January 2020

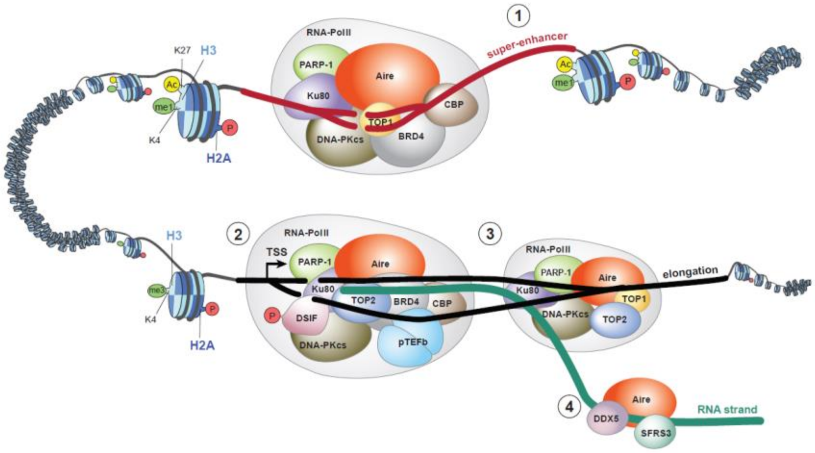
An Introduction to Deep-Sequencing Data Analysis

Exercise 9: ATAC-seq analysis

Dena Leshkowitz and Bareket Dassa

In this assignment, you will practice several steps common to ChIP-seq or ATAC-seq data analysis. You will run an ATAC-seq analysis pipeline using the UTAP workflow, and examine its output at the different processing steps, starting from reads to peak calling. Among all possible downstream analyses, we will practice peak annotation with an online web program (GREAT).

The data that you will analyze was obtained from a research on the Aire transcription factor in Medullary thymic epithelial cells which are involed in the selection of effector and regulatory T cells (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5310976/). Aire complex binds and activates regions on the promotor or enhancer.



Note- running the pipeline may take a few hours.

Part I: Set-up an ATAC-seq analysis

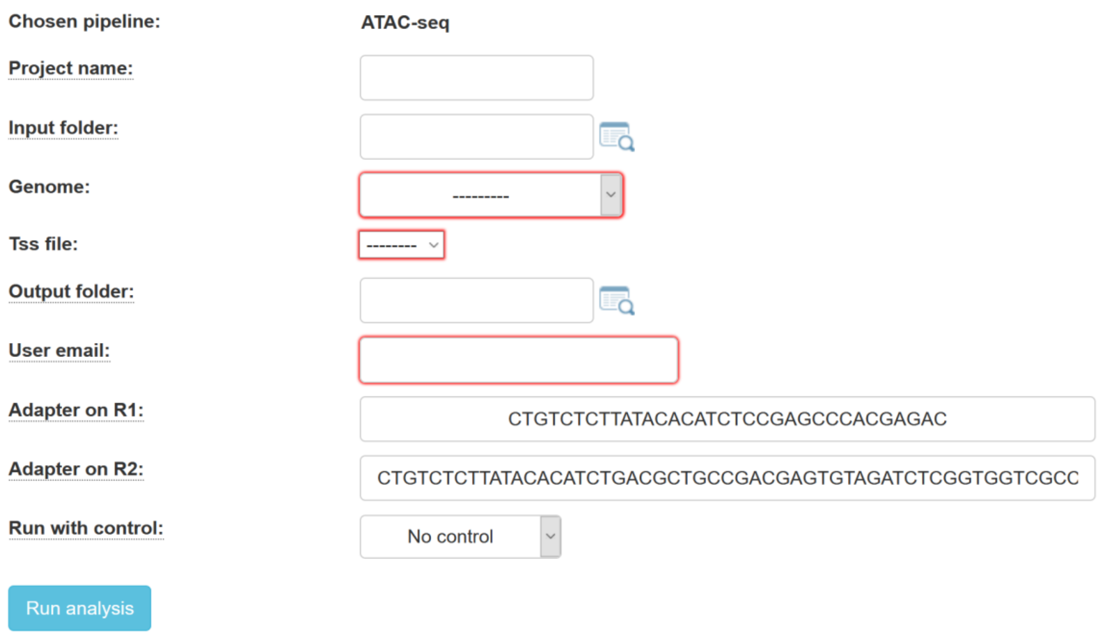
A. Login to "UTAP: User-friendly Transcriptome Analysis Pipeline" at:
   http://ngspipe.wexac.weizmann.ac.il:7000
   Use your class Username and password.
   (The page is accessible only inside the Weizmann network, and is compatible with Firefox and Chrome browser, but not with InternetExplorer).

B. Click "Run pipeline" from the top menu. Choose "ATAC-seq" analysis.

Choose a project name, for example “Class1-exercise”

C. **Select the input folder: The raw fastq files are saved in sub-folders within the folder “/course_2019/ATAC-Seq_aire/”.** Select the entire folder as input. Once you select the input folder, a list of all samples will be displayed.
   In this experiment, input samples are paired-end sequences, duplicates of WT and duplicates of the KO Aire gene.

   Your **output** folder will be selected automatically, **it is important NOT to change it**!



Chosen pipeline:	ATAC-seq
Project name:	
Input folder:	
Genome:	---------
Tss file:	--------
Output folder:	
User email:	
Adapter on R1:	CTGTCTCTTATACACATCTCCGAGCCCACGAGAC
Adapter on R2:	CTGTCTCTTATACACATCTGACGCTGCCGACGAGTGTAGATCTCGGTGGTCGCC
Run with control:	No control

D. **Additional set ups:**
   Select the **reference genome** (Mus musculus (mm10)) and **Tss annotation** of +/-500 for the reads alignment. Use the default adapter sequences.
   This analysis runs without a control.

E. **Change the email to your personal email.**

F. **Run the pipeline** (this may take a few hours)

Finally, submit the run for analysis. Once the analysis is completed, you will be notified by email. You can also check on the status of the run from the “User Datasets” at the UTAP top menu.

Part II: Questions

For this exercise start with creating a folder called “**Exercise_9**” in your home directory. Save the answers to this exercise in a file inside this folder.

There is **no report** for this pipeline, but once the run has ended, you will explore the output files created in the output folder (mount to WEXAC and open your output folder within the “utap_data”).

See the following link for additional information on the pipeline.

1. In ATAC-seq we apply several processing steps for selecting uniquely mapped reads, filter out duplicate reads and reads mapped to mitochondrial genome. This step removes large proportions of the raw reads.

   Follow the processing of sample `WT_1`. The initial number of paired-end reads in this sample (which passed the QC) was 74,370,986.

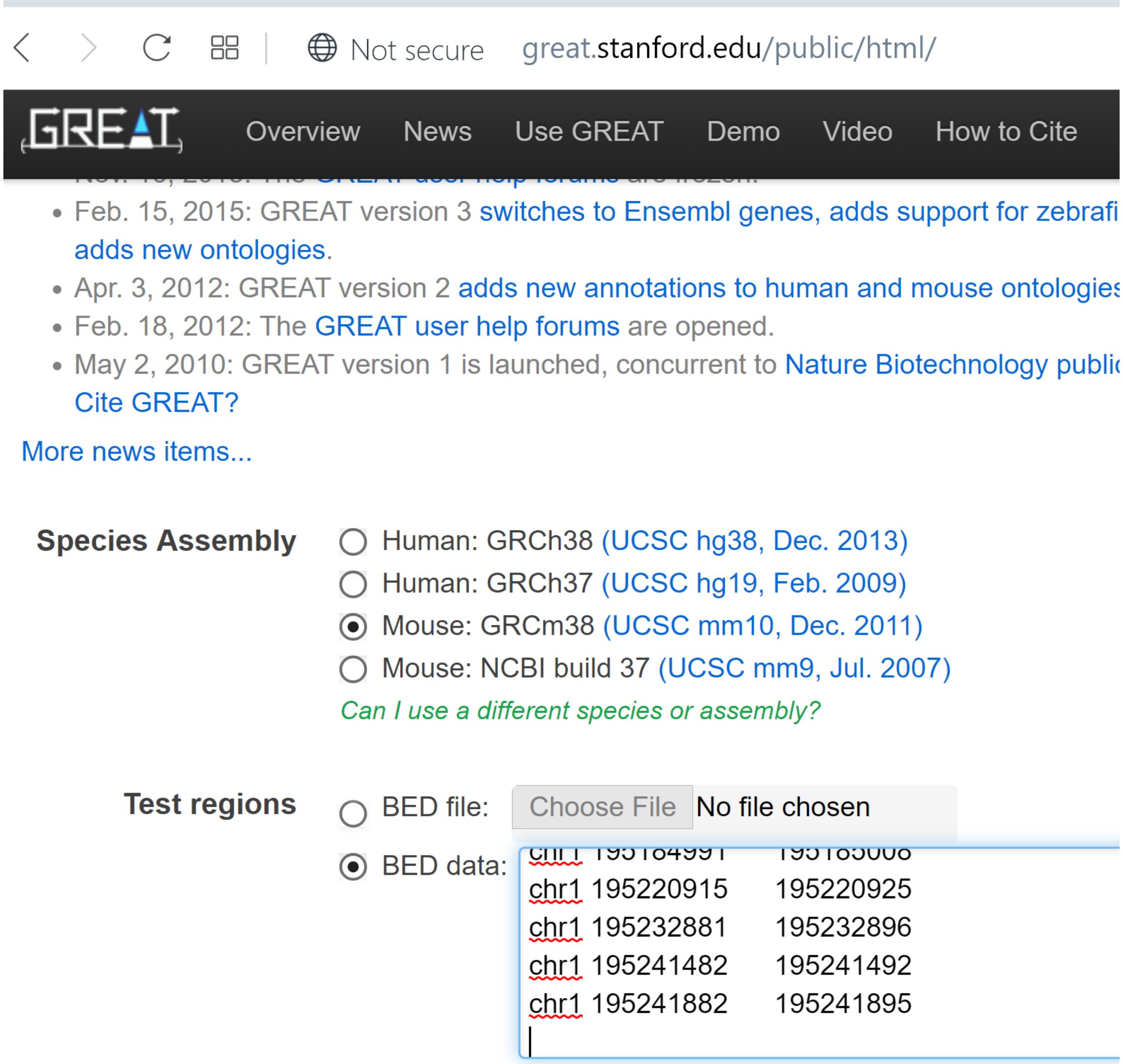
   a. How many reads were paired and uniquely aligned? This information is found in the statistics text file (ending with “.stat”) in folder “4_mapping”.
   b. How many reads remain after removal of duplicates, mitochondrial DNA, and selection of nucleosome-free fragments? Use the statistics file ending with “_nucl_free.statistics” in folder “6_nucleosome_free”.
2. Fragment size distribution is a good indication for the quality of an ATAC-seq experiment. Look at the plot of fragment size distribution, created by picard-tools for sample `WT_1`. This file is in the folder “ngs_plot”, ending with “_picard.pdf”.
   a. Why are there two main distributions of fragment sizes?
   b. Which distribution corresponds to the nucleosome free?
3. In order to compare peaks called by MACS2 from the different samples, we merged the peaks called in each sample into one file: “/course_2019/ATAC-Seq_aire/report_multi_intersectBed.xlsx”. See the headers for more information.

   For simplicity, only peaks from chromosome 1 were saved.

   a. Use the Excel filters to answer how many peaks appear in BOTH duplicates of the WT and not in any of the KO?
      Create a text file (tab delimited) with the first three columns and save it as WT_unique.txt in folder called “**Exercise_9**”.
      Use the Excel filters to answer how many peaks appear in BOTH duplicates of the KO and not in any of the WT?
      Create a text file (tab delimited) with the first three columns and save it as KO_unique.txt in folder called “**Exercise_9**”.

4. In order to associate between the genomic regions of the peaks with their putative target genes, you will use the tool GREAT (Genomic Regions Enrichment of Annotations Tool).

   Paste the chr, start and end filtered rows which you saved in 3a as an input, as shown below.



- Feb. 15, 2015: GREAT version 3 switches to Ensembl genes, adds support for zebrafi adds new ontologies.
- Apr. 3, 2012: GREAT version 2 adds new annotations to human and mouse ontologies
- Feb. 18, 2012: The GREAT user help forums are opened.
- May 2, 2010: GREAT version 1 is launched, concurrent to Nature Biotechnology public Cite GREAT?

More news items...

Species Assembly

○ Human: GRCh38 (UCSC hg38, Dec. 2013)
○ Human: GRCh37 (UCSC hg19, Feb. 2009)
● Mouse: GRCm38 (UCSC mm10, Dec. 2011)
○ Mouse: NCBI build 37 (UCSC mm9, Jul. 2007)
Can I use a different species or assembly?

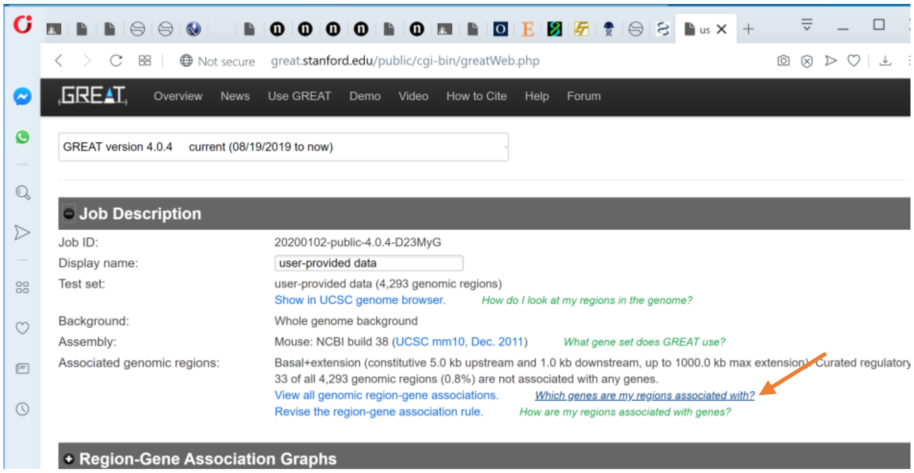
Test regions

○ BED file: No file chosen
● BED data:

```
chr1 195184991     195185008
chr1 195220915     195220925
chr1 195232881     195232896
chr1 195241482     195241492
chr1 195241882     195241895
```

a. Use the result graphs to explain where are most of the peaks found relative to the TSS of genes (is the coverage symmetric)?

b. Explain how does GREAT associates genes to regions? (use the information from the link below)



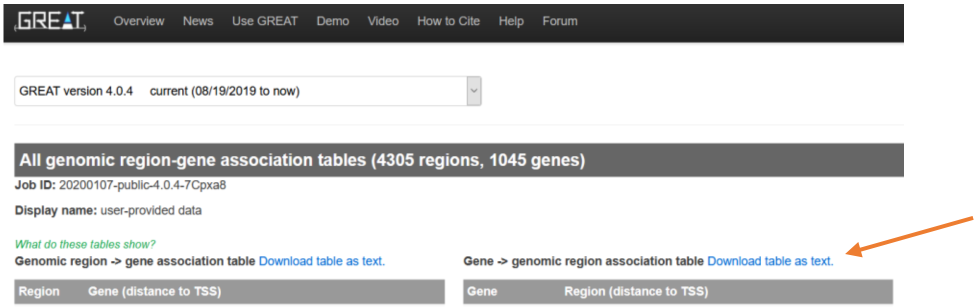
GREAT version 4.0.4 current (08/19/2019 to now)

Job Description

Job ID: 20200102-public-4.0.4-D23MyG
Display name: user-provided data
Test set: user-provided data (4,293 genomic regions)
Show in UCSC genome browser. *How do I look at my regions in the genome?*
Background: Whole genome background
Assembly: Mouse: NCBI build 38 (UCSC mm10, Dec. 2011) *What gene set does GREAT use?*
Associated genomic regions: Basal+extension (constitutive 5.0 kb upstream and 1.0 kb downstream, up to 1000.0 kb max extension). Curated regulatory
33 of all 4,293 genomic regions (0.8%) are not associated with any genes.
View all genomic region-gene associations. *Which genes are my regions associated with?*
Revise the region-gene association rule. *How are my regions associated with genes?*

Region-Gene Association Graphs

c. Download the gene annotation information from "Global Controls" > "View all region-gene associations", and then see below:



GREAT version 4.0.4 current (08/19/2019 to now)

All genomic region-gene association tables (4305 regions, 1045 genes)

Job ID: 20200107-public-4.0.4-7Cpxa8
Display name: user-provided data

What do these tables show?

Genomic region -> gene association table Download table as text.

Region	Gene (distance to TSS)

Gene -> genomic region association table Download table as text.

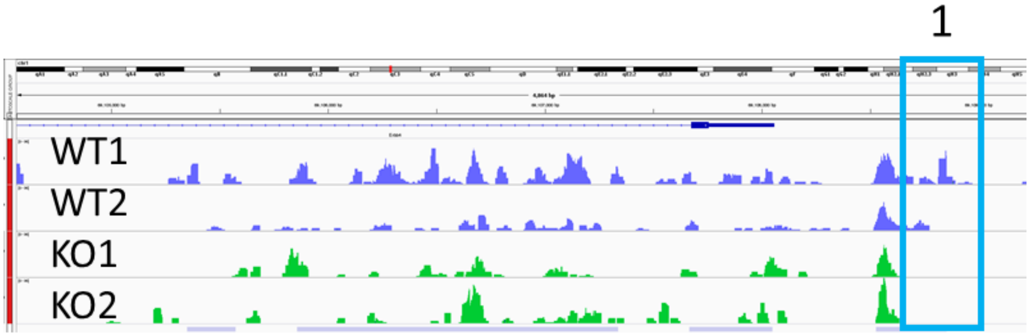
Gene	Region (distance to TSS)

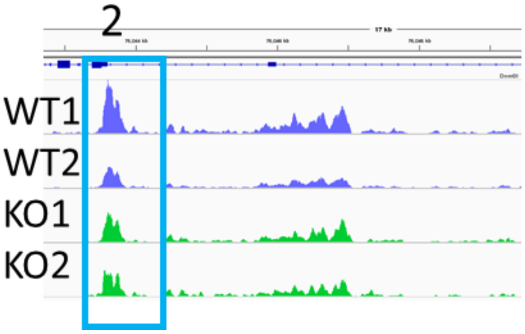
Repeat for the filtering done in 3b.

Using a tool for creating Venn diagrams: Venny (https://bioinfogp.cnb.csic.es/tools/venny/index2.0.2.html), compare the overlap of gene names for both (WT unique and KO unique)

a. How many genes are shared?
b. How many genes are unique?
c. Copy the venny plot in your answer.
d. Explain how it is possible that there are shared genes?

5. Resulting peaks are often visualized with IGV.
Below are two example of nucleosome-free coverage tracks. In both examples, the marked regions are more accessible in the WT than in the KO. Suggest how it is possible to detect these type of peaks?





Good luck!