Strand-seq analysis Day5 scNOVA

Hyobin Jeong

July 14, 2021

Contents

1 Introduction

In this practical session, we will analyze 77 cells of T-ALL P1 which has subclonal chromothripsis event.

```
## Download the scNOVA pipeline from the github and setup R environment
git lfs install
git clone https://github.com/jeongdo801/scNOVA.git

## We need to specify Renvironment for R_conda
mkdir /g/korbel2/StrandSeq/Test_HJ/R_conda
vi ~/.Renviron
R_LIBS=/g/korbel2/StrandSeq/Test_HJ/R_conda
```

2 Preparation of input files for scNOVA

After copying the pipeline, we need to put input files to the input bam and input user folder. Inside of input bam, you need to copy good quality bam files from selected folder of mosaicatcher. In the input user folder, 1) Add key result files from mosaicatcher output in the input user folder 2) Add the subclonality information 3) Add the genes within copy number changed region to mask in the infer differential expression result, if it's not provided, genes will not be masked.

```
## Change the project name in the Snakefile
cd scNOVA
vi Snakefile

## Create input_bam folder, copy bam and index files
mkdir input_bam
cp /scratch/jeong/pipeline_20190625/bam/TALL03-DEA5/selected/*.bam input_bam
cp /scratch/jeong/pipeline_20190625/bam/TALL03-DEA5/selected/*.bai input_bam

## Create input_user folder, prepare input
mkdir input_user
cp /scratch/jeong/pipeline_20190625/strand_states/TALL03-DEA5/  \
100000_fixed_norm.selected_j0.1_s0.5_scedist20/strandphaser_output.txt input_user

cp /scratch/jeong/pipeline_20190625/sv_calls/TALL03-DEA5/100000_fixed_norm.selected_j0.1_s0.5_scedist20/  \
simpleCalls_llr4_poppriorsTRUE_haplotagsFALSE_gtcutoff0.05_regfactor6_filterTRUE.txt input_user

cp input_user_example/input_subclonality_TALL03_DEA5_77cells.txt input_user/input_subclonality.txt

cp input_user_example/input_SV_affected_genes_DEA5.txt input_user/input_SV_affected_genes.txt
```

3 Install dependencies

scNOVA implements multiple R packages in the pipeline, we will install dependencies using conda environment.

```
## Let's go inside the pipeline folder
vi ~/.condarc
pkgs_dirs:
  - /g/korbel2/jeong/envs/pkgs/
envs_dirs:
  - /g/korbel2/jeong/envs/

snakemake -j 4 --use-conda --conda-create-envs-only
```

4 Running scNOVA

Now we are ready to execute scNOVA pipeline

```
## Let's go inside of the folder with output files from plotting pipeline
sbatch -t 90:00:00 -N 1 -n 1 --mem=50000 --mail-type=FAIL,BEGIN,END \
--mail-user=hyobin.jeong@embl.de -o output.txt ./run_pipeline.sh
```

5 Explore the output folders and plots

After the function has finished, you will find several output files including **result plot** which are the most frequently used outputs for the downstream analysis.

```
##Let's explore output files
Main output (DEA5 example data)

Heatmap and UMAP visualization of significant hits : result_plots/Result_scNOVA_plots_TALL3_DEA5.pdf

Infer differential expression table : result/Result_scNOVA_infer_expression_table.txt

Single-cell level NO table : result/TALL3_DEA5_sort_geneid.txt
```

6 Session Info

```
toLatex(sessionInfo())
```

- R version 4.0.3 (2020-10-10), `x86_64-apple-darwin17.0`
- Locale: `en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8`
- Running under: `macOS High Sierra 10.13.3`
- Random number generation:
- RNG: `Mersenne-Twister`
- Normal: `Inversion`
- Sample: `Rounding`
- Matrix products: default

- BLAS:
  `/System/Library/Frameworks/Accelerate.framework/Versions/A/Frameworks/vecLib.framework/Versions/A/libBL`
- LAPACK:
  `/Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib`
- Base packages: base, datasets, graphics, grDevices, methods, parallel, stats, stats4, utils
- Other packages: BiocGenerics 0.36.1, breakpointR 1.8.0, breakpointRdata 1.8.0, cowplot 1.1.1, GenomeInfoDb 1.26.7, GenomicRanges 1.42.0, IRanges 2.24.1, knitr 1.33, S4Vectors 0.28.1
- Loaded via a namespace (and not attached): assertthat 0.2.1, beachmat 2.6.4, Biobase 2.50.0, BiocManager 1.30.16, BiocNeighbors 1.8.2, BiocParallel 1.24.1, BiocSingular 1.6.0, BiocStyle 2.18.1, Biostrings 2.58.0, bitops 1.0-7, bluster 1.0.0, codetools 0.2-18, colorspace 2.0-2, compiler 4.0.3, crayon 1.4.1, DBI 1.1.1, DelayedArray 0.16.3, DelayedMatrixStats 1.12.3, digest 0.6.27, doParallel 1.0.16, dplyr 1.0.7, dqrng 0.3.0, edgeR 3.32.1, ellipsis 0.3.2, evaluate 0.14, fansi 0.5.0, foreach 1.5.1, generics 0.1.0, GenomeInfoDbData 1.2.4, GenomicAlignments 1.26.0, ggplot2 3.3.5, glue 1.4.2, grid 4.0.3, gtable 0.3.0, gtools 3.9.2, highr 0.9, htmltools 0.5.1.1, igraph 1.2.6, irlba 2.3.3, iterators 1.0.13, lattice 0.20-44, lifecycle 1.0.0, limma 3.46.0, locfit 1.5-9.4, magrittr 2.0.1, Matrix 1.3-4, MatrixGenerics 1.2.1, matrixStats 0.59.0, munsell 0.5.0, pillar 1.6.1, pkgconfig 2.0.3, purrr 0.3.4, R6 2.5.0, Rcpp 1.0.6, RCurl 1.98-1.3, rlang 0.4.11, rmarkdown 2.9, Rsamtools 2.6.0, rsvd 1.0.5, scales 1.1.1, scran 1.18.7, scuttle 1.0.4, SingleCellExperiment 1.12.0, sparseMatrixStats 1.2.1, statmod 1.4.36, stringi 1.6.2, stringr 1.4.0, SummarizedExperiment 1.20.0, tibble 3.1.2, tidyselect 1.1.1, tinytex 0.32, tools 4.0.3, utf8 1.2.1, vctrs 0.3.8, xfun 0.24, XVector 0.30.0, yaml 2.2.1, zlibbioc 1.36.0