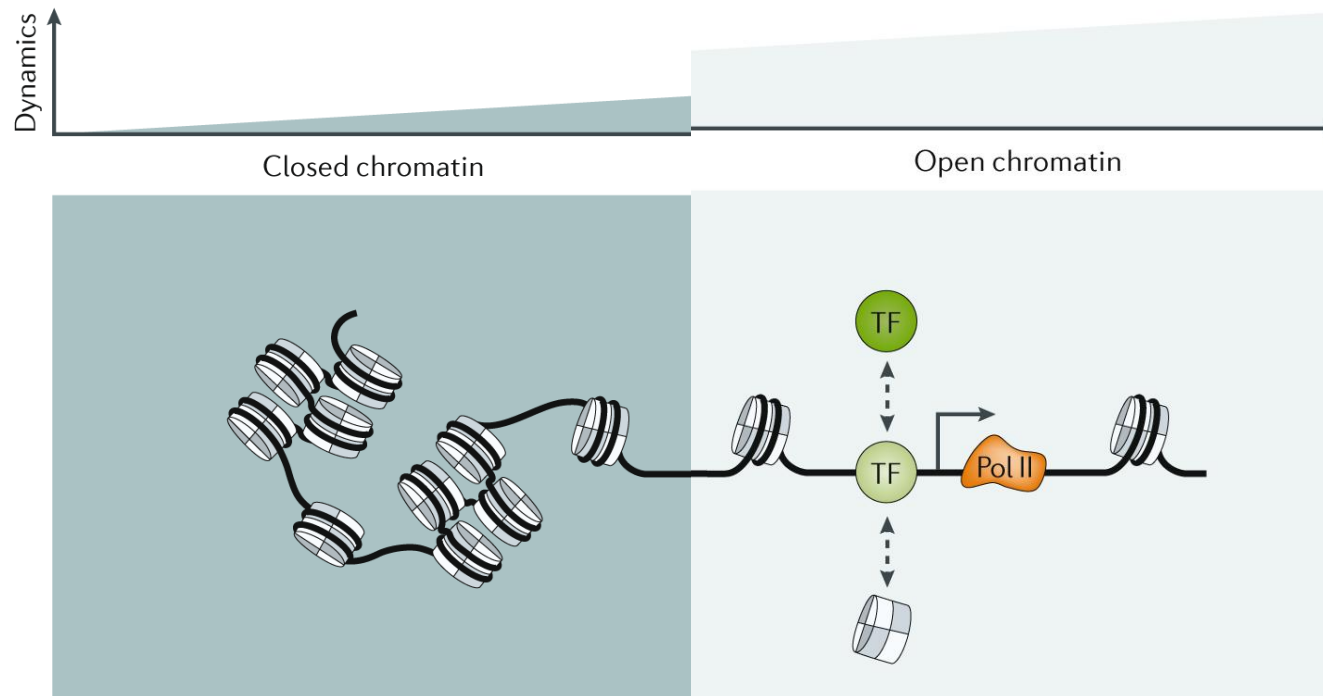


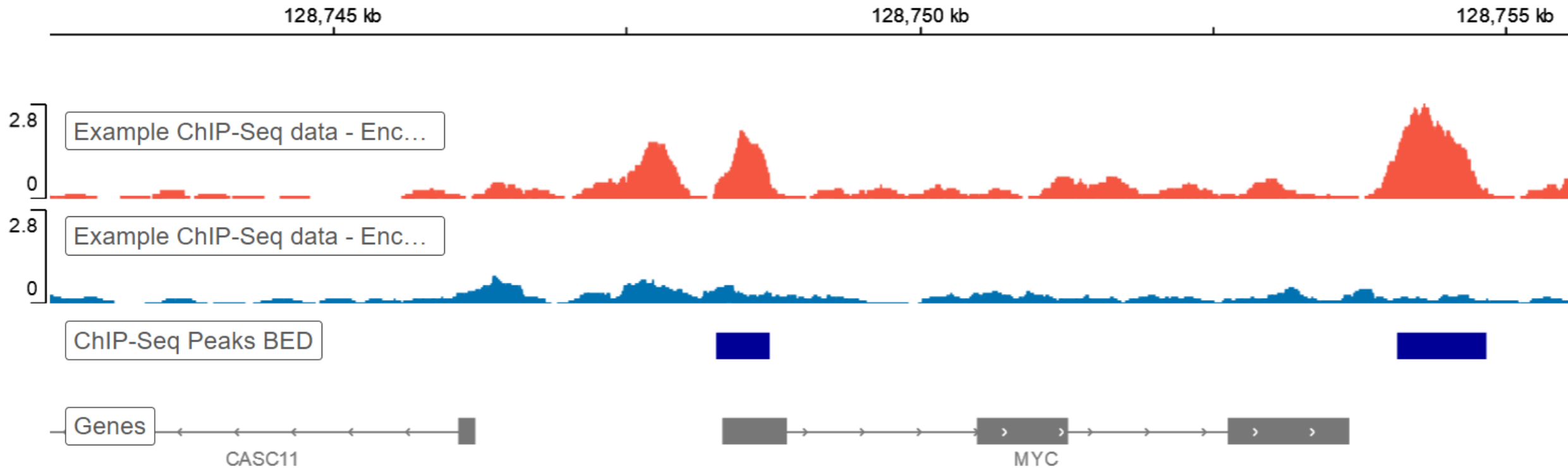
HW-4 (ENCODE) Guide

What is this all about?

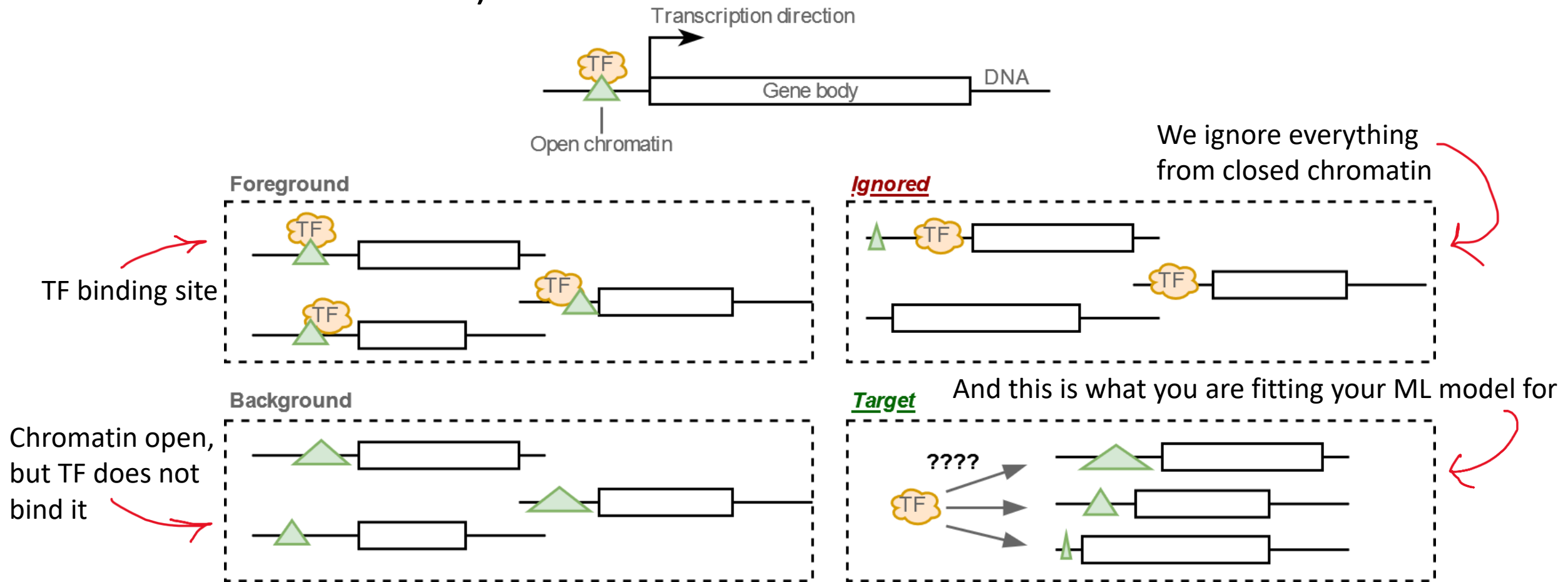
- Genome is a mess, but actually organized mess
- Some parts of the genome are available for proteins, some are not (open/closed chromatin)
- ATAC-seq experiment can determine, which parts of the chromatin were open (basically gives you coordinates)
- If it's open, then protein (Transcription Factors,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

ChIP-seq Matrix





Enter filter terms to filter the experiments included in the matrix.

▼ Enter any text string such as lung or musc or H9 to filter biosample Biosample ▾

Facet ▾

BIOSAMPLE →

TARGET ↓

 Homo sapiens	 Mus musculus	 Caenorhabditis elegans	 Drosophila melanogaster	
Histone	Transcription Factor			
cell line	in vitro differentiated cells	organoid	primary cell	tissue

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:

Experiment search

Clear all selections ✕

Assay

Assay type

Assay title

TF ChIP-seq 7

ChIA-PET 1

TF ChIP-seq ✕

Target category

Target of assay

CTCF 7

NR3C1 6

CREB1 3

POLR2A 3

USF1 3

CEBPB 2

FOSL2 2

FOXA1 2

MAX 2

RAD21 2

CTCF ✕

Hide control experiments

Showing 7 of 7 results

Report

Experiment matrix

Download

Visualize

{;}

Number of displayed results:

25

50

100

200

Add all items to cart

TF ChIP-seq in A549

Homo sapiens A549

Target: [CTCF](#) (Factorbook)

Lab: Michael Snyder, Stanford

Project: ENCODE

Experiment Series: [ENCSR803IXV](#)

Experiment

ENCSR000DYD

● released

● 5

TF ChIP-seq in A549

Homo sapiens A549

Target: [CTCF](#) (Factorbook)

Lab: John Stamatoyannopoulos, UW

Project: ENCODE

Experiment Series: [ENCSR803IXV](#)

Experiment

ENCSR000DPF

● released

▲ 1 2 ● 6

TF ChIP-seq in A549

Homo sapiens A549

Target: [CTCF](#) (Factorbook)

Lab: Vishwanath Iyer, UTA

Project: ENCODE

Experiment Series: [ENCSR803IXV](#)

Reference Epigenome: [ENCSR809EFN](#)

candidate Cis-Regulatory Elements (cCREs): [SCREEN](#)

Experiment

ENCSR000DNA

● released

● 3

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

hide control experiments

Biosample ●

Library

Analysis

Provenance ●

Lab

Project

GGR 12

ENCODE 7

ENCODE ✕

RFA

Date range selection

Quality ●

Other filters ●

- 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*

Experiment search

Clear all selections x

Assay

Assay type

Assay title

Search

TF ChIP-seq 7

ChIA-PET 1

TF ChIP-seq x

Target category

Target of assay

Search

CTCF 7

NR3C1 6

CREB1 3

POLR2A 3

USF1 3

CEBPB 2

FOSL2 2

FOXA1 2

MAX 2

RAD21 2

CTCF x

Hide control experiments

Showing 7 of 7 results

Report

Experiment matrix

Download

Visualize

Number of displayed results:

25

50

100

200

TF ChIP-seq in A549

Homo sapiens A549

Target: CTCF (Factorbook)

Lab: Michael Snyder, Stanford

Project: ENCODE

Experiment Series: ENCSR803IXV

TF ChIP-seq in A549

Homo sapiens A549

Target: CTCF (Factorbook)

Lab: John Stamatoyannopoulos, UW

Project: ENCODE

Experiment Series: ENCSR803IXV

TF ChIP-seq in A549

Homo sapiens A549

Target: CTCF (Factorbook)

Lab: Vishwanath Iyer, UTA

Project: ENCODE

Experiment Series: ENCSR803IXV

Reference Epigenome: ENCSR809EFN

candidate Cis-Regulatory Elements (cCREs): SCREEN

Add all items to cart

Experiment

ENCSR000DYD

released

5

Experiment

ENCSR000DPF

released

1 2 6

Experiment

ENCSR000DNA

released







3

- 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



Summary		Attribution	
Status:	● released	Lab:	Michael Snyder, Stanford
Assay:	ChIP-seq (TF ChIP-seq)	Award:	U54HG004558 (Michael Snyder, Stanford)
Target:	CTCF	Project:	ENCODE
Biosample summary:	Homo sapiens A549	External resources:	UCSC-ENCODE-hg19:wgEncodeEH003384 ↗ GEO:GSM1003606 ↗
Biosample Type:	cell line	Date submitted:	July 1, 2012
Replication type:	isogenic	Date released:	August 20, 2012
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		
		Annotation (gkmsvm-model):	ENCSR059URG
		Annotation (BPNet-model):	ENCSR702YGF
		Tags:	<div></div>
Encyclopedia Integration			
		Factorbook:	view motifs and integrative analysis ↗

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

Isogenic replicates

Isogenic replicate	Technical replicate	Summary
1	1	<i>Homo sapiens</i> A549 cell line
2	1	<i>Homo sapiens</i> A549 cell line

Files

Choose analysis

ENCODE4 v1.5.1 GRCh38

Filter files

File format

<

Genome browser

Association graph

File details

Q Search for a gene

Enter gene name here

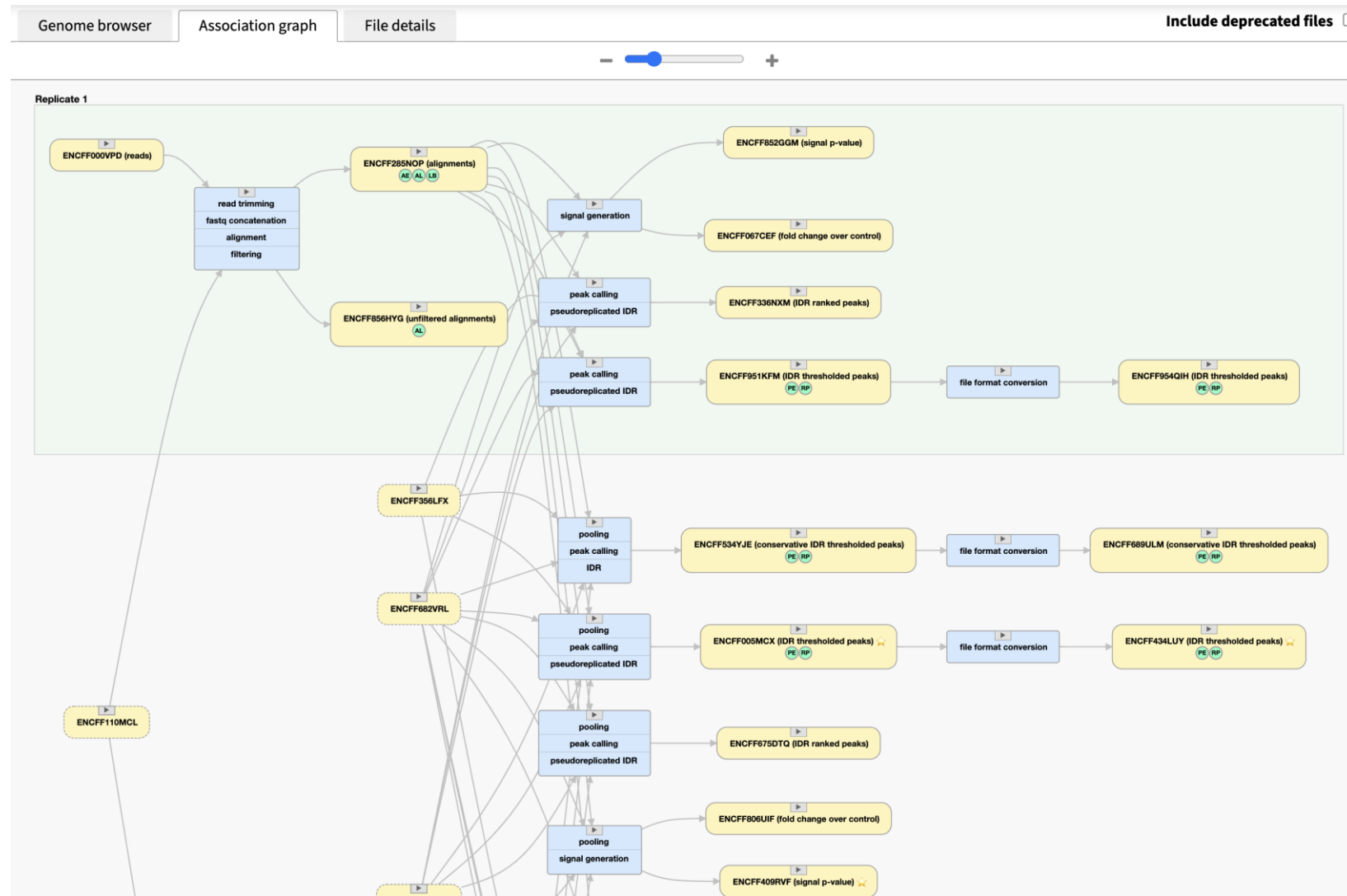
Sort by:

Replicates

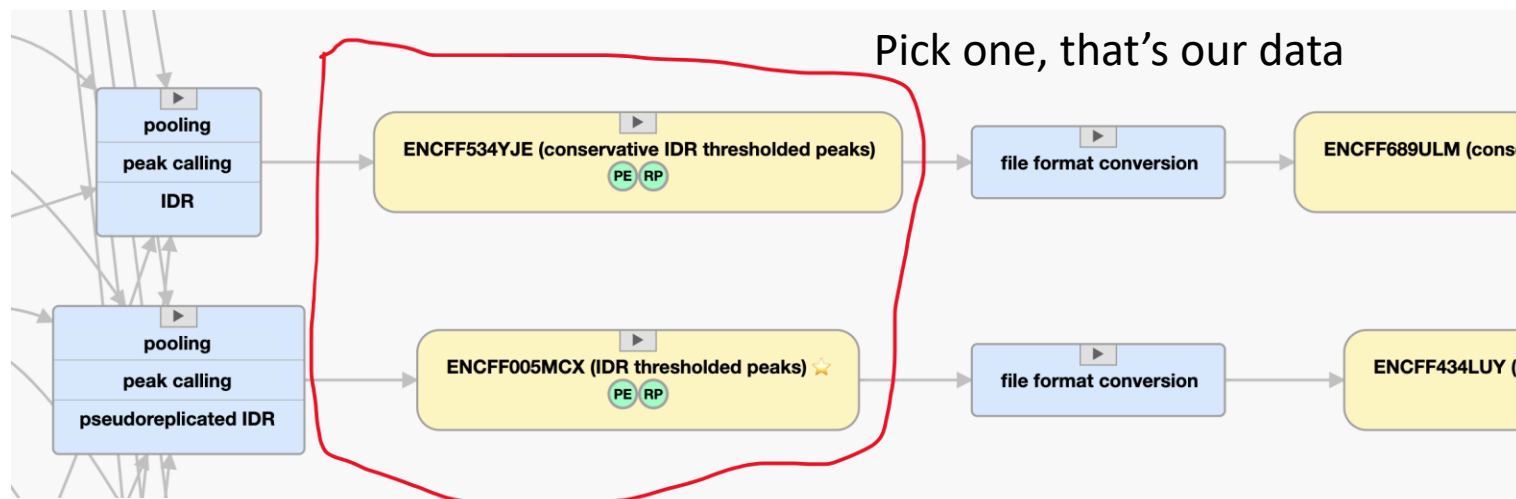
Output type

Reset coordinates

- 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)



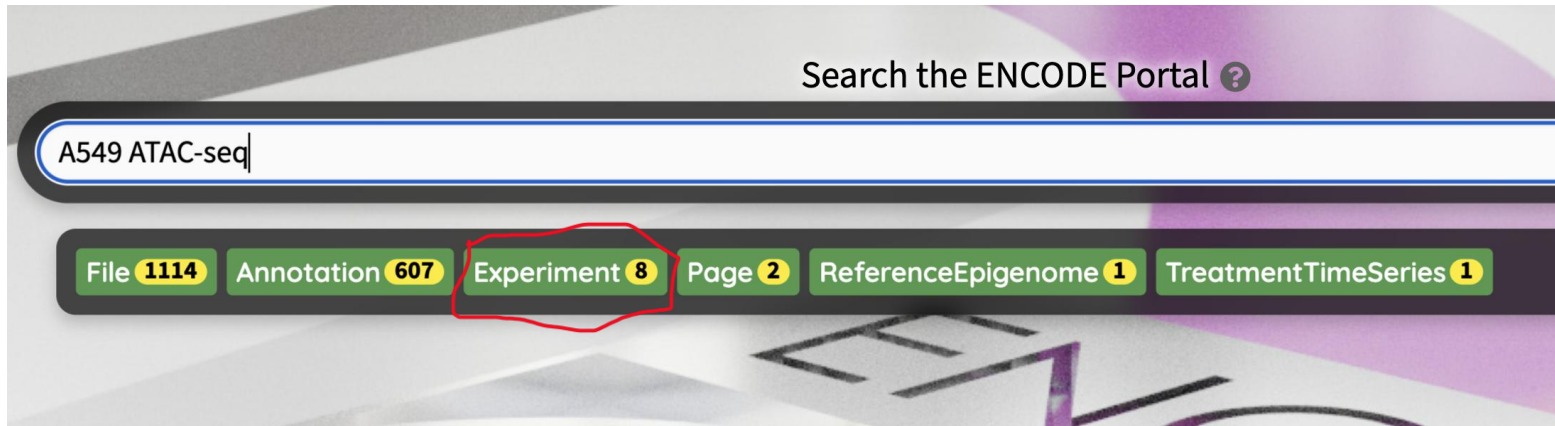
- Zoom in on the are BETWEEN the replicas (gray area). This is the peaks that were called using both replicas, meaning they are more statistically significant.
- Your .bed file is in the yellow box before file format conversion (the right one has BigBed format, which we don't need)
- Choose the box, 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.



bed narrowPeak ENCF005MCX	
Status:	● released
Output:	IDR thresholded peaks
Biological replicate(s):	[1,2]
Technical replicate(s):	[1_1,2_1]
Mapping assembly:	GRCh38
Lab:	ENCODE Processing Pipeline
Date added:	2020-09-25
Software:	idr 2.0.4.2 macs 2.2.4 bedtools 2.29.0 phantompeaks 2.10.0
File size:	774 kB
File download:	ENCF005MCX

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

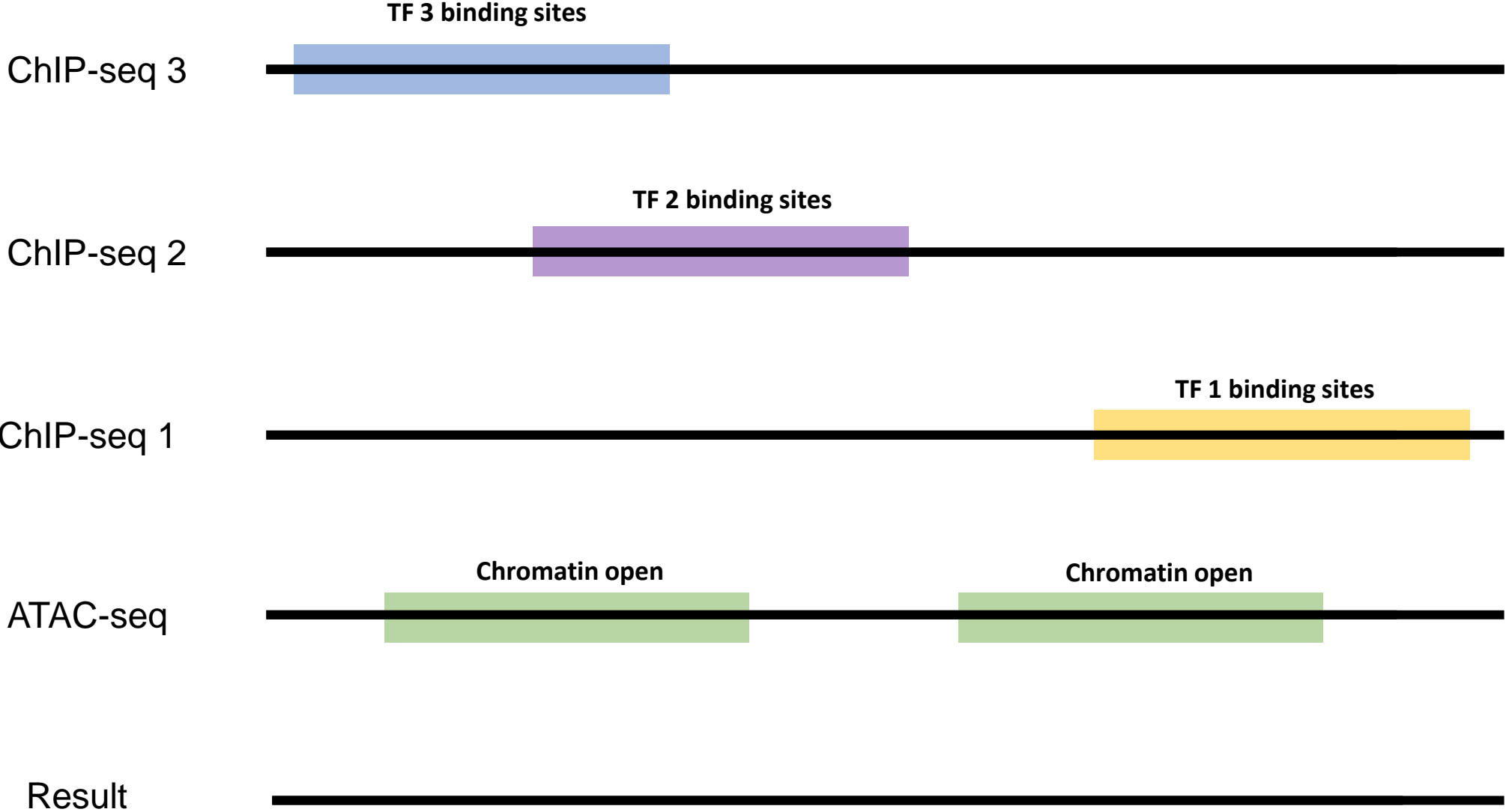


ATAC-seq in A549 <i>Homo sapiens</i> A549 Lab: Michael Snyder, Stanford Project: ENCODE Reference Epigenome: ENCSR809EFN	✓
snATAC-seq in A549 <i>Homo sapiens</i> A549 nuclear fraction Lab: Michael Snyder, Stanford Project: ENCODE Cellular component: nucleus Library construction platform: 10X Genomics Chromium Controller	✗
snATAC-seq in A549 <i>Homo sapiens</i> A549 nuclear fraction Lab: Michael Snyder, Stanford Project: ENCODE Cellular component: nucleus Library construction platform: 10X Genomics Chromium Controller	✗

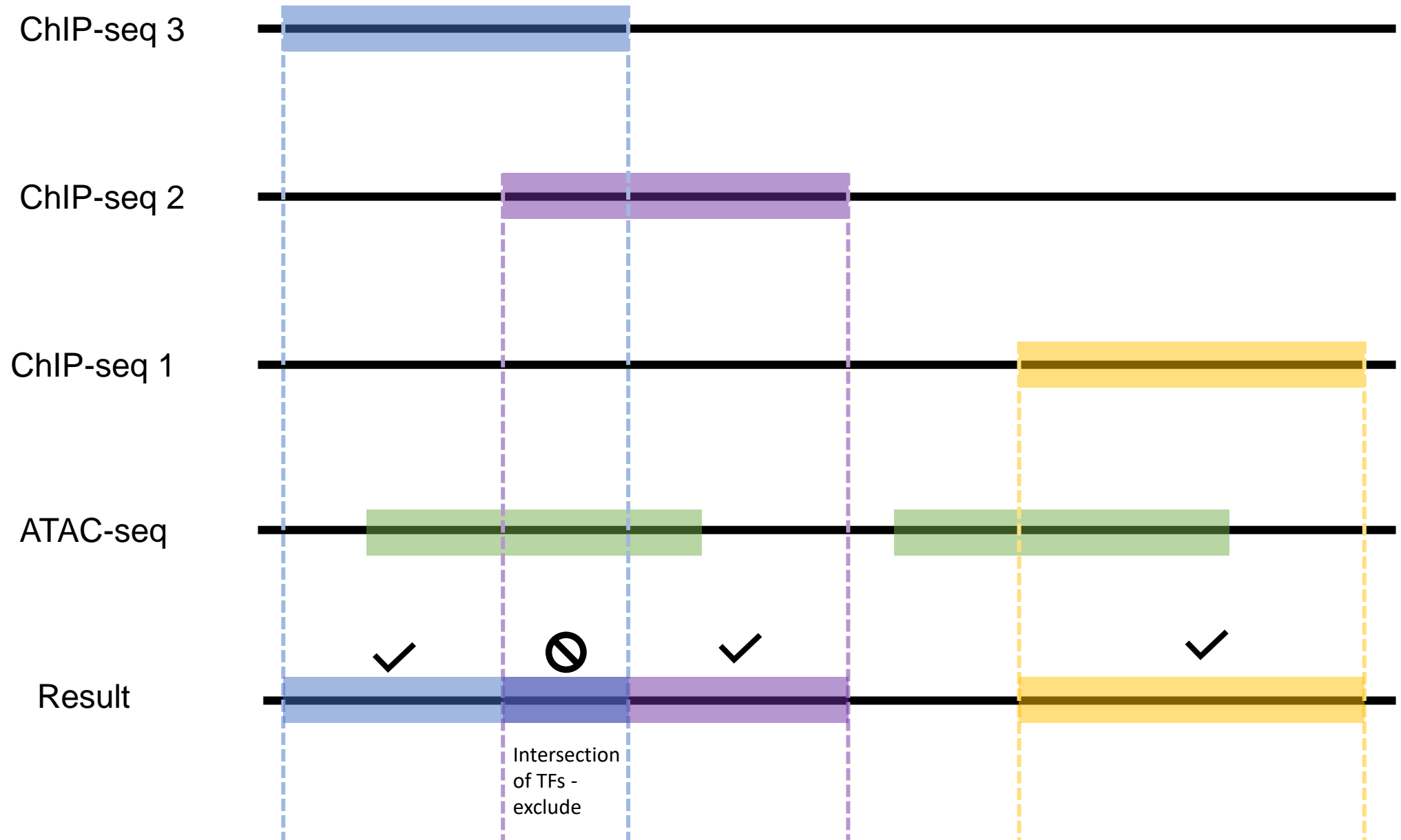
Putting it all together

- 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

ChIP-seq 3



ChIP-seq 2



ChIP-seq 1



ATAC-seq



Result



Intersect with ATAC-seq regions

ChIP-seq 3



ChIP-seq 2



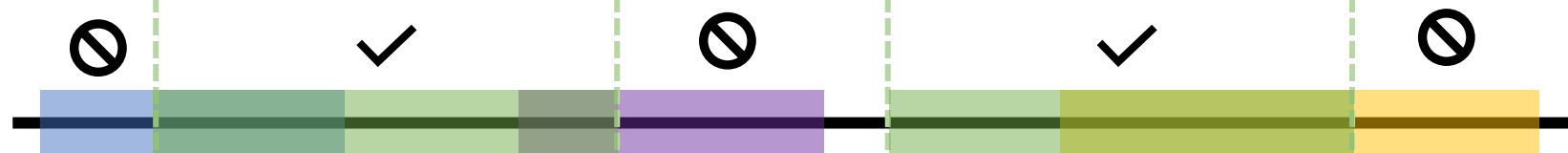
ChIP-seq 1



ATAC-seq



Result



TF doesn't
intersect with
ATAC - exclude

Result: only open chromatin, no intersections between TF

