UGWU PASCHAL

+2348166207095 | ugwupaschal@gmail.com

https://www.linkedin.com/in/paschal-ugwu-52abb6229/

Galaxy training by Lucille Delisle, Maria Doyle and Florian Hey

Topic: ATAC-Seq data analysis.

September, 2020

Abstract

With the help of nucleosomes, the genome is organized and tightly packed in many eukaryotic organisms, including humans (chromatin). Numerous factors, such as the chromatin structure, the position of the nucleosomes, and histone modifications, have a significant impact on the organization and accessibility of the DNA. As a result, these components affect how genes are activated and deactivated. The Assay for Transposase-Accessible Chromatin using Sequencing (ATAC-Seq) technique investigates the accessibility of chromatin to determine the regulatory mechanisms of gene expression. I aimed to contrast the projected open chromatin regions with the recognized CTCF binding sites, a DNA-binding protein thought to be involved in 3D structure. In order to complete this project, I employed the following techniques: peak calling, preprocessing, mapping, filtering mapped reads, filtering duplicate reads, and visualization of coverage. I identified open chromatin regions using MACS2, a method for identifying genomic enrichment regions (peaks). I investigated the read coverage close to the TSS using computeMatrix and plotHeatmap. Putative enhancer regions in open chromatin regions that did not coincide with CTCF sites or TSS may have been discovered by the ATAC-Seq experiment.

Background of Project

From to the tutorial on ATAC-Seq data analysis (Galaxy Training Materials) by Lucille *et al.,* 2021, I gathered that the genome is densely packed and ordered with the aid of nucleosomes in many eukaryotic species, including humans (chromatin). A nucleosome is a structure made up of eight histone proteins and ~147 bp of DNA. The DNA will be split open and released from the nucleosome complex when it is actively being translated into RNA. The arrangement and accessibility of the DNA are greatly influenced by a variety of elements, including the chromatin structure, the location of the nucleosomes, and histone modifications. As a result, these elements play a role in

the activation and inactivation of genes. To ascertain the regulatory mechanisms of gene expression, the Assay for Transposase-Accessible Chromatin utilizing sequencing (ATAC-Seq) approach examines the accessibility of chromatin. The technique can be used to find possible enhancers, silencers, and promoter regions. The DNA region around the transcription start point is known as a promoter (TSS). It has transcription factor binding sites that will draw in the RNA polymerase. A DNA region known as an enhancer can be found up to 1 Mb upstream or downstream of the promoter. The transcription of the gene is boosted when transcription factors bind an enhancer and come into touch with a promoter region. A silencer, on the other hand, reduces or prevents the gene's expression. Because it is simpler, quicker, and uses less cells than competing techniques like FAIRE-Seq and DNase-Seq, ATAC-Seq has gained popularity for discovering accessible areas of the genome. Using ATAC-Seq, a hyperactive Tn5 transposase derivative is applied to the genome in order to identify accessible (open) chromatin areas. A transposable element, which is a DNA sequence that may move (transpose) within a genome, can bind to a transposase. At the same time that the DNA is being sheared by the transposase activity during ATAC-Seq, the modified Tn5 inserts DNA sequences corresponding to shortened Nextera adapters into open regions of the genome. The read library is then ready for sequencing, which includes purification procedures and PCR amplification using complete Nextera adapters.

High-throughput data production has revolutionized molecular biology. However, massive increases in data generation capacity require analysis approaches that are more sophisticated, and often very computationally intensive. Thus, making sense of high-throughput data requires informatics support. Galaxy (http://galaxyproject.org) is a software system that provides this support through a framework that gives experimentalists simple interfaces to powerful tools, while automatically managing the computational details. Galaxy is distributed both as a publicly available Web service, which provides tools for the analysis of genomic, comparative genomic, and functional genomic data, or a downloadable package that can be deployed in individual laboratories. Either way, it allows experimentalists without informatics or programming expertise to perform complex large-scale analysis with just a Web browser (Blankenberg *et al.,* 2010).

Aim of Study

I aimed to contrast the projected open chromatin regions with the recognized

CTCF binding sites, a DNA-binding protein thought to be involved in 3D structure. Data from the study by Buenrostro *et al.* (2013), the first paper on the ATAC-Seq technology, will be used in this project. The information comes from GM12878, a human cell line of pure CD4+ T cells. I downsampled the original dataset to 200,000 randomly chosen reads because the original dataset, which included 2 × 200 million reads, would have been too large to analyze in this project. To have a solid profile on chromosome 22 equivalent to what is obtainable with a typical ATAC-Seq sample (2 x 20 million reads in original FASTQ), I also added roughly 200,000 read pairs that will map to this chromosome. Additionally, I aim to contrast the projected open chromatin regions with the recognized CTCF binding sites, a DNA-binding protein thought to be involved in 3D structure. CTCF can be used as a positive control to determine the efficacy of the ATAC-Seq experiment because it is known to bind to hundreds of locations throughout the genome. For instance, at some CTCF binding locations, good ATAC-Seq data would contain accessible regions both inside and outside of TSS. Due to this, I will download CTCF binding sites from ENCODE that were discovered by ChIP in the same cell line (ENCSR000AKB, dataset ENCFF933NTR).

Method

1. Preprocessing

a. Get Data: I downloaded the sequenced reads (FASTQs) as well as other annotation files. Then, to increase the number of reads that will map to the reference genome (here human genome version 38, GRCh38/hg38), I preprocessed the reads.

b. Quality Control: The first step was to check the quality of the reads and the presence of the Nextera adapters. When I performed ATAC-Seq, I got DNA fragments of about 40 bp since two adjacent Tn5 transposases cut the DNA (Adey *et al.* 2010). This can be smaller than the sequencing length so I expected to have Nextera adapters at the end of those reads and assess the reads with **FastQC**.

c. Trimming Reads: To trim the adapters I provided the Nextera adapter sequences to **Cutadapt**. The forward and reverse adapters were slightly different. I also trimmed low quality bases at the ends of the reads (quality less than 20). I only kept reads that were at least 20 bases long. I remove short reads (< 20bp) as they

were not useful, they would either be thrown out by the mapping or may interfere with our results at the end.

2. Mapping

a. Mapping Reads to Reference Genome: Next I mapped the trimmed reads to the human reference genome. Here I used **Bowtie2**. I extended the maximum fragment length (distance between read pairs) from 500 to 1000 because I know some valid read pairs are from this fragment length. I used the --very-sensitive parameter to have more chance to get the best match even if it takes a bit longer to run. I ran the **end-to-end** mode because I trimmed the adapters so I expected the whole read to map, no clipping of ends was needed. Regarding the genome I chose, the hg38 version of the human genome contains alternate loci. This means that some region of the genome are present both in the canonical chromosome and on its alternate loci. The reads that map to these regions would map twice. To be able to filter reads falling into repetitive regions but keep reads falling into regions present in alternate loci, I mapped on the Canonical version of hg38 (only the chromosome with numbers, chrX, chrY, and chrM).

3. Filtering Mapped Reads

a. Filter Uninformative Reads: I applied some filters to the reads after the mapping. ATAC-Seq datasets can have a lot of reads that map to the mitchondrial genome because it is nucleosome-free and thus very accessible to Tn5 insertion. The mitchondrial genome is uninteresting for ATAC-Seq so I removed the reads. I also removed reads with low mapping quality and reads that were not properly paired.

b. Filter Duplicate Reads: Because of the PCR amplification, there might be read duplicates (different reads mapping to exactly the same genomic region) from overamplification of some regions. As the Tn5 insertion was random within an accessible region, I did not expect to see fragments with the same coordinates. I considered such fragments to be PCR duplicates. I removed them with **Picard MarkDuplicates**.

c. Check Insert Sizes: I checked the insert sizes with **Paired-end histogram** of insert size frequency. The insert size is the distance between the R1 and R2 read pairs. This tells us the size of the DNA fragment the read pairs came from. The fragment length distribution of a sample gives a very good indication of the quality of the ATAC-Seq.

4. Peak calling

a. Call Peaks: I finished the data preprocessing. Next, in order to find regions corresponding to potential open chromatin regions, I wanted to identify regions where reads have piled up (peaks) greater than the background read coverage. The tools which are currently used are Genrich and MACS2. MACS2 is more widely used. Genrich has a mode dedicated to ATAC-Seq but is still not published and the more reads you have, the less peaks you get (see the issue here).
At this step, two approaches existed:

- The first one was to select only paired whose fragment length was below 100bp corresponding to nucleosome-free regions and to use a peak calling like for a ChIP-seq, joining signal between mates. The disadvantages of this approach is that one can only use it if he/she has paired-end data and will miss small open regions where only one Tn5 bound.
- The second one chosen here was to use all reads to be more exhaustive. In this approach, it was very important to re-center the signal of each reads on the 5' extremity (read start site) as this is where Tn5 cuts. Indeed, you want your peaks around the nucleosomes and not directly on the nucleosome:

b. Using MACS2: I converted the BAM file to BED format because when I set the extension size in MACS2, it would only consider one read of the pair while here we would like to use the information from both.

5. Visualisation of Coverage

a. Prepare the Datasets

i. Extract CTCF peaks on chr22 in intergenic regions: As our training dataset was focused on chromosome 22 I only used the CTCF peaks from chr22. I expected to have ATAC-seq coverage at TSS but only good ATAC-seq had coverage on intergenic CTCF. Indeed, the CTCF protein was able to position nucleosomes and creates a region depleted of nucleosome of around 120bp (Fu *et al.* 2008). This is smaller than the 200bp nucleosome-free region around TSS and also probably not present in all cells. Thus it is more difficult to get enrichment. In order to get the list of intergenic CTCF peaks of chr22, I will first select the peaks on chr22 and then exclude the one which overlap with genes.

b. Convert bedgraph from **MACS2** to bigwig: The bedgraph format is easily readable for human but it can be very large and visualising a specific region is quite slow. I changed it to bigwig format which is a binary format, so I can visualise any region of the genome very quickly.

c. Create heatmap of coverage at TSS with deepTools: I was interested in checking the coverage on specific regions. For this, I computed a heatmap. I used the **deepTools plotHeatmap**. As an example, I made a heatmap centered on the transcription start sites (TSS) and another one centered on intergenic CTCF peaks. First, on the TSS:

i. Generate computeMatrix: The input of **plotHeatmap** is a matrix in a hdf5 format. To generate it I used the tool **computeMatrix** that evaluated the coverage at each locus I was interested in.

ii. Plot with **plotHeatmap: I** now generated a heatmap. Each line will be a transcript. The coverage was summarized with a color codes. All TSS were aligned in the middle of the figure and only the 2 kb around the TSS were displayed. Another plot, on top of the heatmap, showed the mean signal at the TSS. There was one heatmap per bigwig.

d. Visualise Regions with pyGenomeTracks: In order to visualise a specific region (e.g. the gene RAC2), I can used a genome browser like **IGV** or **UCSC browser**, or use **pyGenomeTracks** to make publishable figures. I used **pyGenomeTracks**.

Results

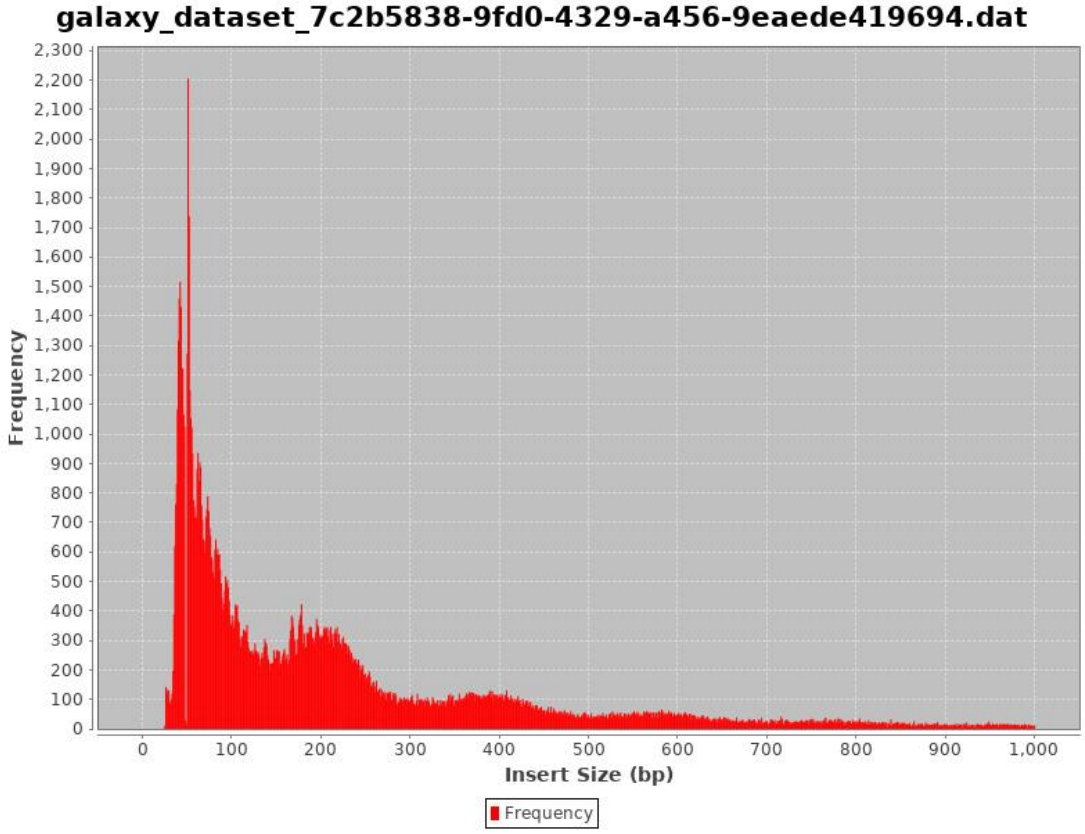

Figure 1: Fragment size distribution

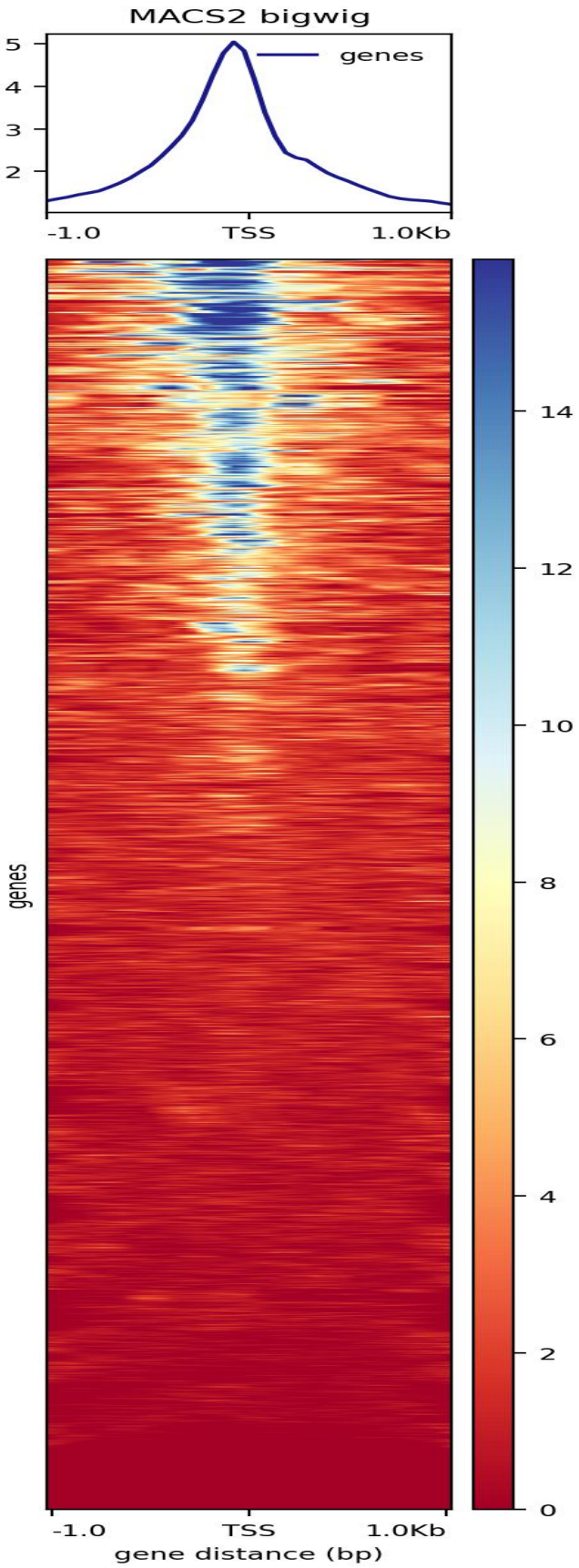

Figure 2: plotHeatmap output

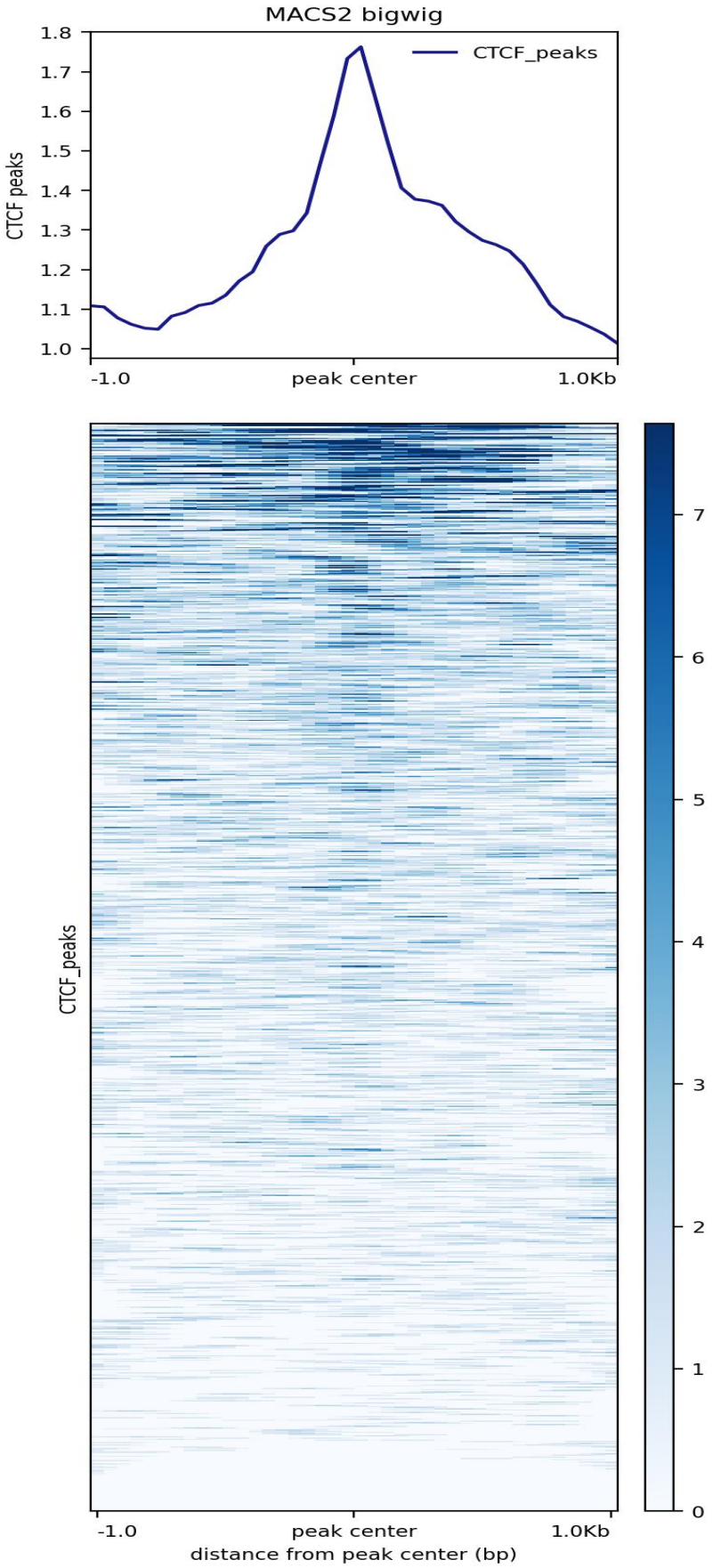

Figure 3: plotHeatmap output on CTCF

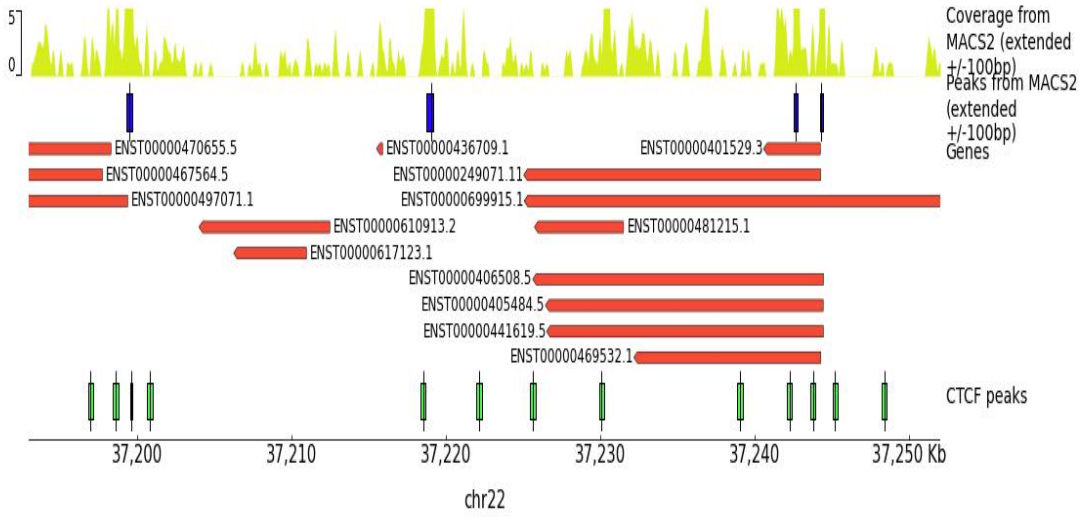

Figure 4: pyGenomeTracks output

Conclusion

I have gained knowledge of the fundamentals of ATAC-Seq data analysis through this project. A transposase (enzyme) known as Tn5 is used to treat the genome as part of the ATAC-Seq procedure, a technique to examine the accessibility of chromatin. By removing and re-inserting adapters, it identifies open chromatin areas for sequencing. I gained knowledge about data quality control through the literature review. Low-quality bases, adapter contamination, the right insert size, and PCR duplicates are things to watch out for (duplication level). If FastQC issues a warning regarding adapters or duplicate PCR products, I demonstrated how to remove them. I used Bowtie2 to map the reads, and I screened them for properly paired, high-quality, and reads that didn't map to the mitochondrial genome. With the use of MACS2, a technique for locating genomic enrichment regions, I discovered open chromatin regions (peaks). Using computeMatrix and plotHeatmap, I looked into the read coverage near the TSS. Last but not least, I used pyGenomeTracks to visualize the peaks and other useful tracks, like CTCF binding areas and hg38 genes. The ATAC-Seq experiment may have identified putative enhancer regions in open chromatin areas that were not overlapping with CTCF sites or TSS.

References

Adey, A., Morrison, H. G., Asan, Xun, X., Kitzman, J. O., Turner, E. H., Stackhouse, B., MacKenzie, A. P., Caruccio, N. C., Zhang, X., & Shendure, J. (2010). Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome biology*, *11*(12), R119. https://doi.org/10.1186/gb-2010-11-12-r119

Batut, B., Hiltemann, S., Bagnacani, A., Baker, D., Bhardwaj, V., Blank, C., Bretaudeau, A., Brillet-Guéguen, L., Čech, M., Chilton, J., Clements, D.,

Doppelt-Azeroual, O., Erxleben, A., Freeberg, M. A., Gladman, S., Hoogstrate, Y., Hotz, H. R., Houwaart, T., Jagtap, P., Larivière, D., … Grüning, B. (2018). Community-Driven Data Analysis Training for Biology. *Cell systems*, *6*(6), 752–758.e1. https://doi.org/10.1016/j.cels.2018.05.012

Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A., & Taylor, J. (2010). Galaxy: a web-based genome analysis tool for experimentalists. *Current protocols in molecular biology*, *Chapter 19*, Unit–19.10.21. https://doi.org/10.1002/0471142727.mb1910s89

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods*, *10*(12), 1213–1218. https://doi.org/10.1038/nmeth.2688

Corces, M. R., Trevino, A. E., Hamilton, E. G., Greenside, P. G., Sinnott-Armstrong, N. A., Vesuna, S., Satpathy, A. T., Rubin, A. J., Montine, K. S., Wu, B., Kathiria, A., Cho, S. W., Mumbach, M. R., Carter, A. C., Kasowski, M., Orloff, L. A., Risca, V. I., Kundaje, A., Khavari, P. A., Montine, T. J., … Chang, H. Y. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nature methods*, *14*(10), 959–962. https://doi.org/10.1038/nmeth.4396

Fu, Y., Sinha, M., Peterson, C. L., & Weng, Z. (2008). The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS genetics*, *4*(7), e1000138. https://doi.org/10.1371/journal.pgen.1000138

Green, B., Bouchier, C., Fairhead, C., Craig, N. L., & Cormack, B. P. (2012). Insertion site preference of Mu, Tn5, and Tn7 transposons. *Mobile DNA*, *3*(1), 3. https://doi.org/10.1186/1759-8753-3-3

Kia, Amirali & Gloeckner, Christian & Osothprarop, Trina & Gormley, Niall & Bomati, Erin & Stephenson, Michelle & Goryshin, Igor & He, Molly. (2017). Improved genome sequencing using an engineered transposase. *BMC Biotechnology*. 17. 10.1186/s12896-016-0326-1.

Litzenburger, U. M., Buenrostro, J. D., Wu, B., Shen, Y., Sheffield, N. C., Kathiria, A., Greenleaf, W. J., & Chang, H. Y. (2017). Single-cell epigenomic variability reveals functional cancer heterogeneity. *Genome biology*, *18*(1), 15. https://doi.org/10.1186/s13059-016-1133-7

Lucille Delisle, Maria Doyle, Florian Heyl, 2021 ATAC-Seq data analysis (Galaxy Training Materials). https://training.galaxyproject.org/training-material/topics/epigenetics/tutorials/atac-seq/tutorial.html Online; accessed Mon Sep 19 2022