CFCE
CENTER for FUNCTIONAL CANCER EPIGENETICS

ChiP-seq QC Report For

BCL3 t00

djy3

July 5, 2017

ChiLin Summary

I Summary Table **2**

II FastQC **4**

II.1 FastQC score distribution 5

II.2 FastQC GC content distribution 6

III QC result instruction **7**

Part I

Summary Table

Metrics	treat1	control1
FastQC	36	36

Part II

FastQC

II.1 FastQC score distribution

We draw the cumulative percentage plot of the FastQC sequence quality scores of all historic data and show how your new data compare(Figure 1).

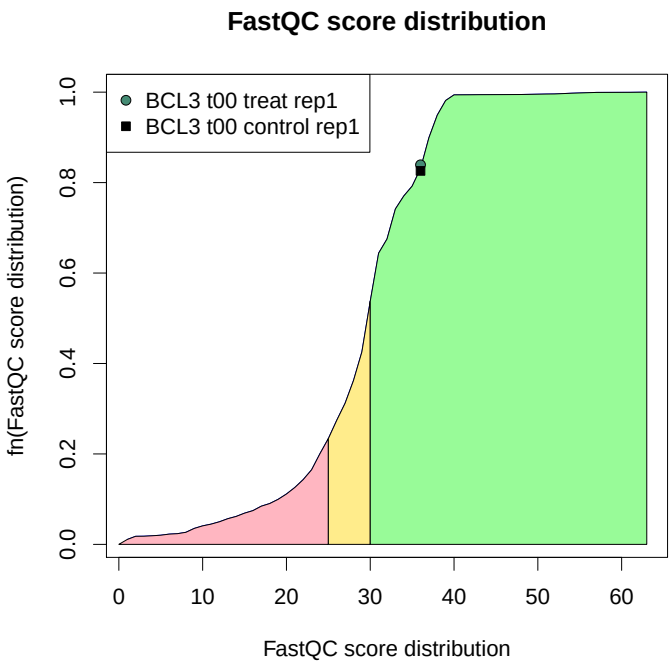
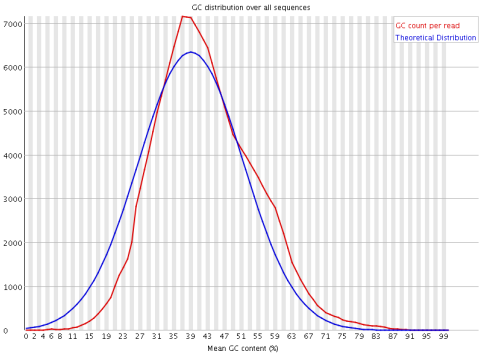


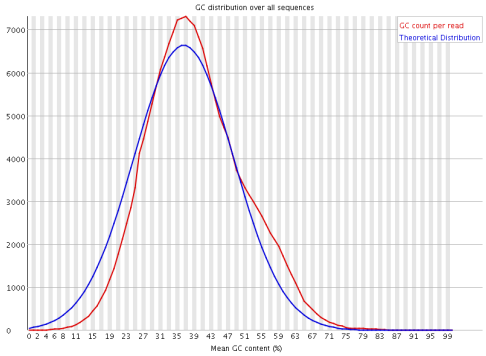
Figure 1: FastQC score distribution plot

II.2 FastQC GC content distribution

We draw the GC content of each of replicates and controls. The expected GC content distribution (i.e. theoretical) is unimodal (Figure: 2). A bi-modal distribution may indicate sample contamination.



BCL3 t00 treat rep1



BCL3 t00 control rep1

Figure 2: sequence GC content

Part III

QC result instruction

1. **FastQC** is the sample's median sequence quality scores. ChiLin calculates these scores using the FastQC software[1]. A good sequence quality score is ≥ 25.
2. **Original total reads** is the sample's raw total reads number.
3. **Uniquely mapped reads** is the number of reads with mapping quality above 1. First, ChiLin aligns reads onto user-specified genomes. Then, it filters the SAM files. The uniquely mapped RATIO is the uniquely mapped reads divided by the total reads. A good uniquely mapped ratio is $\geq 60\%$.
4. **Unique locations of 4M reads** is the number of genomic locations with one or more uniquely mapped reads (unique locations) from sub-sampled 4M reads. Unique locations ratio unique locations number divided by total number of uniquely mapped reads. ChiLin estimates NRF by dividing the number of unique locations by 4M sampled uniquely mapped reads. If reads are less than 4M, then ChiLin uses the total reads instead. ChiLin reports number of unique locations and the unique locations ratio. A good unique locations of 4M reads should be $\geq 70\%$.
5. **Locations with only 1 read from 4M reads number (ratio)** is the number of locations with read number equal to 1 (N1). The ratio is N1 divided by 4M reads unless the total reads is less than 4M, in which case the total reads is used. A good score for this metric is $> 70\%$.
6. **PBC of 4M reads** is N1 (see 5) divided by unique locations (see 4). A good PBC score is $\geq 80\%$.
7. **Fragment size of 4M reads** is in silico estimation of your size selection through maximum cross correlation. The estimation should to be close to the size selected in your experiment.
8. **Exon/DHS/Promoter ratio of 4M reads** is the estimated ratio of reads falling in these regions (from a 4M reads sub-sample). Exons regions are defined as the merged exons regions from the RefSeq gene table. Promoter regions are defined as the RefSeq TSS +/- 2kb regions. Union DHS regions are called from ENCODE II UW DNase-seq Hypersensitive regions. The IP group samples should have higher reads ratios than the control group samples.
9. **FRiP[2] of 4M non-chrM reads** is used for evaluating the signal to noise ratio. First, ChiLin removes chrM reads from the total reads. Then ChiLin sub-samples 4M of these reads. Finally, it calculates the ratio of the sub-sample which fall under the called peaks. A good FRiP score is $\geq 1\%$.
10. **Replicates total peaks** are the total peaks number called by MACS2 with fixed extension size and q value cutoff. A good peaks number depends on your experiment.
11. **Replicates 10 fold confident peaks** are the number of peaks called by MACS2 where the fold change is ≥ 10.
12. **Replicates 20 fold confident peaks** are the number of peaks called by MACS2 where the fold change is ≥ 20.
13. **Replicates reads correlation** is the whole genome reads pearson correlation for all replicates with resolution 146. A good correlation score is ≥ 0.6.
14. **Replicates peaks overlap** is the replicates peaks overlapping number.
15. **Top peaks not overlap with blacklist regions ratio** is the ratio of the merged top 5000 peaks (ordered by MACS2 -log (qvalue)) which do not overlap with blacklist region[3]. This is expected to be $\geq 90\%$.

[1] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
[2] ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia
[3] https://sites.google.com/site/anshulkundaje/projects/blacklists

16. **Top peaks overlap with union DHS number (ratio)** is the ratio of the merged top 5000 peaks (ordered by MACS2 -log (qvalue)) which overlap with union DHS regions. Union DHS regions are obtained from ENCODE II UW DNase-seq Hypersensitive regions. The union DHS regions was collected from 122 human datasets or 53 mouse datasets, we do not have union DHS of other species. Union DHS generation methods is consisted of three steps: 1.for peaks length longer than 300bp, trim macs2 peaks length to 300bp around macs2 summits, 2.if less than 300bp, preserve the original length, 3.merge the peaks overlap each other. This is expected to be $\geq$ 70%.

17. **Exon/Intron/Integenic/Promoter ratio of peak summits** is calculated using the summits of the merged peaks. ChiLin reports the ratio of overlap with exon, intron, intergenic, and promoter regions for these summits.
We list the background Exon/Intron/Integenic/Promoter ratio here:

Assembly	Exon	Intron	Intergenic	Promoter
hg19	1.92%	36.39%	58.37%	3.32%
hg38	1.95%	36.30%	58.27%	3.47%
mm9	1.91%	32.28%	62.38%	3.44%
mm10	1.91%	32.48%	62.14%	3.46%

18. **Top peaks conservation plot** is the Phastcons conservation scores distribution around +/- 2kb of the top 5000 merged peak summits. Phastcons conservation scores are from placental mammals multiple alignment. For TFs and active histone mark the plot should be shown as a sharp peak in the center.

19. **Top peaks motif analysis** is the motif analysis performed on the top 5000 merged peak summits. These summits are used for discovering highly enriched motifs with MDSeqPos., **Z-score**: a statistical measure of the motif credibility. A good motif will usuaull have a z-score of less than -15. **Hits**: number of times the motif occurs in the top 1000 regions.