Bioinformatic analysis of Native Chromatin Immunoprecipitation with high throughput Sequencing (N-ChIP-seq) in Sheep Alveolar Macrophages

Written by:

Alisha T. Massa

Washington State University, Pullman, WA, USA

September 20, 2020

#Examples of script usage are given in the gray boxes on lines that do not have a hashtag (#). Hashtag lines include notes on optional parameters. You can always choose your own meaningful file names for output. The scripts provided here were run on the bash command line in CentOS Linux release 7.8.2003. It may be more effective for large numbers of files if you modify the scripts and use a compute cluster, please ask your friendly neighborhood bioinformatician for assistance. Similar tasks may also be completed in the graphical user interface of Galaxy available at https://usegalaxy.org/

De-multiplexing sequencing data

1) .bcl files were converted to .fastq format and the adapter sequences were trimmed from reads using bcl2fastq2 (https://support.illumina.com/downloads/bcl2fastq-conversion-software-v2-20.html Illumina, San Diego, CA, USA).
2) FastQC (Andrews, 2016) was used to check duplication rate, read quality, and ensure adapters were removed. Read files for each sheep and each immunoprecipitation target were saved in compressed format as "Sheep1_K27ac.fastq.gz"

Mapping the trimmed reads to the reference genome

3) Reads were mapped using BWA v0.7.17 (Li & Durbin, 2009) each of the following reference genome assemblies after first creating the reference genome index in BWA:
    a. Oar_rambouillet_v1.0 excluding the mitochondrial genome (GCA_002742125.1, Worley, K., personal communication, 2019, (Worley, 2017)). Shown below as "Oar_rambouillet.fna"
    b. Oar_v3.1 (GCA_000298735.1)
    c. Oar_v4.0 (GCA_000298735.2)

#creates the index for alignment, optional parameters q 15 is used to specify the quality threshold for read trimming down to 35bp, and t 24 is specifying the number of threads to use to increase speed of processing (adjust this based on your system).

```
bwa aln -q 15 -t 24 Oar_rambouillet.fna Sheep1_K27ac.fastq.gz > Sheep1_K27ac.sai
```

```
#creates the alignment output file, default parameters

bwa samse Oar_rambouillet.fna Sheep1_K27ac.sai Sheep1_K27ac.fastq.gz > Sheep1_K27ac.bam
```

4) BAM files were sorted using Picard v2.9.2 (http://broadinstitute.github.io/picard/).

```
#be sure to adjust the validation_stringency to lenient if you want to see errors during this process which can help catch file formatting problems. You can also create the index at this time in a single step which will be a .bai file.

java -jar /opt/modules/biology/picard/2.9.2/bin/picard.jar SortSam I=Sheep1_K27ac.bam O=Sheep1_K27acSORT.bam SORT_ORDER=coordinate VALIDATION_STRINGENCY=LENIENT CREATE_INDEX=True
```

5) Mapped reads can now be filtered as preferred to prepare them for peak analysis. SAMtools v1.9 (Li et al., 2009) was used to remove reads that did not map, but retain reads that had a single best mapping location (primary mapping), any mapQ cutoff greater than 3 should remove reads that map equally to more than one location. Relaxed mapQ cutoffs will retain reads that contain SNP but increases the chance of retaining reads with sequencing errors. We filtered reads between mapQ cutoff score of 1-30 for comparison. Filter mapQ cutoff of 5 was chosen.

```
#Filtering to remove unmapped reads and reads that map equally to more than one location with q 5 parameter, b parameter tells SAMtools to print the output in BAM format.

samtools view -b -q 5 Sheep1_K27acSORT.bam > Sheep1_filterq5_K27acSORT.bam
```

Peak calling with MACS

6) The effective genome size must be calculated or empirically determined based on known references such as the Golden Path Length from ENSEMBL. This value can be calculated based on the length of your sequencing reads. For sheep we chose 2.62e9 based on the Golden Path Length. Peaks were called in individual animals at multiple false discovery rates (FDR) of less than 1%, 5%, and 10% with MACS2 v2.1.1 (Feng et al., 2012; Zhang et al., 2008). Check that MACS model is determining the correct fragment length (d) from bioanalyzer fragment analysis of your chromatin or from other methods of fragment length analysis. Note that MACS2 broad mode uses a default setting of 0.1 FDR to call broad peaks and the specified FDR to call narrow peaks. Broad peak calling option was used for H3K4me1 and H3K27me3 ChIP data. Input DNA from the same chromatin digestion batch was used as the control and is named "Sheep1_filterq5_inputSORT.bam." Next reads were pooled from filtered BAM files of both animals and used to call pooled peaks.

```
#call peaks with MACS2, -g parameter specifies effective genome size, -q specifies the FDR. Parameter -B specifies to include a bedGraph file in the output, can chose to save signal per million reads with this option, add --SPMR

#narrow peaks

macs2 callpeak -t Sheep1_filterq5_K27acSORT.bam -c Sheep1_filterq5_inputSORT.bam -f BAM -g 2.62e9 -n Sheep1_filterq5_FDR05_K27ac -B -q 0.05

#broad peaks (H3K27me3 and H3K4me1)

macs2 callpeak -t Sheep1_filterq5_K4m1SORT.bam -c Sheep1_filterq5_inputSORT.bam -f BAM -g 2.62e9 -n Sheep1_filterq5_FDR05_K4m1broad -B -q 0.05 --broad
```

Calling consensus peaks present in both replicates

7) MACS2 outputs from each animal and for the pooled peaks were renamed from .narrowPeak and .broadPeak files to .bed files.
8) BED files were sorted for comparative analyses.

```
#renaming and sort the peak files into .bed files

sort -k 1,1 -k2,2n Pooled_filterq5_FDR05_K4m1broad_peaks.broadPeak > Pooled_filterq5_FDR05_K4m1.bed
```

9) Called peaks that overlap between each animal and were called as significant in the pooled peaks were considered reproducible consensus peaks. These were determined with bedtools v2.26.0 and bedops v2.4.38 (Neph et al., 2012; Quinlan & Hall, 2010). These tools were also used to determine overlapping regions between multiple ChIP targets.

```
#to determine consensus peaks. Note when using bedops merge to collapse duplicated regions it may only retain the first three required BED fields in the output file with chromosome, start, and end positions.  You may need to add back additional information to the peaks if needed for downstream analysis.  UCSC provides some useful tools in the their binary utilities directory https://genome.ucsc.edu/goldenPath/help/bigWig.html and available here http://hgdownload.soe.ucsc.edu/admin/exe/


bedops -u Pooled_filterq5_FDR05_CTCF.bed Sheep1_filterq5_FDR05_CTCF.bed Sheep2_filterq5_FDR05_CTCF.bed \
   | bedmap --echo --echo-map-id-uniq --max-element - \
   | awk -F"|" '(split($2, a, ";") > 2)' \
   > bedopsunion_all3overlap_CTCF.bed

bedops --merge --ec bedopsunion_all3overlap_CTCF.bed > Consensus_CTCF.bed
```

#to determine peak regions from multiple ChIP targets that overlapped, this is directional, so only the regions in the first file are reported that overlap with the second file. Optional parameter -u reports each sequence in the first file only once, even if the second file has multiple regions that overlap. We then counted how many regions were reported in the output.

intersectBed -a Sheep1_filterq5__FDR05_K4m1.bed -b Sheep1_filterq5_FDR05_K27ac.bed -wa -u -sorted > Sheep1_K4m1peakswithK27ac.bed

wc -l Sheep1_K4m1peakswithK27ac.bed

Convert BAM files to normalized wiggle tracks for visualization

10) BAM files are often very large and can be difficult to view on your local computer. BigWig format has much reduced file size and during conversion we can normalize the data and subtract the input control to remove noise for visualization of ChIP-seq signal tracks. This requires two steps first conversion of the BAM file to BigWig (.bw) format and second removal of the input signal both with tools within deepTools v3.3.0 (Ramírez et al., 2016).

#convert BAM files to .bw format, with smoothing over 10 bp binSize, and normalization based on reads per genomic coverage (RPGC, signal normalized to 1X genome sequencing coverage), effective genome size for sheep was specified as previously mentioned. Reads were extended to the average digested fragment length of 150 bp (single nucleosome). You can adjust to use the optimal number of processors for your system and amount of time you can let the process run.

bamCoverage --bam Pooled_filterq5_K27acSORT.bam -o Pooled_K27ac_norm.bw --binSize 10 --normalizeUsing RPGC --effectiveGenomeSize 2620000000 --extendReads 150 --numberOfProcessors max/2

#remove signal from the input control. Note that this likely will create some areas in your ChIP signal tracks with negative signal (captured in the input, but de-enriched in high quality immunoprecipitation data) which you can chose to display or not when viewing. Name your output something meaningful to you with the -o parameter.

bigwigCompare --bigwig1 Pooled_K4m1_norm.bw --bigwig2 Pooled_input_norm.bw --binSize 10 -p max/2 -o Pooled_K4m1-input.bw -of bigwig --operation subtract

Annotate peaks with nearest gene information

11) Annotate peak regions with the nearest gene information using HOMER v4.10.4 (Heinz et al., 2010) annotatePeaks.pl program. It can be very helpful to also chose to output a file with a summary of the annotation statistics in addition to the annotation output. The NCBI *Ovis aries* Refseq Annotation Release 103 for the Rambouillet sheep genome was used (GCF_002742125.1) (O'Leary et al., 2016).

#annotate the peaks with nearest gene and output to a text file, output summary annotation statistics to a second text file. You can name them something meaningful to you. Note you may need to specify the full path to the file location and you can optionally include a directory location where you want the output files to be placed.

annotatePeaks.pl Consensus_K4m3.bed Oar_rambouillet.fna -gtf Oar_rambouillet.gtf -annStats Consensus_K4m3annStats.txt > Consensus_K4m3geneannotations.txt

References

Andrews, S. (Babraham B. (2016). *FastQC: a quality control tool for high throughput sequence data.* (v0.11.3). available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Feng, J., Liu, T., Qin, B., Zhang, Y., & Liu, X. S. (2012). Identifying ChIP-seq enrichment using MACS. *Nature Protocols*, *7*(9), 1728–1740. https://doi.org/10.1038/nprot.2012.101

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell*, *38*(4), 576–589. https://doi.org/10.1016/j.molcel.2010.05.004

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K., Rynes, E., Maurano, M. T., Vierstra, J., Thomas, S., Sandstrom, R., Humbert, R., & Stamatoyannopoulos, J. A. (2012). BEDOPS: high-performance genomic feature operations. *Bioinformatics*, *28*(14), 1919–1920. https://doi.org/10.1093/bioinformatics/bts277

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733-45. https://doi.org/10.1093/nar/gkv1189

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F., & Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, *44*(W1), W160–W165. https://doi.org/10.1093/nar/gkw257

Worley, K. C. (Baylor C. of M. H. G. S. C. (2017). *Oar_rambouillet_v1.0*. GCA_002742125.1. https://www.ncbi.nlm.nih.gov/assembly/GCF_002742125.1/

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., & Liu, X. S. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, *9*(9), R137. https://doi.org/10.1186/gb-2008-9-9-r137