Evolution of H3K27me3-marked chromatin is linked to gene expression evolution and to patterns of gene duplication and diversification

Robert K. Arthur,[1,2] Lijia Ma,[2,3] Matthew Slattery,[2,3,4] Rebecca F. Spokony,[2,3,5] Alexander Ostapenko,[2,3] Nicolas Nègre,[2,3,6] and Kevin P. White[1,2,3,7]

[1]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA; [2]Institute for Genomics and Systems Biology, University of Chicago and Argonne National Laboratory, Chicago, Illinois 60637, USA; [3]Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA; [4]Department of Biomedical Sciences, University of Minnesota Medical School, Duluth, Minnesota 55455, USA; [5]Department of Natural Sciences, Baruch College, City University of New York, New York 10010, USA; [6]Université de Montpellier 2 and INRA, UMR1333 DGIMI, F-34095 Montpellier, France

Histone modifications are critical for the regulation of gene expression, cell type specification, and differentiation. However, evolutionary patterns of key modifications that regulate gene expression in differentiating organisms have not been examined. Here we mapped the genomic locations of the repressive mark histone 3 lysine 27 trimethylation (H3K27me3) in four species of *Drosophila,* and compared these patterns to those in *C. elegans*. We found that patterns of H3K27me3 are highly conserved across species, but conservation is substantially weaker among duplicated genes. We further discovered that retropositions are associated with greater evolutionary changes in H3K27me3 and gene expression than tandem duplications, indicating that local chromatin constraints influence duplicated gene evolution. These changes are also associated with concomitant evolution of gene expression. Our findings reveal the strong conservation of genomic architecture governed by an epigenetic mark across distantly related species and the importance of gene duplication in generating novel H3K27me3 profiles.

[Supplemental material is available for this article.]

While transcriptional regulation has long been recognized as a significant target of evolutionary change (King and Wilson 1975), the specific mechanisms behind regulatory divergence have been difficult to dissect (Carroll 2008). In most cases, research has focused on recognizable *cis*-elements where transcription factors bind in a sequence-specific manner (Borneman et al. 2007; Bradley et al. 2010; Dowell 2010; Ni et al. 2012). Changes to either a transcription factor's DNA-binding properties or transcription factor binding sites (TFBS) enable the evolution of differential regulation, and numerous examples of this phenomenon have been observed (Ludwig et al. 2005; Shultzaberger et al. 2012).

Whereas changes in *cis*-regulatory elements have been implicated in phenotypic and specifically morphological changes (Carroll 2008; Stern and Orgogozo 2008), other components of transcriptional regulation such as the chromatin environment have been scarcely explored (Lenhard et al. 2012). Histone modifications, which are chemical alterations of the histone spools upon which DNA is threaded, constitute one of the best-described elements of chromatin state (Zhou et al. 2011). These modifications can act directly or indirectly (through recruited enzymes) to alter DNA accessibility, thereby controlling other DNA–protein interactions (Campos and Reinberg 2009). Unlike TFBSs, however, histone modifications are not necessarily easily localizable to particular sequence elements, making their evolution difficult to study (Delest et al. 2012).

Histone 3 lysine 27 trimethylation (H3K27me3) is one of the best-known histone modifications, in terms of both its biogenesis and effects. H3K27me3 is associated with complex *cis*-regulatory elements called Polycomb Response Elements (PREs) (Delest et al. 2012). Unlike TFBSs, PREs are compound regulatory elements composed of multiple, sometimes partially redundant, sequence elements (Delest et al. 2012). Although the exact biochemical binding site composition and structural constraints upon PREs are not completely resolved, a group of coordinately binding proteins called Polycomb group (PcG) factors recruit Polycomb Repressive Complexes, which are necessary to deposit and maintain H3K27me3 and have been associated genome wide with H3K27me3 domains (Lanzuolo and Orlando 2012).

The effect of H3K27me3 is unambiguously and strongly repressive (Filion et al. 2010; Kharchenko et al. 2011; Li et al. 2014). H3K27me3-associated loci have been proposed to congregate in silenced foci called Polycomb factories, which inhibit transcription by preventing access to RNA polymerase II and other *trans*-factors (Bantignies et al. 2011; Sexton et al. 2012). This repressive state is both temporally and spatially dynamic (Akkers et al. 2009; Filion et al. 2010; Kharchenko et al. 2011; Nègre et al. 2011).

Accordingly, the PcG proteins and H3K27me3 have been shown to be necessary for differentiation and maintenance of cell type identity in organisms across eukaryotes (Bernstein et al. 2006; Feng and Jacobsen 2011). Dysregulation of H3K27me3 has also been implicated in the genesis and progression of cancer (Chi et al.

[7]**Corresponding author**
E-mail kpwhite@uchicago.edu

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.162008.113. Freely available online through the *Genome Research* Open Access option.

2010; Ellinger et al. 2012). These results indicate that H3K27me3 plays a role in the creation and maintenance of cell type-specific programs of transcriptional control for a wide variety of species and cell fates.

Given its close association with the process of differentiation, the extent to which H3K27me3 domains are conserved across species is of great interest (Shubin et al. 2009; Cain et al. 2011). We sought to examine the rate of evolutionary change in H3K27me3 between four species of the well-characterized *Drosophila* clade, the mechanisms by which such change might occur, and the possible consequences of H3K27me3 evolution for nearby gene expression.

In order to investigate the evolution of H3K27me3 patterns across the genome, we performed chromatin immunoprecipitation followed by sequencing (ChIP-seq) for H3K27me3 in four species, *D. melanogaster*, *D. simulans*, *D. yakuba*, and *D. pseudoobscura*, with divergence times ranging from less than 5 million years (Myr) to more than 35 Myr (Tamura et al. 2004; Obbard et al. 2012). To determine whether and how H3K27me3 changes might alter gene expression, we also performed RNA sequencing (RNA-seq) in all species. For all analyses we used white prepupae for ease of developmental synchronization between species and because there has been extensive previous work characterizing genome-wide regulatory evolution at this developmental stage (Rifkin et al. 2003; Gu et al. 2004; Ni et al. 2012). We find that for single-copy orthologous genes, H3K27me3 signal is strongly conserved in even distantly related species. However, duplicated gene orthologs exhibit much greater divergence in H3K27me3. Moreover, different kinds of duplicates appear to have very different rates of H3K27me3 and expression divergence, indicating functional distinctions in the epigenetic consequences of different gene duplication mechanisms.

Results

Evolutionary stasis of H3K27me3 levels across single-copy genes in *Drosophila*

We examined H3K27me3 signal at the white prepupal stage of development, a tightly defined, 20-min window at the beginning of pupariation. Two biological replicates were collected for each experiment and one control. H3K27me3 presents unique challenges for ChIP-seq analysis. Unlike transcription factors and certain other histone modifications, H3K27me3 forms broad patterns of enrichment that are of indeterminate length (genome-wide analysis has shown these to be up to ∼100 kb in *Drosophila*) (Papp and Müller 2006; Kharchenko et al. 2011; Nègre et al. 2011), and therefore simple peak-based analysis is not a suitable analysis method (Xiao et al. 2012). Because of the many distinct cell types of a *Drosophila* puparium, it is also not appropriate to treat H3K27me3 as a binary modification of chromatin (Zang et al. 2009). Therefore we chose to analyze the data quantitatively, using the number of sequence reads falling within prespecified intervals (e.g., the exons of a gene) as an indication of the overall level of H3K27me3 signal within a region (Supplemental Fig. S1). This measure is both highly correlated between biological replicates (Supplemental Fig. S2) and robustly anticorrelated with gene expression (Supplemental Fig. S3), suggesting that the metric is biologically meaningful. Using the modENCODE-curated list of orthologous genes in *Drosophila* (AP Boyle, CL Araya, C Brdlik, P Cayting, C Cheng, Y Cheng, K Gardner, L Hillier, J Janette, L Jiang, et al., in prep.), we directly compared each gene in the *melanogaster* genome to its ortholog in *D. simulans*, *D. yakuba*, and *D. pseudoobscura* (abbreviated Sim, Yak, and Pse). In total, we compared 12,017 orthologs from *D. melanogaster* to *D. simulans*; 11,018 from *D. melanogaster* to *D. yakuba*; and 11,881 from *D. melanogaster* to *D. pseudoobscura.*

Overall we found high conservation of H3K27me3 signal between *Drosophila* genomes, ranging from Spearman correlation coefficients (SCC) of 0.78 to 0.88 (Fig. 1A; a representative locus is depicted in Supplemental Fig. S4). Given that any two pairs of biological replicate experiments show a SCC of ∼0.95 (Supplemental Fig. S2), the observed between-species correlations indicate relatively slow evolutionary change for H3K27me3 patterns across the genome. Simulations showed that a relatively small level of ortholog misidentification or misannotation error (10%–20%; similar to published estimates) (Chen et al. 2007; Creevey et al. 2011) is sufficient to explain the observed decrease in correlation

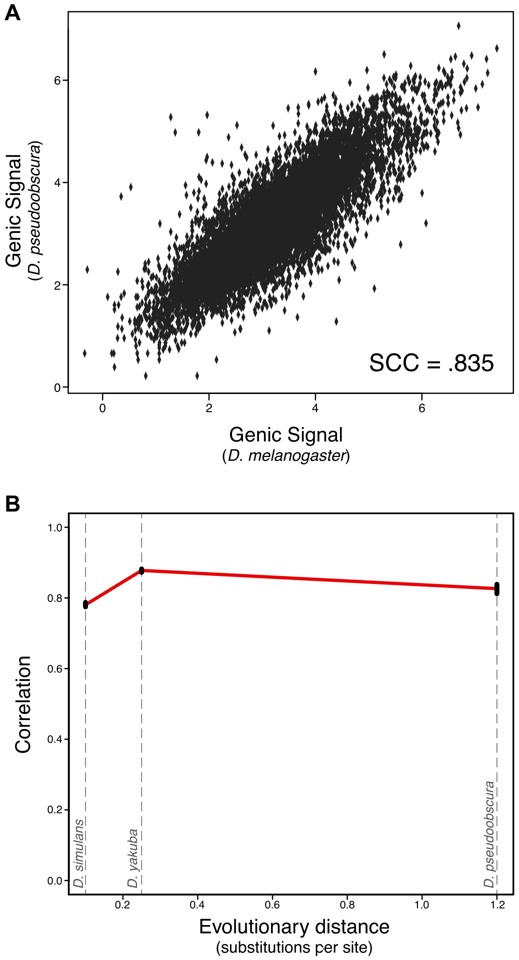

Figure 1. Strong conservation of H3K27me3 in *Drosophila* orthologs. (*A*) Example graph of orthologous gene conservation for the comparison of *melanogaster* to *pseudoobscura*. Each dot is a single-copy orthologous gene pair, and the position on the *x*-axis represents the log *melanogaster* genic signal (see Methods), while the *y*-axis represents the log *pseudoobscura* genic signal; each is the mean of two experiments. The overall rank correlation coefficient between species is 0.835. (*B*) Overall trend of single-copy ortholog conservation within *Drosophila*. Each point is the Spearman rank correlation of one pairwise between-species comparison, plotted against the evolutionary distance from *Drosophila melanogaster* (in substitutions per neutral site). Black bars represent bootstrapped 95% confidence intervals. From *left* to *right* the species are *D. simulans* (SCC = 0.781), *D. yakuba* (SCC = 0.878), and *D. pseudoobscura* (SCC = 0.835).

between species relative to the technical variation within species. Notably, we do not see evidence for a linearly decreasing trend of conservation over phylogenetic distance within the three examined pairwise comparisons (Fig. 1B), in contrast to what has been observed for transcription factor binding sites in similarly low numbers of species (Dowell 2010; He et al. 2011; Ni et al. 2012).

Fold change in H3K27me3 of single-copy orthologs showed a weak correlation with accelerated sequence evolution. For example, change in H3K27me3 is positively associated with elevated d_N/d_S (SCC = 0.045, permutation test, $P < 10^{-5}$) in a comparison of *D. melanogaster* and *D. simulans*. These results are consistent with observations of mammalian stem cells (Xiao et al. 2012). Surprisingly, there is no strong correlation between change in H3K27me3 signal and change in gene expression among single-copy orthologous genes (SCC: Sim: 0.06; Yak: 0.043; Pse: −0.027). As noted above, this observation is consistent with most differences between single-copy orthologs being a result of technical variation. Alternatively, differences in H3K27me3 signal might be compensated for by other mechanisms of transcriptional regulation, resulting in little ultimate difference in gene expression.

Conservation of H3K27me3 between *Drosophila* and *Caenorhabditis* orthologs

Given the extremely high conservation of the H3K27me3 epigenetic mark between *Drosophila* orthologs, we investigated whether H3K27me3 signal is conserved at greater phylogenetic distances. We chose to examine pairwise conservation between *D. melanogaster* and the nematode worm *Caenorhabditis elegans* (abbreviated Cel), whose most recent common ancestor dates to the Cambrian divergence (>500 Myr) (Nei et al. 2001). We compared our *D. melanogaster* prepupal data with corresponding H3K27me3 data generated by the modENCODE Consortium at the L3 stage in worms (results were similar when comparing embryonic stages as well; Supplemental Fig. S5). There were a total of 3157 single-copy orthologs between these two species.

We found substantial evidence of H3K27me3 conservation between these two highly diverged species (Fig. 2A) (SCC = 0.487); H3K27me3 is nearly as conserved as gene expression for these genes (SCC = 0.497). Assuming the lowest observed rate of loss in correlation observed within *Drosophila* and a linear relationship between divergence time and loss of correlation, one would expect no significant correlation given the phylogenetic distance separating worm and fly. H3K27me3 has previously been compared between cell lines from human, mouse, and pig, three species separated by 50–100 million years (Xiao et al. 2012). To our knowledge this is the first study to show significant conservation of an epigenetic mark between species as diverged as *D. melanogaster* and *C. elegans*. There is a moderate, significant negative correlation between gain of H3K27me3 and loss of expression (Fig. 2B) (SCC = −0.263). This result stands in contrast to the relative lack of correlation between expression change and H3K27me3 change within *Drosophila*, and may be a consequence of greater fold change when comparing *D. melanogaster*–*C. elegans* (median, absolute $\log_2$ fold change: 0.699) than *D. melanogaster* with, for example, *D. pseudoobscura* (median, absolute $\log_2$ fold change: 0.469).

Evolution of H3K27me3 is more rapid in gene duplicates

Next we sought to examine how H3K27me3 signal might vary in the aftermath of gene duplication events. Gene duplication has been found to be a significant driver of regulatory and gene expression divergence in other interspecific comparisons (Gu et al. 2004; Kaessmann et al. 2009). While we found above that H3K27me3 is highly conserved among single-copy orthologs, we hypothesized that gene duplication might allow the creation of novel H3K27me3 regulation among duplicated gene orthologs. We therefore compared H3K27me3 levels for each newly duplicated gene to its corresponding single-copy ortholog in another genome (see Supplemental Fig. S6).

Because duplicated genes tend to share substantial nucleotide similarity, it is possible that some ChIP-seq reads might have difficulty mapping accurately to paralogs, either by mapping to the wrong paralog or failing to map uniquely and thus being discarded. To investigate the extent of this issue, we performed read simulation studies (Huang et al. 2012) which showed that most paralogs are accurately mappable, and we excluded from further analysis those genes that are not mappable (see Supplemental Methods; Supplemental Fig. S7). As a further precaution, we also

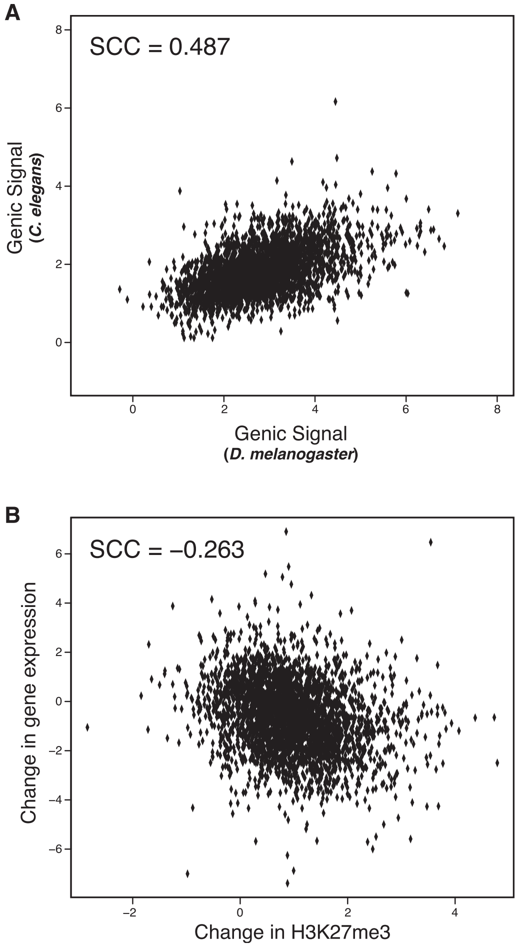

Figure 2. Conservation of H3K27me3 extends to *Caenorhabditis elegans*. (*A*) Substantial conservation of H3K27me3 between *D. melanogaster* and *C. elegans*. As in Figure 1A, each dot represents a gene, with the *x*-axis position corresponding to log H3K27me3 signal in *D. melanogaster* and the *y*-axis position corresponding to log H3K27me3 signal in *C. elegans*. The Spearman correlation coefficient is 0.487 between them. (*B*) Change in H3K27me3 levels is associated with changes in gene expression. Each dot is an individual single-copy orthologous gene, where the *x*-axis is the change in log H3K27me3 and the *y*-axis is the change in log expression. The overall Spearman correlation coefficient is −0.263.

performed read mapping in which up to two valid alignments are accepted for the pairwise comparison of *D. melanogaster* to *D. yakuba*; all results remained significant.

To call individual genes, either duplicated or single-copy orthologs, as significantly diverged in H3K27me3 signal (relative to the orthologous copy in the other genome), we quantified the degree of H3K27me3 divergence and compared to a permutation-based null (see Supplemental Methods). Significantly evolved genes (FDR of 0.05) showed greater absolute differences in expression between species (Permutation test: Sim: $P = 0.11$; Yak: $P < 10^{-5}$; Pse: $P < 10^{-5}$), and higher d_N/d_S (Permutation test: Sim: $P < 10^{-5}$).

H3K27me3 conservation differs systematically depending upon the occurrence of gene duplications. Duplicated gene orthologs generally have lower conservation of H3K27me3 signal, relative to the single-copy orthologs (Fig. 3A–C) (this is true even after accounting for differences in number of duplicated vs. single-copy orthologs). In each species comparison, duplicated genes constitute three to four times more of the H3K27me3-diverged set than expected based on their overall frequency in the compared genomes (Table 1) (Fisher's exact test: Sim: $P = 3 \times 10^{-12}$; Yak: $P < 2.2 \times 10^{-16}$; Pse: $P = 6.9 \times 10^{-11}$). Based on these results, we infer that H3K27me3 signal is more labile following gene duplication events.

While duplicated gene orthologs tended to have less confidence in orthology assignment, we did not find a significant association between ortholog bootstrap confidence and magnitude of change in H3K27me3, indicating that poor ortholog identification is not responsible for the observed trend (Supplemental Fig. S8). In addition, excluding large ortholog families that had experienced many duplications did not change the trend significantly (Supplemental Fig. S9). To further verify that read mismapping within paralogs was not the cause of our results, we analyzed the promoters of each orthologous gene similarly. Because promoter

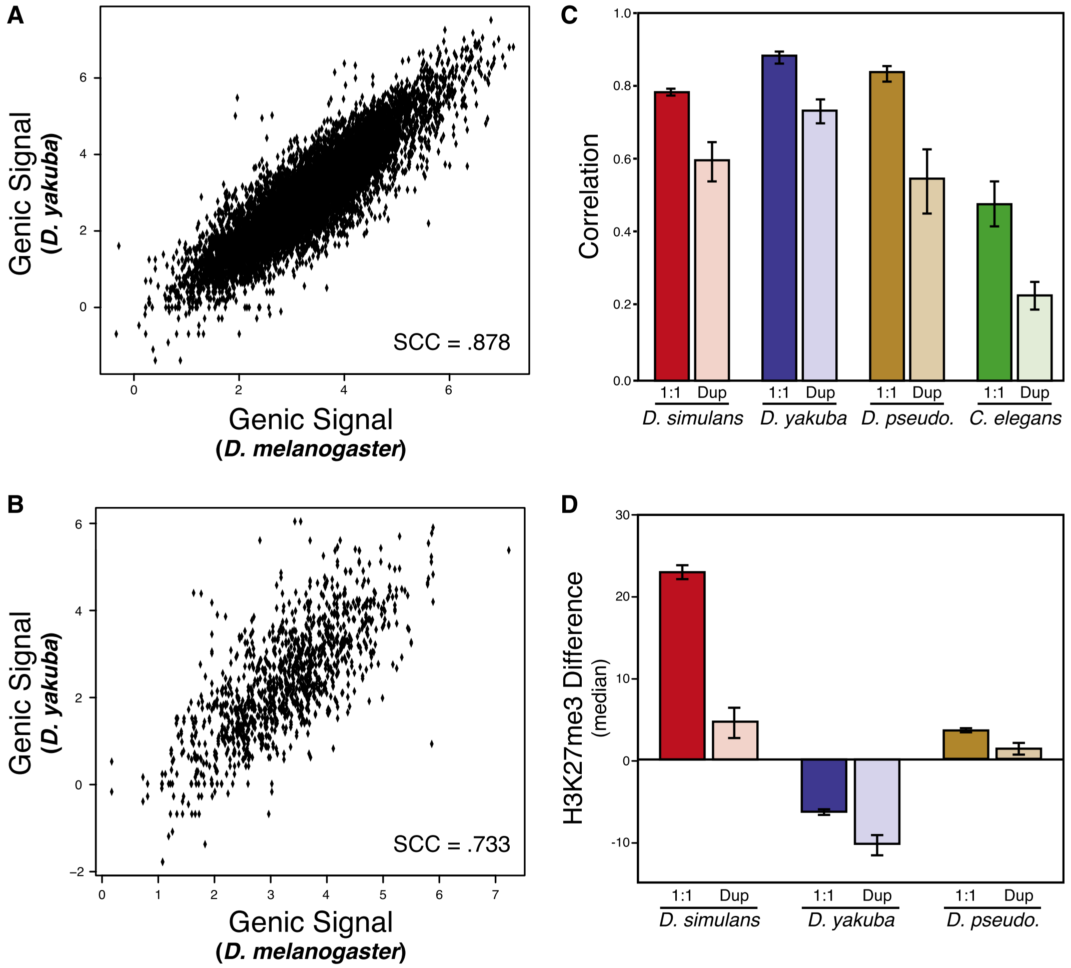

Figure 3. Duplicated genes have less conserved H3K27me3 signal. (*A*) As in Figure 1A, single-copy orthologous gene H3K27me3 conservation between *melanogaster* and *yakuba*. (*B*) As in *A* and Figure 1A, but depicting duplicated gene orthologs. (*C*) Overall pattern of H3K27me3 conservation in different gene sets. Each pair of bars represents one species comparison; bars on the *left* are the correlation coefficient of single-copy orthologs; on the *right*, the correlation coefficient of duplicated gene orthologs. Each bar represents a single pairwise comparison's Spearman rho, compared against *Drosophila melanogaster*: (red) *simulans*; (purple) *yakuba*; (brown) *pseudoobscura*; (green) *C. elegans*. Overall correlation decreases in each case. Lines are 95% bootstrapped confidence intervals for each correlation; we account for the differences in sample size (single-copy orthologs are more common than duplicated gene orthologs) by resampling the number of duplicated gene orthologs in each case. (*D*) Duplicated gene orthologs are more likely to lose H3K27me3. Each pair of barplots is one species (*simulans*, *yakuba*, *pseudoobscura*); the *left* in each pair is the median H3K27me3 change in duplicated genes relative to the single gene in the other genome; while the *right* is the median in single-copy orthologs. Each pair shows that duplicated genes are biased toward the loss of H3K27me3 signal. For each pair (by permutation test): *simulans*, $P < 10^{-5}$; *yakuba*, $P < 10^{-5}$; and *pseudoobscura*, $P < 10^{-5}$.

sequences evolve more rapidly than exonic sequence, promoters should suffer less from read mismapping. We found that, as with exonic H3K27me3 signal, promoters' methylation signal conservation was less for duplicated gene orthologs than for single-copy orthologs (Supplemental Fig. S10).

While change in H3K27me3 was associated with accelerated sequence evolution in single-copy orthologs, this effect was much more pronounced in duplicated gene orthologs (SCC = 0.099 for duplicated genes vs. SCC = 0.045 for single-copy orthologs; this difference is significant [by bootstrap, $P < 0.05$]). Similarly, duplicated gene orthologs showed stronger correlations between gain of H3K27me3 and loss of expression (Permutation test: Sim, $P < 10^{-4}$; Yak, $P < 10^{-4}$; Pse, $P < 10^{-4}$). The more robust associations of H3K27me3 evolution, gene expression evolution, and d_N/d_S in duplicated gene orthologs may indicate a tighter relationship between H3K27me3 evolution and neofunctionalization in duplicated genes than in single-copy orthologs.

Comparing duplicated genes to each other (rather than orthologous genes in another genome), we found that duplicated gene paralogs are tightly correlated in the direction and magnitude of H3K27me3 evolution after a duplication event. Evolution of H3K27me3 in duplicated gene paralogs showed significant positive correlations, indicating that both duplicates tended to undergo similar epigenetic changes after their duplication (Sim: 0.264, Yak: 0.362, Pse: 0.348, Cel: 0.447). The positive correlation indicates that the duplication event allows not only the new gene to alter its epigenetic state, but also the parent.

The median H3K27me3 change for duplicated orthologs is significantly lower than single-copy orthologous gene H3K27me3 differences in all cases (Fig. 3D), indicating an overall loss of H3K27me3 following duplication (Permutation test: Sim, $P < 10^{-5}$; Yak, $P < 10^{-5}$; Pse, $P < 10^{-5}$). However, we observed that there were significant positive correlations between gene duplicate age and evolutionary change of H3K27me3 levels, indicating that gene duplicates regain lost H3K27me3 signal as they age (SCC: Yak: 0.109, $P < 10^{-5}$; Pse: 0.064, $P = 0.015$). The loss of H3K27me3 was also related to an average increase in expression of both duplicates (Permutation test: Sim, $P < 10^{-5}$; Yak, $P < 10^{-5}$; Pse, $P < 10^{-5}$). We examined expression patterns in newly duplicated genes using the Berkeley *Drosophila* Genome Project's in situ database (Tomancak et al. 2002) and found that such genes are enriched for tissue-specific expression (Fisher's exact test, $P = 0.0006$). Since these recent duplicates generally lose H3K27me3, our results imply that such genes may acquire tissue specificity through DNA-binding *trans*-regulators.

A notable exception to this pattern is the gene *Zeus* (FBgn0032089; also called *Rcd-1r*). Previously, *Zeus* has been shown to be a recent duplicate of *CAF40* (FBgn0031047; also called *Rcd-1*), which has acquired an essential role in gonadal development (Chen et al. 2012). *Zeus* duplicated sometime before the common ancestor of *D. melanogaster* and *D. simulans*; it is not present in *D. yakuba*. *Zeus* has acquired a substantial increase in H3K27me3 in *D. melanogaster* relative to both its paralog, *CAF40*, and the location into which it duplicated in the *D. yakuba* genome (Fig. 4).

Duplicated genes relocate to regions of low H3K27me3 signal

Novel genes, once duplicated, must rapidly acquire unique functionality or otherwise suffer pseudogenization (Force et al. 1999). However, to acquire functionality a gene must first be expressed and exposed to selection. We predicted that newly duplicated genes in repressive chromatin environments (bearing high H3K27me3 signal) would be less expressed and thus more likely to become pseudogenes.

We examined the relative H3K27me3 profiles in *D. simulans* for the locations into which both *D. melanogaster*-specific novel genes and pseudogenes had duplicated. Using synteny, we located the regions in *D. simulans* that novel genes moved into to infer the ancestral H3K27me3 levels before the duplication occurred (see Supplemental Methods). Importantly, this analysis relies on the assumption that the extant *D. simulans* H3K27me3 signal is, on average, similar to the H3K27me3 signal in the ancestor of *D. melanogaster* in which the gene duplication event occurred. We found that locations into which pseudogenes had duplicated were characterized by threefold more H3K27me3 signal than the locations into which novel, protein-coding genes had duplicated (Permutation test, $P = 0.0015$) (Fig. 5). This finding is consistent with our hypothesis, and suggests that chromatin-mediated silencing of novel gene expression can prevent the acquisition of function in newly duplicated genes, adding a new dimension to the duplication–degeneration–complementation model (Force et al. 1999). Furthermore, this result provides a mechanistic explanation for the observed bias toward loss of H3K27me3 in recent duplicated genes: Only those duplicates which move into regions of low H3K27me3 signal are likely to remain functional.

Table 1. Duplicated genes are more likely to be significantly diverged than expected

Species comparison	Total number of diverged genes	Diverged single copy genes	Diverged duplicated genes	Expected diverged duplicated genes
melanogaster–simulans	589	466	123***	37
melanogaster–yakuba	569	343	226***	59
melanogaster–pseudoobscura	496	331	165***	66

Using a permutation-based approach, we called individual genes as significantly evolved with respect to H3K27me3 (both single-copy and duplicated gene orthologs; see Methods). We find that genes that have undergone duplication are much more likely to be significantly diverged than expected by chance. From *left* to *right*, columns are as follows: the species being compared; the total number of genes called as significantly diverged; the number of single-copy orthologous genes called as diverged; the number of duplicated genes called as diverged; and the expected number of duplicated genes called as diverged. Duplicated genes are more likely to be H3K27me3-diverged than expected, based on their proportion in each genome (Fisher's exact test: Sim: $P = 3 \times 10^{-12}$; Yak: $P < 2.2 \times 10^{-16}$; Pse: $P = 6.9 \times 10^{-11}$).

Patterns of H3K27me3 evolution depend on duplication mechanism

Gene duplication events are known to occur via diverse mechanisms (Hastings et al. 2009), which lead to different consequences for the resulting duplicated gene's location. Because H3K27me3 is spatially localized (Kharchenko 2011; Nègre et al. 2011), we might expect that localized gene duplication events (e.g., tandem duplication) would lead to the newly duplicated gene falling within or close to the parent gene's H3K27me3 domain. Alternatively, a mechanism such as retroposition, which is able to deposit novel genes as far from their parent as different chromosomes (Kaessmann et al.

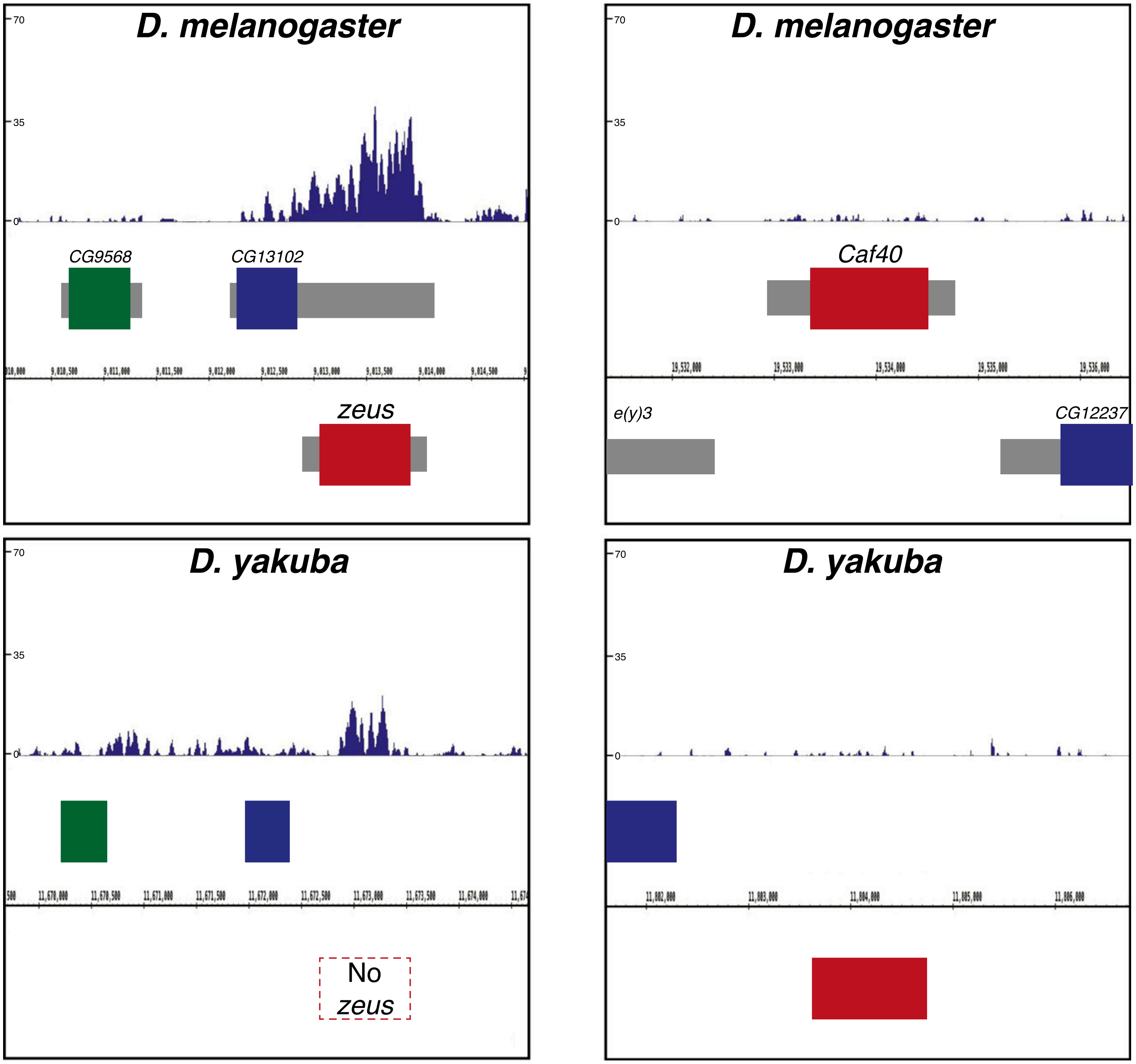

Figure 4. The novel gene duplicate *Zeus* has undergone rapid gain of H3K27me3. (*Left*) A Genome Browser snapshot indicating H3K27me3 signal at the *Zeus* (*Rcd-1r*) locus in *D. melanogaster*. The first window indicates the normalized, input-subtracted H3K27me3 signal pooled from two separate biological replicates in *D. melanogaster*. (*Below*) Boxes indicate protein-coding genes: CG9568 (green), CG13102 (blue), and *Zeus* (red). The next window is as above, but in an outgroup species (*D. yakuba*), which does not possess the *Zeus* retroposition. Browser snapshots are aligned such that orthologous genes match position. Note that the H3K27me3 level around *Zeus* is substantially higher in *D. melanogaster* relative to the equivalent region in *D. yakuba*. (*Right*) As above, but focusing instead on the parental gene *CAF40* (*Rcd-1*; red), flanked on the *left* by *e(y)3* (gray), and on the *right* by CG12237 (blue). Note that *CAF40* possesses little to no H3K27me3 signal, in strong contrast to *Zeus*. The sequence identity between *Zeus* and *CAF40* is 71% (Quezada-Díaz et al. 2010).

2009), might be able to completely change a novel gene's epigenetic regulation by delivering the novel gene to a location with an entirely different chromatin profile.

We examined whether the distribution of H3K27me3 changes varies by the location of the duplicated gene relative to its parent. We found that interchromosomal translocations show greater H3K27me3 divergence than intrachromosomal translocations (Fig. 6A) (Sim, $P < 10^{-5}$; Yak, $P < 10^{-5}$). As expected, interchromosomal translocations also show greater divergence in gene expression (Sim, $P = 0.017$; Yak, $P < 10^{-5}$). These results show that interchromosomal duplications drive greater change in H3K27me3 signal and concomitant divergence in gene expression.

The hypothesis that localization affects H3K27me3 divergence also predicts that intrachromosomal duplicates are more likely to undergo more extensive H3K27me3 evolution the further they are from the parent gene. Indeed, we see evidence within the intrachromosomal population that the distance between the old gene and the new gene is related to the subsequent H3K27me3 divergence (Fig. 6B). There is a significant correlation between distance from the parental gene and the magnitude of H3K27me3

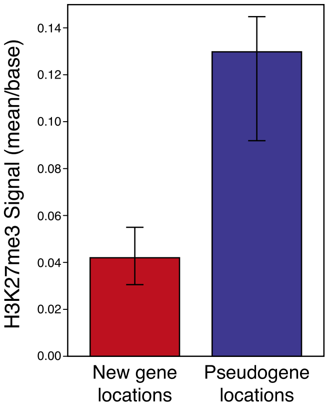

Figure 5. Locations into which duplicates move affects eventual pseudogenization fate. Regions in the *D. simulans* genome into which new genes and pseudogenes had moved in *D. melanogaster* show differences in mean H3K27me3 signal. On the *left*, functional new genes' duplication locations show significantly less H3K27me3 signal than pseudogenes' equivalent locations. Bars represent bootstrapped 95% confidence intervals. Note that for the purposes of this analysis, we assume that extant *D. simulans* H3K27me3 signal is representative of the H3K27me3 signal occurring in the ancestor of *D. melanogaster* in which new gene duplications occurred (see Methods).

divergence (Yak: SCC = 0.094; Permutation test, $P = 0.001$), but it is worth noting that this effect saturates after ~100 kb, suggesting that moving beyond that threshold does little to further change H3K27me3 regulation. This saturation effect coincides with a loss of H3K27me3 autocorrelation in approximately the same range (data not shown).

Discussion

Previous studies have discovered relatively slow evolution of histone modifications (Cain et al. 2011). Our results reinforce the slow evolution of a histone modification, but we find very different evolutionary regimes between single-copy orthologs and orthologs that have undergone duplication events. While single-copy orthologs exhibit relatively little divergence in terms of genic H3K27me3, duplicated genes evolve comparatively rapidly.

These differences are exaggerated in the most distant comparison we make, between *Drosophila melanogaster* and *Caenorhabditis elegans*. Orthologous genes in these species show substantial similarity in H3K27me3 signal, even though the two species are deeply diverged. Meanwhile, orthologous gene sets in which duplication events have occurred are significantly less conserved.

Among duplicated genes, we find interesting dynamics of H3K27me3 change. Gene duplicates show highly correlated gain or loss of H3K27me3 signal, indicating that both parent and duplicate undergo similar changes to their epigenetic status following duplication. Duplicated genes are biased with respect to the H3K27me3 change they are subject to: Both duplicates are more likely to lose methylation than single-copy orthologous genes. By examining the regions into which functional new genes localize relative to pseudogenes, we show that new genes tend to move to regions of low H3K27me3 signal. This result comports with the duplication–degeneration–complementation theory (Force et al. 1999), and suggests the importance of chromatin state as a determinant of novel gene fate. We speculate that the removal of H3K27me3 could allow paralogs to acquire new tissue specificity via *trans*-regulatory factors instead of histone modifications (Huerta-Cepas et al. 2011).

We find evidence that gene duplications on different chromosomes and at greater distances are more likely to acquire H3K27me3 signal and gene expression changes. The novel gene *Zeus* exemplifies this pattern, and has gained a substantial level of H3K27me3 relative to its parent gene *CAF40*. These results strongly imply that gene duplication mechanisms that can move the resulting duplicate further from the original gene are associated with more H3K27me3 evolution in the duplicate, which could have important consequences for understanding the eventual fate of gene duplicates. We do not know to what extent these differ-

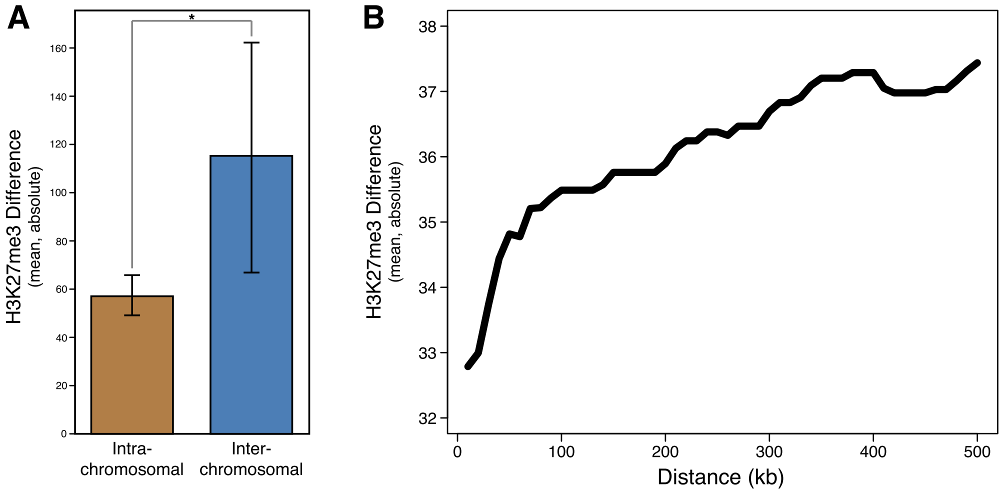

Figure 6. H3K27me3 changes in paralogs differ depending on duplicate location. (*A*) Within *Drosophila simulans*, interchromosomal translocations have significantly greater H3K27me3 changes (in terms of absolute value; permutation test, $P < 10^{-5}$). To accurately identify chromosomal translocations, we limited comparison to the major chromosome arms (2R, 2L, 3R, 3L, and X), and considered any gene that moved between chromosome arms an interchromosomal translocation. (*B*) On the same chromosome, duplicated gene orthologs' H3K27me3 changes increase with distance from the parent gene. On the *y*-axis, the cumulative mean change in H3K27me3 signal as a function of distance (on the *x*-axis). There is a significant correlation between distance and magnitude of H3K27me3 change (SCC = 0.103, $P = 0.002$).

ences would manifest for other histone marks, which tend to be relatively more compact than H3K27me3 domains (Kharchenko et al. 2011; Nègre et al. 2011).

Significant evolution of H3K27me3 signal was associated with evolution of gene expression and sequence evolution (as measured by d_N/d_S). In a separate comparison, independent of genes, a bin-based approach also found that evolution of H3K27me3 was also associated with an elevated rate of sequence substitution between species (see Supplemental Material; Supplemental Fig. S11). It appears that evolution of this mark is associated with evolution in the underlying sequence, although the causal basis of that association is ambiguous in the present study (Xiao et al. 2012). Whether H3K27me3 evolution allows the sequence to evolve, or whether H3K27me3 divergence is the result of new Polycomb regulation due to the creation or destruction of PcG binding sites is unknown.

To this end, future work is necessary to clarify exactly how the evolution of H3K27me3 is mediated genetically. Although many of the key enzymes responsible for deposition and maintenance of H3K27me3 are known, it is likely that we do not have a complete catalog of all the elements of all PREs (Delest et al. 2012), nor do we know the exact spatial or orientation requirements of the involved binding sites (if indeed there are any). However, the extreme conservation of H3K27me3 across distantly related species leads to the question of how a *cis*-regulatory element composed of relatively degenerate sequence motifs could persist functionally over such a long period of time (Ruvinsky and Ruvkun 2003; Fisher et al. 2006).

It is worth noting caveats that apply to the above analysis. Our results are gathered at only one strictly defined time period, which was chosen for its ease of use and developmental significance. Genes for which we observe no change in H3K27me3 signal may instead be highly diverged at other developmental times. Furthermore, it is possible that the dynamics of histone modification evolution differ depending on the stage of development (Kalinka et al. 2010). Additionally, we believe the observed differences in H3K27me3 reflect a lower bound estimate of the possible differences between species. Because of the heterogeneous nature of the sample material, it may be difficult to detect small-scale, tissue-specific differences in the epigenetic markings over loci. It may well be that orthologous genes undergo many changes of this nature between the species examined.

In summary, we have shown that H3K27me3 signal diverges more rapidly in duplicated genes. When diverged, duplicated genes most often show correlated loss of H3K27me3 relative to the single ortholog in the other genome, but the extent to which this mark changes depends on how far the duplicates are moved from the parent gene.

These results are indicative of an interesting dichotomy in the regulation of epigenetic states. Whereas single gene orthologs remain locked into their epigenetic status, duplicates are able to undergo rewiring of H3K27me3 signal and resulting gene expression. Our results indicate an interplay between gene duplication and the evolution of chromatin state as a mechanism for generating evolutionary novelty.

Methods

ChIP-seq

Strains were maintained at room temperature until collection at the white prepupa stage. ChIP-seq was carried out in duplicate using the standard modENCODE protocols. We mapped 36-bp reads to the appropriate reference genome using Bowtie (v.0.12.7) (Langmead et al. 2009). *Caenorhabditis elegans* H3K27me3 ChIP-seq data were generated as described in AP Boyle, CL Araya, C Brdlik, P Cayting, C Cheng, Y Cheng, K Gardner, L Hillier, J Janette, L Jiang, et al. (in prep.). Data from Nègre et al. (2011) are available under accession numbers GSE27111 and GSE23537.

RNA-seq

Single-end reads were trimmed and mapped using TopHat, and FPKM values were called with Cufflinks (Trapnell et al. 2012). *Caenorhabditis elegans* RNA-seq data were processed as described in Li et al. (2014).

Quantitative analysis

We used BEDTools (Quinlan and Hall 2010) to count the number of reads occurring in either (1) the exons of each gene as determined by the reference annotation, or (2) regular sliding windows 1 kb in width (Quinlan and Hall 2010). Genic H3K27me3 was normalized by the number of exons (see Supplemental Methods). We compared orthologous genes using the list of modENCODE orthologs, and we compared orthologous windows by using liftOver to map windows between species.

Pseudogene location analysis

In order to examine the H3K27me3 signal occurring in the locations of recently duplicated genes/pseudogenes, we extracted the two flanking regions (both 500 bp) around each recently duplicated gene or pseudogene, which were then lifted into the *D. simulans* genome using liftOver. To limit our analysis to very recent, confident duplication events, we only analyzed the comparison of *D. melanogaster–D. simulans*, and considered only cases in which both flanking regions mapped, with a distance of <5 kb between them. We examined H3K27me3 levels in the region between flanks as representative of the ancestral state (see Supplemental Methods).

Statistical analysis

To call individual intervals as significantly diverged in different species, we used DESeq (Anders and Huber 2010) on the control-subtracted read counts within bins. To estimate a false discovery rate (FDR), we used the *Q*-value package in R (Storey and Tibshirani 2003). To estimate the significance of differences between two populations' sample means, we used permutation tests as described in Sokal and Rohlf (1994) (p. 808; called therein "sampled randomization" tests). All statistical analysis was performed in R.

Data access

Drosophila RNA-seq data and ChIP-seq data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE49945. *Caenorhabditis* ChIP-seq data have been submitted to GEO under accession numbers GSE49724 and GSE49738.

Acknowledgments

The modENCODE Project was supported by NIH grant U01HG004264 awarded to K.P.W. R.K.A. was supported by an NSF

Graduate Research Fellowship and an NIH training grant (T32GM007197). We thank Jason Lieb and his laboratory for use of the *C. elegans* H3K27me3 ChIP-seq data. We are grateful to Benjamin Krinsky and Manyuan Long for interesting discussions on gene duplication. We thank Kacy Gordon, Aashish Jha, and Xiaochun Ni for helpful comments and critical review of our manuscript. We are indebted to many members of the White and Ruvinsky laboratories for useful criticism, wonderful discussions, and technical help. Finally, we appreciate the comments of three anonymous reviewers whose feedback greatly improved the paper.

References

Akkers RC, Van Heeringen SJ, Jacobi UG, Janssen-Megens EM, Françoijs K-J, Stunnenberg HG, Veenstra GJC. 2009. A hierarchy of H3K4me3 and H3K27me3 acquisition in spatial gene regulation in *Xenopus* embryos. *Dev Cell* **17:** 425–434.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11:** R106.

Bantignies F, Roure V, Comet I, Leblanc B, Schuettengruber B, Bonnet J, Tixier V, Mas A, Cavalli G. 2011. Polycomb-dependent regulatory contacts between distant Hox loci in *Drosophila*. *Cell* **144:** 214–226.

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, et al. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125:** 315–326.

Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, Seringhaus MR, Wang LY, Gerstein M, Snyder M. 2007. Divergence of transcription factor binding sites across related yeast species. *Science* **317:** 815–819.

Bradley RK, Li XY, Trapnell C, Davidson S, Pachter L, Chu HC, Tonkin LA, Biggin MD, Eisen MB. 2010. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related *Drosophila* species. *PLoS Biol* **8:** e1000343.

Cain CE, Blekhman R, Marioni JC, Gilad Y. 2011. Gene expression differences among primates are associated with changes in a histone epigenetic modification. *Genetics* **187:** 1225–1234.

Campos EI, Reinberg D. 2009. Histones: annotating chromatin. *Annu Rev Genet* **43:** 559–599.

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134:** 25–36.

Chen F, Mackey AJ, Vermunt JK, Roos DS. 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE* **2:** e383.

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:** 2798–2809.

Chi P, Allis CD, Wang GG. 2010. Covalent histone modifications–miswritten, misinterpreted and mis-erased in human cancers. *Nat Rev Cancer* **10:** 457–469.

Creevey CJ, Muller J, Doerks T, Thompson JD, Arendt D, Bork P. 2011. Identifying single copy orthologs in Metazoa. *PLoS Comput Biol* **7:** e1002269.

Delest A, Sexton T, Cavalli G. 2012. Polycomb: a paradigm for genome organization from one to three dimensions. *Curr Opin Cell Biol* **24:** 405–414.

Dowell RD. 2010. Transcription factor binding variation in the evolution of gene regulation. *Trends Genet* **26:** 468–475.

Ellinger J, Kahl P, Von der Gathen J, Heukamp LC, Gütgemann I, Walter B, Hofstädter F, Bastian PJ, Von Ruecker A, Müller SC, et al. 2012. Global histone H3K27 methylation levels are different in localized and metastatic prostate cancer. *Cancer Invest* **30:** 92–97.

Feng S, Jacobsen SE. 2011. Epigenetic modifications in plants: an evolutionary perspective. *Curr Opin Plant Biol* **14:** 179–186.

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:** 212–224.

Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS. 2006. Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science* **312:** 276–279.

Force A, Lynch M, Pickett FB, Amores A, Yan Y, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151:** 1531–1545.

Gu Z, Rifkin SA, White KP, Li W-H. 2004. Duplicate genes increase gene expression diversity within and between species. *Nat Genet* **36:** 577–579.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. *Nat Rev Genet* **10:** 551–564.

He Q, Bardet AF, Patton B, Purvis J, Johnston J, Paulson A, Gogol M, Stark A, Zeitlinger J. 2011. High conservation of transcription factor binding and evidence for combinatorial regulation across six *Drosophila* species. *Nat Genet* **43:** 414–420.

Huang W, Li L, Myers JR, Marth GT. 2012. ART: a next-generation sequencing read simulator. *Bioinformatics* **28:** 593–594.

Huerta-Cepas J, Dopazo J, Huynen MA, Gabaldón T. 2011. Evidence for short-time divergence and long-time conservation of tissue-specific expression after gene duplication. *Brief Bioinform* **12:** 442–448.

Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet* **10:** 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:** 811–814.

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:** 480–485.

King M, Wilson A. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188:** 107–116.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25.

Lanzuolo C, Orlando V. 2012. Memories from the polycomb group proteins. *Annu Rev Genet* **46:** 561–589.

Lenhard B, Sandelin A, Carninci P. 2012. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet* **13:** 233–245.

Li JJ, Huang H, Bickel PJ, Brenner SE. 2014. Comparison of *D. melanogaster* and *C. elegans* developmental stages, tissues, and cells by modENCODE RNA-seq data. *Genome Res* (this issue). doi: 10.1101/gr.170100.113.

Ludwig MZ, Palsson A, Alekseeva E, Bergman CM, Nathan J, Kreitman M. 2005. Functional evolution of a *cis*-regulatory module. *PLoS Biol* **3:** e93.

Nègre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, Kheradpour P, Eaton ML, Loriaux P, Sealfon R, et al. 2011. A *cis*-regulatory map of the *Drosophila* genome. *Nature* **471:** 527–531.

Nei M, Xu P, Glazko G. 2001. Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *Proc Natl Acad Sci* **98:** 2497–2502.

Ni X, Zhang YE, Nègre N, Chen S, Long M, White KP. 2012. Adaptive evolution and the birth of CTCF binding sites in the *Drosophila* genome. *PLoS Biol* **10:** e1001420.

Obbard DJ, Maclennan J, Kim K-W, Rambaut A, O'Grady PM, Jiggins FM. 2012. Estimating divergence dates and substitution rates in the *Drosophila* phylogeny. *Mol Biol Evol* **29:** 3459–3473.

Papp B, Müller J. 2006. Histone trimethylation and the maintenance of transcriptional ON and OFF states by trxG and PcG proteins. *Genes Dev* **20:** 2041–2054.

Quezada-Díaz JE, Muliyil T, Río J, Betrán E. 2010. *Drcd-1 related*: a positively selected spermatogenesis retrogene in *Drosophila*. *Genetica* **138:** 925–937.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842.

Rifkin SA, Kim J, White KP. 2003. Evolution of gene expression in the *Drosophila melanogaster* subgroup. *Nat Genet* **33:** 138–144.

Ruvinsky I, Ruvkun G. 2003. Functional tests of enhancer conservation between distantly related species. *Development* **130:** 5133–5142.

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. 2012. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* **148:** 458–472.

Shubin N, Tabin C, Carroll S. 2009. Deep homology and the origins of evolutionary novelty. *Nature* **457:** 818–823.

Shultzaberger RK, Maerkl SJ, Kirsch JF, Eisen MB. 2012. Probing the informational and regulatory plasticity of a transcription factor DNA-binding domain. *PLoS Genet* **8:** e1002614.

Sokal RR, Rohlf FJ. 1994. *Biometry: The principles and practices of statistics in biological research*. W.H. Freeman, New York.

Stern DL, Orgogozo V. 2008. The loci of evolution: how predictable is genetic evolution? *Evolution* **62:** 2155–2177.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci* **100:** 9440–9445.

Tamura K, Subramanian S, Kumar S. 2004. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Mol Biol Evol* **21:** 36–44.

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:** 88.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7:** 562–578.

Xiao S, Xie D, Cao X, Yu P, Xing X, Chen C-C, Musselman M, Xie M, West FD, Lewin HA, et al. 2012. Comparative epigenomic annotation of regulatory DNA. *Cell* **149:** 1381–1392.

Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. 2009. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* **25:** 1952–1958.

Zhou VW, Goren A, Bernstein BE. 2011. Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet* **12:** 7–18.

Received June 12, 2013; accepted in revised form January 17, 2014.