Supplementary Figures and Tables for Kondo and Vedanayagam, "New genes often acquire male-specific functions but rarely become essential in Drosophila"

Supplementary Figure 1. Spatiotemporal expression patterns of recently emerged "RNAi-lethal" genes.

Chen et al (Science 2010) used RNAi assays to deplete 195 genes that had emerged recently in the lineage towards *D. melanogaster*, resulting in classification of 59 "RNAi-lethal" and 136 "RNAi-viable" young genes. We analyzed their expression patterns across the diversity of developmental stage, dissected tissue, and cell line RNA-seq datasets reported by the modENCODE project. The 59 "RNAi-lethal" loci are summarized in here and the 136 "RNAi-viable" young genes are presented in Supplementary Figure 2. (Boxed legend) Selected tissues and stages are colored to enable browsing of spatiotemporal expression patterns. Mean RPKM values and standard deviation from replicate RNA-seq datasets is potted. In particular, the datasets that are most relevant to the male reproductive system are placed at the far right of each gene summary; e.g. testis, accessory gland, male body. Also included with these is unsexed larval imaginal disc/gonad library data, which we have previously observed is rich in testis-specific transcripts and likely represents developmental expression of genes that are also expressed in the adult male reproductive system. For this reason, we also highlight larval and pupal libraries, which although unsexed, undoubtedly include expression from developing male gonads. (Below) Spatiotemporal expression patterns of 59 "RNAi-lethal" young genes. More than half of these genes are highly restricted to the male reproductive system, which is not needed for viability. Many other genes exhibit expression patterns and/or levels that are not overtly expected for lethal phenotypes. These spatiotemporal expression patterns mirror those of the 136 "RNAi-viable" genes (Supplementary Figure 2), suggesting that the RNAi lethality might not be due to suppression of the intended loci.

Supplementary Figure 2. Spatiotemporal expression patterns of recently emerged "RNAi-viable" genes.

Chen et al (Science 2010) used RNAi assays to deplete 195 genes that had emerged recently in the lineage towards *D. melanogaster*, resulting in classification of 59 "RNAi-lethal" and 136 "RNAi-viable" young genes. We analyzed their expression patterns across the diversity of developmental stage, dissected tissue, and cell line

RNA-seq datasets reported by the modENCODE project. The 59 “RNAi-lethal” loci are presented in Supplementary Figure 1 and the 136 “RNAi-viable” young genes are summarized here. (Boxed legend) Selected tissues and stages are colored to enable browsing of spatiotemporal expression patterns. Mean RPKM values and standard deviation from replicate RNA-seq datasets is potted. In particular, the datasets that are most relevant to the male reproductive system are placed at the far right of each gene summary; e.g. testis, accessory gland, male body. Also included with these is unsexed larval imaginal disc/gonad library data, which we have previously observed is rich in testis-specific transcripts and likely represents developmental expression of genes that are also expressed in the adult male reproductive system. For this reason, we also highlight larval and pupal libraries, which although unsexed undoubtedly include expression from developing male gonads. (Below) Spatiotemporal expression patterns of 136 “RNAi-viable” young genes. More than half of these genes are highly restricted to the male reproductive system, which is not needed for viability. Many other genes exhibit expression patterns and/or levels that are not overtly expected for lethal phenotypes. These spatiotemporal expression patterns mirror those of the 59 “RNAi-lethal” genes (Supplementary Figure 1), suggesting that the RNAi lethality might not be due to suppression of the intended loci.

Supplementary Figure 3. Examples of chains and nets evidence for young gene emergence. Many recently-emerged genes are close paralogs of existing loci. The global multi-genome alignments visualized in the UCSC Genome Browser often represent paralogous sequences together that can obscure true gene age. In the example of CG17268 (*Prosα4T1*), Multiz alignments give the impression that it is conserved throughout the Drosophilid phylogeny. However, inspection of underlying chains and nets shows that this is not the case. While all alignments from species between *D. melanogaster* and *D. ficusphila* in the lineage place CG17268 and its neighboring genes on the same level, all of the more distant species place the CG17268-homologous sequence on a different level. In fact, the parental copy *Prosα4* has been misaligned in all of these other species. Note that *Prosα4* is ubitquitously expressed, including in the male reproductive system, whereas *Prosα4T1* is largely restricted to the male reproductive system.

Supplementary Figure 4. Expanding the set of young *D. melanogaster* gene annotations. Previous efforts by Long and colleagues defined 566 D. melanogaster genes that were born since the last common ancestor with *D. ananassae* ("mel-group"), which is intermediate in the sequenced Drosophilid phylogeny; about 3/4 of these emerged more recently within mel-subgroup species. By systematic assessment of modENCODE gene models and stringent inspection of genome syntenies, we were able to double the number of young *D. melanogaster* genes.

Supplementary Figure 5. Bimodal distribution of tissue specificity of *Drosophila* genes. (A) We calculated tissue specificity index (τ) for 12974 genes based on modENCODE RNA-seq data from 80 developmental stages, dissected tissues and cell lines. (B) Box and whiskers distribution plot of τ values; median is 0.614.

Supplementary Figure 6. Summary of expression biases of young *Drosophila* genes to the male reproductive system. These barplots summarize the library type of maximal expression ("xmax") from ~80 developmental stages/tissues/cell lines RNA-seq datasets, for sets of evolutionarily young *Drosophila* genes across various age groups. These data represent summaries of individual gene expression plots provided in Supplementary Figures 7-10 for mel-group genes (present in *Dmel/sim/sech/ere/yak/ana*), mel-subgroup genes (present in *Dmel/sim/sech/ere/yak*), mel-complex genes (present in *Dmel/sim/sech*) and genes that are present only in *Dmel*. Note the preponderance of genes whose "xmax" is in testis or accessory glands.

Supplementary Figure 7. Spatiotemporal expression patterns for individual *D. melanogaster* genes lacking orthologs in other *Drosophila* species.

Supplementary Figure 8. Spatiotemporal expression patterns for individual *D. melanogaster* genes with only mel-complex orthologs (present in *Dmel/sim/sech*).

Supplementary Figure 9. Spatiotemporal expression patterns for individual *D. melanogaster* genes with only mel-subgroup orthologs (present in *Dmel/sim/sech/ere/yak*).

Supplementary Figure 10. Spatiotemporal expression patterns for individual *D. melanogaster* genes with only mel-group orthologs (present in *Dmel/sim/sech/ere/yak/ana*).

Supplementary Figure 11. Transcriptome age indices (TAI) for Drosophila tissues, stages and cells. The TAI is weighted expression metric that utilizes evolutionary information to convey the relative age of the transcriptome. The libraries with the highest outlier TAI scores are testis, adult male and accessory glands, while the next highest cohort includes unsexed developmental stages that include male gonad material (larval imaginal disc/gonad, pupal stages, and certain larval stages). Note that the lowest TAI cohort includes all the different cell lines, as well as ovaries and early embryos that contain maternally deposited transcriptome.

Supplementary Figure 12. Comparison of RNAi-induced lethality for recently-emerged genes using independent triggers. We show two pie charts, representing 59 "young, essential" and 136 "young, non-essential" genes classified by Chen and Long (*8*). Although the vast majority of RNA-induced lethality phenotypes were originally reported to be confirmed by an additional transgene, nearly all of these validations involved insertions of the same trigger. We compiled data from actin5C-Gal4 tests using different RNAi triggers, where available. Genes for which no independent lines are available, and concordance on lethal/viable calls, are shown in white; genes for which tests using the same trigger as in Chen and Long but resulted in a discordant phenotypic call are shown in blue as "discordant 0 different lines". Red/orange slices depict phenotypic concordance between Chen 2010 and other studies, Blue/Green slices represent discordant lethal/viable calls between Chen 2010 and other studies using different lines. Grey slices represent situations with 3 or more RNAi lines, comprising outcomes that are concordant as well as discordant with the results originally reported by Chen and Long 2010.

Supplementary Figure 13. Viable BACE CRISPR mutants reproduce glial degeneration in known *BACE* alleles. (A) Western blot analysis shows that newly-generated CRISPR alleles lack detectable BACE protein. (B-C) Sections of 32 day old fly retinas. (B) Control *w[1118]* eye. In between the retina and the lamina neuropil lies the lamina cortex, which contains the glial and neuronal cell bodies; the lamina neuropil contains only fibers and

no cell bodies. (C) CRISPR *BACE* mutant shows degeneration of glial cells in the lamina cortex (holes marked by arrows). Note that some retinal degeneration is also visible in the *BACE* mutant, but this is usually seen to some extent in *white* mutants, although it may be exacerbated by the glial death in the lamina cortex of *BACE* mutants.

Supplementary Figure 14. Spermatogenesis in *prosα4T1[SK2]* homozygotes. (A-E) In wild-type testes, the nuclei of differentiated spermatids reside in the basal region of the testis, and flagella extend apically. (A) Each nuclear bundle forms an actin-rich individualization complex (IC), and the ICs (red, phalloidin, arrows) move away from the nuclear bundles (blue, DAPI), traversing the sperm tails and eventually forming waste bags in the apical region (arrowheads). (B) Individualizing cysts stain positively for active caspase (green, cleaved Caspase-3) as they enact an apoptosis-like program to degrade extruded cytoplasm. (C) Differentiated spermatid nuclei incorporate protamines (green, protamine-GFP). (D) ICs (red, phalloidin, arrows) are visible forming around bundled basal nuclei (D', arrows). Following individualization, mature sperm are coiled in the base of the testis, and mature sperm nuclei become tightly compacted (D' arrowheads). (E) Progressing ICs are highly organized during their transit. (F-J) In *prosα4T1* mutants, normal Caspase-3 staining is seen in individualizing spermatids (G), protamines are incorporated normally into the maturing spermatid nuclei (H), and ICs are formed around bundled nuclei of individualizing spermatids (I, arrows). However, progressing ICs (J) and nuclei in the coiling region (I', arrowheads) are disorganized. (K) In the germinal proliferation center at the apical tip of the testis, a-spectrin staining reveals the round spectrosomes (green, arrowheads) of the germline stem cells at the hub (asterisk) and of the gonialblasts. Branched fusomes are observed in mitotic spermatogonia (green, arrows). (L) Phase contrast imaging of testis squashes reveals the post-meiotic (pre-individualization) spermatids, each with one phase-dark mitochondrial aggregate per phase-light nucleus. (M) Mature sperm are deposited in the seminal vesicle for storage until mating (DAPI). (N-P) *prosα4T1* mutants show normal a-spectrin staining (N) and normal post-meiotic spermatids (O). However, mutant seminal vesicles are devoid of mature sperm (P). Scale bars are 20 μm.

Supplementary Figure 15. Mutations in Spn77Bc display spermatid individualization defects. (A) Western Blot showing that two different alleles of CG6289/Spn77Bc are strong loss-of-protein alleles. (B-D) At the apical hub of the testis (asterisk), stem cells

and gonialblasts, marked by round spectrosomes (anti-spectrin, green, arrowheads), are normal in the two mutant alleles of Spn77Bc (SK1 and SK4). Mitotic spermatogonia, marked by branched fusomes (a-spectrin, green,arrows), are also normal. (E-G) Meiotic cytokineses are normal, as evidenced by the appearance of one phase dark mitochondrial aggregate for each phase light nucleus in squashed testis preparations. (H-M) In Spn77Bc mutants, Caspase-3 staining is normal (H-J, cleaved Caspase-3, green), as are nuclear condensation and bundling (K-M, DAPI, blue) and individualization complex formation (K-M, phalloidin, red). (N-P) In Spn77Bc mutants, individualization complexes (phalloidin) disperse abnormally as they progress away from the nuclei towards the apical region of the testis. (Q-S) In the weaker mutant allele, SK1, mature needle shaped sperm nuclei (DAPI) are visible in the seminal vesicle (R), while in the stronger allele, SK4, fewer mature sperm are seen (S). In R, only the round nuclei of the seminal vesicle wall are evident. Scale bars are 20 μm. (T) Spn77Bc mutants have individualization defects, showing significantly more abnormal individualization complexes (ICs) than controls. *, $p < 0.005$. **, $p < 0.0001$. (U) Despite testis cytology defects, Spn77Bc mutants have normal fertility.

Supplementary Figure 16. Summary of dN/dS analyses and DoS statistics
(A) Genomewide distribution of dN/dS for *D. melanogaster* genes. The distribution is skewed to the left with a genome average of 0.16. (B, C) dN/dS ratios of tissue-restricted genes ($\tau>0.61$, see Figure 1C) were calculated for loci that were highest expressed in the male reproductive system (B) or in other tissues/stages/cell types (C). Genome average of dN/dS is indicated with a dotted red line, the dots in each box-plot represent means and values are indicated; the median is shown as a midline in the box plots. dN/dS for older genes are much below the genome average indicating purifying selection, while genes in the younger age groups show higher dN/dS relative to genome average, a pattern indicative of positive selection. However, high dN/dS ratio can result from either positive selection or relaxed selective constraint; and dN/dS cannot distinguish these two alternatives due to the lack of within-species polymorphism data in the analysis, which is a limitation of this analysis. While dN/dS has limitations to detect positive selection, synonymous changes inferred from dN/dS do not confound patterns of purifying selection. (D) dN/dS ratios according to gene age for broadly expressed genes (τ <=0.614) that happen to be highest (Xmax) in male reproductive system. (E) dN/dS ratios for broadly expressed genes with Xmax outside of the male reproductive

system. While the dN/dS ratios of younger genes tends to be higher than the genome average in both (D) and (E), there are relatively few genes in this categories and the skew is lower than for young genes with high tissue-restriction to the male reproductive system (B). On the other hand, highly conserved, ubiquitously expressed older genes tend to show a strong signature of purifying selection, as the average dN/dS ratio for these age categories is well below the genome average.

(F) Genomewide distribution of DoS values for D. melanogaster genes. The distribution follows a normal distribution with an average of -0.04. For DoS statistics divided according to gene age classes, we observe the average DoS of different classes of broadly expressed young genes with Xmax male reproductive system (G) vs Xmas in other cells/tissues (H) are mostly below the genome average, indicative of purifying selection. Two exceptions are mel-subgroup in Panel G, and mel-group in Panel H for which DoS is above the genome average, but still close to neutral evolution (0.01 and 0.06, respectively). Overall, the trend of DoS in broadly expressed genes is consistent with purifying selection. This contrasts with the behavior of young genes with high tissue-restriction to the male reproductive system exhibits strongly positive DoS (see main Figure 3).

Supplementary Table 1. Evolutionary ages and spatiotemporal expression of *D. melanogaster* genes. Summary of gene ages and expression values in FPKM across 93 developmental stages, tissues and cell lines comprising 195 RNA-seq datasets. For each gene, the inferred evolutionary age, gene names, aliases (FBgn and CGID) and tissue specificity index (τ) are listed. We highlighted genes from old multi-gene families, for which age assessments amongst the Cellular organisms/Eukaryota/Bilateria classifications cannot be made with high degree of certainty; the age assignment by ProtHistorian used in Fig. 1 is given.

Supplementary Table 2. Summary of RNAi-induced phenotypes. The initial study of Chen and Long (Science 2010) reported lethality phenotypes from RNAi-mediated suppression of 59 evolutionary young genes. This table summarizes data for these 59 genes from published screens (*18-20*) of the VDRC "KK" and "GD" collections, the Harvard "TRiP" collection and the NIG RNAi collection. We parsed them for their agreement of viability/lethality calls across different insertions of the same RNAi trigger,

and for the agreement of different triggers against the same gene, using the ubiquitous driver Act5C-Gal4, and summarize information from other available Gal4 drivers. We highlight where discrepant lethal/viable phenotypes were obtained for the same or for different RNAi triggers for the same gene. The "CRISPR mutants" column indicates loci that we have mutagenized to obtain viable frameshift alleles. Note that for three genes with viable frameshift CRISPR mutants, there was apparent evidence of independent RNAi triggers yielding lethality.

Supplementary Table 3. CRISPR mutagenesis of 21 conserved, known essential genes

This table summarizes the molecular details for CRISPR/Cas9-mediated mutagenesis for 21 conserved, known essential genes on chromosome 2L of D. melanogaster. The molecular lesions and inferred mutant proteins are produced are listed. Excepting the two alleles of *BACE* which were viable, all 33 alleles of the remaining 20 loci were lethal. Note that further experimental tests indicate that viable CRISPR-*BACE* alleles are protein-null and recapitulate retinal glial degeneration phenotypes of known *BACE* mutants, indicating that *BACE* is likely not an essential gene.

Supplementary Table 4. CRISPR mutagenesis of 24 young, "RNAi essential" genes

This table summarizes the molecular details for CRISPR/Cas9-mediated mutagenesis for 22 young *D. melanogaster* genes that were proposed to be essential based on transgenic RNAi assays. The molecular lesions and inferred mutant proteins are produced are listed. Excepting single alleles of CG7594[SK3], CG17802[SK2], and CG17268[SK13], all other alleles of all other loci were homozygous viable on their initially isolated chromosomes. However, CG7594[SK3], CG17802[SK2], and CG17268[SK13] were viable over deficiencies, indicating that lethality was not due to loss-function of the mutagenized young gene. Note that CG17268[SK13]/Df was completely male sterile, as was the case for homozygous and hemizygous combinations of two other different frameshift alleles of CG17268, [SK2] and [SK11].