Testing the adaptive walk model of gene evolution

Ana Filipa Moutinho[*,1,2], Adam Eyre-Walker[2], Julien Y. Dutheil[1,3]

[1]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Biology, Plön, Germany
[2]School of Life Sciences, University of Sussex, Brighton, United Kingdom
[3]Unité 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 fitted models of the distribution of fitness effects to population genomic datasets of *Arabidopsis* and *Drosophila*. 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 fitness effects. Our study, therefore, provides strong evidence that molecular evolution follows an adaptive walk model across 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 towards 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. The challenge lies in studying adaptation across long evolutionary time scales: how does the distribution of beneficial mutations vary in 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 phylostratigraphy [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_N) to synonymous (d_S) substitutions rates, ω, and gene age in the divergence between humans and mouse, with young genes presenting a higher ω. Cai and Petrov [22] reported similar findings using human-chimpanzee divergence data. By looking at polymorphism data, 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 showed that these genes are more often positively selected. Similar correlations between ω and gene age have been observed in 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_N/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. Additionally, 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 ω_a and ω_{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 Brassicas *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

We tested the adaptive walk model of sequence evolution by assessing the impact of gene age on the rate of adaptive (ω_a) and non-adaptive (ω_{na}) non-synonymous substitutions. To assess whether the effect of gene age persisted when controlling for multiple confounding factors, we applied a non-parametric measure of correlation between gene age and ω_a and ω_{na} for each category of the co-factors analysed. The overall effect of gene age on ω_a and ω_{na} in each co-factor was assessed by combining significance values across tests in both species using the weighted Z-method [49].

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 (ω_a) and non-adaptive (ω_{na}) non-synonymous substitutions. We found that gene age significantly impacts estimates of ω, ω_a and ω_{na} in both species' pairs (Table 1 and Figure 1b). This result suggests that the higher ω 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 [50–52], 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 low support for a chromosomal effect (p = 0.041 for ω_{na} and p = 0.094 for ω_a). We, therefore, combined data from all chromosomes for subsequent analyses.

148 <u>The effect of gene age on the rate of molecular adaptation is robust to multiple confounding factors</u>
149 Genes of different ages intrinsically differ in their features [16,22,36]. As such traits significantly
150 impact the rate of molecular evolution [39], they may be confounding the faster adaptive rates
151 observed in young genes. Here, we assessed whether the effect of gene age on the rate of molecular
152 adaptation persisted after controlling for multiple confounding factors. To do so, we first assessed the
153 correlation of gene age with the rates of molecular evolution in distinct categories of genes, according
154 to a putative confounding factor. As estimates of the rate of adaptive substitutions for a small number
155 of genes exhibit large sampling variances [53,54], we could only assess each confounding factor
156 individually with this apporach.
157 Previous studies reported that younger genes encode shorter proteins [24,35,55] and are
158 expressed at lower levels [17,22,24,26], a pattern that we also observed in our data set (gene age vs.
159 protein length: Kendall's τ = -0.485, p = 2.82e-02; τ = -8.48, p = 1.06e-05, Figure S1a; gene age vs.
160 gene expression: τ = -0.595, p = 7.35e-03; τ = -0.790, p = 4.00e-05, Figure S1b in supplementary data;
161 for *D. melanogaster* and *A. thaliana,* respectively). As younger proteins are shorter than older ones,
162 they have a higher proportion of exposed residues [39]: gene age is significantly positively correlated
163 with the average RSA per gene (τ = 0.636, p = 3.98e-03; τ = 0.695, p = 3.03e-04, for *D. melanogaster*
164 and *A. thaliana* respectively; Figure S2a in supplementary data). Because exposed residues are more
165 flexible [56], young genes tend to encode proteins with a higher degree of intrinsic disorder, a pattern
166 previously reported in mice [36]. We confirmed this pattern in *D. melanogaster* (τ = 0.606, p = 6.10e-
167 03; Figure S2b in supplementary data) and *A. thaliana* (τ = 0.467, p = 1.53e-02; Figure S2b in
168 supplementary data).
169 We split our data into two roughly equal sized groups according to protein length, expression
170 level, average RSA, and average intrinsic disorder and reran the analysis within the "high" and "low"
171 groups, combining probabilities from the two analyses using the weighted Z-method [49]. Some
172 phylostrata were further combined when under-represented in some gene categories (see Material and
173 Methods). We found that ω, ω_{na} and ω_a remain significantly correlated to gene age, except when
174 controlling for protein length and gene expression for ω_a in *Arabidopsis* (Figure 2 and Table 1). This
175 weaker effect may be a consequence of how the most recent clades were combined in these analyses,
176 as there was little data available for those genes (see Material and Methods). Nonetheless, when
177 combining probabilities across the two species, we observed a significant correlation between all
178 measures of evolutionary rate and gene age controlling each of the co-factors (Table 1).
179 We further jointly estimated the combined effect of the potential confounding factors by
180 applying a recently developed method that extends the MK test with a generalized linear model [57].
181 This approach allows us to disentangle between direct and indirect effects on the rate of adaptive
182 substitutions by analysing multiple factors simultaneously. However, this method does not account for
183 segregating slightly deleterious mutations, which can bias estimates of the rate of adaptive
184 substitutions [58]. Hence, following the approach suggested in [57], we removed sites for which the

allele frequency was below 50% to minimize any potential bias. Despite the large reduction in the data set, this analysis revealed a strong effect of gene age (Table S1 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.

The effect of gene age on the molecular rate of adaptation is robust to BLAST's false negative rates

The physlostratigraphy approach has been previously used to date the emergence of new genes, and some studies have pointed out its potential limitations [28,59–62]. Because BLAST homology searches might fail to identify homologs in short or rapidly evolving genes, such genes could be mistakenly classified as young. To assess whether BLAST's false negative rate could explain the correlation between gene age and the rate of adaptive evolution, we analysed the gene age's effect by correcting the variation in E-value estimates from BLAST's searches between each gene and their respective outgroups. 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 with a subset of genes for which the correlation between the E-value and gene age was no longer significant (see Material and Methods and supplementary file S2) ($\tau = 0.408$, $p = 0.111$; $\tau = 0.354$, $p = 0.141$, for *D. melanogaster* and *A. thaliana* respectively; Figure S3b in supplementary data). We observed that the effect of gene age prevailed for all estimates in the two species (ω: $\tau = 0.929$, $p = 1.30e-03$; ω_{na}: $\tau = 0.786$, $p = 6.49e-03$; ω_a: $\tau = 0.643$, $p = 2.59e-02$ in *A. thaliana*; and ω: $\tau = 0.697$, $p = 1.61e-03$; ω_{na}: $\tau = 0.636$, $p = 3.98e-03$; ω_a: $\tau = 0.636$, $p = 3.98e-03$ 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 [63] and immune and stress responses [17,34,63]. As proteins encoding such functions tend to have higher molecular rates of adaptation [39,42,43,46,47,64], 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 assessed the effect of gene age separately for several GO-annotated genes, when a sufficient number of annotated genes in each age class was available (see Material and Methods). In *A. thaliana*, we found that the impact of gene age on ω_a 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 [65,66]. In *D. melanogaster*, we observed a strong effect of gene age on ω_a 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 ω_{na}, our analyses revealed a strong influence of gene age in 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 [42,43,47,64], gene age still has an impact on the efficiency of selection acting upon a protein.

<u>Substitutions in young genes have larger effect sizes</u>

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 [67] as a proxy for the fitness effects of amino-acid substitutions. We looked at the fixed differences separated by one mutational step between each pair of 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 = 2.00e-07; *Drosophila*: $\tau = 0.788$, p = 3.628e-04; Figure 3 and supplementary file S4), suggesting that substitutions in these genes have larger fitness effects than in old ones. To further test whether these larger fitness 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 amongst 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

258 individually significant in *Drosophila* (p = 0.006) and significant across species when we combine
259 probabilities (p = 0.009) (Figure S7). We observed the same pattern in the analysis of $\bar{G}_{na}$ with a
260 significant correlation in *Arabidopsis* (p = 0.004) and a significant correlation when combining the
261 probabilities from both species (p = 0.0007) (Figure S7). These findings suggest that substitutions in
262 young genes tend to be more adaptive and under less constraint, although the strength of selection
263 differs between species.
264
265 **Discussion**
266 We used a population genomic approach to disentangle the effects of positive and negative selection
267 on the rate of non-synonymous substitutions. Using complete genome data from two *Arabidopsis* and
268 *Drosophila* species, we showed that the higher rate of non-synonymous substitutions in younger genes
269 results both from relaxed purifying selection (higher ω_{na}) and a higher rate of adaptive substitutions
270 (higher ω_a) (Figure 1b). By looking at the magnitude effect of gene age, we observed that young
271 genes present a 25-fold higher rate of adaptation than older genes in *Drosophila* species and around
272 30-fold higher in *Arabidopsis*. The magnitude of this effect is higher than that observed for
273 recombination rate and solvent exposure in these species, two other factors strongly correlated to the
274 rate of adaptive evolution [39,68]. We also observe that young genes undergo substitutions that are
275 larger in terms of physicochemical properties than older genes. A question remains: what are the
276 drivers of these effects?
277
278 <u>The magnitude effect of gene age on adaptive evolution is species-specific</u>
279 Although we observed a strong impact of gene age on the molecular adaptive rate in both species
280 pairs, the shape of the relation between these two variables differs. While the relationship between
281 gene age and ω_a is monotonously increasing in *Arabidopsis*, it has several peaks in *Drosophila*
282 (Figure 1b). This pattern is particularly evident if we discard the two youngest clades. In *Drosophila*,
283 the correlation becomes much weaker and non-significant for ω_a (ω: τ = 0.600, p = 0.016; ω_{na}: τ
284 =0.556, p = 0.025; ω_a: τ = 0.467, p = 0.060), whereas, in *Arabidopsis,* the effect of gene age persists
285 (ω: τ = 0.9487, p = 6.342e-06; ω_{na}: τ = 0.872, p = 3.345e-05; ω_a: τ = 0.692, p = 9.86e-04).
286 Intriguingly, this multimodal distribution of ω_a observed in *Drosophila* resembles the pattern of
287 emergence of young genes in this species [17]. The peak in the adaptive substitution rate observed for
288 clades 6 and 7 (Figure 1b) coincided with the animal phyla's major radiation at the time of extensive
289 periods of glacial cycles [69]. When looking at the functions coded by these proteins, we found that
290 they are linked to a wide range of vital cellular and biological processes, such as defence mechanisms
291 and cell differentiation (Figure S8 in supplementary data). This pattern suggests that these genes might
292 be experiencing higher molecular adaptive rates due to their role in such vital processes. However, for
293 these genes to keep such high rates of adaptive substitutions until recent times (*i.e.*, in the branch
294 between *D. melanogaster* and *D. simulans*), epistatic interactions might be at play. Studies across taxa

295 have proposed that functional epistasis is an important factor in the evolution of genes involved in
296 defence mechanisms and adaptation to new environmental stresses [70–74]. We posit that such gene
297 interactions keep these proteins further from their optimum throughout time due to the rugged shape of
298 the fitness landscape, leading to the high molecular adaptive rates observed in the branch between *D.*
299 *melanogaster* and *D. simulans*. To further test this hypothesis, we used the degree of protein-protein
300 interactions (PPI) as a proxy for epistatic interactions and analysed its relationship with gene age. We
301 observed that genes in clade 7 have a slightly higher degree of PPI than other strata (Figure S9 in
302 supplementary data and supplementary file S5), suggesting that these genes might be experiencing
303 relatively more epistatic interactions. These findings are consistent with epistasis influencing the
304 evolution of these genes, potentially explaining their continued higher rates of molecular adaptation.
305 In contrast, the burst of the emergence of new genes in *Arabidopsis* coincided with the plant-specific
306 radiation right before the emergence of Brassicaceae [17,75]. This trend is consistent with our results
307 from *A. thaliana*, where the bursts of ω_a occur in younger clades (after clades 11 and 12 in Figure 1b).
308 These distinct patterns observed between species suggest that the role of a gene's age in molecular
309 adaptation is complex, as also evidenced by the lack of a significant correlation with ω_a previously
310 reported in humans [22]. The authors proposed that this result may be a consequence of the generally
311 low molecular adaptive rates observed in primates [22,48].
312 Despite these species-specific trends, our analyses revealed a strong correlation between ω_a
313 and gene age extending through hundreds of millions of years (Figures 1 and 2). These findings
314 suggest a consistent effect of a gene's age on the rate of molecular adaptation across taxa.
315
316 <u>An adaptive walk model of gene evolution</u>
317 Our study highlighted that, after their emergence, young genes evolve through relaxed selection, as
318 first proposed by Ohno [76], but also by acquiring beneficial mutations, as described in the "adaptive-
319 conflict" model [37,77]. Ohno's idea of evolution was "non-Darwinian" in its nature, as he believed
320 that "natural selection merely modified while redundancy created" [76]. He proposed that new genes
321 evolve by accumulating "forbidden" mutations, where they are only preserved if the development of a
322 formerly non-existent function occurs, a process known as neo-functionalisation. In this scenario,
323 natural selection only acts at the stage of acquiring a new function. Further extensions of this theory
324 suggested that the preservation of a new gene can also occur through sub-functionalisation, where the
325 accumulation of deleterious mutations leads to a complementary loss of function in both copies of the
326 gene [78,79].
327 In contrast, the "adaptive-conflict" model assumes that the ancestral gene could carry more
328 than two pleiotropically constrained functions [37,77]. Once the duplication event occurs, each copy
329 then becomes specialised in one of the ancestral functions. In this case, the ancestral gene's split
330 proceeded through positive Darwinian selection [37,77]. These theories are based on the evolution of
331 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 [67] 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, by looking at the proportion of adaptive substitutions across Grantham's distances categories in each age class, we observed that adaptive substitutions in young genes are under stronger positive selection (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 only weak 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. 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 to characterize the specificities 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 S2 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 S3 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 [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].

<u>Estimation of the adaptive and non-adaptive rate of non-synonymous substitutions</u>

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 ω, ω_{na} and ω_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).

406 Gene age vs. protein length and gene expression
407 Data on protein length and gene expression were included in the datasets of Moutinho et al. [39]
408 (Tables S2 and S3 in supplementary data online). Gene expression data were downloaded from the
409 database Expression Atlas [100], taking one baseline experiment for each species [39]. Mean gene
410 expression levels were then obtained by averaging across samples and tissues for each gene. To
411 correct for these two factors, we divided our dataset into two equally sized groups based on the factor
412 we wished to control. Short proteins had a size up to 366 and 389 amino-acids, and long proteins had a
413 size up to 4,674 and 5,098 amino-acids in *A. thaliana* and *D. melanogaster*, respectively. We further
414 merged phylostrata containing a low number of genes. For *D. melanogaster,* we categorised gene age
415 into 6 main clades by combining clades 3-4, 5-6, 7-10, and 11-12, keeping the others unchanged. In *A.*
416 *thaliana*, we combined the 15 clades in 6 main groups by merging clades 5-8 and 9-15. For gene
417 expression, we used a total of 17,126 and 6,247 genes for *A. thaliana* and *D. melanogaster,*
418 respectively, being categorised as lowly and highly expressed. Genes were classified as lowly
419 expressed if the mean expression levels were up to 10.3 and 6.8, and highly expressed genes were the
420 ones with expression up to 6,632.8 and 4,237.0 in *A. thaliana* and *D. melanogaster*, respectively. For
421 *D. melanogaster*, we categorised gene age in 6 categories by combining clades 3-5, 6-9, and 10-12. In
422 *A. thaliana*, we combined the data in 6 clades, merging clades 4-7, 8-11 and 12-15.
423
424 Gene age vs. protein structure
425 Since most young genes lack a defined three-dimensional structure [36], they do not have information
426 on the residue's solvent accessibility. Hence, we used a deep learning approach, NetSurfP-2.0, that
427 predicts the RSA of each residue from the amino-acid sequence [101] by using as a training model the
428 HH-suite sequence alignment tool for protein similarity searches [102]. To assess whether this
429 approach provided reliable results, we compared the RSA estimates of NetSurfP-2.0 with those
430 obtained from the PDB structures in our dataset [39]. We found a good correlation between the two
431 approaches for both species (Kendall's $\tau = 0.571$, $p < 2e\text{-}216$; $\tau = 0.462$, $p < 2e\text{-}216$, for *D.*
432 *melanogaster* and *A. thaliana* respectively). Using NetSurfP-2.0, RSA estimates were successfully
433 obtained for a total of 4,238,686 (88% of the total codon sites) and 7,479,807 (99% of the total codon
434 sites) amino-acid residues for *D. melanogaster* and *A. thaliana,* respectively. To assess the impact of
435 RSA at the gene level, we analysed the total number of genes in both species by making two
436 categories of genes according to the average RSA value per gene. Genes with lower RSA had mean
437 values between 0.127-0.389 in *Drosophila* and 0.217-0.386 in *Arabidopsis*. Genes with a higher RSA
438 had mean values between 0.390-0.894 in *Drosophila* and 0.387-0.898 in *Arabidopsis*. The phylostrata
439 groups were defined by combining clades 7-8 in *D. melanogaster*, and 8-11, 12-15 in *A. thaliana*.
440 For the analysis on the residue intrinsic disorder, we used the estimates included in the
441 datasets of Moutinho et al. [39]. These data were obtained with the software DisEMBL [103], where
442 intrinsic disorder was estimated per site and classified according to the degree of "hot loops" *i.e.*,

443 highly mobile loops. This analysis was performed for 17,732 and 7,410 genes for *A. thaliana* and *D.*
444 *melanogaster,* respectively. Genes were combined into two categories according to the mean value of
445 their residue's intrinsic disorder. Genes with a low level of intrinsic disorder had values between
446 0.029-0.080 in *Drosophila* and among 0.041-0.084 in *Arabidopsis*. Genes with a higher degree of
447 intrinsic disorder had values between 0.081-0.554 in *Drosophila* and among 0.085-0.551 in
448 *Arabidopsis*. In *D. melanogaster*, all the 12 phylostrata could be used. In *A. thaliana*, the 15 strata
449 were combined in 12 categories by merging clades 9-10, 11-12 and 13-14.
450
451 Correcting for BLAST e-values
452 We analysed the robustness of the gene age's effect by correcting the variation in the Expect (E) value
453 estimates in BLAST's searches between our focus species and their respective outgroups. By reducing
454 the variation in E-values estimates, we could correct for potential failures in BLAST's homology
455 searches. To do so, we used a subset of genes for which the correlation between the E value and gene
456 age was no longer significant: 12,472 genes with an E value lower than 1e-150 for *A. thaliana* and
457 7,104 genes with an E value lower than 1e-100 for *D. melanogaster* (supplementary file S2). For *A.*
458 *thaliana*, analyses were carried by combining clades 8-13, with no genes left in clades 14 and 15. For
459 *D. melanogaster*, analyses were performed with the 12 strata.
460
461 Gene age vs. protein function
462 Gene ontology terms were obtained from the "dmelanogaster_gene_ensembl" and the
463 "athaliana_eg_gene" tables in the Ensembl database (version 103), through the R package "biomaRt"
464 [104]. A total of 7,253 (~70% of the genes) and 15,604 (~80% of the genes) genes were mapped in *D.*
465 *melanogaster* and *A. thaliana*, respectively. To check whether the effect of gene age prevailed across
466 functional protein classes, we analysed the GO terms with the highest number of young genes
467 mapped: more than 50 genes present in Clades 11 and 12 in *D. melanogaster*; and more than 30 genes
468 present in Clades 12 to 15 in *A. thaliana*. This filtering step resulted in 6,637 genes across 23 GO
469 categories in *D. melanogaster* (Table S2 in Supplementary Data online), and 15,410 genes across 10
470 GO categories in *A. thaliana* (Table S3 in Supplementary Data online). To analyse the effect of gene
471 age, we compared three age classes. In *D. melanogaster*, the first age category spanned over
472 phylostrata 1-3, the second category covered clades 4-7, and the third one included clades 8-12. In *A.*
473 *thaliana*, the first category comprised genes belonging to clades 1-6, the second category spanned over
474 clades 7-11, and the third one included the phylostrata between clades 12-15 (Figure 1a).
475
476 Gene Age vs. Protein-protein interactions (PPI)
477 We obtained PPI data for *D. melanogaster* from the STRING database [105], which includes both
478 physical and functional interactions (https://string-db.org/). This database included 13,046 proteins

with annotated interactions, which were used to analyse the distribution of protein networks across phylostrata.

The fitness effects of amino-acid substitutions

We used Grantham's physicochemical distances between amino-acids [67] 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 ($\bar{G}_{na}$) non-synonymous substitutions in 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 <> C mutation, with the synonymous SFS of 4-fold degenerate sites separated only by A <> C mutations ($SFS_{4F(A<>C)}$). 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 <> T and C <> G mutations. The synonymous SFS for this amino acid pair was estimated as $SFS_{4F(weighted)} = \left[(f_{AGT} + f_{TGT} + f_{AGC} + f_{TGC}) * SFS_{4F(A<>T)} + (f_{TCT} + f_{TGT} + f_{TCC} + f_{TGC}) * SFS_{4F(C<>G)}\right] / (f_{AGT} + f_{TGT} + f_{AGC} + f_{TGC} + f_{TGT} + f_{TGC})$. 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 ω_a and ω_{na} 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 ($\bar{G}_{na}$) non-synonymous substitutions in each age stratum as $\bar{G}_a = \sum_i \omega_{ai} \bar{g}_i N_i / \sum_i N_i$ and $\bar{G}_{na} = \sum_i \omega_{nai} \bar{g}_i N_i / \sum_i N_i$ where the sum is across the groups of amino acids grouped by Grantham's distance, ω_{ai} and ω_{nai} are the rates of adaptive and non-adaptive substitution in the *i*th group, $\overline{g_i}$ is the average grantham's distance amongst the amino acid pairs in the group and N_i is the number of non-synonymous sites in the *i*th group.

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 ω_a and ω_{na} were estimated for each resampling, allowing to compute the rate differences between categories. A one-tailed P-value can be obtained using the formula P = (2k + 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 * 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 posthoc 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 ω_a and ω_{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 ω_a and ω_{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, the MK regression analysis was performed per site, where we analysed the combined effect of RSA, intrinsic protein disorder, protein length, gene expression, protein function and the sex chromosome in *Drosophila*. As this method does not deal with categorical variables, we created a binary variable for protein function based on the gene's presence (1.0) or absence (0.0) in stress-related proteins (*i.e.*, response to stress, response to external stimulus and signal transduction). We used the same rationale for the chromosome X 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%.

Acknowledgments

The authors thank Diethard Tautz, Tal Dagan and Chaitanya Gokhale for fruitful discussions. J.Y.D. acknowledges funding from the Max Planck Society.

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.

553 4. Smith JM. Natural selection and the concept of a protein space. Nature. 1970;225(5232):563–4.
554 5. Gillespie JH. A Simple Stochastic Gene Substitution Model. Theor Popul Biol. 1983;23:202–
555 15.
556 6. Gillespie JH. Molecular evolution over the mutational landscape. Evolution. 1984;38(5):1116–
557 29.
558 7. Gillespie JH. The Causes of Molecular Evolution. Oxford University Press; 1991.
559 8. Kimura M. The Neutral Theory of Molecular Evolution. Cambridge Univeristy Press; 1983.
560 9. Orr HA. The Population Genetics of Adaptation: The Distribution of Factors Fixed during
561 Adaptive Evolution. Evolution. 1998;52(4):935.
562 10. Orr AH. The evolutionary genetics of adaptation: A simulation study. Genet Res.
563 1999;74(3):207–14.
564 11. Lenski RE, Rose MR, Simpson SC, Tadler SC. Long-Term Experimental Evolution in
565 Escherichia coli. I. Adaptation and Divergence During 2,000 Generations. Am Nat.
566 1991;138(6):1315–41.
567 12. Cooper VS, Lenski RE. The population genetics of ecological specialization in evolving
568 Escherichia coli populations. Nature. 2000;407(6805):736–9.
569 13. Gerrish P. The rhythm of microbial adaptation. Nature. 2001;413(6853):299–302.
570 14. Rozen DE, De Visser JAGM, Gerrish PJ. Fitness effects of fixed beneficial mutations in
571 microbial populations. Curr Biol. 2002;12(12):1040–5.
572 15. Schoustra SE, Bataillon T, Gifford DR, Kassen R. The properties of adaptive walks in evolving
573 populations of fungus. PLoS Biol. 2009;7(11): e1000250.
574 16. Lynch M. Genomics: Gene duplication and evolution. Science. 2002;297(5583):945–7.
575 17. Tautz D, Domazet-Lošo T. The evolutionary origin of orphan genes. Nat Rev Genet.
576 2011;12(10):692–702.
577 18. Cohen O, Ashkenazy H, Belinky F, Huchon D, Pupko T. GLOOME: Gain loss mapping
578 engine. Bioinformatics. 2010;26(22):2914–5.
579 19. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, et al. Gapped blast and psi-
580 blast: a new generation of protein database search programs. FASEB J. 1998;12(8):3389–402.
581 20. Domazet-Lošo T, Brajković J, Tautz D. A phylostratigraphy approach to uncover the genomic
582 history of major adaptations in metazoan lineages. Trends Genet. 2007;23(11):533–9.
583 21. Thornton K, Long M. Rapid divergence of gene duplicates on the Drosophila melanogaster X
584 chromosome. Mol Biol Evol. 2002;19(6):918–25.
585 22. Cai JJ, Petrov DA. Relaxed purifying selection and possibly high rate of adaptation in primate
586 lineage-specific genes. Genome Biol Evol. 2010;2(1):393–409.
587 23. Domazet-Loso T, Tautz D. An evolutionary analysis of orphan genes in Drosophila. Genome
588 Res. 2003;13(10):2213–9.
589 24. Vishnoi A, Kryazhimskiy S, Bazykin GA, Hannenhalli S, Plotkin JB. Young proteins

590 experience more variable selection pressures than old proteins. Genome Res.
591 2010;20(11):1574–81.
592 25. Cai JJ, Woo PCY, Lau SKP, Smith DK, Yuen KY. Accelerated evolutionary rate may be
593 responsible for the emergence of lineage-specific genes in Ascomycota. J Mol Evol.
594 2006;63(1):1–11.
595 26. Wolf YI, Novichkov PS, Karev GP, Koonin E V., Lipman DJ. The universal distribution of
596 evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent
597 ages. Proc Natl Acad Sci. 2009;106(18):7273–80.
598 27. Albà MM, Castresana J. Inverse relationship between evolutionary rate and age of mammalian
599 genes. Mol Biol Evol. 2005;22(3):598–606.
600 28. Domazet-Lošo T, Carvunis AR, Albà MM, Šestak MS, Bakarić R, Neme R, et al. No evidence
601 for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution.
602 Mol Biol Evol. 2017;34(4):843–56.
603 29. Zhang YE, Vibranovski MD, Krinsky BH, Long M. Age-dependent chromosomal distribution
604 of male-biased genes in Drosophila. Genome Res. 2010;20(11):1526–33.
605 30. Daubin V, Ochman H. Bacterial genomes as new gene homes: The genealogy of ORFans in E.
606 coli. Genome Res. 2004;14(6):1036–42.
607 31. García-Vallvé S, Alonso Á, Bravo IG. Papillomaviruses: Different genes have different
608 histories. Trends Microbiol. 2005;13(11):514–21.
609 32. Cui X, Lv Y, Chen M, Nikoloski Z, Twell D, Zhang D. Young genes out of the male: An
610 insight from evolutionary age analysis of the pollen transcriptome. Mol Plant. 2015;8(6):935–
611 45.
612 33. Arendsee ZW, Li L, Wurtele ES. Coming of age: Orphan genes in plants. Trends Plant Sci.
613 2014;19(11):698–708.
614 34. Kuo CH, Kissinger JC. Consistent and contrasting properties of lineage-specific genes in the
615 apicomplexan parasites Plasmodium and Theileria. BMC Evol Biol. 2008;8(1):1–16.
616 35. Neme R, Tautz D. Phylogenetic patterns of emergence of new genes support a model of
617 frequent de novo evolution. BMC Genomics. 2013;14(1):1-13.
618 36. Wilson BA, Foy SG, Neme R, Masel J. Young genes are highly disordered as predicted by the
619 preadaptation hypothesis of de novo gene birth. Nat Ecol Evol. 2017;1(6):1-6.
620 37. Hughes AL. The evolution of functionally novel proteins after gene duplication. Proc R Soc B.
621 1994;256(1346):119–24.
622 38. Zhang J, Zhang Y ping, Rosenberg HF. Adaptive evolution of a duplicated pancreatic
623 ribonuclease gene in a leaf-eating monkey. Nat Genet. 2002;30(4):411–5.
624 39. Moutinho AF, Trancoso FF, Dutheil JY. The impact of protein architecture on adaptive
625 evolution. Mol Biol Evol. 2019;36(9): 2013-2028.
626 40. Afanasyeva A, Bockwoldt M, Cooney CR, Heiland I, Gossmann TI. Human long intrinsically

627 disordered protein regions are frequent targets of positive selection. Genome Res.
628 2018;28(7):975–82.
629 41. Subramanian S, Kumar S. Gene expression intensity shapes evolutionary rates of the proteins
630 encoded by the vertebrate genome. Genetics. 2004;168(1):373–81.
631 42. Stukenbrock EH, Bataillon T, Dutheil JY, Hansen TT, Li R, Zala M, et al. The making of a
632 new pathogen: Insights from comparative population genomics of the domesticated wheat
633 pathogen Mycosphaerella graminicola and its wild sister species. Genome Res.
634 2011;21(12):2157–66.
635 43. Enard D, Cai L, Gwennap C, Petrov DA. Viruses are a dominant driver of protein adaptation in
636 mammals. Elife. 2016;5: e12469.
637 44. Rocha EPC, Danchin A. An Analysis of Determinants of Amino Acids Substitution Rates in
638 Bacterial Proteins. Mol Biol Evol. 2004;21(1):108–16.
639 45. Pal C, Papp B, Hurst LD. Highly Expressed Genes in Yeast Evolve Slowly. Genetics.
640 2001;158(1998):927–31.
641 46. Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG. Dynamic evolution
642 of the innate immune system in Drosophila. Nat Genet. 2007;39(12):1461–8.
643 47. Obbard DJ, Welch JJ, Kim KW, Jiggins FM. Quantifying adaptive evolution in the Drosophila
644 immune system. PLoS Genet. 2009;5(10): e1000698.
645 48. Galtier N. Adaptive Protein Evolution in Animals and the Effective Population Size
646 Hypothesis. PLoS Genet. 2016;12(1):1–23.
647 49. Whitlock MC, Whitlock MC. Combining probability from independent tests: the weighted Z-
648 method is superior to Fisher's approach. Wiley Online Libr. 2005;18(5):1368–73.
649 50. Vicoso B, Charlesworth B. Evolution on the X chromosome: Unusual patterns and processes.
650 Nat Rev Genet. 2006;7(8):645–53.
651 51. Vicoso B, Charlesworth B. Effective population size and the faster-X effect: An extended
652 model. Evolution. 2009;63(9):2413–26.
653 52. Charlesworth B, Campos JL, Jackson BC. Faster-X evolution: Theory and evidence from
654 Drosophila. Mol Ecol. 2018;27(19):3753–71.
655 53. Stoletzki N, Eyre-Walker A. Estimation of the neutrality index. Mol Biol Evol. 2011;28(1):63–
656 70.
657 54. Smith NGC, Eyre-Walker a. Adaptive protein evolution in Drosophila . Nature. 2002;
658 415(6875):1022-1024.
659 55. Ding Y, Zhou Q, Wang W. Origins of New Genes and Evolution of Their Novel Functions.
660 Annu Rev Ecol Evol Syst. 2012;43(1):345–63.
661 56. Choi SS, Vallender EJ, Lahn BT. Systematically assessing the influence of 3-dimensional
662 structural context on the molecular evolution of mammalian proteomes. Mol Biol Evol.
663 2006;23(11):2131– 2133.

664 57. Huang Y-F. Dissecting Genomic Determinants of Positive Selection with an Evolution-Guided
665 Regression Model. Mol Biol Evol. 2021(msab291).
666 58. Eyre-Walker A, Keightley PD. Estimating the rate of adaptive molecular evolution in the
667 presence of slightly deleterious mutations and population size change. Mol Biol Evol.
668 2009;26(9):2097–108.
669 59. Elhaik E, Sabath N, Graur D. The "inverse relationship between evolutionary rate and age of
670 mammalian genes" is an artifact of increased genetic distance with rate of evolution and time
671 of divergence. Mol Biol Evol. 2006;23(1):1–3.
672 60. Moyers BA, Zhang J. Phylostratigraphic bias creates spurious patterns of genome evolution.
673 Mol Biol Evol. 2015;32(1):258–67.
674 61. Moyers BA, Zhang J. Evaluating Phylostratigraphic Evidence for Widespread de Novo Gene
675 Birth in Genome Evolution. Mol Biol Evol. 2016;33(5):1245–56.
676 62. Albà MM, Castresana J. On homology searches by protein Blast and the characterization of the
677 age of genes. BMC Evol Biol. 2007;7(1):1–8.
678 63. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. More than just orphans: are
679 taxonomically-restricted genes important in evolution? Trends Genet. 2009;25(9):404–13.
680 64. Slotte T, Bataillon T, Hansen TT, St. Onge K, Wright SI, Schierup MH. Genomic determinants
681 of protein evolution and polymorphism in arabidopsis. Genome Biol Evol. 2011;3(1):1210–9.
682 65. Geldner N, Robatzek S. Plant receptors go endosomal: A moving view on signal transduction.
683 Plant Physiol. 2008;147(4):1565–74.
684 66. Groen AJ, De Vries SC, Lilley KS. A proteomics approach to membrane trafficking. Plant
685 Physiol. 2008;147(4):1584–9.
686 67. Grantham R. Amino Acid Difference Formula to Help Explain Protein Evolution. Science.
687 1974;185(4154):862–4.
688 68. Castellano D, Coronado-Zamora M, Campos JL, Barbadilla A, Eyre-Walker A. Adaptive
689 evolution is substantially impeded by hill-Robertson interference in drosophila. Mol Biol Evol.
690 2016;33(2):442–55.
691 69. Hoffman PF, Kaufman AJ, Halverson GP, Schrag DP. A neoproterozoic snowball earth.
692 Science. 1998;281(5381):1342–6.
693 70. Montooth KL, Marden JH, Clark AG. Mapping Determinants of Variation in Energy
694 Metabolism, Respiration and Flight in Drosophila. Genetics. 2003;165(2):623–35.
695 71. Paixão T, Barton NH. The effect of gene interactions on the long-term response to selection.
696 Proc Natl Acad Sci. 2016;113(16):4422–7.
697 72. Hansen TF. Why Epistasis is Important for Selection and Adaptation. Evolution.
698 2013;67(12):3501–11.
699 73. Behrman EL, Howick VM, Kapun M, Staubach F, Bergland AO, Petrov DA, et al. Rapid
700 seasonal evolution in innate immunity of wild Drosophila melanogaster. Proc R Soc B Biol

701 Sci. 2018; 285(1870): 20172599.
702 74. Lagator M, Colegrave N, Neve P. Selection history and epistatic interactions impact dynamics
703 of adaptation to novel environmental stresses. Proc R Soc B Biol Sci. 2014;281(1794).
704 75. Wang H, Moore MJ, Soltis PS, Bell CD, Brockington SF, Alexandre R, et al. Rosid radiation
705 and the rapid rise of angiosperm-dominated forests. Proc Natl Acad Sci. 2009;106(10):3853–8.
706 76. Ohno S. Evolution by gene duplication. Springer Science & Business Media; 1970.
707 77. Piatigorsky J, Wistow G. The recruitment of crystallins: new functions precede gene
708 duplication. Science. 1991;252(5009):1078–9.
709 78. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. Preservation of duplicate
710 genes by complementary, degenerative mutations. Genetics. 1999;151(4):1531–45.
711 79. Prince VE, Pickett FB. Splitting pairs: The diverging fates of duplicated genes. Nat Rev Genet.
712 2002;3(11):827–37.
713 80. Jacob F. Evolution and Tinkering. Science. 1977;196(4295):1161–6.
714 81. Cai J, Zhao R, Jiang H, Wang W. De novo origination of a new protein-coding gene in
715 Saccharomyces cerevisiae. Genetics. 2008;179(1):487–96.
716 82. Heinen TJAJ, Staubach F, Häming D, Tautz D. Emergence of a New Gene from an Intergenic
717 Region. Curr Biol. 2009;19(18):1527–31.
718 83. Zhao L, Saelao P, Jones CD, Begun DJ. Origin and spread of de novo genes in Drosophila
719 melanogaster populations. Science. 2014;343(6172):769–72.
720 84. Neme R, Tautz D. Evolution: Dynamics of de novo gene emergence. Curr Biol.
721 2014;24(6):R238–40.
722 85. Palmieri N, Kosiol C, Schlötterer C. The life cycle of Drosophila orphan genes. Elife. 2014;3:
723 e01311.
724 86. Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, et al. Proto-
725 genes and de novo gene birth. Nature. 2012;487(7407):370–4.
726 87. Smith JM. Natural selection and the concept of a protein space. Nature. 1970;225:563–4.
727 88. Orr HA. the Population Genetics of Adaptation: the Adaptation of Dna Sequences. Evolution.
728 2002;56(7):1317.
729 89. Bustamante CD, Nielsen R, Sawyer SA, Olsenk KM, Purugganank MD, Hartl DL. The Cost of
730 Inbreeding in Arabidopsis. Nature. 2002; 416(6880), 531-534.
731 90. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral
732 molecular variation. Genetics. 1993;134(4):1289–303.
733 91. Assaf ZJ, Petrov DA, Blundell JR. Obstruction of adaptation in diploids by recessive, strongly
734 deleterious alleles. Proc Natl Acad Sci. 2015;112(20):E2658–66.
735 92. Di C, Murga Moreno J, Salazar-Tortosa DF, Lauterbur ME, Enard D. Decreased recent
736 adaptation at human mendelian disease genes as a possible consequence of interference
737 between advantageous and deleterious variants. Elife. 2021;10:e69026.

738 93. Gossmann TI, Keightley PD, Eyre-Walker A. The effect of variation in the effective population
739 size on the rate of adaptive molecular evolution in eukaryotes. Genome Biol Evol.
740 2012;4(5):658–67.
741 94. Eyre-Walker A. The genomic rate of adaptive evolution. Trends Ecol Evol. 2006;21(10):569–
742 575.
743 95. Rousselle M, Simion P, Tilak M-K, Figuet E, Nabholz B, Galtier N. Is adaptation limited by
744 mutation? A timescale-dependent effect of genetic diversity on the adaptive substitution rate in
745 animals. PLOS Genet. 2020;16(4):e1008668.
746 96. Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, Crepeau MW, et al.
747 Population Genomics of Sub-Saharan Drosophila melanogaster: African Diversity and Non-
748 African Admixture. PLoS Genet. 2012;8(12): e1003080.
749 97. Weigel D, Mott R. The 1001 Genomes Project for Arabidopsis thaliana. Genome Biol. 2009;
750 10(5):1-5.
751 98. Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette NC, et al. Bio++: Efficient
752 extensible libraries and tools for computational molecular evolution. Mol Biol Evol.
753 2013;30(8):1745–50.
754 99. Wickham H. ggplot2 - Elegant Graphics for Data Analysis (2nd Edition). J Stat Softw.
755 2017;77:2–5.
756 100. Petryszak R, Keays M, Tang YA, Fonseca NA, Barrera E, Burdett T, et al. Expression Atlas
757 update - An integrated database of gene and protein expression in humans, animals and plants.
758 Nucleic Acids Res. 2016; 44(D1): D746-D752.
759 101. Klausen MS, Jespersen MC, Nielsen H, Jensen KK, Jurtz VI, Sønderby CK, et al. NetSurfP-
760 2.0: Improved prediction of protein structural features by integrated deep learning. Proteins
761 Struct Funct Bioinforma. 2019;87(6):520–7.
762 102. Remmert M, Biegert A, Hauser A, Söding J. HHblits: Lightning-fast iterative protein sequence
763 searching by HMM-HMM alignment. Nat Methods. 2012;9(2):173–5.
764 103. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction:
765 Implications for structural proteomics. Structure. 2003;11(11): 1453-1459.
766 104. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and
767 Bioconductor: A powerful link between biological databases and microarray data analysis.
768 Bioinformatics. 2005;21(16):3439–40.
769 105. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, et al. STRING 8 - A global
770 view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res.
771 2009;37(SUPPL. 1): D412-D416.
772 106. Bergman J, Eyre-Walker A. Does Adaptive Protein Evolution Proceed by Large or Small Steps
773 at the Amino Acid Level? Mol Biol Evol. 2019; 36(5):990-998..
774 107. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate : A Practical and Powerful

Approach to Multiple Testing. J R Stat Soc. 1995;57(1):289–300.

108. 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.

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

111. Zeileis A, Hothorn T. Diagnostic checking in regression relationships.. R News. 2002;2(3):7–10.

Figures

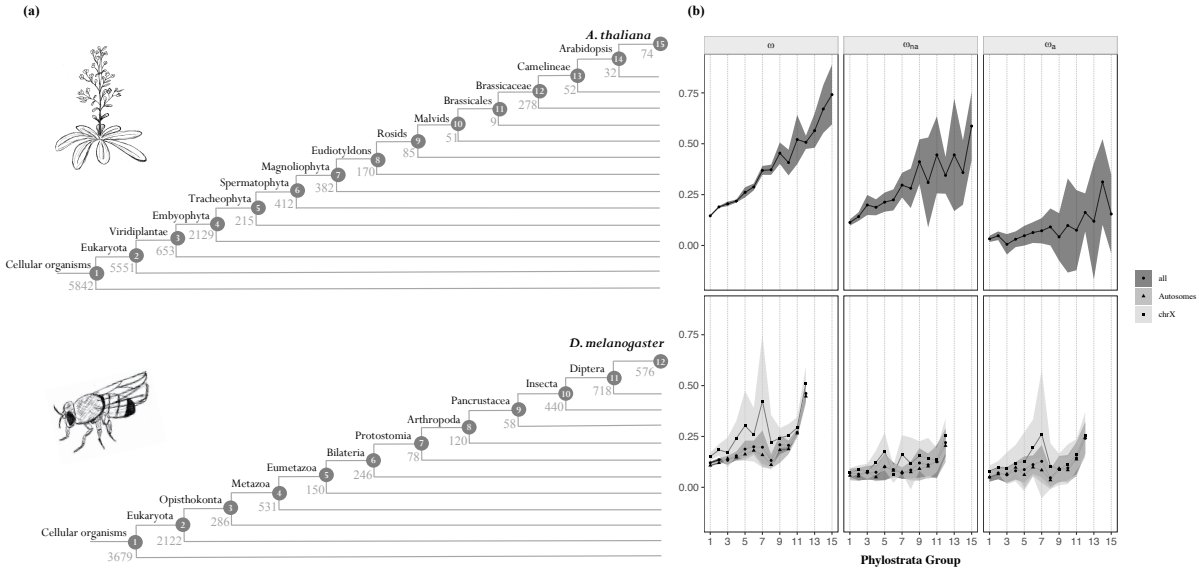

Figure 1. (a) Phylogenetic definition of the strata used in the analyses for *A. thaliana* (top) and *D. melanogaster* (bottom). The number of genes mapped to each clade is shown. **(b)** Relationship between the rate of protein evolution (ω), non-adaptive non-synonymous substitutions (ω_{na}) and adaptive 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 shown. 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.

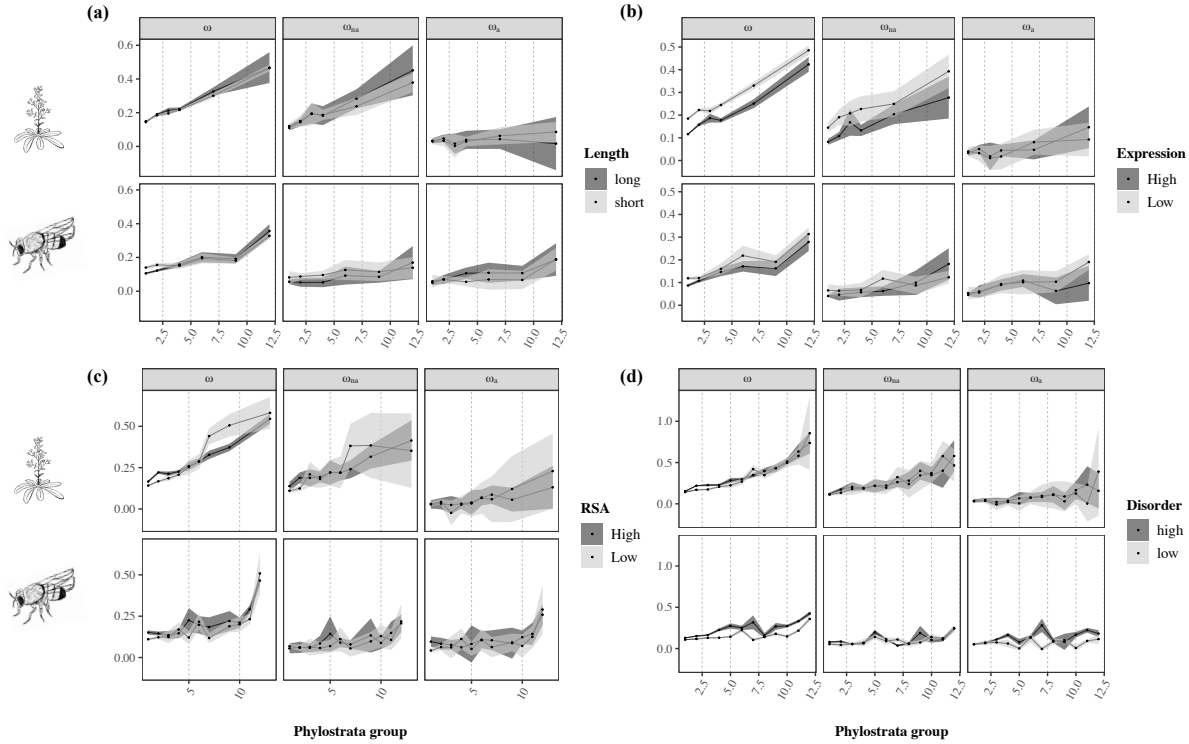

Figure 2. Estimates of ω, ω_{na} and ω_a plotted as a function of **(a)** protein length and **(b)** mean expression levels, **(c)** relative solvent accessibility, and **(d)** protein intrinsic disorder with gene age in *A. thaliana* (top) and *D. melanogaster* (bottom). Analyses were performed by comparing 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.

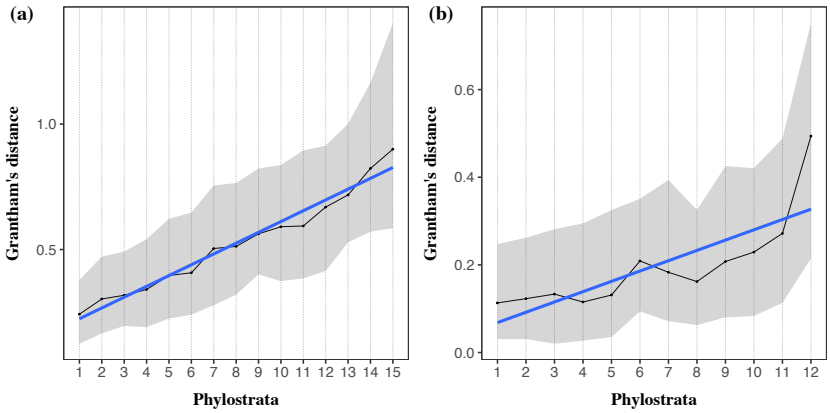

Figure 3. Relationship between gene age and Grantham's distance between amin-acids for *A. thaliana* (a) and *D. melanogaster* (b). A linear model was fitted between gene age and Grantham's distances values and is represented with the blue line. For each clade, the median value of the Grantham's distance between residues is depicted with the black dot. The shaded area represents the physicochemical distances within the 1st and 3rd quartile.

Tables

Table 1. Kendall's correlation coefficients for the relationship between ω, ω_{na} and ω_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 ω, ω_{na} and ω_a.

		Arabidopsis			*Drosophila*			**Weighted Z across species**		
		ω	$\boldsymbol{\omega_{na}}$	$\boldsymbol{\omega_a}$	**ω**	$\boldsymbol{\omega_{na}}$	$\boldsymbol{\omega_a}$	**ω**	$\boldsymbol{\omega_{na}}$	$\boldsymbol{\omega_a}$
Gene Age		0.962 ***	0.848 ***	0.733 ***	0.727 ***	0.697 **	0.636 **			
Protein Length	**Long**	1.000 **	0.867 *	-0.200	0.867 *	0.600 (.)	0.867 *	1.56e-04 ***	7.71e-05 ***	7.98e-03 **
	Short	1.000 **	0.867 *	0.600 (.)	0.733 *	0.867 *	0.467			
	Weighted Z	6.46e-04 ***	1.61e-03 **	0.133	2.64e-03 **	5.29e-03 **	0.0105 *			
Gene Expression	**High**	0.867 *	0.867 *	0.467	0.867 *	1.000 **	0.600 (.)	6.93e-05 ***	6.89e-06 ***	3.53e-03 **
	Low	0.867 *	1.000 **	0.333	0.867 *	0.733 *	1.000 **			
	Weighted Z	1.51e-03 **	3.71e-04 ***	0.186	1.09e-03 **	1.68e-03 **	2.24e-03 **			
Protein Intrinsic Disorder	**High**	1.000 ***	0.939 ***	0.636 **	0.670 **	0.303	0.515 *	<2e-216 ***	6.60e-06 ***	2.53e-03 **
	Low	0.970 ***	0.909 ***	0.454 *	0.630 **	0.576 **	0.273			
	Weighted Z	< 2e-216 ***	< 2e-216 ***	1.20e-03 **	3.85e-05 ***	5.80e-03 **	4.18e-02 *			
Mean Relative Solvent Accessibility	**High**	0.944 ***	0.889 ***	0.722 **	0.636 **	0.673 **	0.564 *	1.00e-07 ***	9.00e-07 ***	1.37e-05 ***
	Low	1.000 ***	0.778 **	0.667 *	0.636 **	0.491 *	0.564 *			
	Weighted Z	6.20e-06 ***	1.41e-05 ***	1.24e-03 **	3.67e-04 ***	7.76e-04 ***	1.55e-03 **			

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 ω, ω_{na} and ω_a in *Arabidopsis* and *Drosophila*.