Day 10 Worksheet | Nascent Sequencing : Tfit

Author: Rutendo F. Sigauke (Adapted from Margaret Gruca SR2019)

Overview

This worksheet assumes the samples have gone through QC processing. We will learn how to annotate our nascent data using FStitch and Tfit. The bidirectional regions called can be used as input into downstream analysis pipelines. We will cover the basics of Motif Displacement Score (MDS) and Transcription Factor Enrichment Analysis (TFEA).

Preprocessing alignment files for FStitch and Tfit

- Filter out low-quality reads and multi mapping reads from bam files
- Generate input bedgraph coverage files

Introduction to Tfit

Below is the basic command you will need to use to run Tfit. You will need to specify a few files:

- Specify path to Tfit executable
- Add path to the Tfit configuration file
- Add path to input bedgraph files

The basic argument to run Tfit is:

```
$ mpirun -np 1 -host ${SLURM_JOB_NODELIST} tfit_path bidir
-config tfit_config -ij sample.bedGraph -N sample_id -o
output_dir
```

Tfit returns four output files:

- `sample_id.log`
    - Log file with summaries of the run
- `sample_id_prelim_bidir_hits.bed`
    - Preliminary bidirectional calls
- `sample_id_bidir_predictions.bed`
    - Final bidirectional regions called Tfit
- `sample_id_models_MLE.tsv`
    - Specific parameter values for all bidirectionals

For each predicted bidirectional, information about the center (`mu`), strand bias (`pi`), pausing probability (`omega`) and variance (`sigma`) are outputted in the `sample_id_bidir_predictions.bed` and `sample_id_bidir_predictions.bed` files.

```
mu : center of the bidirectional transcript
sigma : the variance in RNA polymerase II loading
lambda : the entry length or amount of skew
pi : the strand bias, probability of forward strand data point
omega : the pausing probability, how much bidirectional signal to
elongation/noise signal
```

Running Tfit

1. Log into the AWS
2. Create a working directory for day10 in `scratch`

```
$ mkdir -p /scratch/Users/<YourUsername>/day10
```

3. Create a scrpts directory inside day10

```
$ mkdir -p /scratch/Users/<YourUsername>/day10/scripts
```

```
[rutendos@ip-172-31-38-192 ~]$ mkdir -p /scratch/Users/rutendos/day10
[rutendos@ip-172-31-38-192 ~]$ mkdir -p /scratch/Users/rutendos/day10/scripts
```

```
[rutendos@ip-172-31-38-192 ~]$ ls /scratch/Users/rutendos/day10/
scripts
```

4. Moving and editing Tfit script

```
$ cp
/scratch/Shares/public/sread2021/scripts/day10/Tfit_ex
ample.sbatch
/scratch/Users/<YourUsername>/day10/scripts
```

```
[rutendos@ip-172-31-38-192 scripts]$ ls
[rutendos@ip-172-31-38-192 scripts]$ cp /scratch/Shares/public/sread2021/scripts/
day10/Tfit_example.sbatch /scratch/Users/rutendos/day10/scripts
[rutendos@ip-172-31-38-192 scripts]$ ls
Tfit_example.sbatch
```

5. Edit the script with paths to your respective day10 scratch directories

```
#!/bin/bash
#SBATCH -p compute
#SBATCH --mail-type=ALL # Mail events (NONE, BEGIN, END, FAIL, ALL)
#SBATCH --mail-user=rusi2317@colorado.edu # Where to send mail
#SBATCH --nodes=1 # Run on a single node
#SBATCH --ntasks=4     # Number of CPU (processer cores i.e. tasks)
#SBATCH --mem=5gb # Memory limit
#SBATCH --time=00:15:00 # Time limit hrs:min:sec
#SBATCH --output=/scratch/Users/rutendos/eofiles/dmso_rep1.chr1_Tfit.%j.out
#SBATCH --error=/scratch/Users/rutendos/eofiles/dmso_rep1.chr1_Tfit.%j.err
#SBATCH --job-name=Tfit_run # Job name

#load modules
module load mpi/openmpi-x86_64
module load gcc/7.1.0
module load bedtools/2.25.0

#Initiate paths and variables
outdir=/scratch/Users/rutendos/day10
rootname=dmso_rep1.chr21
bdgraphdir=/scratch/Shares/public/sread2021/cookingShow/day8/bedgraph_groseq

############################################################
################## print job info ##########################

printf "\nfastq Directory: $INDIR"
printf "\nOutput Directory: $OUTDIR"
printf "\nRun on: $(hostname)"
printf "\nRun from: $(pwd)"
printf "\nScript: $0\n"
printf "\nYou've requested $SLURM_CPUS_ON_NODE core(s).\n"
date

export OMP_NUM_THREADS=4
echo $OMP_NUM_THREADS
echo $SLURM_JOB_NODELIST

#make output directories
mkdir -p $outdir
mkdir -p ${outdir}/tfit_out

tfit_outdir=${outdir}/tfit_out
bg_file=${bdgraphdir}/${rootname}.bedGraph

#Initiate Tfit executables
Tfitdir=/scratch/Shares/public/sread2021/algorithms/Tfit_2017
src=${Tfitdir}/src/Tfit
config_file=${Tfitdir}/config_files/config_file.txt

#=================================================================
#calling tfit command
cmd="mpirun -np 1 -host ${SLURM_JOB_NODELIST}"

$cmd $src bidir -ij $bg_file -o $tfit_outdir -N ${rootname} -MLE 1
#=================================================================
```

6. Run the tfit script (the script should be done running in 3 minutes).

```
$ sbatch Tfit_example.sbatch
```

Your output files will be in
`/scratch/Users/<YourUsername>/day10/tfit_out`

7. Once the run is done, we can visualize the called regions in IGV
    a. Open X2GO
    b. Load hg38 genome
    c. Load bedgraph file from
    `/scratch/Shares/public/sread2021/cookingShow/day8/bedgraph_groseq/dmso_rep1.chr21.bedGraph`
    d. Load Tfit bidirectionals from
    `/scratch/Users/<YourUsername>/day10/tfit_out/dmso_rep1.chr21-1_`**`prelim_bidir_hits`**`.bed`

    `/scratch/Users/<YourUsername>/day10/tfit_out/dmso_rep1.chr21-1_`**`bidir_predictions`**`.bed`