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 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 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 mice, with young genes presenting a higher ω. Similar correlations between ω 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_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. 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

111 macromolecular level, some studies showed that highly disordered [39,40] and exposed residues [39]
112 present higher rates of adaptive evolution. Finally, there is evidence that proteins involved in the
113 immune and stress response have higher molecular adaptive rates [39,43,46,47]. Thus, controlling for
114 these confounding factors when assessing the impact of gene age on the rate of molecular adaptation is
115 crucial.
116 Here, we used a population genomic approach to test the adaptive walk model. We make two
117 predictions: first, that younger genes are undergoing faster rates of adaptive evolution, and second, the
118 evolutionary steps they make are larger. We tested the first prediction by estimating rates of adaptive
119 and non-adaptive protein evolution using an extension of the MacDonald-Kreitman (MK) test [48],
120 which uses counts of polymorphism and substitution at selected and neutral sites. We quantified the
121 rates of adaptive and non-adaptive evolution using the statistics ω_a and ω_{na}, which denote the rates of
122 adaptive and non-adaptive non-synonymous substitution relative to the mutation rate. We investigated
123 whether protein length, gene expression, relative solvent accessibility (RSA), intrinsic protein
124 disorder, BLAST's false-negative rate, and protein function act as confounding factors of the effect of
125 gene age. To test the second prediction, we considered the rates of substitution between amino acids
126 separated by different physicochemical distances as a function of gene age. We tested our hypotheses
127 in two pairs of species with different life-history traits: the Diptera *Drosophila melanogaster* and *D.*
128 *simulans* and the Brassica *Arabidopsis thaliana* and *A. lyrata*. In each species pair, we compared their
129 most recent genes with those dating back to the origin of cellular organisms.
130
131 **Results**
132 Young genes have a higher rate of adaptive substitutions
133 We tested the adaptive walk model of sequence evolution by assessing the impact of gene age on the
134 rate of adaptive (ω_a) and non-adaptive (ω_{na}) non-synonymous substitutions. We used Grapes [48] to
135 estimate ω_a and ω_{na}, which accounts for segregating slightly deleterious and advantageous mutations
136 by specifically modelling the distribution of fitness effects (DFE). Gene age data were obtained from
137 published data sets [28,33]. We found that gene age is significantly correlated to estimates of ω, ω_a
138 and ω_{na} in both species' pairs (Table 1 and Figure 1b). This result suggests that the higher ω ratio of
139 more recently evolved genes is due to a higher rate of adaptive and non-adaptive non-synonymous
140 substitutions. As X-linked genes are known to evolve faster due to the male hemizygosity [49–51], we
141 assessed whether the relationship between evolutionary rates and gene age differed between
142 chromosomes in *Drosophila* (Figure 1b). We compared models with and without the chromosome's
143 effect (see Material and Methods and supplementary file S1) and found only low support for a
144 chromosomal effect (p = 0.041 for ω_{na} and p = 0.094 for ω_a). We, therefore, combined data from all
145 chromosomes for subsequent analyses.
146
147 The effect of gene age on the rate of molecular adaptation is robust to multiple confounding factors

148 Genes of different ages intrinsically differ in their features [16,22,36]. As such traits significantly
149 impact the rate of molecular evolution [39], they may be confounding the faster adaptive rates
150 observed in young genes. Previous studies reported that younger genes encode shorter proteins
151 [24,35,52] and are expressed at lower levels [17,22,24,26], a pattern that we also observed in our data
152 set (gene age vs. protein length for *D. melanogaster* and *A. thaliana,* respectively: Kendall's τ = -
153 0.485, p = 2.82e-02; τ = -0.848, p = 1.06e-05, Figure S1a; gene age vs. gene expression for *D.*
154 *melanogaster* and *A. thaliana,* respectively: τ = -0.595, p = 7.35e-03; τ = -0.790, p = 4.00e-05, Figure
155 S1b in supplementary data). As younger proteins are shorter than older ones, they have a higher
156 proportion of exposed residues [39]: gene age is significantly positively correlated with the average
157 RSA per gene (τ = 0.636, p = 3.98e-03; τ = 0.695, p = 3.03e-04, for *D. melanogaster* and *A. thaliana*
158 respectively; Figure S2a in supplementary data). Because exposed residues are more flexible [53],
159 young genes tend to encode proteins with a higher degree of intrinsic disorder, a pattern previously
160 reported in mice [36]. We confirmed this pattern in *D. melanogaster* (τ = 0.606, p = 6.10e-03; Figure
161 S2b in supplementary data) and *A. thaliana* (τ = 0.467, p = 1.53e-02; Figure S2b in supplementary
162 data). Given that ω_a and ω_{na} are negatively correlated to length and gene expression, and positively
163 correlated to RSA and protein disorder [39], these could be driving the effect of gene age on ω_a and
164 ω_{na}; *i.e.*, the correlation between ω_a and ω_{na} and age could be incidental.
165 To control for the impact of each confounding factor using Grapes, we split our data into two
166 roughly equal-sized groups according to protein length, expression level, average RSA, and average
167 intrinsic disorder and reran the analysis within the "high" and "low" groups, combining probabilities
168 from the two analyses using the weighted Z-method [54]. As estimates of the rate of adaptive
169 substitutions for a small number of genes exhibit large sampling variances [55,56], each putative
170 confounding factor could only be tested separately. Some phylostrata were further combined when
171 under-represented in some gene categories (see Material and Methods). We found that ω, ω_{na} and ω_a
172 remain significantly correlated to gene age, except when controlling for protein length and gene
173 expression for ω_a in *Arabidopsis* (Figure 2 and Table 1). This weaker effect may be a consequence of
174 how the most recent clades were combined in these analyses, as there was little data available for those
175 genes (see Material and Methods). Nonetheless, when combining probabilities across the two species,
176 we observed a significant correlation between all measures of evolutionary rate and gene age
177 controlling each of the co-factors (Table 1).
178 We next asked whether the observed effect could be explained by the residual correlation of
179 the cofactors with gene age within each "high" and "low" group (Table S1 in supplementary data). We
180 performed three different tests: (1) the correlation between ω_a and the mean value of each co-factor in
181 each age class within the "high" and "low" groups; (2) the correlation between the co-factor and gene
182 age in each "high" and "low" group; and (3) a linear model where ω_a is the response variable, and
183 gene age, category, species (*Arabidopsis* or *Drosophila*), and the within category cofactor values are
184 explanatory variables. While we do observe a general correlation of the co-factor with gene age within

185 each group, we found that nearly all correlations between ω_a and the co-factor are non-significant
186 (Table S1). Moreover, the linear model analyses showed that in all cases but gene length, gene age is
187 the most significant variable, with an effect consistent with that observed in Table 1. For gene length,
188 gene age was only significant as an interaction with the species variable, having a positive effect only
189 in *Drosophila* (Table S2). These findings, therefore, suggest that the effect of gene age on ω_a is
190 independent of the co-factor.
191 To jointly estimate the effect of the potential confounding factors, we applied a recently
192 developed method that extends the MK test with a generalized linear model [57]. This approach
193 disentangles the effects of each factor on the rate of adaptive substitutions per nucleotide site.
194 However, this method does not model the distribution of fitness effects and hence cannot account for
195 segregating slightly deleterious mutations, which can bias estimates of the rate of adaptive
196 substitutions [58]. Hence, following the approach suggested in Huang [57], we removed sites for
197 which the derived allele frequency was below 50% to minimize any potential bias. Despite the large
198 reduction in the data set, this analysis revealed a significant effect of gene age (Table S3 in
199 supplementary data). Our findings, therefore, suggest that the effect of gene age on rates of protein
200 evolution is robust to the tested confounding factors and that a gene's age acts as a significant
201 determinant of the rate of adaptive and non-adaptive evolution in both species.
202 Lastly, we aimed at assessing the effect size of gene age on ω_a relative to other factors.
203 Because correlation coefficients were computed from values averaged over multiple genes and genes
204 were categorized differently for each analysis, the comparison of correlation coefficients does not
205 provide a reliable estimate of relative effect sizes. Therefore, we assessed the relative contribution of
206 each variable to the MK regression model by computing likelihood pseudo-R^2s [59]. These were
207 estimated by comparing the log-likelihood of the full model (all factors included) with the log-
208 likelihood of reduced models, where each single factor was removed independently from the full
209 model. These estimates suggest that, despite its very significant effect, gene age contributes
210 comparatively little to the explained variance compared to factors such as protein function (Table S4).
211 To note, however, that these estimates refer to the relative contribution of each variable per nucleotide
212 site and not per gene as this is the unit used by the MK regression, which results in generally low R^2
213 values.
214
215 <u>The effect of gene age on the molecular rate of adaptation is robust to BLAST's false negative rates</u>
216 The phylostratigraphy approach has been previously used to date the emergence of new genes, and
217 some studies have pointed out its potential limitations [28,60–63]. Because BLAST homology
218 searches might fail to identify homologs in short or rapidly evolving genes, such genes could be
219 mistakenly classified as young. As expected, we observed that genes in younger phylostrata present
220 higher E-values in both species ($\tau = 0.564$, $p = 0.025$; $\tau = 0.951$, $p = 1.20e-06$, for *D. melanogaster*
221 and *A. thaliana* respectively; Figure S3a in supplementary data). To control this effect, we reran our

222 analyses using genes with very low E-values that are likely to be detected in most age strata and for
223 which the correlation with gene age was no longer significant (see Material and Methods and
224 supplementary file S2) (τ = 0.408, p = 0.111; τ = 0.354, p = 0.141, for *D. melanogaster* and *A.*
225 *thaliana* respectively; Figure S3b in supplementary data). We observed that the effect of gene age
226 remained for all estimates in the two species when using this restricted data set (ω: τ = 0.929, p =
227 1.30e-03; ω_{na}: τ = 0.786, p = 6.49e-03; ω_a: τ = 0.643, p = 2.59e-02 in *A. thaliana*; and ω: τ = 0.697, p
228 = 1.61e-03; ω_{na}: τ = 0.636, p = 3.98e-03; ω_a: τ = 0.636, p = 3.98e-03 in *D. melanogaster*, Figure S4
229 in supplementary data). These results suggest that the correlation of gene age with the rate of adaptive
230 evolution cannot be attributed to errors in dating the emergence of a gene stemming from the failure of
231 identifying homologs in older taxa.
232
233 <u>The effect of gene age on the rate of molecular adaptation does not depend on protein function</u>
234 Lineage-specific genes are known to be involved in species-specific adaptive processes, such as the
235 evolution of morphological diversity [64] and immune and stress responses [17,34,64]. As proteins
236 encoding such functions tend to have higher molecular rates of adaptation [39,42,43,46,47,65], we
237 further assessed whether the observed effect of gene age could be due to younger genes being enriched
238 in functions with higher evolutionary rates. We first examined which functions are encoded by young
239 genes in *A. thaliana* and *D. melanogaster*. In *A. thaliana*, young genes (Clades 12 to 15 in Figure 1a)
240 are mostly involved in a large variety of cellular processes, stress response and external stimulus,
241 protein binding, and signal transduction (Figure S5a in supplementary data). In *D. melanogaster*,
242 young genes (Clades 11 and 12 in Figure 1a) encode mostly functions involved in the cell's anatomic
243 structure, stress response, nervous system processes, enzyme regulators, immune system mechanisms,
244 and a wide range of metabolic processes (Figure S5b in supplementary data).
245 To further correct for the potential bias of protein function, we ran Grapes across categories of
246 GO-annotated genes while simultaneously controlling for the effect of gene age. As some gene
247 functions were biased towards some age categories (Figure S5), we could not do this analysis for all
248 GO terms. We, therefore, only used the GO terms with a sufficient number of annotated genes in each
249 age class (see Material and Methods). To assess the effect of gene age, we grouped the genes in each
250 GO category into three age groups, trying to keep a similar number of genes in each (see Material and
251 Methods). In *A. thaliana*, we found that the impact of gene age on ω_a is stronger in proteins linked to
252 stress response and cellular components, where younger genes present higher molecular adaptive rates
253 (Figure S6a and supplementary file S3 in supplementary data). Although the GO term cellular
254 component represents a comprehensive annotation, it denotes the cellular compartments where
255 processes such as signal transduction and membrane trafficking occur, essential for maintaining the
256 cell homeostasis [66,67]. In *D. melanogaster*, we observed a strong effect of gene age on ω_a for
257 proteins encoding chromosomal organisation, protein complex, stress response, signal transduction,
258 and involved in the cell cycle (Figure S6b and supplementary file S3 in supplementary data in

259 supplementary data). Even though these functions cover a wide range of molecular processes, they are
260 involved in DNA replication, genome stability, and immune and stress responses, which are critical
261 functions for the co-evolutionary arms race between hosts and parasites [47]. When looking at the
262 non-adaptive rate of evolution, our analyses revealed a strong influence of gene age in ω_{na} within
263 most functions analysed in both species, where young genes present higher rates of non-adaptive
264 substitutions (Figure S6 and supplementary file S3 in supplementary data in supplementary data).
265 These results suggest that, when restricting the analysis to proteins involved in defence mechanisms,
266 which are known to adapt faster [42,43,47,65], gene age still has an impact on the efficiency of
267 selection acting upon a protein.
268
269 Substitutions in young genes have larger effect sizes
270 Our second prediction under the adaptive walk model is that substitutions in young genes should have
271 larger fitness effects than in older genes. To test this prediction, we used Grantham's physicochemical
272 distances between amino acids [68] as a proxy for the fitness effects of amino-acid mutations. We
273 looked at the fixed differences separated by one mutational step between each pair of species (*i.e.,*
274 fixed differences in the ingroup when compared to the outgroup species) and reported the average
275 Grantham's distances between residues within each age stratum. We observed that substitutions in
276 young genes tend to occur between less biochemically similar residues (*Arabidopsis*: $\tau = 1$, p = 2.00e-
277 07; *Drosophila*: $\tau = 0.788$, p = 3.628e-04; Figure 3 and supplementary file S4), suggesting that
278 substitutions in these genes have larger physicochemical effects than in old ones. To further test
279 whether these larger effects had a positive or negative impact on younger genes, we estimated the
280 average Grantham's distance amongst adaptive ($\bar{G}_a$) and non-adaptive ($\bar{G}_{na}$) non-synonymous
281 substitutions (see Material and Methods and supplementary file S4). We expect a positive correlation
282 between $\bar{G}_a$ and $\bar{G}_{na}$ and gene age, *i.e.,* the average Grantham's distance for adaptive and non-adaptive
283 substitutions should increase towards the present. This is indeed what we find: we observed a positive
284 correlation between $\bar{G}_a$ and gene age in both species, which is individually significant in *Drosophila*
285 (p = 0.006) and significant across species when we combine probabilities (p = 0.009) (Figure S7). We
286 observed the same pattern in the analysis of $\bar{G}_{na}$ with a significant correlation in *Arabidopsis* (p =
287 0.004) and a significant correlation when combining the probabilities from both species (p = 0.0007)
288 (Figure S7).
289
290 **Discussion**
291 We used a population genomic approach to disentangle the effects of positive and negative selection
292 on the rate of non-synonymous substitutions. Using complete genome data from two *Arabidopsis* and
293 *Drosophila* species, we showed that the higher rate of non-synonymous substitutions in younger genes
294 results both from relaxed purifying selection (higher ω_{na}) and a higher rate of adaptive substitutions
295 (higher ω_a) (Figure 1b). Although gene age is not the variable contributing most to the rate of

296 molecular adaptation (Table S4), its effect is consistent across comparisons and robust to multiple
297 confounding factors (Figure 2, Tables 1, S2 and S3). Moreover, we observed that young genes
298 undergo substitutions that are larger in terms of physicochemical properties than older genes. A
299 question remains: what are the drivers of these effects?
300
301 The magnitude effect of gene age on adaptive evolution is species-specific
302 Although we observed a strong impact of gene age on the molecular adaptive rate in both species
303 pairs, the shape of the relation between these two variables differs. While the relationship between
304 gene age and ω_a is monotonously increasing in *Arabidopsis*, it has several peaks in *Drosophila*
305 (Figure 1b). This pattern is particularly evident if we discard the two youngest clades. In *Drosophila*,
306 the correlation becomes much weaker and non-significant for ω_a (ω: τ = 0.600, p = 0.016; ω_{na}: τ
307 =0.556, p = 0.025; ω_a: τ = 0.467, p = 0.060), whereas, in *Arabidopsis,* the effect of gene age persists
308 (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).
309 Intriguingly, this multimodal distribution of ω_a observed in *Drosophila* resembles the pattern of gene
310 emergence in this species [17]. The peak in the adaptive substitution rate observed for clades 6 and 7
311 (Figure 1b) coincided with the animal phyla's major radiation at the time of extensive periods of
312 glacial cycles [69]. When looking at the functions coded by these proteins, we found that they are
313 linked to a wide range of vital cellular and biological processes, such as defence mechanisms and cell
314 differentiation (Figure S8 in supplementary data). This pattern suggests that these genes might be
315 experiencing higher molecular adaptive rates due to their role in such vital processes. However, for
316 these genes to keep such high rates of adaptive substitutions until recent times (*i.e.*, in the branch
317 between *D. melanogaster* and *D. simulans*), epistatic interactions might be at play. Studies across taxa
318 have proposed that functional epistasis is an important factor in the evolution of genes involved in
319 defence mechanisms and adaptation to new environmental stresses [70–74]. We posit that such gene
320 interactions keep these proteins further from their optimum throughout time due to the rugged shape of
321 the fitness landscape, leading to the high molecular adaptive rates observed in the branch between *D.*
322 *melanogaster* and *D. simulans*. To further test this hypothesis, we used the degree of protein-protein
323 interactions (PPI) as a proxy for epistatic interactions and analysed its relationship with gene age. We
324 observed that genes in clade 7 have a slightly higher degree of PPI than other strata (Figure S9 in
325 supplementary data and supplementary file S5), suggesting that these genes might be experiencing
326 relatively more epistatic interactions. These findings are consistent with epistasis influencing the
327 evolution of these genes, potentially explaining their continued higher rates of molecular adaptation.
328 In contrast, the burst of the emergence of new genes in *Arabidopsis* coincided with the plant-specific
329 radiation right before the emergence of Brassicaceae [17,75]. This trend is consistent with our results
330 from *A. thaliana*, where the bursts of ω_a occur in younger clades (after clades 11 and 12 in Figure 1b).
331 These distinct patterns observed between species suggest that the role of a gene's age in molecular
332 adaptation is complex, as also evidenced by the lack of a significant correlation with ω_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 ω_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

370 tend to occur between more dissimilar residues (Figure 3), suggesting that the evolution of young
371 genes proceeds in larger steps compared to old ones. Moreover, we observed that adaptive
372 substitutions have greater physicochemical effects (*i.e.,* the average Grantham's distance of adaptive
373 substitutions ($\bar{G}_a$) is correlated to gene age; Figure S7 in supplementary data). However, the strength
374 of this signal differs between species. While we observed a strong positive correlation between $\bar{G}_a$ and
375 gene age in *Drosophila*, this relationship is weaker in *Arabidopsis*. The opposite pattern is observed
376 for the correlation between $\bar{G}_{na}$ and gene age, where this relationship is only significant in
377 *Arabidopsis* (Figure S7 in supplementary data). These patterns suggest that, in *Arabidopsis*, there is a
378 stronger influence of relaxed purifying selection when compared to positive selection. This weaker
379 adaptive signal may be a consequence of their mating system. The transition from out-crossing to self-
380 fertilization in *A. thaliana* exposed these plants to less efficient purifying selection due to their
381 reduced effective population sizes [89,90], thus accumulating more deleterious mutations. This higher
382 load of segregating deleterious mutations can interfere with nearby beneficial mutations, preventing
383 their fixation and slowing down adaptation. Indeed, previous studies have shown that segregating
384 recessive deleterious mutations can substantially decrease the fixation rate of advantageous mutations,
385 leading to a weaker adaptive signal in some parts of the genome [91,92]. In contrast, the large
386 effective population sizes of *Drosophila* species make them more efficient in removing deleterious
387 mutations and prone to stronger positive selection [93–95]. These patterns could explain the observed
388 differences in the strength of selection in young genes between species.
389 An alternative hypothesis to the adaptive-walk theory is a static model of adaptation. This
390 model is based on the level of constraint of a gene: as some genes are less constrained than others,
391 they accumulate both advantageous and deleterious mutations, some of which may lead to the loss of a
392 gene. Under this hypothesis, less constrained genes tend to be young because they are more likely to
393 be lost. Additionally, as such genes accumulate a larger number of deleterious mutations, they also
394 represent a bigger mutational target for adaptive mutations, which could also explain the overall
395 correlation between ω_a and ω_{na}. We can test this hypothesis by assessing whether the correlation
396 between ω_a and gene age is independent of ω_{na}. We controlled for the effect of ω_{na} by running
397 Grapes on a subset of genes for which the correlation between p_n/p_s and gene age was no longer
398 significant. We found that gene age was still significantly correlated with ω_a whereas $\omega_{na}$was not
399 (Figure S10), thus suggesting that the rate of adaptive evolution is independent of the non-adaptive
400 rate. While these results do not completely discard the possibility of a stationary view of adaptation,
401 they agree with our adaptive walk model of gene evolution.
402 Overall, our findings support an adaptive walk model hypothesis of gene evolution over time,
403 where protein-coding genes adapt following a pattern of diminishing returns in *Drosophila* and
404 *Arabidopsis*. The increasing availability of population genomic datasets will allow the detailed
405 characterization of the adaptive walk model of gene evolution in an increasingly diverse range of taxa.
406

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 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).

Gene age vs. protein length and gene expression

444 Data on protein length and gene expression were included in the datasets of Moutinho et al. [39]
445 (Tables S2 and S3 in supplementary data online). Gene expression data were downloaded from the
446 database Expression Atlas [100], taking one baseline experiment for each species [39]. Mean gene
447 expression levels were then obtained by averaging across samples and tissues for each gene. To
448 correct for these two factors, we divided our dataset into two equally sized groups based on the factor
449 we wished to control. Short proteins had a size up to 366 and 389 amino acids, and long proteins had a
450 size up to 4,674 and 5,098 amino acids in *A. thaliana* and *D. melanogaster*, respectively. We further
451 merged phylostrata containing a low number of genes. For *D. melanogaster,* we categorised gene age
452 into 6 main clades by combining clades 3-4, 5-6, 7-10, and 11-12, keeping the others unchanged. In *A.*
453 *thaliana*, we combined the 15 clades into 6 main groups by merging clades 5-8 and 9-15. For gene
454 expression, we used a total of 17,126 and 6,247 genes for *A. thaliana* and *D. melanogaster,*
455 respectively, being categorised as lowly and highly expressed. Genes were classified as lowly
456 expressed if the mean expression levels were up to 10.3 and 6.8, and highly expressed genes were the
457 ones with expression up to 6,632.8 and 4,237.0 in *A. thaliana* and *D. melanogaster*, respectively. For
458 *D. melanogaster*, we categorised gene age into 6 categories by combining clades 3-5, 6-9, and 10-12.
459 In *A. thaliana*, we combined the data in 6 clades, merging clades 4-7, 8-11 and 12-15.
460
461 Gene age vs. protein structure
462 Since most young genes lack a defined three-dimensional structure [36], they do not have information
463 on the residue's solvent accessibility. Hence, we used a deep learning approach, NetSurfP-2.0, that
464 predicts the RSA of each residue from the amino-acid sequence [101] by using as a training model the
465 HH-suite sequence alignment tool for protein similarity searches [102]. To assess whether this
466 approach provided reliable results, we compared the RSA estimates of NetSurfP-2.0 with those
467 obtained from the PDB structures in our dataset [39]. We found a good correlation between the two
468 approaches for both species (Kendall's $\tau = 0.571$, $p < 2e\text{-}216$; $\tau = 0.462$, $p < 2e\text{-}216$, for *D.*
469 *melanogaster* and *A. thaliana* respectively). Using NetSurfP-2.0, RSA estimates were successfully
470 obtained for a total of 4,238,686 (88% of the total codon sites) and 7,479,807 (99% of the total codon
471 sites) amino-acid residues for *D. melanogaster* and *A. thaliana,* respectively. To assess the impact of
472 RSA at the gene level, we analysed the total number of genes in both species by making two
473 categories of genes according to the average RSA value per gene. Genes with lower RSA had mean
474 values between 0.127-0.389 in *Drosophila* and 0.217-0.386 in *Arabidopsis*. Genes with a higher RSA
475 had mean values between 0.390-0.894 in *Drosophila* and 0.387-0.898 in *Arabidopsis*. The phylostrata
476 groups were defined by combining clades 7-8 in *D. melanogaster*, and 8-11, 12-15 in *A. thaliana*.
477 For the analysis of the residue intrinsic disorder, we used the estimates included in the datasets
478 of Moutinho et al. [39]. These data were obtained with the software DisEMBL [103], where intrinsic
479 disorder was estimated per site and classified according to the degree of "hot loops" *i.e.*, highly mobile
480 loops. This analysis was performed for 17,732 and 7,410 genes for *A. thaliana* and *D. melanogaster,*

481 respectively. Genes were combined into two categories according to the mean value of their residue's
482 intrinsic disorder. Genes with a low level of intrinsic disorder had values between 0.029-0.080 in
483 *Drosophila* and among 0.041-0.084 in *Arabidopsis*. Genes with a higher degree of intrinsic disorder
484 had values between 0.081-0.554 in *Drosophila* and among 0.085-0.551 in *Arabidopsis*. In *D.*
485 *melanogaster*, all the 12 phylostrata could be used. In *A. thaliana*, the 15 strata were combined in 12
486 categories by merging clades 9-10, 11-12 and 13-14.
487
488 Correcting for BLAST e-values
489 We analysed the robustness of the gene age's effect by correcting the variation in the Expect (E) value
490 estimates in BLAST's searches between our focus species and their respective outgroups. By reducing
491 the variation in E-values estimates, we could correct for potential failures in BLAST's homology
492 searches. To do so, we limited the analysis to genes with particularly low E-values that are more likely
493 to be identified in all age strata: 12,472 genes with an E value lower than 1e-150 for *A. thaliana* and
494 7,104 genes with an E value lower than 1e-100 for *D. melanogaster* (supplementary file S2). For *A.*
495 *thaliana*, analyses were carried out by combining clades 8-13, with no genes left in clades 14 and 15.
496 For *D. melanogaster*, analyses were performed with the 12 strata. By using these genes, we found no
497 correlation between E-value and age.
498
499 Gene age vs. protein function
500 Gene ontology terms were obtained from the "dmelanogaster_gene_ensembl" and the
501 "athaliana_eg_gene" tables in the Ensembl database (version 103), through the R package "biomaRt"
502 [104]. A total of 7,253 (~70% of the genes) and 15,604 (~80% of the genes) genes were mapped in *D.*
503 *melanogaster* and *A. thaliana*, respectively. To check whether the effect of gene age prevailed across
504 functional protein classes, we analysed the GO terms with the highest number of young genes
505 mapped: more than 50 genes present in Clades 11 and 12 in *D. melanogaster*; and more than 30 genes
506 present in Clades 12 to 15 in *A. thaliana*. This filtering step resulted in 6,637 genes across 23 GO
507 categories in *D. melanogaster* (Table S5 in Supplementary Data online), and 15,410 genes across 10
508 GO categories in *A. thaliana* (Table S6 in Supplementary Data online). To analyse the effect of gene
509 age, we compared three age classes. In *D. melanogaster*, the first age category spanned over
510 phylostrata 1-3, the second category covered clades 4-7, and the third one included clades 8-12. In *A.*
511 *thaliana*, the first category comprised genes belonging to clades 1-6, the second category spanned over
512 clades 7-11, and the third one included the phylostrata between clades 12-15 (Figure 1a).
513
514 Gene Age vs. Protein-protein interactions (PPI)
515 We obtained PPI data for *D. melanogaster* from the STRING database [105], which includes both
516 physical and functional interactions (https://string-db.org/). This database included 13,046 proteins

517 with annotated interactions, which were used to analyse the distribution of protein networks across
518 phylostrata.
519
520 The fitness effects of amino-acid substitutions
521 We used Grantham's physicochemical distances between amino-acids [68] as a proxy for the fitness
522 effects of amino-acid substitutions and performed two analyses. In the first one, we calculated the
523 average Grantham's distance between amino acid substitutions for each phylostrata. In the second
524 analysis, we estimated the average Grantham's distance for adaptive ($\bar{G}_a$) and non-adaptive ($\bar{G}_{na}$)
525 non-synonymous substitutions within each age stratum. We restricted our analysis to amino acid pairs
526 separated by a single mutational step. To estimate the rate of adaptive and non-adaptive non-
527 synonymous amino acid substitutions between pairs of amino acids we used the same approach as
528 Bergman and Eyre-Walker [106]. This method compares the non-synonymous SFS for a particular
529 amino acid pair, for example, alanine and aspartic acid, which are separated by a A <> C mutation,
530 with the synonymous SFS of 4-fold degenerate sites separated only by A <> C mutations
531 ($SFS_{4F(A<>C)}$). For amino acid pairs separated by more than one mutation type, we estimated the
532 weighted average of the synonymous SFS of 4-fold sites for the different types of mutations,
533 weighting by the frequency of the respective codons. For example, serine and cysteine are separated
534 by A <> T and C <> G mutations. The synonymous SFS for this amino acid pair was estimated as
535 $SFS_{4F(weighted)} = [(f_{AGT} + f_{TGT} + f_{AGC} + f_{TGC}) * SFS_{4F(A<>T)} + (f_{TCT} + f_{TGT} + f_{TCC} + f_{TGC}) *$
536 $SFS_{4F(C<>G)}] / (f_{AGT} + f_{TGT} + f_{AGC} + f_{TGC} + f_{TGT} + f_{TGC})$, where *f* corresponds to the frequency of
537 each codon (*e.g.,* f_{AGT} is the frequency of the codon AGT). It is not possible to estimate the rates of
538 adaptive and non-adaptive substitution for each pair of amino acids for each phylostratum because we
539 have insufficient data. So, we combined amino acid pairs into ten categories based on their
540 Grantham's distance, and we combined clades 5-6, 8-11 and 12-15 in *Arabidopsis*, and clades 3-4, 5-7
541 and 8-10 in *Drosophila*. We used Grapes to estimate ω_a and ω_{na} for the ten categories of Grantham's
542 physicochemical distances for each combined age stratum. We calculated the average Grantham's
543 distance for adaptive ($\bar{G}_a$) and non-adaptive ($\bar{G}_{na}$) non-synonymous substitutions as $\bar{G}_a =$
544 $\sum_i \omega_{ai}\bar{g}_i N_i / \sum_i N_i$ and $\bar{G}_{na} = \sum_i \omega_{nai}\bar{g}_i N_i / \sum_i N_i$ where the sum is across the ten categories of
545 Grantham's distances within each age class, ω_{ai} and ω_{nai} are the rates of adaptive and non-adaptive
546 substitution in the *i*th category for each age class, $\overline{g_i}$ is the average Grantham's distance amongst the
547 amino acid pairs within each Grantham's distance category, and N_i is the number of non-synonymous
548 sites in the *i*th category in each age class. Higher values of $\bar{G}_a$ (the same for $\bar{G}_{na}$ with non-adaptive
549 substitutions) mean that there is a higher proportion of adaptive substitutions across Grantham's
550 distance categories.
551
552 Assessing the correlation between ω_a and ω_{na}

553 To test whether the impact of gene age on ω_a was independent of ω_{na}, we ran Grapes for a subset of
554 genes for which the correlation between ω_{na} and gene age was no longer significant. We controlled
555 for the variation in ω_{na} by discarding genes according to their p_n/p_s values. To ensure a sufficient
556 number of genes, we kept the genes that were within 0.02 and 0.05 standard deviation from the modal
557 value of the p_n/p_s distribution in *Arabidopsis* (mode: 0.05) and *Drosophila* (mode: 0.01),
558 respectively. The subset of data included 8,652 and 2,570 genes for *Arabidopsis* and *Drosophila*,
559 respectively.
560
561 Statistical analyses
562 Assessing the effect of gene age within each protein functional class was performed by comparing rate
563 estimates between all pairs of age categories. 100 bootstrap replicates were generated and ω_a and ω_{na}
564 were estimated for each resampling, allowing us to compute the rate differences between categories. A
565 one-tailed P-value can be obtained using the formula P = (k + 1)/ (N + 1), where N=100 is the number
566 of bootstrap replicates and k is the number of times the computed difference was greater (resp. lower)
567 than 0. Here, we used a two-tailed version of this test, computing the P-value as P = [2 * min (k^-, k^+) +
568 1]/(N+1), where k^- is the number of times the difference was negative, and k^+ is the number of times
569 the difference was positive. P-values for all pairwise comparisons were corrected for multiple testing
570 using the FDR method [107] as implemented in R [108] (see detailed R script in supplementary file
571 S3). For the analysis with PPI and gene age, statistical significance was assessed using non-parametric
572 *post hoc* tests, as implemented in the "Kruskal" method of the R package "agricolae" using the FDR
573 method to correct for multiple testing [109] (see detailed R script in supplementary file S5). For the
574 rest of the analyses, statistical significance was assessed with Kendall's correlation tests using the
575 mean value of the 100 bootstrap replicates for each category (see detailed script in supplementary file
576 S6). To estimate the combined P-value for each co-factor we used the weighted-Z method using the R
577 package "metap" [110]. We estimated the weight of each P-value using a linear modelling approach
578 with ω_a and ω_{na} as response variables, and gene age and potential co-factors as explanatory variables
579 and inferred the reciprocal of the squared standard error of the residuals in each model (see detailed R
580 scripts in supplementary file S7). To determine whether the chromosome impacted gene age's effect
581 on estimates of ω_a and ω_{na}, we performed an analysis of covariance (ANCOVA) by comparing a
582 model M1 that included the impact of the chromosome, age, and their interaction, with a model M0
583 that included age only (see detailed R script in supplementary file S1). Normality, homoscedasticity,
584 and independence of the error terms of the model were assessed with the package "lmtest" [111] in R.
585 Finally, to further assess the effect of residual correlations between gene age and each cofactor, we
586 fitted linear models with the rates of protein evolution (*i.e.*, ω, ω_{na} and ω_a) as response variables and
587 the average value of the cofactor for each age class as an explanatory variable (nested within the
588 cofactor category, low or high), together with gene age, species (*Drosophila* or *Arabidopsis*), cofactor
589 category, and their interaction. A model selection was conducted using the R function 'step' and the

590 significance of the model coefficients was assessed for the selected models (see detailed R script in
591 supplementary file S8).
592
593 MK regression
594 The MK regression analysis was performed per nucleotide site, where we analysed the combined
595 effect of RSA, intrinsic protein disorder, protein length, gene expression, protein function, Grantham's
596 distances, and the sex chromosome in *Drosophila*. As this method does not automatically handle
597 categorical variables, we created a binary variable for protein function based on the gene's presence
598 (1) or absence (0) in stress-related proteins (*i.e.*, response to stress, response to external stimulus and
599 signal transduction). We used the same rationale for the X chromosome effect in *Drosophila*. We
600 performed the MK regression analysis by comparing polymorphic and divergence sites between non-
601 synonymous and 4-fold degenerated sites in both species. As this method does not account for the
602 effects of weak selection [57], we analysed only sites with an allele frequency above 50%. To assess
603 the relative contribution of each variable to the regression model, we compared the likelihood pseudo-
604 R^2 following the Cox and Snell method [59]. These were estimated by comparing the log-likelihood of
605 the full model (all factors included) with that from each reduced model (each factor removed) using
606 the following equation:

607 $$\mathrm{R2} = 1 - (\exp(-\frac{2}{N} * (ln_{full} - ln_{reduced}))$$

608 where N represents the number of sites analysed with the MK regression, ln_{full} represents the log-
609 likelihood of the model with all factors included, and $ln_{reduced}$ represents the log-likelihood of the
610 model excluding each factor (*e.g.*, to assess the contribution of the gene age variable, the full model
611 was compared to a model including all variables except gene age).
612
613 **Acknowledgements**
614 The authors thank Diethard Tautz, Tal Dagan and Chaitanya Gokhale for fruitful discussions. J.Y.D.
615 acknowledges funding from the Max Planck Society. We thank two anonymous reviewers and Nicolas
616 Lartillot for their comments and suggestions to this manuscript.
617
618 **References**
619 1. Fisher R. The Genetical Theory of Natural Selection. Oxford Univ. Press, Oxford; 1930.
620 2. Wright S. Evolution in Mendelian Populations. Genetics. 1931;16(2):97–159.
621 3. Wright S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. Vol. 1,
622 Sixth International Congress on Genetics. 1932. p. 356–66.
623 4. Smith JM. Natural selection and the concept of a protein space. Nature. 1970;225(5232):563–4.
624 5. Gillespie JH. A Simple Stochastic Gene Substitution Model. Theoretical Population Biology.
625 1983;23:202–15.
626 6. Gillespie JH. Molecular evolution over the mutational landscape. Evolution (N Y).
627 1984;38(5):1116–29.
628 7. Gillespie JH. The Causes of Molecular Evolution. Oxford University Press; 1991.
629 8. Kimura M. The Neutral Theory of Molecular Evolution. Cambridge Univeristy Press; 1983.

9. Orr HA. The Population Genetics of Adaptation: The Distribution of Factors Fixed during Adaptive Evolution. Evolution. 1998;52(4):935.
10. Orr AH. The evolutionary genetics of adaptation: A simulation study. Genetical Research. 1999;74(3):207–14.
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.
12. Cooper VS, Lenski RE. The population genetics of ecological specialization in evolving Escherichia coli populations. Nature. 2000;407(6805):736–9.
13. Gerrish P. The rhythm of microbial adaptation. Nature. 2001;413(6853):299–302.
14. Rozen DE, De Visser JAGM, Gerrish PJ. Fitness effects of fixed beneficial mutations in microbial populations. Current Biology. 2002;12(12):1040–5.
15. Schoustra SE, Bataillon T, Gifford DR, Kassen R. The properties of adaptive walks in evolving populations of fungus. PLoS Biology. 2009;7(11).
16. Lynch M. Genomics: Gene duplication and evolution. Science, 2002;297(5583):945–947.
17. Tautz D, Domazet-Lošo T. The evolutionary origin of orphan genes. Nature Reviews Genetics. 2011;12(10):692–702.
18. Cohen O, Ashkenazy H, Belinky F, Huchon D, Pupko T. GLOOME: Gain loss mapping engine. Bioinformatics. 2010;26(22):2914–5.
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.
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.
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.
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.
23. Domazet-Loso T, Tautz D. An evolutionary analysis of orphan genes in Drosophila. Genome Research. 2003;13(10):2213–9.
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.
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.
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.
27. Albà MM, Castresana J. Inverse relationship between evolutionary rate and age of mammalian genes. Molecular Biology and Evolution. 2005;22(3):598–606.
28. Domazet-Lošo T, Carvunis AR, Albà 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.
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.
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.
31. García-Vallvé S, Alonso Á, Bravo IG. Papillomaviruses: Different genes have different histories. Trends in Microbiology. 2005;13(11):514–21.
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.
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, Heiland I, Gossmann TI. 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. Stukenbrock 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. 2006;7(8):645–53.
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.

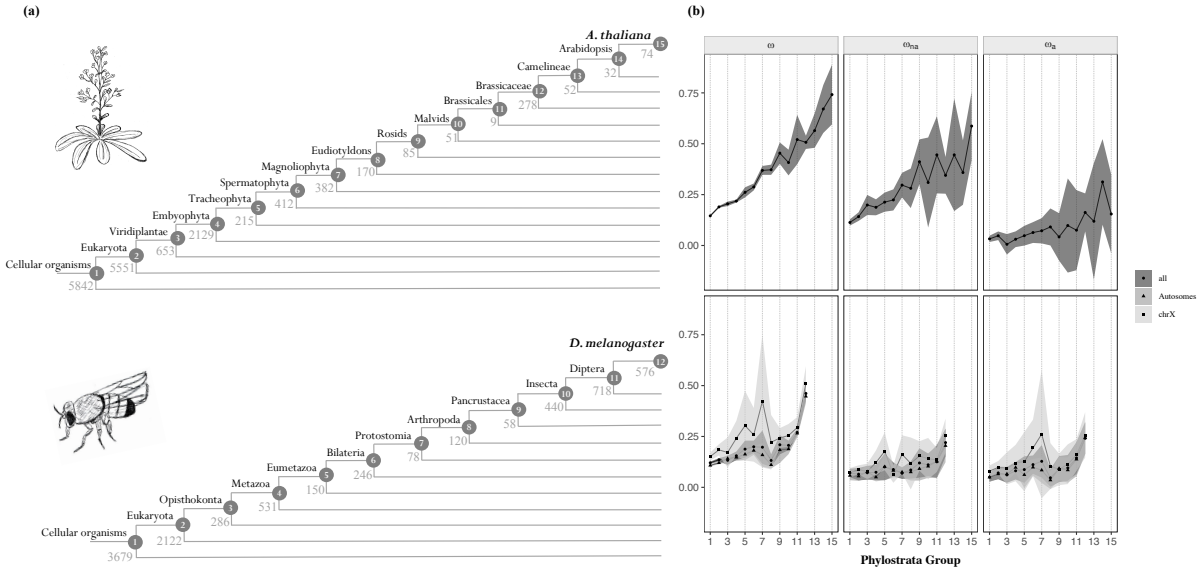
59. Cox DR, Snell EJ. Analysis of Binary Data (2nd ed.). London: Chapman and Hall; 1970.
60. Elhaik E, Sabath N, Graur D. The "inverse relationship between evolutionary rate and age of mammalian genes" is an artifact of increased genetic distance with rate of evolution and time of divergence. Molecular 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. Albà 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. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 1999;151(4):1531–45.
79. Prince VE, Pickett FB. Splitting pairs: The diverging fates of duplicated genes. Nature Reviews Genetics. 2002;3(11):827–37.
80. Jacob F. Evolution and Tinkering. Science (1979). 1977;196(4295):1161–6.
81. 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.
82. Heinen TJAJ, Staubach F, Häming D, Tautz D. Emergence of a New Gene from an Intergenic Region. Current Biology. 2009;19(18):1527–31.
83. 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.
84. Neme R, Tautz D. Evolution: Dynamics of de novo gene emergence. Current Biology. 2014;24(6):R238–40.
85. Palmieri N, Kosiol C, Schlötterer C. The life cycle of Drosophila orphan genes. Elife. 2014;3:1–21.

86. Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, et al. Proto-genes and de novo gene birth. Nature. 2012;487(7407):370–4.
87. Smith JM. Natural selection and the concept of a protein space. Nature. 1970; 225(5232), 563-564.
88. Orr HA. the Population Genetics of Adaptation: the Adaptation of Dna Sequences. Evolution. 2002;56(7):1317.
89. Bustamante CD, Nielsen R, Sawyer SA, Olsenk KM, Purugganank MD, Hartl DL. The Cost of Inbreeding in Arabidopsis. Nature. 2002; 416(6880): 531-534.
90. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993;134(4):1289–303.
91. 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.
92. 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
93. Gossmann TI, Keightley 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.
94. Eyre-Walker A. The genomic rate of adaptive evolution. Trends Ecol Evol. 2006;21:569–75.
95. 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.
96. 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).
97. Weigel D, Mott R. The 1001 Genomes Project for Arabidopsis thaliana. Genome Biology. 2009; 10(5):1-5.
98. 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.
99. Wickham H. ggplot2 - Elegant Graphics for Data Analysis (2nd Edition). Journal of Statistical Software. 2017;77:2–5.
100. 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. Nucleic Acids Research. 2016; 44(D1):D746-D752.
101. Klausen MS, Jespersen 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.
102. 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.
103. 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.
104. 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.
105. 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.
106. 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.
107. 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.
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(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.

858 **Figures**

859



860

861 **Figure 1. (a)** Phylogenetic definition of the strata used in the analyses for *A. thaliana* (top) and *D.*
862 *melanogaster* (bottom). The number of genes mapped to each clade is shown. **(b)** Relationship
863 between the rate of protein evolution (ω), non-adaptive non-synonymous substitutions (ω_{na}) and
864 adaptive non-synonymous substitutions (ω_a) with gene age in *A. thaliana* (top) and in *D.*
865 *melanogaster* (bottom). Clades are ordered according to (a). In *D. melanogaster*, the results for X-
866 linked, autosomal, and total genes are shown. Mean values of ω, ω_{na} and ω_a for each category are
867 represented with the black points. Error bars denote for the 95% confidence interval for each category,
868 computed over 100 bootstrap replicates.

869

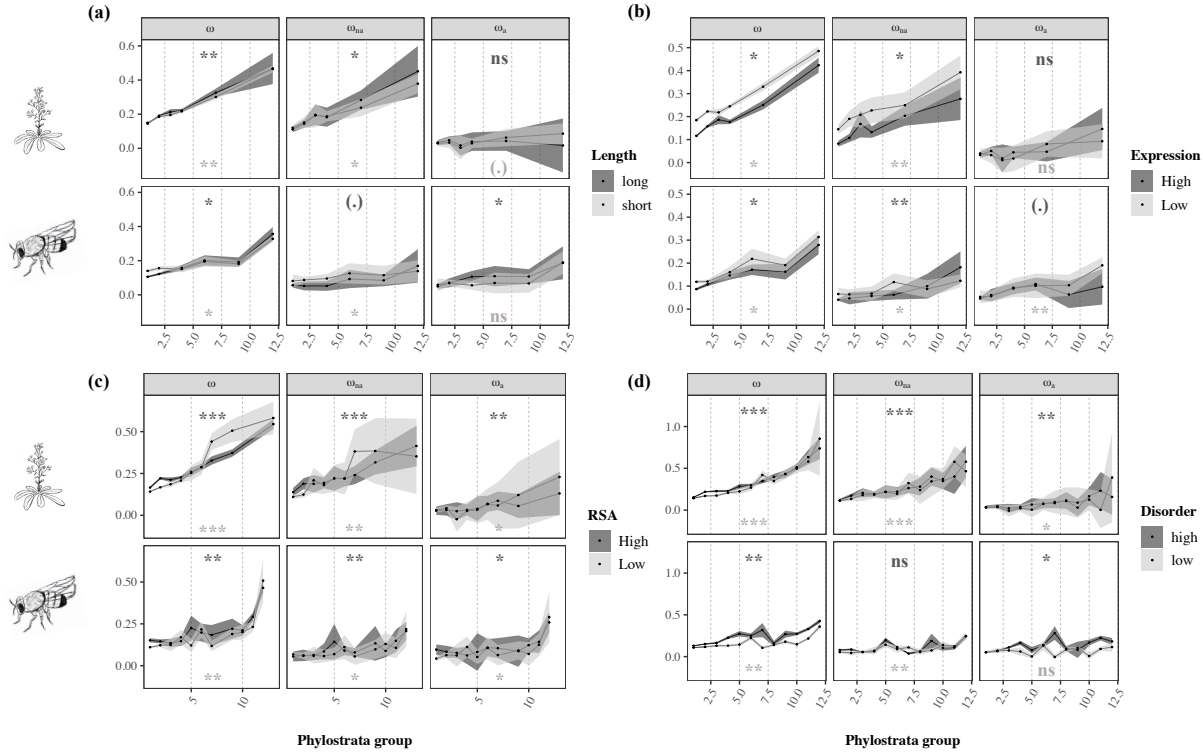


Figure 2. Estimates of ω, ω_{na} and ω_a plotted as a function of **(a)** protein length and **(b)** mean expression levels, **(c)** relative solvent accessibility, and **(d)** protein intrinsic disorder with gene age in *A. thaliana* (top) and *D. melanogaster* (bottom). Analyses were performed by comparing 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. 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" $P > 0.10$).

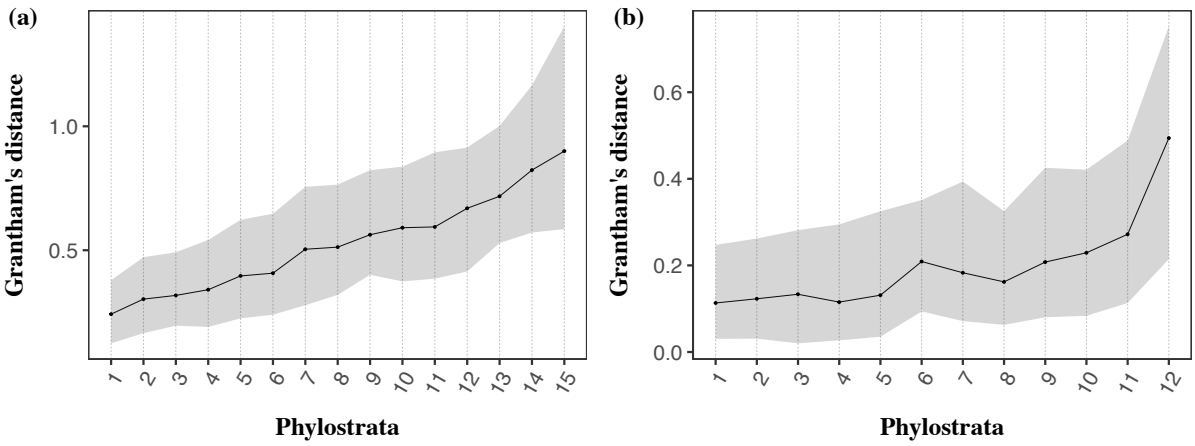


Figure 3. Relationship between gene age and Grantham's distance between amino acids for *A. thaliana* (a) and *D. melanogaster* (b). For each clade, the median value of Grantham's distance between residues is depicted with the black dot. The shaded area represents the physicochemical distances within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

885 **Tables**
886

887 **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
888 gene age with the respective co-factors: protein length, gene expression, protein intrinsic disorder and the mean relative solvent accessibility per gene. The
889 combined probabilities for each co-factor within and across species are presented in the fields "Weighted Z" and "Weighted Z across species", respectively, for
890 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 **			

891
892 **Note**. For each variable, the correlation coefficient and the combined probabilities are shown with the respective significance (*$P < 0.05$; **$P < 0.01$;
893 ***$P < 0.001$; "." $0.05 \leq P < 0.10$) for ω, ω_{na} and ω_a in *Arabidopsis* and *Drosophila*. As the effect of gene age was assessed by combining genes in each age
894 class, these correlation coefficients do not measure the intrinsic gene-level strength of correlation between gene age and molecular rates of evolution.