# Supplementary material for single cell ATAC-Seq

Single-cell ATAC-Seq datasets were provided by 10xGenomics. We used Cell Ranger ATAC v 1.1.0 to process them. At first, we applied count procedure to process individual samples (old and young mice), later these were aggregated with dedicated aggregate Cell Ranger ATAC procedure with library depth normalization.

Count and aggregate command used:

cellranger-atac count --id=<id> --fastqs=<folder> --sample <sample> --reference refdata-cellranger-atac-mm10-1.1.0

cellranger-atac aggr --id=<id> --csv merged.csv --normalize=depth --reference=refdata-cellranger-atac-mm10-1.1.0

Overall pipeline consisted of the following steps: filtration and quality control, cells calling, normalization, feature selection, dimensionality reduction, t-SNE and UMAP projection, clustering, differential markers analysis including annotation and visualization, clusters visualization for genome browser, and data export to Single Cell Explorer The source code of the pipeline is available at <https://github.com/JetBrains-Research/sc-atacseq-explorer>.

Cell Ranger ATAC output file fragments.tsv.gz contains high-quality fragments which passed all the filters. We used this file as a main reference for UMI reads processing.

Fragments were filtered against blacklist regions to avoid possible PCR or alignment artifacts. Blacklist regions by Kundaje lab were downloaded from <http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/mm10-mouse/mm10.blacklist.bed.gz>. Any fragment overlapping this blacklist were omitted from analysis.

Raw reads profiles were visualized by deeptools v 3.1.3 as a part of a technical pipeline for NGS data using Snakemake workflow tool (<https://github.com/JetBrains-Research/chipseq-smk-pipeline>). Deeptools were used to create bigwig tracks for bulk alignment tracks. Both bulk reads, as well as individual cell populations, can be explored online in JBR Genome Browser viewer (<https://research.jetbrains.org/groups/biolabs/tools/jbr-genome-browser>) at <http://artyomovlab.wustl.edu/jbr/2019_gxfer1_DT1634_Denis>.

Cell calling was performed almost the same way as in Cell Ranger ATAC. On the figure **S1\_cell\_peaks** we showed distribution of UMI fragments overlapping any peak from raw data. We used thresholds of 200 fragments per UMI and less as noise, and 8000 and more as “barcode multiplets”. Removing noise and multiplets significantly improves both analysis robustness and overall computational costs. Noise and duplets quantities are shown on figure **S2\_cell\_calling** with distribution per UMI after duplets and noise clipping.

Even though some single-cell ATAC-Seq analysis approaches do not require predefined peaks regions (e.g. SnapATAC [ref]), correct bulk signal peak calling is a crucial part of most of the existing methods. Cell Ranger processing includes bulk signal peak calling with their own proprietary algorithm, without any additional parameters, and some meaningful biological peaks were missing in the result track. Cell Ranger ATAC BED file with bulk peaks is available in file **S0\_cell\_ranger\_peaks.bed**. To overcome this limitation, we reprocessed bulk signal with SPAN v 0.11.0.4882 [ref] to ensure the presence of these locations in downstream analysis. BAM alignment file possorted\_bam.bam was used for bulk peak calling, we used SPAN in semi-supervised mode, and with a small extent of annotations the following parameters were used:

--fdr 1e-10 gap=2. Peaks file contained 80914 peaks, which perfectly agrees with expected peaks number for ATAC-Seq dataset. Peaks are available in file **S0\_merged\_200\_1.0E-10\_2.peak.0E-10\_2.peak** in Encode broad peak format (ref here).

We applied standard Cell Ranger ATAC normalization to median count per UMI, followed by peak length normalization to make DNA accessible sites of different widths have the same impact on analysis. Sites with extremely high coverage (>99%) were ignored as they likely represent sequencing errors or housekeeping genes. Filtering out sites with low variation (<1%) improved the robustness of real biological signal. (ref SnapATAC and Seurat). Standard Cell Ranger ATAC method were applied for dimensionality reduction and clustering: Inverse Document Frequency normalization, IRLBA Singular Vector Decomposition, graph clustering, t-SNE projection, and differential analysis. Also, we used a novel UMAP method for data visualization. The only hyperparameters are the size of low dimensional space and number of clusters. D=15 was used as a default value used by Cell Ranger ATAC.

Clustering was performed on L2 normalized dataset after singular vector decomposition as well as t-SNE projection, this explains why some clusters may look mixed when displayed on a two-dimensional plot. N=15 was used, number of clusters was chosen to be sufficient for good data separation but still feasible for interpretation. Clusters were sorted by size in descending order. On figure **S3\_cluster\_sizes** we demonstrate cluster sizes. Overall UMI coverage per cluster is shown on figure **S4\_clusters\_umi**, all the barcode multiplets are successfully filtered out. On figures **S5\_tsne\_clusters\_split** and **S6\_tsne\_age\_split** we show individual clusters and age in projection coordinates. Table with cluster and age information for each UMI is available in table **S9\_umi\_cluster\_age.csv**. T-SNE coordinates are given in table **S10\_umi\_tsne.csv**.

We annotated peaks with corresponding genes with bedtools [ref] command:

bedtools closest -D b to associate arbitrary genome location to closest gene reporting distance to TSS (transcription start site) upstream or downstream with respect to strand. Genes markup v M22 was downloaded from the Gencode portal (ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode\_mouse/release\_M22/gencode.vM22.annotation.gtf.gz). With annotated peaks and clusters, we created a table of mean values per cluster, this table is available as table **S7\_clusters\_peak\_values**. This file contains information about chromosome, start and end offsets of each individual peak, closest gene, with distance relative to transcription start site. The rest columns contain information about average fragments number per UMI in cluster. This table was used to create figure 5b. We created dedicated bigwig file for each file using the pipeline and deeptools. Visualization files are available at: <http://artyomovlab.wustl.edu/publications/supp_materials/4Oleg/2019_sc_ATAC_seq_DT1634_Denis/bigwig/>

We applied procedure from Cell Ranger ATAC to compute clusters differential markers. This analysis was performed for each peak individually. We performed test of mean value in particular cluster versus others with Negative Binomial GLM. Multiple hypothesis testing was performed with Benjamini-Hochberg [ref] procedure. For each cluster we picked 50 top significant markers and plotted z-score among all the cells, the plot is shown on figure **S8\_markers\_z**.

Individual markers for each cluster are available at <http://artyomovlab.wustl.edu/publications/supp_materials/4Oleg/2019_sc_ATAC_seq_DT1634_Denis/markers/>. Each file marker\_N.bed is a BED6 file and contains information about chromosome, offsets, marker number, minus logarithm10 of adjusted p-value.

Motif enrichment was performed with Homer v4.10. 500 most significant markers were picked for each cluster. Homer was launched with the following parameters:

findMotifsGenome.pl markers\_N\_500.bed mm10 motifs\_N -size 200 -mask. For each cluster Homer produced motifs and associated transcription factors, some of those are shown in figure 5e.

The last but not the least step was data preparation for Single Cell Explorer <https://artyomovlab.wustl.edu/sce/?token=2019_scatacseq_Denis>. Single cell explorer requires the following files: data.h5 – information about mean UMI per peak per cluster in binary HDF5 format, dataset.json – dataset description and metainformation, exp\_data.json – all UMI and peaks identifiers with total fragments per UMI, markers.json – results of differential clusters analysis, and plot\_data.json – t-SNE coordinates for visualization.