Strand-seq analysis Part1 scTRIP using Mosaicatcher

Hyobin Jeong

June 29, 2021

Contents

1 Introduction

In this practical session, we will analyze 47 cells of GM20509 from HGSVC which has unexpected heterogeneity.

```
## Copy the Mosaicatcher pipeline (for the practical session)
cd /scratch/name_abcd/
mkdir Practical_singlecell
cd Practical_singlecell
cp /g/korbel2/StrandSeq/Test_HJ/pipeline_20190625.tar.gz ./
tar -xvzf pipeline_20190625.tar.gz
cd pipeline_20190625/
```

2 Preparation of input files for Mosaicatcher

After copying the pipeline, we need to put input files to the 'bam' folder Inside of 'bam', you need to make two subfolders 'all' and 'selected'. You can copy the bam files to those folder, but to save the space we will bind the files rather than copying the files to this folder.

```
## Copy the folder containing bam files
cd bam
mkdir all
mkdir selected
cp /g/korbel2/StrandSeq/20180726_TALL03-DEA5/bam/*.bam all
cp /g/korbel2/StrandSeq/20180726_TALL03-DEA5/bam/*.bai all

cp /g/korbel2/StrandSeq/Test_HJ/copy_cells.pl ./
vi selected_cells.txt
perl copy_cells.pl
```

3 snv sites to genotype

As single-cell sequencing can be sparse to call single-nucleotide variants (SNV), we need the help of known heterozygous SNV sites to get more accurate haplotype information. Those snv sites information is in the 'snv sites to genotype' folder. If the index file is older than main file, it can make error during the run, so that we need to make sure the time points of copied files.

```
## Let's go inside the snv_sites_to_genotype folder
cd ../

## Check the time points the files were generated
ls -lh snv_sites_to_genotype

## If the index file is same or older than main file, we can try following code
cp snv_sites_to_genotype/*.tbi ./
cp *.tbi snv_sites_to_genotype
rm *.tbi
```

4 Running Mosaicatcher

Now we are ready to execute mosicatcher pipeline

```
## Let's go inside of the folder with output files from plotting pipeline
cd /scratch/name_abcd/Practical_singlecell/pipeline_20190625
chmod +x run_pipeline_singularity.sh
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_singularity.sh


##If you encounter permission error, following will help to resolve it.

-bash-4.2$ more run_pipeline_singularity.sh
#!/bin/bash

# Set these two paths to link large external data (reference genomes) to the respective places within the ima
REF="/scratch/name_abcd/Practical_WT/practical/utils/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna"
R_REF="/scratch/name_abcd/single_sequences.2bit"

snakemake \
    -j 6 \
    --configfile Snake.config-singularity.json \
    --use-singularity \
    --singularity-args "-B ${REF}:/reference.fa:ro \
                        -B ${REF}.fai:/reference.fa.fai:ro \
                        -B ${R_REF}:/usr/local/lib/R/site-library/BSgenome.Hsapiens.UCSC.hg38/extdata/single_
    --latency-wait 60 \
    --printshellcmd
```

5 Explore the output folders and plots

After the function has finished, you will find several output files including **sv call plots**, **sv call list**, and **haplotype info** which are the most frequently used for the downstream analysis. There are two different parameter settings which are stringent and lenient. For general purpose stringent setting is recommended by default, and to find single-cell rearrangement lenient setting is recommended.

```
##Two different parameter settings
stringent: simpleCalls_llr4_poppriorsTRUE_haplotagsFALSE_gtcutoff0.05_regfactor6_filterTRUE.txt
lenient: simpleCalls_llr4_poppriorsTRUE_haplotagsTRUE_gtcutoff0_regfactor6_filterFALSE.txt
```

- **sv call plots** These output plots show the sv call result with different visualizations.

```
##Single-cell SV detection result for each chromosome separately
sv_calls/RPE/100000_fixed_norm.selected_j0.1_s0.5_scedist20/plots/sv_calls

##Variant allele frequency and genomic positions of all the potential SVs
sv_calls/RPE/100000_fixed_norm.selected_j0.1_s0.5_scedist20/plots/sv_consistency

##Heatmap of SVs clustered by log likelihood score of SVs
```

```
sv_calls/RPE/100000_fixed_norm.selected_j0.1_s0.5_scedist20/plots/sv_clustering
```

- **sv call list** These lists can be found from two different folders.

```
##List of SVs after merging consecutive SVs (to prevent oversegmentation)
postprocessing/merge/RPE/100000_fixed_norm.selected_j0.1_s0.5_scedist20/

##List of SVs before merging
sv_calls/RPE/100000_fixed_norm.selected_j0.1_s0.5_scedist20/
```

- **haplotype info** Additionally you can get VCF files of haplotyping result based on your Strand-seq data.

```
##Haplotyping result VCF files of particular chromosome (for example chr1)
strand_states/RPE/100000_fixed_norm.selected_j0.1_s0.5_scedist20/StrandPhaseR_analysis.chr1/VCFfiles/

##Strand state of every chromosome and single cells
strand_states/RPE/100000_fixed_norm.selected_j0.1_s0.5_scedist20/strandphaser_output.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