Differential TF activity (diffTF) Tool Pipeline for MS

INSTALLING diffTF PROGRAMS

*note: any file highlighted in ***bold*** is provided

1. Install the following programs:
    1. Snakemake
        i. $ conda install – c conda-forge mamba
        ii. $ mamba create -c conda-forge -c bioconda -n snakemake
        iii. $ conda activate snakemake
        iv. $ snakemake --help

    2. Subread
        i. Download the Subread source package
            1. (snakemake) [prvaldes@tscc-login11 Subread]$ wget https://sourceforge.net/projects/subread/files/subread-2.0.1/subread-2.0.1-source.tar.gz
        ii. Uncompress the source package
            1. (snakemake) [prvaldes@tscc-login11 Subread]$ tar zxvf subread-2.0.1-source.tar.gz
        iii. Build the Makefile
            1. (snakemake) [prvaldes@tscc-login11 src]$ make -f Makefile.Linux
            2. # Installation successfully completed.
            3. # Generated executables were copied to directory ../bin/ Location: /home/prvaldes/programs/Subread/subread-2.0.1-source/bin

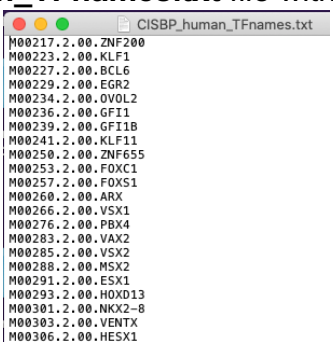
II. Clone the git respository
    1. (snakemake) [prvaldes@tscc-login11 ~]$ git clone https://git.embl.de/grp-zaugg/diffTF.git

CREATING HUMAN CIS-BP TFBS DATABASE

III. Create the cisBP TFBS .tar.gz file using the Find Individual Motif Occurrences (FIMO) from the MEME suite (https://meme-suite.org/meme/doc/fimo.html)
    1. Download 1,078 motifs from FIMO using the **CIS-BP_FIMO_allmotifs.sh** script
    2. Convert motif FIMO files in gff3 format to .bed format using the **CIS-BP_FIMO_convert2bed.sh** script
    3. Format the files names for each .bed file -> {TF}_TFBS.bed using the **CISBP_human_diffTF_nomenclature.sh** script

IV. Download and format the cisBP Transcription factor binding site (TFBS) database called CIS-BP_MEME_TFBS_human.tar.gz (n = 1,078 total TFs):
    1. (base) [prvaldes@tscc-3-12 CIS-BP_MEME_TFBS_human]$ tar -zxvf CIS-BP_MEME_TFBS_human.tar.gz

2. Replace underscores with periods (.) using the **CISBP_human_diffTF_replace_underscores.sh** script made
3. Remove any non Homosapien TF's manually (end product is n = 988 total TFs)
4. Create the **CISBP_human_TFnames.txt** file with the list of CISBP TF's


CISBP_human_TFnames.txt

```
M00217.2.00.ZNF200
M00223.2.00.KLF1
M00227.2.00.BCL6
M00229.2.00.EGR2
M00234.2.00.OVOL2
M00236.2.00.GFI1
M00239.2.00.GFI1B
M00241.2.00.KLF11
M00250.2.00.ZNF655
M00253.2.00.FOXC1
M00257.2.00.FOXS1
M00260.2.00.ARX
M00266.2.00.VSX1
M00276.2.00.PBX4
M00283.2.00.VAX2
M00285.2.00.VSX2
M00288.2.00.MSX2
M00291.2.00.ESX1
M00293.2.00.HOXD13
M00301.2.00.NKX2-8
M00303.2.00.VENTX
M00306.2.00.HESX1
```

5. Format the files names for each .bed file -> {TF}_TFBS.bed using the **CISBP_human_diffTF_nomenclature.sh** script
6. Indicate which .bed files are empty .bed files found (n = 65 total TFs found)
    i. (base) [prvaldes@tscc-3-12 CIS-BP_MEME_TFBS_human]$ find /home/prvaldes/scratch/diffTF/cisBP/CIS-BP_MEME_TFBS_human -type f -empty -print >> /home/prvaldes/scratch/diffTF/cisBP/CIS-BP_MEME_TFBS_human/empty.txt
7. Remove the empty .bed files found in the empty.txt file using the **CISBP_human_diffTF_remove.sh** file (n = 923 total TFs to use)
8. Change the following TFBS names with hyphens to dots:
    i. (snakemake) [prvaldes@tscc-login12 CIS-BP_MEME_TFBS_human]$ mv M00301.2.00.NKX2-8_TFBS.bed M00301.2.00.NKX2.8_TFBS.bed
    ii. (snakemake) [prvaldes@tscc-login12 CIS-BP_MEME_TFBS_human]$ mv M00320.2.00.NKX2-5_TFBS.bed M00320.2.00.NKX2.5_TFBS.bed
    iii. (snakemake) [prvaldes@tscc-login12 CIS-BP_MEME_TFBS_human]$ mv M05012.2.00.NKX3-2_TFBS.bed M05012.2.00.NKX3.2_TFBS.bed
    iv. (snakemake) [prvaldes@tscc-login12 CIS-BP_MEME_TFBS_human]$ mv M05042.2.00.NKX2-3_TFBS.bed M05042.2.00.NKX2.3_TFBS.bed
    v. (snakemake) [prvaldes@tscc-login12 CIS-BP_MEME_TFBS_human]$ mv M05242.2.00.NKX6-3_TFBS.bed M05242.2.00.NKX6.3_TFBS.bed
    vi. (snakemake) [prvaldes@tscc-login12 CIS-BP_MEME_TFBS_human]$ mv M05255.2.00.NKX3-1_TFBS.bed M05255.2.00.NKX3.1_TFBS.bed
    vii. (snakemake) [prvaldes@tscc-login12 CIS-BP_MEME_TFBS_human]$ mv M05558.2.00.BORCS8-MEF2B_TFBS.bed M05558.2.00.BORCS8.MEF2B_TFBS.bed

V. Create TF-gene translation table for cisBP called **translationTable_cisBP.csv** file using Excel

SYMBOL ENSEMBL HOCOID
ZNF200 ENSG00000010539 M00217.2.00.ZNF200
KLF1 ENSG00000105610 M00223.2.00.KLF1
BCL6 ENSG00000113916 M00227.2.00.BCL6
EGR2 ENSG00000122877 M00229.2.00.EGR2
OVOL2 ENSG00000125850 M00234.2.00.OVOL2

VI. Convert double type raw RNA-seq counts to integer only raw RNA-seq counts, called **RNA-seq-counts2.tsv** using the **diffTF_Convert_to_Integers.R** script file

OBTAINING THE HUMAN (hg38) REFERENCE GENOME

VII. Download human (hg38) reference genome to use in the diffTF program
1. (base) [prvaldes@tscc-login11 hg38]$ wget ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz
2. (base) [prvaldes@tscc-login11 hg38]$ gunzip GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz

RUNNING diffTF for the APP^{V717I} vs. NDC comparison

VIII. Create a tab-separated file that summarizes the input data for APP^{V717I} and NDC samples called **sampleData_V717I.tsv**

IX. Create the configuration file (config.json) that defines various parameters of the pipeline APP^{V717I} vs. NDC (**config_V717I_cisBP.json**)

X. Create the **startAnalysis_V717I_cisBP.sh** script file

XI. Run diffTF analysis for APP^{V717I} mutation vs. NDC using the **startAnalysis_V717I_cisBP.sh** script file
1. (base) [prvaldes@tscc-login12 CIS-BP_MEME_bed]$ qsub -I -q condo -l walltime=8:00:00 -l nodes=2:ppn=24:mem128
2. qsub: waiting for job 24760541.tscc-mgr7.local to start
3. qsub: job 24760541.tscc-mgr7.local ready
4. (base) [prvaldes@tscc-2-58 input]$ conda activate snakemake
5. (snakemake) [prvaldes@tscc-2-58 input]$ sh startAnalysis_V717I_cisBP.sh

XII. When the program stops with the following error (also do for $PSEN1^{A79V}$ vs. NDC and $PSEN2^{N141I}$ vs. NDC comparisons):

```
[Wed Feb  3 17:35:41 2021]
Error in rule filterSexChromosomesAndSortPeaks:
    jobid: 7
    output: /home/prvaldes/scratch/diffTF/output_V717I_cisBP/PEAKS/NDCvsV717I.all.consensusPeaks.filtered.sorted.bed
    shell:

            grep ^chr /home/prvaldes/scratch/diffTF/output_V717I_cisBP/TEMP/NDCvsV717I.all.consensusPeaks.bed | grep -v "^chrX\|^chrY\|^chrM\|^chrUn"  | sort -k1,1
 -k2,2n > /home/prvaldes/scratch/diffTF/output_V717I_cisBP/PEAKS/NDCvsV717I.all.consensusPeaks.filtered.sorted.bed

        (one of the commands exited with non-zero exit code; note that snakemake uses bash strict mode!)

Removing output files of failed job filterSexChromosomesAndSortPeaks since they might be corrupted:
/home/prvaldes/scratch/diffTF/output_V717I_cisBP/PEAKS/NDCvsV717I.all.consensusPeaks.filtered.sorted.bed
```

rerun the startAnalysis_V717I_cisBP.sh script under the TSCC login node where Subread is stored(/home/prvaldes/programs/Subread/subread-2.0.1-source/bin) using 128GB of RAM.

XIII. When finished the program output for **startAnalysis_V717I_cisBP.sh** should look like this:

```
[Fri Feb  5 19:20:09 2021]
Finished job 0.
943 of 943 steps (100%) done
Complete log: /oasis/tscc/scratch/prvaldes/diffTF/input/.snakemake/log/2021-02-05T140024.367313.snakemake.log


################################
#  Workflow finished, no error  #
# Check the FINAL_OUTPUT folder #
################################



Running time in minutes: 327.9
```

RUNNING diffTF for the PSEN1^{A79V} vs. NDC comparison

XIV. Create a tab-separated file that summarizes the input data for *PSEN1^{A79V}* and NDC samples called **sampleData_A79V.tsv**

XV. Create the configuration file (config.json) that defines various parameters of the pipeline *PSEN1^{A79V}* vs. NDC (**config_A79V_cisBP.json**)

XVI. Create the **startAnalysis_A79V_cisBP.sh** script file

XVII. Run diffTF analysis for *PSEN1^{A79V}* mutation vs. NDC using the **startAnalysis_A79V_cisBP.sh** script file

1. (base) [prvaldes@tscc-login2 ~]$ qsub -I -q condo -l walltime=8:00:00 -l nodes=2:ppn=16:mem256:sandy (make sure to increase memory usage here from 128GB to 256GB of RAM under this job run)
2. qsub: waiting for job 24830344.tscc-mgr7.local to start
3. qsub: job 24830344.tscc-mgr7.local ready
4. (base) [prvaldes@tscc-1-9 ~]$ conda activate snakemake
5. (snakemake) [prvaldes@tscc-1-9 input]$ sh startAnalysis_A79V_cisBP.sh

XVIII. When finished the program output for **startAnalysis_A79V_cisBP.sh** should look like this:

```
[Mon Feb  8 13:53:18 2021]
Finished job 0.
3 of 3 steps (100%) done
Complete log: /oasis/tscc/scratch/prvaldes/diffTF/input/.snakemake/log/2021-02-08T125619.552948.snakemake.log


################################
#  Workflow finished, no error  #
# Check the FINAL_OUTPUT folder #
################################



Running time in minutes: 57.0
```

RUNNING diffTF for the PSEN2^{N141I} vs. NDC comparison

I. Create a tab-separated file that summarizes the input data for *PSEN2^{N141I}* and NDC samples called **sampleData_N141I.tsv**

II. Create the configuration file (config.json) that defines various parameters of the pipeline *PSEN1^{A79V}* vs. NDC (**config_N141I_cisBP.json**)

III. Create the **startAnalysis_N141I_cisBP.sh** script file

IV. Run diffTF analysis for *PSEN2^{N141I}* mutation vs. NDC using the **startAnalysis_N141I_cisBP.sh** script file

1. (base) [prvaldes@tscc-login12 ~]$ qsub -I -q condo -l walltime=8:00:00 -l nodes=2:ppn=16:mem256:sandy (make sure to increase memory usage here from 128GB to 256GB of RAM under this job run)
2. qsub: waiting for job 24840443.tscc-mgr7.local to start
3. qsub: job 24840443.tscc-mgr7.local ready
4. (base) [prvaldes@tscc-1-9 ~]$ conda activate snakemake
5. (snakemake) [prvaldes@tscc-1-9 input]$ sh startAnalysis_N141I_cisBP.sh

XIX. When finished the program output for **startAnalysis_N141I_cisBP.sh** should look like this:

```
[Tue Feb  9 23:33:25 2021]
Finished job 0.
365 of 365 steps (100%) done
Complete log: /oasis/tscc/scratch/prvaldes/diffTF/input/.snakemake/log/2021-02-09T172039.394927.snakemake.log


################################
#  Workflow finished, no error  #
# Check the FINAL_OUTPUT folder #
################################



Running time in minutes: 376.8
```