Computational and Structural Biotechnology Journal 20 (2022) 353–367

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Epigenomic signatures on paralogous genes reveal underappreciated universality of active histone codes adopted across animals

Kuei-Yuan Lan, Ben-Yang Liao [*]

Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Miaoli County 350, Taiwan, ROC

ARTICLE INFO

Article history:
Received 17 September 2021
Received in revised form 15 December 2021
Accepted 18 December 2021
Available online 28 December 2021

Keywords:
Epigenetics
Evolutionary conservation
Gene duplication
Transcriptional initiation
Transcriptional elongation
Homogeneity

ABSTRACT

The results of conventional gene-based analyses which combine epigenome and transcriptome data, including those conducted by the ENCODE/modENCODE projects, suggest various histone modifications performing regulatory functions in controlling mRNA expression (referred to as a histone code) in several model animals. While some histone codes were found to be universally adopted across organisms, "species-specific" histone codes have also been defined. We found that the characterization of these histone codes was confounded by factors (e.g. gene essentiality, expression breadth) that are independent of, but correlated with, gene expression levels. Hence, we attempted to decode histone marks in mouse (*Mus musculus*), fly (*Drosophila melanogaster*), and worm (*Caenorhabditis elegans*) genomes by examining ratios of RNA sequencing (and chromatin immunoprecipitation sequencing) intensities between paralog genes to remove confounding effects that would otherwise be present in a gene-based approach. With this paralog-based approach, associations between four histone modifications (H3K4me3, H3K27ac, H3K9ac, and H3K36me3) and gene expression are substantially revised. For example, we demonstrate that H3K27ac and H3K9ac represent universal active marks in promoters, rather than worm-specific marks as previously reported. Second, acting regions of the studied active marks that are common across species (and across a wide range of tissues at different developmental stages) were found to extend beyond the previously defined regions. Thus, it appears that the active histone codes analyzed have a universality that has previously been underappreciated. Our results suggested that these universal codes, including those previously considered species-specific, could have an ancient origin, and are important in regulating animal gene expression abundance.

1. Introduction

Eukaryotic chromosomes are formed from a repeating structural unit, the nucleosome, which has146 base pairs of DNA wrapped around an octamer of four histone proteins (H2A, H2B, H3, and H4) [1,2]. Histone proteins are crucial to gene accessibility and can be altered by post-translational modifications (PTMs), especially at their N-terminal tails [3]. PTMs included methylation, acetylation, and other types of modification which are established by specific enzymes [4]. The "histone code hypothesis" suggests that specific combinations of histone modifications would be read by certain proteins or protein complexes to result in specified biological outcome, including the state of genomic DNA transcription. [5–8]. Accordingly, histone modifications that mark actively transcribed genes or transcriptionally repressed genes has been referred to as an "active histone code (or mark)" [9,10] or "repressive histone code (or mark)" [11,12], respectively. Histone modifications are critically important for organism development [13], for disease progression [14], and are key contributors to central processes of molecular evolution [15]. However, deciphering how histone codes translate into a gene expression derived biological response(s) which have phenotypic consequences has been a challenge [6,16].

The ENCODE/modENCODE projects have compared chromatin immunoprecipitation sequencing (ChIP-seq) based histone modification signals among different regions of genes that are differentially transcribed (i.e., activated genes vs. suppressed genes) [4,17]. One of the goals of these projects has been to identify histone marks which potentially regulate the mRNA transcription of genes in various genomes. Histone modifications that are enriched in expressed genes (or silent genes), yet are depleted in silent genes (or expressed genes), have been defined as active marks

[*] Corresponding author at: Institute of Population Health Sciences, National Health Research Institutes, Miaoli County 350, Taiwan, ROC.

E-mail address: liaoby@nhri.edu.tw (B.-Y. Liao).

https://doi.org/10.1016/j.csbj.2021.12.027
2001-0370/

(or repressive marks) [15,18]. As a result, both universal and species-specific histone marks (or "codes") have been identified. For example, according to patterns of scaled ChIP-fold enrichment of histone modifications on regions of expressed or repressed genes of various species (Extended Data Figure 1 of ref. [17]), H3K4me3 (trimethylation of histone H3 at lysine 4) represents an active mark (a mark that activates gene transcription) at the 5'-genic regions of fly, worm, and mouse genes [15]. Meanwhile, H3K27ac (acetylation of histone H3 lysine 27) and H3K9ac (acetylation of histone H3 lysine 9) have been found to be active marks specifically in the promoter regions of worm genes (see Supplemental Figure S3 of ref. [15] and below for details). However, there may be biases in conventional approaches by treating genes as units in defining histone codes. These biases could derive from factors (e.g., gene essentiality, expression breadth, and local genomic environments) that correlate with mRNA level, yet the directions or strengths of correlations differs between species. For instance, while introns of genes are marked by specific histone modifications such as H3K36me3 (trimethylation of histone H3 at lysine 36) [19], mRNA expression level is positively correlated with intron density in human and fly, but negatively correlated with intron density in worm [20]. Consequently, it remains unclear if "species-specificit y" of species-specific histone codes is a true effect. We hypothesized that this "species-specificity" is partially an artefact from gene-based observations that were confounded by organism-specific factors unrelated to the transcriptional status of the marked genes (see below and Fig. S1 for the framework of testing this hypothesis).

To address this issue and test our hypothesis, we examined if gene essentiality and expression breadth could have confounded the identification of histone codes, and if these two biological properties are more homogeneous between pairs of paralogous genes (paralogs) originating from tandem duplication events than the other gene pairs (Fig. S1). According to the result, we proposed to define histone codes by treating pairs of tandemly duplicated paralogs as units, and then simultaneously comparing ChIP-seq and RNA sequencing (RNA-seq) signals of such paralog pairs (see Methods) (Fig. S1). With this proposed approach, the influences of potential confounding factors in gene-based approaches were reduced. Here, we focused on four active histone modifications (H3K4me3, H3K27ac, H3K9ac, and H3K36me3) whose regulatory mechanisms have been proposed. For example, H3K4me3 in a promoter may facilitate transcriptional initiation by recruiting transcriptional machinery [21–23]. H3K27ac proximal to the transcription start sites (TSSs) of genes may potentially modulate transcription factor binding or chromatin structure to enhance transcription [24,25]. H3K9ac [26] and H3K27ac [27] in promoter regions may facilitate the progression of RNA polymerase II from transcriptional initiation to transcriptional elongation by recruiting the super elongation complex to chromatin. Meanwhile, a hallmark of active transcription, H3K36me3 at a gene body may enhance suppression of cryptic transcriptional initiation sites during the

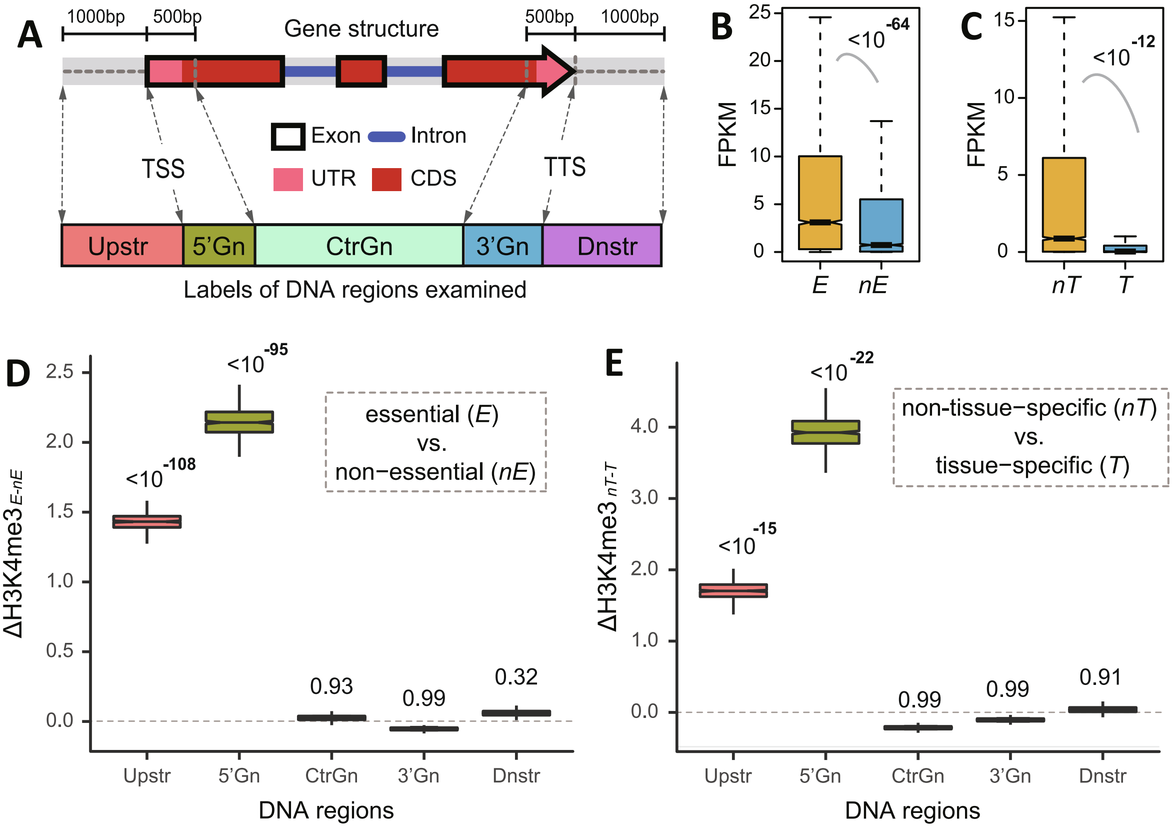

Fig. 1. The presence of bias in gene-based approaches for defining histone codes. (A) Five potential acting regions of histone marks: "TSS-1000 to TSS" ("Upstr"), "TSS to TSS + 500" ("5'Gn"), "TSS + 500 to TTS-500" ("CtrGn"), "TTS-500 to TTS" ("3'Gn"), and "TTS to TTS + 1000" ("Dnstr"). (B) Distributions of mRNA expression level (in FPKM) of essential genes (*E*) versus non-essential genes (*nE*), and (C) non-tissue-specific genes (*nT*, genes with $\tau \leq 0.1$) versus tissue-specific genes (*T*, genes with $\tau \geq 0.6$). The distributions of (D) $\Delta H3K4me3_{E\text{-}nE}$ and (E) $\Delta H3K4me3_{nT\text{-}T}$ indicate that essential (or non-tissue-specific) genes have higher levels of H3K4me3 than nonessential (or tissue-specific) genes in "Upstr" and "5'Gn" regions after controlling for FPKM differences between the two groups. *P*-values are from (B, C) the Mann-Whitney *U* test under the null hypothesis of equal median or (D, E) the sign test under the null hypothesis of the negative median. (B-E) Upper quartile, median, and lower quartile values are indicated in each box, while the bars outside each box indicate quartile ranges. CDS, coding sequence; UTR, untranslated region; TSS, transcription start site; TTS, transcription termination site.

elongation process [28,29]. We used the proposed approach to redefine acting regions of H3K4me3, H3K27ac, H3K9ac, and H3K36me3 in mice (at various developmental stages in different tissues), flies (L3 larva, adult head), and worms (L3 larva, adult). We also investigated H3K27me3 (trimethylation of histone H3 lysine 27), a known repressive mark. For comparative purposes, we additionally generated results of the gene-based method in defining histone marks using criteria analogous to the paralog-based method proposed here. When these results were compared, the species-specificity of the code for each histone mark was updated, and a surprisingly high universality of active histone codes was found to be utilized by the three animal model organisms examined.

2. Results and discussion

2.1. Gene essentiality, expression breadth, and nucleosome density influence the identification of histone codes in gene-based studies

To understand if organism-specific factors correlated with, but independent of, the transcriptional status of the genes have influenced the identification of histone codes in previous studies, it is important to know how active or repressive histone marks were defined previously and what factors are correlated with gene expression abundance. It has been proposed that histone modifications determine the "on versus off" status of target genes [30,31]. The actions of histone modifications can also depend on the DNA regions in which they are located [32]. For each of the histone marks investigated for this study, potential acting regions were specified (Fig. 1A): 1) the region encompassing 1000 bp (base pairs) upstream of the transcriptional start site (TSS), referred to as the "Upstream region" (i.e., "TSS-1000 to TSS" or "Upstr"); 2) the region from the TSS to 500 bp downstream, referred to as the "5'-genic region" (i.e., "TSS to TSS + 500" or "5'Gn"); 3) the region encompassing 500 bp downstream of the TSS to 500 bp upstream of the transcriptional termination site (TTS), referred to as the "central genic region" (i.e., "TSS + 500 to TTS-500" or "CtrGn"); 4) the region including 500 bp upstream of the TTS, referred to as the "3'-genic region" (i.e., "TTS-500 to TTS" or "3'Gn"); and 5) the region including 1000 bp downstream of the TTS, referred to as the "downstream region" (i.e. TTS to TTS + 1000 or "Dnstr"). Previously, ChIP-seq-based signals for histone marks present in these DNA regions were subjected to z-score transformation to compare "expressed genes" (defined as genes with a fragments per kilobase of exon per million fragments mapped (FPKM) value ≥ 1 or other thresholds; see Methods for the calculation of FPKM) versus "silent genes" (defined as genes with a FPKM value <1 or other thresholds). Histone modifications that are enriched in expressed genes (or silent genes), yet are depleted in silent genes (or expressed genes), have been defined as active marks or repressive marks, respectively [15,18]. More sophisticated methods which implement correlation-based statistics and mathematical modeling have been applied to define histone codes [4,33–37]. However, in these studies, genes were treated as units and the potential confounding effects of factors linked with transcriptional abundances (e.g.., gene essentiality, tissue-specificity, etc.) were not considered.

Genes that are highly expressed in animal genomes tend to be essential [38,39] and more ubiquitously expressed [40]. Consistent with these observations, we observed that essential genes (essential genes are genes that are required for the survival or reproduction of an organism, and the essentiality of a gene could be defined based on the phenotypic consequences of gene deletion experiments; see Methods) and non-tissue-specific genes ($\tau \leq 0.1$, see Methods; τ ranges from 0 to 1; a greater τ value indicates greater specificity) tend to be highly expressed in mouse liver tissue (Fig. 1B and 1C). These results suggest that these two types of genes are more likely to be defined as "expressed genes". We measured the difference in average density of H3K4me3 between essential genes ("*E*", see Methods) and nonessential genes ("*nE*", see Methods) (ΔH3K4me3$_{E\text{-}nE}$), and between non-tissue-specific genes ("*nT*", $\tau \leq 0.1$, see Methods for τ) and tissue-specific genes ("*T*", $\tau \geq 0.6$) (ΔH3K4me3$_{nT\text{-}T}$). After controlling for differences in mRNA abundance between the two groups by matching their mRNA abundance distributions (see Methods), the average density of H3K4me3 was found to be significantly higher in the essential genes and non-tissue-specific genes in upstream and 5'-genic regions, respectively (as indicated by the values of ΔH3K4me3$_{E\text{-}nE}$ and ΔH3K4me $_{nT\text{-}T}$ which were significantly > 0 in the "Upstr" and "5'Gn" regions in Fig. 1, D and E, respectively, according to sign tests). These results suggest that the "active" roles for 5'-genic H3K4me3 previously reported based on the result of ENCODE/modENCODE [17] may be partially contributed by the confounding effect of gene essentiality (Fig. 1D) and tissue-specificity (Fig. 1E). It has also been suggested that uneven genomic distribution of nucleosomes may represent a bias when interpreting ChIP-seq signals for histone modifications. For example, genes located in genomic segments occupied by nucleosomes with greater spacing (i.e., with lower nucleosome density), tend to have lower raw ChIP-seq signals for any histone modification [41]. However, experiments to measure nucleosome density that are required to normalize ChIP-seq signals have seldom been performed simultaneously. Consequently, if genes associated with different densities of nucleosomes are directly compared when identifying histone codes, the results obtained may be distorted.

Considering the abovementioned issues, an approach to minimize the influence of confounding effects from gene essentiality, expression breadth, and other potential factors such as genomic environments and gene structure in defining histone codes are desired.

2.2. Homogeneous essentiality, expression breadth, and nucleosome occupancy of paralogous genes

Because of the shared ancestry, paralogs, especially those which arise from tandem duplications, often encode proteins with similar sequences and domain architecture, produce products of the same functional category, and reside in similar genomic environments [42,43]. We therefore proposed to utilize these inherited similarities of paralogs to minimize the influences of confounding factors in decoding regulatory roles of histone modifications.

To examine if our proposed strategy has the potential to test the hypothesis and control for the abovementioned confounding effects, we first examined whether paralogs tend to be homogeneous in terms of tissue-specificity or nucleosome spacing (Fig. S1). Based on the orthology information available for genes annotated by Ensembl, we obtained 7489 multiple-exon duplicated protein coding genes in the mouse genome. Moreover, each of the genes has at least one paralog. Single-exon genes were excluded because they could possibly represent retrogenes (genes duplicated by mRNA-mediated retrotransposition), which often diverge from their progenitor genes in sequence, regulatory patterns, genomic environments, and functions [44,45]. From this pool of genes, we identified 120,869 and $>5.6 \times 10^7$ nonredundant gene pairs (gene pairs consisting of "gene A and gene B" and "gene B and gene A" were considered the same and were not double counted). These two sets of gene pairs were defined as paralog and non-paralog gene pairs, respectively. Differences in tissue-specificity τ (see Methods) (or nucleosome density proximal to the TSS, *nuc*; see Methods) of each of the above gene pair groups were calculated as: $\Delta\tau$ (or Δnuc) = $|X_A - X_B|/|X_A + X_B|$, where X_A and X_B represent τ (or *nuc*) for genes A and B of a focal gene pair, respectively. The paralog gene pairs were also compared with the non-paralog gene pairs after controlling for mRNA expression abundance in liver tissue, from which *nuc* was estimated (see Methods). Briefly, 500 sets of gene pairs were selected, each of which consisted of one paralog pair and one non-paralog pair. For each set, the mRNA expression abundances of both gene copies of the non-paralog pair were matched with those of paralog pair. Then, the average $\Delta\tau$ (or Δnuc) values for both the paralog pairs and the non-paralog pairs were calculated. This process was repeated 1000 times, and distributions for average $\Delta\tau$ (or Δnuc) for the paralog pairs and non-paralog pairs after controlling for mRNA abundance were obtained (Fig. 2, A & B). Significantly smaller $\Delta\tau$ (or Δnuc) values were obtained for the paralog pairs versus the non-paralog pairs. This result suggests that greater homogeneous tissue-specificity (or nucleosome density) exists between paralogs than between non-paralogs.

Next, we examined if paralogous genes tend to be more similar in gene essentiality than non-paralog pairs. Among the paralog gene pairs defined in the above section, 3324 gene pairs had both copies of genes phenotyped to define their gene essentiality (see Methods). According to essentiality data of mouse genes defined based on the phenotypic data of gene deletion experiments (see Methods), it was determined that these 3324 gene pairs include 824 (40.6%) essential genes and 1204 (59.4%) nonessential genes. Considering these proportions within this subset of genes, the expected number of randomly selected gene pairs with equal essentiality (both essential or both nonessential) was calculated as: $(p^2 + q^2) \times 3324 = 1720.35$, where $p = 0.406$ and $q = 0.594$. The observed number of gene pairs with equal essentiality was 2176, a value significantly greater than the expected value ($P < 10^{-55}$; χ^2 test). In several previous studies, gene essentiality, expression level, and gene duplication were found to be interrelated [43]. To determine whether this higher-than-expected number is a byproduct from this type of interrelationship, we randomly selected 500 pairs of non-paralog genes with matched mRNA abundance distributions to paralog gene pairs (see Methods). After repeating this process 1000 times and calculating the proportion of gene pairs with equal essentiality (P_{eqESS}) each time, a distribution of P_{eqESS} for the nonparalog gene pairs was obtained. This same procedure was also applied to the paralog gene pairs. While gene essentiality could change after a gene duplication event [46], paralog pairs still exhibited a significantly greater chance of being equally essential (or equally nonessential) than non-paralog gene pairs, when the difference in mRNA abundance was controlled ($P < 10^{-300}$, Mann-Whitney U test) (Fig. 2C).

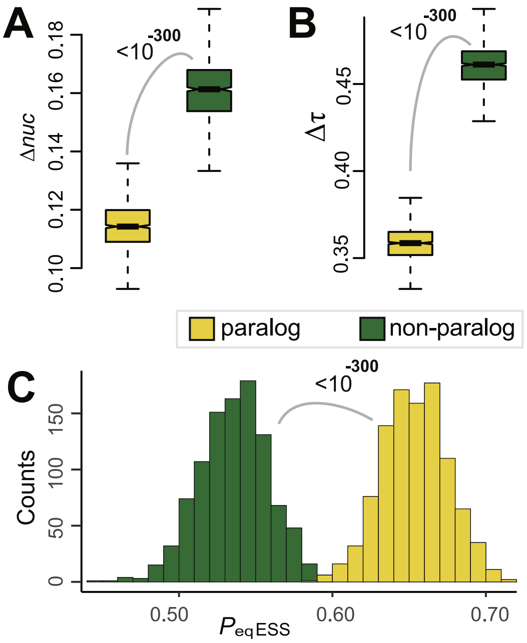

Fig. 2. Homogeneity of paralog genes in regard to (A) nucleosome density, (B) tissue-specificity, and (C) essentiality. In comparison with randomly selected non-paralog gene pairs with matched mRNA expression abundances, paralog pairs have significantly lower Δnuc (A) and $\Delta\tau$ (B) values, and significantly greater P_{eqESS} (C). In each box of the boxplot (A) or (B), upper quartile, median, and lower quartile values are indicated. The bars outside each box indicate quartile ranges. In (C), the distribution of P_{eqESS} for each of the paralog or non-paralog pairs is shown in the histogram. *P*-values are from the Mann-Whitney *U* test under the null hypothesis of equal median of the two compared gene pair groups.

Due to significant homogeneity in expression breadth, nucleosome density, and gene essentiality among paralogs, we expected that the biases associated with these three properties (and potentially many others) in defining histone codes can be autocorrected by comparing and contrasting ChIP-seq histone modification signals and RNA-seq signals between pairs of paralogs which originated from tandem duplication events (Fig. S1).

2.3. Contrasting ChIP-seq and RNA-seq signals between paralogs to define regulatory roles of animal histone modifications

To examine how universal an active or repressive histone code is, data collected from multiple species under different conditions are needed for the analysis. In addition to the 7489 multiple-exon duplicated mouse genes mentioned above, we further obtained multiple-exon duplicated genes from fly and worm genomes. Consequently, 120869, 2624, and 69,412 paralog pairs from 7489 mouse genes, 1379 fly genes, and 6034 worm genes, respectively, were included in subsequent analyses to reassess histone codes in each of these model organisms.

To quantify histone modification status and expression status of these genes, ChIP-seq histone modification data for H3K4me3, H3K27ac, H3K9ac, and H3K36me3 were obtained from: mouse organs (heart, liver, forebrain/cerebellum) at three different stages (embryonic day 11.5, embryonic day 16.5, and 8-week-old adult), fly (whole organism of L3 larval, and adult head), and worm whole organism (at L3 larval and adult stages), except for H3K9ac and H3K36me3 data for 8-week-old mouse brain and H3K4me3 and H3K9ac data for adult worm which were unavailable. The four active histone marks were selected based on the regulatory roles previously proposed for each (see Introduction). In addition, the three organs selected for analysis from mouse (heart, liver, forebrain/cerebellum) were intended to represent differentiated cell lineages derived from mesoderm, endoderm, and ectoderm germ layers in mammals, respectively [47]. While ChIP-seq histone modification data and RNA-seq data for the selected tissues/stages of the model organisms were profiled simultaneously by ENCODE/modENCODE [17], data for the different organisms were independently generated (see Methods). To integrate these data for meta-analysis, both RNA-seq and ChIP-seq data were reprocessed (see Methods).

For each paralog pair from each species, the ratio of mRNA abundance was calculated as $E_{\mathrm{S}}/E_{\mathrm{W}}$, where E_{S} or E_{W} represents the FPKM of the strongly expressed copy, or the weakly expressed copy, of the focal paralog pair, respectively (Fig. 3). All of the paralog pairs were classified into four groups: "$1 \leq E_{\mathrm{S}}/E_{\mathrm{W}} < 2$", "$2 \leq E_{\mathrm{S}}/E_{\mathrm{W}} < 4$", "$4 \leq E_{\mathrm{S}}/E_{\mathrm{W}} < 8$", or "$E_{\mathrm{S}}/E_{\mathrm{W}} > 8$". Meanwhile, the ratio of histone modification signals was calculated as $H_{\mathrm{S}}/H_{\mathrm{W}}$, where H_{S} or H_{W} represents the ChIP-seq signals of the strongly expressed copy, or the weakly expressed copy, respectively (Fig. 3). Because H_{W} could have a value of zero, a pseudo count value of "0.5" was added to both H_{S} and H_{W} when calculating $H_{\mathrm{S}}/H_{\mathrm{W}}$. $H_{\mathrm{S}}/H_{\mathrm{W}}$ was calculated for "Upstr", "5'Gn", "CtrGn", "3'Gn", and "Dnstr" regions for each gene (Fig. 1A). Paralogs consisting of

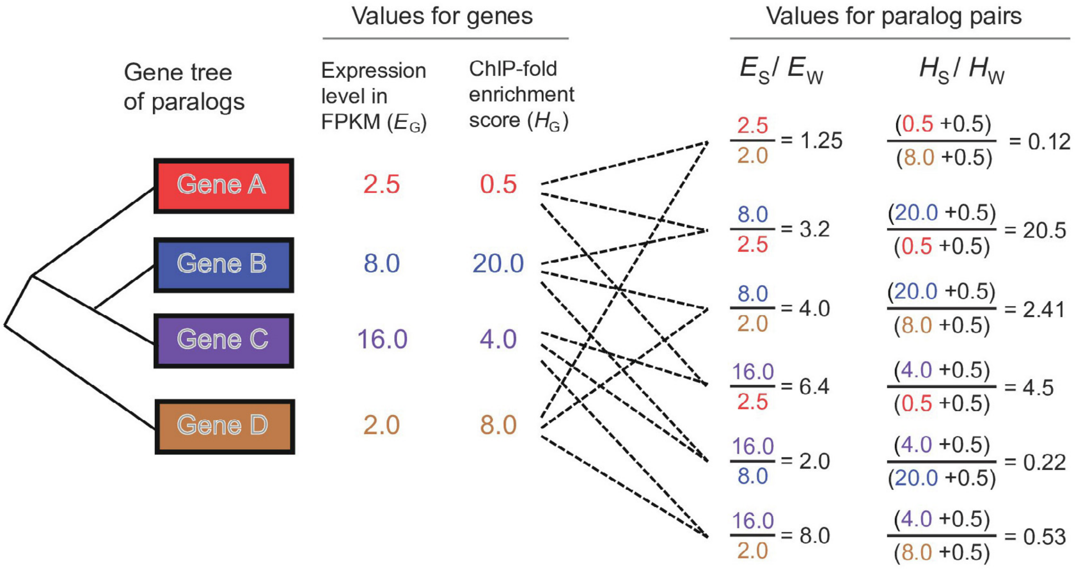

Fig. 3. Example of the calculation of E_S/E_W versus H_S/H_W of a histone modification for hypothesis testing. In this example, the paralog group has four member genes and therefore 6 pairs of paralogs. The mRNA abundance in FPKM (E_G) and ChIP-fold enrichment score of a given region (H_G) of each gene, and the values of E_S/E_W and H_S/H_W of each paralog pair, are indicated.

genes with a length <1500 bp from the TSS to the TTS were excluded. Region-specific H_S/H_W values were compared among the groups with different ranges of E_S/E_W values.

Histone modifications which exhibited the following patterns in specific genic regions were defined as "active marks": *Pattern A-I*: groups with increased E_S/E_W showed an increased value of mean H_S/H_W; and *Pattern A-II*: H_S/H_W of the group with "$E_S/E_W > 8$" was statistically greater than those of the group "$1 \leq E_S/E_W < 2$" according to the Mann-Whitney *U* test under the null hypothesis of equal medians of the two compared groups. Those showing opposite patterns (i.e., *Pattern R-I*: groups with increased E_S/E_W showed a decreased value of mean H_S/H_W; and *Pattern R-II*: H_S/H_W of the group "$1 \leq E_S/E_W < 2$" was statistically greater than those of the group "$E_S/E_W > 8$") were defined as "repressive marks". Based on the results obtained from the three animal models, the histone marks that could be consistently identified in all three species (and all conditions) were labeled "universal marks", whereas the others were labeled "species-specific marks".

2.4. Extended universal regions of H3K4me3 as an active mark

It is known that H3K4me3 tends to be present in the promoters of actively transcribed genes, and its high intensity around a TSS positively correlates with the corresponding gene's mRNA abundance. This distribution feature, and the enzymes that methylate H3K4, have been found to be evolutionarily conserved across fungi, plants, and animals [48,49]. Since H3K4me3 is able to recruit transcriptional machinery and serve as a substrate to facilitate initiation of transcription [21–23], divergence in H3K4me3 nearby TSSs has been linked with divergence of mRNA expression level of primate orthologous genes [50].

Based on enriched (and depleted) signals of H3K4me3 in specific regions of expressed genes (and silent genes) reported by modENCODE [17], H3K4me3 is an active mark commonly used in the 5′ genic regions of genes in human, fly, and worm genomes, and not in other regions [15]. The results from our paralog-based analyses show that the H_S/H_W and E_S/E_W distributions which characterize H3K4me3 are consistent with *Pattern A-I* and *Pattern A-II* not only in 5′-genic regions, but also in promoter regions, central genic, and 3′-genic regions of worm genes (Fig. 4A), in promoter regions, central genic regions, and downstream regions of fly genes (Fig. 4B, by considering only consensus parts across two conditions), and in promoter regions and central genic regions of mouse genes (Fig. 4C, by considering only consensus parts across tissues and developmental stages). Thus, the acting regions of H3K4me3 which are found to be universally associated with gene upregulation in all three models extend the previously described acting region of H3K4me3 to both upstream and central genic regions. It is possible that the localization of H3K4me3 at central genic regions of highly expressed animal genes is related to assurance of transcriptional consistency (or stability) [48,49,51]. However, further studies are needed to better understand this possibility.

2.5. Revised association of H3K27ac and H3K9ac with gene expression and species-specificity

Based on the data of modENCODE [17], it has been suggested that H3K27ac and H3K9ac mark genes that are transcriptionally active in worm, fly, and mouse genomes when they are located in "5′-genic regions" [15]. In the present study, we consistently observed that these two histone modification marks exhibited ChIP-seq intensities consistent with *Pattern A-I* and *Pattern A-II* across the three model organisms investigated (H3K27ac, Fig. 5; H3K9ac, Fig. 6) and across different mouse organs at various developmental stages (H3K27ac, Fig. 5C; H3K9ac, Fig. 6C). Taken together, these results suggest a universally active role for H3K27ac and H3K9ac in 5'-genic regions of animal genes.

In addition to the "5′-genic region", H3K27ac and H3K9ac also exhibited *Pattern A-I* or *Pattern A-II* in upstream and central genic regions across the same species and stages/tissues (Figs. 5 & 6, respectively). These results suggest that H3K27ac and H3K9ac modifications in these regions are associated with gene activation. Previously, H3K27ac and H3K9ac were identified to be enriched in the upstream regions of activated *C. elegans* genes, yet not in fly or mammal genes, and were labeled as "worm-specific" active marks according to the gene-based analysis [15]. In the present study, both upstream H3K27ac and H3K9ac were identified as active marks universally adopted by a wide range of animal species (and by divergent cell lineages in the same organism) (Figs. 5 & 6). It has been hypothesized that the presence of H3K9ac [26]

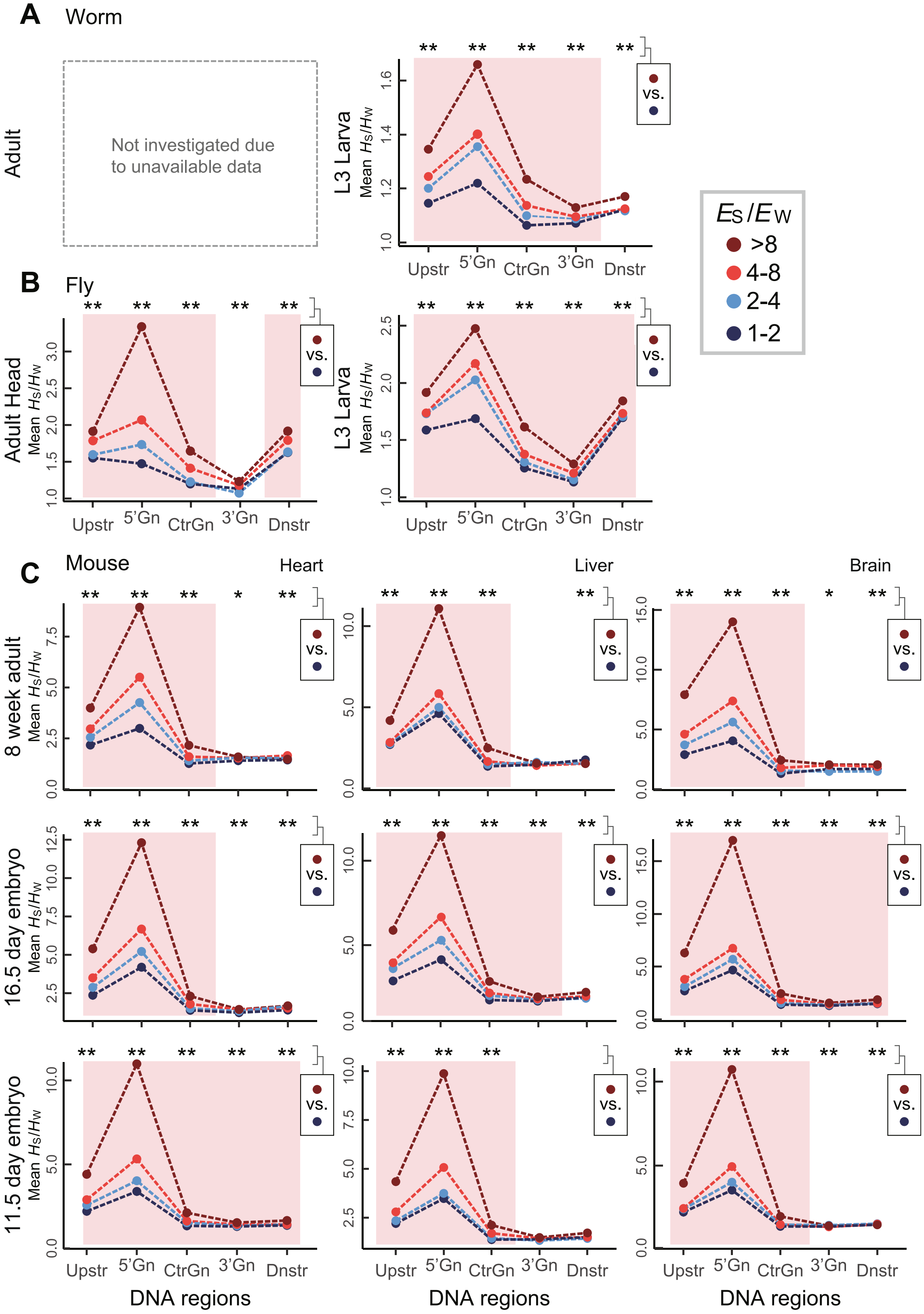

Fig. 4. The relationship of E_S/E_W versus H3K4me3 H_S/H_W in various regions of (A) worm genes (at L3 larva stage), (B) fly genes (adult head or L3 larva), and (C) mouse genes (heart, liver, or cerebellum/forebrain at developmental stages of day 11.5, day 16.5, or 8-weeks, as indicated). Regions of the corresponding genes showing both *Patterns A-I* and *A-II* which support an active role for the histone mark examined are marked with a pink background. The symbols, * or **, above each panel indicate $0.01 < P < 0.05$ or $P < 0.01$, respectively, according to the Mann-Whitney *U* test in examining *Pattern A-II*. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

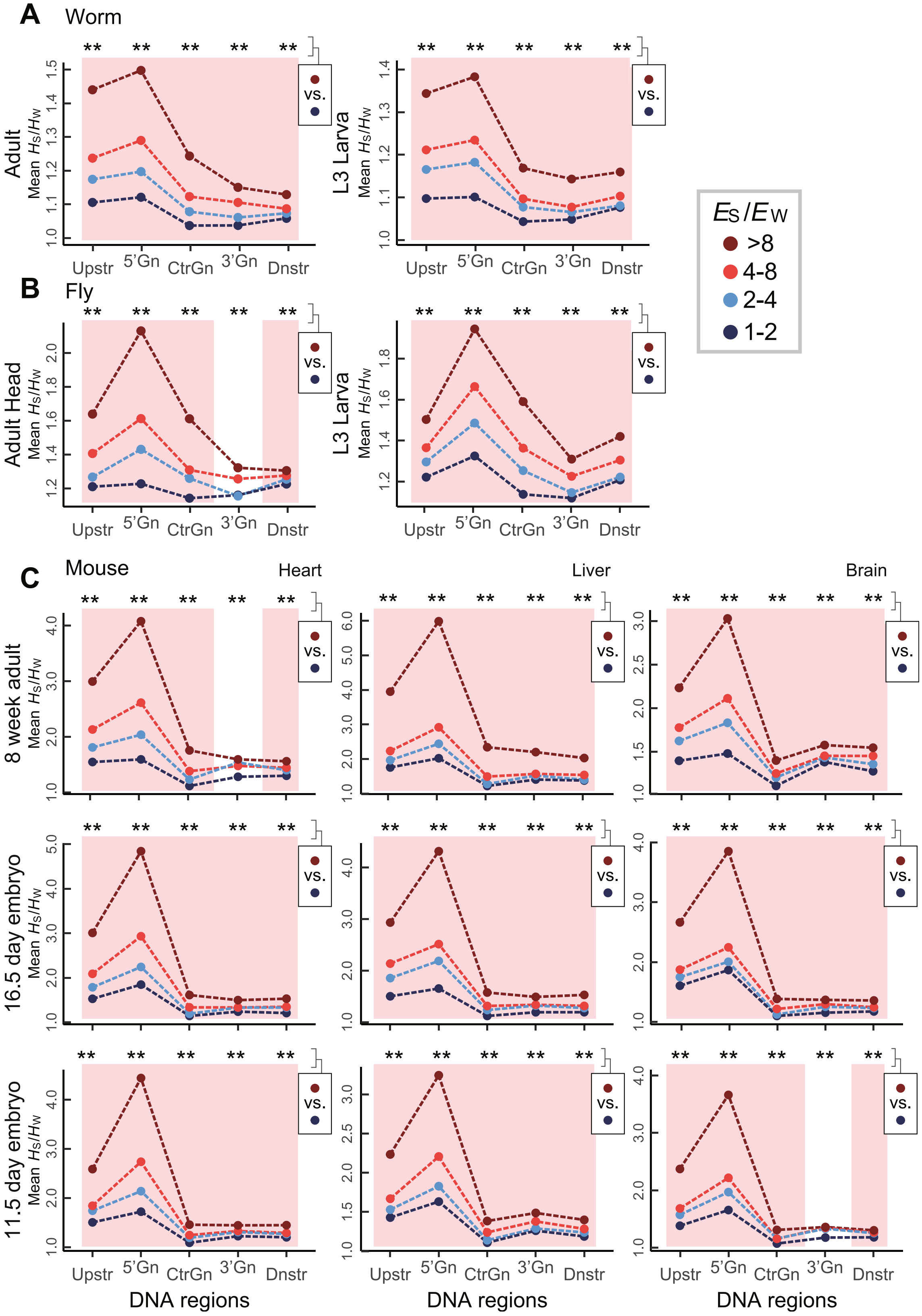

Fig. 5. The relationship of E_S/E_W versus H3K27ac H_S/H_W in various regions of (A) worm genes (at adult or L3 larva stage), (B) fly genes (adult head, or L3 larva), and (C) mouse genes (heart, liver, or cerebellum/forebrain at developmental stages of day 11.5, day 16.5, or 8-weeks, as indicated). Regions of the corresponding genes showing both *Patterns A-I* and *A-II* which support an active role for the histone mark examined are marked with a pink background. The symbols, * or **, above each panel indicate $0.01 < P < 0.05$ or $P < 0.01$, respectively, according to the Mann-Whitney *U* test in examining *Pattern A-II*. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

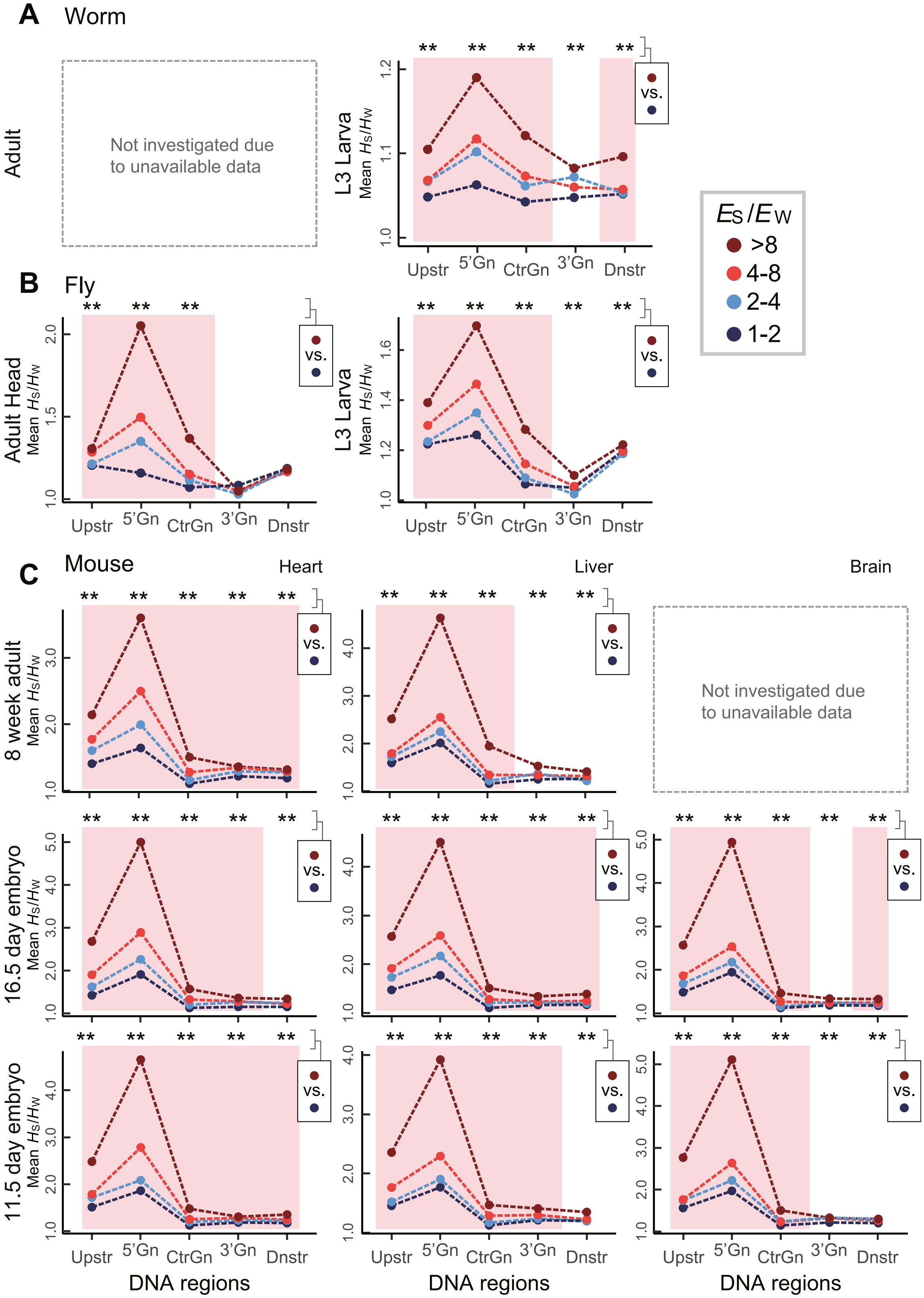

Fig. 6. The relationship of E_S/E_W versus H3K9ac H_S/H_W in various regions of (A) worm genes (at L3 larva stage), (B) fly genes (adult head, or L3 larva), and (C) mouse genes (heart, liver, or cerebellum/forebrain at developmental stages of day 11.5, day 16.5, or 8-weeks, as indicated). Regions of the corresponding genes showing both *Patterns A-I* and *A-II* which support an active role for the histone mark examined are marked with a pink background. The symbols, * or **, above each panel indicate $0.01 < P < 0.05$ or $P < 0.01$, respectively, according to the Mann-Whitney *U* test in examining *Pattern A-II*. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and H3K27ac [27] in gene promoters mediate progression from initiation of transcription to elongation during the transcriptional process. Moreover, glioma-amplified sequence 41 (GAS41) is able to detect H3K27ac in promoters of actively transcribed genes, and this has been associated with cancer cell proliferation [52]. Interestingly, H3K27ac [53] and H3K9ac [54] have been shown to mark promoters of expressed genes in fungal genomes. Thus, it appears that H3K9ac and H3K27ac modifications in promoters are important for guiding gene expression in not only various animal species but also fungal species, and the role of guiding transcription of both of these two modifications could have a common origin dated before the divergence of fungi and animals.

Investigations in several plants, including rice (*Oryza sativa*) [55], maize (*Zea mays*) [56], and *Arabidopsis* [57], have shown that H3K9ac and H3K27ac are overrepresented across the gene body. Moreover, the level of modifications at genic regions proximal to TSSs has been found to positively correlate with mRNA expression levels. This trend in plants is remarkably consistent with the results of genic H3K9ac and H3K27ac in the three animal models examined in the present study (Figs. 5 & 6). However, the underlying mechanistic implications for this trend are not entirely clear. We also observed that the intensities of H3K27ac and H3K9ac were highly correlated in all of the genic regions, independent of the organisms or tissues examined (except for adult worm due to unavailable H3K9ac data) (Table S1). It will be of interest to determine whether these two modifications which are associated with gene upregulation mediate the same mechanism across species. The universal acting regions of H3K27ac were also found to extend to the downstream regions of genes (Fig. 5). This result is consistent with a general function observed for H3K27ac in mouse fibroblast cells to control transcriptional readthrough of animal genes [58]. However, this possible role requires further exploration.

2.6. H3K36me3 as a universal active mark

H3K36me3 influences alternative splicing [19] and guides N6-methyladenosine modification of mRNAs [59]. As a hallmark of active transcription, H3K36me3 prevents cryptic transcription initiation during the elongation process which has been associated with aging in fungal and animal cells [28,60,61]. The ChIP-fold enrichment status of H3K36me3 characterized by modENCODE through a gene-based analysis [17] has demonstrated that H3K36me3 is commonly observed in central genic regions of mammalian, fly, and worm genes (acting as a universal active mark), while association of H3K36me3 with other gene regions (i.e., "3'-genic regions" and "downstream regions") is organism-specific [15] (Fig. 7).

The presence of H3K36me3 in "central genic regions", "3'-genic regions", and "downstream regions" was consistent with distributions of H_S/H_W and E_S/E_W according to *Pattern A-I* and *Pattern A-II* in genes of the three model organisms investigated (Fig. 7, A-C, across organisms; Fig. 7C, across different mouse organs at various developmental stages). Hence, our paralog-based analysis suggests that the universal acting regions of H3K36me3 can extend downstream of "central genic regions". The universal active role of H3K36me3 in the "3'-genic region" may further imply a common need for animal cells to prevent spurious transcription initiation from cryptic promoters, even in this region.

The phenomena of stop codon readthrough is prevalent in metazoans [62]. Based on the observations that highly expressed genes tend to not have readthrough motifs and have a decreased rate of readthrough, it has been proposed that stop codon readthrough is largely non-adaptive [63]. However, in the present study, H_S/H_W of both H3K27ac and H3K36me3 increased as E_S/E_W increased universally in the 3′-genic and downstream regions (Fig. 7). Enrichment of these two modifications in these two regions has previously characterized genes which produce mRNAs by transcriptional readthrough [58]. Further investigation is needed to determine whether the presence of H3K27ac and H3K36me3 in "downstream regions" is related to maintenance of an open chromatin state for "functional" transcriptional readthrough.

2.7. Assessing H3K27me3 as a repressive mark

In addition to the four active histone marks examined above, we also conducted a paralog-based analysis of H3K27me3. It has been proposed that this histone mark recruits PRC1 and induces monoubiquitination of H2A at lysine 119 to inhibit RNA polymerase II elongation [64–66]; meanwhile, depositions of H3K27me3 at gene bodies may be the consequence of gene silencing [67]. Previous gene-based analyses have indicated that intensities of H3K27me3 negatively correlate with mRNA expression levels at upstream and 5′ genic regions of genes in human cell lines, fly, and worm [15,17]. However, in our paralog-based approach, H3K27me3 only exhibited *Patterns R-I* and *R-II* in both stages of worm and adult mouse tissues in the upstream region (Fig. 8), and not in fly and the rest of the mouse tissue/stages examined. Moreover, although H3K27me3 generally exhibited *Patterns R-I* and *R-II* in the 5′-genic region, *Pattern R-I* was not observed in mouse embryo day 16.5 heart tissue due to a higher mean H_S/H_W of "$4 \leq E_S/E_W < 8$" paralogs than mean H_S/H_W of "$2 \leq E_S/E_W < 4$" paralogs (Fig. 8).

It should be noted that for H3K27me3 in the 5′ genic region, the E_S/E_W between the two groups of "$4 \leq E_S/E_W < 8$" and "$2 \leq E_S/E_W < 4$" paralogs did not statistically differ from each other in heart tissue from the day 16.5 mouse embryo ($P = 0.560$, U test) (Fig. 8C). Similarly, for H3K27me3 in the promoter regions of fly genes, the E_S/E_W between the "$4 \leq E_S/E_W < 8$" paralogs and the "$2 \leq E_S/E_W < 4$" paralogs also did not statistically differ (L3 larva: $P = 0.497$; adult head: $P = 0.099$; U test) (Fig. 8B). Thus, if we used a less stringent criteria (i.e., by pooling "$4 \leq E_S/E_W < 8$" and "$2 \leq E_S/E_W < 4$" paralogs into a single category ["$2 \leq E_S/E_W < 8$"] to examine if H_S/H_W decreases with increased E_S/E_W) in defining repressive marks, H3K27me3 in 5'-genic regions could be considered a universal mark. In a study of mouse tissues, acquisition of H3K27me3 kept a substantial proportion of genes silent after early embryogenesis [68]. Therefore, if we only consider mouse adult tissues and L3 larva stage of worm and fly as done in ref. [17] in our analysis, and if we pooled "$4 \leq E_S/E_W < 8$" and "$2 \leq E_S/E_W < 4$" into a single group, upstream H3K27me3 could also be considered a universal mark across mouse, fly, and worm genomes. Hence, the present results are not contradictory to previously reported associations of H3K27me3 with gene activities, if a less stringent criterion is applied.

2.8. The analogous gene-based analysis

Although in the above sections, the results inferred from modENCODE gene-based data and the results based on our proposed paralog-based method were compared and discussed, the methods used in generating these two sets of results did not differ only in the units used for decoding histone modifications. For the purpose of a fair comparison, we additionally conducted a gene-based analysis using criteria analogous to our proposed paralog-based method as follows. In this gene-based method, E_G represents the FPKM of each gene (see Methods). For each condition of each species, genes were categorized into four equal sized bins according to the FPKM values, from low to high as E_G = Q_1 (0–25 percentile rank), Q_2 (25–50 percentile rank), Q_3 (50–75 percentile rank) or Q_4 (75–100 percentile rank). H_G represents the ChIP-fold enrich-

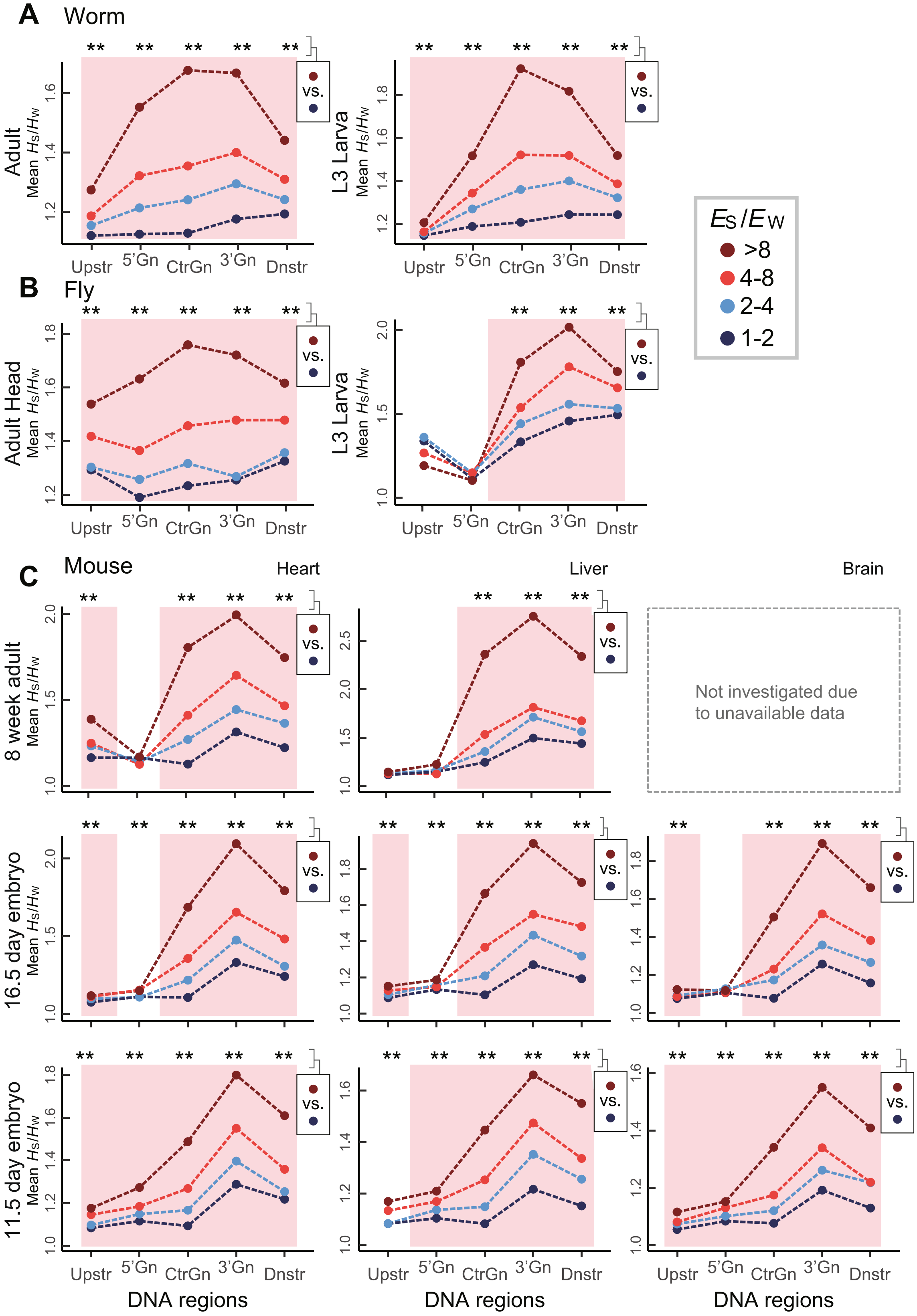

Fig. 7. The relationship of E_S/E_W versus H3K36me3 H_S/H_W in various regions of (A) worm genes (at adult or L3 larva stage), (B) fly genes (adult head, or L3 larva), and (C) mouse genes (heart, liver, or cerebellum/forebrain at developmental stages of day 11.5, day 16.5, or 8-weeks, as indicated). Regions of the corresponding genes showing both *Patterns A-I* and *A-II* which support an active role for the histone mark examined are marked with a pink background. The symbols, * or **, above each panel indicate $0.01 < P < 0.05$ or $P < 0.01$, respectively, according to the Mann-Whitney *U* test in examining *Pattern A-II*. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

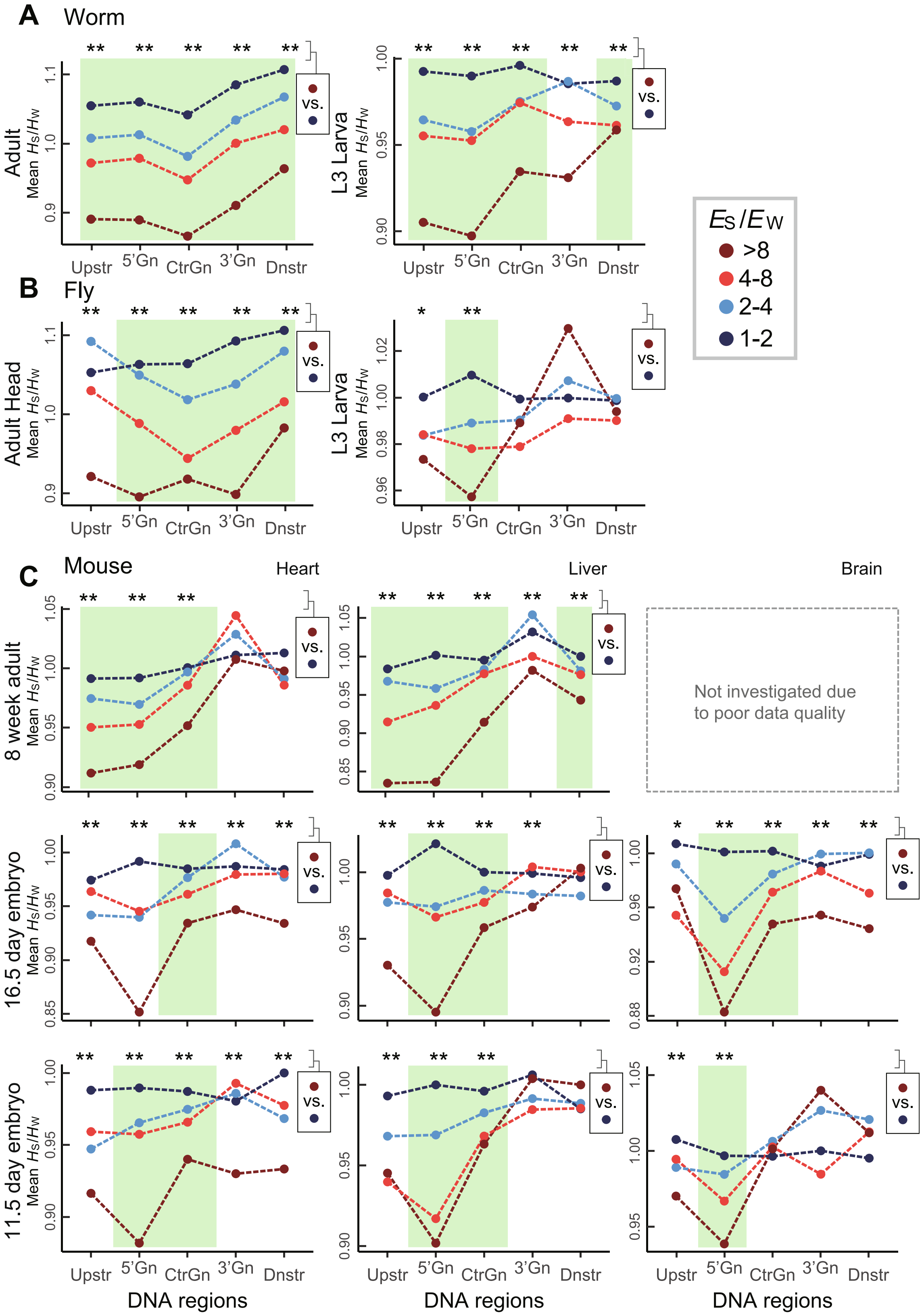

Fig. 8. The relationship of E_S/E_W versus H3K27me3 H_S/H_W in various regions of (A) worm genes (L3 larva stage), (B) fly genes (L3 larva stage), and (C) mouse genes (heart, liver, or cerebellum/forebrain at developmental stages of day 11.5, day 16.5, or 8-weeks, as indicated). Regions of the corresponding genes showing both *Patterns R-I* and *R-II* to support the repressive role of H3K27me3 are marked with a green background. The symbols, * or **, above each panel indicate $0.01 < P < 0.05$ or $P < 0.01$, respectively, according to the Mann-Whitney *U* test in examining *Pattern R-II*. Data from 8-week mouse brain were not analyzed due to poor quality (Pearson's correlation coefficients for H3K27me3 signals of two replicates in "Upstr", "5'Gn", "CtrGn", "3'Gn", or "Dnstr" region of mouse genes were only: 0.394, 0.354, 0.590, 0.476, or 0.559, respectively). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

ment score of the focal region for each gene (see Methods). Histone modifications which exhibited the following patterns in specific genic regions were defined as "active marks": *Pattern A-I*: groups with increased E_G showed an increased value of mean H_G; and *Pattern A-II*: H_G of the group with "E_G = Q_4" was statistically greater than those of the group "E_G = Q_1" according to the *U* test under the null hypothesis of equal medians of the two compared groups. Those showing opposite patterns (i.e., *Pattern R-I*: groups with increased E_G showed a decreased value of mean H_G; and *Pattern R-II*: H_G of the group "E_G = Q_1" was statistically greater than those of the group "E_G = Q_4") were defined as "repressive marks".

We found the universal regions of active histone marks defined by this analogous gene-based method to be: H3K4me3 at central genic region (Fig. S2); H3K27ac at promoter, 5′-genic, and central genic regions (Fig. S3); H3K9ac at 5′-genic and central genic regions (Fig. S4); H3K36me3 at 3′-genic region (Fig. S5). For the repressive histone modification mark, H3K27me3 was found to be universal at gene downstream region (Fig. S6). Only universal regions of two of the five histone modifications (i.e., H3K27ac and H3K9ac) are overlapped with universal regions defined according to the modENCODE data, as summarized in Fig. S7. Because this analogous gene-based approach is expected to produce results confounded by factors correlated with but independent of expression level, as previously discussed, the "histone codes" identified by this approach may not reflect the direct association between histone modifications and transcriptional control of genes and are not discussed further.

2.9. Concluding remarks

In the present study, we developed a novel approach for defining histone codes by contrasting ChIP-seq histone modification signals and RNA-seq signals between paralog pairs. Our method has the potential to reduce the confounding effects of gene properties which influence the determination of histone codes in a conventional gene-based approach. We also demonstrated that simultaneously controlled factors include gene essentiality, tissue-specificity, and nucleosome density (Figs. 1 & 2). In theory, these controlled factors could have included any of the factors which exhibited a higher similarity between tandemly duplicated paralogs than between randomly sampled gene pairs from a genome (i.e., intron density, protein domain architecture, functional categories of genes, etc.). With our approach, we redefined and substantially revised the associations of several histone modifications with gene expression (Figs. 4-8), and the results are summarized in Fig. 9.

Histone modifications at specific regions that were previously identified as active marks according to the modENCODE data (regions marked with red empty boxes in Fig. 9) were all verified by our approach (as shown with red boxes marked with orange solid boxes in Fig. 9). This remarkable consistency, was not observed in the gene-based method using equivalent criteria in defining active or repressive histone marks (Fig. S7). More importantly, we substantially extended the universal active regions for the four active marks examined (H3K4me3, H3K27ac, H3K9ac, H3K36me3) beyond the regions previously defined by ENCODE/modENCODE. The extension of universal active regions for active histone marks were not obvious in the results produced by the gene-based approach (Fig. S7). According to the results for H3K27me3, the criteria we used to define the action of histone marks are stringent (see the previous section). Moreover, the universality of each of the active histone marks was inferred not only from multiple species, but also under various conditions for the species examined. Therefore, the proposed extension of universal acting regions for H3K4me3, H3K27ac, H3K9ac, and H3K36me3 are unlikely to be artifacts. Indeed, of the many histone codes previously considered to be "species-specific", we found them to be universally adopted in fly, worm, and mouse genomes (Fig. 9). Therefore, these marks, including H3K27ac and H3K9ac in the upstream regions of genes and others shown in Fig. 9, may have contributed to regulation of mRNA expression in the genome of a common ancestor of metazoans, and may be of ancient origin.

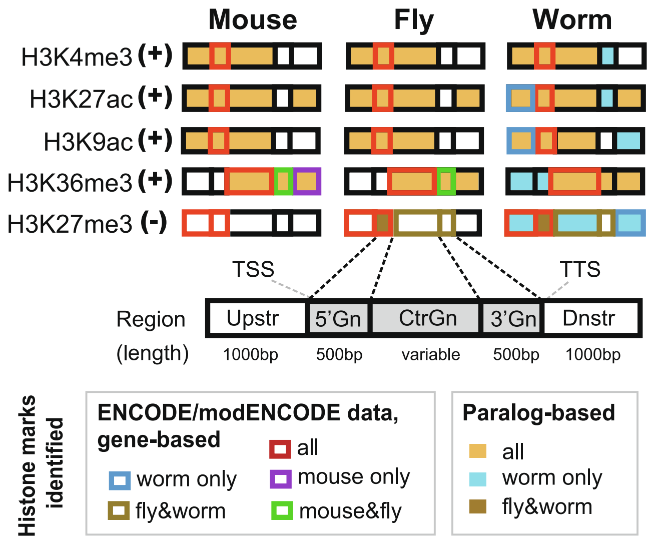

Fig. 9. Summary of our proposed histone mark revisions. The empty colored boxes (red, blue, purple, bronze, or green) mark the acting regions identified based on the modENCODE ChIP-fold enrichment status which is gene-based, while the solid colored boxes (orange, blue, or bronze) mark the acting regions identified using the paralog-based approach described in the present study. The colors used for the boxes indicate the species-specificity (or universality) of the focal histone marks in the specific regions examined. Universal histone marks identified with a gene-based approach based on modENCODE results versus a paralog-based approach are indicated with red empty boxes versus orange solid boxes, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3. Methods

3.1. Paralogous genes and their genome coordinates

Annotated genome assemblies from mouse (*M. musculus*, GRCm38.p6), fly (*D. melanogaster*, BDGP6.28), and worm (*C. elegans*, WBcel235) were obtained from Ensembl [69]. Orthology information and chromosomal coordinates of genes for each of the genomes were downloaded from Ensembl BioMart (http://asia.ensembl.org/biomart/martview/; accessed Jul 2020). In total, 13907, 7876, and 11,818 genes with at least one paralog were identified from the mouse, fly, and worm genomes, respectively. For cases where a single gene may be transcribed into multiple isoforms, TSS and TTS of the longest isoform were used to define "Upstr", "5'Gn", "CtrGn", "3'Gn", and "Dnstr" regions of the gene (Fig. 1A). When definitions of TSS and TTS were based on the foremost 5'-end and hindmost 3'-end positions of all of mapped isoforms, virtually identical results were obtained (Fig. S8). To ensure our analyses were based on genes with sufficient length in the central genic region (>500 nucleotides), genes with a length ≤1500 nucleotides between their TSS and TTS were discarded. Single-exon genes were also discarded to avoid the inclusion of retrogenes in our analyses. After removing these two types of genes, 7489, 1379, and 6034 genes with at least one paralog in the mouse, fly, and worm genomes, respectively, were included in our subsequent analyses.

3.2. RNA-seq and ChIP-seq data

RNA-seq and ChIP-seq data were obtained from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/; accessed Jul 2020) and modENDCODE (http://www.modencode.org/; accessed Jul 2020), respectively. Focusing on developmental stages or tissues which had both RNA-seq and ChIP-seq data available for H3K4me3, H3K27ac, H3K9ac, and H3K36me3, raw RNA-seq and ChIP-seq data (for the four focal histone marks) were downloaded for various tissues (heart, liver, and forebrain/cerebellum) at various developmental stages (day 11.5 embryo, day 16.5 embryo, and 8-week-old adult) of mouse (*M. musculus*), the whole organism at third larval stage (L3) or adult stage of worm (*C. elegans*), and the whole third-instar larva (L3) and adult head of fly (*D. melanogaster*). Accession numbers for the RNA-seq and ChIP-seq data of the abovementioned tissues or stages that were downloaded are listed in Table S2 (worm and fly) and S3 (mouse).

According to standard data preprocessing and quality assessment procedures of ENCODE and modENCODE [17], the RNA-seq and ChIP-seq raw sequences were mapped to their respective genomes (GRCm38.p6 for mouse; BDGP6.28 for fly; WBcel235 for worm) by Bowtie2 (2.2.5) [70] with no allowance for sequence mismatch (-N 0). SAMtools [71] (http://samtools.sourceforge.net; accessed Jul 2020) was used to transform the resulting .sam files into .bam files for subsequent analyses. Paralogous genes could be highly similar in sequence. To accurately estimate RNA-seq and ChIP-seq signals of individual genes with a closely related paralog, Sambamba (0.7.0) was used with the parameter "XS == null" to remove alignment records that derived from reads mapped to multiple regions of the genome. Filtered mapping data of RNA-seq were processed by using the "cufflinks" function of Cufflinks (2.2.1) [72] (http://cole-trapnell-lab.github.io/cufflinks/manual/, accessed Jul 2020) with default parameters used to estimate the mRNA expression abundance for each gene measured as FPKM (values for E_S, E_W or E_G). To estimate the intensity of each histone mark, filtered mapping data generated by ChIP-seq were processed with the "callpeak" function of MACS2 (2.1.2) [73] and the parameters of "--nomodel, --SPMR". The resulting .bdg files were processed according to the "bdgcmp" function of MACS2 with "-m FE" applied to calculate the fold enrichment of histone mark intensities relative to background inferred from the control sample (values for H_S, H_W or H_G) (Tables S2 & S3). We added a pseudo count value of "0.5" to both H_S and H_W when calculating H_S/H_W. We found that the use of alternative pseudo count values did not change the results obtained (Fig. S9).

3.3. Gene essentiality, tissue-specificity, and nucleosome density

A gene is essential when mutations in it cause premature death or infertility of an organism. Otherwise, a gene is considered nonessential. Based on these definitions, 4341 essential genes and 4701 nonessential genes in the mouse genome were obtained from the Online Gene (OGEE) database [74] (http://ogee.medgenius.info; access at Jul 2020). Tissue-specificity calculated as τ based on a gene's expression profile across 26 mouse tissues was computed for each of the mouse genes selected in our previous study [40]. According to these precomputed τ values [40], 3347 tissue-specific genes ($\tau \geq 0.6$) and 4715 non-tissue-specific genes ($\tau \leq 0.1$) from the mouse genome were identified. Nucleosome positioning data for mouse liver were obtained from NucMap [75] (http://bigd.big.ac.cn/nucmap/NucMap_FTP_Directory/; accessed Jul 2020) (represented as iNPS peaks in the sample mmNuc0410101, 3 month adult liver) [76]. The nucleosome density value, *nuc*, for each gene was defined according to the number of "peak(s)" identified between 1 kilobase upstream of the TSS and 1 kilobase downstream of the TSS of the gene.

3.4. Eliminating biases associated with mRNA abundance in comparisons of gene groups

Essential genes tend to be more highly expressed than nonessential genes (Fig. 1B). Similarly, non-tissue-specific genes tend to be more highly expressed than tissue-specific genes (Fig. 1C). Therefore, in order to compute $\Delta H3K4me3_{E\text{-}nE}$ (or $\Delta H3K4me3_{nT\text{-}T}$), the intrinsic difference in mRNA abundance between essential and nonessential genes (or between non-tissue-specific and tissue-specific genes) has to be eliminated by matching the FPKM distributions of the two focal gene groups. To achieve this, we categorized genes according to their FPKM into seven bins: 1) FPKM < 1, 2) 1 ≤ FPKM < 5, 3) 5 ≤ FPKM < 10, 4) 10 ≤ FPKM < 20, 5) 20 ≤ FPKM < 40, 6) 40 ≤ FPKM < 80, and 7) FPKM ≥ 80. Genes assigned to the first (FPKM < 1) and last (FPKM ≥ 80) bins were discarded. The numbers of essential and nonessential genes (or non-tissue-specific and tissue-specific genes) in each FPKM category were subsequently calculated (Fig. S10, A & B). Equivalent numbers of genes for each bin of the compared gene groups were randomly sampled without replacement, as shown in Figure S10A (or Fig. S10B). $\Delta H3K4me3_{E\text{-}nE}$ (or $\Delta H3K4me3_{nT\text{-}T}$) for the genic region of interest was calculated by taking the average H3K4me3 intensity of the resampled essential genes (or non-tissue-specific genes) and subtracting the average H3K4me3 intensity of the resampled nonessential genes (or tissue-specific genes). This process was repeated 500 times to obtain the data presented in Fig. 1D (or Fig. 1E).

Paralog pairs could be similar in their expressed mRNA abundance due to a shared ancestry. Therefore, eliminating the potential influence of mRNA expression similarity was also required when examining homogeneity in τ, nucleosome density, and essentiality, and between paralog pairs versus non-paralog pairs. To do this, we selected 500 sets of gene pairs, each of which consisted of one paralog pair and one non-paralog pair (the members of the non-paralog pairs that were selected had to have a mRNA expression abundance that matched the corresponding paralog pair, according to the assigned FPKM categories [FPKM < 1, 1 ≤ FPKM < 5, 5 ≤ FPKM < 10, 10 ≤ FPKM < 20, 20 ≤ FPKM < 40, 40 ≤ FPKM < 80, or FPKM ≥ 80] of the gene members). The average Δτ (or Δ*nuc*) value for both the paralog pairs and the non-paralog pairs were calculated. When examining homogeneity in essentiality, we calculated the proportion of gene pairs whose gene members are equally essential (P_{eqESS}) for the paralog pairs, based on the subset of paralog pairs that have gene essentiality data for both copies. The FPKM distribution for gene members of this subset was calculated (Fig. S11A) and compared with the reference distribution for the selected non-paralog pairs with matched expression abundance (Fig S11B) (Fig. 2C).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by intramural funding from the National Health Research Institutes [grant number: PH-110-PP-03]; and research grants from the Ministry of Science and Technology (MOST 105-2314-B-400-021-MY4 and 109-2628-B-400-005) to B.Y.L.

Author contributions

B-Y.L. conceived the study. K.-Y.L. performed analyses. K.-Y.L and B-Y.L. designed experiments and wrote the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.12.027.

References

[1] Van Hoide KE, Sahasrabuddhe CG, Shaw BR. A model for particulate structure in chromatin. Nucleic Acids Res 1974;1(11):1579–86.
[2] Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 1997;389:251–60.
[3] Kouzarides T. Chromatin modifications and their function. Cell 2007;128 (4):693–705.
[4] Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, et al. Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet 2008;40(7):897–903.
[5] Strahl BD, Allis CD. The language of covalent histone modifications. Nature 2000;403(6765):41–5.
[6] Wargo MJ, Rizzo PJ. Exception to eukaryotic rules. Science 2001;294 (5551):2477.
[7] Rando OJ. Combinatorial complexity in chromatin structure and function: revisiting the histone code. Curr Opin Genet Dev 2012;22(2):148–55.
[8] Turner BM. Cellular memory and the histone code. Cell 2002;111(3):285–91.
[9] Smith AE, Chronis C, Christodoulakis M, Orr SJ, Lea NC, et al. Epigenetics of human T cells during the G0->G1 transition. Genome Res 2009;19:1325–37.
[10] Zhou G-L, Liu D-P, Liang C-C. Memory mechanisms of active transcription during cell division. BioEssays 2005;27(12):1239–45.
[11] Pannell D, Osborne CS, Yao SY, Sukonnik T, Pasceri P et al. (2000) Retrovirus vector silencing is de novo methylase independent and marked by a repressive histone code. EMBO J 19: 5884-5894.
[12] Wu M, Zhang Ye, Wu N-H, Shen Y-F. Histone marks and chromatin remodelers on the regulation of neurogenin1 gene in RA induced neuronal differentiation of P19 cells. J Cell Biochem 2009;107(2):264–71.
[13] Dattani A, Kao D, Mihaylova Y, Abnave P, Hughes S, Lai A, et al. Epigenetic analyses of planarian stem cells demonstrate conservation of bivalent histone modifications in animal stem cells. Genome Res 2018;28(10):1543–54.
[14] Portela A, Esteller M. Epigenetic modifications and human disease. Nat Biotechnol 2010;28(10):1057–68.
[15] Chang A-F, Liao B-Y. Recruitment of histone modifications to assist mRNA dosage maintenance after degeneration of cytosine DNA methylation during animal evolution. Genome Res 2017;27(9):1513–24.
[16] Prakash K, Fournier D. Evidence for the implication of the histone code in building the genome structure. Biosystems 2018;164:49–59.
[17] Ho JWK, Jung YL, Liu T, Alver BH, Lee S, Ikegami K, et al. Comparative analysis of metazoan chromatin organization. Nature 2014;512(7515):449–52.
[18] Riddle NC, Minoda A, Kharchenko PV, Alekseyenko AA, Schwartz YB, Tolstorukov MY, et al. Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin. Genome Res 2011;21 (2):147–63.
[19] Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J. Differential chromatin marking of introns and expressed exons by H3K36me3. Nat Genet 2009;41(3):376–81.
[20] Carmel L, Koonin EV. A universal nonmonotonic relationship between gene compactness and expression levels in multicellular eukaryotes. Genome Biol Evol 2009;1:382–90.
[21] Ruthenburg AJ, Allis CD, Wysocka J. Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark. Mol Cell 2007;25 (1):15–30.
[22] Vermeulen M, Mulder KW, Denissov S, Pijnappel WWMP, van Schaik FMA, Varier RA, et al. Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. Cell 2007;131(1):58–69.
[23] Lauberth S, Nakayama T, Wu X, Ferris A, Tang Z, Hughes S, et al. H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. Cell 2013;152(5):1021–36.
[24] Chen H, Li H, Liu F, Zheng X, Wang S, Bo X, et al. An integrative analysis of TFBS-clustered regions reveals new transcriptional regulation models on the accessible chromatin landscape. Sci Rep 2015;5(1):8465.
[25] Wang Su, Zang C, Xiao T, Fan J, Mei S, Qin Q, et al. Modeling cis-regulation with a compendium of genome-wide histone H3K27ac profiles. Genome Res 2016;26(10):1417–29.
[26] Gates LA, Shi J, Rohira AD, Feng Q, Zhu B, Bedford MT, et al. Acetylation on histone H3 lysine 9 mediates a switch from transcription initiation to elongation. J Biol Chem 2017;292(35):14456–72.
[27] Gao Y, Chen L, Han Y, Wu F, Yang W-S, Zhang Z, et al. Acetylation of histone H3K27 signals the transcriptional elongation for estrogen receptor alpha. Commun Biol 2020;3(1). https://doi.org/10.1038/s42003-020-0898-0.
[28] Venkatesh S, Workman JL. Set2 mediated H3 lysine 36 methylation: regulation of transcription elongation and implications in organismal development. Wiley Interdiscip Rev Dev Biol 2013;2(5):685–700.
[29] Teissandier A, Bourc'his D. Gene body DNA methylation conspires with H3K36me3 to preclude aberrant transcription. EMBO J 2017;36:1471–3.
[30] Jenuwein T, Allis CD. Translating the histone code. Science 2001;293 (5532):1074–80.
[31] Zentner GE, Henikoff S. Regulation of nucleosome dynamics by histone modifications. Nat Struct Mol Biol 2013;20(3):259–66.
[32] Suganuma T, Workman JL. Signals and combinatorial functions of histone modifications. Annu Rev Biochem 2011;80(1):473–99.
[33] Litt MD, Simpson M, Gaszner M, Allis CD, Felsenfeld G. Correlation between histone lysine methylation and developmental changes at the chicken beta-globin locus. Science 2001;293:2453–5.
[34] Karlic R, Chung H-R, Lasserre J, Vlahovicek K, Vingron M. Histone modification levels are predictive for gene expression. Proc Natl Acad Sci U S A 2010;107 (7):2926–31.
[35] Cheng C, Yan K-K, Yip KY, Rozowsky J, Alexander R, Shou C, et al. A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. Genome Biol 2011;12(2):R15.
[36] Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, et al. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. Nature 2011;471(7339):480–5.
[37] Lai WK, Buck MJ (2013) An integrative approach to understanding the combinatorial histone code at functional elements. Bioinformatics 29: 2231-2237.
[38] Stolc V, Gauhar Z, Mason C, Halasz G, van Batenburg MF, Rifkin SA, et al. A gene expression map for the euchromatic genome of Drosophila melanogaster. Science 2004;306(5696):655–60.
[39] Chen H, Zhang Z, Jiang S, Li R, Li W et al. (2020) New insights on human essential genes based on integrated analysis and the construction of the HEGIAP web-based platform. Brief Bioinform 21: 1397–1410.
[40] Liao B-Y, Weng M-P. Unraveling the association between mRNA expressions and mutant phenotypes in a genome-wide assessment of mice. Proc Natl Acad Sci U S A 2015;112(15):4707–12.
[41] van Leeuwen F, van Steensel B. Histone modifications: from genome-wide maps to functional insights. Genome Biol 2005;6:113.
[42] Diss G, Ascencio D, DeLuna A, Landry CR. Molecular mechanisms of paralogous compensation and the robustness of cellular networks. J Exp Zool B Mol Dev Evol 2014;322:488–99.
[43] Kabir M, Wenlock S, Doig AJ, Hentges KE. The essentiality status of mouse duplicate gene pairs correlates with developmental co-expression patterns. Sci Rep 2019;9:3224.
[44] Baertsch R, Diekhans M, Kent WJ, Haussler D, Brosius J. Retrocopy contributions to the evolution of the human genome. BMC Genomics 2008;9 (1):466. https://doi.org/10.1186/1471-2164-9-466.
[45] Kang LF, Zhu ZL, Zhao Q, Chen LY, Zhang Z. Newly evolved introns in human retrogenes provide novel insights into their evolutionary roles. BMC Evol Biol 2012;12:128.
[46] Li J, Yuan Z, Zhang Z, Zhang J. The cellular robustness by genetic redundancy in budding yeast. PLoS Genet 2010;6(11):e1001187.
[47] Tseng WC, Munisha M, Gutierrez JB, Dougan ST. Establishment of the vertebrate germ layers. Adv Exp Med Biol 2017;953:307–81.
[48] Benayoun BA, Pollina E, Ucar D, Mahmoudi S, Karra K, Wong E, et al. H3K4me3 breadth is linked to cell identity and transcriptional consistency. Cell 2014;158 (3):673–88.
[49] Howe FS, Fischl H, Murray SC, Mellor J. Is H3K4me3 instructive for transcription activation? BioEssays 2017;39(1):1–15.
[50] 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.
[51] Pérez-Lluch S, Blanco E, Tilgner H, Curado J, Ruiz-Romero M, Corominas M, et al. Absence of canonical marks of active chromatin in developmentally regulated genes. Nat Genet 2015;47(10):1158–67.
[52] Hsu C-C, Shi J, Yuan C, Zhao D, Jiang S, Lyu J, et al. Recognition of histone acetylation by the GAS41 YEATS domain promotes H2A.Z deposition in non-small cell lung cancer. Genes Dev 2018;32(1):58–69.
[53] Tang J, Chisholm SA, Yeoh LM, Gilson PR, Papenfuss AT, Day KP, et al. Histone modifications associated with gene expression and genome accessibility are dynamically enriched at Plasmodium falciparum regulatory sequences. Epigenetics & Chromatin 2020;13(1). https://doi.org/10.1186/s13072-020-00365-5.
[54] Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, Lee TI, et al. Genome-wide map of nucleosome acetylation and methylation in yeast. Cell 2005;122(4):517–27.
[55] Du Z, Li H, Wei Q, Zhao X, Wang C, Zhu Q, et al. Genome-wide analysis of histone modifications: H3K4me2, H3K4me3, H3K9ac, and H3K27ac in Oryza sativa L. Japonica. Mol Plant 2013;6(5):1463–72.
[56] Zhang W, Garcia N, Feng Y, Zhao H, Messing J. Genome-wide histone acetylation correlates with active transcription in maize. Genomics 2015;106(4):214–20.

[57] Charron JB, He H, Elling AA, Deng XW (2009) Dynamic landscapes of four histone modifications during deetiolation in Arabidopsis. Plant Cell 21: 3732–3748.

[58] Vilborg A, Sabath N, Wiesel Y, Nathans J, Levy-Adam F, Yario TA, et al. Comparative analysis reveals genomic features of stress-induced transcriptional readthrough. Proc Natl Acad Sci U S A 2017;114(40):E8362–71.

[59] Huang H, Weng H, Zhou K, Wu T, Zhao BS, Sun M, et al. Histone H3 trimethylation at lysine 36 guides m(6)A RNA modification co-transcriptionally. Nature 2019;567(7748):414–9.

[60] Pu M, Ni Z, Wang M, Wang X, Wood JG, Helfand SL, et al. Trimethylation of Lys36 on H3 restricts gene expression change during aging and impacts life span. Genes Dev 2015;29(7):718–31.

[61] Sen P, Dang W, Donahue G, Dai J, Dorsey J, Cao X, et al. H3K36 methylation promotes longevity by enhancing transcriptional fidelity. Genes Dev 2015;29 (13):1362–76.

[62] Jungreis I, Lin MF, Spokony R, Chan CS, Negre N, Victorsen A, et al. Evidence of abundant stop codon readthrough in Drosophila and other metazoa. Genome Res 2011;21(12):2096–113.

[63] Li C, Zhang J, Duret L. Stop-codon read-through arises largely from molecular errors and is generally nonadaptive. PLoS Genet 2019;15(5):e1008141.

[64] Wang H, Wang L, Erdjument-Bromage H, Vidal M, Tempst P, Jones RS, et al. Role of histone H2A ubiquitination in Polycomb silencing. Nature 2004;431 (7010):873–8.

[65] Stock JK, Giadrossi S, Casanova M, Brookes E, Vidal M, Koseki H, et al. Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells. Nat Cell Biol 2007;9(12):1428–35.

[66] Kimura H. Histone modifications for human epigenome analysis. J Hum Genet 2013;58(7):439–45.

[67] Hosogane M, Funayama R, Shirota M, Nakayama K. Lack of transcription triggers H3K27me3 accumulation in the gene body. Cell Rep 2016;16 (3):696–706.

[68] Jadhav U, Nalapareddy K, Saxena M, O'Neill NK, Pinello L, Yuan G-C, et al. Acquired tissue-specific promoter bivalency is a basis for PRC2 necessity in adult cells. Cell 2016;165(6):1389–400.

[69] Yates AD, Achuthan P, Akanni W, Allen J, Allen J, et al. Ensembl 2020. Nucleic Acids Res 2020;48:D682–8.

[70] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9(4):357–9.

[71] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25(16):2078–9.

[72] Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 2012;7(3):562–78.

[73] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol 2008;9(9). https://doi.org/10.1186/gb-2008-9-9-r137.

[74] Chen W-H, Lu G, Chen X, Zhao X-M, Bork P. OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. Nucleic Acids Res 2017;45(D1):D940–4.

[75] Zhao Y, Wang J, Liang F, Liu Y, Wang Q et al. (2019) NucMap: a database of genome-wide nucleosome positioning map across species. Nucleic Acids Res 47: D163-D169.

[76] Chen W, Liu Yi, Zhu S, Green CD, Wei G, Han J-D. Improved nucleosome-positioning algorithm iNPS for accurate nucleosome positioning from sequencing data. Nat Commun 2014;5(1). https://doi.org/10.1038/ncomms5909.