# Protein evidence of unannotated ORFs in *Drosophila* reveals unappreciated diversity in the evolution of young proteins

Eric B Zheng, Li Zhao*

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

*Correspondence to: lzhao@rockefeller.edu

## Abstract

*De novo* gene origination, where a previously non-genic genomic sequence becomes genic through evolution, has been increasingly recognized as an important source of evolutionary novelty across diverse taxa. Many *de novo* genes have been proposed to be protein-coding, and in several cases have been experimentally shown to yield protein products. However, the systematic study of *de novo* proteins has been hampered by doubts regarding the translation of their transcripts without the experimental observation of protein products. Using a systematic, ORF-focused mass-spectrometry-first computational approach, we identify almost 1000 unannotated open reading frames with evidence of translation (utORFs) in the model organism *Drosophila melanogaster*, 371 of which have canonical start codons. To quantify the comparative genomic similarity of these utORFs across *Drosophila* and to infer phylostratigraphic age, we further develop a synteny-based protein similarity approach. Combining these results with reference datasets on tissue- and life-stage-specific transcription and conservation, we identify different properties amongst these utORFs. Contrary to expectations, the fastest-evolving utORFs are not the youngest evolutionarily. We observed more utORFs in the brain than in the testis. Most of the identified utORFs may be of *de novo* origin, even accounting for the possibility of false-negative similarity detection. Finally, sequence divergence after an inferred *de novo* origin event remains substantial, raising the possibility that *de novo* proteins turn over frequently. Our results suggest that there is substantial unappreciated diversity in *de novo* protein evolution: many more may exist than have been previously appreciated; there may be divergent evolutionary trajectories; and *de novo* proteins may be gained and lost frequently. All in all, there may not exist a single characteristic model of *de novo* protein evolution, but rather complex origins and evolutionary trajectories for *de novo* proteins.

## Impact statement

The analysis of mass-spectrometry data for all possible open reading frames reveals protein evidence for evolutionarily young, unannotated proteins with distinct characters.

## Introduction

While the study of the evolution of entire species or the fixation of individual mutations has matured over decades, our understanding of how genes evolve is relatively less mature. The dominant mechanism is duplication and divergence (Ohno, 1970; Long et al., 2003), but the early years of the genomic era revealed a seeming paradox of species-specific "orphan genes". These genes lacked homologs even in closely related species, which made it necessary to invoke non-standard duplication and divergence mechanisms (Domazet-Loso and Tautz, 2003). Advances in both the sequencing of genomes and in comparative genomics soon led to an explosion of work on the origins of orphan genes, generally identified through lineage-specificity.

Lineage-specific orphan genes can originate from a variety of mechanisms, ranging from extreme, rapid divergence after duplication (Chen et al., 2010; Domazet-Loso and Tautz, 2003) to horizontal gene transfer (Verster et al., 2019) to *de novo* gene birth. In the latter mechanism, previously non-genic genomic sequence acquires genic function, thus yielding a *de novo* gene that is characterized by its lineage-specificity (Begun et al., 2006; Levine et al., 2006; McLysaght and Hurst, 2016; Schlötterer, 2015; Van Oss and Carvunis, 2019). *De novo* genes are identified from diverse taxa from fruit flies (Begun et al., 2006; Levine et al., 2006), to yeast (Cai et al., 2008) and even humans (Knowles and McLysaght, 2009; Xie et al., 2012). While it would seem unlikely for such *de novo* genes to gain viable function from presumably unoptimized sequence, empirical and population genetics studies have demonstrated that they have important functions (Cai et al., 2008; Lange et al., 2021; Levine et al., 2006; Reinhardt et al., 2013; Xie et al., 2019; Zhao et al., 2014). However, exactly what proportion of lineage-specific genes are contributed through *de novo* gene birth remains actively studied (Tautz and Domazet-Lošo, 2011; McLysaght and Guerzoni, 2015; Vakirlis et al., 2020).

Despite significant interest, the study of evolutionarily young genes is surprisingly conceptually difficult. While an empirical investigation of biological effect is ideal, designing such experiments for a specific evolutionarily young gene can be challenging due to the likely lack of homology to previously studied genes. Even identifying a given candidate genetic locus as an evolutionarily young gene can be fraught with challenges: evolutionary youth is functionally equivalent to concluding lineage-specificity, but methods like phylostratigraphy via BLAST (Domazet-Loso et al., 2007), while useful and intuitive, can be vulnerable to false-negative detections, especially for short genes (Moyers and Zhang, 2015, 2016). Moreover, while researchers generally agree that a gene should be functional, exactly how this should be defined and assessed can vary substantially (Van Oss and Carvunis, 2019). As an example, whether evolutionarily young genes function as RNAs or as proteins is often an area of debate. One of the earliest recognized young genes is an RNA (Wang et al., 2002), other studies also provided such support

(Cridland et al., 2022; Ruiz-Orera et al., 2015); in contrast, in several cases even *de novo* genes have been empirically shown to yield protein products (Bungard et al., 2017; Cai et al., 2008; Knowles and McLysaght, 2009; Li et al., 2010; Zhang et al., 2019). While defining a gene as one that can code for a translated protein is convenient and intuitive, evidence for the translation into polypeptides of lineage-specific genes and in particular, *de novo* genes as a class, is often lacking (Van Oss and Carvunis, 2019). As a result, much effort remains needed for systematic approaches to studying evolutionarily young genes and proteins.

Proteomics approaches like shotgun mass spectrometry (MS) can be a powerful tool for the rapid identification of translational evidence for many potential proteins at once, and would be a good fit for such a systematic study. However, standard shotgun MS analysis approaches require *a priori* assumptions about the sequences of the proteins to be detected (Sinitcyn, Rudolph, and Cox 2018). In the case of discovering possible lineage-specific proteins, this *a priori* requirement can be confounding, as lineage-specific proteins may be systematically under-annotated due to their lack of homology.

Here, we employ a systematic, ORF-focused MS-first computational approach, to identify unannotated open reading frames with evidence of translation (utORFs) in *Drosophila melanogaster*. We further develop a synteny-based protein similarity approach to quantify the comparative genomic similarity of these utORFs across *Drosophila*, to estimate their phylostratigraphic age, and finally to infer possible *de novo* origin for the vast majority of these utORFs. Combined with reference datasets on tissue- and life stage-specific transcription and conservation, this data suggests differences amongst these utORFs with respect to their evolutionary and functional properties. Finally, we independently support a subset of these utORFs via additional MS empirical data and ribosome profiling data.

## Results

### A comprehensive database of all possible open reading frames in the *Drosophila melanogaster* genome

As a first step towards the systematic discovery of unannotated proteins in *Drosophila melanogaster*, we generated a comprehensive database of all possible open reading frames (ORFs) in the *D. melanogaster* genome. This comprehensive database is the result of translating the entire annotated genome in all six frames; we also included the annotated transcriptome to enable discovery of non-canonical translation phenomena such as alternate frame ORFs (de Klerk and ’t Hoen 2015). In this database, each potential ORF is defined as a contiguous span of amino acid-coding codons; so, these potential ORFs do not necessarily begin with or contain a canonical start codon. This is necessary in order to account for unannotated splicing events, as a six-frame translation of the genome cannot correctly associate the exons

96 of a gene. Since no coding exon within the coding sequence of a gene can contain a stop codon, the set of
97 all codon spans between stop codons is a superset of the set of all exons in the genome. In addition to
98 annotated ORFs (FlyBase r6.15 release), this comprehensive potential ORF database contains 4,582,998
99 unique potential ORFs that are not otherwise annotated. Of these additional potential ORFs, the median
100 length is 32 amino acids and the 95th percentile is 96 amino acids. Considering solely the 2,597,794
101 potential ORFs that contain a canonical start codon, the median length following the start codon remains
102 32 amino acids, and the 95th percentile is 112 amino acids. In total, this suggests that the *D. melanogaster*
103 genome has substantial latent potential for generating new protein-coding genes. Of course, the
104 overwhelming majority of these potential ORFs are neither transcribed nor translated and are thus not
105 functional (Durand et al. 2019), but considering these will allow us to scan the potential proteomic
106 diversity in *D. melanogaster*.

107 **Two-round MS search improves discovery of unannotated ORFs without compromising FDR**

108 A comprehensive proteome of *D. melanogaster* was recently published (Casas-Vila et al. 2017), including
109 data from 15 stages of the whole lifecycle of the fruit fly (ranging from 0-2h embryos to adults, including
110 larval and pupal stages), as well as sharper time-resolution data covering 14 stages of embryogenesis
111 (hourly timepoints from 0-6 hours; bi-hourly timepoints from 6 to 20 hours). We analyzed the *Casas-*
112 *Villa et al.* data by searching for the complete potential ORF database described above to discover
113 evidence of translation of unannotated ORFs in *D. melanogaster*. Because the number of potential ORFs
114 is several orders of magnitude greater than the number of annotated proteins, the search space is
115 enormously expanded, and thus a traditional one-round search is likely to miss a very large number of
116 identifications at a given false-discovery rate (FDR). Accordingly, to improve total sensitivity while
117 maintaining an acceptable FDR, we used two rounds of analysis (**Methods**). We identify dramatically
118 more novel unannotated protein groups with statistical significance in this two-round approach than in a
119 traditional single-round search, while the number of false positive identifications is comparable
120 (**Supplemental File 1A**). The substantial enrichment (around ten-fold) of unannotated ORFs relative to
121 either the observed number of false positives or the desired FDR strongly supports the validity of the two-
122 round approach.

123 In total, after additional filtering (see **Methods**), our two-round search identified 3123 unannotated
124 unique peptides that underlie 993 unannotated ORFs without homology to annotated proteins (**Figure 1A;**
125 **Figure 1B; Figure 1–Source Data 1**; **Figure 1–Source Data 2**), which we hereafter refer to as
126 unannotated translated ORFs (utORFs). utORFs range in length from 14 amino acids (aa) to over 100 aa,
127 with the longest being 1748 aa (**Figure 1C**). The median length of a utORF is 37 aa, and the mean is 48.6
128 aa. Among these utORFs, 331 proteins have canonical start codons with 1027 peptides that originate from

129 sequence after the canonical start codon, implying that on average each protein has 3 supporting peptide-
130 spectrum matches. The mean and median lengths of a utORF starting with a canonical start codon are
131 53.5 aa and 37 aa, respectively, which is very similar to utORFs overall. Unless specifically mentioned,
132 we analyze henceforth the set of 993 utORFs and not the 331 with canonical start codons.

133 Compared to the annotated *D. melanogaster* proteome, utORF sequences have similar amino acid
134 composition (**Figure 1D**; Spearman rank-order correlation of amino acid frequencies, utORFs to
135 annotated proteome: 0.767, $p$ = 7.98 E -5). The amino acid composition for those utORFs with start
136 codons is also highly similar. Some hydrophobic and positively-charged amino acids, such as leucine,
137 arginine, and tryptophan are overrepresented relative to annotated proteins, while negatively-charged
138 amino acids, such as aspartate and glutamate are underrepresented. Overall, the distribution of positively-
139 charged to negatively-charged amino acids is significantly different among utORFs versus annotated *D.*
140 *melanogaster* protein sequences (Chi-squared test with Yates's correction, Chi-sq = 1463.55, df=1, $p$ <
141 machine epsilon). In addition, the distribution of hydrophobic amino acids versus polar amino acids is
142 also highly significantly different (same test, Chi-sq = 44.96, df=1, $p$ < 1E -10). These differences may be
143 explained by the amino acid composition of potential ORFs, as the implied amino acid composition of all
144 potential ORFs is more similar to utORFs than annotated proteins (Spearman rank-order correlation of
145 amino acid frequencies, potential ORFs to utORFs: 0.970, $p$ = 1.71 E -12; potential ORFs to annotated
146 proteome: 0.785, $p$ = 4.15 E -5).

147 All 993 utORFs have unique genomic locations (**Figure 1–Source Data 3**). Unannotated translated ORFs
148 reside in a range of genomic locations, including intergenic, intronic, or UTRs (**Figure 1E**). About one-
149 third of utORFs are located intergenically; of these, about one-third again are longer than 50 aa and
150 contain a canonical start codon. We propose that these may be single-exon unannotated proteins. A
151 considerable number of utORFs overlap with annotated genes, with an even distribution between sense
152 and anti-sense directions relative to the annotation. Of the former, several are located either upstream or
153 downstream of the canonical CDS; these may be instances of uORFs or translation read-through, which
154 have been previously described as having important biological functions (Andrews and Rothnagel 2014).
155 Notably, with the exception of a modest reduction in the relative proportion of sense utORFs, the
156 distribution of locations is not different for those utORFs with a canonical start codon. The genomic
157 sequences that make up utORFs on average have greater GC content compared that of the overall *D.*
158 *melanogaster* genome (X and autosomes; mean utORF GC content 43.4% vs. X and autosome GC
159 content 42.15%; 1-sample two-tailed $t$-test, $p$<1E-4), but they have wide variety, with some with as much
160 as 62% GC content and others as few as 21% (**Figure 1G**). Intriguingly, the GC content of annotated
161 genes (annotated gene GC content 42.9%) is between that of the genome and utORFs, and utORFs do not

162 have significantly different GC content compared to annotated genes (1-sample two-tailed *t*-test, $p >$
163 0.05). This suggests either that a selective pressure may act towards increasing the GC content in utORFs,
164 or that utORFs may be more likely to originate from regions with higher GC content.

165 Approximately half of utORFs (566 of 993; 57.0%) have a majority coiled-coil secondary structure, with
166 the remainder a mixture of beta sheets and alpha helices. Notably, this remains true even when
167 normalizing for length, as the weighted proportion of sequence in a predicted coiled-coil structure is on
168 average about half (**Figure 1F**). Predictions of structural disorder of utORFs suggest that while they are
169 rather disordered, most retain a substantial proportion that is ordered. The median proportion of
170 disordered utORFs is 24.5%.

171 **Novel peptides identified are not artifacts caused by polymorphism in conserved genes**

172 It is possible that polymorphisms within a population can yield unique peptides due to amino acid
173 substitutions, thus resulting in inaccurate attribution of such peptides to utORFs (Faridi, Purcell, and Croft
174 2018). However, such polymorphisms can contribute at most a small fraction of the peptides that support
175 the identified utORFs. To quantify this, using a trie, we calculated the minimum distance between the
176 peptides that uniquely support utORFs and the expected set of peptides from the reference proteome (see
177 **Methods**). Of the 3123 tryptic peptides that are uniquely identified from utORFs, only 1 could possibly
178 result from a single amino acid substitution of a tryptic peptide from an annotated protein; none could
179 result from a single amino acid insertion or deletion. A simultaneous double amino acid substitution could
180 explain an additional 5 peptides; 24 more peptides could result from a double insertion or deletion. Thus,
181 single or double amino acid substitutions can explain at most 30 of 3123 unique tryptic peptides (< 1%).
182 As a result, it is exceedingly likely that the peptides detected here are the product of utORFs rather than
183 polymorphisms at annotated genes.

184 **Inferring ages of unannotated translated ORFs**

185 In order to compare the conservation of utORFs across species, it is necessary to find the orthologous
186 sequence with near base-pair-level resolution. Due to the algorithmic compromises necessary for whole-
187 genome searches, BLAST homology searches do not guarantee that they will return optimum results,
188 especially for short sequences that may be highly diverged. Moreover, any homology search that solely
189 relies on searching for a given query sequence necessarily neglects broader genomic context like synteny.
190 Conversely, common genomic utilities like UCSC's *liftOver* are explicitly not recommended for fine-
191 detailed genomic coordinate conversions. To address these issues, we used a pre-existing multiple-
192 sequence-alignment (MSA) to guide a maximally-optimal local genome alignment (**Figure 2 AB**). This

193 allows the combination of both whole genomic contexts as well as the algorithmic guarantees of the
194 Smith-Waterman alignment.

195 To estimate a null model of protein similarity scores, we calculated control distributions with varying
196 assumptions. Assuming amino acids drawn from the distribution of amino acids in the annotated *D.*
197 *melanogaster* proteome, 10000 simulated pairs of randomly-generated protein sequences have a mean
198 expected similarity score ranging from 1.7 to 2.0 (**Figure 2C**). We chose a conservative threshold of 2.5
199 points as a threshold for determining significant sequence similarity, as it is approximately at least 2
200 standard deviations above the mean expectation.

201 We used the above-described protein similarity score to identify significant orthologs across related
202 *Drosophila* and insect species using the UCSC 27-way MSA (**Figure 3A**). We inferred gene ages for
203 each utORF, where the inferred origin of a utORF corresponded to the last common ancestor between
204 *Drosophila melanogaster* and the most distantly related species where a significant ortholog was
205 identified. We find that utORFs have a wide range of inferred gene ages, ranging from utORFs specific
206 just to *D. melanogaster* to utORFs that may be conserved broadly through ancient insect taxa (**Figure**
207 **3B**). The plurality of utORFs likely predate the last common ancestor of *Drosophila* with the most-distant
208 significant ortholog in *D. grimshawi* or any of its sister species; the next largest group predate the last
209 common ancestor of the *Drosophila* species group, with the most-distant significant ortholog in *D.*
210 *ficusphila* or its sister species. Notably, gene age inferences are not dramatically affected by choice of
211 significance threshold (**Figure 3–figure supplement 1**).

212 **Latent class analysis reveals subpopulations of unannotated translated ORFs**

213 Latent class analysis (LCA) is a statistical method that supposes that the population under study is
214 comprised of a mixture of distinct subpopulations – the "latent" classes (Collins and Lanza 2009).
215 Importantly, this is the only major assumption necessary, so the analysis is robust to the shape of
216 distributions or correlations between variables. Moreover, LCA can be used both deductively and
217 inductively; its primary output is a table of probabilities for categorical variables conditional on class
218 membership. Thus, these conditional probabilities can be used deductively to describe differences in the
219 inferred classes, and they can be used inductively to predict class membership given an individual set of
220 observations.

221 Due to the unbiased strategy of identification, it is likely that the unannotated translated ORFs identified
222 here consist of a mixture of loci resulting from different, distinct evolutionary processes, ranging from
223 intergenic *de novo* genes to alternative open reading frames of conserved annotated genes (Samandi et al.
224 2017). In addition, certain key variables, like inferred gene age and genomic location, are inherently

225 categorical. Rather than simple inter-variable correlations, LCA is a valuable lens for understanding the
226 unannotated loci identified here.

227 Accordingly, latent class analysis suggests that the population of high-confidence unannotated translated
228 ORFs can be subdivided into five classes (Methods; **Table 1; Figure 4-Source Data 1**) and
229 interpretations assigned based off differences in the various factor probabilities conditional on class
230 membership (Collins and Lanza 2009).

231 Class 1 loci are notable for minimal transcription (**Figure 4C**) as well as a marked bias of being
232 intergenic or antisense (**Figure 4A**). They also are almost exclusively of intermediate length between 20
233 and 50 residues (**Figure 4B**). This supports an interpretation as putatively nonfunctional loci. Class 2 loci
234 are remarkable for monophyly (**Figure 4C**), likely due to very recent inferred emergence (**Figure 4D**),
235 suggesting that these may be *melanogaster*-specific loci. Dramatic differences in overall conservation
236 (**Figure 4E**) as well as subtler ones in inferred emergence times (**Figure 4D**) distinguish classes 3, 4, and
237 5. Class 5 is also remarkable for strong bias towards locations overlapping existing, annotated genes, as
238 well as high conservation, supporting an interpretation as ORFs in alternative reading frames. Finally,
239 class 3 can be distinguished from class 4 through a more recent inferred age of emergence approximately
240 at the last common ancestor of the *melanogaster* subgroup (**Figure 4D**) and through remarkably low
241 conservation that is even more frequent than that of *melanogaster*-specific ORFs. This supports a
242 distinction between class 3 as fast-evolving ORFs and class 4 as general unannotated ORFs.

243 With inferred classes in hand, it is possible to use the classes for forward inferences by assigning each
244 utORF to a class and comparing the classes. Importantly, while binning continuous variables is necessary
245 for conducting latent class analysis, comparisons between classes can be done on the original continuous
246 variables, removing that potential source of bias, and additional variables not originally used for assigning
247 classes can be used as well. The posterior probabilities of class membership for each utORF suggest
248 generally clean partitioning between classes, as each utORF generally has a high posterior probability of
249 belonging to only a single class (**Figure 4–figure supplement 1**). The exceptions are utORFs with
250 membership in either class 4 (unannotated ORFs) or class 5 (alternative-frame ORFs). However, allowing
251 an additional sixth class leads to overfitting (see **Methods**). Accordingly, we use the maximum posterior
252 probability per utORF to assign class membership for further analysis.

253 For certain parameters, differences by inferred class membership recapitulate expected trends. For
254 example, phastCons conservation scores vary by class, with the lowest conservation in fast-evolving and
255 *melanogaster*-specific utORFs (**Figure 5A**). Alternative-frame utORFs are very well conserved due to the
256 additional constraints imposed by their overlapping annotated genes. Despite not being included as a

257 variable in LCA, phyloP conservation scores also vary significantly by class (**Figure 5B**). Negative
258 phyloP scores are consistent with faster evolution; 61.4% (27 of 44) are from fast-evolving utORFs and
259 24.1% (15 of 44) are from *melanogaster*-specific utORFs. Importantly. while phastCons and phyloP
260 scores are significantly correlated, only phyloP scores can differentiate between relative low conservation
261 and accelerated evolution. This recapitulation of faster evolution supports the overall latent class analysis
262 approach. Finally, consistent with the assigned interpretation, *melanogaster*-specific and fast-evolving
263 utORFs are the youngest according to phylostratigraphy (**Figure 5C**). Thus, the fastest-evolving utORFs,
264 which are the fast-evolving utORFs with negative phyloP scores, are surprisingly not the youngest
265 utORFs, as the distinct *melanogaster*-specific utORFs are younger.

266 For other parameters, differences between inferred classes reveals surprising trends. Fast-evolving
267 utORFs and putatively nonfunctional utORFs are significantly shorter than all others (one-sided Brunner-
268 Munzel nonparametric test of lengths vs. all other utORFs, $p << 0.001$ for both). Interestingly,
269 *melanogaster*-specific utORFs are not significantly shorter (same test, $p$=0.11), despite prior reports that
270 younger genes tend to be shorter. With regards to maximum observed transcription, alternative-frame
271 utORFs tend to be significantly better-transcribed than all other utORFs (one-sided Brunner-Munzel
272 nonparametric test of log-transformed TPMs vs. all other utORFs, $p << 0.001$). This is likely due to active
273 transcription of the overlapping pre-existing loci. In contrast, fast-evolving utORFs are significantly less
274 well-transcribed (same test, $p < 0.001$), while differences in *melanogaster*-specific and utORFs are
275 statistically insignificant. Consistent with these trends, alternative-frame utORFs are less tissue- and
276 developmentally-specific (one-sided Brunner-Munzel test of tau across FlyAtlas2 tissues, $p < 0.001$; same
277 test of tau across modENCODE developmental stages, $p < 0.001$) while fast-evolving utORFs are more
278 specifically-expressed (tau across FlyAtlas2 tissues, $p = 0.0039$; tau across modENCODE developmental
279 stages, $p < 0.001$). Unusually, a plurality of fast-evolving utORFs (33 of 96, 34.4%) have their highest
280 expression in adult fly tissues in the brain, with the next most common location the testis (10 of 96,
281 10.4%). Moreover, while transcription of annotated genes is high in both the brain and testis, fast-
282 evolving utORFs are dramatically less-transcribed in the testis than in the brain (75th percentile $\log_{10}$ of
283 TPM in brain, fast-evolving utORFs: 0.172; vs. annotated genes: 1.320; in testis, fast-evolving utORFs: -
284 0.662; vs. annotated genes 1.177; delta $\log_{10}$ of TPM in testis versus in brain, fast-evolving utORFs
285 vs. annotated genes: one-sided Mann-Whitney U test $p$ = 7.5E-6; **Figure 5–figure supplement 1**).
286 Indeed, with the exception of alternate-frame utORFs, most utORFs are not transcribed in the testis but
287 are transcribed in the brain, while most annotated genes are highly transcribed in both (**Figure 5–figure**
288 **supplement 1**). This is somewhat intriguing as previous work suggests that evolutionarily young genes –
289 particularly *de novo* genes – emerge in the testis.

290 When considering only utORFs with canonical start codons, the results are generally qualitatively similar
291 (**Figure 4–figure supplement 2, Figure 5–figure supplement 2**). Five latent classes remain the best fit,
292 as allowing a sixth class has unfavorable AIC and BIC effects. Class assignments can proceed with
293 similar logic; however, utORFs with start codons are longer overall (**Figure 4–figure supplement 2B**),
294 so some class interpretations change somewhat. Class 3 appears biased towards intergenic locations and
295 longer length, so we assign it an interpretation of "intergenic ORFs" (**Supplemental File 1B**).
296 Partitioning between classes remains good (**Figure 4–figure supplement 3**). Turning once again towards
297 forward inferences, we recapitulate the patterns seen for the conservation of *melanogaster*-specific and
298 alternative-frame ORFs. While differences in length are apparent between classes (**Figure 5–figure**
299 **supplement 2D**), they are not statistically significant; the *melanogaster*-specific class still is not
300 significantly shorter (one-sided Brunner-Munzel nonparametric test of lengths vs. all other utORFs:
301 intergenic ORFs, longer, $p = 0.11$; alternate-frame, shorter, $p = 0.15$; *melanogaster*-specific, shorter, $p =$
302 $0.36$; putatively nonfunctional, shorter, $p = 0.20$). Finally, utORFs with canonical start sites remain less
303 transcribed in the testis than the brain (delta log base 10 of TPM in testis versus in brain, utORFs with
304 canonical start sites vs. annotated genes: one-sided Mann-Whitney U test $p = 7.3E-21$). Non-canonical
305 start initiation is increasingly recognized, especially for short ORFs; evidence for translation initiation at
306 sites other than AUG has come from both ribosome profiling and mass spectrometry approaches from a
307 variety of model species (Chen et al., 2020; Ingolia et al., 2011; Ma et al., 2014; Wu et al., 2020). Since
308 we recapitulated similar results for utORFs with canonical start sites as for all utORFs, we continued with
309 all utORFs.

310 **Many unannotated translated ORFs have evidence consistent with a *de novo* origin**

311 A genetic locus emerging through *de novo* gene birth from nongenic sequence would be expected to be
312 detected in a lineage-specific manner with detectable genomic homology despite evidence inconsistent
313 with functional conservation. That is, a utORF may be of *de novo* origin if there exists at least one species
314 where an orthologous ORF is not significant despite genomic homology. Ideally, the strongest inferences
315 of *de novo* origin would rely on multiple outgroup species. This protects against an erroneous *de novo*
316 inference due to a homology detection failure rather than true functional changes, which has been
317 extensively described as a potential source of erroneous identifications (Moyers and Zhang 2015, 2016,
318 2018). Many of our utORFs have such stringent evidence consistent with a *de novo* origin (**Figure 6A**),
319 with at least two supporting outgroups originating from different phyla. Between 40 and 75% of
320 identified loci across all inferred classes may be *de novo*, with the greatest proportion in the fast-evolving,
321 *melanogaster*-specific, and unannotated ORFs. These differences between classes are statistically
322 significant (Pearson's Chi-squared test, 4 degrees of freedom, $X^2 = 31.941$, $p = 1.97E-6$). For all utORFs

323 with potential *de novo* origin, 394 (67.8%) have 3 or more such supporting outgroup species. The number
324 of such supporting outgroup species appears to be more affected by the availability of such than by
325 genomic divergence, as the youngest classes (*melanogaster*-specific and fast-evolving) have more
326 supporting outgroup species than the comparatively older alternative-frame class (**Figure 6B**).

327 Finally, we considered the frequency by which utORFs of potential *de novo* origin is lost. Since we
328 consider the protein sequence to be equivalent to the function, we use similar logic as for finding support
329 for *de novo* origin, but looking only at species descended from the inferred potential de novo origin event.
330 We find that losses of utORFs of potential *de novo* origin are extremely common; for all utORFs with
331 potential *de novo* origin, 470 (80.9%) have at least 1 potential inferred loss (**Figure 6C**).

332 **Observed retention times of many peptides underlying unannotated translated ORFs are similar to**
333 **those of peptides underlying annotated proteins**

334 One possible weakness of the methodology used in this study is that peptide-spectrum matches (PSMs)
335 are prone to false identifications (Nesvizhskii and Aebersold 2005). One method that has been proposed
336 to evaluate PSMs is to use the retention time of identified peptides on the liquid chromatography system
337 upstream of the mass spectrometer; retention time should vary as a function of the peptide's sequence.
338 While calculating retention time of a given peptide *a priori* remains an unsolved problem, substantial
339 progress has been made on predicting retention times of sets of peptides *a posteriori*. To evaluate the
340 validity of PSMs of utORFs, we used AutoRT (Wen *et al*. 2020) to predict retention times of peptides
341 supporting annotated FlyBase proteins as well as utORFs. While some peptides supporting utORFs have
342 drastically larger differences from the predicted retention times (**Figure 7–figure supplement 1**), a
343 sizeable number have observed retention times that are similar to peptides from annotated FlyBase
344 proteins. Specifically, 385 of 3123 peptides (12.3%) supporting utORFs have at least one PSM with a
345 delta retention time that is smaller than the 95th percentile of the delta retention times of all PSMs that
346 support annotated FlyBase proteins; within this subset, the difference between predicted and observed
347 retention times is modest (**Figure 7A**). These peptides are sufficient to support 270 utORFs, where
348 melanogaster-specific ORFs are slightly more likely and nonfunctional ORFs somewhat less likely to be
349 supported (**Supplemental File 1C**); however, this observation is not statistically significant (Chi-squared
350 test with 4 degrees of freedom; Chi-sq = 4.45; $p > 0.05$).

351 **utORFs can be independently validated**

352 To further validate that the discovered utORFs are not artifactual, we collected our own data to develop
353 independent empirical support for these utORFs. We collected targeted MS data from mixed 0-24 hour *D.*
354 *melanogaster* embryos (RAL517 strain), targeting a manually-curated set of 15 target utORFs chosen
355 both for suitability for targeted MS (e.g. few missed cleavages; no post-translational modifications) as

356 well as expression in embryo from modENCODE data. Of these 15 targeted utORFs, we were able to
357 detect supporting evidence for 1 through the observation of 3 transitions (**Figure 7B**). We also collected
358 shotgun MS data from mixed 0-24 hour embryos; we were able to detect supporting evidence for 3
359 additional utORFs from the shotgun MS data at a FDR of 1% (**Figure 7-Source Data 1**); this shotgun
360 dataset was fairly shallow, as we identified 1963 annotated proteins compared to the approximately 8000
361 protein groups reported by Casas-Vila et al. (2017). Given the low sensitivity of MS, having evidence for
362 utORFs from independent datasets supports an unneglectable translation level for at least a subset of the
363 utORFs. Furthermore, we re-analyzed published ribosome profiling data (Patraquim et al. 2020; Zhang et
364 al. 2018) and identified supporting evidence from an additional 243 utORFs. Of these, alternate-frame
365 ORFs are unsurprisingly the most frequently supported, perhaps due to translation of the canonical ORF,
366 but the other classes are well-represented with the exception of the putatively nonfunctional utORFs
367 (**Figure 7–figure supplement 2**). The biases across classes are statistically significant (Pearson's Chi-
368 squared test, 4 degrees of freedom, $X$^2 = 41.569, $p$ << 0.001), and they are consistent with our proposed
369 interpretation of the classes. Finally, we specifically looked for evidence of framing in ribosome profiling
370 data by implementing the binomial statistical test used for the first report of framing in *Drosophila*
371 (Patraquim et al., 2020). Of 81 utORFs with more than 2 unique footprints, 17 utORFs (21.0%) have
372 statistically-significant biases towards the observed frame in at least 1 stage of embryogenesis, supporting
373 the notion of active translation (alpha 0.05; 11 utORFs at alpha 0.001).

374 Together, this evidence supports our identification of the pool of utORFs overall. We accordingly
375 integrated all data that we analyzed with respect to additional lines of evidence supporting our utORFs
376 and factors of interest in order to identify utORFs of highest priority for downstream functional analyses;
377 we used a combination of high transcription from RNAseq atlases, supporting validation from either our
378 MS experiments or framed ribosome profiling reads, and conservation metrics. By calculating ranks
379 across each of these factors and averaging ranks per utORF, we generated a priority list for our utORFs
380 (**Figure 7 - Source Data 2**). Of the top 50 utORFs, 24 are alternate-frame ORFs, 14 unannotated ORFs, 7
381 melanogaster-specific ORFs, and 4 fast-evolving ORFs. This is similar to the overall proportion of class
382 assignments, except alternate-frame ORFs are somewhat over-represented due a higher incidence of
383 framed ribosome profiling reads. These utORFs may be the most promising candidates for future
384 experimental work.

385 **Discussion**

386 **Identification of hundreds of potential protein products of *de novo* origin**

387 In this work, we identify almost one thousand possible utORFs from a systematic, ORF-focused MS-first
388 computational approach. These utORFs are unlikely to be annotation failures due to the lack of sequence
389 homology to annotated proteins. Instead, like many similar "orphan genes", these may be of *de novo*
390 origin: the vast majority have a syntenic sequence conservation pattern that is consistent with *de novo*
391 origin (**Figure 6AB**). The total number of identifications is larger than previous studies of *de novo* genes
392 within a relatively short timescale in *Drosophila* (Begun et al. 2006; Chen, Zhang, and Long 2010; Zhao
393 et al. 2014), and is consistent with the identification of thousands of possible *de novo* polypeptides from
394 systematic studies in other biological systems (Ruiz-Orera et al. 2018; Stein et al. 2018; Durand et al.
395 2019). Moreover, while evidence of translation of proposed *de novo* genes is generally lacking
396 (McLysaght and Hurst 2016), in this case, the method of identification allows a stronger inference of
397 translation.

398 Calling *de novo* origin for a given gene is challenging, as the call is almost wholly dependent on negative
399 results in the outgroup lineages. For example, in small phylogenies with a single outgroup, it is possible
400 for a single false negative call to lead to an incorrect inference of *de novo* origin. Recent studies have
401 demonstrated that false negatives may be alarmingly likely, especially when using simple BLAST-based
402 phylostratigraphy (Moyers and Zhang 2015, 2016). Here, our method of calling *de novo* origin employs
403 proposed best practices of incorporating syntenic information (Arendsee et al. 2019; Vakirlis and
404 McLysaght 2019) and is robust to individual sequence homology calls, improving our confidence in our
405 inferences. Importantly, our inferences of gene age and potential *de novo* origin do not rely at all on the
406 absence of sequence from the MSA, which would be analogous to the lack of a significant BLAST hit.
407 Instead, our inferences of *de novo* origin rely on protein similarity scores below our threshold of
408 significance (**Figure 6B**), which ensures that these inferences are not simply due to extensive genomic
409 divergence beyond the limit of detection. If anything, our formulation of the protein similarity score may
410 be vulnerable to calling spurious similarity due to shared genomic ancestry between closely related
411 species. However, such errors would cause us to infer systematically older *de novo* origin dates for each
412 gene; given that we still conclude that many utORFs are of potential *de novo* origin, the impact of such
413 possible errors is low.

414 Other mechanisms beyond *de novo* origin could potentially produce "orphan" / lineage-specific genes.
415 For example, horizontal gene transfer from microorganisms like viruses has been reported in *Drosophila*
416 (Verster et al., 2019), and much attention has been focused on sequence divergence (e.g. after
417 pseudogenization) as an alternative explanation for lineage-specific genes (Moyers and Zhang, 2015

418 Weisman et al., 2020). When, as is often the case for our utORFs, there are many supporting outgroups,
419 multiple pseudogenization events are necessary by parsimony to produce the observed pattern; while this
420 seems unlikely, we cannot exclude the possibility of repeated substantial divergence. However, work
421 explicitly focusing on the possibility that sequence divergence is responsible for lineage-specific genes
422 recently concluded that extensive divergence was not likely the dominant mechanism (Vakirlis et al.,
423 2020).

424 Our confidence in our identifications of utORFs is supported by our confirmatory results from mass
425 spectrometry and published ribosome profiling, including the observation of statistically significant
426 framing from ribosome-profiling data. While the absolute number of recapitulated utORF detections is
427 relatively low, mass spectrometry has notoriously low power, even in well-studied proteomes like humans
428 (Baker et al. 2017). Thus, it is not particularly surprising that we were unable to recapitulate most utORFs
429 in a relatively heterogenous sample, as we expect utORFs to be of low abundance and potentially only
430 expressed under specific time- and tissue- limited conditions. Unlike sequencing approaches, the
431 empirical readout of mass spectrometry is not directly mappable to information (i.e. the underlying
432 nucleotide sequence), so false identifications of peptide-spectrum matches are a notorious problem in the
433 field (Nesvizhskii and Aebersold, 2005). However, the peptides underlying our identification of utORFs
434 are not artifacts caused by polymorphism in conserved genes. Moreover, our results suggest that the
435 reverse-decoy approach for controlling the number of false-positive identifications remains valid for our
436 two-round method: the power for the identification of annotated Drosophila genes is nearly the same
437 between a standard MS approach and our two-round approach (**Supplemental File 1A**). Further
438 functional investigation of utORF products would be a potential extension of this work (**Figure 7 -**
439 **Source Data 2**).

440 **Class inference reveals heterogenous evolutionary dynamics**

441 In contrast to previous work, our approach of applying latent class analysis allows additional refinement
442 beyond that achievable by simple per-variable analyses. For example, previous studies have described an
443 association between gene age and length: namely, that younger genes tend to be shorter (Carvunis et al.
444 2012; Neme and Tautz 2013; Wilson et al. 2017). Here, we find that the youngest class of utORFs (the
445 *melanogaster* specific class) are not significantly shorter; instead, the fast-evolving and the nonfunctional
446 classes are the shortest overall (**Figure 5C**). Importantly, while length is an input into the latent class
447 analysis, it is not the most obvious differentiator between the classes, nor is it the primary basis upon
448 which we interpreted the classes. Since the fast-evolving class is still quite young, it is possible that
449 previous reports of an association between gene age and length are related to a conflation of fast evolution
450 with gene age. Importantly, our analysis is not able to evaluate whether there is a direction in the

historical evolutionary trajectory of the length of a given utORF. Thus, we cannot distinguish between the possibilities that shorter utORFs are faster evolving or that faster evolving ORFs require a shorter length. This question merits further examination in future.

Previous studies, including our own, have suggested that evolutionarily young genes tend to be expressed in the testis (Levine et al. 2006; Begun et al. 2007; Soumillon et al. 2013; Zhao et al. 2014; Luis Villanueva-Cañas et al. 2017; Witt et al. 2019), though other tissues, like the brain, can give rise to evolutionarily young genes (Chen et al., 2012). In contrast to that, the utORFs identified here are expressed more in the brain than in the testis (**Figure 5–figure supplement 1**) in each inferred class. Moreover, while most annotated genes are highly expressed in the testis, most utORFs are not expressed at all in the testis, yet they are expressed in the brain (**Figure 5–figure supplement 1**). Considering that most previously-reported *de novo* genes also tend to be expressed in the testis (Levine et al. 2006; Begun et al. 2007; Zhao et al. 2014; Luis Villanueva-Cañas et al. 2017; Witt et al. 2019), our results may suggest that the proportion of protein-coding *de novo* genes might be higher in brain than in testis, or that our MS-first approach allows the detection of otherwise-missed protein-coding *de novo* genes. Furthermore, these utORFs may have biologically significant effects in the brain. For example, the maximum observed expression of a plurality of the relatively short, fast-evolving ORFs is in the brain, suggesting that some could act as neuropeptides.

Despite the success of our latent class analysis approach in elucidating distinct evolutionary trajectories, half of our identified utORFs of interest remain consolidated in a single class. However, additional classes cannot be safely supported due to overfitting. Additional approaches may be necessary in order to further characterize these utORFs.

**Turnover in utORFs may be frequent**

While *de novo* gene birth has been increasingly well-studied over recent years, *de novo* gene death has been examined less-closely. It is obvious that *de novo* gene death must be relatively frequent, as otherwise the net size of proteomes would be unrealistically large relative to the annotation. Consistent with this intuition, we observe that most *de novo* genes diverge by enough to be indistinguishable from random sequences in at least part of the clade descended from the last common ancestor after the *de novo* gene birth event (**Figure 6C**). Since we consider conservation of the protein-coding sequence to be approximately equivalent to the function of the locus, we consider such sequence divergence as evidence of gene death.

An implication of most models of *de novo* gene origins is that *de novo* gene birth is rare at the individual gene level and thus occurs only once in evolutionary history. However, the very large number of gene

483 death events suggests a second possibility. It is possible that the rates of gene birth and death may be
484 more balanced; that is, instead of being born once and recurrently lost, a subset of *de novo* genes may be
485 born, die, and subsequently reborn (Schmitz et al., 2018; Stein et al., 2018; Durand et al., 2019). Thus,
486 some *de novo* gene evolution may not proceed unidirectionally, but rather like a random walk along a
487 continuum. This is consistent with the proto-gene model (Carvunis et al. 2012) or the pre-adaptation
488 model (Wilson et al. 2017). Under the proto-gene model, the frequent deaths of our utORFs would imply
489 that they are proto-genes and not genes. However, with the exception of the *melanogaster*-specific
490 utORFs, our utORFs are often conserved across species with significant divergence times, which implies
491 that some may have evolutionary constraints and may in fact be functional genes. It is difficult to make
492 strong conclusions here without getting trapped in the philosophical quagmire of the nature of "gene-
493 ness"; nevertheless, it is possible that no single "grand unified theory" of *de novo* gene origination can
494 describe all phenomena.

495 **Outlook**

496 Together, our results show that evolution of young proteins may progress along different, distinct
497 trajectories in *Drosophila*. Whether similarly distinct trajectories are apparent in other model species such
498 as yeast or mammals remains to be seen. Of note, *Drosophila* is a taxon of multicellular organisms with a
499 large effective population size, so selective processes are more efficient; mammals – especially primates
500 and *Homo* – are evolutionarily young and have a smaller effective population size, while yeasts are
501 unicellular. If these factors affect general evolutionary properties, such as the selective cost of translation
502 of lowly functional proteins and the probability of fixation by drift, it is possible that they may affect the
503 evolution of *de novo* proteins. In the case of *Homo*, all these factors may be more favorable to the fixation
504 of new *de novo* proteins, and the availability of broad and varied -omics data is unparalleled. It would
505 therefore be an obvious extension to employ a similar approach to investigate possible utORFs and *de*
506 *novo* proteins in humans.

507 **Methods**

508 **Comprehensive *Drosophila melanogaster* ORFeome**

509 To generate the comprehensive database of potential ORFs in *Drosophila melanogaster*, we translated the
510 repeat-masked genome (UCSC Aug. 2014 BDGP Release 6 + ISO1 MT/dm6) and transcriptome
511 (FlyBase, r6.15). We used the repeat-masked genome to control the number of transposable elements and
512 other degenerate candidates. We retained only ORFs of at least 8 residues; for ORFs not containing a
513 canonical start codon, only those of at least 20 residues were retained. We deduplicated potential ORF
514 sequences, leaving 4,583,941 unique potential ORFs.

515 **Discovery of utORFs in published shotgun MS data**

516 To identify evidence of utORFs, we reanalyzed a published proteome of *D. melanogaster* (PRIDE
517 accession numbers PXD005691 and PXD005713; respectively, whole lifecycle and hourly/bihourly
518 embryogenesis stages; (Casas-Vila et al. 2017)) using MaxQuant v. 1.6.1.0 (Tyanova, Temu, and Cox
519 2016). To speed computation, we divided analysis of the overall proteome into 4 subsets (**Figure 1 -**
520 **Source Data 4**). As in (Casas-Vila et al. 2017), we used standard parameters (namely, methionine
521 oxidation and N-terminal acetylation as variable modifications; carbamidomethylation as a fixed
522 modification; tryptic digestion; minimum peptide length 7).

523 *Two-round MS search*

524 In the two-round MS search, we used two rounds of analysis per data subset to improve total sensitivity
525 while maintaining an acceptable false discovery rate (**Figure 1A**). In the first "discovery" round, we set
526 the peptide and protein group FDRs in MaxQuant to 0.2, with the protein search sequences set to be the
527 annotated *D. melanogaster* proteome (FlyBase, r6.15), supplemented with the comprehensive potential
528 ORF database. We then identified candidate ORFs from any of the four data subsets (**Figure 1-Source**
529 **Data 4**) to develop a filtered subset of potential ORFs with possible mass spectrometry evidence. These
530 relaxed parameters allow us to identify many more potential ORFs, but with a corresponding increase in
531 false positive identifications. In the second "verification" round, we set the peptide and protein group
532 FDRs to 0.01, with the protein search sequences set to be the annotated *D. melanogaster* proteome
533 (FlyBase, r6.15), supplemented with only that filtered subset of potential ORFs identified in the first
534 round (around 1000-fold fewer). We subsequently performed additional manual curation of the identified
535 hits to arrive at our final set of utORFs; in brief, we dropped utORFs that were shorter than 14 residues,
536 that had detectable sequence homology to annotated proteins via BLASTP, and the shorter of multiple
537 possible utORFs that were identified from the same protein group by MaxQuant.

538 *Evaluation of peptide and protein polymorphisms*

539 Evaluating whether a given peptide could have arisen from a polymorphism is equivalent to calculating
540 the number of substitutions from said peptide to the nearest peptide expected from the annotated
541 proteome. Variations of this problem are well-studied in computer science, and trie data structures are
542 commonly used to solve them efficiently. We used a Python implementation of tries
543 (https://github.com/jfjlaros/dict-trie) to build a trie of all tryptic peptides that would be expected from the
544 annotated *D. melanogaster* proteome. For each peptide uniquely identified from a utORF, we calculated
545 Hamming (the number of substitutions without the possibility of indels) and Levenshtein (the combined
546 number of substitutions and indels) distances.

547 *Genomic analysis of utORFs*

548 To map utORF sequences back to their genomic locations, we used BLAT version 36 (Hinrichs et al.
549 2006) to identify the locations of utORF sequences that were supported in the discovery round. We
550 reformatted the output to a GTF and made certain manual corrections; this file, containing the final
551 locations, is appended as **Figure 1–Source Data 3**. We excluded those loci whose genomic location
552 could not be determined from genomic analyses. For those loci where multiple genomic locations were
553 possible, we used the first location identified by BLAT.

554 We subsequently classified the genomic locations of these utORF sequences using custom scripts relying
555 on the Bioconductor libraries in R (Huber et al. 2015). Specifically, we classified a location as sense or
556 antisense if any part of the location overlapped with an annotated gene in the appropriate direction. We
557 used the FlyBase r6.15 annotations. We classified a location as intergenic if and only if it was neither
558 sense nor antisense as defined above.

559 We computed conservation scores by using the 27-way phastCons and phyloP bigWig tracks provided by
560 the UCSC Genome Browser (Hinrichs et al. 2006). For a given genomic location, we calculated the
561 conservation score by averaging the score through the relevant interval.

562 *Published sequencing data*

563 We used a number of published sequencing datasets to characterize utORFs. To evaluate developmental
564 stage- and tissue-specific transcription levels, we used modENCODE (modENCODE Consortium et al.
565 2010) and FlyAtlas2 (Leader et al. 2018) data.

566 **Gene age inference**
567 *Multiple sequence alignment and genomic similarity scoring*

568 We used BioPython v. 1.69 (Cock et al. 2009) to subset the UCSC 27-way insect multiple-sequence
569 alignment (MSA) to a smaller MSA that corresponded to the location in the *D. melanogaster* genome of
570 each unannotated translated ORF plus the flanking 9bp on each side. In a pairwise fashion for each
571 species against *D. melanogaster*, we extracted genomic sequences from the MSA, dropped gaps in the
572 sequences, and calculated identity scores.

573 *Modified Smith-Waterman protein similarity scoring*

574 We calculated protein similarity scores using the Smith-Waterman algorithm with the BLOSUM62
575 matrix (Henikoff and Henikoff 1992), no gap or extension penalties, and we normalized by the length of
576 the ORF in *D. melanogaster*. To ensure score stability with respect to minor indels, we calculated protein

577 similarity scores in all three reading frames, keeping only the frame that scored best. In doing so, we
578 considered the entirety of the syntenic block irrespective of the formal boundaries of ORFs in the related
579 species; this allows us to capture the locally-optimal (within the block) similarity and tolerate recent
580 acquisition of nonsense or frameshift mutations.

581 *Phylostratigraphy*

582 Using the reference phylogeny corresponding to the UCSC 27-way insect MSA, we inferred approximate
583 gene emergence dates from the pattern of modified Smith-Waterman protein similarity scores. We
584 defined significance as a score greater than 2.5 points. In turn, we estimated the inferred age of a given
585 ORF as the age of the last common ancestor between *D. melanogaster* and the most distant insect species
586 for which a significant modified Smith-Waterman protein similarity score was calculated.

587 **Latent Class Analysis**

588 We performed latent class analysis (Collins and Lanza 2009) using the *poLCA* R package (Linzer and
589 Lewis 2011). Specifically, we used 100 iterations of a latent class model with no covariates and with
590 factors as follows: inferred gene age; monophyly; genomic location (classified as solely intergenic,
591 antisense, sense, or any combination); phastCons score (binned as [0, 0.2), [0.2, 0.8), and [0.8, 1.0]);
592 length (binned as [0, 20), [20, 50), and [50, infinity); transcription (binned as TPM > 0.1 or ⩽ 0.1 from
593 maximum per-sample TPM) tissue specificity (binned as *tau* > 0.8 or ⩽ 0.8 from FlyAtlas tissues)
594 developmental specificity (binned as *tau* > 0.8 or ⩽ 0.8 from modENCODE developmental stages). We
595 excluded loci with missing values from latent class analysis. We chose the number of classes (five) based
596 on a combination of minimizing the BIC and AIC (**Supplemental File 1D and E**) as well as model
597 interpretation, similar to recommended procedures (Collins and Lanza 2009). For the analysis of all
598 utORFs, models with 6 classes did not converge on a single maximum-likelihood solution, suggesting
599 potential under-identification, and were thus not considered.

600 We subsequently used the resulting five-class latent class model to predict the posterior class membership
601 probabilities for each classified candidate locus. We used the modal probability to assign class
602 membership. We also performed a parallel analysis of utORFs with canonical start codons, which led to
603 similar conclusions.

604 **Validation of utORFs**
605 *Retention time validation*

606 We used AutoRT (Wen et al. 2020) to predict retention times of identified peptides from both annotated
607 FlyBase proteins and utORFs. To train the model, we conducted transfer learning per each source run

608 using the published AutoRT model (originally trained on PXD006109) with 40 epochs on either an
609 NVIDIA K80 GPU or an NVIDIA V100 GPU; the training set consisted of only those annotated FlyBase
610 peptides with posterior error probabilities of 0.01 or smaller as estimated by MaxQuant. We then used the
611 per-run trained models with default parameters to predict the retention times of the peptides resulting
612 from either the annotated FlyBase proteins or utORFs.

613 *Empirical validation*

614 We collected protein samples from *D. melanogaster* 0–24-hour embryos (RAL517 strain) for validation
615 of utORFs via mass spectrometry. In brief, we homogenized dechorionated embryos with a mortar and
616 pestle in ice-cold lysis buffer (50mM Tris HCl, pH7.5; 137 mM NaCl; 0.25% NP-40; Roche cOmplete
617 protease inhibitors) before centrifugation at 18,000 rpm at 4°C and filtration through a 0.45 micron PVDF
618 membrane.

619 Purified protein was acetone-precipitated, reduced, and alkylated before digestion with LysC and trypsin,
620 acidified, and solid-phase extraction. The resulting peptides were analyzed by LC-MS/MS (120min
621 gradient, 25cm EasySprayer column in high res./high mass accuracy mode) on a ThermoFisher Lumos
622 instrument. In addition, the embryo sample was analyzed on the same instrument by PRM-MS, targeting
623 a set of 16 tryptic peptides. We manually chose these targeted peptides based on suitability of the peptide
624 for targeting (i.e., few/no missed cleavages or modifications), expression of the utORF in embryos, and
625 the strength of the prior detection from the Casas-Vila datasets.

626 We analyzed the shotgun data using ProteomeDiscoverer 1.4 and searched via Mascot against the Uniprot
627 D. melanogaster database concatenated with common contaminants and our set of utORFs with a 1%
628 peptide FDR. We analyzed the targeted data using SkylineMS (v. 21.2.0.369 (2efacf038)). We deposited
629 the raw data in PRIDE under accession PXD032197.

630 We additionally reanalyzed published ribosome profiling data downloaded from the NCBI Sequence
631 Read Archive (SRA) project PRJNA306373 (Zhang et al. 2018) and project SRP254283 (Patraquim et al.
632 2020). For the Zhang dataset, we downloaded raw reads from the above SRA accessions, aligned them to
633 the genome using hisat2 v.2.2.1 (Kim et al. 2019), and calculated TPM using stringtie v.1.3.4d (Pertea et
634 al. 2016). We handled the Patraquim dataset similarly, except that we additionally trimmed the raw reads
635 using trimmomatic 0.38 (Bolger, Lohse, and Usadel 2014) and filtered out reads aligning to annotated
636 RNAs using bowtie2 v. 2.3.5 (Langmead and Salzberg 2012) before alignment. For each utORF, we
637 calculated the maximum TPM observed across either dataset and called the utORF as supported if the
638 TPM was greater than 0.2.

To identify utORFs with statistically significant framing of ribosome-profiling reads, we used a binomial test as in Patraquim et al. (2020). We aligned the trimmed, filtered reads to the set of sequences corresponding to our utORFs, plus an additional 30bp up- and down-stream, using bowtie v. 1.3.1 (Langmead et al., 2009) with options --tryhard --suppress 1,6,7,8. We analyzed the resulting aligned footprints using the R package riboSeqR v 1.20.0. For each utORF with at least two unique footprints, we calculated the binomial probability of observing the number of footprints in the utORF's frame versus the other two frames at an uncorrected alpha of 0.05. We pooled footprints across replicates and across read lengths for a given utORF in each sample.

**Data Availability**

Raw MS data are deposited in PRIDE under accession number PXD032197. Relevant scripts and intermediate files can be found in our Github repository https://github.com/LiZhaoLab/utORF_mass_spec.

**Acknowledgements**

We thank members of the Zhao laboratory for helpful discussions during the work. We thank Henrik Molina and the Proteomics Resource Center at the Rockefeller University for the help with mass spectrometry data.

**Funding**

This work was supported by National Institutes of Health (NIH) MIRA R35GM133780, the Robertson Foundation, a Rita Allen Foundation Scholar Program, and a Vallee Scholar Program (VS-2020-35), a Monique Weill-Caulier Career Scientist Award, and an Alfred P. Sloan Research Fellowship (FG-2018-10627) to L.Z. E.B.Z. was supported by a Medical Scientist Training Program grant from the National Institute of General Medical Sciences of the National Institutes of Health under award number T32GM007739 to the Weill Cornell/Rockefeller/Sloan Kettering Tri-Institutional MD-PhD Program. The content of this study is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

**Competing interest statement**

The authors declare no competing interests.

667 **Table and Figure Legends**

668 **Table 1: Latent class analysis of all unannotated translated ORFs.**

| Class | Interpretation | Estimated Percent | Number |
|---|---|---|---|
| 1 | putatively nonfunctional loci | 4.35% | 41 |
| 2 | melanogaster-specific ORFs | 5.71% | 54 |
| 3 | fast-evolving ORFs | 12.03% | 96 |
| 4 | unannotated ORFs | 57.61% | 591 |
| 5 | alternative frame ORFs | 20.30% | 161 |

669

670 **Figure 1. Identification of utORFs and their properties**. **A**. We detected utORFs by searching via a
671 two-round approach through a comprehensive database of potential ORFs that was generated from a six-
672 frame translation of the transcriptome and genome (see Methods). **B**: an example utORF. The full
673 sequence of utORF 633.1 is shown. The peptide supporting the inference of this utORF is bolded and
674 colored red. **C**: distribution of lengths of utORFs. **D**: distribution of amino acids in utORFs (blue) and in
675 the reference D. melanogaster proteome (orange). Standard single-letter abbreviations for amino acids are
676 used. **E**: genomic locations of utORFs. Since a utORF can overlap multiple items, e.g., an ORF that
677 overlaps with an annotated intron, exon, and another gene antisense to the first, categories are not
678 exclusive. **F**: left: secondary structure composition of utORFs. Approximately half of utORFs have a
679 coiled-coil secondary structure, with the remainder a mixture of beta sheets and alpha helices. right:
680 predicted structural disorder of utORFs. **G**: Percent GC content of utORFs and GC content of the X and
681 autosomes of the *D. melanogaster* genome (red line).

682 **Figure 2. Synteny-based orthology detection and protein sequence similarity quantification. A**:
683 When performing a simple homology search for a locus of interest (red arrow) across a given genome
684 (blue line), the search space is orders of magnitude larger, requiring heuristic shortcuts to reduce the
685 computational complexity. **B**: Using a multiple sequence alignment, we can simply find the block that
686 contains the locus of interest and use that to evaluate potential orthologs in other genomes. The search
687 space is approximately a similar size as that of the original locus, so an optimal search is computationally
688 feasible. Syntenic information (i.e., flanking genes, gray arrows) is encoded within the multiple sequence
689 alignment. **C**: We calculated the pairwise protein similarity score (see **Methods**) across 10000 pairs of
690 randomly-generated sequences of lengths 20-150 with amino acid frequencies matching that of the
691 annotated *D. melanogaster* proteome. The mean score (blue line) and 2 standard deviations (shading) are
692 shown. Our significance threshold of 2.5 lies beyond 2 standard deviations from expectation.

693 **Figure 3: Inferred gene ages of the utORFs. A**: The reference phylogenetic tree used for these analyses
694 (UCSC 27-way insect alignment). Abbreviations are as follows: *D. mel*: *Drosophila melanogaster*, *D.*
695 *sim*: *D. simulans*, *D. sec*: *D. sechellia*, *D. yak*: *D. yakuba D. ere*: *D. erecta*, *D. bia*: *D. biarmipes*, *D. suz*:
696 *D. suzukii*, *D. ana*: *D. ananassae*, *D. bip*: *D. bipectinata*, *D. eug*: *D. eugracilis*, *D. ele*: *D. elegans*, *D. kik*:
697 *D. kikkawai*, *D. tak*: *D. takahashii*, *D. rho*: *D. rhopaloa*, *D. fic*: *D. ficusphila*, *D. pse*: *D. pseudoobscura*,
698 *D. per*: *D. persimilis*, *D. mir*: *D. miranda*, *D. wil*: *D. willistoni*, *D. vir: D. virilis*, *D. moj: D. mojavensis*,
699 *D. alb: D. albomicans*, *D. gri*: *D. grimshawi*, *M. dom*: *Musca domestica*, *A. gam*: *Anopheles gambiae*,
700 *Apis mel*: *Apis mellifera*, *T. cas*: *Tribolium castaneum*. **B**: The most distantly related species in which a
701 significant ortholog of a utORF exists varies. The red line illustrates the cumulative distribution of loci.
702 For convenience, sister species are grouped together under one species (e.g., *D. virilis* with *D. grimshawi*;
703 *D. eugracilis*, *D. elegans*, *D. kikkawai*, and *D. takahashii* with *D. rhopaloa*; etc.). Abbreviations as in **A**.

704 **Figure 4. Latent class analysis of the utORFs reveals differences between classes**. **A**: Class 1 is
705 notably distinct for strong bias towards intergenic and antisense locations at the expense of sense
706 locations. Class 2 is notable for being relatively unbiased, and for being the only class with appreciable
707 members in a combination of locations. Class 5 is strongly skewed towards antisense and sense locations.
708 **B**: Class 1 is almost exclusively of intermediate length. Class 4 has the greatest length bias, followed by
709 class 2. tiny: fewer than 20 residues; medium: from 20 to 49 residues; long: 50 or more residues. **C**: Class
710 1 is notably distinct from the others for minimal transcription. Low tissue and developmental specificity
711 may be an effect of minimal transcription. Class 2 is remarkable for being entirely monophyletic. Class 5
712 has slightly lower tissue and developmental specificities than classes 2-4. sigTPM: maximum per-sample
713 TPM 0.1; tau: tissue specificity 0.8; dev_tau: developmental specificity 0.8. **D**: Class 2 is by far the
714 youngest. Class 3 tends to be of an intermediate age, with inferred emergence at around the LCA of the
715 *melanogaster* subgroup (droRho). In contrast, class 4 emergence is distributed throughout the LCA of the
716 *melanogaster* subgroup and the *Drosophila* genus. **E**: Class 2 is notable for overall low conservation.
717 Class 3 is remarkably even less conserved. Class 4 and 5 are distinguished through differences in
718 intermediate vs. significant conservation. unconserved: phastCons score < 0.2; amb: phastCons score ≥
719 0.2 and < 0.8; significant: phastCons score ≥ 0.8.

720 **Figure 5. Differences between inferred classes recapitulate expected trends in age and conservation**
721 **and reveal surprising trends in lengths and expression**. **A**: As expected, phastCons conservation scores
722 vary by class. Scores near 0 indicate low conservation while scores near 1 indicate high conservation.
723 Note that fast-evolving and mel-specific loci have dramatically lower conservation whereas alt-frame loci
724 are very well conserved. **B**: Despite not being included as a variable in LCA, phyloP conservation scores
725 also vary by class. Scores below 0 indicate potential fast evolution (acceleration) while scores above 0

726 indicate conservation. Note that the lowest scores predominantly occur in fast-evolving and mel-specific
727 loci, and that alt-frame loci remain the best-conserved. **C**: Distribution of phylostratigraphically-inferred
728 gene ages by inferred latent class. **D**: Class variations in length are generally minimal, with the exception
729 of utORFs being slightly longer than average, and fast-evolving and nonfunctional loci being shorter than
730 average. **E**: Maximum observed transcription across all FlyAtlas2 and modENCODE samples varies by
731 class. **F**: Tissue specificity (tau) calculated from FlyAtlas2 data shows that nonfunctional and fast-
732 evolving loci are the most specific, with alternate frame loci the least. **G**: Developmental specificity (tau)
733 calculated from modENCODE data similarly shows that nonfunctional and fast-evolving loci are the most
734 specific, with alternate frame loci the least.

735 **Figure 6. Many utORFs have evidence consistent with a de novo origin. A**: Proportion of utORFs by
736 inferred class with genomic conservation consistent with *de novo* origin. Box widths correlate with size of
737 class (**Table 1**). **B**: Number of supporting outgroups by inferred class for loci with potential *de novo*
738 origin. **C**: Number of losses inferred after potential de novo origin.

739 **Figure 7. Independent validation of utORF identification. A:** Cumulative distribution of differences
740 between observed and predicted retention times for peptide-spectrum matches of peptides supporting
741 annotated FlyBase proteins (orange) and peptide-spectrum matches of peptides supporting utORFs that
742 are no worse than the 95$^{th}$ percentile of annotated FlyBase PSMs (blue). **B:** Transitions detected via PRM
743 of the peptide GPGAAISQR from protein extracts of *D. melanogaster* embryos.

744

745 **Supplemental Figure Legends**

746 **Figure 3–figure supplement 1:** Robustness of gene age inferences with respect to significance threshold.
747 **A**: Figure 3A, but also showing the *melanogaster* species subgroup, species group, and *Drosophila* taxa.
748 **B**: Change in furthest significant ortholog (using a significance threshold of 2.3 instead of 2.5). Coloring
749 corresponds to panel A. Lowering the threshold causes more orthologs of utORFs to be called as
750 significant, increasing the inferred gene age. **C**: As in panel B, but with a significance threshold of 2.7
751 instead of 2.5. Increasing the threshold causes fewer orthologs of utORFs to be called as significant,
752 reducing the inferred gene age.

753 **Figure 4–figure supplement 1:** Posterior probabilities per unannotated translated ORF of class
754 membership inferred from latent class analysis for all utORFs.

755 **Figure 5–figure supplement 1:** Transcription of utORFs in selected tissues. Top panel: utORFs,
756 separated by inferred LCA class. Mean TPMs across the given tissue in FlyAtlas2 are log10-transformed
757 with a pseudocount of 1E-3. Horizontal line marks an expression threshold of TPM 0.1. Bottom panel: As
758 top, but for, in order, annotated FlyBase genes, separated between those with names, those with "CG"
759 identifiers, and those with "CR" identifiers, and all utORFs.

760 **Figure 4–figure supplement 2:** Latent class analysis of utORFs with canonical start sites reveals
761 differences between classes. **A-E:** same as Figure 4, but examining utORFs with canonical start sites.

762 **Figure 4–figure supplement 3:** Posterior probabilities per unannotated translated ORF of class
763 membership inferred from latent class analysis for utORFs with canonical start sites.

764 **Figure 5–figure supplement 2:** Differences between inferred classes for utORFs with canonical start
765 sites recapitulate expected trends. **A-G**: same as Figure 5, but examining utORFs with canonical start
766 sites.

767 **Figure 7–figure supplement 1**: Cumulative distribution of differences between observed and predicted
768 retention times for every peptide-spectrum match of peptides supporting all annotated FlyBase proteins
769 (blue) and all utORFs (orange).

770 **Figure 7–figure supplement 2**: Proportion of utORFs by inferred class with supporting evidence from
771 ribosome profiling. Box widths correlate with size of class (Table 1).

772 **Supplemental File 1A:** Two-round analysis improves identifications.

773 **Supplemental File 1B**: Latent class analysis of unannotated translated ORFs with canonical start sites.

774 **Supplemental File 1C**: AutoRT-supported identifications by inferred gene class.

775 **Supplemental File 1D**: Summary statistics used for selecting the number of latent classes when
776 considering all utORFs. Five latent classes were chosen as a balance of minimizing the AIC and BIC and
777 model interpretation. AIC: Akaike information criterion; BIC: Bayesian information criterion; $G^2$:
778 likelihood-ratio statistic; Chi-sq: Chi-squared. * : Models with 6 classes did not converge on a single
779 maximum-likelihood solution within 100 replicates.

780 **Supplemental File 1E**: Summary statistics used for selecting the number of latent classes when
781 considering only utORFs with canonical start codons. Five latent classes were chosen as a balance of
782 minimizing the AIC and BIC and model interpretation. AIC: Akaike information criterion; BIC: Bayesian
783 information criterion; $G^2$: likelihood-ratio statistic; Chi-sq: Chi-squared.

784 **Source Data Files**

785 Figure 1–Source Data 1 (csv): all utORFs sequences.

786 Figure 1–Source Data 2 (csv): utORF supporting peptides.

787 Figure 1–Source Data 3 (gtf): utORF locations.

788 Figure 4 –Source Data 1 (csv): utORF inferred LCA classes.

789 Figure 7–Source Data 1 (xlsx): MASCOT search results from embryo MS data.

790 Figure 7–Source Data 2 (csv): rank of potentially biologically significant targets.

791 Figure 1–Source Data 4 (xlsx) - dataset subset mappings.

792 **References**


793 Andrews, S.J., and Rothnagel, J.A. (2014). Emerging evidence for functional peptides encoded by short
794 open reading frames. Nat Rev Genet *15*, 193–204. *https://doi.org/10.1038/nrg3520*.

795 Arendsee, Z., Li, J., Singh, U., Bhandary, P., Seetharam, A., and Wurtele, E.S. (2019). fagin: synteny-
796 based phylostratigraphy and finer classification of young genes. BMC Bioinformatics *20*, 440.
797 *https://doi.org/10.1186/s12859-019-3023-y*.

798 Baker, M.S., Ahn, S.B., Mohamedali, A., Islam, M.T., Cantor, D., Verhaert, P.D., Fanayan, S., Sharma,
799 S., Nice, E.C., Connor, M., et al. (2017). Accelerating the search for the missing proteins in the human
800 proteome. Nature Communications *8*, 1–13. *https://doi.org/10.1038/ncomms14271*.

801 Begun, D.J., Lindfors, H.A., Thompson, M.E., and Holloway, A.K. (2006). Recently evolved genes
802 identified from Drosophila yakuba and D. erecta accessory gland expressed sequence tags. Genetics *172*,
803 1675–1681. *https://doi.org/10.1534/genetics.105.050336*.

Begun, D.J., Lindfors, H.A., Kern, A.D., and Jones, C.D. (2007). Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade. Genetics *176*, 1131–1137. *https://doi.org/10.1534/genetics.106.069245*.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. *https://doi.org/10.1093/bioinformatics/btu170*.

Bungard, D., Copple, J.S., Yan, J., Chhun, J.J., Kumirov, V.K., Foy, S.G., Masel, J., Wysocki, V.H., and Cordes, M.H.J. (2017). Foldability of a Natural De Novo Evolved Protein. Structure *25*, 1687–1696.e4. *https://doi.org/10.1016/j.str.2017.09.006*.

Cai, J., Zhao, R., Jiang, H., and Wang, W. (2008). De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. Genetics *179*, 487–496. *https://doi.org/10.1534/genetics.107.084491*.

Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M.A., Yildirim, M.A., Simonis, N., Charloteaux, B., Hidalgo, C.A., Barbette, J., Santhanam, B., et al. (2012). Proto-genes and de novo gene birth. Nature *487*, 370–374. *https://doi.org/10.1038/nature11184*.

Casas-Vila, N., Bluhm, A., Sayols, S., Dinges, N., Dejung, M., Altenhein, T., Kappei, D., Altenhein, B., Roignant, J.-Y., and Butter, F. (2017). The developmental proteome of Drosophila melanogaster. Genome Research *27*, 1273–1285. *https://doi.org/10.1101/gr.213694.116*.

Chen, J., Brunner, A.-D., Cogan, J.Z., Nuñez, J.K., Fields, A.P., Adamson, B., Itzhak, D.N., Li, J.Y., Mann, M., Leonetti, M.D., et al. (2020). Pervasive functional translation of noncanonical human open reading frames. Science *367*, 1140–1146. *https://doi.org/10.1126/science.aay0262*.

Chen, S., Zhang, Y.E., and Long, M. (2010). New genes in Drosophila quickly become essential. Science *330*, 1682–1685. *https://doi.org/10.1126/science.1196380*.

Chen, S., Spletter, M., Ni, X., White, K.P., Luo, L., and Long, M. (2012). Frequent recent origination of brain genes shaped the evolution of foraging behavior in Drosophila. Cell Rep *1*, 118–132. *https://doi.org/10.1016/j.celrep.2011.12.010*.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423. *https://doi.org/10.1093/bioinformatics/btp163*.

Collins, L.M., and Lanza, S.T. (2009). *Latent Class and Latent Transition Analysis* (Hoboken, NJ, USA: John Wiley & Sons, Inc.).

Cridland, J.M., Majane, A.C., Zhao, L., and Begun, D.J. (2022). Population biology of accessory gland-expressed de novo genes in Drosophila melanogaster. Genetics *220*, iyab207. *https://doi.org/10.1093/genetics/iyab207*.

de Klerk, E., and ’t Hoen, P.A.C. (2015). Alternative mRNA transcription, processing, and translation: Insights from RNA sequencing. Trends in Genetics *31*, 128–139. *https://doi.org/10.1016/j.tig.2015.01.001*.

Domazet-Loso, T., and Tautz, D. (2003). An evolutionary analysis of orphan genes in Drosophila. Genome Research *13*, 2213–2219. *https://doi.org/10.1101/gr.1311003*.

842 Domazet-Loso, T., Brajković, J., and Tautz, D. (2007). A phylostratigraphy approach to uncover the
843 genomic history of major adaptations in metazoan lineages. Trends Genet *23*, 533–539.
844 *https://doi.org/10.1016/j.tig.2007.08.014*.

845 Durand, É., Gagnon-Arsenault, I., Hallin, J., Hatin, I., Dubé, A.K., Nielly-Thibault, L., Namy, O., and
846 Landry, C.R. (2019). Turnover of ribosome-associated transcripts from de novo ORFs produces gene-like
847 characteristics available for de novo gene emergence in wild yeast populations. Genome Res *29*, 932–
848 943. *https://doi.org/10.1101/gr.239822.118*.

849 Faridi, P., Purcell, A.W., and Croft, N.P. (2018). In Immunopeptidomics We Need a Sniper Instead of a
850 Shotgun. Proteomics *18*, e1700464. *https://doi.org/10.1002/pmic.201700464*.

851 Henikoff, S., and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. Proc Natl
852 Acad Sci U S A *89*, 10915–10919. *https://doi.org/10.1073/pnas.89.22.10915*.

853 Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M.,
854 Furey, T.S., Harte, R.A., Hsu, F., et al. (2006). The UCSC Genome Browser Database: update 2006.
855 Nucleic Acids Research *34*, D590–8. *https://doi.org/10.1093/nar/gkj144*.

856 Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S.,
857 Gatto, L., Girke, T., et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor.
858 Nat Methods *12*, 115–121. *https://doi.org/10.1038/nmeth.3252*.

859 Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome profiling of mouse embryonic stem
860 cells reveals the complexity and dynamics of mammalian proteomes. Cell *147*, 789–802.
861 *https://doi.org/10.1016/j.cell.2011.10.002*.

862 Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment
863 and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol *37*, 907–915.
864 *https://doi.org/10.1038/s41587-019-0201-4*.

865 Knowles, D.G., and McLysaght, A. (2009). Recent de novo origin of human protein-coding genes.
866 Genome Research *19*, 1752–1759. *https://doi.org/10.1101/gr.095026.109*.

867 Lange, A., Patel, P.H., Heames, B., Damry, A.M., Saenger, T., Jackson, C.J., Findlay, G.D., and
868 Bornberg-Bauer, E. (2021). Structural and functional characterization of a putative de novo gene in
869 Drosophila. Nat Commun *12*, 1667. *https://doi.org/10.1038/s41467-021-21667-6*.

870 Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*,
871 357–359. *https://doi.org/10.1038/nmeth.1923*.

872 Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient
873 alignment of short DNA sequences to the human genome. Genome Biology *10*, R25.
874 *https://doi.org/10.1186/gb-2009-10-3-r25*.

875 Leader, D.P., Krause, S.A., Pandit, A., Davies, S.A., and Dow, J.A.T. (2018). FlyAtlas 2: a new version
876 of the Drosophila melanogaster expression atlas with RNA-Seq, miRNA-Seq and sex-specific data.
877 Nucleic Acids Research *46*, D809–D815. *https://doi.org/10.1093/nar/gkx976*.

878 Levine, M.T., Jones, C.D., Kern, A.D., Lindfors, H.A., and Begun, D.J. (2006). Novel genes derived from
879 noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression.

880 Proceedings of the National Academy of Sciences of the United States of America *103*, 9935–9939.
881 *https://doi.org/10.1073/pnas.0509809103*.

882 Li, D., Dong, Y., Jiang, Y., Jiang, H., Cai, J., and Wang, W. (2010). A de novo originated gene depresses
883 budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. Cell Res
884 *20*, 408–420. *https://doi.org/10.1038/cr.2010.31*.

885 Linzer, D.A., and Lewis, J.B. (2011). poLCA : An R Package for Polytomous Variable Latent Class
886 Analysis. Journal of Statistical Software *42*. *https://doi.org/10.18637/jss.v042.i10*.

887 Long, M., Betrán, E., Thornton, K., and Wang, W. (2003). The origin of new genes: glimpses from the
888 young and old. Nature Reviews Genetics *4*, 865–875. *https://doi.org/10.1038/nrg1204*.

889 Luis Villanueva-Cañas, J., Ruiz-Orera, J., Agea, M.I., Gallo, M., Andreu, D., and Albà, M.M. (2017).
890 New Genes and Functional Innovation in Mammals. Genome Biol Evol *9*, 1886–1900.
891 *https://doi.org/10.1093/gbe/evx136*.

892 Ma, J., Ward, C.C., Jungreis, I., Slavoff, S.A., Schwaid, A.G., Neveu, J., Budnik, B.A., Kellis, M., and
893 Saghatelian, A. (2014). Discovery of Human sORF-Encoded Polypeptides (SEPs) in Cell Lines and
894 Tissue. Journal of Proteome Research *13*, 1757–1765. *https://doi.org/10.1021/pr401280w*.

895 McLysaght, A., and Guerzoni, D. (2015). New genes from non-coding sequence: the role of de novo
896 protein-coding genes in eukaryotic evolutionary innovation. Philosophical Transactions of the Royal
897 Society B: Biological Sciences *370*, 20140332. *https://doi.org/10.1098/rstb.2014.0332*.

898 McLysaght, A., and Hurst, L.D. (2016). Open questions in the study of de novo genes: what, how and
899 why. Nature Reviews. Genetics *17*, 567–578. *https://doi.org/10.1038/nrg.2016.78*.

900 modENCODE Consortium, Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., Negre, N., Eaton, M.L.,
901 Landolin, J.M., Bristow, C.A., Ma, L., et al. (2010). Identification of functional elements and regulatory
902 circuits by Drosophila modENCODE. Science (New York, N.Y.) *330*, 1787–1797.
903 *https://doi.org/10.1126/science.1198374*.

904 Moyers, B.A., and Zhang, J. (2015). Phylostratigraphic bias creates spurious patterns of genome
905 evolution. Mol Biol Evol *32*, 258–267. *https://doi.org/10.1093/molbev/msu286*.

906 Moyers, B.A., and Zhang, J. (2016). Evaluating Phylostratigraphic Evidence for Widespread De Novo
907 Gene Birth in Genome Evolution. Mol Biol Evol *33*, 1245–1256.
908 *https://doi.org/10.1093/molbev/msw008*.

909 Moyers, B.A., and Zhang, J. (2018). Toward Reducing Phylostratigraphic Errors and Biases. Genome
910 Biol Evol *10*, 2037–2048. *https://doi.org/10.1093/gbe/evy161*.

911 Neme, R., and Tautz, D. (2013). Phylogenetic patterns of emergence of new genes support a model of
912 frequent de novo evolution. BMC Genomics *14*, 117. *https://doi.org/10.1186/1471-2164-14-117*.

913 Nesvizhskii, A.I., and Aebersold, R. (2005). Interpretation of shotgun proteomic data: the protein
914 inference problem. Mol Cell Proteomics *4*, 1419–1440. *https://doi.org/10.1074/mcp.R500012-MCP200*.

915 Ohno, S. (1970). *Evolution by Gene Duplication* (Berlin, Heidelberg: Springer Berlin Heidelberg).

Patraquim, P., Mumtaz, M.A.S., Pueyo, J.I., Aspden, J.L., and Couso, J.-P. (2020). Developmental regulation of canonical and small ORF translation from mRNAs. Genome Biol *21*, 128. *https://doi.org/10.1186/s13059-020-02011-5*.

Pertea, M., Kim, D., Pertea, G.M., Leek, J.T., and Salzberg, S.L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature Protocols *11*, 1650–1667. *https://doi.org/10.1038/nprot.2016.095*.

Reinhardt, J.A., Wanjiru, B.M., Brant, A.T., Saelao, P., Begun, D.J., and Jones, C.D. (2013). De Novo ORFs in Drosophila Are Important to Organismal Fitness and Evolved Rapidly from Previously Non-coding Sequences. PLoS Genetics *9*. *https://doi.org/10.1371/journal.pgen.1003860*.

Ruiz-Orera, J., Hernandez-Rodriguez, J., Chiva, C., Sabidó, E., Kondova, I., Bontrop, R., Marqués-Bonet, T., and Albà, M.M. (2015). Origins of De Novo Genes in Human and Chimpanzee. PLoS Genet *11*, e1005721. *https://doi.org/10.1371/journal.pgen.1005721*.

Ruiz-Orera, J., Verdaguer-Grau, P., Villanueva-Cañas, J.L., Messeguer, X., and Albà, M.M. (2018). Translation of neutrally evolving peptides provides a basis for de novo gene evolution. Nature Ecology and Evolution *2*, 890–896. *https://doi.org/10.1038/s41559-018-0506-6*.

Samandi, S., Roy, A.V., Delcourt, V., Lucier, J.-F., Gagnon, J., Beaudoin, M.C., Vanderperre, B., Breton, M.-A., Motard, J., Jacques, J.-F., et al. (2017). Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. Elife *6*, e27860. *https://doi.org/10.7554/eLife.27860*.

Schlötterer, C. (2015). Genes from scratch - the evolutionary fate of de novo genes. Trends in Genetics *31*, 215–219. *https://doi.org/10.1016/j.tig.2015.02.007*.

Schmitz, J.F., Ullrich, K.K., and Bornberg-Bauer, E. (2018). Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. Nature Ecology and Evolution *2*, 1626–1632. *https://doi.org/10.1038/s41559-018-0639-7*.

Sinitcyn, P., Rudolph, J.D., and Cox, J. (2018). Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data. Annual Review of Biomedical Data Science *1*, 207–234. *https://doi.org/10.1146/annurev-biodatasci-080917-013516*.

Soumillon, M., Necsulea, A., Weier, M., Brawand, D., Zhang, X., Gu, H., Barthès, P., Kokkinaki, M., Nef, S., Gnirke, A., et al. (2013). Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. Cell Rep *3*, 2179–2190. *https://doi.org/10.1016/j.celrep.2013.05.031*.

Stein, J.C., Yu, Y., Copetti, D., Zwickl, D.J., Zhang, L., Zhang, C., Chougule, K., Gao, D., Iwata, A., Goicoechea, J.L., et al. (2018). Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. Nat Genet *50*, 285–296. *https://doi.org/10.1038/s41588-018-0040-0*.

Tautz, D., and Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. Nature Reviews. Genetics *12*, 692–702. *https://doi.org/10.1038/nrg3053*.

Tyanova, S., Temu, T., and Cox, J. (2016). The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nature Protocols *11*, 2301–2319. *https://doi.org/10.1038/nprot.2016.136*.

954 Vakirlis, N., and McLysaght, A. (2019). Computational Prediction of De Novo Emerged Protein-Coding
955 Genes. Methods Mol Biol *1851*, 63–81. *https://doi.org/10.1007/978-1-4939-8736-8_4*.

956 Vakirlis, N., Carvunis, A.-R., and McLysaght, A. (2020). Synteny-based analyses indicate that sequence
957 divergence is not the main source of orphan genes. Elife *9*, e53500. *https://doi.org/10.7554/eLife.53500*.

958 Van Oss, S.B., and Carvunis, A.-R. (2019). De novo gene birth. PLoS Genet *15*, e1008160.
959 *https://doi.org/10.1371/journal.pgen.1008160*.

960 Verster, K.I., Wisecaver, J.H., Karageorgi, M., Duncan, R.P., Gloss, A.D., Armstrong, E.E., Price, D.K.,
961 Menon, A.R., Ali, Z.M., and Whiteman, N.K. (2019). Horizontal Transfer of Bacterial Cytolethal
962 Distending Toxin B Genes to Insects. Mol Biol Evol *36*, 2105–2110.
963 *https://doi.org/10.1093/molbev/msz146*.

964 Wang, W., Brunet, F.G., Nevo, E., and Long, M. (2002). Origin of sphinx, a young chimeric RNA gene
965 in Drosophila melanogaster. Proc Natl Acad Sci U S A *99*, 4448–4453.
966 *https://doi.org/10.1073/pnas.072066399*.

967 Weisman, C.M., Murray, A.W., and Eddy, S.R. (2020). Many, but not all, lineage-specific genes can be
968 explained by homology detection failure. PLoS Biol *18*, e3000862.
969 *https://doi.org/10.1371/journal.pbio.3000862*.

970 Wen, B., Li, K., Zhang, Y., and Zhang, B. (2020). Cancer neoantigen prioritization through sensitive and
971 reliable proteogenomics analysis. Nat Commun *11*, 1759. *https://doi.org/10.1038/s41467-020-15456-w*.

972 Wilson, B.A., Foy, S.G., Neme, R., and Masel, J. (2017). Young genes are highly disordered as predicted
973 by the preadaptation hypothesis of de novo gene birth. Nature Ecology and Evolution *1*, 1–19.
974 *https://doi.org/10.1038/s41559-017-0146*.

975 Witt, E., Benjamin, S., Svetec, N., and Zhao, L. (2019). Testis single-cell RNA-seq reveals the dynamics
976 of de novo gene transcription and germline mutational bias in Drosophila. Elife *8*, e47138.
977 *https://doi.org/10.7554/eLife.47138*.

978 Wu, Q., Wright, M., Gogol, M.M., Bradford, W.D., Zhang, N., and Bazzini, A.A. (2020). Translation of
979 small downstream ORFs enhances translation of canonical main open reading frames. EMBO J *39*,
980 e104763. *https://doi.org/10.15252/embj.2020104763*.

981 Xie, C., Zhang, Y.E., Chen, J.-Y., Liu, C.-J., Zhou, W.-Z., Li, Y., Zhang, M., Zhang, R., Wei, L., and Li,
982 C.-Y. (2012). Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs.
983 PLoS Genet *8*, e1002942. *https://doi.org/10.1371/journal.pgen.1002942*.

984 Xie, C., Bekpen, C., Künzel, S., Keshavarz, M., Krebs-Wheaton, R., Skrabar, N., Ullrich, K.K., and
985 Tautz, D. (2019). A de novo evolved gene in the house mouse regulates female pregnancy cycles. Elife *8*,
986 e44392. *https://doi.org/10.7554/eLife.44392*.

987 Zhang, H., Dou, S., He, F., Luo, J., Wei, L., and Lu, J. (2018). Genome-wide maps of ribosomal
988 occupancy provide insights into adaptive evolution and regulatory roles of uORFs during Drosophila
989 development. PLoS Biol *16*, e2003903. *https://doi.org/10.1371/journal.pbio.2003903*.

990 Zhang, L., Ren, Y., Yang, T., Li, G., Chen, J., Gschwend, A.R., Yu, Y., Hou, G., Zi, J., Zhou, R., et al.
991 (2019). Rapid evolution of protein diversity by de novo origination in Oryza. Nat Ecol Evol *3*, 679–690.
992 *https://doi.org/10.1038/s41559-019-0822-5*.

Zhao, L., Saelao, P., Jones, C.D., and Begun, D.J. (2014). Origin and spread of de novo genes in Drosophila melanogaster populations. Science (New York, N.Y.) *343*, 769–772. *https://doi.org/10.1126/science.1248286*.

Figure 1

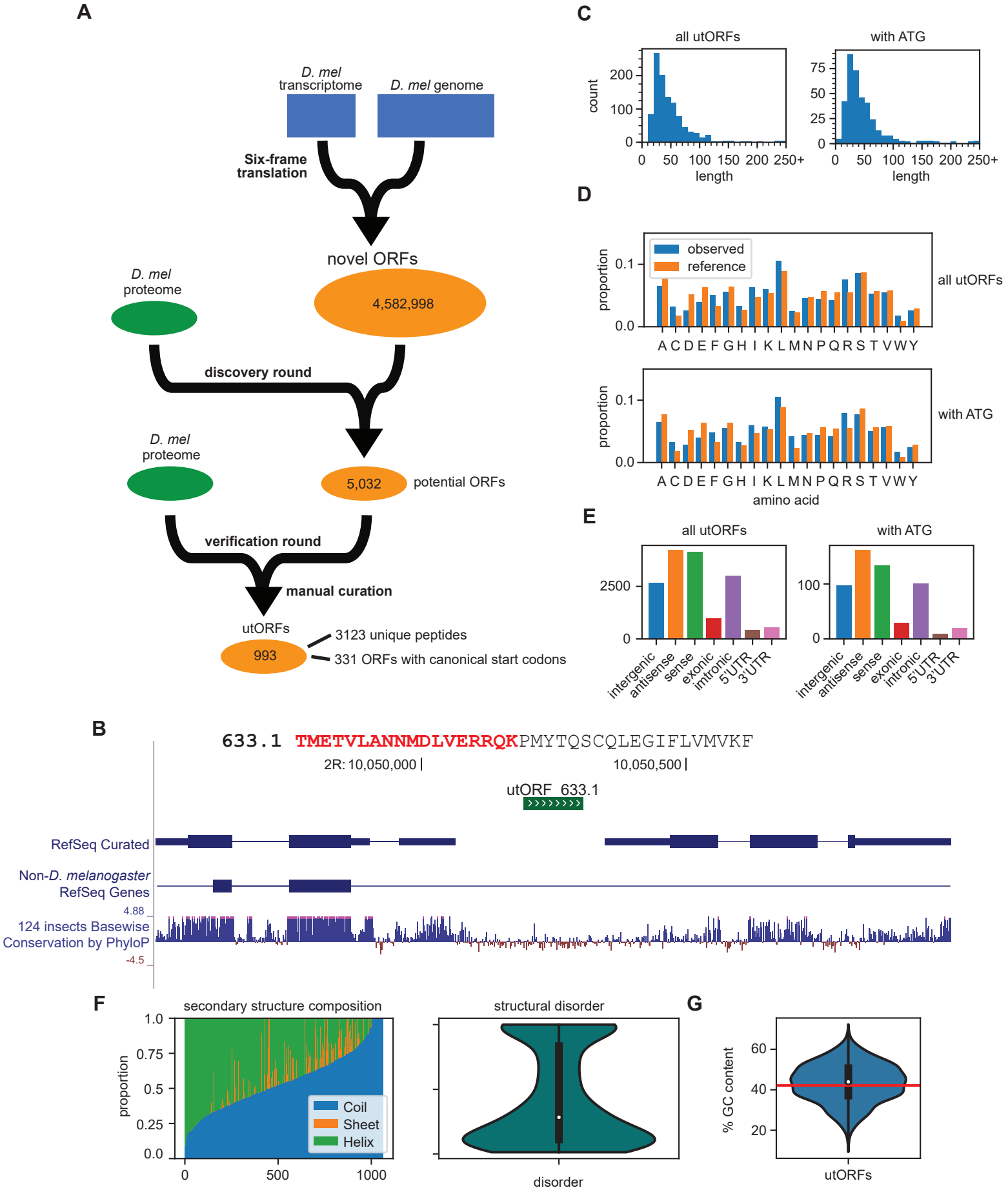

A
D. mel transcriptome
D. mel genome
Six-frame translation
novel ORFs
4,582,998
D. mel proteome
discovery round
D. mel proteome
5,032
potential ORFs
verification round
manual curation
utORFs
993
3123 unique peptides
331 ORFs with canonical start codons
B
633.1 TMETVLANNMDLVERRQKPMYTQSCQLEGIFLVMVKF
2R: 10,050,000
10,050,500
utORF_633.1
RefSeq Curated
Non-D. melanogaster RefSeq Genes
124 insects Basewise Conservation by PhyloP
4.88
-4.5
C
all utORFs
with ATG
count
length
D
observed
reference
proportion
all utORFs
with ATG
amino acid
E
all utORFs
with ATG
intergenic
antisense
sense
exonic
intronic
5'UTR
3'UTR
F
secondary structure composition
proportion
Coil
Sheet
Helix
structural disorder
disorder
G
% GC content
utORFs

Figure 2

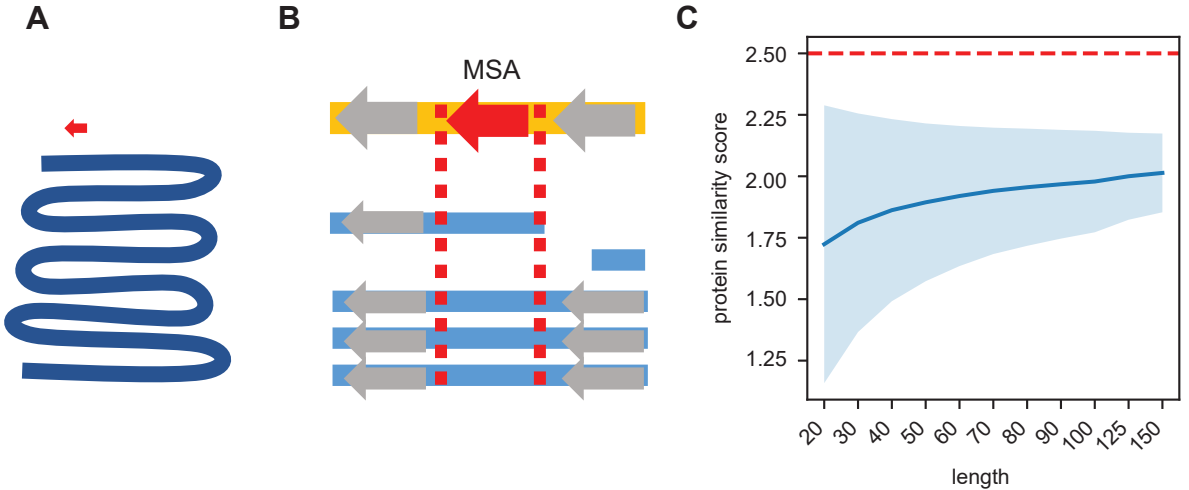

A
B
MSA
C
2.50
2.25
2.00
1.75
1.50
1.25
protein similarity score
20
30
40
50
60
70
80
90
100
125
150
length

Figure 3

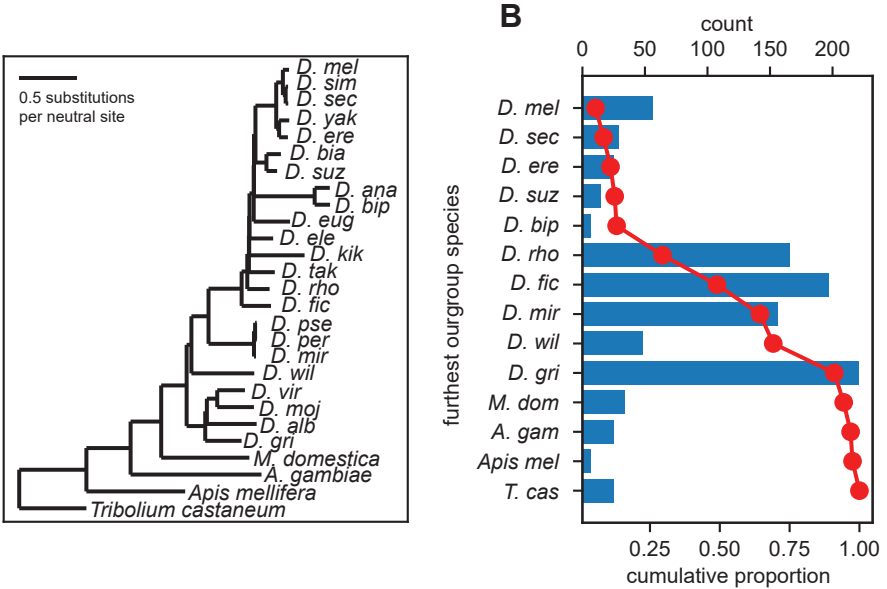

A
0.5 substitutions
per neutral site
D. mel
D. sim
D. sec
D. yak
D. ere
D. bia
D. suz
D. ana
D. bip
D. eug
D. ele
D. kik
D. tak
D. rho
D. fic
D. pse
D. per
D. mir
D. wil
D. vir
D. moj
D. alb
D. gri
M. domestica
A. gambiae
Apis mellifera
Tribolium castaneum
B
count
0
50
100
150
200
furthest ourgroup species
D. mel
D. sec
D. ere
D. suz
D. bip
D. rho
D. fic
D. mir
D. wil
D. gri
M. dom
A. gam
Apis mel
T. cas
0.25
0.50
0.75
1.00
cumulative proportion

Figure 4

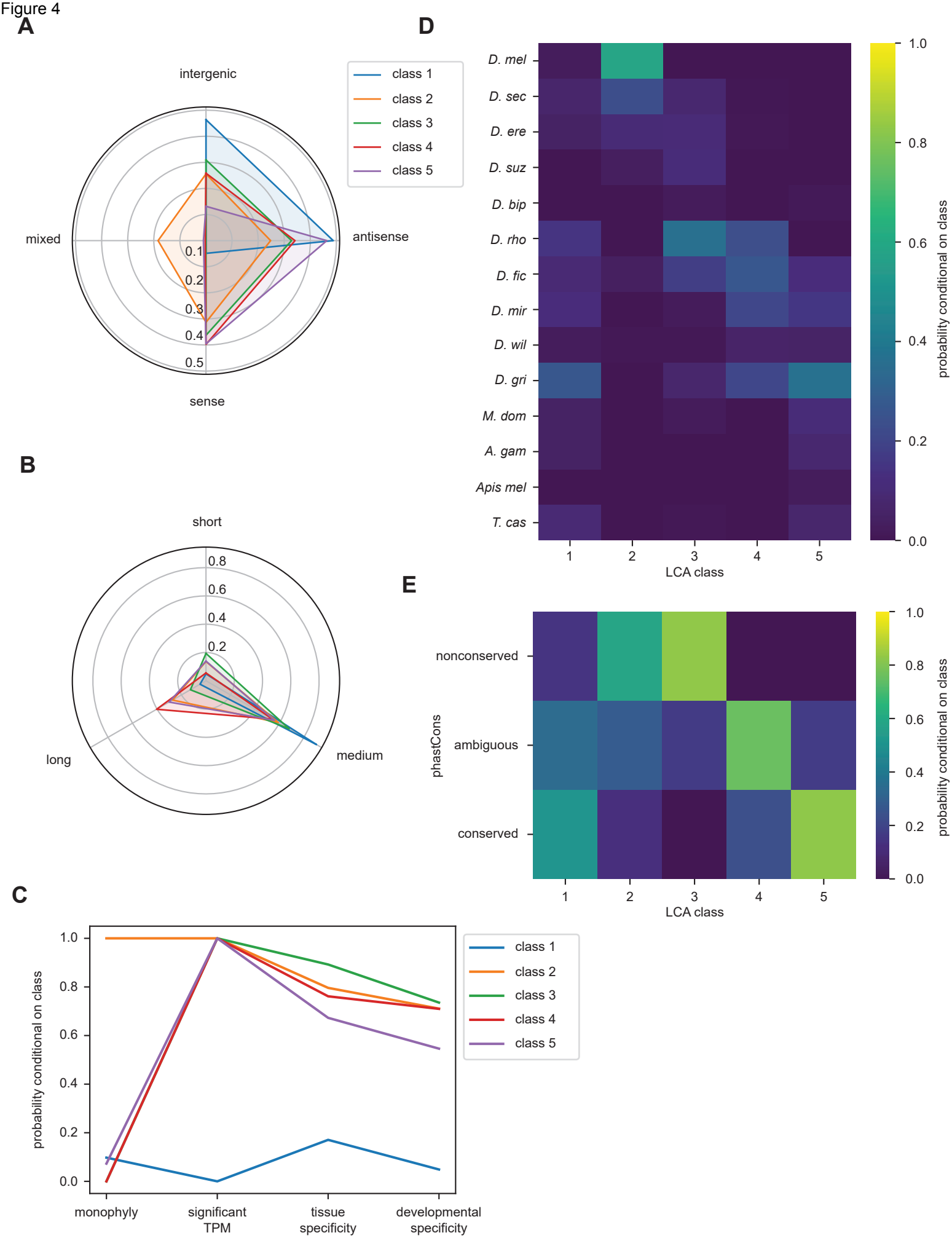

A
intergenic
antisense
sense
mixed
0.1
0.2
0.3
0.4
0.5
class 1
class 2
class 3
class 4
class 5
B
short
medium
long
0.8
0.6
0.4
0.2
C
probability conditional on class
1.0
0.8
0.6
0.4
0.2
0.0
monophyly
significant TPM
tissue specificity
developmental specificity
class 1
class 2
class 3
class 4
class 5
D
D. mel
D. sec
D. ere
D. suz
D. bip
D. rho
D. fic
D. mir
D. wil
D. gri
M. dom
A. gam
Apis mel
T. cas
1
2
3
4
5
LCA class
probability conditional on class
1.0
0.8
0.6
0.4
0.2
0.0
E
phastCons
nonconserved
ambiguous
conserved
1
2
3
4
5
LCA class
probability conditional on class
1.0
0.8
0.6
0.4
0.2
0.0

Figure 5

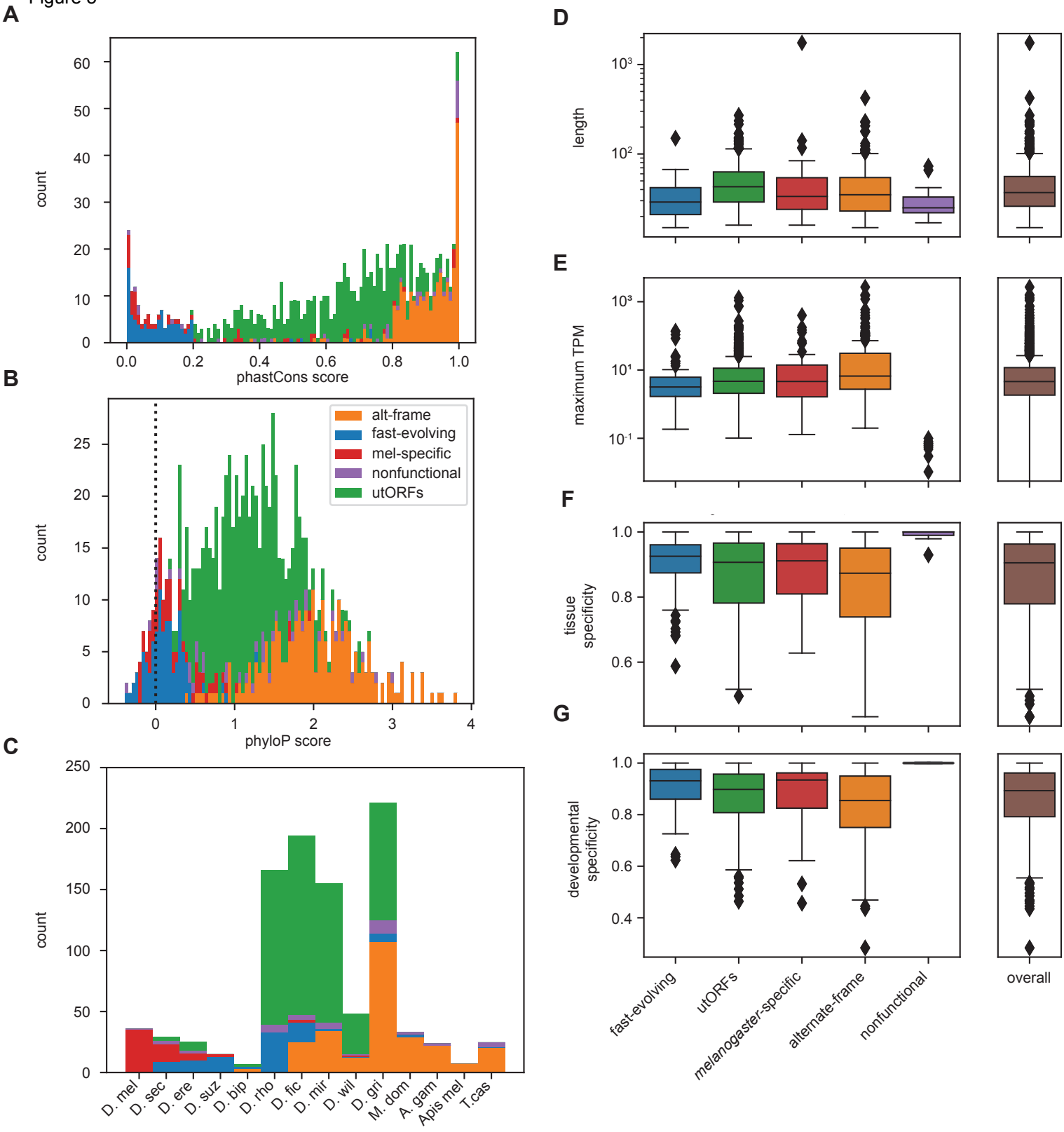

A
count
phastCons score
B
alt-frame
fast-evolving
mel-specific
nonfunctional
utORFs
count
phyloP score
C
count
D. mel
D. sec
D. ere
D. suz
D. bip
D. rho
D. fic
D. mir
D. wil
D. gri
M. dom
A. gam
Apis mel
T.cas
D
length
E
maximum TPM
F
tissue specificity
G
developmental specificity
fast-evolving
utORFs
melanogaster-specific
alternate-frame
nonfunctional
overall

Figure 6

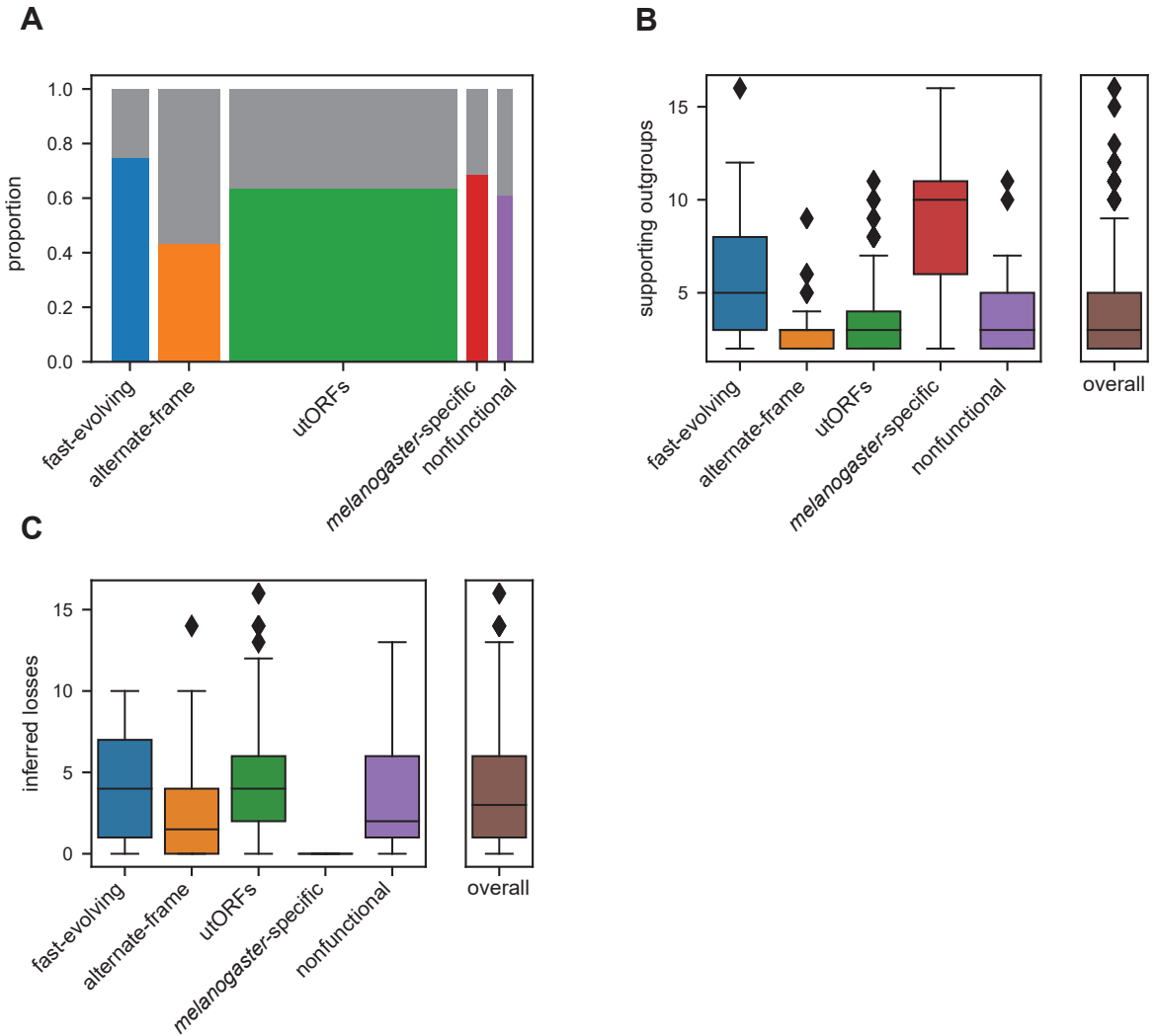

A
proportion
1.0
0.8
0.6
0.4
0.2
0.0
fast-evolving
alternate-frame
utORFs
melanogaster-specific
nonfunctional
B
supporting outgroups
15
10
5
fast-evolving
alternate-frame
utORFs
melanogaster-specific
nonfunctional
overall
C
inferred losses
15
10
5
0
fast-evolving
alternate-frame
utORFs
melanogaster-specific
nonfunctional
overall

Figure 7

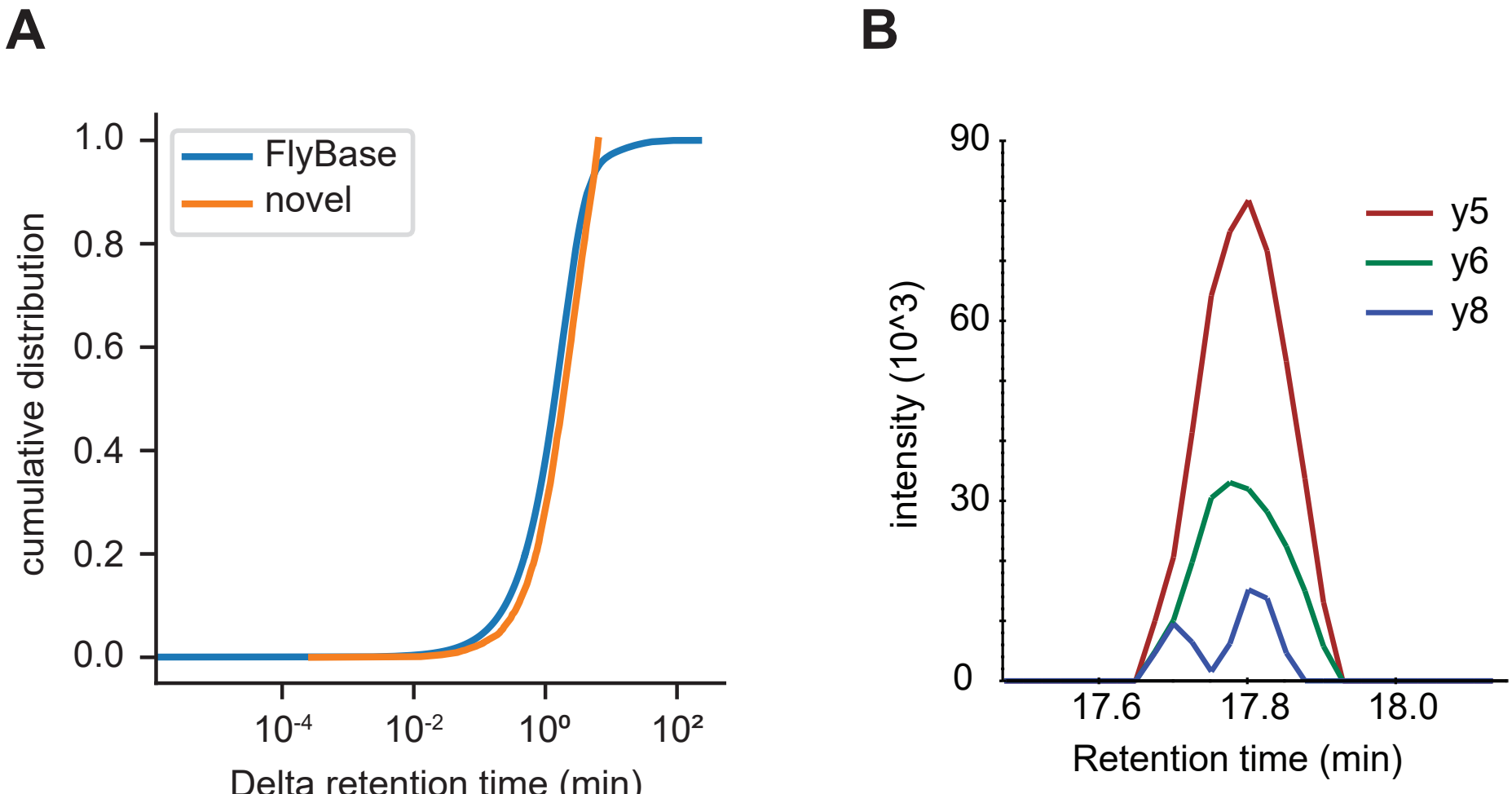

A
B
FlyBase
novel
cumulative distribution
1.0
0.8
0.6
0.4
0.2
0.0
10-4
10-2
10⁰
10²
Delta retention time (min)
y5
y6
y8
intensity (10^3)
90
60
30
0
17.6
17.8
18.0
Retention time (min)

Figure 3 - Figure supplement 1

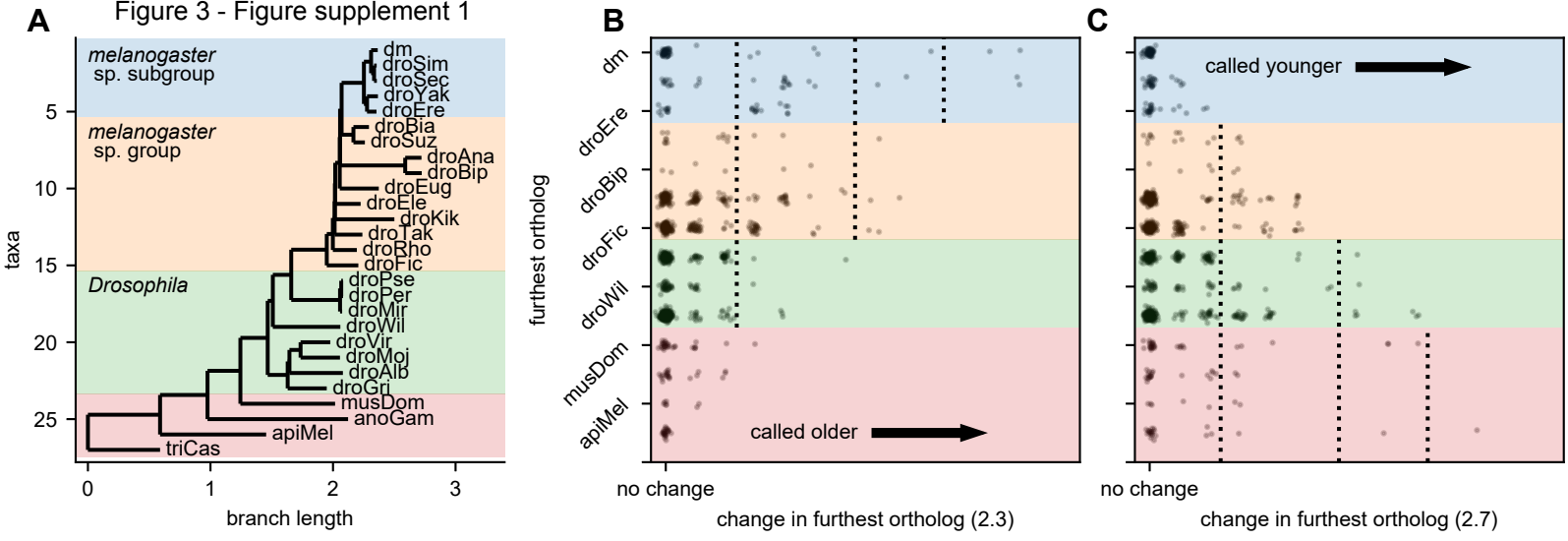


Figure 4 - Figure supplement 1

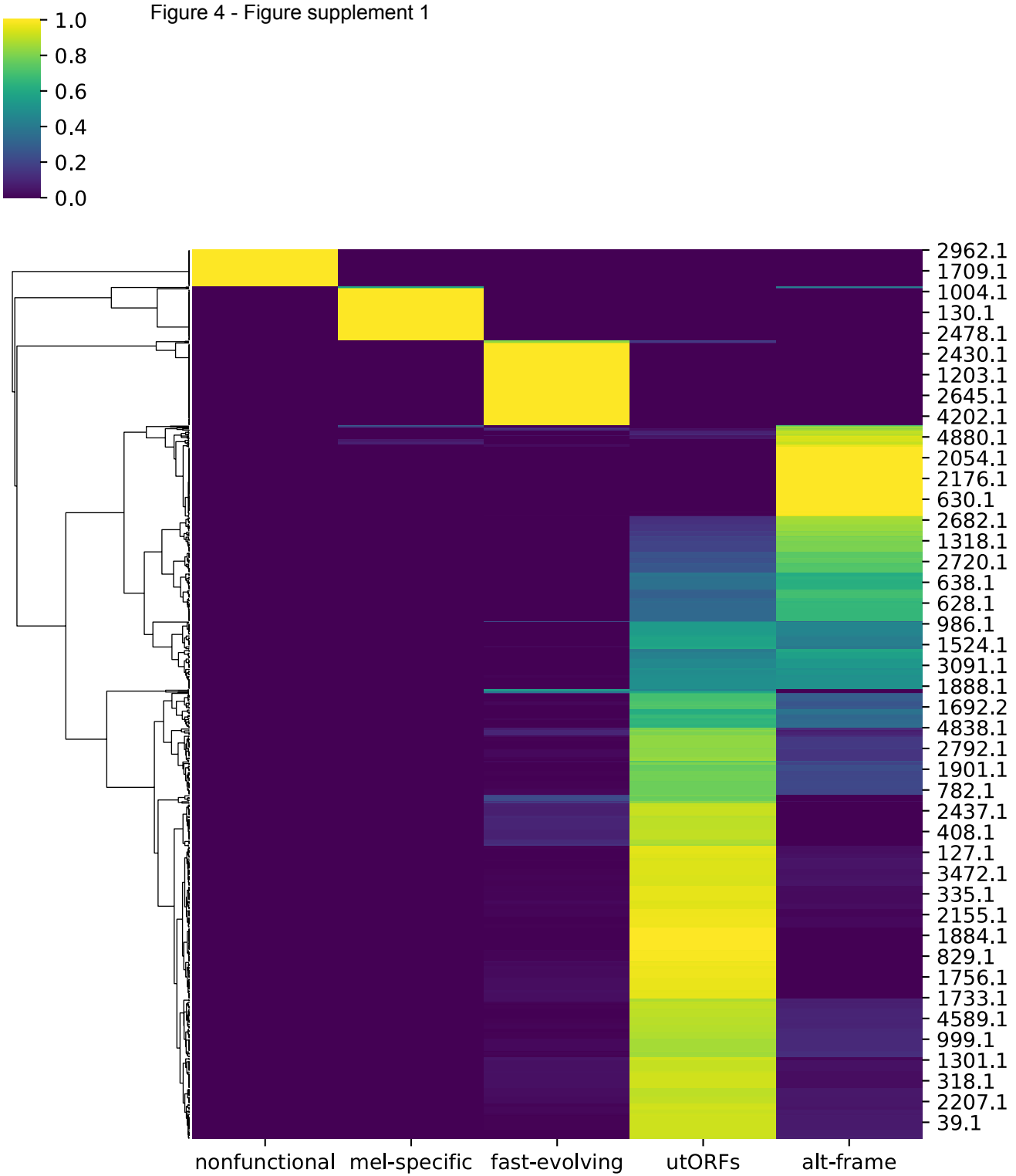


Figure 5 - Figure supplement 1

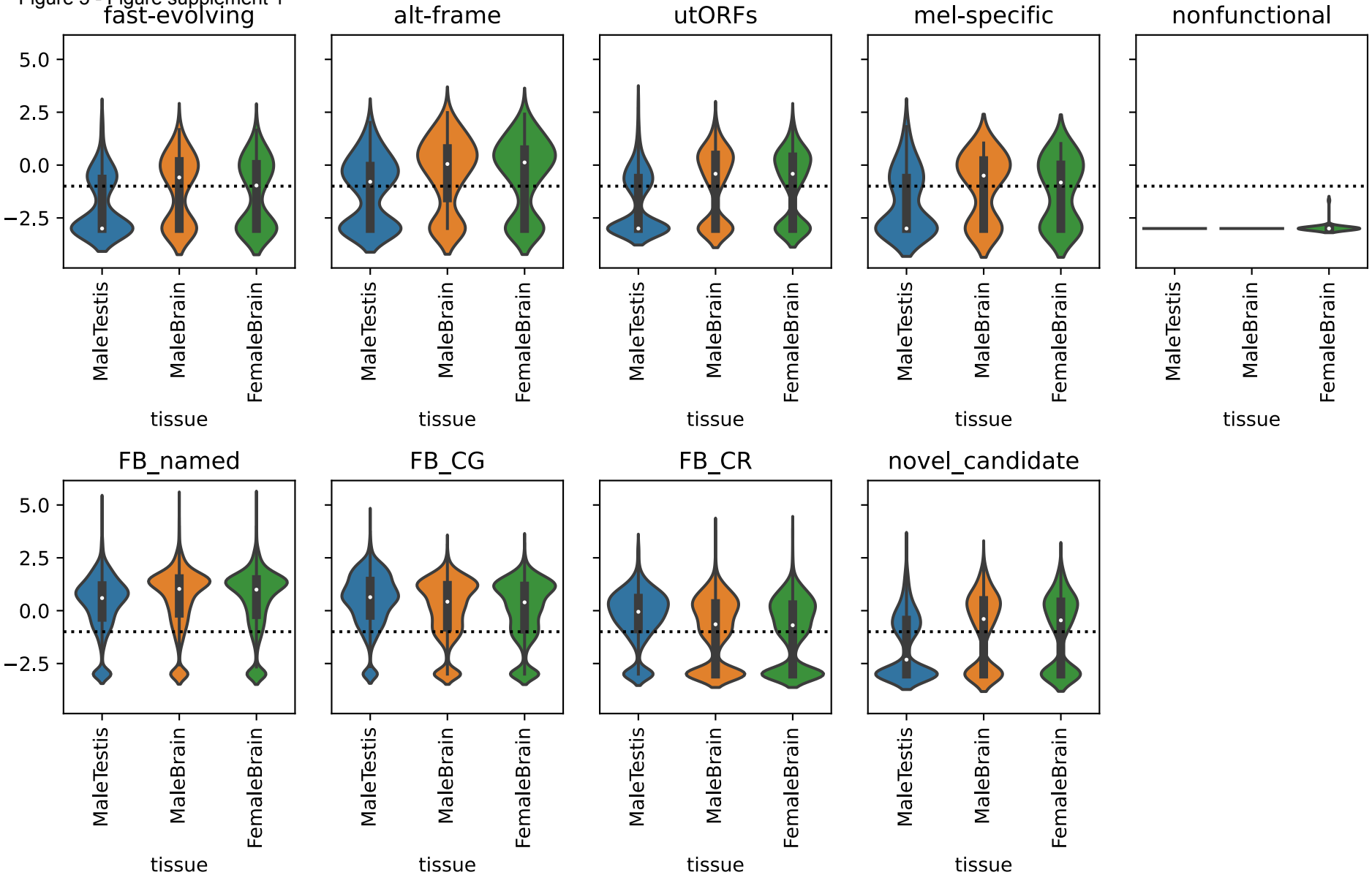


Figure 4 - Figure supplement 2

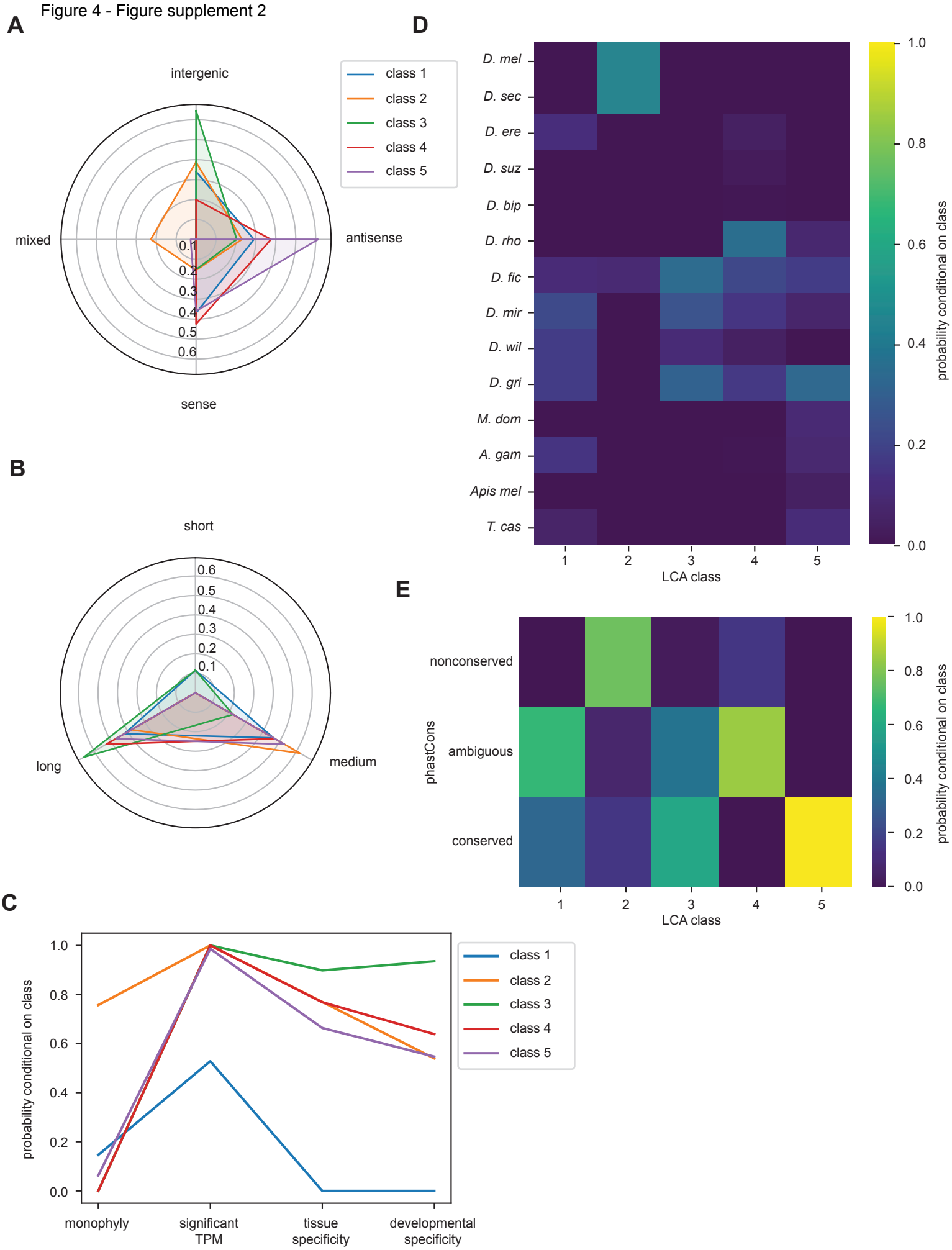


Figure 4 - Figure supplement 3

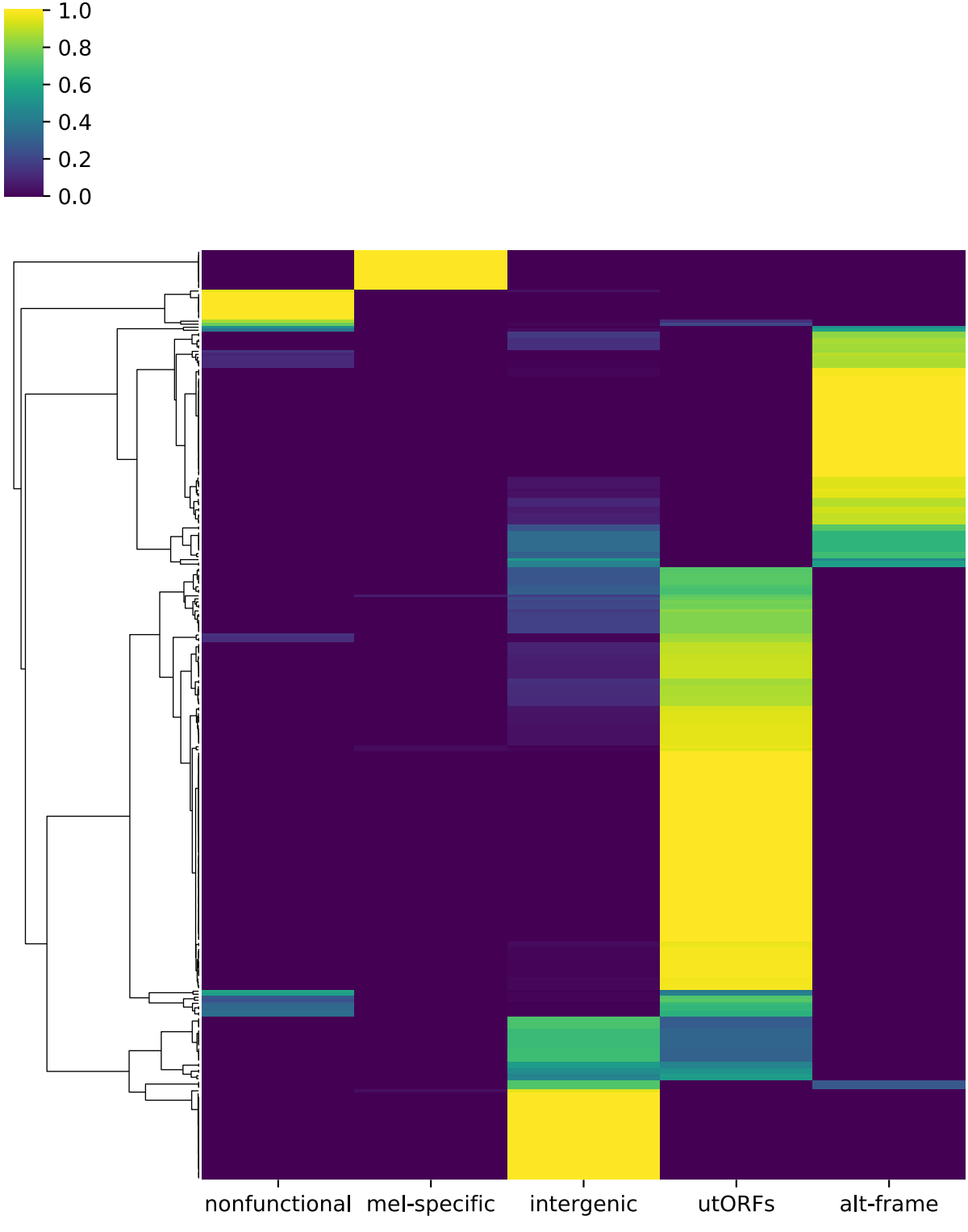
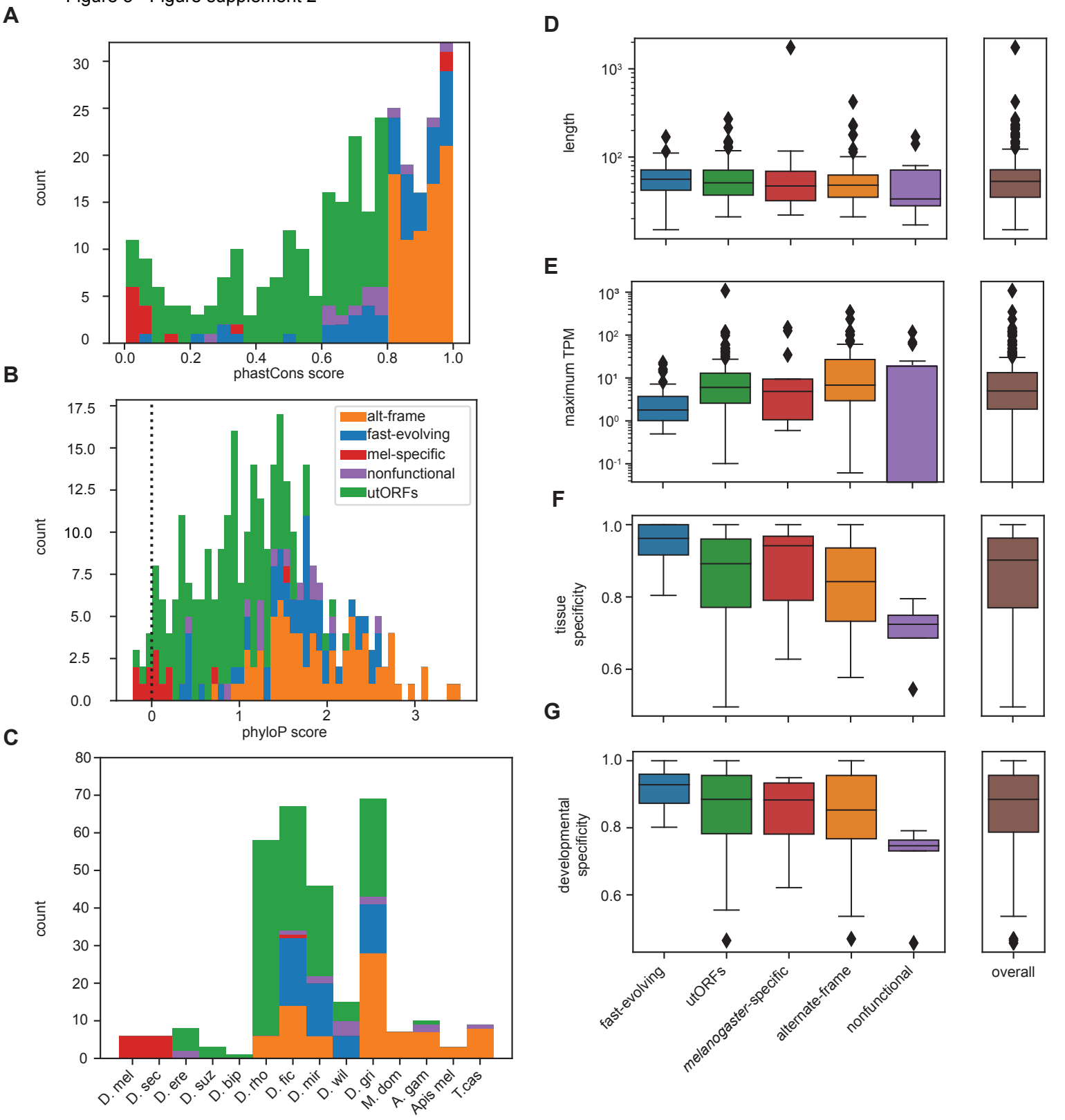


Figure 5 - Figure supplement 2



# Figure 7 - Figure supplement 1

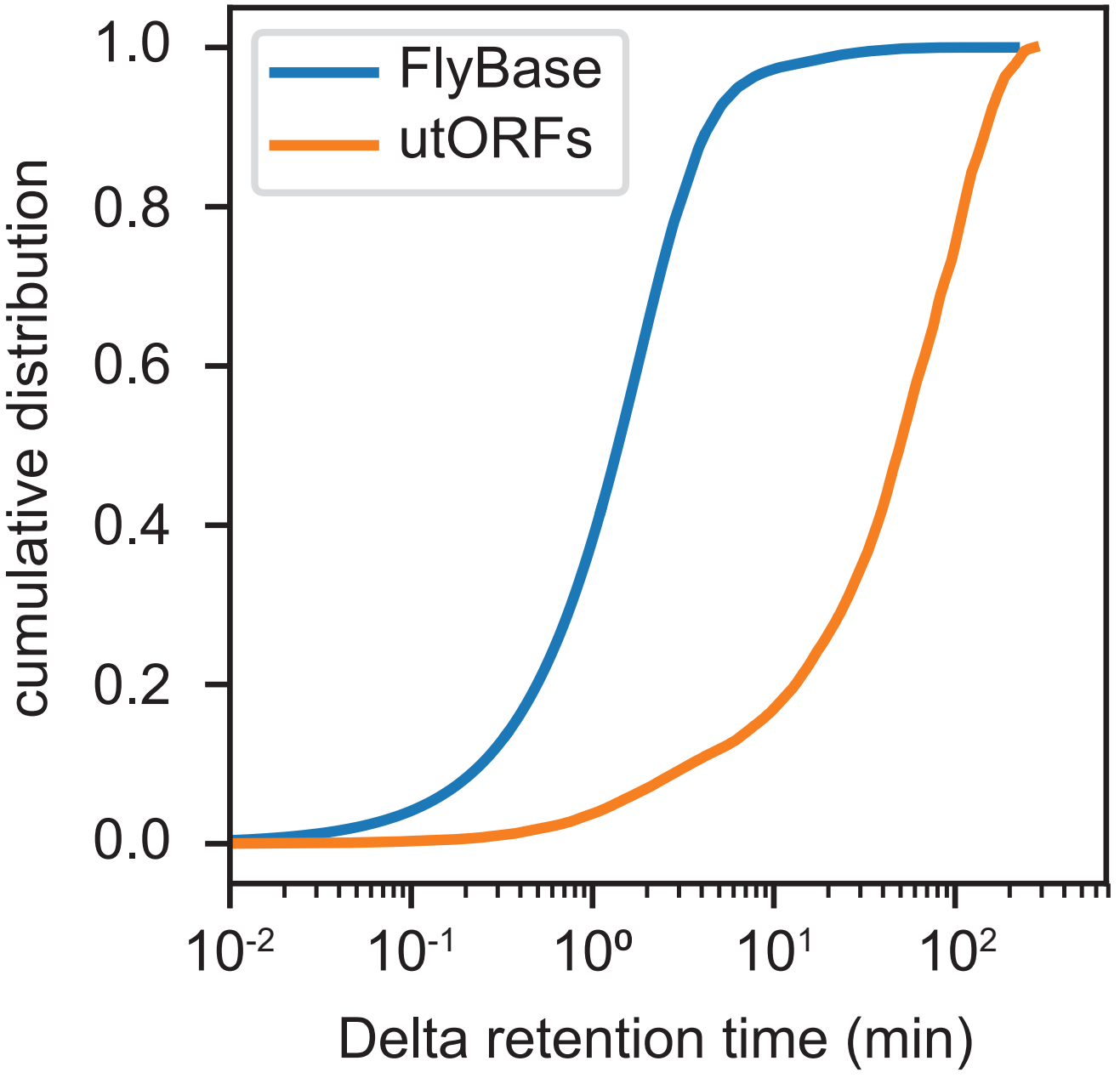
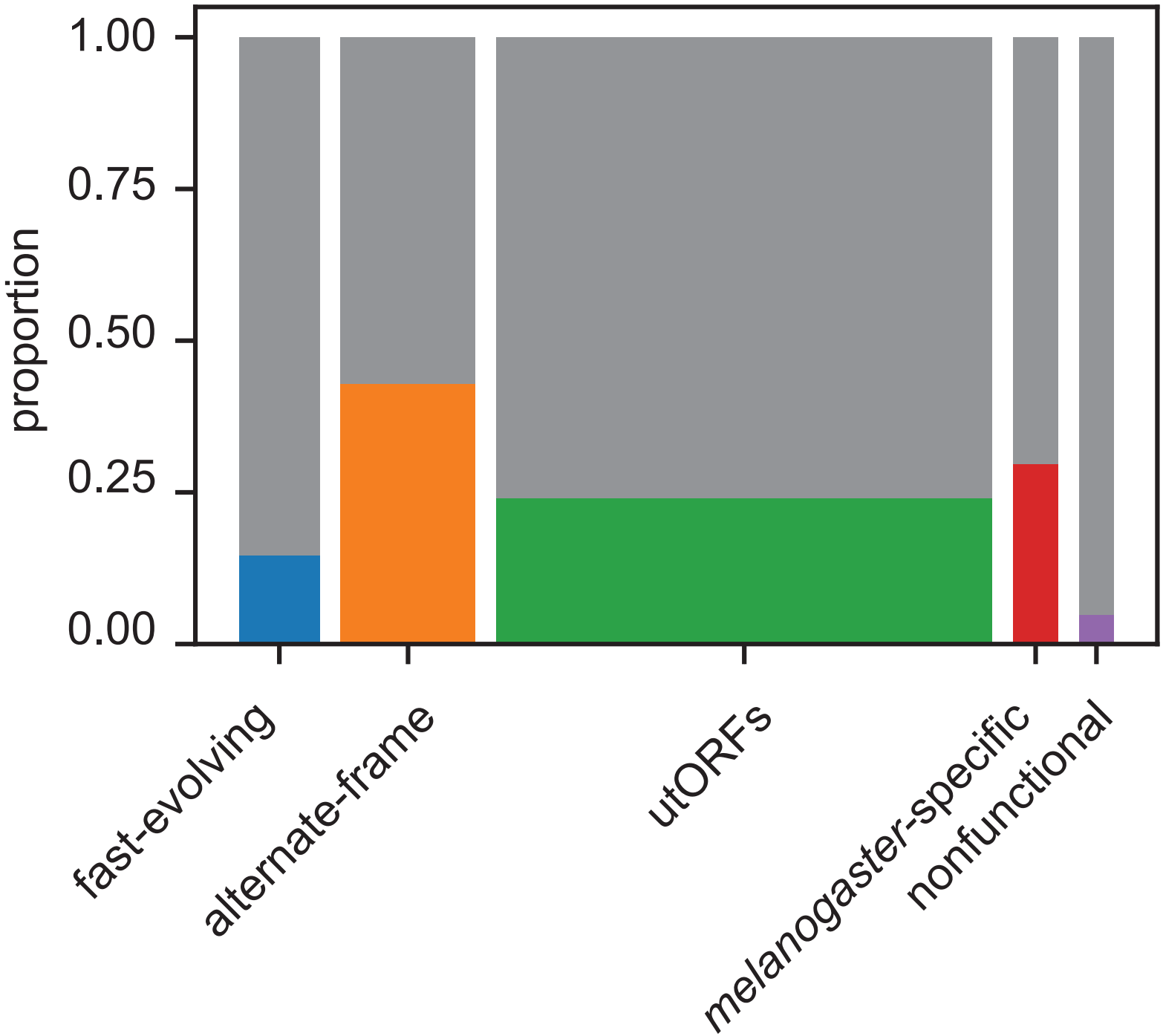



Figure 7 - Figure supplement 2