Correspondence: Yong E. Zhang and Wenwen Shi, State Key Laboratory of Integrated Management of Pest Insects and Rodents & Key Laboratory of the Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China. Tel: +86 10 64806339, +86 10 64806606; fax: +86 10 64806339; email: zhangyong@ioz.ac.cn, shiwenwen@ioz.ac.cn

†Daqi Yu and Wenwen Shi contributed equally to this research.

ORIGINAL ARTICLE

Underrepresentation of active histone modification marks in evolutionarily young genes

Daqi Yu†, Wenwen Shi† and Yong E. Zhang

This is an Accepted Article that has been peer-reviewed and approved for publication in the Insect Science but has yet to undergo copy-editing and proof correction. Please cite this article as doi: 10.1111/1744-7917.12299.

State Key Laboratory of Integrated Management of Pest Insects and Rodents & Key Laboratory of the Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China

Abstract

It is known that evolutionarily new genes can rapidly evolve essential roles in fundamental biological processes. Nevertheless, the underlying molecular mechanism of how they acquire their novel transcriptional pattern is less characterized except for the role of *cis*-regulatory evolution. Epigenetic modification offers an alternative potential possibility. Here, we examined how histone modifications have changed among different gene age groups in *Drosophila melanogaster* by integrative analyses of an updated new gene dataset and published epigenomic data. We found a robust pattern across various datasets where both the coverage and intensity of active histone modifications, Histone 3 lysine 4 trimethylation and lysine 36 trimethylation, increased with the evolutionary age. Such a temporal correlation is negative and much weaker for the repressive histone mark, lysine 9 trimethylation, which is expected given its major association with heterochromatin. By the further comparison with neighboring old genes, the depletion of active marks of new genes could be only partially explained by the local epigenetic context. All these data are consistent with the observation

that older genes bear relatively higher expression level and suggest that the evolution of histone modifications could be implicated in transcriptional evolution after gene birth.

Key words *Drosophila*; H3K4me3; H3K36me3; H3K9me3; epigenetic evolution; new gene evolution

Introduction

New gene refers to a novel genetic locus, a physically distinct and derived segment of DNA, which pops out throughout the whole history of life (Long *et al.*, 2003). Emerging evidence from both animals and plants has shown that new genes play diverse functional roles in development or reproduction (Long *et al.*, 2003; Chen *et al.*, 2013). These genetic novelties usually originated either via duplication of preexisting genes or *de novo* from non-coding DNAs (Kaessmann, 2010; Chen *et al.*, 2013; Andersson *et al.*, 2015). Possibly because duplication may lead to the incomplete inclusion of the original regulatory sequences and *de novo* origination may be associated with suboptimal regulatory elements, younger genes tend to be expressed in a more tissue-specific manner compared to older genes (Zhang *et al.*, 2012; Schlotterer, 2015). For example, human-specific new genes are often only transcribed in two or three tissues with moderate abundance while genes predating vertebrate split are highly transcribed in more than 20 tissues (Zhang *et al.*, 2012).

The transcriptional enhancement demonstrated by the increase of expressional breadth and abundance suggests that the underlying regulatory changes are associated with the age of genes. Accordingly, numerous studies have uncovered that the evolution of *cis*-regulatory elements in new genes, especially promoters, often contributes to the adjustments of expression patterns by co-opting or modifying preexisting sequences (Zaiss & Kloetzel, 1999; Fablet *et al.*, 2009; Xie *et al.*, 2012; Wu & Sharp, 2013; Sorourian *et al.*, 2014; Ruiz-

Orera *et al.*, 2015). By contrast, although epigenetic mechanisms are also critical to gene expression regulation (Kouzarides, 2007; Li *et al.*, 2007; Brown & Bachtrog, 2014), little is known about their roles in expressional evolution of new genes. It was not until very recently that two important epigenetic mechanisms, *i.e.*, DNA methylation and histone modifications, were found to be implicated in the transcriptional divergence of duplicated new genes and their parental copies (Arthur *et al.*, 2014; Keller & Yi, 2014; Wang *et al.*, 2014).

Herein this study, we updated the age-dating data of *Drosophila melanogaster* (*D. melanogaster*) which were generated previously (Zhang *et al.*, 2010) and examined how epigenetic marks evolve with the increase of gene ages. Since the general functionality of DNA methylation in *Drosophila* remains controversial (Lyko *et al.*, 2000; Raddatz *et al.*, 2013; Zhang *et al.*, 2015), we focused on histone modifications. Given the availability of public data, we analyzed three widely studied marks: tri-methylation of lysine 4, 9 and 36 of Histone 3, *i.e.* H3K4me3, H3K9me3 and H3K36me3, respectively. H3K4me3 is predominantly associated with locus activation and enriched around the transcription start site (TSS) (Greer & Shi, 2012; Taniguchi & Moore, 2014). Similarly, H3K36me3 is also associated with active gene transcription and mainly deposited on the gene body region (Wagner & Carpenter, 2012; Pu *et al.*, 2015). Reversely, H3K9me3 is mainly involved in the formation of large-scale heterochromatin domain, but also appears to be capable of silencing euchromatic genes (Ebert *et al.*, 2006; Kouzarides, 2007).

Because older genes appear to be widely and abundantly transcribed compared to younger genes (Zhang *et al.*, 2012; Schlotterer, 2015), it is logically expected to observe the overrepresentation of active histone marks and underrepresentation of repressive histone marks in the former group relative to the latter group. In order to test this hypothesis, we analyzed genome-wide histone modification data of *D. melanogaster* head considering the data availability. We first confirmed that in this specific organ older genes show higher expression level and genes with higher expression level are associated with significantly more active marks (H3K4me3 and H3K36me3) and slightly less repressive mark (H3K9me3). We then discovered that H3K4me3 and H3K36me3 marks indeed show more enrichment in older genes relative to younger genes in terms of both binding intensity and binding coverage. Moreover, the increase of active marks present a significant linear correlation with the gene age. Similarly, the decrease of repressive chromatin mark H3K9me3 shows a weaker temporal pattern compared to active marks, possibly because most of H3K9me3 marks localize in the constitutive heterochromatin region (Ebert *et al.*, 2004). Additional analyses in larva revealed exactly the same pattern suggesting the wide applicability of this age-associated pattern. Thus, our findings suggest that histone modification may be also subject to fine-tuning by recruiting more and more active histone marks after the birth of new genes.

Materials and methods

Gene age dating

We followed our previous pipeline and dated *D. melanogaster* coding genes along the *Drosophila* genus tree (Zhang *et al.*, 2010). In brief, for each gene annotated in Ensembl version 78 (Flicek *et al.*, 2014), we took advantage of UCSC whole genome syntenic alignment and inferred the phylogenetic distribution of its orthologs. Then, based on the parsimony rule, we made an inference on when this gene originated. However, since the genome-wide synteny may degenerate for species outside of the *Drosophila* genus due to the vast divergence, we can only trace gene origination to the common ancestor of *Drosophila* genus 63 million years ago (Hedges *et al.*, 2006).

In order to cover an even longer evolutionary period, we followed the widely used phylostratigraphy strategy (Tautz & Domazet-Loso, 2011; Yang *et al.*, 2015), which only relied on homology of a single gene rather than a long synteny. We classified genes shared by *Drosophila* genus into five age groups based on the homolog distribution information in the following outgroup species including *Anopheles gambiae*, *Aedes aegypti*, *Bombyx mori*, *Tribolium castaneum* and *Apis mellifera*. We chose these insects due to the corresponding taxonomy breadth and relatively better genome assembly quality. Homolog information is directly retrieved from Ensembl Metazoa release 26 (Flicek *et al.*, 2014). A gene is deemed to

belong to a specific age branch by following the parsimony rule implemented in the phlostratigraphy strategy (Tautz & Domazet-Loso, 2011; Yang *et al.*, 2015).

Chromatin Immunoprecipitation followed by sequencing (ChIP-seq) data analyses

We retrieved the *D. melanogaster* ChIP-seq data from modENCODE project (Celniker *et al.*, 2009). We focused on samples from heads of both male and female *D. melanogaster* adults and examined three types of histone modifications, namely H3K4me3, H3K36me3 and H3K9me3. The corresponding modENCODE dataset IDs are 5098, 5091 and 4933, respectively. The reason that we focused on head samples only is simply due to the data availability across different profiling platform (see next section). To further test whether the pattern we discovered is robust across different samples, we also analyzed whole body data of *D. melanogaster* larvae 3rd instar. The modENCODE dataset IDs are 5096, 4950 and 4939, corresponding to H3K4me3, H3K36me3 and H3K9me3, respectively.

By following the previous practice (Karlic *et al.*, 2010), we used the well-developed strategy in RNA-sequencing (RNA-seq) quantification field in order to generate a gene level statistic of histone modification enrichment based on ChIP-seq data. First, we built a feature annotation file (gtf) as routinely performed in RNA-seq. Herein, we defined a gene region as the gene body and the corresponding flanking 1 kb region accounting for the uncertainty in the annotation of untranslated regions, a TSS region as the upstream 1 kb and downstream 1

kb around the TSS, and all the remaining regions as the non-overlapping background (Fig. S1). The reason that we extended the TSS region with 1 kb plus is: the H3K4me3 signal or the tag density tends to reach a sharp peak around the TSS and then descends down until at least 1 kb (Barski *et al.*, 2007; Cheng & Gerstein, 2012). For genes with more than one TSS, the 2 kb windows are merged if these TSS are not far from each other with the distance smaller than 2 kb; otherwise, the final quantification is averaged across different TSS. We constructed a pseudo gene annotation gtf file by including all the gene regions and remaining intervals. Analogously, we built a pseudo TSS annotation file. In the case of histone modifications distributing broadly such as H3K36me3 and H3K9me3, the gene gtf file is used in the subsequent analyses. By contrast, for H3K4me3 histone modification which is prominently concentrated around the TSS (Barski *et al.*, 2007), the TSS gtf file is taken.

In order to differentiate reads derived from paralogous duplicates which may share high sequence similarity, we then implemented the Kallisto package (Bray *et al.*, 2015). Given the gene or TSS gtf files, we extracted the corresponding sequences from the latest Flybase dm6 genome assembly by *gffread* command wrapped in Cufflinks (Trapnell *et al.*, 2010), which were further indexed and quantified by Kallisto. We trimmed raw reads with Trimmomatic with the remaining fragment not less than 32 basepairs (bp) (Bolger *et al.*, 2014). We mapped all trimmed reads to the genome via Novoalign and randomly assigned multi-mapping reads (reads mapping equally well to more than one genes) with the '-r Random' parameter (http://www.novocraft.com). By comparing all mapped reads with the previously built index,

Kallisto reassigned multi-mapping reads via a statistical expectation-maximization procedure based on both unique-mapping reads and multi-mapping reads onto the same feature (gene or TSS) and then inferred the raw read count for each feature.

ChIP-seq is different from RNA-seq where the background signal quantified by the control or mock experiment should be subtracted from the case experiment. So, we normalized the case and the control data by calculating the scaling index with the NCIS package after transforming the alignment files (in SAM format) into genomic interval files (in BED format) with samtools and bedtools (Li *et al.*, 2009; Quinlan & Hall, 2010; Liang & Keles, 2012). In case the mock signal is stronger than the case for a gene or TSS of interest, the read count is reset as 0. After normalization and subtraction, we calculated $\log_2$basedFPKM (Fragments Per Kilobase of gene or TSS per Million mapped reads) after adding a tiny offset 0.01 to compensate zero values. So, the minimum expression will be $\log_2(0.01)$ or -6.7 .

Across three type of histone modification data, we divided each age group into three equal-size classes based on $\log_2$FPKM, *i.e.*, 'bottom', 'middle' and 'top'. For each class and each type of histone modification, we calculated the Spearman's σ and *P* between the median binding intensity and age class, which is followed by Bonferroni correction.

To compare new genes with the nearest old genes, we divided all genes into the old and new gene group including entries predating *Drosophila* genus split (branch -4 to branch 0) and postdating *Drosophila* genus split (branch 1 to branch 6), respectively. For every focal

new gene, we treat its TSS as an anchor and search for a closest TSS that is associated with an old gene. Some genes encode multiple TSS and we chose the shortest TSS pair.

RNA-seq data analyses

We retrieved *D. melanogaster* adult female and male head RNA-seq data from modENCODE project with IDs of 3083 and 3084 (Celniker *et al.*, 2009). Similarly, we trimmed raw reads with Trimmomatic (Bolger *et al.*, 2014). We then took the widely used pipeline: to map reads against the genome via HISAT and to generate gene level quantification via cufflinks (Trapnell *et al.*, 2010; Kim *et al.*, 2015). To be comparable with the ChIP-seq data of modENCODE, which is from heads of a mixture of male and female adults, we took the average of the individual FPKM values of male and female heads. Finally, we transferred the raw value to $\log_2$FPKM after adding an offset 0.01.

Chromatin immunoprecipitation (ChIP) coupled with tiling microarrays (ChIP-chip) data analyses

ChIP-chip offers at least one magnitude lower dynamic range compared to ChIP-seq because of the inherent difference between the array-based and sequencing-based platforms (Marioni *et al.*, 2008), it is thus less capable of binding intensity quantification. Considering

such a shortcoming, we took advantage of ChIP-chip datasets to analyze a different feature, *i.e.*, histone binding coverage across genes. Specifically, we obtained ChIP-chip raw data from NCBI GEO database with the accession number GSE22438 (Edgar *et al.*, 2002; Wood *et al.*, 2010). Samples are from heads of female *D. melanogaster* adults, which are 10 days and 40 days old, respectively. Since the pattern is essentially the same in either 10-day or 40-day heads, we only presented the result on the basis of 10-day heads in the main text. Raw case and control CEL files are loaded into Starr package in Bioconductor (Gentleman *et al.*, 2004; Zacher *et al.*, 2010). Probe logarithmic densities are normalized with loess method by *getRatio* command and exported by *writewig* command (Zacher *et al.*, 2011). Probe sequences are extracted through *featureData* command in Starr package. All 25 bp probes are aligned to the dm6 assembly with bowtie and no mismatch is allowed (Langmead, 2010). Only uniquely mapped probes are reserved for downstream analyses in order to differentiate paralogous duplicates. Peak calling procedures are conducted with Chipotle by changing the step size from 250 bp to 100 bp to achieve a higher sensitivity (Buck *et al.*, 2005).

Similar to ChIP-seq data analyses, we next performed gene level analyses. Given the probe length of 25 bp and the spacer length of 15 bp on the tiling array, genes with unique probe coverage less than 80% (20 probes per kb) were filtered out. For the remaining genes, we defined a parameter called peak coverage to quantify the extent of binding which is simply the overall percentage of sequence of interest covered by ChIP-chip binding peaks. Herein, the target region or 'sequence of interest' is defined as we did in ChIP-seq data analyses: 2 kb

windows around TSS covered by peaks for H3K4me3 marks and gene body with flanking 1 kb for H3K36me3 and H3K9me3 (Fig. S1).

To get a more comprehensive view of binding coverage changes across different age groups, we further divided all genes into four or five classes based on the genome-wide distribution of the peak coverage. In the case of H3K4me3 or H3K36me3, we first specified two classes, *i.e.*, '0' and '100%' because a significant proportion of genes are not bound or completely bound, respectively (Fig. S2). The remaining genes are equally divided into three classes, *i.e.*, 'low', 'moderate' and 'high'. By contrast, only a few genes are completely bound by H3K9me3 peaks (Fig. S2). Thus, we only specified the '0' class followed by dividing the remaining genes into three equal-size classes ('low', 'moderate' and 'high'). For each class, we calculated Spearman's σ to quantify its relationship with gene age. Bonferroni correction is added to account for the multiple testing issue.

Results

Young genes are lowly transcribed and histone modification are correlated with transcription for fly head

We followed our previous methodology and dated Flybase annotated protein-coding genes along the insect phylogenetic tree (Zhang *et al.*, 2010; Yang *et al.*, 2015) (Fig. 1; Materials and methods). Considering the data availability, we took *Drosophila* head as a major model

system to examine the correlation between transcriptional level and histone modification level across different age groups. However, although it is known that older genes tend to show higher expression level than younger genes (Zhang *et al.*, 2012; Schlotterer, 2015) and histone modification is correlated with transcriptional regulation (Martin & Zhang, 2005; Li *et al.*, 2007), these two patterns may not always hold for any samples (*e.g. Drosophila* head). In order to lay a solid foundation for subsequent analyses, we first tested these assumptions.

We reanalyzed both histone modification data (ChIP-seq) and transcriptome data (RNA-seq) generated by the modENCODE project (Celniker *et al.*, 2009). As introduced previously, we focused on two active modification marks, H3K4me3 and H3K36me3, and one repressive mark, H3K9me3. To serve as an overall statistic of histone modification enrichment on TSS (for H3K4me3) or gene body (for H3K36me3 or H3K9me3), we calculated $\log_2$FPKM (Fragments Per Kilobase of gene or TSS per Million mapped reads) after accounting for sequence mapping uncertainty between recently duplicated paralogous genes (Materials and Methods). In parallel, we quantified gene-level transcriptional intensity as $\log_2$FPKM too (Materials and Methods). As expected, consistent with previous reports (Zhang *et al.*, 2012; Schlotterer, 2015), older genes show higher transcriptional level compared to young genes with the median transcriptional level significantly correlated with the gene age (Spearman's σ = -0.95, $P < 2.2 \times 10^{-16}$, Fig. 2A). Moreover, histone modification intensity is indeed significantly correlated with transcriptional level across all three marks ($P < 2.2 \times 10^{-16}$, Fig. 2B–D). Notably, the Spearman's σ is lowest in case of H3K9me3 (|-0.26|) compared to

H3K4me3 (0.69) and H3K36me3 (0.47) suggesting the limited gene-level regulatory role of this repressive mark, which is consistent with its major role in heterochromatin (Ebert *et al.*, 2004; Ebert *et al.*, 2006).

Older genes show higher binding intensity of active marks relative to younger genes

Then, we examined how binding intensity evolves by comparing different age groups. In order to reach a higher sensitivity over different binding intensity ranges, we divided genes in each age group into three classes of equal sizes, *i.e.*, 'top', 'middle' and 'low' binding classes. Consistent with the trend that older genes bear broader and stronger expression relative to younger genes (Fig. 2), the active marks including H3K4me3 and H3K36me3 are gradually overrepresented with the increase of gene age, while the repressive mark, H3K9me3 is gradually underrepresented in this process (Fig. 3). The trend is roughly the same across top, middle and bottom binding intensity categories although the last category could not reach the statistical significance because majority of genes in many age groups are not bound at all. Moreover, active marks show strong correlation between median intensity and evolutionary age in top and middle binding categories ($|\sigma| > 0.8$, $P < 0.01$) while H3K9me3 shows relatively weaker correlation in the corresponding categories ($0.3 < \sigma < 0.8$, $P > 0.01$, Fig. 3). Such a difference is consistent with the pattern that H3K9me3 binding intensity is less correlated with transcriptional level (Fig. 2).

Epigenetic context partially predicts the histone modification levels of new genes

Next we are wondering whether the underrepresentation of active histone marks of new genes is actually a characteristic of new genes themselves or a characteristic of the chromatin environment in the region where new genes originate. To address this issue, we examined histone marks of the nearest old genes (branch -4~0) relative to the focal new genes (branch 1~6) (Materials and methods).

we found that the binding intensity of new genes tend to be positively correlated with the neighboring old genes across all three types of histone marks and the correlation is statistically significant for multiple age groups (Spearman's $P < 0.05$, Bonferroni correction; Fig. 4A). However, such a correlation does not necessarily mean the aforementioned histone modification pattern of new genes is fully explained by the local epigenetic context. Actually, we directly performed a pairwise comparison between new genes and nearest old genes and found that new genes are statistically significantly less bound by both H3K4me3 and H3K36me3 marks across all age groups (Fig. 4B). Even for the repressive histone mark, H3K9me3, which are less correlated with the expression level (Fig. 2), new genes tend to be bound more compared to nearby old genes (Fig. 4B).

Overrepresentation of active marks in older genes also occurs with respect to the binding coverage

So far, our ChIP-seq based analyses demonstrate that the intensity of histone modification marks is positively correlated with gene age, *i.e.*, younger genes are often significantly depleted with active histone modification marks, which could not be fully explained by the local epigenetic context. However, it remains unknown how the binding coverage changes during evolution. More than that, a pattern discovered based a single platform, *i.e.*, ChIP-seq, may be generated due to some technical artefacts. To address these two issues, we further analyzed ChIP-chip datasets profiling heads of *D. melanogaster* female adults (Wood *et al.*, 2010). After removing multi-mapping probes, we called binding peaks and calculated the binding coverage or peak coverage, which is defined as the overall percentage of sequences of interest covered by all peaks (Materials and Methods). Based on such a parameter, we examined how the binding coverage differs across different gene age groups.

Consistent with the age-associated temporal pattern of binding intensity, older genes tend to be bound by more H3K4me3 active marks (Fig. 5A). Actually, except for the low and moderate peak coverage groups, the proportion of all other three categories are statistically correlated with the evolutionary age (|Spearman's $\sigma| \geq 0.81$, $P < 0.05$, Table 1). Exactly like the case of H3K4me3, the H3K36me3 active marks show the pattern where older genes are more extensively modified (Table 1, Fig. S3). Thus, older genes are associated with various

types of active marks across larger percentage of TSS or gene bodies compared to younger genes.

In contrast to the active marks H3K4me3 and H3K36me3 but consistent with the result of binding intensity analyses (Fig. 3), the repressive mark H3K9me3 only shows limited correlation with gene age. Actually, only the moderate group declines with the increase of gene age (Spearman's σ = -0.84, P = 0.018, Table 1; Fig. 5B), while all the other three categories appear irrelevant with gene age (Table 1). Such a pattern is again consistent with low correlation between H3K9me3 binding intensity and gene expression (Fig. 2).

Discussion

To be preserved in the genome, new genes especially incompletely duplicated new genes or *de novo* new genes must acquire their own expression patterns by either co-opting preexisting regulatory context or evolving such context *de novo*. From this perspective, how *cis*-regulatory regions of new genes evolve has been attracting wide interest for more than one decade (Zaiss & Kloetzel, 1999; Fablet *et al.*, 2009; Xie *et al.*, 2012; Wu & Sharp, 2013; Sorourian *et al.*, 2014; Ruiz-Orera *et al.*, 2015). However, as introduced previously, although epigenetic modifications are also important in regulating gene expression and are well studied in various biological processes, it is rarely explicitly studied in the new gene origination process. Nevertheless, some exciting results just emerged as demonstrated in a few very

recent studies on how epigenetic evolution accompanied transcriptional divergence of duplicated genes (Arthur *et al.*, 2014; Keller & Yi, 2014; Wang *et al.*, 2014). Different from these pioneering works, we directly examined how various histone modifications evolve dynamically across hundreds of million years by classifying genes into more than 10 different age groups (Fig. 1).

Across ChIP-seq and ChIP-chip platforms, we discovered analogous patterns where both the coverage and intensity of active histone modifications, H3K4me3 and H3K36me3, are positively correlated with gene age (Figs. 3, 5). By contrast, there is a relatively weaker trend that H3K9me3 marks decline in the older gene groups compared to the young gene groups. Actually, since it is uncertain whether H3K9me3 preferentially binds gene body or TSS, we performed a TSS-centric analysis for H3K9me3 as did for H3K4me3 and rediscovered an analogous weak age-associated pattern (Fig. S4). Such a contrast between active and repressive marks is consistent with the previous reports where H3K4me3 and H3K36me3 are mainly associated with genic regions while H3K9me3 tends to be deposited to constitutive heterochromatins (Ebert *et al.*, 2006). More importantly, although our analyses are mainly based on 10-day *Drosophila* heads, the age-associated pattern is reproducible in both 40-day *Drosophila* heads (Figs. S3, S5) and larvae whole body (Fig. S6), suggesting that the pattern is widely applicable. Thus, these serials of data strongly support our initial hypothesis which proposes the evolution of histone modification accompanies expressional evolution of new genes (Zhang *et al.*, 2012).

Further analysis showed that histone modification level of young genes was positively correlated with their nearest old genes (Fig. 4A), suggesting that new genes' histone modification was affected by their chromatin environments. However, the histone modification level is significantly different between new genes and their nearest old genes (Fig. 4B), demonstrating that the age-associated pattern of new genes is not only a characteristic of their chromatin environment but also of themselves. Actually, RNA-based duplicated new genes or retrogenes always lost their preexisting regulatory sequence and they can be fused with the host gene in the insertion site as a new chimeric gene (Vinckenbosch *et al.*, 2006). *De novo* originated new genes generally overlap with a preexisting gene or hitchhike a bi-directional promoter (Xie *et al.*, 2012). As for DNA-based duplicated new genes, majority of them are linked in tandem (Zhou *et al.*, 2008), which often leads to the dosage gain effect (Chang & Liao, 2012). Thus, across all these three major mechanisms of new gene origination, new genes can co-opt preexisting regulatory context and thus show some similarity with the neighboring old gene with respect to histone modification binding. However, similar to retrogene and *de novo* gene, DNA-based duplicated gene generated by partial duplication may also have incomplete promoter region, which needs to be evolved. Thus, the regulation of new genes could be generally suboptimal, which causes the overall underrepresentation of active marks. In this regard, either duplicated new genes or *de novo* new genes are subject to subsequent epigenetic evolution, which leads to the age-dependent pattern discovered here.

When we established the pattern, we carefully handled the multi-mapping issues, *i.e.*, read or probe mapping uncertainty due to the sequence similarity between young duplicates. In ChIP-seq analyses we made a quantification via expectation-maximization based on all reads derived from the same gene or the same TSS; while in ChIP-chip analyses we simply excluded genes with at least 20% regions covered by multi-mapping probes (Materials and methods). These two different strategies revealed qualitatively similar pattern (Figs. 3, 5) suggesting the robustness of our result. More than that, the multi-mapping issue mainly exists for the youngest age group (5/6) enriched with recent duplicates (Yang *et al.*, 2015). Even if we removed this age group, the pattern will largely hold.

Such a robust age-associated pattern suggests that new genes may gradually accumulate active histone modifications through sequence alterations after their birth. However, this pattern could be also explained by preferential loss of young genes with relatively less active marks. These genes may be only lowly or narrowly transcribed and thus have a low pleiotropy, which means a relatively higher tolerance of loss-of-function (LoF) mutations (Yang *et al.*, 2015). This alternative hypothesis would predict roughly constant upper-bound or maximum binding intensity across different age groups, which is incompatible with the data (Figs. 3, S6). Thus, biased retention of new genes with active marks could not serve as a major mechanism to create the age-associated pattern observed in our data.

If genes gradually adjust their histone marks with increase of ages, the next question would be whether this epigenetic rewiring directly contributes to transcriptional evolution. Although evolution of *cis*-regulatory elements appear to directly lead to transcriptional enhancement (Zaiss & Kloetzel, 1999; Fablet *et al.*, 2009; Xie *et al.*, 2012; Wu & Sharp, 2013; Sorourian *et al.*, 2014; Ruiz-Orera *et al.*, 2015), our correlation-based analyses could not establish a direct causality between epigenetic modification evolution and transcriptional evolution of new genes. Although the significance of epigenetic mechanism for gene regulation is widely appreciated (Kouzarides, 2007; Li *et al.*, 2007; Brown & Bachtrog, 2014), it has been argued that recruitment of histone modification marks are merely a downstream response of transcription factor binding (Zhou *et al.*, 2014). In other words, changes of histone modification binding could be viewed as a passive layer of gene regulation. From this aspect, it is more likely that new genes evolve *cis*-regulatory regions first and then recruit active histone modification marks to further enhance their transcription. Nevertheless, regardless of the correlation and causality, our evolutionary and functional genomic analyses unambiguously support that epigenetic context of genes continuously remodels during their evolutionarily aging process. Such a change may be directly or indirectly involved in the transcriptional difference between old and young genes.

Acknowledgments

We thank the modENCODE project for releasing the *D. melanogaster* ChIP-seq and RNA-seq data. We thank Drs. De-Xing Zhang, Bin He and Chuan-Yun Li for insightful discussions. Computing was supported by the HPC Platform, Scientific Information Center, Institute of Zoology, Chinese Academy of Sciences. This research is jointly supported by grants from the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB13000000) and China Postdoctoral Science Foundation.

Disclosure

The authors declare that there are no potential conflicts of interest, including specific financial interests and relationships and affiliations relevant to the subject of this manuscript.

References

Andersson, D.I., Jerlstrom-Hultqvist, J. and Nasvall, J. (2015) Evolution of new functions *de novo* and from preexisting genes. *Cold Spring Harbor Perspectives in Biology*, 7

Arthur, R.K., Ma, L., Slattery, M., Spokony, R.F., Ostapenko, A., Negre, N. and White, K.P. (2014) Evolution of H3K27me3-marked chromatin is linked to gene expression

evolution and to patterns of gene duplication and diversification. *Genome Research*, 24, 1115–1124.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, 129, 823–837.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30, 2114–2120.

Bray, N., Pimentel, H., Melsted, P. and Pachter, L. (2015) Near-optimal RNA-Seq quantification. *arXiv preprint*, 1505.02710.

Brown, E.J. and Bachtrog, D. (2014) The chromatin landscape of *Drosophila*: comparisons between species, sexes, and chromosomes. *Genome Research*, 24, 1125–1137.

Buck, M.J., Nobel, A.B. and Lieb, J.D. (2005) ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. *Genome Biology*, 6, R97.

Celniker, S.E., Dillon, L.A., Gerstein, M.B., Gunsalus, K.C., Henikoff, S., Karpen, G.H., Kellis, M., Lai, E.C., Lieb, J.D., MacAlpine, D.M., Micklem, G., Piano, F., Snyder, M., Stein, L., White, K.P., Waterston, R.H. and mod, E.C. (2009) Unlocking the secrets of the genome. *Nature*, 459, 927–930.

435 Chang, A.Y. and Liao, B.Y. (2012) DNA methylation rebalances gene dosage after
436 mammalian gene duplications. *Molecular Biology and Evolution*, 29, 133–144.

437 Chen, S., Krinsky, B.H. and Long, M. (2013) New genes as drivers of phenotypic evolution.
438 *Nature Reviews Genetics*, 14, 645–660.

439 Cheng, C. and Gerstein, M. (2012) Modeling the relative relationship of transcription factor
440 binding and histone modifications to gene expression levels in mouse embryonic stem
441 cells. *Nucleic Acids Research*, 40, 553–568.

442 Ebert, A., Lein, S., Schotta, G. and Reuter, G. (2006) Histone modification and the control of
443 heterochromatic gene silencing in *Drosophila*. *Chromosome Research*, 14, 377–392.

444 Ebert, A., Schotta, G., Lein, S., Kubicek, S., Krauss, V., Jenuwein, T. and Reuter, G. (2004)
445 Su (var) genes regulate the balance between euchromatin and heterochromatin in
446 *Drosophila*. *Genes & Development*, 18, 2973–2983.

447 Edgar, R., Domrachev, M. and Lash, A.E. (2002) Gene Expression Omnibus: NCBI gene
448 expression and hybridization array data repository. *Nucleic Acids Research*, 30, 207–
449 210.

450 Fablet, M., Bueno, M., Potrzebowski, L. and Kaessmann, H. (2009) Evolutionary origin and
451 functions of retrogene introns. *Molecular Biology and Evolution*, 26, 2147–2156.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Giron, C.G., Gordon, L., Hourlier, T., Hunt, S., Johnson, N., Juettemann, T., Kahari, A.K., Keenan, S., Kulesha, E., Martin, F.J., Maurel, T., McLaren, W.M., Murphy, D.N., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H.S., Ruffier, M., Sheppard, D., Taylor, K., Thormann, A., Trevanion, S.J., Vullo, A., Wilder, S.P., Wilson, M., Zadissa, A., Aken, B.L., Birney, E., Cunningham, F., Harrow, J., Herrero, J., Hubbard, T.J., Kinsella, R., Muffato, M., Parker, A., Spudich, G., Yates, A., Zerbino, D.R. and Searle, S.M. (2014) Ensembl 2014. *Nucleic Acids Research*, 42, D749–D755.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y.C., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y.H. and Zhang, J.H. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5, R80.

Greer, E.L. and Shi, Y. (2012) Histone methylation: a dynamic mark in health, disease and inheritance. *Nature Reviews Genetics*, 13, 343–357.

Hedges, S.B., Dudley, J. and Kumar, S. (2006) TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, 22, 2971–2972.

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

Karlic, R., Chung, H.-R., Lasserre, J., Vlahovicek, K. and Vingron, M. (2010) Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 107, 2926–2931.

Keller, T.E. and Yi, S.V. (2014) DNA methylation and evolution of duplicate genes. *Proceedings of the National Academy of Sciences of the United States of America*, 111, 5932–5937.

Kim, D., Langmead, B. and Salzberg, S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*, 12, 357–360.

Kouzarides, T. (2007) Chromatin modifications and their function. *Cell*, 128, 693–705.

Langmead, B. (2010) Aligning short sequencing reads with Bowtie. *Current Protocols in Bioinformatics*, 32, 11.

Li, B., Carey, M. and Workman, J.L. (2007) The role of chromatin during transcription. *Cell*, 128, 707–719.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25, 2078–2079.

489 Liang, K. and Keles, S. (2012) Normalization of ChIP-seq data with control. *BMC*
490 *Bioinformatics*, 13, 199.

491 Long, M., Betran, E., Thornton, K. and Wang, W. (2003) The origin of new genes: glimpses
492 from the young and old. *Nature Reviews Genetics*, 4, 865–875.

493 Lyko, F., Ramsahoye, B.H. and Jaenisch, R. (2000) DNA methylation in *Drosophila*
494 *melanogaster*. *Nature*, 408, 538–540.

495 Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. and Gilad, Y. (2008) RNA-seq: An
496 assessment of technical reproducibility and comparison with gene expression arrays.
497 *Genome Research*, 18, 1509–1517.

498 Martin, C. and Zhang, Y. (2005) The diverse functions of histone lysine methylation. *Nature*
499 *Reviews Molecular Cell Biology*, 6, 838–849.

500 Pu, M., Ni, Z., Wang, M., Wang, X., Wood, J.G., Helfand, S.L., Yu, H. and Lee, S.S. (2015)
501 Trimethylation of Lys36 on H3 restricts gene expression change during aging and
502 impacts life span. *Genes & Development*, 29, 718–731.

503 Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing
504 genomic features. *Bioinformatics*, 26, 841–842.

505 Raddatz, G., Guzzardo, P.M., Olova, N., Fantappié, M.R., Rampp, M., Schaefer, M., Reik,
506 W., Hannon, G.J. and Lyko, F. (2013) Dnmt2-dependent methylomes lack defined

DNA methylation patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 8627–8631.

Ruiz-Orera, J., Hernandez-Rodriguez, J., Chiva, C., Sabidó, E., Kondova, I., Bontrop, R., Marqués-Bonet, T. and Albà, M. (2015) Origins of *de novo* genes in human and chimpanzee. *arXiv preprint*, 1507.07744.

Schlotterer, C. (2015) Genes from scratch–the evolutionary fate of *de novo* genes. *Trends in Genetics*, 31, 215–219.

Sorourian, M., Kunte, M.M., Domingues, S., Gallach, M., Oezdil, F., Rio, J. and Betran, E. (2014) Relocation facilitates the acquisition of short cis-regulatory regions that drive the expression of retrogenes during spermatogenesis in *Drosophila*. *Molecular Biology and Evolution*, 31, 2170–2180.

Taniguchi, H. and Moore, A.W. (2014) Chromatin regulators in neurodevelopment and disease: Analysis of fly neural circuits provides insights. *BioEssays*, 36, 872–883.

Tautz, D. and Domazet-Loso, T. (2011) The evolutionary origin of orphan genes. *Nature Reviews Genetics*, 12, 692–702.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. and Pachter, L. (2010) Transcript assembly and quantification by

RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28, 511–515.

Vinckenbosch, N., Dupanloup, I. and Kaessmann, H. (2006) Evolutionary fate of retroposed gene copies in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 3220–3225.

Wagner, E.J. and Carpenter, P.B. (2012) Understanding the language of Lys36 methylation at histone H3. *Nature Reviews Molecular Cell Biology*, 13, 115–126.

Wang, J., Marowsky, N.C. and Fan, C. (2014) Divergence of gene body DNA methylation and evolution of plant duplicate genes. *PLoS ONE*, 9, e110357.

Wood, J.G., Hillenmeyer, S., Lawrence, C., Chang, C., Hosier, S., Lightfoot, W., Mukherjee, E., Jiang, N., Schorl, C., Brodsky, A.S., Neretti, N. and Helfand, S.L. (2010) Chromatin remodeling in the aging genome of *Drosophila*. *Aging Cell*, 9, 971–978.

Wu, X. and Sharp, P.A. (2013) Divergent transcription: a driving force for new gene origination? *Cell*, 155, 990–996.

Xie, C., Zhang, Y.E., Chen, J.Y., Liu, C.J., Zhou, W.Z., Li, Y., Zhang, M., Zhang, R., Wei, L. and Li, C.Y. (2012) Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. *PLoS Genetics*, 8, e1002942.

Yang, H., He, B.Z., Ma, H., Tsaur, S.C., Ma, C., Wu, Y., Ting, C.T. and Zhang, Y.E. (2015) Expression profile and gene age jointly shaped the genome-wide distribution of premature termination codons in a *Drosophila melanogaster* population. *Molecular Biology and Evolution*, 32, 216–228.

Zacher, B., Kuan, P.F. and Tresch, A. (2010) Starr: Simple Tiling ARRay analysis of Affymetrix ChIP-chip data. *BMC Bioinformatics*, 11, 194.

Zacher, B., Torkler, P. and Tresch, A. (2011) Analysis of Affymetrix ChIP-chip data using starr and R/Bioconductor. *Cold Spring Harbor Protocols*, 5, 475–492.

Zaiss, D.M. and Kloetzel, P.M. (1999) A second gene encoding the mouse proteasome activator PA28beta subunit is part of a LINE1 element and is driven by a LINE1 promoter. *Journal of Molecular Biology*, 287, 829–835.

Zhang, G., Huang, H., Liu, D., Cheng, Y., Liu, X., Zhang, W., Yin, R., Zhang, D., Zhang, P., Liu, J., Li, C., Liu, B., Luo, Y., Zhu, Y., Zhang, N., He, S., He, C., Wang, H. and Chen, D. (2015) N6-methyladenine DNA modification in *Drosophila*. *Cell*, 161, 893–906.

Zhang, Y.E., Vibranovski, M.D., Krinsky, B.H. and Long, M. (2010) Age-dependent chromosomal distribution of male-biased genes in *Drosophila*. *Genome Research*, 20, 1526–1533.

559 Zhang, Y.E., Landback, P., Vibranovski, M. and Long, M. (2012) New genes expressed in
560 human brains: implications for annotating evolving genomes. *Bioessays*, 34, 982–991.

561 Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., Li, X., Ding, Y., Yang, S. and
562 Wang, W. (2008) On the origin of new genes in *Drosophila*. *Genome Research*, 18,
563 1446–1455.

564 Zhou, X., Cain, C.E., Myrthil, M., Lewellen, N., Michelini, K., Davenport, E.R., Stephens,
565 M., Pritchard, J.K. and Gilad, Y. (2014) Epigenetic modifications are associated with
566 inter-species gene expression variation in primates. *Genome Biology*, 15, 547.

567

568 Accepted November 12, 2015

569

570

571

572

573

574

575

Fig. 1 Age distribution of *D. melanogaster* protein coding genes. Divergence times within *Drosophila* genus and those between *Drosophila* genus and other insects are based on Tamura *et al.* (2004) and TimeTree (Hedges *et al.*, 2006), respectively. The age of branches is represented by the number above each branch. The count of protein coding genes belonging to a specific age branch annotated in *D. melanogaster* is given in parentheses. Thus, '-4' represents the oldest branch or the oldest gene age group in *D. melanogaster* and '6' represents the youngest group. For example, -4 (7635) indicates that 7635 protein coding

genes in *D. melanogaster* are shared by all major insect orders included in the figure, while 6 (44) stands for a group of 44 protein coding genes specific to *D. melanogaster* which originated in the last few million years.

Fig. 2 Transcriptome and histone modification in fly head. **(**A) Distribution of transcriptional level across different age groups is shown as boxplot where the black lines and dots mark the median and outliers, respectively. The Spearman's test is performed between median expressional levels and corresponding evolutionary ages. (B–D) Both H3K4me3 and H3K36me3 are positively correlated with gene expression level, while H3K9me3 is negatively correlated with gene expression level. Both transcription and histone modification abundance is quantified as $\log_2$FPKM. Across all four panels, the Spearman's σ is shown with P always smaller than 2.2×10^{-16}.

Fig. 3 Older genes tend to bear more H3K4me3, H3K36me3 and less H3K9me3 histone modification marks with respect to binding intensity measured by $\log_2$FPKM. Across three marks, the binding profile data across each gene age group were divided into three classes of equal size ordered by $\log_2$FPKM, namely, 'top', 'middle' and 'bottom'. For each class, boxplots were made across different gene age groups to summarize the distribution. The Spearman's σ (upper lane) and the corresponding P (lower lane, Bonferroni correction) is

presented except the 'bottom' class where the median of many age groups are at the minimum level (log2(0.01)), namely, many genes are not bound at all in the corresponding age groups.

Fig. 4 The histone modification levels of new genes are positively correlated with nearest old genes while new genes show less H3K4me3 and H3K36me3 modification level, and more H3K9me3 modification level. For each type of histone modification marks and each new gene age groups (1~5/6), we calculated the Spearman's rank correlation between new genes and nearest old genes (Panel A), which is followed by Boxplot view and pairwise Wilcoxon signed-rank test (Panel B). Asterisks represent different levels of significance after Bonferroni correction with '*', '**' and '***' denoting '$P < 0.05$', '$P < 0.01$' and '$P < 0.001$' respectively. The solid black line in Panel A shows the linear fitting result.

Fig. 5 Older genes tend to be bound by more H3K4me3 and less H3K9me3 histone modification marks across a broader region. (A) For each TSS, we examined how much the upstream 1 kb and downstream 1 kb window was covered by H3K4me3 binding peaks. Given the overall percentage, we divided genes into five coverage groups (0/low/moderate/high/100%, respectively; see Materials and methods). (B) This panel follows the same convention as Panel A except that the coverage is calculated based on the

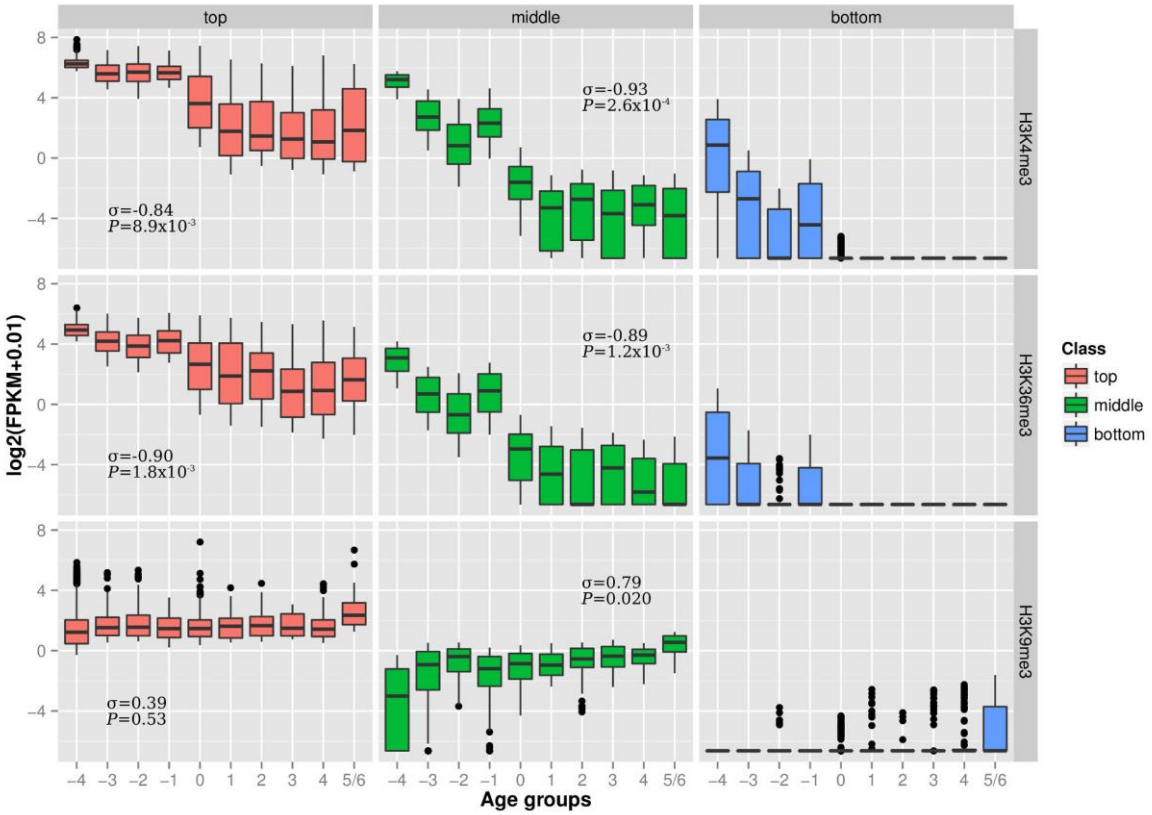
gene body with upstream 1 kb and downstream 1 kb window and the '100%' class were merged into 'high' class since there are so few genes fully covered by H3K9me3 marks (Fig. S2).

635 **Table 1** Correlation between gene age and the percentage of peak coverage of histone
636 modification.

		0	Low	Moderate	High	100%
H3K4me3	σ	0.81	0.38	-0.54	-0.92	-0.90
	P	0.041	1	0.53	2.3×10^{-3}	4.4×10^{-3}
H3K36me3	σ	0.94	-0.81	-0.55	-0.90	-0.92
	P	$<2.2\times10^{-16}$	0.041	0.52	4.4×10^{-3}	2.3×10^{-3}
H3K9me3	σ	0.26	-0.70	-0.84	0.47	
	P	1	0.12	0.018	0.71	

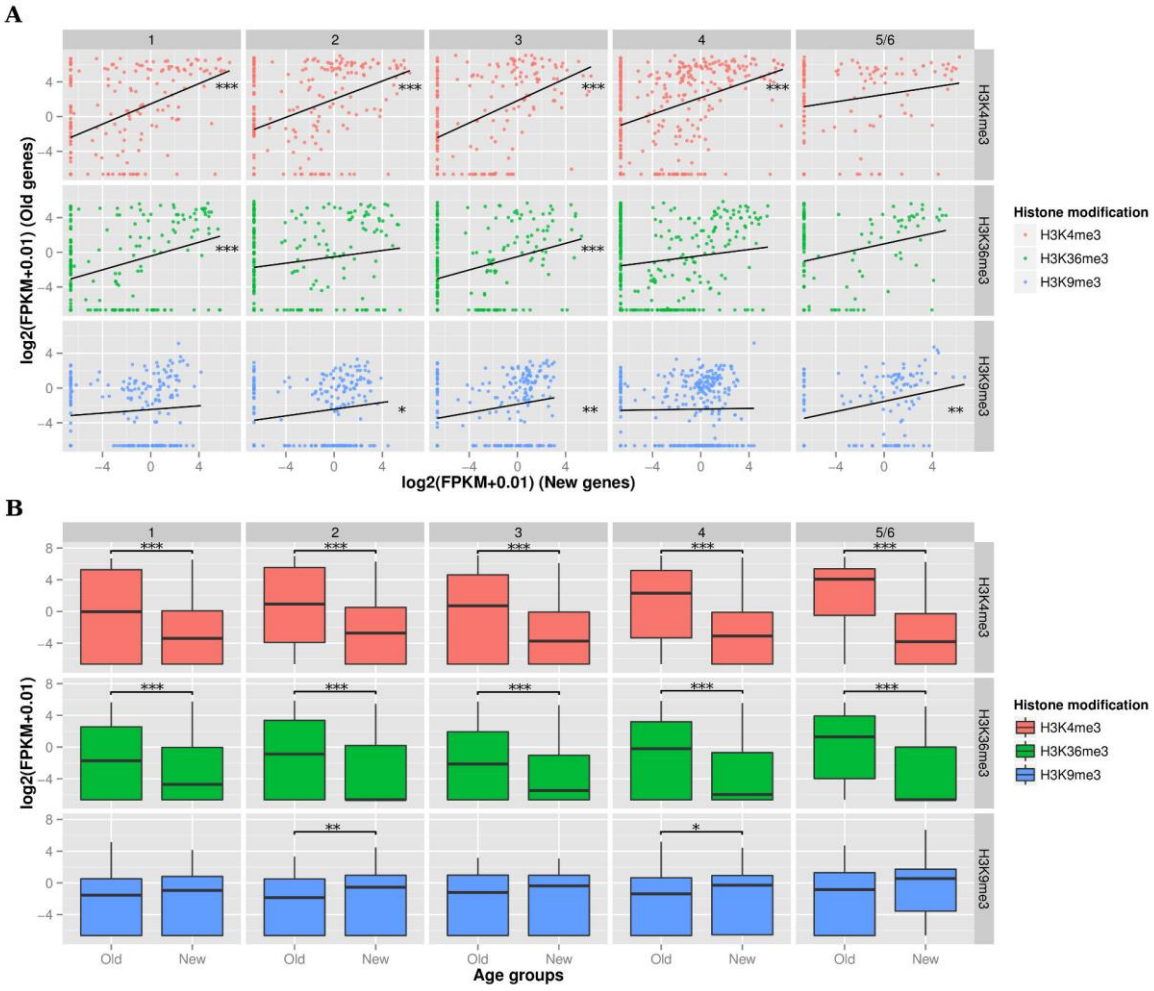
637 For each type of histone modification and each class of peaks coverage, we calculated the
638 Spearman's σ (upper lane) and the corresponding *P* (lower lane, Bonferroni correction)
639 between percentage of peaks coverage and the gene age (Materials and methods). Notably,
640 for H3K9me3, there are not many entries with '100%' coverage and thus a few genes with
641 '100%' coverage are grouped into the 'high' class (Fig. S2).

642



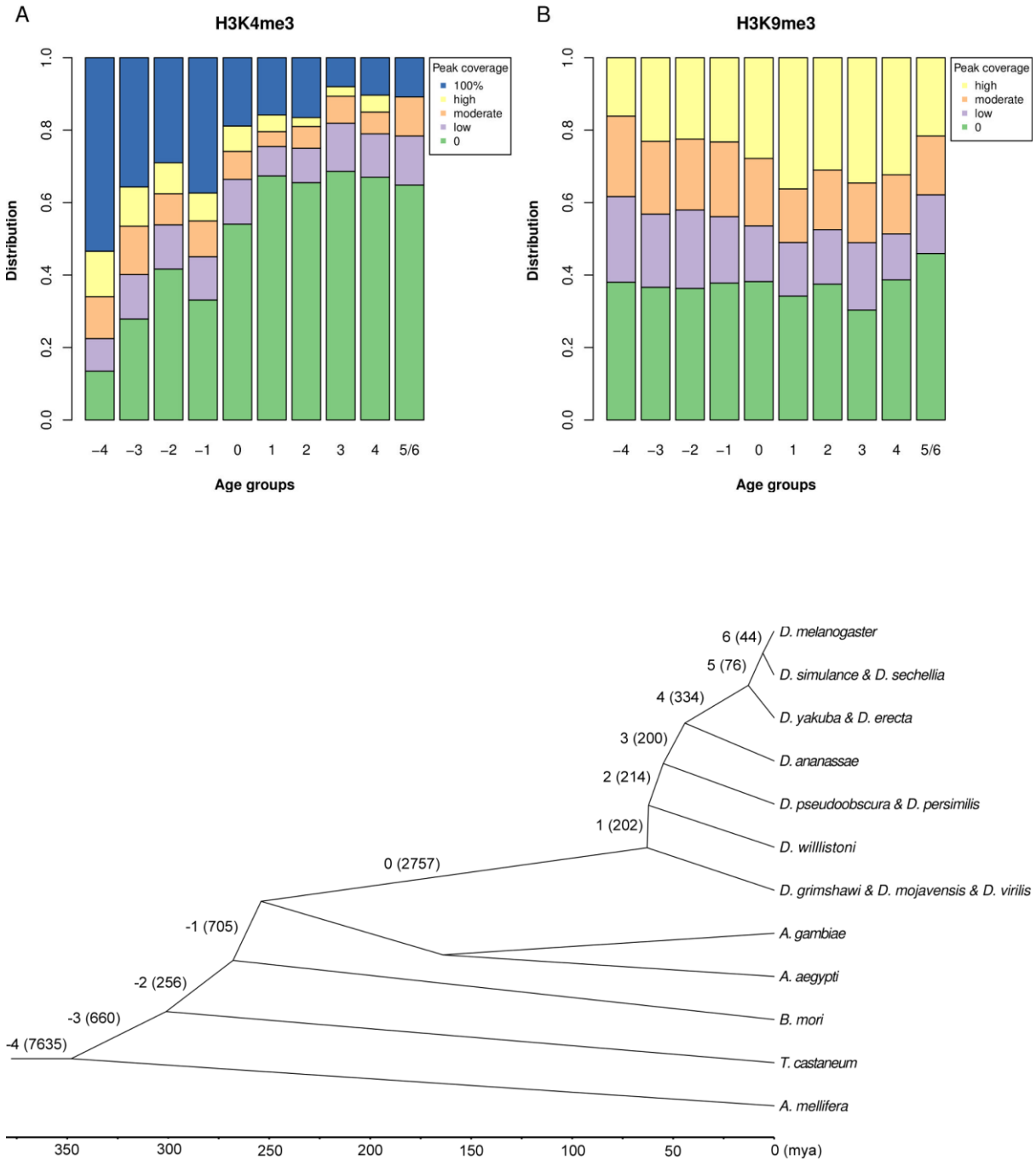
643

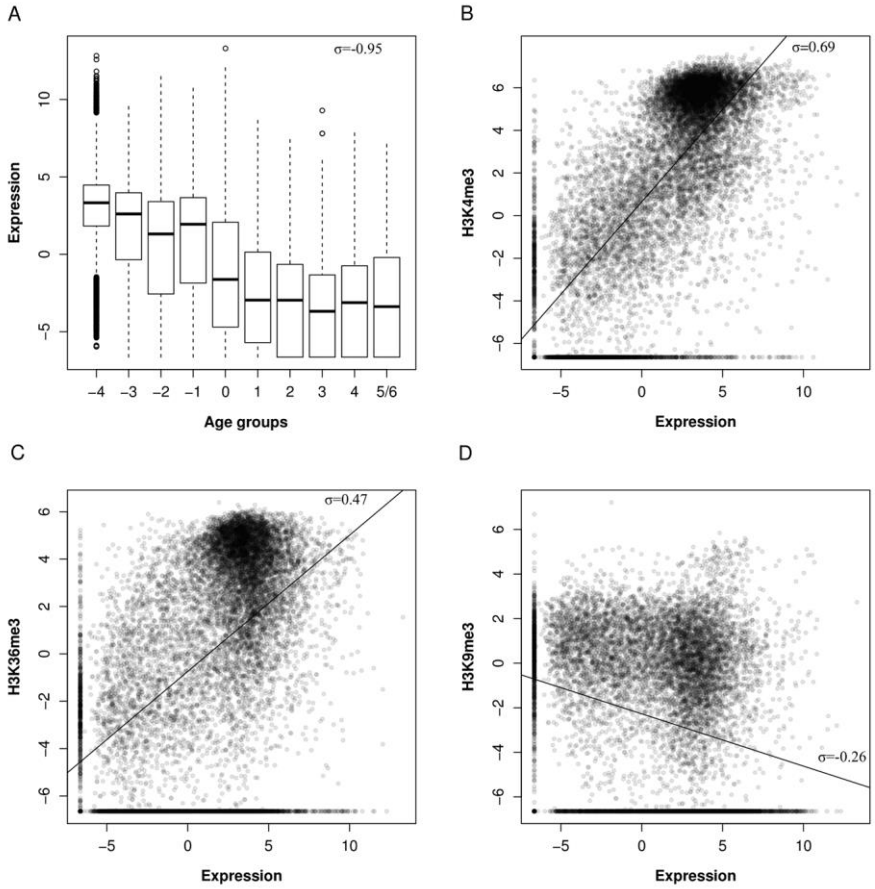
644



645

646





651

652