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

Analysis of *Drosophila* spatial expression preference

modENCODE poly(A)+ RNA profiling data in *D. melanogaster* for 189 libraries was obtained from SRA Bioproject ID: PRJNA75285 (Brown et al. 2014). Since individual *Drosophila* strains can exhibit wide variation in testis gene expression (Zhao et al. 2014), we also included 6 *D. melanogaster* testis RNA-seq datasets (Raleigh lines) from SRA Bioproject ID: PRJNA210329. The paired-end RNA-seq data was mapped to r6.13 version of the *D. melanogaster* genome using subread-aligner from Subread software suite (Liao et al. 2013). Expression profiling was then obtained for the r6.13 annotation using featureCounts software from the Subread package; and gene expression values were normalized to FPKM using the DEseq package from R Bioconductor. A summary of *D. melanogaster* expression data across all RNA-seq datasets is provided in **Supplementary Table 1**.

Tissue specificity index

The breadth of gene expression in a given tissue *i,* was estimated using the tissue specificity index (Yanai et al. 2005), where in the equation given below, *N* is the number of tissues, x_i is the expression in tissue i, and x_{max} is the maximum expression of the gene in all tissues. τ ranges from 0 to 1; values closer to 1 indicate high tissue specificity. τ values were estimated for 13190 *D. melanogaster* genes for which expression data was available in our dataset. Genome-wide τ distribution shows a bimodal trend, with broadly two classes of genes: ubiquitously expressed genes (n=6490) and genes expressed in a tissue-specific setting (n=6484).

$$\tau = \frac{\sum_{i=1}^{N} 1 - \frac{x_i}{x_{max}}}{N-1}$$

Analysis of *Drosophila* gene ages

Assignment of gene ages and orthologs is challenging process to conduct on the genome scale (Hu et al. 2011). Often a deep assessment must be carried out on a individual case basis to fully understand evolutionary history. We have used the

following hybrid strategies to assign ages to *D. melanogaster* genes, where we (1) incorporated careful and systematic gene-by-gene inspection to guarantee confidence in the birthdate of young genes, and (2) a series of steps to assign ages to genes that already existed in the pan-Drosophilid ancestor.

(1) Evolutionarily recent genes from (Chen et al. 2010; Chen et al. 2012) (566 genes) were re-assessed using UCSC Genome Browser chains and nets alignments. We supplemented this catalog with additional genes from *D. melanogaster* Release r6.13, paying particular attention to genes for which no known orthologs were listed (358), genes for which orthologs were not listed in all 12 *Drosophila* reference genome species (495) (Clark et al. 2007), and genes for which one or more duplicates were listed in *D. melanogaster* (225) as potential genes to be recently-emerged. These genes were manually verified for contiguity in chains and nets alignments. Genomic loci ambiguous to determine ages by syntenic alignment such as histone genes (*n*=112) and mod(mdg4) loci (n=23) were excluded from our analysis.

(2) We set aside the bona fide young fly genes and attempted to assign ages to the remainder of the Drosophila genes using the ProtHistorian package (Capra et al. 2012). The ProtHistorian gene-age estimation employs “Dollo” parsimony in age determination, which is a common gain-loss phylogenetic analysis method with an assumption that multiple losses across lineages are common, but gains along a phylogenetic tree are rare, and restricted to a few nodes. In brief, *D. melanogaster* genes were queried for orthologs from PPODv4 (Princeton protein orthology prediction) and OrthoMCL database using age_proteins.py script from the ProtHistorian pipeline. The PPODv4 and OrthoMCL databases encompass protein orthologs from 48 species across the animal phylogeny onto which the queried genes were placed on a node depending on the presence/absence of an ortholog in the phylogeny.

ProtHistorian assigns orthologs based on protein homology, and therefore, genes from multi-gene families with high degree of similarity are likely to be misclassified. Indeed, this was the reason for filtering out all the established young genes from this pipeline, since many young paralogs are prone to misclassification. Nevertheless, we still recognized that ProtHistorian assigned a population of genes assigned to old (i.e., Cellular organisms, Eukaryota and Bilateria) age groups that exhibited high tissue-restriction (τ), which we later determined were actually younger (Dipteran/pan-Drosophilid) paralogs that had been misassigned based on similarity to older parental copies. Since proteins from large gene-families are potential candidates for inaccuracies

in gene age assignment, we performed three additional assessments to improving the accuracy of age classification.

First, we determined homologs using reciprocal best BLAST (RBB), followed by phylogenetic clustering of homologs. In the cases where homologs from a multi-gene family clusters with an updated catalogue of young genes, and lack an ortholog in other insects, we reassigned the multi-gene family member to Pan-Drosophilid age group. Alternatively, if these genes lacked an ortholog in worms, but had an insect-inclusive ortholog, we reassigned the gene to Diptera age group. Second, we performed a fuzzy reciprocal BLAST (Hahn et al. 2007), with relaxed identity parameters to obtain an inclusive list of all likely gene-family members that may have been missed by best-reciprocal BLAST alone, and re-assigned genes to Diptera/Pan-drosophila age class as described above for RBB. Third, to determine if older multi-gene family homologs were misassigned to younger (Diptera/Pan-drosophila) age groups, we queried the ProtHistorian classification to the DIOPT database (Hu et al. 2011) encompassing ortholog predictions from 14 databases for best-consensus orthologs in Bilateria or Eukaryota. On this basis, we reassigned 616 genes initially classified as Dipteran/Pan-Drosophilid to older age groups.

The age assignments are provided in **Supplementary Table S1**. We note that for 1183 old genes that are members of multigene families there remains potential uncertainty as to age assignments (Cellular organisms, Eukaryota and Bilateria), due to high similarity among homologs. In Supplementary Table S1, these are highlighted as old multi-gene family genes likely belonging to Cellular organisms/Eukaryota/Bilateria categories, although the age categories assigned and used in Figure 1 are given.

Population genomic analyses

D. melanogaster whole-genome resequence data was obtained from *Drosophila* Genome Nexus (DGN), Version 1.1. We downloaded DPGP2 (Drosophila population genomics project; *n*=115) (Lack et al. 2015) and DGRP (Drosophila genome reference panel; *n*=204) (Mackay et al. 2012) data as genome assemblies, aligned to dm3 version of the *D. melanogaster* genome. Genome assembly data from DGN represents assemblies only from Chr X, 2L, 2R, 3L and 3R. Since data from DGN do not have assemblies for 2LHet, 2RHet, 3LHet, 3RHet, XHet and Y, we excluded genes from unassembled heterochromatic regions from our population genomic analyses.

Gene alignments

Gene alignments from 319 *D. melanogaster* DPGP and DGRP lines were obtained using the longest isoform for a given gene from *D. melanogaster* r5.45 gene feature format (gff) file. A custom shell script along with GFF and BED files used to obtain gene alignments are made available at the Lai Lab github site (https://github.com/Lai-Lab-Sloan-Kettering)

Outgroup data

D. simulans was used as an outgroup to perform evolutionary analyses. Orthologs and protein coding genes (CDS) for *D. melanogaster* genes in *D. simulans* were obtained from FlyBase. *D. simulans* orthologs were then aligned with *D. melanogaster* sequence using the MUSCLE aligner (http://www.drive5.com/muscle/). To ensure a strict codon-by-codon alignment with respect to *D. melanogaster*, gaps resulting from alignments in the *D. simulans* orthlog were removed.

Divergence (D_N/D_S) analysis

D_N/D_S analysis is an evolutionary test for neutrality (Nei and Gojobori 1986). Defined as the D_N/D_S ratio, when two diverged protein-coding DNA sequences from closely related species are inspected, under a neutral model of molecular evolution, the ratio of the number of non-synonymous (replacement) substitutions per non-synonymous site to the number of synonymous (silent) substitutions per synonymous site is expected to be unity, given as,

$$\frac{D_N}{D_S} = 1 \qquad (1),$$

where, D_N is the non-synonymous (replacement) divergence and D_S is the synonymous (silent) divergence. Unity is interpreted as neutral evolution. However, two other scenarios are possible. First, when

$$\frac{D_N}{D_S} > 1 \qquad (2),$$

the result is interpreted as positive selection, as a result of an excess of non-synonymous divergence. Second, when

$$\frac{D_N}{D_S} < 1 \qquad (3),$$

the result is interpreted as purifying selection, as a result of an excess of synonymous divergence. In *D. melanogaster*, the distribution of dN/dS is left-skewed with a genome average of 0.16, with very few genes exceed the theoretical expectation > 1 to be experiencing positive selection. A genome average of 0.16 indicates that the majority of genes in the genome undergoing purifying selection. Therefore, significant deviations from the genome average with values nearing 1 can be interpreted as evidence for an excess of non-synonymous substitutions—an hallmark of positive selection. However, dN/dS cannot distinguish positive selection from relaxed constraint due to lack of within-species polymorphisms in the analysis. This can be overcome by alternative tests of neutrality such as a McDonald-Kreitman test, which incorporates polymorphism data to contrast positive selection from relaxed evolutionary constraint.

McDonald-Kreitman tests

McDonald-Kreitman (MK) test is a statistical test for neutrality, which incorporates both within-species polymorphisms, and between-species divergence in the evolutionary analysis. Under a neutral expectation of molecular evolution, for a given protein-coding gene, the ratio of non-synonymous (replacement) polymorphism to synonymous (silent) polymorphism is expected to be equal to the ratio of non-synonymous (replacement) divergence to synonymous (silent) divergence, given as

$$\frac{P_N}{P_S} = \frac{D_N}{D_S} \qquad (4),$$

where, P_N is the non-synonymous (replacement) polymorphism and P_S is the synonymous (silent) polymorphism, while D_N is the non-synonymous (replacement) divergence and D_S is the synonymous (silent) divergence. Deviations from this expected neutrality can be statistically evaluated using a 2 X 2 contingency table (McDonald and Kreitman 1991). Statistical rejection of neutrality can be driven by either excess of polymorphisms or divergence. When there is a statistical rejection of neutrality due to excess of non-synonymous divergence, the result is interpreted as positive/directional selection. MK tests are performed on a gene-by-gene case, so to analyze evolutionary

patterns of a cohort of genes, an extension of MK test is usually performed, where using the site types *(P_N ,P_S, D_N ,D_S)* a metric like Direction of Selection (see below) is obtained to compare and contrast groups of genes.

Direction of selection analysis

Direction of selection (DoS) is a statistic, which is an extension of the MK test to determine the direction in which the MK test rejects neutrality. DoS is the difference between the ratio of non-synonymous divergence to total divergence and the ratio of non-synonymous polymorphisms to total polymorphisms for a given gene, given as

$$DoS = \frac{D_N}{(D_N+D_S)} - \frac{P_N}{(P_N+P_S)} \quad (5),$$

where, P_N is the non-synonymous (replacement) polymorphism and P_S is the synonymous (silent) polymorphism, while D_N is the non-synonymous (replacement) divergence and D_S is the synonymous (silent) divergence, derived from the MK test (Stoletzki and Eyre-Walker 2011). Positive DoS values result from an excess of non-synonymous divergent sites, which is interpreted as evidence for positive selection, while negative DoS values result from an excess of non-synonymous polymorphisms, interpreted as purifying selection.

CRISPR/Cas9-mediated mutagenesis

We used the transgenic Cas9/gRNA system (Kondo and Ueda 2013) to perform mutagenesis in the *yw* background. gRNAs were directed against early portions of common coding exons and we conducted the primary mutagenesis with the aim of inducing two different frameshift alleles for each locus. The transgenic gRNA system is sufficiently effective that additional mutants were often recovered in the primary screen; however, due to the large number of alleles being handled, we did not usually keep these other mutants. We provide a full accounting of the mutagenesis pipeline and results in **Supplementary Tables 3** and **4**, which summarize efforts on "old, known lethal genes" and "young, RNAi-lethal genes", respectively. These include primers for gRNA synthesis, the nature of the frameshift alleles obtained, the predicted protein products of the mutant alleles, and lethality/viability of mutants.

We emphasize that CRISPR is not exempt from off-target effects, and highlight the importance of evaluating multiple mutants and hemizygous allelic combinations. For

example, individual alleles of young genes *CG7594*, *CG31882* and *CG17268* were lethal within the initially isolated chromosomes, but these were attributable to second-site aberrations since both were viable over deficiencies, and independent *CG7594* and *CG17268* frameshift alleles were viable (**Supplementary Table 4**). On the other hand, it is also clear that transgenic RNAi phenotypes need to be evaluated carefully. Even having evidence from multiple RNAi triggers can be insufficient to guarantee on-target effects. Amongst the small minority of young genes with independent RNAi triggers yielding lethality (**Supplementary Table 2**), three of these (*CG13559*, *CG10474* and *Prosα4T1*) were viable as CRISPR-induced mutants. Different RNAi triggers may share off-target effects, a scenario that may be challenging to avoid for genes with close paralogs (such as *Prosα4T1*, **Supplementary Figure 2**).

Fertility tests

We crossed individual males to two individual *yw* virgins at 25ºC, all flies at 3-5 days old. The males were discarded after one day, and the females were transferred three more times at three day intervals. This allowed us to assess total fertility as well as progressive decline in progeny yield over time. Each trial typically involved testing 10 males per genotype, and multiple independent trials were run to assess different mutant allelic combinations. For female tests, we conducted bulk assays of at least two vials of virgin females crossed to *yw* males, to obtain a qualitative assessment of their fertility.

Western blotting

We prepared lysates from five 2-3 day old males (to detect CG6289) or from 10 heads (to detect dBACE) by homogenization in cold lysis buffer (10mM Tris-HCl, 300mM NaCl, 1mM EDTA, 1%Triton-X100) containing fresh EDTA free protease inhibitor cocktail (Sigma-Aldrich). Electrophoresis (7µg of total protein per sample) was carried out in a 4–20% Mini-PROTEAN® TGX™ Precast Protein Gels and blotted onto PVDF membranes. The membranes were then blocked for 1hr at RT using 5% blocking solution before incubating with primary antibodies (mouse anti-tubulin at 1:500 DSHB, rabbit anti-CG6289 at 1:1000 (Ravi Ram et al. 2005) or rabbit anti-dBACE at 1:2000 (Bolkan et al. 2012) at 4°C overnight, followed by HRP-conjugated secondary antibodies (Jackson ImmunoResearch #115-035-062) used at 1:10,000. The chemiluminescence signals were generated using Amersham ECL Prime Western Blotting Detection Reagent and detected on a Fujifilm LAS-3000 Imager.

Supplemental References

Bolkan BJ, Triphan T, Kretzschmar D. 2012. beta-secretase cleavage of the fly amyloid precursor protein is required for glial survival. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **32**: 16181-16192.

Brown JB, Boley N, Eisman R, May G, Stoiber M, Duff M, Booth B, Park S, Suzuki A, Wan K et al. 2014. Diversity and dynamics of the Drosophila transcriptome. *Nature* **512**: 393-399.

Capra JA, Williams AG, Pollard KS. 2012. ProteinHistorian: tools for the comparative analysis of eukaryote protein origin. *PLoS computational biology* **8**: e1002567.

Chen S, Spletter M, Ni X, White KP, Luo L, Long M. 2012. Frequent recent origination of brain genes shaped the evolution of foraging behavior in Drosophila. *Cell reports* **1**: 118-132.

Chen S, Zhang YE, Long M. 2010. New genes in Drosophila quickly become essential. *Science* **330**: 1682-1685.

Clark AG Eisen MB Smith DR Bergman CM Oliver B Markow TA Kaufman TC Kellis M Gelbart W Iyer VN et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**: 203-218.

Hahn MW, Han MV, Han SG. 2007. Gene family evolution across 12 Drosophila genomes. *PLoS genetics* **3**: e197.

Hu Y, Flockhart I, Vinayagam A, Bergwitz C, Berger B, Perrimon N, Mohr SE. 2011. An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC bioinformatics* **12**: 357.

Kondo S, Ueda R. 2013. Highly improved gene targeting by germline-specific Cas9 expression in Drosophila. *Genetics* **195**: 715-721.

Lack JB, Cardeno CM, Crepeau MW, Taylor W, Corbett-Detig RB, Stevens KA, Langley CH, Pool JE. 2015. The Drosophila genome nexus: a population genomic resource of 623 Drosophila melanogaster genomes, including 197 from a single ancestral range population. *Genetics* **199**: 1229-1241.

Liao Y, Smyth GK, Shi W. 2013. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic acids research* **41**: e108.

Mackay TF, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM et al. 2012. The Drosophila melanogaster Genetic Reference Panel. *Nature* **482**: 173-178.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**: 652-654.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular biology and evolution* **3**: 418-426.

Ravi Ram K, Ji S, Wolfner MF. 2005. Fates and targets of male accessory gland proteins in mated female Drosophila melanogaster. *Insect Biochem Mol Biol* **35**: 1059-1071.

Stoletzki N, Eyre-Walker A. 2011. Estimation of the neutrality index. *Molecular biology and evolution* **28**: 63-70.

Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E et al. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**: 650-659.

Zhao L, Saelao P, Jones CD, Begun DJ. 2014. Origin and spread of de novo genes in Drosophila melanogaster populations. *Science* **343**: 769-772.