DOI: https://doi.org/10.53555/nnas.v5i12.576 **Publication URL: http://nnpub.org/index.php/AS/article/view/576**

CHIP-Seq PEAK CALLING AND DATA ANALYSIS IN WOODY PLANTS

Wang Lina[1]*

*[*1] State Key Laboratory of Tree Genetics and Breeding, Northeast Forestry University, 51 Hexing Road, Harbin ,150040, China*

Corresponding Author: -
Email: wanglina@nefu.edu.cn

Abstract: -

Chromatin immunoprecipitation (ChIP) followed by highthroughput sequencing (ChIP-Seq) provides an important tool to study genome-wide protein-DNA interactions to help understand gene regulation in the context of native chromatin. A complete ChIP-Seq workflow including peak calling, peak classification, transcription factor binding site overrepresentation and definition of new TFBS motifs. In this paper, ChIP-Seq peak calling and data analysis is described in detail.

1. INTRODUCTION

ChIP-Seq is a convenient technique to identify, characterize and map occupancy of specific DNA fragments with proteins against which specific antibodies exist or which can be epitope-tagged. All the resulting ChIP-DNA fragments are sequenced simultaneously and the resulting sequencing reads are mapped back to the reference genome. This will effectively give you an idea of where your target DNA interactions occur throughout the genome.

We used the MACS2 version 2.1.0 peak finding algorithm to identify regions of IP enrichment over background (Y. Zhang et al., 2008). A q value threshold of 0.05 f or enrichment was used for all datasets (Bailey et al.,1994). MEME and DREME wer e used to detect the sequence motifs, which were determined to detect the long and s hort consensus sequences (Bailey et al., 2011). After the motif detection, Tomtom soft ware was used to annotate the motifs based on the sequences' similarity (S.Gupta et al., 2007). The original sequence needs to be filtered to remove the connector, decont aminate, If there is a reference genome of the related species and the reference geno me is compared, the analysis can be carried out through the following process and th e analysis process is shown in Figure 1. The detailed analysis process is as follows：

(1) Use the cutadaptprogram to remove the street sequence from the original offline d ata (Martin M et al., 2011). (2) Use Trimmomatic program to get clean data by remo ving low-quality sequences (Bolger A M et al., 2014). (3) The Fastqcprogram was used to count the amount of cleandata and detect the proportion of Q20 and Q30 (Planet E et al., 2010). (4) Using the bowtie2 program to compare cleandata to the referenc e genome (Langmead B et al., 2012). (5) Calculate the capture efficiency of sequenci ng experiment, the coverage of ChIP region and the average sequencing depth. (6) Pe ak calling was performed on the genome using MACS2 (Zhang et al., 2008). (7) ME ME was used for motif detection of binding peaks (Timothy L et al., 2009). (8) Diff erences and notes of peak between samples to be carried out (Wang S et al., 2013).

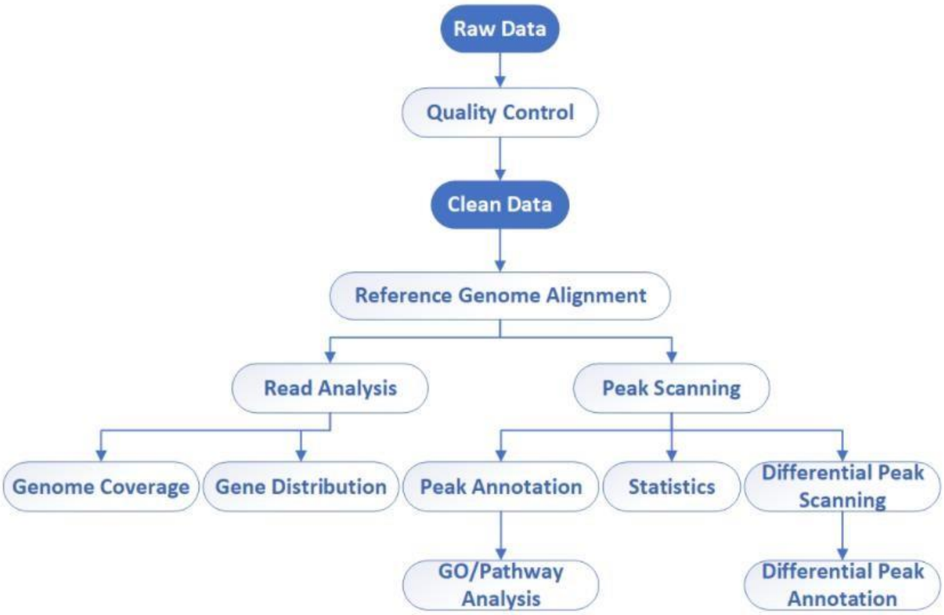

Fig.1 Bioinformatics workflow, Extensive quality control is performed at the conclusion of each step of the process.

2. Results

2.1 Quality Control of Fast QC

To check the quality of original reads after sequencing, a common tool is used FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), through FastQC tools to remove joint and low quality of the clean data. It can be seen from figure 2 that the theoretical values of GC range of samples are similar, and there is no obvious PCR amplification deviation (Fig. 2). From Fig. 2A, we can see the mass map of each base: the horizontal axis represents the position, the vertical axis quality, the red represents the median, and the yellow is 25%-75%. If the lower quartile of any position is less than 10 or the median is less than 25, the WARN is reported. If the lower quartile of any position is less than 5 or the median is less than 20 and the FAIL is reported. Figure 2B is the GC content graph: the horizontal axis is the average GC content of the sequence, and the vertical axis is the number of reads. The GC content represents the deviation of PCR amplification. Figure 2C is the mass diagram of the sequence: the base mass represents the probability of base recognition error. The higher the base mass value is, the more reliable the base recognition is, and the smaller the possibility of base error is. The horizontal axis is the base mass value, and the vertical axis is the number of reads. 20 or Q20 means that 1 out of 100 bases can be recognized incorrectly. Q30 is one out of 1,000 bases, and Q40 is one out of 10,000 bases. Figure 2D is the sequence length distribution diagram: the horizontal axis is the length of the sequence, and the vertical axis is the number of reads. This diagram can detect whether most of the sequence lengths obtained by sample sequencing are in line with expectations.

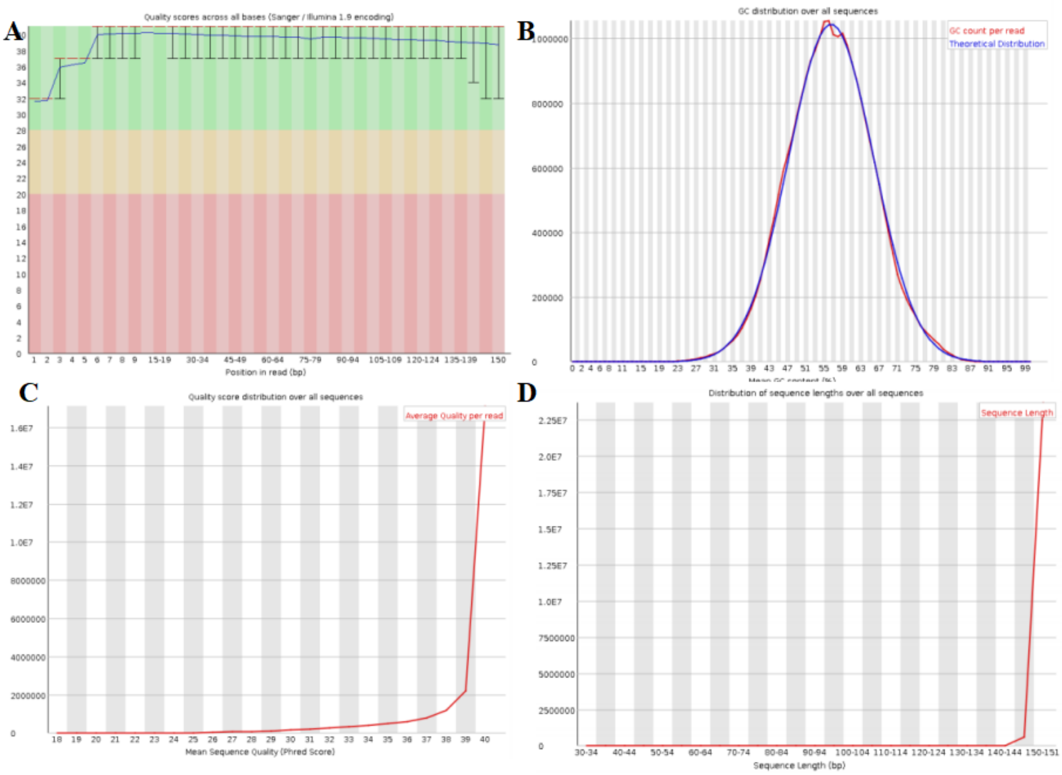

Fig.2 (A) The mass map of each base, different colors represent different percentage ranges. (B) GC content diagram, from which the deviation of PCR amplification can be seen. (C) The quality diagram of the sequence, indicating the probability of base recognition errors. (D) Sequence length distribution map.

2.2 Reference sequence alignment analysis

Reads Mapping and Peak Calling are the number of tests according to the direct standards of quality standards. Reads Mapping refers to the alignment of original data that have been processed offline to the reference genome. Reads mapping is mainly used to test the proportion of unique mapped Reads. The higher the specific response rate is, the better the data quality is. The following table is the reference sequence alignment analysis data: “All” refers to the total participating alignment sequence, that is, the filtered sequencing sequence (Clean Data). Mapped Reads refers to the sequence that can be aligned to the genome in the total participating alignment sequence. Mapping Rate refers to the genome alignment Rate, that is, the ratio of sequencing sequence alignment to genome proportion in the total sequence after sequencing filtration (Mapped Reads/All Reads). Unique Mapping refers to the sequence that can only have a Unique alignment position on the genome in the total alignment sequence.Unique Mapping Rate refers to the Unique Mapped Reads/ Mapped Reads ratio, that is, the proportion of Unique Mapped Reads in the sequence after being filtered from sequencing. Alignment analysis data are shown in table 1.

Table.1 Reference sequence alignment analysis data

Statistics	Result
All	19081830
UnMapped	165277
Mapped	18916553
MappedRate	0.991
UniqueMapped	14685129
UniqueMappedRate	0.77
RepeatMapped	4231424
AllBasE	954088153
UnMappedBaseMappedBase	8263850
MappedBase	945824303

During ChIP specific crosslinked DNA-protein complexes are enriched by using an antibody targeting your protein of interest. All the resulting ChIP-DNA fragments are sequenced simultaneously and the resulting sequencing reads are mapped back to the reference genome. This will effectively give you an idea of where your target protein-DNA interactions occur throughout the entire genome. Genome Inspector is a tool to correlate peaks with genomic regions, such as promoters, primary transcripts or microRNAs. In addition, different analyses can be compared to each other. For instance, replicates can be compared for their overlap or ChIP-Seq peaks can be compared to expressed genes from RNA-Seq experiment. Correlate data can be exported for futher analyses. To generate networks of regulated genes, Genome Inspector can be used to extract target genes. These target genes can be loaded into the Genomatix Pathway System (GePS) to obtain gene enrichments and generate signaling networks or visualize relevant cannonical pathways together with all available gene annotation. If there is serious chloroplast pollution, there will be obvious enrichment of special chromosomes.

2.3 Peak Calling

Peak Calling is a statistical method to detect characteristic peaks in chromosome regions as candidate protein binding sites,and it is the key to the subsequent analysis. After the comparison result is obtained, the peak is calibrated using MACS2 tool, and the enrichment position (Peak) of ChIP experiment relative to input can be compared.

The BAM file generated by reads sequence of genome is used as input file, and MACS2 software is used to call peak and Q-value cutoff is 0.05. Through call peak analysis, the specific location information of binding sites on genome can be obtained, and subsequent analysis of Motiff and annotation can be carried out. The detail information of peak are shown in table 2.

Table.2 Starting position of peak on chromosome

chr	start	end	length	abs_summit	pileup	p-value	fold_enrichment	q-value
6	81766003	81768027	2025	81766946	256	1.48E-54	39.98377	1.6363E-318
1	151987982	151988869	888	151988546	238	5.43E-24	47.13347	6.7644E-318
1	221393494	221394690	1197	221394228	212	1.43E-7	61.10661	1.814E-317
12	12538077	12539992	1916	12539564	231	1.87E-7	50.28303	2.3606E-317
13	73481995	73482687	693	73482236	217	3.3102E-5	57.85439	4.2249E-317

The specific meaning of the data in Table 2 is shown in Table 3, which includes the Chr, Start, End, Length, Abs_summit, Pileup, Fold_enrichment, Q-value, Name.

Table 3 Specific Significance of Various Parameters

Name	Meaning
Chr	Peak chromosome
Start	The starting position of the peak
End	Termination position of peak
Length	Width of peak area
Abs_summit	Absolute position of peak
Pileup	Height of Peak Position
P-value	P-value of the peak
Fold_enrichment	Enrichment of the peak
Q-value	Q-value of peak
Name	Name of peak

2.4 Statistical analysis of annotation results

The distribution of binding sites in the genome of each sample was analyzed, as shown in Figure 3. Through analysis, we found that the distribution of binding sites in genome, and through the distribution of results, we can understand the possible functions of binding sites. Promoter represents the promoter region, UTR represents 3'UTR and 5'UTR, 1st Intron represents the first exon and intron of the gene, Other Exon and Other Intron represent the exon and intron of the gene except the first, Downstream represents the downstream region of the gene termination location, Distal Intergenic represents the intergenic region besides the promoter and downstream region of the gene.

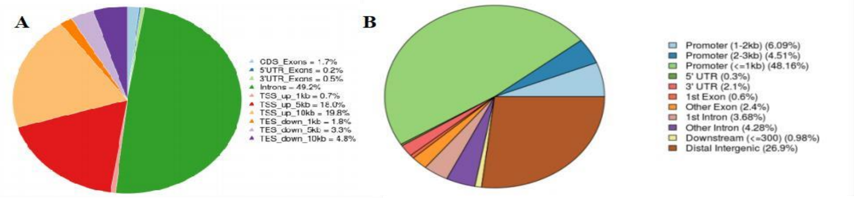

Fig 3. Reads classification pie chart, from which we can see the proportion of each type of Reads in this sequencing. Promoter-TSS represents the proportion of distribution in promoter zone. TTS representation distribution in TTS ratio. Distal Intergenic is the proportion distributed in the distal region. Exon is the proportion distributed in exon.

2.5 Analysis of gene function and signaling pathway

Based on the hierarchical structure of GO, all the mutual regulation and subordination relations between GO were arranged into a database. By constructing a functional relationship network, the functional groups affected by the experiment and the internal subordination relations of significant functions were summarized. The significance of go-term (p-value<0.01) in go-analysis of annotated genes was used for functional regulation Analysis, and a functional regulation network was constructed.

In order to accurately classify the functions of genes, GO analysis assigns annotated genes to different functional classifications. GO classification can describe the functions of genes in all aspects. GO can be divided into three main groups: biological process (BP), molecular function (MF) and cellular component (CC). The annotated genes were annotated by GO from BP, MF and CC in the database, and the genes involved in all GO were obtained. Fisher test was used to calculate the P-Value of each GO, and the significance GO of annotated gene enrichment was screened out.

From the perspective of biological significance, Biological Process is more close to the phenotype and can often describe the actual situation of samples. Cellular components are often used to describe cellular localization of genes, which makes sense when looking at specific subcellular locations. As a description of the analytical function, the actual mode of action of proteins is often described. When we pay attention to the changes in the way proteins act in the whole biological event, we can consider this point of view. The results of these three parts are relatively close to each other in terms of analysis strategy and result presentation when analyzing data, so most analysis chooses biological process as the presentation form.

GO analysis of annotated genes: GOID is the ID information of GO-Term in the database, GOTerm represents the GO details enriched by genes, DifGene refers to the number of genes that are annotated into the GO entry in all of the annotated genes. AllDifGene refers to the number of genes with GO database annotations in all of the annotated genes. GenelnGO refers to the number of gene annotations recorded in the database in this go-term, AllGene is the number of genes annotated with the GO database, P-value refers to the significance of enrichment in GO of the genes obtained by Fisher’s exact test. FDR is the result corrected by BH algorithm for P-value.

2.6 Metabolic pathway analysis

Pathway-Analysis, based on the gene annotation database, detects the significant pathway of the annotated genes. Therefore, the key to Pathway-Analysis lies in the fact that it has a complete database and relatively complete pathway annotations. Pathway annotation was conducted based on the KEGG database for all pathway terms involved in the annotated genes. Fisher test was used to calculate the significance level (p-value) of pathway, and the significant pathway terms enriched by the annotated genes were screened out. From the perspective of biological significance. Pathway can directly reflect the effect of genes on phenotypes, and the relationship between significantly enriched signaling pathways and phenotypes can reach the maximum and Path enrichment analysis of annotated genes was seen in Fig 4.

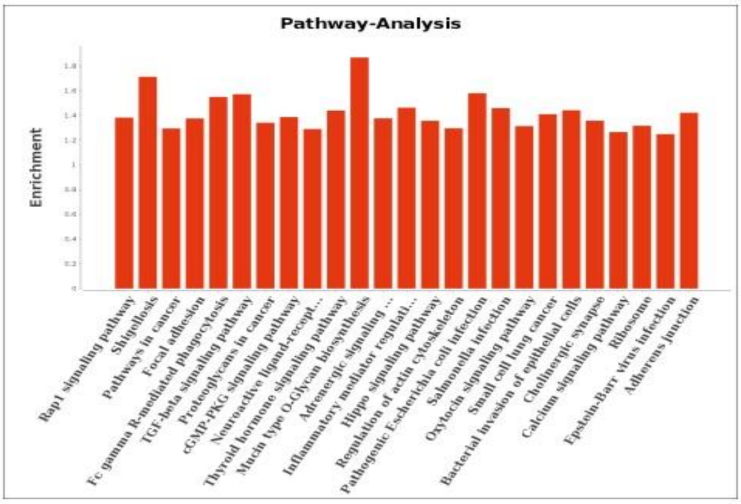

Fig 4. Path enrichment analysis of annotated genes. The aim of this map was to select the first 20 items from the prominent Pathway-Term for display. The coordinate axis Y is Enrichment, the coordinate axis X is Pathway-Term, and the red bar graph represents significance items.

2.7 Motif analysis

Motif is a specific base sequence with high affinity to some proteins. Motif analysis is carried out by using FindMotifsGenome tool of HOMER. Reliable peaks are screened according to peak information obtained from peak calibration and annotation. Sequences near the peak are analyzed by using MEME tool. The input file was the peak file and the genome fasta file. DNA sequences were extracted from the peak file and compared with the motif database to obtain motif, Motif analysis results was seen in Fig 5.

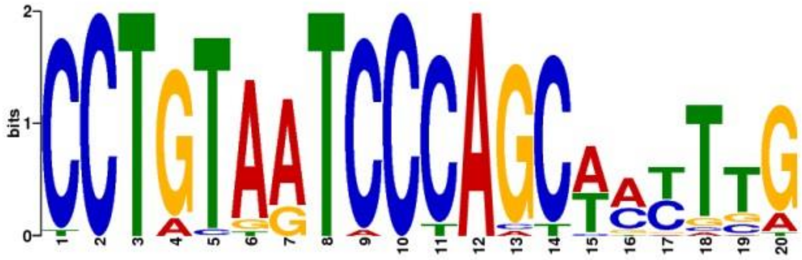

Fig.5 Motif analysis results in Chip-Seq experiment. Motif is the transcription factor corresponding to the predicted Motif. Letter size is proportional to the frequency of this nucleotide in Motif, and the P-value of the predicted Motif is predicted. The smaller the Pvalue is, the more reliable the predicted factor will be.

Funding

This work was supported by the The National Natural Science Foundation of China [3180030530].

References

[1].Bolger A M , Lohse M , Usadel B . Trimmomatic: a flexible trimmer for Illumin a sequence data[J]. Bioinformatics, 2014, 30 (15):2114-2120.

[2].Gupta S , Stamatoyannopoulos J A , Bailey T L , et al. Quantifying similarity be tween motifs[J]. Genome Biology, 2007, 8(2):R24-R24.

[3].J. Ye , Coulouris G , Zaretskaya I , et al. Primer-BLAST: A tool to design target -specific primers for polymerase chain reaction[J]. BMC Bioinformatics, 2012, 13(1):1 34.

[4].Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Meth ods. 2012, 9:357-359.

[5].Martin R C , Vining K , Dombrowski J E . Genome-wide (ChIP-seq) identificatio n of target genes regulated by BdbZIP10 during paraquat-induced oxidative stress[J]. Bmc Plant Biology, 2018, 18(1):58.

[6].Martin M . Cutadapt removes adapter sequences from high-throughput sequencing read[J]. Embnet Journal, 2011, 17(1). 2114–2120.

[7].Planet E , Attolini S O , Reina O , et al. htSeqTools: high-throughput sequencing quality control, processing and visualization in R[J]. Bioinformatics, 2012, 28(4):589590.

[8].T.L. Bailey, DREME: motif discovery in transcription factor ChIP-Seq data,Bioinfo rmatics 27 (2011) 1653e1659.

[9].Timothy L. Bailey, Mikael Bodén, Fabian A. Buske, Martin Frith, Charles E. Gra nt, Luca Clementi, Jingyuan Ren, Wilfred W. Li, William S. Noble, "MEME SUITE: tools for motif discovery and searching", Nucleic Acids Research, 37:W202-W208, 2009.

[10]. T.L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to di scover motifs in biopolymers, Proc. Int. Conf. Intell. Syst. Mol. Biol. 2 (1994) 28-36.

[11]. Wang S, Sun H, Ma J, et al. Target analysis by integration of transcriptome and ChIP-Seq data with BETA[J]. Nature Protocols, 2013, 8(12):2502-25.

[12]. Zhang Y , Liu T , Meyer C A, et al. Model-based analysis of chip-seq (MACS) [J]. Genome biology, 2008, 9(9):R137.