1 Library Construction and Sequencing

After DAP DNA was extracted, the enriched DNAs were fragmented into short fragments by ultrasonic. Next, the DNA fragments were end repaired, 3’A added, and ligated to Illumina sequencing adapters. DNA fragments with proper size (usually 100-300bp, including adapter sequence) were selected for PCR amplified. Finally we got qualified library for sequencing. PCR amplified and sequenced using Illumina HiSeq^{TM} 4000 (or other platforms) by Gene Denovo Biotechnology Co. (Guangzhou, China)

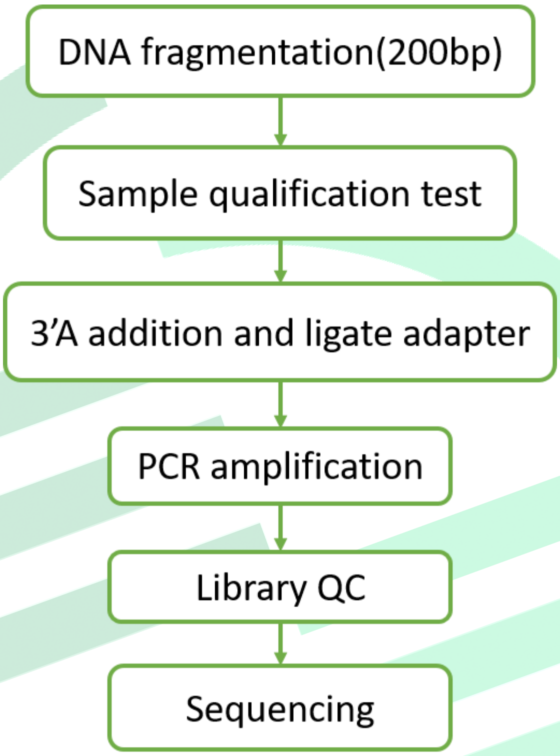

Fig 1 Flow chart of DAP-seq Library Construction

2 Flow Chart of Bioinformatics Analysis

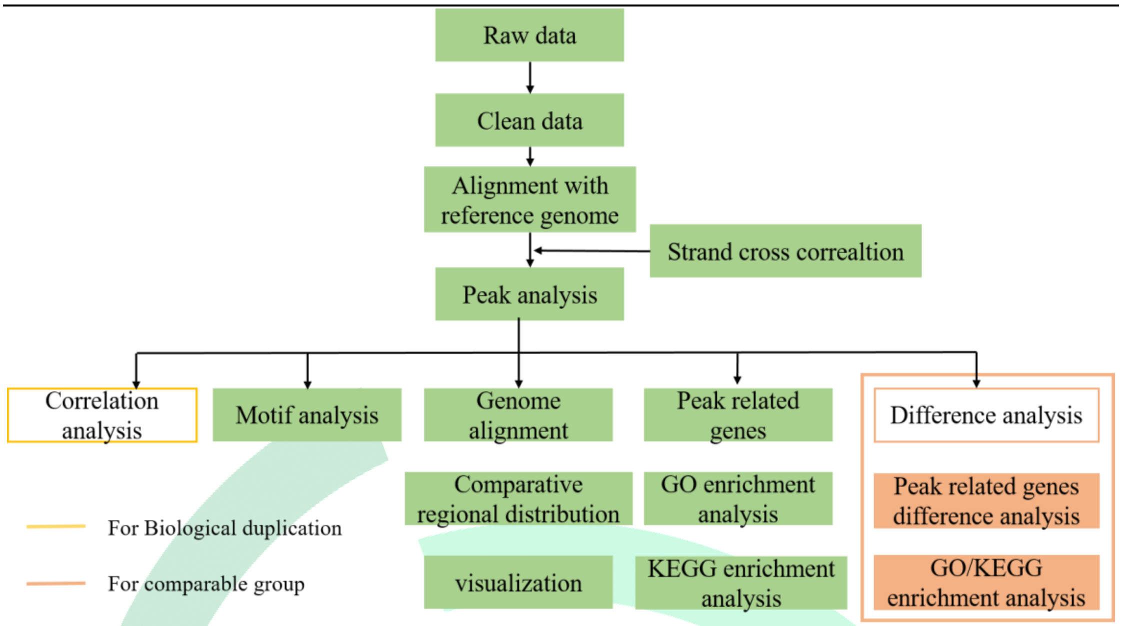

Fig 2 Flow chart of DAP-seq bioinformatic analysis

2.1 Clean Reads Filtering

Reads obtained from the sequencing machines included raw reads containing adapters or low quality bases which would affect the following assembly and analysis. Raw reads would be processed to get high quality clean reads according to three stringent filtering standards:

i Removing reads containing adapters;

ii Removing reads containing more than 10% of unknown nucleotides (N);

iii Removing low quality reads containing more than 40% of low quality (Q-value≤10) bases.

2.2 Reads Alignment

Bowtie2[1] (version: 2.2.5) was used to align the clean reads from each sample against the reference genome. All reads from transcriptional initiation site (TSS) to transcriptional termination site (TES) interval and upstream and downstream 2k interval were counted by deepTools[2] (version: 3.2.0) software.

2.3 Strand Cross Correlation

Strand cross-correlation of tag density providing a quick assessment of IP-seq data set quality and binding characteristics. It is based on the fact that a high quality IP-set experiment produces significant clustering of enriched DNA sequence tags at locations bound by the protein of interest, and that the sequence tag density accumulates on forward and reverse strands centered on the binding site. Phantompeakqualtools[3] R package was used to calculate the correlation between forward and reverse strands, which can reflect whether the chromatin immunoprecipitation effect is optimal.

2.4 Peak Calling

MACS2 [4] (version: 2.1.2) software was designed to identify read-enriched regions from DAP-seq data. Dynamic Poisson Distribution was used to calculated p-value of the specific region based on the unique mapped reads. The region would be defined as a peak when q-value<0.05.

2.5 Peak Related Genes Annotation

Peak related genes were annotated by the ChIPseeker [5] R package. According to the genomic location information and gene annotation information of peak, peak related genes can be confirmed.

Besides, the distribution of peak on different genome regions, such as intergenic, introns, downstream, upstream and exons was performed.

2.6 Peak Related Genes GO Enrichment Analysis

Gene Ontology (GO) is an international standardized gene functional classification system which offers a dynamic-updated controlled vocabulary and a strictly defined concept to comprehensively describe properties of genes and their products in any organism. GO has three ontologies: molecular function, cellular component and biological process. The basic unit of GO is GO-term. Each GOterm belongs to a type of ontology. GO enrichment analysis provides all GO terms that significantly enriched in peak related genes comparing to the genome background, and filter the peak related genes that correspond to biological functions. Firstly all peak related genes were mapped to GO terms in the Gene Ontology database (http://www.geneontology.org/), gene numbers were calculated for every term, significantly enriched GO terms in peak related genes comparing to the genome background were defined by hypergeometric test. The calculating formula of P-value is:

$$P = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

Here N is the number of all genes with GO annotation; n is the number of peak related genes in N; M is the number of all genes that are annotated to the certain GO terms; m is the number of peak related genes in M. The calculated p-value were gone through Bonferroni Correction, taking corrected-pvalue≤0.05 as a threshold. GO terms meeting this condition were defined as significantly enriched GO terms in peak related genes. This analysis was able to recognize the main biological that peak related genes exercise.

2.7 Peak Related Genes Pathway Enrichment Analysis

Genes usually interact with each other to play roles in certain biological functions. Pathway- based analysis helps to further understand genes biological functions. KEGG is the major public pathway-related database [6] (Release 87.0). Pathway enrichment analysis identified significantly enriched metabolic pathways or signal transduction pathways in peak related genes comparing with the whole genome background. The calculating formula is the same as that in GO analysis.

$$P = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

Here N is the number of all transcripts that with KEGG annotation, n is the number of peak related genes in N, M is the number of all transcripts annotated to specific pathways, and m is number of peak related genes in M. The calculated p-value was gone through FDR Correction, taking FDR ⩽ 0.05 as a threshold. Pathways meeting this condition were defined as significantly enriched pathways in peak related genes.

2.8 Motif Analysis

The DNA binding site for specific transcription factors or histone modifications was not random, while they show conserved DNA sequence pattern. MEME suit (http://meme-suite.org/) was used to detect the motifs. MEME (http://meme-suite.org/tools/meme) and DREME (http://memesuite.org/tools/dreme) were used to detect the sequence motif, which determined to detect long and short consensus sequence.

2.9 Peak Combination and Multi-sample Clustering

The DiffBind [7] (version 2.8) software was used to combine peaks of every group, and obtained the union of peaks among groups. Peak abundance was shown by calculating rpm-values of each sample.

2.10 PCA analysis

For comparison between samples, the unsupervised dimensionality reduction method principal component analysis (PCA) was applied in all samples using R package models (http://www.r-project.org/). PCA is a statistical procedure that converts thousands of correlated variables into a set of values of linearly uncorrelated variables called principal components.

2.11 Common and Specific Peak Related Genes Analysis between Groups

In different group, peaks with overlap were defined as the common peak. The genes closest to each peak were defined as peak-related genes. Peaks distant from peak-related gene (which locates farther upstream than 2k or downstream 300bp) were removed, and the remaining peak related genes were used for subsequent GO and KO enrichment analysis.

2.12 Peak Difference Analysis between Groups

the DiffBind [7] (version 2.8) software was used to analyses peak differences between groups. Diffbind was designed to work with multiple peak sets simultaneously, representing different DAP experiments (antibodies, transcription factor and/or histone marks, experimental conditions, replicates) as well as managing the results of multiple peak callers. We identified peaks with log 2|M|≥1 and FDR≤0.05 in a comparison as significant differential peaks. Then genes associated with different peaks were annotated, and enrichment analysis of GO functions and KEGG pathways were identified.

3 Reference

[1] Langmead B, Salzberg S L. Fast gapped-read alignment with Bowtie 2[J]. Nature methods, 2012, 9(4): 357

[2] Fidel Ramírez, Ryan D P , Björn Grüning, et al. Deeptools2: A next generation web server for deep-sequencing data analysis[J]. Nucleic Acids Research, 2016, 44(Web Server issue):gkw257.

[3] Landt SG1, Marinov GK, Kundaje A et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012 Sep;22(9):1813-31. doi: 10.1101/gr.136184.111.

[4] Zhang Y, Liu T, Meyer CA, et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biology,2008, 9: R137

[5]Yu G, Wang L G, He Q Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization[J]. Bioinformatics, 2015, 31(14): 2382-2383

[6] Kanehisa, M., M. Araki, et al. KEGG for linking genomes to life and the environment. Nucleic Acids Res. 2008.36 (Database issue): D480-4.

[7] Stark R, Brown G. DiffBind: differential binding analysis of ChIP-Seq peak data [J]. R package version, 2011, 100: 4.3.