Testing the adaptive walk model of gene evolution

Ana Filipa Moutinho*,1,2, Adam Eyre-Walker2, Julien Y. Dutheil1,3

1Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Biology, Plön, Germany
2School of Life Sciences, University of Sussex, Brighton, United Kingdom
3Unité Mixte de Recherche 5554 Institut des Sciences de l’Evolution, CNRS, IRD, EPHE, Université de Montpellier, Montpellier, France

*Corresponding author:
Ana Filipa Moutinho

Current address: School of Life Sciences, University of Sussex, Brighton, United Kingdom
E-Mail: a.f.moutinho@sussex.ac.uk

Author Contributions:
Conceptualization: AFM (equal), AEW (equal), JYD (equal)
Data curation: AFM (lead)
Formal analysis: AFM (lead)
Investigation: AFM (lead), AEW (contributing), JYD (contributing)
Methodology: AFM (equal), JYD (equal), AEW (contributing)
Project administration: JYD (lead)
Software: AFM (lead)
Supervision: JYD (lead)
Visualization: AFM (lead)
Writing – original draft: AFM (lead), JYD (contributing)
Writing – review & editing: AFM (equal), AEW (equal), JYD (equal)

Preprint servers: bioRxiv (doi: https://doi.org/10.1101/2021.04.28.441765)

Classification: Major - Biological Sciences; Minor – Evolution, Genetics

Keywords: adaptive walk, molecular evolution, adaptive evolution, gene age, distribution of fitness effects

This PDF file includes:
- Main text
- Figures 1 to 3
- Table 1
Abstract

Understanding the dynamics of species adaptation to their environments has long been a central focus of the study of evolution. Theories of adaptation propose that populations evolve by "walking" in a fitness landscape. This "adaptive walk" is characterised by a pattern of diminishing returns, where populations further away from their fitness optimum take larger steps than those closer to their optimal conditions. Hence, we expect young genes to evolve faster and experience mutations with stronger fitness effects than older genes because they are further away from their fitness optimum. Testing this hypothesis, however, constitutes an arduous task. Young genes are small, encode proteins with a higher degree of intrinsic disorder, are expressed at lower levels, and are involved in species-specific adaptations. Since all these factors lead to increased protein evolutionary rates, they could be masking the effect of gene age. While controlling for these factors, we used population genomic datasets of Arabidopsis and Drosophila and estimated the rate of adaptive substitutions across genes from different phylostrata. We found that a gene's evolutionary age significantly impacts the molecular rate of adaptation. Moreover, we observed that substitutions in young genes tend to have larger physicochemical effects. Our study, therefore, provides strong evidence that molecular evolution follows an adaptive walk model across a large evolutionary timescale.

Introduction

How does adaptive evolution proceed in space and in time? This question has long intrigued evolutionary biologists. Fisher [1] proposed that adaptation relies on mutations with small effect sizes at the phenotypic level. He presented the geometric model of adaptation where phenotypic evolution occurs continuously and gradually toward an optimum fitness [1]. At the molecular level, Wright [2,3] first introduced the idea that populations evolve in the space of all possible gene combinations to acquire higher fitness. He characterised this model of evolution as a "walk" in an adaptive landscape. Wright consequently proposed the shifting balance theory of adaptation, which combines the effects of drift and selection. Drift acts by moving the population away from its local peak, while natural selection directs the population to higher fitness, the so-called "global optimum" in the fitness landscape. With the rise of molecular genetics, Maynard Smith [4] extended this idea to a sequence-based model of adaptation. He introduced the idea 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. Wright's and Maynard Smith's adaptation model was further extended by Gillespie [5–7], who presented the "move rule" in an adaptive landscape. Following Kimura's work on the effect sizes of mutations [8], Gillespie suggested that adaptation proceeds in large steps, where mutations with higher fitness effects are more likely to reach fixation. The adaptive walk model was later fully developed by Allen Orr [9,10], who extended Fisher's geometric model of adaptation and showed that adaptive walks lead to a pattern of diminishing returns. A sequence further away from its local optimum tends to accumulate large-effect mutations at the beginning of the walk, small-effect mutations being then
only fixed when the sequence is approaching its optimum fitness. Therefore, under this model, adaptation relies both on mutations of large and small fitness effects, in a time-dependent manner. Experimental studies tracing the evolution of bacteria [11–14] and fungi [15] have supported this model. Experimental studies, however, can only assess patterns of adaptation at relatively short time scales in artificial environments and are limited to certain organisms. The challenge lies in studying adaptation across long evolutionary time scales: how does the distribution of beneficial mutations vary over time?

While long-term evolutionary processes are not directly observable, they leave a signature in genome sequences. Species share genes to variable extents, according to their degree of divergence. Some ancient genes, which evolved early, are shared by many species. Others, which evolved more recently, are only shared by a few related species [16,17]. Consequently, the age of a gene can be inferred from its phyletic pattern, that is, its presence or absence across the phylogeny [18]. These phyletic patterns are reconstructed using sequence similarity searches performed by tools like BLAST [19]. 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 recognised homologs are only present in closely related species. This approach is known as phylolstratigraphy [20].

Previous studies suggested that young or lineage-specific protein-coding genes evolve faster than older ones [17,21–27]. Albà and Castresana [27] showed a negative correlation between the ratio of the non-synonymous \( d_M \) to synonymous \( d_S \) substitutions rates, \( \omega \), and gene age in the divergence between humans and mice, with young genes presenting a higher \( \omega \). Similar correlations between \( \omega \) and gene age have been observed in primates [22], fungi [25], *Drosophila* [23,28,29], bacteria [30], viruses [31], plants [32,33], and protozoan parasites [34]. Despite the observed consistency across taxa, the underlying causal mechanisms remain debated [21]. As the average \( d_M / d_S \) ratio integrates both positive and negative selection, a comparatively higher ratio can result from a less stringent purifying selection, the occurrence of positive selection, or both. By looking at polymorphism data, Cai and Petrov [22] suggested that the faster evolution in young primate genes may be due to the lack of selective constraint posed by purifying selection, while analyses of adaptive evolution were inconclusive.

Analysing the effect of gene age on adaptive evolution is a complex task, as young and old genes differ in their structural properties, expression level, and protein function. Young genes tend to be smaller [22,24,35], have a higher degree of intrinsic disorder [36], and are expressed at lower levels [17,22,24,26]. Moreover, young genes tend to encode proteins involved in developing species-specific characteristics and immune and stress responses [16,37,38]. As the macromolecular structure [39,40], gene expression levels [39,41] and protein function [39,42,43] are known determinants of the rate of protein adaptation, they could be confounding 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 adaptation rates [39,41,44,45]. At the
macromolecular level, some studies showed that highly disordered [39,40] and exposed residues [39]
present higher rates of adaptive evolution. Finally, there is evidence that proteins involved in the
immune and stress response have higher molecular adaptive rates [39,43,46,47]. Thus, controlling for
these confounding factors when assessing the impact of gene age on the rate of molecular adaptation is
crucial.

Here, we used a population genomic approach to test the adaptive walk model. We make two
predictions: first, that younger genes are undergoing faster rates of adaptive evolution, and second, the
evolutionary steps they make are larger. We tested the first prediction by estimating rates of adaptive
and non-adaptive protein evolution using an extension of the MacDonald-Kreitman (MK) test [48],
which uses counts of polymorphism and substitution at selected and neutral sites. We quantified the
rates of adaptive and non-adaptive evolution using the statistics $\omega_a$ and $\omega_{na}$, which denote the rates of
adaptive and non-adaptive non-synonymous substitution relative to the mutation rate. We investigated
whether protein length, gene expression, relative solvent accessibility (RSA), intrinsic protein
disorder, BLAST’s false-negative rate, and protein function act as confounding factors of the effect of
gene age. To test the second prediction, we considered the rates of substitution between amino acids
separated by different physicochemical distances as a function of gene age. We tested our hypotheses
in two pairs of species with different life-history traits: the Diptera Drosophila melanogaster and D.
simulans and the Brassica Arabidopsis thaliana and A. lyrata. In each species pair, we compared their
most recent genes with those dating back to the origin of cellular organisms.

Results

Young genes have a higher rate of adaptive substitutions

We tested the adaptive walk model of sequence evolution by assessing the impact of gene age on the
rate of adaptive ($\omega_a$) and non-adaptive ($\omega_{na}$) non-synonymous substitutions. We used Grapes [48] to
estimate $\omega_a$ and $\omega_{na}$, which accounts for segregating slightly deleterious and advantageous mutations
by specifically modelling the distribution of fitness effects (DFE). Gene age data were obtained from
published data sets [28,33]. We found that gene age is significantly correlated to estimates of $\omega_a$, $\omega_{na}$
and $\omega_{na}$ in both species’ pairs (Table 1 and Figure 1b). This result suggests that the higher $\omega$ ratio of
more recently evolved genes is due to a higher rate of adaptive and non-adaptive non-synonymous
substitutions. As X-linked genes are known to evolve faster due to the male hemizygosity [49–51], we
assessed whether the relationship between evolutionary rates and gene age differed between
chromosomes in Drosophila (Figure 1b). We compared models with and without the chromosome’s
effect (see Material and Methods and supplementary file S1) and found only low support for a
chromosomal effect ($p = 0.041$ for $\omega_{na}$ and $p = 0.094$ for $\omega_a$). We, therefore, combined data from all
chromosomes for subsequent analyses.

The effect of gene age on the rate of molecular adaptation is robust to multiple confounding factors.
Genes of different ages intrinsically differ in their features [16,22,36]. As such traits significantly impact the rate of molecular evolution [39], they may be confounding the faster adaptive rates observed in young genes. Previous studies reported that younger genes encode shorter proteins [24,35,52] and are expressed at lower levels [17,22,24,26], a pattern that we also observed in our data set (gene age vs. protein length for D. melanogaster and A. thaliana, respectively: Kendall’s $\tau = -0.485$, $p = 2.82e-02$; $\tau = -0.848$, $p = 1.06e-05$, Figure S1a; gene age vs. gene expression for D. melanogaster and A. thaliana, respectively: $\tau = -0.595$, $p = 7.35e-03$; $\tau = -0.790$, $p = 4.00e-05$, Figure S1b in supplementary data). As younger proteins are shorter than older ones, they have a higher proportion of exposed residues [39]: gene age is significantly positively correlated with the average RSA per gene ($\tau = 0.636$, $p = 3.98e-03$; $\tau = 0.695$, $p = 3.03e-04$, for D. melanogaster and A. thaliana respectively; Figure S2a in supplementary data). Because exposed residues are more flexible [53], young genes tend to encode proteins with a higher degree of intrinsic disorder, a pattern previously reported in mice [36]. We confirmed this pattern in D. melanogaster ($\tau = 0.606$, $p = 6.10e-03$; Figure S2b in supplementary data) and A. thaliana ($\tau = 0.467$, $p = 1.53e-02$; Figure S2b in supplementary data). Given that $\omega_n$ and $\omega_{na}$ are negatively correlated to length and gene expression, and positively correlated to RSA and protein disorder [39], these could be driving the effect of gene age on $\omega_n$ and $\omega_{na}$; i.e., the correlation between $\omega_n$ and $\omega_{na}$ and age could be incidental.

To control for the impact of each confounding factor using Grapes, we split our data into two roughly equal-sized groups according to protein length, expression level, average RSA, and average intrinsic disorder and reran the analysis within the “high” and “low” groups, combining probabilities from the two analyses using the weighted Z-method [54]. As estimates of the rate of adaptive substitutions for a small number of genes exhibit large sampling variances [55,56], each putative confounding factor could only be tested separately. Some phylostrata were further combined when under-represented in some gene categories (see Material and Methods). We found that $\omega$, $\omega_{na}$ and $\omega_n$ remain significantly correlated to gene age, except when controlling for protein length and gene expression for $\omega_n$ in Arabidopsis (Figure 2 and Table 1). This weaker effect may be a consequence of how the most recent clades were combined in these analyses, as there was little data available for those genes (see Material and Methods). Nonetheless, when combining probabilities across the two species, we observed a significant correlation between all measures of evolutionary rate and gene age controlling each of the co-factors (Table 1).

We next asked whether the observed effect could be explained by the residual correlation of the cofactors with gene age within each “high” and “low” group (Table S1 in supplementary data). We performed three different tests: (1) the correlation between $\omega_n$ and the mean value of each co-factor in each age class within the “high” and “low” groups; (2) the correlation between the co-factor and gene age in each “high” and “low” group; and (3) a linear model where $\omega_n$ is the response variable, and gene age, category, species (Arabidopsis or Drosophila), and the within category cofactor values are explanatory variables. While we do observe a general correlation of the co-factor with gene age within...
each group, we found that nearly all correlations between $\omega_a$ and the co-factor are non-significant (Table S1). Moreover, the linear model analyses showed that in all cases but gene length, gene age is the most significant variable, with an effect consistent with that observed in Table 1. For gene length, gene age was only significant as an interaction with the species variable, having a positive effect only in *Drosophila* (Table S2). These findings, therefore, suggest that the effect of gene age on $\omega_a$ is independent of the co-factor.

To jointly estimate the effect of the potential confounding factors, we applied a recently developed method that extends the MK test with a generalized linear model [57]. This approach disentangles the effects of each factor on the rate of adaptive substitutions per nucleotide site. However, this method does not model the distribution of fitness effects and hence cannot account for segregating slightly deleterious mutations, which can bias estimates of the rate of adaptive substitutions [58]. Hence, following the approach suggested in Huang [57], we removed sites for which the derived allele frequency was below 50% to minimize any potential bias. Despite the large reduction in the data set, this analysis revealed a significant effect of gene age (Table S3 in supplementary data). Our findings, therefore, suggest that the effect of gene age on rates of protein evolution is robust to the tested confounding factors and that a gene's age acts as a significant determinant of the rate of adaptive and non-adaptive evolution in both species.

Lastly, we aimed at assessing the effect size of gene age on $\omega_a$ relative to other factors. Because correlation coefficients were computed from values averaged over multiple genes and genes were categorized differently for each analysis, the comparison of correlation coefficients does not provide a reliable estimate of relative effect sizes. Therefore, we assessed the relative contribution of each variable to the MK regression model by computing likelihood pseudo-$R^2$'s [59]. These were estimated by comparing the log-likelihood of the full model (all factors included) with the log-likelihood of reduced models, where each single factor was removed independently from the full model. These estimates suggest that, despite its very significant effect, gene age contributes comparatively little to the explained variance compared to factors such as protein function (Table S4).

To note, however, that these estimates refer to the relative contribution of each variable per nucleotide site and not per gene as this is the unit used by the MK regression, which results in generally low $R^2$ values.

The effect of gene age on the molecular rate of adaptation is robust to BLAST’s false negative rates. The phylostratigraphy approach has been previously used to date the emergence of new genes, and some studies have pointed out its potential limitations [28,60–63]. Because BLAST homology searches might fail to identify homologs in short or rapidly evolving genes, such genes could be mistakenly classified as young. As expected, we observed that genes in younger phylostrata present higher E-values in both species ($\tau = 0.564, p = 0.025; \tau = 0.951, p = 1.20e-06$, for *D. melanogaster* and *A. thaliana* respectively; Figure S3a in supplementary data). To control this effect, we reran our
analyses using genes with very low E-values that are likely to be detected in most age strata and for which the correlation with gene age was no longer significant (see Material and Methods and supplementary file S2) (τ = 0.408, p = 0.111; τ = 0.354, p = 0.141, for *D. melanogaster* and *A. thaliana* respectively; Figure S3b in supplementary data). We observed that the effect of gene age remained for all estimates in the two species when using this restricted data set (ω: τ = 0.929, p = 1.30e-03; ωna: τ = 0.786, p = 6.49e-03; ωa: τ = 0.643, p = 2.59e-02 in *A. thaliana*; and ω: τ = 0.697, p = 1.61e-03; ωna: τ = 0.636, p = 3.98e-03; ωa: τ = 0.636, p = 3.98e-03 in *D. melanogaster*, Figure S4 in supplementary data). These results suggest that the correlation of gene age with the rate of adaptive evolution cannot be attributed to errors in dating the emergence of a gene stemming from the failure of identifying homologs in older taxa.

The effect of gene age on the rate of molecular adaptation does not depend on protein function

Lineage-specific genes are known to be involved in species-specific adaptive processes, such as the evolution of morphological diversity [64] and immune and stress responses [17,34,64]. As proteins encoding such functions tend to have higher molecular rates of adaptation [39,42,43,46,47,65], we further assessed whether the observed effect of gene age could be due to younger genes being enriched in functions with higher evolutionary rates. We first examined which functions are encoded by young genes in *A. thaliana* and *D. melanogaster*. In *A. thaliana*, young genes (Clades 12 to 15 in Figure 1a) are mostly involved in a large variety of cellular processes, stress response and external stimulus, protein binding, and signal transduction (Figure S5a in supplementary data). In *D. melanogaster*, young genes (Clades 11 and 12 in Figure 1a) encode mostly functions involved in the cell's anatomic structure, stress response, nervous system processes, enzyme regulators, immune system mechanisms, and a wide range of metabolic processes (Figure S5b in supplementary data).

To further correct for the potential bias of protein function, we ran Grapes across categories of GO-annotated genes while simultaneously controlling for the effect of gene age. As some gene functions were biased towards some age categories (Figure S5), we could not do this analysis for all GO terms. We, therefore, only used the GO terms with a sufficient number of annotated genes in each age class (see Material and Methods). To assess the effect of gene age, we grouped the genes in each GO category into three age groups, trying to keep a similar number of genes in each (see Material and Methods). In *A. thaliana*, we found that the impact of gene age on ωa is stronger in proteins linked to stress response and cellular components, where younger genes present higher molecular adaptive rates (Figure S6a and supplementary file S3 in supplementary data). Although the GO term cellular component represents a comprehensive annotation, it denotes the cellular compartments where processes such as signal transduction and membrane trafficking occur, essential for maintaining the cell homeostasis [66,67]. In *D. melanogaster*, we observed a strong effect of gene age on ωa for proteins encoding chromosomal organisation, protein complex, stress response, signal transduction, and involved in the cell cycle (Figure S6b and supplementary file S3 in supplementary data in...
supplementary data). Even though these functions cover a wide range of molecular processes, they are involved in DNA replication, genome stability, and immune and stress responses, which are critical functions for the co-evolutionary arms race between hosts and parasites [47]. When looking at the non-adaptive rate of evolution, our analyses revealed a strong influence of gene age in $\omega_{na}$ within most functions analysed in both species, where young genes present higher rates of non-adaptive substitutions (Figure S6 and supplementary file S3 in supplementary data in supplementary data).

These results suggest that, when restricting the analysis to proteins involved in defence mechanisms, which are known to adapt faster [42,43,47,65], gene age still has an impact on the efficiency of selection acting upon a protein.

**Substitutions in young genes have larger effect sizes**

Our second prediction under the adaptive walk model is that substitutions in young genes should have larger fitness effects than in older genes. To test this prediction, we used Grantham’s physicochemical distances between amino acids [68] as a proxy for the fitness effects of amino-acid mutations. We looked at the fixed differences separated by one mutational step between each pair of species (i.e., fixed differences in the ingroup when compared to the outgroup species) and reported the average Grantham’s distances between residues within each age stratum. We observed that substitutions in young genes tend to occur between less biochemically similar residues (Arabidopsis: $\tau = 1$, $p = 2.00\times 10^{-07}$; Drosophila: $\tau = 0.788$, $p = 3.628\times 10^{-04}$; Figure 3 and supplementary file S4), suggesting that substitutions in these genes have larger physicochemical effects than in old ones. To further test whether these larger effects had a positive or negative impact on younger genes, we estimated the average Grantham’s distance amongst adaptive ($\bar{G}_a$) and non-adaptive ($\bar{G}_{na}$) non-synonymous substitutions (see Material and Methods and supplementary file S4). We expect a positive correlation between $\bar{G}_a$ and $\bar{G}_{na}$ and gene age, i.e., the average Grantham’s distance for adaptive and non-adaptive substitutions should increase towards the present. This is indeed what we find: we observed a positive correlation between $\bar{G}_a$ and gene age in both species, which is individually significant in Drosophila ($p = 0.006$) and significant across species when we combine probabilities ($p = 0.009$) (Figure S7). We observed the same pattern in the analysis of $\bar{G}_{na}$ with a significant correlation in Arabidopsis ($p = 0.004$) and a significant correlation when combining the probabilities from both species ($p = 0.0007$) (Figure S7).

**Discussion**

We used a population genomic approach to disentangle the effects of positive and negative selection on the rate of non-synonymous substitutions. Using complete genome data from two Arabidopsis and Drosophila species, we showed that the higher rate of non-synonymous substitutions in younger genes results both from relaxed purifying selection (higher $\omega_{na}$) and a higher rate of adaptive substitutions (higher $\omega_a$) (Figure 1b). Although gene age is not the variable contributing most to the rate of
molecular adaptation (Table S4), its effect is consistent across comparisons and robust to multiple confounding factors (Figure 2, Tables 1, S2 and S3). Moreover, we observed that young genes undergo substitutions that are larger in terms of physicochemical properties than older genes. A question remains: what are the drivers of these effects?

The magnitude effect of gene age on adaptive evolution is species-specific

Although we observed a strong impact of gene age on the molecular adaptive rate in both species pairs, the shape of the relation between these two variables differs. While the relationship between gene age and \( \omega_a \) is monotonously increasing in Arabidopsis, it has several peaks in Drosophila (Figure 1b). This pattern is particularly evident if we discard the two youngest clades. In Drosophila, the correlation becomes much weaker and non-significant for \( \omega_a \) (\( \omega: \tau = 0.600, p = 0.016; \omega_{na}: \tau = 0.556, p = 0.025; \omega_a: \tau = 0.467, p = 0.060 \)), whereas, in Arabidopsis, the effect of gene age persists (\( \omega: \tau = 0.9487, p = 6.342\times10^{-6}; \omega_{na}: \tau = 0.872, p = 3.345\times10^{-5}; \omega_a: \tau = 0.692, p = 9.86\times10^{-4} \)).

Intriguingly, this multimodal distribution observed in Drosophila resembles the pattern of gene emergence in this species [17]. The peak in the adaptive substitution rate observed for clades 6 and 7 (Figure 1b) coincided with the animal phyla's major radiation at the time of extensive periods of glacial cycles [69]. When looking at the functions coded by these proteins, we found that they are linked to a wide range of vital cellular and biological processes, such as defence mechanisms and cell differentiation (Figure S8 in supplementary data). This pattern suggests that these genes might be experiencing higher molecular adaptive rates due to their role in such vital processes. However, for these genes to keep such high rates of adaptive substitutions until recent times (i.e., in the branch between D. melanogaster and D. simulans), epistatic interactions might be at play. Studies across taxa have proposed that functional epistasis is an important factor in the evolution of genes involved in defence mechanisms and adaptation to new environmental stresses [70–74]. We posit that such gene interactions keep these proteins further from their optimum throughout time due to the rugged shape of the fitness landscape, leading to the high molecular adaptive rates observed in the branch between D. melanogaster and D. simulans. To further test this hypothesis, we used the degree of protein-protein interactions (PPI) as a proxy for epistatic interactions and analysed its relationship with gene age. We observed that genes in clade 7 have a slightly higher degree of PPI than other strata (Figure S9 in supplementary data and supplementary file S5), suggesting that these genes might be experiencing relatively more epistatic interactions. These findings are consistent with epistasis influencing the evolution of these genes, potentially explaining their continued higher rates of molecular adaptation.

In contrast, the burst of the emergence of new genes in Arabidopsis coincided with the plant-specific radiation right before the emergence of Brassicaceae [17,75]. This trend is consistent with our results from A. thaliana, where the bursts of \( \omega_a \) occur in younger clades (after clades 11 and 12 in Figure 1b).

These distinct patterns observed between species suggest that the role of a gene’s age in molecular adaptation is complex, as also evidenced by the lack of a significant correlation with \( \omega_a \) previously.
reported in humans [22]. The authors proposed that this result may be a consequence of the generally low molecular adaptive rates observed in primates [22,48].

Despite these species-specific trends, our analyses revealed a strong correlation between \( \omega_a \) and gene age extending through hundreds of millions of years (Figures 1 and 2, Table S2 and S3).

These findings suggest a consistent effect of a gene’s age on the rate of molecular adaptation across taxa.

An adaptive walk model of gene evolution

Our study highlighted that, after their emergence, young genes evolve through relaxed selection, as first proposed by Ohno [76], but also by acquiring beneficial mutations, as described in the "adaptive-conflict" model [37,77]. Ohno's idea of evolution was "non-Darwinian" in its nature, as he believed that "natural selection merely modified while redundancy created" [76]. He proposed that new genes evolve by accumulating "forbidden" mutations, where they are only preserved if the development of a formerly non-existent function occurs, a process known as neo-functionalisation. 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-functionalisation, where the accumulation of deleterious mutations leads to a complementary loss of function in both copies of the gene [78,79].

In contrast, the "adaptive-conflict" model assumes that the ancestral gene could carry more than two pleiotropically constrained functions [37,77]. Once the duplication event occurs, each copy then becomes specialised in one of the ancestral functions. In this case, the ancestral gene’s split proceeded through positive Darwinian selection [37,77]. These theories are based on the evolution of gene duplicates and agree with the idea of evolution as a "tinkerer" proposed by Jacob [80], 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 [17,81,82]. 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 [83–86].

When looking at the fundamental ideas behind these theories, one can draw one prominent feature that portrays the evolution of new genes: young genes are further away from their fitness optimum. Hence, we posit that these genes follow an adaptive walk model of gene evolution to reach their fitness peak [3,87,88]. As their full potential has yet to be met, more consecutive beneficial mutations are theoretically needed to reach their fitness optimum, leading to the higher molecular adaptive rates observed in these genes. In turn, older genes are closer to their optimal features and less robust to large effects’ mutations, thus only accumulating mutations with small fitness effects. Such slightly advantageous mutations are more difficult to select for, leading to lower adaptive rates in these proteins. We further tested this hypothesis using Grantham’s physicochemical distances [68] as a proxy for the fitness effect of substitutions. This analysis showed that substitutions in young genes...
tend to occur between more dissimilar residues (Figure 3), suggesting that the evolution of young genes proceeds in larger steps compared to old ones. Moreover, we observed that adaptive substitutions have greater physicochemical effects (i.e., the average Grantham’s distance of adaptive substitutions ($\bar{G}_a$) is correlated to gene age; Figure S7 in supplementary data). However, the strength of this signal differs between species. While we observed a strong positive correlation between $\bar{G}_a$ and gene age in Drosophila, this relationship is weaker in Arabidopsis. The opposite pattern is observed for the correlation between $\bar{G}_{na}$ and gene age, where this relationship is only significant in Arabidopsis (Figure S7 in supplementary data). These patterns suggest that, in Arabidopsis, there is a stronger influence of relaxed purifying selection when compared to positive selection. This weaker adaptive signal may be a consequence of their mating system. The transition from out-crossing to self-fertilization in A. thaliana exposed these plants to less efficient purifying selection due to their reduced effective population sizes [89,90], thus accumulating more deleterious mutations. This higher load of segregating deleterious mutations can interfere with nearby beneficial mutations, preventing their fixation and slowing down adaptation. Indeed, previous studies have shown that segregating recessive deleterious mutations can substantially decrease the fixation rate of advantageous mutations, leading to a weaker adaptive signal in some parts of the genome [91,92]. In contrast, the large effective population sizes of Drosophila species make them more efficient in removing deleterious mutations and prone to stronger positive selection [93–95]. These patterns could explain the observed differences in the strength of selection in young genes between species.

An alternative hypothesis to the adaptive-walk theory is a static model of adaptation. This model is based on the level of constraint of a gene: as some genes are less constrained than others, they accumulate both advantageous and deleterious mutations, some of which may lead to the loss of a gene. Under this hypothesis, less constrained genes tend to be young because they are more likely to be lost. Additionally, as such genes accumulate a larger number of deleterious mutations, they also represent a bigger mutational target for adaptive mutations, which could also explain the overall correlation between $\omega_a$ and $\omega_{na}$. We can test this hypothesis by assessing whether the correlation between $\omega_a$ and gene age is independent of $\omega_{na}$. We controlled for the effect of $\omega_{naa}$ by running Grapes on a subset of genes for which the correlation between $p_n/p_s$ and gene age was no longer significant. We found that gene age was still significantly correlated with $\omega_a$ whereas $\omega_{na}$ was not (Figure S10), thus suggesting that the rate of adaptive evolution is independent of the non-adaptive rate. While these results do not completely discard the possibility of a stationary view of adaptation, they agree with our adaptive walk model of gene evolution.

Overall, our findings support an adaptive walk model hypothesis of gene evolution over time, where protein-coding genes adapt following a pattern of diminishing returns in Drosophila and Arabidopsis. The increasing availability of population genomic datasets will allow the detailed characterization of the adaptive walk model of gene evolution in an increasingly diverse range of taxa.
Material and Methods

We assessed the role of gene age on adaptive evolution using the divergence and polymorphism data published in Moutinho et al. [39]. For *Drosophila melanogaster*, the data included 10,318 protein-coding genes in 114 genomes for one chromosome arm of the two large autosomes (2L, 2R, 3L, and 3R) and one sex chromosome (X) pooled from an admixed sub-Saharan population from Phase 2 of the *Drosophila* Genomics Project [DPGP2] [96] and divergence estimates from *D. simulans* (Table S5 in supplementary data online). For *Arabidopsis thaliana*, we analysed 18,669 protein-coding genes in 110 genomes comprising polymorphism data from a Spanish population (1001 Genomes Project) [97] and divergence out to *A. lyrata* (Table S6 in supplementary data online). We used these data sets to infer the synonymous and non-synonymous unfolded site frequency spectrum (SFS), and synonymous and non-synonymous divergence from the rate of synonymous and nonsynonymous substitutions.

Sites with a missing outgroup allele were considered as missing data. All estimates were obtained using the “bppPopStats” program from the Bio++ Program Suite (version 2.4.0) [98]. Gene age data were obtained from published data sets, wherein 9,004 *Drosophila* [28] and 17,732 *Arabidopsis* [33] genes were used. Analyses were performed by dividing genes into 12 and 15 phylostrata for *D.* melanogaster and *A. thaliana* (Figure 1), respectively, numbered from the oldest (stratum 1) to the most recent (strata 12 and 15 in *D.* melanogaster and *A. thaliana*, respectively). The most recent clades include orthologous genes present in each species and their respective outgroups. The analyses on the X-linked and autosomal genes in *D. melanogaster* were performed with 1,478 and 7,526 genes, respectively. We fitted models of the distribution of fitness effects (DFE) across different age classes and gene categories to estimate the molecular rate of adaptation [48].

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

The rates of adaptive non-synonymous substitutions were estimated with the Grapes program [48], using the Gamma-Exponential DFE, as this model was previously shown to best fit the data [39]. As this method does not accommodate SNPs with missing information in some individuals, we reduced the original data set to n=110 and n=105 genomes for *D. melanogaster* and *A. thaliana*, respectively, by randomly down-sampling sites with at least n genotypes available and discarding positions where less than n genotypes were present. Estimates of substitution rates and their confidence intervals were obtained with a bootstrap analysis by sampling genes in each category, with replacement. We performed a total of 100 replicates, and the DFE model was fitted for each replicate with Grapes. Results for $\omega$, $\omega_{nA}$ and $\omega_{A}$ were plotted using the R package "ggplot2" [99] by taking the mean value and the 95% confidence interval of the 100 bootstrap replicates performed for each category (see detailed R scripts in the supplementary files in the supplementary data online, https://gitlab.gwdg.de/molsysevol/supplementarydata_geneage).

Gene age vs. protein length and gene expression
Data on protein length and gene expression were included in the datasets of Moutinho et al. [39] (Tables S2 and S3 in supplementary data online). Gene expression data were downloaded from the database Expression Atlas [100], taking one baseline experiment for each species [39]. Mean gene expression levels were then obtained by averaging across samples and tissues for each gene. To correct for these two factors, we divided our dataset into two equally sized groups based on the factor we wished to control. Short proteins had a size up to 366 and 389 amino acids, and long proteins had a size up to 4,674 and 5,098 amino acids in *A. thaliana* and *D. melanogaster*, respectively. We further merged phylostrata containing a low number of genes. For *D. melanogaster*, we categorised gene age into 6 main clades by combining clades 3-4, 5-6, 7-10, and 11-12, keeping the others unchanged. In *A. thaliana*, we combined the 15 clades into 6 main groups by merging clades 5-8 and 9-15. For gene expression, we used a total of 17,126 and 6,247 genes for *A. thaliana* and *D. melanogaster*, respectively, being categorised as low and highly expressed. Genes were classified 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*, we categorised gene age into 6 categories by combining clades 3-5, 6-9, and 10-12. In *A. thaliana*, we combined the data in 6 clades, merging clades 4-7, 8-11 and 12-15.

**Gene age vs. protein structure**

Since most young genes lack a defined three-dimensional structure [36], 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 [101] by using as a training model the HH-suite sequence alignment tool for protein similarity searches [102]. To assess whether this approach provided reliable results, we compared the RSA estimates of NetSurfP-2.0 with those obtained from the PDB structures in our dataset [39]. We found a good correlation between the two approaches for both species (Kendall’s τ = 0.571, p < 2e-216; τ = 0.462, p < 2e-216, for *D. melanogaster* and *A. thaliana* respectively). Using NetSurfP-2.0, RSA estimates were successfully obtained for a total of 4,238,686 (88% of the total codon sites) and 7,479,807 (99% of the total codon sites) amino-acid residues for *D. melanogaster* and *A. thaliana*, respectively. To assess the impact of RSA at the gene level, we analysed the total number of genes in both species by making two categories of genes according to the average RSA value per gene. Genes with lower RSA had mean values between 0.127-0.389 in *Drosophila* and 0.217-0.386 in *Arabidopsis*. Genes with a higher RSA had mean values between 0.390-0.894 in *Drosophila* and 0.387-0.898 in *Arabidopsis*. The phylostrata groups were defined by combining clades 7-8 in *D. melanogaster*, and 8-11, 12-15 in *A. thaliana*.

For the analysis of the residue intrinsic disorder, we used the estimates included in the datasets of Moutinho et al. [39]. These data were obtained with the software DisEMBL [103], where intrinsic disorder was estimated per site and classified according to the degree of “hot loops” i.e., highly mobile loops. This analysis was performed for 17,732 and 7,410 genes for *A. thaliana* and *D. melanogaster*.
respectively. Genes were combined into two categories according to the mean value of their residue’s intrinsic disorder. Genes with a low level of intrinsic disorder had values between 0.029-0.080 in Drosophila and among 0.041-0.084 in Arabidopsis. Genes with a higher degree of intrinsic disorder had values between 0.081-0.554 in Drosophila and among 0.085-0.551 in Arabidopsis. In D. melanogaster, all the 12 phylostrata could be used. In A. thaliana, the 15 strata were combined in 12 categories by merging clades 9-10, 11-12 and 13-14.

Correcting for BLAST e-values

We analysed the robustness of the gene age’s effect by correcting the variation in the Expect (E) value estimates in BLAST's searches between our focus species and their respective outgroups. By reducing the variation in E-values estimates, we could correct for potential failures in BLAST’s homology searches. To do so, we limited the analysis to genes with particularly low E-values that are more likely to be identified in all age strata: 12,472 genes with an E value lower than 1e-150 for A. thaliana and 7,104 genes with an E value lower than 1e-100 for D. melanogaster (supplementary file S2). For A. thaliana, analyses were carried out by combining clades 8-13, with no genes left in clades 14 and 15. For D. melanogaster, analyses were performed with the 12 strata. By using these genes, we found no correlation between E-value and age.

Gene age vs. protein function

Gene ontology terms were obtained from the "dmelanogaster_gene_ensembl" and the "athaliana_eg_gene" tables in the Ensembl database (version 103), through the R package "biomaRt" [104]. A total of 7,253 (~70% of the genes) and 15,60 (~80% of the genes) genes were mapped in D. melanogaster and A. thaliana, respectively. To check whether the effect of gene age prevailed across functional protein classes, we analysed the GO terms with the highest number of young genes mapped: more than 50 genes present in Clades 11 and 12 in D. melanogaster; and more than 30 genes present in Clades 12 to 15 in A. thaliana. This filtering step resulted in 6,637 genes across 23 GO categories in D. melanogaster (Table S5 in Supplementary Data online), and 15,410 genes across 10 GO categories in A. thaliana (Table S6 in Supplementary Data online). To analyse the effect of gene age, we compared three age classes. In D. melanogaster, the first age category spanned over phylostrata 1-3, the second category covered clades 4-7, and the third one included clades 8-12. In A. thaliana, the first category comprised genes belonging to clades 1-6, the second category spanned over clades 7-11, and the third one included the phylostrata between clades 12-15 (Figure 1a).

Gene Age vs. Protein-protein interactions (PPI)

We obtained PPI data for D. melanogaster from the STRING database [105], which includes both physical and functional interactions (https://string-db.org/). This database included 13,046 proteins
The fitness effects of amino-acid substitutions

We used Grantham's physicochemical distances between amino-acids [68] as a proxy for the fitness effects of amino-acid substitutions and performed two analyses. In the first one, we calculated the average Grantham’s distance between amino acid substitutions for each phylostrata. In the second analysis, we estimated the average Grantham’s distance for adaptive ($\bar{G}_a$) and non-adaptive ($\bar{G}_n$) non-synonymous substitutions within each age stratum. We restricted our analysis to amino acid pairs separated by a single mutational step. To estimate the rate of adaptive and non-adaptive non-synonymous amino acid substitutions between pairs of amino acids we used the same approach as Bergman and Eyre-Walker [106]. This method compares the non-synonymous SFS for a particular amino acid pair, for example, alanine and aspartic acid, which are separated by a A $\leftrightarrow$ C mutation, with the synonymous SFS of 4-fold degenerate sites separated only by A $\leftrightarrow$ C mutations ($SFS_{4F(A$ $\leftrightarrow$ $C)}$). For amino acid pairs separated by more than one mutation type, we estimated the weighted average of the synonymous SFS of 4-fold sites for the different types of mutations, weighting by the frequency of the respective codons. For example, serine and cysteine are separated by A $\leftrightarrow$ T and C $\leftrightarrow$ G mutations. The synonymous SFS for this amino acid pair was estimated as

\[
SFS_{4F(\text{weighted})} = \left[ (f_{\text{AGT}} + f_{\text{TGT}} + f_{\text{AGC}} + f_{\text{TGC}}) \ast SFS_{4F(A \leftrightarrow C)} + (f_{\text{TCT}} + f_{\text{TGT}} + f_{\text{TCC}} + f_{\text{TGC}}) \ast SFS_{4F(C \leftrightarrow G)} \right] / (f_{\text{AGT}} + f_{\text{TGT}} + f_{\text{AGC}} + f_{\text{TGC}} + f_{\text{TCT}} + f_{\text{TGT}} + f_{\text{TCC}} + f_{\text{TGC}}),
\]

where $f$ corresponds to the frequency of each codon (e.g., $f_{\text{AGT}}$ is the frequency of the codon AGT). It is not possible to estimate the rates of adaptive and non-adaptive substitution for each pair of amino acids for each phylostratum because we have insufficient data. So, we combined amino acid pairs into ten categories based on their Grantham’s distance, and we combined clades 5-6, 8-11 and 12-15 in Arabidopsis, and clades 3-4, 5-7 and 8-10 in Drosophila. We used Grapes to estimate $\omega_a$ and $\omega_n$ for the ten categories of Grantham’s physicochemical distances for each combined age stratum. We calculated the average Grantham’s distance for adaptive ($\bar{G}_a$) and non-adaptive ($\bar{G}_n$) non-synonymous substitutions as

\[
\bar{G}_a = \sum_i \omega_{ai} \bar{g}_i N_i / \sum_i N_i \quad \text{and} \quad \bar{G}_n = \sum_i \omega_{nai} \bar{g}_i N_i / \sum_i N_i \quad \text{where the sum is across the ten categories of Grantham’s distances within each age class, $\omega_{ai}$ and $\omega_{nai}$ are the rates of adaptive and non-adaptive substitution in the $i$th category for each age class, $\bar{g}_i$ is the average Grantham’s distance amongst the amino acid pairs within Grantham’s distance category, and $N_i$ is the number of non-synonymous sites in the $i$th category in each age class. Higher values of $\bar{G}_a$ (the same for $\bar{G}_n$ with non-adaptive substitutions) mean that there is a higher proportion of adaptive substitutions across Grantham’s distance categories.}

Assessing the correlation between $\omega_a$ and $\omega_n$
To test whether the impact of gene age on $\omega_n$ was independent of $\omega_{na}$, we ran Grapes for a subset of genes for which the correlation between $\omega_{na}$ and gene age was no longer significant. We controlled for the variation in $\omega_{na}$ by discarding genes according to their $p_n/p_s$ values. To ensure a sufficient number of genes, we kept the genes that were within 0.02 and 0.05 standard deviation from the modal value of the $p_n/p_s$ distribution in *Arabidopsis* (mode: 0.05) and *Drosophila* (mode: 0.01), respectively. The subset of data included 8,652 and 2,570 genes for *Arabidopsis* and *Drosophila*, respectively.

**Statistical analyses**

Assessing the effect of gene age within each protein functional class was performed by comparing rate estimates between all pairs of age categories. 100 bootstrap replicates were generated and $\omega_a$ and $\omega_{na}$ were estimated for each resampling, allowing us to compute the rate differences between categories. A one-tailed P-value can be obtained using the formula $P = (k + 1)/(N + 1)$, where $N=100$ is the number of bootstrap replicates and $k$ is the number of times the computed difference was greater (resp. lower) than 0. Here, we used a two-tailed version of this test, computing the P-value as $P = [2 \times \min (k^-, k^+) + 1]/(N+1)$, where $k^-$ is the number of times the difference was negative, and $k^+$ is the number of times the difference was positive. P-values for all pairwise comparisons were corrected for multiple testing using the FDR method [107] as implemented in R [108] (see detailed R script in supplementary file S3). For the analysis with PPI and gene age, statistical significance was assessed using non-parametric post hoc tests, as implemented in the “Kruskal” method of the R package “agricolae” using the FDR method to correct for multiple testing [109] (see detailed R script in supplementary file S5). For the rest of the analyses, statistical significance was assessed with Kendall’s correlation tests using the mean value of the 100 bootstrap replicates for each category (see detailed script in supplementary file S6). To estimate the combined P-value for each co-factor we used the weighted-Z method using the R package “metap” [110]. We estimated the weight of each P-value using a linear modelling approach with $\omega_a$ and $\omega_{na}$ as response variables, and gene age and potential co-factors as explanatory variables and inferred the reciprocal of the squared standard error of the residuals in each model (see detailed R scripts in supplementary file S7). To determine whether the chromosome impacted gene age’s effect on estimates of $\omega_a$ and $\omega_{na}$, we performed an analysis of covariance (ANCOVA) by comparing a model M1 that included the impact of the chromosome, age, and their interaction, with a model M0 that included age only (see detailed R script in supplementary file S1). Normality, homoscedasticity, and independence of the error terms of the model were assessed with the package “lmtest” [111] in R. Finally, to further assess the effect of residual correlations between gene age and each cofactor, we fitted linear models with the rates of protein evolution (i.e., $\omega$, $\omega_{na}$ and $\omega_a$) as response variables and the average value of the cofactor for each age class as an explanatory variable (nested within the cofactor category, low or high), together with gene age, species (*Drosophila* or *Arabidopsis*), cofactor category, and their interaction. A model selection was conducted using the R function ‘step’ and the
The significance of the model coefficients was assessed for the selected models (see detailed R script in supplementary file S8).

**MK regression**

The MK regression analysis was performed per nucleotide site, where we analysed the combined effect of RSA, intrinsic protein disorder, protein length, gene expression, protein function, Grantham’s distances, and the sex chromosome in *Drosophila*. As this method does not automatically handle categorical variables, we created a binary variable for protein function based on the gene’s presence (1) or absence (0) in stress-related proteins (*i.e.*, response to stress, response to external stimulus and signal transduction). We used the same rationale for the X chromosome effect in *Drosophila*. We performed the MK regression analysis by comparing polymorphic and divergence sites between non-synonymous and 4-fold degenerated sites in both species. As this method does not account for the effects of weak selection [57], we analysed only sites with an allele frequency above 50%. To assess the relative contribution of each variable to the regression model, we compared the likelihood pseudo-$R^2$ following the Cox and Snell method [59]. These were estimated by comparing the log-likelihood of the full model (all factors included) with that from each reduced model (each factor removed) using the following equation:

$$R^2 = 1 - \frac{\exp \left( - \frac{2}{N} \times (ln_{full} - ln_{reduced}) \right)}{}$$

where $N$ represents the number of sites analysed with the MK regression, $ln_{full}$ represents the log-likelihood of the model with all factors included, and $ln_{reduced}$ represents the log-likelihood of the model excluding each factor (*e.g.*, to assess the contribution of the gene age variable, the full model was compared to a model including all variables except gene age).

**Acknowledgements**

The authors thank Diethard Tautz, Tal Dagan and Chaitanya Gokhale for fruitful discussions. J.Y.D. acknowledges funding from the Max Planck Society. We thank two anonymous reviewers and Nicolas Lartillot for their comments and suggestions to this manuscript.

**References**

1. Fisher R. The Genetical Theory of Natural Selection. Oxford Univ. Press, Oxford; 1930.
2. Wright S. Evolution in Mendelian Populations. Genetics. 1931;16(2):97–159.
3. Wright S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. Vol. 1, Sixth International Congress on Genetics. 1932. p. 356–66.
4. Smith JM. Natural selection and the concept of a protein space. Nature. 1970;225(5232):563–4.
5. Gillespie JH. A Simple Stochastic Gene Substitution Model. Theoretical Population Biology. 1983;23:202–15.
6. Gillespie JH. Molecular evolution over the mutational landscape. Evolution (N Y). 1984;38(5):1116–29.
7. Gillespie JH. The Causes of Molecular Evolution. Oxford University Press; 1991.
8. Kimura M. The Neutral Theory of Molecular Evolution. Cambridge Univeristy Press; 1983.
630 9. Orr HA. The Population Genetics of Adaptation: The Distribution of Factors Fixed during Adaptive Evolution. Evolution. 1998;52(4):935.
631 10. Orr AH. The evolutionary genetics of adaptation: A simulation study. Genetical Research. 1999;74(3):207–14.
632 11. Lenski RE, Rose MR, Simpson SC, Tadler SC. Long-Term Experimental Evolution in Escherichia coli. I. Adaptation and Divergence During 2,000 Generations. The American Naturalist. 1991;138(6):1315–41.
633 12. Cooper VS, Lenski RE. The population genetics of ecological specialization in evolving Escherichia coli populations. Nature. 2000;407(6805):736–9.
634 13. Gerrish P. The rhythm of microbial adaptation. Nature. 2001;413(6853):299–302.
635 14. Rozen DE, De Visser JAGM, Gerrish PJ. Fitness effects of fixed beneficial mutations in microbial populations. Current Biology. 2002;12(12):1040–5.
636 15. Schoustra SE, Batalion T, Gifford DR, Kassen R. The properties of adaptive walks in evolving populations of fungus. PLoS Biology. 2009;7(11).
637 16. Lynch M. Genomics: Gene duplication and evolution. Science. 2002;297(5583):945–947.
638 17. Tautz D, Domazet-Lošo T. The evolutionary origin of orphan genes. Nature Reviews Genetics. 2011;12(10):692–702.
639 18. Cohen O, Ashkenazy H, Belinky F, Hucho D, Pupko T. GLOOME: Gain loss mapping engine. Bioinformatics. 2010;26(22):2914–5.
640 19. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, et al. Gapped blast and psi-blast: a new generation of protein database search programs. FASEB Journal. 1998;12(8):3389–402.
641 20. Domazet-Lošo T, Brajković J, Tautz D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends in Genetics. 2007;23(11):533–9.
642 21. Thornton K, Long M. Rapid divergence of gene duplicates on the Drosophila melanogaster X chromosome. Molecular Biology and Evolution. 2002;19(6):918–25.
643 22. Cai JJ, Petrov DA. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. Genome Biology and Evolution. 2010;2(1):393–409.
644 23. Domazet-Loso T, Tautz D. An evolutionary analysis of orphan genes in Drosophila. Genome Research. 2003;13(10):2213–9.
645 24. Vishnoi A, Kryazhimskiy S, Bazykin GA, Hannenhalli S, Plotkin JB. Young proteins experience more variable selection pressures than old proteins. Genome Research. 2010;20(11):1574–81.
646 25. Cai JJ, Woo PCY, Lau SKP, Smith DK, Yuen KY. Accelerated evolutionary rate may be responsible for the emergence of lineage-specific genes in Ascomycota. Journal of Molecular Evolution. 2006;63(1):1–11.
647 26. Wolf YI, Novichkov PS, Karev GP, Koonin E V., Lipman DJ. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. Proceedings of the National Academy of Sciences. 2009;106(18):7273–80.
648 27. Albá MM, Castresana J. Inverse relationship between evolutionary rate and age of mammalian genes. Molecular Biology and Evolution. 2005;22(3):598–606.
649 28. Domazet-Lošo T, Carvunis AR, Albá MM, Šestak MS, Bakarić R, Neme R, et al. No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. Molecular Biology and Evolution. 2017;34(4):843–56.
650 29. Zhang YE, Vibranovski MD, Krinsky BH, Long M. Age-dependent chromosomal distribution of male-biased genes in Drosophila. Genome Research. 2010;20(11):1526–33.
651 30. Daubin V, Ochman H. Bacterial genomes as new gene homes: The genealogy of ORFans in E. coli. Genome Research. 2004;14(6):1036–42.
652 31. García-Vallvé S, Alonso Á, Bravo IG. Papillomaviruses: Different genes have different histories. Trends in Microbiology. 2005;13(11):514–21.
653 32. Cui X, Lv Y, Chen M, Nikoloski Z, Twell D, Zhang D. Young genes out of the male: An insight from evolutionary age analysis of the pollen transcriptome. Molecular Plant. 2015;8(6):935–45.
654 33. Arendsee ZW, Li L, Wurtele ES. Coming of age: Orphan genes in plants. Trends in Plant Science. 2014;19(11):698–708.
34. Kuo CH, Kissinger JC. Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites Plasmodium and Theileria. BMC Evolutionary Biology. 2008;8(1):1–16.

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

36. Wilson BA, Foy SG, Neme R, Masel J. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. Nature Ecology and Evolution. 2017;1(6).

37. Hughes AL. The evolution of functionally novel proteins after gene duplication. Proceedings of the Royal Society B. 1994;256(1346):119–24.

38. Zhang J, Zhang Y. Ping, Rosenberg HF. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nature Genetics. 2002;30(4):411–5.

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

40. Afanasyeva A, Bockwoldt M, Cooney CR, Helland I, Gossmann Tl. Human long intrinsically disordered protein regions are frequent targets of positive selection. Genome Research. 2018;28(7):975–82.

41. Subramanian S, Kumar S. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics. 2004;168(1):373–81.

42. Stukkenbrock EH, Bataillon T, Dutheil JY, Hansen TT, Li R, Zala M, et al. The making of a new pathogen: Insights from comparative population genomics of the domesticated wheat pathogen Mycosphaerella graminicola and its wild sister species. Genome Research. 2011;21(12):2157–66.

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

44. Rocha EPC, Danchin A. An Analysis of Determinants of Amino Acids Substitution Rates in Bacterial Proteins. Molecular Biology and Evolution. 2004;21(1):108–16.

45. Pal C, Papp B, Hurst LD. Highly Expressed Genes in Yeast Evolve Slowly. Genetics. 2001;158(1998):927–31.

46. Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG. Dynamic evolution of the innate immune system in Drosophila. Nature Genetics. 2007;39(12):1461–8.

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

48. Galtier N. Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. PLoS Genetics. 2016;12(1):1–23.

49. Vicoso B, Charlesworth B. Evolution on the X chromosome: Unusual patterns and processes. Nature Reviews Genetics. 2008;9(1):1–16.

50. Vicoso B, Charlesworth B. Effective population size and the faster-X effect: An extended model. Evolution (N Y). 2009;63(9):2413–26.

51. Charlesworth B, Campos JL, Jackson BC. Faster-X evolution: Theory and evidence from Drosophila. Molecular Ecology. 2018 Oct 1;27(19):3753–71.

52. Ding Y, Zhou Q, Wang W. Origins of New Genes and Evolution of Their Novel Functions. Annual Review of Ecology, Evolution, and Systematics. 2012;43(1):345–63.

53. Choi SS, Vallender EJ, Lahn BT. Systematically assessing the influence of 3-dimensional structural context on the molecular evolution of mammalian proteomes. Molecular Biology and Evolution. 2006;23(11):2131–3.

54. Whitlock MC, Whitlock MC. Combining probability from independent tests: the weighted Z-method is superior to Fisher’s approach. Wiley Online Library. 2005 Sep;18(5):1368–73.

55. Stoletzki N, Eyre-Walker A. Estimation of the neutrality index. Molecular Biology and Evolution. 2011;28(1):63–70.

56. Smith NGC, Eyre-Walker a. Adaptive protein evolution in Drosophila. Nature. 2002;415(6875):1022-1024.

57. Huang Y-F. Dissecting Genomic Determinants of Positive Selection with an Evolution-Guided Regression Model. Molecular Biology and Evolution. 2022; 39(1): msab291.

58. Eyre-Walker A, Keightley PD. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Molecular Biology and Evolution. 2009;26(9):2097–108.
59. Cox DR, Snell EJ. Analysis of Binary Data (2nd ed.). London: Chapman and Hall; 1970.

60. Elhaik E, Sabath N, Graur D. 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. Molecular Biology and Evolution. 2006;23(1):1–3.

61. Moyers BA, Zhang J. Phylostratigraphic bias creates spurious patterns of genome evolution. Molecular Biology and Evolution. 2015;32(1):258–67.

62. Moyers BA, Zhang J. Evaluating Phylostratigraphic Evidence for Widespread de Novo Gene Birth in Genome Evolution. Molecular Biology and Evolution. 2016;33(5):1245–56.

63. Albá MM, Castresana J. On homology searches by protein Blast and the characterization of the age of genes. BMC Evolutionary Biology. 2007;7:1–8.

64. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. More than just orphans: are taxonomically-restricted genes important in evolution? Trends in Genetics. 2009;25(9):404–13.

65. Slotte T, Bataillon T, Hansen TT, St. Onge K, Wright SI, Schierup MH. Genomic determinants of protein evolution and polymorphism in arabidopsis. Genome Biology and Evolution. 2011;3(1):1210–9.

66. Geldner N, Robatzek S. Plant receptors go endosomal: A moving view on signal transduction. Plant Physiology. 2008;147(4):1565–74.

67. Groen AJ, De Vries SC, Lilley KS. A proteomics approach to membrane trafficking. Plant Physiology. 2008;147(4):1584–9.

68. Grantham R. Amino Acid Difference Formula to Help Explain Protein Evolution. Science. 1974;185(4154):862–4.

69. Hoffman PF, Kaufman AJ, Halverson GP, Schrag DP. A neoproterozoic snowball earth. Science. 1998;281(5381):1342-1346.

70. Montooth KL, Marden JH, Clark AG. Mapping Determinants of Variation in Energy Metabolism, Respiration and Flight in Drosophila. Genetics. 2003 Oct 1;165(2):623–35.

71. Paixão T, Barton NH. The effect of gene interactions on the long-term response to selection. Proc Natl Acad Sci. 2016;113(16), 4422-4427.

72. Hansen TF. Hansen. T. F. (2013). Why epistasis is important for selection and adaptation. Evolution. 2013; 67(12): 3501-3511.

73. Behrman EL, Howick VM, Kapun M, Staubach F, Bergland AO, Petrov DA, et al. Rapid seasonal evolution in innate immunity of wild Drosophila melanogaster. Proceedings of the Royal Society B: Biological Sciences. 2018; 285(1870): 20172599.

74. Lagator M, Colegrave N, Neve P. Selection history and epistatic interactions impact dynamics of adaptation to novel environmental stresses. Proceedings of the Royal Society B: Biological Sciences. 2014; 281(1794): 20141679.

75. Wang H, Moore MJ, Soltis PS, Bell CD, Brockington SF, Alexandre R, et al. Rosid radiation and the rapid rise of angiosperm-dominated forests. Proc Natl Acad Sci. 2009;106(10):3853–8.

76. Ohno S. Evolution by gene duplication. Springer Science & Business Media; 1970.

77. Piatigorsky J, Wistow G. The recruitment of crystallins: new functions precede gene duplication. Science. 1991;252(5009):1078–9.

78. Prince VE, Pickett FB. Splitting pairs: The diverging fates of duplicated genes. Nature Reviews Genetics. 2002;3(11):827–37.

79. Jacob F. Evolution and Tinkering. Science (1979). 1977;196(4295):1161-6.

80. Cai J, Zhao R, Jiang H, Wang W. De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. Genetics. 2008;179(1):487–96.

81. Heinen TJAJ, Staubach F, Häming D, Tautz D. Emergence of a New Gene from an Intergenic Region. Current Biology. 2009;19(18):1527–31.

82. Zhao L, Saelao P, Jones CD, Begun DJ. Origin and spread of de novo genes in Drosophila melanogaster populations. Science. 2014;343(6172):769–72.

83. Neme R, Tautz D. Evolution: Dynamics of de novo gene emergence. Current Biology. 2013;24(6):R238–40.

84. Palmieri N, Kosiol C, Schlötterer C. The life cycle of Drosophila orphan genes. Elife. 2014;3:1–21.
856. Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, et al. Proteogenomes and de novo gene birth. Nature. 2012;487(7407):370–4.
857. Smith JM. Natural selection and the concept of a protein space. Nature. 1970;225(5232), 563-564.
858. Orr HA. The Population Genetics of Adaptation: the Adaptation of DNA Sequences. Evolution.
859. 2002;56(7):1317.
860. Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL. The Cost of Inbreeding in Arabidopsis. Nature. 2002; 416(6880): 531-534.
861. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993; 134(4):1289–303.
862. Assaf ZJ, Petrov DA, Blundell JR. Obstruction of adaptation in diploids by recessive, strongly deleterious alleles. Proceedings of the National Academy of Sciences. 2015;112(20):E2658–66.
863. Di C, Murga Moreno J, Salazar-Tortosa DF, Lauterbur ME, Enard D. Decreased recent adaptation at human mendelian disease genes as a possible consequence of interference between advantageous and deleterious variants. Elife. 2021; 10: e69026
864. Gossmann TI, Kightley PD, Eyre-Walker A. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. Genome Biology and Evolution. 2012;4(5):658–67.
865. Eyre-Walker A. The genomic rate of adaptive evolution. Trends Ecol Evol. 2006;21:569–75.
866. Rousselle M, Simion P, Tilak M-K, Figuet E, Nabholz B, Galtier N. Is adaptation limited by mutation? A timescale-dependent effect of genetic diversity on the adaptive substitution rate in animals. PLOS Genetics. 2020;16(4):e1008668.
867. Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, Crepeau MW, et al. Population Genomics of Sub-Saharan Drosophila melanogaster: African Diversity and Non-African Admixture. PLoS Genetics. 2012;8(12).
868. Weigel D, Mott R. The 1001 Genomes Project for Arabidopsis thaliana. Genome Biology. 2009;10(5):1-5.
869. Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette NC, et al. Bio++: Efficient extensible libraries and tools for computational molecular evolution. Molecular Biology and Evolution. 2013;30(8):1745–50.
870. Wickham H. ggplot2 - Elegant Graphics for Data Analysis (2nd Edition). Journal of Statistical Software. 2017;77:2–5.
871. PetryszaK R, Keays M, Tang YA, Fonseca NA, Barrera E, Burdett T, et al. Expression Atlas update - An integrated database of gene and protein expression in humans, animals and plants.
872. Nucleic Acids Research. 2016; 44(D1):D746-D752.
873. Klausen MS, Jenspersen MC, Nielsen H, Jensen KK, Jurtz VI, Sønderby CK, et al. NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. Proteins: Structure, Function and Bioinformatics. 2019;87(6):520–7.
874. Remmert M, Biegert A, Hauser A, Söding J. HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature Methods. 2012;9(2):173–5.
875. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: Implications for structural proteomics. Structure. 2003; 11(11): 1453-1459.
876. DurincK S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;21(16):3439–40.
877. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, et al. STRING 8 - A global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Research. 2009;37.
878. Bergman J, Eyre-Walker A. Does Adaptive Protein Evolution Proceed by Large or Small Steps at the Amino Acid Level? Molecular Biology and Evolution. 2019; 36(5): 990-998.
879. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. 1995;57(1):289–300.
880. R Core Team. A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2017.
109. Mendiburu F, Simon R, De Mendiburu F. Agricolae—Ten years of an Open source Statistical tool for experiments in Breeding, agriculture and biology. PeerJ PrePrints. 2015(No. e1748).

110. Dewey M. metap: meta-analysis of significance values. R package version 1.4. 2020;

111. Zeileis A, Hothorn T. Diagnostic Checking in Regression Relationshipslmtest citation info. R News. 2002;2(3):7–10.
**Figures**

Figure 1. (a) Phylogenetic definition of the strata used in the analyses for *A. thaliana* (top) and *D. melanogaster* (bottom). The number of genes mapped to each clade is shown. (b) Relationship between the rate of protein evolution (\(\omega\)), non-adaptive non-synonymous substitutions (\(\omega_{na}\)) and adaptive non-synonymous substitutions (\(\omega_a\)) 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 shown. Mean values of \(\omega\), \(\omega_{na}\) and \(\omega_a\) 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 2. Estimates of \( \omega \), \( \omega_{na} \), and \( \omega_a \) plotted as a function of (a) protein length and (b) mean expression levels, (c) relative solvent accessibility, and (d) protein intrinsic disorder with gene age in *A. thaliana* (top) and *D. melanogaster* (bottom). Analyses were performed by comparing short and long genes (a), lowly and highly expressed genes (b), proteins with low and high mean RSA values (c), and proteins with low and high average intrinsic disorder (d) across age categories (see Material and Methods). Legend as in Figure 1. Significance levels are shown for each correlation between gene age and the rates of protein evolution (*\( P < 0.05 \); **\( P < 0.01 \); ***\( P < 0.001 \); ." 0.05 \( \leq P < 0.10 \); “ns” \( P > 0.10 \)).
Figure 3. Relationship between gene age and Grantham’s distance between amino acids for *A. thaliana* (a) and *D. melanogaster* (b). For each clade, the median value of Grantham’s distance between residues is depicted with the black dot. The shaded area represents the physicochemical distances within the 1st and 3rd quartile.
Table 1. Kendall’s correlation coefficients for the relationship between $\omega$, $\omega_{na}$ and $\omega_a$ and gene age, for the analysis of gene age and the combined analyses of gene age with the respective co-factors: protein length, gene expression, protein intrinsic disorder and the mean relative solvent accessibility per gene. The combined probabilities for each co-factor within and across species are presented in the fields “Weighted Z” and “Weighted Z across species”, respectively, for $\omega$, $\omega_{na}$ and $\omega_a$.

| Gene Age          | Arabidopsis | Drosophila | Weighted Z across species |
|-------------------|-------------|------------|---------------------------|
|                   | $\omega$    | $\omega_{na}$ | $\omega_a$ | $\omega$ | $\omega_{na}$ | $\omega_a$ | $\omega$ | $\omega_{na}$ | $\omega_a$ | $\omega$ | $\omega_{na}$ | $\omega_a$ |
| Long              | 1.00 **     | 0.867 *    | -0.200        | 0.867 *  | 0.600 (0)     | 0.867 *   | 1.56e-04 | 7.71e-05 | 7.98e-03     |
| Short             | 1.00 **     | 0.867 *    | 0.600 (0)     | 0.733 *  | 0.867 *       | 0.467     | 6.46e-04 | 1.61e-03 | 2.64e-03 **  |
|                   | 6.46e-04 ***| 1.61e-03 **| 0.133         | 2.64e-03 ** | 5.29e-03 ** | 0.0105 *  |
| Weighted Z        | 9.00 **     | 3.71e-03 ***| 0.186         | 1.09e-03 | 1.68e-03 **   | 2.24e-03 **|
| High              | 0.867 *     | 0.867 *    | 0.467         | 0.867 *  | 1.00 **       | 0.600 (0) |
| Low               | 0.867 *     | 1.000 **   | 0.333         | 0.733 *  | 1.00 **       |           |
|                   | 1.51e-03 ** | 3.71e-04 ***| 0.186         | 1.09e-03 | 1.68e-03 **   | 2.24e-03 **|
| Weighted Z        | 1.41e-05 ** | 1.24e-03 ** | 3.67e-04 ***  | 7.76e-04 *** | 1.55e-03 ** |
| Protein Length    | 0.970 ***   | 0.909 ***  | 0.454 *       | 0.630 *  | 0.576 **      | 0.273     |
| High              | 1.000 ***   | 0.939 ***  | 0.636 **      | 0.670 ** | 0.303        | 0.515 *   |
| Low               | 0.970 ***   | 0.909 ***  | 0.454 *       | 0.630 *  | 0.576 **      | 0.273     |
|                   | <2e-216 *** | <2e-216 *** | 1.20e-03 **   | 3.85e-05 *** | 5.80e-03 ** | 4.18e-02 * |
| Protein Intrinsic Disorder |           |            |              |           |             |           |
| High              |              |             |              |           |             |           |
| Low               |              |             |              |           |             |           |
|                   |              |             |              |           |             |           |
| Mean Relative Solvent Accessibility |           |            |              |           |             |           |
| High              | 0.944 ***   | 0.889 ***  | 0.722 **      | 0.636 ** | 0.673 **     | 0.564 *   |
| Low               | 1.000 ***   | 0.778 **   | 0.667 *       | 0.636 ** | 0.491 *      | 0.564 *   |
|                   | 6.20e-06 ***| 1.41e-05 ***| 1.24e-03 **   | 3.67e-04 *** | 7.76e-04 *** | 1.55e-03 **|
| Weighted Z        |              |             |              |           |             |           |
|                   |              |             |              |           |             |           |

Note. For each variable, the correlation coefficient and the combined probabilities are shown with the respective significance (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; "," $0.05 \leq P < 0.10$) for $\omega$, $\omega_{na}$ and $\omega_a$ in *Arabidopsis* and *Drosophila*. As the effect of gene age was assessed by combining genes in each age class, these correlation coefficients do not measure the intrinsic gene-level strength of correlation between gene age and molecular rates of evolution.