ChIP-Seq Analysis Report

Demo Report

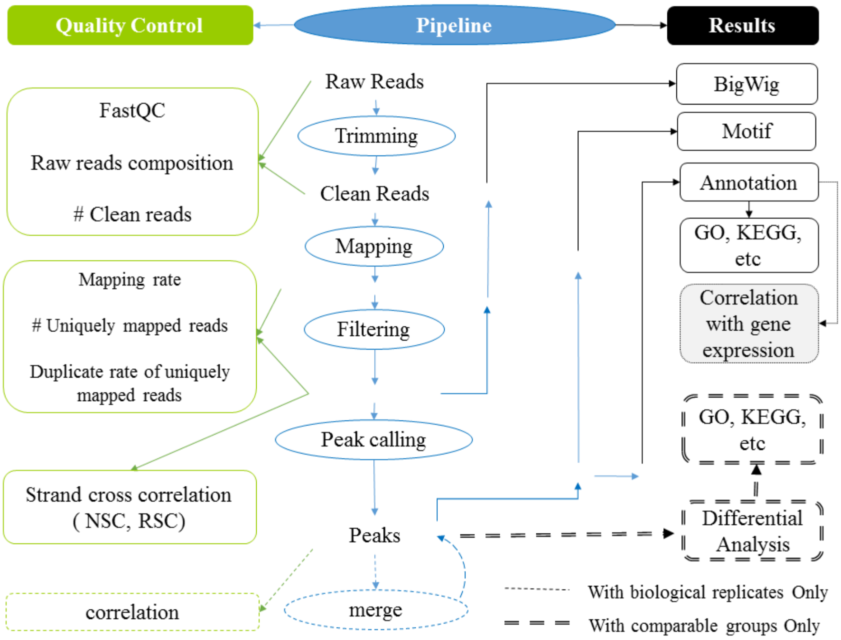
Overseas Department

Sep 21, 2016

Contents

I Pipeline

Chromatin Immunoprecipitation Sequencing (ChIP-seq) is a method which combines Chromatin Immunoprecipitation (ChIP) and the next generation sequencing (NGS) technology. This method can identify genome regions that interact with transcription factors and chromatin-associated proteins in the whole genome region efficiently. By mapping the sequencing results onto the genome accurately, researchers can obtain genome wide information about the DNA region interacting with histone and transcription factor.



Pipeline

II Results

1 Data quality control

1.1 FastQC of Raw_reads

As the high-throughput sequencing technology getting mature, we can obtain a mass amount of data easily. The key is how to extract the information that we need from the data. The first step is to check the raw data quality before doing follow-up data analysis. The popular tool for the quality control is software FastQC. Summary of the raw data quality control is in the following:

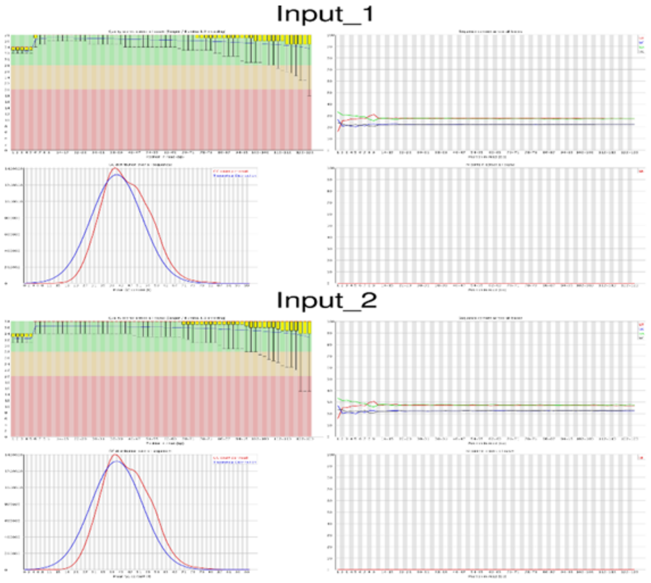


Figure 1-1 raw reads of FastQC

Top-left: horizontal axis represents position in read (bp), vertical axis represents quality scores. Different colors represent different quality range;

Top-middle: horizontal axis represents position in read (bp), vertical axis represents base percentage;

Top-right: the distribution of read length;

Bottom-left:GC distribution over all sequences. Red line is the actual GC content distribution, blue line is the theoretical distribution;

Bottom-middle: N-ratio across all bases;

Bottom-right: x-axis is the sequence duplication level. y-axis is the percent of reads.

1.2 Trimming of raw data

For the raw data which pass the quality control, primer mismatch may result in nucleotidic composition bias at first several positions of the reads, which could lead to wrong bases insertion during sequencing. Because of the small fragment size of ChIP-Seq, there are often some adapter-appended reads. Trimming off the adapter sequences and low-quality bases is necessary. To make sure the quality of data analysis, raw data need to be filtered to get clean data, and all the follow-up analysis are based on these clean data. The procedure for data trimming is in the following:

(1)Discard the reads with low quality (proportion of low quality bases larger than 50%)

(2) Discard the reads with N ratio (unsure base) larger than 15%;

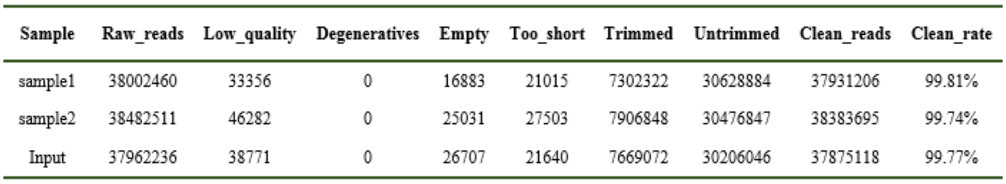
(3) Discard the reads with adaptor at the 5'-end;

(4) Discard the reads without adaptor and inserted fragment at the 3'-end;

(5) Trim the adapter sequence at the 3'-end;

(6) Discard the reads whose length are less than 18nt after trimming.

Table 1-1 Summary of raw data quality control



Sample	Raw_reads	Low_quality	Degeneratives	Empty	Too_short	Trimmed	Untrimmed	Clean_reads	Clean_rate
sample1	38002460	33356	0	16883	21015	7302322	30628884	37931206	99.81%
sample2	38482511	46282	0	25031	27503	7906848	30476847	38383695	99.74%
Input	37962236	38771	0	26707	21640	7669072	30206046	37875118	99.77%

While trimming, the bases that match the adapter or with quality value less than 20 at the 3'-end are removed, and only the reads whose length is longer than 18nt after trimming are kept.

Raw_reads: reads from the base-calling. Click on the number to check the result of FastQC;

Low_quality: reads with mean quality lower than 20 before trimming;

Degeneratives: reads with at least 15% N before trimming;

Empty: reads with all bases from adapter (s);

Too_short: reads shorter than 18nt that are discarded after trimming;

Trimmed: reads with at least 18nt that are kept after trimming;

Untrimmed: reads that are kept untrimmed;

Clean_reads: kept reads after trimming, including both trimmed and untrimmed. Click on the number to check the result of FastQC;

Clean_rate: the ratio of Clean_reads to Raw_reads.

1.3 FastQC of Clean_reads

Quality control of the clean reads after trimming is in the following:

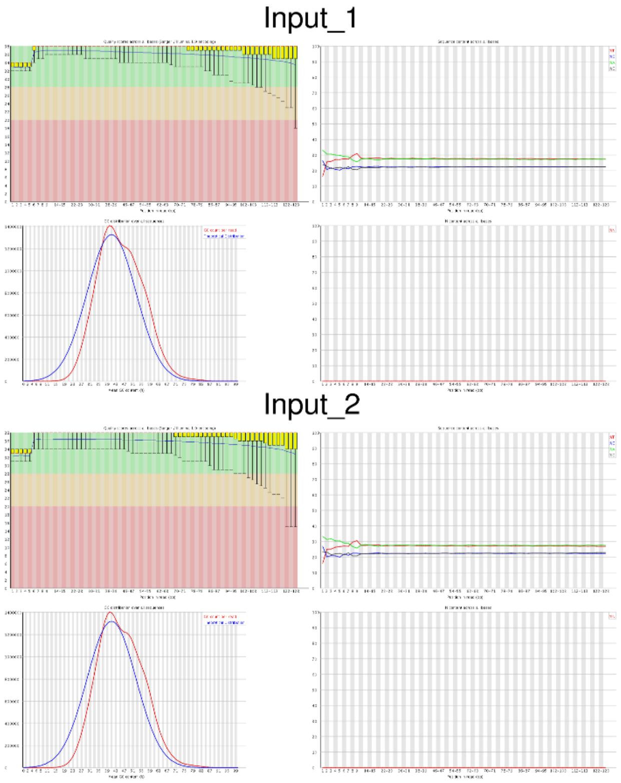


Figure 1-2 FastQC of clean reads

Top-left: horizontal axis represents position in read (bp), vertical axis represents quality scores, different colors represent different quality range.

Top-middle: horizontal axis represents position in read (bp), vertical axis represents base percentage.

Top-right: the distribution of read length.

Bottom-left: mean GC content (%). Red line is the actual GC content distribution, blue line is the theoretical distribution.

Bottom-middle: N-ratio across all bases.

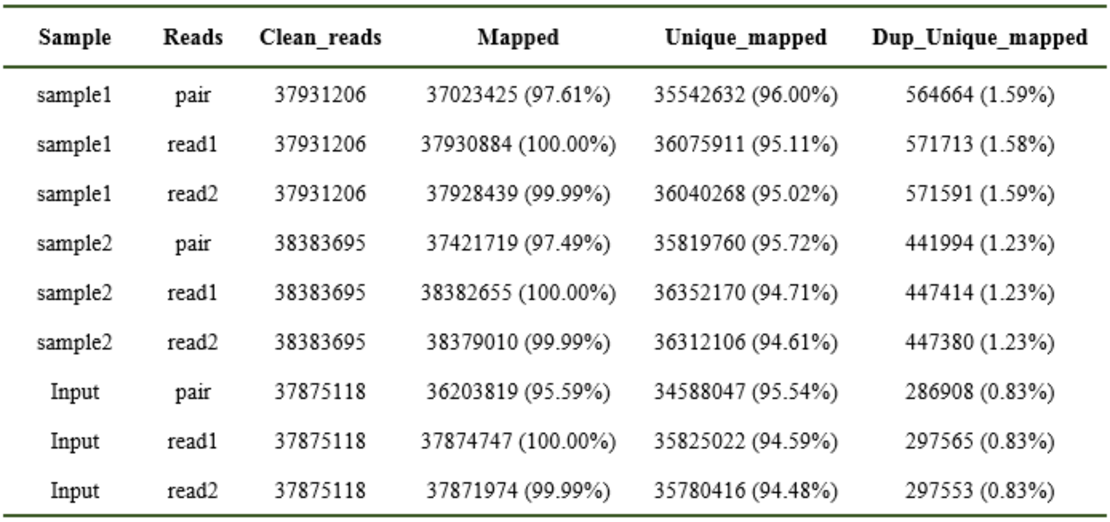
Bottom-right: x-axis is the sequence duplication level. y-axis is the percent of the deduplicated reads.

2 Mapping

2.1 Summary of mapping

The common tools for mapping are Bowtie, BWA, MAQ, TOPhat, etc. We choose proper softwares and parameters according to different genome characters to do the genome mapping analysis for the filtered reads. Considering the small fragment size of ChIP-Seq, and the percentage of the unique sequence in the total sequence is the most important information, thus, we can map the reads to the reference genome using BWA much more accurately (Li, H. and R. Durbin, 2009). The summary of mapping is in the following table:

Table 2-1 Summary of mapping



Sample	Reads	Clean_reads	Mapped	Unique_mapped	Dup_Unique_mapped
sample1	pair	37931206	37023425 (97.61%)	35542632 (96.00%)	564664 (1.59%)
sample1	read1	37931206	37930884 (100.00%)	36075911 (95.11%)	571713 (1.58%)
sample1	read2	37931206	37928439 (99.99%)	36040268 (95.02%)	571591 (1.59%)
sample2	pair	38383695	37421719 (97.49%)	35819760 (95.72%)	441994 (1.23%)
sample2	read1	38383695	38382655 (100.00%)	36352170 (94.71%)	447414 (1.23%)
sample2	read2	38383695	38379010 (99.99%)	36312106 (94.61%)	447380 (1.23%)
Input	pair	37875118	36203819 (95.59%)	34588047 (95.54%)	286908 (0.83%)
Input	read1	37875118	37874747 (100.00%)	35825022 (94.59%)	297565 (0.83%)
Input	read2	37875118	37871974 (99.99%)	35780416 (94.48%)	297553 (0.83%)

Unique_mapped: reads with MAPQ (Li and Ruan et al., 2008) no lower than 13; can be interpreted as the chance of non-accurate mapping (same score for the random mapping) is 0.05.

Duplicates: the reads mapped to the exact same position of the genome;

Mapped is relatively to Clean_reads;

Unique_mapped is relatively to Mapped;

Dup_Unique_mapped is relatively to Unique_mapped; For human and point-source factors, the recommended Unique_mapped region should be at least 10M, and the repetition rate should be less than 20% (Landt, S. G. and G. K. Marinov, et al, 2012).

2.2 MAPQ

The most important thing during Chip-Seq analysis is the percentage of the unique sequence in the total sequence number. Duplicates were labeled using SAMBLAST (Faust and Hall, 2014) and mapping quality value was calculated (MAPQ). Proper quality value was chosen as the only threshold for mapping. Here we choose 13 as the threshold, which means that the mapping chance of the accordingly non-unique region is only 0.05. Only keep one reads for the duplicates in the followed peak calling.

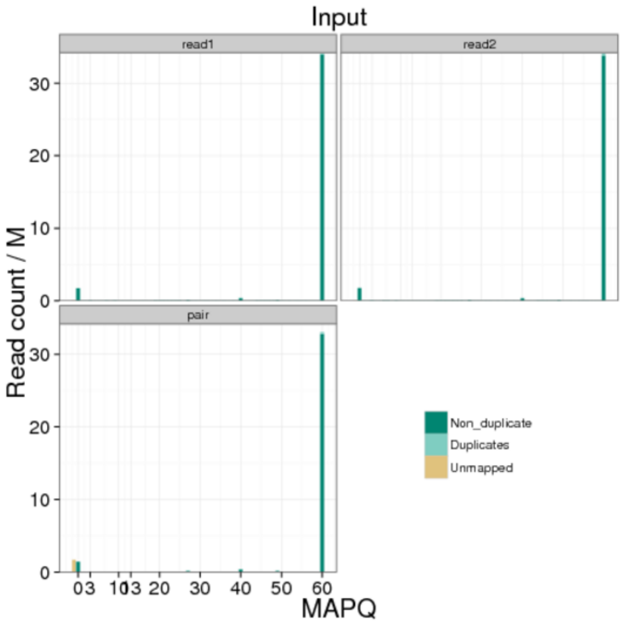


Figure 2-1 Distribution of MAPQ

Horizontal axis is MAPQ, vertical axis is the reads count.

2.3 Genome-wide distribution of the mapped reads

Summary of the density of total mapped reads in different chromosomes (plus and minus) is shown in the followed figure. With 5k slide window size, calculate the medium of the number of reads mapped to each base within window, and convert it to log2. The longer the whole chromosome is, the more reads are mapped (Marquez et al. 2012). From the correlation of the number of the mapped reads and the length of the chromosome in the figure, we can see the correlation of the total number of the reads and the length of the chromosome much easier.

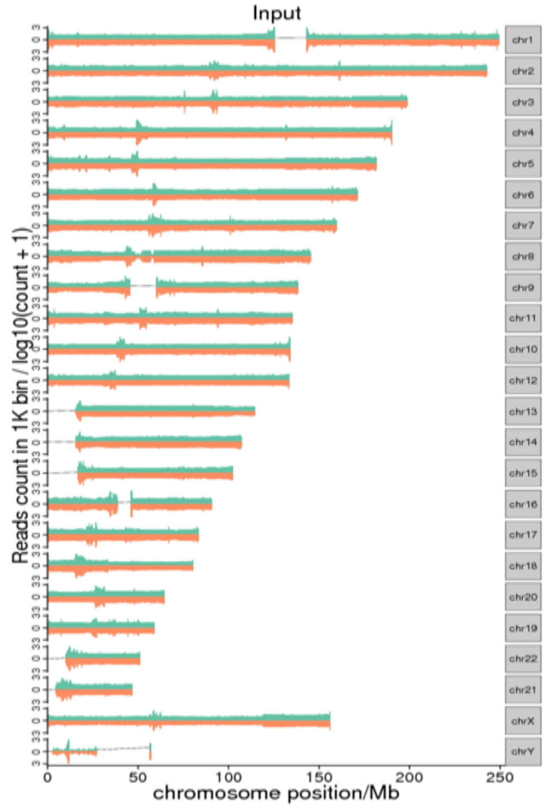


Figure 2-2 Genomewide distribution of the mapped reads

Horizontal axis represents the postition of the chromosome, vertical axis represents the number of the reads mapped to 1000nt window size. Here is the unique mapping and deduplication results.

2.4 Distribution of the reads mapped to the gene

Since the binding sites of transcription factor and histone protein are important for gene regulation, thus, analysis of relative mapping position distribution can help us predict the protein function. Divide each gene and its 2kb upstream and 2kb downstream into 100 equal parts. Calculate mapped reads in each part, and the percentage ratio of the reads in each part to total reads as reads density.

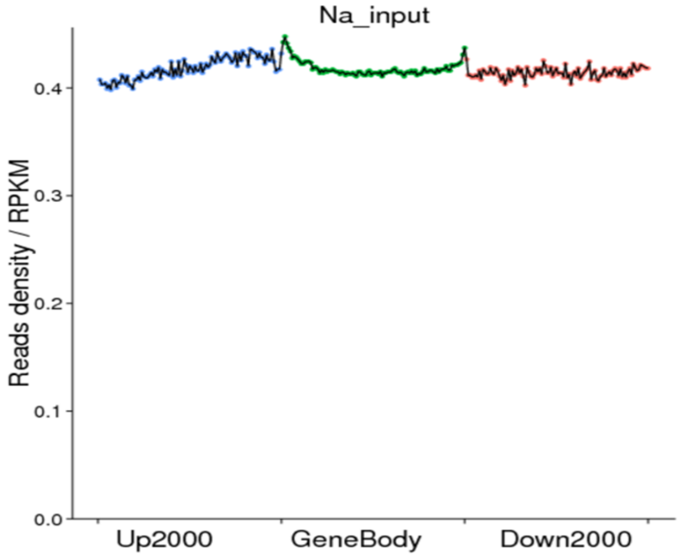


Fig2-3 Distribution of the reads mapped to gene

Horizontal axis: relative position of the gene.
vertical axis： reads density.

2.5 Sample correlation detection

Biological replicate is necessary for every experiment, same for high-throughput technology (Hansen et al.). Biological replicate mainly has two applications: One is to prove that the biological experiment can be replicated and there is no large variance. The other one is to make sure that following differential gene analysis can get reliable results. The correlation among samples is an important index to see whether the experiment design is reliable and whether the sampling is right. The correlation coefficient is much closer to 1, the similarity of the expression pattern among samples is much higher.

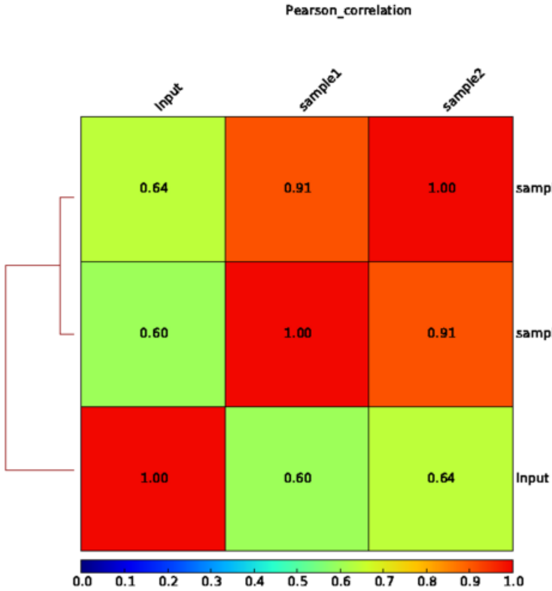


Figure 2-4 Pearson test among samples

Heat-map of sample correlation test, correlation coefficient among samples (Pearson correlation coefficient). The darkness of the color represents how large is the correlation coefficient.

2.6 Visualization of pileup signal

We provide the visualization results of genome wide reads mapping in bam format. IGV (Integrative Genomics Viewer) browser is recommended to view the bam file. IGV browser has the following characters: (1) can reveal single or multiple mapping positions in the genome in different scales, including the distribution of the reads in different chromosomes, and the distribution of annotated exons, introns, splicing junctions and inter-gene region; (2) can reveal reads abundance in different region under different scales which reflect the expression level; (3) can reveal the annotation information of the gene and alternative splicing isoforms; (4) can reveal other annotation information; (5) can download annotation information from remote and local server. Please check the IGV manual for detail instruction (IGVQuickStart.pdf).

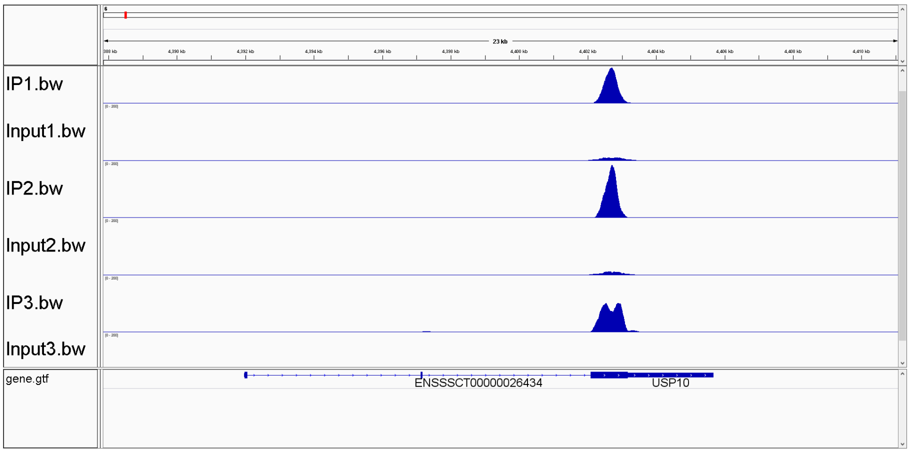


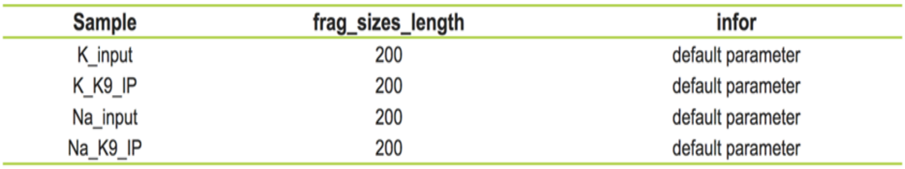
Figure 2-5 Visualization of pileup signal by IGV (demo)

3 Fragment size prediction

3.1 Summary of fragment size

For a specific binding site, there is a significant reads enrichment in the binding site. For single-end sequencing, we use MACS2 software to predict the frag_sizes of IP experiment. MACS2 scan the whole genome using certain window size and calculate the enrichment level of the reads in each window. Then extract (eg.1000) proper windows as the samples to build the enrichment model to predict the length of frag_sizes. For double-end sequencing, we use RSeQC software to predict the frag_sizes for the mapping results. Use the predicted frag_sizes for later peak calling.

Table 3-1 summary of frag_size



Sample	frag_sizes_length	infor
K_input	200	default parameter
K_K9_IP	200	default parameter
Na_input	200	default parameter
Na_K9_IP	200	default parameter

Sample: sample name

frag_sizes_length: frag_sizes length

infor: default parameter

3.2 K distribution of fragment size

Length distribution using kernel density method is in the following:

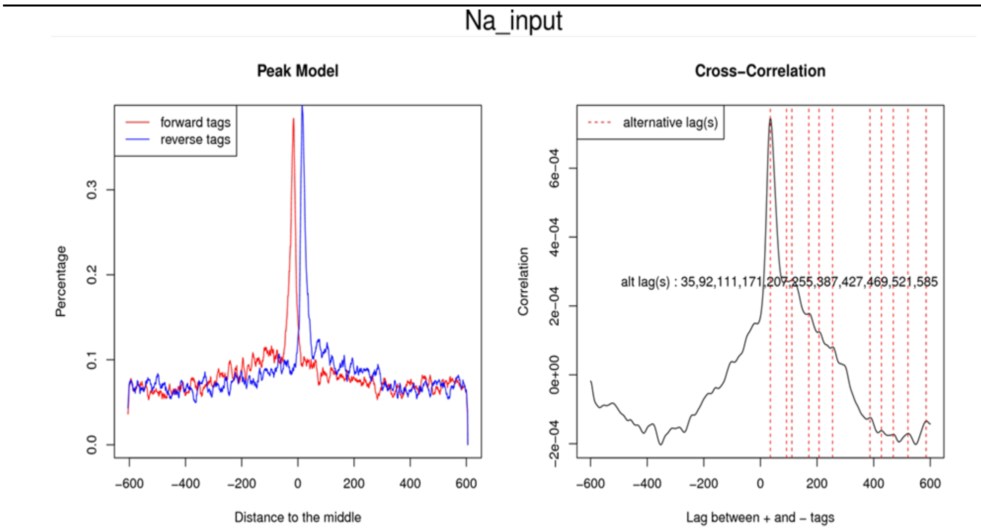


Figure 3-1 frag_size distribution

Single end:

left: horizontal axis is the distance to the center of peak model. vertical axis is the percentage of the reads chosen for modeling.

Red color represents plus strand, blue color represents minus strand.

Right: horizontal axis is the distance to the middle of the peak model. vertical axis is the correlation between plus strand and minus strand. The distance from the red dash line to the middle is the predicted frag_sizes.

Double end:

horizontal axis is the length of the predicted length of frag_sizes, vertical axis is the value of kernel density.

4 Strand cross correlation

4.1 Summary of strand cross correlation

As we know that the measured reads will be approximately distributed to plus and minus tags in average. In this way, by calculating the correlation of plus and minus strand (SCC), we can test the best distance between two strands. By testing the SCC of IP and input data, we can not only obtain the correlation coefficient between plus and minus strand, but can also test the effect of IP experiment.

Table4-1 Summary of strand cross correlation

Sample	Median_read_length	Predicted_fragment_length	CC_min	CC_read_length	CC_fragment_length	NSC	RSC	Description
sample1	125	250	0.4282	0.4357	0.4568	1.0668	3.7965	The immunoprecipitation seems to be successful.
sample2	125	240	0.4312	0.4465	0.4671	1.0831	2.3444	The immunoprecipitation seems to be successful.
Input	125	240	0.4259	0.4380	0.4541	1.0661	2.3268	The immunoprecipitation seems to be successful.

Sample: sample name

Median_read_length: mean value of reads length

Predicted_fragment_length: predicted length of fragment sizes

CC_min: the lowest SCC

CC_read_length: the SCC of the longest reads

CC_fragment_lengt: fragment sizes relative SCC

NSC: normalized strand coefficient, no less than 1.05

RSC: relative strand correlation, no less than 0.8

4.2 Plots of strand cross correlation

SCC curve for all the samples reveal the enrichment from different IP or different experiments.

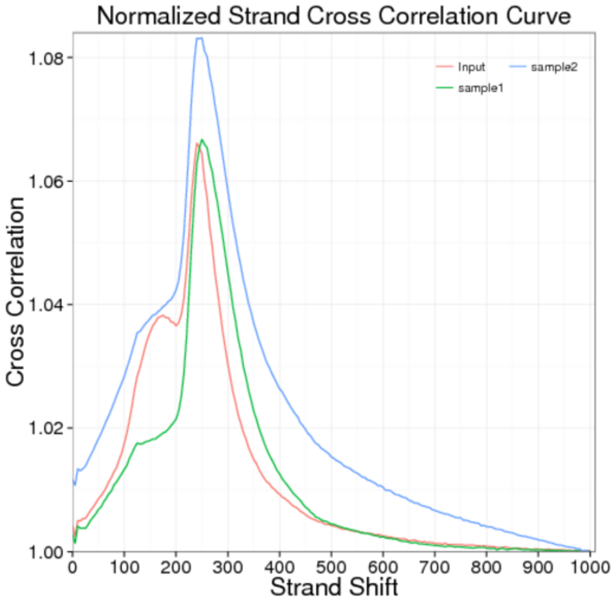


Figure4-1 Plots of strand cross correlation

Different color represents different samples.

4.3 SCC distribution between experimental groups

The SCC curve of successful IP in the same experimental group (including IP and Input) in the frag_size has a peak. The ratio of the CC (Cross Correlation) value of this peak to the lowest CC value of the whole SCC curve (NSC) should be no less than 1.05. Besides, there is another peak (shadow peak) in the reads. RSC value (should be no less than 0.8) need consider the ratio of two differences, the difference between the CC value in the frag_size and the lowest CC value, and the difference between CC value in the shadow peak and the lowest CC value.

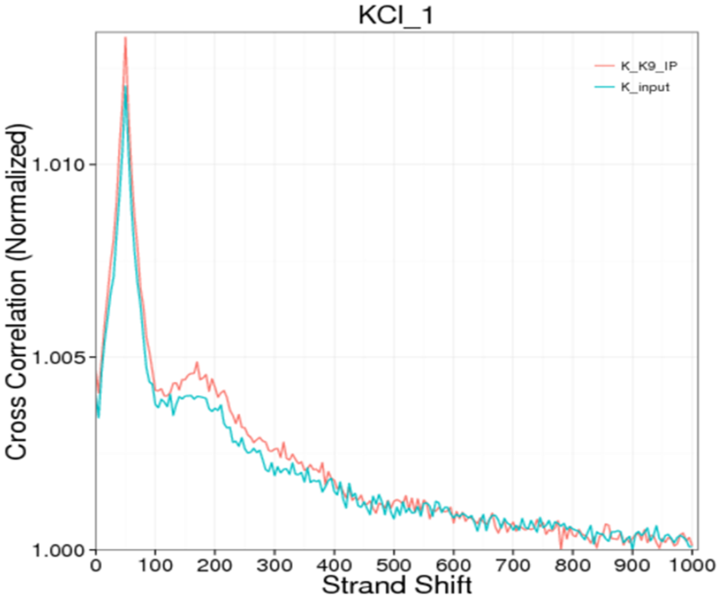


Figure4-2 SCC curve

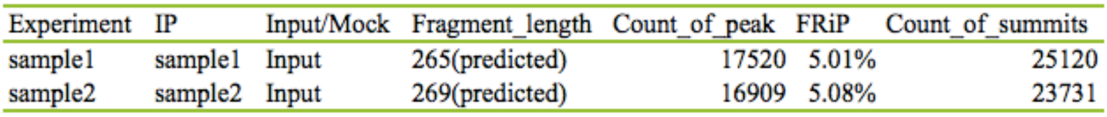
Horizontal axis is the lag between plus and minus tags during Pearson Correlation Coefficient calculation. Vertical axis is the Pearson Correlation Coefficient.

5 Peak Calling

The annotation of transcription factor binding sites, histone binding sites is the important information for understanding the regulation mechanism and function. Recent developed next generation sequencing technology sequence DNA after Chromatin Immunoprecipitation directly. By mapping to the reference sequence can obtain the information of protein-DNA binding sites directly. By making use of MACS2 software (Yong Zhang,Tao Liu et al., 2008) (threshold q value=0.05) to finish the peak calling, we can calculate the number of peaks, the peak width and its distribution, and find the peak related genes. The results are in the following:

5.1 Summary of peak calling

Table5-1 Summary of peak calling



Experiment	IP	Input/Mock	Fragment_length	Count_of_peak	FRiP	Count_of_summits
sample1	sample1	Input	265(predicted)	17520	5.01%	25120
sample2	sample2	Input	269(predicted)	16909	5.08%	23731

Experiment: experimental group name(one ChIP Experiment includes an IP and a control, eg. Input or Mock or no control);

IP:experiment name after chip handling.

Input/Mock: control group

Frag_length: predicted value of frag_sizes length;

Count_of peak: the number of peak (narrow). If would like to test broad peak, need annotate in the information collection form.

FRiP: the ratio of the numer of the reads in the peak to the total reads, which can test the effect of IP experiment.

Count_of_summits: number of summits. Some peaks can have multiple summits due to close position.

5.2 Genome wide distribution of peaks

Summary of genome wide distribution of peaks is shown in the following figure. From the number of the peak mapping to the chromosome and its distribution can reflect the distribution of the protein binding sites. When the number of the chromosomes including the peak is larger than 30, only show the peak distribution in 15 chromosomes.

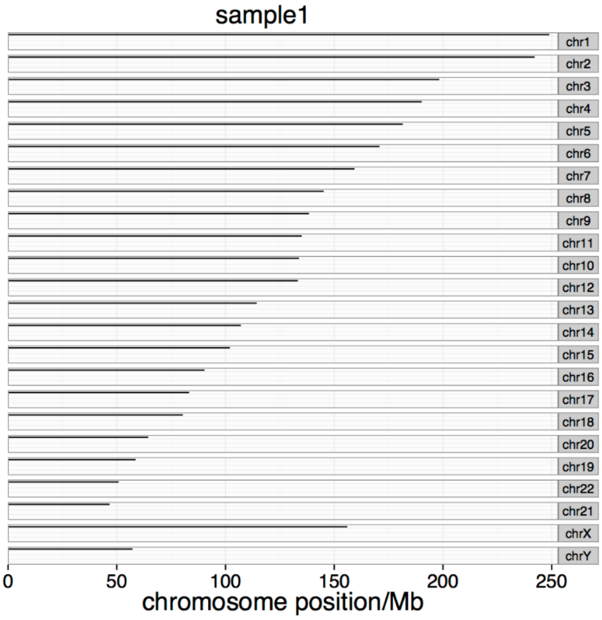


Figure 5-1 Genome wide distribution of peaks

Horizontal axis is the coordinate of the peak in the chromosome. Vertical is the chromosome. Every blue line represents a peak. When the number of the chromosomes including the peak is larger than 30, only show the peak distribution in 15 chromosomes.

5.3 Distribution of peak width

The peak width represents the length of the DNA that is bound by protein. Calculate the number of peaks using peak width as the measurement. The distribution of peak width is in the following:

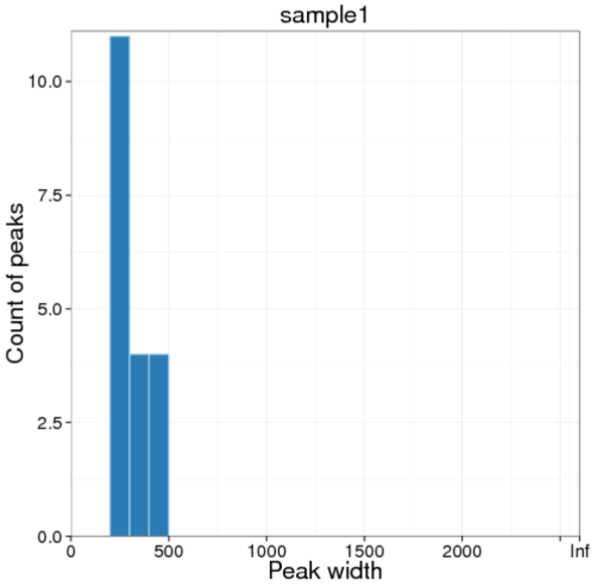


Figure 5-2 Distribution of peak width

Horizontal axis is the width of the peak (nt), vertical axis is the number of the corresponding peaks.

5.4 Distribution of fold enrichment

The fold enrichment value here can be also called as signal value, which is the digital display of the peak signal during peak calling. The larger the value, the more reads are enriched to this peak. Calculate the number of the peaks using fold change. The distribution of fold enrichment is in the following:

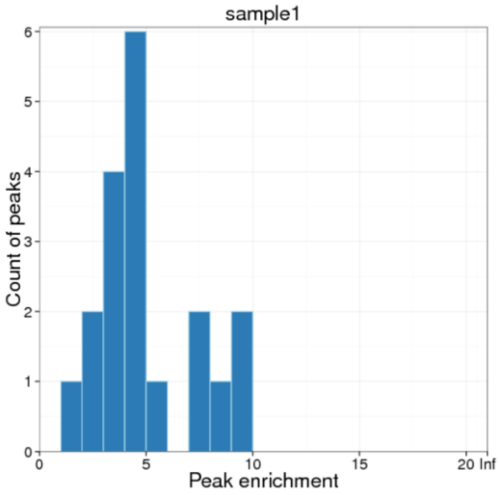


Figure 5-3 Distribution of fold enrichment

Horizontal axis represents the enrichment fold change of the peak. Vertical axis represents the number of the peaks.

5.5 Distribution of q values

The significance of the peak is the measurement of the confidence level. Calculate the q value for eah peak. Calculate the number of peaks using the significance of peak as the measurement. The significance distribution is in the following:

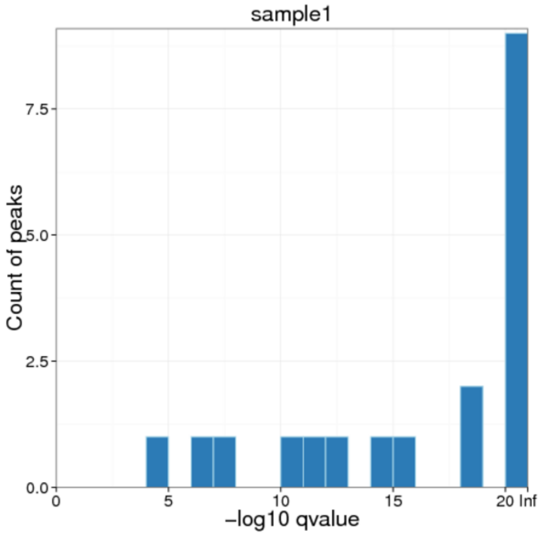


Figure 5-4 Distribution of q values

Calculate the q value in each peak; Horizontal axis represents -log10 q value of the peak; Vertical axis represents the number of the peaks.

5.6 Count of summits in peaks

Calculate the number of summits in each peak, and infer the type of peak in IP experiment.

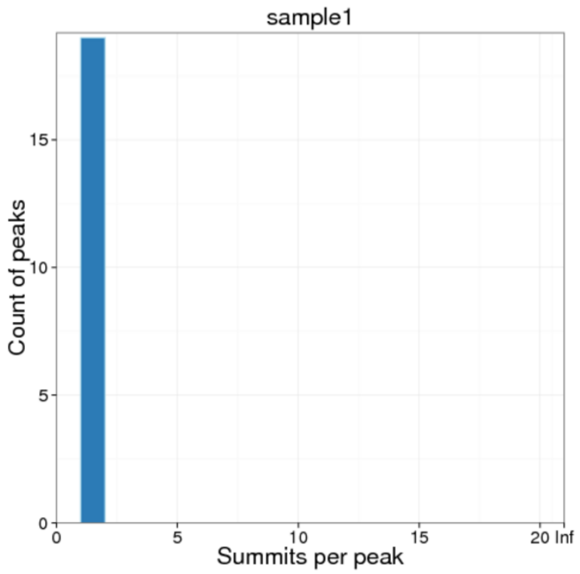


Figure 5-5 Count of summits in peaks

Horizontal axis represents the number of summit in each peak. Vertical axis is the corresponding number of the peak.

5.7 Summits distribution

Each peak was divided to 100bp windows and the summits in each window of all the peaks were counted in the following:

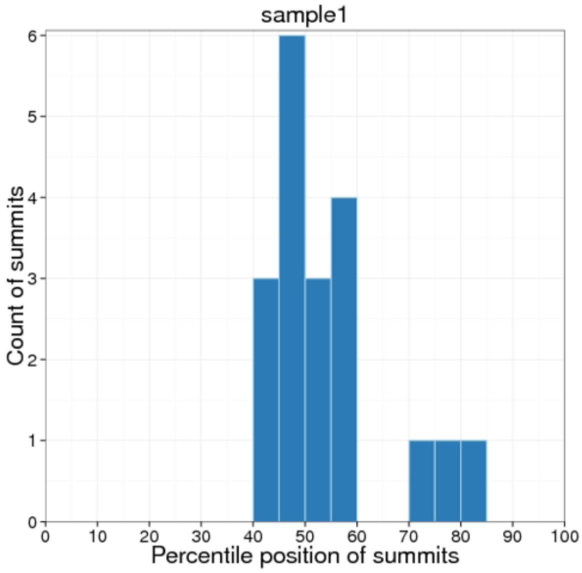


Figure 5-6 Percentile position of summits in peaks

Horizontal axis is the position of sumits, vertical axis is the count of the summits.

6 Motif analysis

The binding of protein such as transcription factor, histone etc. and DANN is not random, instead, has sequence preference. Motif analysis can not only detect protein specific binding sites but can also obtain the annotated motif and it's binding site, motif sequence information etc. By using MEME(Timothy L. Bailey and Charles Elkan,1994)and Dreme (Timothy L. Bailey,2011) softwares to detect significant motif sequence in the peak. By using Tomtom (Shobhit Gupta, JA Stamatoyannopolous,2007) software can annotate the motif by mapping it to the annotated Motif database.

Use sequence logo to show the base bias in different position in the binding sites in long Motif (8~30) (Fig. 6-1) and short Motif (~8) (Fig. 6-2). The results are in the following:

(Note: because of binding site specification, motif sequence can only show in one region (<=8 or >=9), so that part of the figure in the following has empty result).

6.1 Motif searching

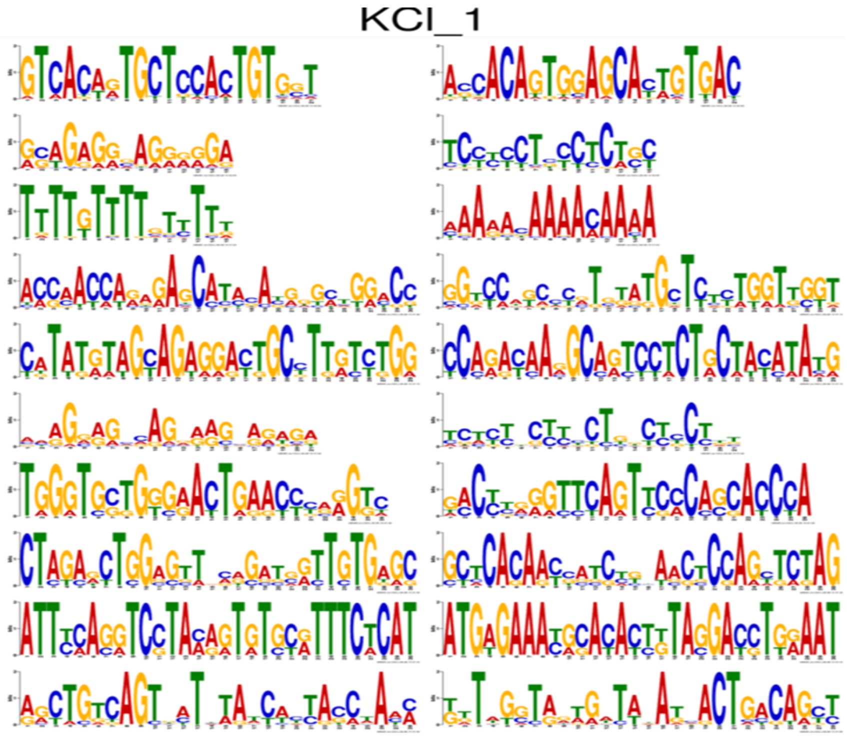


Figure 6-1 long conservative sequence

Logo is listed in order. The figure in the right is the reverse complement sequence.

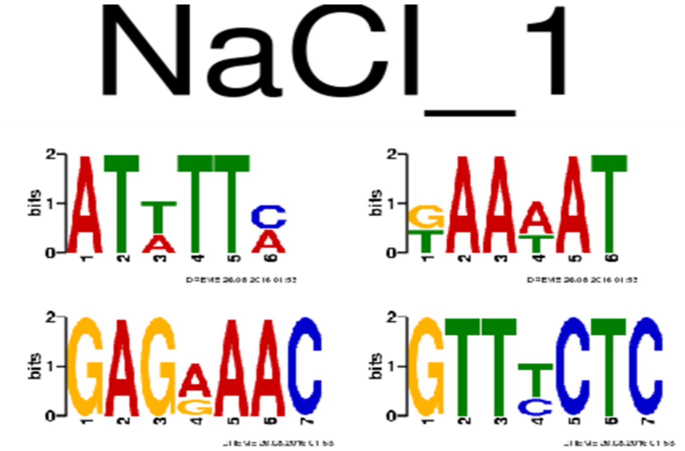


Fig. 6-2 short conservative sequence

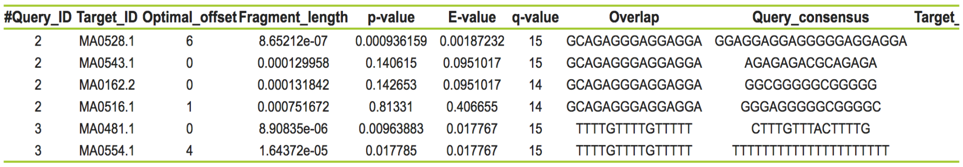
Logo is listed in order. The figure in the right is the reverse complement sequence.

6.2 MEME motif annotation

Map the deteced motif sequence from MEME to the annotated motif using Tomtom. The result is in the following:

(Note: since the conservative sequence in the binding site is short, motif can be only in one region (<=8), resulting in no result for meme test).

Table 6-1 MEME detection Motif calculation



#Query_ID	Target_ID	Optimal_offset	Fragment_length	p-value	E-value	q-value	Overlap	Query_consensus	Target_
2	MA0528.1	6	8.65212e-07	0.000936159	0.00187232	15	GCAGAGGGAGGAGGA	GGAGGAGGAGGGGGAGGAGGA	
2	MA0543.1	0	0.000129958	0.140615	0.0951017	15	GCAGAGGGAGGAGGA	AGAGAGACGCAGAGA	
2	MA0162.2	0	0.000131842	0.142653	0.0951017	14	GCAGAGGGAGGAGGA	GGCGGGGGCGGGGG	
2	MA0516.1	1	0.000751672	0.81331	0.406655	14	GCAGAGGGAGGAGGA	GGGAGGGGGCGGGGC	
3	MA0481.1	0	8.90835e-06	0.00963883	0.017767	15	TTTTGTTTTGTTTTT	CTTTGTTTACTTTTG	
3	MA0554.1	4	1.64372e-05	0.017785	0.017767	15	TTTTGTTTTGTTTTT	TTTTTTTTTTTTTTTTTTTTT	

#Query ID: detected motif;

Target ID: known motif ID in the database;

Optimal offset: the number of lag bases;

p-value:probability of MCMC;

E-value:false positive probability;

q-value: FDR value;

Overlap: overlapping base pair between two sequences;

Query consensus: detected motif sequence;

Target consensus: motif sequence in the target database;

Orientation: plus or minus strand for the target sequence;

Using sequence logo to show the comparison results between MEME detected motif and known motif. The result is in the following:

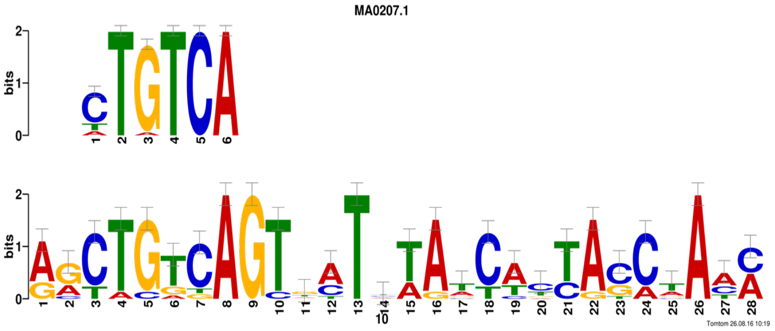


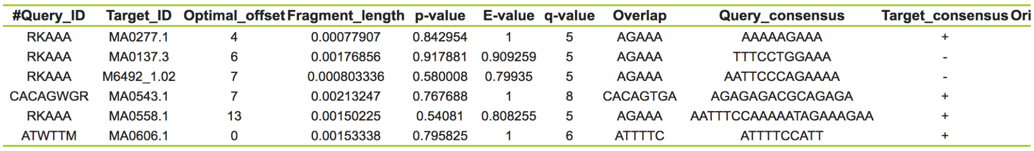
Figure 6-3 MEME motif annotation

6.3 Dreme motif annotation

The comparison between the motif detected by Dreme use Tomtom and known motif:

(Note: Since the conservative sequence in the binding sites is long, the motif could only show up in one region (>=8), which could show empty result with Dreme detection)

Table 6-2 Summary of motif detected by Dreme



#Query_ID	Target_ID	Optimal_offset	Fragment_length	p-value	E-value	q-value	Overlap	Query_consensus	Target_consensus	Ori
RKAAA	MA0277.1	4	0.00077907	0.842954	1	5	AGAAA	AAAAAGAAA	+	
RKAAA	MA0137.3	6	0.00176856	0.917881	0.909259	5	AGAAA	TTTCCTGGAAA	-	
RKAAA	M6492_1.02	7	0.000803336	0.580008	0.79935	5	AGAAA	AATTCCCAGAAAA	-	
CACAGWGR	MA0543.1	7	0.00213247	0.767688	1	8	CACAGTGA	AGAGAGACGCAGAGA	+	
RKAAA	MA0558.1	13	0.00150225	0.54081	0.808255	5	AGAAA	AATTTCCAAAAATAGAAAGAA	+	
ATWTTM	MA0606.1	0	0.00153338	0.795825	1	6	ATTTTC	ATTTTCCATT	+	

#Query ID: detected motif ID;

Target ID: known motif ID in the database;

Optimal offset: lag base number;

p-value: probability of MCMC;

E-value: false positive probability;

q-value: FDR value;

Overlap: overlapping base between two sequences;

Query consensus: detected motif base sequence;

Target consensus: motif base sequence in the target database;

Orientation: plus or minus tag for the targeted sequence;

Sequence logo showing the comparison result between detected motif with Dreme and known motif is in the following:

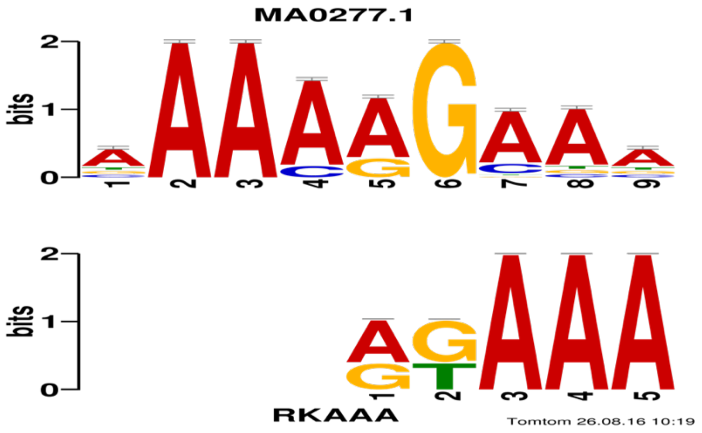


Figure 6-4 Dreme motif annotation

7 Peak annotation

7.1 Peak-TSS distance

Peak-TSS distance distribution can predict protein binding sites. One can estimate IP effect according to protein binding sites. One can predict protein regulatory mechanism or function according to the protein binding character. TSS (transcription start site) of every peak related gene TSS are detected using PeakAnnotator (Salmon-Divon and Dvinge et al., 2010). Calculate peak numbers according to peak-TSS distance, and analyze peak-TSS distance distribution.

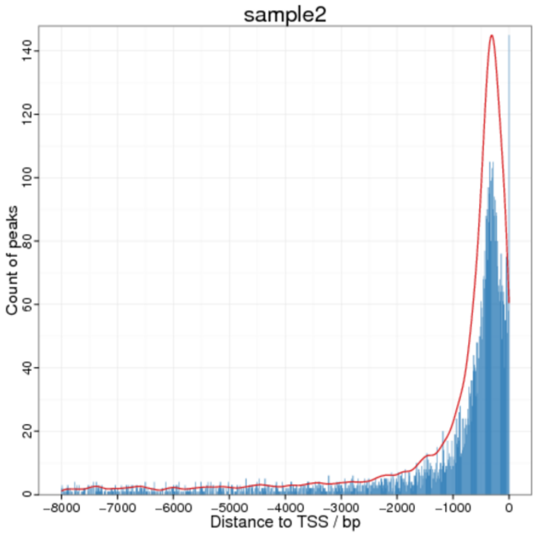


Figure 7-1 Distribution of Peak-TSS Distance

Horizontal axis represents the distance from peak to TSS. Vertical axis represents the number of the peak.

7.2 Distribution of peaks in functional region

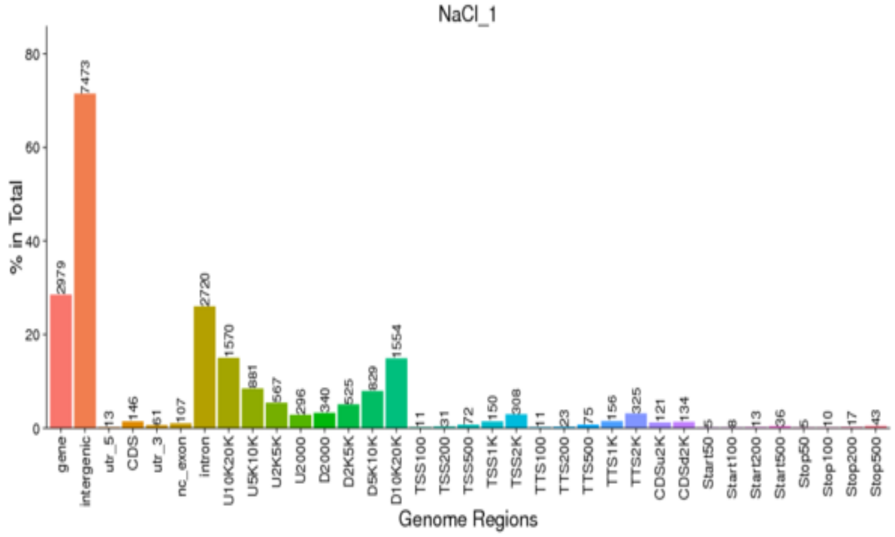


Figure 7-2 peak distribution in functional gene region

Distribution of peak in different functional area. Horizontal axis represents different functional area, vertical axis represents the ratio of the peak in the functional region to the total peaks. The number on the top of the functional region represents peak number.

U2000 means 2000bp in the upstream region, D2000 means 2000bp in the downstream region;
CDSu2K and CDSd2K means upstream and downstream 2kb of CDS;
TSS100, Stop100, start100 means 100bp centered with TSS,TTS, Start-codon and stop-codon.

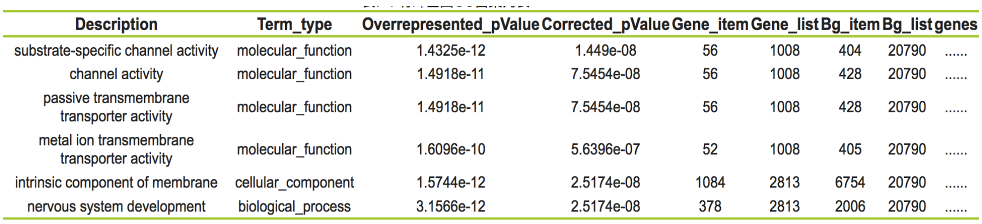
7.3 GO enrichment analysis

Gene Ontology (GO, http://www.geneontology.org/) is international standard classification system for gene function attributes. As the database built by Gene Ontology Consortium, it aims to build up an updating language vocabulary standard to describe gene and protein function for all species. GO covers three domains: Molecular Function, Biological Process and Cellular Component.

Gene or protein can find the corresponding GO accession number by ID access or sequence annotation. And every GO number can find its corresponding Term, which is the function classification or cellular localization.

Any gene whose position is overlapped with peak is peak related gene. Results of GO enrichment is in the following:

Table 7-1 peak related gene GO enrichment



Description	Term_type	Overrepresented_pValue	Corrected_pValue	Gene_item	Gene_list	Bg_item	Bg_list	genes
substrate-specific channel activity	molecular_function	1.4325e-12	1.449e-08	56	1008	404	20790	......
channel activity	molecular_function	1.4918e-11	7.5454e-08	56	1008	428	20790	......
passive transmembrane transporter activity	molecular_function	1.4918e-11	7.5454e-08	56	1008	428	20790	......
metal ion transmembrane transporter activity	molecular_function	1.6096e-10	5.6396e-07	52	1008	405	20790	......
intrinsic component of membrane	cellular_component	1.5744e-12	2.5174e-08	1084	2813	6754	20790	......
nervous system development	biological_process	3.1566e-12	2.5174e-08	378	2813	2006	20790	......

GO_accession: the unique GO ID;

Description: Function description;

Term_type: Including cellular_component,   biological_process and molecular_function;

Over_represented_pValue: significance of enrichment analysis;

Corrected_pValue: P-Value after correction, normally, with padj< 0.05 are enriched;

Gene_with_peak_item: the number of peak related genes with this GO term;

Gene_with_peak_list: the number of peak related gene;

Bg_item: the number of background gene with this GO term;

Bg_list: the number of background genes;

Genes: Annotated peak related gene ID (not show in this table);

Choose the first five results for each experiment.

7.4 GO enrichment

Peak overlapping gene GO enrichment bar, which directly reflect the distribution of the number of the peak overlapping genes enrichment on the biological process, cellular component and molecular function.

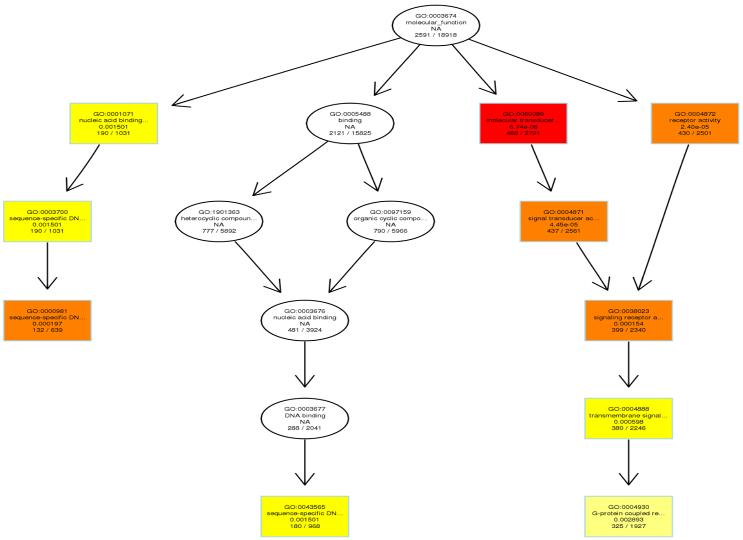


Fig 7-3 GO enrichment

Bar: horizontal axis is the number of the peak overlapping gene. Different color is used to differentiate biological process, cellular component and molecular function.

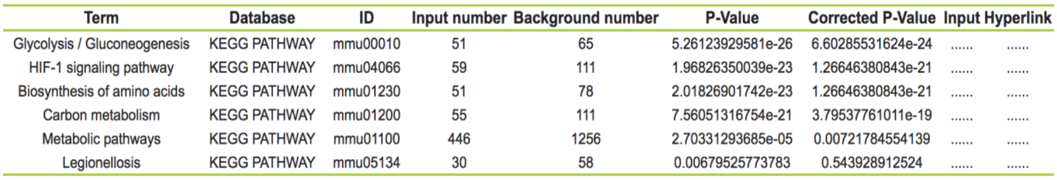
DAG figure: every note represents a GO project. Rectangular represents top 10 enrichment GO. Darkness represents the enrichment level. The darker of the color, the higher of the enrichment.

The name of the project and the significance (padj) are shown in every note.

7.5 KEGG enrichment analysis

Different genes coordinate to each other to realize the function in organism. Pathway enrichment can help identify the main biochemical metabolic pathway and signaling pathway of differential expressed genes. KEGG (Kyoto Encyclopedia of Genes and Genomes) is a database which can analyze gene function and genome information. It helps researchers study the genes and their expression as a whole network. As the main public database related to pathway (Kanehisa,2008), KEGG provides an excellent pathway search, including metabolism of carbohydrates, nucleotide, amino acids, etc. and organism biodegration. This not only provide the possible metabolism pathway, but also give a comprehensive annotation of the enzyme in each catalytic reaction, including amino acid sequence, PDB link etc. It is a powerful tool to study organism metabolism and do network analysis. Pathway enrichment makes use of pathway in the KEGG database as the unit, with super geometric test to find the significant enriched pathway in the differential expressed genes with the whole genome as the background.

Table 7-2 KEGG enrichment list



Term	Database	ID	Input number	Background number	P-Value	Corrected P-Value	Input	Hyperlink
Glycolysis / Gluconeogenesis	KEGG PATHWAY	mmu00010	51	65	5.26123929581e-26	6.60285531624e-24	......	......
HIF-1 signaling pathway	KEGG PATHWAY	mmu04066	59	111	1.96826350039e-23	1.26646380843e-21	......	......
Biosynthesis of amino acids	KEGG PATHWAY	mmu01230	51	78	2.01826901742e-23	1.26646380843e-21	......	......
Carbon metabolism	KEGG PATHWAY	mmu01200	55	111	7.56051316754e-21	3.79537761011e-19	......	......
Metabolic pathways	KEGG PATHWAY	mmu01100	446	1256	2.70331293685e-05	0.00721784554139	......	......
Legionellosis	KEGG PATHWAY	mmu05134	30	58	0.00679525773783	0.543928912524	......	......

(1) Term: Description information of KEGG pathway.

(2) Database: KEGG PATHWAY.

(3) ID: The only pathway identifier in the KEGG database.

(4) Input number: the number of peak overlapping gene in the corresponding pathway.

(5) Background number: The number of the genes in the corresponding pathway.

(6) P-value: statistical significant level of enrichment analysis.

(7) Corrected p-value: statistical significant level after correction. Normally, if the corrected P-value < 0.05 this pathway is enriched.

(8) Input: The peak overlapping genes in the corresponding pathway. Because of too much genes, here we ignore it.

(9) Hyperlink: see result file.

7.6 Scatterplot of KEGG enrichment

Peak overlapping gene KEGG enrichment scatter (Fig7-4) is the visualization of KEGG enrichment analysis. In this figure, the extent of KEGG enrichment is measured by Rich factor, qvalue and the number of the genes that are enriched in this pathway. Rich factor is the ratio of the number of the genes in the corresponding pathway of the peak overlapping gene to the total number of all annotated genes in this pathway. qvalue is the pvalue after multiple hypothesis testing correction. The range of qvalue is from 0 to 1, with more close to 0, representing more significant for the enrichment. Here only shows the top 20 significant enriched pathways. If the enriched pathways are less than 20, then all of them are shown.

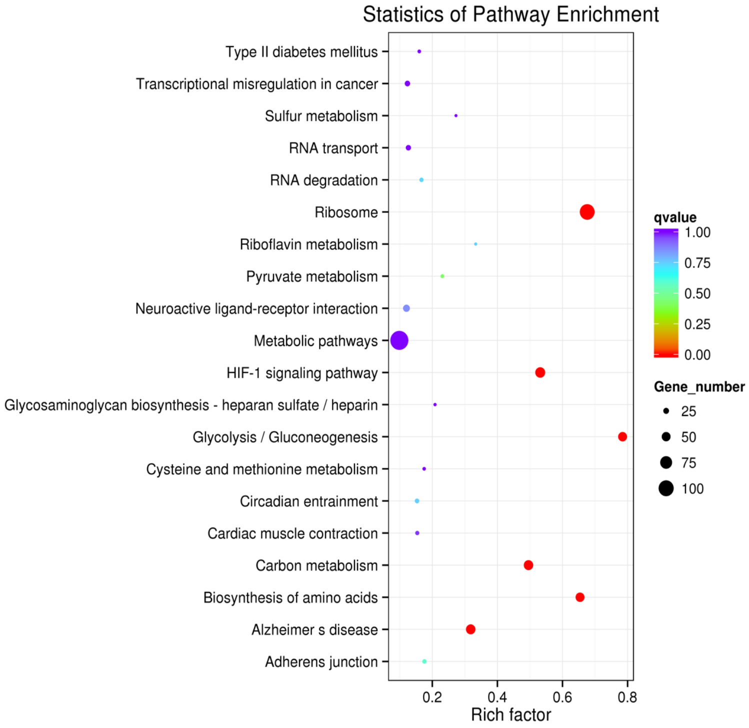


Figure 7-4 KEGG enrichment scatter

Vertical axis represents pathway name, horizontal axis represents Rich factor, the size of the dot represents the numbers of the overlapping peak genes in the pathway, and the color of the dots corresponds to different qvalue range. The size of the dots represents the number of the genes whose peaks are overlapping in this pathway.

8 Differential analysis

Only do differential analysis between groups when the number of groups is no less than two.

Calculate differential analysis using FoldEnrich of peak in different experimental groups (the ratio of RPM value in group A to group B). Finding differerntial binding sites in different groups by finding differential peak when the ratio of FoldEnrich is larger than 2. From which, we can find the differential binding sites related gene to do the follow-up annotation and enrichment.

8.1 Summary of differential peaks

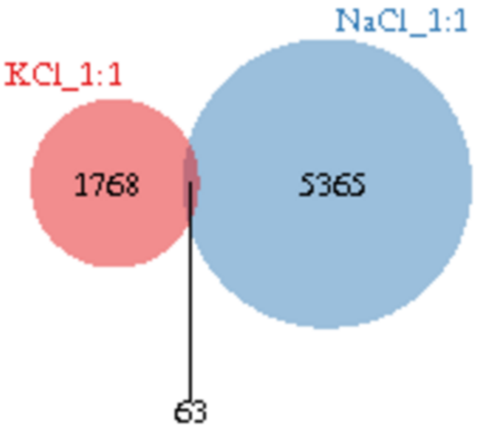


Figure 8-1-1 Venn diagram of differential peak in comparable groups

Here shows the peak number within comparable group (eg. A and B). Different colors represent different comparable groups (A or B). The sum of the number in the pie is the peak number of this comparable group. Single color is group specific peak number. Overlapping part is the common peak number between two comparable groups. Colon represents the number of appearance within the group.

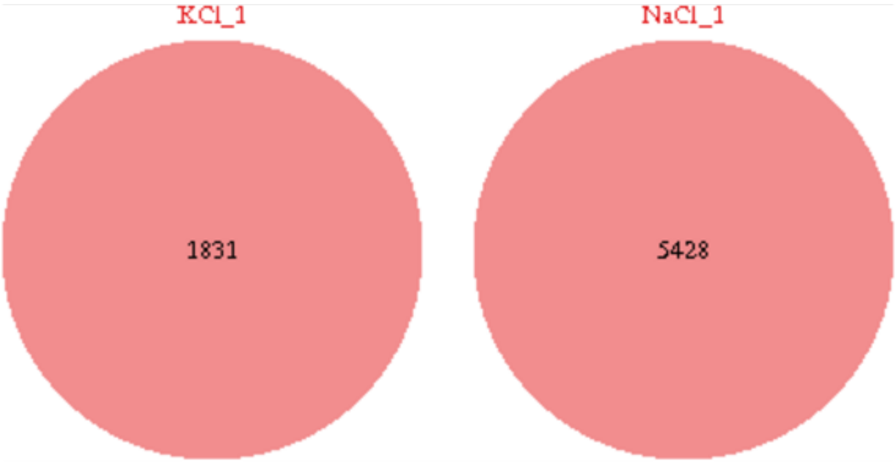


Figure 8-1-2 Venn diagram of differential peak in comparable gruops

Since some comparable groups (eg. A or B) include multiple experimental groups (eg. A1, A2, A3), here only show the number of peak among experimental groups within one comparable group (eg. A1&A2, A1&A3). Different colors represent different experimental groups (A1 or A2). The sum of the number in the pie is the peak number of this experimental group. Single color is group specific peak number. Overlapping part is the common peak number between two experimental groups.

8.2 Enrichment level analysis among different samples

Use RPM value of the peak in different samples (the ratio of 1M reads that enriched to the peak in a single sample) to do clustering analysis to determine the enrichment pattern of same peak in different samples or the enrichment difference of different peak in the same sample. At the same time, Enrichment comparison between IP and Input in the group can show the peak enrichment in IP experiment. To do hierarchical clustering based on RPM value in different samples. Different colors represent different clustering information. The more close between the IP sample enrichment, the more similar function or biological process they have.

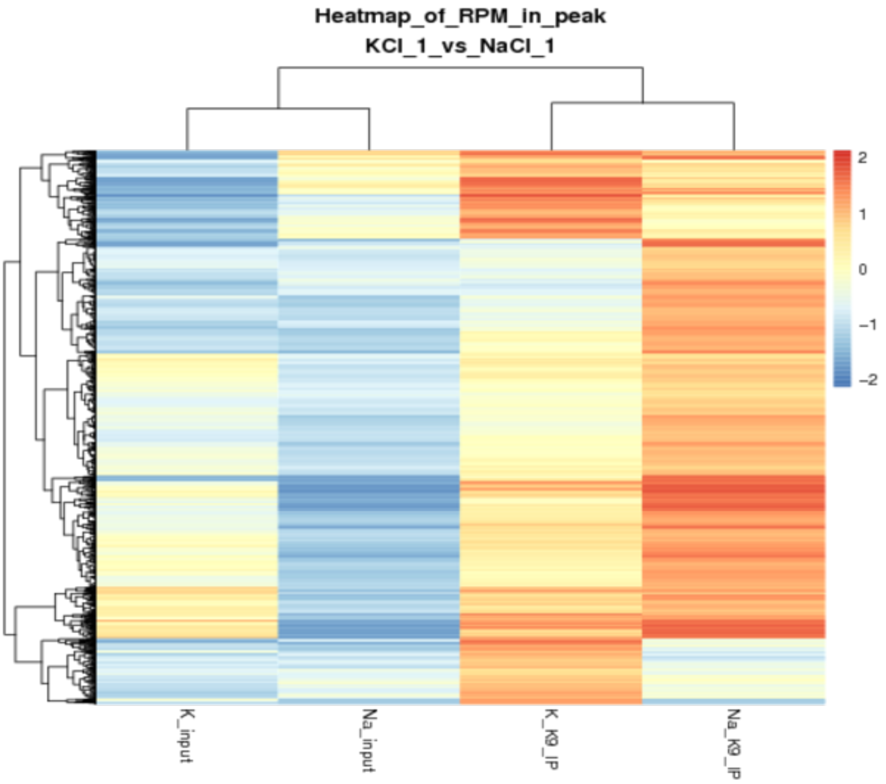


Figure 8-2 Enrichment analysis among samples

Top is the clustering among different samples in comparable groups.

Bottom is the correlation among different samples in comparable groups.

8.3 Reads density distribution among comparable groups for the annotated peaks

Calculate RPM value of each sample within peak annotated area (the ratio of 1M reads enriched to the peak in one sample). Using boxplot to show reads enrichment in the annotated peak area in different samples.

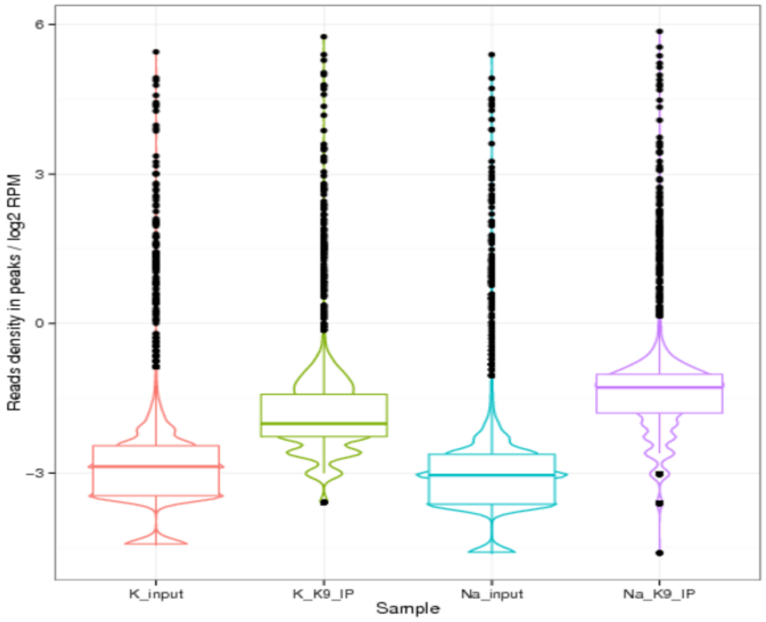


Figure 8-3 Boxplot of RPM value among samples in different groups

8.4 Enrichment level of different experimental peak analysis

Enrichment for the peaks by using the FoldEnrich value (the ratio of RPM from IP and Input) of the peaks from different experimental group. This is used to determine the enrichment pattern of the same peak in different experimental groups, or different peaks in the same experimental group. At the same time, by comparing the fold change of peak enrichment between different experimental groups, reveal the condition of peak enrichment among experimental groups to find differential enriched peak, which is the differential protein binding sites. Based on the FoldEnrich value from different experimental groups to do hierarchical clustering analysis. Different colors represent different clustering information.

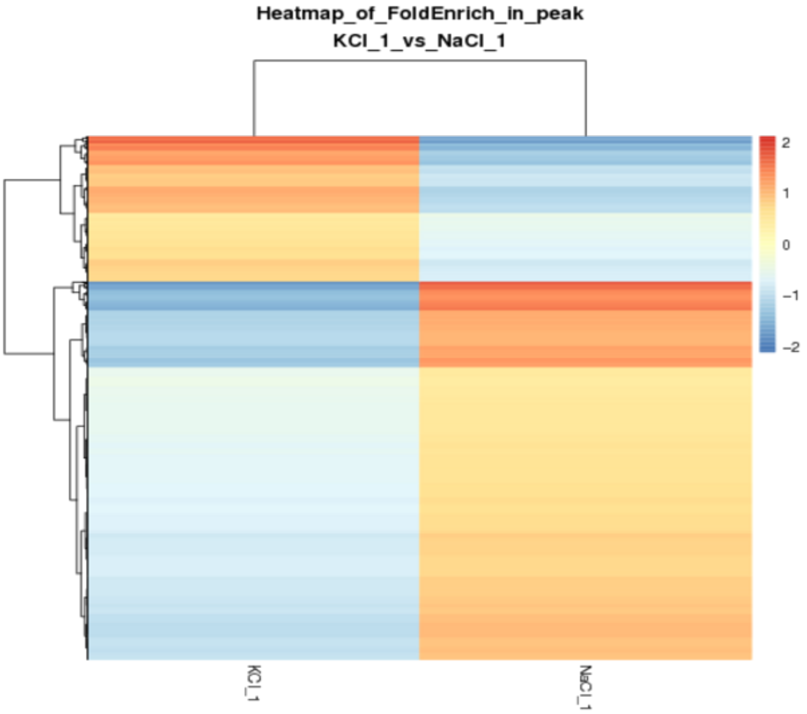


Figure 8-4 Peak enrichment analysis among different experimental groups

Top is the clustering analysis of peaks from different experimental groups.

Bottom is the correlation among different experimental groups.

8.5 Reads density distribution among different experimental groups for the annotated peaks

Calculate the FoldEnrich (the ratio of RPM from IP and Input) value in every comparable group in the peak annotation region. Enrichment level from different IP was shown by boxplot.

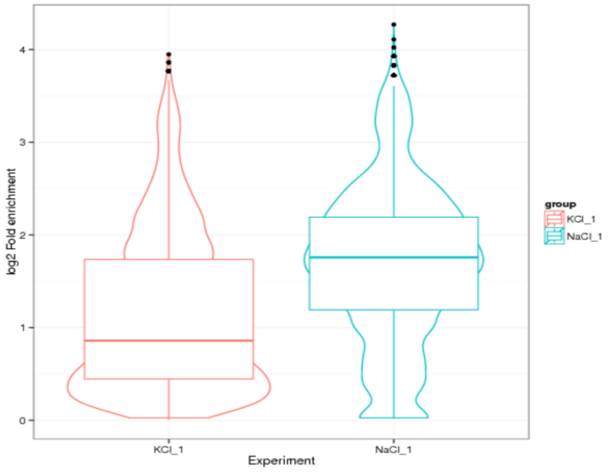


Figure 8-5 Reads density distribution in peak annotation region from different experiment

8.6 GO enrichment analysis of differential peak related genes

GO enrichment analysis for the differential peak related genes. Result is in the following:

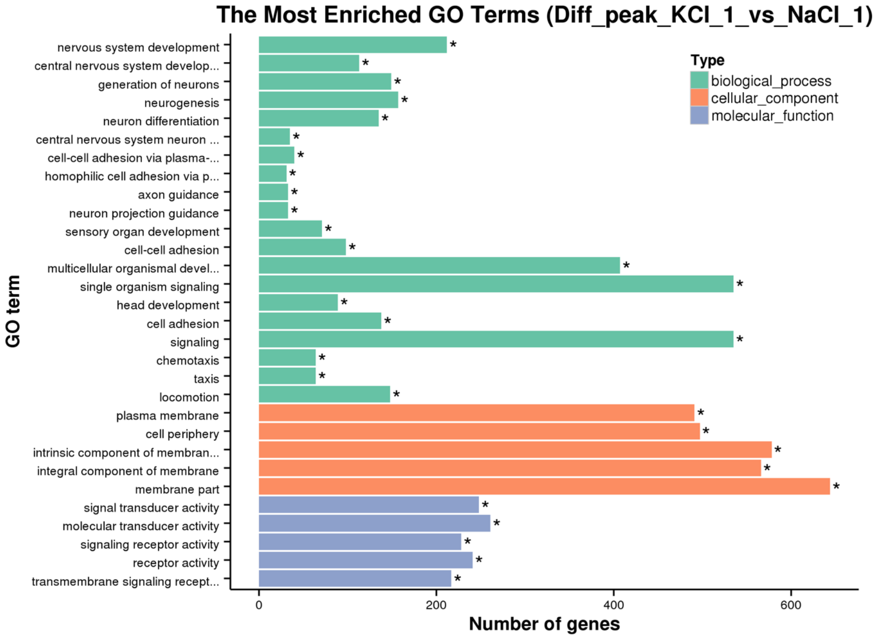


Figure 8-6 GO enrichment analysis for the differential peak genes

8.7 KEGG enrichment analysis of differential peak related genes

KEGG enrichment analysis for the differential peak genes. Result is in the following:

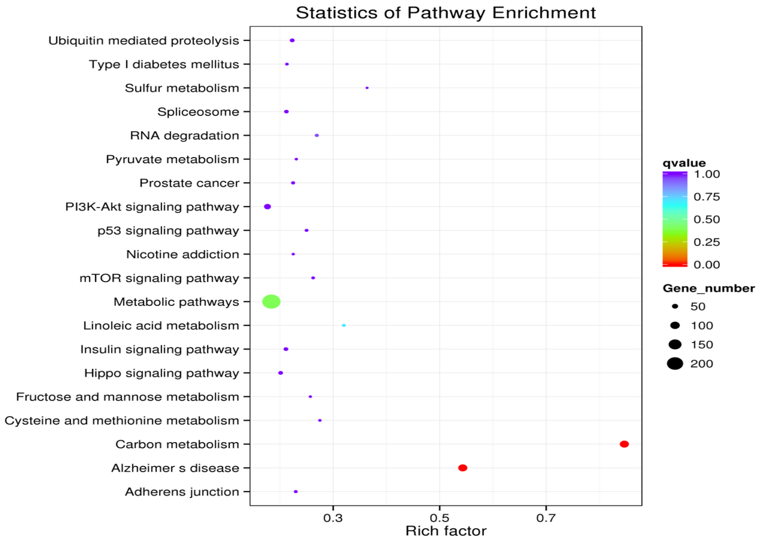


Figure 8-7 KEGG enrichment analysis for differential peak genes

III References

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at:http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Ashburner, M. and C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet 25 (1): 25-9.

Bailey, T. L. and N. Williams, et al. (2006). "MEME: discovering and analyzing DNA and protein sequence motifs." Nucleic Acids Res 34 (Web Server issue): W369-73.

Bailey, T. L. and M. Boden, et al. (2009). "MEME SUITE: tools for motif discovery and searching." Nucleic Acids Res 37 (Web Server issue): W202-8.

Bailey, T. and P. Krajewski, et al. (2013). "Practical guidelines for the comprehensive analysis of ChIP-seq data." PLoS Comput Biol 9 (11): e1003326.

Faust, G. G. and I. M. Hall (2014). "SAMBLASTER: fast duplicate marking and structural variant read extraction." Bioinformatics 30 (17): 2503-5.

Jiang, H. and R. Lei, et al. (2014). "Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads." BMC Bioinformatics 15: 182.

Kanehisa, M. and S. Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes." Nucleic Acids Res 28 (1): 27-30.

Kent, W. J. and A. S. Zweig, et al. (2010). "BigWig and BigBed: enabling browsing of large distributed datasets." Bioinformatics 26 (17): 2204-7. Available online at: http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/

Kharchenko, P. V. and M. Y. Tolstorukov, et al. (2008). "Design and analysis of ChIP-seq experiments for DNA-binding proteins." Nat Biotechnol 26 (12): 1351-9. Available online at: http://compbio.med.harvard.edu/Supplements/ChIP-seq/

Li, H. and R. Durbin (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform." Bioinformatics 25 (14): 1754-60.

Li, H. and J. Ruan, et al. (2008). "Mapping short DNA sequencing reads and calling variants using mapping quality scores." Genome Res 18 (11): 1851-8.

Li, H. and B. Handsaker, et al. (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics 25 (16): 2078-9.

Landt, S. G. and G. K. Marinov, et al. (2012). "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia." Genome Res 22 (9): 1813-31.

Mao, X. and T. Cai, et al. (2005). "Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary." Bioinformatics 21 (19): 3787-93.

Nicol, J. W. and G. A. Helt, et al. (2009). "The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets." Bioinformatics 25 (20): 2730-1. Available online at: http://bioviz.org/igb/index.html

Peter J.Park(2009). ChIP-seq: advantages and challenges of a maturing technology. Nature Reviews Genetics 10, 669-679.

Ramirez, F. and F. Dundar, et al. (2014). "deepTools: a flexible platform for exploring deep-sequencing data." Nucleic Acids Res 42 (Web Server issue): W187-91.

R Core Team (2015). R: A Language and Environment for Statistical Computing. Available online at: https://www.r-project.org/

Shirley Pepke, Barbara Wold and Ali Mortazavi (2009).Computation for ChIP-seq and RNA-seq studies. Nature methods, VOL.6 NO.11s

Salmon-Divon, M. and H. Dvinge, et al. (2010). "PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci." BMC Bioinformatics 11: 415.

Yong Zhang,Tao Liu et al. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biology, 9:R137

Young, M. D. and M. J. Wakefield, et al. (2010). "Gene ontology analysis for RNA-seq: accounting for selection bias." Genome Biol 11 (2): R14. Available online at: https://bioconductor.org/packages/release/bioc/html/goseq.html

Thorvaldsdottir, H. and J. T. Robinson, et al. (2013). "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration." Brief Bioinform 14 (2): 178-92. Available online at: https://www.broadinstitute.org/igv/

Young M D, Wakefield M J, Smyth G K, et al. (2010).Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biology, doi:10.1186/gb-2010-11-2-r14. (GOseq)

Kanehisa, M., M. Araki, et al. (2008). KEGG for linking genomes to life and the environment. Nucleic acids research.(KEGG)

IV Contact us

The pipeline and the ways to present the results are under sustaining updating. If you have any suggestions or questions about this report, contact us.

Tel: 0086-10-82837801 ext 849

Email: oversea_support@novogene.com