ChIP-seq

Contents

1 Introduction

Transcription regulation: Transcription factors (TF) bind to chromatin in sequence-specific manner (TF binding motif) to regulate gene expression. Genes co-expressed in different conditions are likely under the control of the same TB.

ChIP-Seq: A method used to analyze protein interactions with DNA. ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify genome-wide binding site (BS) of TFs. It is often conducted in millions of cells, where protein-DNA interactions are fixed, then chromatin is sheared into small pieces. Then a factor-specific antibody is used to pull down the factor as well as the DNA attached to it.

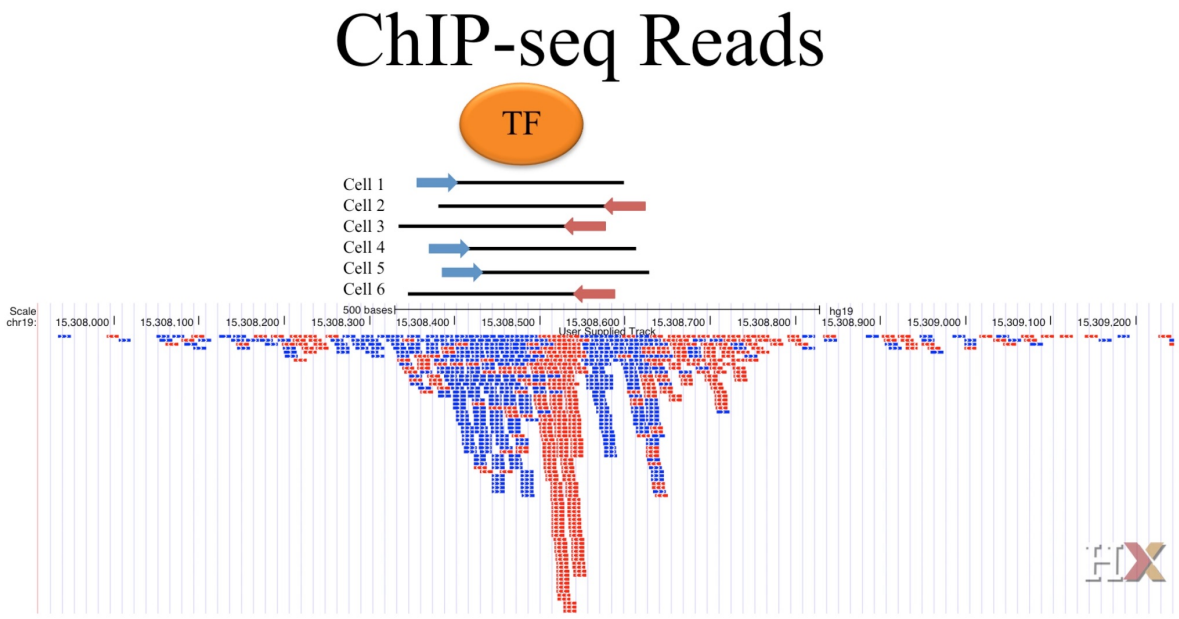

Figure 1: 9-1

2 ChIP-seq technique

The technique consists of the following steps:

1- Crosslink proteins and DNA 2- Break DNA into pieces with sonication 3- Add TF-specific antibody 4- Perform immunoprecipitation 5- Reverse crosslinks 6- Purify and amplify DNA 7- Sequence DNA

A typical control sample for a ChIP-seq experiment is a separate sample from the same experimental condition that has not been enriched by antibody treatment and immunoprecipitation.

3 ChIP-seq quality control

We use *FASTQC* app for QC:

- A score of < 30 usually means very poor control.
- We should also check for mapability. The higher the uniquely mapped reads, the better.
- Also, the higher uniquely mapped locations, the better.
- PBC score is another indicator of data quality. The higher the better.
- FRiP score is another quality score. It is an indicator of signal to noise ratio. The higher the better.
- ChIP-seq peak overlap with DNase hypersensitivity regions is also an indicator of good data quality.

Other measures of high quality include:

- If the ChIP-seq experiment of a TF had a good quality, then it should capture all the BSs of the TF in that cell condition. Because TF BSs are under more evolutionary constraint, they will have overall better conservation than genome background. However, non-conserved sites might still be functional, and very often only a small portion of the total binding peaks have good evolutionary conservation.
- When replicates of the same individuals are performed, under good quality conditions, signal correlation > 0.6 and peak overlap over 60% are considered good replicate agreement.

- For TF ChIP-seq, the correct TF motif should be enriched more in the stronger peaks, and more enriched in the peak center (summit).
- Even though only small % of all the TF BSs are at the gene promoters, there is still an enrichment of overall ChIP-seq peak and signal enrichment near the transcription start sites of genes.

In the following table you can see an example of a QC report of high quality data:

Field	Result
Raw sequence median quality score	30
% Reads uniquely mapped	91.5%
PCR bottleneck coefficient (PBC)	95.3%
Fraction of reads in peaks (FRiP)	49.1%

4 ChIP-seq alignment

For aligning sequence reads to the genome, one can use the following algorithms:

- Bowtie and BWA: specifically useful for single-end seq data.
- STAR: for paired-end read data.

5 ChIP-seq peak calling

ChIP-seq peak: We call a ChIP-seq BS a ChIP-seq peak if the number of reads in the treatment sample are significantly different from that in the control samples.

MACS: TF BSs are enriched in the ChIP protocol, and in sequencing the plus-strand reads will be mapped to the left of the precise binding location, while minus-strand reads will be mapped to the right of binding. ChIP fragments are often 100-300bp long, and sequencing only sequences the ends of the fragments. If plus-strand reads (to the left of binding) and minus-strand reads (to the right of the binding) are separated by 120bp, the precise binding site should be right in the middle of the two. So shifting reads by 60bp in 3' direction will find the precise binding location. Model-based analysis for ChIP-seq (MACS), is an algorithm that models the shift size of ChIP-Seq tags, and uses it to improve the spatial resolution of predicted BSs.

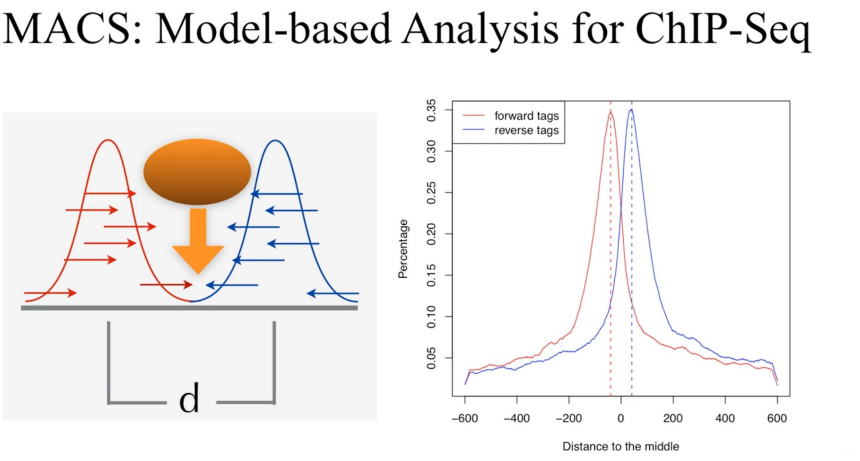

Figure 2: 9-2

6 Target genes

Wherever a TF binds to a promoter of a gene, we assume that gene is being regulated by that TF. There are different regulatory rules:

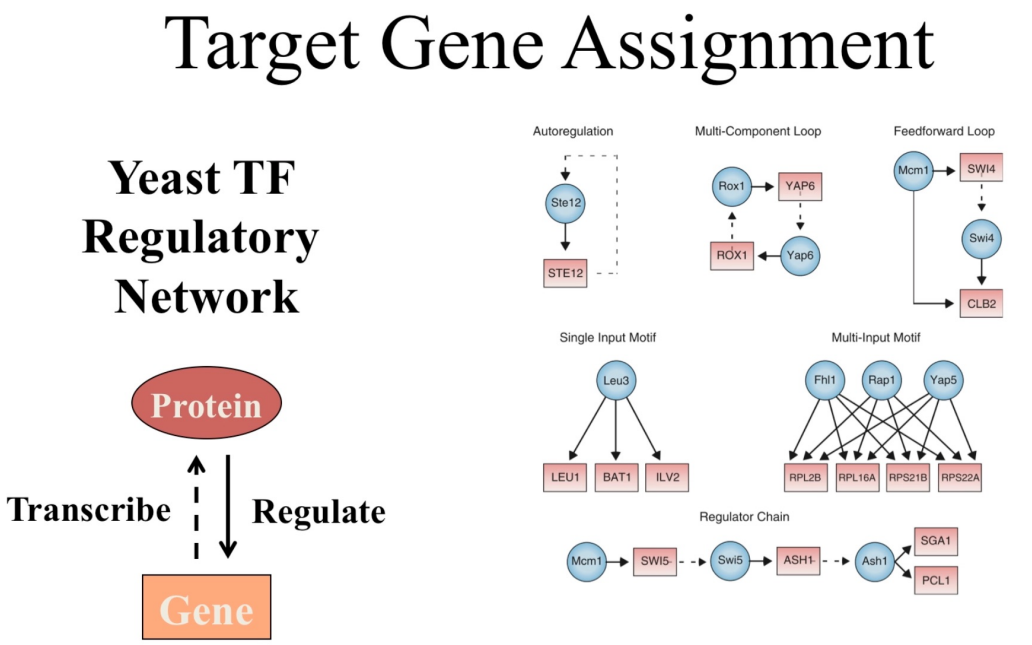

Figure 3: 9-3

However, only 3% of TFs in the human genome bind to the promoter of a gene. This makes the process of finding the target gene challenging. To tackle this problem, we look at the transcription start site (TSS) of a gene and evaluate the BSs within 100kb from it, weighting them by their distance, also checking if these genes show differential expression. Note that TFs could cause activation as well as repression of a gene, or both. Also note, TF binding and target gene is a many-by-many relationship: one BS can regulate the expression of multiple genes (stochastic so might be different in different cells) and one gene is regulated by multiple nearby enhancers.

BETA: Binding expression target analysis (BETA) is an algorithm for finding the target gene of a TF, as well as evaluating the activating or repressing role of a TF on a gene. The TF is a transcriptional activator if genes with better regulatory potential of the TF binding (i.e. more BSs and binding closer to the TSS of the genes) are more activated in expression than random genome background, and vice versa.

7 Cistrome data browser

Cistrome DB in an online platform (available at http://cistrome.org/db) that can be used to check what factors regulate a gene of interest and what factors bind in a interval or have a significant binding overlap with a peak set. It specifically:

- Collects and processes public ChIP-seq and DNase-seq data in human and mouse.
- Can visualize the processed data in genome browsers.
- Can find similar datasets to a particular dataset in the DB.
- Shows quality metrics of the data.

8 Cistrome analysis pipeline

Cistrome functions can be divided into three categories: data preprocessing, gene expression and integrative analysis. A general workflow using Cistrome is to upload datasets, preprocess them using peak calling tools to generate peak locations in BED format and signal profiles in WIGGLE format, upload gene expression data to produce specific gene lists, and then use various integrative analysis tools to generate figures and reports. An overview of the pipeline is presented bellow:

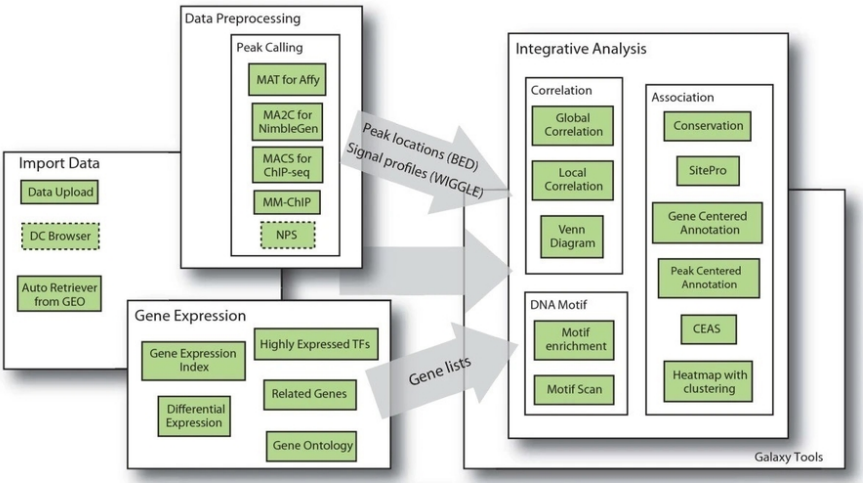

Figure 4: 9-4; The left panel shows available tools, the middle panel shows messages, tool options, or result details, and the right panel shows the datasets organized in the user's history, including datasets that have been or are being processed (in green and yellow, respectively), or waiting in the queue (in gray). CEAS,; DC, Data Collection module; GEO, Gene Expression Omnibus; NPS, Nucleosome Positioning from Sequencing.

Correlation and association tools consists of the following:

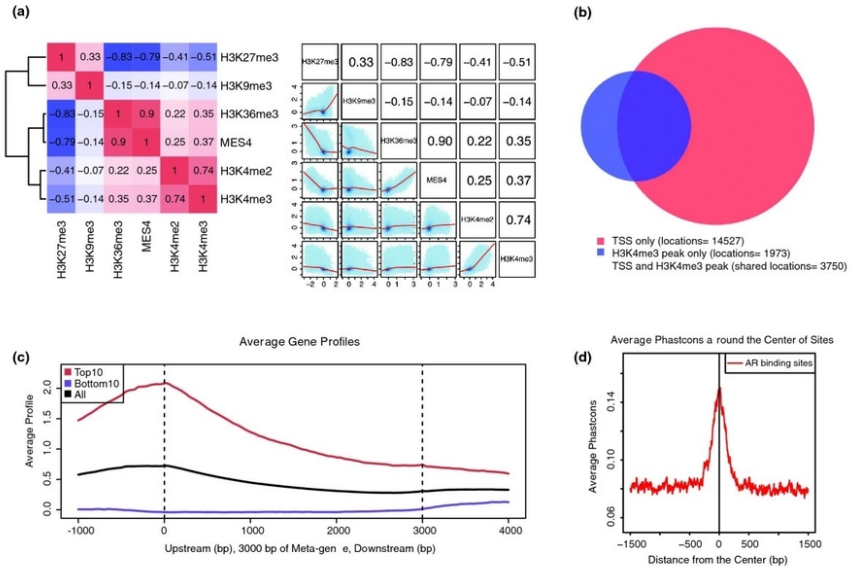

Figure 5: 9-5

(a) Correlation plots using different histone marks in C. elegans early embryos. Cistrome correlation tools

can generate either a heatmap with hierarchical clustering according to pair-wise correlation coefficients or a grid of scatterplots. (b) Venn diagram showing the overlap of H3K4me3 peaks (in blue) with TSS for all the genes (in red) in the C. elegans genome. (c) Meta-gene plot generated by CEAS showing the H3K4me3 signals enriched at gene promoter regions; the top expressed genes (red) have higher H3K4me3 signals than the bottom expressed genes (purple). (d) Conservation plot showing that the human androgen receptor (AR) binding sites from ChIP-chip are more conserved than their flanking regions in placental mammals.

1- Upload BAM files from "Data/Cisterome". Make sure the correct genome (mm9 in this case) is selected.

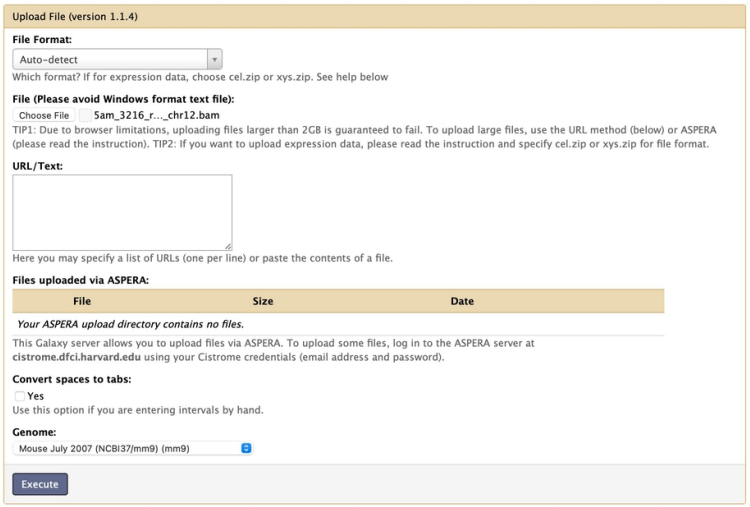

Figure 6: 9-6

2- Call peaks from alignment results:

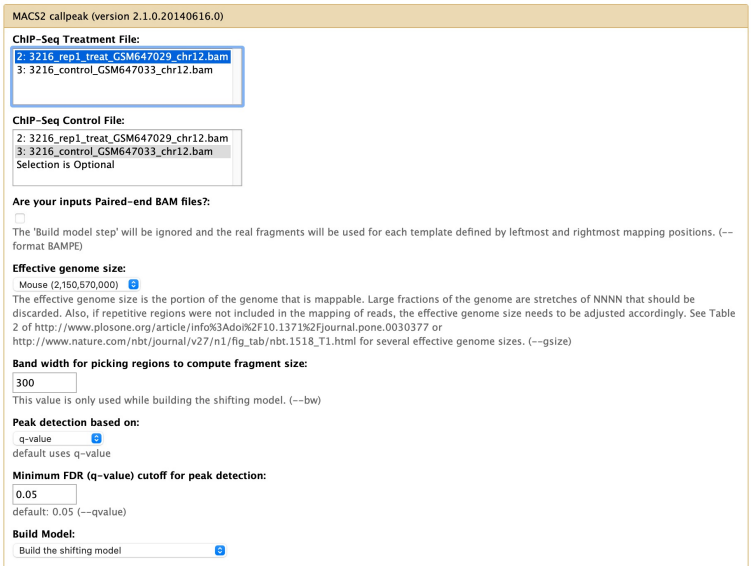

Figure 7: 9-7

3- Check peak calling results:

chr12	3032348	3032442	MACS2_peak_1	17	.	1.98635	4.62897	1.75828	47
chr12	3063356	3063450	MACS2_peak_2	17	.	1.98635	4.62897	1.75828	47
chr12	3117515	3117609	MACS2_peak_3	17	.	1.98635	4.62897	1.75828	47
chr12	3137548	3137642	MACS2_peak_4	17	.	1.98472	4.53023	1.73183	47
chr12	3143187	3143281	MACS2_peak_5	17	.	1.98635	4.62897	1.75828	47
chr12	3150727	3150821	MACS2_peak_6	17	.	1.98472	4.53023	1.73183	47
chr12	3153385	3153479	MACS2_peak_7	17	.	1.98472	4.53023	1.73183	47
chr12	3159047	3159141	MACS2_peak_8	17	.	1.98635	4.62897	1.75828	47
chr12	3163477	3163571	MACS2_peak_9	13	.	1.96967	3.93040	1.38816	47
chr12	3205748	3205842	MACS2_peak_10	13	.	1.96967	3.93040	1.38816	47
chr12	3370894	3370988	MACS2_peak_11	13	.	2.72701	3.80931	1.32295	6
chr12	3585013	3585113	MACS2_peak_12	58	.	6.37901	9.46327	5.85902	17
chr12	3806338	3806449	MACS2_peak_13	96	.	7.95822	13.55825	9.61465	53
chr12	3859517	3859671	MACS2_peak_14	71	.	7.13259	10.90005	7.16678	78
chr12	3860128	3860279	MACS2_peak_15	101	.	8.71762	14.10739	10.11621	68
chr12	3867104	3867212	MACS2_peak_16	50	.	5.30629	8.60031	5.09729	70
chr12	3888118	3888212	MACS2_peak_17	25	.	4.16594	5.64404	2.52106	78
chr12	3912005	3912283	MACS2_peak_18	86	.	7.69573	12.48450	8.60430	185
chr12	3982277	3982392	MACS2_peak_19	40	.	5.35162	7.39922	4.02002	50
chr12	4160472	4160566	MACS2_peak_20	29	.	4.33273	6.19710	2.97570	56
chr12	4349347	4349441	MACS2_peak_21	26	.	3.71401	5.86088	2.68786	46
chr12	4420610	4420704	MACS2_peak_22	22	.	4.03645	5.27596	2.24974	55
chr12	4737202	4737296	MACS2_peak_23	34	.	4.87404	6.76135	3.43876	69
chr12	4805616	4805710	MACS2_peak_24	28	.	3.74075	6.04124	2.83795	33
chr12	5104111	5104205	MACS2_peak_25	40	.	3.88048	7.43662	4.05300	50
chr12	5161818	5161912	MACS2_peak_26	13	.	1.96967	3.93040	1.38816	47
chr12	5211001	5211095	MACS2_peak_27	29	.	4.33273	6.19710	2.97570	42
chr12	5258761	5258855	MACS2_peak_28	32	.	3.79542	6.47248	3.22442	50
chr12	5484150	5484305	MACS2_peak_29	65	.	4.88710	10.26937	6.58812	81
chr12	5485844	5485938	MACS2_peak_30	13	.	1.96967	3.93040	1.38816	47
chr12	5492324	5492418	MACS2_peak_31	13	.	1.96967	3.93040	1.38816	47
chr12	5712078	5712172	MACS2_peak_32	34	.	3.82335	6.73760	3.43876	47

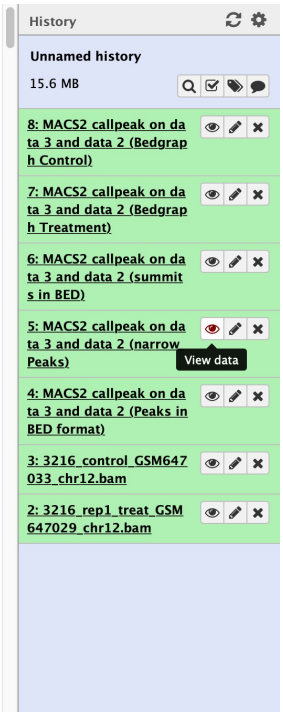

Figure 8: 9-8

9 BETA software suite

BETA integrates ChIP-seq data and gene expression data. It has 3 major functions:

- Predict if a factor has an an activating or repressive function.
- Identify the motifs enriched in a set of ChIP-seq peaks.
- Infer a factor's direct target genes.

Sub-commands include:

- Basic: Predict activating or repressive function of a TF and the direct target genes. Requires differential gene expression data + ChIP-seq peak data.
- Plus: Beta basic + motif analysis on target regions. Requires the same data as basic + genome FASTA data.
- Minus: Predict TF target genes. Requires only ChIP-seq peak data.

Input files include:

- ChIP-seq binding data:
  - Generated by ChIP-seq peak calling algorithm (MACS2).
  - Is in .BED format:
    - 3 cols: chr, start, end
    - 5 cols: chr, start, end, peak_name, score
    - Remove header line if present.
- Expression data:
  - A tab-delimited file with:
    - GeneID (official gene symbol or ref-seq ID)
    - Regulatory status (up- or down-regulated)
    - Statistical value (FDR, p-value, adjusted p-value)
  - Standard options include:

* Limma output (LIM)
* Cuffdiff output (CUF)
* Or custom format.

Output files include: * Function prediction:

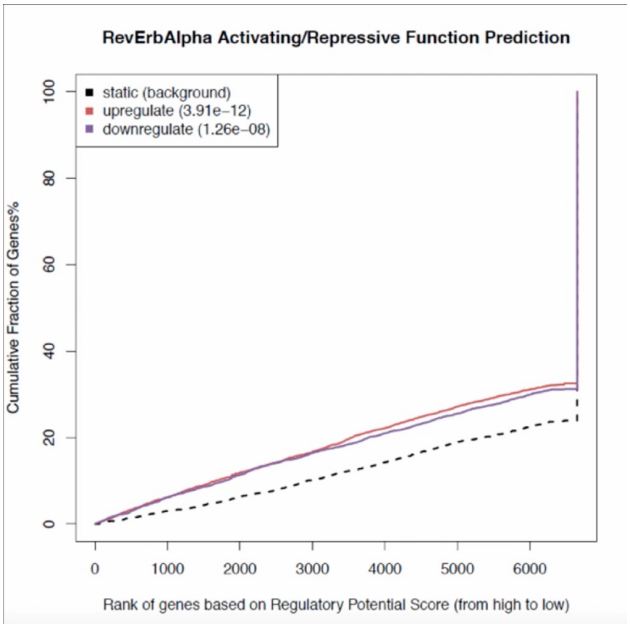

Figure 9: 9-9; In this example, the TF is both activating and repressing.

- Target genes:

Chroms	txStart	txEnd	refseqID	rank	product	Strands	GeneSymbol
chr9	65142066	65178465	NM_001044389	8.057e-07		+	Clpx
chr13	53062577	53076408	NM_017373	1.099e-06		-	Nfil3
chr11	101229043	101239217	NM_008061	2.905e-06		+	G6pc
chr7	120350978	120457640	NM_007489	4.077e-06		+	Arntl
chr1	123200929	123229166	NM_178082	6.055e-06		-	Insig2
chr4	33332397	33335762	NM_001033225	6.185e-06		-	Pnrc1
chr19	36806220	36811094	NM_016854	1.540e-05		-	Ppp1r3c
chr1	74379183	74400266	NM_019999	1.685e-05		+	Pnkd
chr19	30172930	30249931	NM_138595	1.875e-05		-	Gldc
chr16	10959365	10993214	NM_019980	2.030e-05		-	Litaf
chr19	46206388	46210184	NM_007703	2.197e-05		+	Elovl3
chr6	5433350	5446278	NM_013743	2.645e-05		-	Pdk4
chr14	26278670	26486233	NM_183208	2.946e-05		+	Zmiz1
chr18	21103125	21142349	NM_026301	2.988e-05		+	Rnf125
chr9	121823473	121825423	NM_010012	3.086e-05		-	Cyp8b1
chr5	76638892	76733573	NM_007715	3.417e-05		-	Clock
chr16	26356731	26371925	NM_016674	3.605e-05		-	Cldn1
chr14	51563977	51573087	NM_013632	3.887e-05		+	Pnp
chr1	182157134	182176081	NM_011183	4.064e-05		-	Psen2
chr8	121961061	121970156	NM_027950	4.206e-05		+	Osgin1

Figure 10: 9-10

- Associated peaks:

chrom	pStart	pEnd	Refseq	Symbol	Distance	Score
chr9	65065632	65065778	NM_001044389	Clpx	-76361	0.0285973849686
chr9	65081144	65081358	NM_001044389	Clpx	-60815	0.0532583861717
chr9	65137893	65138100	NM_001044389	Clpx	-4070	0.515406174923
chr9	65145057	65145173	NM_001044389	Clpx	3049	0.536891099106
chr9	65151460	65151734	NM_001044389	Clpx	9531	0.414268899624
chr9	65152192	65152389	NM_001044389	Clpx	10224	0.402943066987
chr9	65154526	65154870	NM_001044389	Clpx	12632	0.365942156654
chr9	65155934	65156152	NM_001044389	Clpx	13977	0.346774696341
chr13	53070041	53070430	NM_017373	Nfil3	-6173	0.473823684555
chr13	53071629	53071895	NM_017373	Nfil3	-4646	0.503666971575
chr13	53072220	53072364	NM_017373	Nfil3	-4116	0.514458699506
chr13	53075563	53075702	NM_017373	Nfil3	-776	0.587993138717
chr13	53075873	53076087	NM_017373	Nfil3	-428	0.596235235113
chr13	53076325	53076473	NM_017373	Nfil3	-9	0.606312347974

Figure 11: 9-11; This data tells us what specific peaks contributed to the finding of target genes.

- motifs enriched in ChIP-seq peaks:

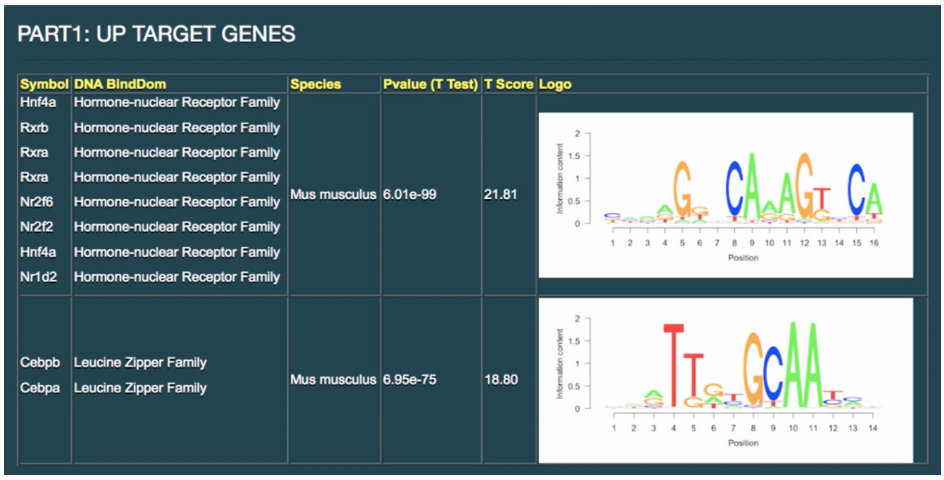

Figure 12: 9-12