***Part1. Work flow***

Last update : 20211223

**Graphical user interface

Description automatically generated**

**Figure1.** Computational pipeline of scNOVA. The scNOVA method employs single cell tri-channel processing (scTRIP)[4](https://paperpile.com/c/kosGbt/NuKN), as realised by the MosaiCatcher software, for haplotype-aware somatic SV discovery of Del, Dup, Inv, InvDup, translocation, and complex SV events. Different Modules of scNOVA have been developed to enable single-cell mulitomics of these somatic SVs.

***Part2. Overall structure***

**Workflow management:**

https://github.com/jeongdo801/scNOVA

snakemake/v5.3.1

**Pre-required step: Preparation of single-cell genetic information:**

Plotting pipeline: https://git.embl.de/meiers/strand-seq-pipeline

Mosaicatcher: https://github.com/friendsofstrandseq/mosaicatcher-pipeline

**Input files:**

1. Single-cell bam files from Strand-seq pipeline: input\_bam/\*.bam
2. Clone assignment: Input\_subclonality.txt
3. CN\_normalization files
4. SV affected genes : input\_user/input\_SV\_affected\_genes.txt

**Module1: Extract NO in gene-bodies and regulatory elements**

1. SAMtools/1.3.1-foss-2016b
2. biobambam2/2.0.76-foss-2016b
3. deeptools/2.5.1-foss-2016b-Python-2.7.12

**Module2: Infer gene expression (CNN + DESeq2)**

1. CNN

perl/v5.16.3

Python/3.7.4-GCCcore-8.3.0

cuDNN/7.6.4.38-gcccuda-2019b

CUDA/10.1.243-GCC-8.3.0

TensorFlow/1.15.0-fosscuda-2019b-Python-3.7.4

scikit-learn/0.21.3-foss-2019b-Python-3.7.4

matplotlib/3.1.1-foss-2019b-Python-3.7.4

R version 4.0.0 (2020-04-24)

R packages : pracma

1. DESeq2

R version 4.0.0 (2020-04-24)

R packages : DESeq2, matrixStats, pheatmap, gplots, umap, Rtsne, factoextra

1. chromVAR

R version 4.0.0 (2020-04-24)

chromVAR, nabor, motifmatchr

**Main output files**

1. Single-cell heatmap of inferred differentially expressed genes between clones
2. Table of probability of expression from CNN (binary classification mode)
3. Table of CNN filtered DESeq2 result

**Other output files**

1. Single-cell diagnostic plot for tSNE and UMAP
2. Single-cell NO matrix for genes

***Part3. Practical aspect***

**Location of pipelines in the server**

Plotting pipeline: /g/korbel2/StrandSeq/Test\_HJ/Plotting\_pipeline.tar.gz

Mosaicatcher: /g/korbel2/StrandSeq/Test\_HJ/pipeline\_20190625.tar.gz

Mosaicatcher(mm10):

scNOVA: /g/korbel2/StrandSeq/Test\_HJ/pipeline\_scNOVA\_20201223.tar.gz

1. Installation: install required R packages in the R 4.0.0version
2. Create input\_bam folder
3. Create input\_user folder
4. Change the project name in the Snakefile
5. Launch the run\_pipeline.sh script (seneca)

sbatch -t 7-00:00:00 -N 1 -n 1 --mem=100000 --mail-type=FAIL,BEGIN,END --mail-user=ID@embl.de -o output.txt ./run\_pipeline.sh

**Example data set and output (GM20509)**

Plotting pipeline: /g/korbel2/StrandSeq/20200109\_U24/20200225\_GM20509B

Mosaicatcher: /g/korbel/jeong/pipeline\_20190625\_GM20509

scNOVA: /g/korbel2/jeong/pipeline\_scNOVA\_20201223\_GM20509

**Optional: Tutorials for plotting and mosaicatcher**

<https://github.com/jeongdo801/SV_practical_computational>

**Optional: Tutorials for scNOVA (in development)**

<https://github.com/jeongdo801/scNOVA>

***Part4. Data sets available***

**RPE-1 WT vs. RPE-1 BM510 (p53 ko)**

scNOVA: /g/korbel2/jeong/pipeline\_scNOVA\_20201223 \_RPE1\_scNOVA01

**RPE-1 WT vs. RPE-1 C7 (another transformed cell line)**

scNOVA: /g/korbel2/jeong/pipeline\_scNOVA\_20201223 \_RPE1\_scNOVA02

**RPE-1**

Fastq:

/g/korbel/shared/data/others/StrandSeq/runs/2018-06-06-H37V3AFXY

**BM510**

Fastq: /g/korbel/shared/data/others/StrandSeq/runs/2018-07-13-H3NKVAFXY

/g/korbel/shared/data/others/StrandSeq/runs/2018-02-20-HWFL7AFXX

**C7**

Fastq: /g/korbel/shared/data/others/StrandSeq/runs/2018-02-13-HWCYNAFXX/ \*\_Raeder\_lane1C7x03PE20\*\*\*\_1\_sequence.txt.gz

/g/korbel/shared/data/others/StrandSeq/runs/2018-02-13-HWCYNAFXX/ \*\_Raeder\_lane1C7x03PE20\*\*\*\_2\_sequence.txt.gz

/g/korbel/shared/data/others/StrandSeq/runs/2018-05-25-H2KT2AFXY

**LCL (GM20509, aka NA20509)**

Fastq: /g/korbel/shared/data/others/StrandSeq/runs/2020-02-25-H2NCTAFX2/\*\_Hasenfeld\_lane1GM20509Bx01PE20\*\*\*\_1\_sequence.txt.gz

/g/korbel/shared/data/others/StrandSeq/runs/2020-02-25-H2NCTAFX2/\*\_Hasenfeld\_lane1GM20509Bx01PE20\*\*\*\_2\_sequence.txt.gz

Plotting pipeline: /g/korbel2/StrandSeq/20200109\_U24/20200225\_GM20509B

Mosaicatcher: /g/korbel/jeong/pipeline\_20190625\_GM20509

scNOVA: /g/korbel2/jeong/pipeline\_scNOVA\_20201223\_GM20509

**T-ALL P1**

Fastq: /g/korbel/shared/data/others/StrandSeq/runs/2018-07-24-H5T3WAFXY

/g/korbel/shared/data/others/StrandSeq/runs/2018-08-08-H5NWGAFXY

Plotting pipeline: /g/korbel2/StrandSeq/20180726\_TALL03-DEA5/

Mosaicatcher: /g/korbel/jeong/pipeline\_20190625\_DEA5\_77cells

scNOVA:

/g/korbel2/jeong/pipeline\_scNOVA\_20210625\_DEA5\_77cells\_mosaic\_77cells

**CLL\_24**

Fastq:

Mosaicatcher:

scNOVA: /g/korbel2/jeong/pipeline\_scNOVA\_20201223\_CLL\_24

**AML vs. Control (CD34+) further data needs to be collected for control**

***Part5. Update plan***

1. scNOVA for mm10 genome (mouse) is in development

/g/korbel/jeong/pipeline\_scNOVA\_mm10\_develop

1. Alternative mode using PLS-DA will be incorporated

/g/korbel/jeong/pipeline\_scNOVA\_develop\_PLSDA\_DEA5\_77cells\_mosaic\_77cells

1. Single-cell level expression prediction will be generated