Computational Molecular Biology 2014, Vol. 4, No. 10, 1-17
http://cmb.biopublisher.ca

BioPublisher©

Research Article **Open Access**

Genomic and functional characterization of histone H3 lysine 4 methylation co-localized marks

Jie Lv[1], Hongbo Liu[1], Hui Liu[1], Qiong Wu[1], Yan Zhang[2]

1. School of Life Science and Technology, Harbin Institute of Technology, Harbin, 150001, China
2. College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China

Corresponding author email: kigo@hit.edu.cn (QW); yanyou1225@gmail.com (YZ)

Computational Molecular Biology, 2014, Vol.4, No.10 doi: 10.5376/cmb.2014.04.0010
Received: 07 Sep., 2014
Accepted: 25 Oct., 2014
Published: 14 Nov., 2014

Preferred citation for this article:
Lv et al., 2014, Genomic and functional characterization of histone H3 lysine 4 methylation co-localized marks, Computational Molecular Biology, Vol.4, No.10, 1-17 (doi: 10.5376/cmb.2014.04.0010)

Abstract Histone modifications play important roles in dynamic transcription regulation. In mammals, methylation of lysine 4 in histone H3 (H3K4) is associated with open chromatin environment. From functional genomic perspective, the combinations of methylation co-localized marks in lysine residue 4 of histone H3 (H3K4me) are little studied. The genomic patterns of specific H3K4me co-localized peaks are highly conserved. Additionally, the proteins encoded by genes with co-localization peaks in promoter regions have more partners in protein-protein interaction network. We also found the unbalanced base composition, that is, AT nucleotide is preferred in genomic regions with co-localization H3K4me modifications. Gene Ontology enrichment analysis revealed that genes with specific co-localization modifications in promoter regions are function-specific. We also found the PolII level for different combinations are correlated with the differential methyl accumulation of H3K4. Me1me2me3, the triplet for H3K4me, is associated with tissue specificity. This study helps understanding the genomic features of H3K4me co-localization and the role of H3K4me co-localization in function genomic regulation.

Keywords Histone modifications; Co-localization; Genomic composition; CpG islands; H3K4me

Introduction

In eukaryote, the chromatin is packed by consecutive octamers comprised by basic histone types H2A, H2B, H3 and H4, around which DNA sequences of 147bp are wrapped. The histones can be altered by different post-translational chemical groups, leading to different biological effects. Acetyl, methyl, phosphoryl and ubiquityl are the most common post-translational chemical group types. Straightforwardly, a common question may be raised by researchers: do different histone modifications bring out distinct biological outcomes? The histone code hypothesis may answer the question (Cosgrove and Wolberger, 2005). According to the hypothesis, specific histone modification combination can act coordinately to form a barcode which is read by other outer proteins to bring about various biological effects. Though the "histone code" hypothesis is debated, arising evidences are emerging to support the hypothesis (Fischle et al., 2003). Histone methylation has been associated with activating and repressive functions. In mouse embryonic stem cells, developmental genes are marked both by the activating H3K4me3 and the repressive H3K27me3 ('bivalent') (Mikkelsen et al., 2007; Bernstein et al., 2006; Meissner et al., 2008).

Besides the patterns of different histone modifications at different residues of histones are complex, patterns for different number of methyl groups that modify the same residues are also complex. The ε-amino group of lysines can be mono-, di-, or trimethylated with potentially distinct effects on chromatin structure (Santos-Rosa et al., 2002). In yeast, a H3K4 methyltransferase (SET1) is identified (Liu et al., 2005) and the kinetics of the separation of SET1 from the elongating RNA polymerase is associated with the differential methylation of H3K4. In *Arabidopsis thaliana*, distinct H3K4 methyltransferase complexes contribute to differential accumulation of H3K4 at specific residues. For example, the dysfunction of H3K4 methyltransferase ATX1 can lead to decreased

H3K4me3 and largely unchanged H3K4me2 (Alvarez-Venegas and Avramova, 2005). In contrast, the dysfunction of H3K4 methyltransferase ATX2 can lead to decreased H3K4me2 and largely unchanged H3K4me3 (Saleh et al., 2008). In the three differential methylation states of H3K4, trimethylation seems to be more stable, while mono- and dimethylation are less stable. JARID1 family includes histone demethylases for H3K4 trimethylation (Agger et al., 2008), and the conversion from H3K4 trimethylation to dimethylation or monomethylation is possible. Mono-, di-, or tri-methyl marks in lysine 4 of histone H3 are key epigenetic modifications for regulating gene expression, especially H3K4me3 mark. In addition, CpG islands (CGIs) enriched with H3K4 methylation are unmethylated to facilitate transcription (Lv et al., 2010a; Zhang et al., 2011; Liu et al., 2013; Su et al., 2012). However, different effects may be associated with mono-, di-, or trimethylation of lysine residues. Mono- and di-methyl marks of H3K4me are enriched in intergenic regions such as enhancers which have indirect regulatory roles on gene expression (Heintzman et al., 2007; Roh et al., 2007). Taken together, significant differences of the function exist for different combinations for H3K4 methylation markers depending on the number of the methyl groups, but little is studied on this issue previously.

Significant technological progress has provided unprecedented resolution for genome-wide histone modification mapping (Barski et al., 2007). Several large-scale studies have provided high-resolution histone modification profiles, the most comprehensive ones are from Barski and Wang *et al.* in CD4+ T cells (Barski et al., 2007; Wang et al., 2008). Based on this dataset, we aim to study the specific genomic and other attributes for both four methylation co-localized marks in lysine residue 4 of histone H3 (H3K4), that is, mono- and di-methylation (me1me2), mono- and tri-methylation (me1me3), mono-, di- and tri-methylation (me1me2me3), lastly, di- and tri-methylation (me2me3), with single-localized marks for me1, me2 and me3 as controls. Principally, the number of mapped tags detected for a particular position is proportional to the specific modification level of the corresponding nucleosome. The enriched genomic fragments are considered as 'true' peaks in genomic scale, either span single nucleosome or multiple nucleosomes.

It is unknown what are the distinctions of underlying genomic features for co-localized and single-localized histone methylation modifications. In this study, we characterize the genomic and functional genomic features for four H3K4me co-localization types. Some but all co-localization combinations are more conserved than single-localization controls at a higher-than-expected frequency in and out of transcriptional start sites (TSSs) proximal regions (TPRs). The proteins encoded by the genes overlapping co-localized peaks in TPRs have more protein partners in protein-protein interaction network than those with single-localized peaks. Moreover, co-localization types are distinct with respect to functional categories revealed by Gene Ontology enrichment analysis, suggesting that genes with similar functions may share similar H3K4me co-localization patterns. CpG depletion is more prominent in co-localization related genes than controls. In addition, AT nucleotide-rich is a general feature for co-localized H3K4 methylation regions. Me1me2me3, the triplet version of H3K4me, is found to be prominently associated with tissue specificity. Overall, this study represents an important contribution to the understanding of histone codes (Lv et al., 2010a) and the role of H3K4me co-localization in function genomic regulation.

1 Methods

1.1 Datasets

The histone modification profile of lysine 4 in histone H3 was from Barski et al. (Barski et al., 2007). It was the most comprehensive genome-scale profiling of histone methylation in human. The histone modification dataset was from human G0/G1 CD4+ T cells. In their studies, ChIP-sequencing (ChIP-seq) was used to sequence tags from two ends of genomic fragments digested from microccocal nuclease (MNase). The technology is quantitative and cost-effective for genome-wide histone modification study. Phylogenetic Conserved Elements (PhastCons) annotation file (hg18), RefSeq gene annotation and reference genomic sequences were downloaded from the UCSC Table Browser (Rhead et al., 2010). The phastCons (pC) score was linearly transformed from [0, 1000] to [0, 1]. If a ChIP-seq peak has no overlap with phast Cons data, the conservation value for that peak is zero.

1.2 Peak finding

The concept of co-localization is based on identified peak in this study. The ChIP-seq peak finding procedure is illustrated as below. A negative binomial model for each modification profile was trained to provide FDR control, for the negative binomial model provides a much better fit to the ChIP-seq data than does the Poisson model. FDRs were estimated by modeling the read count in windows using negative binomial distribution. Each chromosome was scanned with the window size of 100bp with window moving consecutively per 25bp. Under the negative binomial model, windows with read counts greater than a user chosen cutoff for *bona fide* binding regions were identified by controlling FDR < 0.05.

1.3 Co-localization peak identification

To classify genome-wide peaks into different co-localized groups: me1me2, me1me2me3, me1me3, me2me3 and controls (single-localized peaks), genomic intervals were compared exhaustively. Ten Overlap rate (OR) cutoffs were considered in parallel. OR = 1.0 is the most stringent co-localized peak cutoff, likewise, OR = 0.1 generates the loosest. Most analysis in this study took OR = 0.5 as a basis if no explicit statement was declared.

1.4 Gene overlapping analysis

To assess the functional genomic attributes for peaks, we associated the co-localized and single-localized peaks with TPRs defined by upstream 1k and downstream 2k around TSSs of any annotated genes. The boundaries for TPRs were suggested by the study of Barski et al.

1.5 Gene Ontology enrichment analysis

RefSeq mRNA IDs of co-localized peaks overlapping with annotated genes were submitted to the DAVID system (Huang da et al., 2009). Only GO terms with reported *p*-values smaller than 10E-3 and met by Bonferroni multiple testing correction cutoffs were extracted.

1.6 Motif analysis

It was interesting to search for enriched sequence patterns for four classes of co-localized peaks overlapping with genes. We used Gibbs Motif Sampler with 3000 iterations powered by cisGenome suite to perform the analysis (Ji et al., 2008). For the generated motifs, only one key motif was considered as the enriched one by performing motif enrichment. To make a fair control, the matched genomic control sequences simulated from corresponding co-localized peak sequences were used. The software configuration was set according to online tutorial.

2 Results

2.1 Genomic element distribution for different localization types

A total of 82,283 peaks were identified by peak detection. When considering the peak percentages in TSS-proximal regions (TPRs, defined by upstream 1k and downstream 2k around TSSs) and non-TPR regions, we find co-localized peaks vary little in the two regions (Supplementary Table 1). Generally, Me1 localizes less in TPRs, while me2 and me3 localize more in TPRs compared with non-TPRs. The number of me2me3 co-localization is the most relative to other co-localization types (6% overall).

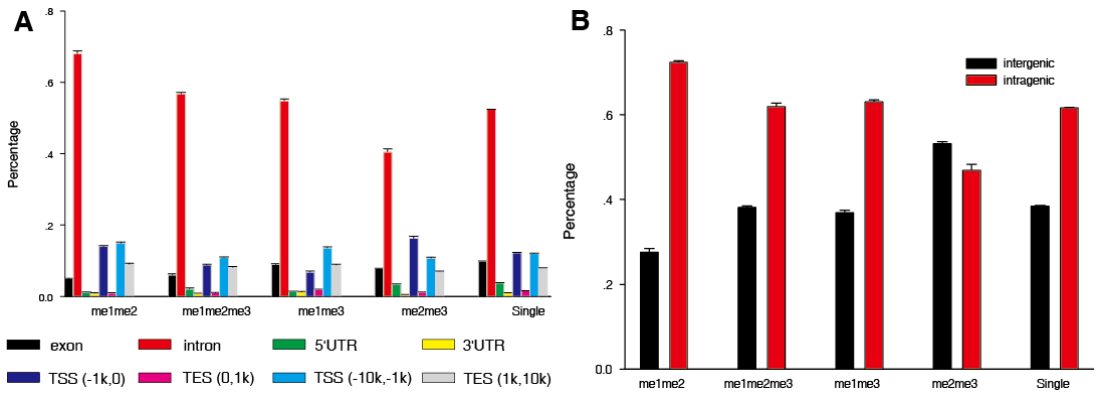


Figure 1 Genomic distribution of co-localized peaks. (A) Distribution of co-localized peaks within gene context. RefSeq genes were used as the gene annotation reference, the genomic annotation is from UCSC Table Browser. (B) Distribution of co-localized peaks and single-localized peaks across genome, where transcriptional Start Sites (TSSs) and Transcriptional End Sites (TESs) are gene boundaries

Then, co- and single-localized peaks overlapping with gene annotation were identified. The summarized box plot was shown in Figure 1 (A). We noted that exon region was not overrepresented in co-localized peaks, compared to single-localized (control) peaks. With respect to intron region, a significant higher percentage than control group could be observed in me1me2 group, suggesting me1me2 was probably a housekeeping mark for gene body and involved in the transcriptional elongation. For any combinations involving me1, the percentage of overlapping introns was higher than others. In addition, 5′ UTR was depleted in co-localized signals, compared with controls. For 3′ UTR element and TES (0k, 10k), all marks followed similar distributions. Compared with single-localized peaks, me1me2 and me2me3 co-localized peaks in regions of 1k upstream of TSSs were found significantly prominent, while me1me2 and me1me3 co-localized peaks were significantly prominent within (-10k, -1k) upstream of TSSs. We found that the me1 and me2 related co-localized groups distributed more than single-localized group within (1k, 10k) downstream of TES, consistent with the previous observation that me1 and me2 signals tended to distribute towards 3′ regions of genes (Zhang et al., 2009). The re-summarized landscape of inter- and intra-gene distributions for Figure 1 (A) was shown in Figure 1 (B). We found that me1me2 located more in intragenic regions and was most overlapped with intron element. Previous studies suggested that the first intron may harbor functional elements to control gene expression (Bradnam and Korf, 2008), which highlighted the potentially regulatory role of me1me2. We noted that co-localized peaks were overrepresented in intron. To unbiasly measure the enrichment of intron in co-localized and single-localized peaks, fold for intron/exon was calculated. The fold for me1me2, me1me2me3, me1me3 was 13.74 ± 0.51, 9.53 ± 0.57, 6.09 ± 0.21 (All p < 1.8E-4), respectively, which were significantly larger than 5.33 ± 0.05 for single-localized peaks. For me2me3, the fold = 5.19 ± 0.15, p=0.0539. From the result, me3-related marks were considered independent of intron localization.

2.2 Co-localized peaks except me1me2 are more phylogenetically conserved than single-localized peaks

It was interesting to explore whether co-localized peaks were more conserved than single-localized peaks. To characterize the conservation of the identified single-localized and co-localized peaks, two phastCons (pC) cutoffs were chosen. For each peak, the average conservation status was averaged for genomic positions with pC score larger than pC cutoff, and finally the peak's conservation was represented by the average pC score. Above all, only peaks overlapping with annotated TPRs were taken into consideration. The high cutoff 0.6 focuses on more conserved peaks, while cutoff of 0.2 just means little conservation. From Table 1, percentages of conserved peaks for different co-localization passed by pC cutoff were shown. To assess the overlap among peaks, Overlap rate (OR) was introduced. OR=1 indicates the given two peaks are completely overlapped. Suppose OR and pC cutoffs be 1.0 and 0.6, respectively. As the cutoff becomes looser, the percentage for all localization types was decreasing. The trend for percentage was straightforward, the varying range with respect to pC cutoff under the same OR cutoff was within [0.05, 0.08], indicating the conservation level was robust. Except me1me2, other types of co-localized peaks were more conserved than single-localized peaks, which was in accord with the fact that Me1 and me2 were not as stable as me3 groups and me3 marks were generally stable. Notably, H3K4me triplet which is the case of three marks co-localizing the same positions was most conserved. As the OR cutoff became stringent, the percentage for H3K4me triplet type became smaller. Even under the most stringent conserved cutoff. When the pC cutoff was chosen as 1.0, the conclusion still held (details see Supplementary Table 2 and Supplementary Table 3). H3K4me3 and H3K27me3 co-localization was previously reported to be even more conserved than either K4 or K27 single-localization in human embryonic stem cells (Zhao et al., 2007), which was consistent with our results.

Table 1 Percentage of overlap PhastCons conservation regions for all H3K4me co-localized peaks in TPRs

OR cutoff	pC cutoff	me1me2	me1me2me3	me1me3	me2me3	Single
0.1	0.2	30%	49%	49%	45%	44%
0.1	0.6	24%	41%	42%	37%	37%
0.5	0.2	30%	48%	48%	45%	44%
0.5	0.6	24%	40%	41%	37%	37%
1.0	0.2	32%	57%	45%	48%	44%
1.0	0.6	25%	49%	40%	41%	36%

Note: OR cutoff is Overlap rate cutoff. pC cutoff is phastCons score cutoff.

Because co-localized peaks except me1me2 were more conserved than single-localized peaks, we used known protein–protein interaction (PPI) data to further confirm the finding. Proteins having multiple partners were considered to be conserved and functionally important. PPI data were collected from a manually curated PPI database: HPRD (Keshava Prasad et al., 2009). From Table 2, we observed that the co-localized peaks had overall more partners than single-localized peaks ($p < 0.05$). Though the average partner number for me1me3 (11.99) was the largest, it was strange that me1me3 was not significant against the control group ($p = 0.07$). We found a protein named with CREBBP with degree of 199 in PPI network, which biased the average degree of me1me3. The *CREBBP* gene was visualized in the UCSC Genome Browser (Rhead et al., 2010), which was shown in Figure 2. In Table 1, H3K4me triplet type seemed the most conserved. But in Table 2, the partner number for triplet type was lower than those of me1me3 and me1me3, which may be caused by the least gene number for the triplet type (gene number : 536).

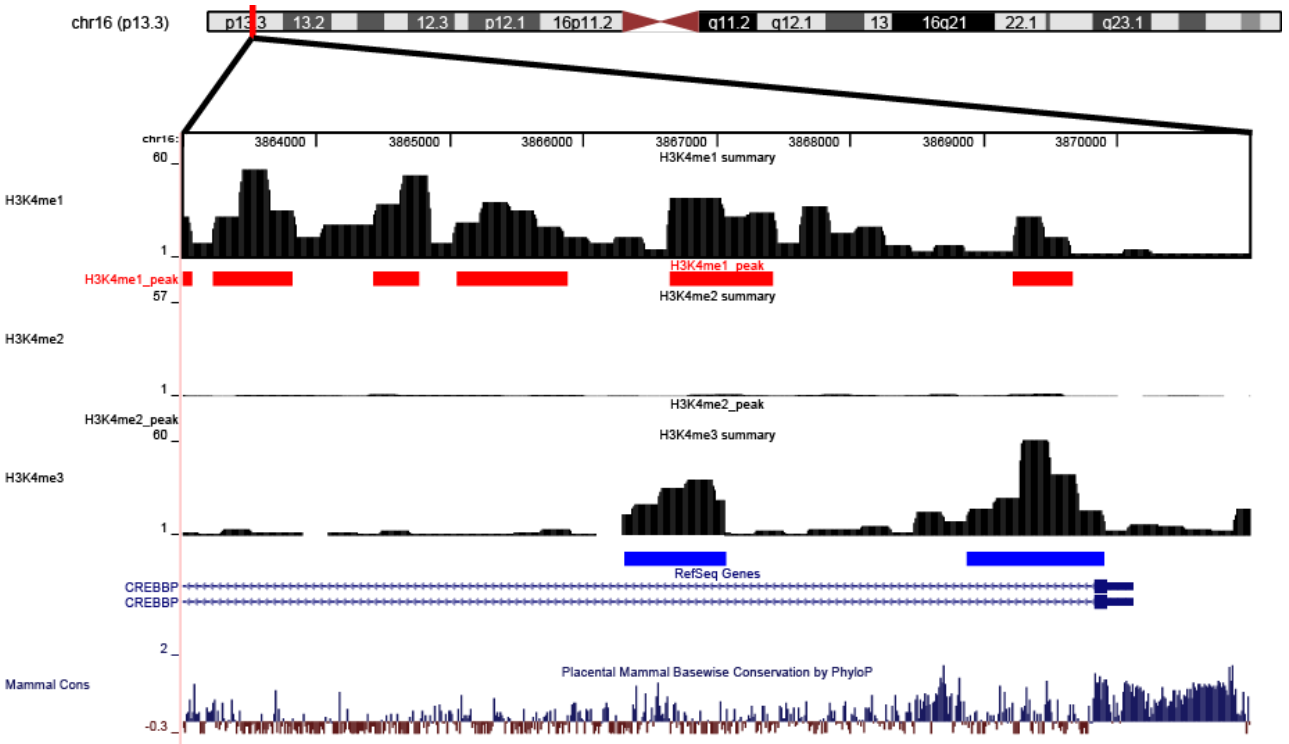


Figure 2 The *CREBBP* gene with surrounding context is displayed as custom tracks on the UCSC genome browser. Red frame indicates the [-1k, 2k] around TSS. The red rectangular track indicates the peaks detected by Cisgenome

Table 2 Number of partners for all H3K4me localization types

	Degree	Gene number with PPI annotation/All gene
me1me2	10.98±20.65	264/688
me1me2me3	10.71±15.94	123/332
me1me3	11.99±21.97	182/451
me2me3	11.01±18.69	401/1021
Average*	11.07±19.57	774/1986
Single	9.38±16.13	4008/10222

Note: * Significant (p=0.02)

2.3 Functional assessment of co-localized peaks using GO terms

Genes associated with co-localized peaks in TPRs (upstream 1k and downstream 2k around TSSs) may be enriched with specific gene functions compared with controls, such as transcription factor binding regulation. To verify this, GO enrichment analysis for genes associated with co-localized peaks in TPRs was performed. We used the DAVID system, which can identify over-represented GO terms for a set of genes (full results in Table 3). A Venn diagram for representing overlapping GO terms was shown in Figure 3A. As was shown in Figure 3A, 11 GO terms are shared by four co-localization types. These terms were supposed to be associated with general cellular component. Among them, two were related with protein binding, three were related with intracellular terms, and another three were related with organelle. Among the four co-localization types, me1me3 was a prominent type in that it has been previously reported that mono-methylation together with marginal tri-methylation is the mark of enhancer (Heintzman et al., 2007). Furthermore, it was also supported by a recent study, in which Transcription Factor Binding Sites (TFBSs) were found to be associated with H3K4me1 and H3K4me3 co-localization (Robertson et al., 2008). In this study, for all six associating GO terms listed in Table 3, all were directly or indirectly related with transcriptional regulation. Interestingly, we observed that the lengths of genes associating with me1me3 (length = 46,521) and me1me2 (length = 45,961) were significantly shorter than the control group (length = 51,615). Our results highlighted the potentially new regulatory role of me1me3 for short genes. In contrast to me1me3, the average gene length was the largest for me2me3 (length = 59,568). Consistent with a previous study in *Arabidopsis thaliana* (Zhang et al., 2009), our result also highlighted me3 mark was a faithful guard for long genes and me1 mark was a guard for short genes.

Table 3 Exclusively enriched GO terms in co-localization types of me1me3, me1me2me3 and me2me3

Localization	GO term	Localization	GO term
me1me3	Transcription, DNA-dependent	me2me3	Cellular component assembly
me1me3	Regulation of metabolic process	me2me3	Macromolecular complex assembly
me1me3	Transcription regulator activity	me2me3	Cell development
me1me3	Regulation of cellular metabolic process	me2me3	Transferase activity, transferring phosphorus-containing groups
me1me3	RNA biosynthetic process	me2me3	Endosome membrane
me1me3	Positive regulation of biological process	me2me3	Endosomal part
me1me2me3	Cell	me2me3	Kinase activity
me1me2me3	Protein kinase cascade	me2me3	Transferase activity
me1me2me3	Transcription repressor activity	me2me3	Cell cycle
me2me3	Protein import into nucleus	me2me3	Protein targeting
me2me3	Endoplasmic reticulum membrane	me2me3	Enzyme binding
me2me3	Positive regulation of programmed cell death	me2me3	Endoplasmic reticulum part
me2me3	Nuclear import	me2me3	Positive regulation of apoptosis
me2me3	Intracellular protein transport	me2me3	Regulation of a molecular function
me2me3	Nuclear envelope-endoplasmic reticulum network	me2me3	Nucleosome
me2me3	Post-translational protein modification	me2me3	Protein import
me2me3	Cell cycle process		

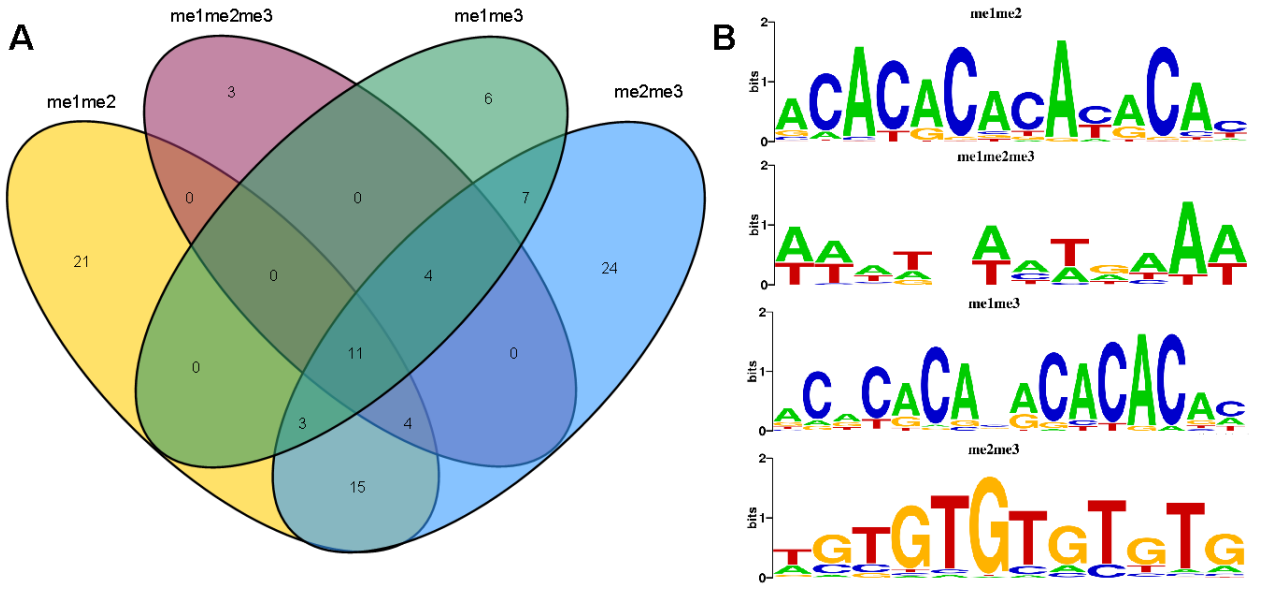


Figure 3 Characterization of co-localized peaks. (A) Venn diagram visualizing the targeting GO terms shared by me1me2, me1me2me3, me1me3 and me2me3 co-localization. (B) Sequence patterns for all types of localized peaks in gene context. Except me1me2me3, all are associated with CA or GT repeat, which is considered to have biological regulatory function

Though mono-, di- and tri-methylation of H3K4 were generally considered to be co-localized in the same regions, the GO term enrichment suggested that the H3K4me triplet type had only three distinct GO terms. The mark was related with signal transduction, cell component, and transcription repression. Contrary to the expectation, 12 out of 332 genes were annotated with transcription repressor activity. We visualized some of them to make a conclusion that several peaks were co-localized in intronic regions, suggesting the regulatory role in intronic regions. Besides introns, other co-localized peaks for these 12 genes were located in TSS upstream regions. Actually, the observation of gene repression by H3K4me mediation was supported by literature. For example, the ING PHD domains provided robust binding modules for H3K4me (Shi et al., 2006). In addition, *PEX14* is one of the 12 genes (see Figure 4 for visualization in UCSC Genome Browser), *PEX14* encoded protein which could function as a transcriptional corepressor and interact with a histone deacetylase (Gavva et al., 2002).

GO terms associated with me2me3 were diverse, which can be interpreted by the fact that me2me3 was a common yet robust signal. We noted a GO term: nucleosome is associated with me2me3 for 13 genes (Supplementary Table 4), implying that me2me3 was a potential signal for activating histone subunits variants and might also be a robust and housekeeping signal for promoters.

Table 4 Nucleotide composition in gene context for different types of localized peaks

Localization type	A	T	G	C	G+C	CpG	TpG	CpG o/e	TpG o/e
me1me2	0.239±0.055	0.241±0.058	0.255±0.061	0.257±0.060	0.514±0.053	0.085±0.032	0.133±0.054	0.670±0.201	1.090±0.300
me1me2me3	0.226±0.050	0.221±0.047	0.272±0.058	0.276±0.058	0.552±0.057	0.103±0.040	0.121±0.051	0.741±0.181	1.022±0.263
me1me3	0.214±0.061	0.216±0.062	0.284±0.070	0.281±0.068	0.563±0.072	0.106±0.053	0.127±0.062	0.749±0.222	1.083±0.304
me2me3	0.229±0.054	0.227±0.054	0.267±0.058	0.267±0.059	0.536±0.064	0.102±0.042	0.116±0.044	0.761±0.207	0.972±0.236
Single	0.205±0.065	0.204±0.065	0.290±0.076	0.289±0.075	0.581±0.095	0.120±0.069	0.114±0.050	0.782±0.229	1.002±0.284

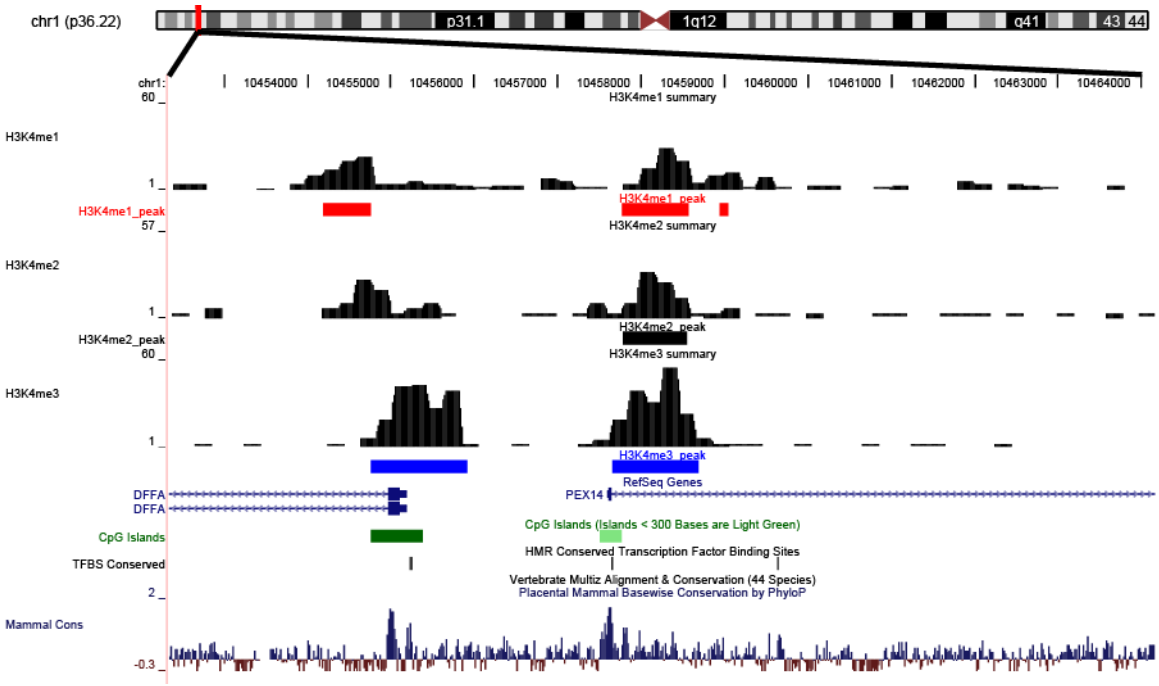


Figure 4 ChIP-seq tag distribution and identified peak information for gene *PEX14* are displayed. Red frame indicates the [-1k, 2k] around TSS. The peak-end track indicates the peaks detected by Cisgenome (Ji et al., 2008). The co-localized peak is localized in intronic region just upstream of a repeat region. The peak in intron is supposed to have regulatory function

2.4 Base composition for co-localized peaks in gene context

Because co-localized peaks differed from single-localized peaks as was analyzed in previous sections, we were curious to investigate the genomic features underlying co-localized peaks. Adenine and thymine nucleotides were more overrepresented in co-localized peaks compared with single-localized peaks (Table 4). Accordingly, guanine and cytosine nucleotides were underrepresented in co-localized peaks. Normally, CpG dinucleotides were rare in vertebrate DNA because the cytosine in such context tended to be methylated then turned into thymines by spontaneous deamination. TpG or CpA would accumulate as deamination products of methylated CpGs, as what was observed in a human α-globin pseudogene (Bird et al., 1987). In Table 4, the TpG content was negatively proportional to the CpG o/e ratio, but the TpG o/e ratio did not have stringently negative tendency against CpG o/e ratio. With respect to TpG o/e ratio, one exception was me2me3 group, which was found to have larger TpG content yet smaller TpG o/e ratio than control. When extending from gene-context to genomic scale, similar results were obtained in Supplementary Table 5. Taken together, co-localization and single-localization peaks had distinct underlying genomic composition.

Table 5 CGI coverage rate for different types of localized peaks

Number (Coverage rate)	TSS-proximal peaks [-1k,2k]	Non-TSS-proximal peaks	All peaks
me1	5948 (0.12±0.3)	33086 (0.03±0.16)	39034(0.04±0.12)
me1me2	1278 (0.18±0.33)	1494 (0.05±0.20)	2772 (0.11±0.13)
me1me2me3	536 (0.34±0.37)	991 (0.06±0.22)	1527 (0.16±0.17)
me1me3	726 (0.35±0.43)	1907 (0.09±0.26)	2633 (0.16±0.12)
me2	3899 (0.35±0.43)	3124 (0.12±0.31)	7023 (0.24±0.20)
me2me3	1860 (0.42±0.33)	2960 (0.10±0.29)	4820 (0.23±0.51)
me3	13212 (0.75±0.41)	11262 (0.33±0.46)	24474 (0.56±0.48)

The definition of CGI is based on some cutoffs including CpG o/e ratio, G+C content and length (Wang and Leung, 2004; Larsen et al., 1992; Gardiner-Garden and Frommer, 1987; Lv et al., 2010b). The overlapping rates and peak overlapping percentages for co-localization and single-localization peaks were both calculated. Consistent with our expectation, the CpG content and CpG o/e ratio for single-localized peaks were all larger than co-localized peaks. In Table 4, CpG o/e ratio was the largest for single-localized peaks. As GC-rich is common in housekeeping genes and GC-poor is common in TPRs of tissue-specific genes, it was straightforward to infer that single- and certain co-localized peaks were general in house-keeping and tissue-specific TSS-proximal regions, respectively. We considered that CGIs were enriched in single-localized peaks, compared with co-localized peaks (Table 5). We did not observe such a trend for single- and co-localized peaks when associating overlapping genes. However, we found the CGI coverage rate increased along with the accumulation of methyl groups both in TPRs and non-TPRs, while TPRs overlapped even more. The observation indicated that H3K4me3 occupied regions overlapped significantly with CGIs. Furthermore, the transcriptional patterns of histone modification combinations were also explored. As PolII is a good proxy for transcription, we used the PolII profile from Barski et al. (Barski et al., 2007) to study the relationship of transcription and H3K4me localization peaks. In Figure 5, we found that genes associated with different combinations of H3K4me localization were expressed at different levels. Consistent with expectation, me1 peaks were least associated with PolII level, me1me3 and me3 were the most (not significant between), while me2 peaks were moderate.

Besides genomic composition, the sequence patterns for co-localized peaks in TPRs context were also explored, which was shown in Figure 3B. We observed that the CA-repeat pattern was overrepresented in the me1me2 and me1me3 types. The GT-repeat pattern, as a complementary type of CA-repeat, was found in me2me3 type. In previous studies, the CA-repeat (GT-repeat) was documented to have regulatory role

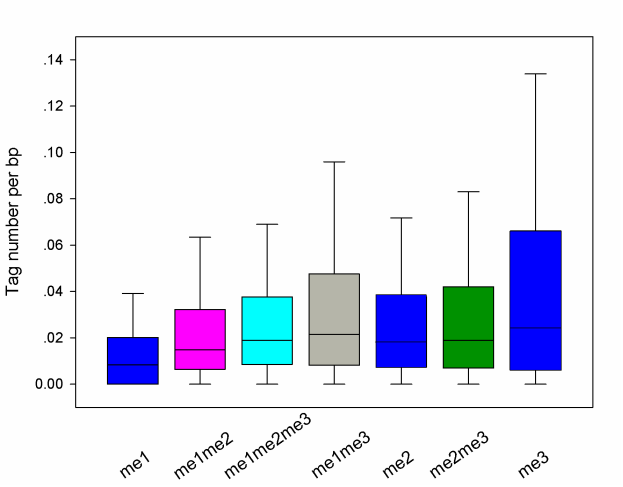


Figure 5 The Average PolII tag number normalized with length for peaks from four types of co-localization and three types of single-localization

that CA RNA elements could function either as splicing enhancers or silencers (Venables, 2007). In a recent study, intronic CA sequences were demonstrated to aid alternative splicing (Hui et al., 2005). Thus, there was a potential association of alternative splicing and histone modifications, especially for specific co-localization peaks. For the H3K4me triplet type, A-rich (T-rich) was found to be associated with tissue-specific genes. From Table 6, the conserved TFBSs for the H3K4me triplet were found more than other combinations. The observation was supported that the H3K4me triplet seemed most conserved (Table 1). Therefore, it was convinced that the H3K4me triplet type was associated with tissue specificity.

Table 6 Conserved TFBS Coverage rates for different types of localizations

Coverage rate	TSS-proximal peaks [-1k,2k]	Non-TSS-proximal peaks	All peaks
me1	0.23	0.13	0.14
me1me2	0.27	0.18	0.22
me1me2me3	0.51	0.29	0.37
me1me3	0.40	0.24	0.28
me2	0.23	0.18	0.21
me2me3	0.42	0.23	0.31
me3	0.43	0.27	0.36

2.5 Co-localization of H3K4me with other histone modification marks

We further correlated all H3K3me co-localization types to other histone modifications derived from the HHMD database (Zhang et al., 2010) to explore chromatin interactions. From Figure 6, we could clearly observe that H3K9me1, H3K27me1 and three H3K79me types were more overlapped with H3K4me than others. Especially, H3K4me1me2 was a more prominent mark to co-localize other modifications, while H3K4me3 was a less likely mark to co-localize with other modifications. H3K9me1 was most likely to co-localize with H3K4me, while H3K27me3 was not found to co-localize with H3K4me. From Figure 6 and Supplementary Figure 1, mono-methylation or mono-/di-methylation for H3K4 were found to be associated more with other mono-methylation modification types.

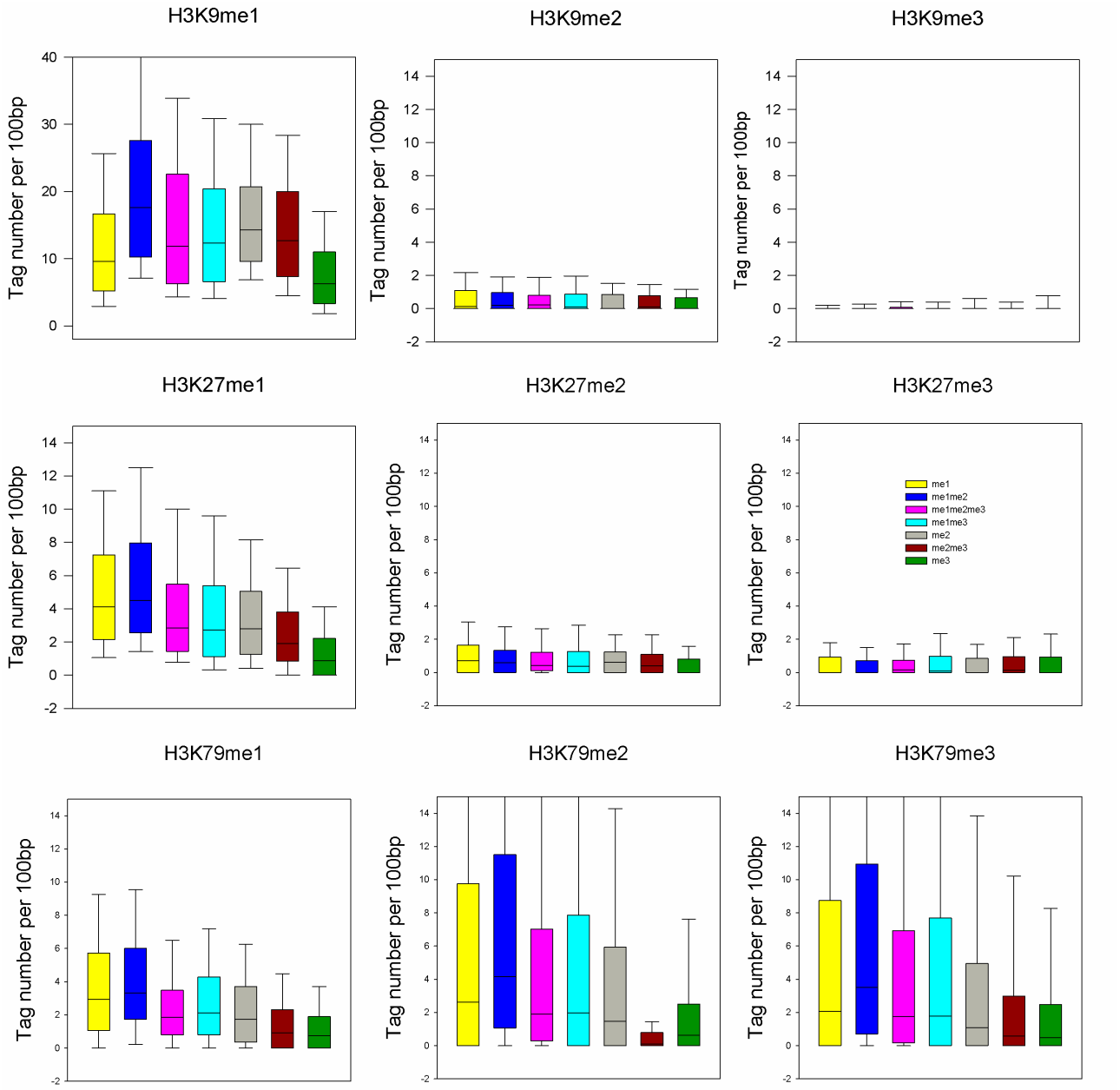


Figure 6 The boxplot for other histone modification tag number normalized with length for peaks from four types of co-localization and three types of single-localization

Previous observations of human stem cells suggested that histone mono-methylation protected the activation potentials required for differentiation (Cui et al., 2009), which seemed to agree with that the data used in this study was sequenced from T cells in G0/G1 phase and such cells were poised for activation, partly explaining why H3K4me1 tended to overlap mono-methylation of other histone modifications.

3 Discussion

This study highlights the specific associations of functional genomic features with different H3K4 methylation modifications. It is known from the study that the differential accumulation of H3K4 at specific genomic loci represents not only the results of enzyme catalysis with dynamic changes, but also specific genomic functions imposed by outer proteins such as histone methyltransferase. Unfortunately, the causal relationship between histone modifications and genomics is still unclear. In this study, we characterize and statistically compare the genomic and functional genomic attributes for different combinations of H3K4 methylation with single-localization types as controls. We find that the distinct H3K4 methylation combinations have distinct underlying genomic backgrounds. Histone modification co-localized peaks tend to mark functionally important regions, such as protein coding regions and regulatory regions. Such redundant placement of histone modifications seems wasteful, but provides a more complex manner to provide extra information beyond importance.

Functionally important regions such as TFBSs and exons tend to be more conserved against the bulk genome, for such regions may be under more selective constraint (Takemaru et al., 1997). To protect such regions, histone modifications can serve as guides for linking outer signals such as enzyme complexes and berried DNA signals. Conserved genomic regions are active regions which need activating epigenetic marks to cross-talk with outer proteins, especially for the three methylation states for H3K4 that can form four distinct types of co-localization. Consistent with prior studies, the co-localization types except me1me2 are more conserved than controls, indicating that co-localizations in H3K4 have underlying conserved genomic contexts, while the causal relationship between co-localization and sequence conservation is elusive. It is possible that distinct methylation modifications can be enriched in peaks within same nucleosome or different nucleosomes within close proximity, from which the recruitment of TrxG and PcG histone methyltransferase complexes are considered frequent. Consistent with our expectation, H3K4me triplet is the most prominent co-localized type, different from three other types of H3K4me in that H3K4me triplet is conserved, related to tissue specificity and is associated with transcription repressor activity, which prompts us that H3K4me triplet may be a novel regulator for balancing activating and repressive transcription, thus H3K4me triplet can be termed as a “trivalent mark”.

An interesting finding from this study is the specific association of co-localization with functional categories which are annotated by GO annotation. me1me3 twins are associated with transcriptional regulation and me2me3 twins involve cell activity. As the me1me2 twins are not robust signals, they are not evident in the GO study. However, most peaks of me1me2 are associated with cell part, binding and cell process. Recent genome-wide histone modification studies indicate that co-localized H3K4me3/H3K27me3 genomic regions exist for various cell types including ES, CD4+ T cells and MEF cells (Barski et al., 2007; Meissner et al., 2008; Roh et al., 2006). As co-localized H3K4me3/H3K27me3 signals specifically pinpoint functional regions poised for differentiation, it is straightforward to propose that co-localized H3K4me signals are largely specific to suppress tissue differentiation owing to the fact that the histone methylation profiles in this study were sequenced from resting T cells and the co-localized H3K4me regions in the T cell lineage do not express. Unfortunately, none of significant GO terms associating with co-localization types is related with tissue differentiation. Generally, H3K4me co-localizations are attributed to repression of T cell-specific genes. But whether the overlapping marks persist when the T cells are activated is not known.

Distinguishing among H3K4me co-localization types that have different functions and fully delimiting the genomic features call for further computational and

experimental efforts. Serving the guide for genome and extra-nucleus information flow, histone modifications can reveal important information for functional genomic studies. Based on the association of histone modification co-localization and gene function, novel measures for gene functional annotation are promising. Gene similarity measures can therefore potentially benefit from epigenetic and expression profiles. Systematic studies of co-localization will hopefully illuminate the mechanisms of the distinct underlying genomic characteristics associating with different co-localizations of H3K4me. Histone modification co-localization may provide cubic targets for chromatin regulation, and further efforts should be paid for functional studies of histone modification co-localization.

Authors' contributions

JL drafted the manuscript and performed the bioinformatics analysis. HBL and HL collected data and pre-preprocessing. QW and YZ conceived of the study, and participated in its design and coordination. All authors read and approved the final manuscript.

Acknowledgments

The authors thank National Natural Science Foundation of China for funding. This work is supported by the National Natural Science Foundation of China [31171383, 31271558, 31371478, 31371334].

References

Agger K., Christensen J., Cloos P.A., and Helin K., 2008, The emerging functions of histone demethylases, Curr Opin Genet Dev, 18**:** 159-168
http://dx.doi.org/10.1016/j.gde.2007.12.003

Alvarez-Venegas R., and Avramova Z., 2005, Methylation patterns of histone H3 Lys 4, Lys 9 and Lys 27 in transcriptionally active and inactive Arabidopsis genes and in atx1 mutants, Nucleic Acids Res, 33**:** 5199-5207
http://dx.doi.org/10.1093/nar/gki830

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
http://dx.doi.org/10.1016/j.cell.2007.05.009

Bernstein B.E., Mikkelsen T.S., Xie X., Kamal M., Huebert D.J., Cuff J., Fry B., Meissner A., Wernig M., Plath K., Jaenisch R., Wagschal A., Feil R., Schreiber S.L., and Lander E.S., 2006, A bivalent chromatin structure marks key developmental genes in embryonic stem cells, Cell, 125**:** 315-326
http://dx.doi.org/10.1016/j.cell.2006.02.041

Bird A.P., Taggart M.H., Nicholls R.D., and Higgs D.R., 1987, Non-methylated CpG-rich islands at the human alpha-globin locus: implications for evolution of the alpha-globin pseudogene, EMBO J, 6**:** 999-1004

Bradnam K.R., and Korf I., 2008, Longer first introns are a general property of eukaryotic gene structure, PLoS ONE, 3**:** e3093
http://dx.doi.org/10.1371/journal.pone.0003093

Cosgrove M.S., and Wolberger C., 2005, How does the histone code work?, Biochem Cell Biol, 83**:** 468-476
http://dx.doi.org/10.1139/o05-137

Cui K., Zang C., Roh T.Y., Schones D.E., Childs R.W., Peng W., and Zhao K., 2009, Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation, Cell Stem Cell, 4**:** 80-93
http://dx.doi.org/10.1016/j.stem.2008.11.011

Fischle W., Wang Y., and Allis C.D., 2003, Binary switches and modification cassettes in histone biology and beyond, Nature, 425**:** 475-479
http://dx.doi.org/10.1038/nature02017

Gardiner-Garden M., and Frommer M., 1987, CpG islands in vertebrate genomes, J Mol Biol, 196**:** 261-282
http://dx.doi.org/10.1016/0022-2836(87)90689-9

Gavva N.R., Wen S.C., Daftari P., Moniwa M., Yang W.M., Yang-Feng L.P., Seto E., Davie J.R., and Shen C.K., 2002, NAPP2, a peroxisomal membrane protein, is also a transcriptional corepressor, Genomics, 79**:** 423-431
http://dx.doi.org/10.1006/geno.2002.6714

Heintzman N.D., Stuart R.K., Hon G., Fu Y., Ching C.W., Hawkins R.D., Barrera L.O., Van Calcar S., Qu C., Ching K.A., Wang W., Weng Z., Green R.D., Crawford G.E., and Ren B., 2007, Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome, Nat Genet, 39**:** 311-318
http://dx.doi.org/10.1038/ng1966

Huang Da W., Sherman B.T., and Lempicki R.A., 2009, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, Nat Protoc, 4**:** 44-57
http://dx.doi.org/10.1038/nprot.2008.211

Hui J., Hung L.H., Heiner M., Schreiner S., Neumuller N., Reither G., Haas S.A., and Bindereif A., 2005, Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing, EMBO J, 24**:** 1988-1998
http://dx.doi.org/10.1038/sj.emboj.7600677

Ji H., Jiang H., Ma W., Johnson D.S., Myers R.M., and Wong W.H., 2008, An integrated software system for analyzing ChIP-chip and ChIP-seq data, Nat Biotechnol, 26**:** 1293-1300
http://dx.doi.org/10.1038/nbt.1505

Keshava Prasad T.S., Goel R., Kandasamy K., Keerthikumar S., Kumar S., Mathivanan S., Telikicherla D., Raju R., Shafreen B., Venugopal A., Balakrishnan L., Marimuthu A., Banerjee S., Somanathan D.S., Sebastian A., Rani S., Ray S., Harrys Kishore C.J., Kanth S., Ahmed M., Kashyap M.K., Mohmood R., Ramachandra Y.L., Krishna V., Rahiman B.A., Mohan S., Ranganathan P., Ramabadran S., Chaerkady R., and Pandey A., 2009, Human Protein Reference Database--2009 update, Nucleic Acids Res, 37**:** D767-772
http://dx.doi.org/10.1093/nar/gkn892

Larsen F., Gundersen G., Lopez R., and Prydz H., 1992, CpG islands as gene markers in the human genome, Genomics, 13**:** 1095-1107
http://dx.doi.org/10.1016/0888-7543(92)90024-M

Liu C.L., Kaplan T., Kim M., Buratowski S., Schreiber S.L., Friedman N., and Rando O.J., 2005, Single-nucleosome mapping of histone modifications in S. cerevisiae, PLoS Biol, 3**:** e328
http://dx.doi.org/10.1371/journal.pbio.0030328

Liu H., Chen Y., Lv J., Zhu R., Su J., Liu X., Zhang Y., and Wu Q., 2013, Quantitative epigenetic co-variation in CpG islands and co-regulation of developmental genes, Sci Rep, 3**:** 2576
http://dx.doi.org/10.1038/srep02576

Lv J., Qiao H., Liu H., Wu X., Zhu J., Su J., Wang F., Cui Y., and Zhang Y., 2010a, Discovering cooperative relationships of chromatin modifications in human T cells based on a proposed closeness measure, PLoS ONE, 5**:** e14219

http://dx.doi.org/10.1371/journal.pone.0014219

Lv J., Su J., Wang F., Qi Y., Liu H., and Zhang Y., 2010b, Detecting novel hypermethylated genes in breast cancer benefiting from feature selection, Comput Biol Med, 40**:** 159-167
http://dx.doi.org/10.1016/j.compbiomed.2009.11.012

Meissner A., Mikkelsen T.S., Gu H., Wernig M., Hanna J., Sivachenko A., Zhang X., Bernstein B.E., Nusbaum C., Jaffe D.B., Gnirke A., Jaenisch R., and Lander E.S., 2008, Genome-scale DNA methylation maps of pluripotent and differentiated cells, Nature, 454**:** 766-770
http://dx.doi.org/10.1038/nature07107

Mikkelsen T.S., Ku M., Jaffe D.B., Issac B., Lieberman E., Giannoukos G., Alvarez P., Brockman W., Kim T.K., Koche R.P., Lee W., Mendenhall E., O'donovan A., Presser A., Russ C., Xie X., Meissner A., Wernig M., Jaenisch R., Nusbaum C., Lander E.S., and Bernstein B.E., 2007, Genome-wide maps of chromatin state in pluripotent and lineage-committed cells, Nature, 448**:** 553-560
http://dx.doi.org/10.1038/nature06008

Rhead B., Karolchik D., Kuhn R.M., Hinrichs A.S., Zweig A.S., Fujita P.A., Diekhans M., Smith K.E., Rosenbloom K.R., Raney B.J., Pohl A., Pheasant M., Meyer L.R., Learned K., Hsu F., Hillman-Jackson J., Harte R.A., Giardine B., Dreszer T.R., Clawson H., Barber G.P., Haussler D., and Kent W.J., 2010, The UCSC Genome Browser database: update 2010, Nucleic Acids Res, 38**:** D613-619
http://dx.doi.org/10.1093/nar/gkp939

Robertson A.G., Bilenky M., Tam A., Zhao Y., Zeng T., Thiessen N., Cezard T., Fejes A.P., Wederell E.D., Cullum R., Euskirchen G., Krzywinski M., Birol I., Snyder M., Hoodless P.A., Hirst M., Marra M.A., and Jones S.J., 2008, Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding, Genome Res, 18**:** 1906-1917
http://dx.doi.org/10.1101/gr.078519.108

Roh T.Y., Cuddapah S., Cui K., and Zhao K., 2006, The genomic landscape of histone modifications in human T cells, Proc Natl Acad Sci U S A, 103**:** 15782-15787
http://dx.doi.org/10.1073/pnas.0607617103

Roh T.Y., Wei G., Farrell C.M., and Zhao K., 2007, Genome-wide prediction of conserved and nonconserved enhancers by histone acetylation patterns, Genome Res, 17**:** 74-81
http://dx.doi.org/10.1101/gr.5767907

Saleh A., Alvarez-Venegas R., Yilmaz M., Le O., Hou G., Sadder M., Al-Abdallat A., Xia Y., Lu G., Ladunga I., and Avramova Z., 2008, The highly similar Arabidopsis homologs of trithorax ATX1 and ATX2 encode proteins with divergent biochemical functions, Plant Cell, 20**:** 568-579
http://dx.doi.org/10.1105/tpc.107.056614

Santos-Rosa H., Schneider R., Bannister A.J., Sherriff J., Bernstein B.E., Emre N.C., Schreiber S.L., Mellor J., and Kouzarides T., 2002, Active genes are tri-methylated at K4 of histone H3, Nature, 419**:** 407-411
http://dx.doi.org/10.1038/nature01080

Shi X., Hong T., Walter K.L., Ewalt M., Michishita E., Hung T., Carney D., Pena P., Lan F., Kaadige M.R., Lacoste N., Cayrou C., Davrazou F., Saha A., Cairns B.R., Ayer D.E., Kutateladze T.G., Shi Y., Cote J., Chua K.F., and Gozani O., 2006, ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression, Nature, 442**:** 96-99
http://dx.doi.org/10.1038/nature04835

Su J., Qi Y., Liu S., Wu X., Lv J., Liu H., Zhang R., and Zhang Y., 2012, Revealing epigenetic patterns in gene regulation through integrative analysis of epigenetic interaction network, Mol Biol Rep, 39**:** 1701-1712
http://dx.doi.org/10.1007/s11033-011-0910-3

Takemaru K., Li F.Q., Ueda H., and Hirose S., 1997, Multiprotein bridging factor 1 (MBF1) is an evolutionarily conserved transcriptional coactivator that connects a regulatory factor and TATA element-binding protein, Proc Natl Acad Sci U S A, 94**:** 7251-7256
http://dx.doi.org/10.1073/pnas.94.14.7251

Venables J.P., 2007, Downstream intronic splicing enhancers, FEBS Lett, 581**:** 4127-4131
http://dx.doi.org/10.1016/j.febslet.2007.08.012

Wang Y., and Leung F.C., 2004, An evaluation of new criteria for CpG islands in the human genome as gene markers, Bioinformatics, 20**:** 1170-1177
http://dx.doi.org/10.1093/bioinformatics/bth059

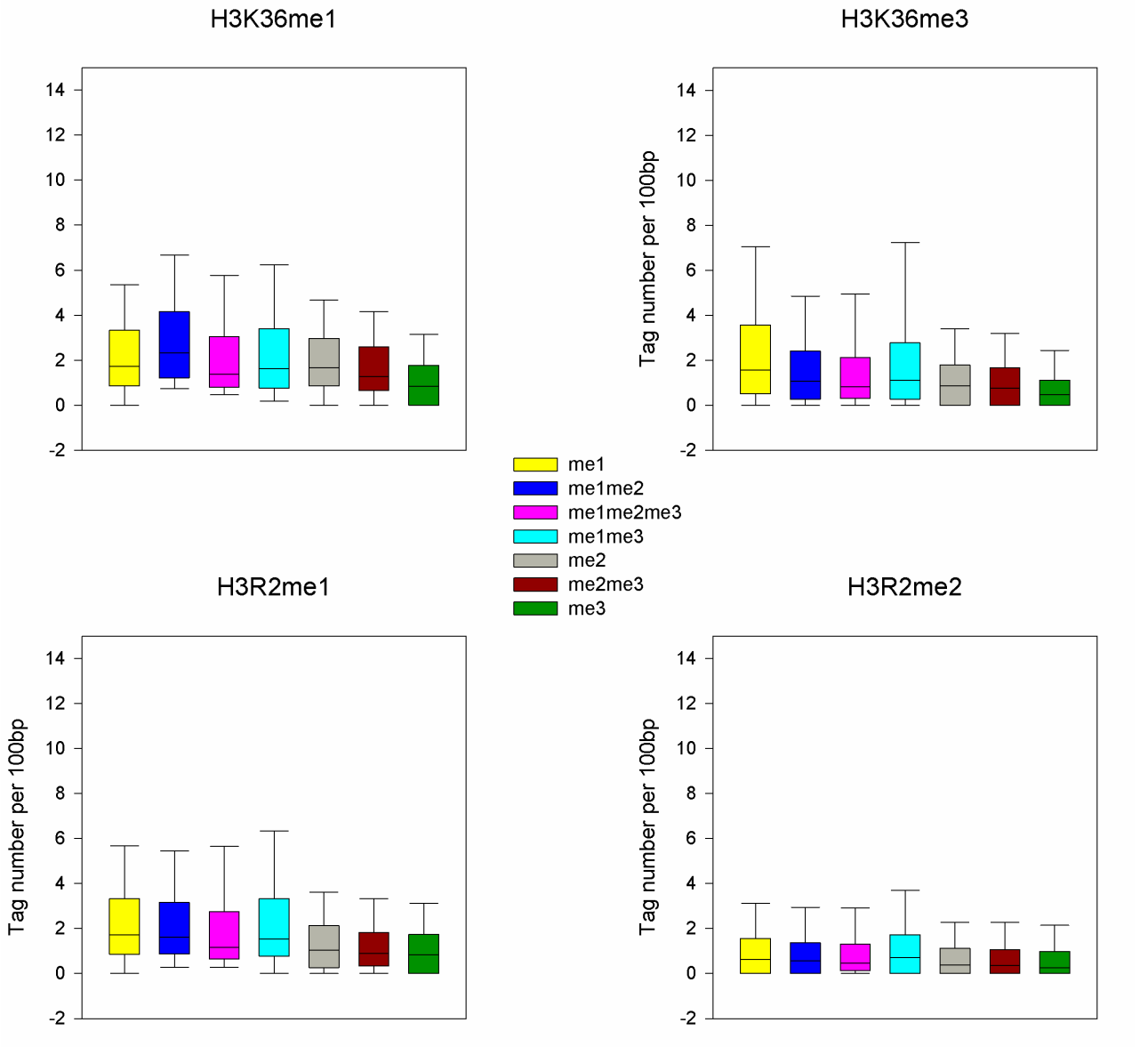
Wang Z., Zang C., Rosenfeld J.A., Schones D.E., Barski A., Cuddapah S., Cui K., Roh T.Y., Peng W., Zhang M.Q., and Zhao K., 2008, Combinatorial patterns of histone acetylations and methylations in the human genome, Nat Genet, 40**:** 897-903
http://dx.doi.org/10.1038/ng.154

Zhang X., Bernatavichute Y.V., Cokus S., Pellegrini M., and Jacobsen S.E., 2009, Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in Arabidopsis thaliana, Genome Biol, 10**:** R62
http://dx.doi.org/10.1186/gb-2009-10-6-r62

Zhang Y., Liu H., Lv J., Xiao X., Zhu J., Liu X., Su J., Li X., Wu Q., Wang F., and Cui Y., 2011, QDMR: a quantitative method for identification of differentially methylated regions by entropy, Nucleic Acids Res, 39**:** e58
http://dx.doi.org/10.1093/nar/gkr053

Zhang Y., Lv J., Liu H., Zhu J., Su J., Wu Q., Qi Y., Wang F., and Li X., 2010, HHMD: the human histone modification database, Nucleic Acids Res, 38**:** D149-154
http://dx.doi.org/10.1093/nar/gkp968

Zhao X.D., Han X., Chew J.L., Liu J., Chiu K.P., Choo A., Orlov Y.L., Sung W.K., Shahab A., Kuznetsov V.A., Bourque G., Oh S., Ruan Y., Ng H.H., and Wei C.L., 2007, Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells, Cell Stem Cell, 1**:** 286-298
http://dx.doi.org/10.1016/j.stem.2007.08.004



Supplementary Figure 1 The boxplot for other histone modification tag number normalized with length for peaks from four types of co-localization and three types of single-localization

Supplementary Table 1 Percentage and peak number for all localization type in TSS-proximal regions, non-TSS-proximal regions and all peaks

	Gene-related peaks	me1	me1me2	me1me2me3	me1me3	me2	me2me3	me3
Distribution of TPRs	Number	5,948	1,278	536	726	3,899	1,860	13,212
	Percentage(Number/27459)	0.22	0.05	0.02	0.03	0.14	0.07	0.48
Distribution of non-TPRs	Number	33,086	1,494	991	1,907	3,124	2,960	11,262
	Percentage(Number/54824)	0.60	0.03	0.02	0.03	0.06	0.05	0.21
Distribution of all peaks	Number	39,034	2,772	1,527	2,633	7,023	4,820	24,474
	Percentage(Number/82283)	0.47	0.03	0.02	0.03	0.09	0.06	0.30

Supplementary Table 2 Percentage of conservation for all localization types based on all TPRs

			OR = 0.1			
pC cutoff	me1me2	me1me2me3	me1me3	me2me3	Average	Single
0.2	0.3	0.49	0.49	0.45	0.4325	0.44
0.4	0.26	0.43	0.45	0.4	0.385	0.39
0.6	0.24	0.41	0.42	0.37	0.36	0.37
0.8	0.21	0.39	0.4	0.35	0.3375	0.34
1.0	0.17	0.33	0.34	0.28	0.28	0.29
			OR = 0.3			
pC cutoff	me1me2	me1me2me3	me1me3	me2me3	Average	Single
0.2	0.3	0.48	0.49	0.45	0.43	0.44
0.4	0.26	0.42	0.45	0.4	0.3825	0.39
0.6	0.24	0.4	0.42	0.37	0.3575	0.37
0.8	0.21	0.38	0.39	0.35	0.3325	0.34
1.0	0.17	0.33	0.34	0.28	0.28	0.29
			OR = 0.5			
pC cutoff	me1me2	me1me2me3	me1me3	me2me3	Average	Single
0.2	0.3	0.48	0.48	0.45	0.4275	0.44
0.4	0.26	0.43	0.44	0.4	0.3825	0.39
0.6	0.24	0.4	0.41	0.37	0.355	0.37
0.8	0.22	0.38	0.39	0.35	0.335	0.34
1.0	0.17	0.33	0.33	0.28	0.2775	0.29
			OR = 0.7			
pC cutoff	me1me2	me1me2me3	me1me3	me2me3	Average	Single
0.2	0.3	0.48	0.47	0.46	0.4275	0.44
0.4	0.26	0.42	0.43	0.41	0.38	0.39
0.6	0.24	0.4	0.41	0.38	0.3575	0.37
0.8	0.21	0.38	0.39	0.35	0.3325	0.34
1.0	0.17	0.32	0.33	0.29	0.2775	0.29
			OR = 1.0			
pC cutoff	me1me2	me1me2me3	me1me3	me2me3	Average	Single
0.2	0.32	0.57	0.45	0.48	0.455	0.44
0.4	0.27	0.51	0.42	0.44	0.41	0.39
0.6	0.25	0.49	0.4	0.41	0.3875	0.36
0.8	0.23	0.46	0.38	0.38	0.3625	0.33
1.0	0.18	0.41	0.32	0.32	0.3075	0.28

Supplementary Table 3 Percentage of conservation for all localization types based on all peaks

				OR=0.1			
pC cutoff	K4/cutoff	me1me2	me1me2me3	me1me3	me2me3	Average	Single
0.2	50	0.26	0.36	0.37	0.37	0.34	0.3
0.4	100	0.22	0.31	0.33	0.33	0.2975	0.26
0.6	150	0.2	0.29	0.3	0.3	0.2725	0.24
0.8	200	0.18	0.26	0.28	0.28	0.25	0.22
1.0	250	0.14	0.22	0.23	0.22	0.2025	0.18
				OR=0.3			
pC cutoff	K4/cutoff	me1me2	me1me2me3	me1me3	me2me3	Average	Single
0.2	50	0.27	0.36	0.36	0.37	0.34	0.3
0.4	100	0.23	0.31	0.32	0.33	0.2975	0.26
0.6	150	0.2	0.28	0.3	0.3	0.27	0.24
0.8	200	0.18	0.26	0.27	0.28	0.2475	0.22
1.0	250	0.15	0.22	0.23	0.22	0.205	0.18
				OR=0.5			
pC cutoff	K4/cutoff	me1me2	me1me2me3	me1me3	me2me3	Average	Single
0.2	50	0.26	0.36	0.36	0.37	0.3375	0.3
0.4	100	0.22	0.31	0.31	0.33	0.2925	0.26
0.6	150	0.2	0.28	0.29	0.3	0.2675	0.24
0.8	200	0.18	0.26	0.26	0.28	0.245	0.22
1.0	250	0.15	0.22	0.22	0.22	0.2025	0.18
				OR=0.7			
pC cutoff	K4/cutoff	me1me2	me1me2me3	me1me3	me2me3	Average	Single
0.2	50	0.26	0.36	0.35	0.37	0.335	0.3
0.4	100	0.22	0.31	0.31	0.33	0.2925	0.26
0.6	150	0.2	0.28	0.29	0.31	0.27	0.24
0.8	200	0.18	0.26	0.26	0.28	0.245	0.22
1.0	250	0.14	0.22	0.22	0.22	0.2	0.18
				OR=1.0			
pC cutoff	K4/cutoff	me1me2	me1me2me3	me1me3	me2me3	Average	Single
0.2	50	0.28	0.45	0.35	0.39	0.3675	0.3
0.4	100	0.23	0.39	0.31	0.35	0.32	0.26
0.6	150	0.21	0.36	0.29	0.32	0.295	0.24
0.8	200	0.19	0.33	0.26	0.3	0.27	0.22
1.0	250	0.15	0.28	0.22	0.24	0.2225	0.18

Supplementary Table 4 Genes associating with *Nucleosome* term in me2me3 colocalization type

RefSeq	Gene symbol	Summary
NM_004893	H2AFY	H2A histone family, member Y
NM_021062	HIST1H2BB	histone cluster 1, H2bb
NM_005633	SOS1	son of sevenless homolog 1 (Drosophila)
NM_001809	C2orf18 CENPA	centromere protein A
NM_033445	HIST3H2A	histone cluster 3, H2a
NM_005320	HIST1H1D	histone cluster 1, H1d
NM_003510	HIST1H2AK	histone cluster 1, H2ak
NM_003540	HIST1H4D HIST1H4F	histone cluster 1, H4f
NM_003518	HIST1H2BG	histone cluster 1, H2bg
NM_003527	HIST1H2BH HIST1H2BO OR2B6	histone cluster 1, H2bo
NM_012412	H2AFV	H2A histone family, member V
NM_003519	HIST1H2BL	histone cluster 1, H2bl
NM_005322	HIST1H1B	histone cluster 1, H1b

Supplementary Table 5 Sequence composition in genomic context

Localization type	A	T	G	C	G+C	CpG	TpG	CpG o/e	TpG o/e
me1me2	0.239±0.055	0.240±0.058	0.256±0.061	0.262±0.061	0.519±0.054	0.082±0.031	0.135±0.058	0.639±0.198	1.122±0.296
me1me2me3	0.234±0.056	0.235±0.054	0.260±0.059	0.264±0.059	0.528±0.061	0.087±0.037	0.133±0.054	0.662±0.191	1.097±0.291
me1me3	0.224±0.060	0.225±0.061	0.274±0.068	0.272±0.067	0.548±0.072	0.095±0.046	0.133±0.062	0.670±0.213	1.129±0.326
me2me3	0.238±0.057	0.236±0.057	0.259±0.060	0.258±0.060	0.519±0.064	0.087±0.039	0.123±0.047	0.688±0.207	1.028±0.266
Single	0.224±0.066	0.224±0.066	0.272±0.073	0.271±0.073	0.542±0.087	0.091±0.059	0.127±0.059	0.675±0.231	1.091±0.324