Intermolecular interactions drive protein adaptive and co-adaptive evolution at both species and population levels

Junhui Peng, Nicolas Svetec, Li Zhao*

Laboratory of Evolutionary Genetics and Genomics, The Rockefeller University, New York, NY 10065, USA

*Correspondence to: lzhao@rockefeller.edu

Abstract

Proteins are the building blocks for almost all the functions in cells. Understanding the molecular evolution of proteins and the forces that shape protein evolution is essential in understanding the basis of function and evolution. Previous studies have shown that adaptation frequently occurs at the protein surface, such as in genes involved in host-pathogen interactions. However, it remains unclear whether adaptive sites are distributed randomly or at regions associated with particular structural or functional characteristics across the genome, since many proteins lack structural or functional annotations. Here, we seek to tackle this question by combining large-scale bioinformatic prediction, structural analysis, phylogenetic inference, and population genomic analysis of *Drosophila* protein-coding genes. We found that protein sequence adaptation is more relevant to function-related rather than structure-related properties. Interestingly, intermolecular interactions contribute significantly to protein adaptation. We further showed that intermolecular interactions, such as physical interactions may play a role in the co-adaptation of fast-adaptive proteins. We found that strongly differentiated amino acids across geographic regions in protein-coding genes are mostly adaptive, which may contribute to the long-term adaptive evolution. This strongly indicates that a number of adaptive sites tend to be repeatedly mutated and selected in evolution, in the past, present, and maybe future. Our results highlight the important roles of intermolecular interactions and co-adaptation in the adaptive evolution of proteins both at the species and population levels.

Introduction

Natural selection plays an important role in the molecular evolution of protein sequences. Recent advances in genome sequencing and reliable inference methods at both phylogenetic and population levels have enabled fast and robust estimation of evolutionary rates and adaptation driven by natural selection. In addition, the increased availabilities of structural and functional data of proteins have made it possible to study how structural and functional constraints affect protein sequence evolution and adaptation. Different proteins and different sites within a protein have varying rates of evolution and adaptation due to both structural and functional constraints (Echave et al., 2016; Kosiol et al., 2008;

Lindblad-Toh et al., 2011; Zhang and Yang, 2015). For example, genes that are highly expressed or perform essential functions are often under strong purifying selection and tend to evolve slowly (Drummond et al., 2005; Moutinho et al., 2019; Pál et al., 2001; Zhang and He, 2005; Zhang and Yang, 2015); genes involved in host-pathogen interactions, e.g., immune responses and antivirus responses, show exceptionally high rates of adaptive changes (Enard et al., 2016; Nielsen et al., 2005; Obbard et al., 2009; Palmer et al., 2018; Sackton et al., 2007; Sironi et al., 2015; Uricchio et al., 2019); and residues that are intrinsically disordered or at the protein surface are fast evolving and proved to be hotspots of adaptive evolution (Afanasyeva et al., 2018; Goldman et al., 1998; Lin et al., 2007; Moutinho et al., 2019; Ramsey et al., 2011). More recently, Slodkowicz & Goldman (Slodkowicz and Goldman, 2020) employed genomic-scale integrated structural and evolutionary phylogenetic analysis in mammals and showed that positively selected residues are clustered near ligand binding sites, especially in proteins that are associated with immune responses and xenobiotic metabolism. However, it remains unclear how adaptive sites are distributed in the genome and how adaptation is related with functions and structures. Moreover, most of the existing literature is focused on protein differences between species, and it remains unclear how much of within-species selective processes like spatially varying selection may contribute to long-term evolution.

Although evidence has shown that adaptation is more likely to occur at intrinsically disordered regions (Afanasyeva et al., 2018) and clustered at the surface of proteins (Dasmeh et al., 2013; Moutinho et al., 2019; Slodkowicz and Goldman, 2020), it remains unclear how functional and structural properties of proteins shape adaptation at the species and population scale. Moreover, due to the lack of structural and functional information of many proteins in the genome, the underlying evolutionary mechanism derived from current studies might be incomplete. Here, we systematically investigated the evolution and adaptation of protein-coding genes in *Drosophila melanogaster* by comparing it to its closely related species and their own populations, to distinguish the main factors that impact the evolution and adaption at the protein-coding level. We applied large-scale bioinformatic and structural analysis to obtain the structural and functional properties of proteins. We then classified residues into different structural and functional sites. By comparing rates of sequence evolution and adaptation between different proteins and sites, we were able to locate hotspots of adaptation at the genome scale. We found that functional properties are better predictors of protein adaptation rates than structural properties. Interestingly, we found that adaptation rates of a protein positively correlate with the fraction of residues that are involved in intermolecular interactions inside the protein. In agreement with this finding, we found that putative binding regions including allosteric sites at protein surface show higher rates of adaptive evolution than other sites. For proteins under fast-adaptive evolution, we showed that they tend to interact with each other more frequently than random expectations, suggesting fast-adaptive genes might undergo co-

69 adaptive evolution. We further discovered that co-adaptation might be universal for many interacting
70 proteins in *D. melanogaster*.
71 Our results suggest that intermolecular interactions in *D. melanogaster* are an important driver of
72 protein adaptive evolution. We further hypothesized and provided evidence that intermolecular
73 interactions, such as physical interactions might be an important mechanism that contribute to the co-
74 adaptive evolution of interacting proteins in *D. melanogaster* genome. One of the intriguing questions is
75 that how the adaptive signals in populations (short-term) and between species (long-term) are correlated.
76 We then asked if those patterns hold for selective process occurring within species. Despite the abundance
77 of literature studying geographic variation in Drosophila species (Bergland et al., 2014; Fabian et al.,
78 2012; Kolaczkowski et al., 2011; Lack et al., 2015; Langley et al., 2012; Pitchers et al., 2013; Reinhardt et
79 al., 2014; Svetec et al., 2016), very little is known for a systematic evaluation of the protein properties
80 affected by spatially varying selection. We thus investigated protein adaptation signals of strongly
81 differentiated amino acids across geographic regions, which were often associated with within-species
82 local adaptations (Matthey-Doret and Whitlock, 2019). We showed that most of the patterns found in the
83 between-species in fact hold at the within species levels. This may partly because most sites contributing
84 to within-species local adaptation tend to also contribute to long-term adaptive evolution in *D.*
85 *melanogaster*, suggesting that a subset of protein-coding loci are constantly or repeatedly utilized for
86 adaptive purpose.
87
88 **Results**
89 **Putative molecular interaction sites are hotspots for protein adaptive evolution**
90 To uncover the main factors that impact the evolutionary rates of genes, we analyzed 13,528 protein-
91 coding genes in *D. melanogaster* using genome data from *melanogaster* subgroup species and *D.*
92 *melanogaster* population genomics data from 205 inbred lines from Drosophila Genetic Reference Panel,
93 Freeze 2.0, DGRP2 (Huang et al., 2014). We applied a maximum likelihood method (Yang, 2007) to
94 compute the dN/dS ratio (ω) using the protein-coding sequences of five closely related melanogaster
95 subgroup species (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba* and *D. erecta*). We estimated
96 the proportions of adaptive changes (α) in each gene by applying an extension of McDonald–Kreitman
97 (MK) test named asymptotic MK (Messer and Petrov, 2013; Uricchio et al., 2019) using *D. yakuba* as
98 outgroup. We then calculated the rate of adaptive changes (ω_a) of each gene by multiplying ω to α ($\omega_a =$
99 $\alpha\omega$) (Moutinho et al., 2019) (See *Material and Methods*). The rate of nonadaptive changes can be further
100 calculated by $\omega_{na}=\omega-\omega_a$. Finally, we successfully assigned ω to 12,118 protein coding genes and ω_a and
101 ω_{na} to 7,192 genes. For each of *D. melanogaster* genes subjecting the same analysis pipeline, we further
102 obtained 17 different structural or functional properties (see *Material and Methods*, supplementary file

103 S1, Table S1-S2). We calculated Pearson's correlations of ω, ω_a and ω_{na} with all these properties (Table
104 S1). Many of these genome-wide correlations were expected (for details, see Supplement Information,
105 Table S1-S2, Fig. S1-S5). Interestingly, we found that some previously unexplored properties, fractions
106 of molecular-interaction sites (including PPI-site ratio, ratio of residues involved in protein-protein
107 interactions, and DNA-site ratio, ratio of residues involved in protein-DNA interactions) were strongly
108 positively correlated with ω, ω_a, and ω_{na} (Supplement Information, section *Molecular interactions*
109 *contribute to the variations of protein sequence evolution and adaptation*, Table S1, Fig. S1-S2). The
110 results indicate that molecular interactions might act as an important factor that drives protein adaptive
111 evolution in the *Drosophila* genome.
112 We then investigated whether residues involved in molecular interactions are targets for adaptive
113 evolution. To tackle this question, we predicted protein-protein interaction sites (PPI-sites) and DNA
114 binding sites (DNA-sites) for each of *D. melanogaster* protein sequences (see Material and Methods). In
115 addition, we characterized allosteric residues as surface and interior critical residues with STRESS model
116 (Clarke et al., 2016) for all the structural models. We also extracted putative binding sites from STRESS
117 Monte Carlo (MC) simulations. We calculated ω, ω_a, and ω_{na} for residues in each of the putative
118 molecular interaction categories. Strikingly, we observed that residues involved in protein-protein
119 interactions, DNA binding and ligand binding exhibited higher rates of adaptive evolution compared to
120 their corresponding null sites (t-test, $p = 7e\text{-}10$, 0.18 and 7e-15, respectively) (Fig. 1A-C). In addition,
121 allosteric residues at protein surface showed higher adaptation rates than allosteric residues at protein
122 interior (t-test, $p = 3e\text{-}10$) or residues that are not involved in ligand binding (t-test, $p = 0.003$) (Fig. 1C,
123 Fig. S6).
124 To gain a better understanding of adaptation in molecular interaction sites, we further visualized
125 positive selections that are associated with molecular interactions. We first investigated whether adaptive
126 evolution is associated with particular protein structures or protein families. To do this, we looked into
127 fast-adaptive proteins with the largest ~15% rates of adaptation ($\omega_a > 0.15$) that are linked to high-quality
128 structural models. Interestingly, among these proteins, we found 45 enriched as trypsin-like
129 cysteine/serine peptidase domain and 17 7TM chemoreceptors, suggesting widespread adaptive evolution
130 acting on these protein families or protein domains in *D. melanogaster* (Table S3). Many of the 7TM
131 chemoreceptors are olfactory and gustatory genes and show adaptive evolution in various species such as
132 *Drosophila* and mosquito (Hill et al., 2002; Lawniczak and Begun, 2007; McBride, 2007; Wu et al.,
133 2009). In addition to these two protein families, previous studies identified recurrent positive selections
134 acting on some other fast-adaptive proteins in *Drosophila* and mammals, and the possible adaptive
135 evolution mechanisms have been linked to exogenous ligand binding, for example, serine protease

136 inhibitors (serpin), Toll-like receptor 4 (TLR-4), and cytochrome P450 (Jiggins and Kim, 2007;
137 Slodkowicz and Goldman, 2020).
138 We used the two representative cases of fast adaptive protein evolution of CG10232 and Or67a -
139 a trypsin-like cysteine/serine peptidase domain and a 7TM chemoreceptor, respectively - to illustrate the
140 link between adaptive evolution and molecular interactions in the two protein families with frequent
141 adaptive evolution. We observed that in both cases, positively selected sites were significantly closer to
142 predicted or inferred binding sites in the protein 3D structure (t-test, p = 8e-133 for CG10232 and 3e-169
143 for Or67a, Fig. 1D-E) and were overlapped with predicted or inferred binding pockets for CG10232
144 (Fisher's exact test, p-values 0.02, Fig. 1D). There might be an overlap with predicted or inferred binding
145 pockets for Or67a, but it is not statistically significant (Fisher's exact test, p = 0.19, Fig. 1E). Specifically,
146 in CG10232, we found clusters of positive selected sites around NAG binding sites that are inferred from
147 a crystal structure of serine protease (PDB code: 2XXL) (Fig. 1D), while in Or67a, positively selected
148 sites expand around the putative odorant binding channel formed by helices S1-S6 in extracellular regions
149 (Butterwick et al., 2018) (Fig. 1E).
150 Besides the examples that are associated with exogenous ligand or exogenous peptide binding, we
151 also identified two previously undescribed examples where adaptive evolution might be linked to
152 endogenous protein binding: Spaztle (spz, Fig. 1F) and Cul6 (Fig. 1G). Spaztle can bind to Toll-like
153 receptors (TLR) and trigger a humoral innate immune response. We built the missing loop in Spaztle in
154 the crystal structure of Toll/Spaztle complex (PDB code 4BV4) according to the dimeric crystal structure
155 of Spaztle (PDB code 3E07). In this complex structural model, we observed several positively selected
156 sites in Toll-4/Spaztle interfaces (Fig. 1F). Cul6, another example, is a protein in the cullins family in *D.*
157 *melanogaster*. The cullins protein family are known as scaffold proteins that assemble multi-subunit
158 Cullin-RING E3 ubiquitin ligase by forming SCF complex with F box and RING-box (Rbx) proteins
159 (Zheng et al., 2002). We constructed the putative Cul6 contained SCF complex by superimposition to the
160 crystal structure of the Cul1-Rbx1-Skp1-F box^{Skp2} SCF ubiquitin ligase complex (Zheng et al., 2002). In
161 the structural model, we observed positively selected sites in Cul6 clustered around the binding sites of
162 RING-box protein, Rbx1, and F-box protein, Skp1 (Fig. 1G). The examples above suggest that
163 intermolecular interactions, including both exogenous and endogenous binding, could contribute to
164 protein adaptive evolution.
165
166 **Frequent adaptive evolution and co-adaptative evolution in genes involved in reproduction,**
167 **immune system, and environmental information processing**
168 To find out whether specific biological functions were associated with fast-adaptive genes, we applied
169 DAVID GO analysis to the genes with the largest rates of adaptation ($\omega_a > 0.15$, top ~15%). The

170 significant GO terms are frequently linked to serine-type endopeptidase activity, reproduction, protein
171 lysis, chemosensory and other related biological functions (Table S4). As these fast-adaptive genes tend
172 to be enriched in similar biological functions, we asked whether these genes evolved co-adaptively, i.e.,
173 whether these proteins are interacting with each other frequently. To test this possibility, we obtained PPI
174 of *D. melanogaster* from STRING database (Szklarczyk et al., 2019) and analyzed protein-protein
175 interactions among fast-adaptive proteins. We found that fast-adaptive proteins tend to interact with each
176 other more frequently than expected (PPI enrichment p-value < 1.0e-16). In the PPI network of fast-
177 adaptive proteins, we observed 7 strongly connected sub-clusters with at least 5 members (Fig. 2A, Table
178 S5, example Fig. 2B-C). Proteins in these sub-clusters are enriched in biological processes such as
179 reproduction, immune response, defense response to bacterium and virus, RNA interference, chitin
180 metabolic, etc., (Table S5-S11) which are in line with the GO analysis of fast-adaptive genes (Table S4),
181 and previous enrichment analysis of positively selected genes identified from genome-wide studies
182 (Begun et al., 2007; Enard et al., 2016; Nielsen et al., 2005).
183 We next asked whether co-adaptation plays a role in the adaptive evolution of interacting proteins
184 to a broader extend, including both fast- and slow-adaptive proteins. To address this question, we
185 analyzed and compared adaptation rates of all D. *melanogaster* PPIs available in STRING database with
186 high confidence, and we found that protein partners of fast-adaptive proteins (ω_a>0.15) have significantly
187 larger maximum/average ω_a compared to slow-adaptive proteins (Fig. 3). We further analyzed and
188 visualized adaptive evolutionary rates of proteins in PPI networks of 9 different biological pathways
189 extracted from KEGG pathways, including the immune system, xenobiotics biodegradation, response to
190 the environment, aging and development, genetic information processing, sensory system, transport and
191 catabolism, cell growth and death and metabolism. We observed that, in these PPI networks, proteins with
192 relatively large ω_a tend to interact with each other (Fig. 4A-B). We also noticed that, for pathways that are
193 previously known as adaptation-hotspots (e.g., immune system), fast-adaptive proteins can act as central
194 nodes and are co-adaptively evolving with other fast-adaptive proteins (Fig. 4A, C). While in pathways
195 such as transport and catabolism, fast-adaptive proteins are mainly at PPI periphery. Specifically, we
196 observed that in pathways related to immune system and environment adaptation, where adaptive
197 evolution often occurs, fast-adaptive proteins have comparable number of interactors as other genes (Fig.
198 4C-D), while in conserved pathways such as transport and catabolism, fast-adaptive genes are often at
199 network peripheries and have significantly fewer interactors (Fig. 4E-F). In line with these findings, we
200 found that ω_a are larger in pathways that harbor fast-adaptive proteins as central nodes than other
201 pathways (Fig. S7).
202
203 **Physical interactions contribute to co-adaptation of fast-adaptive genes**

204 Having established that molecular interactions contribute to the adaptive evolution of protein sequence,
205 we then investigated whether these physical molecular interactions could drive protein-protein co-
206 adaptation. To do this, we looked into interacting fast-adaptive protein pairs that are associated with
207 known or inferred complex structural models. For inferred complex structural models, we superimposed
208 the structural models of the pair of proteins onto their high resolution homologous complex structures.
209 Here we illustrated co-adaptation at PPI interface in two examples: Toll-4/Spatzle and
210 Spn28Db/CG18563 (Fig. 2B-C).
211 ***Toll-4/Spatzle***. Toll-4 is a member of toll-like receptors. Previous studies have shown strong evidence of
212 adaptive evolution of Toll-4 in *Drosophila* and mammals (Levin and Malik, 2017; Slodkowicz and
213 Goldman, 2020), and implied the important roles of ligand binding in the positive selection and
214 coevolution of Toll receptors and Spatzle proteins (Lima et al., 2021). Toll-4 can bind to Spatzle and
215 trigger further innate immune responses with high confidence (inferred from STRING database). In the
216 previous section, we showed that several positively selected sites in Spatzle overlap with Toll-Spatzle
217 interfaces (Fig. 1F). Here, we further showed that, in Toll-4, many significantly positively selected sites
218 were located at interface for Spatzle (Fig. 2B), which is in line with a previous study of Toll-4 in *D.*
219 *willistoni* (Levin and Malik, 2017).
220 ***Spn28Db/CG18563***. Spn28Db is one of the serine protease inhibitors in *D. melanogaster* expressed in
221 male accessory glands, while CG18563 belongs to the protein family of trypsin-like cysteine/serine
222 peptidase domain. The interactions between the two proteins were predicted with high confidence from
223 the STRING database, and the molecular interactions can be inferred from existing crystal structure of
224 serpin and bacteria protease complex (PDB code 1EZX). We observed many positive selected sites at the
225 molecular interface between the two proteins (Fig. 2C), suggesting that physical interactions might play a
226 role in the co-adaptation of the two proteins.
227
228 **Most geographically differentiated non-synonymous SNPs in protein-coding genes are adaptive**
229 To learn more about the relationship between short-term adaptation to local environments and long-term
230 adaptive evolution, we extracted residues with significant allele frequency differentiation across latitudes
231 in North America (Svetec et al., 2016) and Africa (Lack et al., 2015). For the DPGP3 African population
232 data, we followed the same protocol as Svetec et al. 2016 to identify significantly differentiated SNPs
233 (section *Population genetics of DPGP3 African population* in Material and Methods). We then computed
234 evolutionary rates (ω), adaptation rates (ω_a), non-adaptation rates (ω_{na}) and proportions of adaptive
235 changes (α) of these residues as in the previous section (Fig. 5A-B, Fig. S8). We observed that, in both
236 the North American population and the African population, these sites have significantly higher
237 proportions of adaptive changes (Fig. 5A-B) than other SNPs, suggesting that they can be hotspots for

238 adaptive evolution. To find out whether these SNPs are related to even longer-term adaptive evolution,
239 we inferred positive selection sites of each protein-coding gene from phylogenic data (see Material and
240 Methods). We found that geographically differentiated non-synonymous SNPs are significantly enriched
241 for long-term positive selection (Fig. S9). To further characterize structural and functional properties of
242 short-term genetic variations, we mapped geographically differentiated non-synonymous residues to
243 different structural and functional characteristics, such as ISD, RSA, PPI-sites, DNA-sites and ligand-
244 binding sites. We found that these non-synonymous SNPs were significantly enriched in disordered
245 regions and protein surfaces, as well as in protein-protein interactions and ligand-binding (Fig. S9). To
246 better visualize the characteristics of these SNPs, we used *Toll-4* as an example. We mapped its
247 geographically differentiated non-synonymous sites onto its structural model. We found that these sites
248 are either positively selected or are located very close to positively selected sites (Fig. 5C-D). For
249 example, highly differentiated sites in North American population, N279 (FDR 3e-7) and H431 (FDR 3e-
250 6) were predicted to be positively selected both at a probability of p=0.9. While another highly
251 differentiated site, D424 was close to three positively selected sites S401 (p=0.8), H431 (p=0.95) and
252 V448 (p=0.8). We also noticed some differentiated sites that may be located within ligand binding sites,
253 including F297 (FDR 3e-3), S311 (FDR 3e-3), H431 (FDR 3e-6) and H462 (FDR 1e-2). In the structure
254 of Toll4, we also observed 3 highly differentiated sites in African populations, which are S311 (FDR 4e-
255 2), H431 (FDR 2e-2) and S490 (FDR 2e-4). We noticed that all the 3 sites overlapped with differentiated
256 SNPs in North America and two of them (S311 and H431) localized within ligand binding sites (Fig. 5D),
257 further supporting our observation that ligand binding sites in some genes may undergo recurrent adaptive
258 evolution.
259
260 **Discussion**
261 In this study, we systematically studied the impact of structure- and function-related gene properties on
262 protein sequence evolution and adaptation in *D. melanogaster* genome. We found that molecular
263 interactions in proteins contribute to the variation of protein sequence adaptive evolution. A novel
264 discovery of this work is that molecular interaction sites including protein-protein interaction sites and
265 protein-DNA interaction sites are hotspots for adaptative evolution. We revealed that fast-adaptive
266 proteins tend to interact with each other frequently and protein partners of these fast-adaptive proteins
267 tend to have higher adaptation rates, suggesting that co-adaptive evolution might be common in *D.*
268 *melanogaster*. By visualizing examples of interacting fast-adaptive proteins, we further demonstrated that
269 physical interactions may contribute to the co-adaptation of fast-adaptive proteins.
270 Protein surface and intrinsic disorder regions are frequent targets for adaptive evolution and
271 contribute to the variations of protein sequence adaptive evolution (Afanasyeva et al., 2018; Moutinho et

272 al., 2019). However, the detailed mechanisms underlying these observations remains unclear. One
273 possible explanation would be that these regions are frequently linked to intermolecular interactions
274 (Afanasyeva et al., 2018; Moutinho et al., 2019). For example, Moutinho et al. hypothesized that
275 molecular interactions involved in host-pathogen coevolution were the major driver of protein adaptation
276 (Moutinho et al., 2019). Here, we further identified that proportions of possible molecular interaction sites
277 inside proteins contribute to the variations of protein sequence adaptive evolution. These molecular
278 interaction sites or regulatory sites at protein surface can be hotspots of protein adaptation. Indeed, some
279 specific molecular interactions have been linked to adaptive evolution in several case studies (Bachtrog,
280 2008; Hughes and Nei, 1988; Levin and Malik, 2017; Schott et al., 2014), a recent study on sweet taste
281 receptors in the songbird radiation (Toda et al., 2021), and large-scale studies based on proteins with high
282 quality structural models (Slodkowicz and Goldman, 2020). In the latter study, the authors showed that
283 amino acids under positive selection in mammals tend to cluster closer to binding sites of exogenous
284 ligands than expected by chance (Slodkowicz and Goldman, 2020), suggesting an important role of
285 functionally important regions in adaptive evolution. Here, we extend the conclusion *to D. melanogaster*
286 genome, including proteins with or without high resolution structural models. We also showed that in
287 addition to exogenous ligands, endogenous ligands might also contribute to adaptive evolution, while the
288 latter might explain why interacting proteins tend to evolve co-adaptively.
289 Notably, previous studies showed that multi-interface proteins tend to be evolving more slowly
290 than single-interface proteins (Kim et al., 2006) and that proteins with many interactors tend to evolve
291 slowly (Jordan et al., 2003), which seems to be contradictory to our results that proteins with more
292 interaction sites evolve faster and have faster adaptation rates. Here, we argue that, in our study, we used
293 sequence profiles to predict molecular interaction sites in proteins at a genomic scale, rather than only
294 looking into proteins with high resolution structures. In this way, we may capture many weak or transient
295 interactions, which are evolving faster than obligate and conserved interactions (Mintseris and Weng,
296 2005). Meanwhile, we did not exclude intrinsically disordered regions (IDR) or intrinsically disordered
297 proteins (IDP) in our study, which are widespread in *D. melanogaster* genome. It has been suggested that
298 IDR/IDP tend to evolve fast due to the lack of structural restraints (Echave et al., 2016). In the functional
299 aspect, IDR/IDP are thought to be promiscuous binders through many multiple binding mechanisms,
300 including forming static, semi-static, and fuzzy or dynamic complexes (Uversky, 2019), suggesting that
301 the evolution of IDR/IDP cannot be explained merely by the lack of structural restraints. Indeed, IDP and
302 IDR in the human genome were found to be undergoing extensive adaptive evolution (Afanasyeva et al.,
303 2018). At last, it has been recognized that, except for allosteric regulations, encounter complexes
304 (Gabdoulline and Wade, 1999) might also play an important role in mediating intermolecular interactions,
305 such as protein-protein association (Tang et al., 2006) and protein-ligand binding (Re et al., 2019). Since

306 encounter residues that are responsible for encounter complexes do not reside in conserved binding
307 interfaces, these residues could be under relaxed purifying selection or even positive selection, which
308 could be another yet-to-identify mechanism that contributes to protein sequence adaptive evolution.
309 We showed that fast-adaptive proteins are enriched in molecular functions such as reproduction,
310 immunity, and environmental information processing (Begun and Lindfors, 2005; Begun and Whitley,
311 2000; Lazzaro et al., 2004). We further demonstrated that fast-adaptive proteins tend to interact with each
312 other more frequently than random expectations, suggesting co-adaptation might be common among fast-
313 adaptive proteins. Mechanisms contributing to the co-adaptation could be: (1) interacting fast-adaptive
314 proteins are often enriched in similar molecular functions and under similar selective pressure; (2)
315 interacting fast-adaptive undergo co-evolution through physical interactions. In this study, we showed two
316 examples that adaptive evolution could occur at protein-protein interface, which suggest that physical
317 interactions could contribute to the co-adaptation of fast-adaptive proteins in *D. melanogaster*. Moreover,
318 we showed that co-adaptation might exist to a broader extend rather than only among fast-adaptive
319 proteins. Specifically, proteins that interact with fast-adaptive proteins tend to have higher adaptation
320 rates. Since molecular interactions contribute to adaptive evolution, it is reasonable to hypothesize that
321 co-adaptation at a broader extend could be regulated by these interactions. Actually, it has been suggested
322 that interacting proteins tend to have similar evolutionary rates and the possible mechanism would be the
323 co-evolution of physical interactions (Pazos and Valencia, 2008).
324 In this study, we found that amino acids showing strong geographically differences often overlap
325 with sites that show adaptive signals between species. These loci follow similar patterns as adaptive
326 changes, i.e. they are enriched in disordered regions, protein surfaces, and functionally important regions.
327 These results suggest that population differentiation of protein-coding genes can be an important basis for
328 long-term adaptive evolution. In other words, many SNPs are repeatedly selected for the adaptive
329 processes in evolution. Importantly, our results indicate that most of the clinal amino-acid changes are
330 adaptive, suggesting that non-selective forces play a less essential role in the SNPs that show strong
331 geographic differences. Our results also support a large effect of spatially varying selection on protein
332 sequence and structures (Storz and Kelly, 2008). Interestingly, our previous work showed that
333 geographically differentiated SNPs often occur on the same orthologous genes between species but rarely
334 same SNPs (Zhao et al., 2015), it would interesting to extend this work to *D. simulans* to study parallel
335 protein evolution at the structural level.
336 It should be noted that studies at the genomic scale that aim to uncover the function- or structure-
337 related constraints imposed on protein sequence evolution and adaptation share similar limitations that for
338 most of the proteins or residues, structural or functional information would be incomplete or even
339 missing. To overcome this, in this study, we used highly accurate neural-network based tools to predict

340 molecular interactions, secondary structures, intrinsic structural disorder, relative solvent accessibility for
341 each of the proteins. In this way we were able to identify key factors that impact protein sequence
342 evolution and adaptation in a less accurate but rather systematic fashion. Notably, it has been reported that
343 false positives are nonnegligible in methods to the estimation of adaptive evolution (Markova-Raina and
344 Petrov, 2011), and other mechanisms including translational selection were acting on evolution of
345 protein-coding genes (Larracuente et al., 2008), which together hinder our understanding toward protein
346 evolution and protein adaptation. Another limitation of our study is that we did not include indels,
347 especially non-frameshift indels, as it is very difficult to address the adaptive effects of indels. However,
348 with method development and increased knowledge of protein structures, this would be an important
349 question to investigate in the future. In the recently years, deep learning-based predictors and estimators
350 have contributed to our knowledge towards protein structure (Jumper et al., 2021), protein function
351 (Kulmanov and Hoehndorf, 2020), or even protein evolution (Schrider and Kern, 2018). We hope that
352 with the availability of more and more curated structural, functional information and complex structural
353 models of proteins in the near future, we will be able to uncover the precise role of molecular interactions
354 in protein sequence adaptive evolution.
355
356 **Material and Methods**
357 ***d_N/d_S* ratio (ω).**
358 We used a maximum likelihood method to infer d_N/d_S ratio (ω) of *D. melanogaster* protein-coding genes
359 using the genome sequences of five species in *melanogaster* subgroup (*D. melanogaster*, *D. simulans*, *D.*
360 *sechellia*, *D. yakuba*, and *D. erecta*). The protein-coding sequences were extracted from the alignments of
361 26 insects, which were obtained from UCSC Genome Browser
362 (http://hgdownload.soe.ucsc.edu/downloads.html). The sequences were further processed by GeneWise
363 (Birney et al., 2004) to remove possible insertions and deletions using the longest isoforms of the
364 corresponding *D. melanogaster* protein sequences as references (FlyBase version r6.15) (Thurmond et al.,
365 2019). The processed sequences were then realigned by PRANK -codon function (Löytynoja, 2014). The
366 conservation scores and coverages of the alignments can be found in Fig. S10. We used codeml in PAML
367 (Yang, 2007) to compute gene-specific ω using M0 model. We removed sequences that have more than
368 15% of their nucleotides not aligned (gaps) to *D. melanogaster* genes in more than 2 species. To further
369 avoid numeric errors and ensure reasonable estimations, we only retained relatively divergent sequences
370 that are: (1) divergent with dS larger than 0.3, (2) less divergent with dS larger than 0.1 and dN smaller
371 than 0.001 (dS>>dN). At last, there were 12118 genes in total that passed all the criteria and were
372 assigned gene specific ω, containing 6,538,872 amino acids. We also calculated site-specific ω by using

likelihood ratio tests (LRT) comparing M7 model against M8 model and M8fix model and M8 model (Yang et al., 2005).

Rate of adaptive and nonadaptive changes.

We recalled all SNPs of 205 inbred lines from the Drosophila Genetic Reference Panel (DGRP), Freeze 2.0 (Huang et al., 2014) (http://dgrp2.gnets.ncsu.edu). We then generated 410 alternative genomes using all monoallelic and bi-allelic SNP data sets. We extracted the coding sequences of *D. melanogaster* genes from the generated alternative genomes, removed all possible insertions and deletions using GeneWise (Birney et al., 2004) as described above. We then align all the coding sequences to their corresponding aligned CDS sequences using PRANK -codon function (Löytynoja, 2014). We removed polymorphisms segregating at frequencies smaller than 5% to reduce possible slightly deleterious mutations (Charlesworth and Eyre-Walker, 2008). In order to avoid possible effects of low divergence between *D. simulans* and *D. melanogaster* (Keightley and Eyre-Walker, 2012), we used *D. yakuba* as outgroup to estimate nonsynonymous polymorphisms (Pn), synonymous polymorphisms (Ps), nonsynonymous substitutions (Dn) and synonymous substitutions (Ds) by MK.pl (Begun et al., 2007; Langley et al., 2012). Similar as Begun et al. (Begun et al., 2007), we only analyzed genes with at least six variants for each of substitutions, polymorphisms, nonsynonymous changes and synonymous changes. We used an extension of MK test, asymptotic MK (Messer and Petrov, 2013; Uricchio et al., 2019), to estimate the proportions of adaptive changes (α). The rate of adaptive changes (ω_a) was then calculated as $\omega_a = \omega\alpha$ and the rate of non-adaptive changes as $\omega_{na} = \omega - \omega_a$. Details of the asymptotic MK test were as following:

(1) Classical McDonald–Kreitman test. According to Smith and Eyre-Walker (Smith and Eyre-Walker, 2002), the proportions of adaptive changes for protein-coding genes can be calculated as following:

$$\alpha = 1 - \frac{DsPn}{DnPs}$$

According to this equation, we could estimate the proportion of adaptive changes and carried out classical MK test by applying Fisher's exact test.

(2) Asymptotic estimation of α. A known problem of the classical estimation of α above is the accumulation of slightly deleterious mutations at low frequencies. We therefore used an extension of MK test, asymptotic MK test approach (Messer and Petrov, 2013) to estimate the proportions of adaptive changes. As in original aMK, we defined $\alpha(x)$ as a function of derived allele frequency (x):

$$\alpha(x) = 1 - \frac{DsPn(x)}{DnPs(x)}$$

404 where Pn(x) and Ps(x) are number of non-synonymous and synonymous polymorphisms at frequency x,
405 respectively. However, the original approach may suffer from numeric errors when there were very few
406 polymorphic sites, which is quite common in many of *D. melanogaster* genes. To make the estimations
407 more robust while preserving the same asymptote, we further define Pn (x) and Ps(x) as total number of
408 Pn and Ps above frequency x as described in Uricchio et al (Uricchio et al., 2019). We fitted $\alpha(x)$ to an
409 exponential curve of $\alpha(x) \approx \exp(-bx)+c$ using lmfit (Newville and Stensitzki, 2018) and determined the
410 asymptotic value of α at the limit of x, 1.0. We then estimate the rate of adaptive changes (ω_a) as

411 $$\omega_a = \frac{N_a/L_N}{dS} = \frac{dN_a}{dS} = \frac{dN_a}{dN} \cdot \frac{dN}{dS} = \alpha\omega$$

412 where N_a is the number of adaptive changes and $dN_a=N_a/L_N$ is the number of adaptive changes per
413 nonsynonymous site. Finally, we calculated the rate of nonadaptive changes (ω_{na}) as $\omega_{na}=\omega-\omega_a$. The final
414 dataset contains 7192 protein-coding genes, with smallest ω_a being 0.00 and largest being 1.29.
415
416 ***Structure-/function- related properties of D. melanogaster proteins.***
417 We obtained function-related properties mentioned in the main text as following. We derived *D.*
418 *melanogaster* gene ages (Kondo et al., 2017; Zhang et al., 2010) for genes that are specific to *Drosophila*,
419 and from GenTree (Shao et al., 2019) for genes that are beyond *Drosophila* clade. We then assigned a
420 pseudo-age to each of the genes. Specifically, there are 11 age groups from "cellular organisms",
421 assigning to a pseudo age value of 0, to "melanogaster", assigning a pseudo age value of 10. We
422 downloaded *D. melanogaster* protein-protein interaction (PPI) from STRING database (Szklarczyk et al.,
423 2019). A cut-off of combined score larger than 0.7 was used to retain high confident PPI for further
424 analysis. We then used BSpred (Mukherjee and Zhang, 2011) to predict protein-protein interaction (PPI)
425 sites and DRNApred (Yan and Kurgan, 2017) to predict DNA binding sites. For each protein, we
426 calculated ratios of protein interaction residues (PPI-site ratio) and ratios of DNA binding residues (DNA-
427 site ratio) by dividing total predicted protein interaction sites and DNA binding sites over protein length,
428 respectively. For structure-related properties, we used DeepCNF (Wang et al., 2016) to predict these
429 properties for each gene, including three-state secondary structures (helix, sheet, and coil), structural
430 disorder, relative solvent accessibility (RSA). Further, we calculated the ratios of helix, sheet,
431 helix+sheet, and coil residues of each gene from predicted secondary structures. DeepCNF (Wang et al.,
432 2016) is a deep learning method to capture complex sequence–structure relationships and was proved to
433 have high accuracy in the prediction of protein secondary structures (Yang et al., 2018), structural
434 disorder (Necci et al., 2021) and relative solvent accessibility (Wang et al., 2016). For each gene, we
435 computed intrinsic structural disorder (ISD) and relative solvent accessibility (RSA), as protein-length
436 normalized summations of the probabilities of each residue being disorder and exposed, respectively.

437

438 ***Gene expression patterns.***

439 We downloaded gene expression profile from FlyAtlas2 (Leader et al., 2018). We converted FPKM to
440 TPM by normalizing FPKM against the summation of all FPKMs as following:

441 $$\mathrm{TPM_i} = \frac{\mathrm{FPKM_i}}{\sum \mathrm{FPKM_j}} \times 10^6$$

442 After TPM conversion, we only retained genes with expression level larger than 0.1 TPM for further
443 analysis. We treated male and female whole-body TPM as male and female expression levels. We
444 calculated mean expression level by averaging male and female TPM. We used following Z-score to
445 describe male specificities of *D. melanogaster* genes:

446 $$zscore = \frac{TPM(male\ expression) - TPM(female\ expression)}{\sqrt{sd^2(male\ expression) + sd^2(female\ expression)}}$$

447 We calculated tissue specificities of genes using tau values (Yanai et al., 2005) based on the expression
448 profiles of 27 different tissues.

449 ***High quality 3D structures of D. melanogaster proteins.***

450 We downloaded high-quality structures or structural models of *D. melanogaster* proteins from protein
451 data bank (PDB) (Burley et al., 2019), SWISS-MODEL Repository (Bienert et al., 2017), and
452 MODBASE (Pieper et al., 2011), with descending priorities. For example, if there were 3D structures of
453 the same protein or protein region in multiple databases, we first considered high-resolution structures
454 from PDB; if no structures were found in PDB, we then considered SWISS-MODEL Repository; and at
455 last from MODBASE. In addition, we used blastp (Camacho et al., 2009) to search homologs of each *D.*
456 *melanogaster* protein against all PDB sequences with an E-value threshold of 0.001. We further carried
457 out comparative structural modeling using RosettaCM (Song et al., 2013) to model high-quality structural
458 models of proteins or protein regions that were not available in PDB, SWISS-MODEL Repository and
459 MODBASE. For each RosettaCM simulation, we used no more than 5 most significant hits from blastp
460 search. For proteins that are in complex forms, we only extracted monomers for further analysis. At last,
461 we obtained 14543 high quality structural models, corresponding to 11284 genes. These structural models
462 contain 2,691,913 unique amino acids, 41.2% of all the residues in genes that were assigned ω.

463

464 ***Evolutionary rates of different structural/functional sites.***

465 We classified amino acids into different classes of structural/functional properties. Specifically, we
466 classified three classes for both ISD and RSA according to the probability of residues being disordered or
467 exposed: ordered or buried (0.00 to 0.33), medium (0.33 to 0.67), disordered or exposed (0.67 to 1.00).

468 For both PPI and DNA binding, we classified two classes: PPI-site or DNA-site (binding sites), None-PPI
469 or None-DNA (corresponding null sites for PPI or DNA binding). For residues that have 3D structures,
470 we used STRESS (Clarke et al., 2016) to predict putative ligand binding sites and allosteric sites from all
471 the high-quality structures or structural models. We chose STRESS over other programs because it takes
472 both geometry and protein dynamics into account and has been used in genome-wide studies and
473 explained many poorly understood diseases associated variants in humans (Clarke et al., 2016). The
474 allosteric sites were further classified as surface critical or interior critical according to their locations. We
475 then classified these residues into four groups: LIG (ligand binding sites), Surf. Crit. (surface critical
476 sites), Interior Crit. (interior critical sites) and Others (other sites). For each of the site classes, we
477 randomly sampled 100 sequences, each containing 10,000 amino acids. We computed ω, ω_a, and ω_{na} for
478 the randomly sampled sequences similar to the steps described in the above sections.
479
480 ***Population genetics of DPGP3 African population***.
481 We analyzed 20 high-quality genomes in high latitude South Africa and 30 high-quality genomes in low
482 latitude Ethiopia (Lack et al., 2015). We called all biallelic SNPs and removed SNPs segregating at
483 frequencies smaller than 5% to reduce possible slightly deleterious mutations. Similar to Svetec et al.,
484 2016 (Svetec et al., 2016), for each SNP in both populations, we calculated the fixation index, F_{ST}, and
485 used ormidp.test from epitools package in R to perform the odds ratio test for independence. For SNPs at
486 each chromosome arm, we calculated the false discovery rate (FDR) using the Bioconductor *q*-value
487 package (https://github.com/jdstorey/qvalue).
488
489 **Acknowledgements**
490 We thank members of the Zhao Lab for helpful discussions, especially Evan Witt for gene expression
491 analysis and Christopher Langer for critically reading an early version of the manuscript.
492
493 **Author contribution**
494 J.P. and L.Z. conceived the study. J.P. performed all the analysis with the input from L.Z., N.S.
495 conceptualized the population genetic analysis and helped with the interpretation, J.P. and L.Z. wrote the
496 manuscript.
497
498 **Funding**
499 The work was supported by NIH MIRA R35GM133780, the Robertson Foundation, a Monique Weill-
500 Caulier Career Scientist Award, an Alfred P. Sloan Research Fellowship (FG-2018-10627), a Rita Allen
501 Foundation Scholar Program, and a Vallee Scholar Program (VS-2020-35) to L. Z.. J.P. is supported by a

C. H. Li Memorial Scholar Fund Award at The Rockefeller University.

Declaration of interests

The authors declare no competing interests.

Reference

Afanasyeva, A., Bockwoldt, M., Cooney, C.R., Heiland, I., and Gossmann, T.I. (2018). Human long intrinsically disordered protein regions are frequent targets of positive selection. Genome Res. *28*, 975–982.

Bachtrog, D. (2008). Positive selection at the binding sites of the male-specific lethal complex involved in dosage compensation in Drosophila. Genetics *180*, 1123–1129.

Begun, D.J., and Lindfors, H.A. (2005). Rapid evolution of genomic Acp complement in the melanogaster subgroup of Drosophila. Mol. Biol. Evol. *22*, 2010–2021.

Begun, D.J., and Whitley, P. (2000). Adaptive evolution of relish, a Drosophila NF-kappaB/IkappaB protein. Genetics *154*, 1231–1238.

Begun, D.J., Holloway, A.K., Stevens, K., Hillier, L.W., Poh, Y.-P., Hahn, M.W., Nista, P.M., Jones, C.D., Kern, A.D., Dewey, C.N., et al. (2007). Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in Drosophila simulans. PLoS Biol. *5*, e310.

Bergland, A.O., Behrman, E.L., O'Brien, K.R., Schmidt, P.S., and Petrov, D.A. (2014). Genomic Evidence of Rapid and Stable Adaptive Oscillations over Seasonal Time Scales in Drosophila. PLoS Genet. *10*, e1004775.

Bienert, S., Waterhouse, A., De Beer, T.A.P., Tauriello, G., Studer, G., Bordoli, L., and Schwede, T. (2017). The SWISS-MODEL Repository-new features and functionality. Nucleic Acids Res. *45*, D313–D319.

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. Genome Res. *14*, 988–995.

Burley, S.K., Berman, H.M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., Christie, C., Dalenberg, K., Duarte, J.M., Dutta, S., et al. (2019). RCSB Protein Data Bank: Biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res. *47*, D464–D474.

Butterwick, J.A., del Mármol, J., Kim, K.H., Kahlson, M.A., Rogow, J.A., Walz, T., and Ruta, V. (2018). Cryo-EM structure of the insect olfactory receptor Orco. Nature *560*, 447–452.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

Charlesworth, J., and Eyre-Walker, A. (2008). The McDonald-Kreitman test and slightly deleterious

536 mutations. Mol. Biol. Evol. *25*, 1007–1015.
537 Clarke, D., Sethi, A., Li, S., Kumar, S., Chang, R.W.F., Chen, J., and Gerstein, M. (2016). Identifying
538 Allosteric Hotspots with Dynamics: Application to Inter- and Intra-species Conservation. Structure *24*,
539 826–837.
540 Dasmeh, P., Serohijos, A.W.R., Kepp, K.P., and Shakhnovich, E.I. (2013). Positively Selected Sites in
541 Cetacean Myoglobins Contribute to Protein Stability. PLoS Comput. Biol. *9*.
542 Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O., and Arnold, F.H. (2005). Why highly expressed
543 proteins evolve slowly. Proc. Natl. Acad. Sci. U. S. A. *102*, 14338–14343.
544 Echave, J., Spielman, S.J., and Wilke, C.O. (2016). Causes of evolutionary rate variation among protein
545 sites. Nat. Rev. Genet. *17*, 109–121.
546 Enard, D., Cai, L., Gwennap, C., and Petrov, D.A. (2016). Viruses are a dominant driver of protein
547 adaptation in mammals. Elife *5*, e12469.
548 Fabian, D.K., Kapun, M., Nolte, V., Kofler, R., Schmidt, P.S., Schlötterer, C., and Flatt, T. (2012).
549 Genome-wide patterns of latitudinal differentiation among populations of Drosophila melanogaster from
550 North America. Mol Ecol *21*, 4748–4769.
551 Gabdoulline, R.R., and Wade, R.C. (1999). On the protein-protein diffusional encounter complex. J. Mol.
552 Recognit. *12*, 226–234.
553 Goldman, N., Thorne, J.L., and Jones, D.T. (1998). Assessing the impact of secondary structure and
554 solvent accessibility on protein evolution. Genetics *149*, 445–458.
555 Hill, C.A., Fox, A.N., Pitts, R.J., Kent, L.B., Tan, P.L., Chrystal, M.A., Cravchik, A., Collins, F.H.,
556 Robertson, H.M., and Zwiebel, L.J. (2002). G Protein-Coupled Receptors in Anopheles gambiae. Science
557 (80-.). *298*, 176–178.
558 Huang, W., Massouras, A., Inoue, Y., Peiffer, J., Ràmia, M., Tarone, A.M., Turlapati, L., Zichner, T.,
559 Zhu, D., Lyman, R.F., et al. (2014). Natural variation in genome architecture among 205 Drosophila
560 melanogaster Genetic Reference Panel lines. Genome Res. *24*, 1193–1208.
561 Hughes, A.L., and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex
562 class I loci reveals overdominant selection. Nature *335*, 167–170.
563 Jiggins, F.M., and Kim, K.W. (2007). A screen for immunity genes evolving under positive selection in
564 Drosophila. J. Evol. Biol. *20*, 965–970.
565 Jordan, I.K., Wolf, Y.I., and Koonin, E. V. (2003). No simple dependence between protein evolution rate
566 and the number of protein-protein interactions: Only the most prolific interactors tend to evolve slowly.
567 BMC Evol. Biol. *3*.
568 Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates,
569 R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold.

570 Nature *596*, 583–589.
571 Keightley, P.D., and Eyre-Walker, A. (2012). Estimating the rate of adaptive molecular evolution when
572 the evolutionary divergence between species is small. J. Mol. Evol. *74*, 61–68.
573 Kim, P.M., Lu, L.J., Xia, Y., and Gerstein, M.B. (2006). Relating three-dimensional structures to protein
574 networks provides evolutionary insights. Science (80-.). *314*, 1938–1941.
575 Kolaczkowski, B., Kern, A.D., Holloway, A.K., and Begun, D.J. (2011). Genomic Differentiation
576 Between Temperate and Tropical Australian Populations of Drosophila melanogaster. Genetics *187*, 245–
577 260.
578 Kondo, S., Vedanayagam, J., Mohammed, J., Eizadshenass, S., Kan, L., Pang, N., Aradhya, R., Siepel,
579 A., Steinhauer, J., and Lai, E.C. (2017). New genes often acquire male- specific functions but rarely
580 become essential in Drosophila. 1841–1846.
581 Kosiol, C., Vinař, T., da Fonseca, R.R., Hubisz, M.J., Bustamante, C.D., Nielsen, R., and Siepel, A.
582 (2008). Patterns of Positive Selection in Six Mammalian Genomes. PLoS Genet. *4*, e1000144.
583 Kulmanov, M., and Hoehndorf, R. (2020). DeepGOPlus: Improved protein function prediction from
584 sequence. Bioinformatics *36*, 422–429.
585 Lack, J.B., Cardeno, C.M., Crepeau, M.W., Taylor, W., Corbett-Detig, R.B., Stevens, K.A., Langley,
586 C.H., and Pool, J.E. (2015). The drosophila genome nexus: A population genomic resource of 623
587 Drosophila melanogaster genomes, including 197 from a single ancestral range population. Genetics *199*,
588 1229–1241.
589 Langley, C.H., Stevens, K., Cardeno, C., Lee, Y.C.G., Schrider, D.R., Pool, J.E., Langley, S.A., Suarez,
590 C., Corbett-Detig, R.B., Kolaczkowski, B., et al. (2012). Genomic Variation in Natural Populations of
591 Drosophila melanogaster. Genetics *192*, 533–598.
592 Larracuente, A.M., Sackton, T.B., Greenberg, A.J., Wong, A., Singh, N.D., Sturgill, D., Zhang, Y.,
593 Oliver, B., and Clark, A.G. (2008). Evolution of protein-coding genes in Drosophila. Trends Genet. *24*,
594 114–123.
595 Lawniczak, M.K.N., and Begun, D.J. (2007). Molecular population genetics of female-expressed mating-
596 induced serine proteases in Drosophila melanogaster. Mol. Biol. Evol. *24*, 1944–1951.
597 Lazzaro, B.P., Sceurman, B.K., and Clark, A.G. (2004). Genetic basis of natural variation in D.
598 melanogaster antibacterial immunity. Science *303*, 1873–1876.
599 Leader, D.P., Krause, S.A., Pandit, A., Davies, S.A., and Dow, J.A.T. (2018). FlyAtlas 2: A new version
600 of the Drosophila melanogaster expression atlas with RNA-Seq, miRNA-Seq and sex-specific data.
601 Nucleic Acids Res. *46*, D809–D815.
602 Levin, T.C., and Malik, H.S. (2017). Rapidly evolving Toll-3/4 genes encode male-specific Toll-like
603 receptors in drosophila. Mol. Biol. Evol. *34*, 2307–2323.

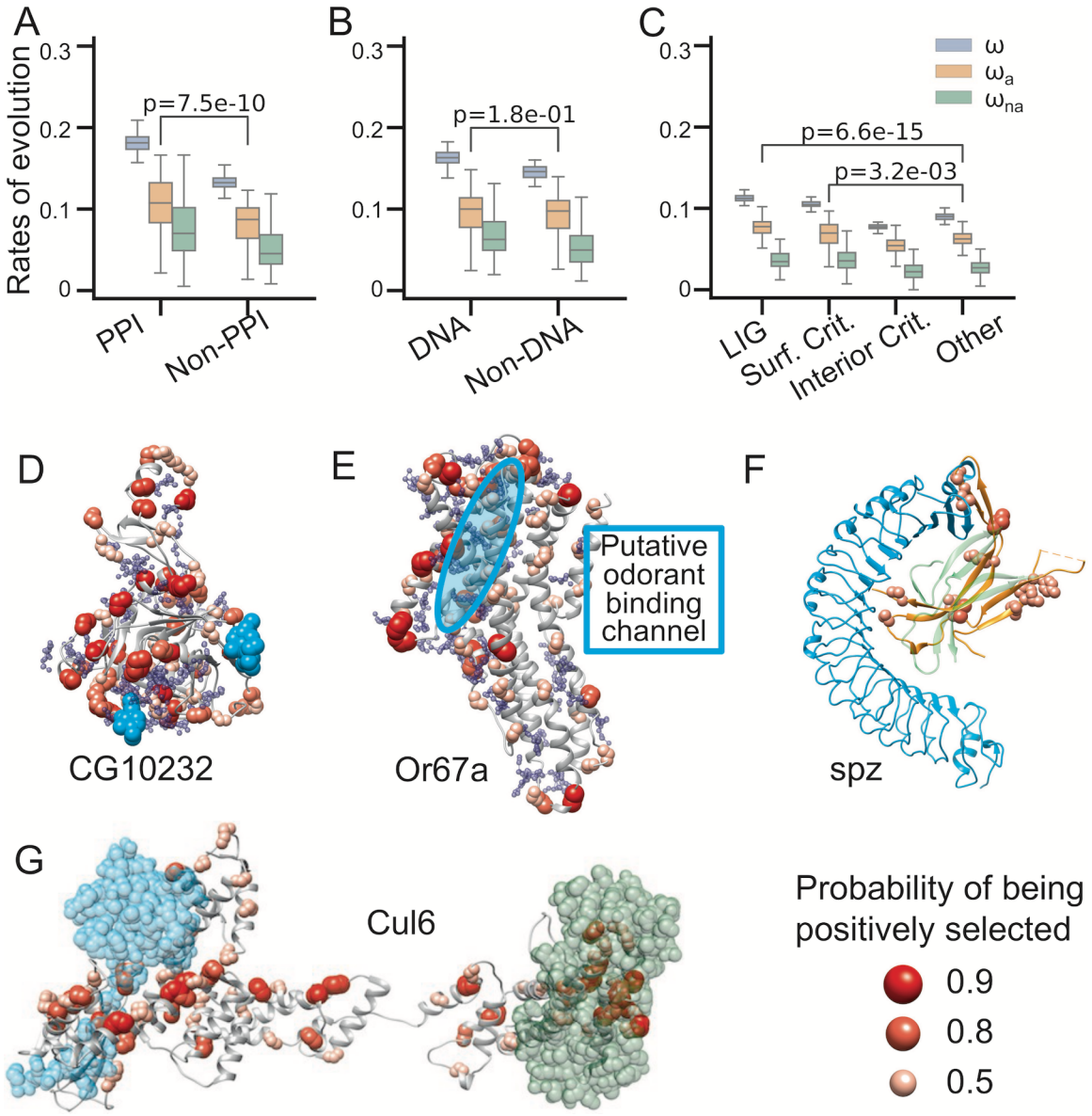
604 Lima, L.F., Torres, A.Q., Jardim, R., Mesquita, R.D., and Schama, R. (2021). Evolution of Toll, Spatzle
605 and MyD88 in insects: the problem of the Diptera bias. BMC Genomics *22*.
606 Lin, Y.S., Hsu, W.L., Hwang, J.K., and Li, W.H. (2007). Proportion of solvent-exposed amino acids in a
607 protein and rate of protein evolution. Mol. Biol. Evol. *24*, 1005–1011.
608 Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J.,
609 Jordan, G., Mauceli, E., et al. (2011). A high-resolution map of human evolutionary constraint using 29
610 mammals. Nature *478*, 476–482.
611 Löytynoja, A. (2014). Phylogeny-aware alignment with PRANK. Methods Mol. Biol. *1079*, 155–170.
612 Markova-Raina, P., and Petrov, D. (2011). High sensitivity to aligner and high rate of false positives in
613 the estimates of positive selection in the 12 Drosophila genomes. Genome Res.
614 Matthey-Doret, R., and Whitlock, M.C. (2019). Background selection and FST: Consequences for
615 detecting local adaptation. Mol. Ecol. *28*, 3902–3914.
616 McBride, C.S. (2007). Rapid evolution of smell and taste receptor genes during host specialization in
617 Drosophila sechellia. Proc. Natl. Acad. Sci. U. S. A.
618 Messer, P.W., and Petrov, D.A. (2013). Frequent adaptation and the McDonald-Kreitman test. Proc. Natl.
619 Acad. Sci. U. S. A. *110*, 8615–8620.
620 Mintseris, J., and Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein-
621 protein interactions. Proc. Natl. Acad. Sci. U. S. A. *102*, 10930–10935.
622 Moutinho, A.F., Trancoso, F.F., Dutheil, J.Y., and Zhang, J. (2019). The Impact of Protein Architecture
623 on Adaptive Evolution. Mol. Biol. Evol. *36*, 2013–2028.
624 Mukherjee, S., and Zhang, Y. (2011). Protein-protein complex structure predictions by multimeric
625 threading and template recombination. Structure *19*, 955–966.
626 Necci, M., Piovesan, D., Hoque, M.T., Walsh, I., Iqbal, S., Vendruscolo, M., Sormanni, P., Wang, C.,
627 Raimondi, D., Sharma, R., et al. (2021). Critical assessment of protein intrinsic disorder prediction. Nat.
628 Methods *18*, 472–481.
629 Newville, M., and Stensitzki, T. (2018). Non-Linear Least-Squares Minimization and Curve-Fitting for
630 Python. Non-Linear Least-Squares Minimization and Curve-Fitting for Python
631 https://doi.org/10.5281/zenodo.11813.
632 Nielsen, R., Bustamante, C., Clark, A.G., Glanowski, S., Sackton, T.B., Hubisz, M.J., Fledel-Alon, A.,
633 Tanenbaum, D.M., Civello, D., White, T.J., et al. (2005). A scan for positively selected genes in the
634 genomes of humans and chimpanzees. PLoS Biol. *3*, e170.
635 Obbard, D.J., Welch, J.J., Kim, K.-W., and Jiggins, F.M. (2009). Quantifying Adaptive Evolution in the
636 Drosophila Immune System. PLoS Genet. *5*, e1000698.
637 Pál, C., Papp, B., and Hurst, L.D. (2001). Highly expressed genes in yeast evolve slowly. Genetics *158*,

638 927–931.
639 Palmer, W.H., Hadfield, J.D., and Obbard, D.J. (2018). RNA-interference pathways display high rates of
640 adaptive protein evolution in multiple invertebrates. Genetics *208*, 1585–1599.
641 Pazos, F., and Valencia, A. (2008). Protein co-evolution, co-adaptation and interactions. EMBO J. *27*,
642 2648–2655.
643 Pieper, U., Webb, B.M., Barkan, D.T., Schneidman-Duhovny, D., Schlessinger, A., Braberg, H., Yang,
644 Z., Meng, E.C., Pettersen, E.F., Huang, C.C., et al. (2011). ModBase, a database of annotated comparative
645 protein structure models, and associated resources. Nucleic Acids Res. *39*, D465–D474.
646 Pitchers, W., Pool, J.E., and Dworkin, I. (2013). Altitudinal clinal variation in wing size and shape in
647 african drosophila melanogaster: One cline or many? Evolution (N. Y). *67*, 438–452.
648 Ramsey, D.C., Scherrer, M.P., Zhou, T., and Wilke, C.O. (2011). The relationship between relative
649 solvent accessibility and evolutionary rate in protein evolution. Genetics *188*, 479–488.
650 Re, S., Oshima, H., Kasahara, K., Kamiya, M., and Sugita, Y. (2019). Encounter complexes and hidden
651 poses of kinaseinhibitor binding on the free-energy landscape. Proc. Natl. Acad. Sci. U. S. A. *116*,
652 18404–18409.
653 Reinhardt, J.A., Kolaczkowski, B., Jones, C.D., Begun, D.J., and Kern, A.D. (2014). Parallel geographic
654 variation in Drosophila melanogaster. Genetics *197*, 361–373.
655 Sackton, T.B., Lazzaro, B.P., Schlenke, T.A., Evans, J.D., Hultmark, D., and Clark, A.G. (2007).
656 Dynamic evolution of the innate immune system in Drosophila. Nat. Genet. *39*, 1461–1468.
657 Schott, R.K., Refvik, S.P., Hauser, F.E., López-Fernández, H., and Chang, B.S.W. (2014). Divergent
658 positive selection in rhodopsin from lake and riverine cichlid fishes. Mol. Biol. Evol. *31*, 1149–1165.
659 Schrider, D.R., and Kern, A.D. (2018). Supervised Machine Learning for Population Genetics: A New
660 Paradigm. Trends Genet. *34*, 301–312.
661 Shao, Y., Chen, C., Shen, H., He, B.Z., Yu, D., Jiang, S., Zhao, S., Gao, Z., Zhu, Z., Chen, X., et al.
662 (2019). GenTree, an integrated resource for analyzing the evolution and function of primate-specific
663 coding genes. Genome Res. *29*, 682–696.
664 Sironi, M., Cagliani, R., Forni, D., and Clerici, M. (2015). Evolutionary insights into host-pathogen
665 interactions from mammalian sequence data. Nat. Rev. Genet. *16*, 224–236.
666 Slodkowicz, G., and Goldman, N. (2020). Integrated structural and evolutionary analysis reveals common
667 mechanisms underlying adaptive evolution in mammals. Proc. Natl. Acad. Sci. *117*, 5977–5986.
668 Smith, N.G.C., and Eyre-Walker, A. (2002). Adaptive protein evolution in Drosophila. Nature *415*, 1022–
669 1024.
670 Song, Y., Dimaio, F., Wang, R.Y.R., Kim, D., Miles, C., Brunette, T., Thompson, J., and Baker, D.
671 (2013). High-resolution comparative modeling with RosettaCM. Structure *21*, 1735–1742.

672 Storz, J.F., and Kelly, J.K. (2008). Effects of Spatially Varying Selection on Nucleotide Diversity and
673 Linkage Disequilibrium: Insights From Deer Mouse Globin Genes. Genetics *180*, 367–379.
674 Svetec, N., Cridland, J.M., Zhao, L., and Begun, D.J. (2016). The Adaptive Significance of Natural
675 Genetic Variation in the DNA Damage Response of Drosophila melanogaster. PLoS Genet. *12*,
676 e1005869.
677 Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva,
678 N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: Protein-protein association networks with
679 increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic
680 Acids Res. *47*, D607–D613.
681 Tang, C., Iwahara, J., and Clore, G.M. (2006). Visualization of transient encounter complexes in protein-
682 protein association. Nature *444*, 383–386.
683 Thurmond, J., Goodman, J.L., Strelets, V.B., Attrill, H., Gramates, L.S., Marygold, S.J., Matthews, B.B.,
684 Millburn, G., Antonazzo, G., Trovisco, V., et al. (2019). FlyBase 2.0: The next generation. Nucleic Acids
685 Res. *47*, D759–D765.
686 Toda, Y., Ko, M.C., Liang, Q., Miller, E.T., Rico-Guevara, A., Nakagita, T., Sakakibara, A., Uemura, K.,
687 Sackton, T., Hayakawa, T., et al. (2021). Early origin of sweet perception in the songbird radiation.
688 Science (80-.). *373*, 226–231.
689 Uricchio, L.H., Petrov, D.A., and Enard, D. (2019). Exploiting selection at linked sites to infer the rate
690 and strength of adaptation. Nat. Ecol. Evol. *3*, 977–984.
691 Uversky, V.N. (2019). Intrinsically disordered proteins and their "Mysterious" (meta)physics. Front.
692 Phys. *7*.
693 Wang, S., Li, W., Liu, S., and Xu, J. (2016). RaptorX-Property: a web server for protein structure
694 property prediction. Nucleic Acids Res. *44*, W430–W435.
695 Wu, D.D., Wang, G.D., Irwin, D.M., and Zhang, Y.P. (2009). A profound role for the expansion of
696 trypsin-like serine protease family in the evolution of hematophagy in mosquito. Mol. Biol. Evol. *26*,
697 2333–2341.
698 Yan, J., and Kurgan, L. (2017). DRNApred, fast sequence-based method that accurately predicts and
699 discriminates DNA- and RNA-binding residues. Nucleic Acids Res. *45*, gkx059.
700 Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-
701 Saban, S., Safran, M., Domany, E., et al. (2005). Genome-wide midrange transcription profiles reveal
702 expression level relationships in human tissue specification. Bioinformatics *21*, 650–659.
703 Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–
704 1591.
705 Yang, Y., Gao, J., Wang, J., Heffernan, R., Hanson, J., Paliwal, K., and Zhou, Y. (2018). Sixty-five years

706 of the long march in protein secondary structure prediction: The final stretch? Brief. Bioinform. *19*, 482–
707 494.
708 Yang, Z., Wong, W.S.W., and Nielsen, R. (2005). Bayes empirical Bayes inference of amino acid sites
709 under positive selection. Mol. Biol. Evol. *22*, 1107–1118.
710 Zhang, J., and He, X. (2005). Significant impact of protein dispensability on the instantaneous rate of
711 protein evolution. Mol. Biol. Evol. *22*, 1147–1155.
712 Zhang, J., and Yang, J.R. (2015). Determinants of the rate of protein sequence evolution. Nat. Rev. Genet.
713 *16*, 409–420.
714 Zhang, Y.E., Vibranovski, M.D., Krinsky, B.H., and Long, M. (2010). Age-dependent chromosomal
715 distribution of male-biased genes in Drosophila. Genome Res. *20*, 1526–1533.
716 Zhao, L., Wit, J., Svetec, N., and Begun, D.J. (2015). Parallel Gene Expression Differences between Low
717 and High Latitude Populations of Drosophila melanogaster and D. simulans. PLOS Genet. *11*, e1005184.
718 Zheng, N., Schulman, B.A., Song, L., Miller, J.J., Jeffrey, P.D., Wang, P., Chu, C., Koepp, D.M., Elledge,
719 S.J., Pagano, M., et al. (2002). Structure of the Cul1-Rbx1-Skp1-F boxSkp2 SCF ubiquitin ligase
720 complex. Nature *416*, 703–709.
721

722 **Figure Legends**
723



724
725 **Fig. 1**. Adaptive evolution in molecular interaction sites. Protein-protein interaction sites (A), DNA
726 binding sites (B) and putative ligand binding sites (C) show higher adaptation rates than none binding
727 sites. The t-test p-values between the adaptation rates of binding sites and non-binding sites were
728 highlighted in (A), (B) and (C). T-test p-values between other evolutionary rates of binding sites and non-
729 binding sites were shown in Fig. S2. Examples of positive selection around molecular interaction sites in
730 high quality structural models of CG10232 (D), Or67a (E), spz (F), and Cul6 (G). Except for spz (PDB
731 code 3e07), the other proteins are obtained from SWISS model repository. Putative ligand binding
732 pockets of CG10232 (D) and Or67a (E) are shown in blue spheres. Ligands including interacting proteins
733 are shown in cyan or green: NAG of CG10232 in cyan (D), Toll receptor of spz in cyan (F), RING-box
734 protein in cyan and F-box protein in green for Cul6 (G). The putative odorant binding channel of Or67a is
735 highlighted in cyan circle (E). The ligand poses in (D, F and G) are obtained by superimposition from
736 structure 2XXL, 4BV4 and 1LDK, respectively.
737

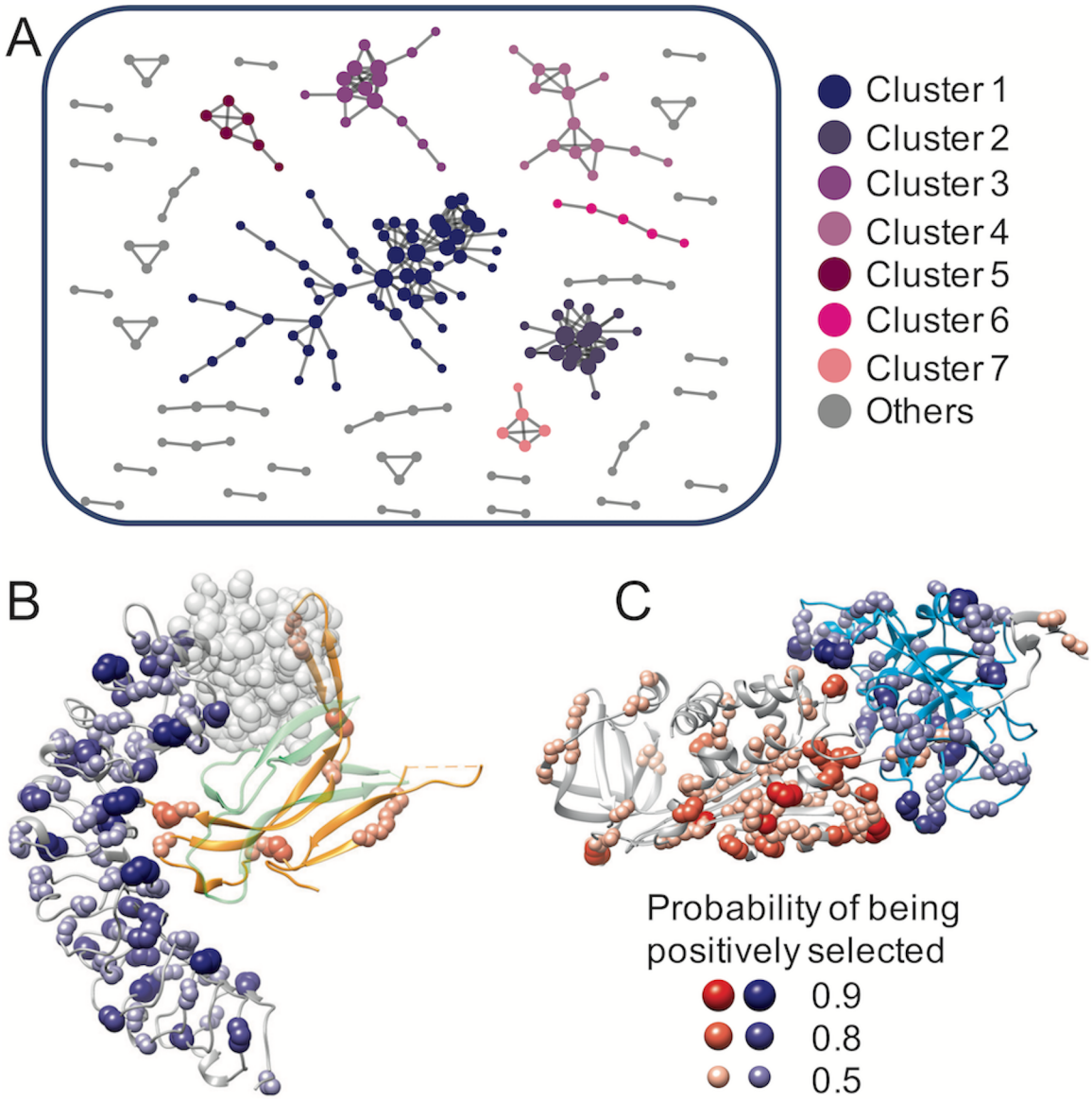


Fig. 2. Co-adaptation of fast-adaptive proteins. (A) Sub-clusters of PPI networks of fast-adaptive proteins. Only proteins with at least one partner were shown. Examples of molecular interactions that might regulate co-adaptation in fast-adaptive proteins: (B) Toll-4 (gray) and spz (orange, with green representing the other spz monomer), (C) Spn28Db (gray, serine protease inhibitor 28Db) and CG18563 (cyan, with Go term "serine-type endopeptidase activity"). A putative N-terminus (transparent beads) of Toll-4 were built by superimposition from 4LXR, since the N-terminus were missing in the structural model. Complex structural model of Spn28Db and CG18563 was inferred from 1EZX.

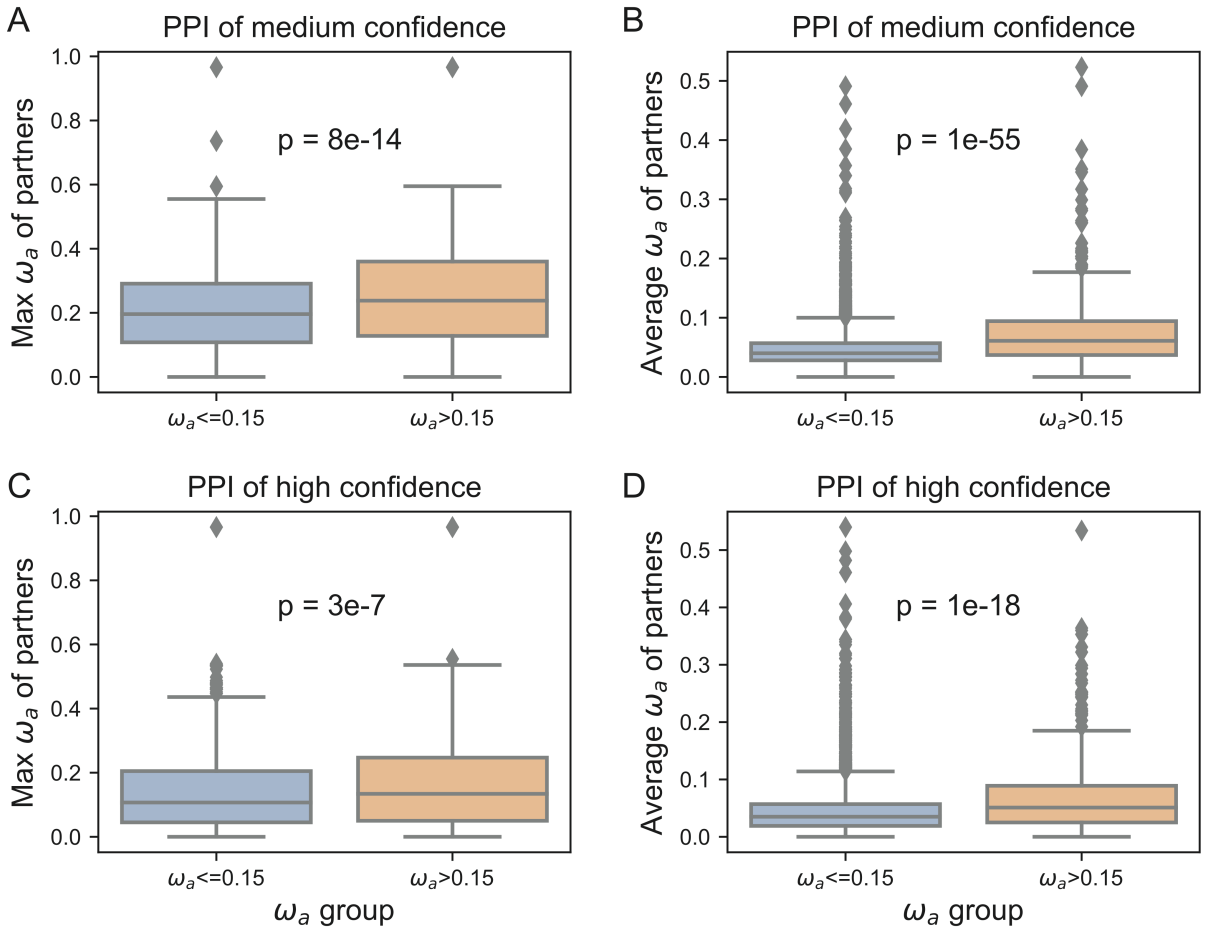


Fig. 3. Co-adaptation of PPIs in *D. melanogaster*. For fast-adaptive proteins, adaptation rates of their partners (orange box plot) are significantly larger compared to slow adaptive proteins (blue box plot). Max ω_a of protein partners are shown in (A and C) and averaged ω_a, of protein partners are shown in (B and D). PPI from STRING with median confidence (combined score larger than 0.4) are shown in (A and B), and PPI with high confidence (combined score larger than 0.7) are shown in (C and D).

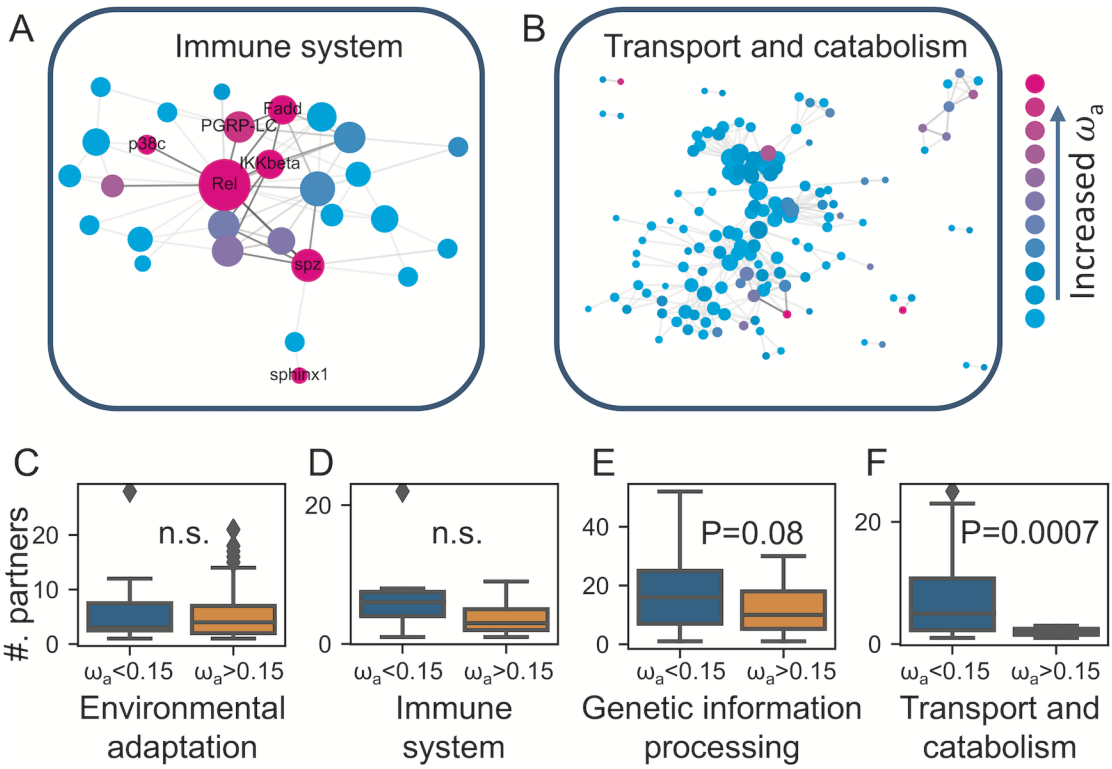


Fig. 4. Rates of protein sequence adaptive evolution in the PPI network of different functional pathways. The PPI networks showed the adaptive evolution in immune system (A) and transport and catabolism (B). In pathways that are hotspots of adaptive evolution, e.g., environmental adaptation (C) and immune system (D), fast-adaptive proteins can act as central nodes. While in conserved pathways, e.g., genetic information processing (E) and transport and catabolism (F), fast-adaptive proteins are often at the periphery of the PPI network.

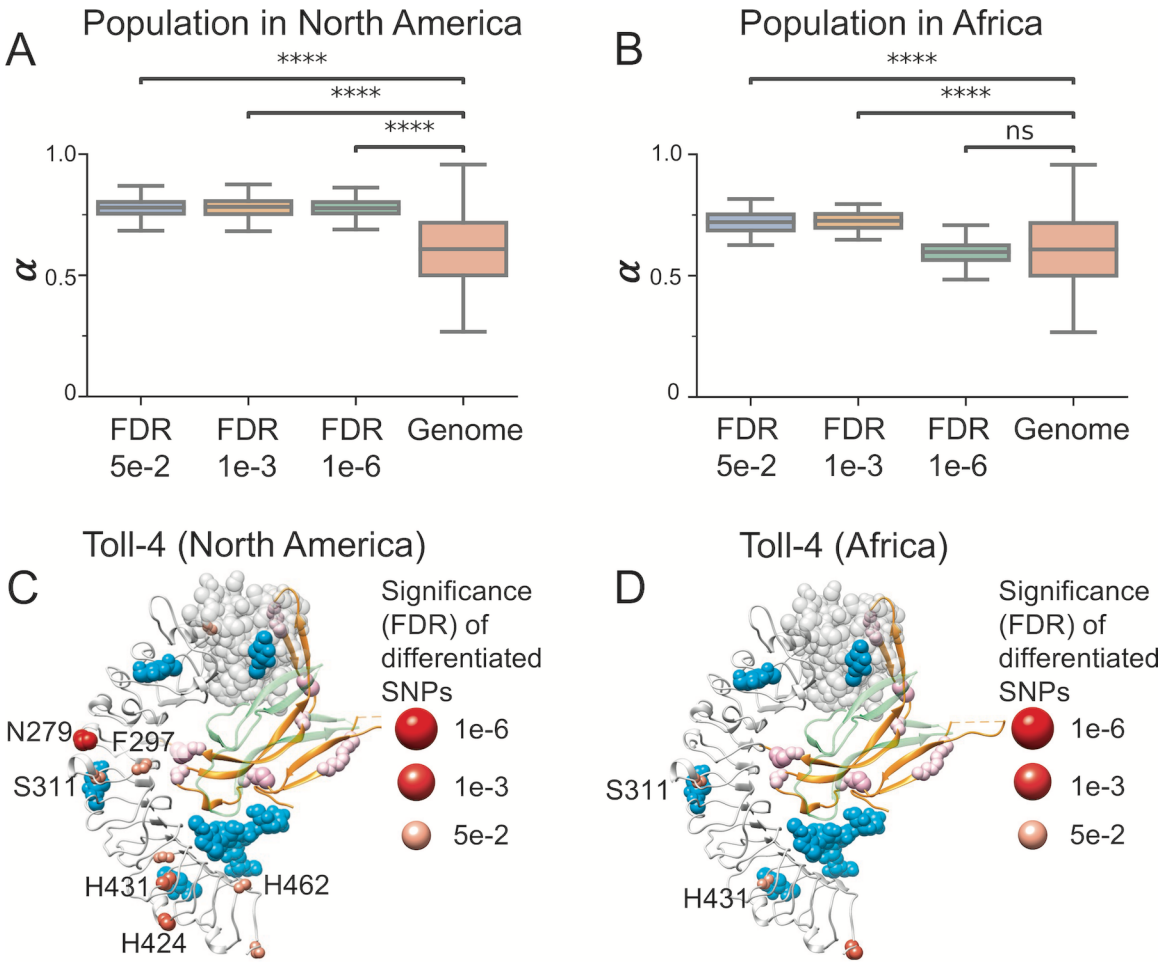


Fig. 5. Adaptive evolution in significantly differentiated SNPs. The significantly differentiated SNPs at different FDR cutoffs all show much higher proportions of adaptative changes (α) than genome-wide expectation in North American population (A) and African population (B) (t-test_ind, ****, p<1e-4). (C) Significant differentiated nonsynonymous SNPs of North American populations in Toll-4. Ligands are shown in cyan by superimposing crystal structure of Toll-Spatzle (PDB code 4BV4) on to Toll-4 structural model. Residues N279 and H431 are both highly differentiated (FDR 3e-7 and 3e-6) and positively selected (both at probability of p=0.9). Other highly differentiated sites, F297, S311, H424, H431 and H462 are located near ligand binding sites or positively selected sites. (D) Significantly differentiated nonsynonymous SNPs of African populations in Toll-4. Two highly differentiated SNPs, S311 (FDR 4e-2) and H431 (FDR 2e-2) exist in North American population and are located near ligand binding sites.