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: moutinho@evolbio.mpg.de, 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:)
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. Early adaptive theories proposed 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. This theory can also be used to understand molecular evolution in time, particularly across genes of different ages. 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 animals and plants. We found that a gene's evolutionary age significantly impacts the molecular adaptive rate. Moreover, we observed that substitutions in young genes tend to have larger fitness effects. Our study, therefore, provides the first evidence of an "adaptive walk" model of molecular evolution in large evolutionary timescales.

Significant statement

How does molecular adaptation occur? John Maynard Smith was one of the first to address this question by introducing the notion of "adaptive walk", which defines the "walk" of a gene towards higher fitness. At the start of this walk, genes tend to experience mutations with larger fitness effects than those closer to their fitness peak. Whilst being well-established, this theory has never been tested on large evolutionary timescales. Here, we achieve this by comparing molecular adaptive rates across genes of different ages in plants and animals. We showed that a gene's age acts as a significant determinant of molecular adaptation, where young genes adapt faster than old ones. We, therefore, provide evidence for an "adaptive walk" through time.

Introduction

How does adaptive evolution proceed in space and in time? This question has long intrigued evolutionary biologists. Fisher (1930) 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 concept of an "adaptive walk," where a protein "walks" in the space of all possible amino-acid sequences towards the ones with increasingly higher fitness values. Wright's and Maynard Smith's adaptation model was further developed by Gillespie (5–7), who presented the "move rule" in an adaptive landscape. 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 (8, 9), who extended Fisher's geometric model of adaptation and demonstrated that, apart from small effect mutations, adaptation relies on mutations of large fitness effects. He, therefore, characterised the adaptive walk with a pattern of diminishing returns. Under this model, a sequence further away from its local optimum will tends to accumulate large-effect mutations at the beginning of the walk. Small-effect mutations will then only be fixed when the sequence is approaching its optimum fitness. Experimental studies tracing the evolution of bacteria (10–13) and fungi (14) provided evidence for this view of adaptation as a walk with diminishing returns. Experimental studies, however, can only assess patterns of adaptation at relatively short time scales. The challenge lies in studying adaptation across time: how does the distribution of beneficial mutations vary across long evolutionary times?

While long-term evolutionary processes are not directly observable, footprints are left in genomes in the form of genes of different ages (15, 16). The age of a gene can be inferred from its phyletic pattern, that is, its presence or absence across the phylogeny (17). This reconstruction is obtained using sequence similarity searches performed by tools like BLAST (18). 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 (19).

Previous studies suggested that young or lineage-specific protein-coding genes evolve faster than older ones (16, 20–26). Albà and Castresana (26) 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 (21) 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 (24), *Drosophila* (22, 27, 28), bacteria (29), viruses (30), plants (31, 32), and protozoan parasites (33).

Despite the observed consistency across taxa, the drivers of such an effect remain debated (21). Besides, young and old genes differ in their structural properties, expression level, and protein function. Young genes tend to be smaller (21, 23, 34), have a higher degree of intrinsic disorder (35), and are expressed at lower levels (16, 21, 23, 25). Moreover, young genes tend to encode proteins involved in developing species-specific characteristics and immune and stress responses (15, 36, 37).

111 As the macromolecular structure (38, 39), gene expression levels (38, 40) and protein function (38, 41,
112 42) are known determinants of the rate of protein adaptation, they could be confounding the effect of
113 gene age. Several studies reported the substantial impact of gene expression on the adaptive rate of
114 proteins, where highly expressed proteins are significantly more constrained and have lower
115 adaptation rates (38, 40, 43, 44). At the macromolecular level, some studies showed that highly
116 disordered (38, 39) and exposed residues (38) present higher rates of adaptive evolution. Finally, there
117 is evidence that proteins involved in the immune and stress response have higher molecular adaptive
118 rates (38, 42, 45, 46). Thus, it is crucial to control for these confounding factors when assessing the
119 impact of gene age on the rate of molecular adaptation.
120 Here, we used a population genomic approach to test the adaptive walk model. We make two
121 predictions: first, that younger genes are undergoing faster rates of adaptive evolution, and second, the
122 evolutionary steps they make are larger. We tested the first prediction by estimating rates of adaptive
123 and non-adaptive protein evolution using an extension of the MacDonald-Kreitman test (47), which
124 uses counts of polymorphism and substitution at selected and neutral sites. We quantified the rates of
125 adaptive and non-adaptive evolution using the statistics ω_a and ω_{na}, which denote the rates of
126 adaptive and non-adaptive non-synonymous substitution relative to the mutation rate. We investigated
127 whether protein length, gene expression, relative solvent accessibility (RSA), intrinsic protein
128 disorder, BLAST's false-negative rate, and protein function act as confounding factors of the effect of
129 gene age. To test the second prediction, we considered the rates of substitution between amino acids
130 separated by different physicochemical distances as a function of gene age. We tested our hypotheses
131 in two pairs of species with different life-history traits: the diptera *Drosophila melanogaster* and *D.*
132 *simulans* and the Brassicas *Arabidopsis thaliana* and *A. lyrata*. In each species pair, we compared their
133 most recent genes with those dating back to the origin of cellular organisms.
134
135 **Results**
136 We tested the adaptive walk model of sequence evolution by assessing the impact of gene age on the
137 rate of adaptive (ω_a) and non-adaptive (ω_{na}) non-synonymous substitutions. To assess whether the
138 effect of gene age persisted when controlling for multiple confounding factors, we applied a non-
139 parametric measure of correlation between gene age and ω_a and ω_{na} for each category of the co-
140 factors analysed. The overall effect of gene age on ω_a and ω_{na} in each co-factor was assessed by
141 combing significance values across tests in both species using the weighted Z-method (48).
142
143 <u>Young genes have a higher rate of adaptive substitutions</u>
144 We tested the adaptive walk model of sequence evolution by assessing the impact of gene age on the
145 rate of adaptive (ω_a) and non-adaptive (ω_{na}) non-synonymous substitutions. We found that gene age
146 significantly impacts estimates of ω, ω_a and ω_{na} in both species' pairs (Table 1 and Figure 1b). This
147 result suggests that the higher ω ratio of more recently evolved genes is due to a higher rate of

148 adaptive and non-adaptive non-synonymous substitutions. As X-linked genes are known to evolve
149 faster (49, 50), we assessed whether the relationship between evolutionary rates and gene age differed
150 between chromosomes in *Drosophila* (Figure 1b). We compared models with and without the
151 chromosome's effect (see Material and Methods and supplementary file S1) and found low support for
152 a chromosomal effect (p = 0.041 for ω_{na} and p = 0.094 for ω_a). We, therefore, combined all
153 chromosomes for subsequent analyses.
154
155 <u>The effect of gene age on the rate of molecular adaptation is robust to multiple confounding factors</u>
156 Genes of different ages intrinsically differ in their features (15, 21, 35). As such traits significantly
157 impact the rate of molecular evolution (38), they may be confounding the faster adaptive rates
158 observed in young genes. Here, we assessed whether the effect of gene age on the rate of molecular
159 adaptation persisted after controlling for multiple confounding factors. To do so, we assessed the
160 correlation of gene age with the rates of molecular evolution in distinct categories of genes, according
161 to a putative confounding factor. As estimates of the rate of adaptive substitutions for a small number
162 of genes exhibit large sampling variances (51, 52), we could only assess each confounding factor
163 individually.
164 Previous studies reported that younger genes encode shorter proteins (23, 34, 53) and are
165 expressed at lower levels (16, 21, 23, 25), a pattern that we also observed in our data set (gene age vs.
166 protein length: Kendall's τ = -0.485, p = 2.82e-02; τ = -8.48, p = 1.06e-05, Figure S1a; gene age vs.
167 gene expression: τ = -0.595, p = 7.35e-03; τ = -0.790, p = 4.00e-05, Figure S1b in supplementary data;
168 for *D. melanogaster* and *A. thaliana,* respectively). As younger proteins are shorter than older ones,
169 they have a higher proportion of exposed residues (38): gene age is significantly positively correlated
170 with the average relative solvent accessibility per gene (τ = 0.636, p = 3.98e-03; τ = 0.695, p = 3.03e-
171 04, for *D. melanogaster* and *A. thaliana* respectively; Figure S2a in supplementary data). Because
172 exposed residues are more flexible (54), young genes tend to encode proteins with a higher degree of
173 intrinsic disorder, a pattern previously reported in mice (35). We confirmed this pattern in *D.*
174 *melanogaster* (τ = 0.606, p = 6.10e-03; Figure S2b in supplementary data) and *A. thaliana* (τ = 0.467,
175 p = 1.53e-02; Figure S2b in supplementary data).
176 We split our data into two roughly equal sized groups according to each of these potentially
177 confounding factors and reran the analysis within the "high" and "low" groups, combining
178 probabilities from the two analyses using the weighted Z-method (48). Some phylostrata were further
179 combined when under-represented in some gene categories (see Material and Methods). We found that
180 ω, ω_{na} and ω_a remain significantly correlated to gene age, except when controlling for protein length
181 and gene expression for ω_a in *Arabidopsis* (Figure 2 and Table 1). This weaker effect may be a
182 consequence of how the most recent clades were combined in these analyses, as there was little data
183 available for those genes (see Material and Methods). Nonetheless, when combining probabilities
184 across the two species, we observed a significant correlation between all measures of evolutionary rate

185 and gene age controlling each of the co-factors (Table 1). Our findings, therefore, suggest that the
186 effect of gene age on rates of protein evolution is robust to the tested confounding factors and that a
187 gene's age acts as a significant determinant of the rate of adaptive and non-adaptive evolution in both
188 species.
189
190 The effect of gene age on the molecular rate of adaptation is robust to BLAST's false negative rates
191 The physlostratigraphic approach has been previously used to date the emergence of new genes, and
192 some studies have pointed out its potential limitations (27, 55–58). Because BLAST homology
193 searches might fail to identify homologs in short or rapidly evolving genes, such genes could be
194 mistakenly classified as young. To assess whether BLAST's false negative rate could explain the
195 correlation between gene age and the rate of adaptive evolution, we analysed the gene age's effect by
196 correcting the variation in E-values estimates from BLAST's searches between each gene and their
197 respective outgroups. As expected, we observed that genes in younger phylostrata present higher E-
198 values in both species (τ = 0.564, p = 0.025; τ = 0.951, p = 1.20e-06, for *D. melanogaster* and *A.*
199 *thaliana* respectively; Figure S3a in supplementary data). To control this effect, we reran our analyses
200 with a subset of genes for which the correlation between the E-value and gene age was no longer
201 significant (see Material and Methods and supplementary file S2) (τ = 0.408, p = 0.111; τ = 0.354, p =
202 0.141, for *D. melanogaster* and *A. thaliana* respectively; Figure S3b in supplementary data). We
203 observed that the effect of gene age prevailed for all estimates in the two species (ω: τ = 0.929, p =
204 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
205 = 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
206 in supplementary data). These results suggest that the correlation of gene age with the rate of adaptive
207 evolution cannot be attributed to errors in dating the emergence of a gene stemming from the failure of
208 identifying homologs in older taxa.
209
210 The effect of gene age on the rate of molecular adaptation does not depend on protein function
211 Lineage-specific genes are known to be involved in species-specific adaptive processes, such as the
212 evolution of morphological diversity (59) and immune and stress responses (16, 33, 59). As proteins
213 encoding such functions tend to have higher molecular rates of adaptation (38, 41, 42, 45, 46, 60), we
214 further assessed whether the observed effect of gene age could be due to younger genes being enriched
215 in functions with higher evolutionary rates. We first examined which functions are encoded by young
216 genes in *A. thaliana* and *D. melanogaster*. In *A. thaliana*, young genes (Clades 12 to 15 in Figure 1a)
217 are mostly involved in a large variety of cellular processes, stress response and external stimulus,
218 protein binding, and signal transduction (Figure S5a in supplementary data). In *D. melanogaster*,
219 young genes (Clades 11 and 12 in Figure 1a) encode mostly functions involved in the cell's anatomic
220 structure, stress response, nervous system processes, enzyme regulators, immune system mechanisms,
221 and a wide range of metabolic processes (Figure S5b in supplementary data). However, it is important

222 to note that these functions represent general terms and not direct gene products due to the difficulty of
223 annotating young genes.
224 To further correct for the potential bias of protein function, we assessed the effect of gene age
225 separately for several GO-annotated genes, when a sufficient number of annotated genes in each age
226 class was available (see Material and Methods). In *A. thaliana*, we found that the impact of gene age
227 on ω_a is stronger in proteins linked to stress response and cellular components, where younger genes
228 present higher molecular adaptive rates (Figure S6a and supplementary file S3 in supplementary data).
229 Although the GO term cellular component represents a comprehensive annotation, it denotes the
230 cellular compartments where processes such as signal transduction and membrane trafficking occur,
231 essential for maintaining the cell homeostasis (61, 62). In *D. melanogaster*, we observed a strong
232 effect of gene age on ω_a for proteins encoding multiple cellular compartments, chromosomal
233 organisation, protein complex, stress response, signal transduction, and involved in the cell cycle
234 (Figure S6b and supplementary file S3 in supplementary data in supplementary data). Even though
235 these functions cover a wide range of molecular processes, they are involved in DNA replication,
236 genome stability, and immune and stress responses, which are critical functions for the co-
237 evolutionary arms race between hosts and parasites (46). When looking at ω_{na}, our analyses revealed
238 a strong influence of gene age in most functions analysed in both species, where young genes present
239 higher rates of non-adaptive substitutions (Figure S6 and supplementary file S3 in supplementary data
240 in supplementary data).
241 These results suggest that, when restricting the analysis to proteins involved in defence
242 mechanisms, which are known to adapt faster (41, 42, 46, 60), gene age still has an impact on the
243 efficiency of selection acting upon a protein.
244
245 <u>Substitutions in young genes have larger effect sizes</u>
246 Our second hypothesis predicts that substitutions in young genes have larger fitness effects than in
247 older genes. To test this prediction, we used Grantham's physicochemical distances between amino-
248 acids (63) as a proxy for the fitness effects of amino-acid substitutions. We looked at the fixed
249 differences separated by one mutational step between each pair of species and reported the average
250 Grantham's distances between residues within each age stratum. We observed that substitutions in
251 young genes tend to occur between less biochemically similar residues (*Arabidopsis*: $\tau = 1$, p = 2.00e-
252 07; *Drosophila*: $\tau = 0.788$, p = 3.628e-04; Figure 3 and supplementary file S4), suggesting that
253 substitutions in these genes have larger fitness effects than in old ones.
254
255 **Discussion**
256 Our population genomic approach successfully disentangled the effects of positive and negative
257 selection on the rate of non-synonymous substitutions. Using complete genome data from two
258 *Arabidopsis* and *Drosophila* species, we showed that the higher rate of non-synonymous substitutions

259 in younger genes results both from a relaxed purifying selection (higher ω_{na}) and a higher rate of
260 adaptive substitutions (higher ω_a) (Figure 1b). By looking at the magnitude effect of gene age, we
261 observed that young genes present a 25-fold higher rate of adaptation than older genes in *Drosophila*
262 species and around 30-fold higher in *Arabidopsis*. The magnitude of this effect is higher than that
263 observed for recombination rate and solvent exposure in these species, two other factors strongly
264 correlated to the rate of adaptive evolution (38, 64). We also observe that young genes undergo
265 substitutions that are larger in terms of physicochemical properties than older genes. A question
266 remains: what are the drivers of these effects?
267
268 <u>The magnitude effect of gene age on adaptive evolution is species-specific</u>
269 Although we observed a strong impact of gene age on the molecular adaptive rate in both species
270 pairs, the shape of their correlations differs. While the relationship between gene age and ω_a is
271 monotonously increasing in *Arabidopsis*, it has several peaks in *Drosophila* (Figure 1b). This pattern
272 is particularly evident if we discard the two youngest clades. In *Drosophila*, the correlation becomes
273 much weaker and non-significant for ω_a (ω: τ = 0.600, p = 0.016; ω_{na}: τ =0.556, p = 0.025; ω_a: τ =
274 0.467, p = 0.060), whereas, in *Arabidopsis,* the effect of gene age persists (ω: τ = 0.9487, p = 6.342e-
275 06; ω_{na}: τ = 0.872, p = 3.345e-05; ω_a: τ = 0.692, p = 9.86e-04). Intriguingly, this multimodal
276 distribution of ω_a observed in *Drosophila* resembles the pattern of emergence of young genes in this
277 species (16). The peak in the adaptive substitution rate observed for clades 6 and 7 (Figure 1b)
278 coincided with the animal phyla's major radiation at the time of extensive periods of glacial cycles
279 (65). When looking at the functions coded by these proteins, we found that they are linked to a wide
280 range of vital cellular and biological processes, such as defence mechanisms and cell differentiation
281 (Figure S7 in supplementary data). This pattern suggests that these genes might be experiencing higher
282 molecular adaptive rates due to their role in such vital processes. However, for these genes to keep
283 such high rates of adaptive substitutions until recent times, epistatic interactions might be at play.
284 Studies across taxa have proposed that functional epistasis is an important factor in the evolution of
285 genes involved in defence mechanisms and adaptation to new environmental stresses (66–70). We
286 posit that such gene interactions keep these proteins further from their optimum throughout time due
287 to the rugged shape of the fitness landscape, leading to the high molecular adaptive rates observed in
288 the branch between *D. melanogaster* and *D. simulans*. To further test this hypothesis, we used the
289 degree of protein-protein interactions (PPI) as a proxy for epistatic interactions and analysed its
290 relationship with gene age. We observed that genes in clade 7 have a slightly higher degree of PPI
291 than other strata (Figure S8 in supplementary data and supplementary file S5), suggesting that these
292 genes might be experiencing relatively more epistatic interactions. These findings are consistent with
293 epistasis influencing the evolution of these genes, potentially explaining their continued higher rates of
294 molecular adaptation. In contrast, the burst of the emergence of new genes in *Arabidopsis* coincided
295 with the plant-specific radiation right before the emergence of Brassicaceae (16, 71). This trend is

consistent with our results from *A. thaliana*, where the bursts of ω_a occur in younger clades (after clades 11 and 12 in Figure 1b). These distinct patterns observed between species suggest that the role of a gene's age in molecular adaptation is complex, as also evidenced by the lack of a significant correlation with ω_a previously reported in humans (21). The authors proposed that this result may be a consequence of the generally low molecular adaptive rates observed in primates (21, 47).

Despite these species-specific trends, our analyses revealed a strong correlation between ω_a and gene age extending through hundreds of millions of years (Figure 2). These findings suggest a consistent effect of a gene's age on the rate of molecular adaptation across taxa.

<u>An adaptive walk model of gene evolution</u>

Our study highlighted that, after their emergence, young genes evolve through relaxed selection, as first proposed by Ohno (1970), but also by acquiring beneficial mutations, as described in the "adaptive-conflict" model (36, 73). Ohno's idea of evolution was "non-Darwinian" in its nature, as he believed that "natural selection merely modified while redundancy created" (Ohno 1970). He proposed that new genes evolve 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 (74, 75).

In contrast, the "adaptive-conflict" model assumes that the ancestral gene could carry more than two pleiotropically constrained functions (36, 73). 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 (36, 73). These theories are based on the evolution of gene duplicates and agree with the idea of evolution as a "tinkerer" proposed by Jacob (1977), where evolution adjusts the already existing elements. In *de novo* evolution, however, new genes emerge by acquiring new functions from the non-coding fragments of the genome (16, 77, 78). 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 (79–82).

When looking at the fundamental ideas behind these theories, one can draw one prominent feature that portraits 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, 83, 84). 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 the Grantham's physicochemical distances (63) 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. Our study, therefore, provides evidence that the adaptive evolution of protein-coding genes follows a pattern of diminishing returns in plants and animals, indicating the potential generality of an adaptive walk model of gene evolution.

Material and Methods

We assessed the role of gene age on adaptive evolution using the divergence and polymorphism data published in Moutinho et al. (38). The data included 10,318 protein-coding genes in 114 *Drosophila melanogaster* individuals from an admixed sub-Saharan population from Phase 2 of the *Drosophila* Genomics Project (DPGP2) (85) and divergence estimates from *D. simulans* (Table S1 in supplementary data online); and 18,669 protein-coding genes in 110 *Arabidopsis thaliana* genomes comprising polymorphism data from a Spanish population (1001 Genomes Project) (86) and divergence out to *A. lyrata* (Table S2 in supplementary data online). These datasets also included data on protein length, gene expression, and residue intrinsic disorder. Gene age data were obtained from published data sets, wherein 9,004 *Drosophila* (27) and 17,732 *Arabidopsis* (32) 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 (47).

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 (47), using the Gamma-Exponential distribution of fitness effects (DFE), as this model was previously shown best to fit the data (38). Estimates 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" (87) 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

370 To correct for the effects of protein length and gene expression, we divided our dataset into two
371 equally sized groups based on the factor we wished to control. Short proteins had a size up to 366 and
372 389 amino-acids, and long proteins had a size up to 4,674 and 5,098 amino-acids in *A. thaliana* and *D.*
373 *melanogaster*, respectively. We further merged phylostrata containing a low number of genes. For *D.*
374 *melanogaster,* we categorised gene age into 6 main clades by combining clades 3-4, 5-6, 7-10, and 11-
375 12, keeping the others unchanged. In *A. thaliana*, we combined the 15 clades in 6 main groups by
376 merging clades 5-8 and 9-15. For gene expression, we used a total of 17,126 and 6,247 genes for *A.*
377 *thaliana* and *D. melanogaster,* respectively, being categorised as lowly and highly expressed. Genes
378 were classified as lowly expressed if the mean expression levels were up to 10.3 and 6.8, and highly
379 expressed genes were the ones with expression up to 6,632.8 and 4,237.0 in *A. thaliana* and *D.*
380 *melanogaster*, respectively. For *D. melanogaster*, we categorised gene age in 6 categories by
381 combining clades 3-5, 6-9, and 10-12. In *A. thaliana*, we combined the data in 6 clades, merging
382 clades 4-7, 8-11 and 12-15.
383
384 Gene age vs. protein structure
385 Since most young genes lack a defined three-dimensional structure (35), they do not have information
386 on the residue's solvent accessibility. Hence, we used a deep learning approach, NetSurfP-2.0, that
387 predicts the RSA of each residue from the amino-acid sequence (88) by using as a training model the
388 HH-suite sequence alignment tool for protein similarity searches (89). To assess whether this approach
389 provided reliable results, we compared the RSA estimates of NetSurfP-2.0 with those obtained from
390 the PDB structures in our dataset (38). We found a good correlation between the two approaches for
391 both species (Kendall's $\tau = 0.571$, $p < 2e\text{-}216$; $\tau = 0.462$, $p < 2e\text{-}216$, for *D. melanogaster* and *A.*
392 *thaliana* respectively). Using NetSurfP-2.0, RSA estimates were successfully obtained for a total of
393 4,238,686 (88% of the total codon sites) and 7,479,807 (99% of the total codon sites) amino-acid
394 residues for *D. melanogaster* and *A. thaliana,* respectively. To assess the impact of RSA at the gene
395 level, we analysed the total number of genes in both species by making two categories of genes
396 according to the average RSA value per gene. Genes with lower RSA had mean values between 0.127-
397 0.389 in *Drosophila* and 0.217-0.386 in *Arabidopsis*. Genes with a higher RSA had mean values
398 between 0.390-0.894 in *Drosophila* and 0.387-0.898 in *Arabidopsis*. The phylostrata groups were
399 defined by combining clades 7-8 in *D. melanogaster*, and 8-11, 12-15 in *A. thaliana*.
400 The intrinsic residue disorder analysis was performed for 7,126,304 and 3,645,645 sites for *A.*
401 *thaliana* and *D. melanogaster,* respectively. Genes were combined into two categories according to
402 the mean value of their residue's intrinsic disorder. Genes with a low level of intrinsic disorder had
403 values between 0.029-0.080 in *Drosophila* and among 0.041-0.084 in *Arabidopsis*. Genes with a
404 higher degree of intrinsic disorder had values between 0.081-0.554 in *Drosophila* and among 0.085-
405 0.551 in *Arabidopsis*. In *D. melanogaster*, all of the 12 phylostrata could be used. In *A. thaliana*, the
406 15 strata were combined in 12 categories by merging clades 9-10, 11-12 and 13-14.

407

408 Correcting for BLAST e-values

409 We analysed the robustness of the gene age's effect by correcting the variation in the Expect (E) value
410 estimates in BLAST's searches between our focus species and their respective outgroups. By reducing
411 the variation in E-values estimates, we could correct for potential failures in BLAST's homology
412 searches. To do so, we used a subset of genes for which the correlation between the E value and gene
413 age was no longer significant: 12,472 genes with an E value lower than 1e-150 for *A. thaliana* and
414 7,104 genes with an E value lower than 1e-100 for *D. melanogaster* (supplementary file S2). For *A.*
415 *thaliana*, analyses were carried by combining clades 8-13, with no genes left in clades 14 and 15. For
416 *D. melanogaster*, analyses were performed with the 12 strata.

417

418 Gene age vs. protein function

419 Gene ontology terms were obtained from the "dmelanogaster_gene_ensembl" and the
420 "athaliana_eg_gene" tables in the Ensembl database (version 103), through the R package "biomaRt"
421 (90). A total of 7,253 (~70% of the genes) and 15,604 (~80% of the genes) genes were mapped in *D.*
422 *melanogaster* and *A. thaliana*, respectively. To check whether the effect of gene age prevailed across
423 functional protein classes, we analysed the GO terms with the highest number of young genes
424 mapped: more than 50 genes present in Clades 11 and 12 in *D. melanogaster*; and more than 30 genes
425 present in Clades 12 to 15 in *A. thaliana*. This filtering step resulted in 6,637 genes across 23 GO
426 categories in *D. melanogaster* (Table S1 in Supplementary Data online), and 15,410 genes across 10
427 GO categories in *A. thaliana* (Table S2 in Supplementary Data online). To analyse the effect of gene
428 age, we compared three age classes. In *D. melanogaster*, the first age category spanned over
429 phylostrata 1-3, the second category covered clades 4-7, and the third one included clades 8-12. In *A.*
430 *thaliana*, the first category comprised genes belonging to clades 1-6, the second category spanned over
431 clades 7-11, and the third one included the phylostrata between clades 12-15 (Figure 1a).

432

433 Gene Age vs. Protein-protein interactions (PPI)

434 We obtained PPI data for *D. melanogaster* from the STRING database (91), which includes both
435 physical and functional interactions (https://string-db.org/). This database included 13,046 proteins
436 with annotated interactions, which were used to analyse the distribution of protein networks across
437 phylostrata.

438

439 Statistical analyses

440 Assessing the effect of gene age within each protein functional class was performed by comparing rate
441 estimates between all pairs of age categories. 100 bootstrap replicates were generated and ω_a and ω_{na}
442 were estimated for each resampling, allowing to compute the rate differences between categories. A
443 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 (91) as implemented in R (92) (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 (92) (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" (93). To obtain the weight of each p-value, we used 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). Finally, 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" (94) in R (95).

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. R. Fisher, *The Genetical Theory of Natural Selection* (Oxford Univ. Press, Oxford, 1930).
2. S. Wright, Evolution in Mendelian Populations. *Genetics* **16**, 97–159 (1931).
3. S. Wright, The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Sixth Int. Congr. Genet.* **1**, 356–366 (1932).
4. J. M. Smith, Natural selection and the concept of a protein space. *Nature* **225**, 563–564 (1970).
5. J. H. Gillespie, A Simple Stochastic Gene Substitution Model. *Theor. Popul. Biol.* **23**, 202–215 (1983).
6. J. H. Gillespie, Molecular evolution over the mutational landscape. *Evolution.* **38**, 1116–1129 (1984).
7. J. H. Gillespie, *The Causes of Molecular Evolution* (Oxford University Press, 1991).
8. H. A. Orr, The Population Genetics of Adaptation: The Distribution of Factors Fixed during

481 Adaptive Evolution. *Evolution.* **52**, 935 (1998).
482 9. A. H. Orr, The evolutionary genetics of adaptation: A simulation study. *Genet. Res.* **74**, 207–
483 214 (1999).
484 10. R. E. Lenski, M. R. Rose, S. C. Simpson, S. C. Tadler, Long-Term Experimental Evolution in
485 Escherichia coli. I. Adaptation and Divergence During 2,000 Generations. *Am. Nat.* **138**, 1315–
486 1341 (1991).
487 11. V. S. Cooper, R. E. Lenski, The population genetics of ecological specialization in evolving
488 Escherichia coli populations. *Nature* **407**, 736–739 (2000).
489 12. P. Gerrish, The rhythm of microbial adaptation. *Nature* **413**, 299–302 (2001).
490 13. D. E. Rozen, J. A. G. M. De Visser, P. J. Gerrish, Fitness effects of fixed beneficial mutations
491 in microbial populations. *Curr. Biol.* **12**, 1040–1045 (2002).
492 14. S. E. Schoustra, T. Bataillon, D. R. Gifford, R. Kassen, The properties of adaptive walks in
493 evolving populations of fungus. *PLoS Biol.* **7** (2009).
494 15. M. Lynch, Genomics: Gene duplication and evolution. *Science.* **297**, 945–947 (2002).
495 16. D. Tautz, T. Domazet-Lošo, The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12**,
496 692–702 (2011).
497 17. O. Cohen, H. Ashkenazy, F. Belinky, D. Huchon, T. Pupko, GLOOME: Gain loss mapping
498 engine. *Bioinformatics* **26**, 2914–2915 (2010).
499 18. S. Altschul, *et al.*, Gapped blast and psi-blast: a new generation of protein database search
500 programs. *FASEB J.* **12**, 3389–3402 (1998).
501 19. T. Domazet-Lošo, J. Brajković, D. Tautz, A phylostratigraphy approach to uncover the
502 genomic history of major adaptations in metazoan lineages. *Trends Genet.* **23**, 533–539 (2007).
503 20. K. Thornton, M. Long, Rapid divergence of gene duplicates on the Drosophila melanogaster X
504 chromosome. *Mol. Biol. Evol.* **19**, 918–925 (2002).
505 21. J. J. Cai, D. A. Petrov, Relaxed purifying selection and possibly high rate of adaptation in
506 primate lineage-specific genes. *Genome Biol. Evol.* **2**, 393–409 (2010).
507 22. T. Domazet-Loso, D. Tautz, An evolutionary analysis of orphan genes in Drosophila. *Genome*
508 *Res.* **13**, 2213–2219 (2003).
509 23. A. Vishnoi, S. Kryazhimskiy, G. A. Bazykin, S. Hannenhalli, J. B. Plotkin, Young proteins
510 experience more variable selection pressures than old proteins. *Genome Res.* **20**, 1574–1581
511 (2010).
512 24. J. J. Cai, P. C. Y. Woo, S. K. P. Lau, D. K. Smith, K. Y. Yuen, Accelerated evolutionary rate
513 may be responsible for the emergence of lineage-specific genes in Ascomycota. *J. Mol. Evol.*
514 **63**, 1–11 (2006).
515 25. Y. I. Wolf, P. S. Novichkov, G. P. Karev, E. V. Koonin, D. J. Lipman, The universal
516 distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of
517 different apparent ages. *Proc. Natl. Acad. Sci.* **106**, 7273–7280 (2009).

518 26. M. M. Albà, J. Castresana, Inverse relationship between evolutionary rate and age of
519 mammalian genes. *Mol. Biol. Evol.* **22**, 598–606 (2005).
520 27. T. Domazet-Lošo, *et al.*, No evidence for phylostratigraphic bias impacting inferences on
521 patterns of gene emergence and evolution. *Mol. Biol. Evol.* **34**, 843–856 (2017).
522 28. Y. E. Zhang, M. D. Vibranovski, B. H. Krinsky, M. Long, Age-dependent chromosomal
523 distribution of male-biased genes in Drosophila. *Genome Res.* **20**, 1526–1533 (2010).
524 29. V. Daubin, H. Ochman, Bacterial genomes as new gene homes: The genealogy of ORFans in
525 E. coli. *Genome Res.* **14**, 1036–1042 (2004).
526 30. S. García-Vallvé, Á. Alonso, I. G. Bravo, Papillomaviruses: Different genes have different
527 histories. *Trends Microbiol.* **13**, 514–521 (2005).
528 31. X. Cui, *et al.*, Young genes out of the male: An insight from evolutionary age analysis of the
529 pollen transcriptome. *Mol. Plant* **8**, 935–945 (2015).
530 32. Z. W. Arendsee, L. Li, E. S. Wurtele, Coming of age: Orphan genes in plants. *Trends Plant Sci.*
531 **19**, 698–708 (2014).
532 33. C. H. Kuo, J. C. Kissinger, Consistent and contrasting properties of lineage-specific genes in
533 the apicomplexan parasites Plasmodium and Theileria. *BMC Evol. Biol.* **8**, 1–16 (2008).
534 34. R. Neme, D. Tautz, Phylogenetic patterns of emergence of new genes support a model of
535 frequent de novo evolution. *BMC Genomics* **14,** 117 (2013).
536 35. B. A. Wilson, S. G. Foy, R. Neme, J. Masel, Young genes are highly disordered as predicted by
537 the preadaptation hypothesis of de novo gene birth. *Nat. Ecol. Evol.* **1**, 0146 (2017).
538 36. A. L. Hughes, The evolution of functionally novel proteins after gene duplication. *Proc. R. Soc.*
539 *B* **256**, 119–124 (1994).
540 37. J. Zhang, Y. ping Zhang, H. F. Rosenberg, Adaptive evolution of a duplicated pancreatic
541 ribonuclease gene in a leaf-eating monkey. *Nat. Genet.* **30**, 411–415 (2002).
542 38. A. F. Moutinho, F. F. Trancoso, J. Y. Dutheil, The impact of protein architecture on adaptive
543 evolution. *Mol. Biol. Evol.*, 560185 (2019).
544 39. A. Afanasyeva, M. Bockwoldt, C. R. Cooney, I. Heiland, T. I. Gossmann, Human long
545 intrinsically disordered protein regions are frequent targets of positive selection. *Genome Res.*
546 **28**, 975–982 (2018).
547 40. S. Subramanian, S. Kumar, Gene expression intensity shapes evolutionary rates of the proteins
548 encoded by the vertebrate genome. *Genetics* **168**, 373–381 (2004).
549 41. E. H. Stukenbrock, *et al.*, The making of a new pathogen: Insights from comparative
550 population genomics of the domesticated wheat pathogen Mycosphaerella graminicola and its
551 wild sister species. *Genome Res.* **21**, 2157–2166 (2011).
552 42. D. Enard, L. Cai, C. Gwennap, D. A. Petrov, Viruses are a dominant driver of protein
553 adaptation in mammals. *Elife* **5**, 1–25 (2016).
554 43. E. P. C. Rocha, A. Danchin, An Analysis of Determinants of Amino Acids Substitution Rates

555 in Bacterial Proteins. *Mol. Biol. Evol.* **21**, 108–116 (2004).
556 44. C. Pal, B. Papp, L. D. Hurst, Highly Expressed Genes in Yeast Evolve Slowly. *Genetics* **158**,
557 927–931 (2001).
558 45. T. B. Sackton, *et al.*, Dynamic evolution of the innate immune system in Drosophila. *Nat.*
559 *Genet.* **39**, 1461–1468 (2007).
560 46. D. J. Obbard, J. J. Welch, K. W. Kim, F. M. Jiggins, Quantifying adaptive evolution in the
561 Drosophila immune system. *PLoS Genet.* **5**, e1000698 (2009).
562 47. N. Galtier, Adaptive Protein Evolution in Animals and the Effective Population Size
563 Hypothesis. *PLoS Genet.* **12**, 1–23 (2016).
564 48. M. C. Whitlock, M. C. Whitlock, Combining probability from independent tests: the weighted
565 Z-method is superior to Fisher's approach. *Wiley Online Libr.* **18**, 1368–1373 (2005).
566 49. B. Vicoso, B. Charlesworth, Evolution on the X chromosome: Unusual patterns and processes.
567 *Nat. Rev. Genet.* **7**, 645–653 (2006).
568 50. B. Vicoso, B. Charlesworth, Effective population size and the faster-X effect: An extended
569 model. *Evolution (N. Y).* **63**, 2413–2426 (2009).
570 51. N. Stoletzki, A. Eyre-Walker, Estimation of the neutrality index. *Mol. Biol. Evol.* **28**, 63–70
571 (2011).
572 52. N. G. C. Smith, a Eyre-Walker, Adaptive protein evolution in Drosophila . *Nature* **415**, 1022–
573 1024 (2002).
574 53. Y. Ding, Q. Zhou, W. Wang, Origins of New Genes and Evolution of Their Novel Functions.
575 *Annu. Rev. Ecol. Evol. Syst.* **43**, 345–363 (2012).
576 54. S. S. Choi, E. J. Vallender, B. T. Lahn, Systematically assessing the influence of 3-dimensional
577 structural context on the molecular evolution of mammalian proteomes. *Mol. Biol. Evol.* **23**,
578 2131–2133 (2006).
579 55. E. Elhaik, N. Sabath, D. Graur, The "inverse relationship between evolutionary rate and age of
580 mammalian genes" is an artifact of increased genetic distance with rate of evolution and time
581 of divergence. *Mol. Biol. Evol.* **23**, 1–3 (2006).
582 56. B. A. Moyers, J. Zhang, Phylostratigraphic bias creates spurious patterns of genome evolution.
583 *Mol. Biol. Evol.* **32**, 258–267 (2015).
584 57. B. A. Moyers, J. Zhang, Evaluating Phylostratigraphic Evidence for Widespread de Novo Gene
585 Birth in Genome Evolution. *Mol. Biol. Evol.* **33**, 1245–1256 (2016).
586 58. M. M. Albà, J. Castresana, On homology searches by protein Blast and the characterization of
587 the age of genes. *BMC Evol. Biol.* **7**, 1–8 (2007).
588 59. K. Khalturin, G. Hemmrich, S. Fraune, R. Augustin, T. C. G. Bosch, More than just orphans:
589 are taxonomically-restricted genes important in evolution? *Trends Genet.* **25**, 404–413 (2009).
590 60. T. Slotte, *et al.*, Genomic determinants of protein evolution and polymorphism in arabidopsis.
591 *Genome Biol. Evol.* **3**, 1210–1219 (2011).

592 61. N. Geldner, S. Robatzek, Plant receptors go endosomal: A moving view on signal transduction.
593 *Plant Physiol.* **147**, 1565–1574 (2008).
594 62. A. J. Groen, S. C. De Vries, K. S. Lilley, A proteomics approach to membrane trafficking.
595 *Plant Physiol.* **147**, 1584–1589 (2008).
596 63. R. Grantham, Amino Acid Difference Formula to Help Explain Protein Evolution. *Science.*
597 **185**, 862–4 (1974).
598 64. D. Castellano, M. Coronado-Zamora, J. L. Campos, A. Barbadilla, A. Eyre-Walker, Adaptive
599 evolution is substantially impeded by hill-Robertson interference in drosophila. *Mol. Biol.*
600 *Evol.* **33**, 442–455 (2016).
601 65. P. F. Hoffman, A. J. Kaufman, G. P. Halverson, D. P. Schrag, A neoproterozoic snowball
602 earth. *Science (80-.).* **281**, 1342–1346 (1998).
603 66. K. L. Montooth, J. H. Marden, A. G. Clark, Mapping Determinants of Variation in Energy
604 Metabolism, Respiration and Flight in Drosophila. *Genetics* **165**, 623–635 (2003).
605 67. T. Paixão, N. H. Barton, The effect of gene interactions on the long-term response to selection.
606 *Proc. Natl. Acad. Sci. U. S. A.* **113**, 4422–4427 (2016).
607 68. T. F. Hansen, WHY EPISTASIS IS IMPORTANT FOR SELECTION AND ADAPTATION.
608 *Evolution (N. Y).* **67**, 3501–3511 (2013).
609 69. E. L. Behrman, *et al.*, Rapid seasonal evolution in innate immunity of wild Drosophila
610 melanogaster. *Proc. R. Soc. B Biol. Sci.* **285**, 20172599 (2018).
611 70. M. Lagator, N. Colegrave, P. Neve, Selection history and epistatic interactions impact
612 dynamics of adaptation to novel environmental stresses. *Proc. R. Soc. B Biol. Sci.* **281**,
613 20141679 (2014).
614 71. H. Wang, *et al.*, Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proc. Natl.*
615 *Acad. Sci. U. S. A.* **106**, 3853–3858 (2009).
616 72. S. Ohno, *Evolution by gene duplication* (Springer Science & Business Media, 1970).
617 73. J. Piatigorsky, G. Wistow, The recruitment of crystallins: new functions precede gene
618 duplication. *Science (80-.).* **252**, 1078–1079 (1991).
619 74. A. Force, *et al.*, Preservation of duplicate genes by complementary, degenerative mutations.
620 *Genetics* **151**, 1531–1545 (1999).
621 75. V. E. Prince, F. B. Pickett, Splitting pairs: The diverging fates of duplicated genes. *Nat. Rev.*
622 *Genet.* **3**, 827–837 (2002).
623 76. F. Jacob, Evolution and Tinkering. *Science (80-.).* **196**, 1161–1166 (1977).
624 77. J. Cai, R. Zhao, H. Jiang, W. Wang, De novo origination of a new protein-coding gene in
625 Saccharomyces cerevisiae. *Genetics* **179**, 487–496 (2008).
626 78. T. J. A. J. Heinen, F. Staubach, D. Häming, D. Tautz, Emergence of a New Gene from an
627 Intergenic Region. *Curr. Biol.* **19**, 1527–1531 (2009).
628 79. L. Zhao, P. Saelao, C. D. Jones, D. J. Begun, Origin and spread of de novo genes in Drosophila

629 melanogaster populations. *Science (80-.).* **343**, 769–772 (2014).
630 80. R. Neme, D. Tautz, Evolution: Dynamics of de novo gene emergence. *Curr. Biol.* **24**, R238–
631 R240 (2014).
632 81. N. Palmieri, C. Kosiol, C. Schlötterer, The life cycle of Drosophila orphan genes. *Elife* **3**, 1–21
633 (2014).
634 82. A. R. Carvunis, *et al.*, Proto-genes and de novo gene birth. *Nature* **487**, 370–374 (2012).
635 83. J. M. Smith, Natural selection and the concept of a protein space. *Nature* **225**, 563–564 (1970).
636 84. H. A. Orr, the Population Genetics of Adaptation: the Adaptation of Dna Sequences. *Evolution.*
637 **56**, 1317 (2002).
638 85. J. E. Pool, *et al.*, Population Genomics of Sub-Saharan Drosophila melanogaster: African
639 Diversity and Non-African Admixture. *PLoS Genet.* **8**, e1003080 (2012).
640 86. D. Weigel, R. Mott, The 1001 Genomes Project for Arabidopsis thaliana. *Genome Biol.* **10**, 107
641 (2009).
642 87. H. Wickham, ggplot2 - Elegant Graphics for Data Analysis (2nd Edition). *J. Stat. Softw.* **77**, 2–
643 5 (2017).
644 88. M. S. Klausen, *et al.*, NetSurfP-2.0: Improved prediction of protein structural features by
645 integrated deep learning. *Proteins Struct. Funct. Bioinforma.* **87**, 520–527 (2019).
646 89. M. Remmert, A. Biegert, A. Hauser, J. Söding, HHblits: Lightning-fast iterative protein
647 sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2012).
648 90. S. Durinck, *et al.*, BioMart and Bioconductor: A powerful link between biological databases
649 and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).
650 91. L. J. Jensen, *et al.*, STRING 8 - A global view on proteins and their functional interactions in
651 630 organisms. *Nucleic Acids Res.* **37**, D412–D416 (2009).
652 92. F. Mendiburu, R. Simon, F. De Mendiburu, Agricolae-Ten years of an Open source Statistical
653 tool for experiments in Breeding, agriculture and biology (2015)
654 https:/doi.org/10.7287/peerj.preprints.1404v1.
655 93. Dewey M, metap: meta-analysis of significance values. R package version 1.4. (2020).
656 94. A. Zeileis, T. Hothorn, Diagnostic Checking in Regression Relationshipslmtest citation info. *R*
657 *News* **2**, 7–10 (2002).
658 95. R Core Team, A language and environment for statistical computing. R Foundation for
659 Statistical Computing, Vienna, Austria. (2017).
660
661

Figures

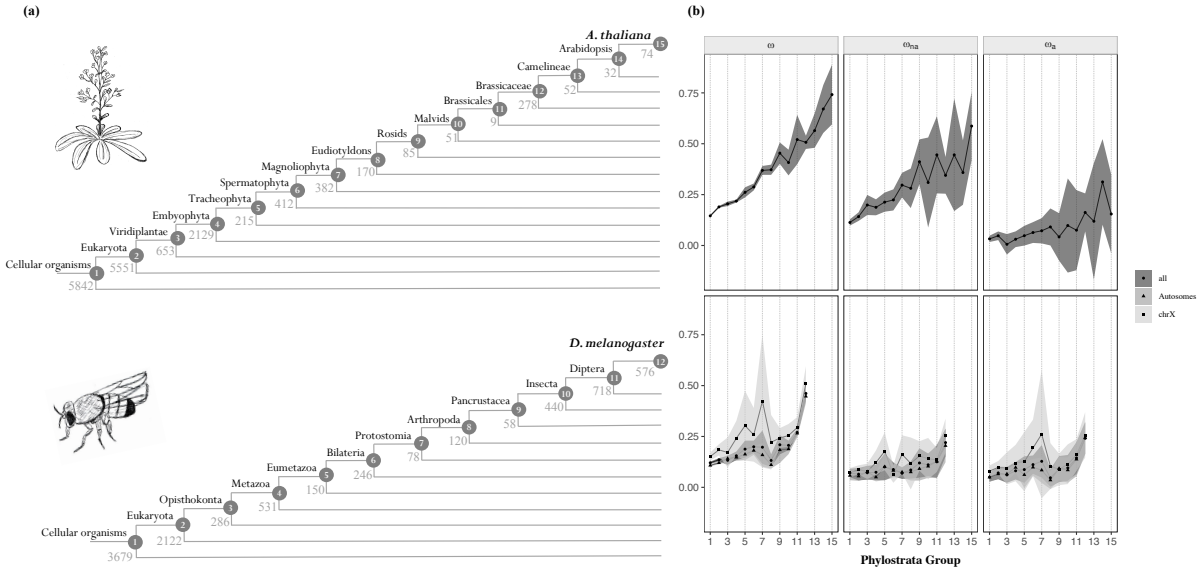
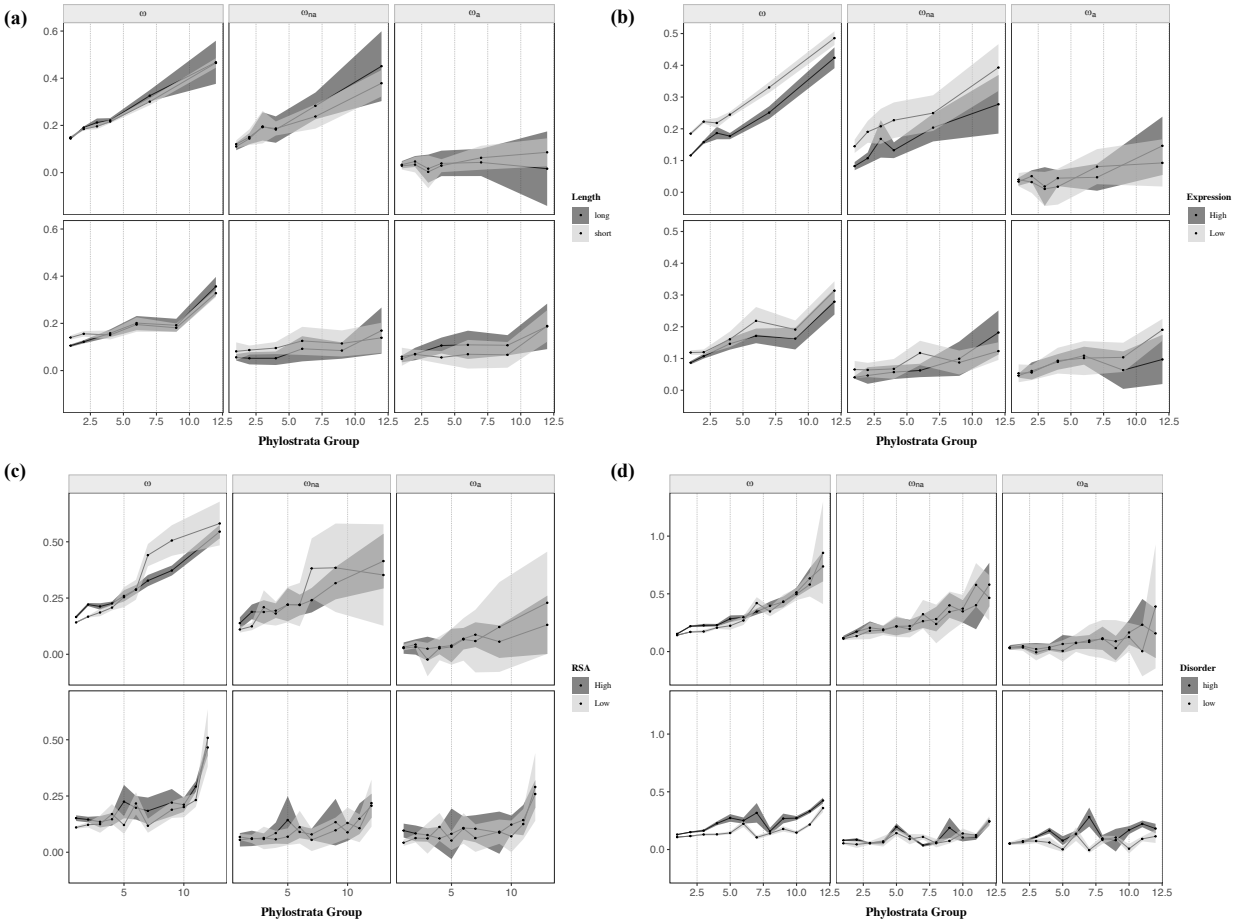


Figure 1. (a) Phylogenetic definition of the strata used in the analyses for *A. thaliana* (top) and *D. melanogaster* (bottom). The number of genes mapped to each clade is shown. **(b)** Relationship between the rate of protein evolution (ω), non-adaptive non-synonymous substitutions (ω_{na}) and adaptive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



674

675 **Figure 2.** Estimates of ω, ω_{na} and ω_a plotted as a function of **(a)** protein length and **(b)** mean
676 expression levels, **(c)** relative solvent accessibility, and **(d)** protein intrinsic disorder with gene age in
677 *A. thaliana* (top) and *D. melanogaster* (bottom). Analyses were performed by comparing short and
678 long genes **(a)**, lowly and highly expressed genes **(b)**, proteins with low and high mean RSA values
679 **(c)**, and proteins with low and high average intrinsic disorder **(d)** across age categories (see Material
680 and Methods). Legend as in Figure 1.

681

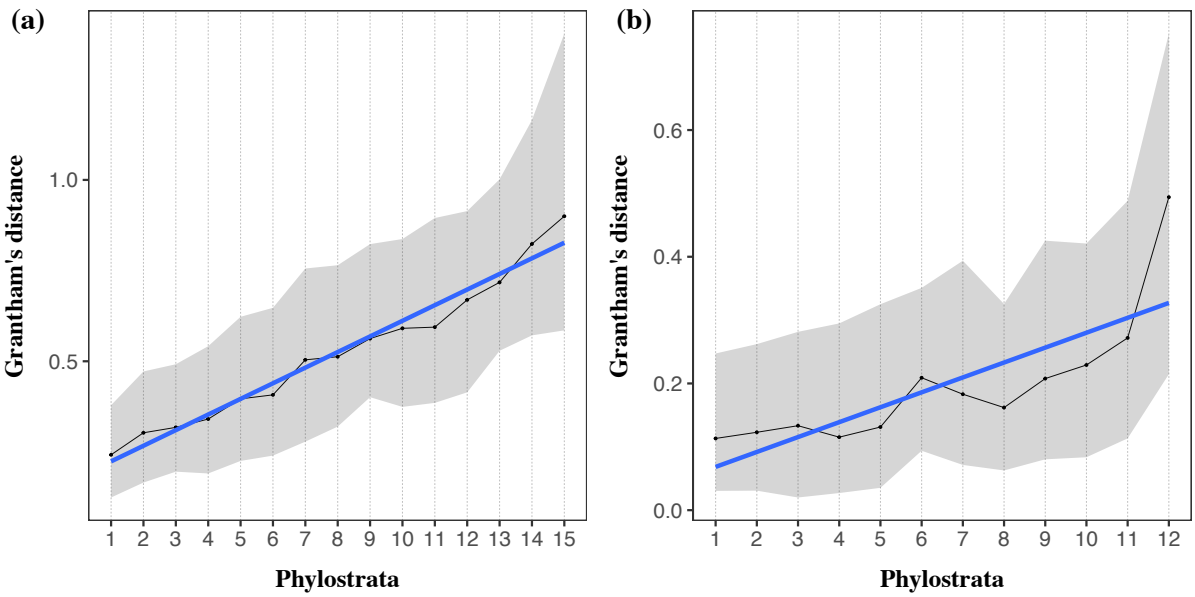


Figure 3. Relationship between gene age and Grantham's distance between amin-acids for *A. thaliana* (a) and *D. melanogaster* (b). A linear model was fitted between gene age and Grantham's distances values and is represented with the blue line. For each clade, the median value of the Grantham's distance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

688 **Tables**
689

690 **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
691 gene age with the respective co-factors: protein length, gene expression, protein intrinsic disorder and the mean relative solvent accessibility per gene. The
692 combined probabilities for each co-factor within and across species are presented in the fields "Weighted Z" and "Weighted Z across species", respectively, for
693 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 **			

694

695 **Note**. For each variable, the correlation coefficient and the combined probabilities are shown with the respective significance (*$P < 0.05$; **$P < 0.01$;
696 ***$P < 0.001$; "." $0.05 \leq P < 0.10$) for ω, ω_{na} and ω_a in *Arabidopsis* and *Drosophila*.