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

Junhui Peng, 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 an essential step in understanding the basis of function and evolution. Previous studies have shown that adaptation occurs frequently 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 that are associated with particular structural or functional characteristics across the genome, since many of the 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. Although adaptation is more relevant to function-related rather than structure-related properties, we observed that physical interactions may play a role in the co-adaptation of fast-adaptive proteins. Importantly, protein-protein and protein-DNA interaction sites are hotspots for protein adaptive evolution, regardless of the levels of intrinsic structural disorder or relative solvent accessibility. 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 are repeatedly mutated and selected in evolution, in the past, present, and maybe future. Our results suggest 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 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. It is now well established that 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;

35 Lindblad-Toh et al., 2011; Zhang and Yang, 2015). For example, genes that are highly expressed or
36 perform essential functions are under strong purifying selection and tend to evolve slowly (Drummond et
37 al., 2005; Moutinho et al., 2019; Pál et al., 2001; Zhang and He, 2005; Zhang and Yang, 2015); genes
38 involved in host-pathogen interactions, e.g., immune responses and antivirus responses, show
39 exceptionally high rates of adaptive changes (Enard et al., 2016; Nielsen et al., 2005; Obbard et al., 2009;
40 Palmer et al., 2018; Sackton et al., 2007; Sironi et al., 2015; Uricchio et al., 2019); and residues that are
41 intrinsically disordered or at the protein surface are fast evolving and has been proved to be hotspots of
42 adaptive evolution (Afanasyeva et al., 2018; Goldman et al., 1998; Lin et al., 2007; Moutinho et al., 2019;
43 Ramsey et al., 2011). More recently, Slodkowicz & Goldman (Slodkowicz and Goldman, 2020)
44 employed genomic-scale integrated structural and phylogenetic evolutionary analysis in mammals and
45 showed that positively selected residues are clustered near ligand binding sites, especially in proteins that
46 are associated with immune responses and xenobiotic metabolism. However, vast majority of the work
47 focused on differences at the species level, it is unclear how much of the polymorphic changes within a
48 species may contribute to long-term evolution.
49 Although evidence have shown that adaptation is more likely to occur at intrinsically disordered
50 regions and clustered at the surface of proteins, the functional properties of adaptation in the genomic and
51 population scale remains unclear. Moreover, due to lack of structural and functional information of many
52 proteins in the genome, the underlying mechanism derived from current studies might be incomplete.
53 Here, we systematically investigated the evolution and adaptation of protein-coding genes in *Drosophila*
54 *melanogaster* by comparing it to its closely related species and their own populations, in order to
55 distinguish the main factors that impact the evolution and adaption at the protein-coding level. We applied
56 large-scale bioinformatic and structural analysis to obtain structural and functional properties of proteins.
57 We then classified residues into different structural and functional sites. By comparing rates of sequence
58 evolution and adaptation between different proteins and different sites, we were able to locate hotspots of
59 adaptation at the genome scale. Although adaptation is more sensitive to functional properties rather than
60 structural properties, we found that putative binding regions including allosteric sites at protein surface
61 show higher rates of adaptation than other sites. For proteins that are under fast-adaptive evolution, we
62 showed that they tend to interact with each other more frequently than random expectations and are often
63 associated with reproduction, immunity, and environmental information processing in *D. melanogaster*.
64 In addition, we showed that interacting proteins in *D. melanogaster* might undergo co-adaptive evolution.
65 Furthermore, we hypothesize that molecular interactions or physical interactions might be an important
66 mechanism that contribute to the adaptive and co-adaptive evolution in *D. melanogaster* genome. At last,
67 we showed that many non-synonymous SNPs contributing to short-term adaptation are overlapped with

68 SNPs contributing to long-term adaptive evolution, suggesting that a subset of SNPs on the genomes are
69 constantly utilized for adaptive purpose.
70
71 **Results**
72 **Putative molecular interaction sites are hotspots for protein adaptive evolution**
73 To uncover the main factors that impact the evolutionary rates of genes, we analyzed 13,528 protein-
74 coding genes in *D. melanogaster* using genome data from *melanogaster* subgroup species and *D.*
75 *melanogaster* population genomics data from 205 inbred lines from Drosophila Genetic Reference Panel,
76 Freeze 2.0, DGRP2 (Huang et al., 2014). We applied a maximum likelihood method (Yang, 2007) to
77 compute dN/dS ratio (ω) using the protein-coding sequences of five closely related melanogaster
78 subgroup species (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba* and *D. erecta*). We estimated
79 the proportions of adaptive changes (α) in each gene by applying an extension of MK test named
80 asymptotic MK (Messer and Petrov, 2013; Uricchio et al., 2019) using *D. simulans* as outgroup. We then
81 calculated the rate of adaptive changes (ω_a) of each gene by multiplying ω to α ($\omega_a = \alpha\omega$) (Moutinho et
82 al., 2019) using *D. yakuba* as the outgroup species (See methods). The rate of nonadaptive changes can be
83 further calculated by $\omega_{na}=\omega-\omega_a$. Finally, we successfully assigned ω to 12,118 protein coding genes and
84 ω_a and ω_{na} to 7,192 genes. For each of *D. melanogaster* genes subjecting the same pipeline of analysis,
85 we further obtained 17 different structural or functional properties (see Methods and supplementary file
86 S1). We calculated Pearson's correlations of ω, ω_a and ω_{na} with all these properties (Table S1). We
87 showed that many of these genome-wide correlations were expected according to previous studies
88 (Supplement Information, section *Impact of gene properties on evolution of protein-coding genes in D.*
89 *melanogaster*, Table S1). Interestingly, among these properties, we found that some previously not
90 reported properties, fractions of molecular-interaction sites (PPI-site ratio, ratio of residues involved in
91 protein-protein interactions, and DNA-site ratio, ratio of residues involved in protein-DNA interactions)
92 strongly positively correlated with ω, ω_a and ω_{na} (Supplement Information, section *Molecular*
93 *interactions contribute to the variations of protein sequence evolution and adaptation*, Table S1, Figure
94 S1). The results indicate that molecular interactions might act as an important factor that drive protein
95 adaptive evolution in *Drosophila* genome.
96 We then investigate whether residues involved in molecular interactions are targets for adaptive
97 evolution. To tackle this question, we predicted protein-protein interaction sites (PPI-sites) and DNA
98 binding sites (DNA-sites) for each of *D. melanogaster* protein sequence (see Methods). In addition, we
99 characterized allosteric residues as surface and interior critical residues with STRESS model (Clarke et
100 al., 2016) for all the structural models. We also extracted putative binding sites from STRESS Monte
101 Carlo (MC) simulations. We calculated ω, ω_a and ω_{na} for residues in each of the putative molecular

102 interaction category. Strikingly, we observed that residues involved in protein-protein interactions, DNA
103 binding and ligand binding exhibited higher rates of adaptive evolution compared to their corresponding
104 null sites (Fig. 1A-C). In addition, allosteric residues at protein surface showed higher adaptation rates
105 than allosteric residues at protein interior or residues that are not involved in ligand binding (Fig. 1C).
106 Since we observed significant positive intercorrelations between PPI and DNA binding with ISD
107 (intrinsic structural disorder) and RSA (relative solvent accessibility) (Table S2), we next asked whether
108 the increase of ω_a in protein-protein interactions sites or DNA binding sites was caused by the increase of
109 disorder or site exposure. We calculated and compared ω, ω_a and ω_{na} for putative PPI and DNA binding
110 sites with different levels of ISD or RSA. Remarkably, we found that ω_a of these binding sites remains
111 similar among different levels of ISD or RSA (Fig. S5AC). The results suggest that PPI or DNA binding
112 events in proteins can result in elevated adaptation rates regardless their structural disorder or site
113 exposure. While for residues that are not associated with putative PPI or DNA binding, we also observed
114 increase in ω_a when increasing ISD or RSA (Fig. S5BD), which could be the result of some other yet
115 unknown underlying mechanisms. In addition, there is possibility that binding sites in disordered regions
116 are not well-predicted. However, given that ISD does not show strong impact to binding sites (Fig.
117 S5AC), we think the inaccuracy of binding sites may not play a significant role.
118 In order to gain better understanding of adaptation in molecular interaction sites, we further
119 visualized positive selections that are associated with molecular interactions. We first investigated
120 whether adaptive evolution is associated with particular protein structures or protein families. To do this,
121 we looked into fast-adaptive proteins with the largest ~15% rates of adaptation ($\omega_a > 0.15$) that are linked
122 to high quality structural models. Interestingly, among these proteins, we found 45 enriched as trypsin-
123 like cysteine/serine peptidase domain and 17 7TM chemoreceptors, suggesting widespread adaptive
124 evolution acting on these protein families or protein domains in *D. melanogaster* (Table S3). Many of the
125 7TM chemoreceptors are olfactory and gustatory genes and show adaptive evolution in various species
126 such as *Drosophila* and mosquito (Hill et al., 2002; Lawniczak and Begun, 2007; McBride, 2007; Wu et
127 al., 2009). In addition to these two protein families, previous studies identified recurrent positive
128 selections acting on some other fast-adaptive proteins in *Drosophila* and mammals, and the possible
129 adaptive evolution mechanisms have been linked to exogenous ligand binding, for example, serine
130 protease inhibitors (serpin), Toll-like receptor 4 (TLR-4), and cytochrome P450 (Jiggins and Kim, 2007;
131 Slodkowicz and Goldman, 2020).
132 In order to visualize the link between adaptive evolution and molecular interactions in the two
133 protein families with frequent adaptive evolution, we showed significant positive selections and
134 molecular interactions in two representatives: CG10232 and Or67a, each for trypsin-like cysteine/serine
135 peptidase domain and 7TM chemoreceptors, respectively. We observed that in both cases, positively

136 selected sites highly overlapped with predicted or inferred binding pockets (Fig. 1D-E). Specifically, in
137 CG10232, we found clusters of positive selected sites around NAG binding sites that are inferred from a
138 crystal structure of serine protease (PDB code: 2XXL) (Fig. 1D), while in Or67a, positively selected sites
139 expand around the putative odorant binding channel formed by helices S1-S6 in extracellular regions
140 (Butterwick et al., 2018) (Fig. 1E).
141 Except for these examples that are associated with exogenous ligand or exogenous peptide
142 binding, we also identified two previously not described examples where adaptive evolution might be
143 linked to endogenous protein binding: Spaztle (spz, Fig. 1F) and Cul6 (Fig. 1G). Spaztle can bind to Toll-
144 like receptors (TLR) and trigger humoral innate immune response. We built the missing loop in Spaztle in
145 the crystal structure of Toll/Spaztle complex (PDB code 4BV4) according to the dimeric crystal structure
146 of Spaztle (PDB code 3E07). In this complex structural model, we observed several positively selected
147 sites in Toll-4/Spaztle interfaces (Fig. 1F). Cul6, another example, is a protein in cullins family in *D.*
148 *melanogaster*. The cullins protein family are known as scaffold proteins that assemble multi-subunit
149 Cullin-RING E3 ubiquitin ligase by forming SCF complex with F box and RING-box (Rbx) proteins
150 (Zheng et al., 2002). We constructed the putative Cul6 contained SCF complex by superimposition to the
151 crystal structure of the Cul1-Rbx1-Skp1-F box^{Skp2} SCF ubiquitin ligase complex (Zheng et al., 2002). In
152 the structural model, we observed positive selected sites in Cul6 clustered around the binding sites of
153 RING-box protein, Rbx1, and F-box protein, Skp1 (Fig. 1G).
154
155 **Frequent adaptive evolution and co-adaptative evolution in genes involved in reproduction,**
156 **immune system, and environmental information processing**
157 To find out whether specific biological functions were associated with fast-adaptive genes, we applied
158 DAVID Go analysis with genes that have largest ~15% rates of adaptation ($\omega_a > 0.15$). The significant Go
159 terms are frequently linked to serine-type endopeptidase activity, reproduction, protein lysis,
160 chemosensory and other related biological functions (Table S4). As these fast-adaptive genes tend to be
161 enriched in similar biological functions, we asked whether these genes are evolved co-adaptively, i.e.,
162 whether these proteins are interacting with each other frequently. To test this possibility, we obtained PPI
163 of *D. melanogaster* from STRING database (Szklarczyk et al., 2019) and analyzed protein-protein
164 interactions among fast-adaptive proteins. We found that fast-adaptive proteins tend to interact with each
165 other more frequently than expected (PPI enrichment p-value < 1.0e-16). In the PPI network of fast-
166 adaptive proteins, we observed 7 strongly connected sub-clusters with at least 5 members (Fig. 2A, Table
167 S5). Proteins in these sub-clusters are enriched in biological processes such as reproduction, immune
168 response, defense response to bacterium and virus, RNA interference, chitin metabolic, etc., which are in
169 line with the Go analysis of fast-adaptive genes (Table S6-S11).

170 We next asked whether co-adaptation plays a role in the adaptive evolution of interacting proteins
171 to a broader extend, including both fast- and slow-adaptive proteins. To address this question, we
172 analyzed and compared adaptation rates of all D. *melanogaster* PPIs available in STRING database with
173 high confidence and we found that protein partners of fast-adaptive proteins (ω_a>0.15) have significantly
174 larger maximum/average ω_a compared to slow-adaptive proteins (Figure 3). We further analyzed and
175 visualized adaptive evolutionary rates of proteins in PPI networks of 9 different biological pathways
176 extracted from KEGG pathways, including immune system, xenobiotics biodegradation, response to
177 environment, aging and development, genetic information processing, sensory system, transport and
178 catabolism, cell growth and death and metabolism. We observed that, in these PPI networks, proteins with
179 relatively large ω_a tend to interact with each other (Figure 4AB). We also noticed that, for pathways that
180 are previously known as adaptation-hotspots, e.g., immune system, fast-adaptive proteins can act as
181 central nodes and are co-adaptively evolved with other fast-adaptive proteins (Figure 4AC). While in
182 pathways such as transport and catabolism, fast-adaptive proteins are mainly at PPI periphery. In line with
183 these findings, we found that ω_a are larger in pathways that harbor fast-adaptive proteins as central nodes
184 than other pathways (Figure S6).
185 ***Physical interactions contribute to co-adaptation of fast-adaptive genes.*** Having established that
186 molecular interactions contribute to adaptive evolution of protein sequence, we then investigated whether
187 these physical molecular interactions could drive protein-protein co-adaptation. To do this, we looked into
188 interacting fast-adaptive protein pairs that are associated known or inferred complex structural models.
189 For inferred complex structural models, we superimposed the structural models of the pair of proteins
190 onto their high resolution homologous complex structures. Here we observed and illustrated co-adaptation
191 at PPI interface in two examples: Toll-4/Spatzle and Spn28Db/CG18563 (Fig. 2BC).
192 ***Toll-4/Spatzle***. Toll-4 is a member of toll-like receptors. Previous studies have shown strong evidence of
193 adaptive evolution of Toll-4 in *Drosophila* and mammals (Levin and Malik, 2017; Slodkowicz and
194 Goldman, 2020). Toll-4 can bind to Spatzle and trigger further innate immune responses with high
195 confidence (inferred from STRING database). In the previous section, we showed that several positively
196 selected sites in Spatzle overlap with Toll-Spatzle interfaces (Fig. 1F). Here, we further showed that, in
197 Toll-4, considerable number of significant positively selected sites were located at interface for Spatzle
198 (Fig. 2B), which is in line with a previous study of Toll-4 in *D. willistoni* (Levin and Malik, 2017).
199 ***Spn28Db/CG18563***. Spn28Db is one of the serine protease inhibitors in *D. melanogaster* that are
200 expressed in male accessory glands, while CG18563 belongs to the protein family of trypsin-like
201 cysteine/serine peptidase domain. The interactions between the two proteins were predicted with high
202 confidence from STRING database, and the molecular interactions can be inferred from existing crystal
203 structure of serpin and bacteria protease complex (PDB code 1EZX). We observed many positive

204 selected sites at the molecular interface between the two proteins (Fig. 2C), suggesting that physical
205 interactions might play a role in the co-adaptation of the two proteins.
206
207 **Most clinally differentiated non-synonymous SNPs in protein-coding genes are adaptive**
208 To find out the relations between short-term adaptation to local environments and long-term adaptive
209 evolution, we extracted residues with significant F_{ST} SNPs from clinal variations (Svetec et al., 2016). We
210 then computed evolutionary rates (ω), adaptation rates (ω_a) and non-adaptation rates (ω_{na}) of these
211 residues as in previous section. We observed that these residues have much higher ratio of adaptation
212 rates over non-adaptation rates than genome-wide random expectations (Fig. 5A), suggesting that these
213 residues have higher proportions of adaptive changes, and that they can be hotspots for adaptive
214 evolution. To find out whether these SNPs are related with even longer-term adaptive evolution, we
215 inferred positive selection sites of each protein-coding gene from phylogenic data (see Methods). We
216 found that the non-synonymous F_{ST} SNPs are significantly enriched in long-term positive selections
217 (Table S12- S13). To further characterize structural and functional properties of short-term genetic
218 variations, we mapped significant nonsynonymous F_{ST} residues to different structural and functional
219 characteristics, such as ISD, RSA, PPI-sites, DNA-sites and ligand-binding sites. We found that these
220 non-synonymous SNPs were enriched in disordered regions and protein surfaces and were significantly
221 more likely to be involved in protein-protein interactions and ligand-binding than expectation (Table S14-
222 S18). To better visualize the characteristics of these SNPs, we used Toll-4 as an example. We mapped
223 significant non-synonymous F_{ST} SNPs in Toll-4 on to its structural model. We showed that F_{ST} SNPs are
224 either positively selected or being very close to positively selected sites (Fig. 5BC). For example, highly
225 differentiated sites, N279 (FDR 3e-7) and H431 (FDR 3e-6) were predicted to be positively selected both
226 at probability at p=0.9. While another highly differentiated site, D424 was close to three positively
227 selected sites S401 (p=0.8), H431 (p=0.95) and V448 (p=0.8). We also noticed some differentiated sites
228 that may be located within ligand binding sites, including F297 (FDR 3e-3), S311 (FDR 3e-3), H431
229 (FDR 3e-6) and H462 (FDR 1e-2).
230
231 **Discussion**
232 In this study, we systematically studied the impact of structure- and function-related gene properties on
233 protein sequence evolution and adaptation in *D. melanogaster* genome. We found that molecular
234 interactions in proteins contribute to the variation of protein sequence adaptive evolution. A novel
235 discovery of this work is that molecular interaction sites including protein-protein interaction sites and
236 protein-DNA interaction sites are hotspots for adaptative evolution. We revealed that fast-adaptive
237 proteins tend to interact with each other frequently and protein partners of these fast-adaptive proteins

238 tend to have higher adaptation rates, suggesting that co-adaptive evolution might be common in *D.*
239 *melanogaster*. By looking at interacting fast-adaptive proteins, we further demonstrated that physical
240 interactions may contribute to the mechanisms of co-adaptative evolution of fast-adaptive proteins.
241 Although our results are in agreement with previous studies on the factors driving protein
242 sequence evolution (Zhang and Yang, 2015), we showed some complex correlations between ω, ω_a and
243 ω_{na} and protein length and male specificity (Supplement information, section *Complex correlations of*
244 *protein length and male expression level with protein evolutionary rates*, Fig. S2-S4, supplement file S2).
245 These complex correlations suggest caveat exists when we looked at protein length and gene expression
246 levels. For example, gene expression level was proved to be a major determinant (Zhang and Yang, 2015)
247 through mechanisms such as the pressure for translational robustness, i.e., robustness to translational
248 missense errors (Drummond et al., 2005). Previous studies have revealed that male biased or female
249 biased genes can be fast evolving (Yang et al., 2016). While on the other hand, many male biased genes
250 can be highly expressed in testis, which results in a complex correlation between protein sequence
251 evolutionary rate and male expression level or even mean expression level of *D. melanogaster*. The
252 unique evolutionary property of these male biased or specific genes could be caused by the unique
253 transcriptional scanning mechanism in testis (Xia et al., 2020). We propose that tissue specificity might be
254 a better quantity when considering the impact of gene expression profile on protein sequence evolution in
255 *D. melanogaster*. In addition to male expression level, a similar complex correlation was observed for
256 protein length. It has been the notion that short proteins tend to evolve faster than long proteins, which
257 may be biologically relevant or byproduct of other factors such as selection on buried and exposed sites
258 (Moutinho et al., 2019). Here, we demonstrated that, in *D. melanogaster*, although protein length is
259 strongly negatively correlated with protein sequence evolutionary rate, genes that have the slowest
260 evolutionary rates tend to be relatively short. This could be caused by the fact that under essential
261 functional constraint, genes can undergo strong purifying selections, while essential genes such as
262 secreted proteins are constrained to be smaller, and that essential genes could be shorter than other genes
263 (Chen et al., 2020).
264 Protein surface and intrinsic disorder regions are frequent targets for adaptive evolution and
265 contribute to the variations of protein sequence adaptive evolution (Afanasyeva et al., 2018; Moutinho et
266 al., 2019), however, the detailed mechanisms underlying these observations remains unclear. One
267 possible explanation would be that these regions are frequently linked to intermolecular interactions
268 (Afanasyeva et al., 2018; Moutinho et al., 2019). For example, Moutinho et al hypothesized that
269 molecular interactions involved in host-pathogen coevolution were the major driver of protein adaptation
270 (Moutinho et al., 2019). Here, we further identified that proportions of possible molecular interaction sites
271 inside proteins contribute to the variations of protein sequence adaptive evolution and that these

molecular interaction sites or regulatory sites at protein surface can be hotspots of protein adaptation. Indeed, some specific molecular interactions have been linked to adaptive evolution in several case studies (Bachtrog, 2008; Hughes and Nei, 1988; Levin and Malik, 2017; Schott et al., 2014) and large-scale studies based on proteins with high quality structural models (Slodkowicz and Goldman, 2020). In the latter study, the authors showed that positive selections in mammals tend to cluster closer to binding sites of exogenous ligands than expected by chance (Slodkowicz and Goldman, 2020), suggesting an important role of function important regions in adaptive evolution. Here, we extend the conclusion *to D. melanogaster* genome, including proteins with or without high resolution structural models. We also showed that except for exogenous ligands, endogenous ligands might also contribution to adaptive evolution, while the latter might explain why interacting proteins tend to evolve co-adaptively.

Notably, previous studies have revealed that multi-interface proteins tend to be evolving more slowly than single-interface proteins (Kim et al., 2006), which seems to be contradictory to our results that proteins with more interaction sites evolve faster and have faster adaptation rates. Here, we argue that, in our study, we used sequence profile to predict molecular interaction sites in proteins at a genomic scale, rather than only looking into proteins with high resolution structures. In this way, we may capture many weak or transient interactions, which are thought to be evolving faster than obligate and conserved interactions (Mintseris and Weng, 2005). Meanwhile, we did not exclude intrinsic disordered regions (IDR) or intrinsic disordered proteins (IDP) in our study, which are widespread in *D. melanogaster* genome. It has been suggested that IDR/IDP tend to evolve fast due to lack of structural restraints (Echave et al., 2016). In the functional aspect, IDR/IDP are thought to be promiscuous binders through many multiple binding mechanisms, including forming static, semi-static, and fuzzy or dynamic complexes (Uversky, 2019), suggesting that the evolution of IDR/IDP cannot be explained merely by the lack of structural restraints. Actually, IDP and IDR in human genome were found to be undergoing extensive adaptive evolution (Afanasyeva et al., 2018). At last, it has been recognized that, except for allosteric regulations, encounter complexes (Gabdoulline and Wade, 1999) might also play an important role in mediating intermolecular interactions, such as protein-protein association (Tang et al., 2006) and protein-ligand binding (Re et al., 2019). Since encounter residues that are responsible for encounter complexes do not reside in conserved binding interfaces, these residues could be under relaxed purifying selections or even positive selections, which could be another yet-to-identify mechanism that contribute to protein sequence adaptive evolution.

We showed that fast-adaptive proteins are enriched in molecular functions such as reproduction, immunity and environmental information processing (Begun and Lindfors, 2005; Begun and Whitley, 2000; Lazzaro et al., 2004). We further demonstrated that fast-adaptive proteins tend to interact with each other more frequently than random expectations, suggesting co-adaptation might be common among fast-

306 adaptive proteins. Mechanisms that contribute to the co-adaptation could be: (1) interacting fast-adaptive
307 proteins are often enriched in similar molecular functions and under similar selective pressure; (2)
308 interacting fast-adaptive undergo co-evolution through physical interactions. In this study we showed two
309 examples that adaptive evolution could occur at protein-protein interface, which suggest that physical
310 interactions could contribute to the co-adaptation of fast-adaptive proteins in *D. melanogaster*. Moreover,
311 we showed that co-adaptation might exist to a broader extend rather than only among fast-adaptive
312 proteins. Specifically, proteins that interact with fast-adaptive proteins tend to have higher adaptation
313 rates. Since molecular interactions contribute to adaptive evolution, it is reasonable to hypothesize that
314 co-adaptation at this broader extend could be regulated by these interactions. Actually, it has been
315 suggested that interacting proteins tend to have similar evolutionary rates and the possible mechanism
316 would be the co-evolution of physical interactions (Pazos and Valencia, 2008).
317 In this study, we found that loci with significant genetic variance among populations harbor
318 higher proportions of long-term adaptive changes and these loci follow similar patterns as adaptive
319 changes, i.e. they are enriched in disordered regions, protein surfaces, and functionally important regions.
320 These results suggest that population differentiation of protein-coding genes can be an important basis for
321 long-term adaptive evolution. In other word, many SNPs are repeatedly selected for adaptive process in
322 evolution. Importantly, our results indicate that most of the clinal amino-acid changes are adaptive,
323 suggesting that non-selective forces play a non-essential role in the SNPs that show strong geographic
324 differences. Our results also support a large effect of spatially varying selection on protein sequence and
325 structures (Storz and Kelly, 2008).
326 It should be noted that studies at the genomic scale that aim to uncover the function- or structure-
327 related constraints imposed on protein sequence evolution and adaptation share similar limitations that for
328 most of the proteins or residues, structural or functional information would be incomplete or even
329 missing. To overcome this, in this study, we used highly accurate neural-network based tools to predict
330 molecular interactions, secondary structures, intrinsic structural disorder, relative solvent accessibility for
331 each of the protein. In this way we were able to identify key factors that impact protein sequence
332 evolution and adaptation in a less accurate but rather systematic fashion. We hope that with the
333 availability of more and more curated structural, functional information and complex structural models of
334 proteins in the near future, we will be able to uncover the precise role of molecular interactions in protein
335 sequence adaptive evolution.
336
337 **Material and Methods**
338 ***d_N/d_S* ratio (ω).** We used a maximum likelihood method to infer d_N/d_S ratio (ω) of *D. melanogaster*
339 protein-coding genes using the genome sequences of five species in *melanogaster* subgroup (*D.*

340 *melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, and *D. erecta*). The protein-coding sequences were
341 extracted from the alignments of 26 insects, which were obtained from UCSC Genome Browser
342 (http://hgdownload.soe.ucsc.edu/downloads.html). The sequences were further processed by GeneWise
343 (Birney et al., 2004) to remove possible insertions and deletions using the longest isoforms of the
344 corresponding *D. melanogaster* protein sequences as references (FlyBase version r6.15) (Thurmond et al.,
345 2019). The processed sequences were then realigned by PRANK -codon function (Löytynoja, 2014). We
346 used codeml in PAML (Yang, 2007) to compute gene-specific ω using M0 model. We removed
347 sequences that have more than 15% of their nucleotides not aligned (gaps) to *D. melanogaster* genes in
348 more than 2 species. To further avoid numeric errors and ensure reasonable estimations, we only retained
349 relatively divergent sequences that are: (1) divergent with dS larger than 0.3, (2) less divergent with dS
350 larger than 0.1 and dN smaller than 0.001 (dS>>dN). At last, there were 12118 genes in total passed all
351 the criteria and were assigned gene specific ω, containing 6,538,872 amino acids. We also calculated site-
352 specific ω by using likelihood ratio tests (LRT) comparing M7 model against M8 model (Yang et al.,
353 2005).
354 ***Rate of adaptive and nonadaptive changes.*** We recalled all SNPs of 205 inbred lines from
355 the Drosophila Genetic Reference Panel (DGRP), Freeze 2.0 (Huang et al., 2014)
356 (http://dgrp2.gnets.ncsu.edu). We then generated 410 alternative genomes using all monoallelic and bi-
357 allelic SNP data sets. We extracted the coding sequences of *D. melanogaster* genes from the generated
358 alternative genomes, removed all possible insertions and deletions using GeneWise (Birney et al., 2004)
359 as described above. We then align all the coding sequences to their corresponding aligned CDS sequences
360 using PRANK -codon function (Löytynoja, 2014). We removed polymorphisms segregating at
361 frequencies smaller than 5% to reduce possible slightly deleterious mutations (Charlesworth and Eyre-
362 Walker, 2008). In order to avoid possible effects of low divergence between *D. simulans* and D
363 melanogaster (Keightley and Eyre-Walker, 2012), we used *D. yakuba* as outgroup to estimate
364 nonsynonymous polymorphisms (Pn), synonymous polymorphisms (Ps), nonsynonymous substitutions
365 (Dn) and synonymous substitutions (Ds) by MK.pl (Begun et al., 2007; Langley et al., 2012). Similar as
366 Begun et al. (Begun et al., 2007), we only analyzed genes with at least six variants for each of
367 substitutions, polymorphisms, nonsynonymous changes and synonymous changes. We used an extension
368 of MK test, asymptotic MK (Messer and Petrov, 2013; Uricchio et al., 2019), to estimate the proportions
369 of adaptive changes (α). The rate of adaptive changes (ω_a) was then calculated as $\omega_a = \omega\alpha$ and the rate of
370 non-adaptive changes as $\omega_{na} = \omega - \omega_a$. Details of the asymptotic MK test were as following:
371 (1) Classical McDonald–Kreitman test. According to Smith and Eyre-Walker (Smith and Eyre-Walker,
372 2002), the proportions of adaptive changes for protein-coding genes can be calculated as following:

373 $$\alpha = 1 - \frac{DsPn}{DnPs}$$

374 According to this equation, we could estimate the proportion of adaptive changes and carried out classical
375 MK test by applying Fisher's exact test.
376 (2) Asymptotic estimation of α. A known problem of the classical estimation of α above is the
377 accumulation of slightly deleterious mutations at low frequencies. We therefore used an extension of MK
378 test, asymptotic MK test approach (Messer and Petrov, 2013) to estimate the proportions of adaptive
379 changes. As in original aMK, we defined α(x) as a function of derived allele frequency (x):

380 $$\alpha(x) = 1 - \frac{DsPn(x)}{DnPs(x)}$$

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

388 $$\omega_a = \frac{N_a/L_N}{dS} = \frac{dN_a}{dS} = \frac{dN_a}{dN} \cdot \frac{dN}{dS} = \alpha\omega$$

389 where N_a is the number of adaptive changes and $dN_a=N_a/L_N$ is the number of adaptive changes per
390 nonsynonymous site. Finally, we calculated the rate of nonadaptive changes (ω_{na}) as $\omega_{na}=\omega-\omega_a$. The final
391 dataset contains 7192 protein-coding genes, with smallest ω_a being 0.00 and largest being 1.29.
392 ***Structure-/function- related properties of D. melanogaster proteins.*** We obtained function-related
393 properties mentioned in main text as following. We derived *D. melanogaster* gene ages (Kondo et al.,
394 2017; Zhang et al., 2010) for genes that are specific to *Drosophila*, and from GenTree (Shao et al., 2019)
395 for genes that are beyond *Drosophila* clade. We then assigned a pseudo-age to each of the genes.
396 Specifically, there are 11 age groups from "cellular organisms", assigning to a pseudo age value of 0, to
397 "melanogaster", assigning a pseudo age value of 10. We downloaded *D. melanogaster* protein-protein
398 interaction (PPI) from STRING database (Szklarczyk et al., 2019). A cut-off of combined score larger
399 than 0.7 was used to retain high confident PPI for further analysis. We then used BSpred (Mukherjee and
400 Zhang, 2011) to predict protein-protein interaction (PPI) sites and DRNApred (Yan and Kurgan, 2017) to
401 predict DNA binding sites. For each protein, we calculated ratios of protein interaction residues (PPI-site
402 ratio) and ratios of DNA binding residues (DNA-site ratio) by dividing total predicted protein interaction
403 sites and DNA binding sites over protein length, respectively. For structure-related properties, we used

404 DeepCNF (Wang et al., 2016) to predict these properties for each gene, including three-state secondary
405 structures (helix, sheet, and coil), structural disorder, relative solvent accessibility (RSA). Further, we
406 calculated the ratios of helix, sheet, helix+sheet, and coil residues of each gene from predicted secondary
407 structures. For each gene, we computed intrinsic structural disorder (ISD) and relative solvent
408 accessibility (RSA), as protein-length normalized summations of the probabilities of each residue being
409 disorder and exposed, respectively.

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

412 $$\mathrm{TPM_i} = \frac{\mathrm{FPKM_i}}{\sum \mathrm{FPKM_j}} \times 10^6$$

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

417 $$zscore = \frac{TPM(male\ expression) - TPM(female\ expression)}{\sqrt{sd^2(male\ expression) + sd^2(female\ expression)}}$$

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

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

Evolutionary rates of different structural/functional sites. We classified amino acids into different classes of structural/functional properties. Specifically, we classified three classes for both ISD and RSA according the probability of residues being disordered or exposed: ordered or buried (0.00 to 0.33), medium (0.33 to 0.67), disordered or exposed (0.67 to 1.00). For both PPI and DNA binding, we classified two classes: PPI-site or DNA-site (binding sites), None-PPI or None-DNA (corresponding null sites for PPI or DNA binding). For residues that have 3D structures, we used STRESS (Clarke et al., 2016) to predict putative ligand binding sites and allosteric sites from all the high-quality structures or structural models. The allosteric sites were further classified as surface critical or interior critical according to their locations. We then classified these residues into four groups: LIG (ligand binding sites), Surf. Crit. (surface critical sites), Interior Crit. (interior critical sites) and Others (other sites). For each of the site classes, we randomly sampled 100 sequences, each containing 10,000 amino acids. We computed ω, ω_a, and ω_{na} for the randomly sampled sequences similar as the steps described in the above sections.

Acknowledgements

We thank members of the Zhao Lab for helpful discussions.

Author contribution

J.P. and L.Z. conceived the study. J.P. performed the analysis with the input from L.Z.. J.P. and L.Z. wrote the manuscript.

Funding

The work was supported by NIH MIRA R35GM133780, the Robertson Foundation, a Monique Weill-Caulier Career Scientist Award, an Alfred P. Sloan Research Fellowship (FG-2018-10627), a Rita Allen 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

468 dosage compensation in Drosophila. Genetics *180*, 1123–1129.
469 Begun, D.J., and Lindfors, H.A. (2005). Rapid evolution of genomic Acp complement in the
470 melanogaster subgroup of Drosophila. Mol. Biol. Evol. *22*, 2010–2021.
471 Begun, D.J., and Whitley, P. (2000). Adaptive evolution of relish, a Drosophila NF-kappaB/IkappaB
472 protein. Genetics *154*, 1231–1238.
473 Begun, D.J., Holloway, A.K., Stevens, K., Hillier, L.D.W., Poh, Y.P., Hahn, M.W., Nista, P.M., Jones,
474 C.D., Kern, A.D., Dewey, C.N., et al. (2007). Population genomics: Whole-genome analysis of
475 polymorphism and divergence in Drosophila simulans. PLoS Biol. *5*, 2534–2559.
476 Bienert, S., Waterhouse, A., De Beer, T.A.P., Tauriello, G., Studer, G., Bordoli, L., and Schwede, T.
477 (2017). The SWISS-MODEL Repository-new features and functionality. Nucleic Acids Res. *45*, D313–
478 D319.
479 Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. Genome Res. *14*, 988–995.
480 Burley, S.K., Berman, H.M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., Christie, C., Dalenberg, K.,
481 Duarte, J.M., Dutta, S., et al. (2019). RCSB Protein Data Bank: Biological macromolecular structures
482 enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic
483 Acids Res. *47*, D464–D474.
484 Butterwick, J.A., del Mármol, J., Kim, K.H., Kahlson, M.A., Rogow, J.A., Walz, T., and Ruta, V. (2018).
485 Cryo-EM structure of the insect olfactory receptor Orco. Nature *560*, 447–452.
486 Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L.
487 (2009). BLAST+: Architecture and applications. BMC Bioinformatics *10*, 421.
488 Charlesworth, J., and Eyre-Walker, A. (2008). The McDonald-Kreitman test and slightly deleterious
489 mutations. Mol. Biol. Evol. *25*, 1007–1015.
490 Chen, H., Zhang, Z., Jiang, S., Li, R., Li, W., Zhao, C., Hong, H., Huang, X., Li, H., and Bo, X. (2020).
491 New insights on human essential genes based on integrated analysis and the construction of the HEGIAP
492 web-based platform. Brief. Bioinform. *21*, 1397–1410.
493 Clarke, D., Sethi, A., Li, S., Kumar, S., Chang, R.W.F., Chen, J., and Gerstein, M. (2016). Identifying
494 Allosteric Hotspots with Dynamics: Application to Inter- and Intra-species Conservation. Structure *24*,
495 826–837.
496 Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O., and Arnold, F.H. (2005). Why highly expressed
497 proteins evolve slowly. Proc. Natl. Acad. Sci. U. S. A. *102*, 14338–14343.
498 Echave, J., Spielman, S.J., and Wilke, C.O. (2016). Causes of evolutionary rate variation among protein
499 sites. Nat. Rev. Genet. *17*, 109–121.
500 Enard, D., Cai, L., Gwennap, C., and Petrov, D.A. (2016). Viruses are a dominant driver of protein
501 adaptation in mammals. Elife *5*, e12469.

502 Gabdoulline, R.R., and Wade, R.C. (1999). On the protein-protein diffusional encounter complex. J. Mol.
503 Recognit. *12*, 226–234.
504 Goldman, N., Thorne, J.L., and Jones, D.T. (1998). Assessing the impact of secondary structure and
505 solvent accessibility on protein evolution. Genetics *149*, 445–458.
506 Hill, C.A., Fox, A.N., Pitts, R.J., Kent, L.B., Tan, P.L., Chrystal, M.A., Cravchik, A., Collins, F.H.,
507 Robertson, H.M., and Zwiebel, L.J. (2002). G Protein-Coupled Receptors in Anopheles gambiae. Science
508 *298*, 176–178.
509 Huang, W., Massouras, A., Inoue, Y., Peiffer, J., Ràmia, M., Tarone, A.M., Turlapati, L., Zichner, T.,
510 Zhu, D., Lyman, R.F., et al. (2014). Natural variation in genome architecture among 205 Drosophila
511 melanogaster Genetic Reference Panel lines. Genome Res. *24*, 1193–1208.
512 Hughes, A.L., and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex
513 class I loci reveals overdominant selection. Nature *335*, 167–170.
514 Jiggins, F.M., and Kim, K.W. (2007). A screen for immunity genes evolving under positive selection in
515 Drosophila. J. Evol. Biol. *20*, 965–970.
516 Keightley, P.D., and Eyre-Walker, A. (2012). Estimating the rate of adaptive molecular evolution when
517 the evolutionary divergence between species is small. J. Mol. Evol. *74*, 61–68.
518 Kim, P.M., Lu, L.J., Xia, Y., and Gerstein, M.B. (2006). Relating three-dimensional structures to protein
519 networks provides evolutionary insights. Science *314*, 1938–1941.
520 Kondo, S., Vedanayagam, J., Mohammed, J., Eizadshenass, S., Kan, L., Pang, N., Aradhya, R., Siepel,
521 A., Steinhauer, J., and Lai, E.C. (2017). New genes often acquire male specific functions but rarely
522 become essential in Drosophila. Genes Dev. *31*, 1841–1846.
523 Kosiol, C., Vinař, T., Da Fonseca, R.R., Hubisz, M.J., Bustamante, C.D., Nielsen, R., and Siepel, A.
524 (2008). Patterns of positive selection in six mammalian genomes. PLoS Genet. *4*, e1000144.
525 Langley, C.H., Stevens, K., Cardeno, C., Lee, Y.C.G., Schrider, D.R., Pool, J.E., Langley, S.A., Suarez,
526 C., Corbett-Detig, R.B., Kolaczkowski, B., et al. (2012). Genomic variation in natural populations of
527 Drosophila melanogaster. Genetics *192*, 533–598.
528 Lawniczak, M.K.N., and Begun, D.J. (2007). Molecular population genetics of female-expressed mating-
529 induced serine proteases in Drosophila melanogaster. Mol. Biol. Evol. *24*, 1944–1951.
530 Lazzaro, B.P., Sceurman, B.K., and Clark, A.G. (2004). Genetic basis of natural variation in D.
531 melanogaster antibacterial immunity. Science *303*, 1873–1876.
532 Leader, D.P., Krause, S.A., Pandit, A., Davies, S.A., and Dow, J.A.T. (2018). FlyAtlas 2: A new version
533 of the Drosophila melanogaster expression atlas with RNA-Seq, miRNA-Seq and sex-specific data.
534 Nucleic Acids Res. *46*, D809–D815.
535 Levin, T.C., and Malik, H.S. (2017). Rapidly evolving Toll-3/4 genes encode male-specific Toll-like

536 receptors in drosophila. Mol. Biol. Evol. *34*, 2307–2323.
537 Lin, Y.S., Hsu, W.L., Hwang, J.K., and Li, W.H. (2007). Proportion of solvent-exposed amino acids in a
538 protein and rate of protein evolution. Mol. Biol. Evol. *24*, 1005–1011.
539 Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J.,
540 Jordan, G., Mauceli, E., et al. (2011). A high-resolution map of human evolutionary constraint using 29
541 mammals. Nature *478*, 476–482.
542 Löytynoja, A. (2014). Phylogeny-aware alignment with PRANK. Methods Mol. Biol. *1079*, 155–170.
543 McBride, C.S. (2007). Rapid evolution of smell and taste receptor genes during host specialization in
544 Drosophila sechellia. Proc. Natl. Acad. Sci. U. S. A. *104*, 4996–5001.
545 Messer, P.W., and Petrov, D.A. (2013). Frequent adaptation and the McDonald-Kreitman test. Proc. Natl.
546 Acad. Sci. U. S. A. *110*, 8615–8620.
547 Mintseris, J., and Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein-
548 protein interactions. Proc. Natl. Acad. Sci. U. S. A. *102*, 10930–10935.
549 Moutinho, A.F., Trancoso, F.F., Dutheil, J.Y., and Zhang, J. (2019). The Impact of Protein Architecture
550 on Adaptive Evolution. Mol. Biol. Evol. *36*, 2013–2028.
551 Mukherjee, S., and Zhang, Y. (2011). Protein-protein complex structure predictions by multimeric
552 threading and template recombination. Structure *19*, 955–966.
553 Newville, M., and Stensitzki, T. (2018). Non-Linear Least-Squares Minimization and Curve-Fitting for
554 Python. Zenodo.
555 Nielsen, R., Bustamante, C., Clark, A.G., Glanowski, S., Sackton, T.B., Hubisz, M.J., Fledel-Alon, A.,
556 Tanenbaum, D.M., Civello, D., White, T.J., et al. (2005). A scan for positively selected genes in the
557 genomes of humans and chimpanzees. PLoS Biol. *3*, 0976–0985.
558 Obbard, D.J., Welch, J.J., Kim, K.W., and Jiggins, F.M. (2009). Quantifying adaptive evolution in the
559 Drosophila immune system. PLoS Genet. *5*, e1000698.
560 Pál, C., Papp, B., and Hurst, L.D. (2001). Highly expressed genes in yeast evolve slowly. Genetics *158*,
561 927–931.
562 Palmer, W.H., Hadfield, J.D., and Obbard, D.J. (2018). RNA-interference pathways display high rates of
563 adaptive protein evolution in multiple invertebrates. Genetics *208*, 1585–1599.
564 Pazos, F., and Valencia, A. (2008). Protein co-evolution, co-adaptation and interactions. EMBO J. *27*,
565 2648–2655.
566 Pieper, U., Webb, B.M., Barkan, D.T., Schneidman-Duhovny, D., Schlessinger, A., Braberg, H., Yang,
567 Z., Meng, E.C., Pettersen, E.F., Huang, C.C., et al. (2011). ModBase,a database of annotated comparative
568 protein structure models,and associated resources. Nucleic Acids Res. *39*, D465–D474.
569 Ramsey, D.C., Scherrer, M.P., Zhou, T., and Wilke, C.O. (2011). The relationship between relative

570 solvent accessibility and evolutionary rate in protein evolution. Genetics *188*, 479–488.
571 Re, S., Oshima, H., Kasahara, K., Kamiya, M., and Sugita, Y. (2019). Encounter complexes and hidden
572 poses of kinaseinhibitor binding on the free-energy landscape. Proc. Natl. Acad. Sci. U. S. A. *116*,
573 18404–18409.
574 Sackton, T.B., Lazzaro, B.P., Schlenke, T.A., Evans, J.D., Hultmark, D., and Clark, A.G. (2007).
575 Dynamic evolution of the innate immune system in Drosophila. Nat. Genet. *39*, 1461–1468.
576 Schott, R.K., Refvik, S.P., Hauser, F.E., López-Fernández, H., and Chang, B.S.W. (2014). Divergent
577 positive selection in rhodopsin from lake and riverine cichlid fishes. Mol. Biol. Evol. *31*, 1149–1165.
578 Shao, Y., Chen, C., Shen, H., He, B.Z., Yu, D., Jiang, S., Zhao, S., Gao, Z., Zhu, Z., Chen, X., et al.
579 (2019). GenTree, an integrated resource for analyzing the evolution and function of primate-specific
580 coding genes. Genome Res. *29*, 682–696.
581 Sironi, M., Cagliani, R., Forni, D., and Clerici, M. (2015). Evolutionary insights into host-pathogen
582 interactions from mammalian sequence data. Nat. Rev. Genet. *16*, 224–236.
583 Slodkowicz, G., and Goldman, N. (2020). Integrated structural and evolutionary analysis reveals common
584 mechanisms underlying adaptive evolution in mammals. Proc. Natl. Acad. Sci. U. S. A. *117*, 5977–5986.
585 Smith, N.G.C., and Eyre-Walker, A. (2002). Adaptive protein evolution in Drosophila. Nature *415*, 1022–
586 1024.
587 Song, Y., Dimaio, F., Wang, R.Y.R., Kim, D., Miles, C., Brunette, T., Thompson, J., and Baker, D.
588 (2013). High-resolution comparative modeling with RosettaCM. Structure *21*, 1735–1742.
589 Storz, J.F., and Kelly, J.K. (2008). Effects of Spatially Varying Selection on Nucleotide Diversity and
590 Linkage Disequilibrium: Insights From Deer Mouse Globin Genes. Genetics *180*, 367–379.
591 Svetec, N., Cridland, J.M., Zhao, L., and Begun, D.J. (2016). The Adaptive Significance of Natural
592 Genetic Variation in the DNA Damage Response of Drosophila melanogaster. PLoS Genet. *12*,
593 e1005869.
594 Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva,
595 N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: Protein-protein association networks with
596 increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic
597 Acids Res. *47*, D607–D613.
598 Tang, C., Iwahara, J., and Clore, G.M. (2006). Visualization of transient encounter complexes in protein-
599 protein association. Nature *444*, 383–386.
600 Thurmond, J., Goodman, J.L., Strelets, V.B., Attrill, H., Gramates, L.S., Marygold, S.J., Matthews, B.B.,
601 Millburn, G., Antonazzo, G., Trovisco, V., et al. (2019). FlyBase 2.0: The next generation. Nucleic Acids
602 Res. *47*, D759–D765.
603 Uricchio, L.H., Petrov, D.A., and Enard, D. (2019). Exploiting selection at linked sites to infer the rate

604 and strength of adaptation. Nat. Ecol. Evol. *3*, 977–984.
605 Uversky, V.N. (2019). Intrinsically disordered proteins and their "Mysterious" (meta)physics. Front.
606 Phys. *7*, 10.
607 Wang, S., Li, W., Liu, S., and Xu, J. (2016). RaptorX-Property: a web server for protein structure
608 property prediction. Nucleic Acids Res. *44*, W430–W435.
609 Wu, D.D., Wang, G.D., Irwin, D.M., and Zhang, Y.P. (2009). A profound role for the expansion of
610 trypsin-like serine protease family in the evolution of hematophagy in mosquito. Mol. Biol. Evol. *26*,
611 2333–2341.
612 Xia, B., Yan, Y., Baron, M., Wagner, F., Barkley, D., Chiodin, M., Kim, S.Y., Keefe, D.L., Alukal, J.P.,
613 Boeke, J.D., et al. (2020). Widespread Transcriptional Scanning in the Testis Modulates Gene Evolution
614 Rates. Cell *180*, 248-262.e21.
615 Yan, J., and Kurgan, L. (2017). DRNApred, fast sequence-based method that accurately predicts and
616 discriminates DNA-and RNA-binding residues. Nucleic Acids Res. *45*.
617 Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-
618 Saban, S., Safran, M., Domany, E., et al. (2005). Genome-wide midrange transcription profiles reveal
619 expression level relationships in human tissue specification. Bioinformatics *21*, 650–659.
620 Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–
621 1591.
622 Yang, L., Zhang, Z., and He, S. (2016). Both Male-Biased and Female-Biased Genes Evolve Faster in
623 Fish Genomes. Genome Biol. Evol. *8*, 3433–3445.
624 Yang, Z., Wong, W.S.W., and Nielsen, R. (2005). Bayes empirical Bayes inference of amino acid sites
625 under positive selection. Mol. Biol. Evol. *22*, 1107–1118.
626 Zhang, J., and He, X. (2005). Significant impact of protein dispensability on the instantaneous rate of
627 protein evolution. Mol. Biol. Evol. *22*, 1147–1155.
628 Zhang, J., and Yang, J.R. (2015). Determinants of the rate of protein sequence evolution. Nat. Rev. Genet.
629 *16*, 409–420.
630 Zhang, Y.E., Vibranovski, M.D., Krinsky, B.H., and Long, M. (2010). Age-dependent chromosomal
631 distribution of male-biased genes in Drosophila. Genome Res. *20*, 1526–1533.
632 Zheng, N., Schulman, B.A., Song, L., Miller, J.J., Jeffrey, P.D., Wang, P., Chu, C., Koepp, D.M., Elledge,
633 S.J., Pagano, M., et al. (2002). Structure of the Cul1-Rbx1-Skp1-F boxSkp2 SCF ubiquitin ligase
634 complex. Nature *416*, 703–709.
635
636
637

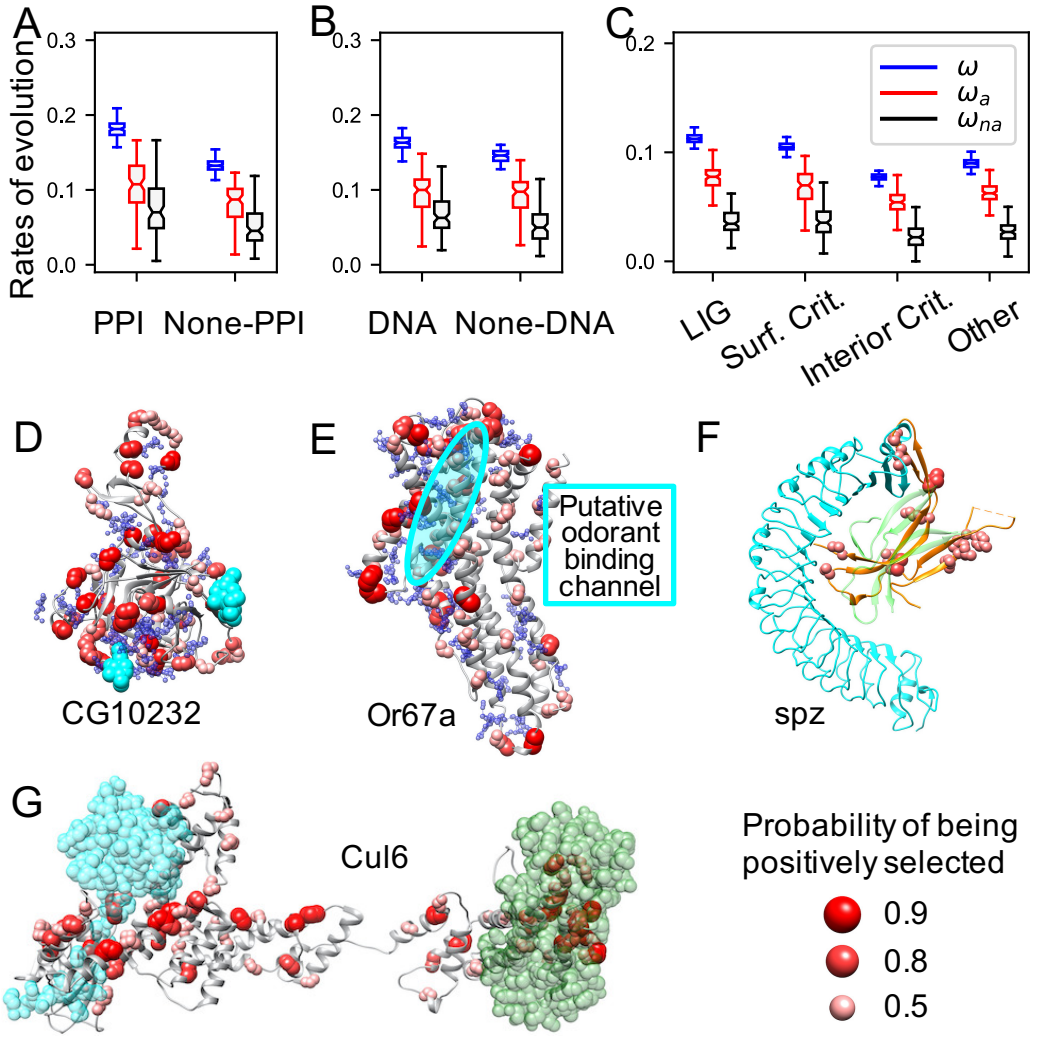

Figure 1. Adaptive evolution in molecular interaction sites. Protein-protein interaction sites (A), DNA binding sites (B) and putative ligand binding sites (C) show higher adaptation rates than none binding sites. Examples of positive selection around molecular interaction sites in high quality structural models of CG10232 (D), Or67a (E), spz (F), and Cul6 (G). Except for spz (PDB code 3e07), the other proteins are obtained from SWISS model repository. Putative ligand binding pockets of CG10232 (D) and Or67a (E) are shown in blue spheres. Ligands including interacting proteins are shown in cyan or green: NAG of CG10232 in cyan (D), Toll receptor of spz in cyan (F), RING-box protein in cyan and F-box protein in green for Cul6 (G). The putative odorant binding channel of Or67a is highlighted in cyan circle (E). The ligand poses in (D, F and G) are obtained by superimposition from structure 2XXL, 4BV4 and 1LDK, respectively.

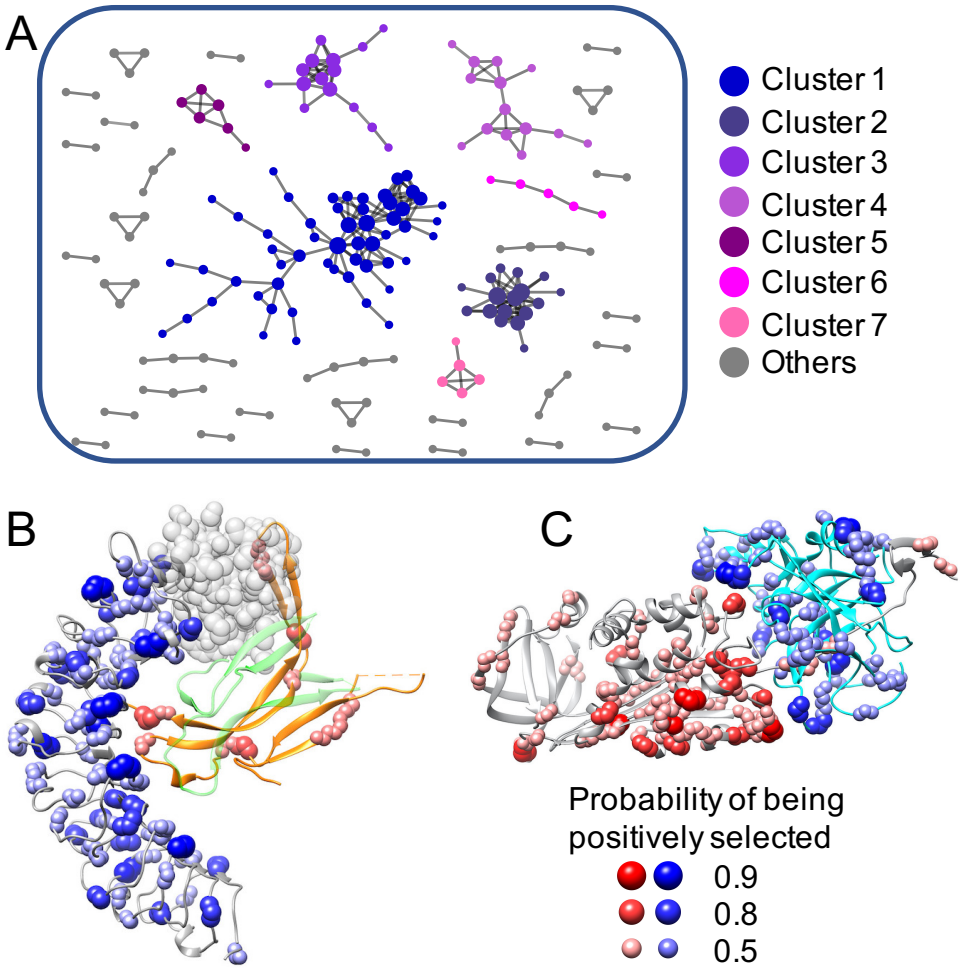

Figure 2. Co-adaptation of fast-adaptive proteins. (A) Sub-clusters of PPI networks of fast-adaptive proteins. Only proteins with at least one partner were shown. Examples of molecular interactions that might regulate co-adaptation in fast-adaptive proteins: (B) Toll-4 (gray) and spz (orange, with green 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.

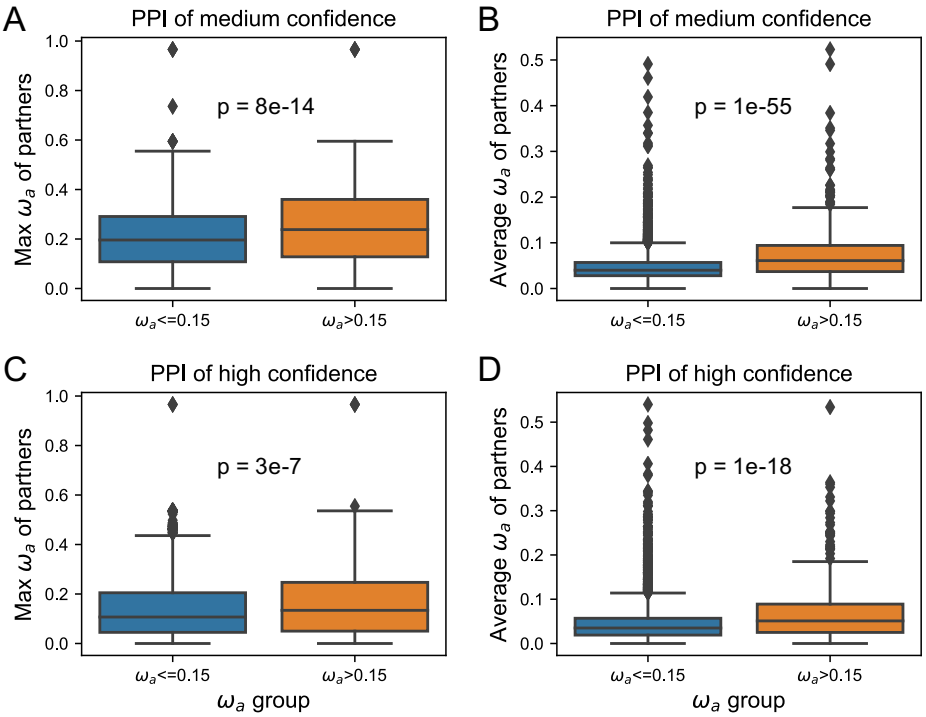

Figure 3. Co-adaptation of PPIs in *D. melanogaster*. For fast-adaptive proteins, adaptation rates of their partners (orange box plot) are significantly larger compared to slow adaptive proteins (blue box plot). Max ω_a of protein partners are shown in (A and C) and averaged ω_a, of protein partners 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).

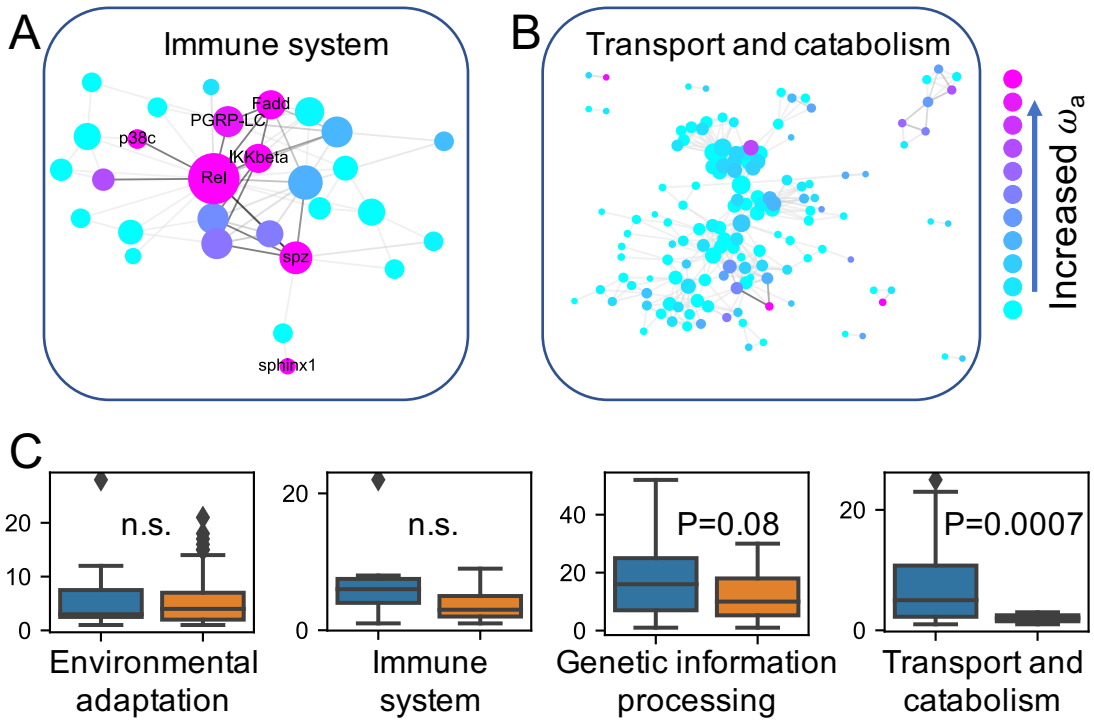

Figure 4. Rates of protein sequence adaptive evolution in the PPI network of different functional pathways. The PPI networks showed the adaptive evolution in immune system (A) and transport and catabolism (B). (C) In pathways that are hotspots of adaptive evolution, fast-adaptive proteins can act as central nodes, while in conserved pathways, fast-adaptive proteins are often at the periphery of the PPI network.

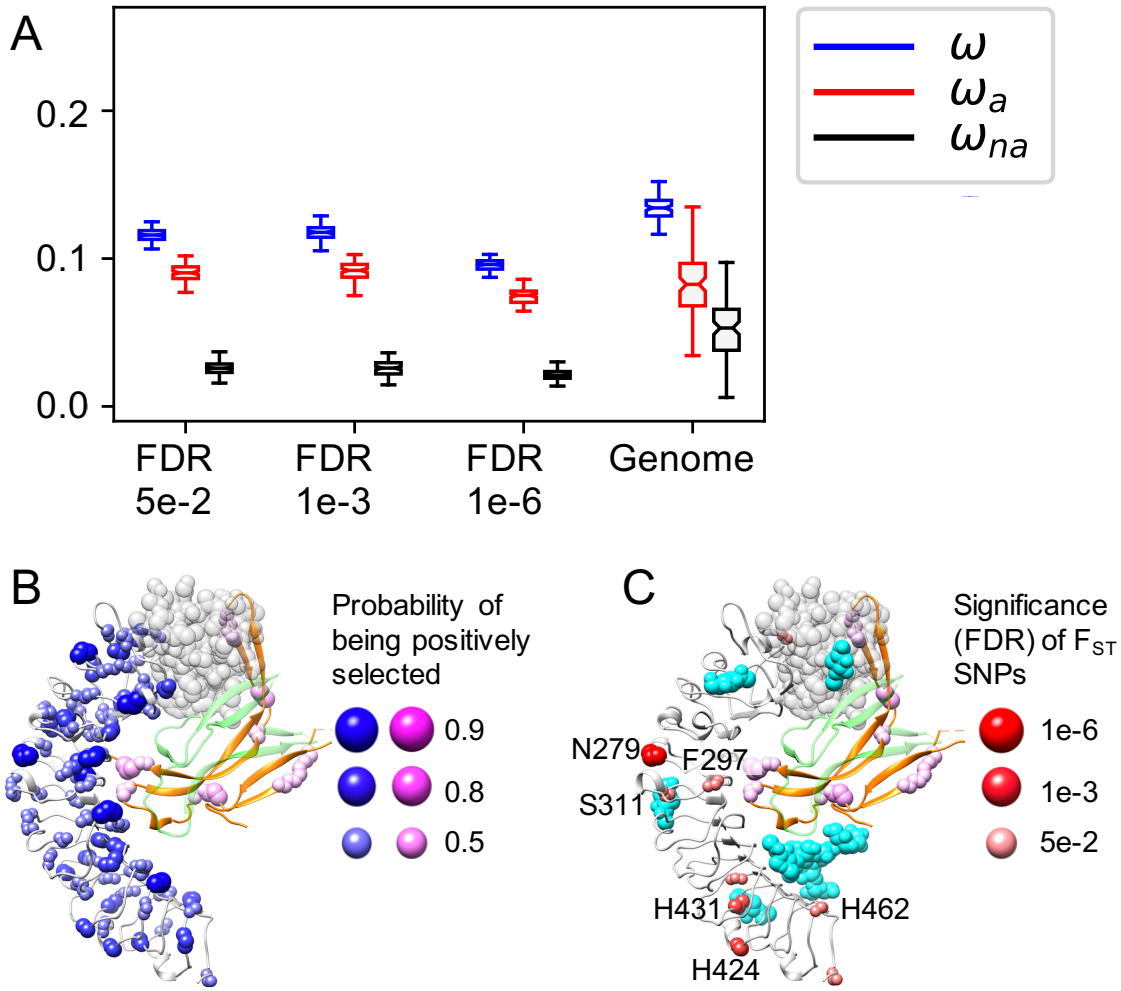

Figure 5. Adaptive evolution in significant nonsynonymous F_{ST} SNPs. (A) The significant SNPs at different FDR cutoffs all show much higher proportions of adaptation than genome-wide expectation. (B) Positive selections in Toll-4 and Spaztle, related to Fig. 2B. (C) Significant nonsynonymous F_{ST} SNPs in Toll-4. Ligands are shown in cyan by superimposing crystal structure of Toll-Spatzle (PDB code 4BV4) on to Toll-4 structural model. N279, H431 are both highly differentiated (FDR 3e-7 and 3e-6) and positively selected (both probability at p=0.9). Other highly differentiated sites, F297, S311, H424, H431 and H462 are located near ligand binding sites or positively selected sites.