June 2015

ChIP-Seq Exercise

Dena Leshkowitz, ***Bioinformatics Unit, WIS***

Introduction

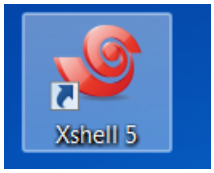
In this workshop we will learn how to analyse ChIP-Seq data. The data is taken from the article Dicken at al. Transcriptional reprogramming of CD11b+Esam(hi) dendritic cell identity and function by loss of Runx3. PLoS One. 2013 Oct 15;8(10). We will use two biological replicate ChIP-Seq experiments that were conducted for detection of Runx3-bound genomic regions using in-house anti Runx3 Ab and $30x10^6$ positive CD11c MACS isolated (Miltenyi Biotec) and classical dendritic cells (DC). For further details please see the article. The BAM files we will be using were done with only a subset of the reads sequenced in the study (1/10 of the reads).

General remark- the commands you need to type in the command line are *in italic* format.

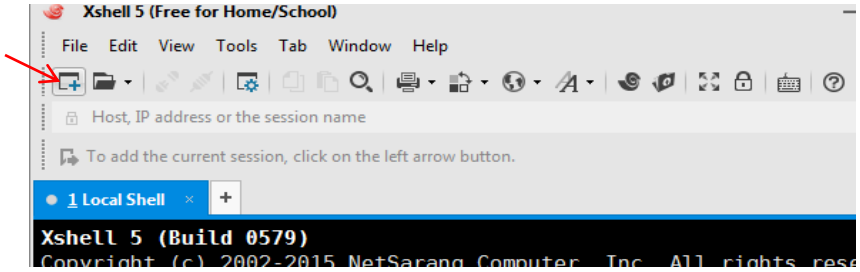
Instructions

1. Accessing the gladia server

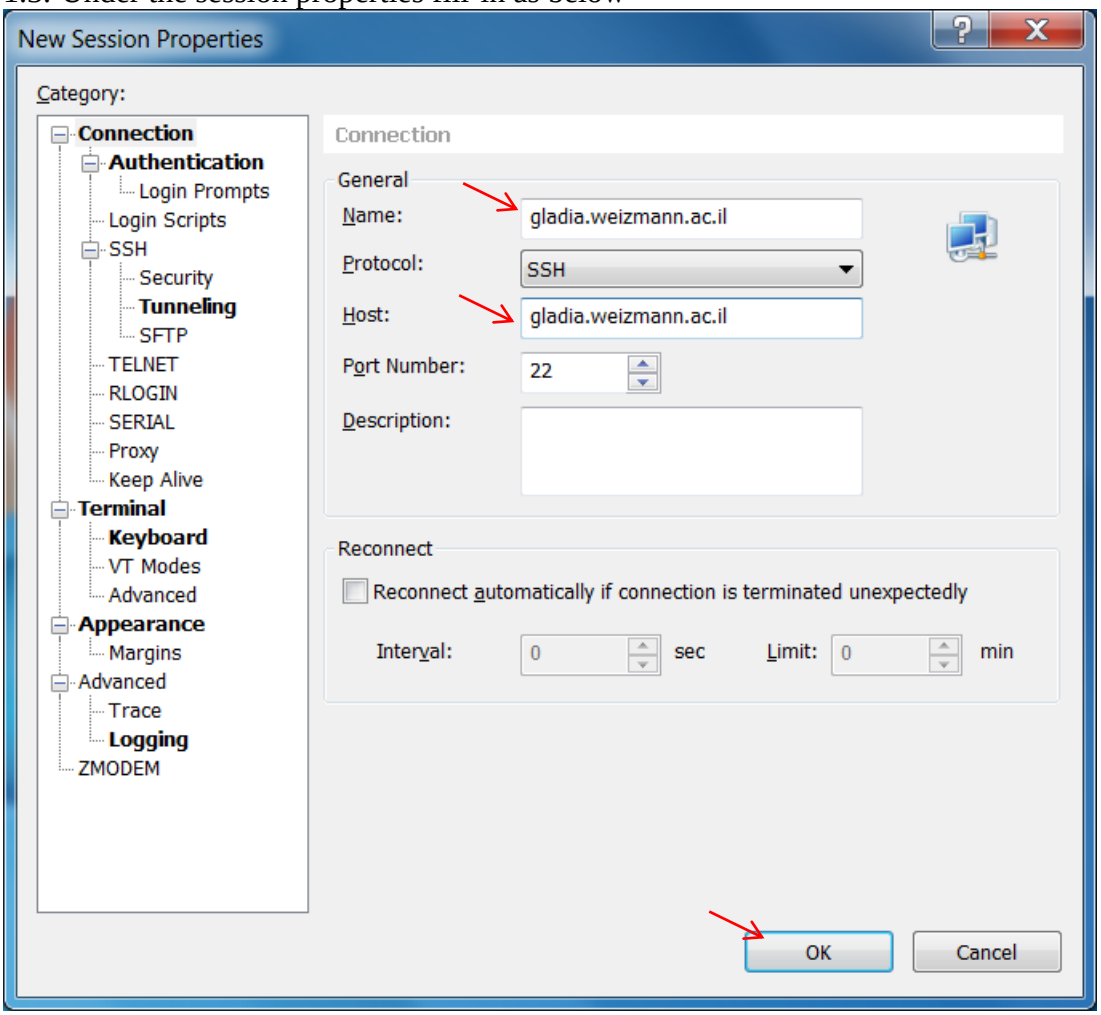
The data is found on a server named gladia, we will access this server using ssh with the tool Xshell. This will only work if you are using a computer within the Weizmann net.



1.1. Double Click on the desktop program icon Xshell5 (as above)
1.2. Open new session (see red arrow below)

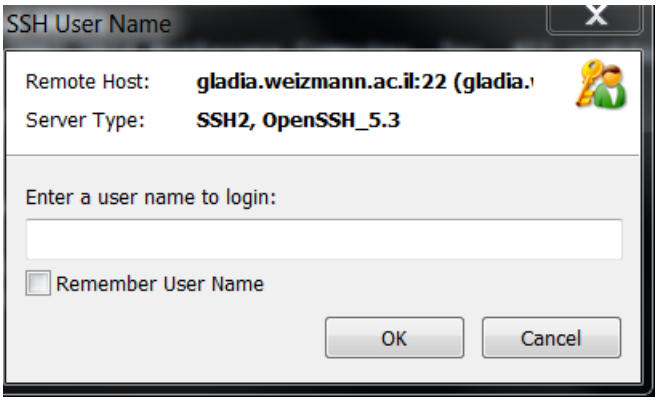


1.3. Under the session properties fill-in as below



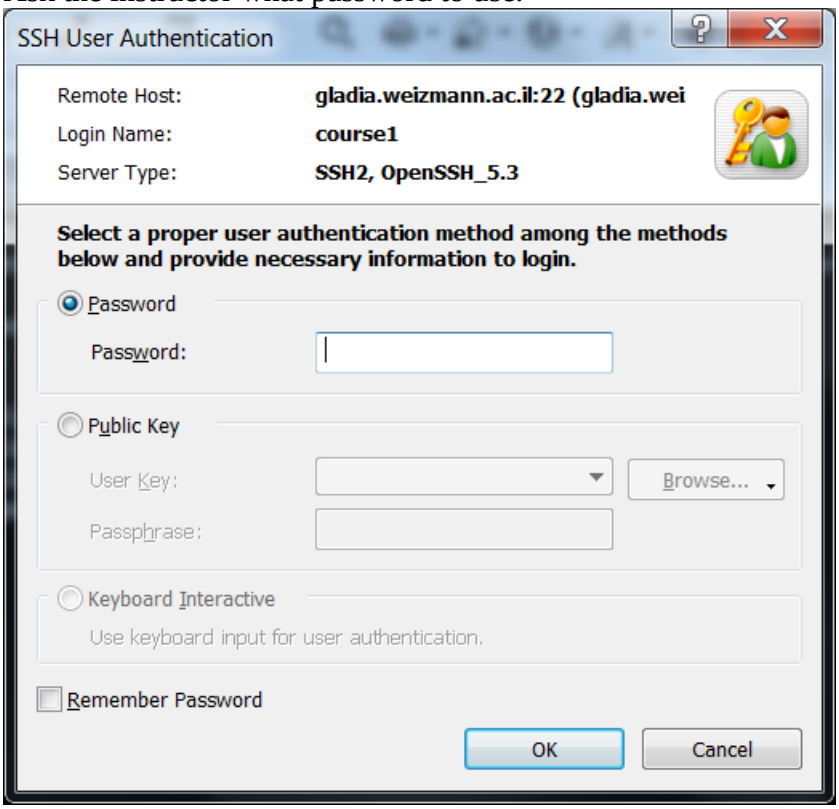
1.4. Name and hostname – gladia.weizmann.ac.il

1.5. Press the OK

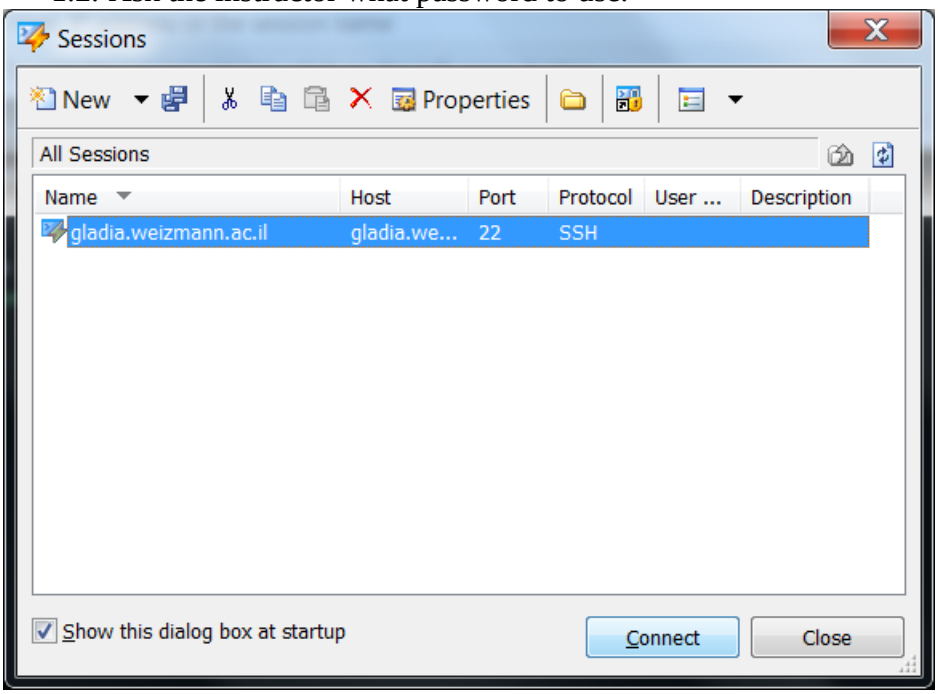


1.6. User's name - your user ID: course# (you need to know your number-ask us!)

1.1. Ask the instructor what password to use.

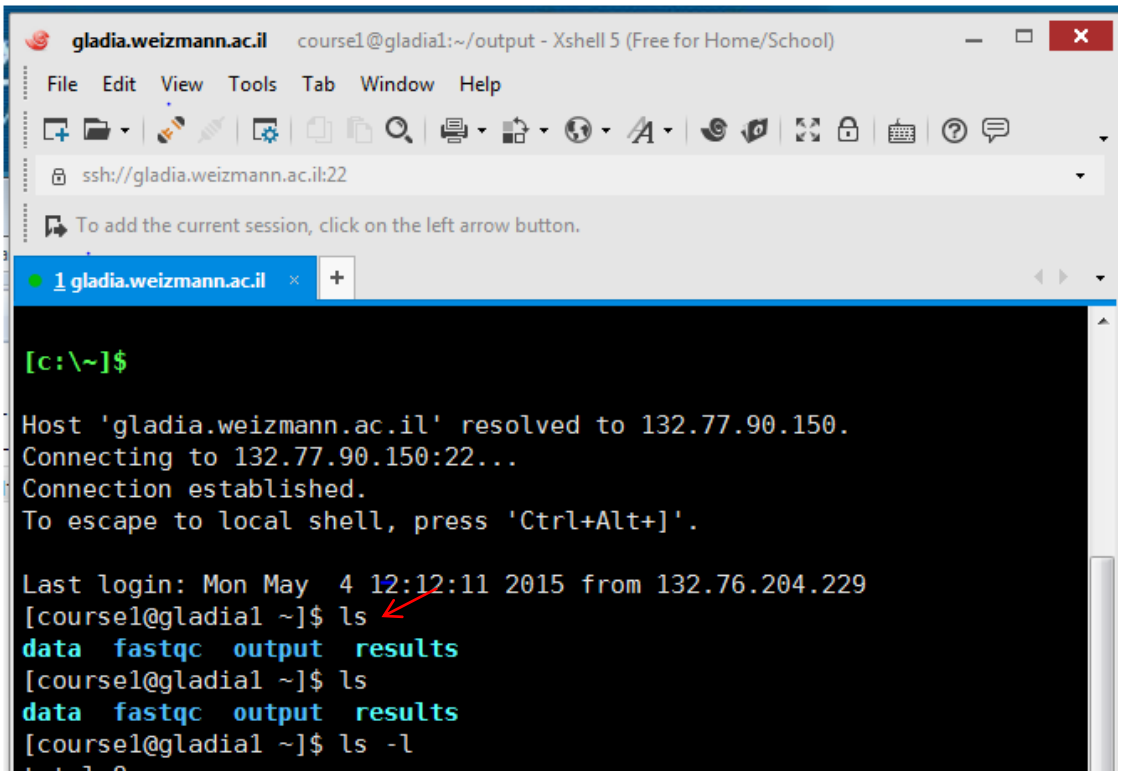


1.2. Ask the instructor what password to use.



You now should have connection to gladia with an ability to write commands in Linux.

2. Basic Unix/Linux commands



In order to see the files and folders in your home directory, type:

ls
ls –l
man ls

ls –l allows you to get more information on the files.

The *man* command gives you information on the command of interest.

To exit the manual,- type *q*.

For more commands, look at the supplementary section (#2).

Move to the output directory

cd output

3. Run MACS command

For more details on the tool look at the site https://github.com/taoliu/MACS/ (scroll down to see the read me).

Typing the macs command with the –h will give an explanation how to run this tool

macs2 callpeak -h

Following is the command to run macs on the first replicate (the sign ~ is a shortcut to my home directory)

macs2 callpeak -t ~/results/IP1.bam -c ~/results/input1.bam -q 0.01 -n macs2_input1_IP1

What is the predicted fragment size?

How many files were created by macs2?

How many reads (tags) were used for peak detection?

Does this number comply with the required number?

Within the outputs produced, one of the files gives us all the peaks detected in a text file, write the following to see the top lines in the file

head macs2_input1_IP1_peaks.bed

To find the number of peaks, count the number of lines in the peaks bed file

wc -l macs2_input1_IP1_peaks.encodePeak

Run macs2 analysis on the second replicate (both IP and input should be for the second samples).

How many peaks were detected in the second replicate experiment?

To find the amount of overlap between the two replicate peak files we will use the program intersectBed. This program comes from BEDtools suite

For the program explanation first type –

intersectBed

What does the option –wa do?

We will redirect this output of intersectBed to wc –l in order to count the lines of overlapping peaks.

intersectBed -wa -a macs2_input1_IP1_peaks.bed -b macs2_input2_IP2_peaks.bed | wc -l

What is the amount of overlap between the replicates?

For good reproducibility we expect to have 50% overlap.

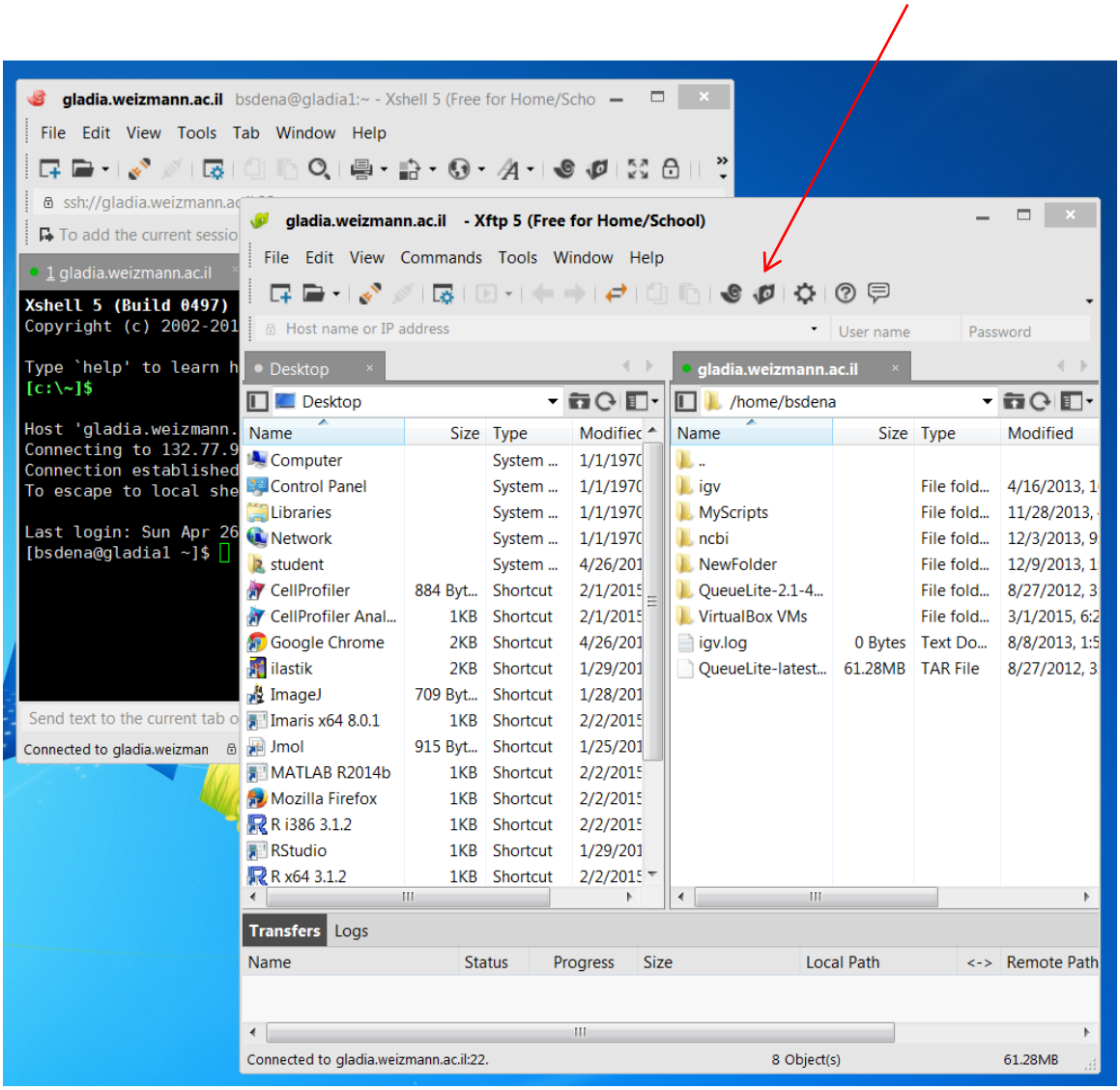
How can we improve the amount of reproducibility for this experiment?

4. Analyse peaks using CEAS: Enrichment of Genome Features

CEAS provides statistics on ChIP enrichment at important genome features such as specific chromosome, promoters, gene bodies, or exons, and infers genes most likely to be regulated by the binding factor.

ceas -b macs2_input1_IP1_peaks.bed --name=replicate1_ceas -g ../results/mm9.refGene

In order to view the ceas outputs (pdf and xls) you will need to open the file transfer.



Go to the **output** directory and click to open the file replicate1_ceas.pdf

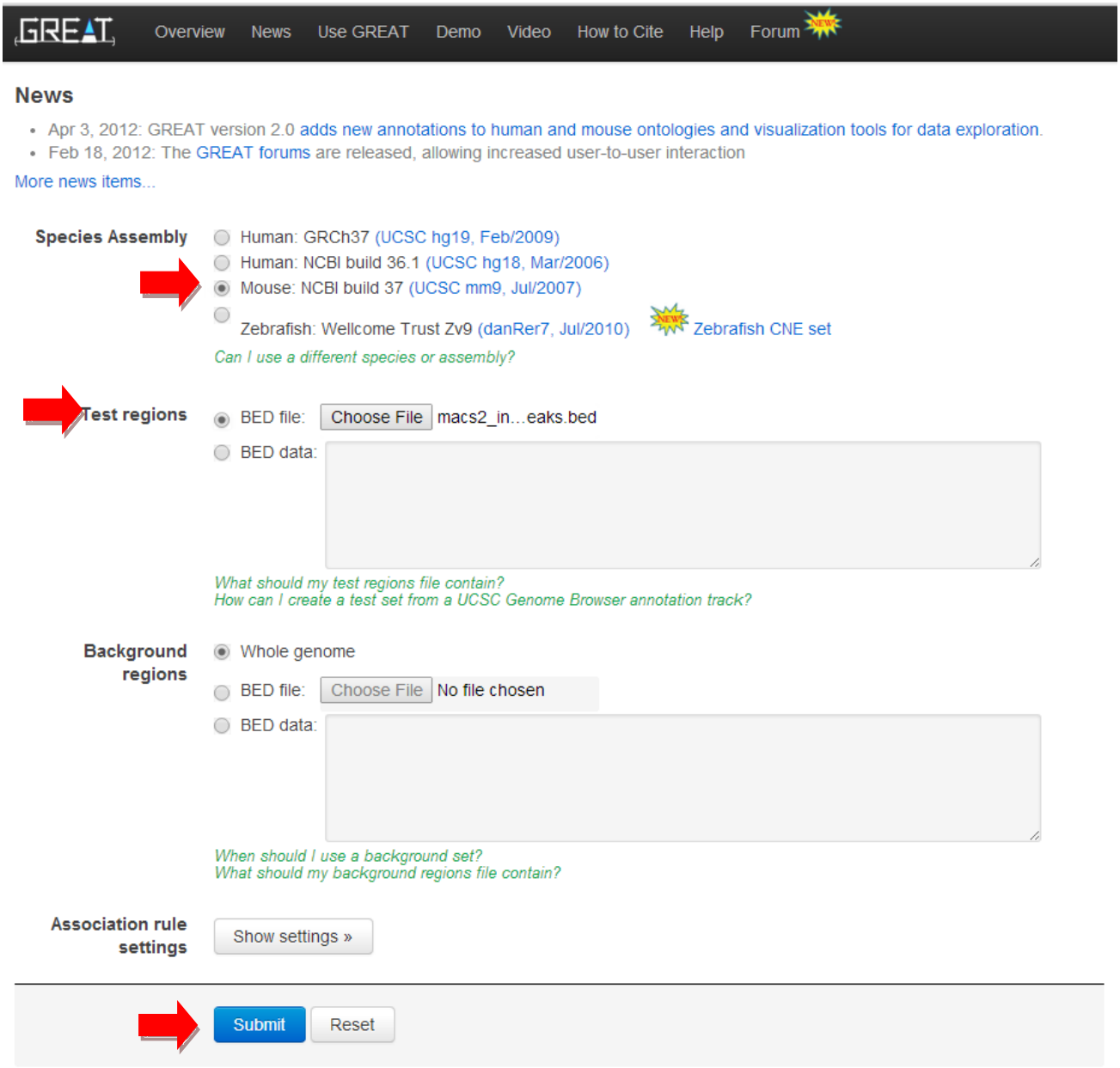
Does our transcription factor - Runx3 preferably bind to promoters?

5. Analyse Peaks using GREAT

The bed file of the peaks is not a "legal" bed file since the fifth column is not an integer and therefore we will extract the first four columns into a new edited bed file.

cut -f1-4 macs2_input2_IP2_peaks.bed > macs2_input2_IP2_peaks.edit.bed

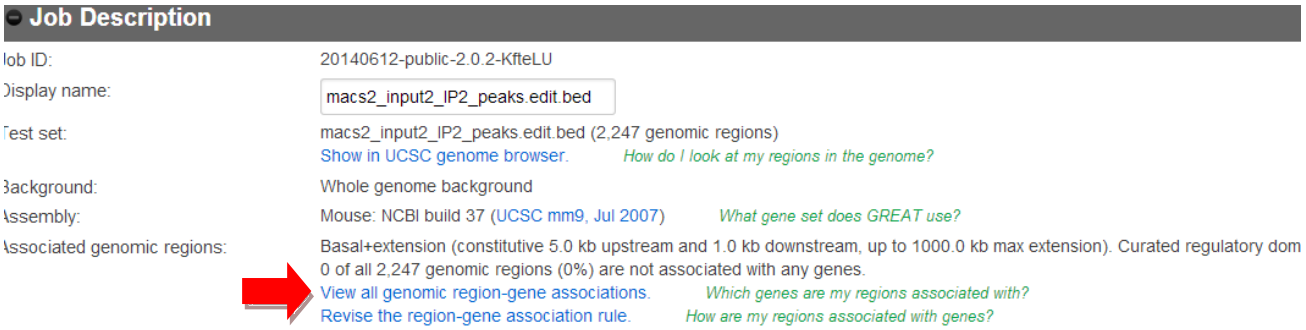
Create a folder on disk D and drag the edited peak bed to this folder. Open GREAT and upload this file. You should select the mm9 mouse built and click submit.



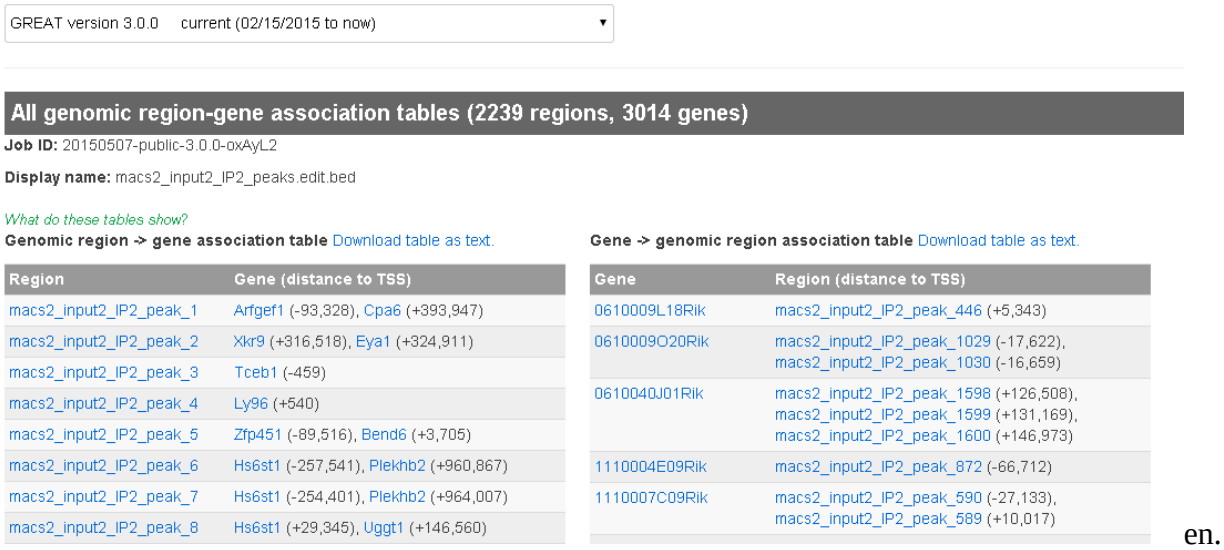
Without going to deep into the researched biological question, the DC are involved in immune reactions.
How many times is the word “immune” found in the enriched terms?

Looking at the Region-Gene Association Graphs – what is the most frequent binned distance from the peak to TSS?

We can view the genes associated with the peaks if we expand the **Jobs Description** (top of the report; press on the +)
Select - View **all genomic region-gene associations**.



The Following window will open -



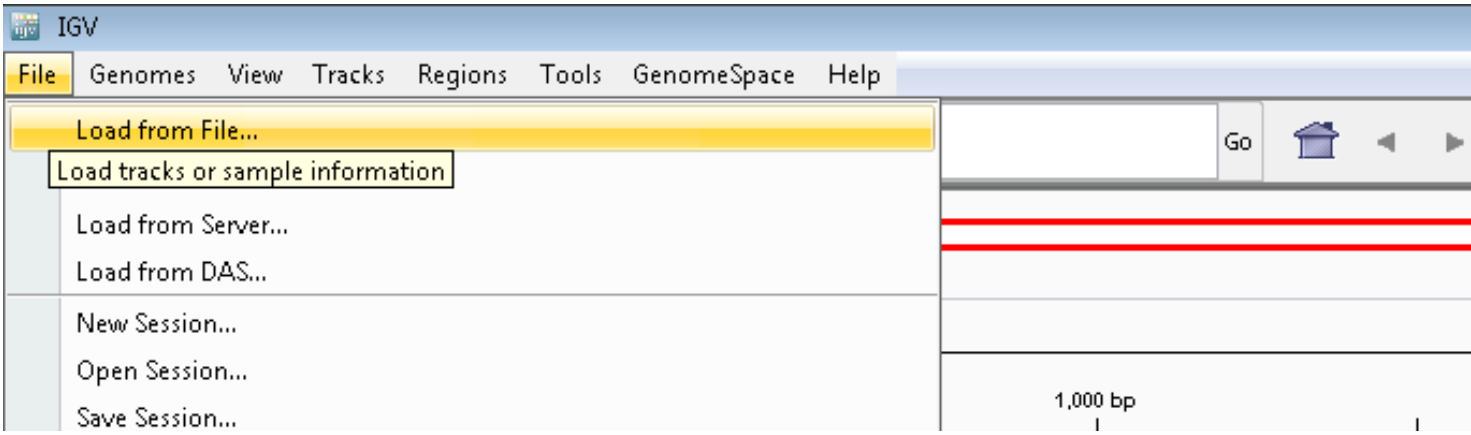
en.

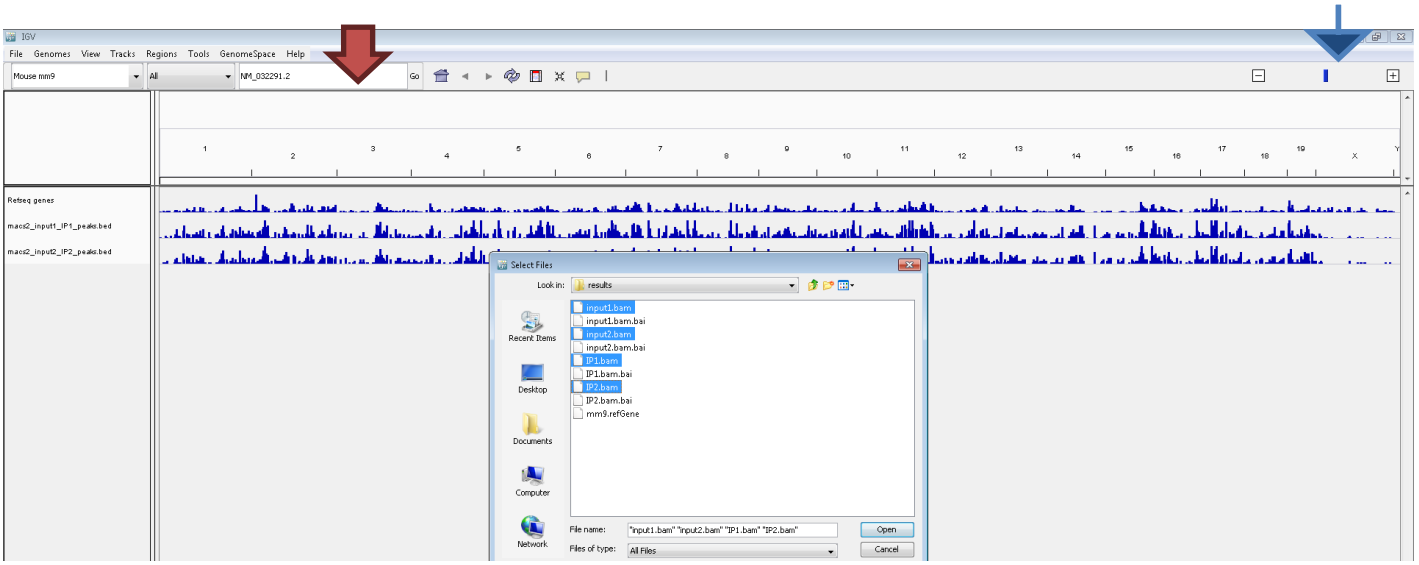
For the next assignment select a gene which has a peak next to it.

6. Browsing the peaks with a genome browser

We will use the Integrative genomics browser - IGV to view the mapped reads and the peaks. Open the IGV tool found on your desktop. Select run on the pop-up window. Once the application opened load the mm9 genome.

Move the alignment files from results directory (ending with bam and with bam.bai) and the bed files from output directory to your folder on D and load them to IGV as described below.





After loading the bam files you can insert the name of the gene you selected from GREAT and write it in the window where there is a red arrow in the picture above.

Where is the peak in regards to the gene (near TSS, upstream…)?
Does this peak occur in both the IP experiments?
Are there other peaks near this gene?
You can zoom out by clicking on the minus sign in the left corner (see blue arrow).

7. Loading to the browser outputs from the complete experiment

You have analysed BAM files that contained a reduced set of the original alignments (10%). In order to give you a flavor of how the peaks and the coverage are when using all the data and after joining the replicates, we have made the output of this analysis available for you. You will find in the "data" directory (above the output directory) the peak bed file and binary wiggle files (normalized shifted read coverage data). Download them and import them to the IGV browser. There are two wiggle files the "treat" is IP1-2_Runx3vsinput1-2_treat_pileup.bw and the "control" is IP1-2_Runx3vsinput1-2_control_lambda.bw. It is best to adjust the coverage of both wiggle files to the same scale, by right click on the wiggle file name and selecting "set data range" adjust the max value to 5. Do this for both wiggle files. You can go to the gene runx3, there is a very nice peak on the runx3 promoter region.
How many peaks were detected using all the data?

Supplementary

1. Basic Linux commands:

```
man (command) ..... shows help on a specific command
ls ................ show directory, in alphabetical order
logout ............. logs off system
mkdir .............. make a directory
rmdir .............. remove directory (rm -r to delete folders with files)
rm ................. remove files
cd ................. change current directory
more  .............. views a file, pausing every screenful
grep ............... search for a string in a file
head ............... show the first few lines of a file
tail ............... show the last few lines of a file
df ................. shows disk space available on the system
du ................. shows how much disk space is being used up by folders
chmod .............. changes permissions on a file
cut  ............... print selected parts of lines
cp   ............... copy file
mv   ............... move file
wc –l .............. print the number of lines
sort   ............. sort lines of text files
```