# ChEA2: Gene-Set Libraries from ChIP-X Experiments to Decode the Transcription Regulome

Yan Kou, Edward Y. Chen, Neil R. Clark, Qiaonan Duan, Christopher M. Tan, Avi Ma'ayan

# ChEA2: Gene-Set Libraries from ChIP-X Experiments to Decode the Transcription Regulome

Yan Kou[1], Edward Y. Chen[1], Neil R. Clark[1], Qiaonan Duan[1], Christopher M. Tan[1], Avi Ma'ayan[1,*]

[1]Department of Pharmacology and Systems Therapeutics, Systems Biology Center New York (SBCNY), Icahn School of Medicine at Mount Sinai, New York, NY 10029
* To whom correspondence should be addressed: avi.maayan[at]mssm.edu

**Abstract.** ChIP-seq experiments provide a plethora of data regarding transcription regulation in mammalian cells. Integrating spurious ChIP-seq studies into a computable resource is potentially useful for further knowledge extraction from such data. We continually collect and expand a database where we convert results from ChIP-seq experiments into gene-set libraries. The manual portion of this database currently contains 200 transcription factors from 221 publications for a total of 458,471 transcription-factor/target interactions. In addition, we automatically compiled data from the ENCODE project which includes 920 experiments applied to 44 cell-lines profiling 160 transcription factors for a total of ~1.4 million transcription-factor/target-gene interactions. Moreover, we processed data from the NIH Epigenomics Roadmap project for 27 different types of histone marks in 64 different human cell-lines. All together the data was processed into three simple gene-set libraries where the set label is either a mammalian transcription factor or a histone modification mark in a particular cell line, organism and experiment. Such gene-set libraries are useful for elucidating the experimentally determined transcriptional networks regulating lists of genes of interest using gene-set enrichment analyses. Furthermore, from these three gene-set libraries, we constructed regulatory networks of transcription factors and histone modifications to identify groups of regulators that work together. For example, we found that the Polycomb Repressive Complex 2 (PRC2) is involved with three distinct clusters each interacting with different sets of transcription factors. Notably, the combined dataset is made into web-based application software where users can perform enrichment analyses or download the data in various formats. The open source ChEA2 web-based software and datasets are available freely online at http://amp.pharm.mssm.edu/ChEA2.

**Keywords:** ChIP-seq, ChIP-chip, Microarrays, Systems Biology, ENCODE, Enrichment Analysis, Transcriptional Networks, Data Integration, Data Visualization, JavaScript D3.

# 1 Introduction

Gene expression in mammalian cells is regulated by transcriptional complexes that include transcription factors, histone modifiers, other chromatin and DNA modifiers, and co-regulators proteins that bring these factors together in a transient manner. These cellular components are increasingly profiled using the ChIP-seq/chip (ChIP-X) technologies to identify the locations on the genome where such cellular components bind. Integrating datasets from many sparse ChIP-X studies applied to mammalian cells into a single computable resource is challenging but can lead to a new level of global understanding of gene expression regulation in mammalian cells. In 2010, a database called ChIP-X Enrichment Analysis (ChEA) was published [1]. This database originally contained integrated data from 87 publications reporting the binding of 92 mammalian transcription factors to their putative targets. Since then, we continually expanded this database where it currently contains 200 transcription factors from 221 publications for a total of 458,471 transcription-factor/target-gene interactions. Such dataset is potentially useful for two types of applications: 1) exploring the global transcriptional regulatory landscape of mammalian cells; 2) identifying the most likely regulators given lists of differentially expressed genes using gene-list enrichment analyses.

While the initial version of the ChEA database has been proven useful for several applications, since its publication in 2010 other similar databases have been published. For example hmChIP provides systematic access to ChIP-X experiments deposited in to GEO [2]. ChIP-Array combines gene expression changes together with ChIP-X experiments to provide querying capabilities [3]. Similarly to ChIP-Array, TranscriptomeBrowser integrates various types of binary regulatory interactions including those extracted from ChIP-X experiments [4]. Two other related tools that aggregate ChIP-X experiments are called CistromeMap [5] and CistromeFinder [6]. CistromeMap and CistromeFinder combine many ChIP-seq and DNase-seq publications for integrative analysis. Complimenting these efforts are databases and tools that identify transcription-factor/target-genes interactions by other computational and experimental methods. For example a database called HTRIdb aggregates mammalian transcription-factor/target-gene interactions from low-throughput studies [7] and there are many tools that detect such interactions computationally using position weight matrices. However, not much meta-analysis has been performed on the networks that such datasets and tools produce to obtain a global view of the transcriptional regulatory landscape in mammalian cells. In addition, the data from these tools is not easily utilized for gene-list enrichment analyses. Here we performed an initial integrative analysis of the ChIP-X data we collected, as well as provide a state-of-the-art HTML5 freely available web application that can be used to query the datasets with lists of differentially expressed genes by performing gene-list enrichment analyses.

## 2 Methods

### 2.1 The Histone Modification Dataset

ChIP-seq for histone modification datasets were collected from the Roadmap Epigenomics project (http://www.roadmapepigenomics.org) deposited to the GEO database (http://www.ncbi.nlm.nih.gov/geo/). Previous studies [8] have indicated that the use of control samples substantially reduces DNA shearing biases and sequencing artifacts, therefore for each experiment, an input control sample was matched according to the description in GEO. ChIP-seq experiments without matched control dataset were not included. BED files were first downloaded and standardized to contain the first six columns as defined by the UCSC genome browser BED file format by adding placeholders if needed. Unlike narrow peaks of transcription factor binding sites, histone modification marks have broad peaks and can be extended to as long as 100kb [8]; therefore the peak-calling software Sicer [9] was used to identify significant peaks. Sicer uses a spatial clustering approach that considers enrichment information from neighboring regions to call broader domains. Each read was shifted by 75bp towards the center of the DNA fragment; window size of 200bp and gap size equals to the window size, or three times the window size, were used according to the authors' guide. However, for large datasets, window size of 500bp and gap size of 1500bp were applied because smaller parameters resulted in memory limit errors. For broad histone marks such as H3K27me3 and H3K36me3, the window size of 500bp and gap size of 1500bp were set to detect significant peaks. Peaks were identified with only one read mapped to the reference genome, not allowing for redundant reads. False discovery rate (FDR) of 0.001 was used to find significant peaks as compared to the control.

The significant peaks for each experiment were collected and associated with annotated human genes. Chromosome coordinates of the human hg19 build were downloaded from NCBI; the longest readout was kept for genes having multiple isoforms. A gene is considered associated with a significant peak if the peak overlaps with the region from -10kb of transcription starting site (TSS) to +5kb of transcription ending site (TES) of the gene. Peaks that overlap with $\pm$2kb of TSS of known genes were defined as enriched at promoter regions. Genes that have significant peaks at promoter regions were ranked according to their distance from the middle point of the associated peak to the transcription starting site. For experiments that identified more than 2000 genes around TSS and TES, the top 2000 genes were taken to generate the gene matrix transpose (GMT) gene-set library file, whereas all genes were taken for experiments that recovered less genes. The resulted dataset contains 27 types of histone modifications for 64 human cell lines and tissue types.

### 2.2 The ENCODE Transcription Factor Dataset

ChIP-seq datasets for transcription factor binding sites were downloaded from the ENCODE project [10] (ftp://encodeftp.cse.ucsc.edu/pipeline/hg19/) on July 9, 2012. A total of 920 experiments applied to 44 cell lines profiling 160 transcription factors were processed. For each experiment, the peak file in broadPeak or narrowPeak format was extracted. For each experiment dataset, the peaks in the broadPeak and

narrowPeak files were mapped to the TSS of human genes from the hg19 build if the peak overlaps with the region of (-5kb, +2kb) surrounding the TSS. Most of the experiments have biological replicates. Therefore, only the unique genes (intersection) with peaks deposited within the defined promoter region from both replicates were extracted for further analysis. The GMT file was generated in the same way as for the histone modification datasets, where the top 2000 genes were taken for experiments that yielded larger number of target genes.

### 2.3 Updating the Manual Part of the ChEA Database

The manual part of the ChEA database is continually updated by manually processing the supporting tables from ChIP-seq publications. Target gene lists are extracted from such publications. Only human or mouse experiments are considered. All gene identifiers are converted to Entrez Gene Symbols. E-mail alerts for abstracts that contain the key term of "ChIP-seq" or "ChIP-chip" point to publications that may contain data that is suitable for this database. In some cases when the authors only provide the peaks' coordinates we process the data to identify and rank the associated genes.

### 2.4 The ChEA2 Web Application

ChEA2 uses HTML5, CSS3, and JavaScript on the frontend to render the page, make the SVG visualizations interactive, and retrieve JSON data from the backend server asynchronously using AJAX. The frontend makes heavy use of the jQuery JavaScript library and other dependent plugins to simplify page manipulation/selection, enable touch gestures, add sortable tables, and create tooltips in combination with using the D3 JavaScript library to generate SVG visualizations. ChEA2 also utilizes JSP to generate some of the more dynamic pages that interact with the backend. The backend server side uses an Apache Tomcat 7 application server to host Java servlet web apps and a Microsoft IIS 6 web server to serve the web pages. Apache Maven is used to simplify the process of compiling the code, processing the web resources, and deploying the web app to the Tomcat server. The application is best viewed with Chrome or Firefox, but it is compatible with all modern browsers that support SVG and HTML5. The code-based is mostly borrowed from the recently published tool Enrichr [11].

### 2.5 Canvas Visualization

The canvas images in Figure 1 represent an alternative way to visualize networks. Instead of connecting nodes with edges, the network nodes are placed onto a square toroidal grid and then clustered together using simulated annealing to maximize local connections. The brightness of a node on the finished grid corresponds to the strength of its connections with its neighbors. The annealing process is performed by swapping two randomly selected nodes on the grid and recalculating the global fitness. After annealing, the canvas is exported as a JavaScript Object Notation (JSON) file which is then sent to the Visualizer module. The Visualizer creates SVG images with HTML5 and the JavaScript Library D3. These images were created with the recently published tool, Network2Canvas [12].

## 3 Results

The manual portion of the ChEA2 database currently contains 200 transcription factors from 221 publications for a total of 458,471 transcription-factor/target interactions. In addition, the data processed from ENCODE includes 920 experiments applied in 44 cell-lines profiling 160 transcription factors for a total of ~1.4 million transcription-factor/target-gene interactions. The data from the NIH Epigenomics Roadmap project contain 27 different types of histone marks in 64 different human cell-lines for a total of 438,064 interactions. The first step in our analysis is to visualize each of the gene-set libraries: the manual ChEA, ENCODE, and histone modifications on a canvas. The canvas visualization is a compact alternative to visualize networks as ball-and-stick diagrams [12]. In these canvases, each ChIP-X experiment is represented as a square on a grid of squares. The location of the experiments/squares is optimized so a gene-set with a similar gene-set are adjacent to each other. The similarity measure for the three grids is based on set similarity computed by the Sets2Networks algorithm [13]. Once such arrangement is determined, the squares are color-coded by the level of similarity each experiment/square has with its neighbors. The brighter the square the more similar it is to its neighboring squares. The canvases clearly identify clusters within the gene-set libraries/networks (Fig. 1). These clusters include a stem-cell/pluripotency set of factors and the PRC2 group in the manual ChEA canvas; the H3K36me3, H3K27me3, mixture of H3K27me3/H3K9me3 clusters in the histone modifications canvas; and CTCF/RAD21, TAF1/TAF7/YY1/HEY1 clusters in the ENCODE canvas; but there other clusters. While some of these clusters are likely functional, where the factors physically interact to regulate the same sets of genes, it should be considered that some factors are more frequently profiled so there is some research bias in the formation of the various clusters on these canvases.

Next, we visualized the manual ChEA gene-set library and the ENCODE gene-set library using multi-dimensional scaling (MDS) with a distance measure similarity computed using the Sets2Networks algorithm [13] (Fig. 2). This visualization identifies the same clusters but has the advantage of showing the labels for most factors as well as provides a visual of the distance similarity among all factors. Then, we visualized the manual ChEA gene-set library as a heatmap using hierarchical clustering (Fig. 3). The heatmap identifies three clusters that include the members of the PRC2 complex: EZH2, EED, SUZ12, and JARID2. Besides these core components of PRC2, the transcription factor MTF2, and the histone methyltransferase SETDB1 as well as RNF2 and PHC1 which are both members of the PRC1 complex, cluster with the canonical PRC2 components. Interestingly, the transcription factor p53 also clusters with these factors. However, these three clusters also include various other factors that are unique for each one of the three clusters. Using known protein-protein interactions from the literature [14], we see that many of these additional factors were previously identified to interact with members of the PRC2 complex, suggesting that these factors directly physically associate with the PRC2 complex members to regulate the same set of genes in different contexts (Fig. 3, right). This potentially identifies multiple roles for the PRC2 complex in different cell types. As we see below, when combining the manual ChEA dataset with the

ENCODE and histone modification datasets the PRC2 complex members strongly associates with the H3K27me3 modification which is already well established.
Next, we visualized the manual ChEA and ENCODE networks as ball-and-stick diagrams. We only connected the transcription factors that were used as immuno-precipitation baits in ChIP-X experiments (Fig. 4). Therefore all nodes are transcription factors and the edges connect transcription factor profiled by ChIP-X to their direct binding targets which are other transcription factors. The manual ChEA network is made of 197 factors connected through 5344 edges whereas the ENCODE network is made of 150 factors connected through 2430 links with clustering coefficients of 0.15 and 0.09 respectively. These clustering coefficients are much higher than the clustering observed for random networks of the same size. Since the density of links in these networks is very high, not much of the network structure is revealed by the ball-and-stick visualization. However, the hubs of the networks clearly stand out. These hubs are likely master regulators or well-studied factors. Interestingly BHLHE40 (also called DEC1, HLHB2, SHARP-2, STRA13, and STRA14) a less known transcription factor appears as a hub in the ENCODE network. This is because this factor is a target of many other factors as well as a regulator of many of the other profiled factors. It is possible that this transcription factor has an important global role: integrating information from many transcription factors and acting as a master regulator of many factors. This central role may have been yet mostly overlooked.

Next, we combined all three gene-set libraries into one heatmap that cluster the entries based on similarity of gene-list content (Fig. 5). Euclidean distance of the Jaccard score was used as a similarity measure and the linkage type was average. Multiple experiments forming clusters are highlighted and listed on the right. In general the histone modification experiments cluster together by modification type and the transcription factors form their own clusters. Focusing on one cluster that contains a mixture of transcription factors and histone modification experiments, we mapped all the known protein-protein interactions that connect the factors, as well as the protein interactions between these factors and the known enzymes that remove or add the chemical modification group on the histone marks from the cluster (Fig. 6). Links connect the enzymes associated with these modifications were created by a literature search and the other links represent known experimentally determined protein-protein interactions collected from available online databases [14]. We find that some factors interact more exclusively with the enzymes that remove the marks, while other factors interact more exclusively with the enzymes that add the mark. Moreover, two writers: EP300 and CREBBP and one remover HDAC1 are highly connected. This could be a literature bias since these factors are well-studied, but also can represent the type of complexes that are formed to regulate the same subset of genes. Finally, we performed enrichment analysis, using all three gene set libraries: ChEA-manual, ENCODE and histone modifications, on the entire gene set library created from the Mouse Genomics Informatics Mammalian Phenotype ontology (MGI-MP) [15]. This mapping identifies relationships between annotated single gene knockout mouse phenotypes and groups of transcriptional regulators. Genes associated with mouse phenotypes (columns) were used for enrichment analysis and the corrected p-values of each term from the three gene-set libraries: manual ChEA,

ENCODE and histone modifications (rows) were used for drawing the hierarchical clustering heatmap. Clusters of mouse phenotypes are highlighted with green labels and clusters of ChEA2 elements are labeled with orange colors (Fig. 7). For example, a group of transcription factors including GATA2, EGR1, BCL11A, FLI1, and NFKB are highly enriched for abnormal immune system as expected. The mouse phenotype clusters of abnormal neuronal system (Cluster 3 in Fig. 7) and abnormal embryonic/early development (Cluster 5) share similar groups of enriched transcription factors and histone modification profiles. This may be due to the observation that neuronal specific genes are suppressed in embryonic stem cells by specific histone modifications. The map includes many potentially interesting small clusters that identify relationships between factors and phenotypes. For example, we highlight two small clusters that connect abnormal pigmentation related phenotypes and abnormal lysosomes to the transcription factors: USF1/2, ATF3, BHLHE40, MAX, NF2, and TFEB. Indeed, TFEB has a well-established role in autophagy [16] and ATF3 is known to be associated with stress responses [17]. Corre and Galibert summarized the roles of USF1/2 in various contexts including the role of these factors in pigmentation determination and stress responses [18]. While there is some related knowledge about the role of some of these factors in determining these phenotypes, our analysis suggests that these factors may work together to regulate the same sets of genes that are responsible for the induction of specific functions that lead to specific phenotypes. In addition to the preliminary meta-analysis, we performed on the three gene-set libraries ChEA2 datasets, we also developed a web-based application that provides access to the data and enables users to query their lists of differentially expressed genes (Fig. 8). The tool is implemented with HTML5 and the JavaScript library D3 and it is freely available online at http://amp.pharm.mssm.edu/ChEA2.

## 4 Conclusions

Here, we assembled a large compendium of ChIP-seq/chip experiments applied to mammalian cells. We began an initial meta-analysis of such complex dataset by abstracting each experiment to a gene list. Such abstraction discards much important information such as peak height and exact binding locations. Peak height and exact binding locations information are likely critical in better assessing cooperation among regulators. In addition, including such quantitative data in enrichment analyses is likely to improve the inference of regulators given sets of differentially expressed genes. Moreover, we must consider the fact the putative target genes determined by ChIP-seq/chip contain many false positives. Integrating results from transcription factor knockdowns followed by genome-wide expression can potentially filter putative binding targets with real functional targets. Another consideration is that our network analysis is very preliminary. We did not consider measuring network properties such as connectivity distributions and identification of network motifs. Since the datasets contain a large-scale directed graphs made of transcription factors regulating other transcription factors, insights can be obtained from structural and dynamical analyses of such graphs. Regardless of these limitations, we have integrated an important large-scale dataset made from most available ChIP-X studies into a computable format and began the global analysis of this potentially useful resource. This project paves the way for more extensive computational studies that

would further unravel the transcriptional networks that are responsible for regulating the complex system of the mammalian cell. In addition, the data collection and organization, as well as the interactive light-weight web-based data visualization of ChEA2 exemplify an application from systems biology that deals with a complex interdisciplinary challenge to extract knowledge from massive datasets [19].

**Figure Legends**

**Fig. 1.** Canvas representation of transcription factors and histone modifications networks using Network2Canvas (N2C) [12].

**Fig. 2.** Multi-dimensional scaling scatter plot of the a) manual ChEA; and b) ENCODE datasets, using the Sets2Networks distance measure [13].

**Fig. 3.** Ball-and-stick representation of the transcription factor regulatory networks from ChEA (a) and ENCODE (b) based on direct regulatory interactions between transcription factors and other transcription factors. Node size and color represents connectivity degree.

**Fig. 4.** Hierarchical clustering heatmap of the ChEA adjacency matrix based on direct targets. Direct protein-protein interactions connecting transcription factors that form clusters on the heatmap are highlighted for several clusters.

**Fig. 5.** Hierarchical clustering of all three gene set libraries: ChEA, ENCODE and histone modifications. Experiments were clustered based on their target-set similarity.

**Fig. 6.** Protein-protein interactions (PPI) between histone modifying enzymes and transcription factors in one of the clusters from Figure 5. The histone modifications (orange square nodes) and transcription factors (green eclipse nodes) are members of a cluster in Fig. 5.

**Fig. 7.** Hierarchical clustering of gene sets causing abnormal phenotypes in mice and enriched transcription factors and/or histone modifications. Generalized descriptions of mouse phenotype clusters are listed at the bottom, with numbers matching to the corresponding clusters.

**Fig. 8.** Screenshots from the ChEA2 web-application. a) Users can enter lists of genes or upload a file containing lists of genes. b) Once the users submit valid gene lists, they are presented with enrichment analysis results that rank transcription factors and histone modifications as bar-graphs, tables, networks of enriched factors and canvases that highlight the enriched factors or enriched histone modification experiments.

## 5 Acknowledgements

This work is supported in part by NIH grants R01GM098316, R01DK088541, U54HG006097-02S1, P50GM071558, and Irma T. Hirschl Career Scientist Award.

# 6 References


1. Lachmann, A., Xu, H., Krishnan, J., Berger, S.I., Mazloom, A.R., and Ma'ayan, A. (2010). ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. Bioinformatics 26, 2438-2444.
2. Chen, L., Wu, G., and Ji, H. (2011). hmChIP: a database and web server for exploring publicly available human and mouse ChIP-seq and ChIP-chip data. Bioinformatics 27, 1447-1448.
3. Qin, J., Li, M.J., Wang, P., Zhang, M.Q., and Wang, J. (2011). ChIP-Array: combinatory analysis of ChIP-seq/chip and microarray gene expression data to discover direct/indirect targets of a transcription factor. Nucleic Acids Research 39, W430-W436.
4. Lepoivre, C., Bergon, A., Lopez, F., Perumal, N., Nguyen, C., Imbert, J., and Puthier, D. (2012). TranscriptomeBrowser 3.0: introducing a new compendium of molecular interactions and a new visualization tool for the study of gene regulatory networks. BMC Bioinformatics 13, 19.
5. Qin, B., Zhou, M., Ge, Y., Taing, L., Liu, T., Wang, Q., Wang, S., Chen, J., Shen, L., and Duan, X. (2012). CistromeMap: a knowledgebase and web server for ChIP-Seq and DNase-Seq studies in mouse and human. Bioinformatics 28, 1411-1412.
6. Sun, H., Qin, B., Liu, T., Wang, Q., Liu, J., Wang, J., Lin, X., Yang, Y., Taing, L., Rao, P.K., et al. (2013). CistromeFinder for ChIP-seq and DNase-seq data reuse. Bioinformatics 29, 1352-1354.
7. Bovolenta, L., Acencio, M., and Lemke, N. (2012). HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. BMC Genomics 13, 405.
8. Pepke, S., Wold, B., and Mortazavi, A. (2009). Computation for ChIP-seq and RNA-seq studies. Nat Meth 6, S22-S32.
9. Zang, C., Schones, D.E., Zeng, C., Cui, K., Zhao, K., and Peng, W. (2009). A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics 25, 1952-1958.
10. The ENCODE Consortium Project. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74.
11. Chen, E.Y., Tan, C., Kou, Y., Duan, Q., Wang, Z., Meirelles, G., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics 14, 128.
12. Tan, C., Chen, E.Y., Dannenfelser R., Clark, N.R., and Ma'ayan, A. (2013). Network2Canvas: Network Visualization on a Canvas with Enrichment Analysis. Bioinformatics. Published online: June 7, 2013
13. Clark, N., Dannenfelser, R., Tan, C., Komosinski, M., and Ma'ayan, A. (2012). Sets2Networks: network inference from repeated observations of sets. BMC Systems Biology 6, 89.
14. Berger, S., Posner, J., and Ma'ayan, A. (2007). Genes2Networks: connecting lists of gene symbols using mammalian protein interactions databases. BMC Bioinformatics 8, 372.
15. Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE; the Mouse Genome Database Group. 2012. The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. Nucleic Acids Res 2012; 40(1):D881-86.
16. Decressac M, Mattsson B, Weikop P, Lundblad M, Jakobsson J, Björklund A. TFEB-mediated autophagy rescues midbrain dopamine neurons from α-synuclein toxicity. Proc Natl Acad Sci U S A. (2013) 110:E1817-26.

17. Hai T, Wolfgang CD, Marsee DK, Allen AE, Sivaprasad U. ATF3 and stress responses. Gene Expr. 1999;7(4-6):321-35.
18. Corre S, Galibert MD. Upstream stimulating factors: highly versatile stress-responsive transcription factors. Pigment Cell Res. 2005 Oct;18(5):337-48.
19. Holzinger, A. (2012) On Knowledge Discovery and interactive intelligent visualization of biomedical data - Challenges in Human–Computer Interaction & Biomedical Informatics. In: M. Helfert, C. Francalanci, J. Filipe (Eds.) Proceedings of the International Conference on Data Technologies and Application, Rome DATA 2012, Setubal (PT), SciTec Press, 3-16

Fig. 1

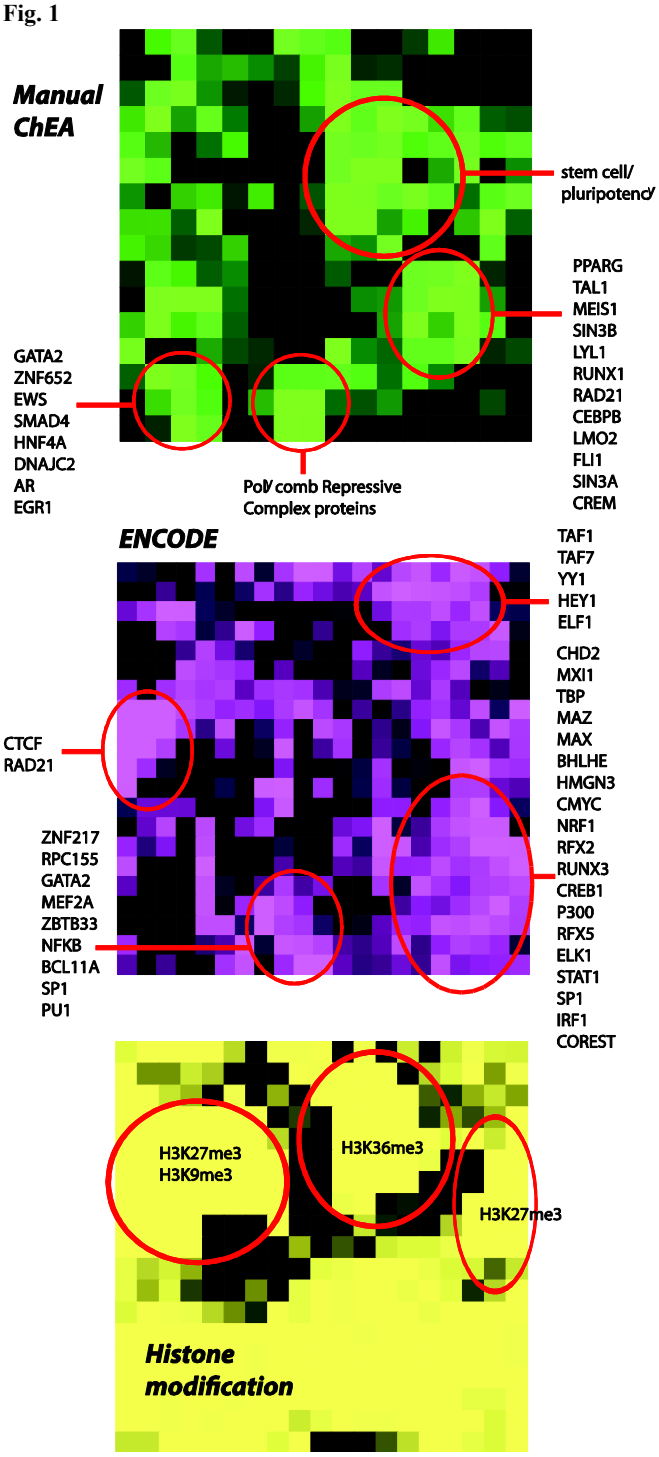
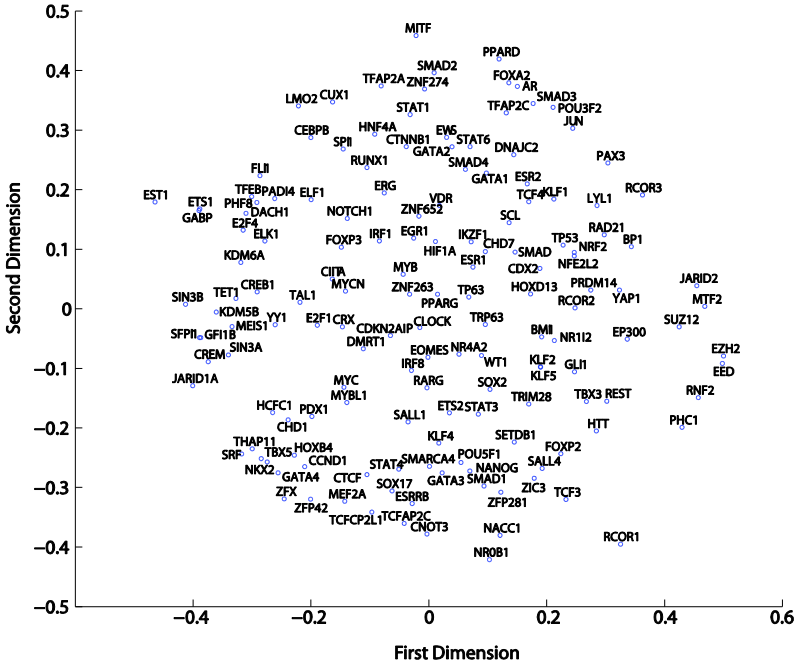

Manual ChEA
stem cell/ pluripotency
PPARG
TAL1
MEIS1
SIN3B
LYL1
RUNX1
RAD21
CEBPB
LMO2
FLI1
SIN3A
CREM
GATA2
ZNF652
EWS
SMAD4
HNF4A
DNAJC2
AR
EGR1
Pol comb Repressive Complex proteins
ENCODE
TAF1
TAF7
YY1
HEY1
ELF1
CHD2
MXI1
TBP
MAZ
MAX
BHLHE
HMGN3
CMYC
NRF1
RFX2
RUNX3
CREB1
P300
RFX5
ELK1
STAT1
SP1
IRF1
COREST
CTCF
RAD21
ZNF217
RPC155
GATA2
MEF2A
ZBTB33
NFKB
BCL11A
SP1
PU1
H3K27me3
H3K9me3
H3K36me3
H3K27me3
Histone modification

**Fig. 2**

**ChEA**




**ENCODE**

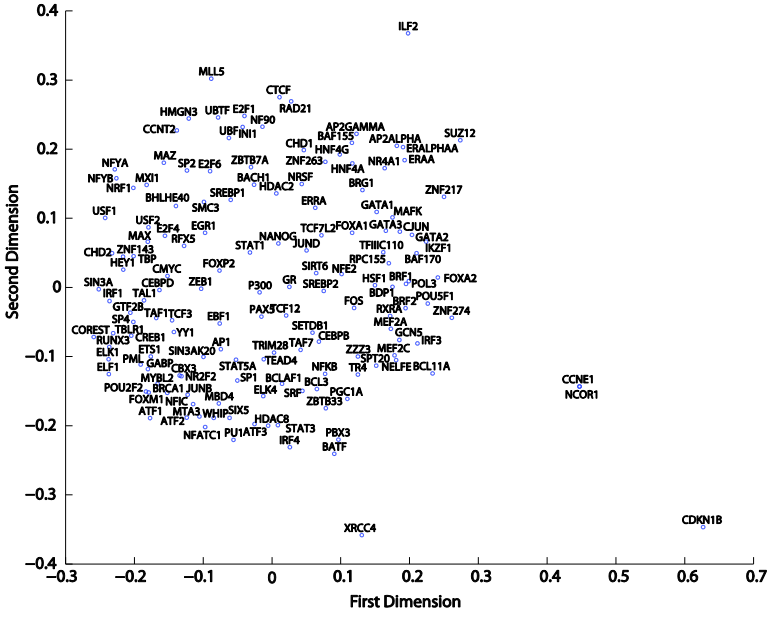


**Fig. 3**

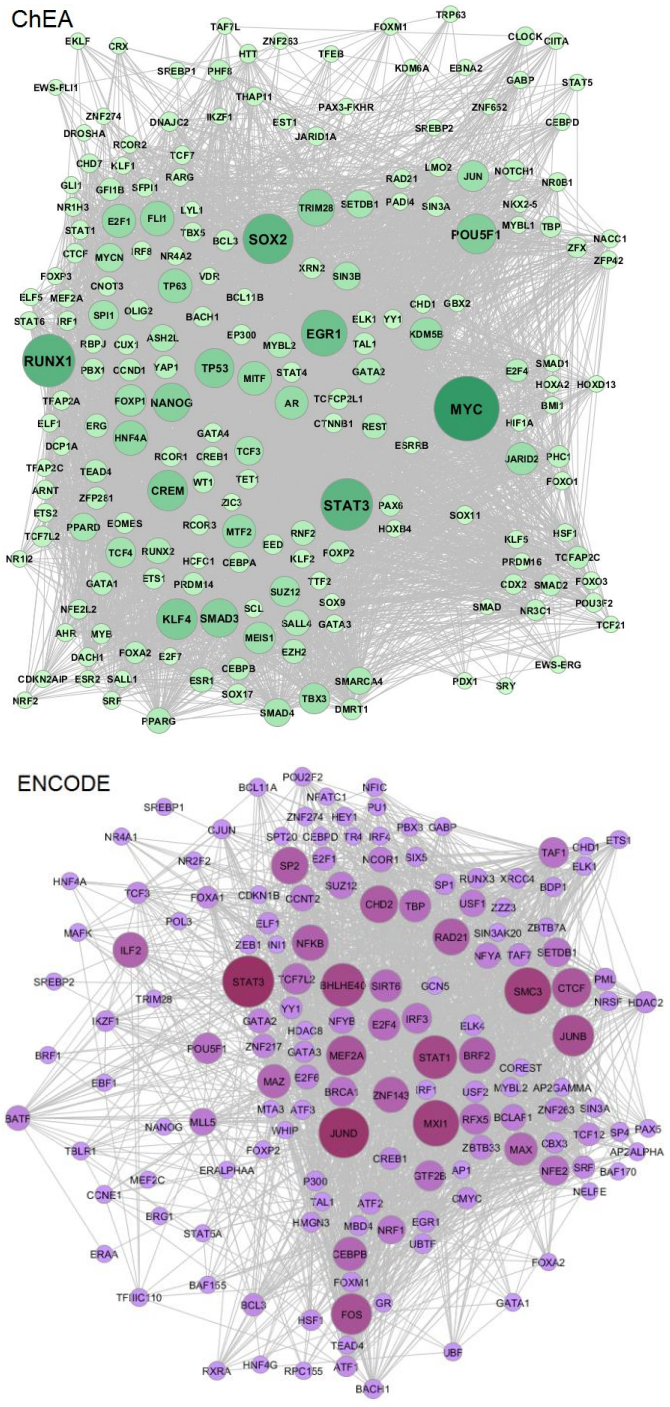

ChEA
SOX2
POU5F1
EGR1
RUNX1
TP53
NANOG
MYC
CREM
STAT3
KLF4
SMAD3
ENCODE
STAT3
SP2
CHD2
TBP
SMC3
CTCF
JUNB
STAT1
BRF2
JUND
MXI1
MAX
FOS
STAT1

**Fig. 4**

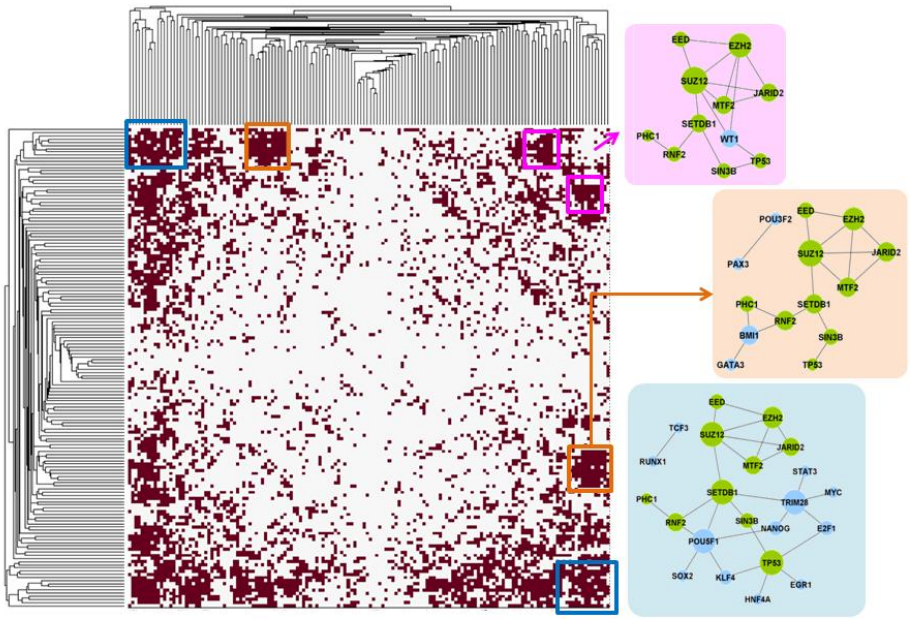

EED
EZH2
SUZ12
JARID2
MTF2
SETDB1
PHC1
WT1
RNF2
TP53
SIN3B
POU3F2
PAX3
BMI1
GATA3
TCF3
RUNX1
STAT3
MYC
TRIM28
NANOG
E2F1
POU5F1
SOX2
KLF4
EGR1
HNF4A

**Fig. 5**

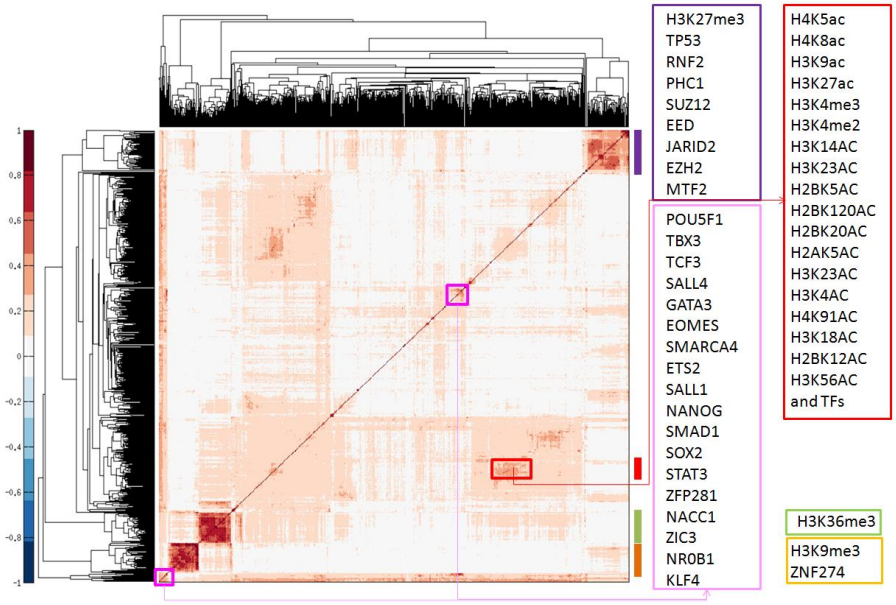

H3K27me3
TP53
RNF2
PHC1
SUZ12
EED
JARID2
EZH2
MTF2
H4K5ac
H4K8ac
H3K9ac
H3K27ac
H3K4me3
H3K4me2
H3K14AC
H3K23AC
H2BK5AC
H2BK120AC
H2BK20AC
H2AK5AC
H3K23AC
H3K4AC
H4K91AC
H3K18AC
H2BK12AC
H3K56AC
and TFs
POU5F1
TBX3
TCF3
SALL4
GATA3
EOMES
SMARCA4
ETS2
SALL1
NANOG
SMAD1
SOX2
STAT3
ZFP281
NACC1
ZIC3
NR0B1
KLF4
H3K36me3
H3K9me3
ZNF274

**Fig. 6**

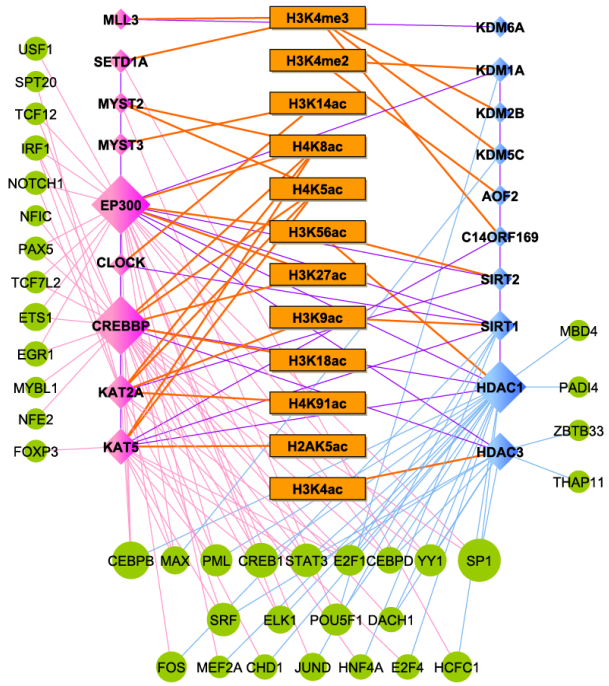

MLL3
SETD1A
MYST2
MYST3
EP300
CLOCK
CREBBP
KAT2A
KAT5
USF1
SPT20
TCF12
IRF1
NOTCH1
NFIC
PAX5
TCF7L2
ETS1
EGR1
MYBL1
NFE2
FOXP3
H3K4me3
H3K4me2
H3K14ac
H4K8ac
H4K5ac
H3K56ac
H3K27ac
H3K9ac
H3K18ac
H4K91ac
H2AK5ac
H3K4ac
KDM6A
KDM1A
KDM2B
KDM5C
AOF2
C14ORF169
SIRT2
SIRT1
HDAC1
HDAC3
MBD4
PADI4
ZBTB33
THAP11
CEBPB
MAX
PML
CREB1
STAT3
E2F1
CEBPD
YY1
SP1
SRF
ELK1
POU5F1
DACH1
FOS
MEF2A
CHD1
JUND
HNF4A
E2F4
HCFC1

Fig. 7

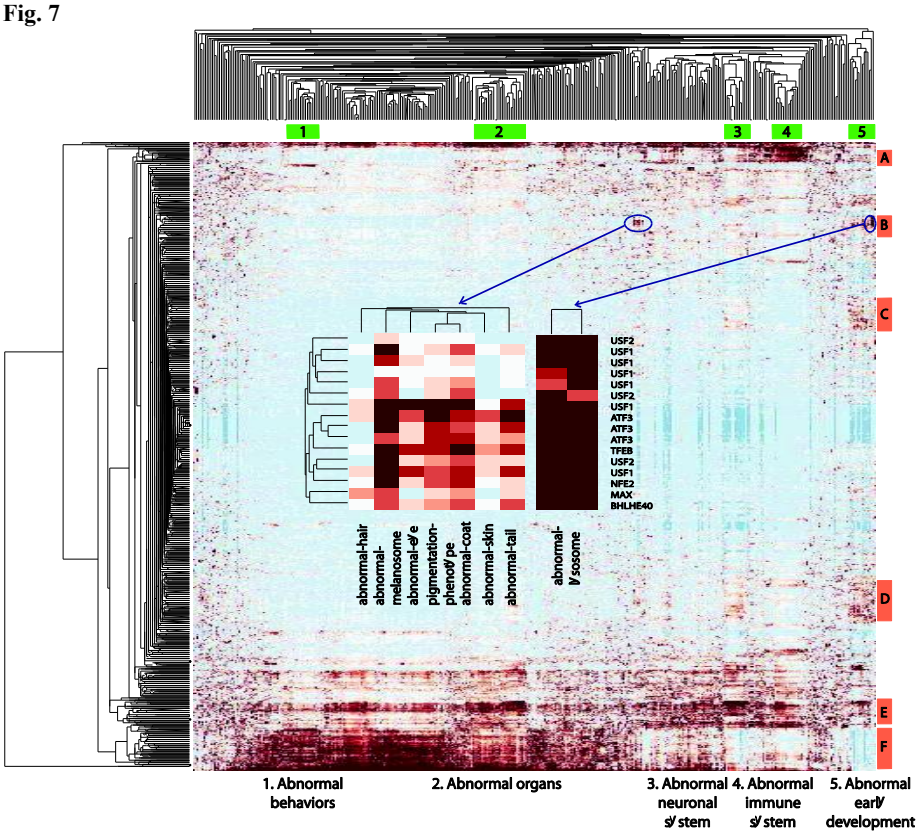
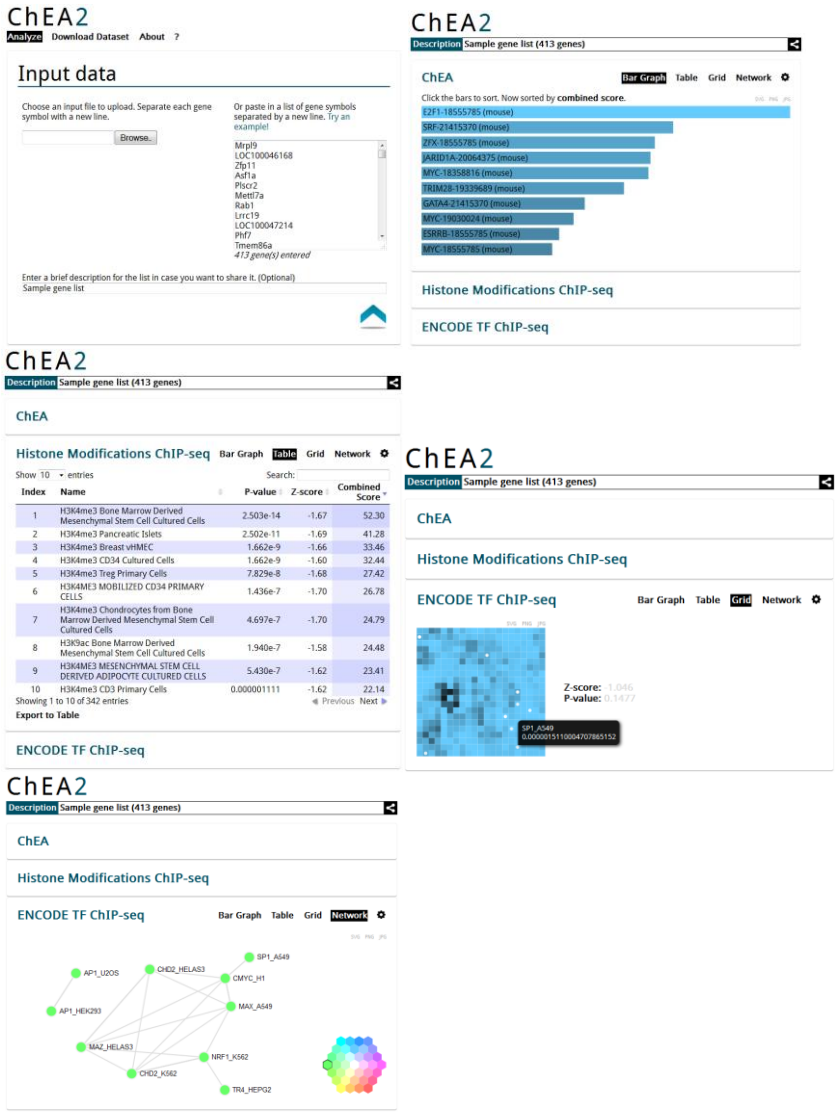

1
2
3
4
5
A
B
C
D
E
F
USF2
USF1
USF1
USF1
USF1
USF2
USF1
ATF3
ATF3
ATF3
TFEB
USF2
USF1
NFE2
MAX
BHLHE40
abnormal-hair
abnormal-melanosome
abnormal-eye pigmentation-phenotype
abnormal-coat
abnormal-skin
abnormal-tail
abnormal-lysosome
1. Abnormal behaviors
2. Abnormal organs
3. Abnormal neuronal system
4. Abnormal immune system
5. Abnormal ear/ development

**Fig. 8**


ChEA2
Analyze Download Dataset About ?
Input data
Choose an input file to upload. Separate each gene symbol with a new line.
Browse..
Or paste in a list of gene symbols separated by a new line. Try an example!
Mrpl9
LOC100046168
Zfp11
Asf1a
Plscr2
Mettl7a
Rab1
Lrrc19
LOC100047214
Phf7
Tmem86a
413 gene(s) entered
Enter a brief description for the list in case you want to share it. (Optional)
Sample gene list
ChEA2
Description Sample gene list (413 genes)
ChEA
Bar Graph Table Grid Network
Click the bars to sort. Now sorted by combined score.
E2F1-18555785 (mouse)
SRF-21415370 (mouse)
ZFX-18555785 (mouse)
JARID1A-20064375 (mouse)
MYC-18358816 (mouse)
TRIM28-19339689 (mouse)
GATA4-21415370 (mouse)
MYC-19030024 (mouse)
ESRRB-18555785 (mouse)
MYC-18555785 (mouse)
Histone Modifications ChIP-seq
ENCODE TF ChIP-seq
ChEA2
Description Sample gene list (413 genes)
ChEA
Histone Modifications ChIP-seq
Bar Graph Table Grid Network
Show 10 entries
Search:
Index Name P-value Z-score Combined Score
1 H3K4me3 Bone Marrow Derived Mesenchymal Stem Cell Cultured Cells 2.503e-14 -1.67 52.30
2 H3K4me3 Pancreatic Islets 2.502e-11 -1.69 41.28
3 H3K4me3 Breast vHMEC 1.662e-9 -1.66 33.46
4 H3K4me3 CD34 Cultured Cells 1.662e-9 -1.60 32.44
5 H3K4me3 Treg Primary Cells 7.829e-8 -1.68 27.42
6 H3K4ME3 MOBILIZED CD34 PRIMARY CELLS 1.436e-7 -1.70 26.78
7 H3K4me3 Chondrocytes from Bone Marrow Derived Mesenchymal Stem Cell Cultured Cells 4.697e-7 -1.70 24.79
8 H3K9ac Bone Marrow Derived Mesenchymal Stem Cell Cultured Cells 1.940e-7 -1.58 24.48
9 H3K4ME3 MESENCHYMAL STEM CELL DERIVED ADIPOCYTE CULTURED CELLS 5.430e-7 -1.62 23.41
10 H3K4me3 CD3 Primary Cells 0.000001111 -1.62 22.14
Showing 1 to 10 of 342 entries
Previous Next
Export to Table
ENCODE TF ChIP-seq
ChEA2
Description Sample gene list (413 genes)
ChEA
Histone Modifications ChIP-seq
ENCODE TF ChIP-seq
Bar Graph Table Grid Network
SP1_A549
CHD2_HELAS3
AP1_U2OS
CMYC_H1
AP1_HEK293
MAX_A549
MAZ_HELAS3
NRF1_K562
CHD2_K562
TR4_HEPG2
ChEA2
Description Sample gene list (413 genes)
ChEA
Histone Modifications ChIP-seq
ENCODE TF ChIP-seq
Bar Graph Table Grid Network
Z-score: 1.046
P-value: 0.1477
SP1_A549