Novogene | Advancing Genomics, Improving Life

From Immunoprecipitation to Data Analysis: A Comprehensive summary of ChIP-seq

ChIP-seq Workflow

From Immunoprecipitation to Bioinformatic Analysis

1 Chromatin immunoprecipitation experiment

ChIP assays begin with covalent stabilization of the protein–DNA complexes. Many protein–DNA interactions are transient and involve multiprotein complexes to orchestrate biological functions. As there is constant movement of proteins and DNA, ChIP captures a snapshot of the protein–DNA complexes that exist at a specific time. In vivo crosslinking covalently stabilizes protein–DNA complexes.

Researchers can often use a combination of crosslinkers to trap interacting proteins and DNA. These crosslinkers permeate directly into intact cells and effectively lock protein–DNA complexes together, allowing even transient complexes to be trapped and stabilized for analysis. These crosslinkers must be reversible to be used for ChIP.

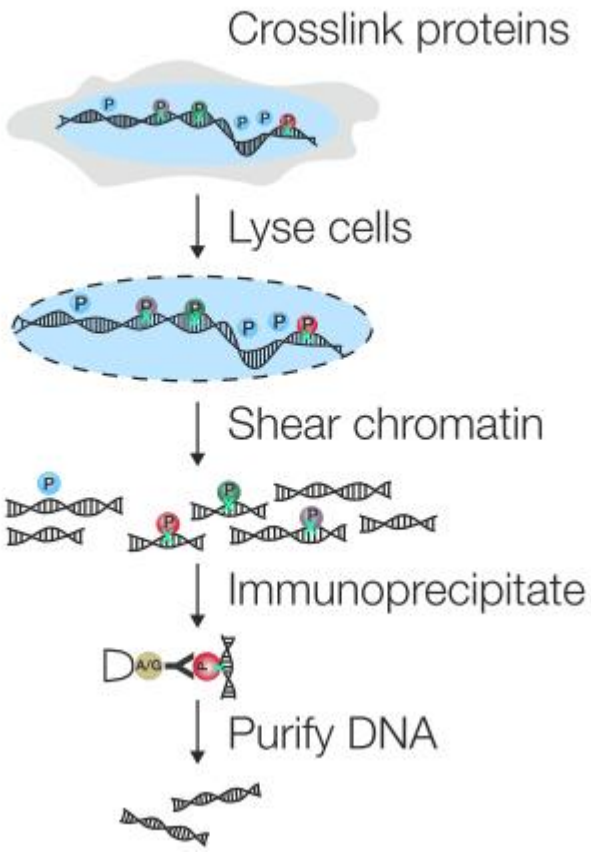

Figure 1. Whole workflow of chromatin immunoprecipitation

1.1 Choices on Antibody

Ideally speaking, Ideally, the research target region should have an antibody that has already worked in ChIP or other IP experiments. All types of antibodies (Monoclonal, oligoclonal, and polyclonal) can work in ChIP. The first key feature of successful ChIP is that specific epitope of interest being exposed during treatment. The second key feature is the specificity of the antibody. Non-specific antibodies will mislead the recognition of region of interests. For example, for researching H3K9me2 protein, you would need very specific antibody that only pulling down only the dimethyl me2 mark and not the monomethyl me1 or trimethyl me3.

1.2 Experiment requirements:

ChIP experiments must have controls. Both comparing same cell lines or different cell lines require a "no-antibody control" (known as Input) for each IP samples. There are other controls to consider to determine if your ChIP experiment worked.
Input role (Essential): Input is referred to fragments of samples without any IP experiments. Theoretically speaking, Input represents random DNA fragments after shearing. A portion of the same IP sample is taken as Input before immunoprecipitation. It can verify the effects of IP assay throughout the experiment. In other words, detecting background "noises" of DNA fragments.
Mock control (Non-essential): Mock control is done through applying non-specific antibodies. For example, IgG. The role of mock sample is to exclude any false positive result and verify the specification of the antibody used.
For each comparison group, we recommend to have >=3 replicates in both control and treatment group. Here is a example format of IP-seq experiment design:

	Group 1 (Control Group, no treatment)	Group 2 (Treatment 1)
IP1	1 sample + 2 replicates	1 sample + 2 replicates
IP2	1 sample + 2 replicates	1 sample + 2 replicates
Input （required）	1 sample	1 sample
IgG Group (optional)	1-2 sample for whole project	

For a standard protocol, approximately 2 x 10^6 cells per immunoprecipitation is

recommended. However, lower cell number to 50000 cells is reported to be possible. Though Novogene does not perform IP experiments, we selected some reference or kits below to help you better understand for IP:
ActiveMotif_kits, MilliporeSigma™ Chromatin Immunoprecipitation (ChIP) Assay Kit, EZ-Magna ChIP™ A/G Chromatin Immunoprecipitation Kit

1.3 Crosslink

ChIP preserves the protein–DNA complexes that exist at a specific time. In vivo crosslinking is traditionally done using a formaldehyde solution. For higher-order interactions, longer crosslinkers from other companies can trap larger protein complexes with more complicated quaternary structures. For some histones, native ChIP can be performed because some protein–DNA interactions (e.g., H3–DNA) are inherently tight, making extra crosslinking unnecessary. Before performing crosslinking, please check your targeted interactions to determine the solution for crosslinking. Note that the duration of crosslinking is important. Overtime crosslinking may hinder downstream lysing and shearing the chromatin to ideal sizes.

1.4 Cell lysis

Cell membranes are dissolved with detergent-based lysis solutions to liberate cellular components, and crosslinked protein–DNA complexes are solubilized. We recommend you to wash off cytosolic proteins as this step would help reduce background signal and increase sensitivity of region detection. The presence of detergents or salts will not affect the protein–DNA complexes, because of the covalent crosslinking within complexes. Protease and phosphatase inhibitors are necessary to keep protein–DNA complexes intact. We suggest you to check status of cell lysing under a microscope constantly, ensuring success rate of experiment. We also suggest you to alter the duration of lysing according to cell types. For cells that find it difficult to detach their nucleus from other components, the duration of lysing can be extended.

1.5 Chromatin Shearing

DNA fragmentation is usually achieved either mechanically by sonication or enzymatically by digestion with MNase. Mechanical fragmentation can generate more random DNA-protein complexes but is usually hard to maneuver over time. While using enzymes is becoming more popular among researchers, keep in mind that such enzymes often have a higher affinity for inter-nucleosome regions and lead to less random results. Note that starting from here, ChIP can be stopped. After shearing/digestion of the chromatin, the complexes can be stored at –80 °C. Ideal chromatin fragment sizes should range from 50 to 700 bp. Novogene, based on prior experience, recommends generating reads ranging from 100 to 500 bp.

1.6 Immunoprecipitation (IP)

This step selectively enriches the protein-DNA complex of interest and eliminates all other unrelated materials. ChIP-validated antibodies are used to immunoprecipitate and isolate the target from other nuclear components. The antibody–protein–DNA complex is affinity purified using an antibody-binding resin such as immobilized protein A, protein G, or both combined. Prior to lysate by first incubating the lysate with the beads for several hours before adding the antibody. The volume of beads used in each ChIP sample can also influence the background. After extended incubation, the bead-antibody-protein-DNA complex must be extensively washed and often purified sequentially with both low- and high-salt buffers. Please remember to preserve 5% of the sample before adding beads. This portion should serve as your Input control.

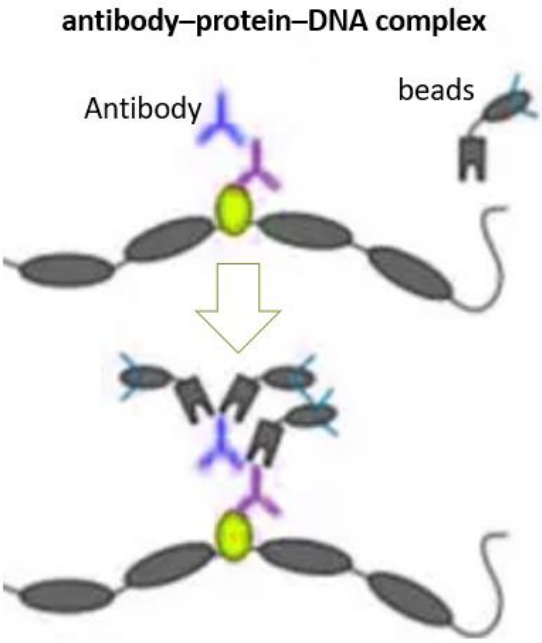

Figure 2. Formation of Antibody-Protein-DNA complex with beads

1.7 Reversal of crosslinking, and DNA clean-up

After IP, the enriched targeted DNA still crosslinks with protein. Therefore, the reversal of such complexes is needed. Reversal is usually executed with extensive heat incubations or digestion of the protein component using Proteinase K. Proteinase K eliminates nucleases from the purified DNA, preventing degradation. Further treatment with RNase A is also recommended to obtain a purer DNA sample. Additional and final purification of the DNA from any remaining proteins should be performed based on phenol-chloroform extraction or spin columns exclusively for DNA purification.

Notice:

The contents above constitute one summary of manuals from different sources provided as suggestions. Novogene only makes recommendations for readers and does not perform IP experiments or take liability for any outsourcing of IP.

2 DNA sample test

The DNA samples are immunoprecipitated and purified on the end of client. Novogene detection of these IPed DNA samples mainly includes two steps:

(1) Qubit accurately quantifies DNA concentration, based on which the total amount is calculated.

(2) Agarose gel electrophoresis analyses of DNA fragments size distribution and detects any contamination.

3 Library Preparation for Sequencing

After the DNA samples have undergone quality control:

(1) The DNA fragments are repaired, dA-tailed.

(2) The DNA fragments with an A tail are ligated to sequencing adaptors.

(3) The final DNA library is obtained by size selection and PCR amplification.

ChIP DNA has two kinds of lib preparation workflows. One is based on regular input and applies our internal kit, provided with a sufficient amount of DNA passing QC. The other is called low input, which applies a kit called NEBNext® Ultra™ II DNA Library Prep Kit for Illumina®.

Both workflows of library construction are as following graph.

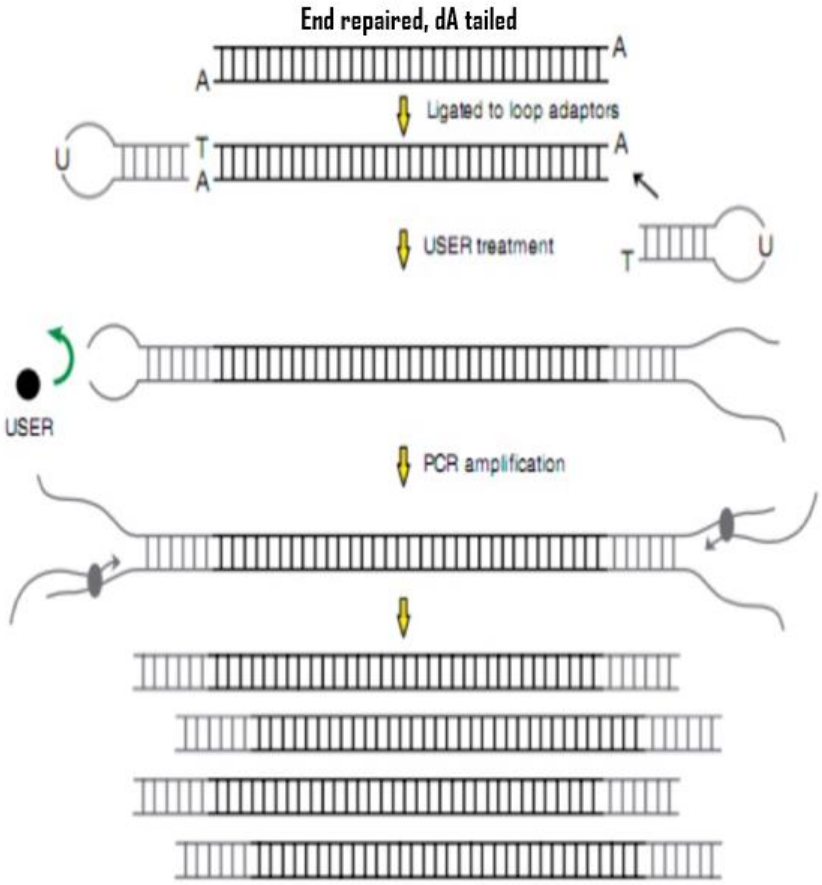

Figure 3. Library construction workflow

4 Quality control of Library

After the construction of the library, the initial quantification was done with Qubit 2.0, and the library was diluted to 1 ng/l. Then the insertion size of the library is detected with NGS3K. If the results meet expectations, the accurate concentration of the library is quantified by Q-PCR (library effective concentration > 2nM) to ensure the accurate molar amount that will be pooled for sequencing.

5 Sequencing

After library quality control, sequencing is performed for different libraries according to the concentration and the demand of data amount on Illumina NovaSeq platform. The basic principle of sequencing is sequencing by synthesis. Four kinds of fluorescent labeled dNTP, DNA polymerase and adapter primers are added to the flow cell for amplification. When each sequence cluster extends the complementary chain, each fluorescent labeled dNTP releases the corresponding fluorescence. The sequencer

captures the fluorescent signal and converts the light signal into the sequencing peak through computer software to obtain the sequence information of the fragment to be detected.

6 Analysis

The analysis of ChIP-seq follows the graph:

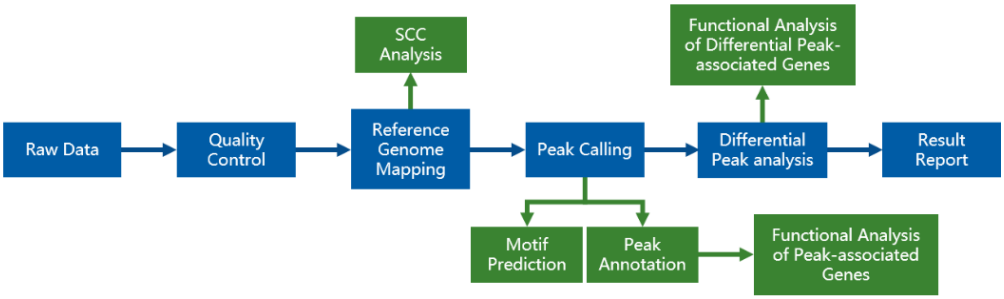

Figure 4. Bioinformatics Analysis Pipeline

6.1 Data QC

The original raw data from Illumina platform is transformed to Sequenced Reads, known as Raw Data or RAW Reads, by base calling of CASAVA. Raw data are recorded in a FASTQ file, which contains sequence information (reads) and corresponding sequencing quality information. Trimming software FastQC would help us get clean data by removing adapter-appended reads and low-quality reads.

6.2 Mapping

We choose proper software with well-tested parameters according to different genome characters to do the genome mapping analysis for filtered reads. Considering the small fragment size of ChIP-Seq and the percentage of unique sequences in the total sequence is most important for Chip-Seq analysis, BWA is considered proper mapping software for DNA based on the fragment size of ChIP-seq. Mapping the reads to the reference genome using BWA gives more accurate results. Duplicates were labeled using SAMBLAST and mapping quality value was calculated as MAPQ. Proper quality value was chosen as the only threshold for mapping.

6.3 Distribution of the reads mapped to the gene

Since the binding sites of transcription factor and histone protein are important for gene regulation, analysis of relative mapping position distribution can help us predict the protein function. We divide each gene and its 2kb upstream and 2kb downstream into 100 equal parts and calculate mapped reads in each part and the

percentage ratio of the reads in each part to total reads as reads density. The mapped reads can be directly visualized through software IGV.

6.4 Fragment size prediction

After mapping, the selection of fragments is vital for IP-seq analysis. For a specific binding site, there is a significant reads enrichment in the binding site we use MACS2 software to predict the frag sizes of IP experiment. MACS 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.

6.5 Summary of strand cross correlation

By testing the SCC of IP and input data, we can obtain the correlation coefficient between two strands and test the effect of IP experiment. The SCC curve of successful IP in the same experimental group (including IP and Input) with 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 and RSC value should be no less than 0.8.

6.6 Peak Calling

After mapped reads are filtered, we need to use these reads to find the potential binding sites between DNA and protein. Since IP has selectively gathered clusters of DNA fragments, reads are supposed to clustered together on these binding sites of genome. In other words, forming peaks on the genome. By making use of MACS2 software (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 fold enrichment value here can be called “signal value”, which is the digital display of the peak signal during peak calling. Larger values indicate more reads are enriched around certain peaks.

6.7 Motif analysis

The bindings of proteins such as transcription factors, histones and others are not random. Such bindings usually have sequence preferences and these conservative sequences are called Motif. Motif analysis can detect protein specific binding sites and obtain annotated motif and motif sequence information. First, we apply MEME and DREME software to detect significant motif sequences around the peak. Then, by using Tomtom software, we can annotate the motif by aligning it to the annotated Motif database.

6.8 Peak Annotation

Peak-TSS distance distribution can help us predict protein binding sites. We can estimate IP effect according to protein binding sites. We can also predict protein regulatory mechanism or function according to its protein binding character. TSS (transcription start site) of every peak related gene is detected by software PeakAnnotator. Next, we calculate peak numbers according to peak-TSS distance, and analyze peak-TSS distance distribution.

6.9 GO/Pathway annotation

Any gene that overlaps with a peak is considered to be a peak-related gene. Gene Ontology (GO, http://www.geneontology.org/) is an international standard classification system for gene function attributes. GO aims to describe gene and protein function for all species. GO covers three domains: Molecular Function, Biological Process, and Cellular Component. Besides GO enrichment, pathway enrichment can help identify the main biochemical, metabolic, and signaling pathways of a list of genes. KEGG (Kyoto Encyclopedia of Genes and Genomes) is a database that can analyze gene function. It is a powerful tool to study organism metabolism and do network analysis. KEGG includes the metabolism of carbohydrates, nucleotides, amino acids, and biodegradation. KEGG enrichment can yield a comprehensive annotation of the enzyme in each catalytic reaction, including the amino acid sequence, PDB link, etc.

Using peak-related genes as input, we can generate the enrichment of GO terms, representing peak's function as annotated in both the GO and pathway databases.

6.10 Differential Peak Analysis

Differential Peak Analysis compares how peaks differ from treatment and control in numbers and signals. Only when the number of groups is greater than two can the difference analysis be performed between groups. During the Analysis, RPM value (the ratio of 1 million reads that enriched the peak in a single sample) of the peak in different samples is used to do clustering analysis to determine the enrichment pattern of the same peak in different samples or the enrichment variety of different peaks in the same sample. At the same time, enrichment comparisons between IP and Input within groups can show the peak enrichment in the IP experiment. FoldEnrich is used for the differential analysis of peaks in different experimental groups (the ratio of RPM value in group A to group B). Differential binding sites are analyzed by finding a differential peak when the ratio of FoldEnrich is >=2. The differential binding sites-related genes are generated to run follow-up annotation and enrichment analysis.

We also present the differential expression venn diagram and boxplot, which directly show how enriched peaks differ among groups and samples.

Customization

Besides standardized results from pipeline, we can offer various results based on your needs. Typical ones include ChIP-seq analysis with spike-in controls and TSS heatmap. Related results are shown below. Any other customization can be evaluated by the professional team from Novogene and we shall deliver results that fit your research goal.

You can always make contact with Novogene here: inquiry_us@novogene.com

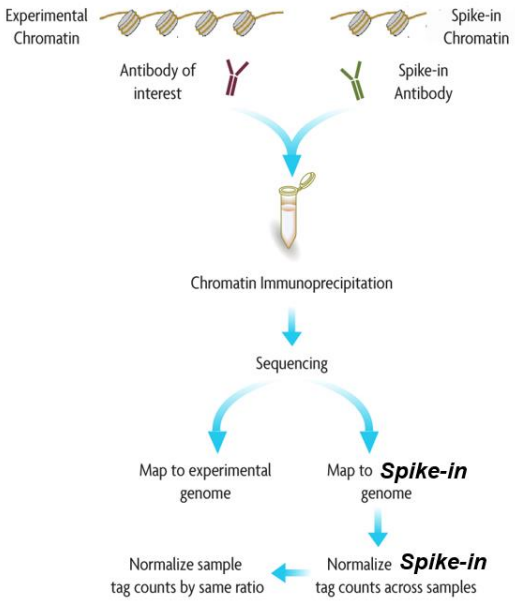

Figure 5. ChIP-seq Pipeline involves with Spike-in Control

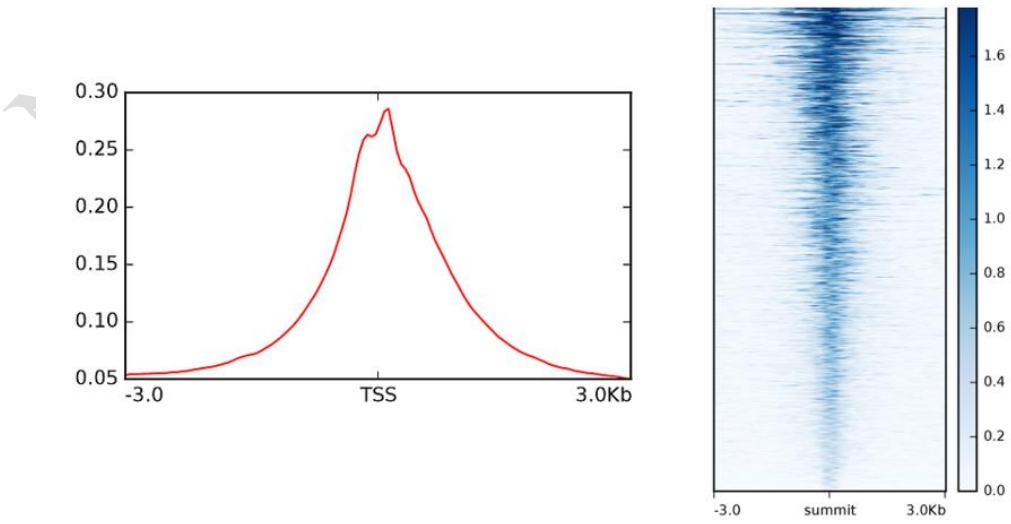

Figure 6. TSS heatmap and profile graphs based on customization

Notice:
All recommendations concerning methods and kits aim at research purposes. All of the recommendations in the context cannot be used for diagnostic purposes. Novogene does not earn profits from or perform advertising for any of the companies listed in the brochure. Novogene does not assume liability for the practical usage of these recommended products.

Reference

Active M. (2023). *How Does ChIP-Seq Spike-In Work?* ChIP-Seq Spike-In Normalization. Available online at: https://www.activemotif.com/catalog/1091/chip-normalization

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. L., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings. *International Conference on Intelligent Systems for Molecular Biology*, 2, 28–36.

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

Brind'Amour J et al. (2015) An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations. Nat Commun 6:6033.

Browne JA et al. (2014) An Optimized protocol for isolating primary epithelial cell chromatin for ChIP. PLoS One 9(6):e100099.

Carey MF et al. (2009) Chromatin immunoprecipitation (ChIP). Cold Spring Harb Protoc 4(9).Brind'Amour J et al. (2015) An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations. Nat Commun 6:6033.

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

Flanagin S et al. (2008) Microplate-based chromatin immunoprecipitation method, Matrix-ChIP: a platform to study signaling of complex genomic events. Nucleic Acids Res 36(3):e17.

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., & Noble, W. S. (2007). Quantifying similarity between motifs. *Genome biology*, 8(2), R24. https://doi.org/10.1186/gb-2007-8-2-r24

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.

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

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/

Sadeh S et al. (2016) Elucidating combinatorial chromatin states at single-nucleosome resolution. Mol Cell 63:1080–1088.

Tehranchi AK et al. (2016) Pooled ChIP-seq links variation in transcription factor binding to complex disease risk. Cell 165(3):730–741.

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

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

Thermo F. S. (2016) *A step-by-step guide to successful chromatin immunoprecipitation (ChIP)*

assays. Available online at: https://tools.thermofisher.com/content/sfs/brochures/Step-by-Step-Guide-to-Successful-ChIP-Assays.pdf

Thorvaldsdóttir, 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/

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

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)