QC and analysis reports for MNase-seq data : GSM907784

March 17, 2019

Contents

1 Data description

Table 1 mainly describe the input file and mapping and analysis parameters.

Table 1: Data description

parameter	value
output name	GSM907784
input file A	GSM907784.bed
input file B	NA
input format	BED
sequencing type	Paired end
genome version (species)	hg19
Q30 filter mapped reads	True
custom region	NA

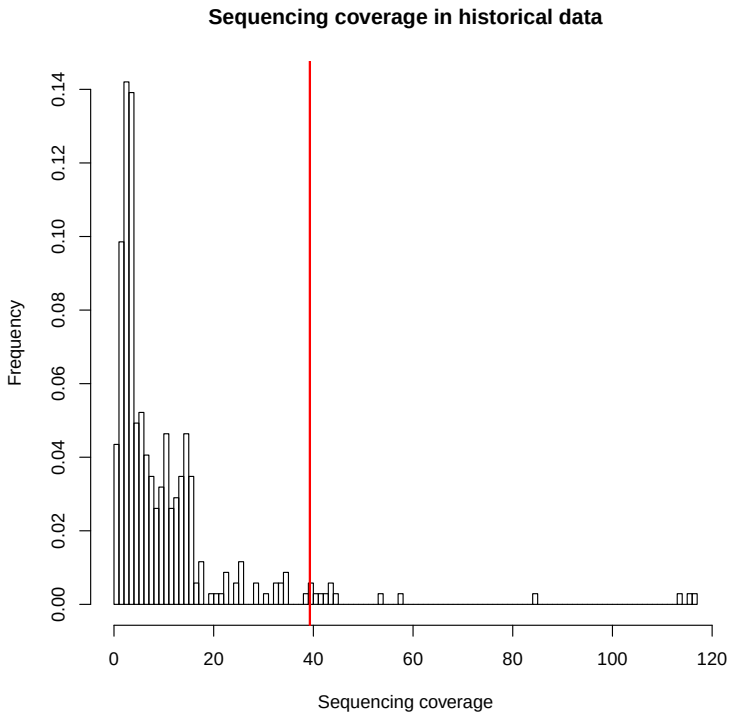
2 QC component

we calculated three key measurements: 1) sequencing coverage, 2) AA/TT/AT dinucleotide frequency and 3) nucleosomal DNA length distribution.

2.1 Sequencing coverage

Sequencing coverage provides a direct measurement of the resolution of two features of nucleosome organization, i.e. occupancy and positioning (Struhl and Segal, 2013). Sequencing coverage is defined as: (Number of reads * 194bp)/(Effective genome size). "Number of reads" is the number of mappable reads after MAPQ filtering (for single end data, for paired end it's the number of fragment). "194bp" is the total length of nucleosome and linker estimated from historical data. "Effective genome size" is defined as 2.7e9 bps for humans and 1.87e9 bps for mice. Below we plotted the distribution of sequencing coverage of historical data; the sequencing coverage of input data was marked by vertical line: 39.3.

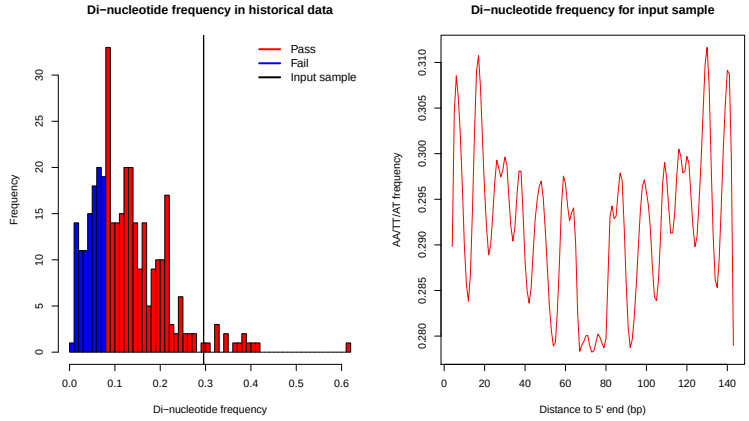
Figure 1: Sequencing coverage



2.2 AA/TT/AT di-nucleotide frequency

The 10-base AA/TT/AT periodicity in nucleosomal DNA provides a measurement of nucleosome rotational positioning, which has been shown to be influenced by DNA sequence (Satchwell, et al., 1986). Mappable reads were sampled down to 10 million and were extended to 147bp in their 3'end direction. Then the aggregate AA/TT/AT di-nucleotide frequency across 4th - 143th bp of the extended reads was calculated (right). We conducted a Fourier transform on the aggregate frequency and used the energy of 10-bp periodicity (defined as rotational score) to show the extent the MNase-seq reads reflect nucleosome organization. Sample with rotational score greater than 0.08 was defined as "Pass" in this measurement, otherwise it's defined as "Fail". The cutoff 0.08 was determined from the distribution of rotational scores from all historical data (left, vertical line marked the rotational score input sample: 0.2957 [Pass]).

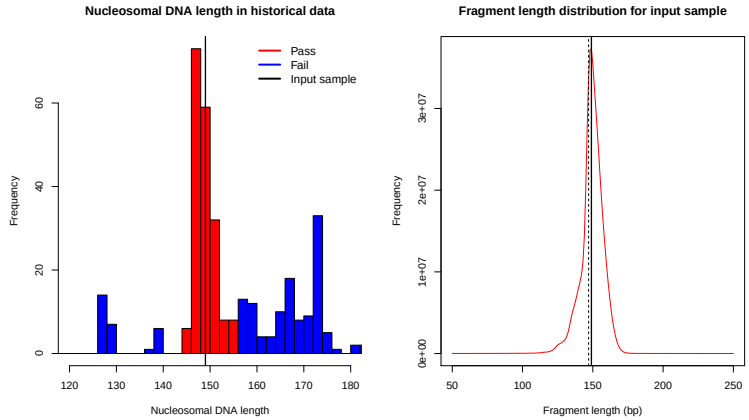
Figure 2: AA/TT/AT di-nucleotide frequency



2.3 Nucleosomal DNA length distribution

Nucleosomal DNA length distribution (refer to fragment length or MNase library size) is closely related and thus can reflect the degree of MNase digestion. For paired end sample, fragment length distribution from all mappable fragments was used directly to infer the nucleosomal DNA length distribution. For single end sample, we calculated a start-to-end distance to estimate the nucleosome length distribution: mappable reads were sampled down to 10 million and then we calculated the distribution of the distance from 5'end of each plus strand read to all 5'end of minus strand reads within 250bp downstream. Duplicate reads were discarded in this calculation. After the distribution of nucleosomal DNA length was generated, the length with highest frequency was defined as the estimated nucleosomal DNA length of the input sample (for both paired end and single end, left). Sample with nucleosomal DNA length within 140bp - 155bp was defined as "Pass", otherwise it's defined as "Fail". The cutoff was determined from the distribution of nucleosomal DNA length from all historical data (left, vertical line marked the nucleosomal DNA length of input sample: 149 [Pass]).

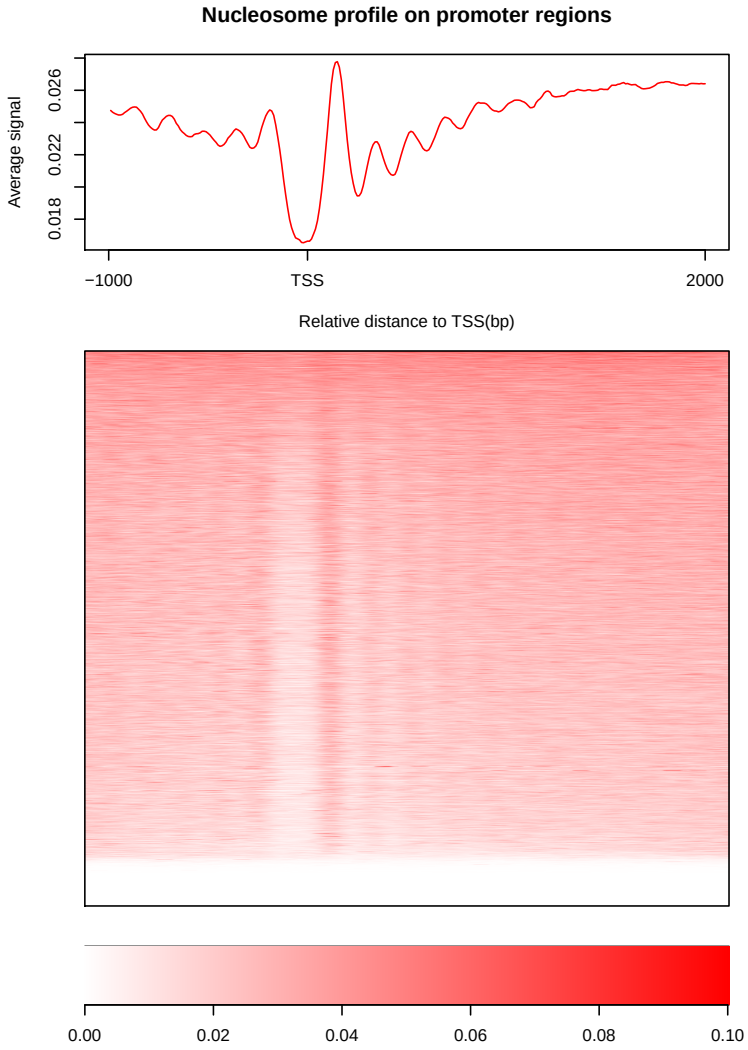
Figure 3: Nucleosomal DNA length distribution



2.4 Nucleosome profile on potential functional regions

CAM generated the average curve and the heatmap of nucleosome organization on promoter regions in 10bp resolution. Signal from minus strand regions were reversed in both heatmap and aggregate curve. The signal for each regions were also outputted as matrix: GSM907784_Tss_profile.txt.

Figure 4: Nucleosome profile on potential functional regions



2.5 Nucleosome depletion level and nucleosome fuzziness around TSS

Based on nucleosome profiles on promoters, CAM generated two scores to describe the nucleosome positioning on promoters. First, nucleosome depletion level described the fold change of the MNase-seq signal of nucleosome free regions compared to the +1 nucleosome and -1 nucleosome. The higher the nucleosome depletion level is, the deeper the nucleosome free region is. Lower nucleosome depletion level associated with weak or none nucleosome free regions, which may indicate reads from open chromatins. Samples with nucleosome depletion level higher than 0.4 was defined as "Pass", otherwise it's defined as "Fail".The cutoff was determined based on the distribution of nucleosome depletion level from all historical data (left, vertical line marked the nucleosome depletion level of input sample: 0.883 [Pass]). Next, nucleosome fuzziness downstream TSS defined whether clear nucleosome positioning pattern was observed from downstream promoters. The nucleosome fuzziness was calculated by the coefficient of variance (CV) of the linker length between the +1, +2, +3 and +4 nucleosomes. The lower the nucleosome fuzziness is, the better nucleosome positioning was observed on promoters. Samples with nucleosome fuzziness lower than 0.4 was defined as "Pass", otherwise it's defined as "Fail".The cutoff was determined based on the distribution of nucleosome fuzziness scores from all historical data (right, vertical line marked the nucleosome fuzziness of input sample: 0.0833 [Pass]).

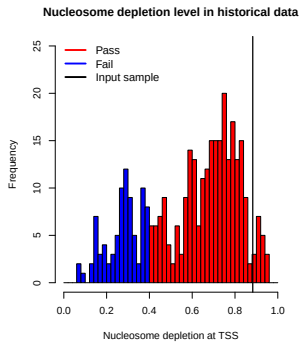


Figure 5: nucleosome depletion

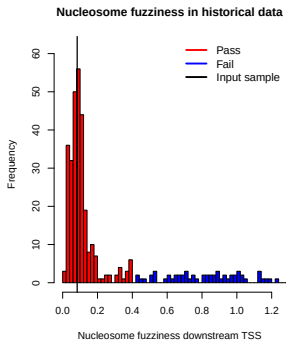


Figure 6: nucleosome fuzziness

2.6 Well-positioned nucleosome arrays

Regions with well-positioned nucleosome arrays are detected as previous described (Zhang, et al., 2014), and the enrichment in potential regulatory regions (downstream promoter and union DNase I hypersensitive sites (DHS sites)) is listed. Enrichment was defined as observed/expected percentage of nucleosome array on promoter ($>$ 1 for enriched). Expected percentage was equal to the percentage of promoter length compared to the total length of effective genome. Similar approach was applied on union DHS sites. For each region with well-positioned nucleosome array, its genomic coordinates together with nucleosome profile values were reported in the output file: GSM907784_profile_on_Nucleosome_Array.bw. Nucleosome arrays with fold enrichment of DHS sites less than 2 is regarded as Fail in this measurement,while fold enrichment of UTR regions less than 1 is regarded as "Fail", indicating the well-positioned nucleosome arrays are more likely to be caused by random rather than the barrier model.

Table 2: Enrichment of well-positioned nucleosome arrays

genomic region(Category)	enrichment
downstream promoter	2.1278 [Pass]
union DHS sites	4.0143 [Pass]

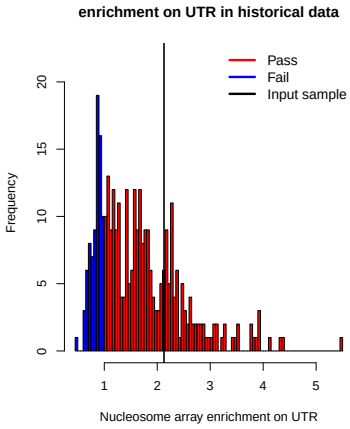


Figure 7: enrichment on UTR

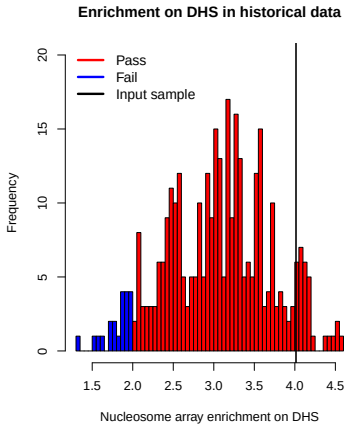


Figure 8: enrichment on DHS

3 Output list

All output files were described in the following table

Table 3: output list

filename	description
GSM907784.bed/GSM907784.bb	mapped reads on the genome
GSM907784_profile.bw	genome-wide nucleosome profile
GSM907784_center.bw	genome-wide nucleosome dyad profile
GSM907784_Tss_profile.txt	nucleosome signal on promoter regions
GSM907784_fraglen.txt	nucleosome fragment length distribution
GSM907784_Nucleosome_Array.bed	well-positioned nucleosome arrays
GSM907784_geneLevel_nucarrayAnnotation.bed	gene level annotation of nuc-arrays
GSM907784_profile_on_Nucleosome_Array.bw	nucleosome signal on well-positioned nucleosome arrays
GSM907784_center_position.bw	nucleosome array score signal on the genome
GSM907784_summary.pdf	summary QC report