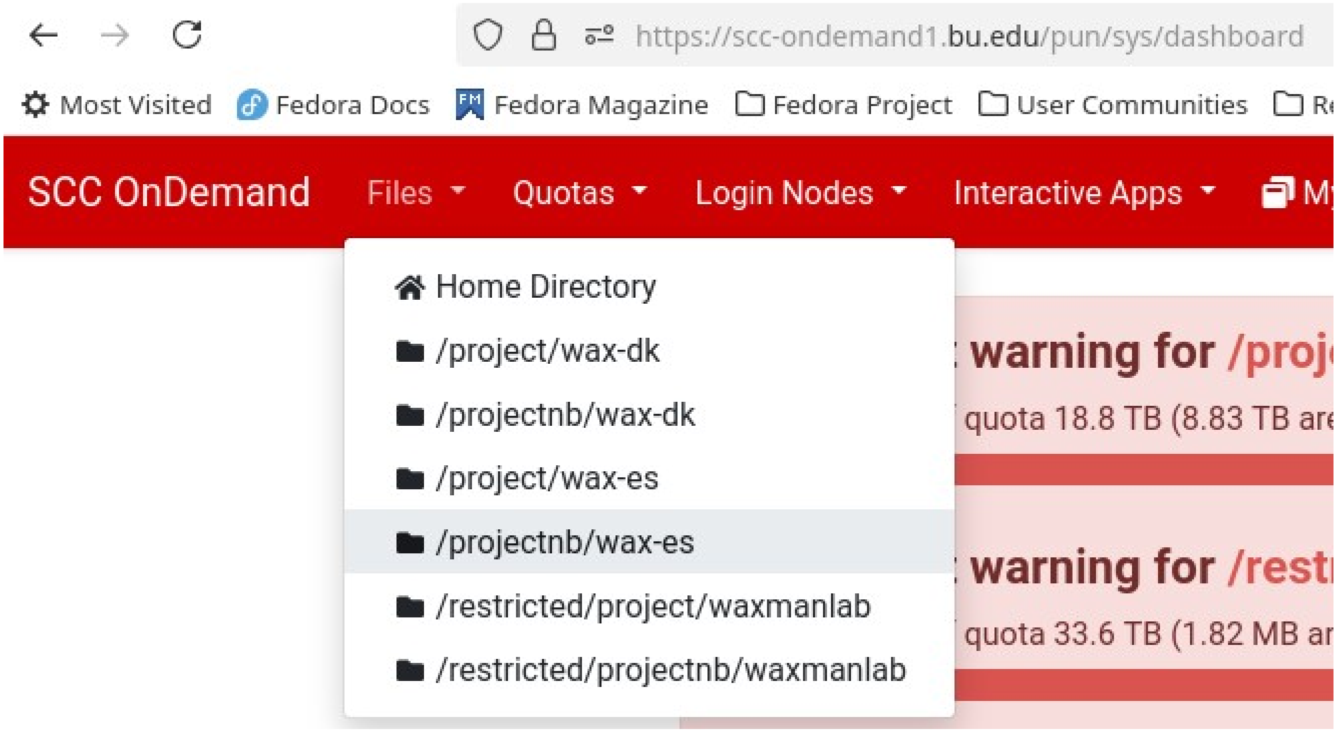
Enrichment analysis script (12/9/2024)

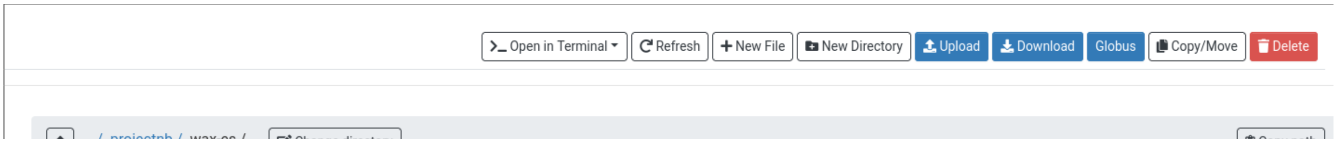
This script identifies enriched peaks in a foreground set of peak regions (set of genomic regions) (**foreground** directory) by comparing them to a set of background peaks (**background** directory) using overlap analysis calculated against a database (**bio_regions** directory) of known sets of peak regions. Enrichment is quantified using a Fisher's exact test, which assesses the statistical significance of the overlap between the foreground peaks and database peaks relative to the background peaks.

Creating directory for enrichment analysis

Login into BU SCC by visiting **scc-ondemand.bu.edu** website. After that, using **Files => /projectnb/wax-es** button on the top of the page to access the project main directory for **wax-es (projectnb)**.

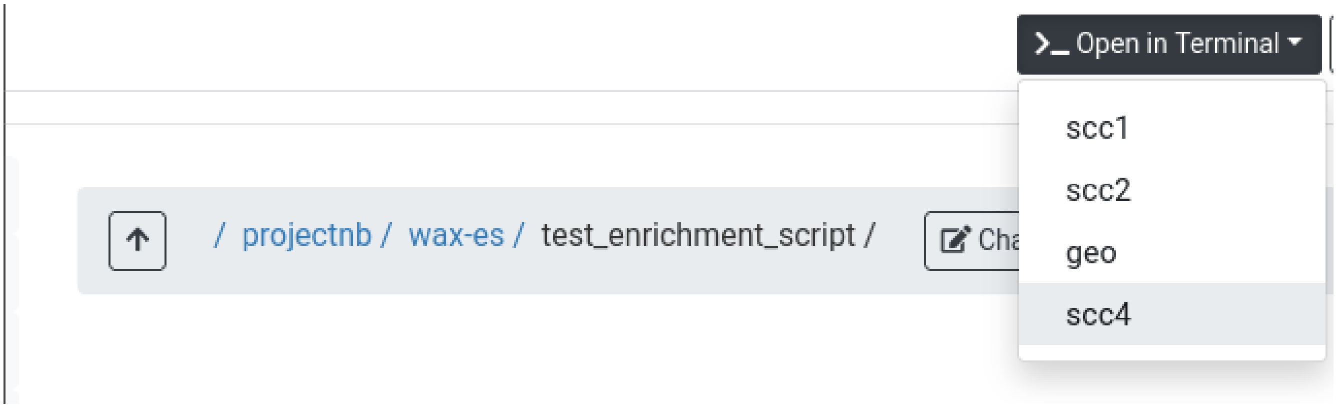


Next, select **your** Directory (or create a new Directory for your projects) using the commands on the top of the page.



Use the terminal to get a fresh copy of the Enrichment script

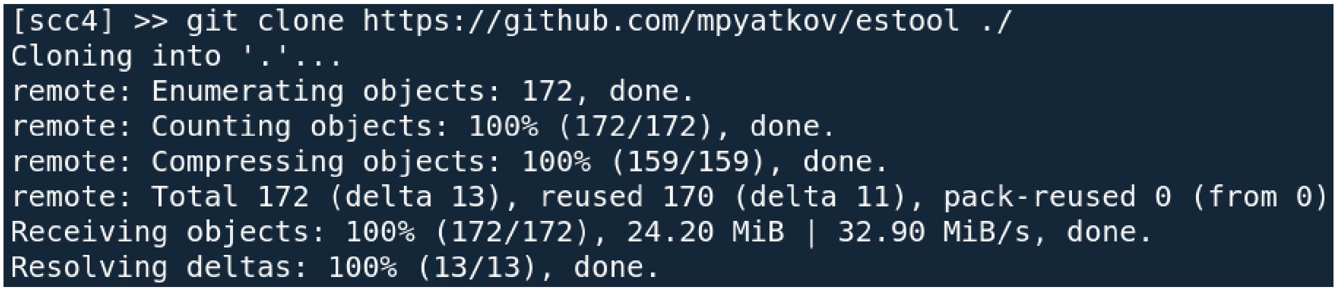
Example: I used the Directory **/projectnb/wax-es/test_enrichment_script** as the starting point for my analysis. To obtain the most recent copy of the enrichment script, open the terminal window inside the working directory. To do that just click "**Open in Terminal => scc4**" on the top of the page



Next, run the following command:

```
git clone https://github.com/mpyatkov/estool ./
```

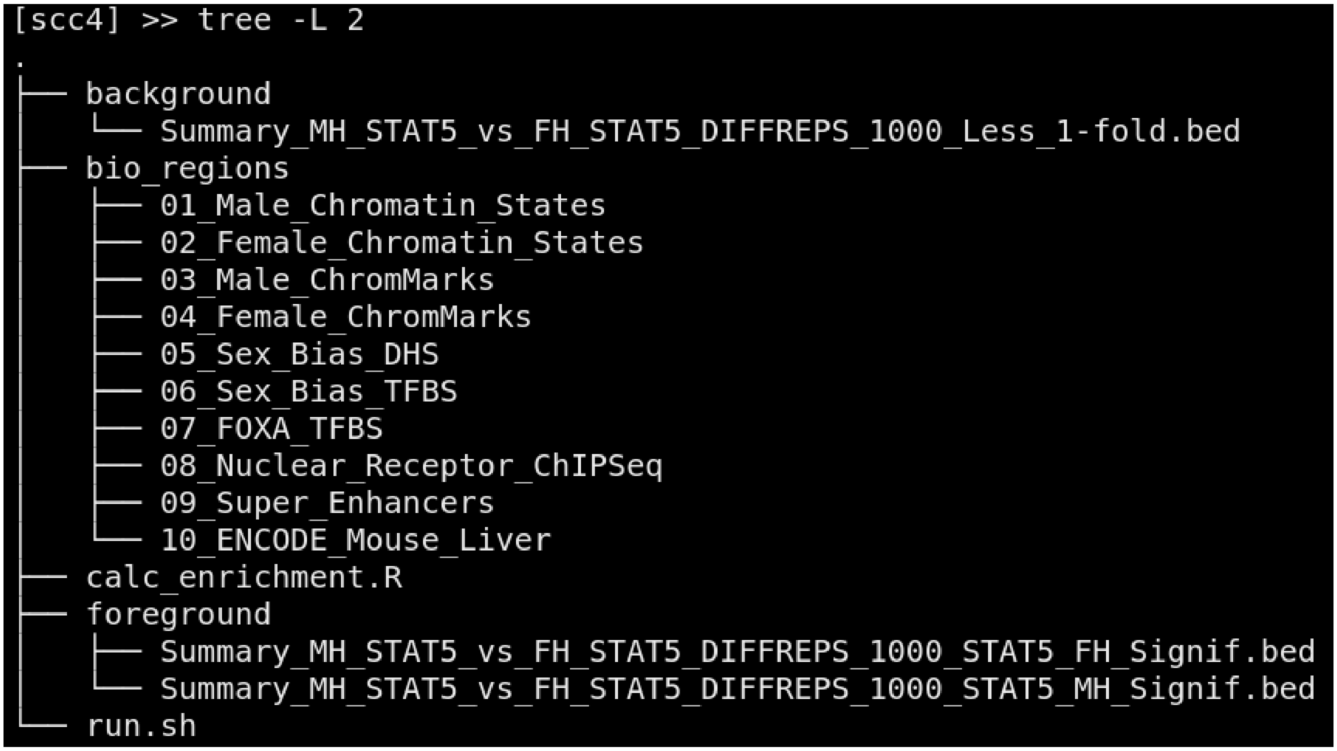
This command should copy all the required files/directories into the current directory



```
[scc4] >> git clone https://github.com/mpyatkov/estool ./
Cloning into '.'...
remote: Enumerating objects: 172, done.
remote: Counting objects: 100% (172/172), done.
remote: Compressing objects: 100% (159/159), done.
remote: Total 172 (delta 13), reused 170 (delta 11), pack-reused 0 (from 0)
Receiving objects: 100% (172/172), 24.20 MiB | 32.90 MiB/s, done.
Resolving deltas: 100% (13/13), done.
```

Structure of the directories

The script and supplementary directories have the following structure:



```
[scc4] >> tree -L 2
.
├── background
│   └── Summary_MH_STAT5_vs_FH_STAT5_DIFFREPS_1000_Less_1-fold.bed
├── bio_regions
│   ├── 01_Male_Chromatin_States
│   ├── 02_Female_Chromatin_States
│   ├── 03_Male_ChromMarks
│   ├── 04_Female_ChromMarks
│   ├── 05_Sex_Bias_DHS
│   ├── 06_Sex_Bias_TFBS
│   ├── 07_FOXA_TFBS
│   ├── 08_Nuclear_Receptor_ChIPSeq
│   ├── 09_Super_Enhancers
│   └── 10_ENCODE_Mouse_Liver
├── calc_enrichment.R
├── foreground
│   ├── Summary_MH_STAT5_vs_FH_STAT5_DIFFREPS_1000_STAT5_FH_Signif.bed
│   └── Summary_MH_STAT5_vs_FH_STAT5_DIFFREPS_1000_STAT5_MH_Signif.bed
└── run.sh
```

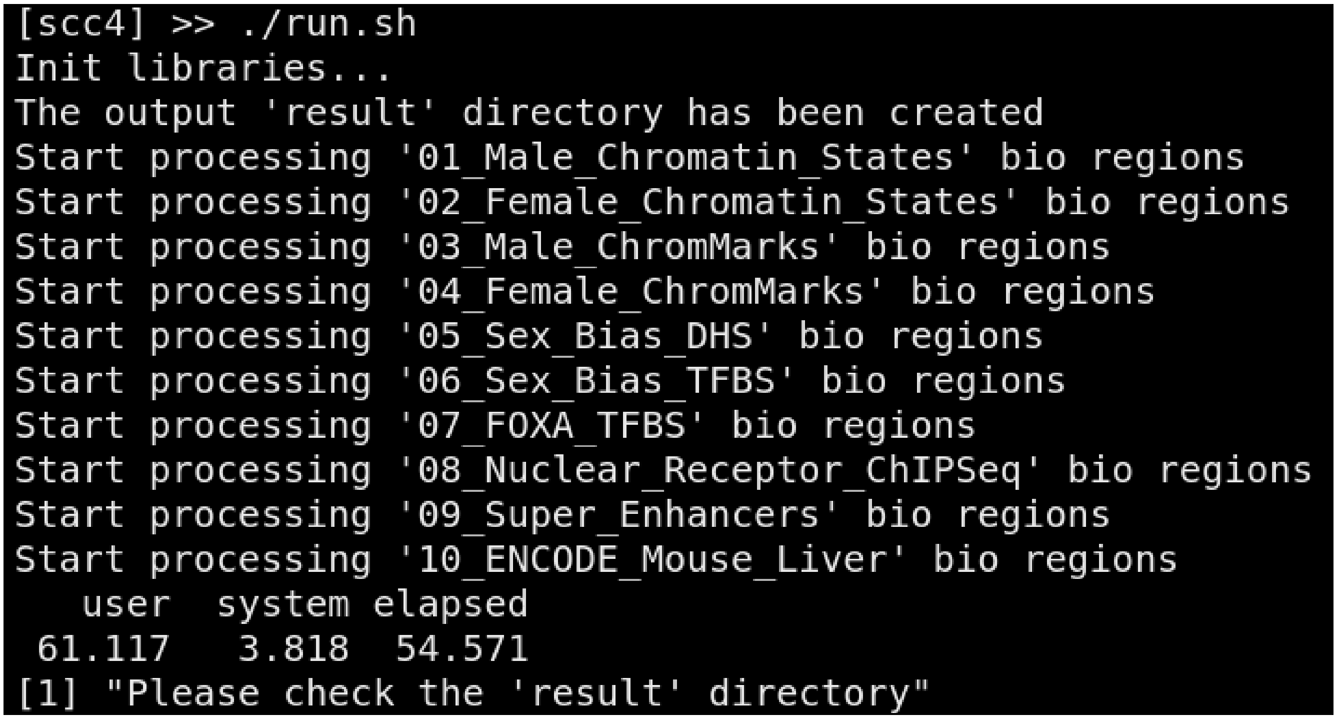
bio_regions – set of the directories which start with a number and contain multiple BED files. The input BED files with lists of peaks from the **foreground** and **background** directories will be tested against this database. The User needs to add to the **foreground** and **background** directories the sets of BED file(s) with peaks that they want to include in their analysis. Procedure to do that: Go back one level to the enclosing folder, then refresh the page using the **“Refresh”** button on the top. After the refreshing, the file browser will show the all the Enrichment Script folders and files that were copied to your working Directory. Navigate to the folder where you want to input your BED files to be used in the analysis: Background and Foreground. The User can now add the new BED files to each of these folders using the **"Upload"** button at the top. To delete the old files in those Directories (assuming you no longer need them for your analyses), the User can use the **"vertical three dots"** icon menu associated with the file/directory.

After downloading all the necessary files to the **foreground** and **background** directories, the user needs to return to the Terminal tab and run the script using the following command:

```
./run.sh
```

Script outputs

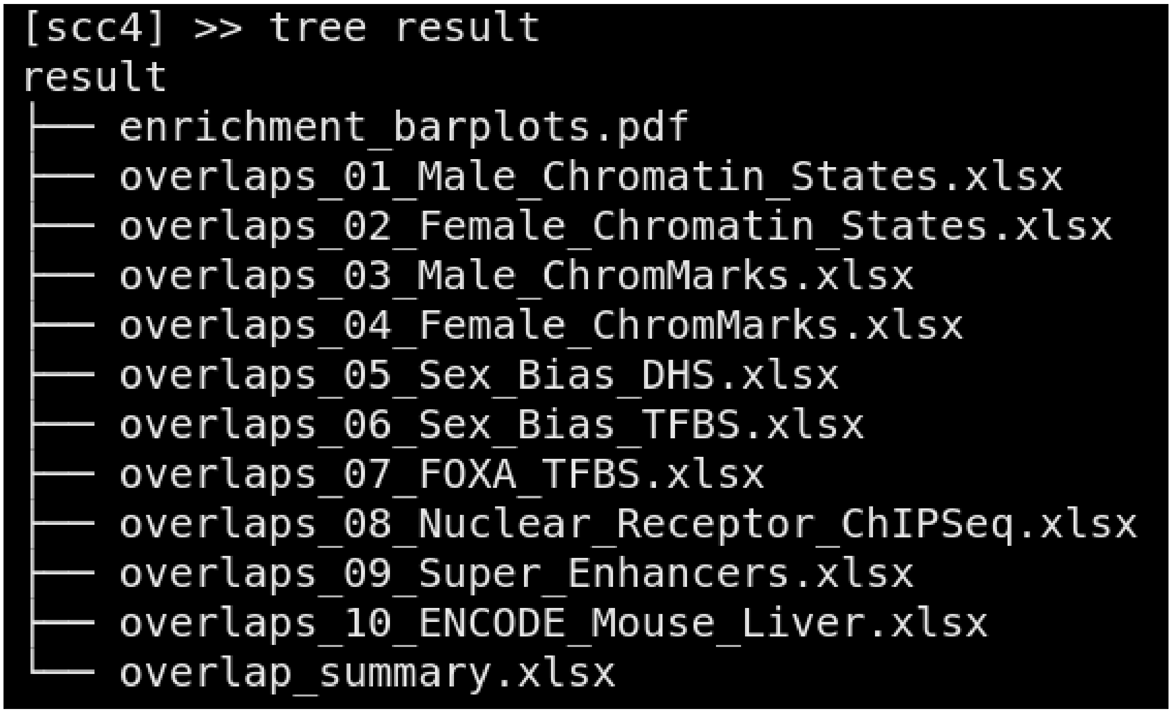
The enrichment script produces the following output (assuming the default configuration, where all **bio_regions** directories are included, and where the bioregion directories have not been modified)



```
[scc4] >> ./run.sh
Init libraries...
The output 'result' directory has been created
Start processing '01_Male_Chromatin_States' bio regions
Start processing '02_Female_Chromatin_States' bio regions
Start processing '03_Male_ChromMarks' bio regions
Start processing '04_Female_ChromMarks' bio regions
Start processing '05_Sex_Bias_DHS' bio regions
Start processing '06_Sex_Bias_TFBS' bio regions
Start processing '07_FOXA_TFBS' bio regions
Start processing '08_Nuclear_Receptor_ChIPSeq' bio regions
Start processing '09_Super_Enhancers' bio regions
Start processing '10_ENCODE_Mouse_Liver' bio regions
   user  system elapsed
 61.117   3.818  54.571
[1] "Please check the 'result' directory"
```

During processing, some output xlsx files may exceed Excel's row limits. In this case, the pipeline will not generate overlap files and will display a warning message **“Too many rows for category table. Skipping creating overlap files for this category”**

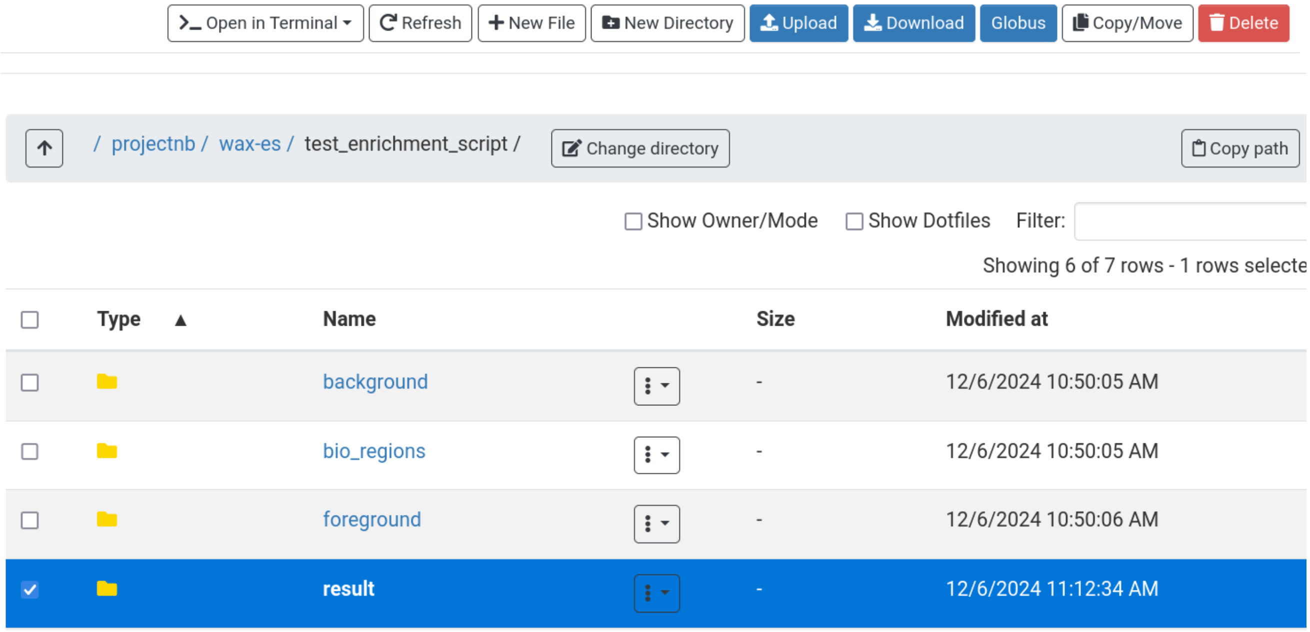
The **result** directory has the following structure and should contain PDF with barplots and .xlsx files:



```
[scc4] >> tree result
result
├── enrichment_barplots.pdf
├── overlaps_01_Male_Chromatin_States.xlsx
├── overlaps_02_Female_Chromatin_States.xlsx
├── overlaps_03_Male_ChromMarks.xlsx
├── overlaps_04_Female_ChromMarks.xlsx
├── overlaps_05_Sex_Bias_DHS.xlsx
├── overlaps_06_Sex_Bias_TFBS.xlsx
├── overlaps_07_FOXA_TFBS.xlsx
├── overlaps_08_Nuclear_Receptor_ChIPSeq.xlsx
├── overlaps_09_Super_Enhancers.xlsx
├── overlaps_10_ENCODE_Mouse_Liver.xlsx
└── overlap_summary.xlsx
```

Downloading data from SCC

To download data from SCC just select checkbox near the required directory and click **Download** button on the top of the page



Showing 6 of 7 rows - 1 rows selecte

□	Type ▲	Name	Size	Modified at
□		background	-	12/6/2024 10:50:05 AM
□		bio_regions	-	12/6/2024 10:50:05 AM
□		foreground	-	12/6/2024 10:50:06 AM
☑		result	-	12/6/2024 11:12:34 AM

Technical notes

By default, the Enrichment script utilizes the Singularity image with all preinstalled R packages. The image located inside the **/projectnb/wax-es/routines/singularity** directory. If you would like to use the script on your local computer, you will need to *uncomment* the code lines on the top of the **calc_enrichment.R** script related to installation of packages locally.