whereas for other taxa, such as primates (Boyko et al. 2008; Hvilsom et al. 2012; Galtier 2016) and many other plants (Gossmann et al. 2010), the rate of adaptive mutations was observed to be very low, wherein amino-acid substitutions are expected to be nearly neutral and fixed mainly through random genetic drift (Boyko et al. 2008). Several authors proposed that this across-species variation in the molecular adaptive rate is explained by an effective population size (N_e) effect, where higher rates of adaptive evolution are observed for species with larger N_e due to a lower impact of genetic drift (Eyre-Walker 2006; Eyre-Walker and Keightley 2009; Gossmann et al. 2012). Galtier (2016), however, reported that N_e had an impact on α and ω_{na} but not ω_a. Hence, he proposed that the relationship with N_e is mainly explained by deleterious effects, wherein slightly deleterious nonsynonymous substitutions accumulate at lower rates in large-N_e species due to the higher efficiency of purifying selection, thus decreasing ω_{na} and consequently inflating α.

The rate of adaptive substitutions, however, was observed to vary extensively along the genome. On a genome-wide scale, it was reported that ω_a correlates positively with both the recombination and mutation rates, but negatively with gene density (Campos et al. 2014; Castellano et al. 2016). When looking at the gene level, previous studies have demonstrated the role of protein function in the rate of adaptive evolution, wherein genes involved in immune defense mechanisms appear with higher rates of adaptive mutations in Drosophila (Sackton et al. 2007; Obbard et al. 2009), humans, and chimpanzees (Nielsen et al. 2005). In Drosophila, sex-related genes also display higher levels of adaptive evolution, being directly linked with species differentiation (Pröschel et al. 2006; Haerty et al. 2007). At the intragenic level, however, the factors impacting the frequency and nature of adaptive mutations remain poorly understood.

There are several structural factors that have been reported to influence the rate of protein evolution but have not been investigated at the population level. Molecular evolution studies of protein families revealed that protein structure, for instance, significantly impacts the rate of amino-acid substitutions, with exposed residues evolving faster than buried ones (Liberles et al. 2012). As a stable conformation is often required to ensure proper protein function, mutations that impair the stability or the structural conformation of the folded protein are more likely to be counter-selected. Moreover, distinct sites in a protein sequence differ in the extent of conformational change they endure upon mutation, a pattern generally well predicted by the relative solvent accessibility (RSA) of a residue (Goldman et al. 1998; Mirny and Shakhnovich 1999; Franzosa and Xia 2009). In this way, residues at the core of proteins evolve slower than the ones at the surface due to their role in maintaining a stable protein structure (Perutz et al. 1965; Overington et al. 1992; Goldman et al. 1998; Bustamante et al. 2000; Dean et al. 2002; Choi et al. 2006; Lin et al. 2007; Conant and Stadler 2009; Franzosa and Xia 2009; Ramsey et al. 2011). Interspecific comparative sequence analyses also revealed that positively selected sites are often found at the surface of proteins (Proux et al. 2009; Adams et al. 2017). Hence, exploring the role that these structural elements play in shaping the rate of adaptive evolution is crucial in order to fully understand what are the main drivers of adaptation within proteomes.

Our study addresses protein adaptive evolution at a fine scale by analyzing the impact of several functional variables among protein-coding regions at the population level. To further assess the potential generality of the inferred effects, we carried our comparison on two model species with distinct life-history traits: the dipter *Drosophila melanogaster* and the brassicaceae *Arabidopsis thaliana*. We fitted models of DFE and estimated the rate of adaptive substitutions, both at the protein and amino-acid residue scale, across several variables and found that solvent exposure is the most significant factor influencing protein adaptation, with exposed residues undergoing ten times faster ω_a than buried ones. Moreover, we observed that the functional class of proteins has also a strong impact on the rate of protein adaptation, with genes encoding for processes of protein regulation and signaling pathways exhibiting the highest ω_a values. We, therefore, hypothesized that intermolecular interactions are the main drivers of adaptive substitutions in proteins. This hypothesis is consistent with the proposal that, at the inter-organism level, coevolution with pathogens constitute a so-far under-assessed component of protein evolution (Sackton et al. 2007; Obbard et al. 2009; Enard et al. 2016; Mauch-Mani et al. 2017).

Results and Discussion

In order to identify the genomic and structural variants driving protein adaptive evolution, we looked at 10,318 protein-coding genes in 114 *Drosophila melanogaster* genomes, analyzing polymorphism data from an admixed sub-Saharan population from Phase 2 of the *Drosophila* Population Genomics Project (DPGP2, Pool et al. 2012) and divergence out to *D. simulans*; and 18,669 protein-coding genes in 110 *Arabidopsis thaliana* genomes, with polymorphism data from a Spanish population (1001 Genomes Project, Weigel and Mott 2009) and divergence to *A. lyrata*. The rate of adaptive evolution was estimated with the Grapes program (Galtier 2016). The Grapes method extends the approach pioneered by the DoFE program (Fay et al. 2001; Smith and Eyre-Walker 2002; Bierne and Eyre-Walker 2004; Eyre-Walker et al. 2006; Eyre-Walker and Keightley 2009; Stoletzki and Eyre-Walker 2011), by explicitly accounting for mutations with slightly advantageous effects. Grapes estimates the rate of nonadaptive nonsynonymous substitutions (ω_{na}), which is then used to estimate the rate of adaptive nonsynonymous substitutions (ω_a) and the proportion of adaptive nonsynonymous substitutions (α). A high α can be potentially explained both by a higher ω_a or a lower ω_{na}, and therefore does not allow to disentangle the two effects. Thus, we explored whether, and how, ω_a and ω_{na}, as well as the total ω, depend on the different functional variables analyzed here.

Results from the model comparison of DFE showed that the Gamma-Exponential model is the one that best fits our data according to Akaike's information criterion (Akaike 1973) (supplementary table S1 in supplementary file S1, Supplementary Material online). This model combines a Gamma distribution of deleterious mutations with an exponential distribution of beneficial mutations. In agreement with previous surveys within animal species, this model suggests the existence of slightly deleterious, as well as slightly beneficial segregating mutations in *D. melanogaster* and *A. thaliana* genomes (Galtier 2016). Genome-wide estimates of ω_a for *A. thaliana* and *D. melanogaster* are 0.05 and 0.09, respectively, and are in the range of previously reported estimates for these species (Smith and Eyre-Walker 2002; Bierne and Eyre-Walker 2004; Gossmann et al. 2012).

In order to investigate the main drivers of protein adaptive evolution, we divided the data sets into sets of genes and amino-acid residues according to the variables analyzed, and fitted models of DFE in each subset independently. We distinguished two types of analyses: gene-based and site-based, where we looked into how the molecular adaptive rate varies across different categories of genes and amino-acid residues, respectively. Gene-based analyses allowed us to explore the impact of the background recombination rate, the number of introns, mean expression levels, and breadth of expression. At the protein level, we investigated the effect of binding affinity to the molecular chaperone *DnaK*, protein length, cellular localization of proteins, protein functional class, and number of protein–protein interactions (PPI). Finally, site-based analyses enabled us to study the effect of the secondary structure (SS) of the protein, by comparing residues present in β-sheets, α-helices, and loops; the tertiary structure, by considering the RSA of a residue and the residue intrinsic disorder; and whether an amino-acid residue participated or not in an annotated active site.

The Impact of Gene and Genome Architecture on Adaptive Evolution

To study the impact of gene and genome architecture on the rate of adaptive evolution, we looked at recombination rate and the number of introns. Recombination rate was previously reported to favor the fixation of adaptive mutations in Drosophila by breaking down linkage disequilibrium (Marais and Charlesworth 2003; Castellano et al. 2016). Our results are consistent with previous observations by showing a significant positive correlation in estimates of ω_a with increasing levels of recombination rate for *D. melanogaster* (table 1 and supplementary fig. S1 and file S2, Supplementary Material online). This was also observed in *A. thaliana* (table 1 and supplementary fig. S1 and file S2, Supplementary Material online), thus corroborating the effect of recombination in the rate of adaptive evolution.

Previous studies proposed that genes containing more introns are under stronger selective constraints due to the high cost of transcription, especially in highly expressed genes (Castillo-Davis et al. 2002). Hence, we would expect regions with more introns to be under stronger purifying selection. Conversely, by increasing the total gene length, introns might also effectively increase the intragenic recombination rate, which could in turn increase the efficacy of positive selection and have a positive impact on ω_a. To disentangle the two

Table 1. Number of Genes and Categories Analyzed for Each Continuous Variable and the Corresponding Kendall's τ with the Respective Significance (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; "." $0.05 \le P < 0.10$) for ω, ω_{na}, and ω_a for *Arabidopsis thaliana* and *Drosophila melanogaster*.

	A. thaliana					*D. melanogaster*				
	Number of Categories	Number of Genes	ω_a	ω_{na}	ω	Number of Categories	Number of Genes	ω_a	ω_{na}	ω
Recombination rate	50	18,668	0.2065 (*)	−0.2212 (*)	0.0857	30	8,485	0.3839 (**)	−0.402 (**)	0.0759
Intron number	13	15,347	−0.1538	−0.3590 (.)	−0.7949 (***)	10	10,318	−0.3333	−0.866 (***)	−0.7333 (**)
Protein length	30	18,669	−0.1310	−0.6735 (***)	−0.6782 (***)	50	10,318	−0.4775 (***)	−0.6963 (***)	−0.7763 (***)
Relative solvent accessibility	28	9,034	0.7513 (***)	0.8466 (***)	0.9841 (***)	19	4,944	0.8129 (***)	0.5789(***)	0.9766 (***)
Protein intrinsic disorder (site)	30	18,668	0.6000 (***)	0.9172 (***)	0.9770 (***)	30	8,485	0.7057 (***)	0.6690(***)	0.9540 (***)
Proportion of disordered residues (gene)	30	18,668	0.1908	0.7333 (***)	0.7517 (***)	20	8,485	0.7263 (***)	0.0631	0.5684 (***)
Breadth of expression	4	17,999	−0.6667	−1.0000 (*)	−1.0000 (*)	6	4,601	−0.7333 (*)	−0.4667	−0.7333 (*)
Mean gene expression	40	17,999	−0.1385	−0.9154 (***)	−0.9282 (***)	15	6,247	−0.5048 (**)	−0.6190 (**)	−0.7714 (***)
Protein–protein interactions	–	–	–	–	–	19	5,628	−0.3099 (.)	−0.1111	−0.3684 (*)

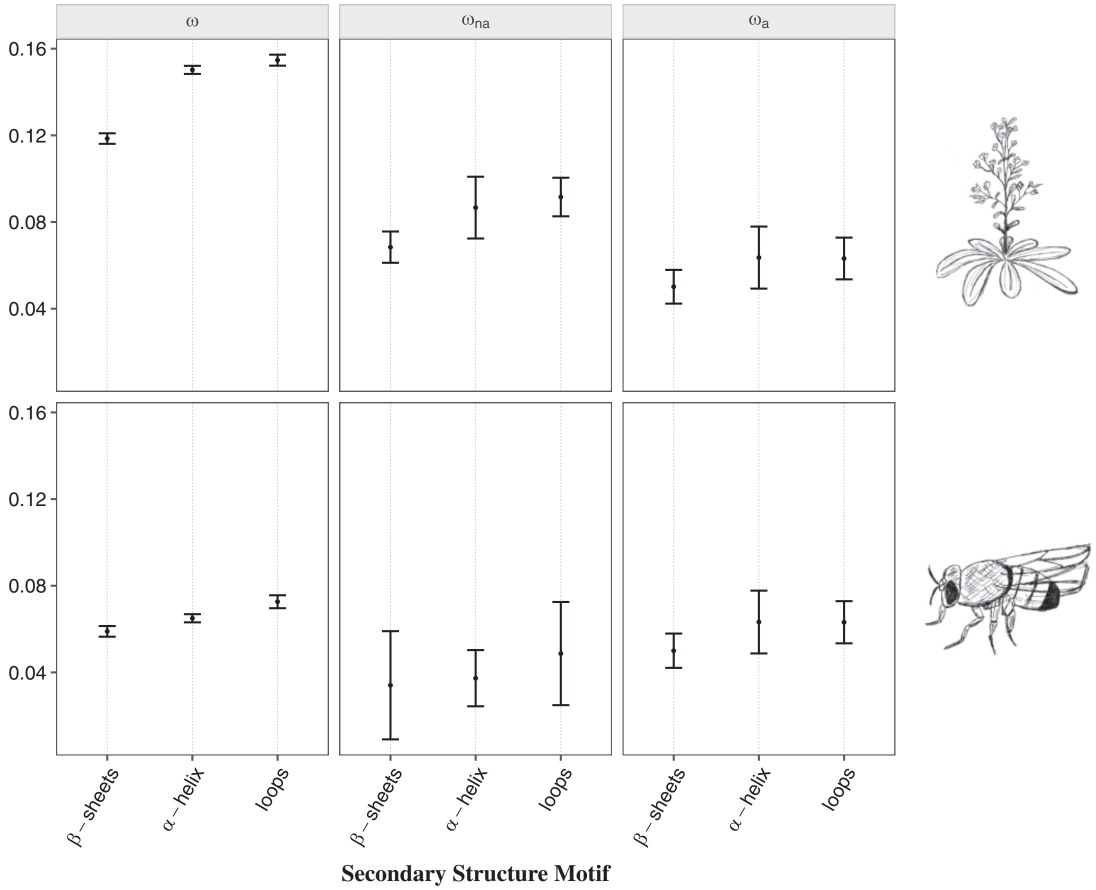

Fig. 1. Estimates of the rate of protein evolution (ω), nondaptive nonsynonymous substitutions (ω_{na}), and adaptive nonsynonymous substitutions (ω_a) for each of the secondary structural motif (β-sheets, α-helices, and loops) in *Arabidopsis thaliana* (top) and *Drosophila melanogaster* (bottom). Mean values of ω, ω_{na}, and ω_a for each motif are represented with the black points. Error bars denote for the 95% confidence interval for each category, computed over 100 bootstrap replicates. The hand-drawings of *A. thaliana* and *D. melanogaster* were made by A.F.M.

effects, analyses were performed by comparing genes with different intron content. Results showed a significant negative correlation of ω_{na} with an increasing number of introns in *D. melanogaster* (table 1 and supplementary fig. S2 and file S2, Supplementary Material online). Conversely, the number of introns did not significantly correlate with ω_a (table 1 and supplementary fig. S2 and file S2, Supplementary Material online). These findings suggest that the effect of the intron content on the rate of protein evolution is essentially due to stronger purifying selection while having a negligible influence on the rate of adaptive substitutions.

The Impact of Protein Structure on Adaptive Evolution

We further explored the impact of three different levels of protein structure (i.e., primary, secondary, and tertiary) on the rate of adaptive evolution. We first looked at the primary structure by categorizing proteins according to their length. Former studies correlating gene length and d_N/d_S have shown that smaller genes evolve more rapidly (Zhang 2000; Lipman et al. 2002; Liao et al. 2006). Here, we investigated whether this faster evolution is followed by a higher rate of adaptive substitutions. Results show significant negative correlations with protein length for values of ω and ω_{na} in both species (table 1 and supplementary fig. S3 and file S2, Supplementary Material online). The same trend was observed for ω_a, although it was only significant in *D. melanogaster* (table 1 and supplementary fig. S3 and file S2, Supplementary Material online). These findings suggest that smaller protein-coding regions are indeed under more relaxed purifying selection but might also evolve, in some cases, under a higher rate of adaptive substitutions.

The analysis at the secondary structural level showed significant differences in the evolutionary rate between the structural motifs, with loops demonstrating the highest values of ω, followed by α-helices and β-sheets (table 2 and fig. 1). When considering adaptive and nonadaptive substitutions separately, β-sheets show significantly lower values of ω_{na} in *A. thaliana* and ω_a in both species, with marginally significant values observed for *D. melanogaster* (table 2, fig. 1 and supplementary file S3, Supplementary Material online). This implies that the structural motif has an impact on the selective constraints in *A. thaliana* and also contributes to the rate of adaptation in the two species. Previous studies investigating protein tolerance to amino-acid change have similarly shown that loops and turns are the most mutable, followed by α-helices and β-sheets (Goldman et al. 1998; Guo et al.

Table 2. Number of Genes and Categories Analyzed for Each Discrete Variable and the Corresponding Difference between the Mean Values of Each Category is Reported for ω, ω_{na}, and ω_a for *Arabidopsis thaliana* and *Drosophila melanogaster*.

	Pairwise Comparisons	*A. thaliana*					*D. melanogaster*				
		Number of Categories	Number of Genes	ω_a	ω_{na}	ω	Number of Categories	Number of Genes	ω_a	ω_{na}	ω
Secondary structure	β-sheets–α-helices	3	9,034	−0.01346 (*)	−0.0182 (.)	−0.0317 (*)	3	4,944	−0.0132 (.)	−0.0033	−0.0060 (*)
	β-sheets–loops			−0.0130 (*)	−0.0231 (*)	−0.0361 (*)			−0.0131 (.)	−0.0146	−0.0137 (*)
	α-helices–loops			0.0004	−0.0049	−0.0045 (*)			0.00009	−0.0114	−0.0076 (*)
Affinity to molecular Chaperone	Binder–Non-Binder	2	17,775	0.0092	0.0260	0.0352 (*)	2	9,420	0.00009	0.0606 (*)	0.0515 (*)
Protein location[a]		7	18,669				7	10,318			
Protein functional class[a]		27	3,780				23	2,948			

Note.—Significance levels as in table 1.

[a] Due to the large amount of comparisons, the detailed pairwise comparisons and the corresponding *P* values are detailed in supplementary files S3 and S4, Supplementary Material online.

2004; Choi et al. 2006). Some authors posed this relationship as an outcome of residue exposure (Goldman et al. 1998; Guo et al. 2004), while others associate it to the degree of structural disorder, where ordered proteins are under stronger selective constraint (Choi et al. 2006). In order to clarify this, we further look into the impact of tertiary structure, by exploring the relationship between residue exposure to solvent and intrinsic protein disorder with the rate of adaptive evolution.

Considering the RSA, several studies previously demonstrated that residues at the surface of proteins evolve faster than the ones at the core (Goldman et al. 1998; Choi et al. 2007; Lin et al. 2007; Franzosa and Xia 2009). This higher substitution rate can be either due to a reduced selective constraint at exposed residues and/or to an increased rate of adaptive substitutions. To disentangle the two effects, we compared the site frequency spectra (SFS) across several categories of RSA. Our results recapitulate those of previous studies on divergence and demonstrate a significant positive correlation with solvent exposure for values of ω (table 1 and fig. 2*a*). Moreover, we demonstrate that both relaxation of the selective constraints (ω_{na}) and a higher rate of adaptive nonsynonymous substitutions (ω_a) explain the higher evolutionary rate at the surface of proteins (table 1, fig. 2*a* and supplementary file S2, Supplementary Material online).

Intrinsically disordered proteins are defined by lacking a well-defined 3D fold (Dunker et al. 2002; Dyson and Wright 2005), more specifically, proteins that have a higher degree of loop dynamics (“hotloops”) (Linding et al. 2003). As these structures are more flexible, we expect them to be under less structural constraint and to accumulate more substitutions (Guo et al. 2004; Wilke et al. 2005; Choi et al. 2006; Afanasyeva et al. 2018), either deleterious and/or beneficial. To test this hypothesis, we asked two different questions: 1) Are intrinsically disordered protein regions more likely to respond to adaptation? 2) Are proteins with more disordered regions undergoing more adaptive substitutions? For the first question, we divided amino-acid residues based on their predicted value of intrinsic disorder. We report a significant positive correlation with ω, ω_a, and ω_{na} with residue intrinsic disorder for both species (table 1, fig. 2*b* and supplementary file S2, Supplementary Material online). For the second question, proteins were categorized according to their proportion of disordered residues (see Materials and Methods). Our results reveal a significant positive correlation of protein disorder with ω in both species, ω_{na} in *A. thaliana* and ω_a in *D. melanogaster* (table 1 and supplementary fig. S4 and file S2, Supplementary Material online). These findings suggest that, at the residue level, intrinsically disordered regions are more likely to respond to adaptation and are also under less selective constraint in both species. However, when considering the whole protein, we observe that intrinsically disordered proteins have different effects between species. In particular, they contribute to the relaxation of purifying selection in *A. thaliana* and to a higher rate of adaptation in *D. melanogaster*. The reason for the difference between species is unclear and will require further analyses.

Finally, we tested whether the rate of adaptive substitutions is affected by the binding affinity of proteins to

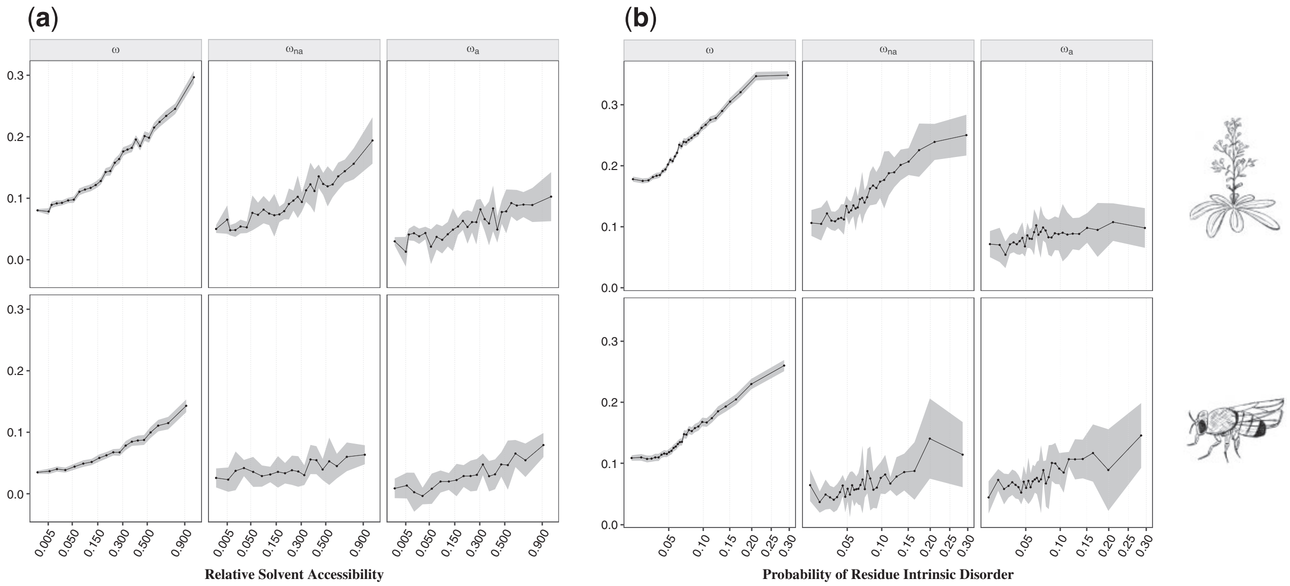

FIG. 2. Relationship between ω, ω_{na}, and ω_a with (*a*) the relative solvent accessibility (RSA) and (*b*) the probability of residue intrinsic disorder for *Arabidopsis thaliana* (top) and *Drosophila melanogaster* (bottom). The *x* axis is scaled using a squared root function. Mean values of each estimate for each category are represented with connected black dots. The shaded area represents the 95% confidence interval of each category, computed over 100 bootstrap replicates.

molecular chaperones. It has been suggested that binding to a chaperone leads to a higher evolutionary rate due to the buffering effect for slightly deleterious mutations (Bogumil and Dagan 2010; Kadibalban et al. 2016). Here, we investigate whether binding to the chaperone *DnaK* could also favor the fixation of adaptive mutations. In agreement with previous studies, we find a higher ω and ω_{na} in proteins binding to *DnaK* in *D. melanogaster* (table 2 and supplementary fig. S5, Supplementary Material online), but no impact on ω_a (table 2 and supplementary fig. S5 and file S3, Supplementary Material online), suggesting that the interaction with a molecular chaperone does not influence the fixation of beneficial mutations.

Protein Function and Adaptive Evolution

We further explored the impact of protein function on sequence evolution. To do so, we analyzed the effect of mean gene expression, breadth of expression, protein location, and protein functional class on the rate of adaptive substitutions. Several studies on both Eukaryote (Pal et al. 2001; Subramanian and Kumar 2004; Wright et al. 2004; Lemos et al. 2005) and Prokaryote (Rocha and Danchin 2004) organisms have shown that highly expressed genes have lower rates of protein sequence evolution. Here, we investigated if the lower evolutionary rate is followed by a reduced rate of adaptive substitutions. Our results support previous findings by displaying a significant negative correlation of mean gene expression with estimates of ω and ω_{na} in both species (table 1, fig. 3 and supplementary file S2, Supplementary Material online). Besides, we find that mean gene expression is also significantly negatively correlated with ω_a in *D. melanogaster* (table 1, fig. 3 and supplementary file S2, Supplementary Material online), suggesting that gene expression also constrains the rate of adaptation, in addition to the well-known effect on purifying selection. It has been hypothesized that the higher selective constraint in highly expressed genes could be driven by the reduced probability of protein misfolding, wherein selection acts by favoring protein sequences that accumulate less translational missense errors (Drummond et al. 2005). Hence, the higher selective pressure to increase stability in highly expressed proteins could also be hampering the fixation of adaptive mutations. Moreover, as mean gene expression is positively correlated with the breadth of expression (Kendall's $\tau = 0.3376$, $P < 2.2\text{e-}16$ in *A. thaliana*; Kendall's $\tau = 0.2170$, $P < 2.2\text{e-}16$ in *D. melanogaster*; supplementary fig. S6, Supplementary Material online), and the latter is a good proxy for the pleiotropic effect of a gene, which is known to impose high selective constraints (i.e., Salvador-Martínez et al. 2018), we also analyzed the impact of the number of tissues where a gene is expressed on the rate of adaptive evolution. We report a significant negative correlation of the breadth of expression (number of tissues) with ω in both species (table 1 and supplementary fig. S7, Supplementary Material online), thus corroborating previous findings (Duret and Mouchiroud 2000; Slotte et al. 2011; Salvador-Martínez et al. 2018). When looking at adaptive and nonadaptive substitutions separately, we observe a significant negative impact on values of ω_a in *D. melanogaster* and ω_{na} in *A. thaliana* (table 1 and supplementary fig. S7 and file S2, Supplementary Material online). This suggests that the breadth of expression is acting together with the mean expression levels, although with an apparently lower magnitude effect both in ω_{na} and ω_a.

In order to assess the impact of protein location, we classified genes into the following cellular categories: cytoplasmic, endomembrane system, mitochondrial, nuclear, plasma membrane, and secreted proteins (supplementary tables S2 and S3 in supplementary file S1, Supplementary Material

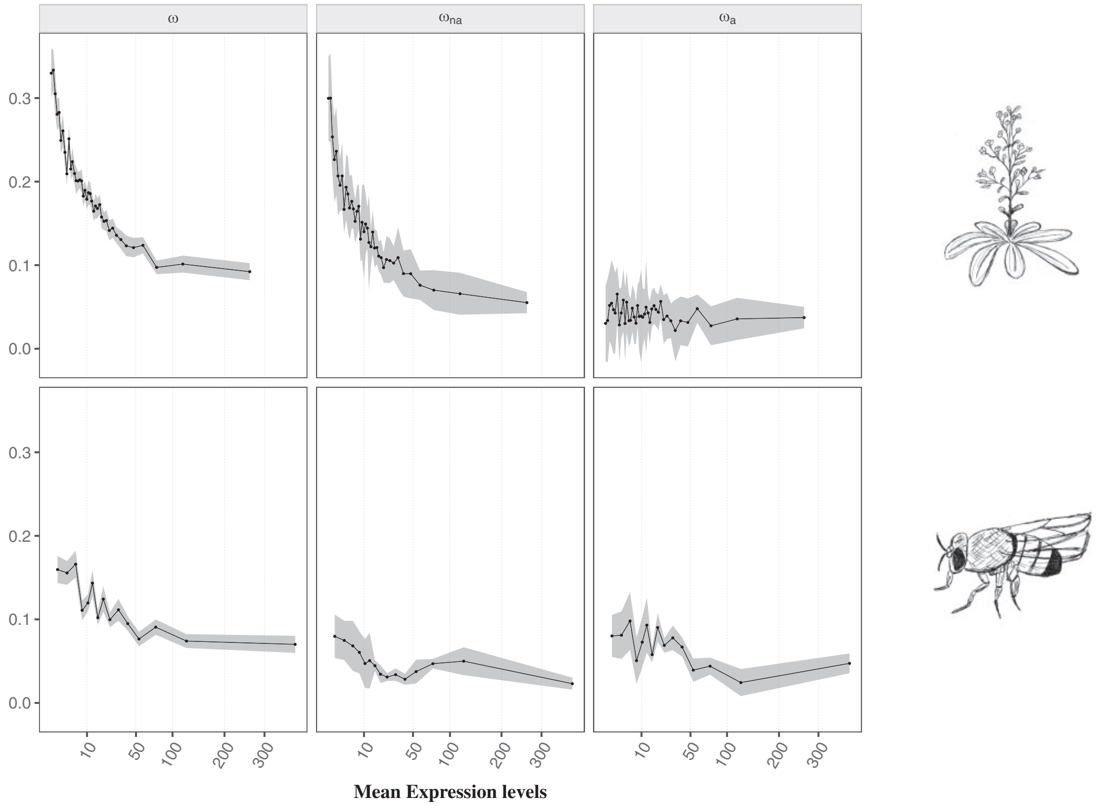

Fig. 3. Estimates of ω, ω_{na}, and ω_{a} for each category of genes with distinct mean gene expression levels for *Arabidopsis thaliana* (top) and *Drosophila melanogaster* (bottom). The *x* axis is scaled using a squared root function. Legend as in figure 2.

online). Results show significantly higher rates of protein evolution in nuclear and secreted proteins, with the lowest values observed in the mitochondria, plasma membrane, and endomembrane system (pairwise comparisons; $P = 0.0128$ in *A. thaliana*; $P = 0.0104$ in *D. melanogaster*; supplementary fig. S8, Supplementary Material online). However, this result seems to be explained by a reduced purifying selection, with significantly higher values of ω_{na} observed in cytoplasmic, nuclear, and secreted proteins (pairwise comparisons; $P = 0.0128$ in *A. thaliana*; $P > 0.0729$ in *D. melanogaster*; supplementary fig. S8, Supplementary Material online), and not by a higher rate of adaptive substitutions, since no significant differences were found between the categories in the estimates of ω_a (supplementary fig. S8 and file S3, Supplementary Material online).

By analyzing the different categories of protein functional class (supplementary tables S2 and S3 in supplementary file S1, Supplementary Material online), we observe that genes involved in protein biosynthesis (i.e., mRNA and ribosome biogenesis and transcription machinery) and signaling for protein degradation (ubiquitin system) exhibit the highest rates of adaptive substitutions (fig. 4 and supplementary file S4, Supplementary Material online), functions coded mostly by nuclear and cytoplasmic proteins. Signal transduction pathways also appear to play a role in adaptation, since protein phosphatases also present high rates of adaptive mutations (Hunter 1995). Moreover, in *A. thaliana*, cytochrome P450 proteins are also in the top categories of ω_a (fig. 4 and supplementary file S4, Supplementary Material online). We fitted a linear model to the ω_{a} values of the shared categories (21 categories in total) to see if results were consistent between the two species and found a positive correlation (Kendall's $\tau = 0.257$, $P = 0.1101$; supplementary fig. S9*a*, Supplementary Material online), which is stronger after discarding the two outliers, mRNA biogenesis and glycosyltransferases (Kendall's $\tau = 0.333$, $P = 0.0490$; supplementary fig. S9*b*, Supplementary Material online). Our findings, therefore, suggest that adaptive mutations occur mainly through processes of protein regulation and signaling pathways.

What Are the Major Drivers of Adaptive Evolution along the Genome?

Overall, we found multiple factors influencing protein adaptive evolution, specifically recombination rate (positive correlation), protein length (negative correlation), secondary structural motif (lower values observed for β-sheets), RSA (positive correlation), protein intrinsic disorder (positive correlation), gene expression levels (negative correlation), and protein functional class. Since some of these variables are intrinsically correlated, we next asked whether some of the inferred effects are spurious. First of all, it is known that protein length and gene expression are negatively correlated, wherein highly expressed genes tend to be shorter, as previously reported for vertebrates (Subramanian and Kumar 2004), yeast (Coghlan and Wolfe 2000; Akashi 2003), and observed in this study (Kendall's $\tau = -0.015$, $P = 1.22\text{e-}02$ in *A. thaliana*; $\tau = -0.093$, $P = 1.70\text{e-}28$ in *D.*

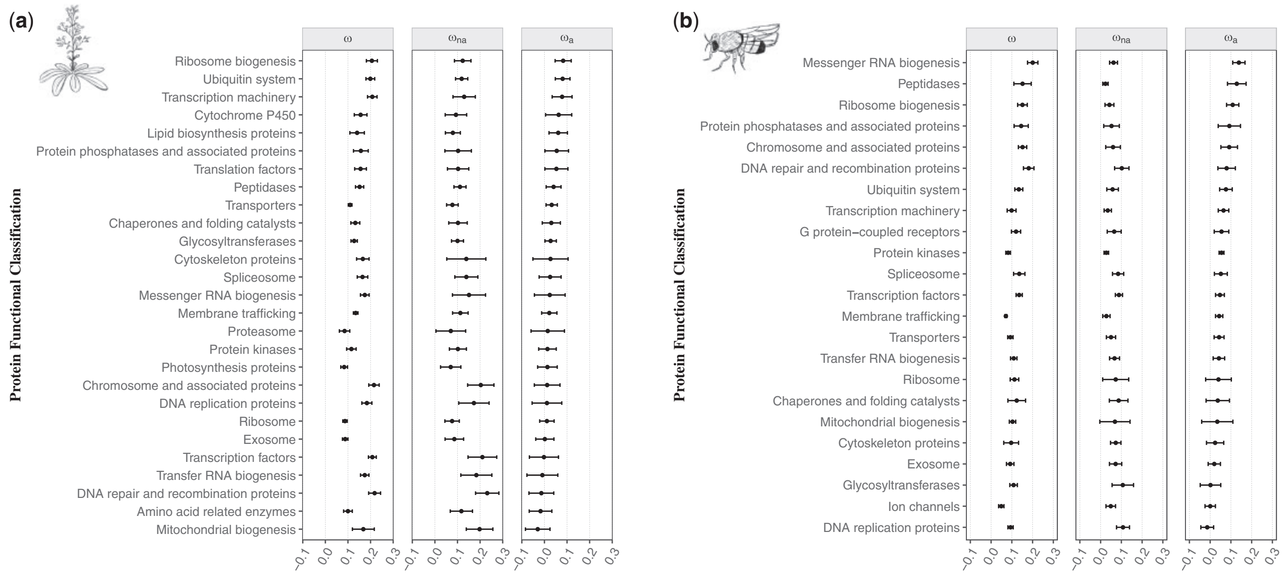

FIG. 4. Estimates of ω, ω_{na}, and ω_a for each category of protein functional class in (*a*) *Arabidopsis thaliana* and (*b*) *Drosophila melanogaster*. Categories are ordered according to the values of ω_a. Mean values of ω, ω_{na}, and ω_a for each class are represented with the black points. Error bars denote the 95% confidence interval for each category, computed over 100 bootstrap replicates.

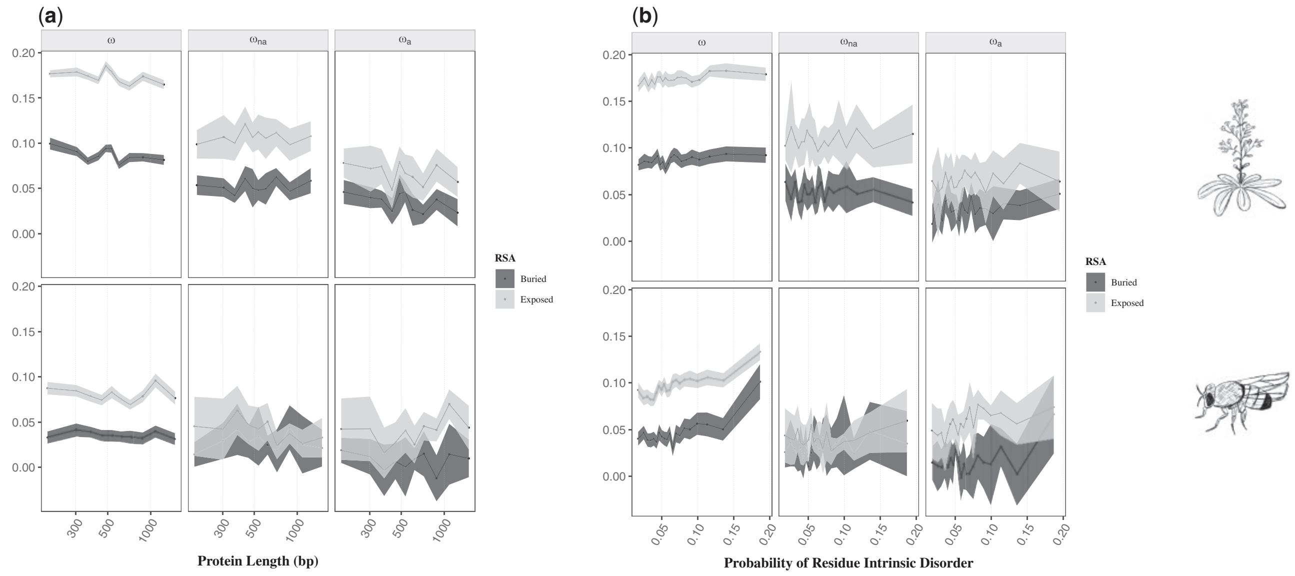

FIG. 5. Estimates of ω, ω_{na}, and ω_a plotted as a function of (*a*) the relative solvent accessibility and protein length and (*b*) the relative solvent accessibility and the probability of residue intrinsic disorder in *Arabidopsis thaliana* (top) and *Drosophila melanogaster* (bottom). The *x* axis is log-scaled. Analyses were performed by comparing buried (RSA <0.05) and exposed (RSA ≥0.05) residues across ten categories of protein length in (*a*) and 20 categories of intrinsic disorder in (*b*) for both species. Legend as in figure 2.

melanogaster; supplementary fig. S10, Supplementary Material online). Since highly expressed genes have lower rates of adaptive substitutions and shorter genes have higher rates of adaptive evolution, we may conclude that these two variables independently impact the rate of adaptation in proteins. Protein length is also negatively correlated with the proportion of exposed residues (Kendall's $\tau = -0.310$, $P = 0.00$ in *A. thaliana*; $\tau = -0.404$, $P = 1.03\text{e-}223$ in *D. melanogaster*; supplementary fig. S11, Supplementary Material online), as the surface/volume ratio of globular proteins decreases when protein length increases (Janin 1979). By estimating the rate of adaptive mutations of buried and exposed sites separately, we observe that the effect of protein length is no longer significant (table 3, fig. 5*a* and supplementary file S5, Supplementary Material online). This suggests that the effect of protein length on the rate of adaptive substitutions is a by-product of the effect of the residue's solvent exposure. Furthermore, mean gene expression is positively correlated with solvent exposure (Kendall's $\tau = 0.016$, $P = 0.1037$ in *A. thaliana*; $\tau = 0.327$, $P = 4.50\text{e-}45$ in *D. melanogaster*; supplementary fig. S12, Supplementary Material online), as expected since highly expressed genes

Table 3. Statistical Results for the Comparisons Performed Including RSA as a Cofactor.

	Categories	Statistics	*Arabidopsis thaliana*		*Drosophila melanogaster*	
			RSA		RSA	
			Buried	Exposed	Buried	Exposed
Protein length	10	ω_a	−0.4222 (.)	−0.2889	−0.0667	0.3333
		ω_{na}	−0.0222	0.0667	−0.0667 (.)	−0.4222 (.)
Protein disorder	20	ω_a	0.2105	0.2105	0.0842	0.5368 (***)
		ω_{na}	−0.0631	−0.0211	0.2947	−0.0316
Secondary structure	B-sheets–α-helices	ω_a	−0.0073	−0.0074	0.0118	−0.0040
		ω_{na}	0.0003	−0.0230 (.)	−0.0063	−0.0006
	B-sheets–loops	ω_a	−0.0021	−0.0078	0.0178	−0.0056
		ω_{na}	0.0050	−0.0173 (*)	−0.0133	−0.0039
	α-helices–loops	ω_a	0.0052	−0.0003	0.0059	−0.0016
		ω_{na}	0.0047	0.0056	−0.0071	−0.0033
Active site	Active–nonactive	ω_a	−0.0004	−0.0048	−0.0078	0.0055
		ω_{na}	−0.0057	0.0070	0.0042	−0.0045

Note.—For each comparison, the value for buried and exposed residues is indicated. For continuous variables (protein length and protein disorder), the Kendall's τ with the respective significance for ω_{na} and ω_a is reported. For discrete variables (secondary structure motif and active site) the difference between the mean values of each category is reported for ω_{na} and ω_a. Significance levels as in table 1.

are shorter and shorter genes have a greater proportion of exposed residues (supplementary figs. S10 and S11, Supplementary Material online). These two variables, however, have opposite effects on ω_a, and we therefore conclude that gene expression is acting independently from solvent exposure on the rate of adaptive protein evolution.

We further note that the SS motif is intrinsically correlated with the degree of intrinsic disorder, where loops and turns represent the most flexible motifs (supplementary fig. S13, Supplementary Material online), consistent with previous studies (Choi et al. 2006). When analyzing different degrees of protein disorder across the structural motifs, we observe that SS has only an impact on estimates of ω, while intrinsic protein disorder is significantly positively correlated with ω within the three motifs in both species, and ω_a within β-sheets in *A. thaliana* and within α-helices in *D. melanogaster* (supplementary fig. S14 and file S5, Supplementary Material online). Moreover, we report that the SS motif is correlated with solvent exposure (supplementary fig. S15, Supplementary Material online), β-sheets being mostly found at the core of proteins, while α-helices and loops have, on an average, higher solvent exposure (Bowie et al. 1990; Guo et al. 2004). By estimating the rate of adaptive substitutions in buried and exposed residues across the three motifs, the impact of SS is no longer noticeable on estimates of ω_a (table 3 and supplementary fig. S16 and file S5, Supplementary Material online), thus suggesting that the effect of SS motif is also a by-product of solvent exposure. When looking at the tertiary structure level, in agreement with Choi et al. (2006), we report that structures with more exposed residues tend to be more flexible (Kendall's $\tau = 0.001$, $P = 0.4726$ in *A. thaliana*; $\tau = 0.015$, $P = 0.0256$ in *D. melanogaster*; supplementary fig. S17, Supplementary Material online). Estimation of the rate of adaptive mutations in buried and exposed sites across different levels of residue intrinsic disorder shows that solvent exposure plays the main role in protein adaptive evolution, with a significant positive impact of protein disorder only observed in values of ω in both species and ω_a in exposed residues for *D. melanogaster* (table 3, fig. 5*b* and supplementary file S5, Supplementary Material online). To further clarify the relative contribution of solvent exposure and protein disorder on the rate of adaptive evolution, we performed an analysis of covariance (ANCOVA), using both measures and their interaction as explanatory variables. Results show that the RSA explains 95% ($P = 3.176\text{e-}14$) and 99% ($P < 2.2\text{e-}16$) of the variation in ω_a and ω_{na}, respectively, in *A. thaliana*; and 87% ($P = 1.011\text{e-}13$) and 62% ($P = 0.00012$) in ω_a and ω_{na}, respectively, in *D. melanogaster*. These findings suggest that the level of exposure of a residue in the protein structure is the main driver of adaptive evolution, and that structural flexibility potentially constitutes a comparatively small, if any, effect to protein adaptation. By comparing the level of exposure of the residues across the different classes of protein function, no differences were observed (supplementary fig. S18, Supplementary Material online), thus suggesting that these two variables independently affect the rate of protein adaptation.

Summarizing, after accounting for potentially confounding effects, our results show that besides population genetic processes such as recombination and mutation rate (Hill and Robertson 1966; Marais and Charlesworth 2003; Castellano et al. 2016), three major protein features significantly impact the rate of protein adaptive evolution: gene expression, RSA, and the protein functional class. When looking at the magnitude effect of each of these variables, we observe that exposed residues have a 10-fold higher rate of adaptive substitutions when compared with completely buried sites (fig. 2*a* and supplementary file S2, Supplementary Material online). The effect of gene expression seems to be of lower magnitude, wherein less expressed genes have a 2-fold higher rate of adaptive substitutions with a significant negative correlation observed only in *D. melanogaster* (fig. 3 and supplementary file S2, Supplementary Material online). As a comparison, genes in highly recombining regions have up to a 10-fold higher rate of adaptive substitutions compared with genes within regions with the lowest recombination rates (supplementary fig. S1 and file S2, Supplementary Material online), being therefore similar to that observed with solvent exposure. Previous studies reported that the type of amino-acid change also plays an important role in protein adaptive evolution, where more similar amino-acids present higher rates of adaptive substitutions (Grantham 1974; Miyata et al. 1979; Bergman and Eyre-Walker 2019). In order to evaluate a potential bias on the type of amino-acid at the surface and at the core of proteins, we computed the proportion of conservative and radical residue changes, according to volume and polarity indices, as defined by Grantham (Grantham 1974). We found similar frequencies of conserved and radical changes in buried and exposed residues, thus suggesting that our results at the structural level are not influenced by the type of amino-acid mutation (97% of conservative and 3% changes on buried residues; 96% of conservative and 4% changes on exposed sites). Our findings therefore suggest that protein architecture strongly influences the rate of adaptive protein evolution, wherein selection acts by favoring a greater accumulation of adaptive mutations at the surface of proteins.

Why Does Adaptation Occur Mainly at the Surface of Proteins?

Our results show that solvent exposure is the protein feature with the strongest impact on the rate of adaptive substitutions at the intramolecular level. To explain this effect, we discuss three hypotheses in which protein adaptive evolution occurs through 1) the acquisition of new biochemical activities at the surface of proteins, 2) the emergence of new functions via network rewiring at the level of PPI, and 3) intermolecular interactions between organisms, as a consequence of host–pathogen coevolution.

We first hypothesized that protein adaptation results from new catalytic activities, wherein adaptive mutations arise within active sites. Bartlett et al. (2002) reported that active sites are mostly present in more intrinsically disordered regions of the protein. Moreover, they proposed that apo-enzymes, which are not yet bound to the substrate or cofactor, present greater residue flexibility, and more exposed catalytic residues, which could favor a higher rate of adaptive substitutions. In order to test this, we estimated the rate of adaptive substitutions on active and nonactive sites, controlling for solvent exposure, and observed only significant differences in ω within buried residues in *A. thaliana* (table 3 and supplementary fig. S19 and file S5, Supplementary Material online), although with higher values observed for nonactive sites. While the nonsignificant differences in the rate of adaptive mutations could result from incomplete annotations, which tend to be biased toward motifs highly conserved across species (De Castro et al. 2006), this suggests that being present in an active site does not influence the rate of adaptation. Active sites, however, are rather mobile, presenting different levels of solvent exposure and residue flexibility according to the stage of the enzymatic reaction (Bartlett et al. 2002). Therefore, it may be arbitrary to assign them a certain solvent exposure class based on the phase the enzymes were crystallized, limiting our capacity to test their role on adaptive evolution.

Several studies discussed the impact of PPI on the rate of protein evolution. Valdar and Thornton (2001) and Caffrey et al. (2004) proposed that PPI may be acting as an inhibitor of protein evolution by enhancing the efficiency of purifying selection due to a higher degree of protein connectivity, typically associated with more complex functions. Mintseris and Weng (2005) supported this assumption but proposed that the proteins evolving slowly are the ones involved in obligate interactions, while proteins involved in transient interactions evolve at faster rates due to higher interface plasticity. Here, we ask whether the higher rate of adaptive mutations at the surface of proteins could have arisen through intermolecular interactions at the protein network level. We addressed this question by estimating the rate of adaptive mutations in genes with different degrees of PPI. This was only possible in *D. melanogaster* since there was limited data available for *A. thaliana*. We report a negative correlation between the number of PPI and ω, ω_{na}, and ω_a, respectively, with only significant values observed for ω (table 1 and supplementary fig. S20 and file S2, Supplementary Material online). These findings suggest that a higher degree of protein connectivity leads to lower rates of protein sequence evolution, but prevent us to assess with confidence whether this effect is due to a stronger purifying selection and/or a slower rate of adaptive substitutions. A potential limitation of this analysis is the low number of genes with PPI information available and the noise associated with the BioGRID annotations. As a physical interaction does not necessarily imply a functional link, we might lack statistical power to detect any putative effect of PPI on ω_a (Chatr-aryamontri et al. 2017).

In support to our third hypothesis, several studies have described the role of the immune and defense responses in molecular evolution across taxa (Sackton et al. 2007; Obbard et al. 2009; Enard et al. 2016; Mauch-Mani et al. 2017). These studies suggest that pathogens could be key drivers of protein adaptation, by acting as a powerful selective pressure through the coevolutionary arms race between hosts and parasites. This could be driving the higher rate of adaptive mutations in protein biosynthesis enzymes (fig. 4), which are the ones typically hijacked by pathogens during host infection (Dangl and Jones 2001; Enard et al. 2016). Moreover, one of the fastest evolving protein class is the ubiquitin system (fig. 4), which is known to be involved in the defense mechanism, both by the host, through processes like the activation of innate immune responses and degradation signaling of pathogenic proteins; and by the pathogen, which inhibits and/or uses this system in order to modulate host responses (Loureiro and Ploegh 2006; Collins and Brown 2010; Dielen et al. 2010; Trujillo and Shirasu 2010; Hiroshi et al. 2014). Membrane trafficking proteins are also well-known for being involved in the immune response mechanisms, a functional class that also presents high values of ω_a, and "DNA replication" together with "mRNA biogenesis" and "transcription machinery" are typical signatures of viruses' activities (fig. 4). Likewise, in *A. thaliana*, cytochrome P450 proteins present a high rate of adaptive mutations (fig. 4), which have been reported to play a crucial role in the defense response in plants (Schuler and Werck-Reichhart 2003). Besides, the reduced selective pressure on nuclear and secreted proteins (supplementary fig. S6, Supplementary Material online) may be also a consequence of their role in disease and pathogen immunity (i.e., Motion et al. 2015; Mosmann et al. 2016), as observed in yeast (Julenius and Pedersen 2006), insects (Sackton et al. 2007; Obbard et al. 2009), and primates (Nielsen et al. 2005).

Our findings, therefore, support the hypothesis that co-evolutionary arms race of the host–pathogen interactions, in particular, intracellular pathogens such as viruses, are a major driver of adaptation in proteins. While we do not rule out that PPI and the acquisition of new biochemical functions could also have an impact, more and better annotation data is required to further evaluate their role. In conclusion, our study reveals that, in addition to genome architecture, protein structure has a substantial impact on adaptive evolution consistent between *D. melanogaster* and *A. thaliana*, unraveling the potential generality of such effect. Our study further emphasizes that the rate of adaptation not only varies substantially between genes but also at the intragenic scale, and we posit that accounting for a fine-scale, intramolecular evolution is necessary to fully understand the patterns of molecular adaptation at the species level.

Materials and Methods

Population Genomic Data and Data Filtering

The *D. melanogaster* data set included alignments of 114 genomes for one chromosome arm of the two large autosomes (2 L, 2 R, 3 L, and 3 R) and one sex chromosome (X) pooled from 22 sub-Saharan populations with a negligible amount of population structure ($F_{ST} = 0.05$; DPGP2, Pool et al. 2012). Release 5 of the Berkeley Drosophila Genome Project (BDGP5, http://www.fruitfly.org/sequence/release5genomic.shtml, last accessed July 2017) was used as the reference genome. Estimations of divergence were performed with *D. simulans*, for which genome alignments with the reference genome were available (http://www.johnpool.net/genomes.html; last accessed July 2017). For *A. thaliana*, analyses were carried out with 110 genomes for the five chromosomes of the Spanish population from the 1001 Genomes Project (Weigel and Mott 2009), using the release 10 from The Arabidopsis Information Resource (TAIR10, ftp://ftp.ensemblgenomes.org/pub/plants/release-40/fasta/arabidopsis_thaliana/dna/; last accessed March 2018) as the reference genome. Divergence estimates were made with *A. lyrata* as an outgroup species, for which a pairwise alignment with the reference genome was available (ftp://ftp.ensemblgenomes.org/pub/plants/release-38/maf; last accessed March 2018). Data processing was conducted with the help of GNU parallel (Tange 2011).

Estimation of the Population Genetic Parameters and Model Selection

Coding DNA sequences (CDS) were extracted from the alignments with MafFilter (Dutheil et al. 2014) according to the General Feature Format (GFF) file of the reference genome of both species. First, a cleaning and filtering process was performed to keep only nonoverlapping genes with the longest transcript, in cases of multiple transcripts per gene. At this stage, 12,801 and 27,072 genes, for *D. melanogaster* and *A. thaliana*, respectively, were kept for further analysis. CDS sequences were then concatenated in order to obtain the full coding region per gene. For the analysis with *A. thaliana*, the alignment of *A. lyrata* with the reference sequence was realigned with each gene alignment of the ingroup using MAFFT v7.38 (Katoh and Standley 2013) with the options *add* and *keeplength* so that no gaps were included in the ingroup. CDS alignments with premature stop codons were excluded and alignment positions lacking a corresponding sequence in the outgroup were discarded. Final data sets included 10,318 genes for *D. melanogaster*/*D. simulans* and 18,669 genes for *A. thaliana*/*A. lyrata*. These data sets were then used to infer both the synonymous and nonsynonymous unfolded and folded SFS, and synonymous and nonsynonymous divergence based on the rate of synonymous and nonsynonymous substitutions. Sites for which the outgroup allele was missing were considered as missing data. All calculations were performed using the BppPopStats program from the Bio++ Program Suite (Guéguen et al. 2013). The Grapes program was then used to compute a genome-wide estimate of the rate of nonadaptive (ω_{na}) and adaptive nonsynonymous substitutions (ω_a) (Galtier 2016). This method assumes that all sites were sampled in the same number of chromosomes and since some sites were not successfully sampled in all individuals, the original data set was reduced to 110 and 105 individuals for *D. melanogaster* and *A. thaliana*, respectively, by randomly down-sampling polymorphic alleles at each site. The following models were fitted and compared using Akaike's information criterion: Neutral, Gamma, Gamma-Exponential, Displaced Gamma, Scaled Beta, and Bessel K. A model selection procedure was conducted on the two data sets using the complete set of genes for comparison (see supplementary table S1 in supplementary file S1, Supplementary Material online). As results were comparable when using the unfolded and folded SFS, subsequent analyses were performed on the unfolded SFS only. Following analyses consist in fitting the selected model on several subsets of the data according to the variables analyzed, comprising sets of genes (see supplementary tables S2 and S3 in supplementary file S1, Supplementary Material online, for detailed information on the genes used for each variable as well as the population genetic parameters estimated per gene for *A. thaliana* and *D. melanogaster*, respectively) and amino-acid residues (see supplementary tables S4 and S5 in supplementary file S1, Supplementary Material online, for detailed information on the amino-acid residues used for each category as well as the population genetic parameters estimated per site for *A. thaliana* and *D. melanogaster*, respectively). We next described the different variables analyzed.

Categorization of Gene and Genome Architecture

Recombination rates were obtained with the R package "MareyMap" (Rezvoy et al. 2007), by using the cubic splines interpolation method. Hereafter, we computed the mean recombination rate in cM/Mb units for each gene. Discretization of the observed distribution of recombination rate was performed in 50 and 30 categories with around 350 and 280 genes each for *A. thaliana* and *D. melanogaster*, respectively. Intronic information was obtained using the GenomeTools from a GFF with exon annotation and the option *addintrons* (Gremme et al. 2013). Genes were discretized into 13 and 10 categories according to their intron content for *A. thaliana* and *D. melanogaster*, respectively.

Categorization of Protein Structure

Genes were discretized according to the total size of the coding region, for which 30 and 50 categories with around 620 and 210 genes each were made for *A. thaliana* and *D. melanogaster*, respectively.

In order to obtain structural information for each protein sequence, blastp (Schaffer 2001) was first used to assign each protein sequence to a PDB structure, and respective chain, by using the "pdbaa" library and an *E*-value threshold of 1e-10. When multiple matches occurred, for instance in cases of multimeric proteins, the match with the lowest *E*-value was kept. This resulted in 5,008 genes for which a PDB structure was available, making a total of 3,834 PDB structures for *D. melanogaster* and 9,121 genes with a total of 3,832 PDB structures for *A. thaliana*. The corresponding PDB structures were then downloaded and further processed to only keep the corresponding chain per polymer. PDB manipulation and analysis were carried on using the R package "bio3d" (Grant et al. 2006). Values for SS and solvent accessibility (SA) per residue were obtained using the "dssp" program with default options and were successfully retrieved for 3,613 PDB files corresponding to 4,944 genes for *D. melanogaster* and 3,806 PDB files for a total of 9,106 genes for *A. thaliana*. Subsequently, to map SS and SA values to each residue of the protein sequence a pairwise alignment between each protein and the respective PDB sequence was performed with MAFFT, allowing gaps in both sequences in order to increase the block size of sites aligned. The final data set comprised a total of 1,397,885 and 1,395,666 sites with SS and SA information, respectively, out of 4,821,113 total codon sites obtained with BppPopStats for the complete set of genes of *D. melanogaster*; and 2,585,468 and 2,585,467 sites mapped with SS and SA information, respectively, out of 7,479,808 codon sites of *A. thaliana*. We computed the RSA by dividing SA by the amino-acid's solvent accessible area (Tien et al. 2013).

Categorization of SS was performed by comparing 460,702, 975,934, and 523,880 amino-acid residues in β-sheets, α-helices, and loops, respectively, in *A. thaliana*, and 258,898, 516,356, and 282,588 sites in β-sheets, α-helices, and loops, respectively, in *D. melanogaster*. RSA values were analyzed with 28 categories with around 85,000 sites each, with the exception of the totally buried residues (RSA = 0) category containing 299,684 sites in *A. thaliana*; and 19 categories with approximately 69,000 residues each, except for 151,417 completely buried residues in *D. melanogaster*. For the analysis of correlation between variables two categories of RSA were considered, comparing buried (RSA <0.05) and exposed (RSA ≥0.05) residues, following Miller et al. (1987).

Estimates of intrinsic protein disorder were acquired via the software DisEMBL (Linding et al. 2003), wherein intrinsic disorder was estimated per site and classified according to the degree of "hot loops," meaning loops with a high degree of mobility. This analysis was successfully achieved for a total of 7,479,807 out of 7,479,808 sites for *A. thaliana* and 3,952,602 out of 4,821,113 sites for *D. melanogaster*. Amino-acid residues were divided into 30 categories with an average of 249,000 and 131,000 sites in *A. thaliana* and *D. melanogaster*, respectively. For the proportion of disordered regions per protein, we considered a residue "disordered" if it was in the top 25% of the measured probabilities of disorder across the proteomes of each species. Analyses were performed with 30 categories with around 620 and 420 genes for *A. thaliana* and *D. melanogaster*, respectively.

Identification of Proteins Binding to a Molecular Chaperone

Prediction of the molecular chaperone *DnaK* binding sites in the protein sequence was estimated with the LIMBO software using the default option *Best overall prediction*. This setting implies 99% specificity and 77.2% sensitivity (Van Durme et al. 2009). Genes were categorized according to this prediction setting, which suggests that every peptide scoring >11.08 is a predicted *DnaK* binder. Genes scoring below that value were not considered as possible binders.

Categorization of Gene Expression

Mean gene expression data were obtained from the database Expression Atlas (http://www.ebi.ac.uk/gxa; last accessed March 2019. Petryszak et al. 2016), wherein one baseline experiment was used for each species (*D. melanogaster*, E-MTAB-4723; *A. thaliana*, E-GEOD-38612). In addition, for *D. melanogaster*, we obtained the breadth of expression data over the embryo anatomy from the BDGP database (Tomancak et al. 2007) and the data were processed and analyzed as in Salvador-Martínez et al. (2018). Mean gene expression levels were obtained by averaging across samples and tissues for each gene, ending up with 40 and 15 categories with around 450 and 430 genes each for *A. thaliana* and *D. melanogaster*, respectively. For the analysis on the breadth of expression, expression patterns in *A. thaliana* were analyzed in four different tissues: roots, flowers, leaves, and siliques; and for *D. melanogaster*, we used the anatomical structures of the embryo development, analyzing 18 structures (see Tomancak et al. 2007 and Salvador-Martínez et al. 2018). Analyses were carried with four and six categories in *A. thaliana* and *D. melanogaster*, respectively, according to the number of tissues/organs a gene is expressed (see supplementary tables S2 and S3 in supplementary file S1, Supplementary Material online, for detailed information).

Protein Cellular Localization and Protein Functional Class

Cellular localization of each protein sequence was predicted with the software ProtComp v9.0 online (from Softberry, http://www.softberry.com/; last accessed May 2018) with the default options and genes were classified into the following cellular categories: cytoplasmic, endomembrane system, mitochondrial, nuclear, peroxisome, plasma membrane, and secreted proteins. The category peroxisome was excluded from further analysis due to the small number of annotated genes (114 and 250 genes in *D. melanogaster* and *A. thaliana*, respectively; detailed information in supplementary tables S2 and S3 in supplementary file S1, Supplementary Material online). Protein functional classes were obtained with the Bioconductor package for R "KEGGREST," using the KEGG BRITE database (Kanehisa et al. 2002). Analysis was carried out with 2,950 and 3,780 genes for *D. melanogaster* and *A. thaliana*, respectively, discretized into the highest levels of each of the three top categories of protein classification: metabolism, genetic information processing and signaling, and cellular processes (see supplementary tables S2 and S3 in supplementary file S1, Supplementary Material online).

Enzymatic Active Sites and PPI

In order to check whether a residue was present in an active site, we used the ScanProsite software (De Castro et al. 2006). Data sets included 1,061,876 and 1,870,166 active sites for *D. melanogaster* and *A. thaliana*, respectively. All sites that were not predicted by the program were considered as nonactive (see supplementary tables S4 and S5 in supplementary file S1, Supplementary Material online). Data on the degree of PPI were obtained with the BioGRID database (Chatr-aryamontri et al. 2017). This was only possible for *D. melanogaster* since the data available for *A. thaliana* was very limited (only 878 annotated genes mapping to our data set). Analyses were carried out with 5,628 genes divided into 19 categories, with 1,114 genes in the first category, and the others ranging from 700 to 130 according to the respective number of interactions (see supplementary tables S2 and S3 in supplementary file S1, Supplementary Material online).

Estimation of the Adaptive and Nonadaptive Rate of Nonsynonymous Substitutions

For all gene and amino-acid sets, 100 bootstrap replicates were generated by randomly sampling genes or sites in each category. The Grapes program was then run on each category and replicate with the Gamma-Exponential DFE (Galtier 2016). The first step included the removal of replicates for which the DFE parameters were not successfully fitted. For this purpose, we discarded 1% in the maximum and minimum values for the mean and shape parameters of the DFE (see supplementary files, Supplementary Material online, for detailed R scripts). Results for ω, ω_{na} and ω_a were plotted using the R package "ggplot2" (Wickham 2017) by taking the mean value and the 95% confidence interval of the 100 bootstrap replicates computed for each category (both for main and supplementary figures, for continuous and discrete variables, see supplementary files, Supplementary Material online).

Statistical Analyses

Significance for all continuous variables, including protein length, number of introns, gene expression, intrinsic residue disorder, proportion of disordered regions, recombination rate, number of PPI, and RSA, was assessed through Kendall's correlation tests. Kendall's correlation test is nonparametric and does not make any assumption on the distribution of the input data. Furthermore, it can be applied to ordinal data, making it appropriate to analyze discretized continuous variables. To do so, the mean value of the 100 bootstrap replicates was taken for each category (see detailed script as well as all statistical results in supplementary file S2, Supplementary Material online). Significance values for discrete variables, comprising binding affinity to *DnaK*, protein location, protein functional class and SS motif, were achieved by estimating the differences between each pair of the categories analyzed, by randomly subtracting each bootstrap replicate. The following steps included counting the number of times the differences between categories were below and above 0, which by taking the minimum of those values gives us a statistic that we call k. The two-tailed *P* value was then estimated by applying the following equation: $P = (2k + 1)/(N + 1)$, where N in the number of bootstrap replicates used. For variables comparing more than two categories, we corrected the *P* value for multiple testing using the FDR method (Benjamini and Hochberg 1995) as implemented in R (R Core Team 2017) (see detailed script and all statistical results in supplementary files S3 and S4, Supplementary Material online). Analyses on the correlations between variables are described in supplementary files S5 and S6, Supplementary Material online. The ANCOVA was performed by applying a linear model to the values of ω_{na} and ω_a with the interaction between RSA and protein disorder following a control for the normality, homoscedasticity, and independence of the corresponding error (supplementary file S5, Supplementary Material online).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

The authors thank Adam Eyre-Walker, Guy Sella, Luis-Miguel Chevin, Hinrich Schulenburg, Tal Dagan, and Chaitanya Gokhale for productive discussions regarding this work. We also thank two anonymous reviewers and David Castellano for constructive comments on the manuscript. We further thank Joel Alves for his writing suggestions on the manuscript, Joost Schymkowitz and Floor Stam for their help with LIMBO software, David Castellano for sharing the recombination rate data used for comparison and Nicolas Galtier for helping with the Grapes software. J.Y.D. acknowledges funding from the Max Planck Society.

References

Adams J, Mansfield MJ, Richard DJ, Doxey AC. 2017. Lineage-specific mutational clustering in protein structures predicts evolutionary shifts in function. *Bioinformatics* 33(9):1338–1345.

Afanasyeva A, Bockwoldt M, Cooney CR, Heiland I, Gossmann TI. 2018. Human long intrinsically disordered protein regions are frequent targets of positive selection. *Genome Res.* 28(7):975–982.

Akaike H. 1973. Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika* 60(2):255–265.

Akashi H. 2003. Translational selection and yeast proteome evolution. *Genetics* 164(4):1291–1303.

Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. 2002. Analysis of catalytic residues in enzyme active sites. *J Mol Biol*. 324(1):105–121.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc*. 57(1):289–300.

Bergman J, Eyre-Walker A. 2019. Does adaptive protein evolution proceed by large or small steps at the amino acid level? *Mol Biol Evol*. 36(5):990–998.

Bierne N, Eyre-Walker A. 2004. The genomic rate of adaptive amino acid substitution in *Drosophila*. *Mol Biol Evol*. 21(7):1350–1360.

Bogumil D, Dagan T. 2010. Chaperonin-dependent accelerated substitution rates in prokaryotes. *Genome Biol Evol*. 2(1):602–608.

Bowie JU, Reidhaar-Olson JF, Lim WA, Sauer RT. 1990. Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science* 247(4948):1306–1310.

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, et al. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet*. 4(5):e1000083.

Bustamante CD, Townsend JP, Hartl DL. 2000. Solvent accessibility and purifying selection within proteins of *Escherichia coli* and *Salmonella enterica*. *Mol Biol Evol*. 17(2):301–308.

Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES. 2004. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci*. 13(1):190–202.

Campos JL, Halligan DL, Haddrill PR, Charlesworth B. 2014. The relation between recombination rate and patterns of molecular evolution and variation in *Drosophila melanogaster*. *Mol Biol Evol*. 31(4):1010–1028.

Carneiro M, Albert FW, Melo-Ferreira J, Galtier N, Gayral P, Blanco-Aguiar JA, Villafuerte R, Nachman MW, Ferrand N. 2012. Evidence for widespread positive and purifying selection across the European rabbit (*Oryctolagus cuniculus*) genome. *Mol Biol Evol*. 29(7):1837–1849.

Castellano D, Coronado-Zamora M, Campos JL, Barbadilla A, Eyre-Walker A. 2016. Adaptive evolution is substantially impeded by hill-Robertson interference in *Drosophila*. *Mol Biol Evol*. 33(2):442–455.

Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. 2002. Selection for short introns in highly expressed genes. *Nat Genet*. 31(4):415–418.

Charlesworth B. 1994. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet Res*. 63(3):213–227.

Charlesworth J, Eyre-Walker A. 2006. The rate of adaptive evolution in enteric bacteria. *Mol Biol Evol*. 23(7):1348–1356.

Chatr-aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, O'Donnell L, Oster S, Theesfeld C, Sellam A, et al. 2017. The BioGRID interaction database: 2017 update. *Nucleic Acids Res*. 45(D1):D369–D379.

Choi SC, Hobolth A, Robinson DM, Kishino H, Thorne JL. 2007. Quantifying the impact of protein tertiary structure on molecular evolution. *Mol Biol Evol*. 24(8):1769–1782.

Choi SS, Vallender EJ, Lahn BT. 2006. Systematically assessing the influence of 3-dimensional structural context on the molecular evolution of mammalian proteomes. *Mol Biol Evol*. 23(11):2131–2133.

Coghlan A, Wolfe KH. 2000. Relationship of codon bias to mRNA and concentration protein length in *Saccharomyces cerevisiae*. *Yeast* 16(12):1131–1145.

Collins CA, Brown EJ. 2010. Cytosol as battleground: ubiquitin as a weapon for both host and pathogen. *Trends Cell Biol*. 20(4):205–213.

Conant GC, Stadler PF. 2009. Solvent exposure imparts similar selective pressures across a range of yeast proteins. *Mol Biol Evol*. 26(5):1155–1161.

Dangl JL, Jones JD. 2001. Plant pathogens and integrated defence responses to infection. *Nature* 411(6839):826–833.

De Castro E, Sigrist CJA, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, Bairoch A, Hulo N. 2006. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res*. 34(Web Server):W362–W365.

Dean AM, Neuhauser C, Grenier E, Golding GB. 2002. The pattern of amino acid replacements in α/β-barrels. *Mol Biol Evol*. 19(11):1846–1864.

Dielen AS, Badaoui S, Candresse T, German-Retana S. 2010. The ubiquitin/26S proteasome system in plant-pathogen interactions: a never-ending hide-and-seek game. *Mol Plant Pathol*. 11(2): 293–308.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A*. 102(40):14338–14343.

Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradović Z. 2002. Intrinsic disorder and protein function. *Biochemistry* 41(21):6573–6582.

Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol*. 17(1):68–74.

Dutheil JY, Gaillard S, Stukenbrock EH. 2014. MafFilter: a highly flexible and extensible multiple genome alignment files processor. *BMC Genomics* 15(1):53.

Dyson HJ, Wright PE. 2005. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol*. 6(3):197–208.

Enard D, Cai L, Gwennap C, Petrov DA. 2016. Viruses are a dominant driver of protein adaptation in mammals. *Elife* 5:e12469.

Eyre-Walker A. 2002. Changing effective population size and the McDonald-Kreitman test. *Genetics* 162(4):2017–2024.

Eyre-Walker A. 2006. The genomic rate of adaptive evolution. *Trends Ecol Evol*. 21(10):569–575.

Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol Biol Evol*. 26(9):2097–2108.

Eyre-Walker A, Woolfit M, Phelps T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* 173(2):891–900.

Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. *Genetics* 158(3):1227–1234.

Franzosa EA, Xia Y. 2009. Structural determinants of protein evolution are context-sensitive at the residue level. *Mol Biol Evol*. 26(10):2387–2395.

Galtier N. 2016. Adaptive protein evolution in animals and the effective population size hypothesis. *PLoS Genet*. 12(1):e1005774.

Goldman N, Thorne JL, Jones DT. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* 149(1):445–458.

Gossmann TI, Keightley PD, Eyre-Walker A. 2012. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. *Genome Biol Evol*. 4(5):658–667.

Gossmann TI, Song BH, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov MV, Filatov DA, Eyre-Walker A. 2010. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Mol Biol Evol*. 27(8):1822–1832.

Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves L. 2006. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* 22(21):2695–2696.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. *Science* 185(4154):862–864.

Gremme G, Steinbiss S, Kurtz S. 2013. Genome tools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans Comput Biol Bioinform*. 10(3):645–656.

Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette NC, Bigot T, Fournier D, Pouyet F, Cahais V, et al. 2013. Bio++: efficient extensible libraries and tools for computational molecular evolution. *Mol Biol Evol*. 30(8):1745–1750.

Guo HH, Choe J, Loeb LA. 2004. Protein tolerance to random amino acid change. *Proc Natl Acad Sci U S A*. 101(25):9205–9210.

Haddrill PR, Loewe L, Charlesworth B. 2010. Estimating the parameters of selection on nonsynonymous mutations in *Drosophila pseudoobscura* and *D. miranda*. *Genetics* 185(4):1381–1396.

Haerty W, Jagadeeshan S, Kulathinal RJ, Wong A, Ram KR, Sirot LK, Levesque L, Artieri CG, Wolfner MF, Civetta A, et al. 2007. Evolution in the fast lane: rapidly evolving sex-related genes in Drosophila. *Genetics* 177(3):1321–1335.

Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD. 2010. Evidence for pervasive adaptive protein evolution in wild mice. *PLoS Genet*. 6(1):e1000825.

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genet Res*. 8(3):269–294.

Hiroshi A, Minsoo K, Chihiro S. 2014. Exploitation of the host ubiquitin system by human bacterial pathogens. *Nat Rev Microbiol*. 12(1):399–413.

Hunter T. 1995. Protein kinases and phosphatases: the Yin and Yang of protein phosphorylation and signaling. *Cell* 80(2):225–236.

Hvilsom C, Qian Y, Bataillon T, Li Y, Mailund T, Salle B, Carlsen F, Li R, Zheng H, Jiang T, et al. 2012. Extensive X-linked adaptive evolution in central chimpanzees. *Proc Natl Acad Sci U S A*. 109(6):2054–2059.

Ingvarsson PK. 2010. Natural selection on synonymous and nonsynonymous mutations shapes patterns of polymorphism in *Populus tremula*. *Mol Biol Evol*. 27(3):650–660.

Janin J. 1979. Surface and inside volumes in globular proteins. *Nature* 277(5696):491.

Julenius K, Pedersen AG. 2006. Protein evolution is faster outside the cell. *Mol Biol Evol*. 23(11):2039–2048.

Kadibalban AS, Bogumil D, Landan G, Dagan T. 2016. DnaK-dependent accelerated evolutionary rate in prokaryotes. *Genome Biol Evol*. 8(5):1590–1599.

Kanehisa M, Goto S, Kawashima S, Nakaya A. 2002. The KEGG databases at GenomeNet. *Nucleic Acids Res*. 30(1):42–46.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 30(4):772–780.

Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. *Mol Biol Evol*. 22(5):1345–1354.

Liao BY, Scott NM, Zhang J. 2006. Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. *Mol Biol Evol*. 23(11):2072–2080.

Liberles D, Teichmann S, Bahar I, Bastolla U, Bloom J, Bornberg-Bauer E, Colwell LJ, De Koning APJ, Dokholyan NV, Echave J, et al. 2012. The interface of protein structure, protein biophysics, and molecular evolution. *Protein Sci*. 21(6):769–785.

Lin YS, Hsu WL, Hwang JK, Li WH. 2007. Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. *Mol Biol Evol*. 24(4):1005–1011.

Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. 2003. Protein disorder prediction: implications for structural proteomics. *Structure* 11(11):1453–1459.

Lipman DJ, Souvorov A, Koonin EV, Panchenko AR, Tatusova TA. 2002. The relationship of protein conservation and sequence length. *BMC Evol Biol*. 2(1):20.

Loureiro J, Ploegh HL. 2006. Antigen presentation and the ubiquitin-proteasome system in host–pathogen interactions. *Adv Immunol*. 92:225

Marais G, Charlesworth B. 2003. Genome evolution: recombination speeds up adaptive evolution. *Curr Biol*. 13(2):68–70.

Mauch-Mani B, Baccelli I, Luna E, Flors V. 2017. Defense priming: an adaptive part of induced resistance. *Annu Rev Plant Biol*. 68(1):485–512.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution ate the Adh locus in Drosophila. *Nature* 351(6328):652–654.

Miller S, Lesk AM, Janin J, Chothia C. 1987. The accessible surface area and stability of oligomeric proteins. *Nature* 328(6133):834.

Mintseris J, Weng Z. 2005. Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc Natl Acad Sci U S A*. 102(31):10930–10935.

Mirny LA, Shakhnovich EI. 1999. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J Mol Biol*. 291(1):177–196.

Miyata T, Miyazawa S, Yasunaga T. 1979. Two types of amino acid substitutions in protein evolution. *J Mol Evol*. 12(3):219–236.

Mosmann VR, Cherwinski H, Bond MW, Giedlin MA, Coffman RL. 2016. Two types of murine helper T cell clone. I. Definition according to profiles of lymphokine activities and secreted proteins. *J Immunol*. 136(7):2348–2357.

Motion GB, Amaro T, Kulagina N, Huitema E. 2015. Nuclear processes associated with plant immunity and pathogen susceptibility. *Brief Funct Genomics*. 14(4):243–252.

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol*. 3(6):0976–0985.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148(3):929–936.

Obbard DJ, Welch JJ, Kim KW, Jiggins FM. 2009. Quantifying adaptive evolution in the Drosophila immune system. *PLoS Genet*. 5(10):e1000698.

Overington J, Donnelly D, Johnson MS, Sali A, Blundell TL. 1992. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci*. 1(2):216–226.

Pal C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. *Genetics* 158(1998):927–931.

Perutz MF, Kendrew JC, Watson HC. 1965. Structure and function of haemoglobin: II. Some relations between polypeptide chain configuration and amino acid sequence. *J Mol Biol*. 13(3):669–678.

Petryszak R, Keays M, Tang YA, Fonseca NA, Barrera E, Burdett T, Füllgrabe A, Fuentes AMP, Jupp S, Koskinen S, et al. 2016. Expression Atlas update – an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res*. 44(D1):D746–D752.

Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, Crepeau MW, Duchen P, Emerson JJ, Saelao P, Begun DJ, et al. 2012. Population genomics of sub-Saharan *Drosophila melanogaster*: African diversity and non-African admixture. *PLoS Genet*. 8(12):e1003080.

Pröschel M, Zhang Z, Parsch J. 2006. Widespread adaptive evolution of Drosophila genes with sex-biased expression. *Genetics* 174(2):893–900.

Proux E, Studer RA, Moretti S, Robinson-Rechavi M. 2009. Selectome: a database of positive selection. *Nucleic Acids Res*. 37(1):404–407.

Ramsey DC, Scherrer MP, Zhou T, Wilke CO. 2011. The relationship between relative solvent accessibility and evolutionary rate in protein evolution. *Genetics* 188(2):479–488.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rezvoy C, Charif D, Guéguen L, Marais G. 2007. MareyMap: an R-based tool with graphical interface for estimating recombination rates. *Bioinformatics* 23(16):2188–2189.

Rocha EPC, Danchin A. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol Biol Evol*. 21(1):108–116.

Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG. 2007. Dynamic evolution of the innate immune system in Drosophila. *Nat Genet*. 39(12):1461–1468.

Salvador-Martínez I, Coronado-Zamora M, Castellano D, Barbadilla A, Salazar-Ciudad I. 2018. Mapping selection within *Drosophila melanogaster* embryo's anatomy. *Mol Biol Evol*. 35(1):66–79.

Sawyer SA, Kulathinal RJ, Bustamante CD, Hartl DL. 2003. Bayesian analysis suggests that most amino acid replacements in drosophila are driven by positive selection. *J Mol Evol*. 57(0):S154–S164.

Schaffer AA. 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res*. 29(14):2994–3005.

Schuler MA, Werck-Reichhart D. 2003. Functional genomics of P450S. *Annu Rev Plant Biol*. 54(1):629–667.

Slotte T, Bataillon T, Hansen TT, St. Onge K, Wright SI, Schierup MH. 2011. Genomic determinants of protein evolution and polymorphism in *Arabidopsis*. *Genome Biol Evol*. 3(1):1210–1219.

Slotte T, Foxe JP, Hazzouri KM, Wright SI. 2010. Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size. *Mol Biol Evol*. 27(8):1813–1821.

Smith NGC, Eyre-Walker A. 2002. Adaptive protein evolution in Drosophila. *Nature* 415(6875):1022.

Stoletzki N, Eyre-Walker A. 2011. Estimation of the neutrality index. *Mol Biol Evol*. 28(1):63–70.

Strasburg JL, Kane NC, Raduski AR, Bonin A, Michelmore R, Rieseberg LH. 2011. Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. *Mol Biol Evol*. 28(5):1569–1580.

Subramanian S, Kumar S. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics* 168(1):373–381.

Tange O. 2011. GNU parallel – the command-line power tool. *USEUNIX Mag*. 36(1):42–47.

Tataru P, Mollion M, Glémin S, Bataillon T. 2017. Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. *Genetics* 207(3):1103–1119.

Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. 2013. Maximum allowed solvent accessibilities of residues in proteins. *PLoS One* 8(11):e80635.

Tomancak P, Berman BP, Beaton A, Weiszmann R, Kwan E, Hartenstein V, Celniker SE, Rubin GM. 2007. Global analysis of patterns of gene expression during Drosophila embryogenesis. *Genome Biol*. 8(7):R145.

Trujillo M, Shirasu K. 2010. Ubiquitination in plant immunity. *Curr Opin Plant Biol*. 13(4):402–408.

Valdar WSJ, Thornton JM. 2001. Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins Struct Funct Genet*. 42(1):108–124.

Van Durme J, Maurer-Stroh S, Gallardo R, Wilkinson H, Rousseau F, Schymkowitz J. 2009. Accurate prediction of DnaK-peptide binding via homology modelling and experimental data. *PLoS Comput Biol*. 5(8):e1000475.

Weigel D, Mott R. 2009. The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biol*. 8(12):e1003080.

Wickham H. 2017. ggplot2: elegant graphics for data analysis. *J Stat Softw*. 35(1):65–88.

Wilke CO, Bloom JD, Drummond DA, Raval A. 2005. Predicting the tolerance of proteins to random amino acid substitution. *Biophys J*. 89(6):3714–3720.

Wright SI, Yau CBK, Looseley M, Meyers BC. 2004. Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Mol Biol Evol*. 21(9):1719–1726.

Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol*. 22(4):1107–1118.

Zhang J. 2000. Protein-length distributions for the three domains of life. *Trends Genet*. 16(3):107–109.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol*. 22(12):2472–2479.