Integrative Cellular Biology & Bioinformatics

CHAIR FOR CLINICAL BIOINFORMATICS
SAARLAND UNIVERSITY

Project 2

In this project you will analyze chromatin accessibility data from human bone marrow and peripheral blood cells. It comprises single-cell ATAC-seq data from 3 samples which are matched to the scRNA-seq data you analyzed in project 1.

The paper that analysed the dataset that also included this data was published by Granja et al. (2019) and the original data can be found here.

Deadline: 03.01.2022 23:59

The same rules apply as for the first project.

Tutorials:

- If you have any Problems, please contact Omar (omla00001@stud.uni-saarland.de) or visit the tutorials.
- During the project, there will be tutorials on some of the tasks.

Publishing of the project	29.11.2021	Introduction
	06.12.2021	Peak-Calling/ Gene-Activity/TF motifs
	13.12.2021	Differential Accessibility and Footprints
	20.12.2021 - 03.01.2022	- No tutorial -
Deadline project	03.01.2022	

Please start early, as there will be no tutorials during the last two weeks of the project!

Submission:

- You will have to submit **one** tar.gz file that includes
  - The code
  - PDF file including all images and responses to the questions
- The code must be well commented and must run without an error to obtain any points
- If you use any other sources for your answers, don't forget to give the reference

Introduction in ArchR

We will use the programming language R and the ArchR software package. You can install ArchR by following the guidelines on this website: https://www.archrproject.com/

The lectures on scATAC analysis and the documentation of the ArchR package provide useful hints for conducting the analysis.

Download the data

You can download the dataset for this project under the following links (see page 8):

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4138888
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4138890
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4138891

Task 1: Preprocessing and quality control (5P)

Task 1.1: Set up the environment

Download the data from the given website. Set the reference genome used in this project to: 'hg19'.

Optionally: You can also use parallel processing (requires package parallel). On windows, parallelization is not possible.

Task 1.2: Read the data into an appropriate data structure and apply filtering

Read the downloaded fragment files into arrow files.

Use lenient filtering on number of fragments and TSS enrichment (e.g., filterFrags > 500, filterTSS > 4). Make sure to compute a tiling matrix with appropriate window size (1000 bp or 500 pb).

You will come back to this step in Task 1.6 to evaluate if stricter thresholds are necessary.

Hint: For the filtering parameters, have a closer look at all the parameters that can be given to the function used for creating the error files.

Task 1.3: Identify doublets

Perform Doublet identification. If you observe doublets, remove them from the data and report your findings.

Task 1.4: Collect all samples into a joint data structure

Create one project that includes all samples. Use table 1 below to add the sample metadata to the project.

Inspect the cell metadata:

- How many cells does your project include?
- What is the median TSS-value and the median of the number of fragments?
- What are the dimensions of your dataset?

Task 1.5: Quality control

Report/plot the following measurements:

- How many cells do you have for each sample?
- Plot the fragment length distribution of all samples in a single plot
- Plot the distribution of TSS enrichment scores in each sample
- For each sample, plot number of fragments vs TSS enrichments

Inspect the plots and report your findings. Do you observe outlier samples/samples with low quality?

Note: If done correctly, some of these plots should have automatically been generated in the creation of arrow files.

Task 1.6: Filter the dataset

Based on your observations in the previous task, apply stricter filtering based on the number of fragments, TSS enrichment score, and remove doublets.

Have a look again at the questions from task 1.4 and answer them again for the new parameters.

Task 2: Dimensionality Reduction (5P)

Task 2.1: Iterative LSI

Apply dimensionality reduction using ArchR's iterative latent semantic indexing approach.

Why do we use LSI and not PCA for the dimensionality reduction of scATAC-data?

Task 2.2: UMAP with sample annotation and QC metrics

Compute data-visualization with UMAP coordinates. Create plots of the cells in UMAP space. Color each UMAP-plot by one of the following variables: Sample, TSS enrichment and number of fragments.

Task 2.3: Dealing with batch effects

Do you observe batch effects between cells in the different samples? If so, correct them. Create UMAP representations and inspect. Have the batch effects been resolved?

Task 3: Clustering (3P)

Apply Louvain clustering to all cells in the dataset. Create a UMAP plot coloring by clusters.

How many cells does each cluster contain? What are the sample proportions in each cluster?

Task 4: Peaks (8P)

Task 4.1: Peak calling

Prepare the data for peak-calling. Explain what you did and why this is necessary.

Compute a joint peak set, grouping by clusters and add the peak Matrix to your project.

Task 4.2: Cluster marker peaks

Task 4.2.1: Identify cluster-specific peaks using reasonable cutoffs. Which parameters are used for which variable and what do they describe?

Task 4.2.2: Create a Heatmap showing accessibility in the marker peaks

Task 4.2.3: Create browser plots around the following genes: CD19, CD34, CD3D, CD69

Task 5: Gene activity (3P)

Task 5.1: Compute gene activity scores using chromatin accessibility

Note: depending on the parameters that you used this might have already been done during data import

Task 5.2: Identify Marker genes

Identify cluster marker genes. Explain the parameters and cut-off values you used.

Task 5.3: Using MAGIC

Plot UMAP embeddings for the first 5 marker genes found in the previous task. Plot the visualization with and without smoothing using Magic.

What is the main idea behind MAGIC (1-2 sentences).

Task 6: Transcription Factor motif activity (8P)

Task 6.1: Compute TF motif activity

Pick a suitable TF motif annotation (database).

Questions:

1. Which annotation did you choose?
2. How was the annotation obtained?

Task 6.2: Plot UMAP embeddings for marker TFs

Which are the TFs that are most variable? Plot the top 2 marker motifs in a separate UMAP plot.

Task 6.3: Motif activity

Plot distribution of motif activity of marker TFs in the different clusters

Task 7: Assign cell types to clusters (6P)

Consider the gene activity and TF activity of the marker genes shown in table 2. Which clusters correspond to each cell type?

Add the cell type annotation to the dataset's metadata. Plot the cell type annotation into a UMAP embedding.

Task 8: Integration with gene expression (7P)

Task 8.1: Data integration

Integrate the annotated scRNA-seq data from Project 1. Link gene expression to accessibility. Plot the scATAC-seq data in the UMAP space and overlay it with gene expression levels for the highlighted marker genes in table 2.

Task 8.2: Correlation Coefficients

Compute the correlation coefficients between the gene expression from the scRNA-seq data and the gene activity from the scATAC-seq data.

Which genes exhibit the highest / lowest agreement?

Task 8.3: Cluster labels from gene expression

Assign cell type labels using the integrated gene expression data and rename the clusters accordingly. Compute a confusion matrix of ATAC clusters with RNA-inferred cell type labels. Report what you observe in terms of agreement.

Task 8.4: Peak-gene linkage

Task 8.4.1: Compute peak-gene linkage using gene expression and chromatin accessibility in nearby enhancers.

Task 8.4.2: Plot side-by side heatmaps of candidate regulatory element accessibility and gene expression

Task 8.4.3: Identify potential enhancers for the highlighted marker genes from Table 2. Plot the links for the marker-genes using genome browser views.

How many potential enhancers do you observe for each gene?

Task 9: Differential accessibility (5P)

Task 9.1: Differential peak accessibility

Compute differential peak accessibility between T cells and monocytes. Create MA and volcano plots of differential accessibility.

Task 9.2: TF motif enrichment

Compute TF motif enrichment in differentially accessible peaks.

Which TFs are enriched in peaks specific to T cells and monocytes respectively?

Plot TF motif activities for the identified TFs as a heatmap or in UMAP space.

Task 10: (Bonus) TF footprinting (+3P)

Task 10.1: Obtaining footprints on feature set

Obtain footprints for the top motifs from your enrichment analysis.

Task 10.2: Normalization for Tn5 Bias

Task 10.2.1: Perform a normalization for the Tn5 Bias.

Questions:

1. Explain the need to correct for Tn5 bias.
2. Which normalization method did you use?

Task 10.2.2: Plot the aggregate footprints for T cells and monocytes. Describe and explain what you see.

Task 11: (Bonus) Co-accessibility (+2P)

Task 11.1: Co-accessibility of peaks

Compute the co-accessibility of the peaks.

Task 11.2: Identify potential enhancers for marker genes

Again, use the highlighted marker genes from Table 2 and plot the linkage of peaks to TSS of these genes.

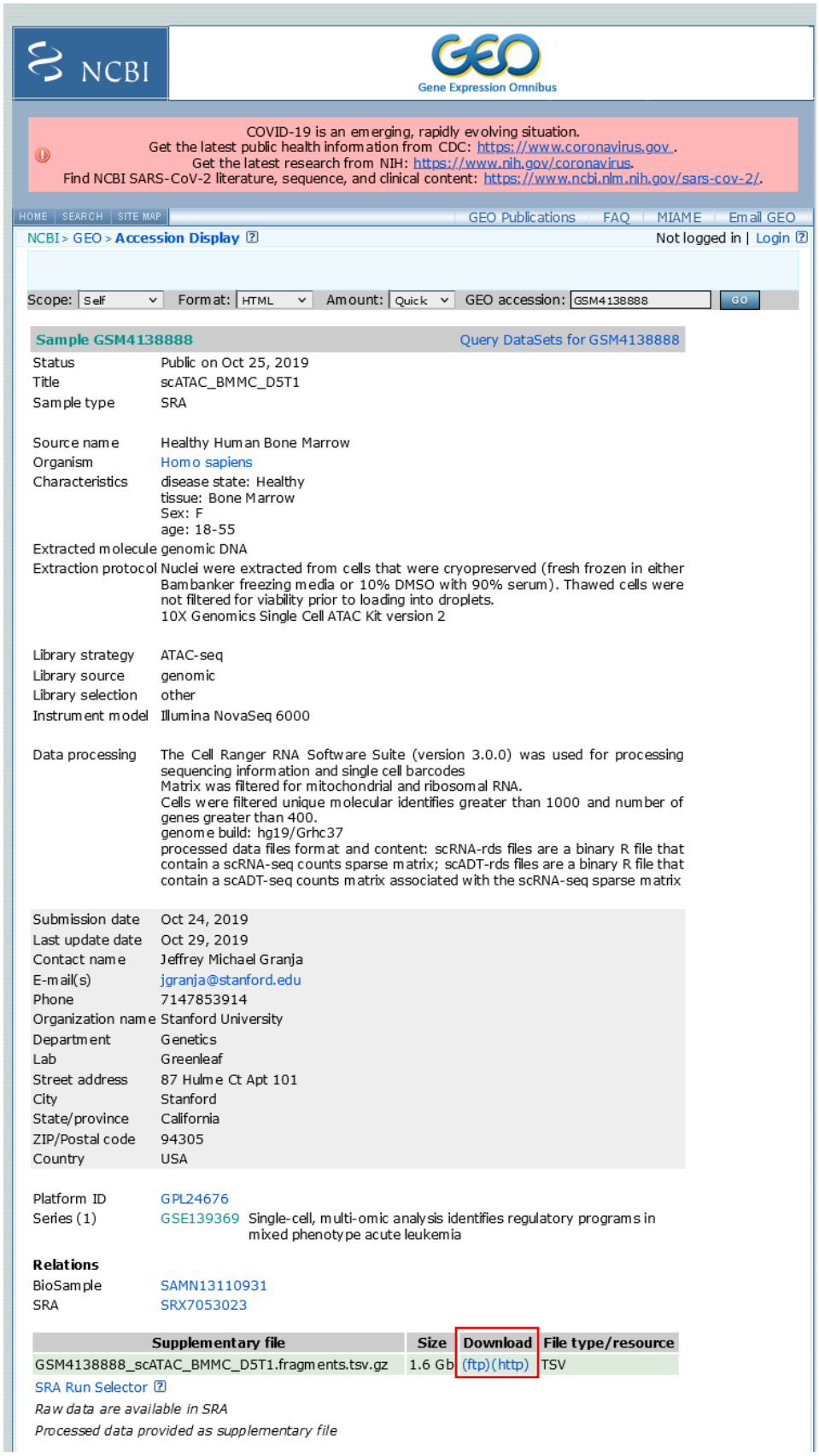
How many potential enhancers do you observe for each gene?

Table 1:

Sample	Type	Donor	Sex
BMMC_D5T1	BMMC	D5	F
CD34_D7T1	CD34	D6	M
CD34_D8T1	CD34	D7	M

Table 2:

Early Progenitor	**CD34**
Erythroid	**GATA1**
B cells	PAX5, MS4A1, MME, **CD19**
T cells	**CD3D**, CD8A, CD4, TBX21, ILR7
Monocytes	**CD14**, MPO, CEBPB
NK	**TBX21**
LMPP	**CD38**, CD52, CSF3R
GMP	ANXA1, **AP351**, APLP2
PRO-B	CD19, CD34



NCBI

Gene Expression Omnibus

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/.

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > **Accession Display** Not logged in | Login

Scope: Self Format: HTML Amount: Quick GEO accession: GSM4138888 GO

Sample GSM4138888 Query DataSets for GSM4138888

Status	Public on Oct 25, 2019
Title	scATAC_BMMC_D5T1
Sample type	SRA
Source name	Healthy Human Bone Marrow
Organism	Homo sapiens
Characteristics	disease state: Healthy tissue: Bone Marrow Sex: F age: 18-55
Extracted molecule	genomic DNA
Extraction protocol	Nuclei were extracted from cells that were cryopreserved (fresh frozen in either Bambanker freezing media or 10% DMSO with 90% serum). Thawed cells were not filtered for viability prior to loading into droplets. 10X Genomics Single Cell ATAC Kit version 2
Library strategy	ATAC-seq
Library source	genomic
Library selection	other
Instrument model	Illumina NovaSeq 6000
Data processing	The Cell Ranger RNA Software Suite (version 3.0.0) was used for processing sequencing information and single cell barcodes Matrix was filtered for mitochondrial and ribosomal RNA. Cells were filtered unique molecular identifies greater than 1000 and number of genes greater than 400. genome build: hg19/Grhc37 processed data files format and content: scRNA-rds files are a binary R file that contain a scRNA-seq counts sparse matrix; scADT-rds files are a binary R file that contain a scADT-seq counts matrix associated with the scRNA-seq sparse matrix
Submission date	Oct 24, 2019
Last update date	Oct 29, 2019
Contact name	Jeffrey Michael Granja
E-mail(s)	jgranja@stanford.edu
Phone	7147853914
Organization name	Stanford University
Department	Genetics
Lab	Greenleaf
Street address	87 Hulme Ct Apt 101
City	Stanford
State/province	California
ZIP/Postal code	94305
Country	USA
Platform ID	GPL24676
Series (1)	GSE139369 Single-cell, multi-omic analysis identifies regulatory programs in mixed phenotype acute leukemia

Relations

BioSample	SAMN13110931
SRA	SRX7053023

Supplementary file	Size	Download	File type/resource
GSM4138888_scATAC_BMMC_D5T1.fragments.tsv.gz	1.6 Gb	(ftp)(http)	TSV

SRA Run Selector

Raw data are available in SRA

Processed data provided as supplementary file