

RESEARCH

Open Access

Redistribution of H3K4me2 on neural tissue specific genes during mouse brain development

Jie Zhang¹, Jeffrey Parvin^{1,2}, Kun Huang^{1,2*}

From The International Conference on Intelligent Biology and Medicine (ICIBM)
Nashville, TN, USA. 22-24 April 2012

Abstract

Background: Histone modification plays an important role in cell differentiation and tissue development. A recent study has shown that the dimethylation of lysine 4 residue on histone 3 (H3K4me2) marks the gene body area of tissue specific genes in the human CD4+ T cells and neural cells. However, little is known of the H3k4me2 distribution dynamics through the cell differentiation and tissue development.

Results: We applied several clustering methods including K-means, hierarchical and principle component analysis on H3K4me2 ChIP-seq data from embryonic stem cell, neural progenitor cell and whole brain of mouse, trying to identify genes with the H3K4me2 binding on the gene body region in different cell development stage and study their redistribution in different tissue development stages.

A cluster of 356 genes with heavy H3K4me2 labeling in the gene body region was identified in the mouse whole brain tissue using K-means clustering. They are highly enriched with neural system related functions and pathways, and are involved in several central neural system diseases. The distribution of H3K4me2 on neural function related genes follows three distinctive patterns: a group of genes contain constant heavy H3K4me2 marks in the gene body from embryonic stem cell stage through neural progenitor stage to matured brain tissue stage; another group of gene have little H3K4me2 marks until cells mature into brain cells; the majority of the genes acquired H3K4me2 marks in the neural progenitor cell stage, and gain heavy labeling in the matured brain cell stage. Gene ontology enrichment analysis also revealed corresponding gene ontology terms that fit in the scenario of each cell developmental stages.

Conclusions: We investigated the process of the H3K4me2 mark redistribution during tissue specificity development for mouse brain tissue. Our analysis confirmed the previous report that heavy labeling of H3K4me2 in the downstream of TSS marks tissue specific genes. These genes show remarkable enrichment in central neural system related diseases. Furthermore, we have shown that H3K4me2 labeling can be acquired as early as the embryonic stem cell stage, and its distribution is dynamic and progressive throughout cell differentiation and tissue development.

Background

Post-translational modifications of histones play important roles in regulating DNA activities and gene expressions in eukaryotic cells [1-4]. In the nucleus, chromatin forms basic units called nucleosomes which constitute an octamer with eight different histone protein molecules

and a 146bp DNA wrapped around it [1]. Each histone molecule has two tails which contain amino acid residues subject to a variety of modifications such as methylation, acetylation, phosphorylation, and ubiquitination. Such modifications can affect (promote or repress) the accessibility of the DNA to RNA polymerase during gene transcription and thus can activate or silence certain genes [1,3,5]. More importantly, some of such modifications (e.g., the methylation on certain lysine and arginine residues on the N-terminal side) are stable during cell

* Correspondence: kun.huang@osumc.edu

¹The CCC Biomedical Informatics Shared Resource, The Ohio State University, USA

Full list of author information is available at the end of the article

division and considered important epigenetic events. In the past few years, with the rapid progress in high throughput technologies such as microarray and massive parallel sequencing, many exciting new discoveries have been made on studying the genome-wide binding landscapes of different histone modification marks using ChIP-chip and ChIP-seq methods [1,6-11]. In these studies, researchers discovered the combination of different histone modifications mark different chromatin states which often correspond to different genomic annotations such as promoters and enhancers [1,11-13]. Such observations have led to discoveries and predictions of many new regulatory sites on the genome [13,14]. In addition, the ChIP-chip and ChIP-seq technologies enable researchers to investigate the distribution of the histone marks (and other proteins) on the genome at a high resolution, which led to the discovery that the binding patterns of the proteins and histone marks over the genome are also highly informative in predicting genomic functions and annotations [13,15-17]. An interesting discovery along this line is the relationship between H3K4me2 binding patterns and tissue specific genes [17].

H3K4me2 (dimethylation of the lysine residue at 4th position on the N-terminal tail of histone 3) is an important histone mark and has been shown to bind to both gene promoter regions and enhancer regions and its binding is often associated with gene activation. A recent study demonstrated that in human CD4+ T cell and neural tissue, H3K4me2 shows specific binding patterns on the tissue-specific genes in their transcribed regions [17]. In this study, the focus is on a 10kb range of DNA covering 2kb upstream of the transcription starting site (TSS) and 8kb downstream of the TSS for the genes. The H3K4me2 binding patterns over this range for all the genes are then extracted and K-means algorithm is applied to cluster these patterns into five groups. Interestingly, a group distinct itself from others as this group not only shows a bimodal distribution around the TSS, it also uniquely shows high binding quantities (or long tail) over the 8kb downstream of the TSS. Genes in this group are then shown to be highly enriched with tissue specific genes.

However, the recent characterization of histone mark distribution on tissue specific genes was carried out in matured tissue cells, little is known about the dynamic change of histone modification during the cell differentiation and development, neither is it known that at what developmental stage the H3K4me2 marking is acquired. It is generally accepted that the histone modifications play an important role in the cell differentiation and tissue development [18]. Here we are using the public available mouse H3K4me2 ChIP-seq data from different developmental stages of neural cells, namely embryonic stem cells, neural progenitor cells, and whole brain cells [19], and try to address the question that if the same tissue-specific

patterns of H3K4me2 present in the neural cells, and if so, how early it emerges and how it redistributes during the developmental process.

Results

Identify tissue specific genes in the whole brain tissue

K-means clustering of H3K4me2 binding in the ± 10 kb-extended TSS region of brain dataset identified seven clusters each with distinctive H3K4me2 binding pattern in this region (Figure 1A). Among them, cluster 2 has a prolonged tail of H3K4me2 binding in the transcribed region. The feature of the prolonged tail of H3K4me2 binding in the transcribed region was shown by the previous study [17] as the marks for tissue-specific genes. Therefore, we focused on this cluster of genes for our study of tracing tissue-specific H3K4me2 marks in neural cell differentiation and neural tissue development. However since the first stage of K-mean clustering is coarse, it was subject to refined clustering ($k = 4$) to search for the gene subset with more uniformed H3K4me2 binding in this region. The further clustering identified one cluster of 356 genes with substantial H3K4me2 labeling in the transcribed region and a small peak of H3K4me2 upstream of TSS (lower left cluster in Figure 1B). GO enrichment analysis revealed that this cluster of genes are highly enriched neural functions such as transmission of nerve impulse (p-value $1.1E-8$) and synapse function (p-value $1.1E-7$). In addition, 55% of this gene set are phosphoproteins, and participate in various neural system functions, such as synaptic transmission (p-value $1.8E-8$), cell-cell signaling (p-value $4.2E-7$), construct presynaptic membrane (p-value $8.6E-8$), etc. These results are consistent with the finding that heavy presence of H3K4me2 downstream of TSS marks the tissue-specific genes, i.e., brain-specific genes in this case.

H3K4me2 distributions are dynamic over different stages of development

Using the same K-means clustering setting, we further separated ES and NP datasets into different H3K4me2 binding patterns in the extended TSS region (Figure 2). Most of them match the binding profiles of brain tissue, but none of them (except cluster 6 of ES) resembles cluster 2 of brain, which has an elevated H3K4me2 in the downstream of TSS. The above identified 356-gene list overlaps significantly with ES clusters 1 (142), 6 (97) and 7 (73), and with NP clusters 2 (140), 4 (78) and 6 (82) (Figure 2). ES clusters 1, 6 and 7 are enriched with processes and networks involved in nerve system development, such as signal transduction of cAMP (p-value $1.686E-6$), neurogenesis (p-value $7.284E-6$) and nerve impulse transmission (p-value $1.284E-7$) (Figure 3). After checking the average H3K4me2 binding profile in the extended TSS region for those genes shared with

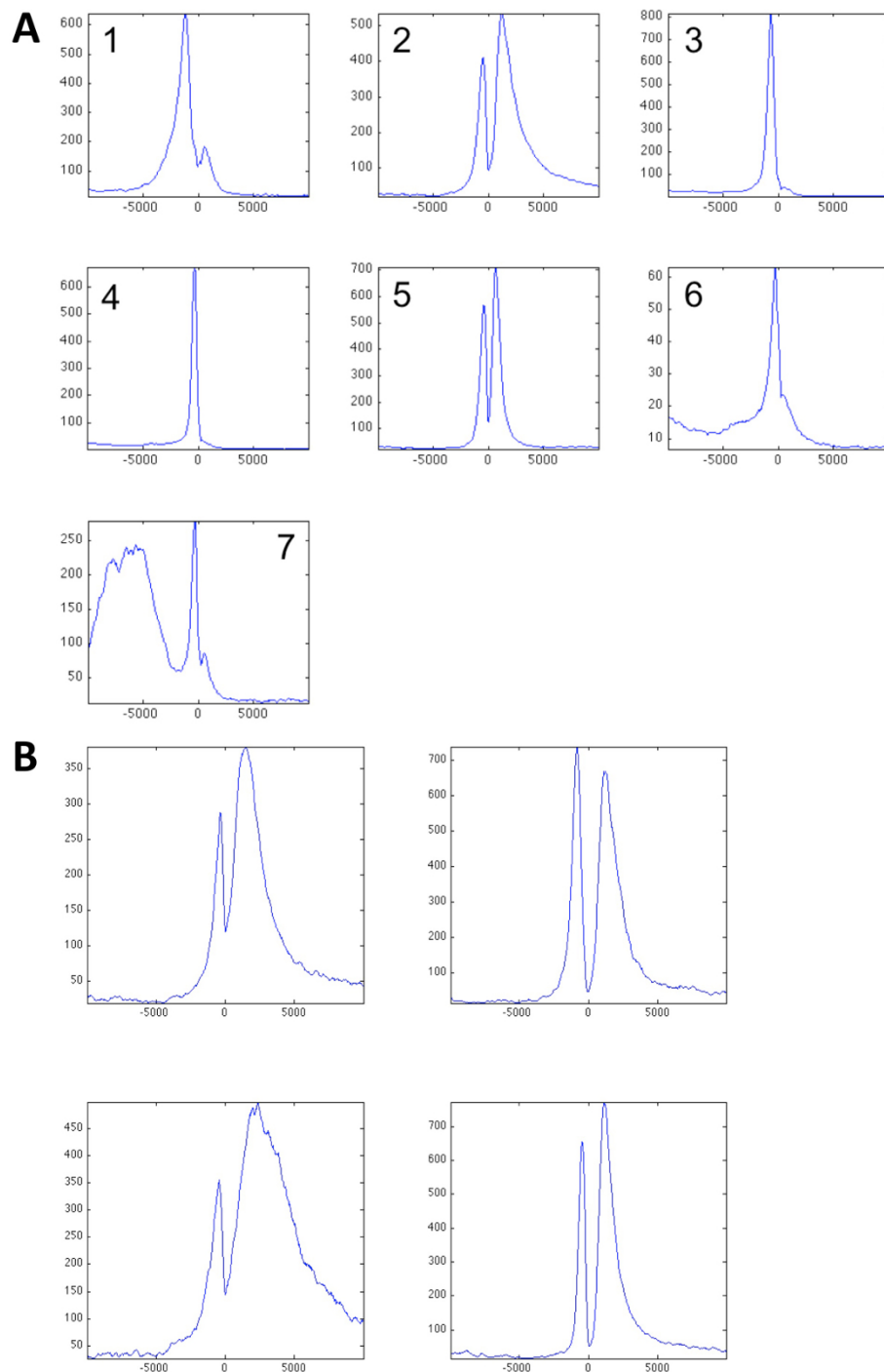
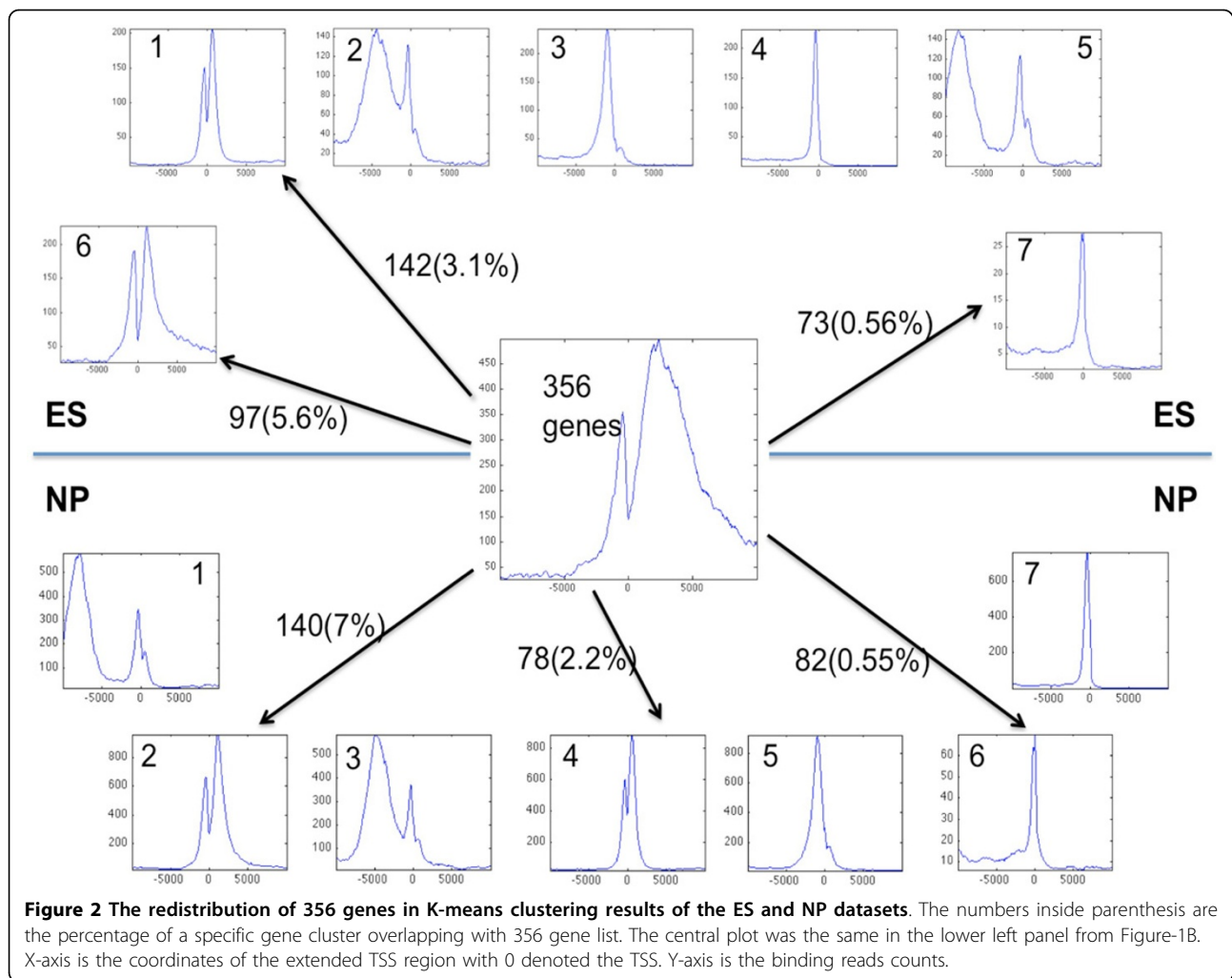


Figure 1 Average H3K4me2 binding showing clusters from K-means clustering of WB in the extended TSS region. A: K-means clustering on the entire gene sets, K=7. B: Further K-means clustering within cluster 2. K=4. The cluster on the lower left of Panel B containing 356 genes was further studied. X-axis is the coordinates of the extended TSS region with 0 denoted the TSS. Y-axis is the binding reads counts. The region extends to +/- 10kbp of TSS. WB: whole brain.

356-gene list (Figure 4), only the cluster 6 of ES (97 common genes) shows H3K4me2 labeling downstream of TSS (panel B in Figure 4), while both the NP cluster 2 (140 common genes) and 6 (82 common genes) show

significant H3K4me2 labeling downstream of TSS (panel D and F in Figure 4). Therefore, the tissue-specific mark is spreading to more genes as cells further differentiate into the neural progenitor stage. The NP cluster 6



contains about half of the entire probeset, so further clustering using K-means was carried out, and it reveals a cluster of 194 genes exhibits heavy H3K4me2 labeling downstream of TSS (data not shown). Interestingly, only 14 were found in the 356-gene list. It demonstrated that the H3K4me2 labeling is a progressive yet dynamic process as cells differentiate from embryonic stem cell through neural progenitor till the final matured brain tissue stage. There are groups of genes with constant H3K4me2 labeling, and groups of genes with H3K4me2 levels change throughout development.

Acquisition of heavy downstream H3K4me2 presence during different stages

To confirm the above observation, we combined the three H3K4me2 TSS binding profiles from ES, NP and WB data for the 356 genes, and performed a hierarchical clustering based on correlation of the binding pattern in this region. Visually the results revealed clearly three distinct H3K4me2 labeling groups as the downstream presence of H3K4me2

of TSS as the development of neural system proceeds from ES through NP to the mature brain stage (Figure 5A). The principle component analysis (PCA) on the same data also confirmed that there exists three well-separated group of genes (Figure 5B). Group 1 contains 74 genes with constant H3K4me2 labeling from ES through NP to matured brain. This group has more diverse H3K4me2 binding features than the other two groups as shown in Figure 5B. GO enrichment analysis shows that this group is enriched with γ -secretase regulation of neuronal cell development and function (p-value 2.313E-9) (Figure 3A). Group 2 contains 73 genes with minimal H3K4me2 labeling in ES and NP stage, but gets heavy H3K4me2 labeling in the matured brain tissue. This group is significantly enriched with cell adhesion of synaptic contact (p-value 5.272E-6) (Figure 6A), and the H3K4me2 binding feature is quite uniform as shown to form a tight cluster in Figure 5B. Group 3 is the largest one containing 209 genes with minimal H3K4me2 labeling in ES stage, but gradually gains H3K4me2 labeling as the cells differentiate into NP cells then mature as brain

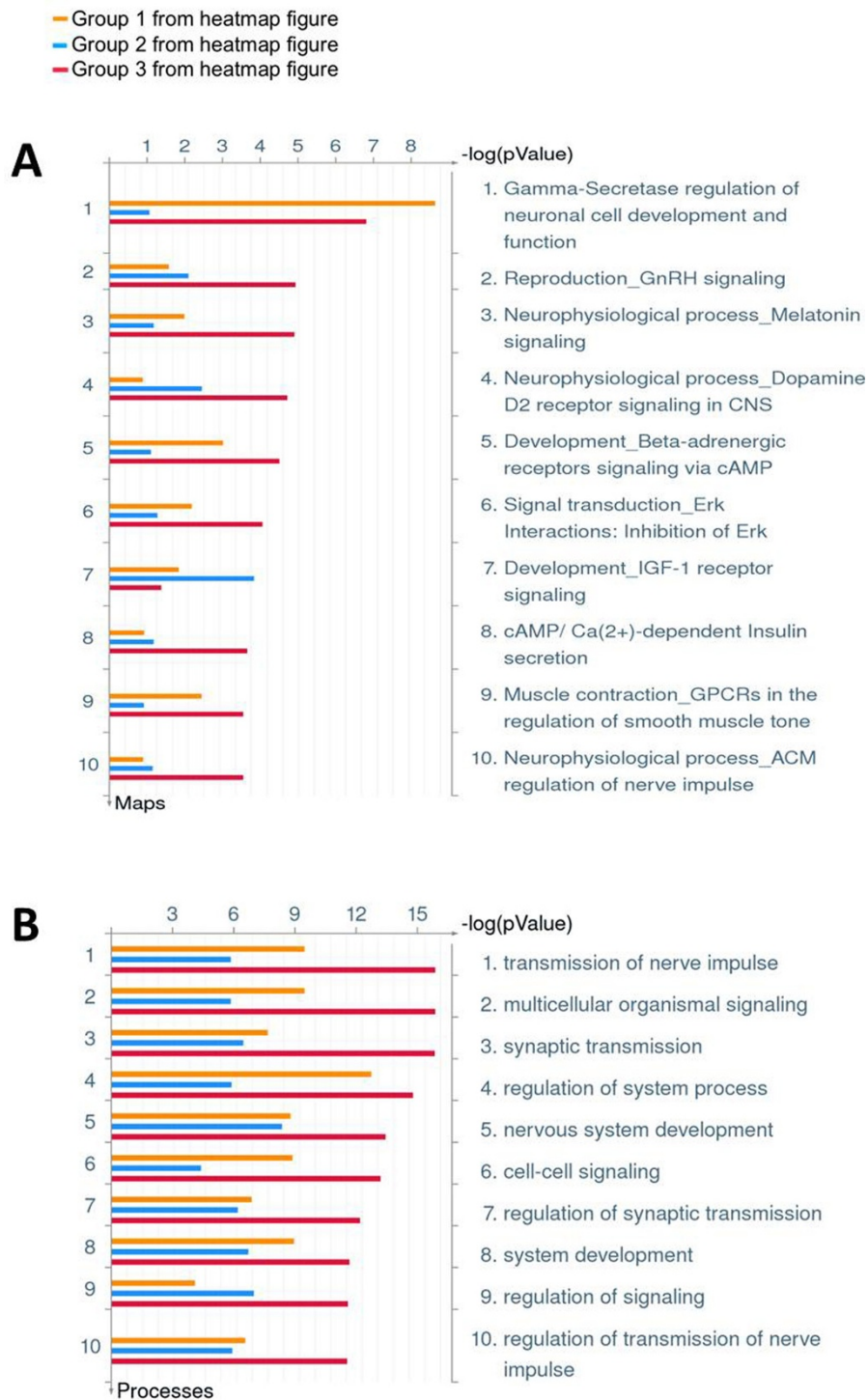
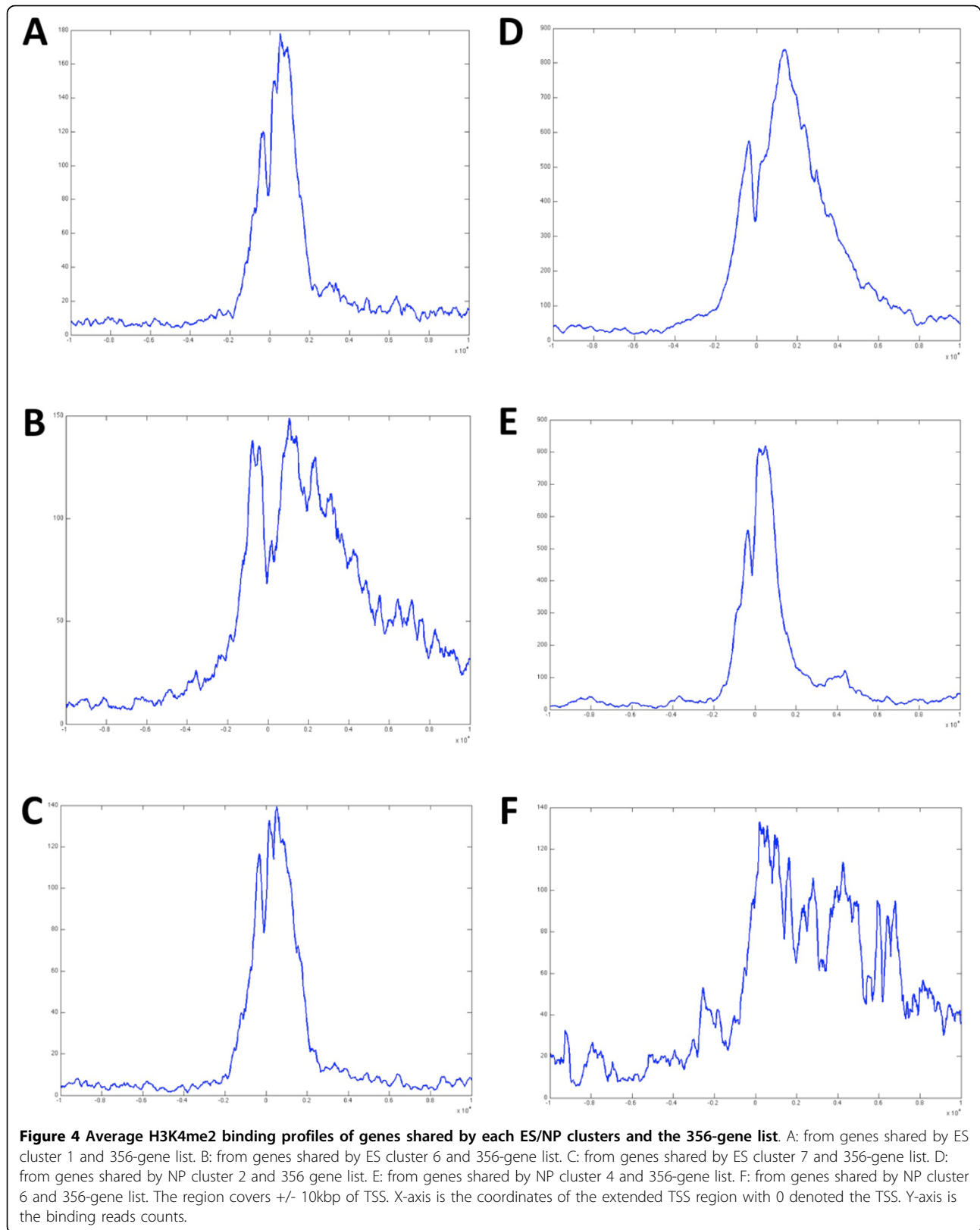


Figure 3 GO enrichment analysis on group 123 identified from hierarchical clustering of the 356-gene H3K4me2 binding of the combined data. A: enriched pathways; B: enriched biological processes. The analysis was performed using MetaCore®.

tissue. It is enriched with several neural transmitter signaling pathways, such as GnRH, melatonin, dopamine D2 receptor, β -adrenergic receptors signaling pathway, and participates in processes such as nerve impulse transmission

(p-value 1.337E-16) and multicellular organismal signaling (p-value 1.337E-16) (Figure 3B) as well as synaptogenesis (p-value 1.451E-7) (Figure 6A). Both groups 1 and 3 are highly enriched with genes involved in several neural degenerative



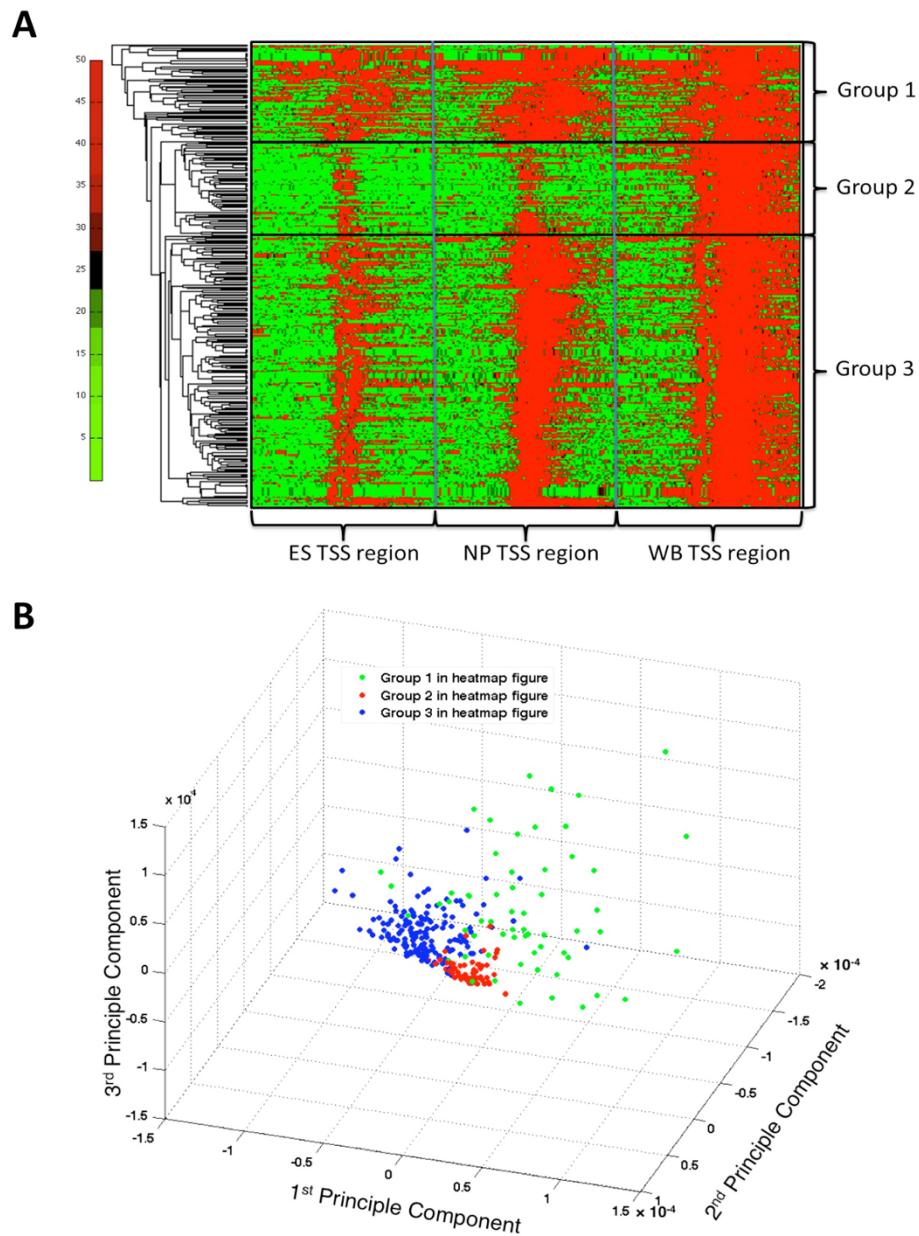


Figure 5 Hierarchical clustering and Principal component analysis (PCA) result on the combined data from ES, NP and WB for H3K4me2 binding for 356-gene list.A: Hierarchical clustering. Each row is a range of +/-10kbp on each side of TSS of a gene. The H3K4me2 binding abundance is shown in the color scale. The H3K4me2 binding patterns of the three groups of genes were indicated by the rectangle boxes. The two vertical bars indicate the range of the extended TSS region from each source. B: PCA analysis. The color dots represent different groups from Panel A. The region covers +/- 10kbp of TSS.

and neurological diseases, such as Alzheimer early onset, epilepsy, tobacco use disorder, schizophrenia, and brain dementia, while cluster 2 genes are the least enriched with such disease genes (Figure 6B).

Discussion

In this paper, we applied detailed clustering analysis on the H3K4me2 distribution patterns around the TSS regions of

genes in the whole genome. Our analysis confirmed the previous report that heavy presence of H3K4me2 in the downstream of TSS marks tissue specific genes. However, more importantly, we have shown that such patterns are not constant. Instead, redistribution of H3K4me2 occurs dynamically in all stages of tissue development.

In addition, our results shed light on understanding the acquisition of tissue specificity during the developmental

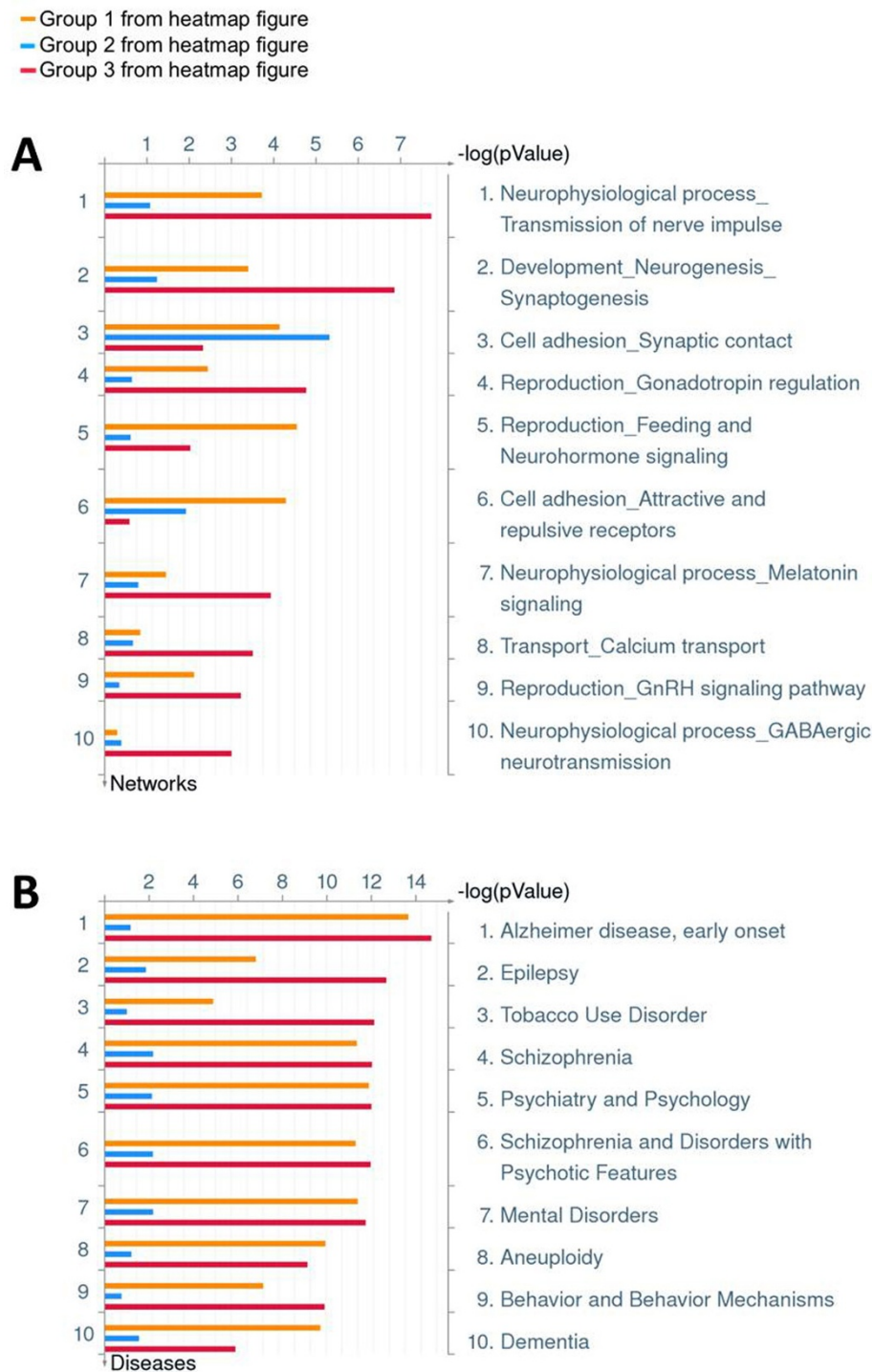


Figure 6 Network and disease enrichment analysis on group 123 identified from hierarchical clustering of the 356-gene H3K4me2 binding of the combined data. A: enriched networks. B: enriched diseases. The analysis was performed using MetaCore®.

process. For the 356 brain tissue specific genes, their H3K4me2 distribution patterns around the TSS regions over the three developmental stages can clearly be divided into three major groups. The GO enrichment

analysis of the three groups of genes indicates that a small portion of neural system-related development and functions was activated or prone to be activated as early as the embryonic stem cell stage, as shown by the

H3K4me2 marks on a small group of genes, whereas most of the neural system-related functions were developed or scheduled to develop through the neural progenitor cell stage, as indicated by the H3K4me2 marks on a larger set of genes in the NP stage. There is a group of genes acquired heavy H3K4me2 labeling downstream of TSS only when the matured brain cell is formed, and consistently cell adhesion activity such as synapse contact formation become the dominate process. This H3K4me2 re-distribution pattern fits very well with the developmental process of neural system.

An interesting observation about these three groups is the gene involvement in neural degenerative and neurological diseases as well as mental disorders including Alzheimer early onset, brain dementia, and schizophrenia. Most of genes associated with these diseases are in the group 1 (with heavy H3K4me2 in ES stage) or group 3 (with heavy H3K4me2 acquired in NP stage). In addition, GO function analysis also suggests that groups 1 and 3 genes are more involved in essential neural functions such as transmission of nerve impulse and synaptic transmission while the group 2 (with heavy H3K4me2 only acquired in the latest stage) are more involved in developing and maintain synapse structures as a mature brain function. These observations imply that these disease-related genes are more involved in the essential neural tissue functions than in the advanced brain activities.

Besides H3K4me2, many other histone marks as well as epigenetic events such as DNA-methylation are also involved and more insight about the differentiation process can be gained with more data available. In addition, the brain tissue data in this study is obtained from whole brain sample even though it is well known that different anatomical and functional regions of the brain have distinctive gene expression patterns and more refined analysis of region and function specific genes could be identified using similar approach as in this paper.

Conclusions

In summary, we investigated the process of the H3K4me2 mark redistribution during tissue specificity development for mouse brain tissue. Our analysis confirmed the previous report that heavy labeling of H3K4me2 in the downstream of TSS marks tissue specific genes. These genes are not only highly enriched with neural system functions and pathways, but also highly involved in CNS diseases. Furthermore, we have shown that such labeling can be acquired as early as the embryonic stem cell stage, and its distribution is dynamic throughout cell differentiation and tissue development. The distribution of H3K4me2 on neural function related genes follows three distinctive patterns: a group of genes contain constant heavy H3K4me2 marks in the gene body from ES through NP to matured

brain tissue stage; another group of gene have little H3K4me2 marks until cells mature into brain cells; the majority of the genes acquired H3K4me2 marks in the NP stage, and gain heavy labeling in the matured brain cell stage. GO enrichment analysis also revealed corresponding GO terms that fit in the scenario of each cell developmental stages.

Methods

Three H3K4me2 ChIP-seq datasets are downloaded from NCBI Gene Expression Omnibus with accession number GSE11172 including murine embryonic stem (ES) cells, ES-derived neural progenitor (NP) cells, and whole brain (WB) tissues [19]. The 36-bp short reads were generated using Illumina GA sequencing platform. In our analysis, we extended the short reads by 200bp, and histograms of the extended reads counts over a 20-kb TSS region (\pm 10kb of TSS) for each entry from the RefSeq mm8 database (obtained from UCSC website) were computed using bin size 20bp. Unsupervised clustering method K-means was used to cluster each individual TSS region profiles from the three origins (ES, NP and WB) using $k = 7$, repeats = 50, distance = square Euclidean as implemented in Matlab. We empirically selected $k = 7$ based on visual inspection of clustering results from different values of k .

In WB tissue, the cluster with an elevated H3K4me2 marking downstream of TSS was further divided into four clusters using K-means ($k = 4$, repeats = 50, distance = square Euclidean) to obtain a cluster of 356 genes with highly elevated H3K4me2 labeling downstream of TSS. This set of genes was compared with clusters obtained from ES and NP samples, and largely overlapped genes and clusters were further subject to GO enrichment analysis using the MetaCore software as well as the NIH DAVID tool.

At the same time, the combined TSS profiles of the 356 genes from all three origins were clustered using hierarchical clustering in Matlab and three clusters were selected based on visual inspection: the cluster with constant H3K4me2 labeling from ES through NP to matured brain; the cluster with minimal H3K4me2 labeling in ES and NP stage, but gets heavy H3K4me2 labeling in brain tissue; and the cluster with minimal H3K4me2 labeling in ES stage, but gradually gains H3K4me2 labeling as the cells differentiate into NP cells and mature as brain tissue. Each cluster of genes was subject to gene set enrichment and network analysis using bioinformatics software MetaCore™ by GeneGo (<http://www.genego.com/meta-core.php>), as well as Database for Annotation, Visualization and Integrated Discovery (DAVID, <http://david.abcc.ncifcrf.gov/>). Principal Component Analysis (PCA) was performed on the same combined data using Matlab code.

List of abbreviations

H3K4me2: dimethylation of lysine 4 residue on histone 3; ES: embryonic stem; NP: neural progenitor; WB: whole brain; GO: gene ontology; TSS: transcription start site; GnRH: Gonadotropin-releasing hormone; ChIP-seq: chromatin immunoprecipitation sequencing; CNS: central nerve system; PCA: principal component analysis

Acknowledgements

This article has been published as part of *BMC Genomics* Volume 13 Supplement 8, 2012: Proceedings of The International Conference on Intelligent Biology and Medicine (ICIBM): Genomics. The full contents of the supplement are available online at <http://www.biomedcentral.com/bmcgenomics/supplements/13/S8>.

This work was supported by NIH 1U01 GM092655-01.

Author details

¹The CCC Biomedical Informatics Shared Resource, The Ohio State University, USA. ²Department of Biomedical Informatics, The Ohio State University, USA.

Authors' contributions

JZ performed all the analysis and wrote the manuscript. JDP participated in the discussion and manuscript editing. KH designed the study and participated in the writing of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Published: 17 December 2012

References

1. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: **High-resolution profiling of histone methylations in the human genome.** *Cell* 2007, **129**(4):823-837.
2. Schmid CD, Bucher P: **ChIP-Seq data reveal nucleosome architecture of human promoters.** *Cell* 2007, **131**(5):831-832, author reply 832-833.
3. Zhang ZD, Paccanaro A, Fu Y, Weissman S, Weng Z, Chang J, Snyder M, Gerstein MB: **Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions.** *Genome research* 2007, **17**(6):787-797.
4. He HH, Meyer CA, Shin H, Bailey ST, Wei G, Wang Q, Zhang Y, Xu K, Ni M, Lupien M, et al: **Nucleosome dynamics define transcriptional enhancers.** *Nature genetics* **42**(4):343-347.
5. Kondo Y, Shen L, Cheng AS, Ahmed S, Bumber Y, Charo C, Yamochi T, Urano T, Furukawa K, Kwabi-Addo B, et al: **Gene silencing in cancer by histone H3 lysine 27 trimethylation independent of promoter DNA methylation.** *Nature genetics* 2008, **40**(6):741-750.
6. Acevedo LG, Iniguez AL, Holster HL, Zhang X, Green R, Farnham PJ: **Genome-scale ChIP-chip analysis using 10,000 human cells.** *BioTechniques* 2007, **43**(6):791-797.
7. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, et al: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447**(7146):799-816.
8. Kim TH, Barrera LO, Ren B: **ChIP-chip for genome-wide analysis of protein binding in mammalian cells.** In *Current protocols in molecular biology* Frederick M Ausubel [et al] 2007, **Chapter 21:Unit 21 13**.
9. Mardis ER: **ChIP-seq: welcome to the new frontier.** *Nature methods* 2007, **4**(8):613-614.
10. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH: **An integrated software system for analyzing ChIP-chip and ChIP-seq data.** *Nature biotechnology* 2008, **26**(11):1293-1300.
11. Yu H, Zhu S, Zhou B, Xue H, Han JD: **Inferring causal relationships among different histone modifications and gene expression.** *Genome research* 2008, **18**(8):1314-1324.
12. Robertson AG, Bilenyk M, Tam A, Zhao Y, Zeng T, Thiessen N, Cezard T, Fejes AP, Wederell ED, Cullum R, et al: **Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding.** *Genome research* 2008, **18**(12):1906-1917.
13. Ucar D, Hu Q, Tan K: **Combinatorial chromatin modification patterns in the human genome revealed by subspace clustering.** *Nucleic acids research* **39**(10):4063-4075.
14. Ernst J, Kellis M: **Discovery and characterization of chromatin states for systematic annotation of the human genome.** *Nature biotechnology* **28**(8):817-825.
15. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al: **Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome.** *Nature genetics* 2007, **39**(3):311-318.
16. Wang G, Wang Y, Shen C, Huang YW, Huang K, Huang TH, Nephew KP, Li L, Liu Y: **RNA polymerase II binding patterns reveal genomic regions involved in microRNA gene regulation.** *PLoS one* **5**(11):e13798.
17. Pekowska A, Benoukraf T, Ferrier P, Spicuglia S: **A unique H3K4me2 profile marks tissue-specific gene regulation.** *Genome research* **20**(11):1493-1502.
18. Rugg-Gunn PJ, Cox BJ, Ralston A, Rossant J: **Distinct histone modifications in stem cell lines and tissue lineages from the early mouse embryo.** *Proceedings of the National Academy of Sciences of the United States of America* **107**(24):10783-10790.
19. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, et al: **Genome-scale DNA methylation maps of pluripotent and differentiated cells.** *Nature* 2008, **454**(7205):766-770.

doi:10.1186/1471-2164-13-S8-S5

Cite this article as: Zhang et al.: Redistribution of H3K4me2 on neural tissue specific genes during mouse brain development. *BMC Genomics* 2012 **13**(Suppl 8):S5.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

