Protocol: annotation of ChIP-seq peaks

Table of Contents

Introduction

Objectives : given a set of identified peaks, interpret the biological meaning of these peaks; look for positional biases, overlapp with other datasets, functional enrichments.

Dataset used: based on the publication "GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility.", Theodorou et al., Genome Research (2012) (PMID: 23172872).
We will use ChIP-seq data for estrogen-receptor alpha (Eralpha) in two conditions:

1. wild-type condition after estrogen stimulation,
2. Eralpha binding after treatment with siRNA anti-GATA3.

Three replicates are available for each condition, and an input dataset for the considered cell-line (MCF-7) (check the dataset description table on the website for a description of the datasets and links to the files).

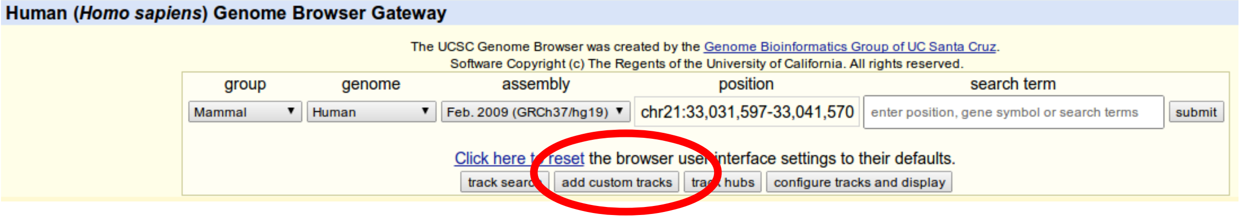
Starting point : BED files containing peak / enriched region locations. In particular, we would like to analyze (I) a genome-wide set of siGATA3 - ESR1 peaks, (ii) siNT - ESR1 peaks.
Question: *do these condition specific peaks exhibit specific features ? What about common peaks to both conditions ?*

Protocol

Step 1 : displaying BED files in the UCSC genome browser

IGV allows to visualise local datasets, such as BAM, BED and WIG files. However, we would like to take advantage of the published genome-wide datasets (e.g. ENCODE datasets) to compare our own data with other datasets. Many of these are publicly available in the UCSC genome browser.

- Download the BED files containing the peak coordinates on your computer (use the links on the course webpage). You should have a set of siGATA3 peaks and siNT peaks.
- Go to the UCSC genome browser : http://genome.ucsc.edu/
- Select Genomes and the hg19 human genome.
- Click on Add custom tracks



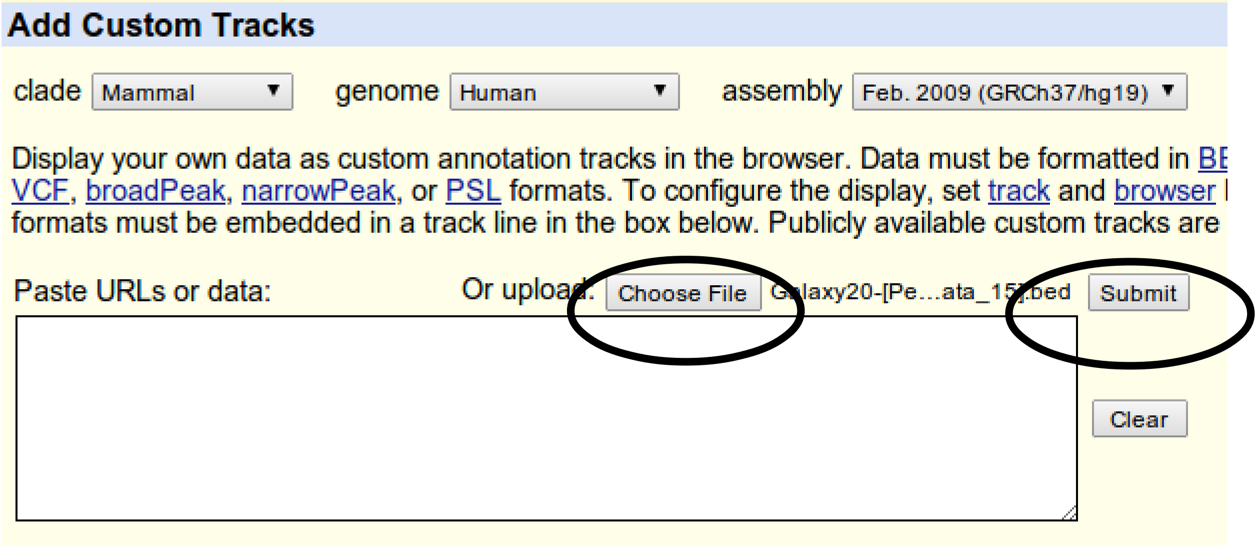
Human (*Homo sapiens*) Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.
Software Copyright (c) The Regents of the University of California. All rights reserved.

group: Mammal
genome: Human
assembly: Feb. 2009 (GRCh37/hg19)
position: chr21:33,031,597-33,041,570
search term: enter position, gene symbol or search terms

Click here to reset the browser user interface settings to their defaults.

- Upload the BED files by clicking *Upload*, then *Submit*



Add Custom Tracks

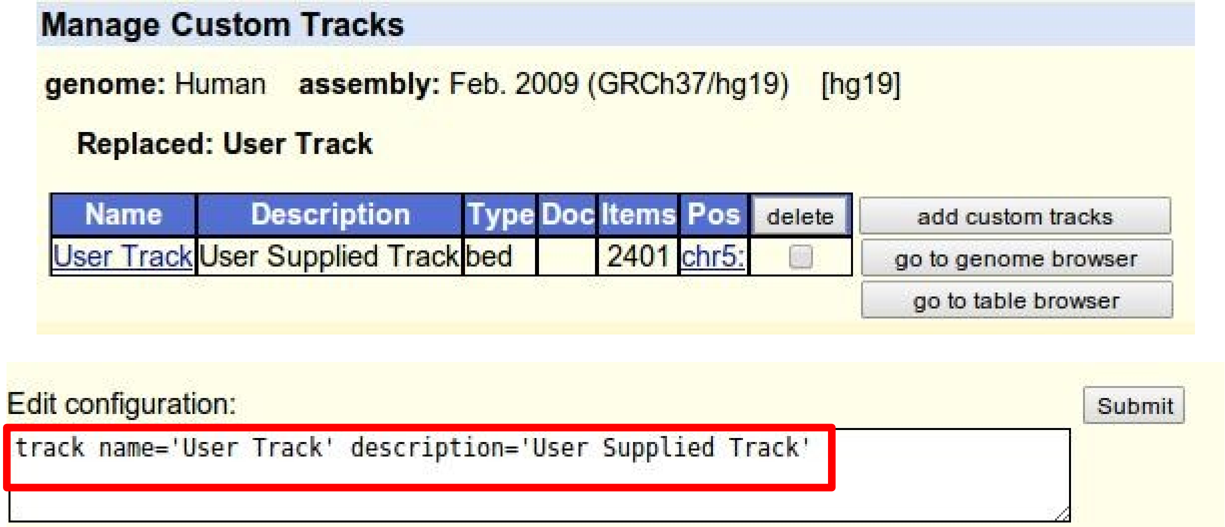
clade: Mammal
genome: Human
assembly: Feb. 2009 (GRCh37/hg19)

Display your own data as custom annotation tracks in the browser. Data must be formatted in BE
VCF, broadPeak, narrowPeak, or PSL formats. To configure the display, set track and browser l
formats must be embedded in a track line in the box below. Publicly available custom tracks are

Paste URLs or data:

Or upload: Choose File Galaxy20-[Pe...ata_15].bed

- Click on the User Track link, and edit the description by changing the name and description of the BED file (Don't forget to put quotes !)



Manage Custom Tracks

genome: Human **assembly:** Feb. 2009 (GRCh37/hg19) [hg19]

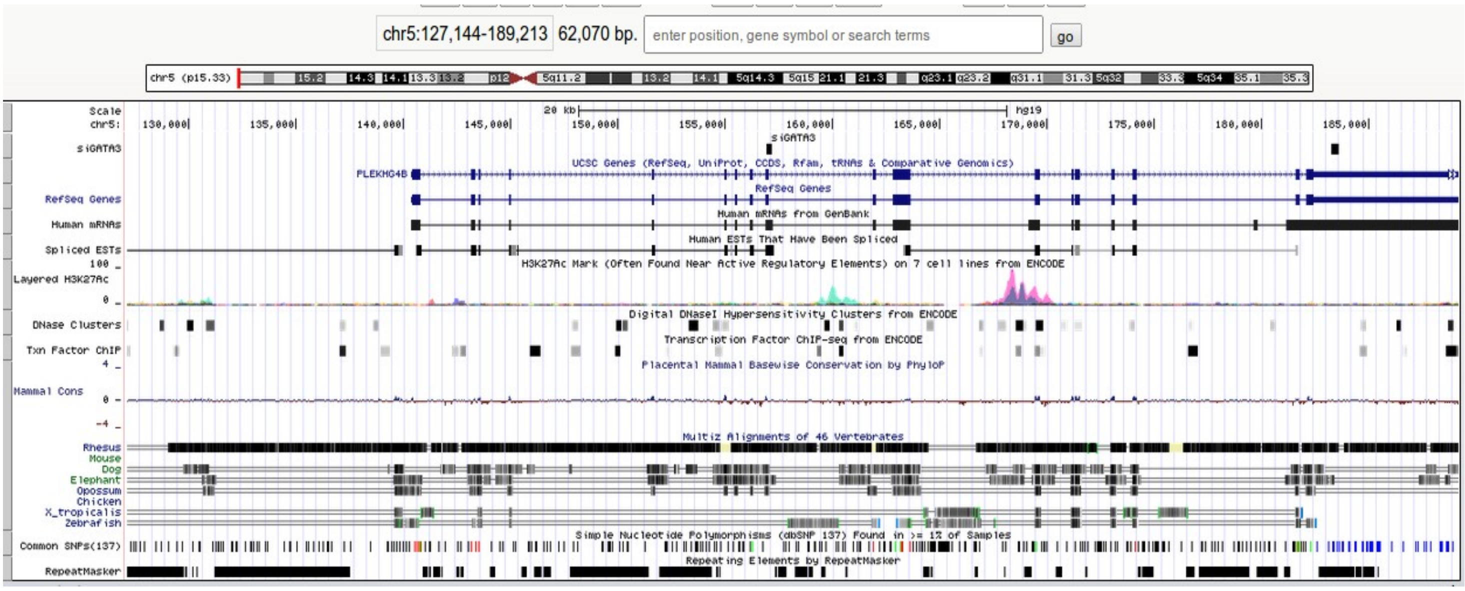
Replaced: User Track

Name	Description	Type	Doc	Items	Pos	delete
User Track	User Supplied Track	bed		2401	chr5:	□

Edit configuration:

```
track name='User Track' description='User Supplied Track'
```

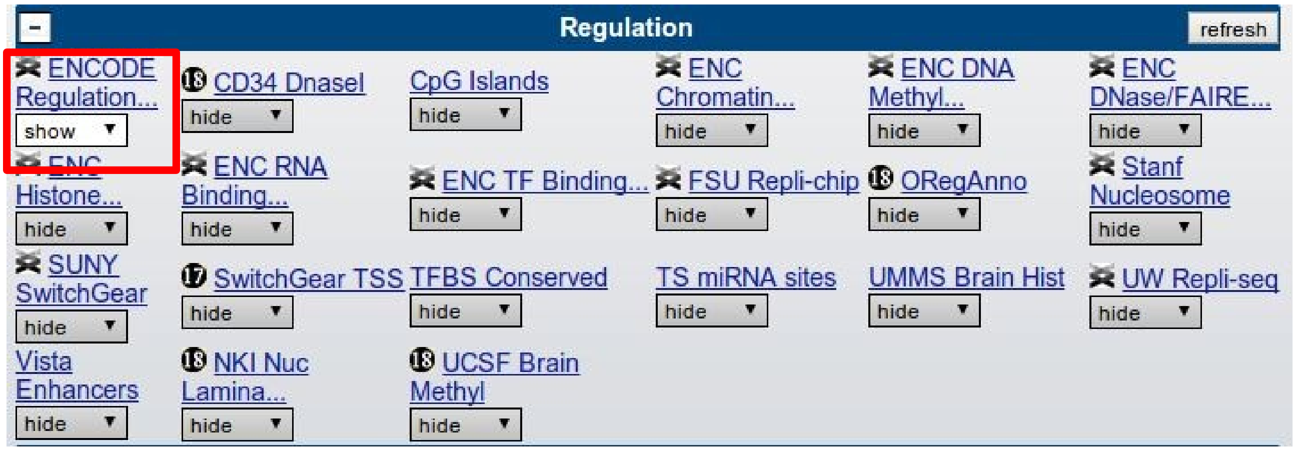
- upload the siGATA3 and siNT BED files to the UCSC genome browser.
- The click *go to genome browser*



chr5:127,144-189,213 62,070 bp.

enter position, gene symbol or search terms
go
chr5 (p15.33)
Scale
chr5:
20 kb
hg19
siGATA3
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
PLEKHG4B
RefSeq Genes
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
Layered H3K27Ac
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
Digital DNaseI Hypersensitivity Clusters from ENCODE
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE
Placental Mammal Basewise Conservation by PhyloP
Mammal Cons
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
Common SNPs(137)
Repeating Elements by RepeatMasker
RepeatMasker

- Zoom / unzoom and move along the genome to see some of your peaks appearing
- Go to the track control panel of the Regulation section, and click on the link ENCODE regulation



Regulation

ENCODE Regulation... show	CD34 DnaseI hide	CpG Islands hide	ENC Chromatin... hide	ENC DNA Methyl... hide	ENC DNase/FAIRE... hide
ENC Histone... hide	ENC RNA Binding... hide	ENC TF Binding... hide	FSU Repli-chip hide	ORegAnno hide	Stanf Nucleosome hide
SUNY SwitchGear hide	SwitchGear TSS hide	TFBS Conserved hide	TS miRNA sites hide	UMMS Brain Hist hide	UW Repli-seq hide
Vista Enhancers hide	NKI Nuc Lamina... hide	UCSF Brain Methyl hide			

- select a display mode for the H3K4me1, H3K27ac and DNase cluster track, then click submit
- select a dataset from the ECODE TF binding track to display ChIP-seq data for other transcription factors

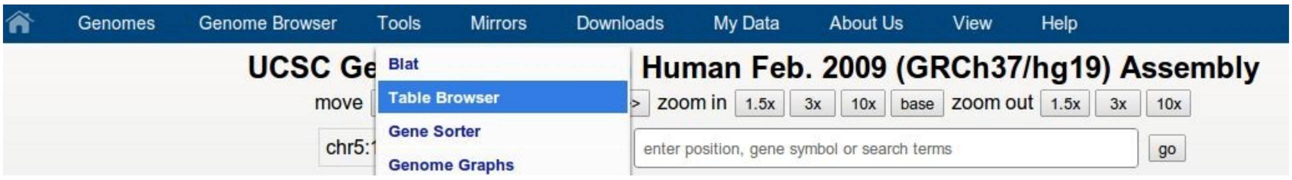
Things to consider:

- do you see coincidence of ESR1 binding sites with histone marks ?
- do you see coincidence of ESR1 binding sites with other TFBS ?

Step 2 : intersecting peak sets in UCSC

Goal : obtain a set of common peaks from the two sets uploaded

- On the top of the page, click on *Tools* > *Table browser*



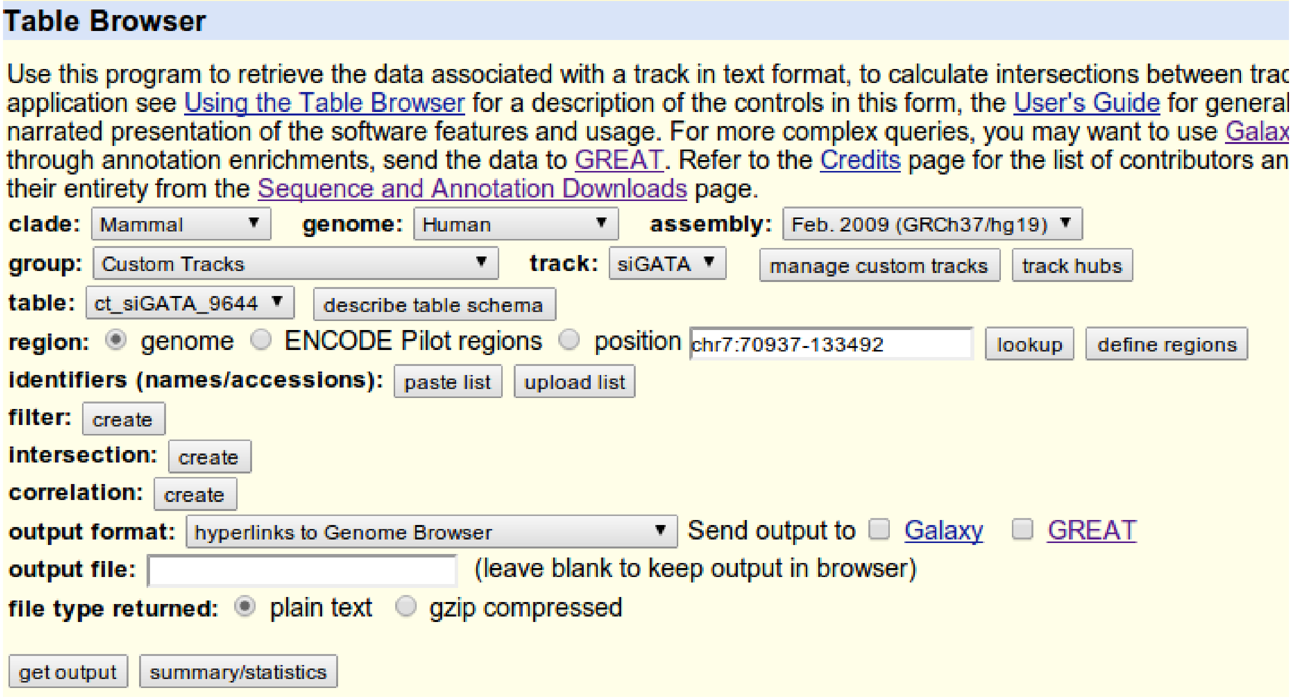


Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between trac application see Using the Table Browser for a description of the controls in this form, the User's Guide for general narrated presentation of the software features and usage. For more complex queries, you may want to use Galax through annotation enrichments, send the data to GREAT. Refer to the Credits page for the list of contributors an their entirety from the Sequence and Annotation Downloads page.

clade: Mammal ▼ **genome:** Human ▼ **assembly:** Feb. 2009 (GRCh37/hg19) ▼

group: Custom Tracks ▼ **track:** siGATA ▼ manage custom tracks track hubs

table: ct_siGATA_9644 ▼ describe table schema

region: ● genome ○ ENCODE Pilot regions ○ position chr7:70937-133492 lookup define regions

identifiers (names/accessions): paste list upload list

filter: create

intersection: create

correlation: create

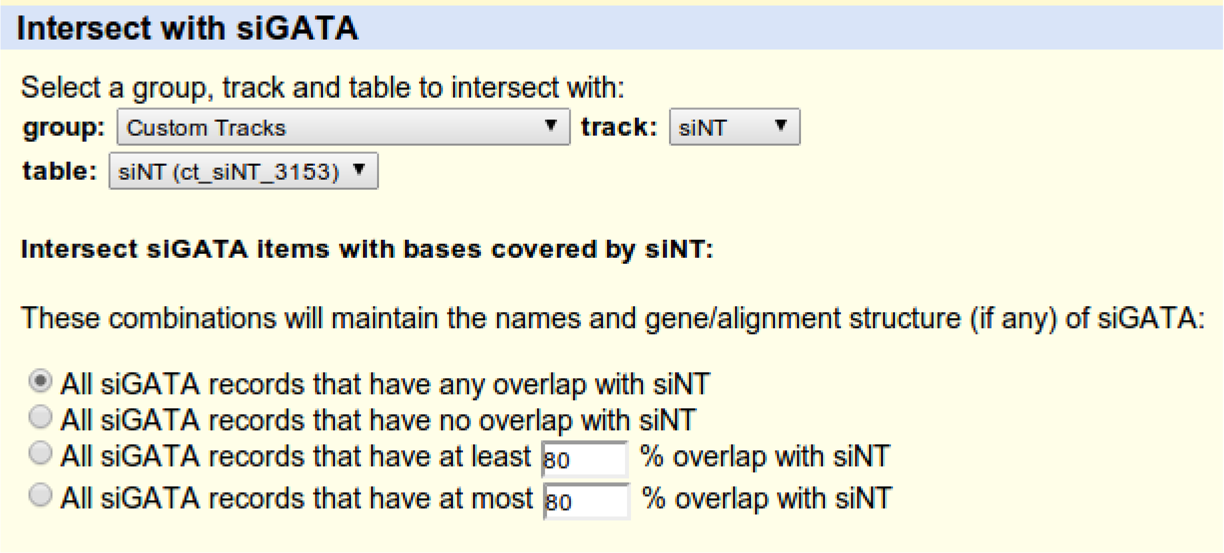
output format: hyperlinks to Genome Browser ▼ Send output to □ Galaxy □ GREAT

output file: (leave blank to keep output in browser)

file type returned: ● plain text ○ gzip compressed

get output summary/statistics

- Click on *intersection,* and select the other custom dataset



Intersect with siGATA

Select a group, track and table to intersect with:
group: Custom Tracks ▾ **track:** siNT ▾
table: siNT (ct_siNT_3153) ▾

Intersect siGATA items with bases covered by siNT:

These combinations will maintain the names and gene/alignment structure (if any) of siGATA:

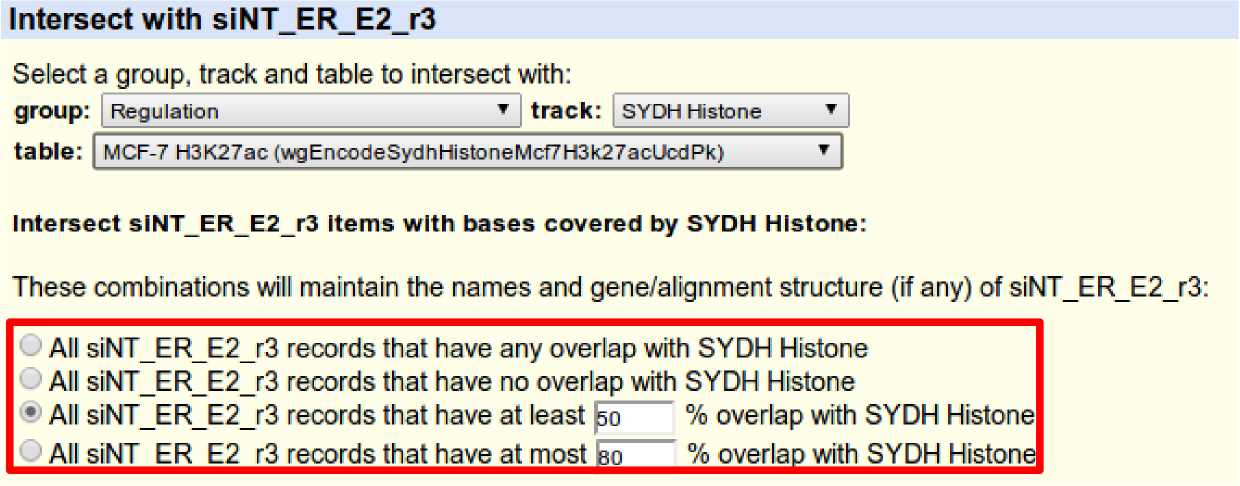
● All siGATA records that have any overlap with siNT
○ All siGATA records that have no overlap with siNT
○ All siGATA records that have at least 80 % overlap with siNT
○ All siGATA records that have at most 80 % overlap with siNT

- Submit, then click on *Summary statistics* to obtain the number of common peaks between 2 datasets.
- In output format, select Custom track, to create a new custom track that will contain common peaks.

You can repeat this procedure to intersect any track with any other (e.g. : obtain the ESR1 peaks that overlapp with H3K4me1 enriched regions, etc...)

As an illustration, we will Intersect one set of peaks with the H3K27ac dataset from ENCODE in the MCF-7 cells

- Go to table browser
- select your set of peaks in the top part
- Click on Intersection > Create
- Select
  - Group : **regulation**
  - Track : **SYDH Histone**
  - Table : **MCF-7 H3K27ac (XXXXXPk)** (attention, il y a 2 choix possible, utilisez le dataset contenant les peaks qui se termine par Pk)



Intersect with siNT_ER_E2_r3

Select a group, track and table to intersect with:
group: Regulation ▾ **track:** SYDH Histone ▾
table: MCF-7 H3K27ac (wgEncodeSydhHistoneMcf7H3k27acUcdPk) ▾

Intersect siNT_ER_E2_r3 items with bases covered by SYDH Histone:

These combinations will maintain the names and gene/alignment structure (if any) of siNT_ER_E2_r3:

○ All siNT_ER_E2_r3 records that have any overlap with SYDH Histone
○ All siNT_ER_E2_r3 records that have no overlap with SYDH Histone
● All siNT_ER_E2_r3 records that have at least 50 % overlap with SYDH Histone
○ All siNT_ER_E2_r3 records that have at most 80 % overlap with SYDH Histone

- Select the third option and indicate that you require at least 50% of the peak to overlapp the H3K27ac.
- Submit, then click *Summary Statistics* to see how many of your peaks intersect the H3K27ac
- in output format, select custom track to create a new custom track containing the peaks which overlapp H3K27ac regions (give a name and a description to your tracks).

Step 3 : functional enrichments

Goal: determine whether the peaks localize close to genes that show a functional enrichment. We will use the tool GREAT for this, which is interfaced with UCSC.

- On the top of the page, click on *Tools* > *Table browser*
- In the drop down menu, select *Custom track* and your track of interest
- check the Send output to GREAT checkbox
- Submit the dataset to GREAT

Things to consider:

- how are the peaks distributed with respect to the TSS ?
- Which functional terms are statistically enriched ?
- Redo the analysis for the siNT and siGATA3 and compare the enrichements.