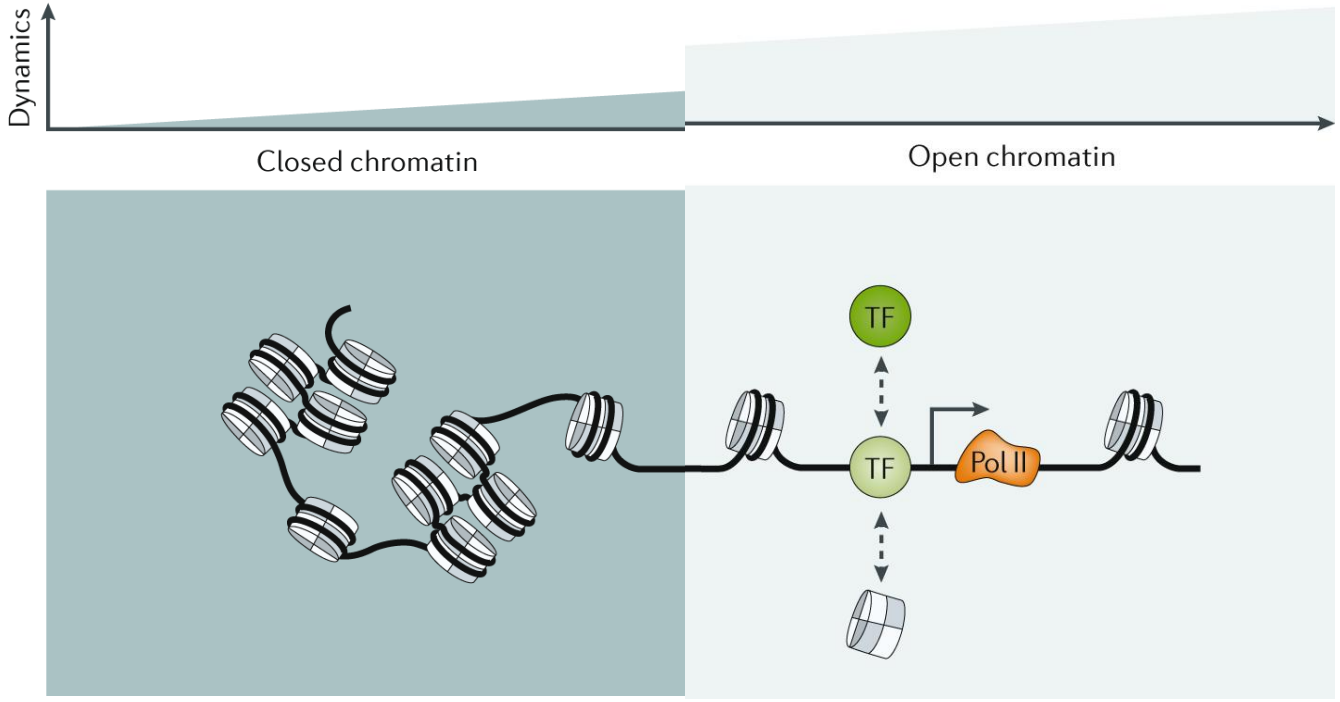
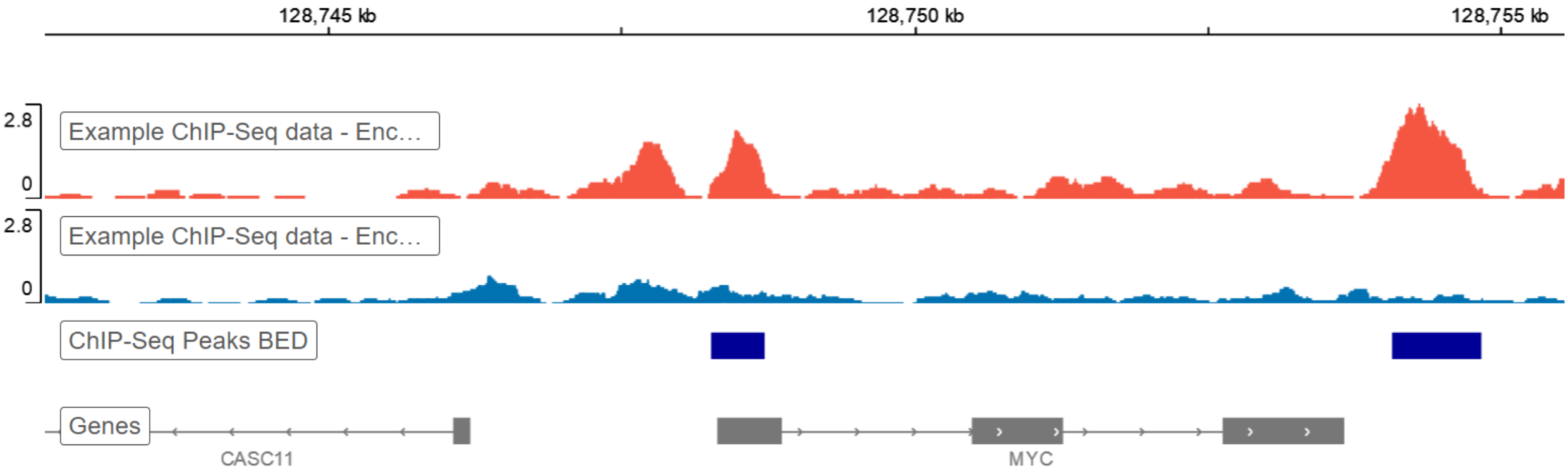
HW-4 (ENCODE) Guide

What is this all about?

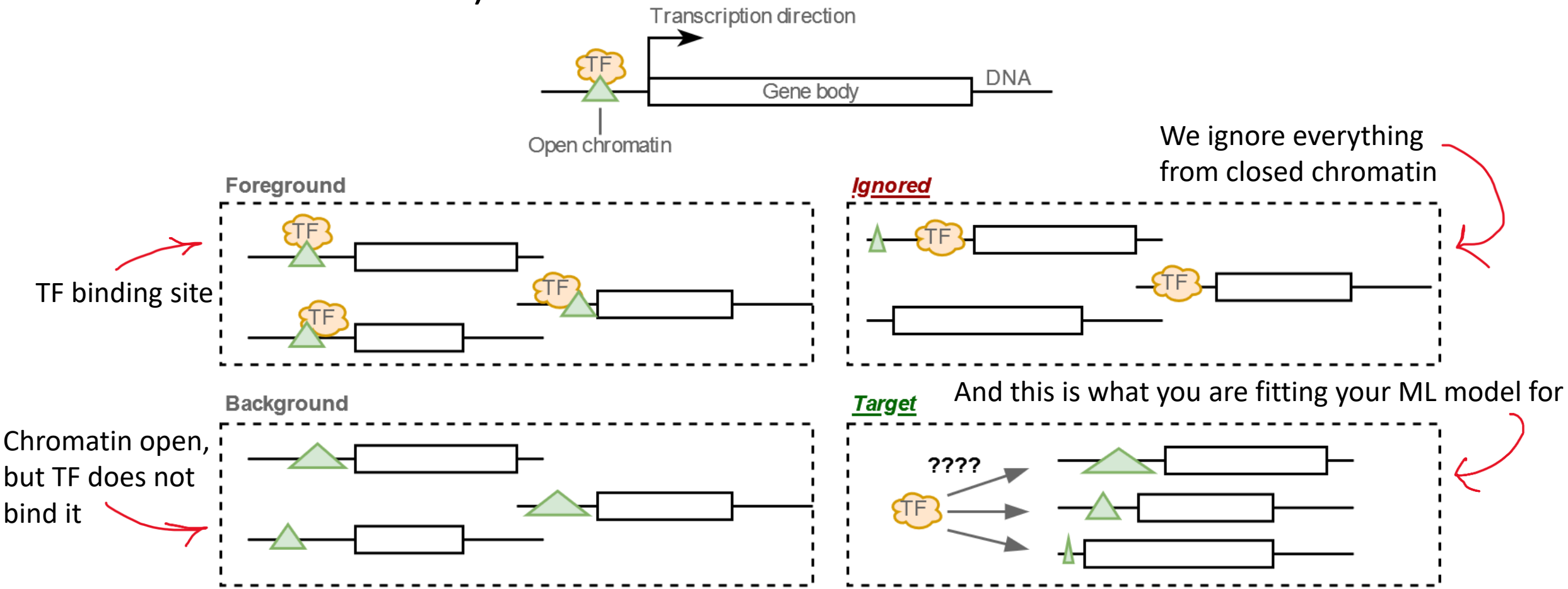
- Genome is a mess, but actually organized mess
- Some parts of the genome are available for proteins, some are not (open/closed chromatin)
- ATAC-seq experiment can determine, which parts of the chromatin were open (basically gives you coordinates)
- If it’s open, then protein (Transcription Factors, TF) can bind it in specific location, called binding sites



- Binding sites are determined by motif (genome sequence)
- ChIP-seq experiment can determine binding sites for specific protein (basically gives you coordinates too, the location of peaks on the image below)



- The task is: given some genome region from open chromatin, predict whether your target TF will bind it or not
- This is what we are doing in seminar 7. In HW-4 & HW-5 you accomplish the same task, but with 3 distinct TFs (i.e. multiclassification).

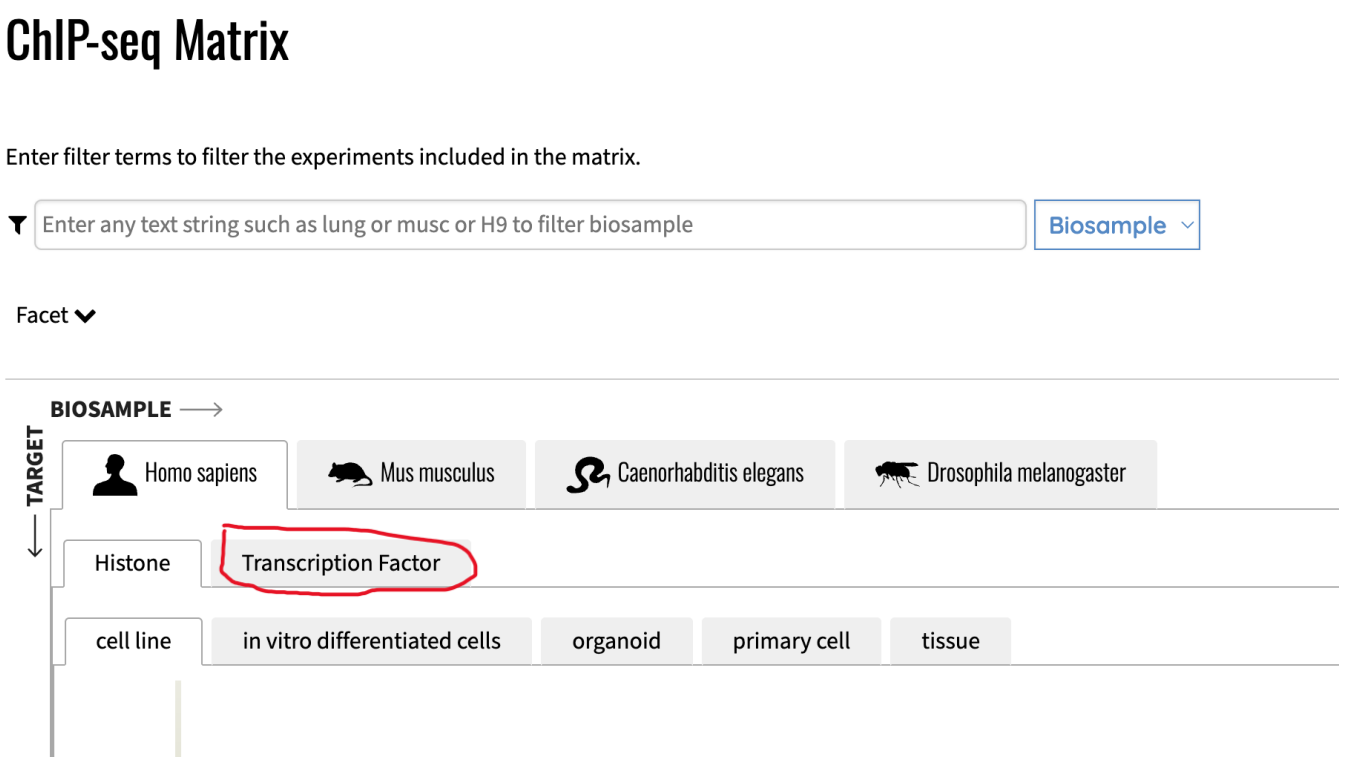


For this homework you will need to:

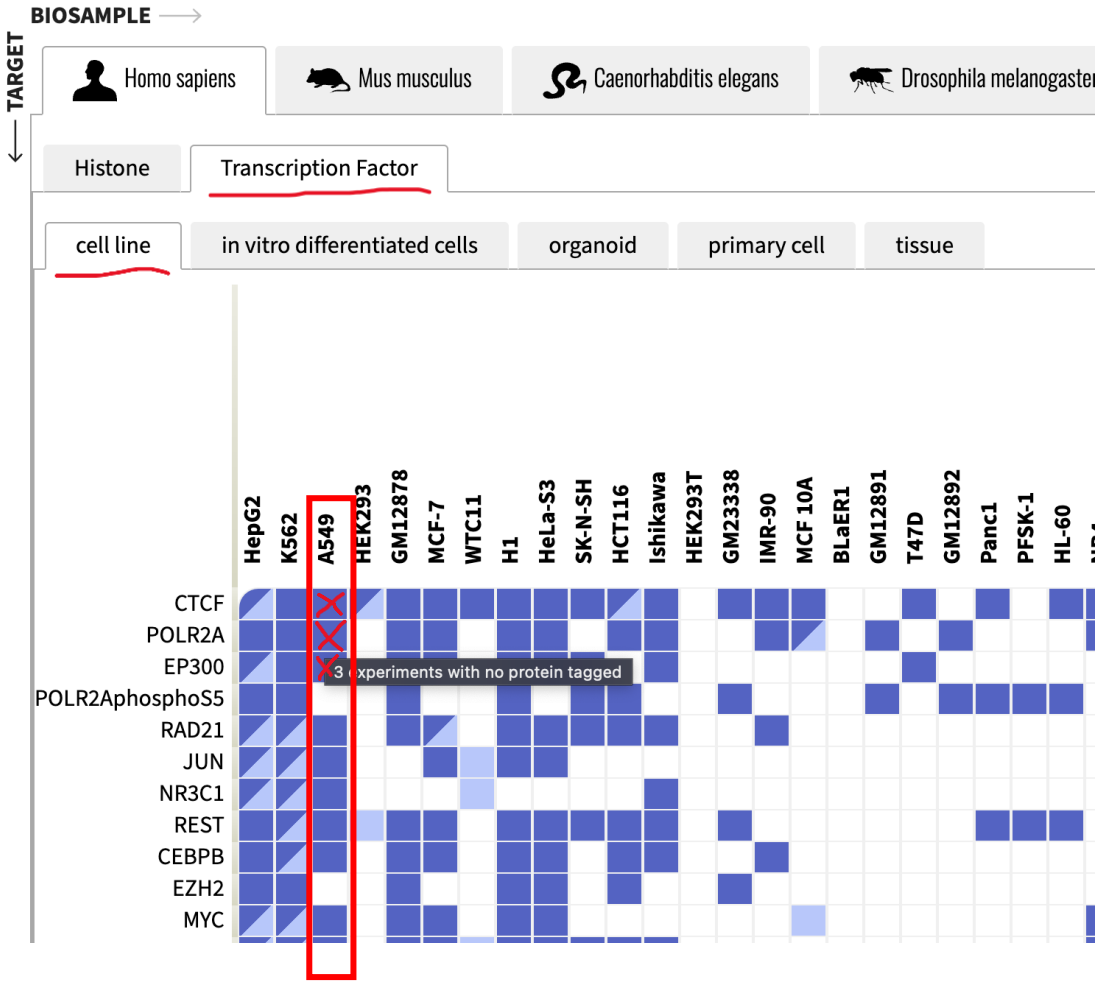
- Choose a **single cell line** in available ENCODE experiment
- Choose a **single ATAC-seq** experiment **with this cell line**
- Choose **three Transcription Factors (TF)** experiments, and download a **ChIP-seq experiment for each from the same cell line**

Choosing cell line and TFs

- Go to https://www.encodeproject.org/
- On the main page, find ChIP-seq experiments
- You will see ChIP-seq matrix
- In the matrix, choose TRANSCRIPTION FACTORS (histones are set by default):

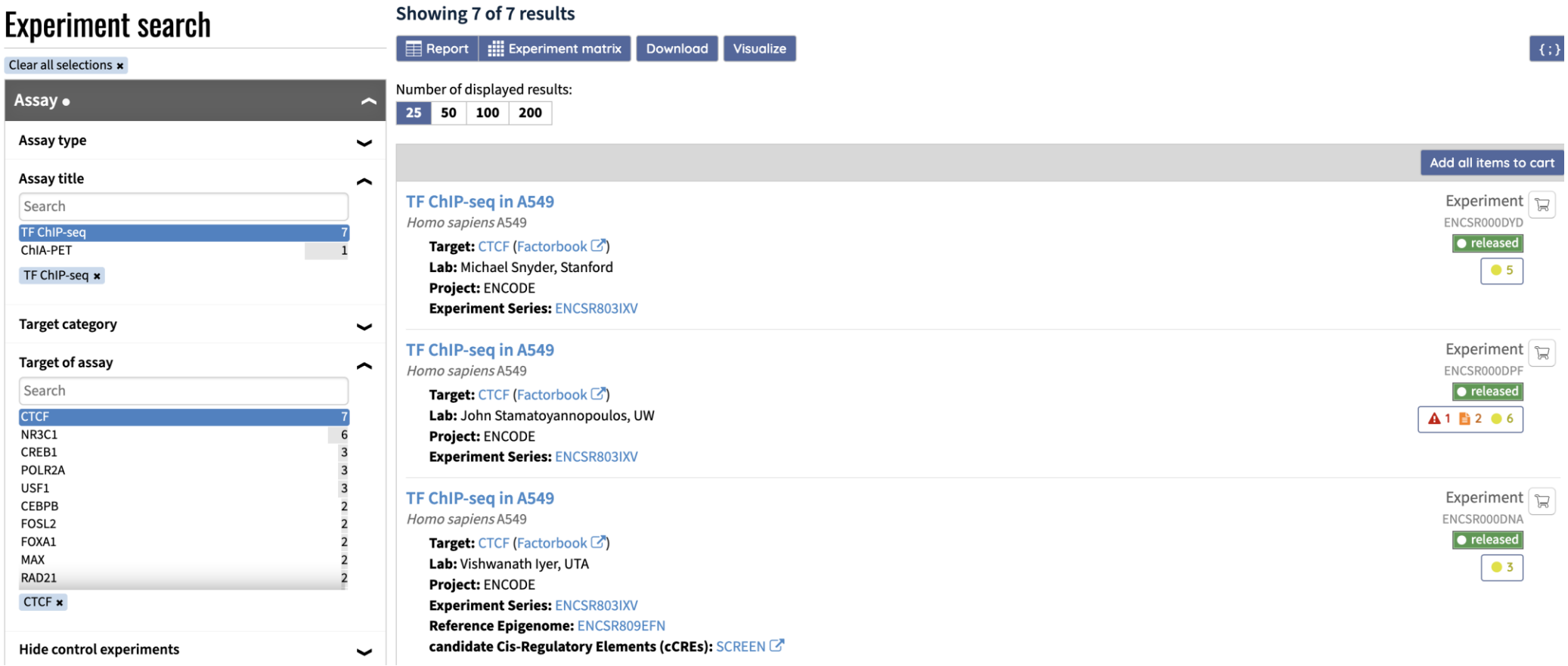


- Now, choose any organism
- Choose ‘cell line’ under ‘histone’ (set by default)
- In the matrix, the columns correspond to cell line names (e.g. A549). Choose ONE.
- For this column, choose three rows/TFs (e.g. CTCF, POLR2A, EP300)
- Be creative and choose different ones
- You can see how many experiments are there by hovering your mouse over the matrix cell

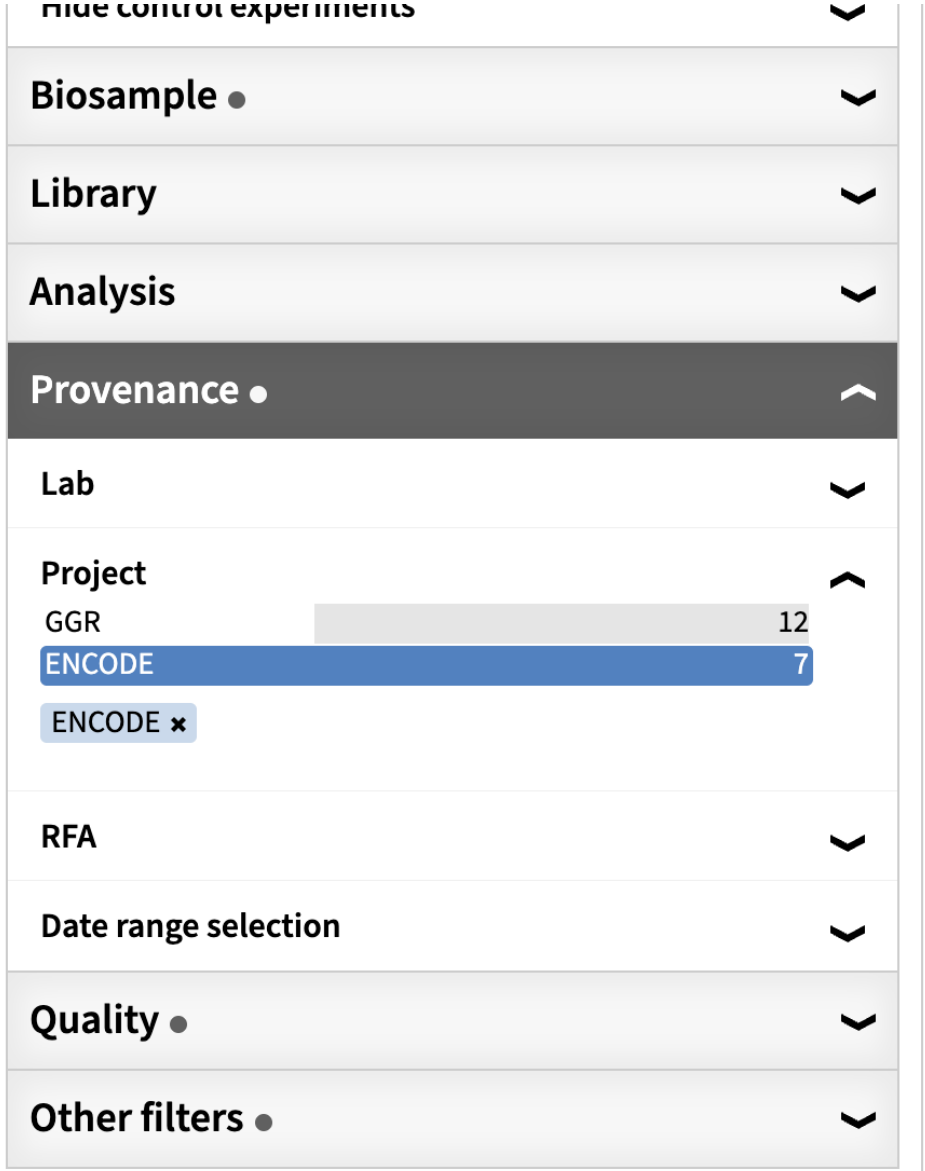


Downloading ChIP-seq data

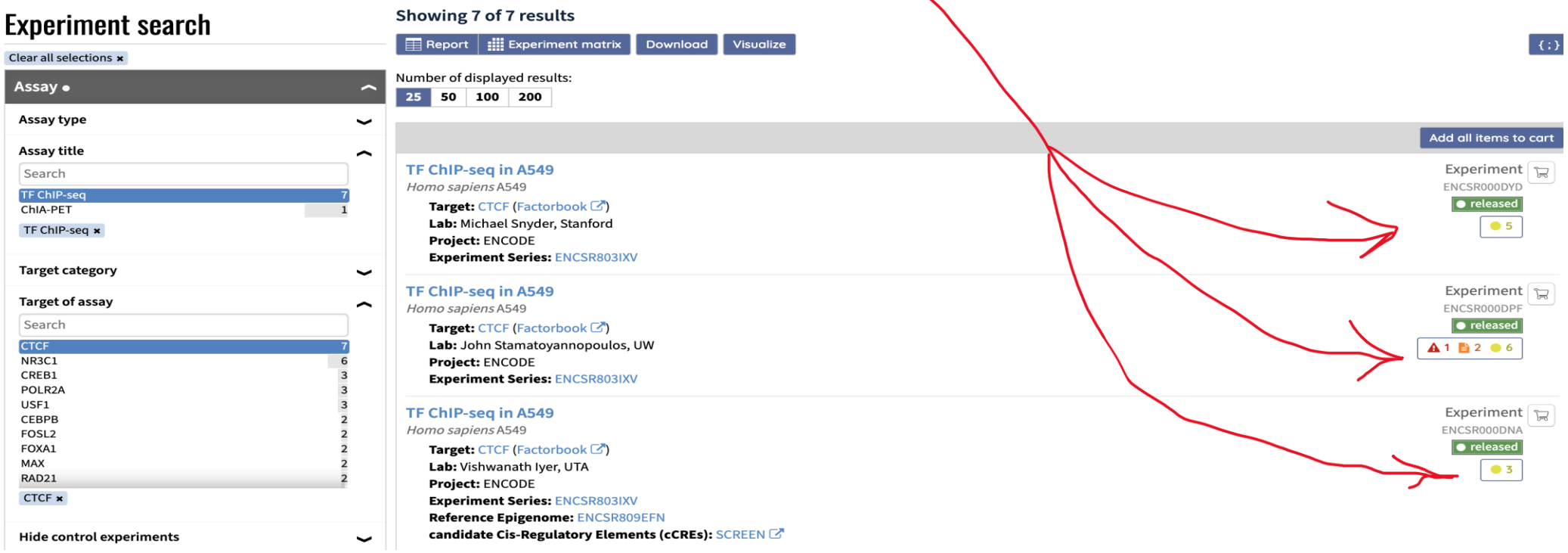
- Let’s say you need to download data for CTCF from A549 cell line. You click on the corresponding matrix cell and see the list of experiments:



- Go to the left part of the page, look at filters, and find PROVENANCE. Go with ENCODE.



- Now choose one of the experiments. Check if there are any warnings. Basically:
  - *Red: don't take it*
  - *Orange: avoid if possible*
  - *Yellow: okay*
  - *No warnings: very good*



- Click on the experiment. Check if everything is good.
- Also check the Lab and Author. If you find experiments from the same lab for other TFs, this might yield better results.

Experiment summary for ENCSR000DYD

doi:10.17989/ENCSR000DYD

5

{;}

Summary

Status:	released
Assay:	ChIP-seq (TF ChIP-seq)
Target:	CTCF
Biosample summary:	*Homo sapiens* A549
Biosample Type:	cell line
Replication type:	isogenic
Description:	CTCF ChIP-seq on human A549 produced by the Snyder lab
Nucleic acid type:	DNA
Fragmentation methods:	see document
Platform:	Illumina Genome Analyzer
Controls:	ENCSR496AXR

Attribution

ENCODE PHASE 2

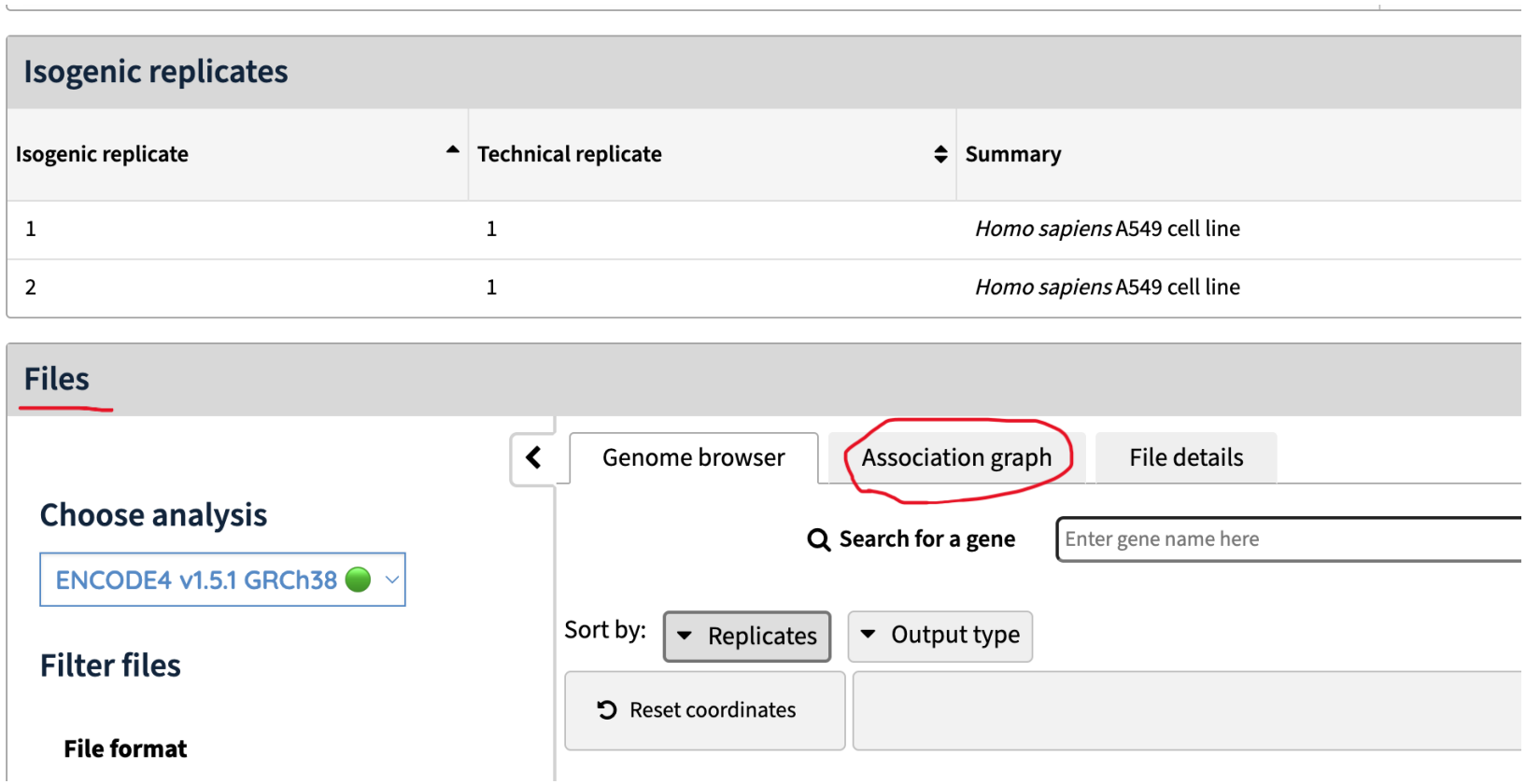
Lab:	Michael Snyder, Stanford
Award:	U54HG004558 (Michael Snyder, Stanford)
Project:	ENCODE
External resources:	UCSC-ENCODE-hg19:wgEncodeEH003384 GEO:GSM1003606
Date submitted:	July 1, 2012
Date released:	August 20, 2012
Annotation (gkmsvm-model):	ENCSR059URG
Annotation (BPNet-model):	ENCSR702YGF

Tags: RegulomeDB 1.0 | RegulomeDB 2.0 | RegulomeDB 2.1 | ENCODE 2 ENCYCLOPEDIA | ENCODE 1 ENCYCLOPEDIA | RegulomeDB 2.2

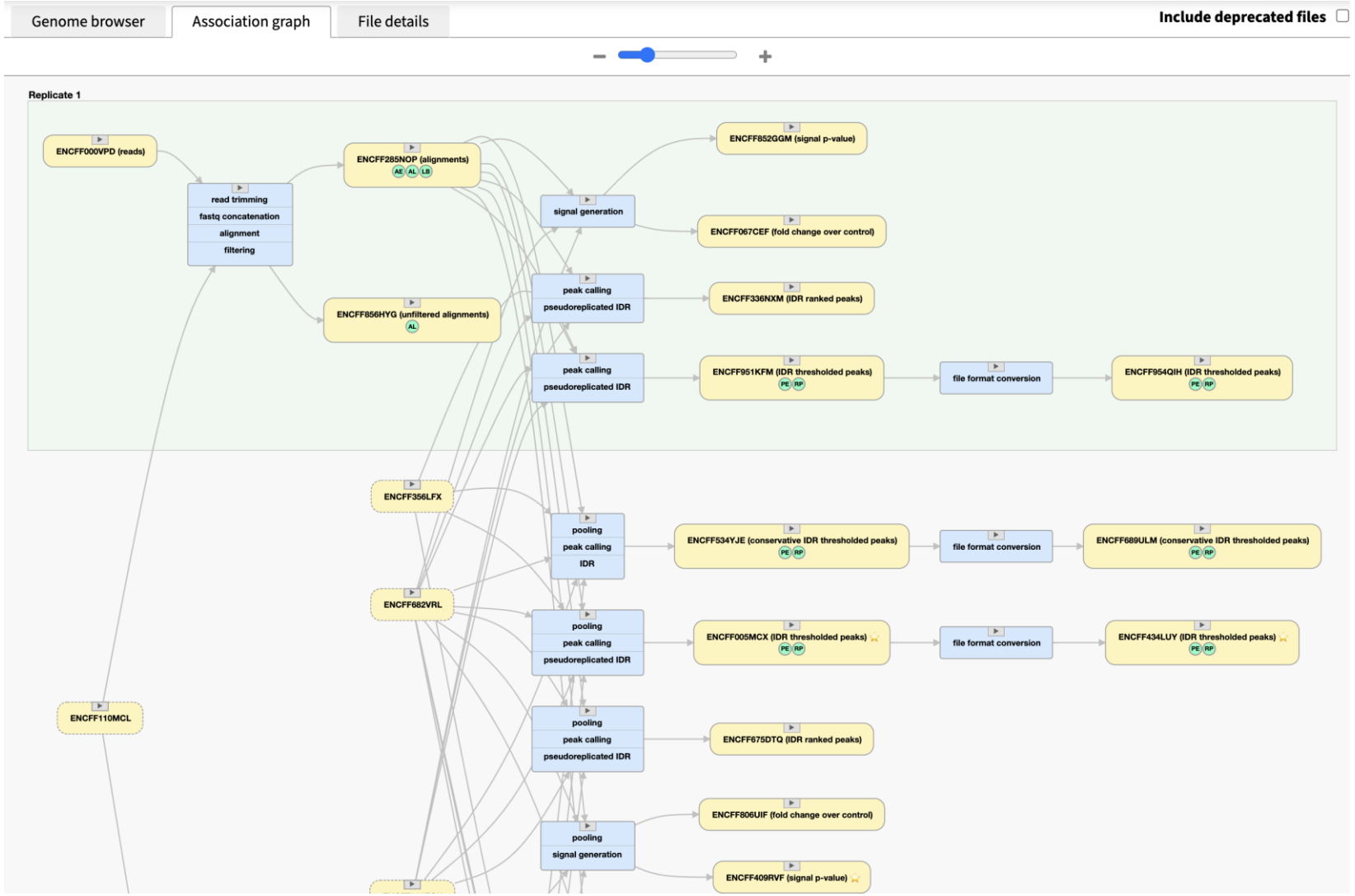
Encyclopedia Integration

Factorbook: view motifs and integrative analysis

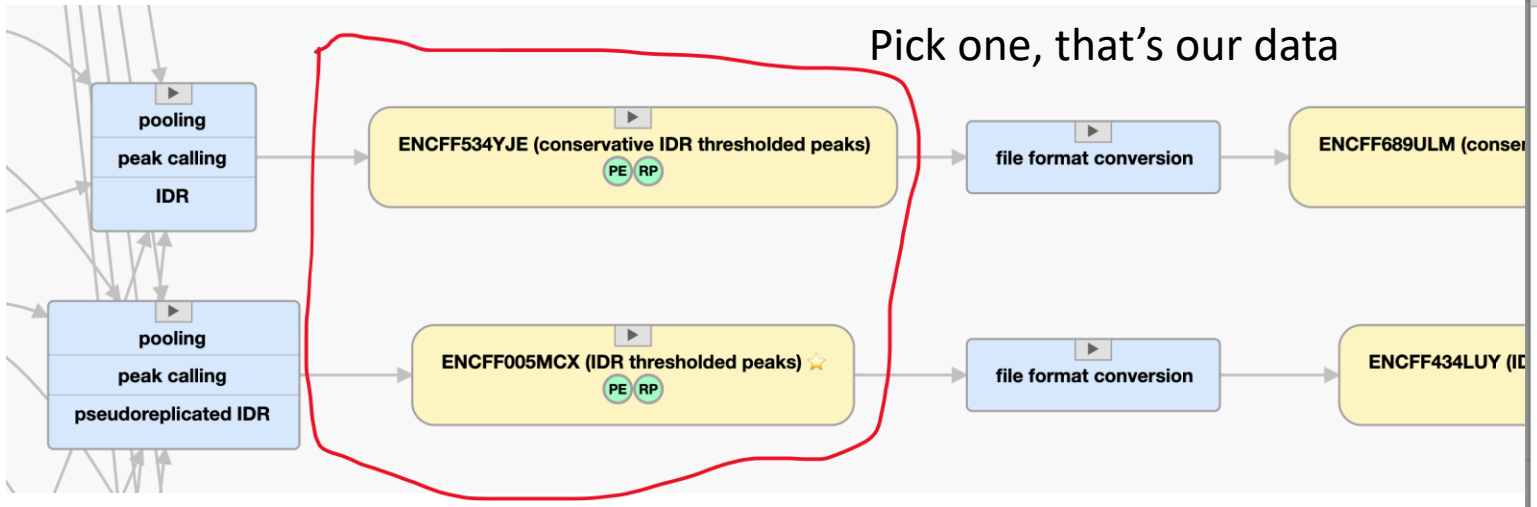
- Now scroll down. Find ’FILES’, then choose ’ASSOCIATION GRAPH’

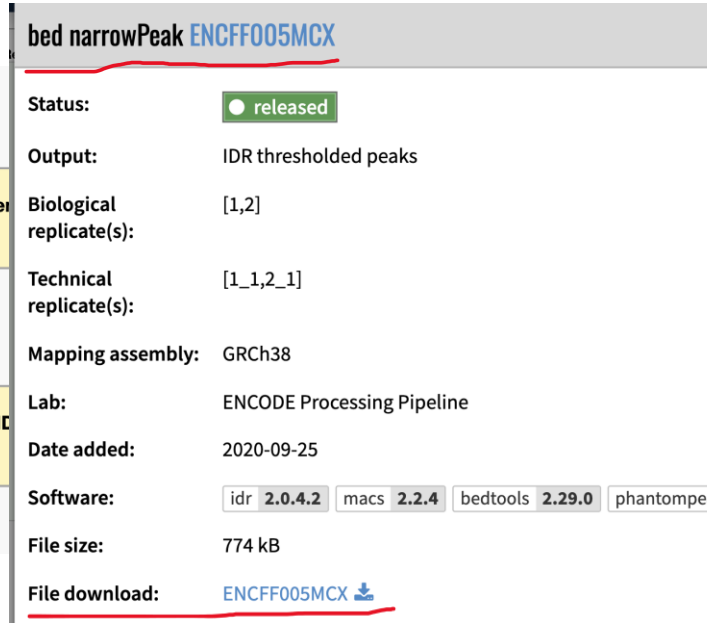


- This is what you see:
- Take your time and research it. Basically, this is a complete cycle of data preparation (from raw reads to IDR peaks). You can check what tools were used to complete each step and in what format.
- Now, there are probably more than one replicas (greenish box, there is another in the lower part of the graph)



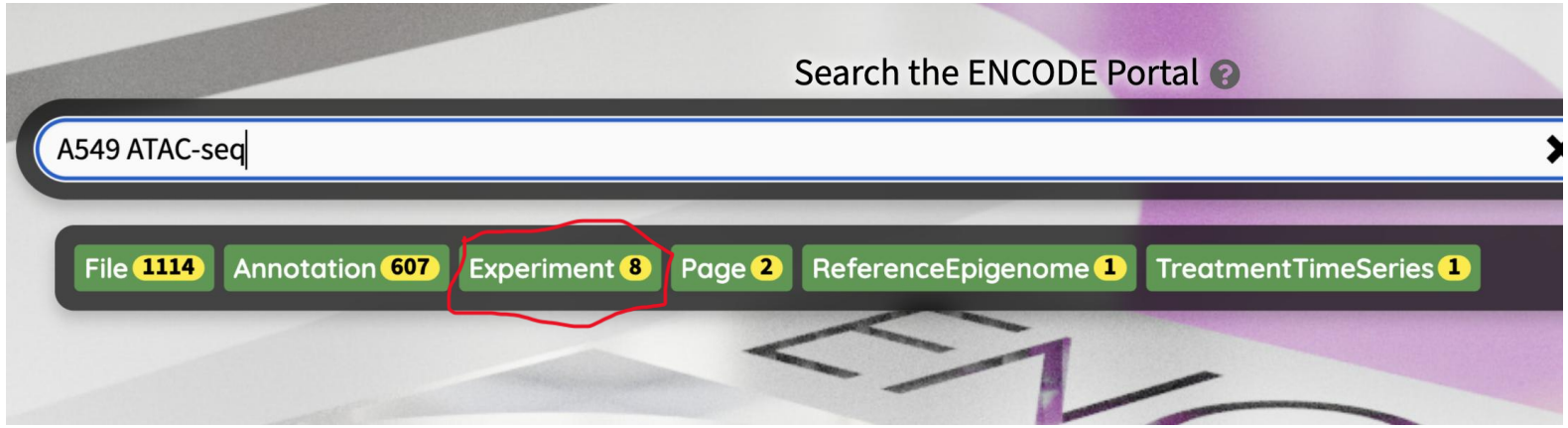
- Zoom in on the are BETWEEN the replicas (gray area). This is the peaks that were called using both replicas, meaning they are more statistically significant.
- Your .bed file is in the yellow box before file format conversion (the right one has BigBed format, which we don't need)
- Choose the box, open it, grab the link for data and insert it into wget like we did on seminar (or download it any way you want)
- Conservative VS Not is your decision. If you choose one, it's probably best to be consistent over different experiments.

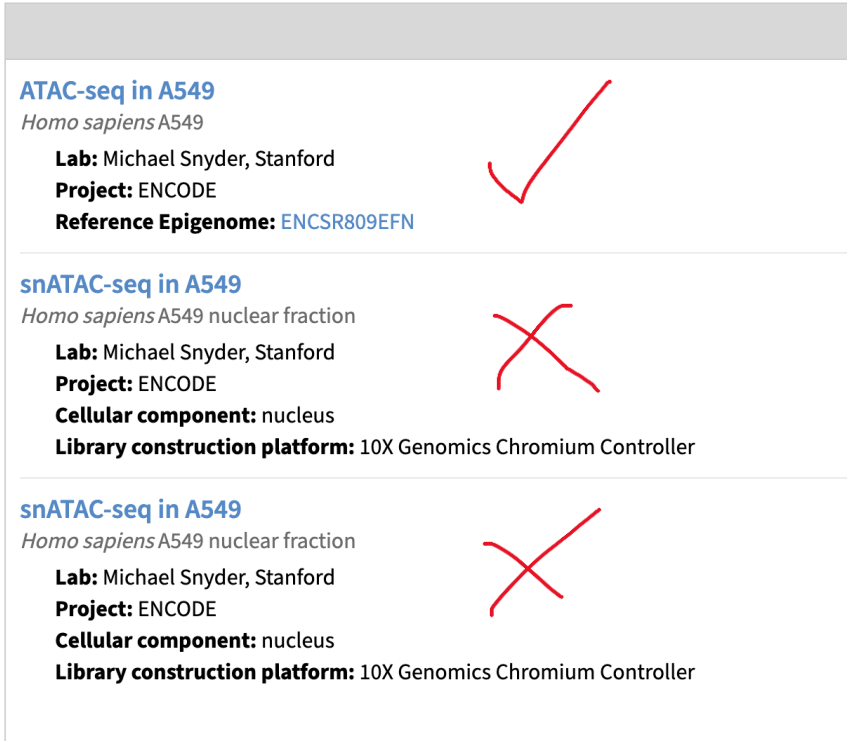




Downloading ATAC-seq data

- Go to search bar
- Insert your cell line and write ATAC-seq
- Choose ’EXPERIMENT’
- Repeat the ChIP-seq steps, with one notion: in the list of experiments, choose ATAC-seq, NOT snATAC-seq

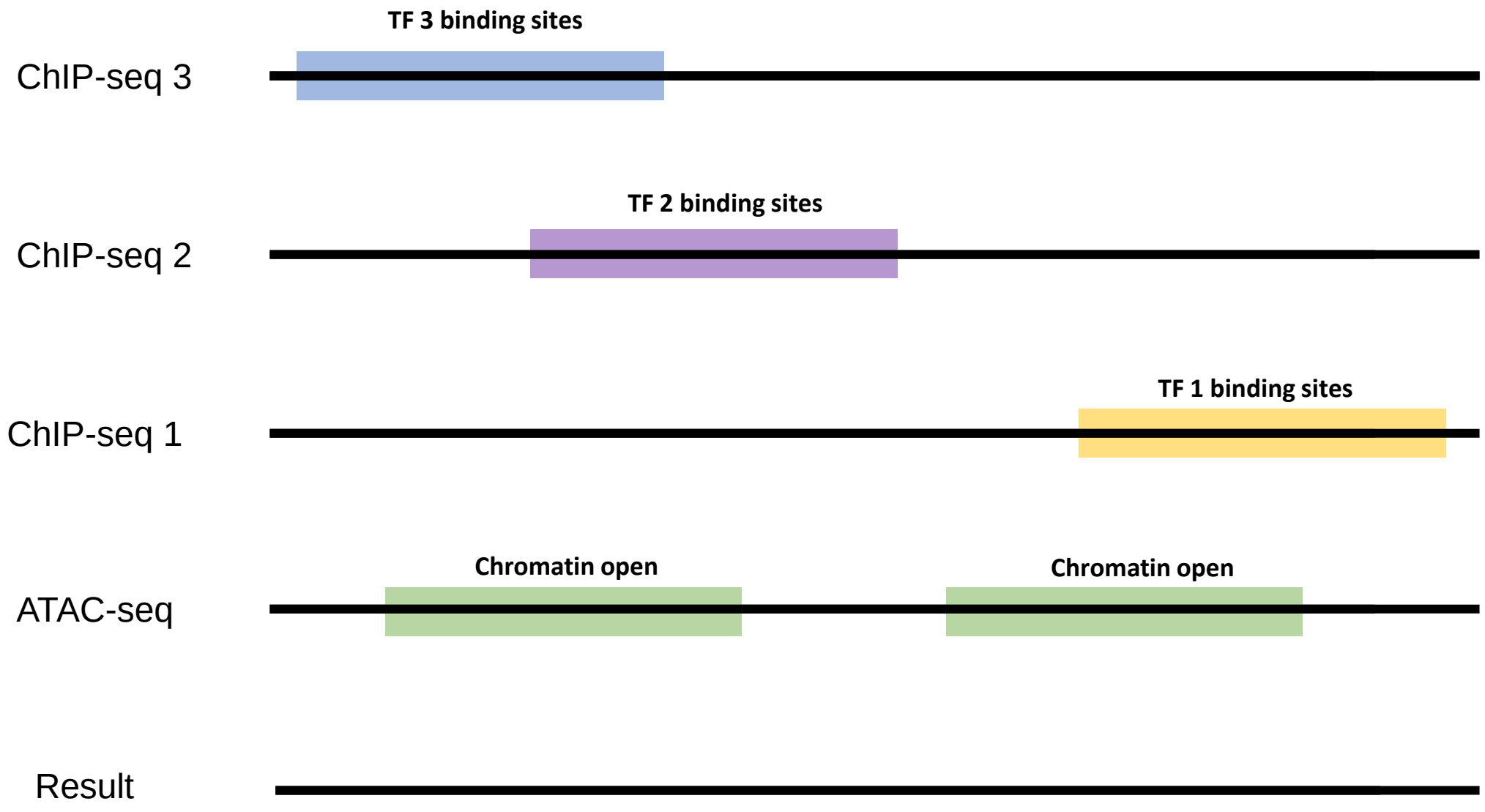




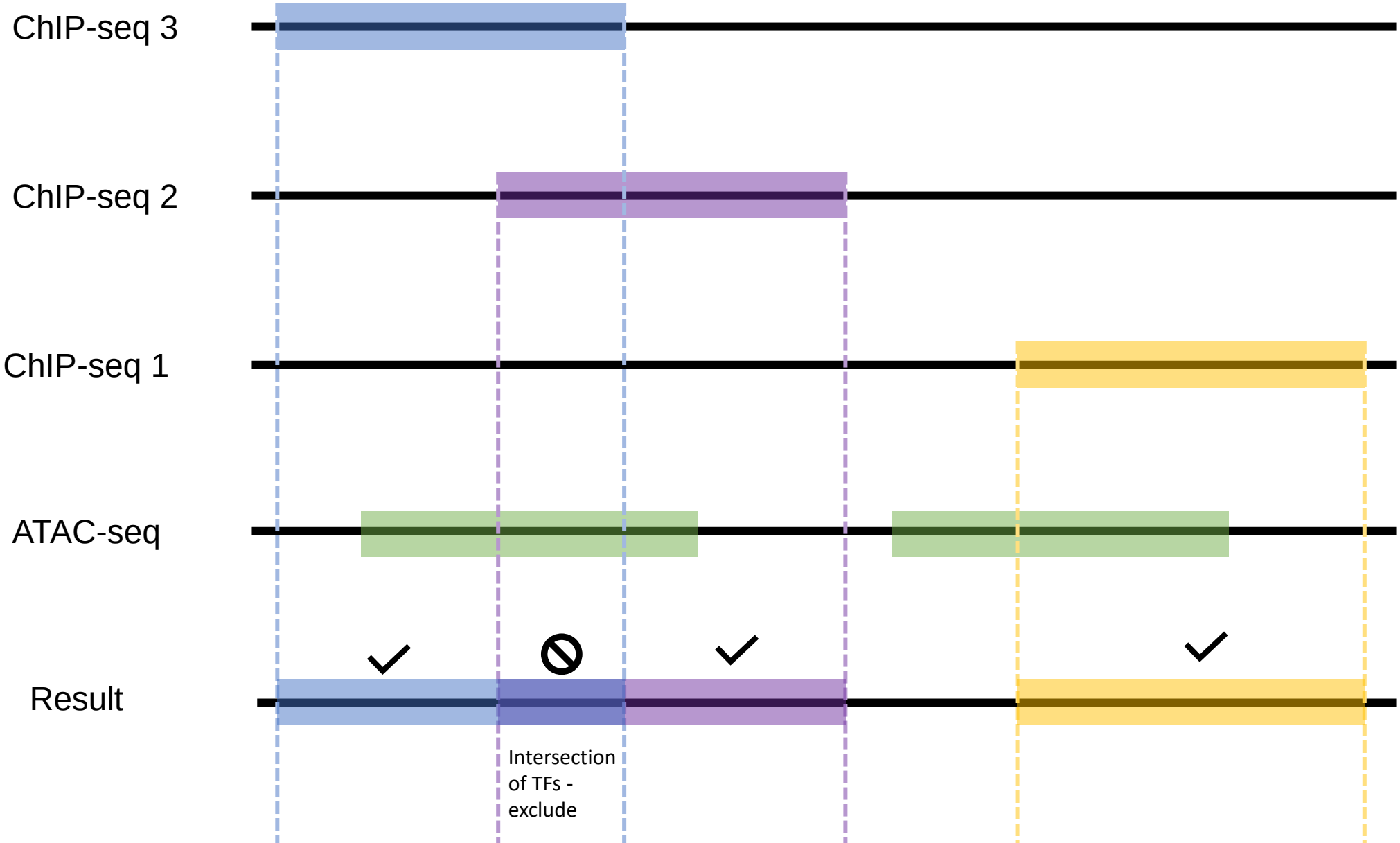
Putting it all together

- Answer the questions in the beginning of the notebook
- Select your cell line and target TFs, download the data according to this guide
- After downloading the data, follow the seminar material to preprocess your data
- In seminar we solve binary classification task, and only work with 1 TF. Here, we work with 3 TFs!
  - You will have four classes: 3 for the TFs of your choice, 1 for the background (usually class 0)
  - You have to exclude overlapping regions. One region – one class. Sanity check yourself by intersecting all regions: if you've done it right, the intersection should be empty.
- Complete all steps of the preprocessing and save your data. You will need it in the next homework, where we will train the model to predict binding sites.

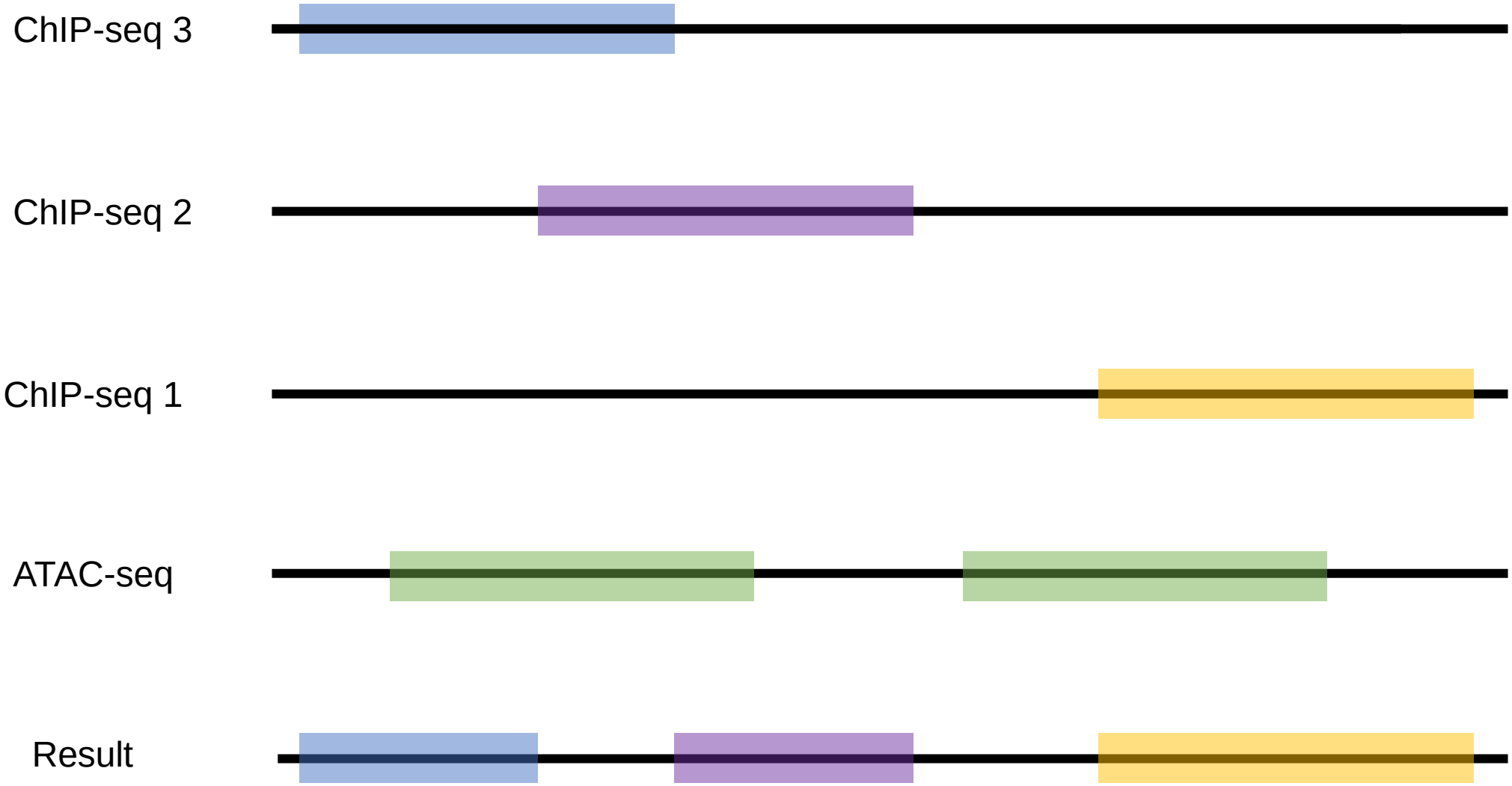
Your data at first



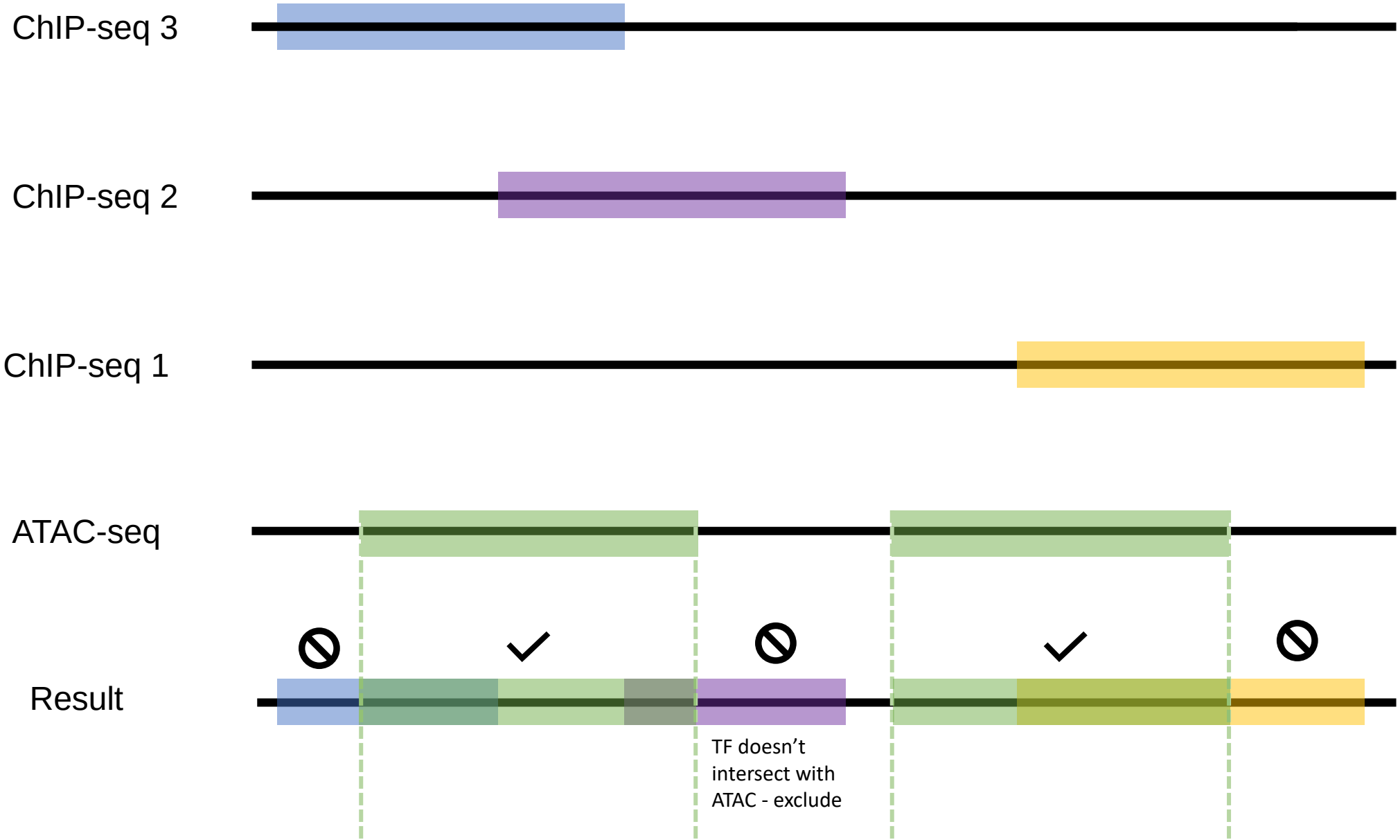
Remove intersection between TFs



Remove intersection between TFs



Intersect with ATAC-seq regions



Result: only open chromatin, no intersections between TF