ChIP-Seq data processing, normalization and visualization

ADITYA SANKAR (adisankara2000@gmail.com)
University of Copenhagen https://orcid.org/0000-0002-1840-3356
Mads Lerdrup (mads.lerdrup@bric.ku.dk)
University of Copenhagen https://orcid.org/0000-0002-7730-8973

Method Article

Keywords: Chromatin, ChIPSeq, Histone modifications

Posted Date: January 27th, 2020

DOI: https://doi.org/10.21203/rs.2.21645/v1

Abstract

This article explains the step-by-step processing of NGS data of ChIPSeq samples for H3K4me3, H3K9me3 and H3K36me3 immunoprecipitations performed in oocytes/embryos. This includes sections on visualization of data, and a comment on normalization.

the associated raw data have been deposited at GSE129735 on Gene Expression Omnibus.

Introduction

This procedure has been applied to process ChIPSeq data from the associated publication. Some aspect of this procedure can be used as inspiration for processing low input ChIPSeq datasets

Equipment

1. Windows PC with atleast 12 GB RAM powerful enough to run EaSeq sessions listed in this article.

2. Installed EaSeq software available freely at www.easeq.net

Procedure

ChIP-seq data processing

1. ChIP-seq libraries were sequenced as 75 and 150 bp paired end reads on an Illumina NextSeq 500.

2. Reads were trimmed using Trimmomatic[1] with the settings “ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 SLIDINGWINDOW:4:20 HEADCROP:15 CROP:120 MINLEN:40” for H3K4me3, “SLIDINGWINDOW:4:20 HEADCROP:5 CROP:50 MINLEN:30” for MII H3K9me3 data, and “SLIDINGWINDOW:4:20 HEADCROP:5 CROP:125 MINLEN:30" for H3K36me3 and 2-cell H3K9me3 data.

3. ChIP-seq reads were aligned as paired ends to the mm10 genome assembly using bowtie2[2] and the settings “–local -X 2000”, sorted and exported as bam-files using samtools[3], and imported into EaSeq[4] (http://easeq.net) for subsequent analyses and visualization.

4. In EaSeq, default settings for ChIP-seq data are used including filtering of reads with identical coordinates. Where nothing else is mentioned, values were normalized to fragments per million reads per kbp (FPKM).

General ChIP-seq data analysis and visualization

1. Unless stated, then all analysis and visualization was performed using EaSeq v.1.05-1.111 (http://easeq.net) and its integrated tools using the default settings.

2. All these analyses, visualizations and used data are available as a session file at http://easeq.net/Sessions/Sankar.eas. Please note that the session requires 12 GB of available RAM.

3. Plots were exported for layout in Adobe Illustrator CS6 as pdfs using the beta tool 'Export snapshot as pdf'.

4. Gene bodies were extracted from the imported RefSeq 'Geneset' using the 'Extract' tool and 'Feature' set to 'Gene', 'Start position' set to 'Start', and 'End position' set to 'End'.

5. To obtain non-redundant genes without multiple identical termini, gene symbols were exported and duplicates were removed. The list of symbols was imported as a 'Regionset' using the 'Geneset' to lookup coordinates. For gene symbols with multiple annotated transcripts, this resulted in the most extreme coordinates being used for subsequent analysis. Possible gene symbols not matched by coordinates were removed using the gate tool.

6. Broad H3K4me3 domains were acquired as described above and imported into EaSeq as 'Regionsets'. The set of broad H3K4me3 regions were subselected for those at or above 20 kbp of size for further analyses.

7. For broad domains and gene bodies, the H3K9me3 levels in replicate 1 were quantified using the ‘Quantify’ tool within windows corresponding to the entire length of the regions, and the order of the regions was changed according to these values using the ‘Sort’ tool.

8. Heatmaps were generated using the ‘HeatMap’ plot type and ratiometric heatmaps showing pseudocolored differences between conditions were generated using the ratiomap plot type.

9. For plots of gene bodies and broad H3K4me3 domains, settings were adjusted to a relative window size of 200% of the size of the visualized feature by checking ‘rel.’ and changing ‘Window size’ to 200 in the plot settings.

10. For all heatmaps ‘Maximum density’ was adjusted as indicated by the FPKM values in the figures.

11. Plots showing the average values at a set of regions were generated using the ‘Average’ tool, where gene bodies and broad H3K4me3 domains adjusted to a relative window size of 200% of the size of the visualized feature by checking ‘rel.’ and changing ‘Window size’ to 200 in the plot settings. ‘Y-axis Maximum’ was adjusted as shown in the figures and multiple plots were overlaid using the ‘Overlay’ tool.

12. Genome browser tracks were generated using the ‘FillTrack’ tool with ‘Y-axis Maximum’ adjusted as shown in the figures and for some tracks loci coinciding with broad H3K4me3 domains were coloured using the ‘Highlight’ option in the plot settings and the broad H3K4me3 domains as the feature to highlight.

13. Tracks showing log2 fold differences were adjusted in the plot settings by setting ‘Ratio’ to the ‘Dataset’ in the denominator and for the Y-axis checking ‘Log scale’, setting the ‘Base’ to 2, and ‘Maximum’ to 2 and ‘Minimum’ to -2.

14. 2D-histograms and associated correlation efficiencies were generated using the ‘Scatter’ tool and adjusted to show 100 bins on both axes in the plot settings. The visualized values were quantified at a set of regions corresponding to 10kbp windows within the entire span of mm10 genome using the

'Quantify' tool and 'Start position offset' and 'End position offset' adjusted to -5000 and 5000 bp, respectively.

15. Visualized ratios were derived by applying the 'Calculate' tool with 'Division' and 'Logarithm' checked to the respective quantified values.

ChIP-seq data normalization (a comment)

1. When assessing differences in H3K9me3 ChIP-seq signal in the *Kdm4a*$^{-/-}$ compared to *Kdm4a*$^{+/+}$ MII oocytes using standard FPKM normalization, we observed a loss of signal at gene bodies, where H3K9me3 was enriched in the *Kdm4a*$^{+/+}$ oocytes.

2. To cross-validate the distribution in the wild-type we used published H3K9me3 data from MII oocytes[5] and found similar enrichment at gene bodies as well as depletion within Broad H3K4me3 Domains (Extended Fig. 4 a, b).

3. We have previously observed that large increases in the fraction of the genome covered by H3K4me3 in MII oocytes compared to e.g. mouse ES cells, results in a much lower local read density due to the sequenced reads being dispersed over a larger fraction of the genome[6].

4. This effect is unaccounted for by the widely used RPKM/FPKM normalization of ChIP-seq samples, and for the previous study we therefore used a normalization strategy based on H3K4me3 at the top-5000 ranked transcription start sites[6].

5. In this study, the widespread gain of H3K9me3 signal at the H3K4me3 enriched regions where it is normally depleted occurs in more than one fifth of the mouse genome[6]. This would not only elevate the signal in the normally depleted regions, but also lead to a considerably reduced signal in the areas of canonical enrichment (Extended Figure 4c).

6. Therefore, we would be highly cautious about interpreting the loss of gene body H3K9me3 in absolute terms, and prefer conservative interpretations. Accordingly, we find it likely that the apparent loss of H3K9me3 is due to the susceptibility of ChIP-seq and FPKM normalization to pronounced increases in the fraction of the genome, which is covered by H3K9me3.

7. As assessments of the extent of this in a quantitatively manner would require a spike-in setup, which have not yet been developed for the very small-scale ChIP-seq used here, we chose the commonly used FPKM-normalization for plots and analyses, and to make conservative statements that are unaffected by the choice of normalization.

8. This conservative approach could lead to underestimations of the extent of H3K9me3 gain in broad domains, and to illustrate the consequences, we have done pairwise comparisons of FPKM normalized plots and similar plots normalized in a manner parallel to what was previously used[6] (Extended Figure 4 d,e).

Anticipated Results

Trimmed, mapped data uploading to genome browser for visualization or direct use for further analysis in EaSeq

References

1. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).

2. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).

3. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

4. Lerdrup, M., Johansen, J.V., Agrawal-Singh, S. & Hansen, K. An interactive environment for agile analysis and visualization of ChIP-sequencing data. *Nat Struct Mol Biol* **23**, 349-357 (2016).

5. Wang, C. *et al.* Reprogramming of H3K9me3-dependent heterochromatin during mammalian embryo development. *Nat Cell Biol* **20**, 620-631 (2018).

6. Dahl, J.A. *et al.* Broad histone H3K4me3 domains in mouse oocytes modulate maternal-to-zygotic transition. *Nature* **537**, 548-552 (2016).

Acknowledgements

John Arne Dahl, Principal Investigator, University of Oslo

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ExtendedFigure4.pdf