On the Regulatory Evolution of New Genes Throughout Their Life History

Jia-Yu Zhang[1] and Qi Zhou*,[1,2]

[1]MOE Key Laboratory of Biosystems Homeostasis & Protection, Life Sciences Institute, Zhejiang University, Hangzhou, China

[2]Department of Molecular Evolution and Development, University of Vienna, Vienna, Austria

*Corresponding author: E-mail: zhouqi1982@zju.edu.cn.

Associate editor: Katja Nowick

Abstract

Every gene has a birthplace and an age, that is, a cis-regulatory environment and an evolution lifespan since its origination, yet how the two shape the evolution trajectories of genes remains unclear. Here, we address this basic question by comparing phylogenetically dated new genes in the context of both their ages and origination mechanisms. In both *Drosophila* and vertebrates, we confirm a clear "out of the testis" transition from the specifically expressed young genes to the broadly expressed old housekeeping genes, observed only in testis but not in other tissues. Many new genes have gained important functions during embryogenesis, manifested as either specific activation at maternal–zygotic transition, or different spatiotemporal expressions from their parental genes. These expression patterns are largely driven by an age-dependent evolution of cis-regulatory environment. We discover that retrogenes are more frequently born in a pre-existing repressive regulatory domain, and are more diverged in their enhancer repertoire than the DNA-based gene duplications. During evolution, new gene duplications gradually gain active histone modifications and undergo more enhancer turnovers when becoming older, but exhibit complex trends of gaining or losing repressive histone modifications in *Drosophila* or vertebrates, respectively. Interestingly, vertebrate new genes exhibit an "into the testis" epigenetic transition that older genes become more likely to be co-occupied by both active and repressive ("bivalent") histone modifications specifically in testis. Our results uncover the regulatory mechanisms underpinning the stepwise acquisition of novel and complex functions by new genes, and illuminate the general evolution trajectory of genes throughout their life history.

Key words: **new gene, histone modification, enhancer, out of the testis.**

Introduction

The great disparity of gene numbers between species indicates that gain and loss of genes is a fundamental evolutionary process. Since the report of the first new gene *jingwei* over two decades ago (Long and Langley 1993), numerous genome-wide and case studies have now demonstrated that origination of functional new genes is one of the main drivers underlying phenotypic innovation (Kaessmann 2010; Chen et al. 2013). The emergence of *jingwei* represents a paradigm rather than an anecdote of new gene evolution: both DNA- and RNA-mediated (retroposition) gene duplications from different parental genes have contributed to the formation of the new chimeric structure of *jingwei*, which acquired a new expression pattern specifically in testis compared with its *Adh* ancestor. Later inspection of multiple *Drosophila* genomes showed that gene duplication accounts for about 80% of species or lineage-specific new genes (Zhou et al. 2008). This conforms to Ohno's hypothesis that gene duplication is the primary source of new genes (Ohno 1970). In addition, at least 30% of *Drosophila* new genes (Zhou et al. 2008), or 50% of *Caenorhabditis elegans* new genes (Katju and Lynch 2006) have been found to incorporate various genomic resources (e.g., partial coding sequences of another gene, or transposable elements) to form a chimeric structure by exon shuffling, potentially facilitating functional innovation. An unexpected finding from genome scans of a broad range of species including yeast (Carvunis et al. 2012), *Drosophila* (Zhao et al. 2014), and human (Knowles and McLysaght 2009; Wu et al. 2011; Ruiz-Orera et al. 2015) is that de novo origination from noncoding sequences has a substantial contribution to new gene origination. Many nascent de novo genes, as well as species-specific gene duplicates are more likely to be still segregating within populations and subjected to random loss than those "older" new genes that have become fixed in populations at an earlier time point and are shared by multiple species (Zhou et al. 2008; Palmieri et al. 2014; Zhao et al. 2014). Similar to *jingwei*, many de novo genes and new gene duplicates have been found to become predominantly or exclusively expressed in testis (Betran and Long 2003; Carelli et al. 2016; Guschanski et al. 2017; Luis Villanueva-Canas et al. 2017). Functional disruption showed some *Drosophila* new genes acquired novel function that is either involved in spermatogenesis (Kondo et al. 2017) (e.g., *nsr* [Ding et al. 2010] gene that originated about 6 million years ago), or associated with male mating behavior (e.g., *sphinx* [Dai et al. 2008] that originated 3 million years ago).

A striking case is *Umbrea*, an older (15 million years ago) new gene that gradually evolved essential centromeric function in comparison with its heterochromatin-binding parental gene *HP1B* (Ross et al. 2013). These case studies suggested that new genes frequently undergo neofunctionalization, and their population dynamics and novel functions are characterized by their age.

Understanding functional evolution of new genes in the context of their ages is critical for illuminating genes' dynamic life history in general (Betran 2015; Carelli et al. 2016). Although it is difficult to reconstruct gene's evolution trajectory, valuable insights have been gained by comparing genes of different ages (Carelli et al. 2016; Guschanski et al. 2017). This is on one hand facilitated by the ongoing effort of functional disruption of identified *Drosophila* new genes using RNAi or CRISPR/Cas-9 technique (Chen et al. 2010; Kondo et al. 2017), and also by the recent development of next-generation sequencing. Transcriptome comparison of multiple *Drosophila* and mammalian adult tissues suggested that younger new gene duplicates, particularly retrogenes are more prone to have a testis-specific expression; whereas the older ones are more often ubiquitously expressed or specifically expressed in other somatic tissues (Assis and Bachtrog 2013; Carelli et al. 2016; Guschanski et al. 2017). This has led to the "out of the testis" hypothesis on the emergence of new genes: it postulates that the permissive chromatin environment of testis provides a haven for nascent genes from natural selection against deleterious effects of the redundant gene dosage when they were born (Dai et al. 2006; Vinckenbosch et al. 2006; Kaessmann et al. 2009; Kaessmann 2010). Other contributing factors to the large number of testis-biased new genes include meiotic sex chromosome inactivation or sexual antagonistic selection, which select against testis-biased genes on the X chromosome. They produce an "out of the X" pattern in both *Drosophila* (Betran et al. 2002; Betran and Long 2003; Vibranovski et al. 2009) and human (Emerson et al. 2004), that X-linked parental genes tend to produce excessive testis-biased new genes located on the autosomes. Such young genes maybe later driven to fixation by intensive sexual selection in testis, or acquisition of novel function beyond the testis by forming new gene structures and/or recruiting new regulatory elements. Such a dynamic life history of new genes is also reflected by the gradually increased connectivity of gene interactions from young genes usually located at the periphery of the network to old genes as an essential hub (Zhang et al. 2015). Overall, most contemporary genome-wide characterizations of new genes take advantage of transcriptome data, which is the output of complex coordinated regulation involving cis-regulatory elements (CREs: promoter, enhancer, etc.) and local epigenomic configuration.

However, little is known about the regulatory mechanisms underlying how a new gene evolves a divergent expression pattern from its ancestor at the genome-wide level. This is because components and principles of transcriptional regulation have not been systematically dissected only until very recently through many consortium projects (e.g., ENCODE and modENCODE) (Roy et al. 2010; The ENCODE Consortium 2012; GTEx Consortium et al. 2017). This question is key to understanding how a new gene can avoid becoming a pseudogene as presumed by the classic model (Ohno 1970). A new gene can either evolve new expression, that is, undergo neofunctionalization by recruiting novel cis-regulatory elements, and/or translocating to a new epigenomic environment as more often occurred with retrogenes (Chen et al. 2013; Arthur et al. 2014; Carelli et al. 2016). Alternatively, a new gene can partition the ancestral expression pattern with its parental gene through complementary degenerative mutations in the regulatory region (subfunctionalization) (Lynch and Force 2000). It is now well established that the epigenomic landscape is shaped by dynamic DNA methylation and various histone modifications. Active and repressive chromatin marks, such as histone H3 lysine 4 trimethylation (H3K4me3), H3K36me3, and H3K27me3, H3K9me3, etc. synergistically or antagonistically bind together to genic or CRE regions to impact the transcription level. In this work, we seek to address the regulatory mechanisms of new gene evolution by analyzing a total of 83 transcriptomic and 281 epigenomic data sets across a broad range of tissues and developmental stages of *Drosophila melanogaster* and human (supplementary fig. S1, Supplementary Material online). We used an updated data set of new genes of *Drosophila* and vertebrates, and paid special attention to bulk and single-cell RNA-seq (scRNA-seq) data during early development, when little is known about new genes' functional roles. By cross investigation of massive epigenomic and CRE profiles of new genes in the context of their origination mechanisms and ages, we unveiled a nonbiased landscape of dynamic regulatory changes throughout new genes' life history.

Results

New Genes Are Becoming Out of the Testis by Age

We acquired a high-confidence data set of new genes following the published pipeline (Chen et al. 2010; Zhang et al. 2010) with the updated genomes and annotations of 12 *Drosophila* species (metazoa release 25) and 14 vertebrate species (Ensembl v73) (fig. 1). Changes to the numbers of new genes comparing with previous studies (Zhang et al. 2012) were mainly caused by the reannotation of many protein-coding genes as noncoding RNAs or pseudogenes, or their direct removal from the updated version of annotation. In brief, we used whole-genome syntenic alignments to inspect the phylogenetic distribution of orthologous genes. We identified species or lineage-specific new genes and inferred their age by parsimony based on their presence/absence of orthologs in multiple outgroups. Using multiple outgroups reduces the chance of misidentification of new genes due to sequencing gaps or independent loss of genes from one certain outgroup, which however does not apply to old age groups (e.g., age group A in fig. 1) in this study. We also inferred the origination mechanisms of new genes as DNA-based gene duplication (referred as "gene duplication" hereafter), retroposition and de novo origination based on each category of a gene's specific feature (e.g., absence of introns in retrogenes, absence of syntenic orthologous genes for de novo genes, see Materials

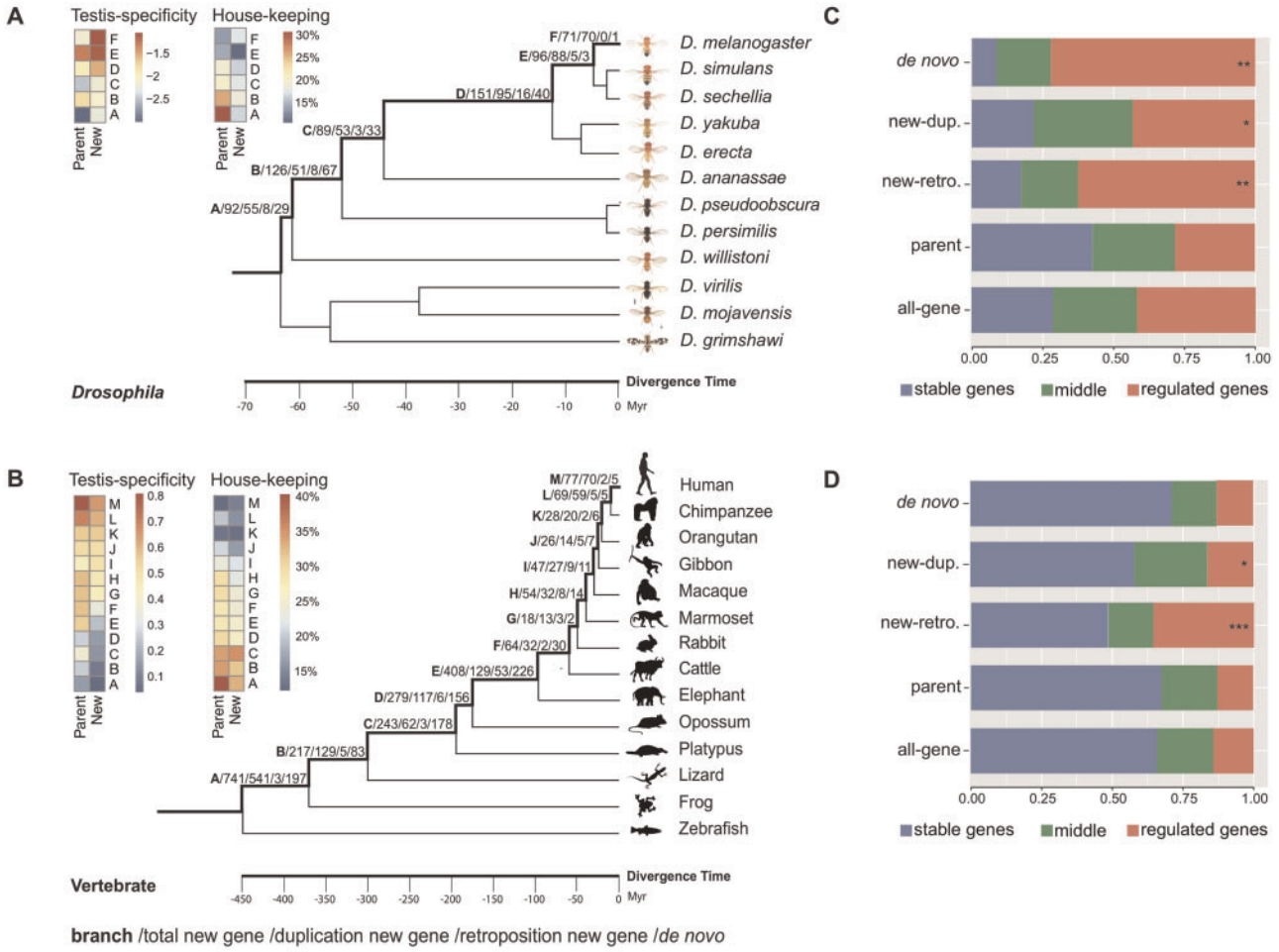

Fig. 1. *Drosophila* and vertebrate new genes and their expression patterns. (*A* and *B*) Different numbers of new genes divided by their age and origination mechanisms. Each age group is designated by an ordered letter from A to F/M shown on each phylogenetic branch. And each group of genes is further divided as total number of new genes, number of DNA-based duplicated genes, retrogenes, and de novo genes, separated by slash. We also showed along the age groups the change of testis-specificity, measured by the log2-based ratio of expression level of testis versus whole male body of *Drosophila* or the sum of all human male adult tissues; and housekeeping gene index, measured by the percentage of tissues/stages which show robust expression. Median values of testis-specificity and housekeeping gene index of each age group are shown by heatmap. The *Drosophila* pictures are from Nicolas Gompel. (*C* and *D*) Regulated genes of human and *Drosophila*, divided by different origination mechanisms. We defined the regulated genes based on the coefficient of expression (CV) level variation across different tissues. CVs of *Drosophila* were derived from Perez-Lluch et al. (2015) and those of human were calculated from GTEx data set (https://www.gtexportal.org/) (Consortium et al. 2017). We also compared the proportion of regulated genes between each category of new genes versus their parental genes, or de novo genes versus the whole genome-wide average, and show the significance levels of Fisher's exact test by asterisks: $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

and Methods section). In total, we annotated 585 *Drosophila* and 3,056 vertebrate new genes. A major technical challenge for any new gene analyses is that except for de novo genes, a large part of the sequences is shared between the new and parental gene pair, which confounds the quantitative comparison of their levels of gene expression or histone modification. To overcome this, we harnessed the sequence divergence sites between the pair, and only counted the reads that can be unambiguously assigned to either new gene or parental gene based on their spanned diagnostic SNPs, for any pairwise comparisons between gene duplications or retrogenes versus their parental genes throughout this work. Normalized gene expression or histone modification level is then measured by the ratios between the informative read counts of RNA-seq/ChIP-seq versus DNA-seq for each gene to correct for any sequencing or mapping bias. Three lines of evidence convinced us that this subset of sequences per gene is able to give us robust and specific estimation of the level of transcription and histone modification: first, the SNP density, as a reflection of divergence level between the gene pair expectedly increases by new genes' age (supplementary fig. S2, Supplementary Material online). We found a substantial number of informative sites (median value 10–30 per 100 bp) even between the youngest group of new genes and their parental genes, given a read length of over 100 bp in most of the analyzed data sets. Second, we observed a significant ($P < 0.05$, Pearson's correlation test) positive/negative correlation between the normalized expression level versus respective active/repressive histone modification level (supplementary fig. S3, Supplementary Material online) across most of the

inspected stages and tissues, based only on informative sites. Third, the proportions of genes showing pronounced histone modifications ("bound" genes), defined by informative sites, increased gradually with the developmental progress of *D. melanogaster* (supplementary fig. S4, Supplementary Material online). This pattern is consistent with that reported by modENCODE consortium which used the entire gene or promoter regions (Kharchenko et al. 2011).

To reveal the global transcriptome change of new genes by age, we focused on two of their features: gene expression specificity, calculated as the ratio of gene expression level in the focal tissue versus that of the whole *D. melanogaster* body or the sum of expression levels of all human adult tissues; and expression breadth, measured by the percentage of tissues/stages with active expression of the focal gene detected among all examined samples, as an indication for housekeeping genes. We found strong evidence supporting the out of the testis hypothesis that new genes decrease their testis-specificity by age, and vertebrate but not *Drosophila* new genes increase their expression breadth, that is, are more likely to be housekeeping genes (fig. 1A and *B*). The different pattern of expression breadth in vertebrate and *Drosophila* species can result from the different numbers of identified new genes and sampled tissues between the two clades, or suggest there is stronger selection on the gene–gene interaction network structure against integrating a new housekeeping gene in *Drosophila*. A similar pattern of expression specificity has not been found among other tissues of *D. melanogaster* and human (supplementary fig. S5, Supplementary Material online), suggesting the unique gene regulation program and evolution forces acting in the testis make it prone for the birth of new genes. If we look into new genes divided by their origination mechanisms, we are still able to find the out of the testis pattern in each category of human new genes, but only in new gene duplicates of *Drosophila*, probably due to its much lower number of retrogenes and de novo genes (supplementary fig. S6, Supplementary Material online). These patterns still hold if we only focus on a shorter evolution time-scale, for example, within the melanogaster group species or the mammalian species (up to branch C in both clades, fig. 1, supplementary fig. S6, Supplementary Material online), where confounding factors like independent gain/loss of genes or misidentifying ancient duplicates as de novo genes are much less likely. Parental genes of human and *Drosophila* new gene duplications also show a similar out of the testis trend, suggesting that the pattern is largely contributed by the duplication of testis CREs along with the gene duplications. As the expression breadth per se does not reflect the degree of expression level variations between tissues/stages, we further calculated the coefficients of variation (CV) (Perez-Lluch et al. 2015) across all the analyzed samples. A gene that shows a highly variable spatial/temporal gene expression pattern, that is, a high CV value, is defined as a "regulated" gene based on the distribution of genome-wide CV values (supplementary fig. S7, Supplementary Material online), otherwise as a "stable" gene. We consistently found new genes, particularly retrogenes and de novo genes (except for human de novo genes), have a significantly larger proportion of regulated genes than the parental genes (fig. 1C and *D*, $P < 0.05$, Fisher's exact test) or the genome-wide level. This can be explained by retrogenes and de novo genes being more likely to recruit novel regulatory elements (see below) than are new genes generated by gene duplication.

Dynamic Expression of New Genes During Early Development

Much effort has been invested in examining new genes' expression in adult tissues, yet little is known about their functional contribution during the embryonic developmental process. The only work that has been done, at a much broader evolutionary time scale than this work, found that in zebrafish and *Drosophila*, younger genes are enriched in early and late embryonic stages, whereas the old genes are enriched at mid-embryonic ("phylotypic") stage (Domazet-Loso and Tautz 2010). This supports a developmental "hourglass" model, where strong constraints on developmental regulation act at the phylotypic stage (Kalinka et al. 2010). Here, we scrutinized both the expression level and spatiotemporal information of new genes in early embryos, by analyzing their scRNA-seq and RNA fluorescence in situ hybridization (RNA-FISH) data. We found in both human and *D. melanogaster* embryos, new genes except for *Drosophila* de novo genes show a similarly robust expression level as parental genes throughout embryonic stages, suggesting some of them have evolved important developmental functions (fig. 2A, "expression level" panel). Human retrogenes and de novo genes are expressed at a significantly higher level ($P < 0.05$, *t*-test and Wilcoxon test) respectively than their parental copies and genome-wide average since the 4 weeks until the 19 weeks of embryonic development, in both germline and somatic cells, in both sexes (supplementary fig. S8, Supplementary Material online). We found a burst of expressed retrogenes (supplementary table S4, Supplementary Material online), but not in new gene duplicates, from the human 4-cell embryonic stage to the 8-cell stage, in contrast to a decrease in numbers and expression levels of active de novo genes from the 2-cell stage until the 8-cell stage. This is particularly interesting as the major maternal–zygotic transition (MZT) occurs during this time window (Braude et al. 1988). It implies that many retrogenes are involved in MZT, and although there is a significant ($P < 0.01$, Fisher's test) excess of de novo genes originally deposited as maternal transcripts in both human and *Drosophila* (supplementary tables S4 and S5, Supplementary Material online), they become degraded during later embryonic development. Indeed, we found that the zygotically activated retrogenes were enriched (Bonferroni corrected $P < 0.05$) for functional categories of "RNA recognition" and "mRNA surveillance pathway" (supplementary fig. S9, Supplementary Material online), which might participate in the post-transcriptional control as shown by many genes at *Drosophila* MZT (Sysoev et al. 2016).

Besides the expression level, we also compared the spatiotemporal embryonic expression between new versus parental genes of *Drosophila*, whose different patterns directly indicate their divergent function. Among the 144 parental genes and

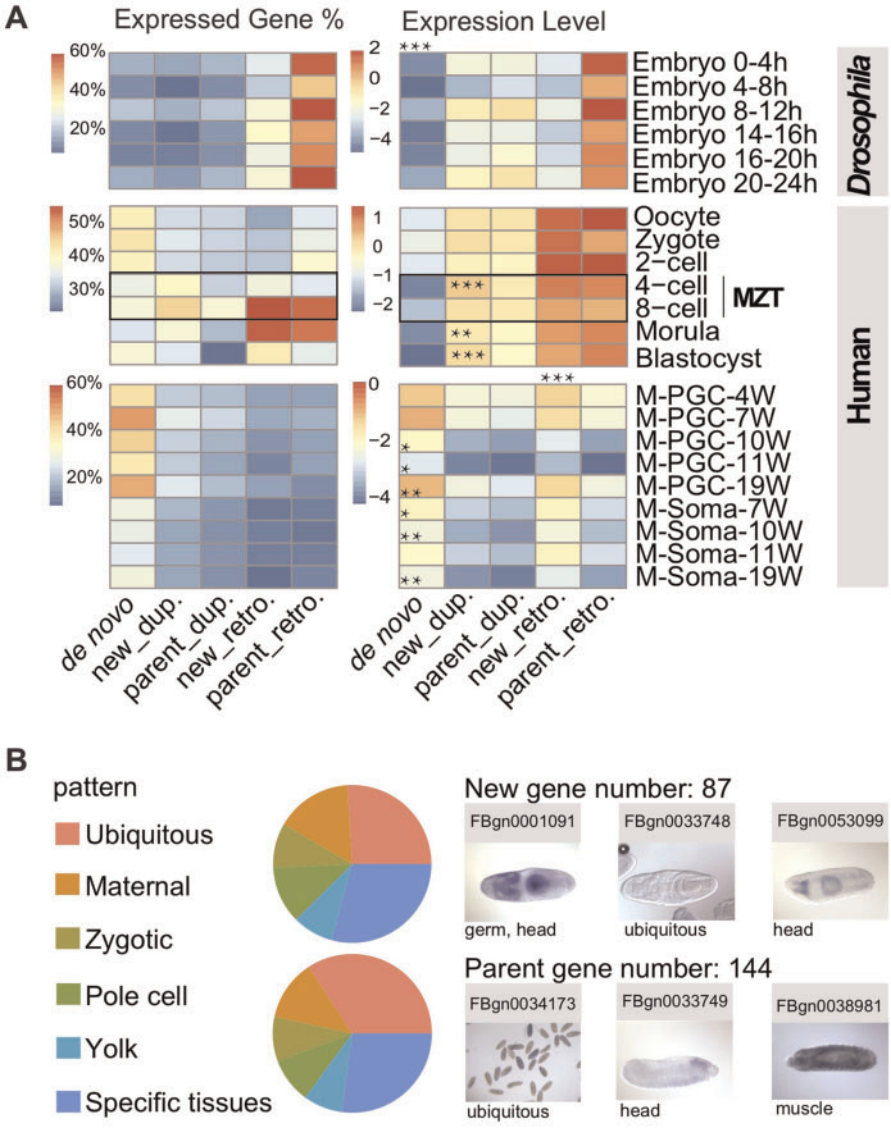

Fig. 2. Dynamic expression of new genes in early embryos. (*A*) We showed the percentage of expressed genes (left columns), and the median values of normalized gene expression levels (right columns) across different embryonic stages of *D. melanogaster* (with bulk RNA-seq) and human (with single RNA-seq), ordered by their developmental course. We also compared the new genes versus parental genes, and de novo genes versus the genome-wide average for their expression levels and show the levels of significance of each pairwise *t*-test (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$). If the entire column is significant, we show the asterisks above the column (e.g., human retrogenes). M-PGC-4W, male primordial germ cell of the 4 weeks' stage; M-Soma-7W, male embryonic somatic cell of the 7 weeks' age. (*B*) Divergent subcellular localizations between new and parental genes in *D. melanogaster* embryos. Pie charts showed the proportion of different expressed locations of new genes and parent genes, with three examples of divergent expression patterns between new and parental genes. Examples were selected to represent different cases as parental gene being ubiquitously expressed and new gene specifically expressed, or the opposite, or both being specifically expressed at different locations. Expression patterns are extracted from the controlled vocabularies of expression annotation from BDGP or Fly-FISH databases.

87 new genes with controlled vocabularies of RNA-FISH annotation available from either BDGP (Tomancak et al. 2002) or Fly-FISH (Lécuyer et al. 2007) database, we found that there are significantly more parental genes show "ubiquitous expression" (fig. 2*B*, 34.25% vs. 25.95% of new genes, $P = 0.0239$, Fisher's exact test); whereas more new genes, although not significantly are expressed at specific tissues (e.g., "pole cells" or "nervous system"; 29.20% vs. 27.09% of parental genes, $P = 0.5412$ Fisher's exact test) in embryos. All the 11 de novo genes with annotated patterns are "maternal" transcripts Lott et al. 2011 that become degraded during later development. This is consistent with what was found for human de novo genes (fig. 2*A*). Among the 40 gene pairs with RNA-FISH data available for both parental and new genes, none of the pairs show exactly the same subcellular localization and expressed time windows between the two, providing strong evidence for functional divergence between the parental and new genes in early embryos. Specifically, 9 gene pairs (e.g., FBgn0034173-FBgn0001091, fig. 2*B*, supplementary fig. S10, Supplementary Material online) show

ubiquitous expression throughout all the investigated stages for the parental gene, but a specific expression pattern for the new gene. Four pairs show the opposite pattern. These cases are strong candidates for neofunctionalization of new genes (supplementary fig. S11, Supplementary Material online). There are 27 gene pairs with parental and new genes expressed in different tissues (e.g., FBgn0038901-FBgn0053099, fig. 2B, supplementary fig. S12, Supplementary Material online), which might be candidates for subfunctionalization, depending on the parental gene's expression in the outgroup species.

Age-Dependent Evolution of Chromatin State

As many more genes become regulated by histone modifications beyond embryonic stages (supplementary fig. S4, Supplementary Material online), we further examined the expression patterns of new genes among all the tissues, divided by different age groups. We found human but not *Drosophila* genes show an age-dependent change of expression: not only new genes but also parental genes become more likely to be expressed in older age groups across all the examined human tissues (supplementary figs. S13C and S14C, Supplementary Material online). The pattern of de novo genes is not as clear, which could be influenced by their much smaller numbers at each age group. Correspondingly, there are gradually less putative pseudogenes, defined as those without robust gene expression throughout all the examined tissues or stages, among older age groups of new genes (supplementary table S1, Supplementary Material online). This suggests that the tendency to become more active and functionally important is a general age-related feature of genes, not just species or lineage-specific new genes.

To test whether the gene expression pattern is driven by age-dependent epigenomic changes, we inspected 14 *D. melanogaster* and 7 human histone modification markers and first compared their binding patterns between new and parental genes. We focused on four active markers H3K4me3, H3K4me1, H3K27ac, H3K36me3, which are strongly associated with active transcription, and promoter, enhancer, or exonic regions; and two repressive markers H3K27me3, H3K9me3 which are associated with gene silencing (supplementary fig. S3, Supplementary Material online). They were chosen because they are among the best-known for their functional association and most broadly studied across almost all tissues and developmental stages in both human and *D. melanogaster* (supplementary fig. S1, Supplementary Material online) (Andersson et al. 2014). We found that in both species, and throughout most stages/tissues, new genes exhibit significantly (Wilcoxon test, $P < 0.05$) lower levels of all active histone modifications and RNA Polymerase II binding at promoter (for H3K4me1/3, H3K27ac) or entire gene regions (for H3K36me3), but a higher level of facultative heterochromatin modification H3K27me3 than their parental genes (fig. 3, supplementary figs. S15 and S16, Supplementary Material online). No significant difference has been observed between new and parental genes for the constitutive heterochromatin modification H3K9me3, which is usually associated with transposable elements. A key distinction between the two repressive markers is that H3K27me3 is strongly associated with spatiotemporal regulation of gene expression, thus more dynamic in its silencing function. In particular, H3K27me3 may form a "bivalent" domain together with H3K4me3 to maintain the influenced gene in a poised chromatin state for later activation of lineage-specific expression. These patterns overall account for a generally lower percentage of expressed genes (supplementary fig. S13, Supplementary Material online) or housekeeping genes, but a higher percentage of regulated genes among new genes (fig. 1), comparing with the parental genes. Specifically, gene duplications and retrogenes have diverged from their parental genes for their active histone modifications to a similar degree, but retrogenes show a more dramatic change of H3K27me3 modifications than gene duplications. We found that there is a larger increase in the percentage of H3K27me3-bound retrogenes compared with their parental genes, than those found between new gene duplications versus their parental genes (supplementary table S2, Supplementary Material online). Retrogenes also more frequently show a significantly ($P < 0.05$, Wilcoxon test) higher level of H3K27me3 modification than their parental genes, compared with gene duplications (supplementary fig. S15, Supplementary Material online). This is probably because that survived retrogenes are often translocated into a preexisting H3K27me3 domain as indicated by their surrounding genes: we found that up and downstream genes of new retrogenes show a significantly ($P < 0.05$, Wilcoxon test) higher level of H3K27me3 modification than those surrounding the parental genes, but the pattern of DNA-based duplicated genes is less pronounced or not as consistent as retrogenes across different tissues or stages (fig. 3, supplementary fig. S17, Supplementary Material online). This does not indicate retrogenes are more likely to be silenced pseudogenes, as we found that retrogenes exhibit a significantly higher proportion of bivalent genes, defined as those bound by both repressive H3K27me3 and active H3K4me3 markers, than their parental genes during larvae stages of *D. melanogaster*, and in some specific tissues of human (e.g., kidney, supplementary figs. S18 and S19, Supplementary Material online).

We further uncovered that both new and parental genes exhibit an age-dependent change of chromatin states, with different trajectories between somatic and germline tissues, and also between human and *D. melanogaster*, the latter of which accounts for their presence or absence of age-dependent gene expression pattern (supplementary fig. S13, Supplementary Material online). For *D. melanogaster*, there is a gradual increase in the percentage of new genes bound by both active and repressive histone marks by age, thus also a higher percentage of bivalent genes in older age groups. In contrast, their parental genes show an opposite trend (fig. 4). These patterns have been observed from the late stage embryos until the second instar larvae, and for gene duplications but not for retrogenes or de novo genes (supplementary figs. S20 and 21A, Supplementary Material online). For human, parental and new genes show a similar pattern to each other: there are generally a higher percentage of genes bound by active marks, but fewer genes bound by repressive marks by

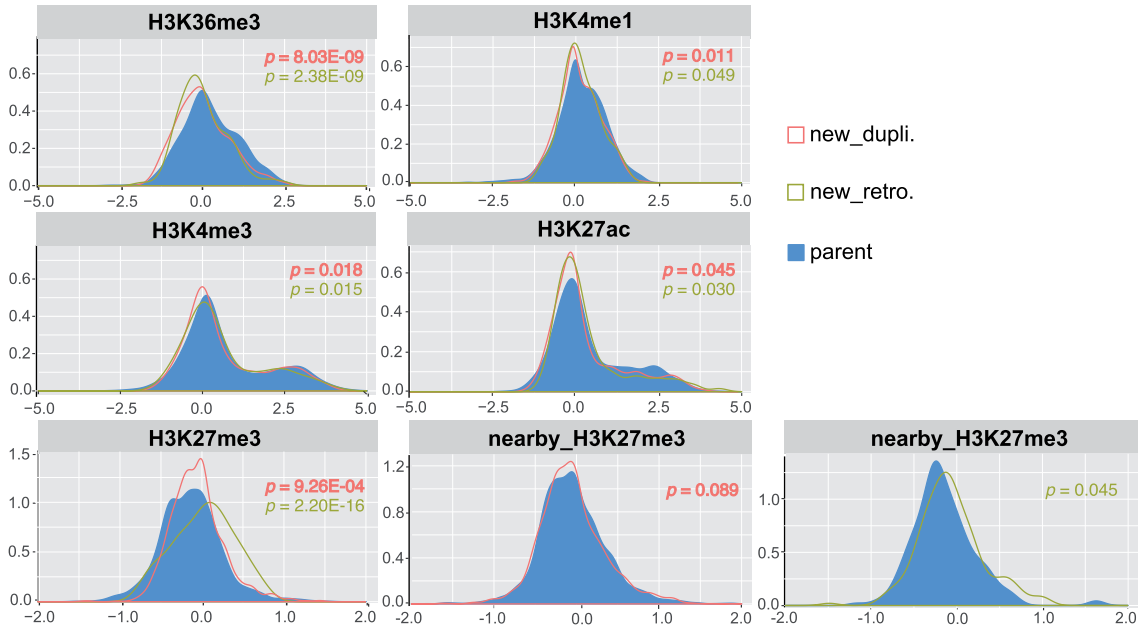

Fig. 3. New genes and parental genes have diverged for their chromatin state. We showed the distributions of normalized histone modification levels measured by log2 ChIP versus input ratio, spanning the gene body, or promoter regions of new gene (different colors of lines) versus parental gene (blue area), and their surrounding genes ("nearby" profiles). We also showed the *P* values of *t*-tests comparing the new genes versus parental genes. For active markers (H3K36me3, H3K4me1, H3K4me3, H3K27ac), both gene duplications and retrogenes show a significantly lower binding level than the parental genes, to a similar degree. For the H3K27me3 repressive marker, retrogenes show a more dramatic increase in binding levels relative to the parental genes, compared with gene duplications.

age, which together result in more active genes in older age groups (supplementary fig. S13, Supplementary Material online). A similar pattern of active or repressive marks has also been observed for human new and parental genes when we compared the levels of histone modifications among different age groups. The patterns are generally consistent among the investigated somatic tissues or stages, but more pronounced in adult tissues (supplementary figs. S21 and S22, Supplementary Material online), and again only observed in gene duplications but not in retrogenes and de novo genes (supplementary fig. S20, Supplementary Material online). This suggests that new gene duplicates have a different evolution trajectory of regulatory changes compared with de novo genes or retrogenes, as the latter two are not likely to inherit the promoters or epigenetic profiles of the parental genes.

Bivalent genes, which are bound by both active and repressive histone marks, show a more complex pattern along the age groups, and between somatic and germline tissues. There is a burst of bivalent new genes at the ancestor of apes (age group J in figs.1 and 4) produced by both gene duplications or retroposition (supplementary fig. S20, Supplementary Material online), after which its proportion reduces by age in somatic tissues. Whereas in germline, there is an intriguing "into the testis" pattern where the proportion of bivalent genes (Lesch et al. 2016) increases by the age of genes, regardless of their origination mechanisms. The opposite trajectory of bivalent genes between somatic and germline tissues makes sense in the light of segregation of novel cell-differentiation related functions of new genes in either type of tissues. This into the testis pattern found at the epigenomic level is also consistent with the out of the testis pattern shown at the transcriptomic level: as old bivalent genes in germ cells usually are expressed in somatic tissues beyond testis and important developmental regulators of embryogenesis (Lesch et al. 2016). These results collectively indicated that it takes young genes some time to gradually evolve active histone modifications, whereas in human but not in *D. melanogaster*, repressive histone modifications have further contributed to the silencing of young genes. They also suggested that strong selection against redundant gene dosage, especially for robustly expressed genes, when a nascent gene copy is born: in human, parental genes of younger new genes tend to be lowly expressed genes with few active marks and many repressive marks; and in *D. melanogaster*, whereas actively expressed parental genes do give birth to new genes of young age, the new genes generally tend to lack active histone modifications to drive robust expression (supplementary figs. S13*A* and *C* and 21*A*, Supplementary Material online).

New Genes and Parental Genes Have Become Divergent for Their CREs

Gene expression is coordinately regulated by epigenomic configuration and CREs. The differential bindings of histone modifications account for the expression level divergence (fig. 3) between parental and new genes. Whereas their spatiotemporal expression differences (fig. 2B) are more likely to be caused by a different composition of CREs. To test this, we further compared the enhancer repertoire between new and parental genes, which is annotated by STARR-seq

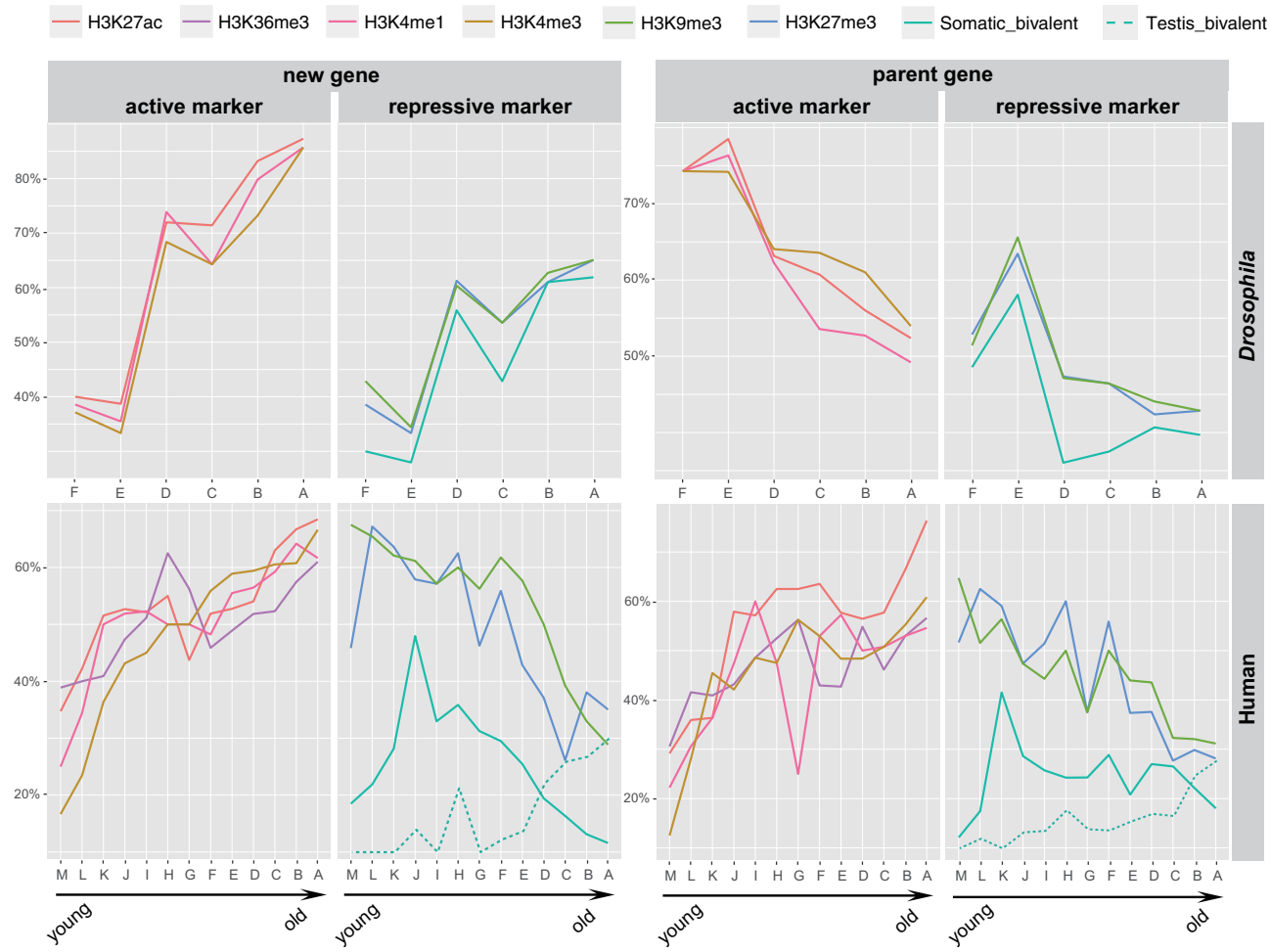

Fig. 4. Age-dependent change of histone modifications. We show here the percentage of bound genes by respective histone modifications across the ordered age groups (x-axis, from left to right, each letter represents one age group shown in fig. 1). Each color of line represents a different histone modification, with warm colors for active markers, and cold colors for repressive markers. The data of human testis bivalent genes are derived from Lesch et al. (2016) shown as dotted line. For the solid lines, we used the ChIP-seq data of second instar larvae for *D. melanogaster*, and adrenal gland adult male tissue for human. The patterns of other *D. melanogaster* or human tissues are consistent and shown in supplementary figures S21 and S22, Supplementary Material online.

(self-transcribing active regulatory region sequencing) in five *Drosophila* species (Arnold et al. 2014) or by CAGE (cap analysis of gene expression) in human (Andersson et al. 2014). As expected, these enhancers show a characteristic enrichment of active (H3K4me1, H3K27ac) histone modifications and depletion of repressive histone modification (H3K27me3) (supplementary fig. S23, Supplementary Material online). *Drosophila* retrogenes but not gene duplications show a significant ($P < 0.05$, Wilcoxon test) decrease in active histone modification (H3K27ac), and an increase in repressive H3K27me3 modification compared with their parental genes (supplementary fig. S24, Supplementary Material online), which together with other gene body histone modifications have contributed to a lower expression level of retrogenes. New genes exhibit gains, losses, or sequence mutations of their enhancers compared with those of their parental genes. And a higher percentage of new genes in older age groups has undergone such turnovers, but rarely show complete retentions of parental genes' enhancers ("enhancer duplication," fig. 5A and B) in both *Drosophila* and human. This indicates a much more diverged cis-regulatory circuit between parental and new genes over evolution. In particular, when searching for the orthologous sequences of specific enhancers ("enhancer gain") that were gained by *Drosophila* new genes in their outgroup species, we found that they are also predominantly enhancers, suggesting new genes have frequently recruited pre-existing enhancers as their new CREs (fig. 5C). In fact, 37 out of 173 analyzed *Drosophila* new gene specific enhancers that do not have an ortholog in outgroup species ("de novo enhancers," fig. 5C, supplementary table S6, Supplementary Material online) all belong to de novo genes. This raises the interesting question about the role of de novo enhancers during the emergence of de novo genes.

Besides the numbers of enhancers, new genes and parental genes have also diverged for their types of enhancers. It has been recently shown that *D. melanogaster* enhancers can be divided into two classes, according to their specificity to core promoters of either housekeeping genes or developmentally

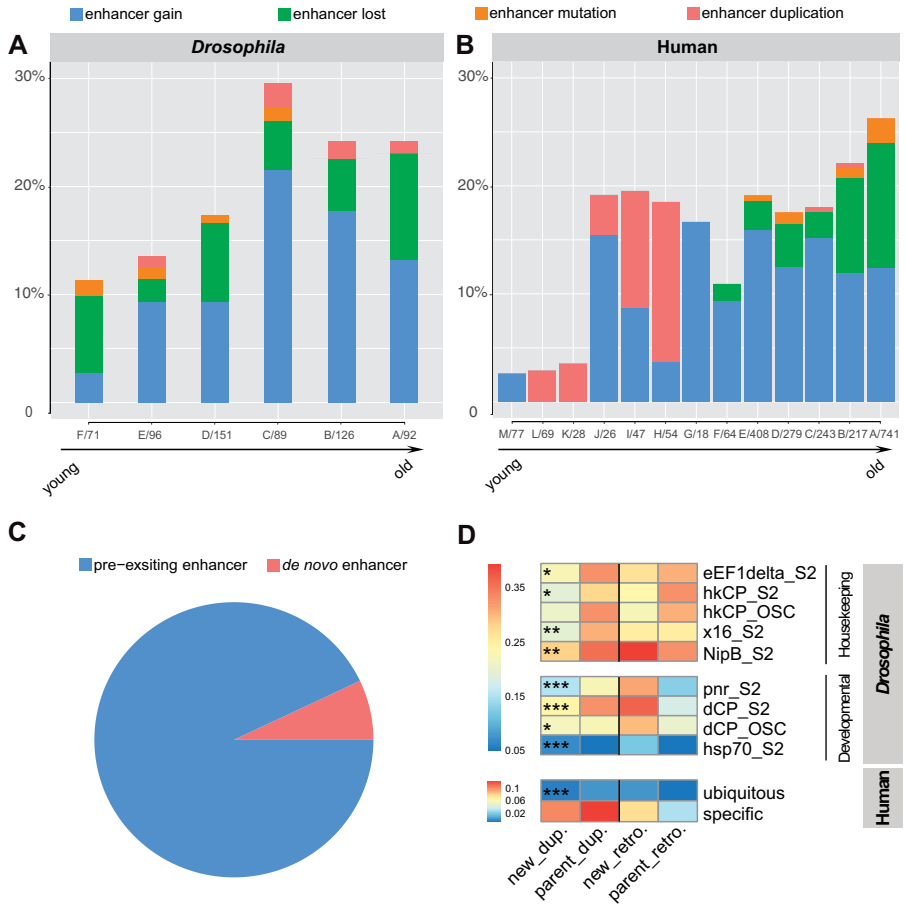

Fig. 5. New genes and parental genes have diverged for their CREs. (*A* and *B*) We compared the enhancers between new and parental genes in *Drosophila* and human, based on the latest STARR-seq or CAGE annotations. The bar plot here shows the proportion of new genes at each age group that have undergone turnovers of enhancers. (*C*) The pie chart shows the source of enhancer gains in *Drosophila* new genes, either from a pre-existing enhancer, as defined by the presence of an orthologous enhancer in the outgroup species, or a de novo enhancer if absent. (*D*) We compared the numbers of different types of enhancers between new and parental genes, and showed here the average numbers (enhancers per gene) for each category. Each row represents one type of *Drosophila* developmental or housekeeping gene enhancers identified from Zabidi et al. (2015), and human ubiquitous or specific enhancers are derived from Andersson et al. (2014). We show the significance level of Fisher's test by different numbers of asterisks (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$).

regulated genes, with each enriched for separate classes of sequence motifs (Zabidi et al. 2015). In parallel, human enhancers have also been annotated to have ubiquitous or cell-type/tissue-specific expression (Andersson et al. 2014). We compared these two types of enhancers' distributions between new and parental genes in both human and *Drosophila*. Indeed, for new genes produced by DNA-based gene duplication, there are significantly ($P < 0.05$, *t*-test) fewer housekeeping/ubiquitous enhancers in new genes than parental genes. Whereas retrogenes possess more, although not significantly ($P > 0.05$, *t*-test), developmental/tissue-specific enhancers than their parental genes (fig. 5*D*). Correspondingly, there are also fewer housekeeping gene-related sequence motifs (e.g., Ohler motif 7) in new gene duplicates, whereas there are more tissue-specific gene-related sequence elements (e.g., TATA box and Initiator element) among new retrogenes, comparing with their parental genes (supplementary fig. S25, Supplementary Material online). These results together demonstrated that the CREs have become diverged for their numbers and types between new and parental genes, which underlies their observed different expression patterns.

Discussion

Genome-wide and experimental case studies have demonstrated that functional new genes have frequently emerged during evolution and constitute a main driving force underlying the evolution of organismal complexity (Kaessmann 2010; Chen et al. 2013). Similar to any other types of mutations, a nascent gene does not usually confer an immediate selective advantage that will drive its rapid fixation throughout the population. Many of them seem to be initially segregating within the population, and to start out as testis-specific genes. This has been observed for new genes of *Drosophila* (Betran and Long 2003; Chen et al. 2012; Assis and Bachtrog 2013), gene duplicates (Guschanski et al. 2017), or retrogenes (Carelli et al. 2016) of mammals, and parallelly in pollen of rice

and *Arabidopsis thaliana* (Cui et al. 2015), suggesting male reproductive tissues are a universal cradle for the birth of new genes. These previous analyses sometimes involved two age groups without knowing the relative contributions of different types of new genes, or only focused on certain type of new genes using a subset of tissue expression data. Here, we dramatically expand the analyzed transcriptome data set and employ a finer division of new genes regarding both their ages and origination mechanisms. We clarified that the out of the testis pattern is only observed in testis, but not in other tissues; and in *Drosophila* it is only observed for DNA-based gene duplications but not in retrogenes and de novo genes (fig. 1, supplementary fig. S6, Supplementary Material online). Several factors probably account for such preference of male reproductive tissues for the new gene birth: first, testes of *Drosophila* and mammals have a distinct epigenomic regulation program from other somatic tissues which may license more promiscuous transcription. This has been recently proposed to act as a "scanning" mechanism for exposing more genes for transcription-coupled DNA repair to reduce germline mutations (Xia et al. 2018), as well as more often exposing the nascent genes to natural selection. Particularly in mammals, RNA polymerase II is enriched in testis (Schmidt and Schibler 1995). And recently a testis-specific histone H3 variant H3t that is essential for spermatogenesis has been identified to form a flexible open chromatin structure for allowing more transcription (Ueda et al. 2017). Whereas the *Drosophila* testis does not show a canonical bivalent (H3K27me3/H3K4me3) chromatin domain on differentiation genes as regularly observed in somatic tissues (Gan et al. 2010), many testis-specific genes instead reside in "BLACK" chromatin (Filion et al. 2010) associated with lamin (Shevelyov et al. 2009), which suppresses their somatic expression. These regulation programs ensure a more robust expression of nascent genes in testis, and also restrict their potentially harmful expression in other tissues. Indeed, the second factor contributing to the out of the testis pattern is probably due to the selection against the redundant gene dosage in somatic tissues. As a response to such selection, it has been shown in yeast and mammals, that the expression level of duplication gene is reduced to maintain the gene dosage (Qian et al. 2010). And the selection against a new pleiotropic or broadly expressed gene is expected to be much stronger than that of a tissue-specific gene. This probably accounts for the pattern that many new genes, particularly retrogenes tend to emerge from a pre-existing silencing/regulatory H3K27me3 domain (fig. 3). For new gene duplications, which seem to have a larger contribution to the out-of-the-testis pattern (supplementary fig. S6, Supplementary Material online), they are much more likely than retrogenes or de novo genes to inherit their parental genes' regulatory elements with few changes (fig. 5), and thus also the expression pattern.

Such a transition beyond the testis can result from rapid turnover of tissue-specific genes or/and acquisition of broader expression patterns during evolution (Betran 2015). A recent study on mammalian duplicated genes suggested that the former is more important for liver-specific genes, whereas the latter is more important for testis-specific genes (Guschanski et al. 2017). The following important question is that how new genes acquired novel and important functions beyond the testis? We addressed the underlying regulatory mechanisms by uncovering an age-dependent acquisition of active histone marks and more turnovers of CREs among both *Drosophila* and human new genes, consistently across somatic tissues and developmental stages (figs. 4 and 5). This suggested that the general evolution trajectory of genes involves becoming more active in chromatin configuration, and more complex in cis-regulatory circuits. The change of repressive histone marks, however show variations between species and between somatic and germline tissues. The interspecific difference may be attributed to the presence and absence of DNA methylation in human and *Drosophila*, respectively. It has been reported that the level of promoter DNA methylation, which is negatively associated with gene expression level, also becomes lower in older human gene duplicate pairs (Keller and Soojin 2014). This indicated that in human, DNA methylation synergistically acts with an age-dependent loss of repressive histone marks and results in more active genes in older age groups observed in this study. However, *Drosophila* lacks DNA methylation except for a very low level of methylation at early embryonic stages (Takayama et al. 2014). Another study recently showed that *Drosophila* and mouse employ different histone modifications for balancing the gene dosage after gene duplication (Chang and Liao 2017). These factors, as well as a mixed cell types (late embryos and larvae) used for *Drosophila* ChIP-seq data in this work probably together account for the different trajectories of repressive histone marks along the age groups between *Drosophila* and human, for both new and parental genes. It is important to note that the major differences between previous studies (Arthur et al. 2014; Keller and Soojin 2014; Chang and Liao 2017) of epigenetic modifications on gene duplications and this work are that the former focused on the comparison between duplicated genes versus single-copy genes, and but not on that between parental and new gene copies. However, as shown here, because parental genes are by definition older than new genes, they can have very different trajectories of epigenetic changes (fig. 4).

Finally, we uncovered that parental and new genes have clearly diverged for their CRE repertoire and become enriched for different types of enhancers or sequence motifs. Despite the much progress that has been made in identifying the enhancers in a high-throughput manner (Andersson et al. 2014; Arnold et al. 2014), assigning them to their downstream genes still remains a great challenge. We conservatively restricted our analyses to enhancers and their nearby genes in this study, which is an underestimate of the CREs. Comparing with promoters, enhancers seem to have a faster evolution rate (Villar et al. 2015). And a pre-existing enhancer might switch its downstream target to the new gene upon its birth, and facilitate its functional innovation (fig. 5). It is therefore of great interest in the future to investigate how the numbers and combination of enhancers evolved across different ages of new genes, when more data (e.g., Hi-C) becomes available. As studying new genes' evolution throughout their life history

provides an entry point into understanding the evolution trajectory of genes in general.

Materials and Methods

Inferring Age and Origination Mechanisms of New Genes

We adopted a published whole genome alignment-based pipeline to identify the new genes and infer their origination time and mechanisms, as described in Zhang et al. (2010). For *Drosophila* (Ensembl metazoa release 26) and human (Ensembl release v73), we took advantage of UCSC whole genome syntenic alignment (https://genome.ucsc.edu/cgi-bin/hgGateway) and inferred the phylogenetic distribution of orthologs of *D. melanogaster* or human genes among the other *Drosophila* or vertebrate genomes. Species or lineage-specific genes were identified as new genes and then assigned into respective age groups, based on their presence/absence in outgroup species and parsimony. We classified new genes' origination mechanisms as DNA-based duplication (gene duplication), RNA-based duplication (retroposition), and de novo origination. We characterized retrogenes as those intronless genes whose parental genes have at least one intron. Otherwise, it will be classified as gene duplication. A gene will be defined as de novo gene if no alignment hit can be found in multiple outgroup protein repertories with a BLAST (Camacho et al. 2009) e-value cut-off as 10^{-6}, an alignment length cut-off as 70%, and a sequence identity cut-off as 50%, also without any annotated paralogs by Ensembl.

Transcriptomic and Epigenomic Analyses

Transcriptomic and epigenomic data of *D. melanogaster* and human were retrieved from databases of NCBI (https://www.ncbi.nlm.nih.gov/sra), ENCODE (https://www.encodeproject.org/), and Roadmap Epigenomics project (http://www.roadmapepigenomics.org/), and published single-cell sequencing data (Xue et al. 2013; Yan et al. 2013) (for all data resources: supplementary table S3, Supplementary Material online). After removing the potential adaptor contaminations, and base pairs of low-quality, we mapped the RNA-seq reads with HISAT2 (v2.0.5) (Kim et al. 2015), and ChIP-seq reads with Bowtie2 (v2.2.9) (Langmead and Salzberg 2012) to the reference genomes of *D. melanogaster* (r6.02) and human (hg19), using a mapping quality cut-off of 20 and taking paired–end relationship into account. To differentiate between the parental and new gene sequences, we used MUSCLE (v.3.8.31) (Zhang et al. 2010) and produced pairwise alignments for 452 *Drosophila* and 1,351 human parental-new gene pairs. Using parental genes as a reference, we recorded the nucleotide and genomic position information of all diagnostic SNPs between new and parental genes, with customized perl codes. For each diagnostic SNP, we counted the number of sequencing reads (at least three as a cutoff) that match the nucleotide of either the parental or the new genes to measure their respective levels of gene expression (from RNA-seq reads) or histone modification (from ChIP-seq reads). After calibrating the difference of total sequenced reads between different samples, the RNA-seq read number of each gene was then divided by the corresponding genomic DNA-seq read number to correct for the mapping bias, and also allow for comparison between genes. Similarly, we calculated the log2 ratio of ChIP (IP) versus input (IN) reads that span the diagnostic SNPs, for the entire gene region for the markers H3K36me1, H3K36me3, H3K9me2, H3K9me3, H3K27me3, H3K79me1, H3K79me2, H4K16ac; or specifically for the putative promoter region ($\pm$2 kb around the transcriptional start sites) for the markers H3K4me1, H3K4me2, H3K4me3, H3K18ac, H3K27ac, H3K9ac, and RNA polymerase II. For de novo genes without any parental genes, we used BEDTools (v2.25.0) (Quinlan and Hall 2010) to count the total read number within the gene regions. To test for the validity of normalization, we performed correlation analyses between the gene expression level versus the histone modification level with R, which showed consistent results with those derived from ENCODE or modENCODE project. We defined a gene as being transcriptionally active or bound by certain histone modifications, based on the distribution of expression level of normalized histone modification level of all genes or promoters in the respective tissue or stage with a cut-off of log2(IP/IN) ratio higher than 0. Bivalent genes were defined as those bound by both H3K27me3 and H3K4me3.

We used two *Drosophila* RNA-FISH databases, Fly-FISH (Lécuyer et al. 2007) and Berkeley *Drosophila* Genome Project (BDGP) (Tomancak et al. 2002) for comparing the localization patterns between new and parental genes during embryogenesis. We combined the two databases and when there were overlapping genes between the two, we selected genes with their parental or new gene's data available in the same database. Then we compared the annotated anatomical terms and embryonic stages with detected expression between new and parental genes.

CRE Data Analysis

We used a nonredundant enhancer data set annotated for *D. melanogaster*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura*, and *D. willistoni* by STARR-seq (Arnold et al. 2014), and a human enhancer data set annotated by FANTOM5 consortium by CAGE (Andersson et al. 2014). For the *Drosophila* housekeeping or developmental gene enhancers, we used the data from (Zabidi et al. 2015), and for human ubiquitous or tissue/cell-specific enhancer, we used the data from FANTOM5 consortium (Andersson et al. 2014). We assigned the enhancer–gene relationship following the same rules as the published work: for STARR-seq enhancers, they are assigned to either parental or new genes when they fall within 2 kb up or downstream of the TSS; for the FANTOM5 enhancers, they are assigned to either parental or new genes when they fall within 5 kb up or downstream of the TSS. For the *Drosophila* housekeeping/developmental gene enhancers, we additionally include those that located within the 5 kb upstream from the TSS, within the gene body itself, 2 kb downstream of the gene, as well as the "closest enhancer" which is assigned to the closest TSS of an annotated gene. To verify the enhancer activities, we calculated the log2 transformed IP/IN ratios at the enhancer regions, after aligning the ChIP-seq reads of H3K27me3, H3K4me1, and H3K27ac derived from *Drosophila* S2 cells

and human K562 cells to the respective reference genomes using bowtie2. We defined an enhancer gain event when the new gene has a specific enhancer that is absent in the parental gene and also outgroup species (see below), and vice versa for "enhancer loss"; whereas enhancer duplication is defined as the case that new and parental genes share the identical enhancer sequence; and "enhancer mutation" refers to the case that new and parental genes have sequence divergences between a pair of homologous enhancers. We examined the candidate cases of enhancer gain/loss using genome alignments between the focal and outgroup species. Once the coordinates of the focal enhancer were translated into those in the outgroup by the UCSC liftOver tool, we further used BEDTools to check the presence/absence of orthologous sequence. For *Drosophila*, we investigated branches E to B, where STARR-seq annotated enhancers are available for the included species. When examining the source of enhancer gain, if the orthologous sequence of the focal enhancer in the outgroup has also been annotated as an enhancer, we defined the gained enhancer as a pre-existing enhancer. Otherwise, it is defined as a de novo enhancer. We used MEME suite (Bailey et al. 2009) to search for the motif occurrences in the new and parental genes, with the published motif matrixes (Zabidi et al. 2015) as queries.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Yong Zhang, Da-qi Yu, Chun-yan Chen for providing the new gene data sets (https://gentree.ioz.ac.cn), and James Howie for his helpful comments. This project is supported by National Natural Science Foundation of China (31722050, 31671319), European Research Council (grant agreement 677696), the Fundamental Research Funds for the Central Universities (2018XZZX002-04), and start-up funds from Zhejiang University to Z.Q.

Author Contributions

Z.Q. conceived the project, Z.J. and Z.Q. performed the analyses and wrote the manuscript.

References

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* 507(7493):455–461.

Arnold CD, Gerlach D, Spies D, Matts JA, Sytnikova YA, Pagani M, Lau NC, Stark A. 2014. Quantitative genome-wide enhancer activity maps for five *Drosophila* species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nat Genet.* 46(7):685–692.

Arthur RK, Ma L, Slattery M, Spokony RF, Ostapenko A, Negre N, White KP. 2014. Evolution of H3K27me3-marked chromatin is linked to gene expression evolution and to patterns of gene duplication and diversification. *Genome Res.* 24(7):1115–1124.

Assis R, Bachtrog D. 2013. Neofunctionalization of young duplicate genes in *Drosophila*. *Proc Natl Acad Sci U S A.* 110(43):17409–17414.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37(Web Server issue):W202–W208.

Betran E. 2015. The "life histories" of genes. *J Mol Evol.* 80:186–188.

Betran E, Long M. 2003. Dntf-2r, a young *Drosophila* retroposed gene with specific male expression under positive Darwinian selection. *Genetics* 164:977–988.

Betran E, Thornton K, Long M. 2002. Retroposed new genes out of the X in *Drosophila*. *Genome Res.* 12(12):1854–1859.

Braude P, Bolton V, Moore S. 1988. Human gene expression first occurs between the four-and eight-cell stages of preimplantation development. *Nature* 332(6163):459.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinform.* 10:421.

Carelli FN, Hayakawa T, Go Y, Imai H, Warnefors M, Kaessmann H. 2016. The life history of retrocopies illuminates the evolution of new mammalian genes. *Genome Res.* 26(3):301–314.

Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, et al. 2012. Proto-genes and de novo gene birth. *Nature* 487(7407):370–374.

Chang AY, Liao BY. 2017. Recruitment of histone modifications to assist mRNA dosage maintenance after degeneration of cytosine DNA methylation during animal evolution. *Genome Res.* 27(9):1513–1524.

Chen S, Krinsky BH, Long M. 2013. New genes as drivers of phenotypic evolution. *Nat Rev Genet.* 14(9):645–660.

Chen S, Ni X, Krinsky BH, Zhang YE, Vibranovski MD, White KP, Long M. 2012. Reshaping of global gene expression networks and sex-biased gene expression by integration of a young gene. *EMBO J.* 31(12):2798–2809.

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

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.

GTEx Consortium. 2017. Genetic effects on gene expression across human tissues. *Nature* 550:204–213.

Cui X, Lv Y, Chen M, Nikoloski Z, Twell D, Zhang D. 2015. Young genes out of the male: an insight from evolutionary age analysis of the pollen transcriptome. *Mol Plant.* 8(6):935–945.

Dai H, Chen Y, Chen S, Mao Q, Kennedy D, Landback P, Eyre-Walker A, Du W, Long M. 2008. The evolution of courtship behaviors through the origination of a new gene in *Drosophila*. *Proc Natl Acad Sci U S A.* 105(21):7478–7483.

Dai H, Yoshimatsu TF, Long M. 2006. Retrogene movement within- and between-chromosomes in the evolution of *Drosophila* genomes. *Gene* 385:96–102.

Ding Y, Zhao L, Yang S, Jiang Y, Chen Y, Zhao R, Zhang Y, Zhang G, Dong Y, Yu H, et al. 2010. A young *Drosophila* duplicate gene plays essential roles in spermatogenesis by regulating several Y-linked male fertility genes. *PLoS Genet.* 6(12):e1001255.

Domazet-Loso T, Tautz D. 2010. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* 468(7325):815–818.

Emerson JJ, Kaessmann H, Betran E, Long M. 2004. Extensive gene traffic on the mammalian X chromosome. *Science* 303(5657):537–540.

Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ, et al. 2010. Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. *Cell* 143(2):212–224.

Gan Q, Schones DE, Eun SH, Wei G, Cui K, Zhao K, Chen X. 2010. Monovalent and unpoised status of most genes in undifferentiated cell-enriched *Drosophila* testis. *Genome Biol.* 11(4):R42.

Guschanski K, Warnefors M, Kaessmann H. 2017. The evolution of duplicate gene expression in mammalian organs. *Genome Res.* 27(9):1461–1474.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* 20(10):1313–1326.

Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet.* 10(1):19–31.
Kalinka AT, Varga KM, Gerrard DT, Preibisch S, Corcoran DL, Jarrells J, Ohler U, Bergman CM, Tomancak P. 2010. Gene expression divergence recapitulates the developmental hourglass model. *Nature* 468(7325):811–814.
Katju V, Lynch M. 2006. On the formation of novel genes by duplication in the *Caenorhabditis elegans* genome. *Mol Biol Evol.* 23(5):1056–1067.
Keller TE, Soojin VY. 2014. DNA methylation and evolution of duplicate genes. *Proc Natl Acad Sci U S A.* 111(16):5932–5937.
Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov AA, Gu T, et al. 2011. Comprehensive analysis of the chromatin landscape in *Drosophila* melanogaster. *Nature* 471(7339):480–485.
Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods.* 12(4):357–360.
Knowles DG, McLysaght A. 2009. Recent de novo origin of human protein-coding genes. *Genome Res.* 19(10):1752–1759.
Kondo S, Vedanayagam J, Mohammed J, Eizadshenass S, Kan LJ, Pang N, Aradhya R, Siepel A, Steinhauer J, Lai EC. 2017. New genes often acquire male-specific functions but rarely become essential in *Drosophila*. *Genes Dev.* 31(18):1841–1846.
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 9(4):357–359.
Lécuyer E, Yoshida H, Parthasarathy N, Alm C, Babak T, Cerovina T, Hughes TR, Tomancak P, Krause HM. 2007. Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell* 131(1):174–187.
Lesch BJ, Silber SJ, McCarrey JR, Page DC. 2016. Parallel evolution of male germline epigenetic poising and somatic development in animals. *Nat Genet.* 48(8):888–894.
Long M, Langley CH. 1993. Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. *Science* 260:91–95.
Lott SE, Villalta JE, Schroth GP, Luo S, Tonkin LA, Eisen MB. 2011. Noncanonical compensation of zygotic X transcription in early *Drosophila* melanogaster development revealed through single-embryo RNA-seq. *PLoS Biol.* 9(2):e1000590.
Luis Villanueva-Canas J, Ruiz-Orera J, Agea MI, Gallo M, Andreu D, Alba MM. 2017. New genes and functional innovation in mammals. *Genome Biol Evol.* 9(7):1886–1900.
Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154(1):459–473.
Ohno S. 1970. Evolution by gene duplication: Berlin, Heidelberg (Germany): Springer.
Palmieri N, Kosiol C, Schlotterer C. 2014. The life cycle of *Drosophila* orphan genes. *Elife* 3:e01311.
Perez-Lluch S, Blanco E, Tilgner H, Curado J, Ruiz-Romero M, Corominas M, Guigo R. 2015. Absence of canonical marks of active chromatin in developmentally regulated genes. *Nat Genet.* 47(10):1158–1167.
Qian W, Liao B-Y, Chang AY-F, Zhang J. 2010. Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends Genet.* 26(10):425–430.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841–842.
Ross BD, Rosin L, Thomae AW, Hiatt MA, Vermaak D, de la Cruz AF, Imhof A, Mellone BG, Malik HS. 2013. Stepwise evolution of essential centromere function in a *Drosophila* neogene. *Science* 340(6137):1211–1214.
Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF, et al. 2010. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* 330(6012):1787–1797.
Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabido E, Kondova I, Bontrop R, Marques-Bonet T, Alba MM. 2015. Origins of de novo genes in human and chimpanzee. *PLoS Genet.* 11(12):e1005721.
Schmidt EE, Schibler U. 1995. High accumulation of components of the RNA polymerase II transcription machinery in rodent spermatids. *Development* 121(8):2373–2383.
Shevelyov YY, Lavrov SA, Mikhaylova LM, Nurminsky ID, Kulathinal RJ, Egorova KS, Rozovsky YM, Nurminsky DI. 2009. The B-type lamin is required for somatic repression of testis-specific gene clusters. *Proc Natl Acad Sci U S A.* 106(9):3282–3287.
Sysoev VO, Fischer B, Frese CK, Gupta I, Krijgsveld J, Hentze MW, Castello A, Ephrussi A. 2016. Global changes of the RNA-bound proteome during the maternal-to-zygotic transition in *Drosophila*. *Nat Commun.* 7:12128.
Takayama S, Dhahbi J, Roberts A, Mao G, Heo SJ, Pachter L, Martin DI, Boffelli D. 2014. Genome methylation in *D. melanogaster* is found at specific short motifs and is independent of DNMT2 activity. *Genome Res.* 24:821–830.
Tomancak P, Beaton A, Weiszmann R, Kwan E, Shu S, Lewis SE, Richards S, Ashburner M, Hartenstein V, Celniker SE, et al. 2002. Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol.* 3(12):RESEARCH0088.
Ueda J, Harada A, Urahama T, Machida S, Maehara K, Hada M, Makino Y, Nogami J, Horikoshi N, Osakabe A, et al. 2017. Testis-specific histone variant H3t gene is essential for entry into spermatogenesis. *Cell Rep.* 18(3):593–600.
Vibranovski MD, Zhang Y, Long M. 2009. General gene movement off the X chromosome in the *Drosophila* genus. *Genome Res.* 19(5):897–903.
Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R, Erichsen JT, Jasinska AJ, et al. 2015. Enhancer evolution across 20 mammalian species. *Cell* 160(3):554–566.
Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene copies in the human genome. *Proc Natl Acad Sci U S A.* 103(9):3220–3225.
Wu DD, Irwin DM, Zhang YP. 2011. De novo origin of human protein-coding genes. *PLoS Genet.* 7(11):e1002379.
Xia B, Baron M, Yan Y, Wagner F, Kim SY, Keefe DL, Alukal JP, Boeke JD, Yanai I. 2018. Widespread transcriptional scanning in testes modulates gene evolution rates. *bioRxiv* https://doi.org/10.1101/282129.
Xue Z, Huang K, Cai C, Cai L, Jiang CY, Feng Y, Liu Z, Zeng Q, Cheng L, Sun YE, et al. 2013. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* 500(7464):593–597.
Yan L, Yang M, Guo H, Yang L, Wu J, Li R, Liu P, Lian Y, Zheng X, Yan J, et al. 2013. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat Struct Mol Biol.* 20(9):1131–1139.
Zabidi MA, Arnold CD, Schernhuber K, Pagani M, Rath M, Frank O, Stark A. 2015. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* 518(7540):556–559.
Zhang W, Landback P, Gschwend AR, Shen B, Long M. 2015. New genes drive the evolution of gene interaction networks in the human and mouse genomes. *Genome Biol.* 16:202.
Zhang YE, Landback P, Vibranovski M, Long M. 2012. New genes expressed in human brains: implications for annotating evolving genomes. *Bioessays* 34(11):982–991.
Zhang YE, Vibranovski MD, Landback P, Marais GA, Long M. 2010. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. *PLoS Biol.* 8(10): e1000494.
Zhao L, Saelao P, Jones CD, Begun DJ. 2014. Origin and spread of de novo genes in *Drosophila* melanogaster populations. *Science* 343:769–772.
Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W. 2008. On the origin of new genes in *Drosophila*. *Genome Res.* 18(9):1446–1455.