# New Genes in *Drosophila* Quickly Evolved Essential Functions in Viability During Development

Shengqian Xia[*1], Nicholas W. VanKuren[1*], Chunyan Chen[2,3*], Li Zhang[1], Clause Kemkemer[1], Yi Shao[2,3], Hangxing Jia[2,3], UnJin Lee[1,4], Alexander S. Advani[4], Andrea Gschwend[5], Maria Vibranovski[6], Sidi Chen[7], Yong E. Zhang[2,3,8] and Manyuan Long[1,4]

1. Department of Ecology and Evolution, The University of Chicago, Chicago, USA.
2. State Key Laboratory of Integrated Management of Pest Insects and Rodents & Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China.
3. University of Chinese Academy of Sciences, Beijing 100049, China.
4. Committee on Genetics, Genomics and Systems Biology, The University of Chicago, Chicago, USA.
5. Department of Horticulture & Crop Science, The Ohio State University, Columbus, Ohio, USA
6. Department of Genetics and Evolutionary Biology, University of São Paulo, Sao Paulo, Brazil.
7. Department of Genetics, Yale School of Medicine, West Haven, USA.
8. Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China.

* Co-first authors

Corresponding author: mlong@uchicago.edu; zhangyong@ioz.ac.cn

**ABSTRACT**

It is a conventionally hold dogma that the genetic basis underling development is conserved in a long evolutionary time scale. Ample experiments based on mutational, biochemical, functional, and complementary knockdown/knockout approaches have revealed the unexpectedly important role of recently evolved new genes in development of *Drosophila*. The recent progresses in the analyses of gene effects and improvements in the computational identification of new genes, which has led to large sample sizes of new genes, open the door to investigate the evolution of gene essentiality with a phylogenetically high resolution. These advancements also raised interesting issues related to phenotypic effect analyses of genes, particularly of those that recently originated. Here we reported our analyses of these issues, including the dating of gene ages, the interpretation of RNAi data that may confuse false positive/false negative rates, and the potential confounding impact of compensation and developmental effects that were not considered during previous CRISPR knockout experiments. We further analyzed new data from knockdowns of a large number of new genes, revealing a similarly high proportion of essential genes from recent evolution, compared to those found in distant ancestors of *D. melanogaster*. Knockout of a few young genes detected analogous essentiality. Furthermore, our experimentally determined distribution and comparison of knockdown efficiency in different RNAi libraries provided valuable data for general functional analyses of genes. Taken together, these data, along with an improved understanding of the phenotypic effect analyses of new genes, provide further evidence to the conclusion that new genes in *Drosophila* quickly evolved essential functions in viability during development.

[*Keywords*: new gene; *Drosophila*; RNAi; the off-target effect; false positive; false negative; CRISPR; essentiality; the compensation effect]

## INTRODUCTION

The question of how often new genes evolve essential functions is a critical problem in understanding the genetic basis of development and general phenotypic evolution. New genes in evolution have recently attracted an increasing amount of discussion (Jacob 1977; Mayr 1982; Ashburner et al. 1999; Chen et al. 2013). Furthermore, the detected large number of new genes with unexpected rate of new gene evolution and revealed important functions of new genes (Zhang et al. 2019; Kasinathan et al. 2020; Long et al. 2013) challenged a widely hold dogma that the genetic basis in control of development is conserved in a long time scale of evolution (Gould, 2002; Carroll, 2005; Carvunis et al. 2012; Krebs et al. 2013; Xie et al. 2019; Vakirlis et al. 2020a; Vakirlis et al. 2020b; Schroeder et al. 2020). Our previous work used the RNAi knockdown in a smaller sample showing that new genes may quickly become essential in *Drosophila* and that potential for a gene to develop an essential function is independent of its age (Chen et al. 2010). This work suggests a tremendous and quickly evolving genetic diversity, which had not been previously anticipated. Since then, genomes of better quality from more species have allowed for more reliable new gene annotation (Shao et al. 2019). In addition, technical progress in the detection of gene effects has increased with better equipped knockdown libraries and direct CRISPR knockout methods. Related scientific discoveries and technical development in knockdown and knockout techniques -- e.g., Green et al (2014) and Kondo et al (2017), respectively -- can be considered when investigating the evolution of gene essentiality.

We will present in this report our recent experiments and computational analyses, examining a few important issues raised in recent years (e.g. by Kondo et al. (2017) and Green et al. (2014)) that we find to be generally relevant for the detection of the phenotypic effects of genes, particularly of those that recently originated. Our investigations include the following: 1) the estimation of new gene ages; 2) an evaluation of the knockdown efficiency distribution in RNAi experiments; 3) an understanding of the differences between different RNAi libraries in phenotyping large samples of new genes for viability effects; 4) an interpretation of knockout data regarding the compensation effect. Our analyses, with additional evidence published recently by our group and others, provide ample and strong evidence to further support

108 a notion suggested by the fitness effect analysis of new genes in *Drosophila*: new genes
109 have quickly evolved essential functions in viability during development.
110
111
112 **RESULTS AND DISCUSSION**
113
114 **Identification of *Drosophila*-specific genes by the age dating.**
115
116 Two new gene datasets are available for *D. melanogaster*, which include the dataset of
117 Kondo *et al* (2017; the K-dataset, the underlying pipeline as the K-pipeline) and the
118 dataset we recently reported (Zhang et al, 2010; Shao et al, 2019 called as the GenTree
119 Fly dataset, the G-dataset). In order to determine which dataset is more accurate and
120 thus could be used in the downstream analyses, we estimated their qualities by
121 performing systematic comparison. Kondo *et al.* (2017) identified 1,182 new genes that
122 postdated the split of *D. melanogaster* and *D. pseudoobscura* (Branches 3~6, ~40
123 Mya（million years ago); Fig. 1A and 1B). They inferred the ages of these genes by
124 incorporating the UCSC DNA-level synteny information, homology information based
125 on comparison of annotated proteins and RNA expression profiling. By contrast, we
126 identified 654 new genes in this same evolutionary period (Fig. 1B) using the same
127 UCSC synteny information (Rhead et al. 2010) and our maximum parsimony-based
128 pipelines (Zhang et al. 2010).
129
130 We investigated why the K-dataset was almost twice the G-dataset over the same
131 evolutionary period. The K-dataset contained 471 new gene candidates in the G-dataset
132 (471/654 = 72%) (Fig. 1B, red), among which 313 are of the exact same ages while 158
133 show minor age differences (Supplemental Table S1). For the remaining 183 genes
134 absent in the K-dataset, we found 101 as authentic new genes (Supplemental Table S2)
135 after extensive validation by manually checking UCSC synteny information and four
136 additional resources including the FlyBase ortholog annotation, Ensembl Metazoa
137 homolog annotation, protein prediction in outgroup species and published literatures
138 (see also Materials and Methods). This result indicates a high false negative rate in the
139 K-dataset. By contrast, only 19 genes are old genes, which represent false positives in
140 the G-dataset. Then, for 45 genes, they are located in clusters of tandemly amplified

141 genes or transposon-rich regions, where synteny is often ambiguous or difficult to build
142 especially when outgroup species genomes are poorly assembled. Their ages are
143 difficult to infer. Analogously, the final remaining 18 candidates (Fig. 1B, green) are
144 dubious, either with inconsistent topology between UCSC synteny information and
145 Ensembl tree, marginal protein similarity between species or gene structure model
146 changes.
147
148 We next examined the 711 (1182-471=711) new gene candidates unique to the K-
149 dataset by manually examining phylogenetic distribution and their syntenic relationship
150 with genes in various species. We could only confirm 49 authentic new genes, which
151 represent false negatives of the G-dataset. By contrast, 318 out of 711 genes were
152 incorrectly dated as new genes due to four problematic practices (Fig. 1B; Supplemental
153 Table S2 lists the 318 false positives): 1) neglecting 275 that have orthologs in outgroup
154 species; 2) taking 32 noncoding or pseudogene models as protein coding genes; 3)
155 treating 6 redundant entries of same genes as different genes; and 4) misdating 5
156 polycistronic coding genes reported by the literatures. In addition, 242 (Fig. 1B, sky
157 blue) genes are located in repetitive regions. To be conservative, the G-dataset excluded
158 these genes from dating. Then, the remaining 102 candidates (Fig. 1B, green) are
159 dubious.
160
161 We conclude, based on above exhaustive manual evaluation, that the G-dataset is of
162 much higher quality compared to the K-dataset: 1) the false negative rate and false
163 positive rate of the G-dataset is estimated as 7.5% (49) and 2.9% (19), respectively; 2)
164 The both parameters of the K-dataset are higher, 8.5% (101) and 26.9% (318),
165 respectively; 3) the G-dataset only contains 9.6% (63) low-quality candidates (genes in
166 repetitive or dubious categories), while the K-dataset consists of 29.1% (344) such
167 candidates. Overall, 56% (662/1,182) new gene candidates in the K-dataset are either
168 false positive or dubious.
169
170
171 **Measuring reproducibility and efficiency of knockdown.**
172
173 We investigated the consistency of RNAi experiments with the same lines and the same
174 drivers in different laboratories, conditions, and years. Zeng et al. (2015) screened

175 16,562 transgenic RNAi lines using an Act5C-Gal4 driver to detect the lethality of
176 12,705 protein-coding genes (~90% of all annotated coding genes) in their study of
177 intestinal stem cell development and maintenance. Their dataset included RNAi lines
178 targeting the same 103 genes that were measured for lethality by Chen et al. (2010).
179 Chen et al. (2010) and Zeng et al. (2015) obtained the same phenotypes for 88 (85.4%)
180 genes, including 30 (29.1%) of the lethal phenotype and 58 (56.3%) of non-lethal
181 phenotype (Fig. 2A, Supplemental Table S3). These data suggest that despite
182 differences in independent observers, lab environments, and years to conduct
183 experiments, the vast majority of RNAi knockdown experiments are reproducible for
184 phenotyping lethality and non-lethality.
185
186 We also tested consistency between RNAi lines with different RNAi drivers (called
187 new drivers) or same drivers in different genome positions. Specifically, the datasets of
188 Chen et al (2010) and Zeng et al (2015) shared 86 new genes in knockdown
189 experiments, mostly (81.4%, 70) with different RNAi drivers and fewer (18.6%, 16)
190 same drivers in different genome positions (Supplemental Table S4). This dataset
191 showed that: 7 genes were consistently lethal; 42 genes were consistently non-lethal;
192 and 37 genes have different phenotypes (Fig. 2B). Thus, the two groups with different
193 drivers or same drivers with different positions show that more genes (57.0%, 49) have
194 the same phenotypes.
195
196 We considered an additional factor in RNAi knockdown, sensitivity, in the two widely
197 used RNAi libraries: the Vienna Drosophila Resource Center's (VDRC's) GD and KK
198 libraries (Dietzl et al., 2007). The GD libraries were constructed using P-elements to
199 randomly insert into the genome long hairpin RNAs (lhRNAs) targeting individual
200 genes, while the KK library inserted constructs carrying short hairpin RNAs (shRNAs)
201 into a specific landing site by ΦC31-mediated homologous recombination. All KK lines
202 carry an insertion at 30B3, but a proportion (23-25%) also carry an insertion at 40D3
203 (*tio* locus) that results in pupal lethality when using constitutive drivers like Act5C-
204 GAL4 (Green et al. 2014; Vissers et al. 2016). Unless specified, no lines discussed
205 below contain 40D3 insertions.
206
207 Given the intrinsic different designs of GD and KK libraries, we hypothesized that they
208 have different false negative or false positive rates, which cause the inconsistency

shown in Fig. 2B. Only GD lines were examined previously, and they have a high false negative rate (39.9%) but low false positive rate (<2%) (Dietzl et al. 2007). The high false negative rate is likely caused by insufficient target gene knockdown, while false positives may be due to off target effects (Dietzl et al. 2007). We thus tested the knockdown efficiency of 75 KK lines targeting randomly selected 75 young genes (Supplemental Table S5, Fig. 3A). We found that the knockdown efficiency of KK lines is generally lower than the efficiency of 64 GD lines as previously reported (Dietzl et al. 2007). Specifically, using the same driver (Act5C), we found that in general, GD lines have significantly higher knockdown efficiency than KK lines, as shown by the knockdown expression as the percentage of the control expression (Fig. 3A). That is, the KK lines have an average knockdown efficiency as 48.6% of control expression while the GD lines show an average efficiency as 38.1% (Fig. 3B and 3C, *t*-test $P$ = 0.031). Notably, the 50~60% expression level of the wide-type level will generally not lead to any fitness loss due to widespread haplosufficiency (Huang et al. 2010). Detecting any fitness effect may be expected when the expression drops to a lower level, for example, 20~30% or lower of the control expression. In this range of knockdown efficiency, we observed that only 29% of KK lines but 41% of GD lines reduced target expression levels to ⩽20% of control levels; 37% of KK lines but 53% of GD lines were seen to reduce target expression levels to ⩽ 30% of control levels (Fig. 3A). Thus, it is expected that GD lines have a significantly higher power in detecting lethal phenotypes as shown in the next section.

To estimate false positive rate of KK lines, we constructed 10 randomly chosen new KK lines targeting one member of a young duplicate gene pair, in addition to one KK lines and 3 TRiP lines (Transgenic RNAi Project, BDSC, Materials and Methods). The rationale is that for each gene of interest its paralog is the most likely off target. The same rationale was also followed by (Dietzl et al. 2007) when false positive rates of GD lines were estimated. We measured the knockdown efficiency and estimated off-target effects using these 14 lines with qPCR experiments in adult whole bodies (Fig. 4). We found that two lines likely produce off-target effects (*CG31958*, *RpS15Ab* (the TRiP line)), for both of which the expression of paralog is down-regulated to similar or even lower level compared to the corresponding gene of interest. 12 other lines have

242 significantly higher target effects than off-target effects, among which 10 genes reduced
243 activity to 20-80% expression level of the control (7 genes reduced activity to 20-40%)
244 and only two genes (*CG32164*, *CG7046*) reach⩽20% of control levels. Thus, if we
245 take 20% as the cutoff of efficient knockdown, only *CG31958* could be counted as the
246 false positive, and *CG32164* and *CG7046* be counted as the true positives. Collectively
247 speaking, the off-target effects are rare while insufficient knockdowns are pervasive.
248
249 These experiments detected a variation of knockdown efficiency among different
250 drivers where newer KK lines have lower efficiency and thus higher false negatives
251 compared to older GD lines. Therefore, these observations offer an alternative
252 interpretation of the incongruence than the false positive-based rationale of Kondo et al
253 (2017): when new RNAi drivers were added to the analysis, insufficient knockdown
254 was also introduced with a high probability. This would create incongruence between
255 old and new drivers if the old and new triggers have significantly different sensitivity.
256
257 **Phenotyping essentiality of new genes in RNAi libraries.**
258
259 We first investigated differences in measured lethality between the KK and GD libraries
260 used in Chen et al (2010). To control for the confounding effect of *tio* insertion in the
261 KK lines, we genotyped these lines using PCR-amplification and found that out of 153
262 KK lines we collected, 47 (30.7%) had two sites and 6 (3.9%) had only 40D3 site (the
263 confounding site) (Green et al. 2014), which all showed lethal phenotypes
264 (Supplemental Table S6). Using the recombination approach (Green et al. 2014), we
265 recovered 41 of the 47 lines into the lines that have only the 30B3 site. The RNAi
266 knockdown of 140 KK lines carrying insertions only at 30B3 identified 12 genes (8.6%)
267 with lethal phenotypes (Supplemental Table S6). Meanwhile, 12 genes in 59 GD lines
268 (20.3%) were detected to have lethal knockdown effects (Chen et al. 2010),
269 significantly higher than the KK libraries ($P$ = 0.0112, Fisher's Exact Test). As
270 aforementioned, this difference is likely due to higher false negative rate of KK lines
271 (Fig. 3).
272
273 By using the essentiality data of 10,652 old genes provided by VDRC
274 (https://stockcenter.vdrc.at/control/library_rnai) that were in branch 0 (Shao et al.

275 2019), we characterized the statistical distribution of essential old genes (Fig. 5). We
276 independently sampled 1000 times, with each randomly sampling 150 old genes and
277 calculating the proportion of essential ones. We found that in the GD library, the
278 probability to obtain a proportion of essential new genes equal or lower than 20.3% is
279 0.780. Meanwhile, in the KK library, the probability to observe a proportion of
280 essential new genes equal or lower than 8.6% is 0.867. These analyses of GD and KK
281 libraries reveal similarly that the proportions of new and old genes with lethal
282 phenotypes are not statistically different.
283
284 Further analysis of gene essentiality data in a recent version of VDRC libraries
285 (retrieved online in April 2019) detected with increased resolution the proportions of
286 essential genes in six detectable ancestral stages of *D. melanogaster*. We reported the
287 analysis of the GD library, which has a significantly higher knockdown efficiency than
288 the KK library. In total, 11,354 genes (72% of 15,682 genes in the species, Ensembl
289 73) have been phenotyped for their lethality or nonlethality, including 702 *Drosophila*
290 genus specific genes (66% of 1,070 detected *Drosophila*-specific genes) (Long et al.
291 2013; Shao et al. 2019) and 10,652 genes that predated the *Drosophila* divergence 40
292 Mya.
293
294 We parsimoniously mapped the 702 *Drosophila*-specific genes on the six ancestral
295 stages by examining their species distribution in the *Drosophila* phylogeny (Shao et al.
296 2019) (Fig. 6A). Of the 702 genes, 19.7% (138) are directly observed to be essential,
297 similar to the proportion of essential old genes, 18.9% ($p = 0.6212$ in the Fisher exact
298 test). We considered a low knockdown efficiency as shown by the 47% of GD lines
299 whose knockdowns are expressed at the level of 30% or higher of the control (Fig. 3A),
300 suggesting that 47% of RNA lines are invalid for the testing and should be subtracted
301 from the total tested lines.
302
303 Thus, the actual proportion of essential genes can be estimated by correcting for the
304 bias of false positives (Fp) and false negatives (Fn) by following formula:
305
306 $$\text{Corrected proportion of essential genes} = [E - (T \cdot F_p)] / [T - (T \cdot F_n)]$$
307

308 Where E and T are observed number of essential genes and total number of genes
309 examined, respectively. $F_p$ was measured as 1.6% (Dietzl et al., 2007) while $F_n$ as 47%
310 as estimated above or 39.9% as measured previously (Dietzl et al., 2007). Thus, the
311 estimated proportion of essential genes after correcting false positives and false
312 negatives can be as high as 36.5% for the estimated false negative rate of 47% in this
313 study. The corrected proportion can be also as high as 32.2% given the previously
314 measured false negative rate of 39.9%. Furthermore, all six stages show a stable
315 proportion of essential genes; none of the proportions is statistically different from the
316 proportion of old genes (Fig. 6A). Meanwhile, lethal rates of new genes which belong
317 to three origin mechanism categories (DNA-based duplication, RNA-based duplication
318 and orphan genes, Shao et al., 2019) also show no significant difference (Fig. 6B).
319 Interestingly, 21.7% of orphan genes, some of which might be *de novo* genes (Long et
320 al., 2013), are essential. These data add new insight into the evolution of essentiality in
321 all ancestral stages: soon after genes originated and fixed in *D. melanogaster*, a stable
322 proportion of new genes is essential throughout entire evolutionary process from
323 ancient ancestors to the speciation of *D. melanogaster*.
324
325 These data of knockdown experiments on a large number of new genes further
326 supported what we proposed before: *Drosophila* new genes rapidly evolve essential
327 functions within the divergence of *Drosophila* genus; knockdown of these genes leads
328 to death of flies.
329
330 **Analyses of mutants identified young essential genes**
331
332 Kondo et al (2017) recommended and used CRISPR/Cas9-mediated mutagenesis to
333 create small frameshift indel mutations in targeted genes. This method has two potential
334 issues. First, it is now well documented that vertebrate cells such as mammalian cells
335 or zebrafish cells recognize such aberrant mRNAs and compensate for their loss by
336 increasing expression of genes with high sequence similarity, such as paralogs in
337 zebrafish, worm and other organisms (Rossi et al. 2015; El-Brolosy and Stainier 2017;
338 El-Brolosy et al. 2019); Ma et al, 2019; Serobyan et al, 2020). This has the effect of
339 producing false negatives especially for recent duplicates. We confirmed that a similar
340 compensation effect exists in *Drosophila*. Specifically, when we induced a one-
341 nucleotide deletion using CRISPR/Cas9 into the ORF region of *vismay* (*vis*), a *D.*

342 *melanogaster*-specific gene duplicated from a parental gene, *achintya* (*achi*), 0.8
343 million years ago, with a sequence similarity of 92% between the two copies. We found
344 that *Achi* in the *vis* mutant was significantly upregulated whereas a randomly selected
345 unrelated gene *CG12608* and the distantly related gene *hth* (9% identity) to *vis*, did not
346 show such an effect (Fig. 7). Second, CRISPR/Cas9-mediated mutagenesis cannot
347 detect the effects of maternal and paternal effect genes, which can be common in
348 *Drosophila* (Perrimon et al. 1989; Raices et al. 2019) and can be detected by RNAi
349 knockdown. Therefore, the two approaches of knockdown and knockout/mutagenesis
350 are complementary to each other given their technical characteristics.
351
352 Actually, in-depth analyses of four cases already provided further evidence supporting
353 essentiality of new genes in development. First, Ross et al (2013) reported a stepwise
354 neofunctionalization evolution in which a centromere-targeting gene in *Drosophila*,
355 *Umbrea*, was generated less than 15 million years ago. Both RNAi knockdown, rescue
356 experiments and P-element mediated gene knockout revealed that *Umbrea* evolved a
357 species-specific essentiality to target centromere in chromosome segregation (Chen et
358 al. 2010; Ross et al. 2013). Second, Lee et al (2019) recently detected stage-specific
359 (embryos/larvae/pupa) lethality associated with RNAi knockdown and CRISPR
360 knockout in *Cocoon,* a gene emerged 4 million years ago in the common ancestor of
361 the clade of *D. melanogaster-simulans.* These data show that *Cocoon* is essential for
362 the survival at multiple developmental stages, including the critical embryonic stage.
363 Third, P-element insertion/excision experiments show the essentiality of *K81* as a
364 paternal element in early development. This gene only exists in the *Drosophila*
365 *melanogaster*-subgroup species that diverged 6 million years ago (Loppin et al. 2005).
366
367 Zeus, a gene that duplicated from the highly conserved transcription factor *CAF40* 4
368 million years ago in the common ancestor of *D. melanogaster* and *D. simulans* rapidly
369 evolved new essential functions in male reproductive functions, as detected in the null
370 mutants and knockdown (Chen et al. 2012; Ventura, 2019). A pair of extremely young
371 duplicates, *Apollo* (*Apl*) and *Artemis* (*Arts*), was found to have been fixed 200,000 years
372 ago in *D. melanogaster* populations (VanKuren and Long, 2018). CRISPR-created
373 gene deletions of these genes showed that both evolved distinct essential functions in
374 gametogenesis and *Apl* critical function in development. *Cocoon* was found to evolve

375 recently in the most recent ancestor of *D. melanogaster* and the *D. simulans* clade (>4
376 Mya). Knockdown and knockout revealed that it is essential for the survival of
377 *Drosophila* at multiple developmental stages, including the critical embryonic stage
378 (Lee et al, 2019).
379

380 In a comprehensive functional and evolutionary analysis of the ZAD-ZNF gene family
381 in *Drosophila* (Kasinathan et al, 2020), 86 paralogous copies were identified with
382 phenotypic effects detected by knockdown and knockout in *D. melanogaster*. It was
383 found that the proportion (17/58 = 29.3%) of lethal copies in old duplicates (>40 MYA)
384 and the proportion (11/28 = 39.3%) of lethal copies in *Drosophila*-specific duplicates
385 (<40MYA) are statistically similar. Further functional analyses of one of the non-
386 essential young copies (*CG17802, Nicknack*) reported by Kondo et al (2017) clearly
387 unveiled an essential function for larval development. These pieces of evidence
388 strongly support the notion that new genes can quickly evolve essential functions in
389 development.

390

391 Despite of the small sample size (4), the generally young ages of these new genes
392 supported the conclusion drawn from the previous large-scale knockdown analysis: the
393 new genes in *Drosophila* rapidly evolve essential functions in development.
394
395
396 **Concluding Remarks**
397

398 We appreciate the extensive experiment, computation and data-compilation by Kondo
399 et al (2017) and their interests in the evolution of gene essentiality. However, we found
400 that the K-pipeline and K-dataset were associated with a high false positive rate.
401 Moreover, their interpretation of RNAi data is problematic due to confusing the false
402 negative and false positive, while they applied an incorrect CRISPR mutagenesis that
403 neglected compensation and parental effects. The data we created in this study, while
404 revealing their errors and technical insufficiencies, increase understanding of technical
405 subtleties for analyzing effects of young duplicate genes and other genes. More data
406 and additional analyses of related scientific issues for the testing of fitness and

407 functional effects of new genes from the two complementary approaches, RNAi
408 knockdown and CRISPR knockout, provided a strong support for the concept: the new
409 genes rapidly evolved essential functions in development in *Drosophila*.
410
411
412 **MATERIALS AND METHODS**
413 ***Comparison of the K-dataset and the G-dataset for Drosophila new genes.***
414 *Overall comparison scheme*
415 To our knowledge, there is no published genome-wide evaluation of gene ages in
416 *Drosophila*. Specifically, around a decade ago, we took advantage of the syntenic
417 genomic (DNA) alignment generated by the UCSC group and performed the genome-
418 wide age dating for the first time in *Drosophila* (Zhang et al. 2010). At that moment,
419 we compared our data to previous studies based on limited number of cases and
420 discovered the general reproducibility across studies. The genome-wide dataset by
421 Kondo *et al.* enabled a systematic large-scale comparison. We began with repeating our
422 pipeline on FlyBase annotation v6.02 and identified 654 new genes originated on the
423 branch toward *D. melanogaster* after the species split of *D. melanogaster* and *D.*
424 *pseudoobscura* (Fig. 1). Concurrently, based on similar FlyBase release v6.13, Kondo
425 *et al.* performed dating according to the same syntenic alignment of UCSC, which is
426 further complemented by v6.02 protein-level BLAST search and filter with testis-
427 specific expression (Kondo et al. 2017). With additional Dollo-parsimonious searches,
428 they identified 1,182 new genes originated in the same period including *melanogaster*-
429 group to *melanogaster*-only, which correspond to age group 3 to 6 in our analysis,
430 respectively (Supplemental Fig. S1). Since the FlyBase annotation version is similar
431 (v6.02 vs. v6.13), only 12 entries out of the G-dataset and K-dataset are not comparable
432 due to "Gene model change" (Supplemental Fig. S1). They represent either expired
433 models or new models in v6.13. Except them, all other genes can be compared across
434 two datasets.
435
436 We found that 471 out of the 654 new gene candidates in the G-dataset are covered in
437 the K-database by comparing the Ensembl IDs of these databases. Moreover, 313
438 (66%) genes show the exact same ages. Since manual curation needs extensive efforts,
439 we did not examine why the remaining 158 genes show minor age difference. Instead,

440 we subsequently only focused on those genes which show conflicting dating results,
441 *i.e.*, included or excluded in new gene dataset across two studies. As a result, we
442 classified the conflicting cases into six major categories, which can be further divided
443 into around 20 more specific sub-categories (Supplemental Fig. S1). We documented
444 how we performed classification as below.
445
446 *Four independent information sources facilitate evaluations of two age datasets*
447
448 The challenges in the dating of gene age largely lie in the ambiguity of calling orthologs
449 across outgroup species (Liebeskind et al. 2016). We found that the conflict of age
450 dating was often due to the difference of DNA-level synteny and protein-level
451 homology search. Specifically, for a gene of interest, *A*, the UCSC best-to-best synteny
452 information shows that its ortholog is present in one outgroup species, *B*. However, the
453 protein-level information may reveal an absence. The opposite scenario can occur too.
454 In these conflicts, we turned to independent resources including FlyBase ortholog
455 annotation, the homolog annotation and gene family tree provided by Ensembl Metazoa
456 (St Pierre et al. 2014; Kersey et al. 2015), protein prediction in outgroup species based
457 on gene models of *D. melanogaster* and literatures. Specifically, FlyBase provided
458 AAA (Assembly/Alignment/Annotation) syntenic ortholog annotation. If species *B*
459 encodes a FlyBase annotated ortholog, gene *A* likely predated the species-split of *D.*
460 *melanogaster* and *B*. Similarly, Ensembl Metazoa provided one-to-one best-to-best
461 ortholog annotation. We used it like FlyBase. Finally, for some cases where synteny
462 predicted orthologous regions of *B* do not harbor an annotated gene, we conceptually
463 translated this region with the protein of *D. melanogaster* as the template. BLAST
464 (Tblastn) was used here. We have two reasons to perform additional annotation: 1)
465 recently evolved genes are often poorly annotated; 2) annotation quality of outgroup
466 species is presumably worse compared to *D. melanogaster* and we need to correct this
467 bias. For particularly interesting cases (*e.g.* polycistronic coding genes), we searched
468 literatures describing their evolutionary history.
469
470 *Conflicting cases could be classified into six categories*
471
472 We implemented a series of customized rules to call the presence of ortholog of gene *A*
473 in species *B*. The first set of rules are used to call presence of ortholog based on gene

474 prediction. For a synteny-predicted candidate orthologous region in *B*, we ran Tblastn
475 to predict whether this region encodes an orthologous protein of *A*. If Tblastn could
476 align the protein of *D. melanogaster* beyond the following thresholds (identity cutoff >
477 70% & coverage cutoff > 30%, identity cutoff > 30% & coverage cutoff > 70%, identity
478 cutoff > 50% & coverage cutoff > 50%), we believed that the ortholog is present. If the
479 alignment meets with the threshold (identity cutoff < 30% & coverage cutoff < 30%),
480 the ortholog of *A* is absent in *B*. For all other cases, we called them as "boundary cases"
481 if there is also no ortholog annotated by FlyBase and Ensembl. A total of 80 candidate
482 new genes fall in this category including 65 cases in the list of Kondo *et al.* and 15 cases
483 in our list (Supplemental Fig. S1).
484
485 Secondly, for 275 out of 318 new genes dated by Kondo *et al* but not by us
486 (Supplemental Fig. S1), we identified orthologs as supported by at least two
487 independent sources (FlyBase, Ensembl homolog, Ensembl gene family tree, and/or
488 prediction). For example, in case of *FBgn0027589*, the ortholog is present across all 12
489 *Drosophila* species, which is supported by both Ensembl and FlyBase. The remaining
490 43 new genes are misidentified due to other types of problems (*e.g.* annotation problem
491 due to polycistronic structure such as *tal-1A*/*tal-2A*/*tal-3A*). All these 318 genes are
492 marked as "Dating problem in Kondo *et al*" (Supplemental Fig. S1). In the opposite
493 scenario, 101 new genes are only identified by us, which could be divided into four
494 cases: 1) for 50 genes called as old genes in the K-dataset, their new gene calling were
495 also supported by lack of one-to-one orthologs annotated by Ensembl or FlyBase; 2)
496 for 5 genes called as old genes in the K-dataset, they are subject to complex
497 evolutionary trajectories (pseudogenization of parental copies), such as *FBgn0032740*
498 and *Cyp6t1*; 3) for 10 cases excluded by the K-dataset, we examined phylogenetic trees
499 provided by Ensembl and confirmed that new gene are derived as suggested by longer
500 branch length; 4) for 36 genes excluded in the K-dataset, FlyBase and Ensembl do not
501 annotate orthologous genes in the outgroup species for most genes, which are also
502 consistent with the lack of Tblastn hits.
503
504 Compared to the K-dataset, the false positives and false negatives are much fewer in
505 the G-dataset (Fig. 1, Supplemental Fig. S1). In the G-dataset, we misidentified only
506 19 new genes with 14 cases caused by double or triple losses in the outgroups (*e.g.*
507 *CG2291*). In our pipeline, we first searched against closely-related species and then

508 went for remotely-related species. If at least two independent losses are needed to
509 explain the phylogenetic distribution of orthologs, we will assign a young age to gene
510 *A* by following the maximum parsimony. The underlying assumption is: 1) the
511 possibility of double or triple losses should be low; 2) the genomic alignment between
512 *D. melanogaster* and remotely-related species is less reliable compared to that between
513 *D. melanogaster* and closely-related species. Consistently, we only identified 14 cases
514 with support by at least one additional source (FlyBase, Ensembl). The remaining 5
515 cases are caused by lack of sensitivity of UCSC genome alignment. Both FlyBase and
516 Ensembl annotate orthologs in the outgroup species, but synteny does not cover the
517 corresponding regions. In opposite, genome alignment built spurious alignment in
518 remotely-related species for 49 new genes identified by Kondo *et al*. However, our
519 Tblastn search could not identify a protein at all. Thus, we referred them as false
520 negatives. All these 68 genes are put into the third category entitled with "Dating
521 problem in this work" (Supplemental Fig. S1).
522
523 A fourth category ("Not applicable or difficult for dating") consists of genes which are
524 most resistant for dating due to their sequence features undesirable for new gene
525 identification. 242 specific new genes claimed by Kondo *et al.* belongs to this category
526 (Supplemental Fig. S1). For 178 out of 242 genes, we found that synteny is in conflict
527 for outgroup species sharing the same phylogenetic relationship relative to *D.*
528 *melanogaster* (e.g. *D. simulans* and *D. sechellia*). These genes are generally located in
529 repetitive regions (*e.g.* tandem amplification). It is thus likely that the orthologous
530 regions may not be equally well assembled or be subject to species-specific gene
531 conversion across these outgroup species. We thus excluded these genes. For example,
532 we masked 19 out of 242 genes including 12 *Ste* genes, 5 Y-linked genes and 2 genes
533 encoded by contigs but not anchored to five major chromosome arms. *Ste* is the X-
534 linked tandem gene families each with redundant copies (Supplemental Fig. S2A). In
535 the UCSC Net track, the most assembles can only reach level 2 of one-way syntenic
536 mapping, rather than the adequate reciprocal syntenic mapping as level 1. The closely
537 related species (*e.g. D. simulans*) also encodes multiple copies, but the corresponding
538 region is not fully assembled and filled with lots of gaps and many of these copies even
539 cannot be assigned to chromosomes (Supplemental Fig. S2B). The size contrast of
540 assemblies between *D. sechellia* and *D. melanogaster* suggests that the region in *D.*
541 *sechellia* might not be properly assembled due to its repetitive structure (Supplemental

542 Fig. S2A and S2C).
543
544 In order to date each member correctly, high-quality outgroup genome must be
545 available first. As for 5 Y-linked genes, Koerich *et al.* (2008) assigned all of them to be
546 old genes (Koerich et al. 2008). The remaining 45 genes consist of three subtypes: 1)
547 21 fast-evolving small proteins (<100 amino acids); 2) 20 polycistronic genes without
548 previous literature support; 3) 4 tandem duplicates. Different from the aforementioned
549 178 genes, the syntenic information is consistent across outgroup species suggesting
550 that they are old genes. The small proteins or polycistronic genes are poorly annotated
551 across outgroups. For 4 duplicates, Ensembl phylogenetic tree could not provide
552 diagnostic information to infer the duplication order. Thus, we believed that all these
553 three subtypes are difficult to date as of now. In our private new gene list, 45 candidates
554 also belong to gene families. Similar to the above 4 tandem duplicates, the synteny is
555 consistent across outgroup species showing that these 45 genes are derived copies in
556 their respective families. However, Ensembl phylogenetic tree could not provide
557 additional support. So, we also put these genes into the same fourth category.
558
559 The fifth category ("Different ortholog definition") only consists of 28 private new
560 genes identified by Kondo *et al* (Supplemental Fig. S1). For 24 out of 28 cases, the
561 UCSC syntenic chain only covers a small portion (<30%) of coding regions or mainly
562 corresponds to untranslated regions (UTRs) in outgroup species. Since the dating of
563 Kondo *et al* is protein-centric, they called these genes as new genes. By contrast, our
564 dating pipeline works on DNA-level and identified these genes as old genes. The reason
565 why we took the age of most conserved exons to represent the age of whole genes is
566 that these exons usually represent most important functional regions. Moreover, by
567 performing dating on DNA-level, our dating does not depend on annotation quality of
568 outgroup species. Actually, for all 24 cases, whether the coding region is accurately
569 annotated is unknown due to the lack of protein evidence. For the remaining 4 cases,
570 they all represent translocated genes. In our terminology, we only referred derived
571 duplicate or orphan genes as new genes. By contrast, Kondo *et al.* interpreted
572 incorrectly translocated genes as new genes.
573
574 The sixth and final category refers to the aforementioned 12 entries not comparable due
575 to "Gene model change" (Supplemental Fig. S1).

576

577 In the main text, we merged “Gene model change”, “boundary cases” and “different
578 ortholog definition” as one dubious (in green color of Fig. 1B) category to simplify.

579

580 *D. melanogaster-specific gene identification*

581

582 Candidate new genes were initially collected from previous studies (Zhou et al. 2008;
583 Zhang et al. 2010; Chen et al. 2012). We removed from this list of 233 candidates: 1)
584 any genes whose *D. melanogaster*-specific release 6.05 (http://flybase.org) annotation
585 status is 'withdrawn', 2) genes not located on the major chromosome arms 2L, 2R , 3L,
586 3R, or X, and 3) members of large tandem arrays, including the *Sperm dynein*
587 *intermediate chain* (Nurminsky et al. 1998; Yeh et al. 2012), *Stellate* (*Ste*), and X:
588 19,900,000-19,960,000 arrays that are *D. melanogaster*-specific but are impossible to
589 specifically study. We checked syntenic whole-genome alignments of the remaining 84
590 genes manually using our multi-species alignments at the UCSC Genome Browser. To
591 be conservative, we required that all outgroups including the *D. simulans*, *D.sechellia*,
592 *D. yakuba*, and *D. erecta* genome assemblies contained no assembly gaps, transposable
593 elements, or repeats corresponding to the flanking regions of the putative *D.*
594 *melanogaster*-specific gene. *D. melanogaster*-specific gene origination mechanisms
595 and parental genes were taken from the original studies and confirmed using BLAT and
596 BLASTp. If a gene had multiple significant ($e < 10^{-10}$) full-length BLASTp hits *in D.*
597 *melanogaster*, the hit that was most similar to the *D. melanogaster*-specific gene was
598 assumed to be the parent. We used available *D. simulans* and *D. yakuba* next-generation
599 sequencing reads to test the presence of putative *D. melanogaster*-specific tandem
600 duplications in these two species (Green et al. 2014; Rogers et al. 2014). We found no
601 breakpoint spanning read pairs supporting *D. melanogaster* tandem duplications in any
602 of 20 *D. simulans* or 20 *D. yakuba* genomes. Thus, these tandem duplications are
603 specifically found in *D. melanogaster* and are not simply missing from the *D. yakuba*
604 and *D. simulans* reference genome assemblies. We checked if any of the duplications
605 in our final set are segregating rather than being fixed within *D. melanogaster* by
606 analyzing 17 whole genome re-sequencing data from the DPGP2 core Rwanda (RG)
607 genomes (Ni et al. 2008). We required tandem duplications to have at least one read
608 uniquely mapped to each of the three unique breakpoints in order to be called as 'present'
609 in a particular line. Ten genes are not found in any of 17 additional *D. melanogaster*

610 genomes we analyzed, suggesting that they are found specifically in the reference stock.
611 Finally, we curated 10 *D. melanogaster*-specific genes. This dataset is actually a subset
612 of G_K new gene data list.
613
614 ***RNAi strain construction***
615
616 Since species-specific new genes are under-represented in public RNAi lines, we
617 generated new RNAi lines following Dietzl et al. (2007). Briefly speaking, we designed
618 RNAi reagents using the E-RNAi server (http://www.dkfz.de/signaling/e-rnai3/) and
619 kept constructs with all possible 19-mers uniquely matching the intended target gene
620 and excluded designs with >1 CAN repeat (simple tandem repeats of the trinucleotide
621 with N indicates any base) (Ma et al. 2006). Constructs were cloned into pKC26
622 following the Vienna *Drosophila* Resource Center's (VDRC's) KK library strategy
623 (http://stockcenter.vdrc.at, last accessed 2 February 2016). We introgressed the X
624 chromosome from Bloomington *Drosophila* Stock Center line 34772, which expresses
625 ΦC31 integrase in ovary under control of the *nanos* promoter, into the VDRC 60100
626 strain. Strain 60100 carries attP sites at 2L:22,019,296 and 2L:9,437,482 (Green et al.
627 2014). We ensured that our RNAi constructs were inserted only at the 2L:9,437,482
628 site using PCR following Green et al. (2014). RNAi constructs were injected into the
629 60100-ΦC31 at 250 ng/μL. Surviving adult flies were crossed to $sna^{Sco}$/CyO balancer
630 flies (BDSC 9325) and individual insertion strains were isolated by backcrossing.
631
632 ***RNAi screen***
633
634 We knocked down target gene expression using driver lines constitutively and
635 ubiquitously expressing GAL4 under the control of either the *Actin5C* or *αTubulin84B*
636 promoter. We replaced driver line's balancer chromosomes with GFP-marked
637 chromosomes to track non-RNAi progeny. Control crosses used flies from the
638 background strains 60100-ΦC31, 25709, or 25710 crossed to driver strains. Five males
639 and five virgin driver females were used in each cross. Crosses were grown at 25°C,
640 40% - 60% humidity, and a 12h:12h light:dark cycle. F1 progeny were counted at day
641 19 after crossing, after all pupae had emerged. We screened F1 RNAi flies for visible
642 morphological defects in 1) wings: vein patterning and numbers, wing periphery; 2)
643 notum: general bristle organization and number, structure and smoothness; 3) legs:

644 number of segments. We monitored survival of RNAi F1s by counting GFP and non-
645 GFP L1, L3 larvae and pupae. We tested RNAi F1 sterility by crossing individual RNAi
646 F1 flies to 60100-ΦC31 and monitoring vials for L1 production. Ten replicates for each
647 sex for each line were performed.
648
649 ***RNAi knockdown specificity and sensitivity***
650
651 We sought to address two known problems of RNAi technology using RT-qPCR. First,
652 since off-target effects are often discussed in RNAi experiments(Dietzl et al. 2007) we
653 need to test whether target gene expression are specifically knocked down, although
654 our constructs are computationally predicted to be specific. Second, since the RNAi
655 knockdown is often incomplete (Dietzl et al. 2007), we need to estimate how many
656 genes are adequately knockdown in expression. We targeted a random dataset of 14 *D.*
657 *melanogaster*-specific genes. We collected qPCR primers from FlyPrimerBank (Hu et
658 al. 2013). For those genes not found in FlyPrimerBank, we took Primer-BLAST to
659 design primers by specifically targeting a ~100 bp region of the gene (Supplemental
660 Table S7). We confirmed primer specificity with PCR and Sanger sequencing.
661
662 We randomly selected 75 KK RNAi lines (no *tio* site insertion) to analyze their knock
663 down efficiency. We cross these 75 KK RNAi lines with same driver which was used
664 in Dietzl et al., 2007 for GD RNAi line knock down efficiency test. We extracted RNA
665 from sets of 8 adult males (2~4 day old) in triplicate from each RNAi cross using
666 TRIzol (Catalog# 15596-026, Invitrogen, USA), treated ~2 μg RNA with RNase-free
667 DNase I (Catalog# M0303S NEW ENGLAND Biolabs, USA), then used 1 μL treated
668 RNA in cDNA synthesis with SuperScript III Reverse Transciptase (Invitrogen, USA)
669 using oligo$(dT)_{20}$ primers. cDNA was diluted 1:40 in water before using 1 μL as
670 template in 10 μL qPCRs with iTaq™ Universal SYBR Green Supermix (Catalog#
671 1725121, Bio-Rad, USA) and 400 nM each primer. Reactions were run on a Bio-Rad
672 C1000 Touch thermal cycler with CFX96 detection system (BioRad, CA). Cycling
673 conditions were 95°C for 30 sec, then 45 cycles of 95°C for 5 sec, 60°C for 30 sec, and
674 72°C for 15 sec. We normalized gene expression levels using the $\Delta\Delta C_T$ method and
675 *RpL32* as the control (Livak and Schmittgen 2001; Dietzl et al. 2007). We tested the
676 specificity and efficiency (90%< qPCR Efficiency<110%) of qPCR primers using an
677 8-$\log_2$ dilution series for each primer pair (VanKuren and Long 2018).

678

679 ***Testing Compensation Effects of New Gene Duplicates***

680

681 We generated the frameshift mutation line of *vis* using the CRISPR protocol previously
682 developed (VanKuren and Long 2018) but with one single sgRNA for one gene as
683 Kondo et al (2017) did. The sgRNA-*vis* primer below was synthesized (the underlined
684 sequence):
685 GAAATTAATACGACTCACTATAGGATGTACGGCAGAACATAAGTTTAAGA
686 GCTATGCTGGAA;
687 We used the following sequence-specific qRT-PCR primers to test the compensatory
688 expression of *achi*, the duplicate of *vis*. Two control genes including *CG12608* and *hth*
689 were examined too. Since *vis*'s expression is largely testis-specific, we extracted RNAs
690 from testis of mated 4-day males and used qRT-PCR with 3 replicates to assess the
691 expression, as developed previously (VanKuren and Long 2018).
692 248bp:
693 Achi-RT1F: AAAGTGACAGGTTTCTCTGTTTG;
694 Achi-RT1R: CTGATCCTCCTCCACGATGAC.
695 237bp:
696 CG12608-RT1F: 5'-CATAGTGGGCACCTACGAG-3';
697 CG12608-RT1R: 5'-TGCGAGAGTATGATCTGCGAC-3'.
698 92bp:
699 hth-RT1F:5'-CCTAGTCATGTATCGCCGGTC-3';
700 hth-RT1R:5'-AGCGGATGTTCATAAATCGCA-3'.
701 Internal control:
702 113bp:
703 RpL32-RT1F: AGCATACAGGCCCAAGATCG;
704 RpL32-RT1R: TGTTGTCGATACCCTTGGGC.

705

706 **ACKNOWLEDGMENTS**

707 We are grateful for valuable discussion with Edwin Ferguson, Urs Schmidt-Ott,
708 Norbert Perrimon, Emily Mortola, the M. Long lab members and Y.E. Zhang lab
709 members for the technical development in RNAi and scientific issues related to this
710 study. We also appreciate Lisa Meadows from VDRC for supplying haplotyping results

for 500 KK lines of the correct insertion site. Y.E. Zhang was supported by the National Key R&D Program of China (2019YFA0802600, 2018YFC1406902) and the National Natural Science Foundation of China (91731302, 31771410, 31970565). M. Long was supported by NSF1026200 and NIH R01GM116113.

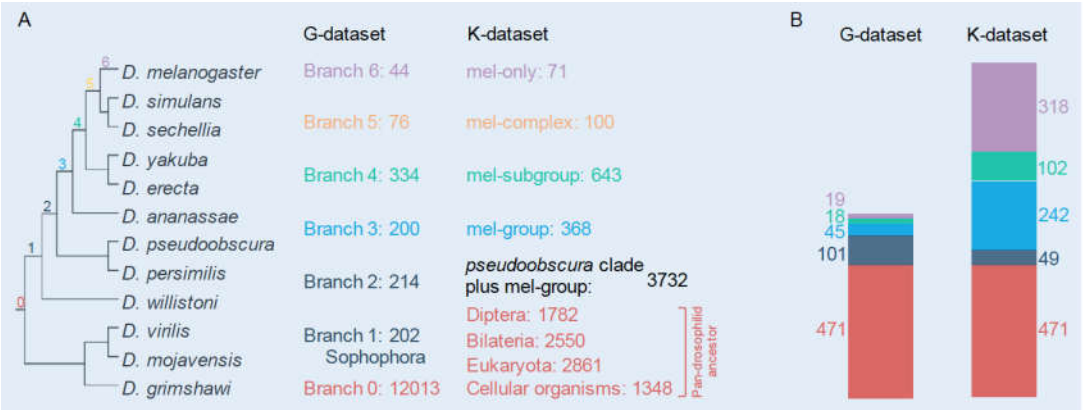


**Figure 1.** Summary of new gene candidates in the K-dataset and G-dataset. A. phylogenetic distribution of gene origination identified by the K-pipeline and the G-pipeline as shown in the two datasets. B. Evaluation of the two datasets based on individual gene analyses. The two datasets share 471 candidates (red). The G-dataset consist of 101 authentic candidates (deep blue) undetected in the K-dataset, 19 false positives (light purple), 18 dubious cases (green) and 45 cases not applicable for dating (sky blue). By contrast, the K-dataset includes 49 bona fide new gene candidate, 318 false positives, 102 dubious cases and 242 difficult cases. Note, the K-dataset mentions 1,182 genes in the main text, however its associated supplemental table includes 1,176 genes with 6 genes listed more than once.

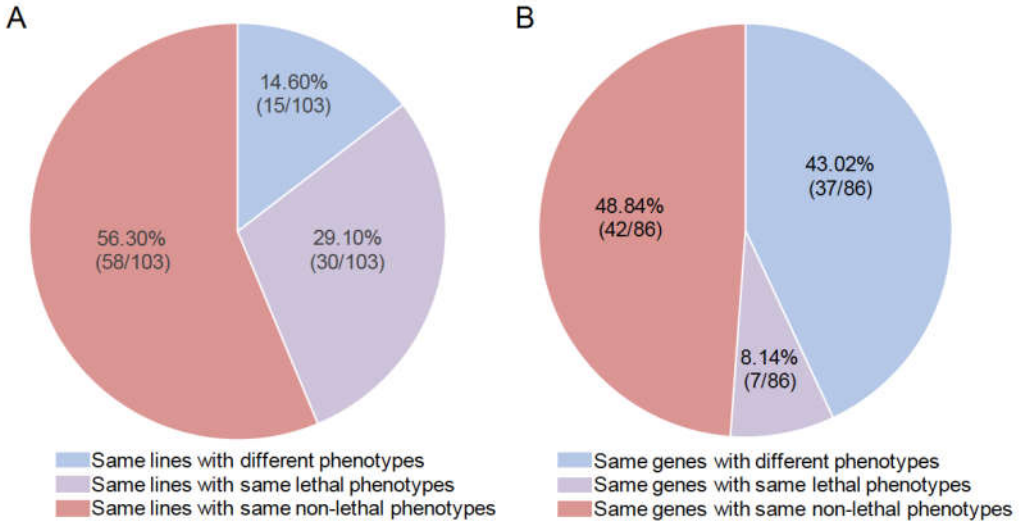


**Figure 2.** The reproducibility analysis of RNAi experiments by comparing two groups of independent experiments by Chen et al (2010) and Zeng et al (2015). A. Phenotypes of same 103 RNAi lines analyzed by Chen et al (2010) and Zeng et al (2015) using same lines; B. Phenotypes of 86 same new genes knocked down by two different drivers or the same drivers with different insertion sites. The old drivers detected 29 genes as lethal while 57 non-lethal; the new drivers detected 20 genes as lethal while 66 non-lethal.

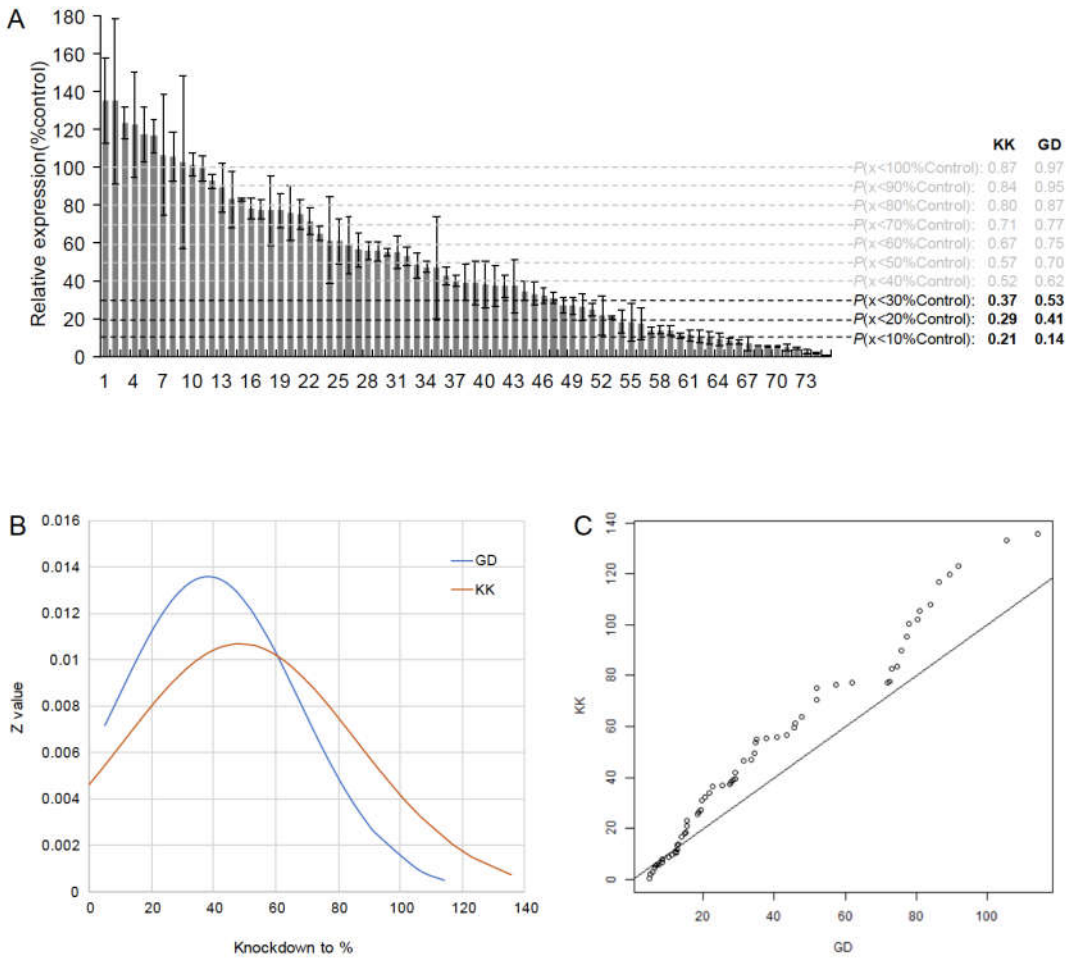


**Figure 3.** Knockdown efficiency in the KK and GD libraries revealed GD lines have significantly higher knockdown efficiency than the KK lines. A. The knockdown efficiency of the 75 KK lines was measured, compared to the expression of the wild-type control and the standard deviation is calculated from the measurement of three repeats; *P* refers to proportion of genes with the expression lower than a certain threshold while the values of KK lines are generated in this work and that of GD lines are extracted from Dietzl et al. (2007). B. The distributions of knockdown efficiency of KK and GD lines. C. The Q-Q Plots between KK and GD lines.

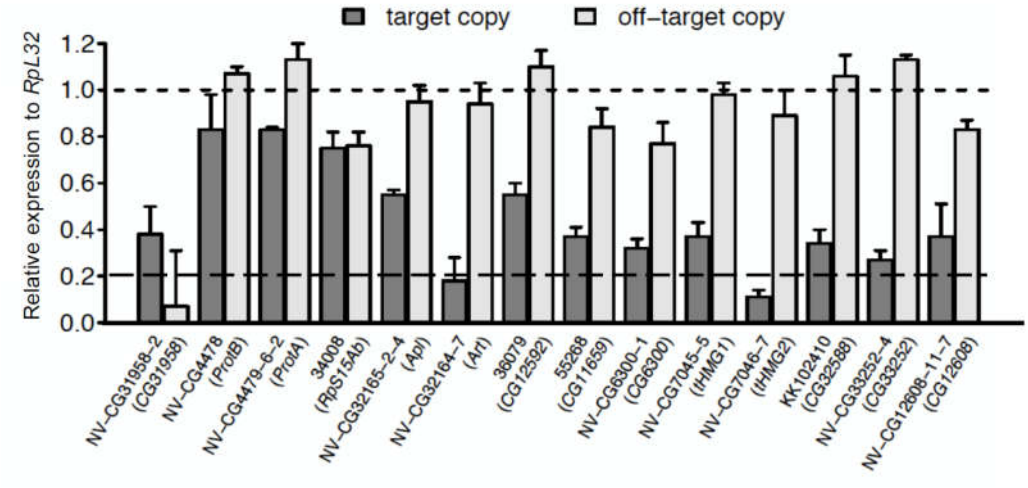


**Figure 4.** Experimental comparison of the efficiency and off-target effects explain the conservative nature of RNAi knockdown experiments and limited off-targets propensity. For each young duplicate gene pair specific for *D. melanogaster* and *melanogaster* species complex, we examined their expression intensity relative to the wide type control in whole body flies with qPCR. The standard deviation is calculated based on three replicates.

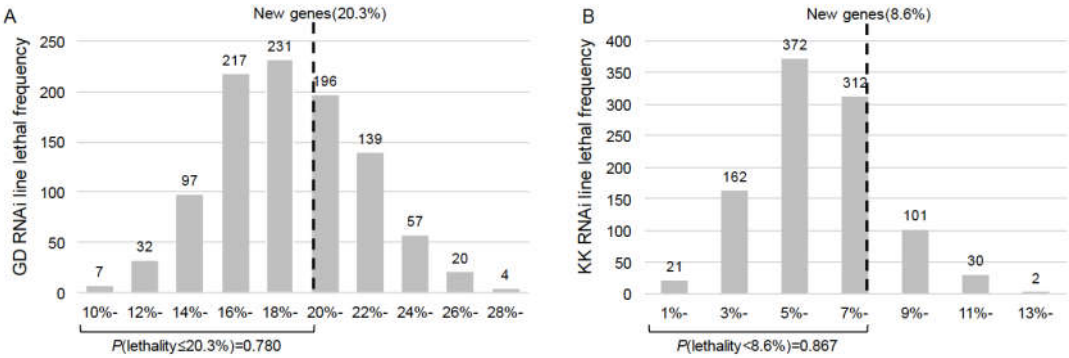


**Figure 5.** Comparison of proportions of lethality between new genes and old genes in GD lines (**A**) and KK lines (**B**) suggests that in both GD and KK lines, new genes have an equally high probability to be lethal as old genes. Since old genes are much more abundant than new genes, we generated 1000 random sample of old genes with the same number of new genes and then plotted the distribution of proportion of essential genes as histograms.

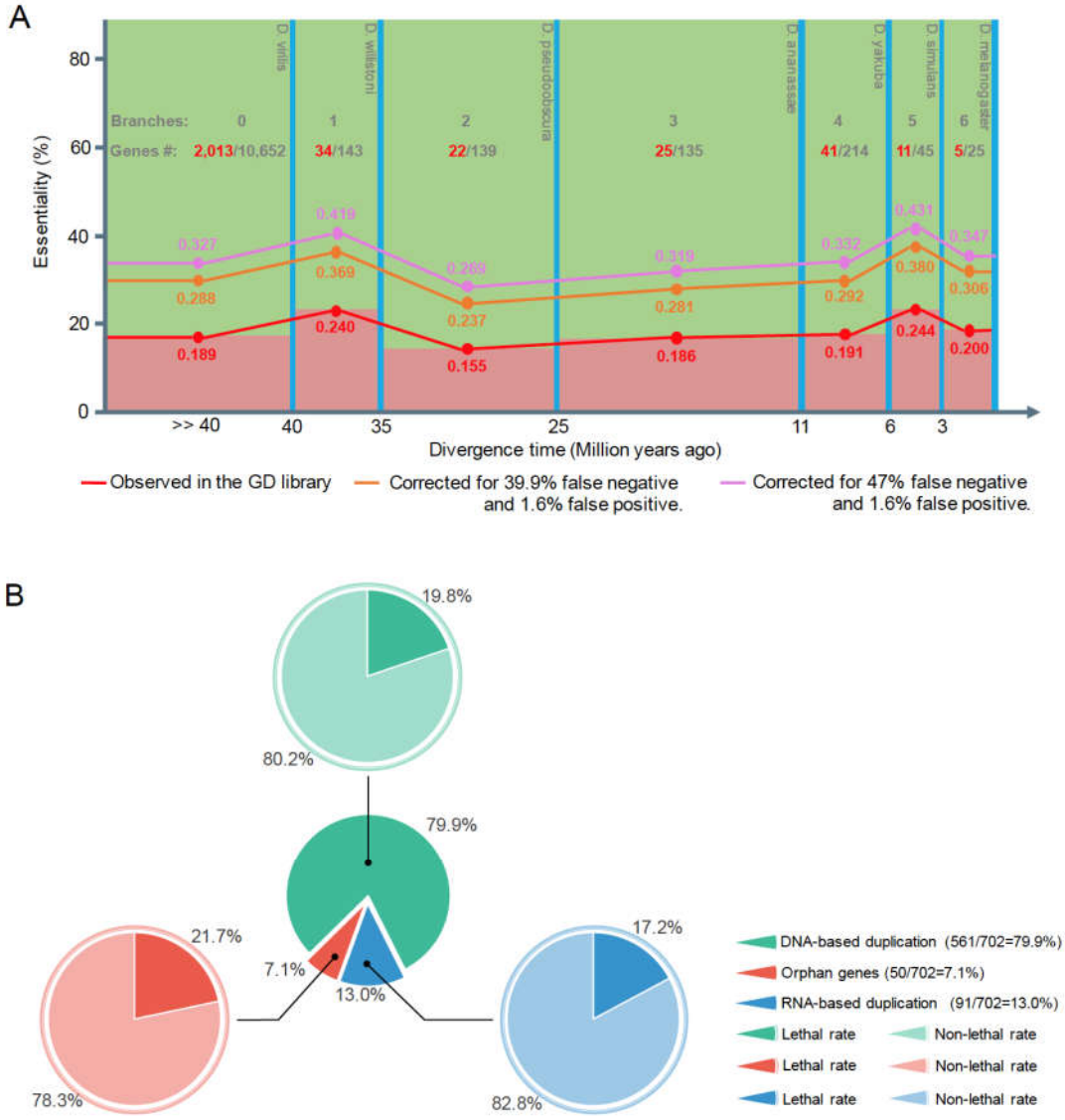


**Figure 6.** Lethality proportion of 702 *Drosophila*-specific genes. A. Lethality proportion of 702 *Drosophila*-specific genes in 6 ancestral stages of extant *D. melanogaster*, compared to the lethality proportion of 10,652 genes older than 40 Mya. No stages show an essentiality proportion significantly different from that of old genes (0.189). B. Lethality proportion of 702 *Drosophila*-specific genes based on three origin mechanism catalogs. No catalog shows a lethality proportion significantly different from that of old genes (0.189).

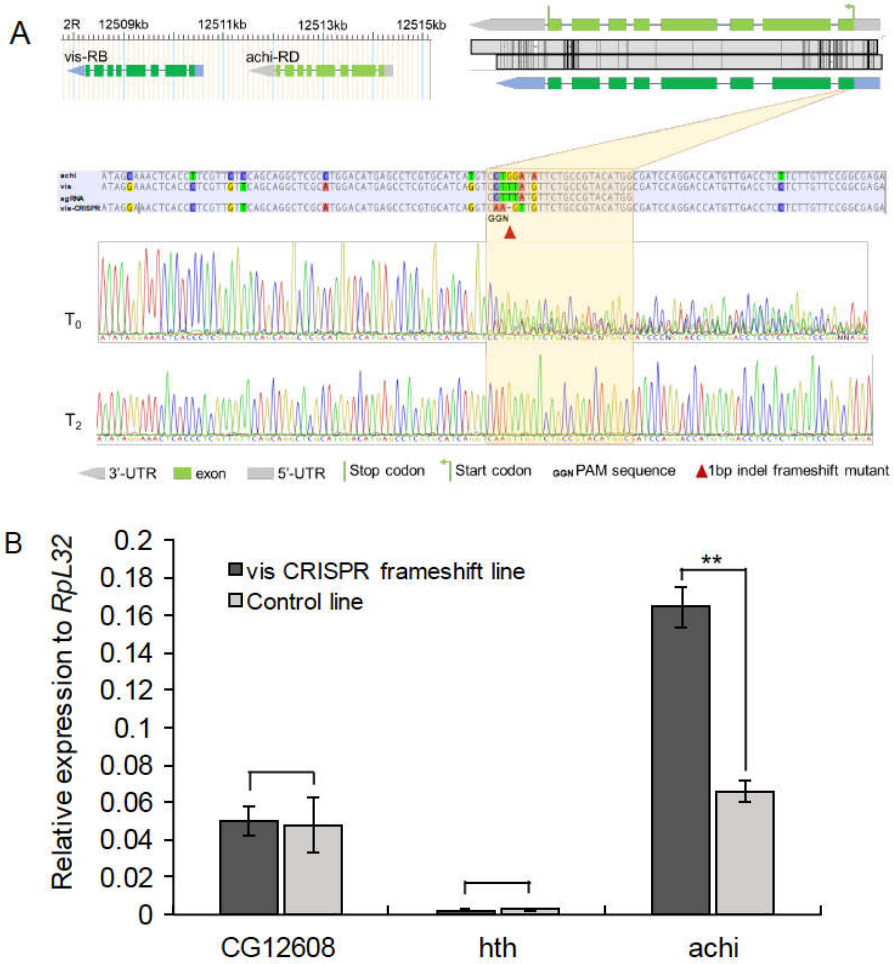


**Figure 7.** CRISPR/Cas9 frameshift mutant could induce compensatory effect in *Drosophila*. A. Design of CRISPR/Cas9 mutant. We targeted a randomly chosen young gene, *vis*, which emerged via duplication of *achi* in the common ancestor of *melanogaster* species complex. The genomic arrangement of two genes are shown in the upper left panel with the boxes referring to exons and connecting lines as introns. The pair shares a high sequence identity (0.92) in their 9 exons, which is schematically shown in the upper right panel. The middle panel shows the diverged site between *vis* and *achi*, which was chosen to design a short guide RNA (sgRNA) specifically targeting *vis*. The mutation (CTTTA→AAGT) was marked with a red triangle. The raw sanger sequencing data for the initial generation (T0) and the second generation of offspring (T2) was shown. B. The compensation effect of *achi*. In the frameshift mutant of *vis*, *achi*'s expression is significantly increased ($P$=0.0003). By contrast, the unrelated *CG12608* and the remotely related *hth* did

837 not show any significant upregulation. *RpL32* was used as a control as in
838 (VanKuren and Long 2018).
839
840 **Supplementary Figures**
841
842 **Figure S1.** Age dating between this work and Kondo *et al.* This figure, following Fig.
843 1 in the main text, adds specific information on how we classified genes into six
844 major categories or dozens of sub categories. For more details, please refer to
845 Materials and Methods.
846
847 **Figure S2.** A representative difficult-to-date locus in the *K*-dataset. A. The syntenic
848 view of *Ste* locus between *D. melanogaster* and *D. simulans* shows fragmented
849 continuity. Due to its multiplicative nature, *Ste* locus is difficult to assemble. In the
850 UCSC Net track, the most assembles can only reach level 2 of one-way syntenic
851 mapping, rather than a better reciprocal syntenic mapping as level 1. B. Some
852 orthologous region in *D. simulans* (lifted from *D. melanogaster*) is not anchored
853 to the chromosome (X) and they are arbitrarily assembled as chrU. C. In *D.*
854 *sechellia*, two scaffolds are assembled with the major scaffold super_20 spanning
855 200 kb, in contrast to the assembly of 15 kb for the orthologous region of *D.*
856 *melanogaster*.
857
858 **Supplementary Tables**
859
860 **Table S1**. The list of genes with the exact ages across the G-dataset and the K-dataset,
861 genes with slightly younger ages in the K-dataset and genes with slightly older
862 ages in the K-dataset, respectively.
863
864 **Table S2**. The list of false negatives and false positives in the K-dataset. Since the *Pan-*
865 *Drosophilid* age group in the K-dataset corresponds to the age group 0, 1 or 2 in
866 the G-dataset (Fig. 1A), we simply replaced the *Pan-Drosophilid* age group as
867 0/1/2 in the table if applicable.
868
869 **Table S3**. 103 knockdown experiments repeated by two independent works (Chen et
870 al. 2010; Zeng et al. 2015). Note, Chen et al (2010) works classified phenotypes as

871 lethal, semi-lethal and viable. Since there are only few genes deemed as semi-
872 lethal, we merged them into lethal gene groups to simplify.
873
874 **Table S4**. For 86 new genes with different RNAi drivers, the consistency between
875 different drivers in Chen et al (2010) and Zeng et al (2015) is listed.
876
877 **Table S5**. The knockdown efficiency data of KK library and GD library.
878
879 **Table S6**. The genotyping results of 153 KK lines, the corrected lines by recombination
880 and knockdown results.
881
882 **Table S7**. Primers for 75 KK lines knockdown efficiency tests.
883
884 **REFERENCES**


885 1. Ashburner M, Misra S, Roote J, Lewis SE, Blazej R, Davis T, Doyle C, Galle
886 R, George R, Harris N et al. 1999. An exploration of the sequence of a 2.9-Mb
887 region of the genome of Drosophila melanogaster: the Adh region. *Genetics*
888 **153**: 179-219.
889 2. Carroll SB. 2005. *Endless Forms Most Beautiful*: The New Science of Evo
890 Devo. W. W. Norton & Company
891 3. Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis
892 N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B et al. 2012. Proto-
893 genes and de novo gene birth. *Nature* **487**: 370-374.
894 4. Chen SD, Krinsky BH, Long MY. 2013. New genes as drivers of phenotypic
895 evolution. *Nat Rev Genet* **14**: 645-660.
896 5. Chen SD, Spletter M, Ni XC, White KP, Luo L, Long M. 2012. Frequent recent
897 origination of brain genes shaped the evolution of foraging behavior in
898 Drosophila. *Cell Rep* **1**: 118-132.
899 6. Chen SD, Zhang YE, Long MY. 2010. New genes in Drosophila quickly
900 become essential. *Science* **330**: 1682-1685.
901 7. Dietzl G, Chen D, Schnorrer F, Su KC, Barinova Y, Fellner M, Gasser B,
902 Kinsey K, Oppel S, Scheiblauer S et al. 2007. A genome-wide transgenic RNAi
903 library for conditional gene inactivation in Drosophila. *Nature* **448**: 151-156.
904 8. El-Brolosy MA, Kontarakis Z, Rossi A, Kuenne C, Gunther S, Fukuda N, Kikhi
905 K, Boezio GLM, Takacs CM, Lai SL et al. 2019. Genetic compensation
906 triggered by mutant mRNA degradation. *Nature* **568**: 193-197.
907 9. El-Brolosy MA, Stainier DYR. 2017. Genetic compensation: A phenomenon in
908 search of mechanisms. *Plos Genet* **13**: e1006780.
909 10. Gould SJ. 2002. The structure of evolutionary theory. Kelknap Press of Harvard
910 University Press. Cambridge, Massashusetts and London, England.
911 11. Green EW, Fedele G, Giorgini F, Kyriacou CP. 2014. A Drosophila RNAi
912 collection is subject to dominant phenotypic effects. *Nat Methods* **11**: 222.

913 12. Hu Y, Sopko R, Foos M, Kelley C, Flockhart I, Ammeux N, Wang X, Perkins
914 L, Perrimon N, Mohr SE. 2013. FlyPrimerBank: an online database for
915 Drosophila melanogaster gene expression analysis and knockdown evaluation
916 of RNAi reagents. *G3* **3**: 1607-1616.
917 13. Huang N, Lee I, Marcotte EM, Hurles ME. 2010. Characterising and predicting
918 haploinsufficiency in the human genome. *Plos Genet* **6**: e1001154.
919 14. Jacob F. 1977. Evolution and tinkering. *Science* **196**: 1161-1166.
920 15. Kasinathan B, Colmenares SU, McConnell H, Young JM, Karpen GH, Malik
921 HS, 2020. Innovation of heterochromatin functions drives rapid evolution of
922 essential ZAD-ZNF genes in *Drosophila* . bioRxiv
923 doi.org/10.1101/2020.07.08.192740.
924 16. Kersey PJ, Allen JE, Armean I, Boddu S, Bolt BJ, Carvalho-Silva D,
925 Christensen M, Davis P, Falin LJ, Grabmueller C. 2015. Ensembl Genomes
926 2016: more genomes, more complexity. *Nucleic Acids Res* **44**: D574-D580.
927 17. Koerich LB, Wang XY, Clark AG, Carvalho AB. 2008. Low conservation of
928 gene content in the Drosophila Y chromosome. *Nature* **456**: 949-951.
929 18. Kondo S, Vedanayagam J, Mohammed J, Eizadshenass S, Kan LJ, Pang N,
930 Aradhya R, Siepel A, Steinhauer J, Lai EC. 2017. New genes often acquire
931 male-specific functions but rarely become essential in Drosophila. *Genes Dev*
932 **31**: 1841-1846.
933 19. Krebs JE, Gildstein ES and Kilpatrick ST. 2013. Lewin's essential genes. Jones
934 & Bartlett Publishers.
935 20. Lee YCG, Ventura IM, Rice GR, Chen D-Y, Colmenares SU, and Long M.
936 2019. Rapid Evolution of Gained Essential Developmental Functions of a
937 Young Gene via Interactions with Other Essential Genes. *Mol Biol Evol* **36**:
938 2212–2226.
939 21. Liebeskind BJ, McWhite CD, Marcotte EM. 2016. Towards Consensus Gene
940 Ages. *Genome Biol Evol* **8**: 1812-1823.
941 22. Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using
942 real-time quantitative PCR and the 2(T)(-Delta Delta C) method. *Methods* **25**:
943 402-408.
944 23. Long MY, VanKuren NW, Chen SD, Vibranovski MD. 2013. New gene
945 evolution: Little did we know. *Annu Rev Genet* **47**: 307-333.
946 24. Loppin B, Lepetit D, Dorus S, Couble P, Karr TL. 2005. Origin and
947 neofunctionalization of a Drosophila paternal effect gene essential for zygote
948 viability. *Current Biology* **15**: 87-93.
949 25. Ma Y, Creanga A, Lum L, Beachy PA. 2006. Prevalence of off-target effects in
950 Drosophila RNA interference screens. *Nature* **443**: 359-363.
951 26. Ma Z, Zhu P, Shi H, Guo L, Zhang Q, Chen Y, Chen S, Zhang Z, Peng J,
952 Chen J. 2019. PTC-bearing mRNA elicits a genetic compensation response via
953 Upf3a and COMPASS components. *Nature* **568**:259–263.
954 27. Mayr EJ. 1982. The Growth of Biological Thought - Diversity, Evolution, and
955 Inheritance. *New York Rev Books* **29**: 41-42.
956 28. Ni JQ, Markstein M, Binari R, Pfeiffer B, Liu LP, Villalta C, Booker M, Perkins
957 L, Perrimon N. 2008. Vector and parameters for targeted transgenic RNA
958 interference in Drosophila melanogaster. *Nat Methods* **5**: 49-51.
959 29. Nurminsky DI, Nurminskaya MV, De Aguiar D, Hartl DL. 1998. Selective
960 sweep of a newly evolved sperm-specific gene in Drosophila. *Nature* **396**: 572-
961 575.

962 30. Perrimon N, Engstrom L, Mahowald AP. 1989. Zygotic Lethals with Specific
963 Maternal Effect Phenotypes in Drosophila-Melanogaster .1. Loci on the X-
964 Chromosome. *Genetics* **121**: 333-352.
965 31. Raices JB, Otto PA, Vibranovski MD. 2019. Haploid selection drives new gene
966 male germline expression. *Genome Res* **29**: 1115-1122.
967 32. Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans
968 M, Smith KE, Rosenbloom KR, Raney BJ et al. 2010. The UCSC Genome
969 Browser database: update 2010. *Nucleic Acids Res* **38**: D613-D619.
970 33. Rogers RL, Cridland JM, Shao L, Hu TT, Andolfatto P, Thornton KR. 2014.
971 Landscape of standing variation for tandem duplications in Drosophila yakuba
972 and Drosophila simulans. *Mol Biol Evol* **31**: 1750-1766.
973 34. Ross BD, Rosin L, Thomae AW, Hiatt MA, Vermaak D, de la Cruz AFA, Imhof
974 A, Mellone BG, Malik HS. 2013. Stepwise evolution of essential centromere
975 function in a Drosophila neogene. *Science* **340**: 1211-1214.
976 35. Rossi A, Kontarakis Z, Gerri C, Nolte H, Holper S, Kruger M, Stainier DYR.
977 2015. Genetic compensation induced by deleterious mutations but not gene
978 knockdowns. *Nature* **524**: 230-235.
979 36. Schroeder, CM, Tomlin, SA, Valenzuela, JR and Malik, HS, 2020. A rapidly
980 evolving actin mediates fertility and developmental tradeoffs in Drosophila.
981 bioRxiv.
982 37. Serobyan V, Kontarakis Z, El-Brolosy MA, Welker JM, Tolstenkov O,
983 Saadeldein AM, Retzer N, Gottschalk A, Wehman AM, Stainier DYR, 2020.
984 Transcriptional adaptation in Caenorhabditis elegans *eLife* **9**: e50014.
985 38. Shao Y, Chen C, Shen H, He BZ, Yu D, Jiang S, Zhao S, Gao Z, Zhu Z, Chen
986 X et al. 2019. GenTree, an integrated resource for analyzing the evolution and
987 function of primate-specific coding genes. *Genome Res* **29**: 682-696.
988 39. St Pierre SE, Ponting L, Stefancsik R, McQuilton P, FlyBase C. 2014. FlyBase
989 102--advanced approaches to interrogating FlyBase. *Nucleic Acids Res* **42**:
990 D780-D788.
991 40. Vakirlis N, Acar O, Hsu B, Castilho Coelho N, Van Oss SB, Wacholder A,
992 Medetgul-Ernar K, Bowman RW, 2nd, Hines CP, Iannotta J et al. 2020a. De
993 novo emergence of adaptive membrane proteins from thymine-rich genomic
994 sequences. *Nat Commun* **11**: 781.
995 41. Vakirlis N, Carvunis AR, McLysaght A. 2020b. Synteny-based analyses
996 indicate that sequence divergence is not the main source of orphan genes. *elife*
997 **9**.
998 42. VanKuren NW, Long MY. 2018. Gene duplicates resolving sexual conflict
999 rapidly evolved essential gametogenesis functions. *Nat Ecol Evol* **31**: 705-712.
1000 43. Ventura IM. 2019. Functional Evolution of Young Retrogenes with Regulatory
1001 Roles in Drosophila. The University of Chicago Ph.D. dissertation
1002 10.6082/uchicago.1799.
1003 44. Vissers JHA, Manning SA, Kulkarni A, Harvey KF. 2016. A Drosophila RNAi
1004 library modulates Hippo pathway-dependent tissue growth. *Nat Commun* **7**:
1005 10368.
1006 45. Xie C, Bekpen C, Kunzel S, Keshavarz M, Krebs-Wheaton R, Skrabar N,
1007 Ullrich KK, Tautz D. 2019. A de novo evolved gene in the house mouse
1008 regulates female pregnancy cycles. *elife* **8**.
1009 46. Yeh SD, Do T, Chan C, Cordova A, Carranza F, Yamamoto EA, Abbassi M,
1010 Gandasetiawan KA, Librado P, Damia E. 2012. Functional evidence that a

1011 recently evolved Drosophila sperm-specific gene boosts sperm competition.
1012 *Proc Natl Acad Sci* **109**: 2043-2048.
1013 47. Zeng XK, Han LL, Singh SR, Liu HH, Neumuller RA, Yan D, Hu YH, Liu Y,
1014 Liu W, Lin XH et al. 2015. Genome-wide RNAi screen identifies networks
1015 involved in intestinal stem cell regulation in Drosophila. *Cell Rep* **10**: 1226-
1016 1238.
1017 48. Zhang L, Ren Y, Yang T, Li G, Chen J, Gschwend AR, Yu Y, Hou G, Zi J,
1018 Zhou R et al. 2019. Rapid evolution of protein diversity by de novo origination
1019 in Oryza. *Nat Ecol Evol* **3**: 679-690.
1020 49. Zhang YE, Vibranovski MD, Krinsky BH, Long M. 2010. Age-dependent
1021 chromosomal distribution of male-biased genes in Drosophila. *Genome Res* **20**:
1022 1526-1533.
1023 50. Zhou Q, Zhang GJ, Zhang Y, Xu SY, Zhao RP, Zhan ZB, Li X, Ding Y, Yang
1024 S, Wang W. 2008. On the origin of new genes in Drosophila. *Genome Res* **18**:
1025 1446-1455.
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036