Published in final edited form as:

Methods Mol Biol. 2014 ; 1150: 175–187. doi:10.1007/978-1-4939-0512-6_11.

Correlating Histone Modification Patterns with Gene Expression Data During Hematopoiesis

Gangqing Hu and Keji Zhao

Abstract

Hematopoietic stem cells (HSC) in mammals are an ideal system to study differentiation. While transcription factors (TFs) control the differentiation of HSCs to distinctive terminal blood cells, accumulating evidence suggests that chromatin structure and modifications constitute another critical layer of gene regulation. Recent genome-wide studies based on next-generation sequencing reveal that histone modifications are linked to gene expression and contribute to hematopoiesis. Here, we briefly review the bioinformatics aspects for ChIP-Seq and RNA-Seq data analysis with applications to the epigenetic studies of hematopoiesis and provide a practical guide to several basic data analysis methods.

Keywords

Hematopoiesis; Epigenetics; Histone modification; RNA-Seq; ChIP-Seq

1 Introduction

Hematopoietic stem cells give rise to all blood cell types while maintaining a capacity of self-renewal [1]. It is known that a core set of transcription factors form a tightly regulated network that controls the spatiotemporal regulation of lineage-specific genes during hematopoiesis [2]. However, the precise molecular mechanisms governing HSC self-renewal and differentiation remain unclear.

There is an increasing awareness of epigenetic mechanisms in controlling the developmental hierarchy of hematopoietic system [3]. Genomic DNA within the eukaryotic nucleus is packaged with histones into a compact form called chromatin. The N-terminal tails of histones are subjected to a variety of posttranslational modifications including acetylation and methylation. Our previous works revealed that histone modifications correlate with gene activities, contribute to T-cell specificity/plasticity, and set stages for hematopoiesis [4–8].

Our knowledge about epigenetic regulation has been greatly advanced by recent development of genome-wide techniques such as ChIP-Seq and RNA-Seq. While ChIP-Seq charts genomic landscapes of transcription factor binding and histone modifications, RNA-Seq quantifies gene expressions. A combinational use of ChIP-Seq and RNA-Seq has been widely applied to the epigenetic studies of hematopoiesis [3]. Here, we have briefly

reviewed the bioinformatical steps commonly used to address epigenetic questions in hematopoiesis by using ChIP-Seq and RNA-Seq.

2 Materials

2.1 Genome Annotation

Genome annotations were downloaded from the online UCSC genome browser [9]:

1. Go to <http://genome.ucsc.edu/cgi-bin/hgGateway>, choose genome and assembly version.
2. Click “tools” at the top of the browser, and then choose “Table Browser.”
3. Specify the source of genome annotation. The “Table Browser” by default provides download of annotation for the UCSC known genes. One may choose other sources of annotation such as RefSeq and Ensembl from the “track” down-drop list.
4. Type a file name in the “output file” text field, click the “get output” button, and save the annotation to a local drive.

2.2 Public ChIP-Seq/RNA-Seq Data

Raw fastq sequence files and/or processed BED6 files included in this review were downloaded from Gene Expression Omnibus (GEO: <http://www.ncbi.nlm.nih.gov/geo/>) [10].

2.3 Software

1. In-house C++ programs (executable files and source code are available upon request):
 - a. Sam2Bed6_Bowtie2: Convert a sam file from Bowtie2 to a BED6 file.
 - b. RemoveRedundantRead: Remove redundant reads from a BED6 file.
 - c. GenerateRPBMBasedSummary: Convert a BED6 file to a BEDGraph file.
 - d. AverageDensityAcrossGenes: Generate average read density across gene features.
 - e. RPKMCalculator: Calculate RPKM at gene level.
 - f. Cat_expr_file: Concatenate gene expression files from different samples.
 - g. SortGeneAnnoByExpr: Sort genome annotations by gene expression level.
 - h. DensityCalculatorPromoters: Calculate normalized read density for promoters.
2. Microsoft Excel: a spreadsheet editing software.
3. Microsoft Access: a database management system.
4. MeV (Multiple Experiment Viewer): a JAVA application for statistic analysis, clustering, and visualization of gene expression data [11].

3 Methods

We first reviewed several common steps for ChIP-Seq and RNA-Seq data analysis, including (1) inspection of data quality, (2) sequence alignment, (3) data visualization, (4) identification of read-enriched regions, and (5) quantification of gene expression. We then introduced a combinational usage of in-house C++ programs, Excel (Microsoft), Access (Microsoft), and MeV (Dana-Farber Cancer Institute) to correlate histone modification with gene expression. We showed examples by using public ChIP-Seq and RNA-Seq data sets generated for epigenetic studies on hematopoiesis.

3.1 Initial Data Quality Inspection

The first step in processing next-generation data is to check the sequence quality. FastQC (Babraham Institute) is a standalone Java application that outputs a summary statistics for fastq files (*see* Note 1). It issues warnings for bases with low quality, for bases with abnormal sequence contents, and for primer/adaptor contaminations.

3.2 Sequence Alignment

There are dozens of algorithms to map short reads to a reference genome. Low-quality bases may be clipped off to increase mapping rate (the % of reads mapped to the reference; *see* Note 2). To minimize ambiguity, reads that are mapped to multiple positions (called multireads) are frequently discarded. Consequently, read enrichments within repetitive regions are underestimated; repetitive sequences are found in constitutive heterochromatin and segmental duplications, both with functions implicated in hematopoiesis [12]. The exclusion of multireads will also underestimate the expression of genes with multiples copies. To address this challenge, several sophisticated probabilistic methods with user-friendly tools have been proposed, for instance, as described in [13].

3.3 Data Visualization

A visualization of the ChIP-Seq and RNA-Seq data in a genome browser such as the UCSC genome browser [9] helps to further inspect the data quality. A local mirror of the UCSC genome browser may be installed. But its maintenance usually requires substantial computational resources. Thus, for a small number of samples, the “custom track” feature from the online UCSC genome browser is recommended (*see* Note 3). An example is illustrated below about how to upload a ChIP-Seq data set (for transcription factor GATA1) and an RNA-Seq data set (for human CD36+ erythrocyte precursor cells 14) to the online UCSC genome browser:

¹FastQC is implemented by script language JAVA and therefore has low run-time performance. In practice, one would extract the first 0.1 million reads to supply to FastQC: “head -n 400000 original_fastq_file>0.1_million_fastq_file”, where the “>” symbol directs the output to the specified file.

²Bowtie2 implements an option “--local” to allow reads to be trimmed at both extremes to optimize the alignment score. If the low-quality bases are known from the initial quality inspection, one could use the option “-5” (“-3”) from Bowtie2 to specify the number of bases to be clipped at the 5’-(3’) end.

³Integrative Genomics Viewer (<http://www.broadinstitute.org/igv/>) is also frequently used by biologists to visualize ChIP-Seq/RNA-Seq data.

1. Download the two data sets (GSM651547 and GSM651555) from GEO. The files are in sra format (short read archive). A sub-module called “fastq-dump” from the SRA Toolkit (NCBI) extracts fastq files from sra files (*see* Note 4).
2. Map the sequences to human genome (hg18) by using Bowtie2, which reports the alignments in sam format [15]. Bowtie2 by default reports the best hit for multireads.
3. Convert sam to BED6 file by using the in-house C++ program “Sam2Bed6_Bowtie2” (*see* Note 5).
4. Remove redundant reads with “RemoveRedundantRead” (*see* Note 6). Since the probability that two reads are mapped to the same genomic position is small for ChIP-Seq data, only one read is retained for each genomic position to minimize biases from amplification (*see* Note 7).
5. Generate genomic distribution of reads with “Generate RPBMBasedSummary.” The program outputs a BEDGraph file, with the first three columns denoting chromosome, starting position and ending position, and the last column denoting the number of reads mapped to the genomic region (*see* Note 8).
6. Customize the BEDGraph file for the UCSC genome browser. The BEDGraph file acceptable by the online UCSC genome browser reserves the first line for parameters of the track (<http://genome.ucsc.edu/goldenPath/help/bedgraph.html>). One needs to edit the BEDGraph file from **step 5** to accommodate this requirement (*see* Note 9).
7. Upload the BEDGraph files to the online UCSC genome browser: (1) Go to <http://genome.ucsc.edu/cgi-bin/hgGate-way>, choose genome (Human) and assembly version (hg18), and click “add custom tracks”; (2) click “Browse” and choose the BEDGraph file and click “Submit” (*see* Note 10); (3) after uploading a file, the user will be redirected to a page called “Manage Custom Tracks”; from there one may choose to upload more files or go to the genome browser.

⁴When running fastq-dump from SRA Toolkit, the whole path of the sra file is recommended to supply as input for beginners to minimize additional configurations. If several sra files are available for one sample, the fastq files can be concatenated with the “cat” command: “cat fastq_file_1 fastq_file_2 ... fastq_file_n>fastq_file_1_2”.

⁵The current version of “Sam2Bed6_Bowtie2” deals with single-end alignment.

⁶The in-house C++ program “RemoveRedundantRead” removes redundant reads for every mapped position and outputs a sorted BED file.

⁷Removing read redundancy is not recommended for RNA-Seq data, since coding regions constitute a small portion of the genome and it is very likely that two reads will hit the same position.

⁸To enable comparison among different samples, the forth column of the BEDGraph file generated by “Generate RPBMBasedSummary” is normalized by library size (in millions) and by window size: reads per base per million reads (RPBM). To run the program, one needs to specify the number of bases to be shifted. Setting the shifting size as half of the length of a nucleosome DNA plus the linker DNA (approximately 200 bps) would work for most ChIP-Seq of histone modifications. It is recommended to set the shifting size to zero for RNA-Seq data set to ensure the shifted positions be within coding regions. One also needs to specify a window size. While a window size of 200 bps works for most ChIP-Seq, the window should be smaller for RNA-Seq data to account for exons less than 200 bps (e.g., 20 bps).

⁹A combinational usage of “echo” and “cat” commands converts a BEDGraph output by “GenerateRPBMBasedSummary” (say file1) to a BEDGraph acceptable by the online UCSC genome browser (say file2): (1) echo track type = bedGraph name = “track name” description = “description of the track”>file2 and (2) cat file1 >>file2. The first command writes the parameters to a newly created file2, and the second command appends the content of file1 to file2.

¹⁰It is highly recommended to compress the BEDGraph files before uploading to the online UCSC genome browser.

8. Save your session (*see* Note 11). Click the “My Data” option on the top of the genome browser and then choose “sessions.” One needs to register, if not have done, to save the session.

Figure 1 shows a screenshot from the online UCSC genome browser for the two data sets. GATA1 not only occupies the promoters of genes *HBB*, *HBD*, and *HBBP1* but also binds to the enhancer sites downstream to gene *HBE1* (top track). Consistent with their important functions in hematopoiesis, both *HBB* and *HBD* are highly expressed (second track). One great advantage of using the online UCSC genome browser is its ability to integrate NGS data sets preexisting in the browser, including those from the ENCODE project. For instance, from the “Encode Regulation ...” section, the enhancer regions in the beta-globin domain in erythrocyte precursor cells are marked by H3K4me1 (an enhancer signature) in the distinctive human mammary epithelial cells (third track). However, the enhancers are not likely active, because the nearby globin genes are all silent (forth track).

3.4 Identification of Read Enriched Regions

A common step in analyzing ChIP-Seq data is to identify the genomic regions enriched with mapped reads. The general idea is to test whether the number of tags with a genomic region is significantly more than those generated from a background model. An initial check of the read distribution from the genome browser helps to tell whether the read-enriched regions are broad or narrow. While different methods have been developed to address each situation, a combinational usage of the methods is not uncommon in literatures [16].

Identification of read-enriched regions is also justified for RNA-Seq data under certain circumstances. It is known that reads from the 3'-end of an RNA molecule are more likely sampled than those from the 5'-end, especially for single cell RNA-Seq [17]. In this situation, normalizing the read count within a gene by the size of read-enriched regions rather than simply by the gene length would improve the quantification of gene expression. The 3'-end- biased sequencing data provide valuable information on the exact ending positions of transcripts, of which the boundaries can be defined by read-enriched islands.

3.5 Gene Expression Quantification

The abundance of mRNA of a gene is quantified by RPKM (the number of reads per kilobases of exon model per million reads) for RNA-Seq, which normalizes the length of RNA species and sequencing depth [18]. The expression level can be measured at both gene and isoform levels and the choice is project specific. The differentially expressed (DE) genes are identified by examining whether or not the difference in read counts between two conditions is significantly higher than expectation. Different probabilistic distributions are proposed to model read count from RNA-Seq data, including Poisson and negative binomial, with representative tools such as edgeR [19].

¹¹In our experiences, the UCSC genome browser only keeps the uploaded tracks for a couple of weeks on the same machine if the session is not saved. Saving the session also allow one to retrieve the tracks from different machines.

3.6 Average Density Profile from ChIP-Seq Data

We previously generated a large number of ChIP-Seq data sets for histone methylations and acetylations in human hematopoietic stem cells, erythrocyte precursors, CD4⁺ T cells, and B cells [5–8]. Analysis of these data sets revealed that different histone modifications show distinctive preferences in genomic localizations. A plot for the average density of reads for a histone modification surrounding and across genic regions helps to reveal its localization preference. Below is an example for how to obtain the average distribution of H3K4me3 across a genic region from human hematopoietic stem cells:

1. Download the BED file for the H3K4me3 ChIP-Seq data from GEO (GSM317587) [6]. Note that the BED file is based on hg18.
2. Download genome annotation from the online UCSC genome browser following instructions in Subheading 2.1 (choose Human and assembly version hg18).
3. Calculate the average density of H3K4me3 across genes by using `anin-houseC++` program “AverageDensityAcrossGenes”: It divides the promoter region (TSS $\pm$ 2 Kbps) into 20 equal size bins, separates gene body region into 10 fractions, and extends to 2 Kbps after TES (10 equal size bins) (*see* Note 12). It outputs the density for each bin/fraction in a `fl` at text file.
4. Visualize the average density with any spreadsheet software such as Excel (Fig. 2a).

3.7 Correlating Histone Modifications with Gene Expression Level

After sorting genes based on their expression levels into equal size groups, two strategies are introduced to visualize the correlation between histone modifications and gene expression: (1) Plot the average density profile of a histone modification (*see* Note 13), and (2) visualize the read density by using heatmap.

Below is a step-by-step guide about how to visualize H3K4me3 densities across groups of genes sorted by gene expression level (both from human hematopoietic stem cells):

1. Download the ChIP-Seq BED file for H3K4me3 (GSM317587) [6] from GEO and genome annotation (hg18) from the online UCSC genome browser.
2. Download the RNA-Seq BED file from GEO (GSM651554) [14].
3. An in-house C++ program “RPKMCalculator” calculates expression at gene level by taking a BED file and a genome annotation file (*see* Note 14).
4. Sort gene annotations with the expression file from **step 3** with “SortGeneAnnoByExpr.”

¹²The last input parameter of the program “AverageDensity AcrossGenes” specifies the number of groups to separate based on the input order of gene annotation file. If the annotation is sorted by for example gene expression level, then by specifying the number of gene groups, the program can be used to correlate the read density with gene expression level.

¹³Visualization based on the average density profile may be sensitive to outliers. For instance, if the read density is extremely high for several genes, then the profile would mostly reflect the features of this gene subset.

¹⁴The “RPKMCalculator” program calculates the number of read mapped to the annotated transcribed regions of an isoform and uses this number to calculate RPKM; it treats different isoforms from one gene independently.

5. Set the desired number of groups as the last input parameter of the in-house C++ program: “AverageDensityAcrossGenes.” It will output a file containing a matrix of average read density with each row corresponding to a bin or fraction and each column corresponding to a gene group. The matrix can be visualized by any spreadsheet software such as Excel (Fig. 2b). If the number of groups is large, the density matrix can be imported into and visualized as a heatmap by MeV (Fig. 2c).

3.8 Visualization of Histone Modifications and Gene Expression During Hematopoiesis

The dynamics of histone modifications and gene expression during early stages of T-cell development was extensively characterized by Dr. Ellen Rothenberg’s laboratory [20]. They generated genome- wide ChIP-Seq data for several histone modifications and RNA- Seq data during the differentiation from “early T-cell precursors” (DN1) to CD4 and CD8 double-positive cells [20]. Using these data sets, we show below an example of a combinational usage of in-house programs, Access/Excel and MeV, to visualize the dynamics of H3K4me2 enrichment at promoters during the early stages of T-cell development. For a concise result, we limited the data analysis to genes that are specifically expressed in DN1 cells. Examples of Access/Excel files are available upon request:

1. Obtain gene expression data. (1) Download RNA-Seq raw sequence data from GEO (GSE31235; “sample1”) [20]. Generate BED6 files as described in **steps 1–3** of Subheading 3.3. (2) Download UCSC genome annotation (mm9) as described in Subheading 2.1. (3) Apply the in-house C++ program “RPKMCalculator” to calculate RPKM values for each BED6 file. (4) Apply the in-house C++ program “Cat_expr_file” to concatenate the expression files from **step 3**. It outputs a file containing a matrix of expression values, with each row corresponding to one gene and each column corresponding to one sample.
2. Define DN1 specific genes. (1) Open the expression matrix file from Excel. (2) Create a new column and define the values for the column by using the formula (Fig. 3a).
3. Create a database using Access (*see* Note 15). Choose “Blank Database” from the templates that appear after running Access to create a blank database, name, and save it to a local drive.
4. Import the expression data into Access. Click the “External Data” panel and choose “Import/Excel File,” which activates the “Import Text Wizard,” to import the excel file generated from **step 2**. During the process, note that (1) the first row of the density file contains field names and (2) attribute “ID” should be specified as primary key of the table.
5. Obtain read density of H3K4me2 at promoters. (1) Download ChIP-Seq data from GEO (GSE31235) and process the sra files to BED6 files. (2) Apply the in-house C++ program “DensityCalculatorPromoters” to calculate the normalized read density at promoters for all samples (*see* Note 16).

¹⁵Access manages data in the forms of tables: Each table contains several columns (or attributes), of which one may be marked as key to distinguish different records. It allows one to intersect different tables through the keys.

6. Import the read density into Access. Similar to **step 4**, click “External data” panel, and chose “Import/Text” to import the text file generated from **step 5** into the database.
7. Intersect the expression and density tables. (1) Click the “create” panel and choose “query design” to activate the “show table” dialog. (2) Add the two tables to the query panel from the dialog. (3) Click the “ID” attribute of one table, hold, drag, and release it to the “ID” attribute of the other (Fig. 3b, dashed arrow 1) to create a join link that implements an intersection operation between the two tables through the specified attributes.
8. Extract read density for DN1 specific genes. (1) Click, hold, drag, and release the attributes associated with read densities to the bottom panel (Fig. 3b, dashed arrow 2). (2) Use “Criteria” in the bottom panel to restrict the query results to DN1-specific gene (Fig. 3b, dashed arrow 3). (3) Execute (“Design/Results/Run”) and save the query (“ctrl+s”). (4) Export the results to a fl at text file (“External Data/Export/Text File”). During this process, choose to include Field name and “Tab” as delimiter.
9. Visualize read density by using MeV (Fig. 3c). (1) Import the fl at text file from **step 6** into MeV (“File/Load Data”). (2) Normalize the read density to highlight changes of H3K4me3 enrichment across samples (“Adjust Data/Gene/Row Adjustments/Normalize Genes/Rows”). (3) Cluster genes based on their read density across samples (“Analysis/Clustering/HCL”; *see* Note 17). A “HCL: Hierarchical clustering” dialog will show up before the clustering for users to set parameters. (4) Choose color theme (“Display/Color Scheme”), adjust color scale (“Display/Set Color Scale Limits”), and set size (“Display/Set Element Size”) of the heatmap.

Acknowledgments

The authors are supported by the Intramural Research Program of the NIH, NHLBI.

References

1. Orkin SH, Zon LI. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell*. 2008; 132:631–644. [PubMed: 18295580]
2. Loose M, Swiers G, Patient R. Transcriptional networks regulating hematopoietic cell fate decisions. *Curr Opin Hematol*. 2007; 14:307–314. [PubMed: 17534154]
3. Cedar H, Bergman Y. Epigenetics of haematopoietic cell development. *Nat Rev Immunol*. 2011; 11:478–488. [PubMed: 21660052]
4. Wei G, Wei L, Zhu J, et al. Global mapping of H3K4me3 and H3K27me3 reveals specificity and plasticity in lineage fate determination of differentiating CD4+ T cells. *Immunity*. 2009; 30:155–167. [PubMed: 19144320]

¹⁶The “DensityCalculatorPromoters” program allows many input BED files, distinguished by different labels from the input. It outputs a matrix of read density for promoters (TSS $\pm$ 2 Kbps), with each row corresponding to one promoter and each column corresponding to one sample.

¹⁷When setting parameters for hierarchical clustering in MeV, one needs to uncheck the “Sample Tree” option if the sample orders are known in prior. The “Normalize Genes/Rows” procedure in MeV subtracts the mean (row) and then divides the standard deviation (row) for each value to be normalized.

5. Barski A, Cuddapah S, Cui K, et al. High-resolution profiling of histone methylations in the human genome. *Cell*. 2007; 129:823–837. [PubMed: 17512414]
6. Cui K, Zang C, Roh TY, et al. Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell*. 2009; 4:80–93. [PubMed: 19128795]
7. Abraham BJ, Cui K, Tang Q, et al. Dynamic regulation of epigenomic landscapes during hematopoiesis. *BMC Genomics*. 2013; 14:193. [PubMed: 23510235]
8. Wang Z, Zang C, Rosenfeld JA, et al. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet*. 2008; 40:897–903. [PubMed: 18552846]
9. Meyer LR, Zweig AS, Hinrichs AS, et al. The UCSC genome browser database: extensions and updates 2013. *Nucleic Acids Res*. 2013; 41:D64–D69. [PubMed: 23155063]
10. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res*. 2013; 41:D991–D995. [PubMed: 23193258]
11. Saeed AI, Sharov V, White J, et al. TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*. 2003; 34:374–378. [PubMed: 12613259]
12. Bailey JA, Gu Z, Clark RA, et al. Recent segmental duplications in the human genome. *Science*. 2002; 297:1003–1007. [PubMed: 12169732]
13. Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat Methods*. 2013; 10:71–73. [PubMed: 23160280]
14. Hu G, Schones DE, Cui K, et al. Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by BRG1. *Genome Res*. 2011; 21:1650–1658. [PubMed: 21795385]
15. Langmead B, Salzberg SL. Fast gapped- read alignment with Bowtie 2. *Nat Methods*. 2012; 9:357–359. [PubMed: 22388286]
16. Kidder BL, Hu G, Zhao K. ChIP-Seq: technical considerations for obtaining high- quality data. *Nat Immunol*. 2011; 12:918–922. [PubMed: 21934668]
17. Ramskold D, Luo S, Wang YC, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol*. 2012; 30:777–782. [PubMed: 22820318]
18. Mortazavi A, Williams BA, McCue K, et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*. 2008; 5:621–628. [PubMed: 18516045]
19. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010; 26:139–140. [PubMed: 19910308]
20. Zhang JA, Mortazavi A, Williams BA, et al. Dynamic transformations of genome- wide epigenetic marking and transcriptional control establish T cell identity. *Cell*. 2012; 149:467–482. [PubMed: 22500808]

Fig. 1.

A screenshot of the uploaded ChIP-Seq and RNA-Seq from the online UCSC genome browser. First track: distribution ChIP-Seq reads of GATA1 at the beta-globin domain for human CD36+ erythrocyte precursor cells. Second track: distribution of RNA-Seq reads from the same cells as in the first track. The Y-axis for the two tracks represents the number of tags normalized by total library size and window size. Third track: genomic distribution of H3K4me1 for human mammary epithelial cell line. Enrichment of H3K4me1 in intergenic regions is a marker of enhancer. Forth track: distribution of RNA-Seq reads for the same cells as in the third track. No read was detected in this region. The last two tracks are preexisting in the online UCSC genome browser. Enhancer regions are marked by *rectangle*

Fig. 2.

H3K4me3 read density at promoters positively correlates with gene expression level. **(a)** Average H3K4me3 read density across promoter region, gene body region, and region 2 Kbps downstream to transcription ending site (TES). The promoter region ($TSS \pm 2$ Kbps) is divided into 20 equal size bins. Gene body region, excluding the first 2 Kbps, is separated into ten fractions. Region downstream of TES is divided into 10 equal size bins. **(b)** Similar to panel **a**, except that the average density is plotted independently for four groups of genes, sorted based gene expression levels. **(c)** The average density is visualized as a heatmap. Genes were sorted into 200 equal size groups by gene expression level. Each row corresponds to a group of genes. Each column corresponds to a bin/fraction of genomic region as defined in **a**

Fig. 3.

A combinational use of Excel/Access and Mev to visualize changes of histone modifications. (a) A new column “DN1-specific” is created and the value is defined by the formula. As shown in the formula, a gene is defined as DN1-specific if the expression is higher than 3 (“B2 > 3,” where “B2” means column “B” row “2”) and is at least twofold higher than any other stages (“B2/(max(C2:F2) + 0.001) > 2,” where 0.001 is a “pseudo count”). (b) Screenshot of the query window. The *dashed arrows* are explained in main text. (c) Hierarchical clustering and heatmap visualization of H3K4me2 density at promoters of DN1-specific genes during the differentiation from DN1 to DP cells. H3K4me2 is highly enriched at the promoters of most genes at DN1 stage and decreases during differentiation. However, about 1/3 of genes also show high enrichments of H3K4me2 at promoters at later development stages