Description of Quality metrics

May 16, 2014

1. **FastQC** is the sample's median sequence quality scores. ChiLin calculates these scores using the FastQC software[2]. A good sequence quality score is $\geq$ 25.

2. **Original total reads** is the sample's raw total reads number.

3. **Uniquely mapped reads** is the reliably-mapped reads ratio. ChiLin calculates this by aligning reads . Then, it filters the SAM files. The uniquely mapped reads is the reads NUMBER with mappping quality set to 1. 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 non-redundant fraction (NRF) of 4M reads. It is also sometimes called the unique locations ratio. Unique locations are the locations in which one or more reads map. 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. **NSC[1] of 4M reads** is the normalized strand cross-correlation coefficient. NSC is calculated as: $NSC = \frac{cc(fragmentlength)}{min(cc)}$, where cc is the cross-correlation. A good NSC score should be $\geq$ 1.1. This score is not useful for histone marks, such as H3K27me3.

8. **RSC[1] of 4M reads** is the relative strand cross-correlation coefficient. RSC is calculated as: $RSC = \frac{cc(fragmentlength)-min(cc)}{cc(readlength)-min(cc)}$, where cc is the cross-correlation. A good RSC score should be $\geq$ 1.0.

9. **Qtag[1] of 4M reads** is a thresholded version of RSC. Qtag is defined by the following step function: RSC $\in 0, 0.25$, QC $=$ -2; RSC $\in$ 0.25, 0.50, QC $=$ -1; RSC $\in 0.50, 1.00$, QC $=$ 0; RSC $\in$ 1, 1.50, QC = 1; RSC $>$ 1.5, QC $=$ 2. A good Qtag score is $\geq$ 1. The Qtag scores of the IP samples should be larger than the control samples.

10. **Fragment size of 4M reads** is in silico estimation of your size selection. The estimation should to be close to the size selected in your experiment.

11. **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.

12. **FRiP[3] 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 $\geq 1\%$.

13. **Replicates total peaks** are the total peaks number called by MACS2. A good peaks number depends on your experiment.

14. **Replicates 10 fold confident peaks** are the number of peaks called by MACS2 where the fold change is $\geq$ 10.

15. **Replicates 20 fold confident peaks** are the number of peaks called by MACS2 where the fold change is $\geq$ 20.

16. **Replicates reads correlation** is the whole genome reads correlation for all replicates. A good correlation score is $\geq$ 0.6.

17. **Replicates peaks overlap** is the replicates peaks overlapping number.

18. **Top peaks not overlap with blacklist regions ratio** is the ratio of the merged top 5000 peaks (ordered by MACS2 scores) which do not overlap with blacklist region. This is expected to be $\geq 90\%$.

19. **Top peaks overlap with union DHS number (ratio)** is the ratio of the merged top 5000 peaks (ordered by MACS2 scores) which overlap with union DHS regions. Union DHS regions are obtained from ENCODE II UW DNase-seq Hypersensitive regions. This is expected to be $\geq 70\%$.

20. **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%
mm10	1.91%	32.48%	62.14%	3.46%
mm9	1.91%	32.28%	62.38%	3.44%

21. **Top peaks conservation plot** is the Phaston conservation scores distribution around +/- 2kb of the top 5000 merged peak summits. For TFs and active histone mark the plot should show a sharp peak in the center.

22. **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.

References

[1] ENCODE consortium. https://genome.ucsc.edu/encode/qualitymetrics.html. 2012.

[2] FastQC. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[3] SG Landt and GK Marinov. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome . . .* , (Park 2009):1813–1831, 2012.