Kai Fu

Analysis pipeline for ChIP-Seq

1. Reads mapping

Use any mapping software (i.e., Bowtie/Bowtie2/BWA) to map the original sequencing file such as .fastq files to its corresponding genome.
In this step, you could obtained the ratio of mapping. Usually, a ChIP-Seq will have mappability larger than 50 percent, which means more than half of reads should be able to be mapped.

2. Peak calling

For both histone ChIP-Seq and transcription factor ChIP-Seq, we could use MACS2 to perform the whole process of peak calling. MACS2 will generate the following results:

2.1 Ratio of non-redundant reads

You could find this in the xls output from the peak results

2.2 Cross-correlation between plus strand reads and minus strand reads

Cross-correlation analysis is useful for figuring out the fragment size of the factors of ChIP-Seq. This could also be served as a quality control of the ChIP-Seq.

2.3 Peak enrichment for plus strand reads and minus strand reads

You could find this in the peak model output. The result will give you a clear view about the enrichment of the antibody.

2.4. Peak information

In the output, you could find the peak in bed format. The first column is the chromosome, second and third column is the start and end of peaks.

For more information, MACS2 has a very comprehensive manual at:
https://github.com/taoliu/MACS

3. IGV track visualization

To better explore the data, you could load the bigwig files generated from MACS2 in the IGV. The manual of IGV could be found: https://www.broadinstitute.org/igv/UserGuide

All of these, you have to get the idea behind each ChIP-Seq datasets. For different histone marks, different transcription factors, the parameters described above will vary. Thus, really understand the key step of experiment (including cells preparation, DNA extraction, PCR amplification, Antibody pull down, Sequencing) is the key to know how the data quality is.

Besides ChIP-Seq, popular application of second generation sequencing, such as DNase-Seq, MNase-Seq, ATTC-Seq, MeDIP-Seq, performs a similar strategy to ChIP-Seq. In this way, the QC part is very similar to ChIP-Seq, but have a slightly different.
For DNase-Seq, the peaks are usually wider than TF ChIP-Seq and each DNase peak usually will have sub-peaks.
For MNase-Seq, you can not call peak because most of genome is covered by nucleosomes. But one can check the quality of MNase-seq by looking at the 10bp periodicity of nucleosomal DNA and the reads relative distances.
For ATTC-Seq, in addition to these, one should expect to observe the fragment size have peaks in mono/di/tri-nucleosomes locations.
For MeDIP-Seq, it’s hard to call peaks and the peak is usually very wide.