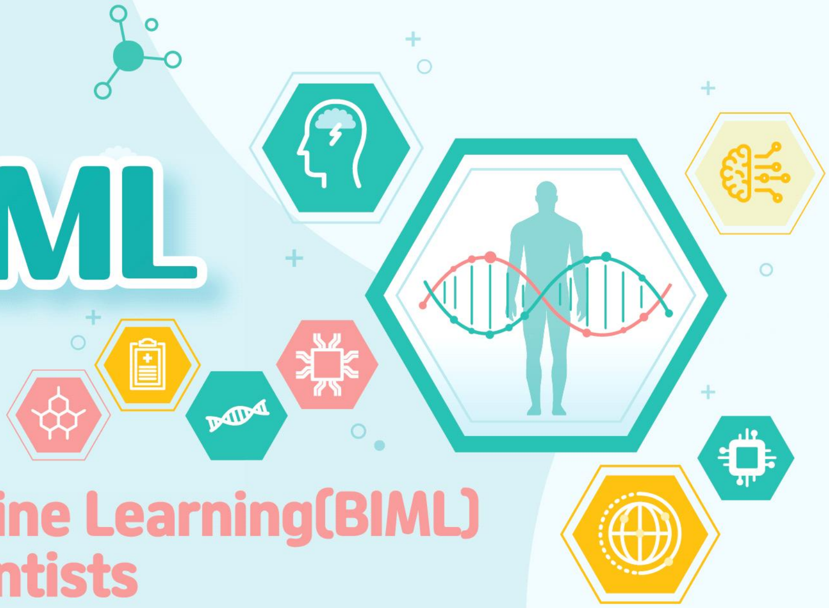
KSBi-BIML 2026

Bioinformatics & Machine Learning(BIML)
Workshop for Life Scientists

생명정보학 & 머신러닝 워크샵 (온라인)



Single-cell multi-omics analysis to study tumor subclones

정효빈 _ 연세대학교

KSBI
KOREAN SOCIETY FOR BIOINFORMATICS
한국생명정보학회

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2026 워크샵을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우 발생하는 **모든 법적 책임은 행위자 본인에게 있음**을 알립니다.

KSBi-BIML 2026

Bioinformatics & Machine Learning (BIML) Workshop for Life Scientists

한국생명정보학회가 주최하는 BIML-2026 동계 Bioinformatics &Machine Learning 교육 워크숍에 여러분을 초대합니다.

BIML 워크숍은 생명정보학 연구자들이 최신 AI바이오 분야의 인공지능 기반 분석 기술과 바이오 데이터 분석 기법을 이론과 실습을 통해 체계적으로 배울 수 있는 전문 교육 프로그램입니다. 2015년에 시작된 BIML 워크숍은 올해로 12년 차를 맞이하며, 국내 생명정보학 분야의 최초이자 최고 수준의 교육 프로그램으로 자리 잡았습니다. 이번 워크숍은 크게 인공지능바이오(AI바이오) 분야와 디지털바이오 분야, 두 분야로 구성됩니다.

AI바이오 분야에서는 생명정보 분석에 폭넓게 응용되고 있는 다양한 인공지능 기반 자료 모델링 기법을 다룰 예정입니다. 특히, 인공지능 심층학습을 활용한 단백질 구조 예측, 유전체 분석, 신약 개발에 대한 이론 및 실습 강의가 진행됩니다.

또한 디지털바이오 분야에서는 단일세포오믹스, 공간오믹스, 멀티오믹스, 메타오믹스에 대한 강의도 마련되어 있어, 연구자들의 분석 역량 강화에 실질적인 도움을 줄 것으로 기대됩니다.

또한 2024년부터 추가된 의료정보 자료 분석을 다루는 강의를 올해도 지속해서 운영하고자 합니다. 이는 최근 의료정보 자료 분석에 관한 연구 수요 증가를 반영한 것으로, 관련 연구를 수행하는 의과학자 및 의료정보 연구자들에게 유용한 지침을 제공할 것입니다.

또한, 올해도 생명정보학 기술의 다양화에 발맞춰 온라인 강좌를 대폭 확대했습니다. 올해는 무료 강좌 10개를 포함한 총 40개 이상의 강좌가 개설되며, 연구 주제에 맞는 강좌 추천과 강연료 할인 혜택도 제공합니다.

BIML-2026는 국내 주요 연구 중심 대학의 전임 교수 및 각 분야 최고 전문가들의 강의로 구성되어 있으며, 기초 이론부터 최신 연구 동향까지 아우르는 심도 있는 교육의 장이 될 것으로 확신합니다.

여러분의 많은 관심과 참여를 기대합니다!

2026년 2월

한국생명정보학회장 류 성 호

강의개요

Single-cell multi-omics analysis to study tumor subclones

암의 종양 내 이질성 (intra-tumor heterogeneity)는 암 조직 내에 다양한 유전체적, 또는 후성 유전체적 특성을 가지는 세포들이 존재하면서 암의 진행을 가속화하고 항암제 내성을 심화시키는 현상을 의미한다. 특히 암의 진화 과정에서 축적되는 유전체 돌연변이와 구조변이들은 새로운 서브클론을 발생시키고, 이러한 서브클론들 각각의 특성을 파악하는 것이 암을 이해하고 치료 전략을 제시하는 데 필요하다. 그렇다면 암에서 이러한 서브클론들을 동정하기 위해 어떤 싱글셀 오믹스 기법들이 개발되어 있을까? 이러한 싱글셀 오믹스 데이터를 분석하기 위해 어떤 생명 정보학적인 도구들을 사용할 수 있을까? 서브클론의 동정 뿐 아니라 그 기능적 특성을 파악하기 위해서는 유전체와 전사체 또는 후성유전체 데이터를 함께 분석하는 싱글셀 멀티 오믹스 분석이 필요하다. 이를 구현하기 위한 생명 정보학적인 방법에는 어떤 것들이 있을까?

본 강의에서는 암에서 서브클론을 동정하기 위해 최근까지 개발되어 있는 다양한 싱글셀 오믹스 기법들에 대해 소개하고, 이들 중 scDNA-seq (Strand-seq, SDR-seq 등)을 이용하는 경우와, scRNA-seq을 이용하는 경우의 데이터 분석을 소개한다. 또한, 서브클론을 동정한 이후에 각각의 기능적인 특성들을 파악할 수 있는 싱글셀 멀티 오믹스를 위해 개발되어 있는 생명정보학 도구들을 소개한다. 이로써, 암의 종양 내 이질성을 심도적으로 탐구하고 의학적 연구에 응용할 수 있는 싱글셀 바이오 데이터 분석 역량을 갖출 수 있도록 하는 것이 최종 목표이다.

강의는 다음의 내용을 포함한다:

- 암에서 서브클론을 동정하기 위한 싱글셀 오믹스 기법들에 대한 소개
- scDNA-seq 기법 중 Strand-seq 데이터에서 서브클론을 동정하는 방법 소개
- Targeted scDNA-seq 기법 중 SDR-seq 소개 및 서브클론을 동정하는 방법 소개
- scRNA-seq 으로부터 서브클론을 유추하기 위한 데이터 분석 방법 소개
- 서브클론을 동정한 후 functional analysis를 위한 싱글셀 멀티오믹스 접근법과 최신 생명정보학 도구 소개

* 교육생준비물: 노트북, R (또는 R studio)

* 강의 난이도: 초급

* 강의: 정효빈 교수 (연세대학교 시스템생물학과)

Curriculum Vitae

Speaker Name: Hyobin Jeong, Ph.D.

▶ Personal Info

Name	Hyobin Jeong
Title	Assistant Professor
Affiliation	Yonsei University, Department of Systems Biology

▶ Contact Information

Address	50 Yonsei-ro, Seodaemun-gu, Seoul, South Korea
Email	hyobinjeong@yonsei.ac.kr

Research Interest

Computational biology, Single-cell Multi-omics, Intra-tumor heterogeneity

Educational Experience

2016.01-2017.11	Postdoc Fellow, Institute of Molecular Biology (IMB), Germany
2015.02-2015.12	Postdoc Fellow, Institute of Basic Science (IBS), Korea
2011.03-2015.02	POSTECH (Ph.D, Interdisciplinary Bioscience and Biotechnology, I-Bio)
2007.03-2011.02	POSTECH (Undergraduate study, Chemical Engineering)

Professional Experience

2024.03-Present	Assistant Professor, Department of Systems Biology, Yonsei University
2022.09-2024.02	Research Professor, Hanyang Institute of Bioscience and Biotechnology (HY-IBB)
2017.12-2022.08	Postdoc Fellow, EMBL, Germany

Selected Publications (3 maximum)

1. 4. Leppä A-M*, Grimes K*, **Jeong H***, Huang F, Andreades A, Waclawiczek A, Boch T, Jauch A, Renders S, Stelmach P, Müller-Tidow C, Karpova D, Sohn M, Grünschläger F, Hasenfeld P, Garagorri E, Thiel V, Dolnik A, Rodriguez-Martin B, Bullinger L, Mrózek K, Eisfeld A-K, Krämer A, Sanders AD, Korbel JO#, Trumpp A#, (2024.12) "Single-cell multiomics analysis reveals dynamic clonal evolution and targetable phenotypes in acute myeloid leukemia with complex karyotype", **Nature Genetics [*co-first]**
2. Grimes K*, **Jeong H***, Amoah A, Niemann J, Raeder B, Hasenfeld P, Benito E, Jann J-C, Nowak D, Ho A, Geiger H, Shui, S, Rausch T, Sanders AD#, Korbel JO#, (2024.05) "Cell type-specific consequences of mosaic structural variants in hematopoietic stem and progenitor cells", **Nature Genetics [*co-first]**
3. **Hyobin Jeong***, Karen Grimes*, Kerstin K. Rauwolf, Peter-Martin Bruch, Tobias Rausch, Patrick Hasenfeld, Eva Benito Garagorri, Tobias Roider, Radhakrishnan Sabarinathan, David Porubsky, Sophie A. Herbst, Büşra Erarslan-Uysal, Johann-Christoph Jann, Tobias Marschall, Daniel Nowak, Jean-Pierre Bourquin, Andreas E. Kulozik, Sascha Dietrich, Beat Bornhauser, Ashley D. Sanders#, Jan O. Korbel#, (2023.06) "Functional analysis of structural variants in single cells using Strand-seq", **Nature Biotechnology [*co-first]**

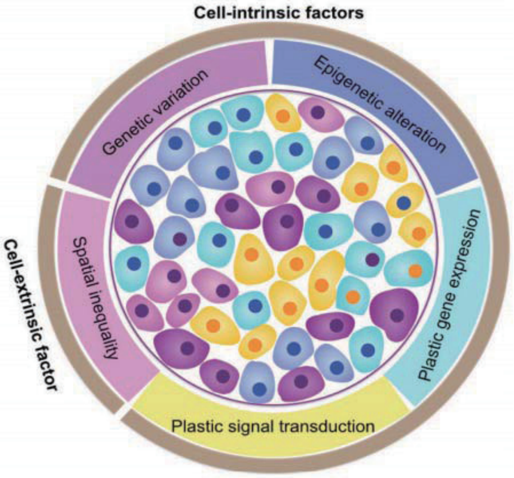
KSBI KOREAN SOCIETY FOR BIOINFORMATICS | 한국생명정보학회

KSBi-BIML 2026

Single-cell multi-omics analysis to
study tumor subclones

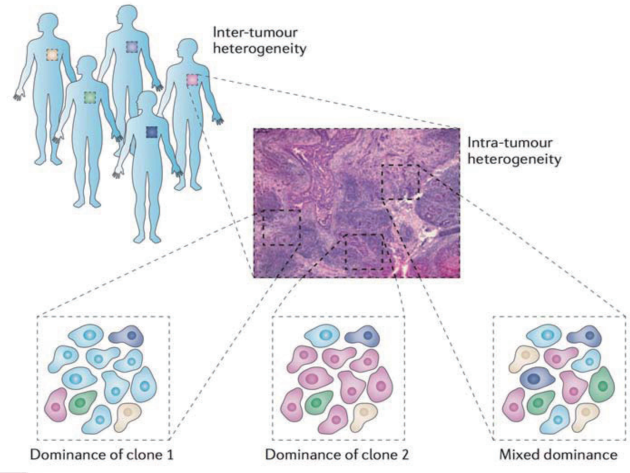
Hyobin Jeong, Assistant professor,
Department of Systems Biology, Yonsei University

Tumor is composed of multiple subclones that makes intra-tumor heterogeneity

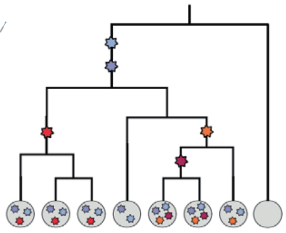


Acta Pharmacologica Sinca (2015)

Multi-layered heterogeneity contributes to therapy failure and cancer progression

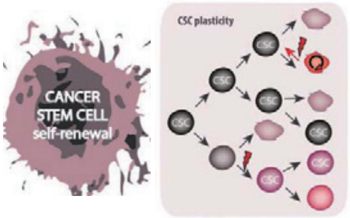


Nature review cancer, 2012



Genome Biology, 2016

Genetic variation



Epigenetic (functional) alteration

Molecular cancer, 2017

How can we tackle the issues with intra-tumor heterogeneity?

- 암에서 이러한 서브클론들을 동정하기 위해 어떤 싱글셀 오믹스 기법들이 개발되어 있을까?

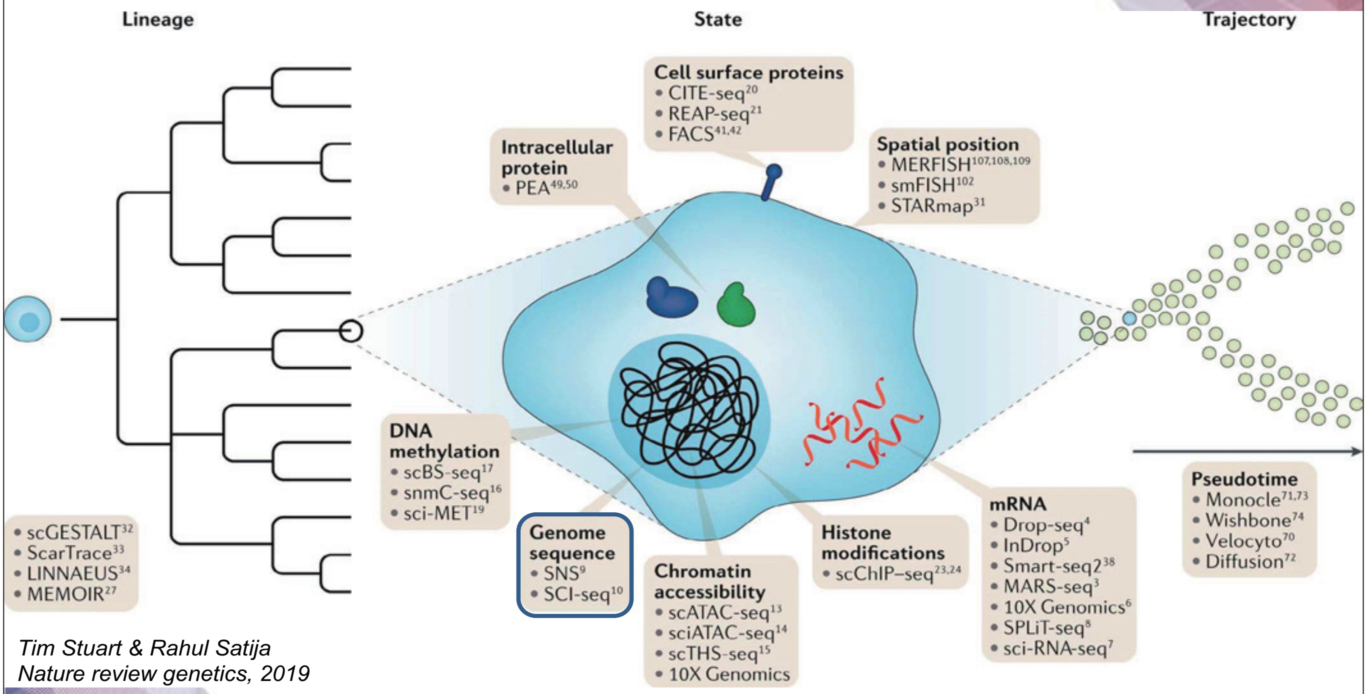
- 이러한 싱글셀 오믹스 데이터를 분석하기 위해 어떤 생명 정보학적인 도구들을 사용할 수 있을까?

- 서브클론의 동정 뿐 아니라 그 기능적 특성을 파악하기 위해서는 유전체와 전사체 또는 후성유전체 데이터를 함께 분석하는 싱글셀 멀티 오믹스 분석이 필요하다. 이를 구현하기 위한 생명 정보학적인 방법에는 어떤 것들이 있을까?

Part1. 암에서 서브클론을 동정하기 위한 싱글셀 오믹스 기법들에 대한 소개

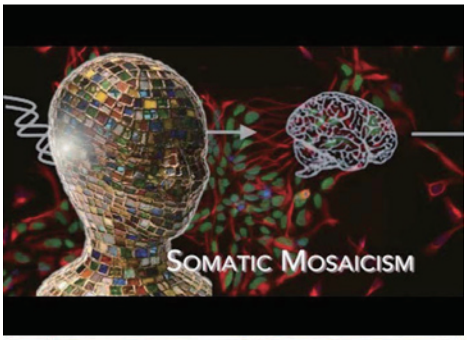
Single-cell multi-omics analysis to study tumor subclones

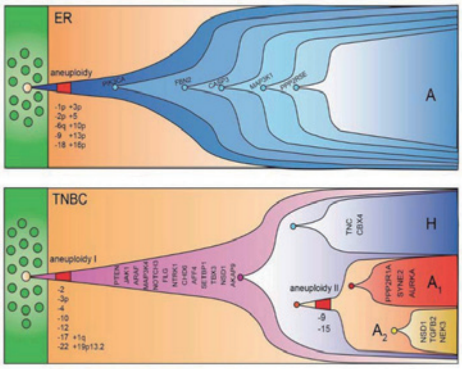
Single-cell technologies to explore cellular heterogeneity



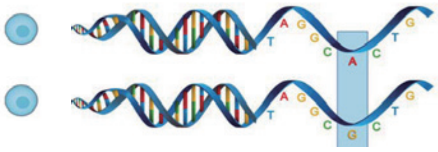
Tim Stuart & Rahul Satija
Nature review genetics, 2019

Genetic changes can happen in nucleotide level and also the form of larger rearrangement

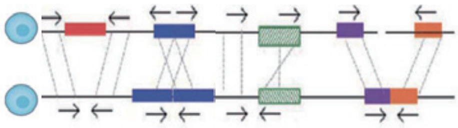




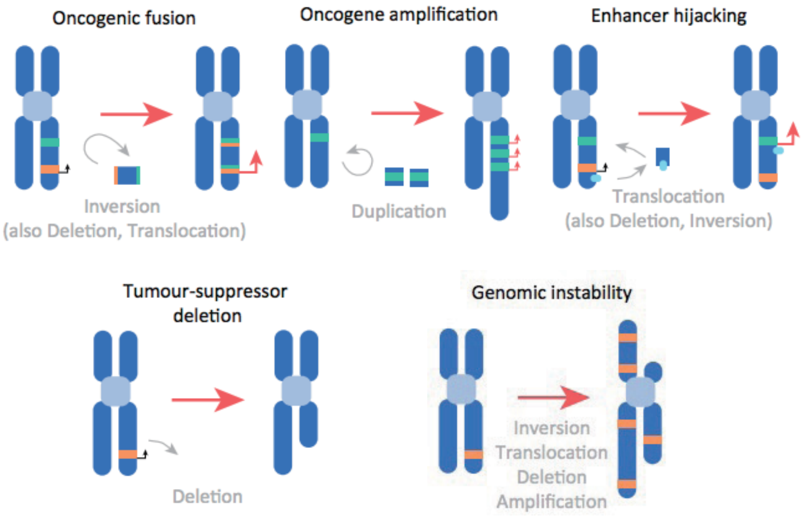
Single nucleotide variation



Structural Variation (SVs) – Strand-seq

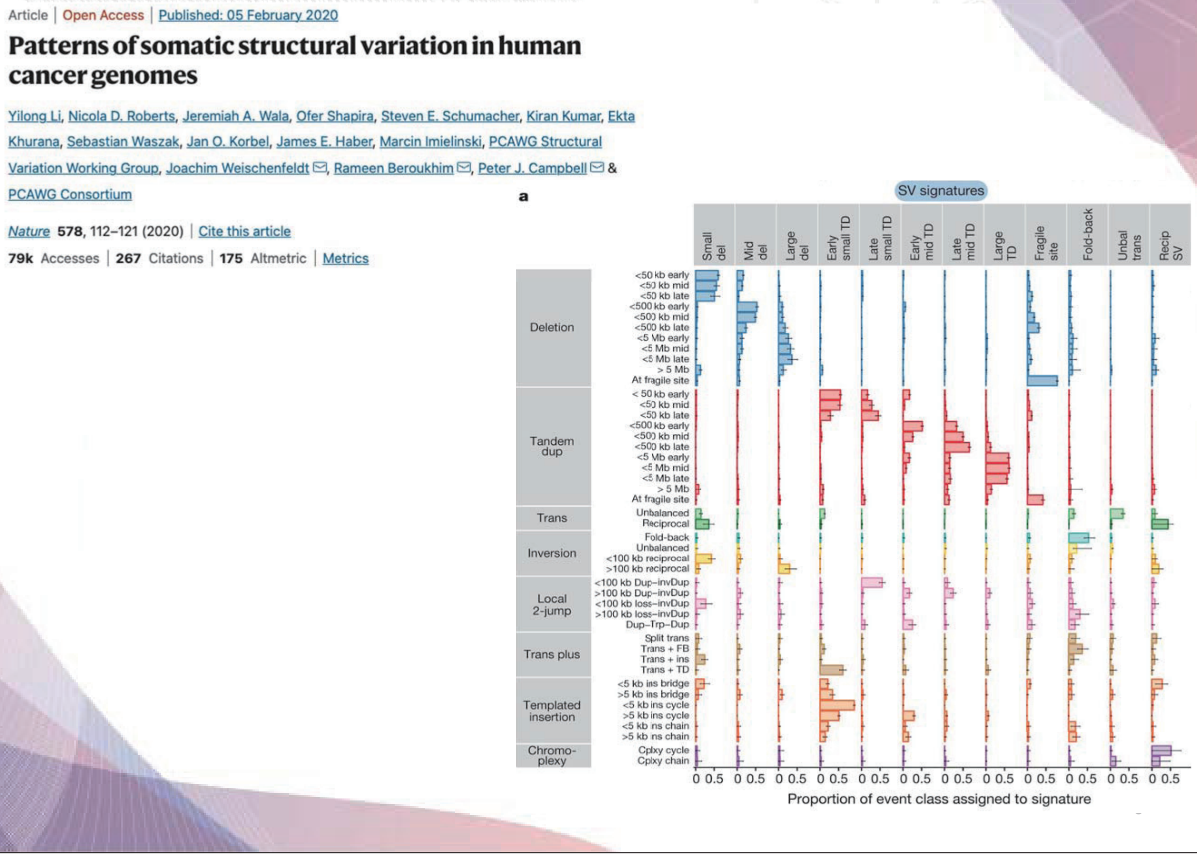


Strauctural variation (SV) is a genomic rearrangement larger than 50bp



Macintyre et al. 2016

Structural variation (SV) is a key mutational process in cancer



Article | Open Access | Published: 05 February 2020

Patterns of somatic structural variation in human cancer genomes

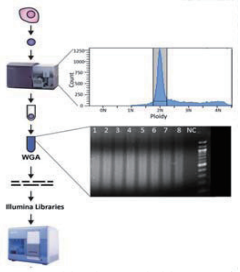
Yilong Li, Nicola D. Roberts, Jeremiah A. Wala, Ofer Shapira, Steven E. Schumacher, Kiran Kumar, Ekta Khurana, Sebastian Waszak, Jan O. Korbel, James E. Haber, Marcin Imielinski, PCAWG Structural Variation Working Group, Joachim Weischenfeldt ✉, Rameen Beroukhim ✉, Peter J. Campbell ✉ & PCAWG Consortium

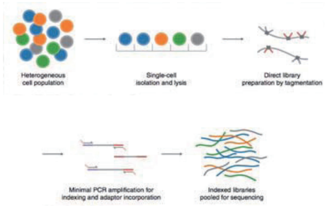
Nature 578, 112–121 (2020) | Cite this article

79k Accesses | 267 Citations | 175 Altmetric | Metrics

Single-cell technologies to explore *genetic* heterogeneity

Single-cell WGS



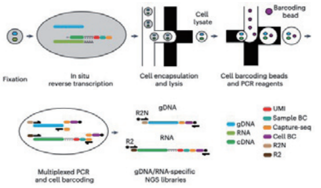


Heterogeneous cell population
Single-cell isolation and lysis
Direct library preparation by tagmentation
Minimal PCR amplification for indexing and adaptor incorporation
Indexed libraries pooled for sequencing

Navin et al. Nature, 2011

Zahn et al.
Nat Methods, 2017

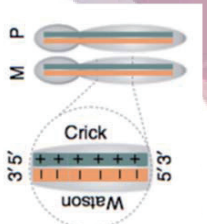
Tapestri (SDR-seq)



Fixation
In situ reverse transcription
Cell lysate
Cell encapsulation and lysis
Barcoding bead
Cell barcoding beads and PCR reagents
Multiplexed PCR and cell barcoding
R2N
gDNA
R2
RNA
gDNA/RNA-specific NGS libraries
gDNA
RNA
cDNA
UMI
Sample BC
Capture-seq
Cell BC
R2N
R2

Linderhofer et al.
Nat Methods,
2025

Strand-seq



Sanders et al.
Nat protocol, 2017

Step1. Alignment - Finding a correct position of reads: BWA

Step2. Remove PCR duplicate: Picard mark duplicate, Biobambam

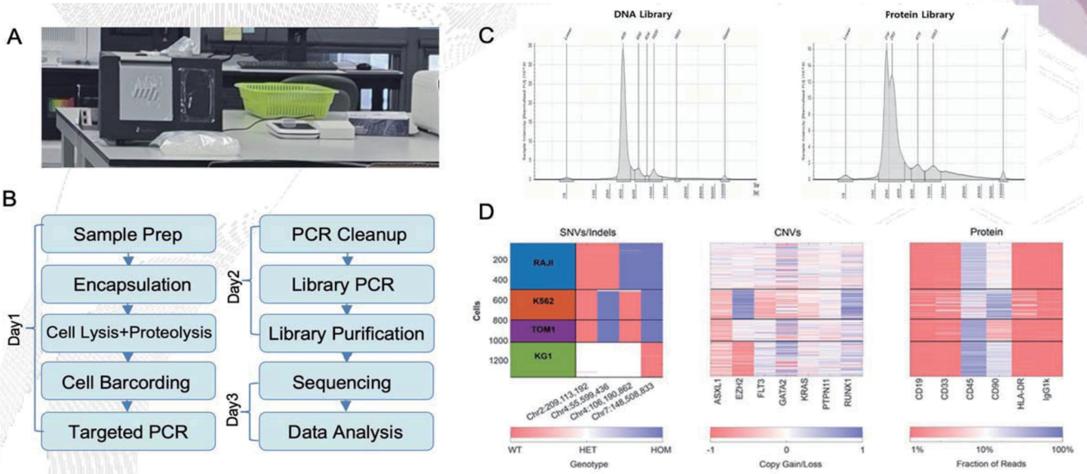
Step3. Genotyping: Freebayes, GATK

Step4. Somatic mutation and CNA calling: SCcaller, Monovar, Aneufinder

Step5. Single-cell clustering and Phylogenetics: SCIPhi, TimeScape

Single-cell technologies to explore *genetic* heterogeneity (Tapestri)

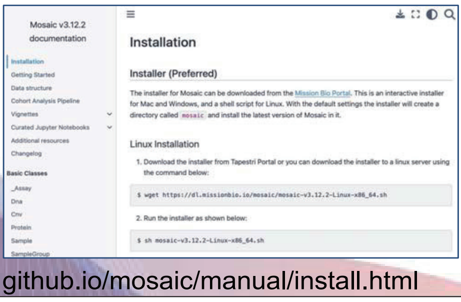
Missionbio Tapestri platform and Mosaic



Sample Prep → Encapsulation → Cell Lysis+Proteolysis → Cell Barcording → Targeted PCR

PCR Cleanup → Library PCR → Library Purification

Sequencing → Data Analysis



Installation

Installer (Preferred)

The installer for Mosaic can be downloaded from the Mission Bio Portal. This is an interactive installer for Mac and Windows, and a shell script for Linux. With the default settings the installer will create a directory called `mosaic` and install the latest version of Mosaic in it.

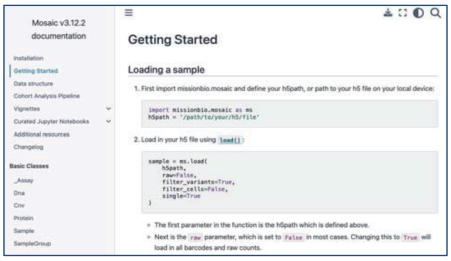
Linux Installation

1. Download the installer from Tapestri Portal or you can download the installer to a linux server using the command below:

```
$ wget https://dl.missionbio.io/mosaic/mosaic-v3.12.2-Linux-x86_64.sh
```

2. Run the installer as shown below:

```
$ sh mosaic-v3.12.2-Linux-x86_64.sh
```



Getting Started

Loading a sample

1. First import missionbio.mosaic and define your h5path, or path to your h5 file on your local device:

```
import missionbio.mosaic as ms
h5path = '/path/to/your/h5/file'
```

2. Load in your h5 file using `load()`:

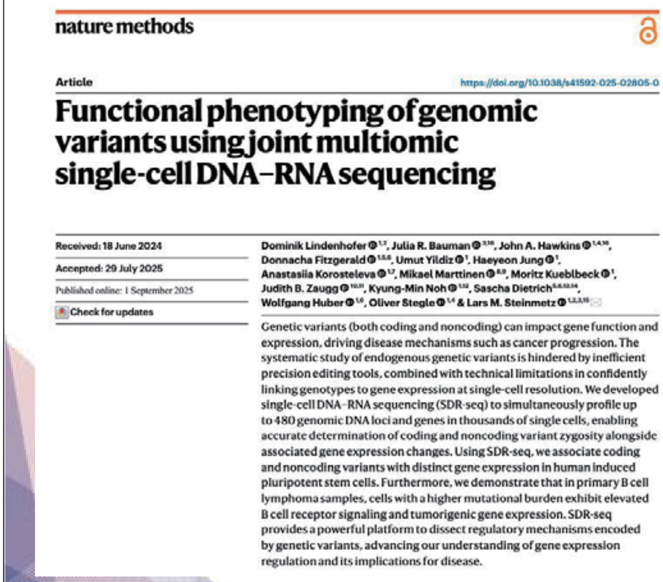
```
sample = ms.load(
    h5path,
    raw=False,
    filter_variants=True,
    filter_cells=False,
    single=True
)
```

- The first parameter in the function is the h5path which is defined above.
- Next is the `raw` parameter, which is set to `False` in most cases. Changing this to `True` will load in all barcodes and raw counts.

https://missionbio.github.io/mosaic/manual/install.html

Single-cell technologies to explore *genetic* heterogeneity (SDR-seq)

SDR-seq (Nature Methods, 2025) and SDRranger



nature methods

Article

https://doi.org/10.1038/s41592-025-02805-0

Functional phenotyping of genomic variants using joint multiomic single-cell DNA–RNA sequencing

Received: 18 June 2024

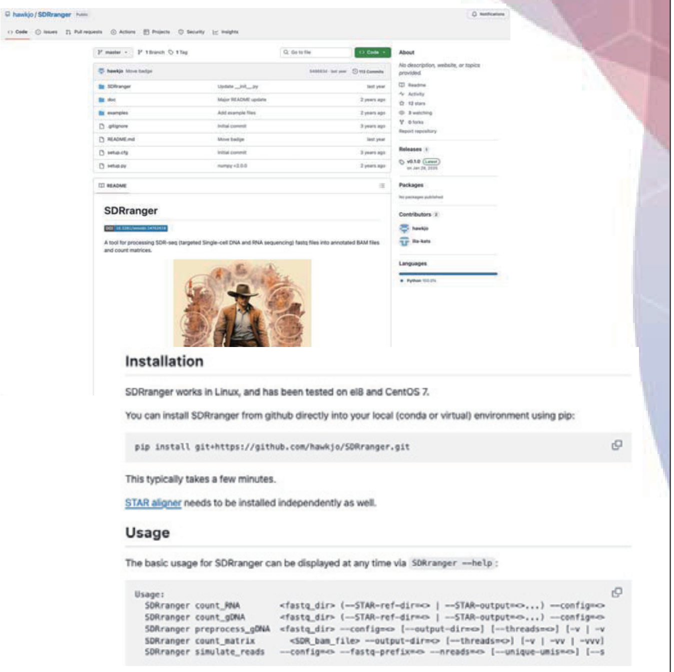
Accepted: 29 July 2025

Published online: 1 September 2025

Check for updates

Dominik Lindenhofer[1,7], Julia R. Bauman[3,16], John A. Hawkins[1,4,16], Donnacha Fitzgerald[1,5,6], Umut Yildiz[1], Haeyeon Jung[1], Anastasiia Korosteleva[1,7], Mikael Marttinen[8,9], Moritz Kueblbeck[1], Judith B. Zaugg[10,11], Kyung-Min Noh[1,12], Sascha Dietrich[5,6,13,14], Wolfgang Huber[1,6], Oliver Stegle[1,4] & Lars M. Steinmetz[1,2,3,15]

Genetic variants (both coding and noncoding) can impact gene function and expression, driving disease mechanisms such as cancer progression. The systematic study of endogenous genetic variants is hindered by inefficient precision editing tools, combined with technical limitations in confidently linking genotypes to gene expression at single-cell resolution. We developed single-cell DNA–RNA sequencing (SDR-seq) to simultaneously profile up to 480 genomic DNA loci and genes in thousands of single cells, enabling accurate determination of coding and noncoding variant zygosity alongside associated gene expression changes. Using SDR-seq, we associate coding and noncoding variants with distinct gene expression in human induced pluripotent stem cells. Furthermore, we demonstrate that in primary B cell lymphoma samples, cells with a higher mutational burden exhibit elevated B cell receptor signaling and tumorigenic gene expression. SDR-seq provides a powerful platform to dissect regulatory mechanisms encoded by genetic variants, advancing our understanding of gene expression regulation and its implications for disease.



No description, website, or topics provided.

hawkjo Move badge		113 Commits
SDRranger	Update __init__.py	last year
doc	Major README update	2 years ago
examples	Add example files	2 years ago
.gitignore	Initial commit	3 years ago
README.md	Move badge	last year
setup.cfg	Initial commit	3 years ago
setup.py	numpy <2.0.0	2 years ago

SDRranger

A tool for processing SDR-seq (targeted Single-cell DNA and RNA sequencing) fastq files into annotated BAM files and count matrices.

Installation

SDRranger works in Linux, and has been tested on el8 and CentOS 7.

You can install SDRranger from github directly into your local (conda or virtual) environment using pip:

```
pip install git+https://github.com/hawkjo/SDRranger.git
```

This typically takes a few minutes.

STAR aligner needs to be installed independently as well.

Usage

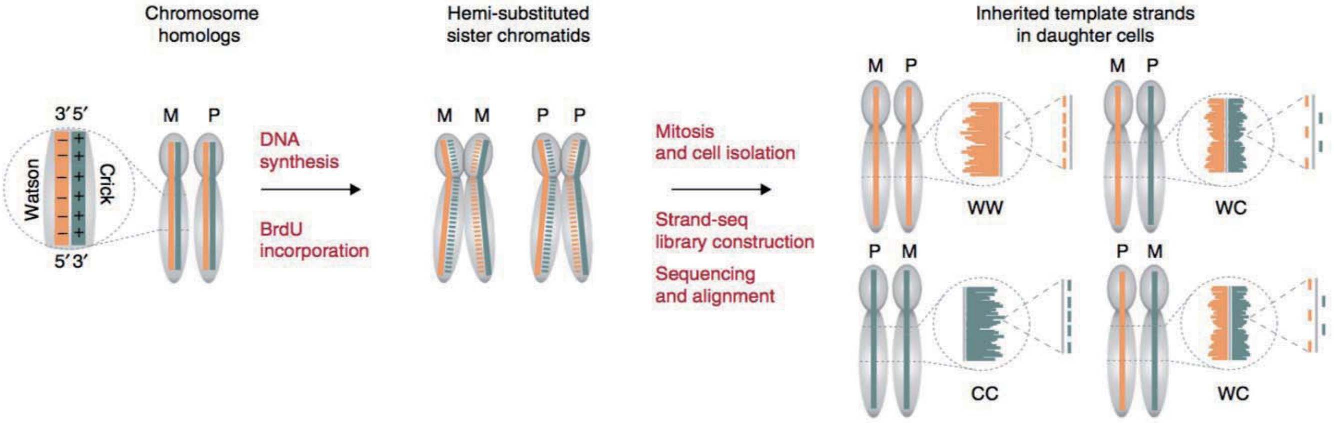
The basic usage for SDRranger can be displayed at any time via `SDRranger --help`:

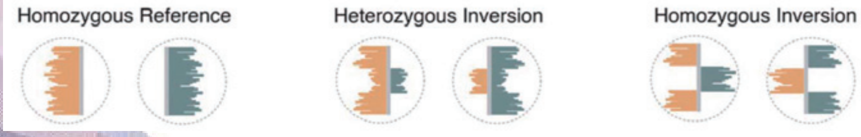
```
Usage:
  SDRranger count_RNA         <fastq_dir> (--STAR-ref-dir=<> | --STAR-output=<>...) --config=<>
  SDRranger count_gDNA        <fastq_dir> (--STAR-ref-dir=<> | --STAR-output=<>...) --config=<>
  SDRranger preprocess_gDNA   <fastq_dir> --config=<> [--output-dir=<>] [--threads=<>] [-v | -v
  SDRranger count_matrix        <SDR_bam_file> --output-dir=<> [--threads=<>] [-v | -vv | -vvv]
  SDRranger simulate_reads    --config=<> --fastq-prefix=<> --nreads=<> [--unique-umis=<>] [--s
```

https://github.com/hawkjo/SDRranger

Single-cell technologies to explore *genetic* heterogeneity (Strand-seq)

Strand-seq (Nat protocol, 2017) and mosaicatcher (Nat biotech, 2020),
scNOVA (Nat biotech, 2023)


Chromosome homologs
Hemi-substituted sister chromatids
Inherited template strands in daughter cells
3′5′
Watson
Crick
5′3′
M
P
DNA synthesis
BrdU incorporation
M M
P P
Mitosis and cell isolation
Strand-seq library construction
Sequencing and alignment
M P
WW
M P
WC
P M
CC
P M
WC



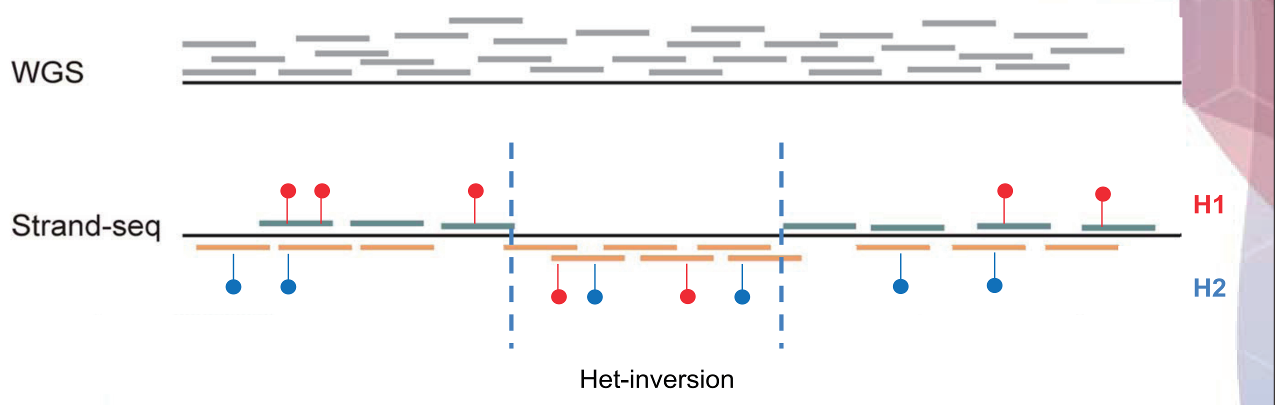
Sanders et al.
Nat protocol, 2017

SBi 한국생명정보학회 Korean Society for Bioinformatics

Part2. scDNA-seq 기법 중 Strand-seq 데이터에서 서브클론을 동정하는 방법 소개

Single-cell multi-omics analysis to
study tumor subclones

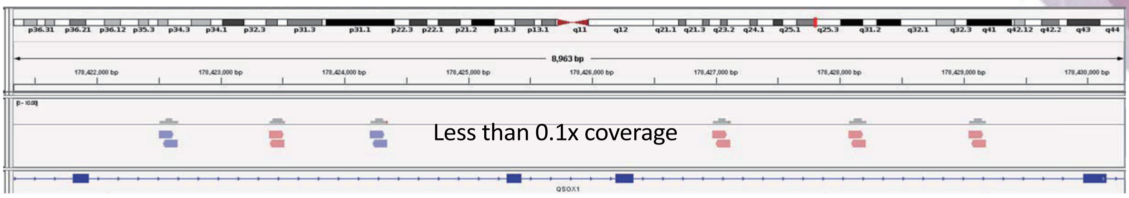
Specialties of the Strand-seq data analysis



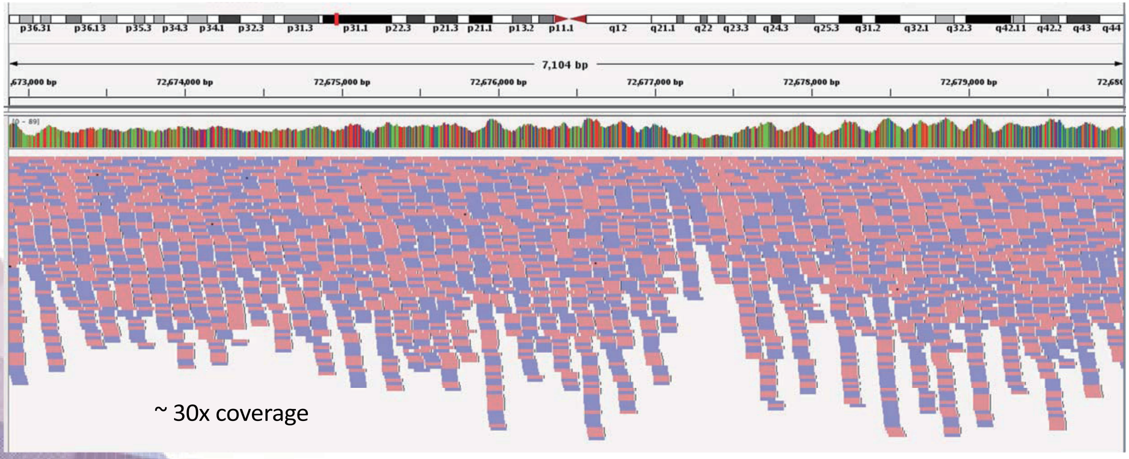
- Sequence orientation is important (Crick or Watson)
- Breakpoint needs to be detected
- Strand state and haplotypes can be assigned
- Multiple types of structural variations need to be classified

Challenges of the Strand-seq data analysis

Strand sequencing



Conventional sequencing (short-read)

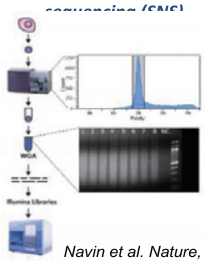


Overview of the single-cell genome data analysis using Strand-seq

Single nucleotide variation (SNVs) - scWGS

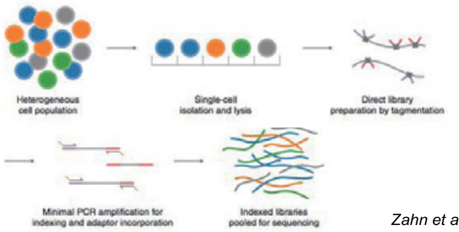
Structural Variation

Single-nucleus sequencing (SNS)



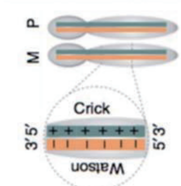
Navin et al. Nature, 2011

Direct library


Heterogeneous cell population
Single-cell isolation and lysis
Direct library preparation by tagmentation
Minimal PCR amplification for indexing and adaptor incorporation
Indexed libraries pooled for sequencing

Zahn et al. Nat Methods, 2017

Strand-seq



Sanders et al. Nat protocol, 2017

Step1. Alignment - Finding a correct position of reads: BWA, sequence orientation

Step2. Remove PCR duplicate: Biobambam

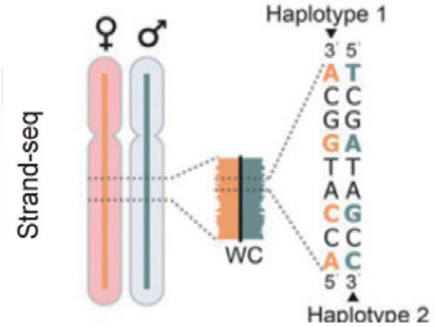
Quality checking!

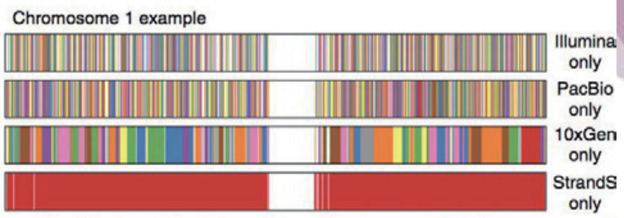
Step3. Genotyping, Haplotyping, Segmentation: StrandPhaseR, breakpointR

Step4. Structural variation calling: MosaiCatcher

Step5. Single-cell clustering and Phylogenetics

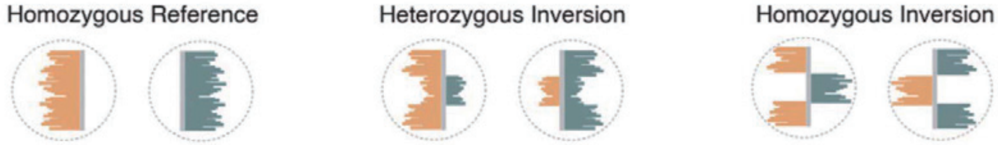
Why the orientation of the reads are important?





Length of the longest haplotype (bp) : Illumina - 15994 bp
PacBio - 1711716 bp
10xGen - 8582136 bp
Strand-seq - 248671482 bp

Porubsky et al. Nat comm, 2017



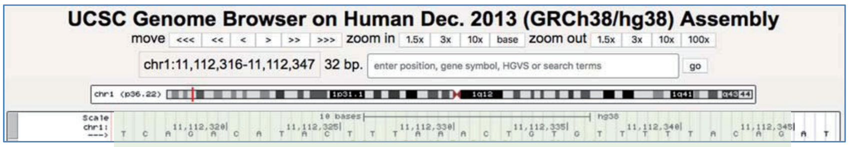
Sanders et al. Genome Res, 2016

How can we assign sequencing reads into Crick and Watson?

ATACTTT

AAAGTAT

- Crick (C) aligns to the plus (forward) strand of the reference assembly
- Watson (W) aligns to the minus (reverse) strand



UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

chr1:11,112,316-11,112,347 32 bp.

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
enter position, gene symbol, HGVS or search terms
go
chr1 (p36.22)
1p31.1
1q12
1q41
q43 44
Scale
chr1:
10 bases
hg38
11,112,320
11,112,325
11,112,330
11,112,335
11,112,340
11,112,345
T C A G A C A T A C T T T A A A C T G T G T T T T T A C A G A T

ATACTTT

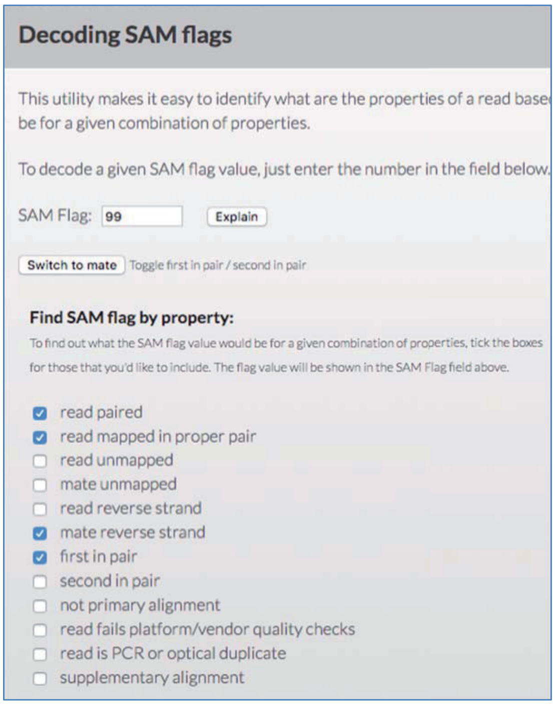
TC AGACATACTTT A A ACTGTGT T TTT ACAG

AAAGTAT

ATACTTT Forward (+) □ Crick (SAMFLAG 0)

AAAGTAT Reverse (-) □ Watson (SAMFLAG 16)

How can we assign sequencing reads into Crick and Watson?



Decoding SAM flags

This utility makes it easy to identify what are the properties of a read base
be for a given combination of properties.

To decode a given SAM flag value, just enter the number in the field below.

SAM Flag: 99 Explain

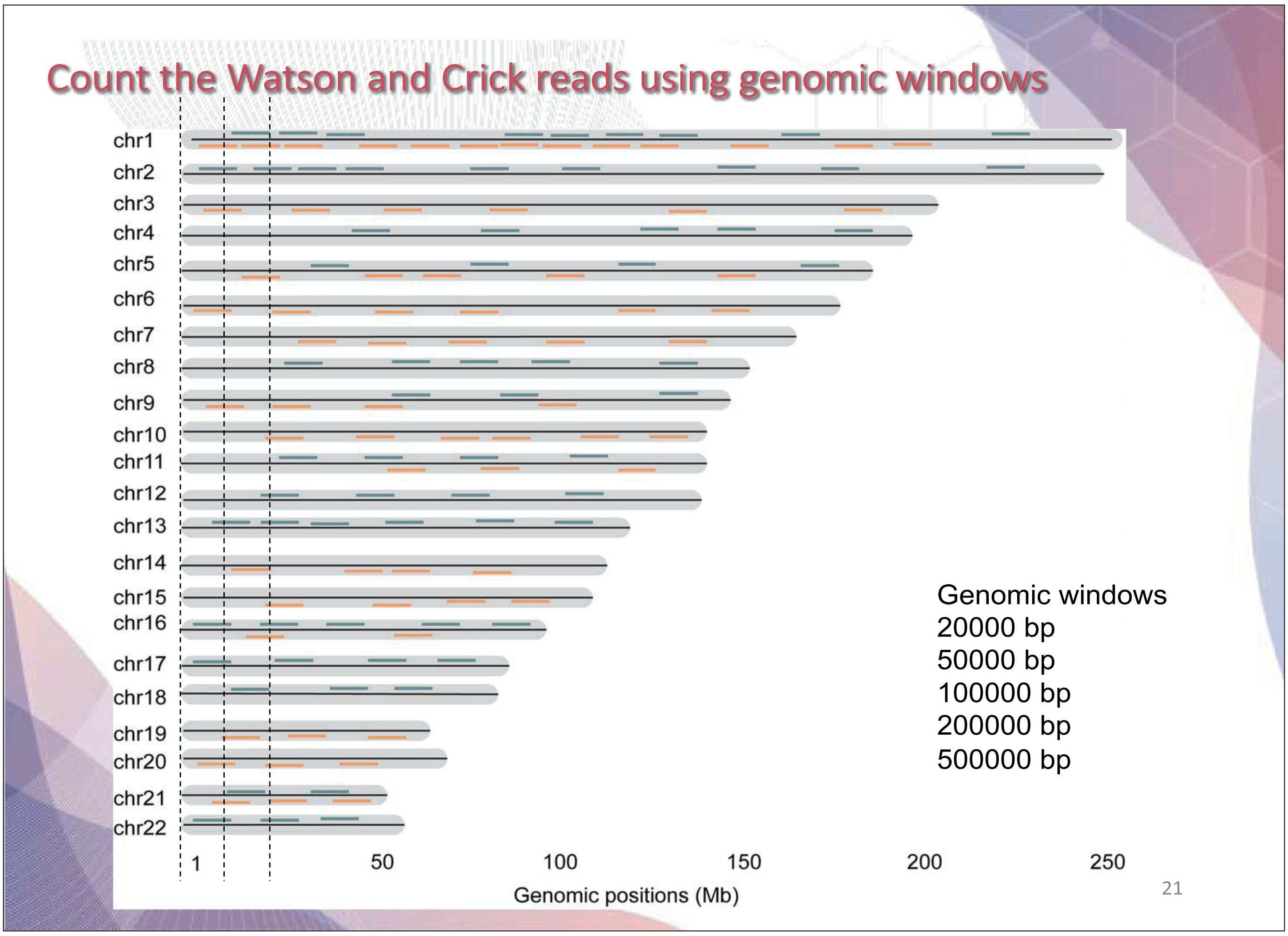
Switch to mate Toggle first in pair / second in pair

Find SAM flag by property:

To find out what the SAM flag value would be for a given combination of properties, tick the boxes for those that you'd like to include. The flag value will be shown in the SAM Flag field above.

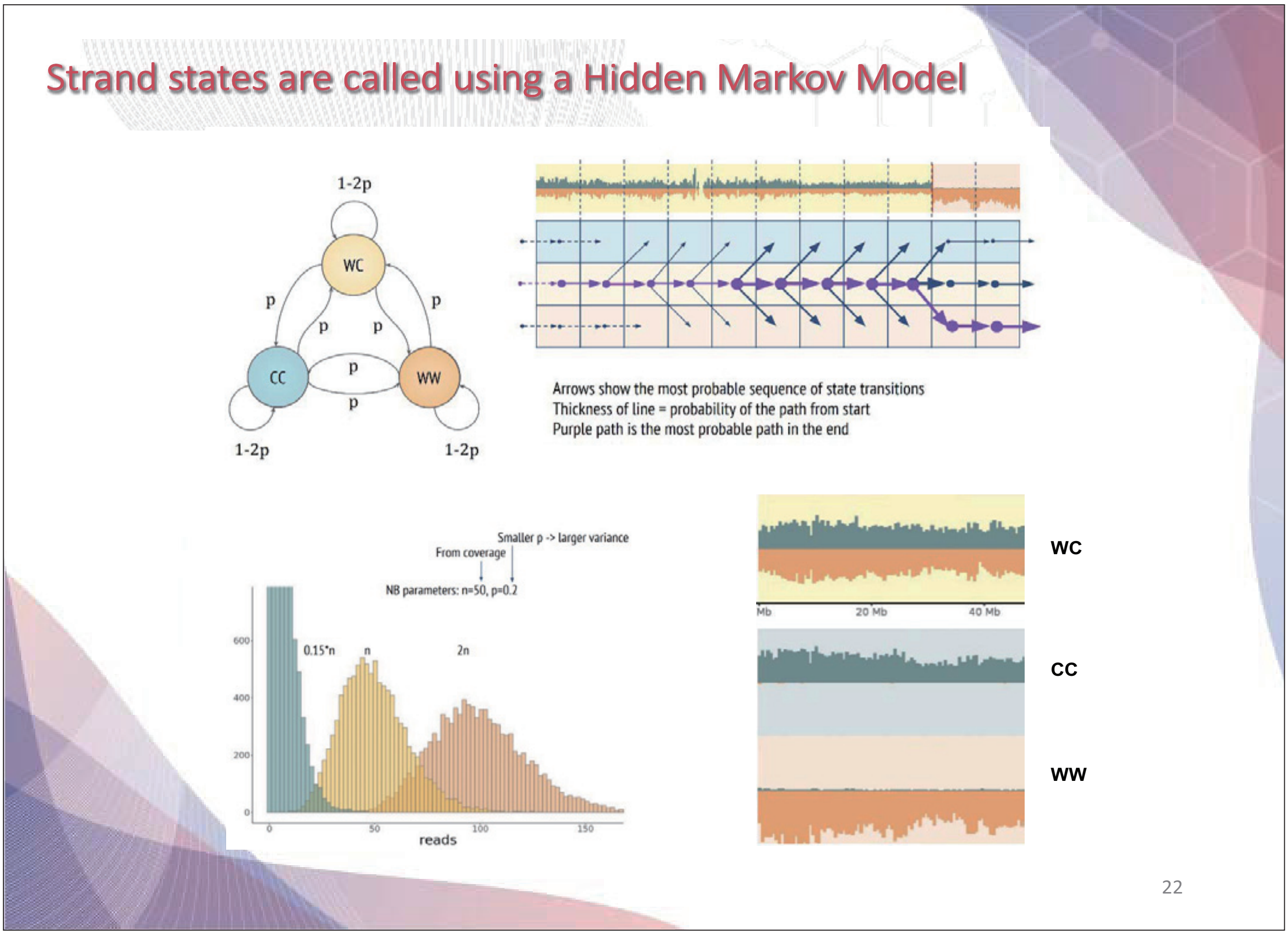
☑ read paired
☑ read mapped in proper pair
☐ read unmapped
☐ mate unmapped
☐ read reverse strand
☑ mate reverse strand
☑ first in pair
☐ second in pair
☐ not primary alignment
☐ read fails platform/vendor quality checks
☐ read is PCR or optical duplicate
☐ supplementary alignment

https://broadinstitute.github.io/picard/explain-flags.html



Count the Watson and Crick reads using genomic windows

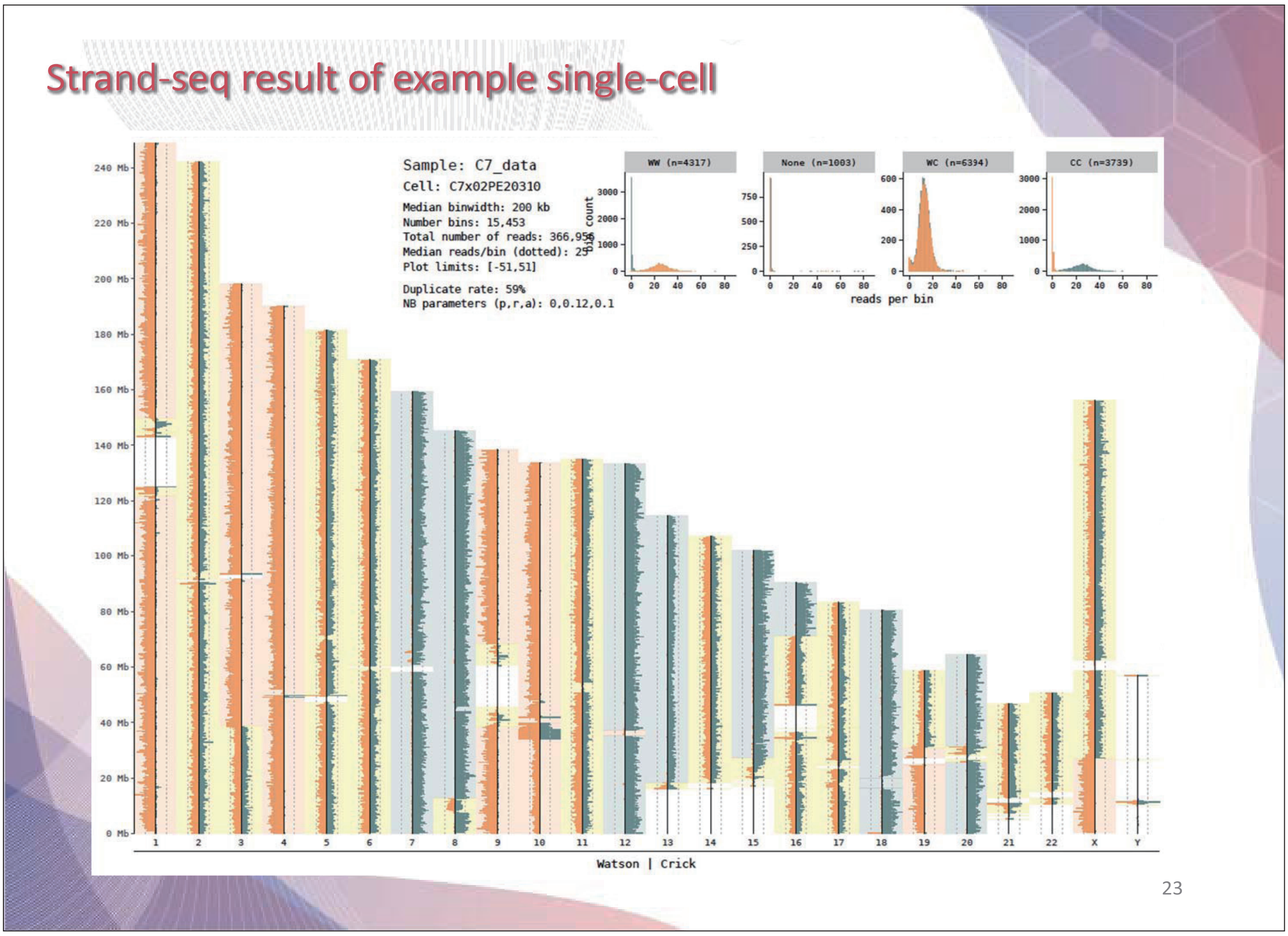
Genomic windows
20000 bp
50000 bp
100000 bp
200000 bp
500000 bp



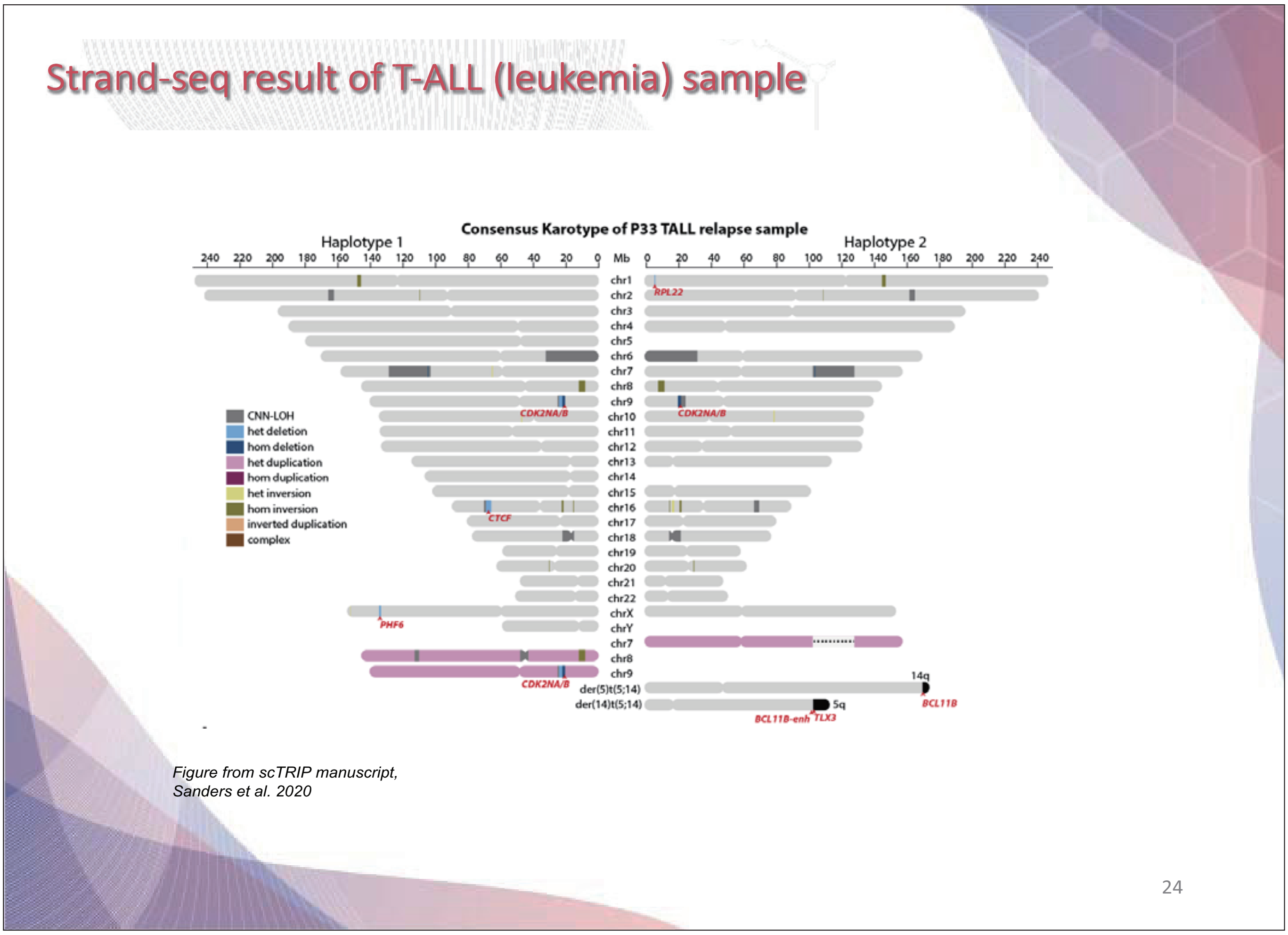
Strand states are called using a Hidden Markov Model

Arrows show the most probable sequence of state transitions
Thickness of line = probability of the path from start
Purple path is the most probable path in the end

Smaller p -> larger variance
From coverage
NB parameters: n=50, p=0.2



Strand-seq result of example single-cell



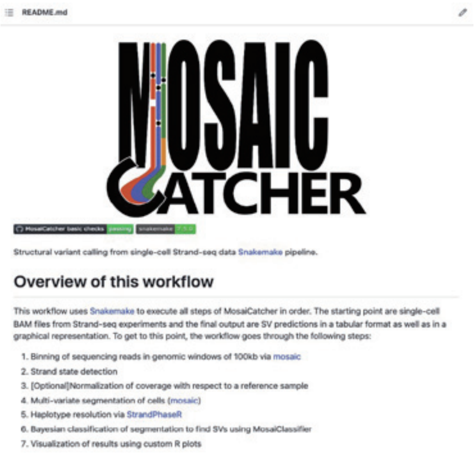
Strand-seq result of T-ALL (leukemia) sample

Consensus Karyotype of P33 TALL relapse sample
Haplotype 1
Haplotype 2
240 220 200 180 160 140 120 100 80 60 40 20 0 Mb 0 20 40 60 80 100 120 140 160 180 200 220 240
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22 chrX chrY chr7 chr8 chr9
der(5)t(5;14)
der(14)t(5;14)
RPL22
CDK2NA/B
CDK2NA/B
CTCF
PHF6
CDK2NA/B
14q
BCL11B
5q
BCL11B-enh TLX3
CNN-LOH
het deletion
hom deletion
het duplication
hom duplication
het inversion
hom inversion
inverted duplication
complex

Figure from scTRIP manuscript, Sanders et al. 2020

Mosaicatcher towards the automatic single-cell SV calling and clustering

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



README.md

MOSAIC CATCHER

MosaiCatcher basic checks passing · snakemake 7.5.0

Structural variant calling from single-cell Strand-seq data Snakemake pipeline.

Overview of this workflow

This workflow uses Snakemake to execute all steps of MosaiCatcher in order. The starting point are single-cell BAM files from Strand-seq experiments and the final output are SV predictions in a tabular format as well as in a graphical representation. To get to this point, the workflow goes through the following steps:

1. Binning of sequencing reads in genomic windows of 100kb via mosaic
2. Strand state detection
3. [Optional]Normalization of coverage with respect to a reference sample
4. Multi-variate segmentation of cells (mosaic)
5. Haplotype resolution via StrandPhaseR
6. Bayesian classification of segmentation to find SVs using MosaiClassifier
7. Visualization of results using custom R plots

Sanders et al. 2020
Weber et al. 2022 (ongoing)

v
1

Korbel group, EMBL

Marschall group, MPI informatics

Ashley Sanders

Sasha Meiers

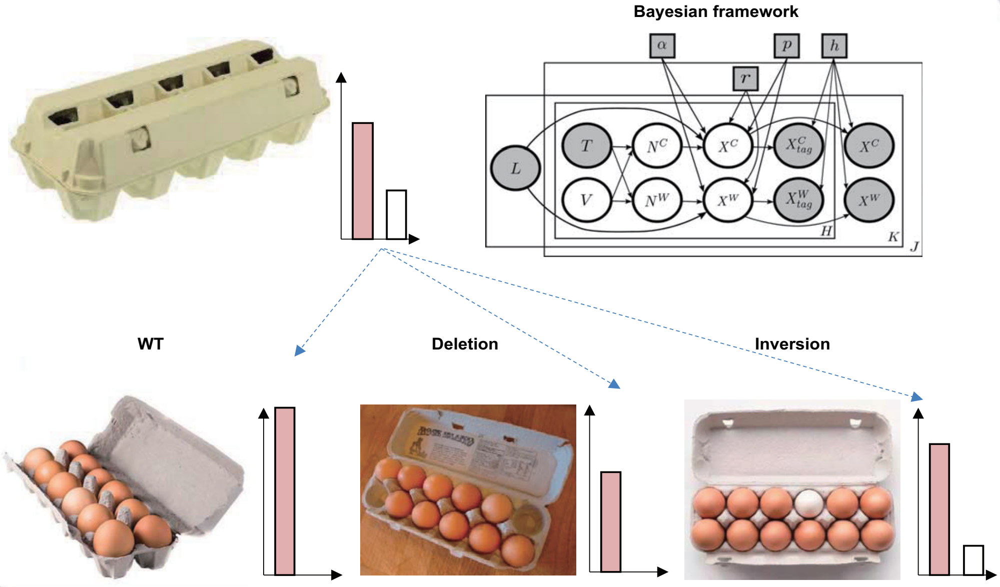
David Porubsky

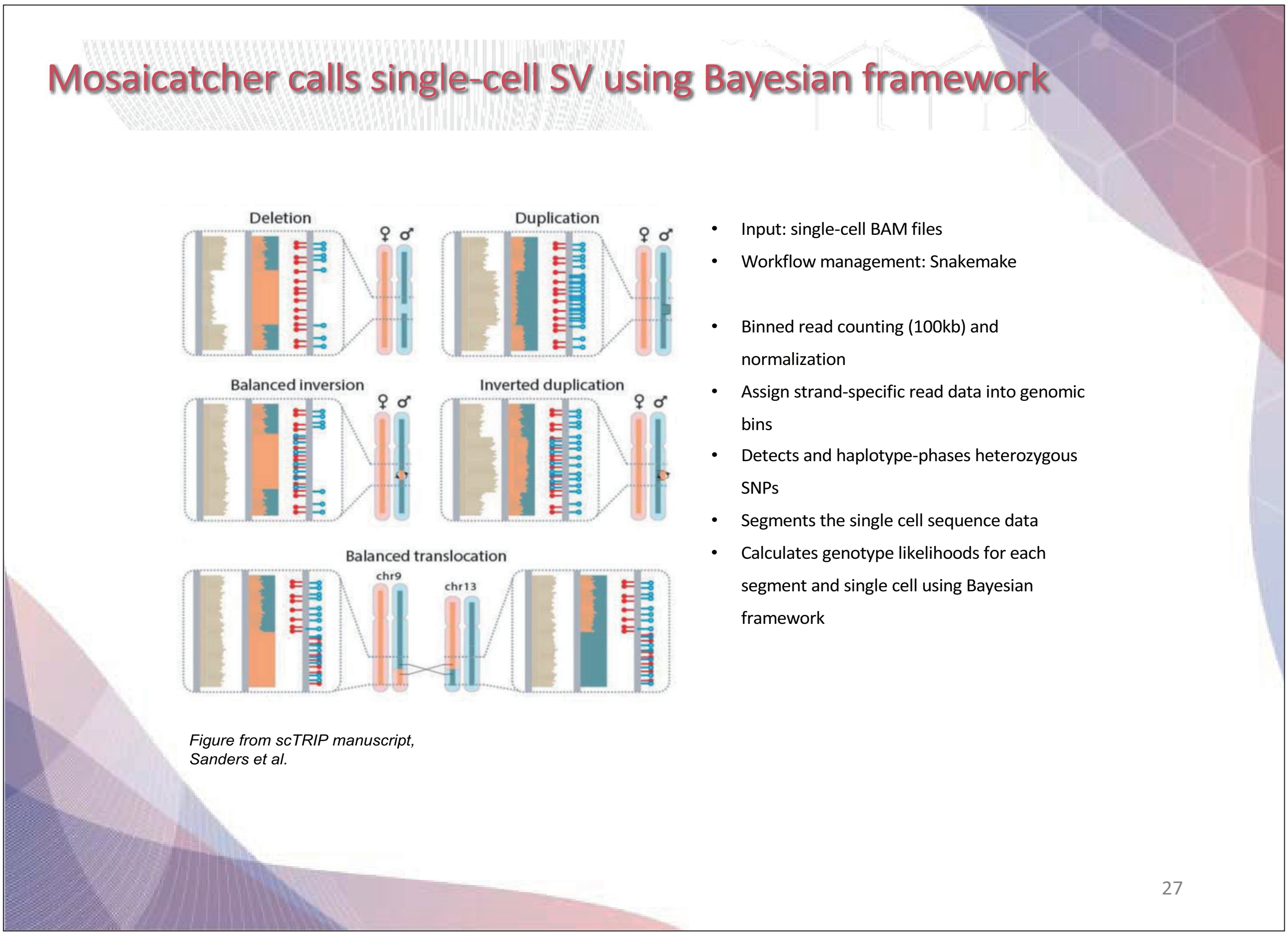
Maryam Ghareghani

v
2

Thomas Weber

Mosaicatcher calls single-cell SV using Bayesian framework





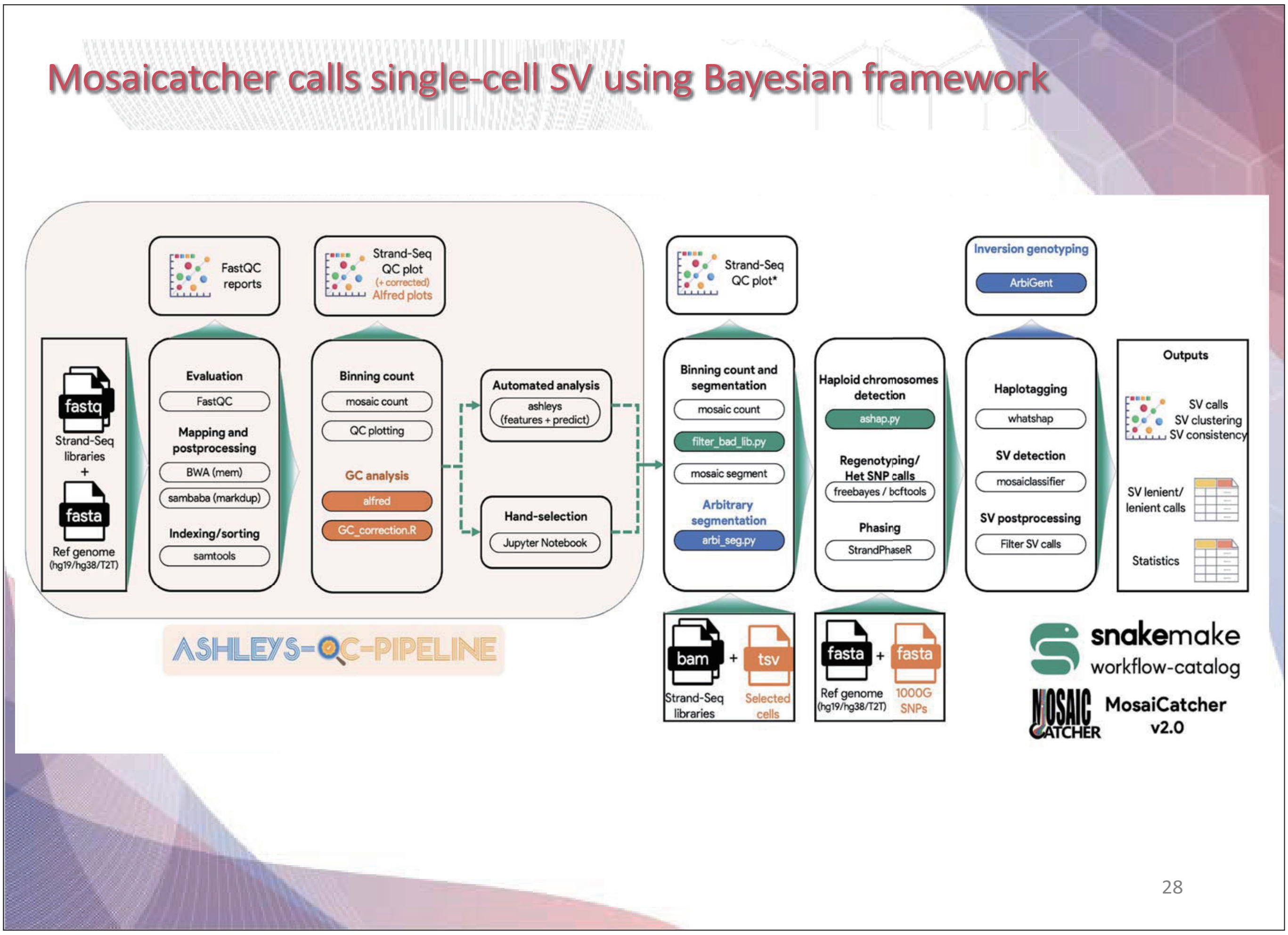
Mosaicatcher calls single-cell SV using Bayesian framework

Figure from scTRIP manuscript, Sanders et al.

- Input: single-cell BAM files
- Workflow management: Snakemake

- Binned read counting (100kb) and normalization
- Assign strand-specific read data into genomic bins
- Detects and haplotype-phases heterozygous SNPs
- Segments the single cell sequence data
- Calculates genotype likelihoods for each segment and single cell using Bayesian framework

27



Mosaicatcher calls single-cell SV using Bayesian framework

fastq Strand-Seq libraries + fasta Ref genome (hg19/hg38/T2T)

FastQC reports

Evaluation: FastQC

Mapping and postprocessing: BWA (mem), sambaba (markdup)

Indexing/sorting: samtools

Strand-Seq QC plot (+ corrected) Alfred plots

Binning count: mosaic count, QC plotting

GC analysis: alfred, GC_correction.R

Automated analysis: ashleys (features + predict)

Hand-selection: Jupyter Notebook

ASHLEYS-QC-PIPELINE

Strand-Seq QC plot*

Binning count and segmentation: mosaic count, filter_bad_lib.py, mosaic segment

Arbitrary segmentation: arbi_seg.py

bam Strand-Seq libraries + tsv Selected cells

Haploid chromosomes detection: ashap.py

Regenotyping/ Het SNP calls: freebayes / bcftools

Phasing: StrandPhaseR

fasta Ref genome (hg19/hg38/T2T) + fasta 1000G SNPs

Inversion genotyping: ArbiGent

Haplotagging: whatshap

SV detection: mosaicclassifier

SV postprocessing: Filter SV calls

Outputs: SV calls, SV clustering, SV consistency; SV lenient/ lenient calls; Statistics

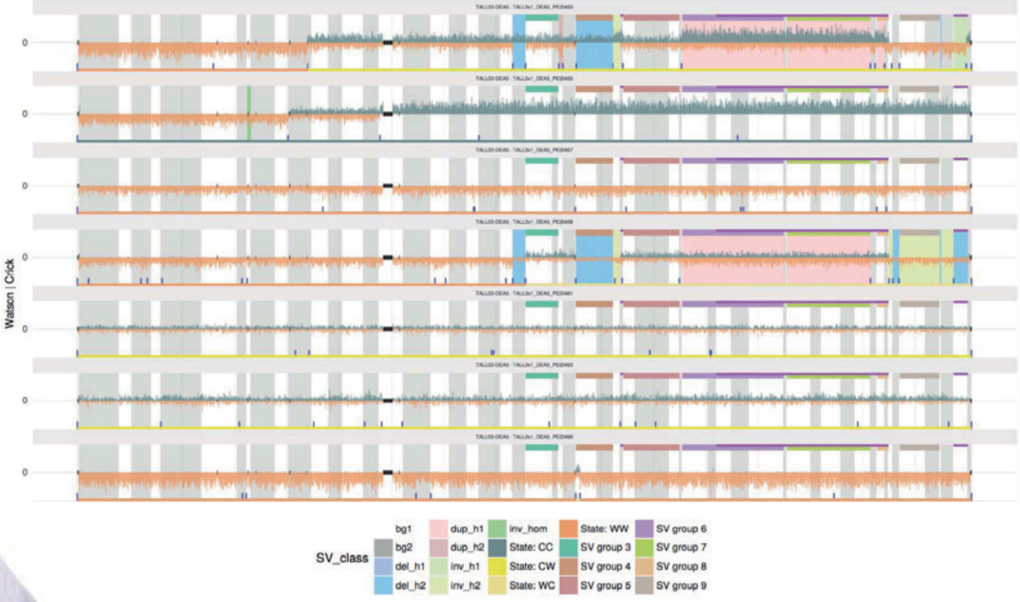
snakemake workflow-catalog

MOSAIC CATCHER MosaiCatcher v2.0

28

Chromosome plot with SVs called by MosaiCatcher framework

Strand-seq from T-ALL PDX (P1) Chromosome6



Heatmap of single-cells based on SVs called by MosaiCatcher framework

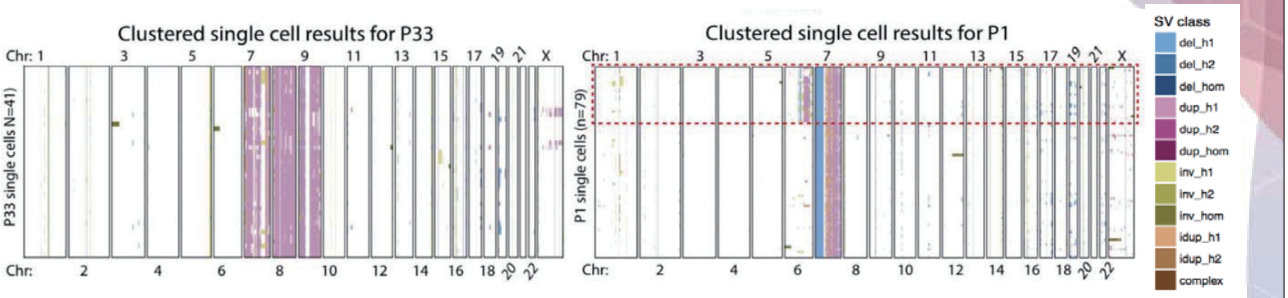
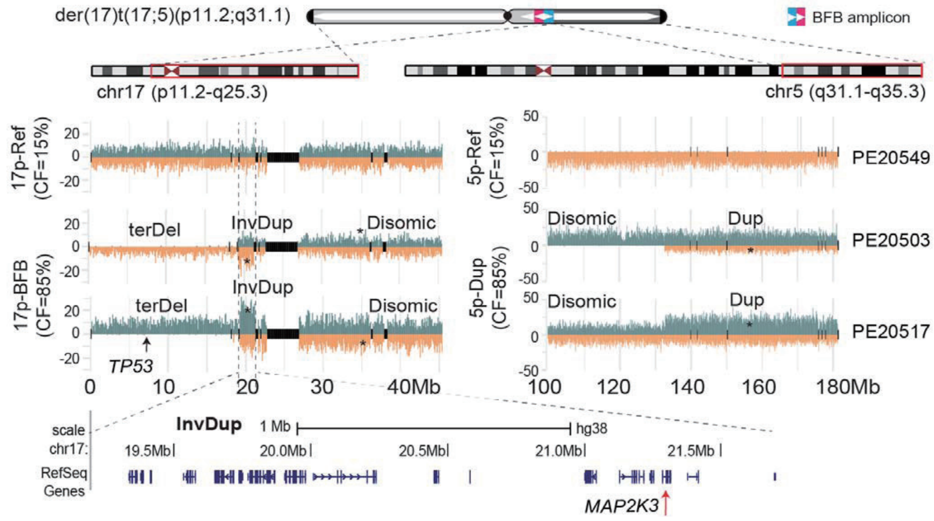


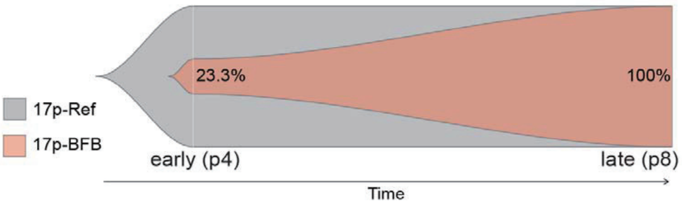
Figure from scTRIP manuscript,
Sanders et al. 2019

- This heatmap was arranged using Ward’s method for hierarchical clustering of SVs genotype likelihoods in two PDX samples
- P33 shows single dominant clone but P1 shows subclonal population in the sample represented by 23 cells

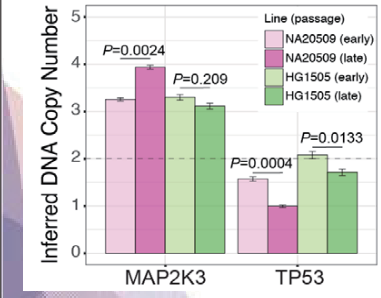
Subclones identified from Strand-seq and MosaiCatcher (Lymphoblastoid cell line, GM20509)

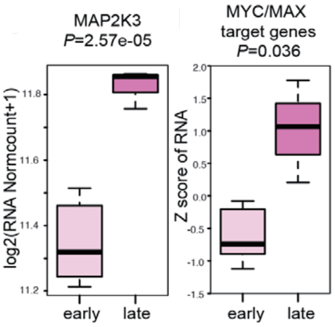


Subclonal evolution can be analyzed using Strand-seq



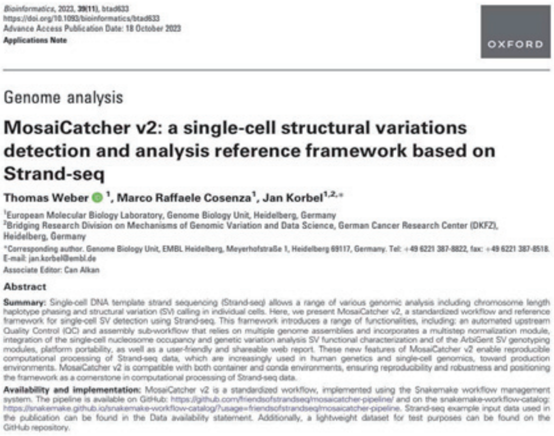
- NA20509 (=GM20509) cell line was in culture for passage 4 (early) and passage 8 (late)





- MAP2K3, and MYC/MAX target genes were increased in late passage
- MYC expression was not changed

Practical session – how to run Mosaicatcher



Bioinformatics, 2023, **39**(11), btad633
https://doi.org/10.1093/bioinformatics/btad633
Advance Access Publication Date: 18 October 2023
Applications Note

OXFORD

Genome analysis

MosaiCatcher v2: a single-cell structural variations detection and analysis reference framework based on Strand-seq

Thomas Weber [1], Marco Raffaele Cosenza[1], Jan Korbel[1,2,*]

[1]European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany
[2]Bridging Research Division on Mechanisms of Genomic Variation and Data Science, German Cancer Research Center (DKFZ), Heidelberg, Germany

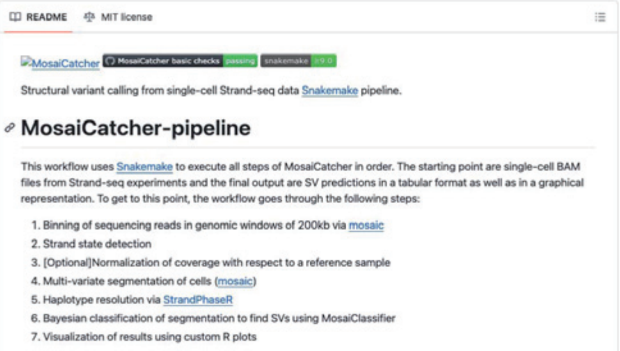
*Corresponding author. Genome Biology Unit, EMBL Heidelberg, Meyerhofstraße 1, Heidelberg 69117, Germany. Tel: +49 6221 387-8822, fax: +49 6221 387-8518. E-mail: jan.korbel@embl.de

Associate Editor: Can Alkan

Abstract

Summary: Single-cell DNA template strand sequencing (Strand-seq) allows a range of various genomic analysis including chromosome length haplotype phasing and structural variation (SV) calling in individual cells. Here, we present MosaiCatcher v2, a standardized workflow and reference framework for single-cell SV detection using Strand-seq. This framework introduces a range of functionalities, including: an automated upstream Quality Control (QC) and assembly sub-workflow that relies on multiple genome assemblies and incorporates a multistep normalization module, integration of the single-cell nucleosome occupancy and genetic variation analysis SV functional characterization and of the ArbiGent SV genotyping modules, platform portability, as well as a user-friendly and shareable web report. These new features of MosaiCatcher v2 enable reproducible computational processing of Strand-seq data, which are increasingly used in human genetics and single-cell genomics, toward production environments. MosaiCatcher v2 is compatible with both container and conda environments, ensuring reproducibility and robustness and positioning the framework as a cornerstone in computational processing of Strand-seq data.

Availability and implementation: MosaiCatcher v2 is a standardized workflow, implemented using the Snakemake workflow management system. The pipeline is available on GitHub: https://github.com/friendsofstrandseq/mosaicatcher-pipeline/ and on the snakemake-workflow-catalog: https://snakemake.github.io/snakemake-workflow-catalog/?usage=friendsofstrandseq/mosaicatcher-pipeline. Strand-seq example input data used in the publication can be found in the Data availability statement. Additionally, a lightweight dataset for test purposes can be found on the GitHub repository.



Structural variant calling from single-cell Strand-seq data Snakemake pipeline.

MosaiCatcher-pipeline

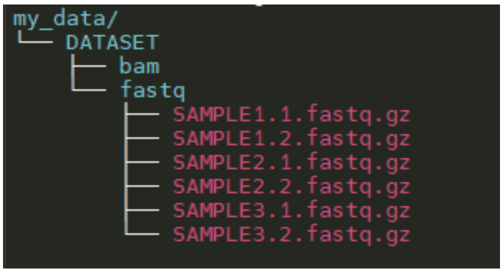
This workflow uses Snakemake to execute all steps of MosaiCatcher in order. The starting point are single-cell BAM files from Strand-seq experiments and the final output are SV predictions in a tabular format as well as in a graphical representation. To get to this point, the workflow goes through the following steps:

1. Binning of sequencing reads in genomic windows of 200kb via mosaic
2. Strand state detection
3. [Optional]Normalization of coverage with respect to a reference sample
4. Multi-variate segmentation of cells (mosaic)
5. Haplotype resolution via StrandPhaseR
6. Bayesian classification of segmentation to find SVs using MosaiClassifier
7. Visualization of results using custom R plots

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

Based on the tutorial written by Chiwon Chung

Data directory tree



```
my_data/
└── DATASET
    ├── bam
    └── fastq
        ├── SAMPLE1.1.fastq.gz
        ├── SAMPLE1.2.fastq.gz
        ├── SAMPLE2.1.fastq.gz
        ├── SAMPLE2.2.fastq.gz
        ├── SAMPLE3.1.fastq.gz
        └── SAMPLE3.2.fastq.gz
```

Data directory must follow a certain format

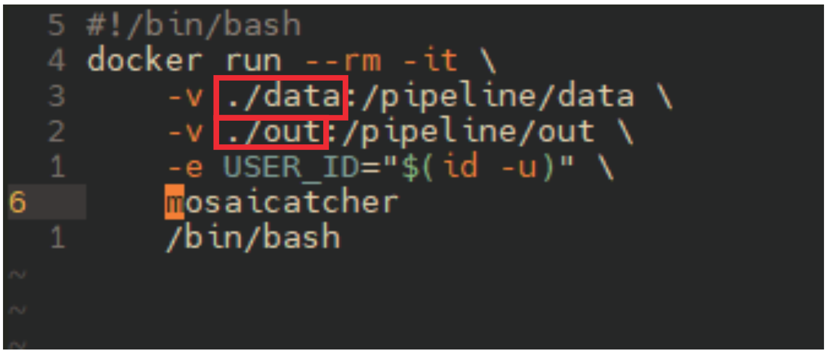
- Pipeline requires you to have either
  1. BAM
  2. Fastq

Placed in like the image on the left.

- Fastq files (for ASHLEYS) must follow the following syntax:

→ [SAMPLE_NAME].x.fastq.gz
→ X is the strand number (coding/non-coding)

Run_mosaicatcher.sh



```
#!/bin/bash
docker run --rm -it \
    -v ./data:/pipeline/data \
    -v ./out:/pipeline/out \
    -e USER_ID="$(id -u)" \
    mosaicatcher
    /bin/bash
```

To start the pipeline

1. Copy the run_mosaicatcher file in /home/shared to directory you desire
2. Run the following command:

   /path/to/run_mosaicatcher.sh

./data: where your test data (and outputs) will be stored

./output: where your ASHLEYS result (and other checkpoint files) will be stored

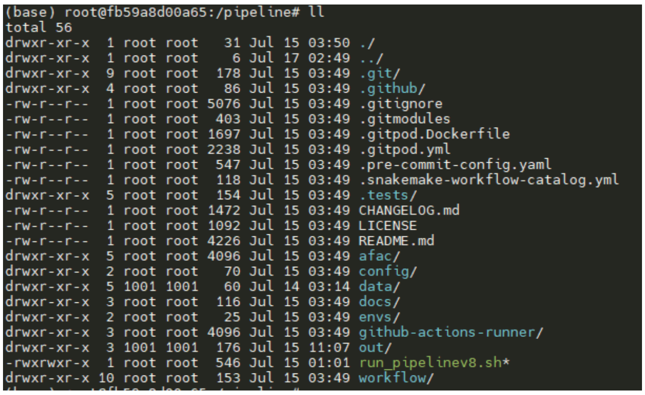
You can change these (red boxes) to whatever you want.

Ex) You have data folder at ~/my_data:

-v ~/my_data:/pipeline/data

Do NOT change the rest of the script!

Within the Docker file



```
(base) root@fb59a8d00a65:/pipeline# ll
total 56
drwxr-xr-x  1 root root   31 Jul 15 03:50 ./
drwxr-xr-x  1 root root    6 Jul 17 02:49 ../
drwxr-xr-x  9 root root  178 Jul 15 03:49 .git/
drwxr-xr-x  4 root root   86 Jul 15 03:49 .github/
-rw-r--r--  1 root root 5076 Jul 15 03:49 .gitignore
-rw-r--r--  1 root root  403 Jul 15 03:49 .gitmodules
-rw-r--r--  1 root root 1697 Jul 15 03:49 .gitpod.Dockerfile
-rw-r--r--  1 root root 2238 Jul 15 03:49 .gitpod.yml
-rw-r--r--  1 root root  547 Jul 15 03:49 .pre-commit-config.yaml
-rw-r--r--  1 root root  118 Jul 15 03:49 .snakemake-workflow-catalog.yml
drwxr-xr-x  5 root root  154 Jul 15 03:49 .tests/
-rw-r--r--  1 root root 1472 Jul 15 03:49 CHANGELOG.md
-rw-r--r--  1 root root 1092 Jul 15 03:49 LICENSE
-rw-r--r--  1 root root 4226 Jul 15 03:49 README.md
drwxr-xr-x  5 root root 4096 Jul 15 03:49 afac/
drwxr-xr-x  2 root root   70 Jul 15 03:49 config/
drwxr-xr-x  5 1001 1001   60 Jul 14 03:14 data/
drwxr-xr-x  3 root root  116 Jul 15 03:49 docs/
drwxr-xr-x  2 root root   25 Jul 15 03:49 envs/
drwxr-xr-x  3 root root 4096 Jul 15 03:49 github-actions-runner/
drwxr-xr-x  3 1001 1001  176 Jul 15 11:07 out/
-rwxrwxr-x  1 root root  546 Jul 15 01:01 run_pipelinev8.sh*
drwxr-xr-x 10 root root  153 Jul 15 03:49 workflow/
```

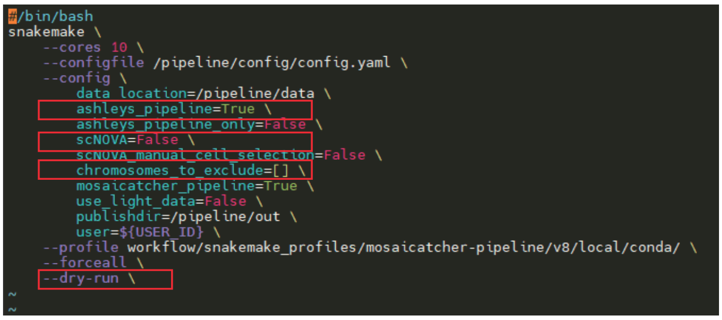
If the command is run correctly, you should now be inside the docker container. Using the ll command as seen in the image, we can see there is a shell script:

→ run_pipelinev8.sh

This is the script we will use to run the actual pipeline.

The data/ and out/ folders should be connected to the data and out folders you set in the run_mosaicatcher.sh file.

Within the Docker file



```
#/bin/bash
snakemake \
    --cores 10 \
    --configfile /pipeline/config/config.yaml \
    --config \
        data_location=/pipeline/data \
        ashleys_pipeline=True \
        ashleys_pipeline_only=False \
        scNOVA=False \
        scNOVA_manual_cell_selection=False \
        chromosomes_to_exclude=[] \
        mosaicatcher_pipeline=True \
        use_light_data=False \
        publishdir=/pipeline/out \
        user=${USER_ID} \
    --profile workflow/snakemake_profiles/mosaicatcher-pipeline/v8/local/conda/ \
    --forceall \
    --dry-run \
~
~
```

Use vim run_mosaicatcherv8.sh to access the script. Please refer to online tutorials for vim controls, as it is out of scope with this manual.

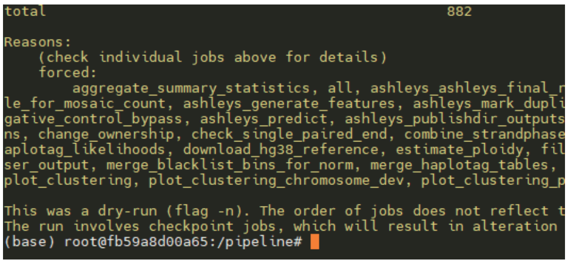
You only need to change the lines accentuated by red boxes.

The pipeline will be run TWICE:

1. First run is for ASHLEYS and mosaicatcher pipeline
2. Second is for scNOVA pipeline

The dry run command is for testing purposes only, and will be removed when the pipeline is run.

Running the pipeline



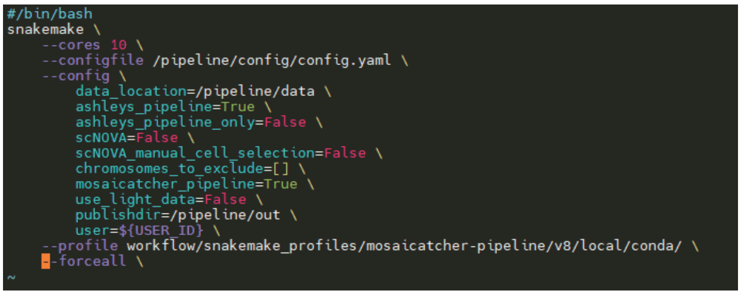
```
total                                                    882

Reasons:
    (check individual jobs above for details)
    forced:
        aggregate_summary_statistics, all, ashleys_ashleys_final_r
le_for_mosaic_count, ashleys_generate_features, ashleys_mark_dupli
gative_control_bypass, ashleys_predict, ashleys_publishdir_outputs
ns, change_ownership, check_single_paired_end, combine_strandphase
aplotag_likelihoods, download_hg38_reference, estimate_ploidy, fil
ser_output, merge_blacklist_bins_for_norm, merge_haplotag_tables,
plot_clustering, plot_clustering_chromosome_dev, plot_clustering_p

This was a dry-run (flag -n). The order of jobs does not reflect t
The run involves checkpoint jobs, which will result in alteration
(base) root@fb59a8d00a65:/pipeline# █
```

First, run the mosaicatcher_v8.sh file as is; it should do a test run.

If all is without issues, you will see no errors and the CLI will show the total number of rules (which will vary depending on your dataset size)

In case you see errors, the most likely reason is that your files are not set up correctly. Make sure your fastq/bam files are placed like it's shown in slide 2.

Running the pipeline

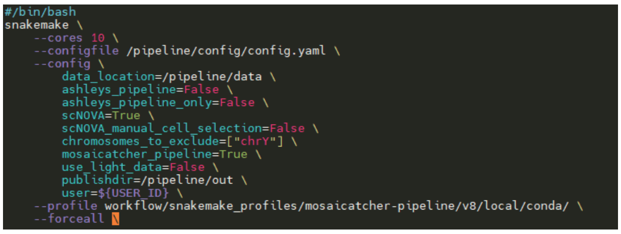


```
#/bin/bash
snakemake \
    --cores 10 \
    --configfile /pipeline/config/config.yaml \
    --config \
        data_location=/pipeline/data \
        ashleys_pipeline=True \
        ashleys_pipeline_only=False \
        scNOVA=False \
        scNOVA_manual_cell_selection=False \
        chromosomes_to_exclude=[] \
        mosaicatcher_pipeline=True \
        use_light_data=False \
        publishdir=/pipeline/out \
        user=${USER_ID} \
    --profile workflow/snakemake_profiles/mosaicatcher-pipeline/v8/local/conda/ \
    --forceall \
~
```

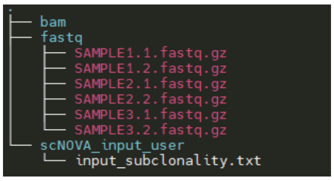
If you see no errors, remove the dry run option that was originally in the run_mosaicatcherv8.sh pipeline, then run it again. The first step of the pipeline should start running.

The duration will vary but expect it to take at least a couple of hours. Depending on server CPU usage, it may even take a full day.

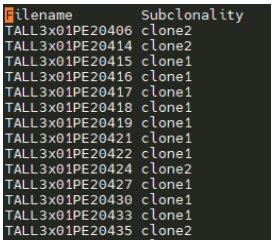
Running the pipeline



```
#/bin/bash
snakemake \
    --cores 10 \
    --configfile /pipeline/config/config.yaml \
    --config \
        data_location=/pipeline/data \
        ashleys_pipeline=False \
        ashleys_pipeline_only=False \
        scNOVA=True \
        scNOVA_manual_cell_selection=False \
        chromosomes_to_exclude=["chrY"] \
        mosaicatcher_pipeline=True \
        use_light_data=False \
        publishdir=/pipeline/out \
        user=${USER_ID} \
    --profile workflow/snakemake_profiles/mosaicatcher-pipeline/v8/local/conda/ \
    --forceall \
```



```
.
├── bam
├── fastq
│   ├── SAMPLE1.1.fastq.gz
│   ├── SAMPLE1.2.fastq.gz
│   ├── SAMPLE2.1.fastq.gz
│   ├── SAMPLE2.2.fastq.gz
│   ├── SAMPLE3.1.fastq.gz
│   └── SAMPLE3.2.fastq.gz
└── scNOVA_input_user
    └── input_subclonality.txt
```



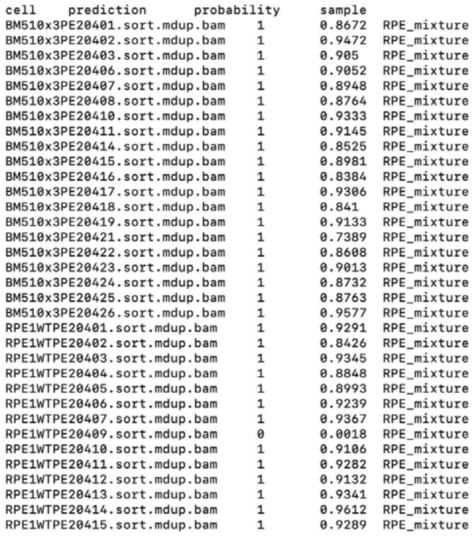
```
Filename         Subclonality
TALL3x01PE20406 clone2
TALL3x01PE20414 clone2
TALL3x01PE20415 clone1
TALL3x01PE20416 clone1
TALL3x01PE20417 clone1
TALL3x01PE20418 clone1
TALL3x01PE20419 clone1
TALL3x01PE20421 clone1
TALL3x01PE20422 clone1
TALL3x01PE20424 clone2
TALL3x01PE20427 clone1
TALL3x01PE20430 clone1
TALL3x01PE20433 clone1
TALL3x01PE20435 clone2
```

Once the pipeline is finished, create a new directory scNOVA_input_user and add your subclonality file. It MUST be named input_subclonality.txt, and it MUST be a tsv file with the correct header names. Otherwise, the pipeline will throw errors.

To run the scNOVA pipeline, please change the run_mosaicatcherv8.sh as shown on the left. Note that scNOVA currently does not support chromosome Y, and must be place in the chromosomes_to_exclude list.

Output – QC result based on ASHLEY algorithm

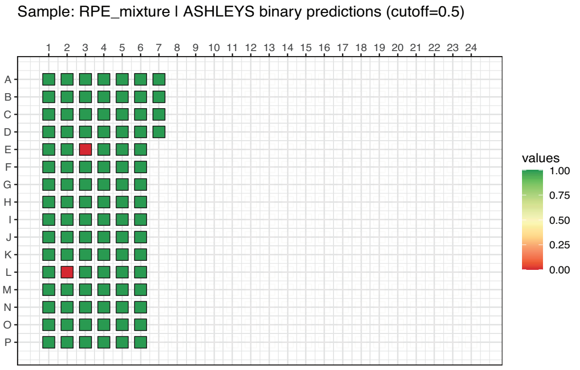
out/data/RPE_mixture/cell_selection/labels_positive_control_corrected.tsv



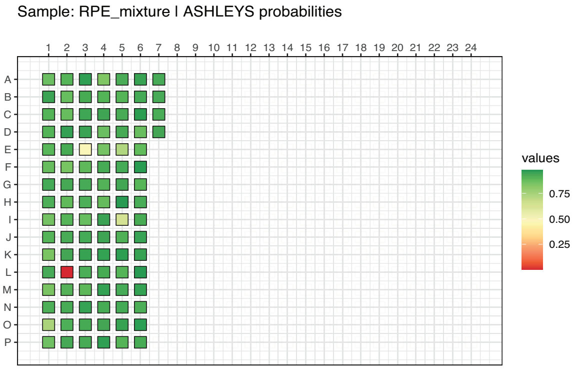
```
cell     prediction       probability      sample
BM510x3PE20401.sort.mdup.bam    1        0.8672  RPE_mixture
BM510x3PE20402.sort.mdup.bam    1        0.9472  RPE_mixture
BM510x3PE20403.sort.mdup.bam    1        0.905   RPE_mixture
BM510x3PE20406.sort.mdup.bam    1        0.9052  RPE_mixture
BM510x3PE20407.sort.mdup.bam    1        0.8948  RPE_mixture
BM510x3PE20408.sort.mdup.bam    1        0.8764  RPE_mixture
BM510x3PE20410.sort.mdup.bam    1        0.9333  RPE_mixture
BM510x3PE20411.sort.mdup.bam    1        0.9145  RPE_mixture
BM510x3PE20414.sort.mdup.bam    1        0.8525  RPE_mixture
BM510x3PE20415.sort.mdup.bam    1        0.8981  RPE_mixture
BM510x3PE20416.sort.mdup.bam    1        0.8384  RPE_mixture
BM510x3PE20417.sort.mdup.bam    1        0.9306  RPE_mixture
BM510x3PE20418.sort.mdup.bam    1        0.841   RPE_mixture
BM510x3PE20419.sort.mdup.bam    1        0.9133  RPE_mixture
BM510x3PE20421.sort.mdup.bam    1        0.7389  RPE_mixture
BM510x3PE20422.sort.mdup.bam    1        0.8608  RPE_mixture
BM510x3PE20423.sort.mdup.bam    1        0.9013  RPE_mixture
BM510x3PE20424.sort.mdup.bam    1        0.8732  RPE_mixture
BM510x3PE20425.sort.mdup.bam    1        0.8763  RPE_mixture
BM510x3PE20426.sort.mdup.bam    1        0.9577  RPE_mixture
RPE1WTPE20401.sort.mdup.bam     1        0.9291  RPE_mixture
RPE1WTPE20402.sort.mdup.bam     1        0.8426  RPE_mixture
RPE1WTPE20403.sort.mdup.bam     1        0.9345  RPE_mixture
RPE1WTPE20404.sort.mdup.bam     1        0.8848  RPE_mixture
RPE1WTPE20405.sort.mdup.bam     1        0.8993  RPE_mixture
RPE1WTPE20406.sort.mdup.bam     1        0.9239  RPE_mixture
RPE1WTPE20407.sort.mdup.bam     1        0.9367  RPE_mixture
RPE1WTPE20409.sort.mdup.bam     0        0.0018  RPE_mixture
RPE1WTPE20410.sort.mdup.bam     1        0.9106  RPE_mixture
RPE1WTPE20411.sort.mdup.bam     1        0.9282  RPE_mixture
RPE1WTPE20412.sort.mdup.bam     1        0.9132  RPE_mixture
RPE1WTPE20413.sort.mdup.bam     1        0.9341  RPE_mixture
RPE1WTPE20414.sort.mdup.bam     1        0.9612  RPE_mixture
RPE1WTPE20415.sort.mdup.bam     1        0.9289  RPE_mixture
```

Output – QC result based on ASHLEY algorithm

out/data/RPE_mixture/plots/plate/ashleys_plate_predictions.pdf
out/data/RPE_mixture/plots/plate/ashleys_plate_probabilities.pdf



Sample: RPE_mixture I ASHLEYS binary predictions (cutoff=0.5)

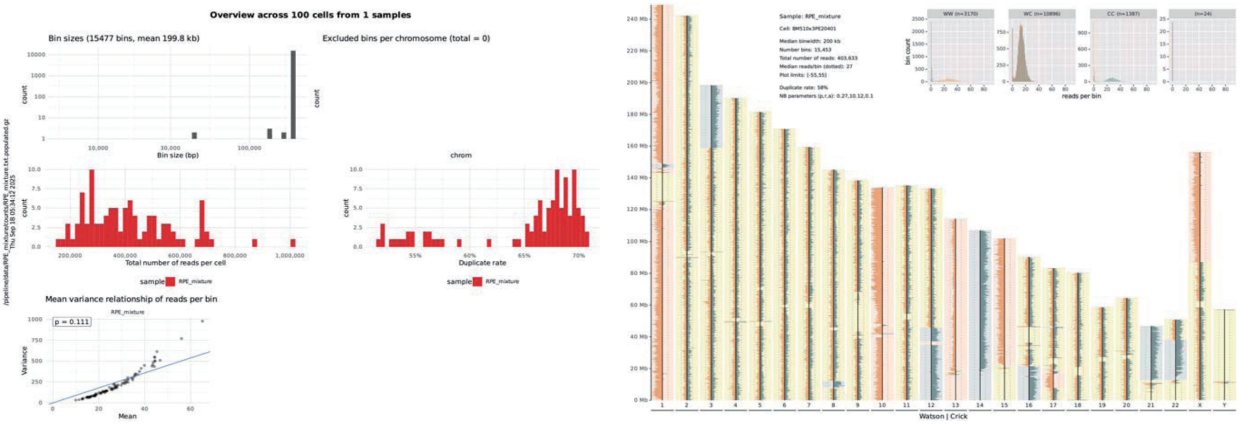


Sample: RPE_mixture I ASHLEYS probabilities

**반올림해서 0.5가 되면 통과되는 결과임 (0.4759 → PASS)

Output – Plotting pipeline

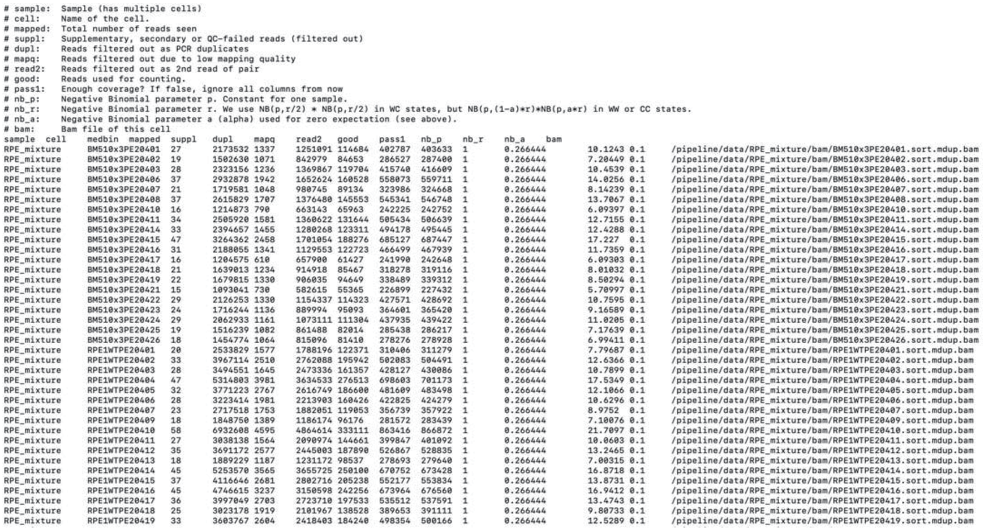
out/data/RPE_mixture/plots/counts/CountComplete.raw.pdf



Output – Plotting pipeline

Plotting pipeline – basic information for each single-cell libraries

out/data/RPE_mixture/counts/RPE_mixture.info_raw

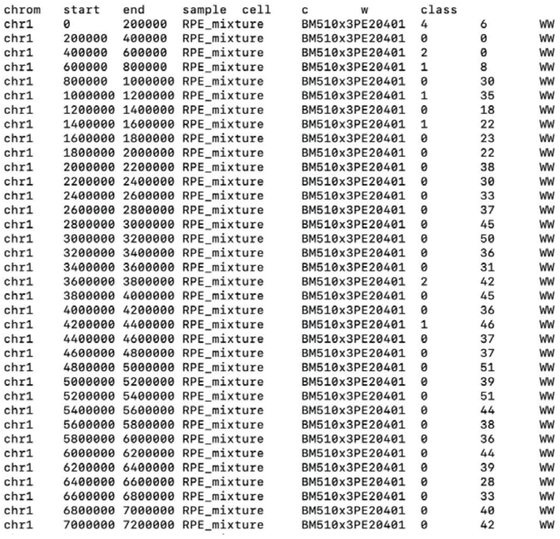


```
# sample:  Sample (has multiple cells)
# cell:    Name of the cell.
# mapped:  Total number of reads seen
# suppl:   Supplementary, secondary or QC-failed reads (filtered out)
# dupl:    Reads filtered out as PCR duplicates
# mapq:    Reads filtered out due to low mapping quality
# read2:   Reads filtered out as 2nd read of pair
# good:    Reads used for counting.
# pass1:   Enough coverage? If false, ignore all columns from now
# nb_p:    Negative Binomial parameter p. Constant for one sample.
# nb_r:    Negative Binomial parameter r. We use NB(p,r/2) * NB(p,r/2) in WC states, but NB(p,(1-a)*r)*NB(p,a*r) in WW or CC states.
# nb_a:    Negative Binomial parameter a (alpha) used for zero expectation (see above).
# bam:     Bam file of this cell
sample  cell    medbin  mapped  suppl   dupl    mapq    read2   good    pass1   nb_p    nb_r    nb_a    bam
RPE_mixture     BM510x3PE20401  27      2173532 1337    1251091 114684  402787  403633  1       0.266444        10.1243 0.1     /pipeline/data/RPE_mixture/bam/BM510x3PE20401.sort.mdup.bam
RPE_mixture     BM510x3PE20402  19      1502630 1071    842979  84653   286527  287400  1       0.266444        7.20449 0.1     /pipeline/data/RPE_mixture/bam/BM510x3PE20402.sort.mdup.bam
RPE_mixture     BM510x3PE20403  28      2323156 1236    1369867 119704  415740  416609  1       0.266444        10.4539 0.1     /pipeline/data/RPE_mixture/bam/BM510x3PE20403.sort.mdup.bam
RPE_mixture     BM510x3PE20406  37      2932878 1942    1652624 160528  558073  559711  1       0.266444        14.0256 0.1     /pipeline/data/RPE_mixture/bam/BM510x3PE20406.sort.mdup.bam
RPE_mixture     BM510x3PE20407  21      1719581 1048    980745  89134   323986  324668  1       0.266444        8.14239 0.1     /pipeline/data/RPE_mixture/bam/BM510x3PE20407.sort.mdup.bam
RPE_mixture     BM510x3PE20408  37      2615829 1707    1376480 145553  545341  546748  1       0.266444        13.7067 0.1     /pipeline/data/RPE_mixture/bam/BM510x3PE20408.sort.mdup.bam
RPE_mixture     BM510x3PE20410  16      1214873 790     663143  65963   242225  242752  1       0.266444        6.09397 0.1     /pipeline/data/RPE_mixture/bam/BM510x3PE20410.sort.mdup.bam
RPE_mixture     BM510x3PE20411  34      2505920 1581    1360622 131644  505434  506639  1       0.266444        12.7155 0.1     /pipeline/data/RPE_mixture/bam/BM510x3PE20411.sort.mdup.bam
RPE_mixture     BM510x3PE20414  33      2394657 1455    1280268 123311  494178  495445  1       0.266444        12.4288 0.1     /pipeline/data/RPE_mixture/bam/BM510x3PE20414.sort.mdup.bam
RPE_mixture     BM510x3PE20415  47      3264362 2458    1701054 188276  685127  687447  1       0.266444        17.227  0.1     /pipeline/data/RPE_mixture/bam/BM510x3PE20415.sort.mdup.bam
RPE_mixture     BM510x3PE20416  31      2188055 1341    1129553 122723  466499  467939  1       0.266444        11.7359 0.1     /pipeline/data/RPE_mixture/bam/BM510x3PE20416.sort.mdup.bam
RPE_mixture     BM510x3PE20417  16      1204575 610     657900  61427   241990  242648  1       0.266444        6.09303 0.1     /pipeline/data/RPE_mixture/bam/BM510x3PE20417.sort.mdup.bam
RPE_mixture     BM510x3PE20418  21      1639013 1234    914918  85467   318278  319116  1       0.266444        8.01032 0.1     /pipeline/data/RPE_mixture/bam/BM510x3PE20418.sort.mdup.bam
RPE_mixture     BM510x3PE20419  22      1679815 1330    906035  94649   338489  339312  1       0.266444        8.50294 0.1     /pipeline/data/RPE_mixture/bam/BM510x3PE20419.sort.mdup.bam
RPE_mixture     BM510x3PE20421  15      1093041 730     582615  55365   226899  227432  1       0.266444        5.70997 0.1     /pipeline/data/RPE_mixture/bam/BM510x3PE20421.sort.mdup.bam
RPE_mixture     BM510x3PE20422  29      2126253 1330    1154337 114323  427571  428692  1       0.266444        10.7595 0.1     /pipeline/data/RPE_mixture/bam/BM510x3PE20422.sort.mdup.bam
RPE_mixture     BM510x3PE20423  24      1716244 1136    889994  95093   364601  365420  1       0.266444        9.16589 0.1     /pipeline/data/RPE_mixture/bam/BM510x3PE20423.sort.mdup.bam
RPE_mixture     BM510x3PE20424  29      2062933 1161    1073111 111304  437935  439422  1       0.266444        11.0205 0.1     /pipeline/data/RPE_mixture/bam/BM510x3PE20424.sort.mdup.bam
RPE_mixture     BM510x3PE20425  19      1516239 1082    861488  82014   285438  286217  1       0.266444        7.17639 0.1     /pipeline/data/RPE_mixture/bam/BM510x3PE20425.sort.mdup.bam
RPE_mixture     BM510x3PE20426  18      1454774 1064    815096  81410   278276  278928  1       0.266444        6.99411 0.1     /pipeline/data/RPE_mixture/bam/BM510x3PE20426.sort.mdup.bam
RPE_mixture     RPE1WTPE20401   20      2533829 1577    1788196 122371  310406  311279  1       0.266444        7.79687 0.1     /pipeline/data/RPE_mixture/bam/RPE1WTPE20401.sort.mdup.bam
RPE_mixture     RPE1WTPE20402   33      3967114 2510    2762088 195942  502083  504491  1       0.266444        12.6366 0.1     /pipeline/data/RPE_mixture/bam/RPE1WTPE20402.sort.mdup.bam
RPE_mixture     RPE1WTPE20403   28      3494551 1645    2473336 161357  428127  430086  1       0.266444        10.7899 0.1     /pipeline/data/RPE_mixture/bam/RPE1WTPE20403.sort.mdup.bam
RPE_mixture     RPE1WTPE20404   47      5314803 3981    3634533 276513  698603  701173  1       0.266444        17.5349 0.1     /pipeline/data/RPE_mixture/bam/RPE1WTPE20404.sort.mdup.bam
RPE_mixture     RPE1WTPE20405   32      3771223 2767    2616749 186600  481609  483498  1       0.266444        12.1066 0.1     /pipeline/data/RPE_mixture/bam/RPE1WTPE20405.sort.mdup.bam
RPE_mixture     RPE1WTPE20406   28      3223414 1981    2213903 160426  422825  424279  1       0.266444        10.6296 0.1     /pipeline/data/RPE_mixture/bam/RPE1WTPE20406.sort.mdup.bam
RPE_mixture     RPE1WTPE20407   23      2717518 1753    1882051 119053  356739  357922  1       0.266444        8.9752  0.1     /pipeline/data/RPE_mixture/bam/RPE1WTPE20407.sort.mdup.bam
RPE_mixture     RPE1WTPE20409   18      1848750 1389    1186174 96176   281572  283439  1       0.266444        7.10076 0.1     /pipeline/data/RPE_mixture/bam/RPE1WTPE20409.sort.mdup.bam
RPE_mixture     RPE1WTPE20410   58      6932608 4595    4864614 333111  863416  866872  1       0.266444        21.7097 0.1     /pipeline/data/RPE_mixture/bam/RPE1WTPE20410.sort.mdup.bam
RPE_mixture     RPE1WTPE20411   27      3038138 1564    2090974 144661  399847  401092  1       0.266444        10.0603 0.1     /pipeline/data/RPE_mixture/bam/RPE1WTPE20411.sort.mdup.bam
RPE_mixture     RPE1WTPE20412   35      3691172 2577    2445003 187890  526867  528835  1       0.266444        13.2465 0.1     /pipeline/data/RPE_mixture/bam/RPE1WTPE20412.sort.mdup.bam
RPE_mixture     RPE1WTPE20413   18      1889229 1187    1231172 98537   278693  279640  1       0.266444        7.00315 0.1     /pipeline/data/RPE_mixture/bam/RPE1WTPE20413.sort.mdup.bam
RPE_mixture     RPE1WTPE20414   45      5253570 3565    3655725 250100  670752  673428  1       0.266444        16.8718 0.1     /pipeline/data/RPE_mixture/bam/RPE1WTPE20414.sort.mdup.bam
RPE_mixture     RPE1WTPE20415   37      4116646 2681    2802716 205238  552177  553834  1       0.266444        13.8731 0.1     /pipeline/data/RPE_mixture/bam/RPE1WTPE20415.sort.mdup.bam
RPE_mixture     RPE1WTPE20416   45      4746615 3237    3150598 242256  673964  676560  1       0.266444        16.9412 0.1     /pipeline/data/RPE_mixture/bam/RPE1WTPE20416.sort.mdup.bam
RPE_mixture     RPE1WTPE20417   36      3997049 2703    2723710 197533  535512  537591  1       0.266444        13.4743 0.1     /pipeline/data/RPE_mixture/bam/RPE1WTPE20417.sort.mdup.bam
RPE_mixture     RPE1WTPE20418   25      3023178 1919    2101967 138528  389653  391111  1       0.266444        9.80733 0.1     /pipeline/data/RPE_mixture/bam/RPE1WTPE20418.sort.mdup.bam
RPE_mixture     RPE1WTPE20419   33      3603767 2604    2418403 184240  498354  500166  1       0.266444        12.5289 0.1     /pipeline/data/RPE_mixture/bam/RPE1WTPE20419.sort.mdup.bam
```

Output – Plotting pipeline

Plotting pipeline – single-cell count matrix using 200kb bins

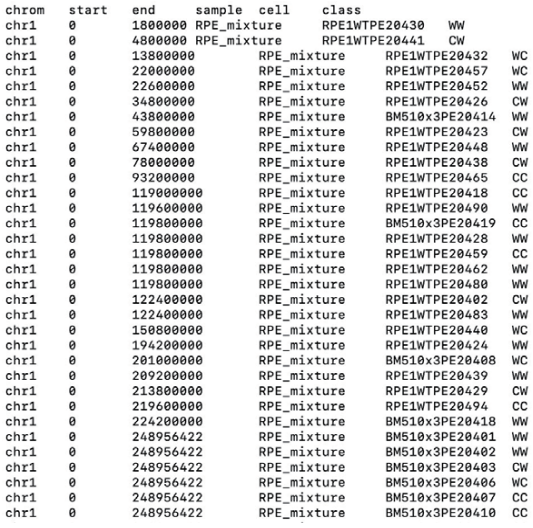
out/data/RPE_mixture/counts/RPE_mixture.txt.raw.gz



```
chrom   start   end     sample  cell    c       w       class
chr1    0       200000  RPE_mixture     BM510x3PE20401  4       6       WW
chr1    200000  400000  RPE_mixture     BM510x3PE20401  0       0       WW
chr1    400000  600000  RPE_mixture     BM510x3PE20401  2       0       WW
chr1    600000  800000  RPE_mixture     BM510x3PE20401  1       8       WW
chr1    800000  1000000 RPE_mixture     BM510x3PE20401  0       30      WW
chr1    1000000 1200000 RPE_mixture     BM510x3PE20401  1       35      WW
chr1    1200000 1400000 RPE_mixture     BM510x3PE20401  0       18      WW
chr1    1400000 1600000 RPE_mixture     BM510x3PE20401  1       22      WW
chr1    1600000 1800000 RPE_mixture     BM510x3PE20401  0       23      WW
chr1    1800000 2000000 RPE_mixture     BM510x3PE20401  0       22      WW
chr1    2000000 2200000 RPE_mixture     BM510x3PE20401  0       38      WW
chr1    2200000 2400000 RPE_mixture     BM510x3PE20401  0       30      WW
chr1    2400000 2600000 RPE_mixture     BM510x3PE20401  0       33      WW
chr1    2600000 2800000 RPE_mixture     BM510x3PE20401  0       37      WW
chr1    2800000 3000000 RPE_mixture     BM510x3PE20401  0       45      WW
chr1    3000000 3200000 RPE_mixture     BM510x3PE20401  0       50      WW
chr1    3200000 3400000 RPE_mixture     BM510x3PE20401  0       36      WW
chr1    3400000 3600000 RPE_mixture     BM510x3PE20401  0       31      WW
chr1    3600000 3800000 RPE_mixture     BM510x3PE20401  2       42      WW
chr1    3800000 4000000 RPE_mixture     BM510x3PE20401  0       45      WW
chr1    4000000 4200000 RPE_mixture     BM510x3PE20401  0       36      WW
chr1    4200000 4400000 RPE_mixture     BM510x3PE20401  1       46      WW
chr1    4400000 4600000 RPE_mixture     BM510x3PE20401  0       37      WW
chr1    4600000 4800000 RPE_mixture     BM510x3PE20401  0       37      WW
chr1    4800000 5000000 RPE_mixture     BM510x3PE20401  0       51      WW
chr1    5000000 5200000 RPE_mixture     BM510x3PE20401  0       39      WW
chr1    5200000 5400000 RPE_mixture     BM510x3PE20401  0       51      WW
chr1    5400000 5600000 RPE_mixture     BM510x3PE20401  0       44      WW
chr1    5600000 5800000 RPE_mixture     BM510x3PE20401  0       38      WW
chr1    5800000 6000000 RPE_mixture     BM510x3PE20401  0       36      WW
chr1    6000000 6200000 RPE_mixture     BM510x3PE20401  0       44      WW
chr1    6200000 6400000 RPE_mixture     BM510x3PE20401  0       39      WW
chr1    6400000 6600000 RPE_mixture     BM510x3PE20401  0       28      WW
chr1    6600000 6800000 RPE_mixture     BM510x3PE20401  0       33      WW
chr1    6800000 7000000 RPE_mixture     BM510x3PE20401  0       40      WW
chr1    7000000 7200000 RPE_mixture     BM510x3PE20401  0       42      WW
```

Output – haplotype phasing result

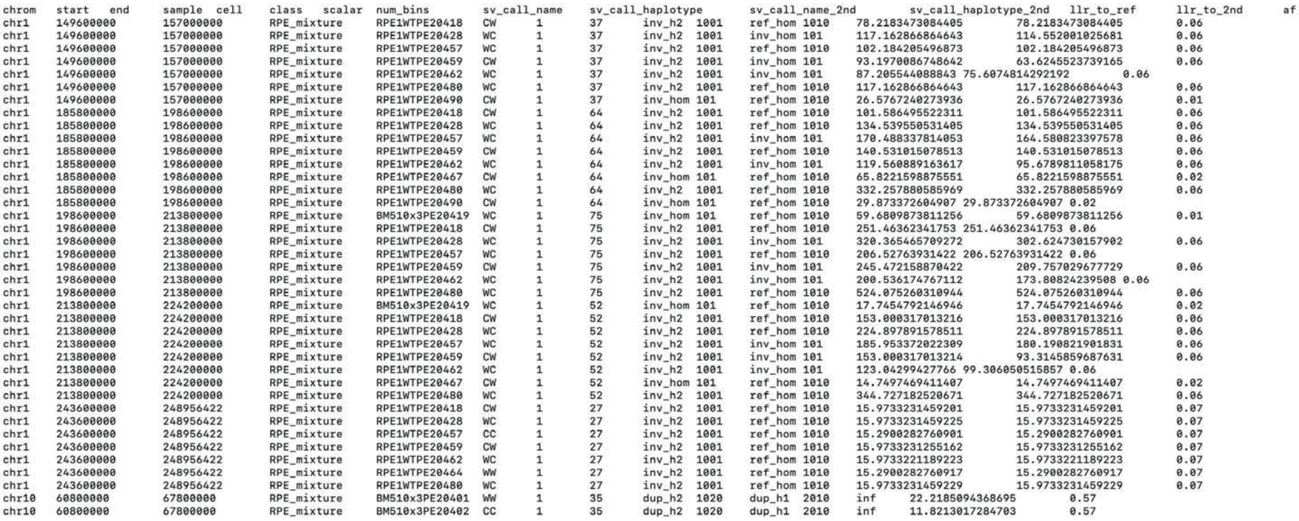
data/RPE_mixture/strandphaser/StrandPhaseR_final_output.txt



```
chrom   start   end     sample  cell    class
chr1    0       1800000 RPE_mixture     RPE1WTPE20430   WW
chr1    0       4800000 RPE_mixture     RPE1WTPE20441   CW
chr1    0       13800000        RPE_mixture     RPE1WTPE20432   WC
chr1    0       22000000        RPE_mixture     RPE1WTPE20457   WC
chr1    0       22600000        RPE_mixture     RPE1WTPE20452   WW
chr1    0       34800000        RPE_mixture     RPE1WTPE20426   CW
chr1    0       43800000        RPE_mixture     BM510x3PE20414  WW
chr1    0       59800000        RPE_mixture     RPE1WTPE20423   CW
chr1    0       67400000        RPE_mixture     RPE1WTPE20448   WW
chr1    0       78000000        RPE_mixture     RPE1WTPE20438   CW
chr1    0       93200000        RPE_mixture     RPE1WTPE20465   CC
chr1    0       119000000       RPE_mixture     RPE1WTPE20418   CC
chr1    0       119600000       RPE_mixture     RPE1WTPE20490   WW
chr1    0       119800000       RPE_mixture     BM510x3PE20419  CC
chr1    0       119800000       RPE_mixture     RPE1WTPE20428   WW
chr1    0       119800000       RPE_mixture     RPE1WTPE20459   CC
chr1    0       119800000       RPE_mixture     RPE1WTPE20462   WW
chr1    0       119800000       RPE_mixture     RPE1WTPE20480   WW
chr1    0       122400000       RPE_mixture     RPE1WTPE20402   CW
chr1    0       122400000       RPE_mixture     RPE1WTPE20483   WW
chr1    0       150800000       RPE_mixture     RPE1WTPE20440   WC
chr1    0       194200000       RPE_mixture     RPE1WTPE20424   WW
chr1    0       201000000       RPE_mixture     BM510x3PE20408  WC
chr1    0       209200000       RPE_mixture     RPE1WTPE20439   WW
chr1    0       213800000       RPE_mixture     RPE1WTPE20429   CW
chr1    0       219600000       RPE_mixture     RPE1WTPE20494   CC
chr1    0       224200000       RPE_mixture     BM510x3PE20418  WW
chr1    0       248956422       RPE_mixture     BM510x3PE20401  WW
chr1    0       248956422       RPE_mixture     BM510x3PE20402  WW
chr1    0       248956422       RPE_mixture     BM510x3PE20403  CW
chr1    0       248956422       RPE_mixture     BM510x3PE20406  WC
chr1    0       248956422       RPE_mixture     BM510x3PE20407  CC
chr1    0       248956422       RPE_mixture     BM510x3PE20410  CC
```

Output – SV calling result

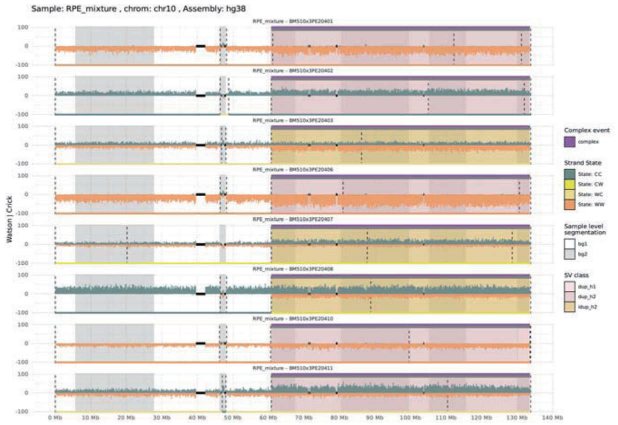
data/RPE_mixture/mosaiclassifier/sv_calls/stringent_filterTRUE.tsv



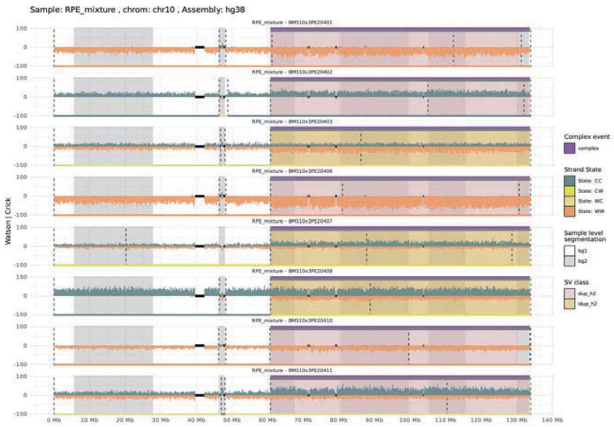
```
chrom   start   end     sample  cell    class   scalar  num_bins        sv_call_name    sv_call_haplotype       sv_call_name_2nd        sv_call_haplotype_2nd   llr_to_ref      llr_to_2nd      af
chr1    149600000       157000000       RPE_mixture     RPE1WTPE20418   CW      1       37      inv_h2  1001    ref_hom 1010    78.2183473084405        78.2183473084405        0.06
chr1    149600000       157000000       RPE_mixture     RPE1WTPE20428   WC      1       37      inv_h2  1001    inv_hom 101     117.162866864643        114.552001025681        0.06
chr1    149600000       157000000       RPE_mixture     RPE1WTPE20457   WC      1       37      inv_h2  1001    ref_hom 1010    102.184205496873        102.184205496873        0.06
chr1    149600000       157000000       RPE_mixture     RPE1WTPE20459   CW      1       37      inv_h2  1001    inv_hom 101     93.1970086748642        63.6245523739165        0.06
chr1    149600000       157000000       RPE_mixture     RPE1WTPE20462   WC      1       37      inv_h2  1001    inv_hom 101     87.205544088843 75.6074814292192        0.06
chr1    149600000       157000000       RPE_mixture     RPE1WTPE20480   WC      1       37      inv_h2  1001    ref_hom 1010    117.162866864643        117.162866864643        0.06
chr1    149600000       157000000       RPE_mixture     RPE1WTPE20490   CW      1       37      inv_hom 101     ref_hom 1010    26.5767240273936        26.5767240273936        0.01
chr1    185800000       198600000       RPE_mixture     RPE1WTPE20418   CW      1       64      inv_h2  1001    ref_hom 1010    101.586495522311        101.586495522311        0.06
chr1    185800000       198600000       RPE_mixture     RPE1WTPE20428   WC      1       64      inv_h2  1001    ref_hom 1010    134.539550531405        134.539550531405        0.06
chr1    185800000       198600000       RPE_mixture     RPE1WTPE20457   WC      1       64      inv_h2  1001    inv_hom 101     170.488337814053        164.580823397578        0.06
chr1    185800000       198600000       RPE_mixture     RPE1WTPE20459   CW      1       64      inv_h2  1001    ref_hom 1010    140.531015078513        140.531015078513        0.06
chr1    185800000       198600000       RPE_mixture     RPE1WTPE20462   WC      1       64      inv_h2  1001    inv_hom 101     119.560889163617        95.6789811058175        0.06
chr1    185800000       198600000       RPE_mixture     RPE1WTPE20467   CW      1       64      inv_hom 101     ref_hom 1010    65.8221598875551        65.8221598875551        0.02
chr1    185800000       198600000       RPE_mixture     RPE1WTPE20480   WC      1       64      inv_h2  1001    ref_hom 1010    332.257880585969        332.257880585969        0.06
chr1    185800000       198600000       RPE_mixture     RPE1WTPE20490   CW      1       64      inv_hom 101     ref_hom 1010    29.873372604907 29.873372604907 0.02
chr1    198600000       213800000       RPE_mixture     BM510x3PE20419  WC      1       75      inv_hom 101     ref_hom 1010    59.6809873811256        59.6809873811256        0.01
chr1    198600000       213800000       RPE_mixture     RPE1WTPE20418   CW      1       75      inv_h2  1001    ref_hom 1010    251.46362341753 251.46362341753 0.06
chr1    198600000       213800000       RPE_mixture     RPE1WTPE20428   WC      1       75      inv_h2  1001    inv_hom 101     320.365465709272        302.624730157902        0.06
chr1    198600000       213800000       RPE_mixture     RPE1WTPE20457   WC      1       75      inv_h2  1001    ref_hom 1010    206.52763931422 206.52763931422 0.06
chr1    198600000       213800000       RPE_mixture     RPE1WTPE20459   CW      1       75      inv_h2  1001    inv_hom 101     245.472158870422        209.757029677729        0.06
chr1    198600000       213800000       RPE_mixture     RPE1WTPE20462   WC      1       75      inv_h2  1001    inv_hom 101     200.536174767112        173.80824239508 0.06
chr1    198600000       213800000       RPE_mixture     RPE1WTPE20480   WC      1       75      inv_h2  1001    ref_hom 1010    524.075260310944        524.075260310944        0.06
chr1    213800000       224200000       RPE_mixture     BM510x3PE20419  WC      1       52      inv_hom 101     ref_hom 1010    17.7454792146946        17.7454792146946        0.02
chr1    213800000       224200000       RPE_mixture     RPE1WTPE20418   CW      1       52      inv_h2  1001    ref_hom 1010    153.000317013216        153.000317013216        0.06
chr1    213800000       224200000       RPE_mixture     RPE1WTPE20428   WC      1       52      inv_h2  1001    ref_hom 1010    224.897891578511        224.897891578511        0.06
chr1    213800000       224200000       RPE_mixture     RPE1WTPE20457   WC      1       52      inv_h2  1001    inv_hom 101     185.953372022309        180.190821901831        0.06
chr1    213800000       224200000       RPE_mixture     RPE1WTPE20459   CW      1       52      inv_h2  1001    inv_hom 101     153.000317013214        93.3145859687631        0.06
chr1    213800000       224200000       RPE_mixture     RPE1WTPE20462   WC      1       52      inv_h2  1001    inv_hom 101     123.04299427766 99.306050515857 0.06
chr1    213800000       224200000       RPE_mixture     RPE1WTPE20467   CW      1       52      inv_hom 101     ref_hom 1010    14.7497469411407        14.7497469411407        0.02
chr1    213800000       224200000       RPE_mixture     RPE1WTPE20480   WC      1       52      inv_h2  1001    ref_hom 1010    344.727182520671        344.727182520671        0.06
chr1    243600000       248956422       RPE_mixture     RPE1WTPE20418   CW      1       27      inv_h2  1001    ref_hom 1010    15.9733231459201        15.9733231459201        0.07
chr1    243600000       248956422       RPE_mixture     RPE1WTPE20428   WC      1       27      inv_h2  1001    ref_hom 1010    15.9733231459225        15.9733231459225        0.07
chr1    243600000       248956422       RPE_mixture     RPE1WTPE20457   CC      1       27      inv_h2  1001    ref_hom 1010    15.2900282760901        15.2900282760901        0.07
chr1    243600000       248956422       RPE_mixture     RPE1WTPE20459   CW      1       27      inv_h2  1001    ref_hom 1010    15.9733231255162        15.9733231255162        0.07
chr1    243600000       248956422       RPE_mixture     RPE1WTPE20462   WC      1       27      inv_h2  1001    ref_hom 1010    15.9733221189223        15.9733221189223        0.07
chr1    243600000       248956422       RPE_mixture     RPE1WTPE20464   WW      1       27      inv_h2  1001    ref_hom 1010    15.2900282760917        15.2900282760917        0.07
chr1    243600000       248956422       RPE_mixture     RPE1WTPE20480   WC      1       27      inv_h2  1001    ref_hom 1010    15.9733231459229        15.9733231459229        0.07
chr10   60800000        67800000        RPE_mixture     BM510x3PE20401  WW      1       35      dup_h2  1020    dup_h1  2010    inf     22.2185094368695        0.57
chr10   60800000        67800000        RPE_mixture     BM510x3PE20402  CC      1       35      dup_h2  1020    dup_h1  2010    inf     11.8213017284703        0.57
```

Output – SV calling result

data/RPE_mixture/plots/sv_calls_dev/stringent_filterTRUE/*.pdf

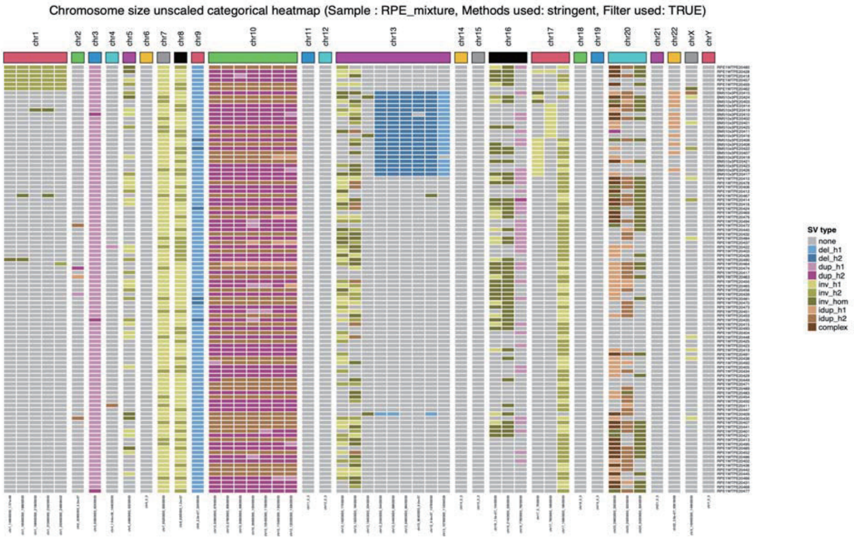


data/RPE_mixture/plots/sv_calls_dev/lenient_filterFALSE/*.pdf



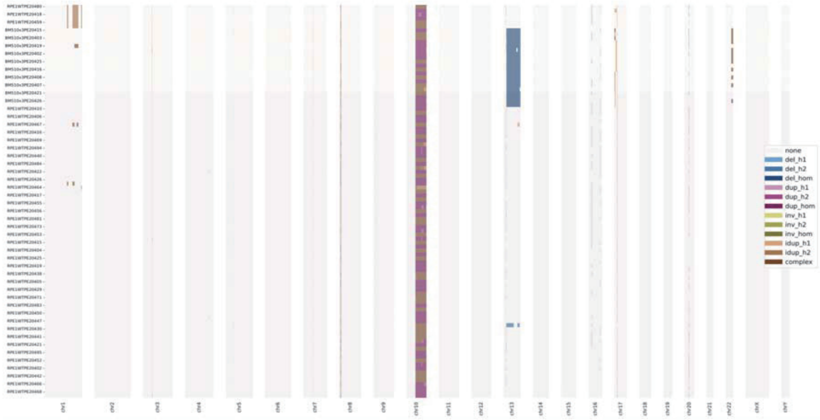
Output – SV calling result

data/RPE_mixture/plots/sv_clustering_dev/stringent-filterTRUE-position.pdf



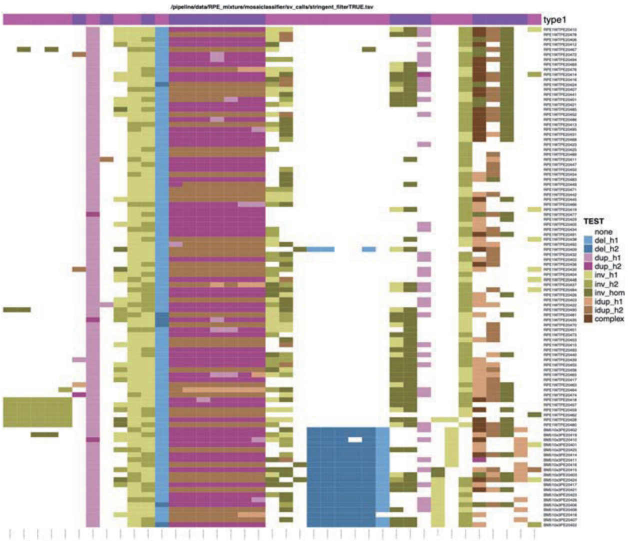
Output – SV calling result

data/RPE_mixture/plots/sv_clustering_dev/stringent-filterTRUE-chromosome.pdf



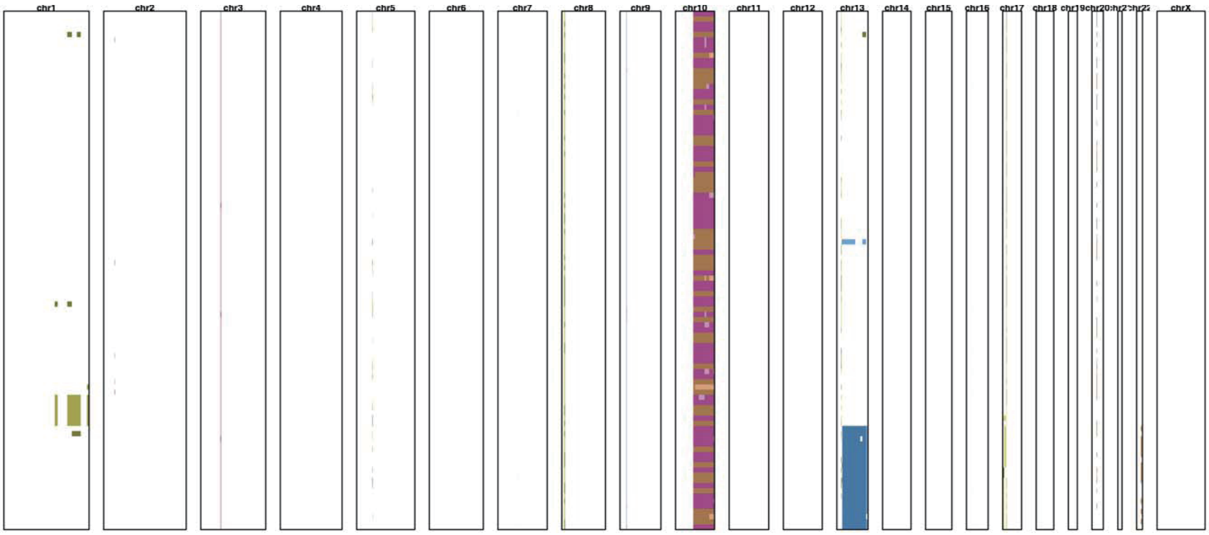
Output – SV calling result

data/RPE_mixture/plots/sv_clustering/stringent-filterTRUE-position.pdf



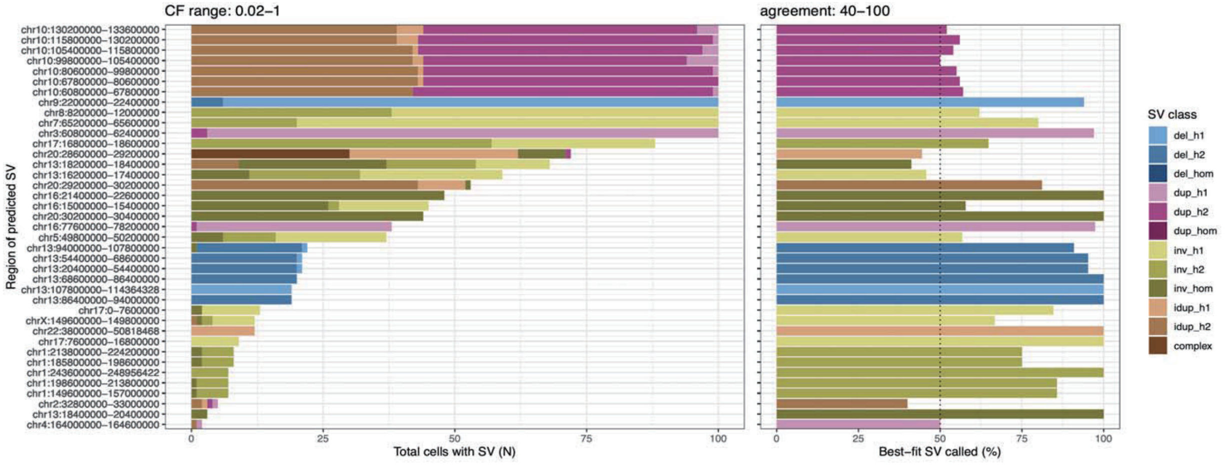
Output – SV calling result

data/RPE_mixture/plots/sv_clustering/stringent-filterTRUE-chromosome.pdf



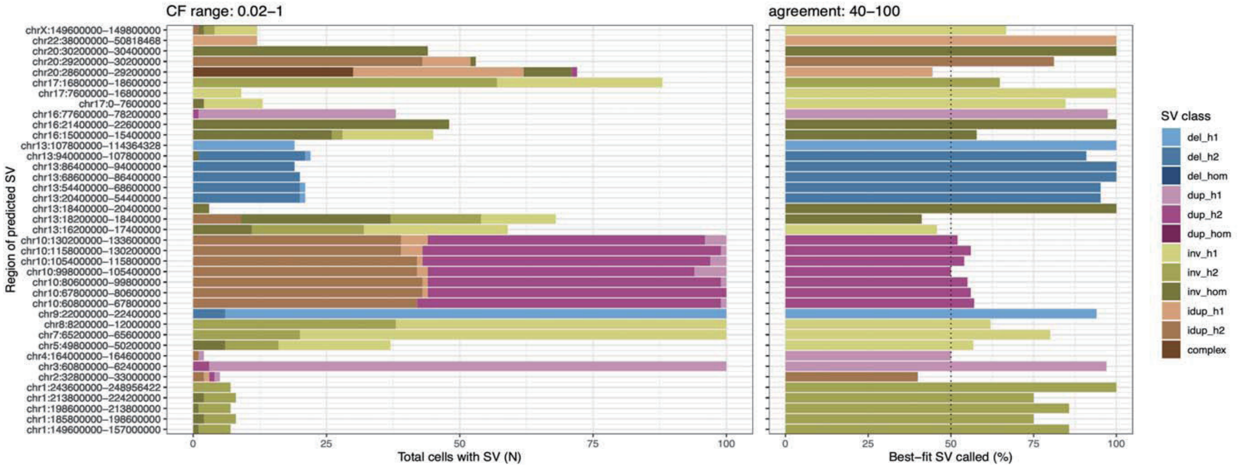
Output – SV calling result

data/RPE_mixture/plots/sv_consistency/stringent_filterTRUE.consistency-barplot-byaf.pdf



Output – SV calling result

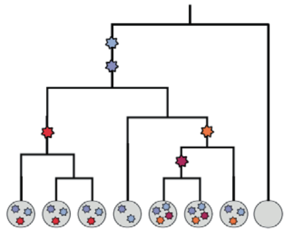
data/RPE_mixture/plots/sv_consistency/stringent_filterTRUE.consistency-barplot-bypos.pdf



Part3. scNOVA – Strand-seq 에서 동정한 서브클론의 기능적 분석을 위한 멀티오믹스 기법

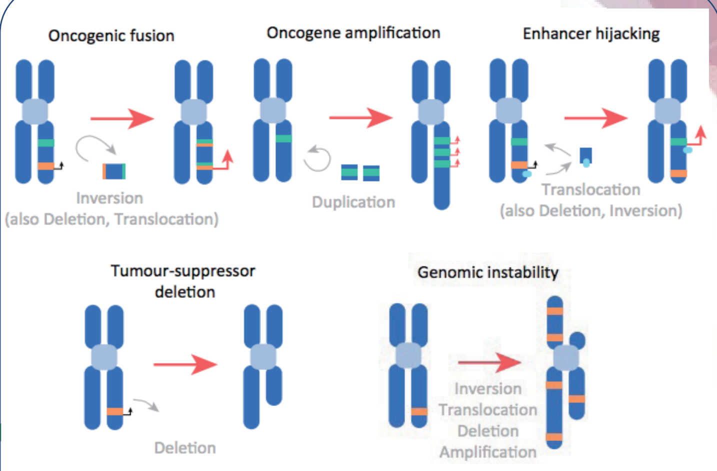
Single-cell multi-omics analysis to study tumor subclones

How can we measure functional consequence of somatic structural variants in different subclones?

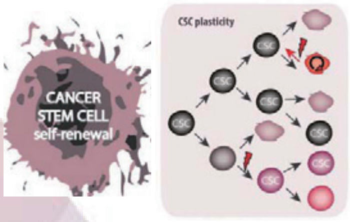


Genome Biology, 2016

Genetic variation



Macintyre et al. 2016

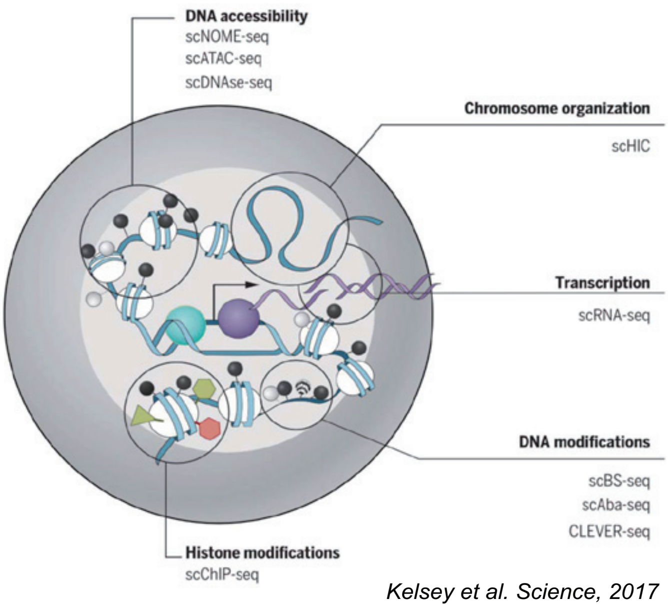


Molecular cancer, 2017

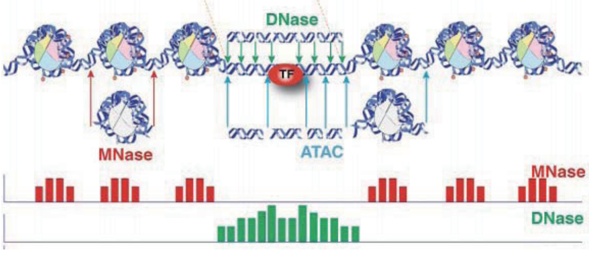
?

Epigenetic (functional) alteration

Single-cell technologies to explore *functional* heterogeneity



Kelsey et al. Science, 2017



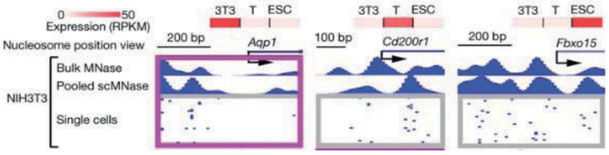
LETTER

scMNase-seq, Lai et al. 2018

https://doi.org/10.1038/s41586-018-0567-3

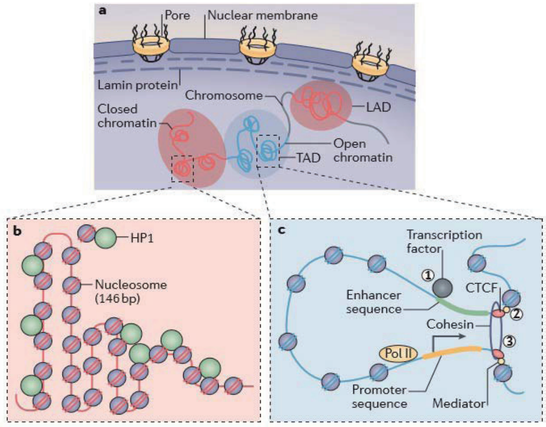
Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing

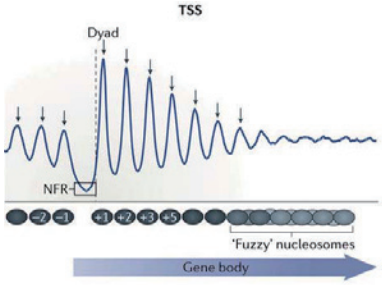
Binbin Lai[1], Weiwu Gao[1,2], Kairong Cui[1], Wanli Xie[1,3], Qingsong Tang[1], Wenfei Jin[4], Gangqing Hu[1], Bing Ni[2] & Keji Zhao[1]*

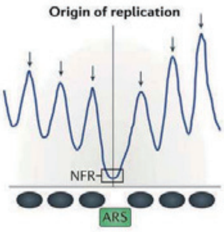


Can we use Nucleosome Occupancy to study functional consequence of SVs ?

Nucleosomes are the basic unit of chromatin which slide along DNA





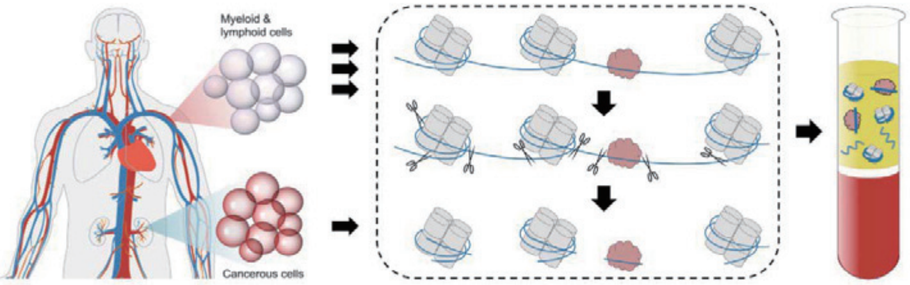


- Nucleosome is composed of two copies of four core histones together with 146~147bp of DNA
- Human diploid genomes have 30 million nucleosomes
- Transcriptionally active gene promoters exhibit a prominent nucleosome-depleted region (NDR) directly upstream of the TSS

Nat Rev Mol Cell Biol, 2017

Nucleosomes pattern is informative for the gene expression and cell type of origin

Cell free DNA protected by nucleosome is secreted to the blood



Nucleosomes pattern is informative for the gene expression and cell type of origin

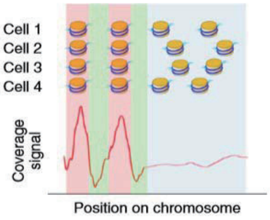
LETTERS

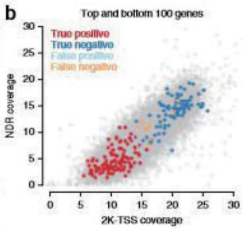
nature genetics

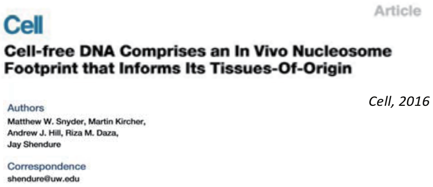
Inferring expressed genes by whole-genome sequencing of plasma DNA

Peter Ulz[1], Gerhard G Thallinger[2,3], Martina Auer[1], Ricarda Graf[1], Karl Kashofer[4], Stephan W Jahn[4], Luca Abete[4], Gunda Pristauz[5], Edgar Petru[5], Jochen B Geigl[1], Ellen Heitzer[1] & Michael R Speicher[1]

Nat Genet, 2016






Cell

Article

Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin

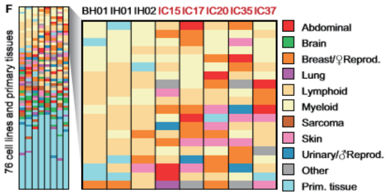
Authors

Matthew W. Snyder, Martin Kircher, Andrew J. Hill, Riza M. Daza, Jay Shendure

Correspondence

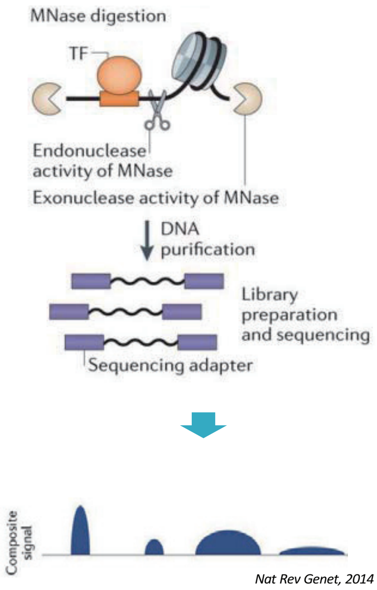
shendure@uw.edu

Cell, 2016



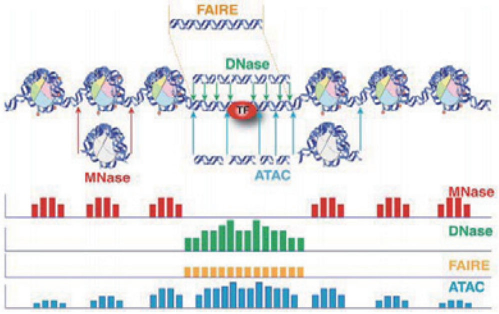
Nucleosome dynamics can be measured by genomic assays

MNase-seq


MNase digestion
TF
Endonuclease activity of MNase
Exonuclease activity of MNase
DNA purification
Library preparation and sequencing
Sequencing adapter
Composite signal

Nat Rev Genet, 2014

- MNase is a secreted glycoprotein with a preference for single-stranded DNA and RNA
- It cleave one strand of DNA when the helix ‘breathes’ and subsequently cleave the other strand to generate a double-strand break
- It then ‘nibbles’ the exposed DNA end until it reaches an obstruction, such as a nucleosome

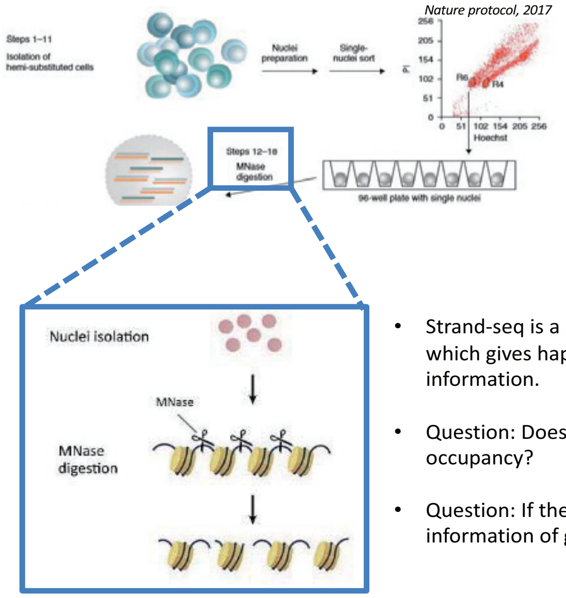


Epigenetics & Chromatin, 2014

Strand-seq protocol involves MNase treatment

Strand-seq protocol

Sanders et al.
Nature protocol, 2017

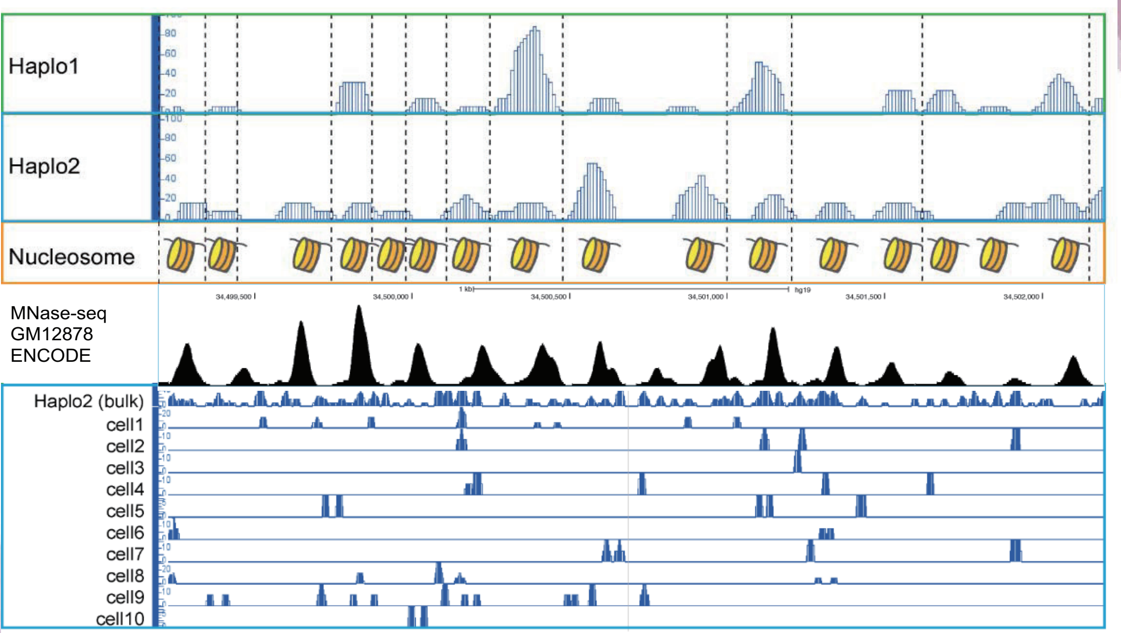


Trends in genetics, 2017

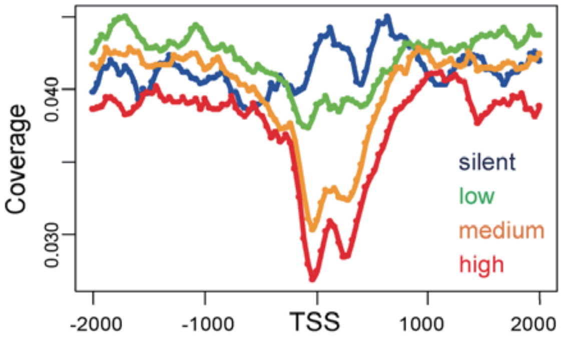
- Strand-seq is a single-cell based DNA sequencing method which gives haplotype-resolved structural variation information.
- Question: Does Strand-seq profile reflects nucleosome occupancy?
- Question: If then, can Strand-seq additionally provides information of gene expression and cell identity?

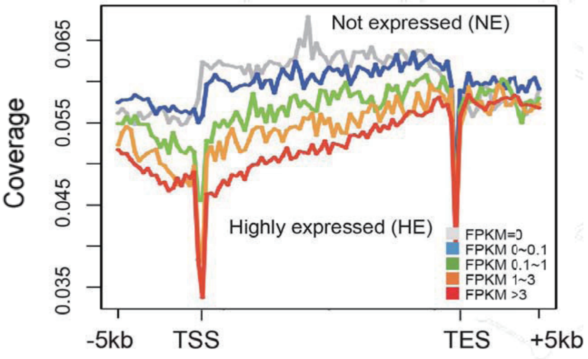
Nucleosome position and occupancy can be detected from Strand-seq data

Han Chinese (CHS) trio (Lymphoblastoid cell line) chr12:34,346,260-34,349,260



Nucleosome occupancy is negatively correlated with gene expression level





Nucleosome occupancy in the genebody is informative for differential expression

Input data (Strand-seq)

RPE-1 (182 cells)

	cell1	cell2	…	cell N
Gene1	10	30	…	5
Gene2	3	2	…	0
…	…	…	…	…
Gene N	30	50	…	80

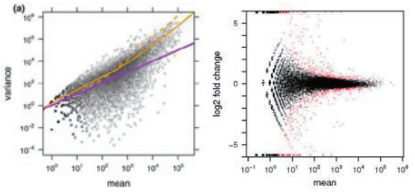
19770 genes

LCL (224 cells)

	cell1	cell2	…	cell N
Gene1	1	2	…	1
Gene2	8	4	…	5
…	…	…	…	…
Gene N	14	25	…	10

19770 genes

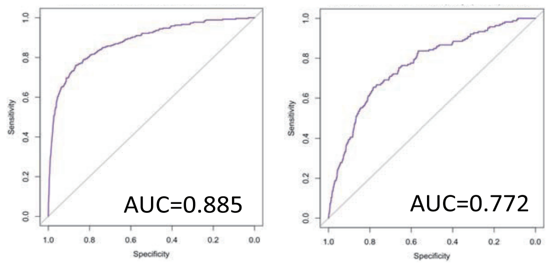
Approach (DESeq of nucleosome occupancy)



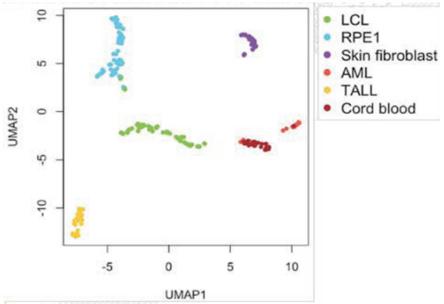
Anders et al. 2010, Love et al. 2014

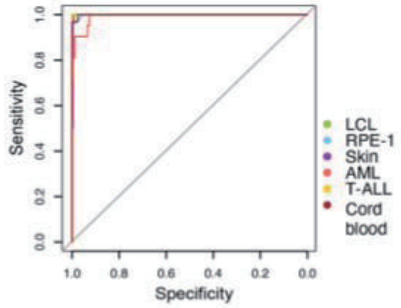
RPE1 up-regulated DEGs

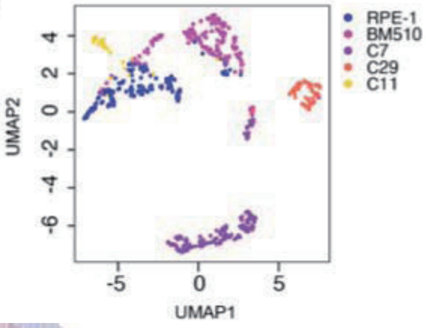
LCL up-regulated DEGs

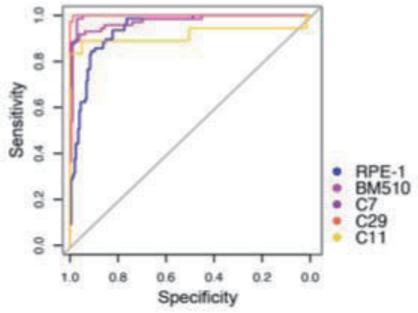


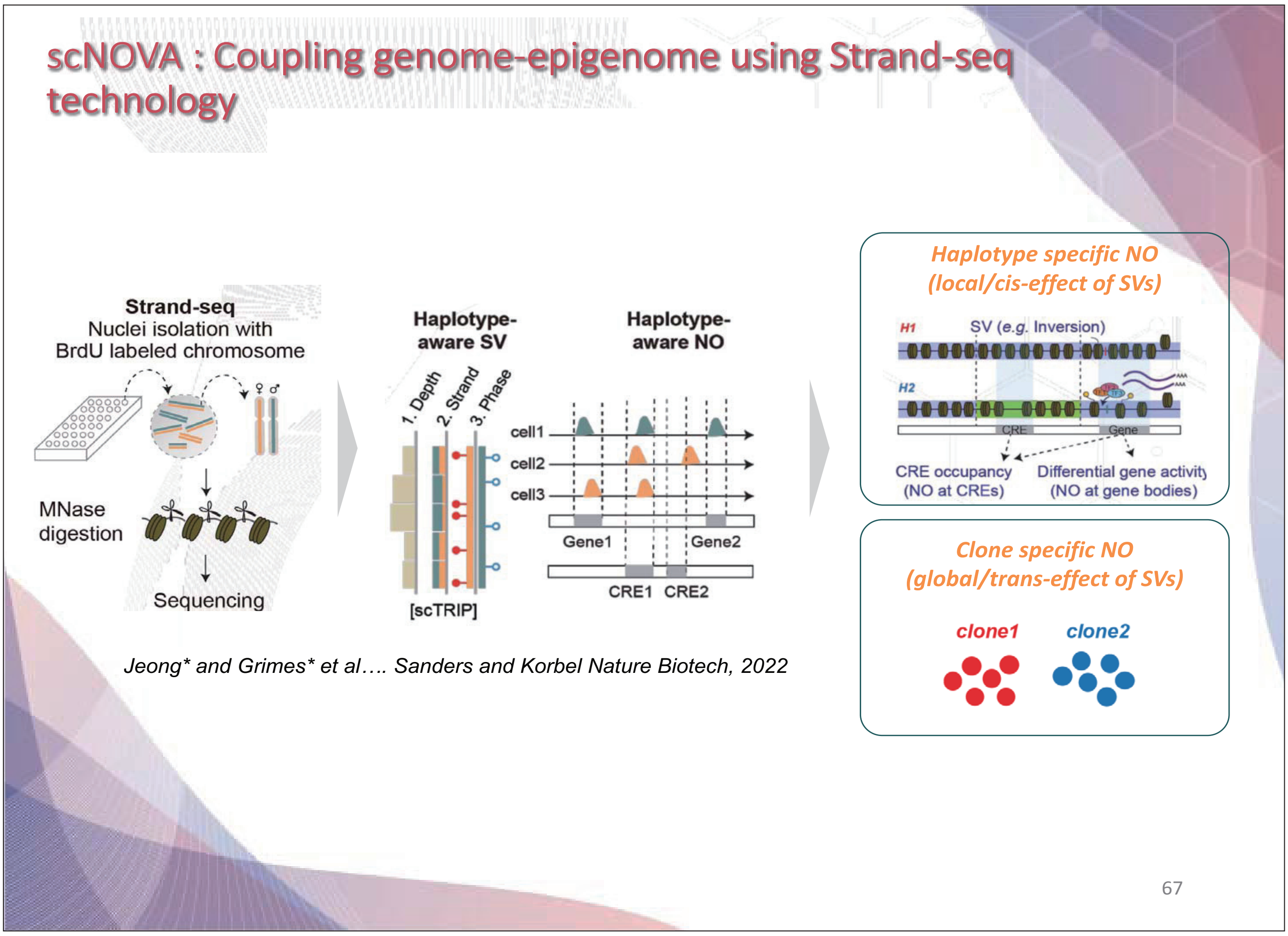
Nucleosome occupancy can be used to classify cell-type









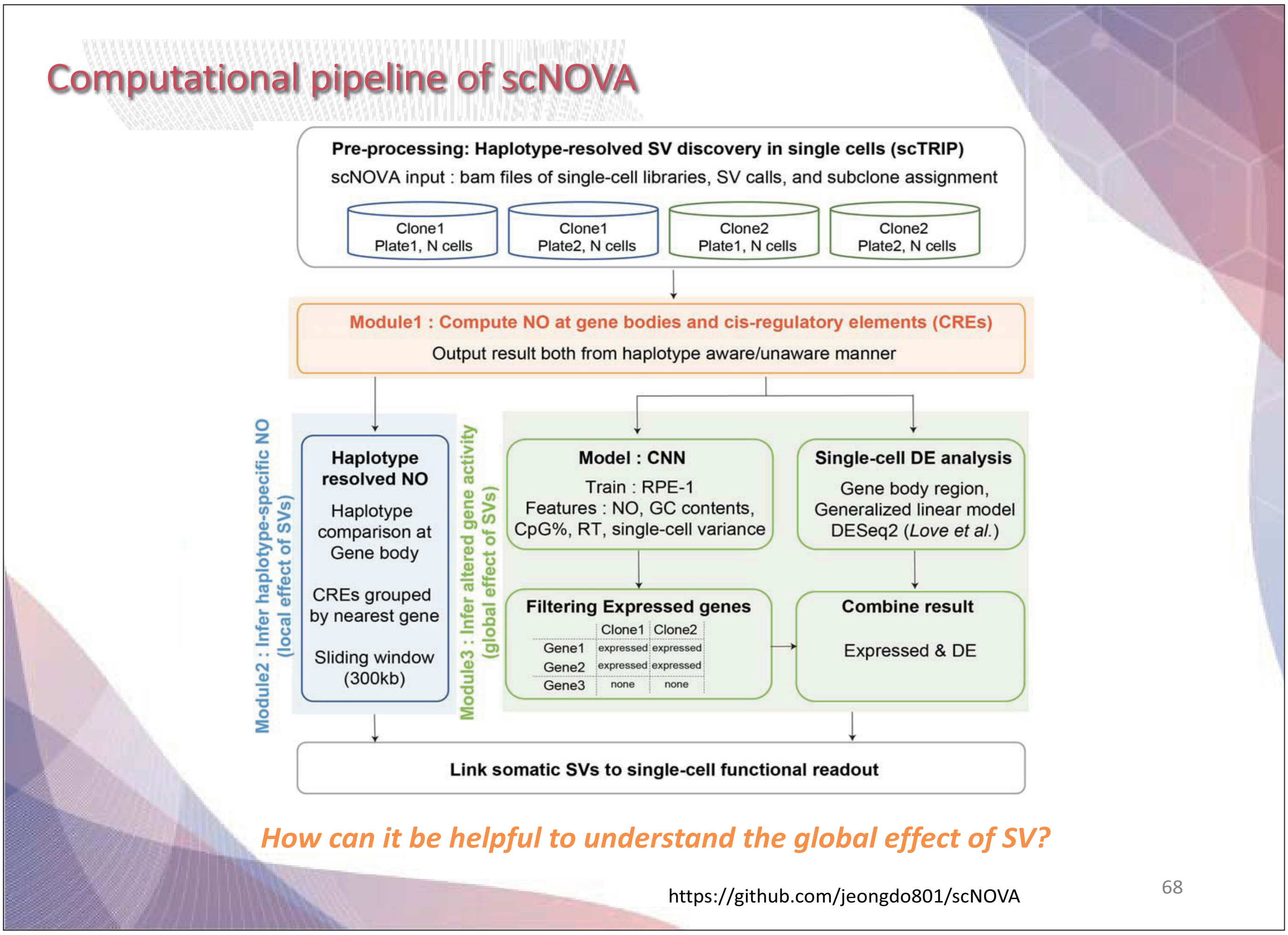


scNOVA : Coupling genome-epigenome using Strand-seq technology

Jeong and Grimes* et al…. Sanders and Korbel Nature Biotech, 2022*

Haplotype specific NO
(local/cis-effect of SVs)

Clone specific NO
(global/trans-effect of SVs)



Computational pipeline of scNOVA

Pre-processing: Haplotype-resolved SV discovery in single cells (scTRIP)

scNOVA input : bam files of single-cell libraries, SV calls, and subclone assignment

Module1 : Compute NO at gene bodies and cis-regulatory elements (CREs)

Output result both from haplotype aware/unaware manner

Module2 : Infer haplotype-specific NO (local effect of SVs)

Haplotype resolved NO

Haplotype comparison at Gene body

CREs grouped by nearest gene

Sliding window (300kb)

Module3 : Infer altered gene activity (global effect of SVs)

Model : CNN

Train : RPE-1

Features : NO, GC contents, CpG%, RT, single-cell variance

Single-cell DE analysis

Gene body region, Generalized linear model DESeq2 (*Love et al.*)

Filtering Expressed genes

	Clone1	Clone2
Gene1	expressed	expressed
Gene2	expressed	expressed
Gene3	none	none

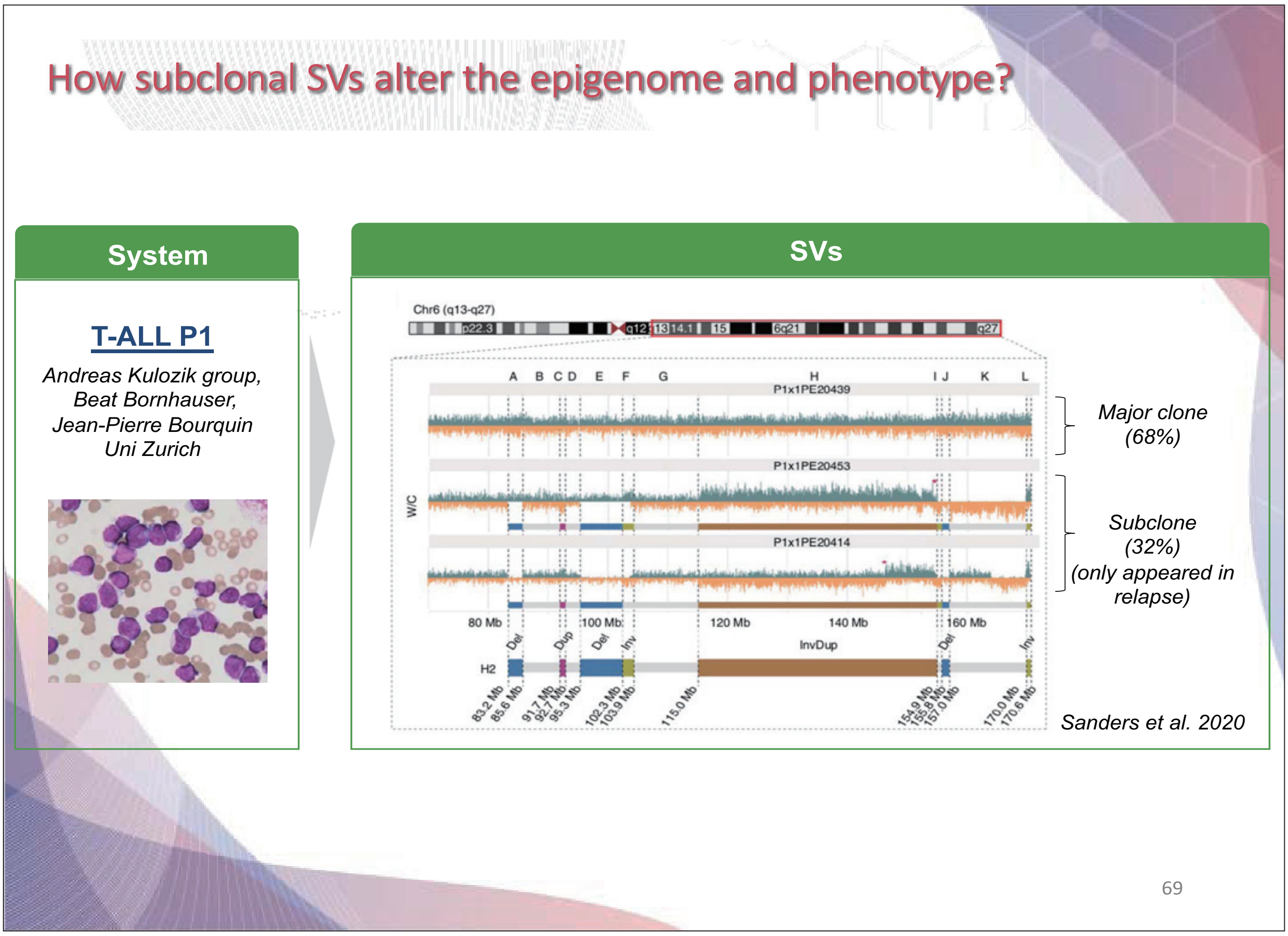
Combine result

Expressed & DE

Link somatic SVs to single-cell functional readout

How can it be helpful to understand the global effect of SV?

https://github.com/jeongdo801/scNOVA



How subclonal SVs alter the epigenome and phenotype?

System

T-ALL P1

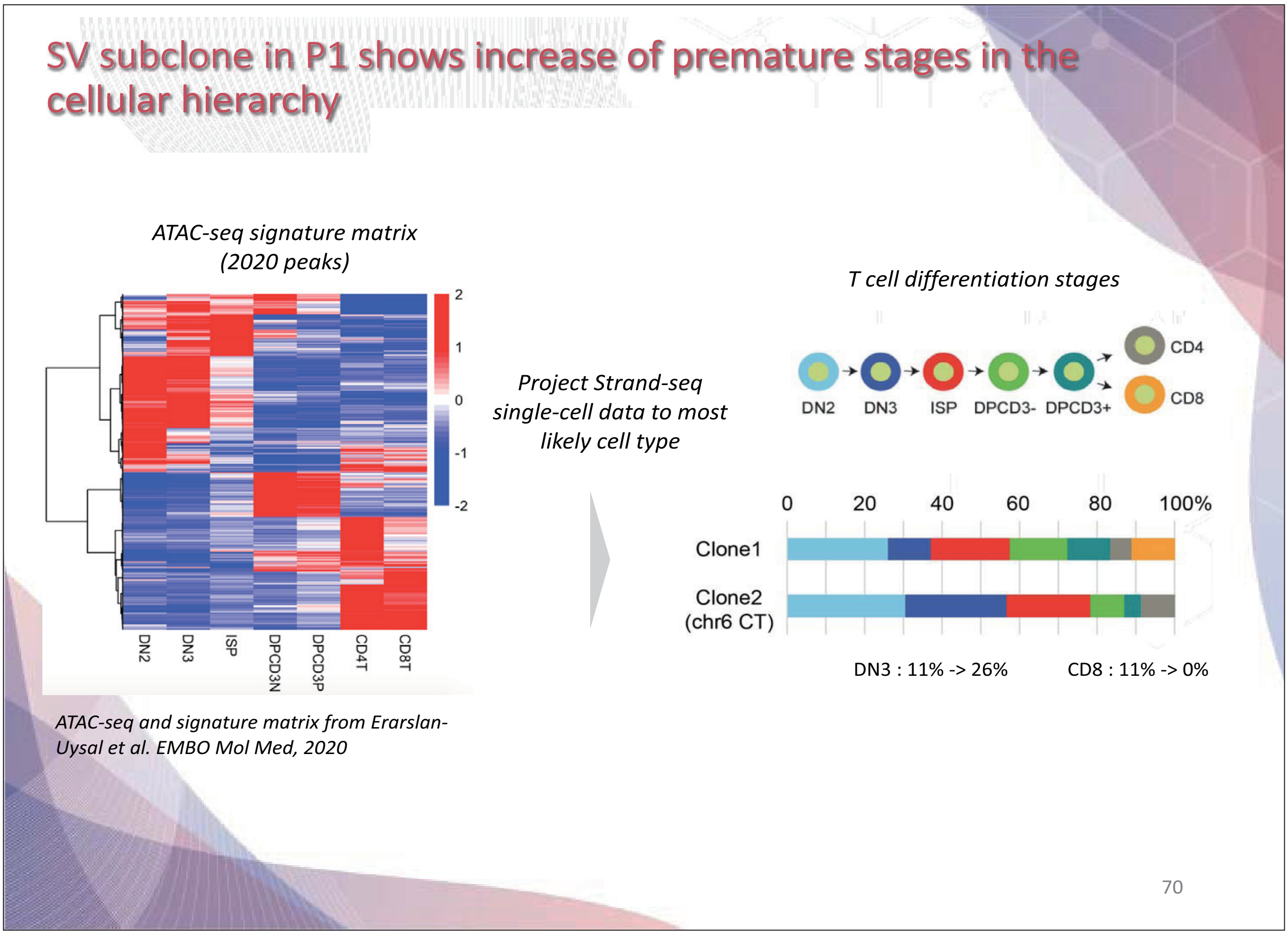
Andreas Kulozik group,
Beat Bornhauser,
Jean-Pierre Bourquin
Uni Zurich

SVs

Major clone (68%)

Subclone (32%) (only appeared in relapse)

Sanders et al. 2020



SV subclone in P1 shows increase of premature stages in the cellular hierarchy

ATAC-seq signature matrix (2020 peaks)

ATAC-seq and signature matrix from Erarslan-Uysal et al. EMBO Mol Med, 2020

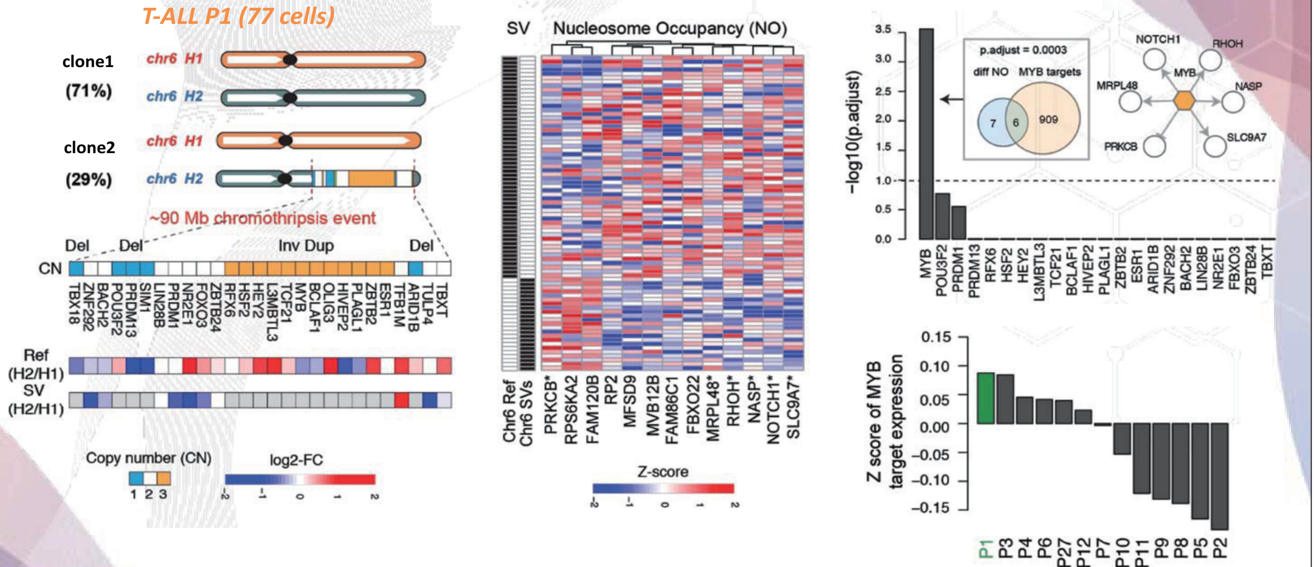
Project Strand-seq single-cell data to most likely cell type

T cell differentiation stages

DN3 : 11% -> 26%

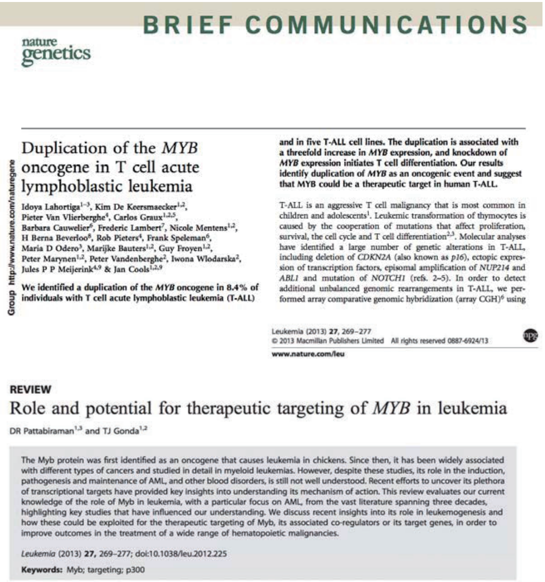
CD8 : 11% -> 0%

SV subclone in T-ALL P1 shows altered MYB target genes including NOTCH1



How the cell type (state) composition different in clone1 and clone2?

Notch signaling and MYB has been reported in T-ALL oncogenesis



BRIEF COMMUNICATIONS

nature genetics

Duplication of the *MYB* oncogene in T cell acute lymphoblastic leukemia

Idoya Lahortiga[1–3], Kim De Keersmaecker[1,2], Pieter Van Vlierberghe[4], Carlos Graux[1,2,5], Barbara Cauwelier[6], Frederic Lambert[7], Nicole Mentens[1,2], H Berna Beverloo[8], Rob Pieters[4], Frank Speleman[6], Maria D Odero[3], Marijke Bauters[1,2], Guy Froyen[1,2], Peter Marynen[1,2], Peter Vandenberghe[2], Iwona Wlodarska[2], Jules P P Meijerink[4,9] & Jan Cools[1,2,9]

Group http://www.nature.com/naturegene

We identified a duplication of the *MYB* oncogene in 8.4% of individuals with T cell acute lymphoblastic leukemia (T-ALL) and in five T-ALL cell lines. The duplication is associated with a threefold increase in *MYB* expression, and knockdown of *MYB* expression initiates T cell differentiation. Our results identify duplication of *MYB* as an oncogenic event and suggest that MYB could be a therapeutic target in human T-ALL.

T-ALL is an aggressive T cell malignancy that is most common in children and adolescents[1]. Leukemic transformation of thymocytes is caused by the cooperation of mutations that affect proliferation, survival, the cell cycle and T cell differentiation[2,3]. Molecular analyses have identified a large number of genetic alterations in T-ALL, including deletion of *CDKN2A* (also known as *p16*), ectopic expression of transcription factors, episomal amplification of *NUP214* and *ABL1* and mutation of *NOTCH1* (refs. 2–5). In order to detect additional unbalanced genomic rearrangements in T-ALL, we performed array comparative genomic hybridization (array CGH)[6] using

Leukemia (2013) **27**, 269–277
© 2013 Macmillan Publishers Limited All rights reserved 0887-6924/13
www.nature.com/leu

REVIEW

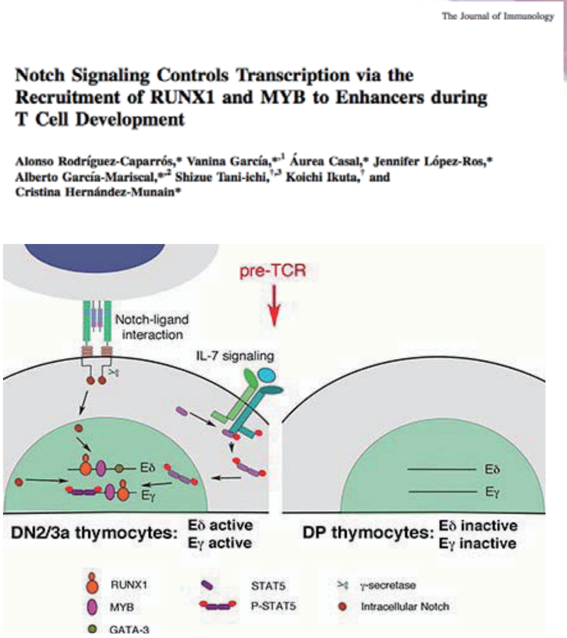
Role and potential for therapeutic targeting of *MYB* in leukemia

DR Pattabiraman[1,3] and TJ Gonda[1,2]

The Myb protein was first identified as an oncogene that causes leukemia in chickens. Since then, it has been widely associated with different types of cancers and studied in detail in myeloid leukemias. However, despite these studies, its role in the induction, pathogenesis and maintenance of AML, and other blood disorders, is still not well understood. Recent efforts to uncover its plethora of transcriptional targets have provided key insights into understanding its mechanism of action. This review evaluates our current knowledge of the role of Myb in leukemia, with a particular focus on AML, from the vast literature spanning three decades, highlighting key studies that have influenced our understanding. We discuss recent insights into its role in leukemogenesis and how these could be exploited for the therapeutic targeting of Myb, its associated co-regulators or its target genes, in order to improve outcomes in the treatment of a wide range of hematopoietic malignancies.

Leukemia (2013) **27**, 269–277; doi:10.1038/leu.2012.225

Keywords: Myb; targeting; p300

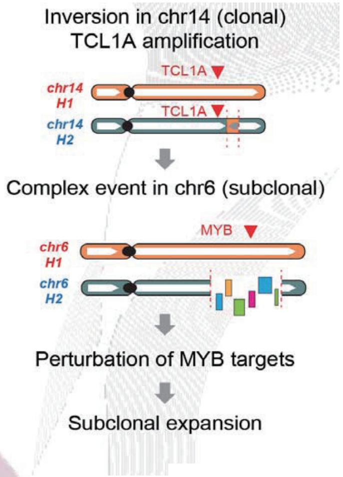


The Journal of Immunology

Notch Signaling Controls Transcription via the Recruitment of RUNX1 and MYB to Enhancers during T Cell Development

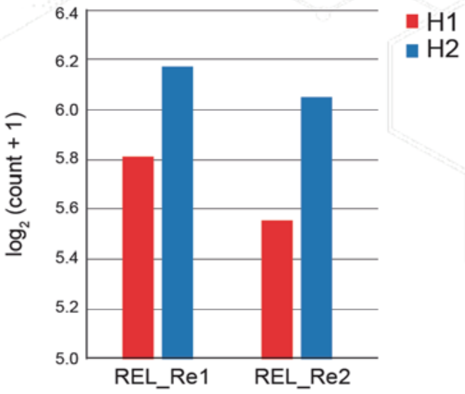
Alonso Rodríguez-Caparrós,* Vanina García,*[1] Áurea Casal,* Jennifer López-Ros,* Alberto García-Mariscal,*[2] Shizue Tani-ichi,†[3] Koichi Ikuta,† and Cristina Hernández-Munain*

Validation of increased dosage of MYB expression in rearranged haplotype



Inversion in chr14 (clonal)
TCL1A amplification
TCL1A
chr14 H1
TCL1A
chr14 H2
Complex event in chr6 (subclonal)
MYB
chr6 H1
chr6 H2
Perturbation of MYB targets
Subclonal expansion

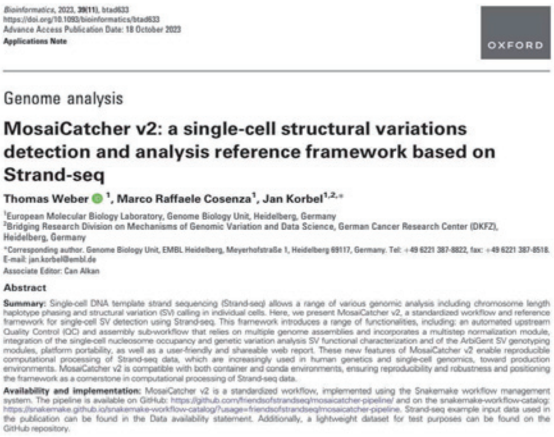
Bulk RNA-seq



MYB H2/H1 relapse
log2-FC = 0.45
(1.37 fold increase)
p-value = 0.0317

Single-cell experiment is needed to confirm subclonal level transcriptome changes

Practical session – how to run scNOVA



Bioinformatics, 2023, **39**(11), btad633
https://doi.org/10.1093/bioinformatics/btad633
Advance Access Publication Date: 18 October 2023
Applications Note

OXFORD

Genome analysis

MosaiCatcher v2: a single-cell structural variations detection and analysis reference framework based on Strand-seq

Thomas Weber [1], Marco Raffaele Cosenza[1], Jan Korbel[1,2,*]

[1]European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany
[2]Bridging Research Division on Mechanisms of Genomic Variation and Data Science, German Cancer Research Center (DKFZ), Heidelberg, Germany

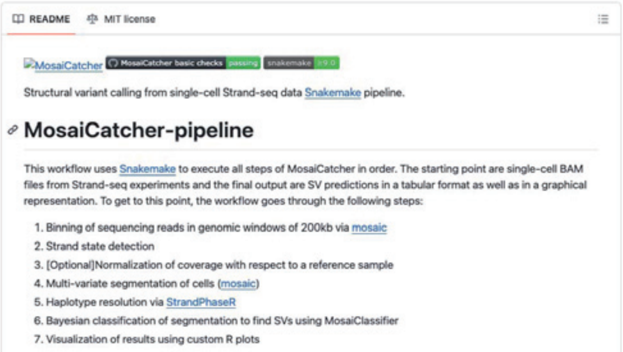
*Corresponding author. Genome Biology Unit, EMBL Heidelberg, Meyerhofstraße 1, Heidelberg 69117, Germany. Tel: +49 6221 387-8822, fax: +49 6221 387-8518. E-mail: jan.korbel@embl.de

Associate Editor: Can Alkan

Abstract

Summary: Single-cell DNA template strand sequencing (Strand-seq) allows a range of various genomic analysis including chromosome length haplotype phasing and structural variation (SV) calling in individual cells. Here, we present MosaiCatcher v2, a standardized workflow and reference framework for single-cell SV detection using Strand-seq. This framework introduces a range of functionalities, including: an automated upstream Quality Control (QC) and assembly sub-workflow that relies on multiple genome assemblies and incorporates a multistep normalization module, integration of the single-cell nucleosome occupancy and genetic variation analysis SV functional characterization and of the ArbiGent SV genotyping modules, platform portability, as well as a user-friendly and shareable web report. These new features of MosaiCatcher v2 enable reproducible computational processing of Strand-seq data, which are increasingly used in human genetics and single-cell genomics, toward production environments. MosaiCatcher v2 is compatible with both container and conda environments, ensuring reproducibility and robustness and positioning the framework as a cornerstone in computational processing of Strand-seq data.

Availability and implementation: MosaiCatcher v2 is a standardized workflow, implemented using the Snakemake workflow management system. The pipeline is available on GitHub: https://github.com/friendsofstrandseq/mosaicatcher-pipeline/ and on the snakemake-workflow-catalog: https://snakemake.github.io/snakemake-workflow-catalog/?usage=friendsofstrandseq/mosaicatcher-pipeline. Strand-seq example input data used in the publication can be found in the Data availability statement. Additionally, a lightweight dataset for test purposes can be found on the GitHub repository.



Structural variant calling from single-cell Strand-seq data Snakemake pipeline.

MosaiCatcher-pipeline

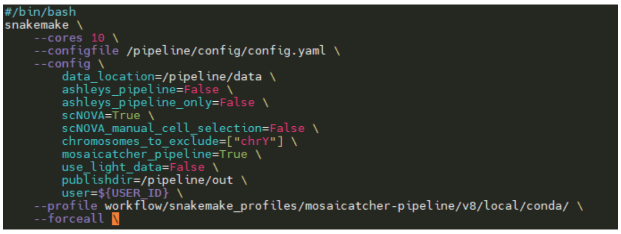
This workflow uses Snakemake to execute all steps of MosaiCatcher in order. The starting point are single-cell BAM files from Strand-seq experiments and the final output are SV predictions in a tabular format as well as in a graphical representation. To get to this point, the workflow goes through the following steps:

1. Binning of sequencing reads in genomic windows of 200kb via mosaic
2. Strand state detection
3. [Optional]Normalization of coverage with respect to a reference sample
4. Multi-variate segmentation of cells (mosaic)
5. Haplotype resolution via StrandPhaseR
6. Bayesian classification of segmentation to find SVs using MosaiClassifier
7. Visualization of results using custom R plots

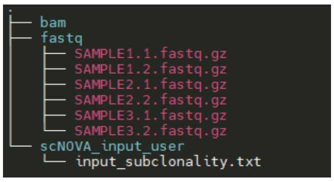
https://github.com/friendsofstrandseq/mosaicatcher-pipeline/

Based on the tutorial written by Chiwon Chung

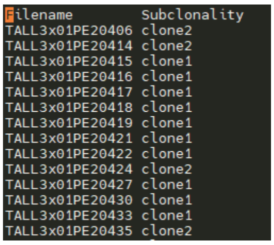
Practical session – running the pipeline



```
#/bin/bash
snakemake \
    --cores 10 \
    --configfile /pipeline/config/config.yaml \
    --config \
        data_location=/pipeline/data \
        ashleys_pipeline=False \
        ashleys_pipeline_only=False \
        scNOVA=True \
        scNOVA_manual_cell_selection=False \
        chromosomes_to_exclude=["chrY"] \
        mosaicatcher_pipeline=True \
        use_light_data=False \
        publishdir=/pipeline/out \
        user=${USER_ID} \
    --profile workflow/snakemake_profiles/mosaicatcher-pipeline/v8/local/conda/ \
    --forceall \
```



```
.
├── bam
├── fastq
│   ├── SAMPLE1.1.fastq.gz
│   ├── SAMPLE1.2.fastq.gz
│   ├── SAMPLE2.1.fastq.gz
│   ├── SAMPLE2.2.fastq.gz
│   ├── SAMPLE3.1.fastq.gz
│   └── SAMPLE3.2.fastq.gz
└── scNOVA_input_user
    └── input_subclonality.txt
```

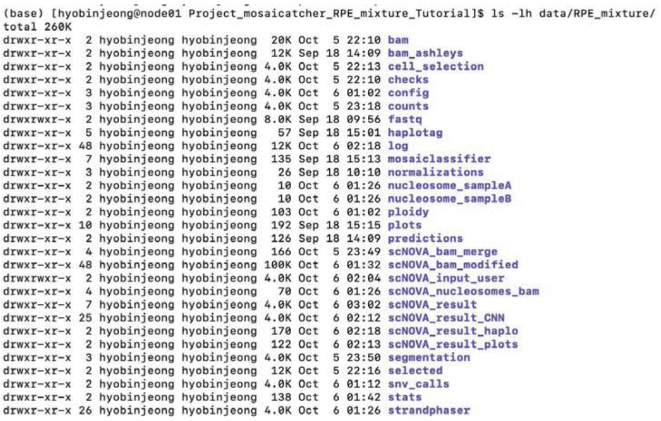


```
Filename           Subclonality
TALL3x01PE20406 clone2
TALL3x01PE20414 clone2
TALL3x01PE20415 clone1
TALL3x01PE20416 clone1
TALL3x01PE20417 clone1
TALL3x01PE20418 clone1
TALL3x01PE20419 clone1
TALL3x01PE20421 clone1
TALL3x01PE20422 clone1
TALL3x01PE20424 clone2
TALL3x01PE20427 clone1
TALL3x01PE20430 clone1
TALL3x01PE20433 clone1
TALL3x01PE20435 clone2
```

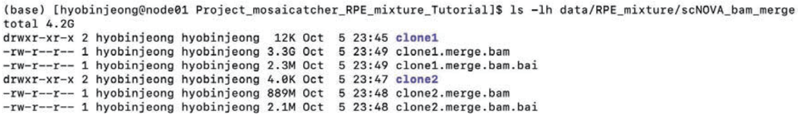
Once the pipeline is finished, create a new directory scNOVA_input_user and add your subclonality file. It MUST be named input_subclonality.txt, and it MUST be a tsv file with the correct header names. Otherwise, the pipeline will throw errors.

To run the scNOVA pipeline, please change the run_mosaicatcherv8.sh as shown on the left. Note that scNOVA currently does not support chromosome Y, and must be place in the chromosomes_to_exclude list.

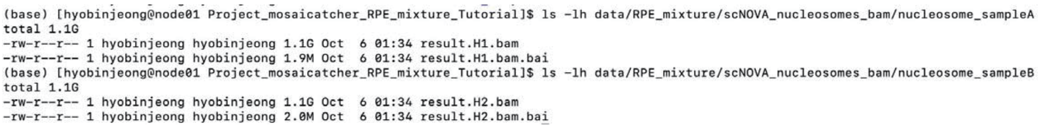
Output – folder structure



```
(base) [hyobinjeong@node01 Project_mosaicatcher_RPE_mixture_Tutorial]$ ls -lh data/RPE_mixture/
total 260K
drwxr-xr-x  2 hyobinjeong hyobinjeong  20K Oct  5 22:10 bam
drwxr-xr-x  2 hyobinjeong hyobinjeong  12K Sep 18 14:09 bam_ashleys
drwxr-xr-x  2 hyobinjeong hyobinjeong 4.0K Oct  5 22:13 cell_selection
drwxr-xr-x  2 hyobinjeong hyobinjeong 4.0K Oct  5 22:10 checks
drwxr-xr-x  3 hyobinjeong hyobinjeong 4.0K Oct  6 01:02 config
drwxr-xr-x  3 hyobinjeong hyobinjeong 4.0K Oct  5 23:18 counts
drwxrwxr-x  2 hyobinjeong hyobinjeong 8.0K Sep 18 09:56 fastq
drwxr-xr-x  5 hyobinjeong hyobinjeong   57 Sep 18 15:01 haplotag
drwxr-xr-x 48 hyobinjeong hyobinjeong  12K Oct  6 02:18 log
drwxr-xr-x  7 hyobinjeong hyobinjeong  135 Sep 18 15:13 mosaiclassifier
drwxr-xr-x  3 hyobinjeong hyobinjeong   26 Sep 18 10:10 normalizations
drwxr-xr-x  2 hyobinjeong hyobinjeong   10 Oct  6 01:26 nucleosome_sampleA
drwxr-xr-x  2 hyobinjeong hyobinjeong   10 Oct  6 01:26 nucleosome_sampleB
drwxr-xr-x  2 hyobinjeong hyobinjeong  103 Oct  6 01:02 ploidy
drwxr-xr-x 10 hyobinjeong hyobinjeong  192 Sep 18 15:15 plots
drwxr-xr-x  2 hyobinjeong hyobinjeong  126 Sep 18 14:09 predictions
drwxr-xr-x  4 hyobinjeong hyobinjeong  166 Oct  5 23:49 scNOVA_bam_merge
drwxr-xr-x 48 hyobinjeong hyobinjeong 100K Oct  6 01:32 scNOVA_bam_modified
drwxrwxr-x  2 hyobinjeong hyobinjeong 4.0K Oct  6 02:04 scNOVA_input_user
drwxr-xr-x  4 hyobinjeong hyobinjeong   70 Oct  6 01:26 scNOVA_nucleosomes_bam
drwxr-xr-x  7 hyobinjeong hyobinjeong 4.0K Oct  6 03:02 scNOVA_result
drwxr-xr-x 25 hyobinjeong hyobinjeong 4.0K Oct  6 02:12 scNOVA_result_CNN
drwxr-xr-x  2 hyobinjeong hyobinjeong  170 Oct  6 02:18 scNOVA_result_haplo
drwxr-xr-x  2 hyobinjeong hyobinjeong  122 Oct  6 02:13 scNOVA_result_plots
drwxr-xr-x  3 hyobinjeong hyobinjeong 4.0K Oct  5 23:50 segmentation
drwxr-xr-x  2 hyobinjeong hyobinjeong  12K Oct  5 22:16 selected
drwxr-xr-x  2 hyobinjeong hyobinjeong 4.0K Oct  6 01:12 snv_calls
drwxr-xr-x  2 hyobinjeong hyobinjeong  138 Oct  6 01:42 stats
drwxr-xr-x 26 hyobinjeong hyobinjeong 4.0K Oct  6 01:26 strandphaser
```



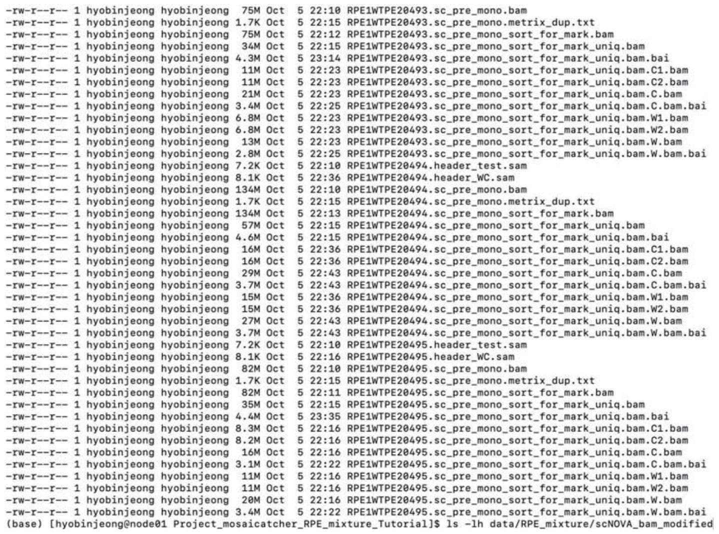
```
(base) [hyobinjeong@node01 Project_mosaicatcher_RPE_mixture_Tutorial]$ ls -lh data/RPE_mixture/scNOVA_bam_merge
total 4.2G
drwxr-xr-x 2 hyobinjeong hyobinjeong  12K Oct  5 23:45 clone1
-rw-r--r-- 1 hyobinjeong hyobinjeong 3.3G Oct  5 23:49 clone1.merge.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong 2.3M Oct  5 23:49 clone1.merge.bam.bai
drwxr-xr-x 2 hyobinjeong hyobinjeong 4.0K Oct  5 23:47 clone2
-rw-r--r-- 1 hyobinjeong hyobinjeong 889M Oct  5 23:48 clone2.merge.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong 2.1M Oct  5 23:48 clone2.merge.bam.bai
```



```
(base) [hyobinjeong@node01 Project_mosaicatcher_RPE_mixture_Tutorial]$ ls -lh data/RPE_mixture/scNOVA_nucleosomes_bam/nucleosome_sampleA
total 1.1G
-rw-r--r-- 1 hyobinjeong hyobinjeong 1.1G Oct  6 01:34 result.H1.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong 1.9M Oct  6 01:34 result.H1.bam.bai
(base) [hyobinjeong@node01 Project_mosaicatcher_RPE_mixture_Tutorial]$ ls -lh data/RPE_mixture/scNOVA_nucleosomes_bam/nucleosome_sampleB
total 1.1G
-rw-r--r-- 1 hyobinjeong hyobinjeong 1.1G Oct  6 01:34 result.H2.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong 2.0M Oct  6 01:34 result.H2.bam.bai
```

Output – Processed single-cell bam files

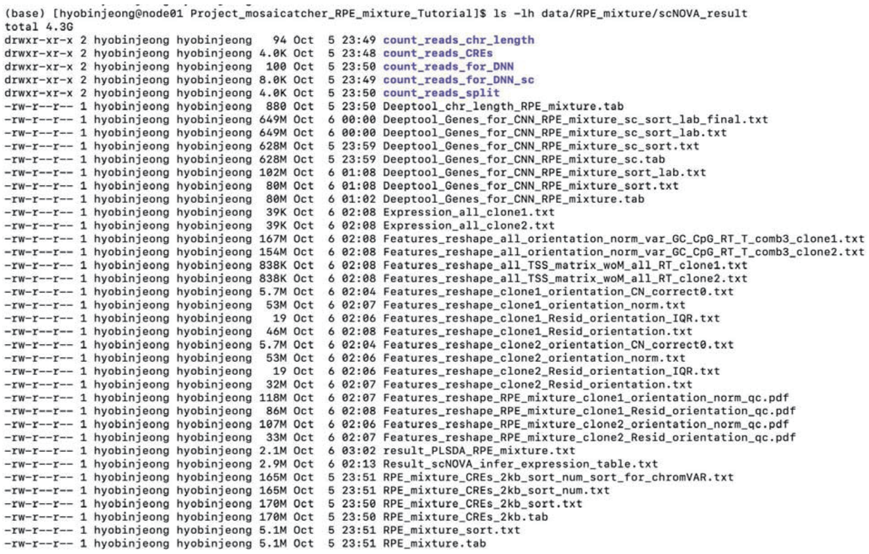
data/RPE_mixture/scNOVA_bam_modified



```
-rw-r--r-- 1 hyobinjeong hyobinjeong  75M Oct  5 22:10 RPE1WTPE20493.sc_pre_mono.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong 1.7K Oct  5 22:15 RPE1WTPE20493.sc_pre_mono.metrix_dup.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong  75M Oct  5 22:12 RPE1WTPE20493.sc_pre_mono_sort_for_mark.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong  34M Oct  5 22:15 RPE1WTPE20493.sc_pre_mono_sort_for_mark_uniq.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong 4.3M Oct  5 23:14 RPE1WTPE20493.sc_pre_mono_sort_for_mark_uniq.bam.bai
-rw-r--r-- 1 hyobinjeong hyobinjeong  11M Oct  5 22:23 RPE1WTPE20493.sc_pre_mono_sort_for_mark_uniq.bam.C1.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong  11M Oct  5 22:23 RPE1WTPE20493.sc_pre_mono_sort_for_mark_uniq.bam.C2.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong  21M Oct  5 22:23 RPE1WTPE20493.sc_pre_mono_sort_for_mark_uniq.bam.C.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong 3.4M Oct  5 22:25 RPE1WTPE20493.sc_pre_mono_sort_for_mark_uniq.bam.C.bam.bai
-rw-r--r-- 1 hyobinjeong hyobinjeong 6.8M Oct  5 22:23 RPE1WTPE20493.sc_pre_mono_sort_for_mark_uniq.bam.W1.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong 6.8M Oct  5 22:23 RPE1WTPE20493.sc_pre_mono_sort_for_mark_uniq.bam.W2.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong  13M Oct  5 22:23 RPE1WTPE20493.sc_pre_mono_sort_for_mark_uniq.bam.W.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong 2.8M Oct  5 22:25 RPE1WTPE20493.sc_pre_mono_sort_for_mark_uniq.bam.W.bam.bai
-rw-r--r-- 1 hyobinjeong hyobinjeong 7.2K Oct  5 22:10 RPE1WTPE20494.header_test.sam
-rw-r--r-- 1 hyobinjeong hyobinjeong 8.1K Oct  5 22:36 RPE1WTPE20494.header_WC.sam
-rw-r--r-- 1 hyobinjeong hyobinjeong 134M Oct  5 22:10 RPE1WTPE20494.sc_pre_mono.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong 1.7K Oct  5 22:15 RPE1WTPE20494.sc_pre_mono.metrix_dup.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong 134M Oct  5 22:13 RPE1WTPE20494.sc_pre_mono_sort_for_mark.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong  57M Oct  5 22:15 RPE1WTPE20494.sc_pre_mono_sort_for_mark_uniq.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong 4.6M Oct  5 22:15 RPE1WTPE20494.sc_pre_mono_sort_for_mark_uniq.bam.bai
-rw-r--r-- 1 hyobinjeong hyobinjeong  16M Oct  5 22:36 RPE1WTPE20494.sc_pre_mono_sort_for_mark_uniq.bam.C1.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong  16M Oct  5 22:36 RPE1WTPE20494.sc_pre_mono_sort_for_mark_uniq.bam.C2.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong  29M Oct  5 22:43 RPE1WTPE20494.sc_pre_mono_sort_for_mark_uniq.bam.C.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong 3.7M Oct  5 22:43 RPE1WTPE20494.sc_pre_mono_sort_for_mark_uniq.bam.C.bam.bai
-rw-r--r-- 1 hyobinjeong hyobinjeong  15M Oct  5 22:36 RPE1WTPE20494.sc_pre_mono_sort_for_mark_uniq.bam.W1.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong  15M Oct  5 22:36 RPE1WTPE20494.sc_pre_mono_sort_for_mark_uniq.bam.W2.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong  27M Oct  5 22:43 RPE1WTPE20494.sc_pre_mono_sort_for_mark_uniq.bam.W.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong 3.7M Oct  5 22:43 RPE1WTPE20494.sc_pre_mono_sort_for_mark_uniq.bam.W.bam.bai
-rw-r--r-- 1 hyobinjeong hyobinjeong 7.2K Oct  5 22:10 RPE1WTPE20495.header_test.sam
-rw-r--r-- 1 hyobinjeong hyobinjeong 8.1K Oct  5 22:16 RPE1WTPE20495.header_WC.sam
-rw-r--r-- 1 hyobinjeong hyobinjeong  82M Oct  5 22:10 RPE1WTPE20495.sc_pre_mono.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong 1.7K Oct  5 22:15 RPE1WTPE20495.sc_pre_mono.metrix_dup.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong  82M Oct  5 22:11 RPE1WTPE20495.sc_pre_mono_sort_for_mark.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong  35M Oct  5 22:15 RPE1WTPE20495.sc_pre_mono_sort_for_mark_uniq.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong 4.4M Oct  5 23:35 RPE1WTPE20495.sc_pre_mono_sort_for_mark_uniq.bam.bai
-rw-r--r-- 1 hyobinjeong hyobinjeong 8.3M Oct  5 22:16 RPE1WTPE20495.sc_pre_mono_sort_for_mark_uniq.bam.C1.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong 8.2M Oct  5 22:16 RPE1WTPE20495.sc_pre_mono_sort_for_mark_uniq.bam.C2.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong  16M Oct  5 22:16 RPE1WTPE20495.sc_pre_mono_sort_for_mark_uniq.bam.C.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong 3.1M Oct  5 22:22 RPE1WTPE20495.sc_pre_mono_sort_for_mark_uniq.bam.C.bam.bai
-rw-r--r-- 1 hyobinjeong hyobinjeong  11M Oct  5 22:16 RPE1WTPE20495.sc_pre_mono_sort_for_mark_uniq.bam.W1.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong  11M Oct  5 22:16 RPE1WTPE20495.sc_pre_mono_sort_for_mark_uniq.bam.W2.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong  20M Oct  5 22:16 RPE1WTPE20495.sc_pre_mono_sort_for_mark_uniq.bam.W.bam
-rw-r--r-- 1 hyobinjeong hyobinjeong 3.4M Oct  5 22:22 RPE1WTPE20495.sc_pre_mono_sort_for_mark_uniq.bam.W.bam.bai
(base) [hyobinjeong@node01 Project_mosaicatcher_RPE_mixture_Tutorial]$ ls -lh data/RPE_mixture/scNOVA_bam_modified
```

Output – Result from scNOVA pipeline

data/RPE_mixture/scNOVA_result

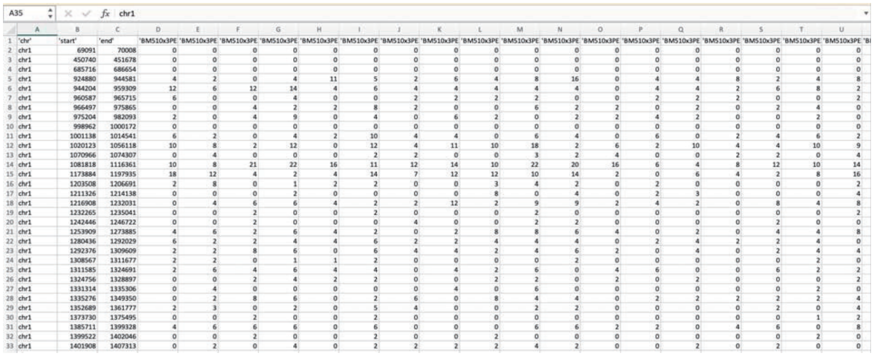


```
(base) [hyobinjeong@node01 Project_mosaicatcher_RPE_mixture_Tutorial]$ ls -lh data/RPE_mixture/scNOVA_result
total 4.3G
drwxr-xr-x 2 hyobinjeong hyobinjeong   94 Oct  5 23:49 count_reads_chr_length
drwxr-xr-x 2 hyobinjeong hyobinjeong 4.0K Oct  5 23:48 count_reads_CREs
drwxr-xr-x 2 hyobinjeong hyobinjeong  100 Oct  5 23:50 count_reads_for_DNN
drwxr-xr-x 2 hyobinjeong hyobinjeong 8.0K Oct  5 23:49 count_reads_for_DNN_sc
drwxr-xr-x 2 hyobinjeong hyobinjeong 4.0K Oct  5 23:50 count_reads_split
-rw-r--r-- 1 hyobinjeong hyobinjeong  880 Oct  5 23:50 Deeptool_chr_length_RPE_mixture.tab
-rw-r--r-- 1 hyobinjeong hyobinjeong 649M Oct  6 00:00 Deeptool_Genes_for_CNN_RPE_mixture_sc_sort_lab_final.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong 649M Oct  6 00:00 Deeptool_Genes_for_CNN_RPE_mixture_sc_sort_lab.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong 628M Oct  5 23:59 Deeptool_Genes_for_CNN_RPE_mixture_sc_sort.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong 628M Oct  5 23:59 Deeptool_Genes_for_CNN_RPE_mixture_sc.tab
-rw-r--r-- 1 hyobinjeong hyobinjeong 102M Oct  6 01:08 Deeptool_Genes_for_CNN_RPE_mixture_sort_lab.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong  80M Oct  6 01:08 Deeptool_Genes_for_CNN_RPE_mixture_sort.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong  80M Oct  6 01:02 Deeptool_Genes_for_CNN_RPE_mixture.tab
-rw-r--r-- 1 hyobinjeong hyobinjeong  39K Oct  6 02:08 Expression_all_clone1.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong  39K Oct  6 02:08 Expression_all_clone2.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong 167M Oct  6 02:08 Features_reshape_all_orientation_norm_var_GC_CpG_RT_T_comb3_clone1.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong 154M Oct  6 02:08 Features_reshape_all_orientation_norm_var_GC_CpG_RT_T_comb3_clone2.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong 838K Oct  6 02:08 Features_reshape_all_TSS_matrix_woM_all_RT_clone1.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong 838K Oct  6 02:08 Features_reshape_all_TSS_matrix_woM_all_RT_clone2.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong 5.7M Oct  6 02:04 Features_reshape_clone1_orientation_CN_correct0.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong  53M Oct  6 02:07 Features_reshape_clone1_orientation_norm.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong   19 Oct  6 02:06 Features_reshape_clone1_Resid_orientation_IQR.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong  46M Oct  6 02:08 Features_reshape_clone1_Resid_orientation.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong 5.7M Oct  6 02:04 Features_reshape_clone2_orientation_CN_correct0.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong  53M Oct  6 02:06 Features_reshape_clone2_orientation_norm.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong   19 Oct  6 02:06 Features_reshape_clone2_Resid_orientation_IQR.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong  32M Oct  6 02:07 Features_reshape_clone2_Resid_orientation.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong 118M Oct  6 02:07 Features_reshape_RPE_mixture_clone1_orientation_norm_qc.pdf
-rw-r--r-- 1 hyobinjeong hyobinjeong  86M Oct  6 02:08 Features_reshape_RPE_mixture_clone1_Resid_orientation_qc.pdf
-rw-r--r-- 1 hyobinjeong hyobinjeong 107M Oct  6 02:06 Features_reshape_RPE_mixture_clone2_orientation_norm_qc.pdf
-rw-r--r-- 1 hyobinjeong hyobinjeong  33M Oct  6 02:07 Features_reshape_RPE_mixture_clone2_Resid_orientation_qc.pdf
-rw-r--r-- 1 hyobinjeong hyobinjeong 2.1M Oct  6 03:02 result_PLSDA_RPE_mixture.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong 2.9M Oct  6 02:13 Result_scNOVA_infer_expression_table.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong 165M Oct  5 23:51 RPE_mixture_CREs_2kb_sort_num_sort_for_chromVAR.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong 165M Oct  5 23:51 RPE_mixture_CREs_2kb_sort_num.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong 170M Oct  5 23:50 RPE_mixture_CREs_2kb_sort.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong 170M Oct  5 23:50 RPE_mixture_CREs_2kb.tab
-rw-r--r-- 1 hyobinjeong hyobinjeong 5.1M Oct  5 23:51 RPE_mixture_sort.txt
-rw-r--r-- 1 hyobinjeong hyobinjeong 5.1M Oct  5 23:51 RPE_mixture.tab
```

Output – Single-cell Nucleosome count matrix

Single-cell Nucleosome count matrix (gene X cell)

data/RPE_mixture/scNOVA_result/RPE_mixture_sort.txt



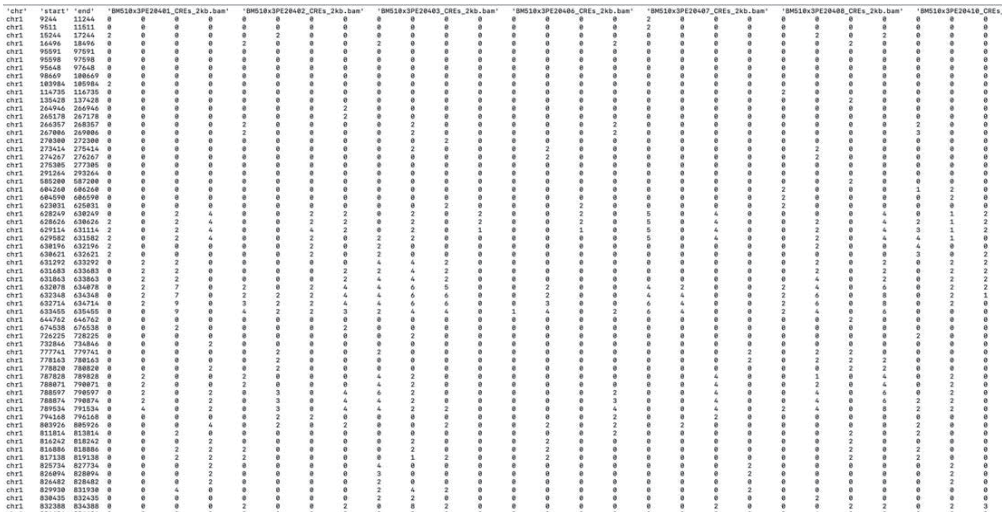
A35 fx chr1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	'chr'	'start'	'end'	'BM510x3PE	'BM510x3PE	'BM510x3PE	'BM510x3PE	'BM510x3PE	'BM510x3PE	'BM510x3PE	'BM510x3PE	'BM510x3PE	'BM510x3PE	'BM510x3PE	'BM510x3PE	'BM510x3PE	'BM510x3PE	'BM510x3PE	'BM510x3PE	'BM510x3PE	'BM510x3PE
2	chr1	69091	70008	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	chr1	450740	451678	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	chr1	685716	686654	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	chr1	924880	944581	4	2	0	4	11	5	2	6	4	8	16	0	4	4	8	2	4	8
6	chr1	944204	959309	12	6	12	14	4	6	4	4	4	4	4	0	4	4	2	6	8	2
7	chr1	960587	965715	6	0	0	4	0	0	2	2	2	2	0	0	2	2	2	0	0	2
8	chr1	966497	975865	0	0	4	2	2	8	2	0	0	6	2	2	0	2	0	2	4	0
9	chr1	975204	982093	2	0	4	9	0	4	0	6	2	0	2	2	4	2	0	0	2	0
10	chr1	998962	1000172	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	chr1	1001138	1014541	6	2	0	4	2	10	4	4	0	6	4	0	6	0	2	4	6	2
12	chr1	1020123	1056118	10	8	2	12	0	12	4	11	10	18	2	6	2	10	4	4	10	9
13	chr1	1070966	1074307	0	4	0	0	0	2	2	0	0	3	2	4	0	0	2	2	0	4
14	chr1	1081818	1116361	10	8	21	22	16	11	12	14	10	22	20	16	6	4	8	12	10	14
15	chr1	1173884	1197935	18	12	4	2	4	14	7	12	12	10	14	2	0	6	4	2	8	16
16	chr1	1203508	1206691	2	8	0	1	2	2	0	0	3	4	2	0	2	0	0	0	0	2
17	chr1	1211326	1214138	0	0	0	2	0	0	0	0	8	0	4	0	2	3	0	0	0	4
18	chr1	1216908	1232031	0	4	6	6	4	2	2	12	2	9	9	2	4	2	0	8	4	8
19	chr1	1232265	1235041	0	0	2	0	0	2	0	0	0	2	0	0	0	0	0	0	0	2
20	chr1	1242446	1246722	0	0	2	0	0	0	4	0	0	2	2	0	0	0	0	2	0	0
21	chr1	1253909	1273885	4	6	2	6	4	2	0	2	8	8	6	4	0	2	0	4	4	8
22	chr1	1280436	1292029	6	2	2	4	4	6	2	2	4	4	4	0	2	4	2	2	4	0
23	chr1	1292376	1309609	2	2	8	6	0	6	4	4	2	4	6	2	0	4	0	2	4	4
24	chr1	1308567	1311677	2	2	0	1	1	2	0	0	0	0	2	0	0	0	0	0	2	0
25	chr1	1311585	1324691	2	6	4	6	4	4	0	4	2	6	0	4	6	0	0	6	2	2
26	chr1	1324756	1328897	0	0	2	4	2	2	0	0	2	2	0	2	0	2	0	0	0	2
27	chr1	1331314	1335306	0	4	0	0	0	0	0	4	0	6	0	0	0	0	0	0	2	0
28	chr1	1335276	1349350	0	2	8	6	0	2	6	0	8	4	4	0	2	2	2	2	2	4
29	chr1	1352689	1361777	2	3	0	2	0	5	4	0	0	2	2	0	0	0	0	2	0	4
30	chr1	1373730	1375495	0	0	2	0	0	2	0	0	0	0	0	0	0	0	0	0	1	2
31	chr1	1385711	1399328	4	6	6	6	0	6	0	0	0	6	6	2	2	0	4	6	0	8
32	chr1	1399522	1402046	0	0	2	0	0	2	0	0	2	0	0	0	0	0	0	0	2	0
33	chr1	1401908	1407313	0	2	0	4	0	2	2	2	2	4	2	0	0	2	0	2	0	0

Output – Single-cell Nucleosome count matrix

Single-cell Nucleosome count matrix (CRE X cell)

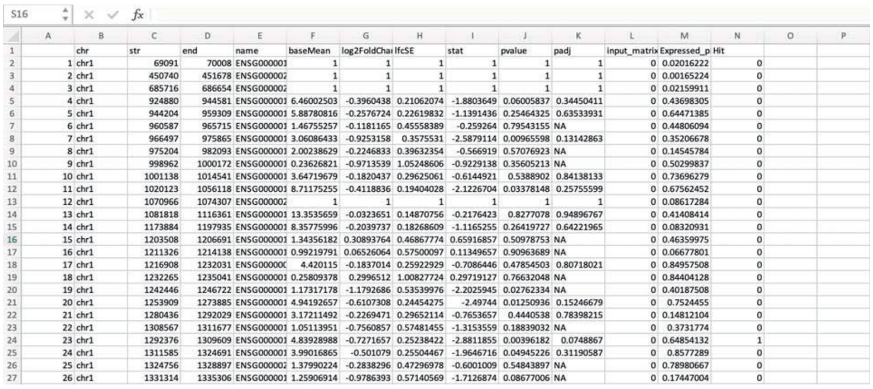
data/RPE_mixture/scNOVA_result/RPE_mixture_CREs_2kb_sort.txt



```
'chr'  'start' 'end'   'BM510x3PE20401_CREs_2kb.bam'   'BM510x3PE20402_CREs_2kb.bam'   'BM510x3PE20403_CREs_2kb.bam'   'BM510x3PE20406_CREs_2kb.bam'   'BM510x3PE20407_CREs_2kb.bam'   'BM510x3PE20408_CREs_2kb.bam'   'BM510x3PE20410_CREs_
chr1   9244    11244   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   2   0   0   0   0   0   0   0   0   0   0
chr1   9511    11511   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   2   0   0   0   0   0   0   0   0   0   0
chr1   15244   17244   2   0   0   0   0   2   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   2   0   2   0   0   0
chr1   16496   18496   0   0   0   0   2   0   0   0   2   0   0   0   0   0   0   0   0   2   0   0   0   0   0   0   2   0   0   0   0
chr1   95591   97591   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
chr1   95598   97598   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
chr1   95648   97648   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
chr1   98669   100669  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
chr1   103984  105984  2   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
chr1   114735  116735  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   2   0   0   0   0   0   0
chr1   135428  137428  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   2   0   0   0   0
chr1   264946  266946  0   0   0   0   0   0   0   2   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
chr1   265178  267178  0   0   0   0   0   0   0   2   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
chr1   266357  268357  0   0   0   0   2   0   0   0   0   2   0   0   0   0   0   0   0   2   0   0   0   0   0   0   0   0   2   0   0
chr1   267006  269006  0   0   0   0   2   0   0   0   0   2   0   0   0   0   0   0   0   2   0   0   0   0   0   0   0   0   3   0   0
chr1   270300  272300  0   0   0   0   0   0   0   0   0   0   2   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
chr1   273414  275414  0   0   0   0   0   0   0   0   0   2   0   0   0   0   0   2   0   0   0   0   0   0   0   2   0   0   0   0   0
chr1   274267  276267  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   2   0   0   0   0   0   0   0   2   0   0   0   0   0
chr1   275305  277305  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
chr1   291264  293264  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
chr1   585200  587200  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   2   0   0   0   0
chr1   604260  606260  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   2   0   0   0   1   2   0
chr1   604590  606590  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   2   0   0   0   0   2   0
chr1   623031  625031  0   0   0   0   0   0   0   0   0   0   2   0   0   0   0   0   2   0   0   0   0   0   2   0   0   0   0   0   0
chr1   628249  630249  0   0   2   4   0   0   2   2   0   2   0   2   0   2   0   0   2   0   5   0   4   0   0   0   0   4   0   1   2
chr1   628626  630626  2   0   2   4   0   0   2   2   0   2   0   2   0   2   0   0   2   0   5   0   4   0   0   2   0   4   2   1   2
chr1   629114  631114  2   0   2   4   0   0   4   2   0   2   0   1   0   1   0   0   1   0   5   0   4   0   0   2   0   4   3   1   2
chr1   629582  631582  2   0   2   4   0   0   2   0   2   2   0   0   0   0   0   0   0   0   5   0   4   0   0   2   0   4   4   1   0
chr1   630196  632196  2   0   0   0   0   0   2   0   2   0   0   0   0   0   0   0   0   0   0   0   0   0   0   2   0   0   4   0   0
chr1   630621  632621  2   0   0   0   0   0   2   0   2   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   3   0   2
chr1   631292  633292  0   2   2   0   0   0   0   0   4   4   2   0   0   0   0   0   0   0   0   0   0   0   0   2   0   2   0   2   2
chr1   631683  633683  0   2   2   0   0   0   0   2   2   4   2   0   0   0   0   0   0   0   0   0   0   0   0   2   0   2   0   2   2
chr1   631863  633863  0   2   2   0   0   0   0   2   4   4   2   0   0   0   0   0   0   0   0   0   0   0   0   4   0   2   0   2   2
chr1   632078  634078  0   2   7   0   2   0   2   4   4   6   5   0   0   0   0   2   0   0   4   2   0   0   2   4   0   6   0   2   2
chr1   632348  634348  0   2   7   0   2   2   2   4   4   6   6   0   0   0   0   2   0   0   4   4   0   0   2   6   0   8   0   2   1
chr1   632714  634714  0   2   9   0   3   2   2   4   4   6   6   0   0   0   0   3   0   0   6   4   0   0   2   6   0   8   0   2   0
chr1   633455  635455  0   0   9   0   4   2   2   3   2   4   4   0   1   0   1   4   0   2   6   4   0   0   2   4   0   6   0   0   0
chr1   644762  646762  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   2   0   0   0   0
chr1   674538  676538  0   0   2   0   0   0   0   2   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
chr1   726225  728225  0   0   0   0   0   0   0   0   0   2   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   2   0   0
chr1   732846  734846  0   0   0   2   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
chr1   777741  779741  0   0   0   0   0   2   0   0   2   0   0   0   0   0   0   0   0   0   0   0   0   2   0   2   2   0   0   0   0
chr1   778163  780163  0   0   0   0   0   2   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   2   0   2   2   2   0   0   0
chr1   778820  780820  0   0   0   2   0   2   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   2   2   0   0   0
chr1   787828  789828  0   2   0   0   2   0   0   0   4   2   0   0   0   0   0   0   0   0   0   0   4   0   0   1   0   4   0   2   0
chr1   788071  790071  0   2   0   0   2   0   0   0   4   2   0   0   0   0   0   0   0   0   0   0   4   0   0   2   0   4   0   2   0
chr1   788597  790597  0   2   0   2   0   3   0   4   6   2   0   0   0   0   0   0   0   2   0   0   4   0   0   4   0   6   0   2   0
chr1   788874  790874  0   2   0   2   0   3   0   4   4   2   0   0   0   0   0   0   0   3   0   0   4   0   0   4   0   6   2   2   0
chr1   789534  791534  0   4   0   2   0   3   0   4   4   2   2   0   0   0   0   0   0   4   0   0   4   0   2   4   0   8   2   2   0
chr1   794168  796168  0   0   0   0   0   2   2   0   0   0   0   0   0   0   0   0   0   2   0   0   2   0   0   0   0   0   0   0   0
chr1   803926  805926  0   0   0   4   0   2   0   2   0   0   2   0   0   0   0   2   0   2   0   2   0   0   0   0   0   0   2   0   0
chr1   811814  813814  0   0   2   0   0   0   0   0   0   0   0   0   0   0   0   0   0   2   0   0   0   0   0   0   2   0   2   0   0
chr1   816242  818242  0   0   0   2   0   0   0   0   0   2   0   0   0   0   0   2   0   0   0   0   0   0   0   0   2   0   0   0   0
chr1   816886  818886  0   0   2   2   2   0   0   0   0   2   0   0   0   0   0   2   0   0   0   0   0   0   0   0   2   0   2   0   0
chr1   817138  819138  0   0   2   2   2   0   0   0   0   1   2   0   0   0   0   2   0   0   0   0   0   0   0   0   2   0   2   0   0
chr1   825734  827734  0   0   0   2   0   0   0   0   4   0   0   0   0   0   0   0   0   0   0   0   0   2   0   0   0   0   0   2   0
chr1   826094  828094  0   0   0   2   0   0   0   0   3   0   0   0   0   0   0   0   0   0   0   0   0   2   0   0   0   0   0   2   0
chr1   826482  828482  0   0   0   2   0   0   0   0   2   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   2   0
chr1   829930  831930  0   0   4   0   0   0   0   0   2   4   2   0   0   0   0   0   0   0   0   0   0   2   0   0   0   0   0   0   0
chr1   830435  832435  0   0   0   0   0   0   0   0   2   2   0   0   0   0   0   0   0   0   0   0   0   0   0   2   0   0   0   0   0
chr1   832388  834388  0   0   0   0   2   0   0   2   0   8   2   0   0   0   0   0   0   0   0   0   2   0   0   0   2   2   0   2   3
```

Output – altered gene activity between clones (DESeq2)

data/RPE_mixture/scNOVA_result/Result_scNOVA_infer_expression_table.txt



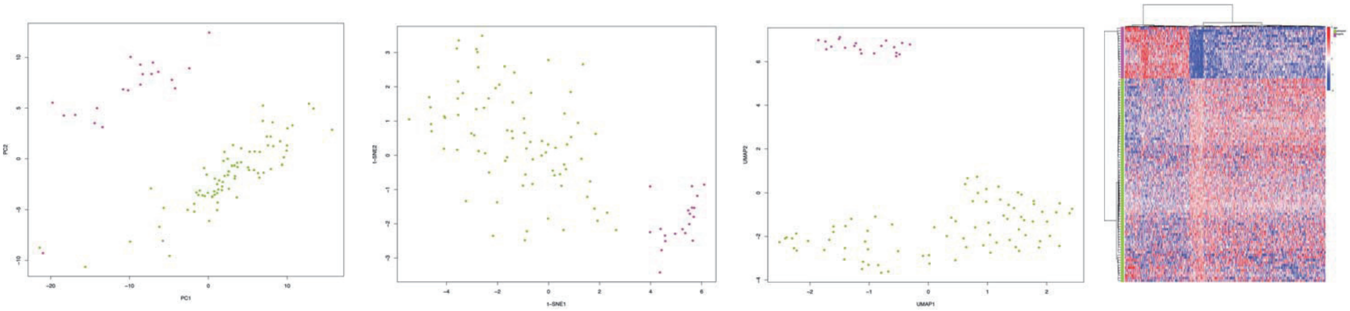
S16

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		chr	str	end	name	baseMean	log2FoldChai	lfcSE	stat	pvalue	padj	input_matrix	Expressed_p	Hit		
2	1	chr1	69091	70008	ENSG000001	1	1	1	1	1	1	0	0.02016222	0		
3	2	chr1	450740	451678	ENSG000002	1	1	1	1	1	1	0	0.00165224	0		
4	3	chr1	685716	686654	ENSG000002	1	1	1	1	1	1	0	0.02159911	0		
5	4	chr1	924880	944581	ENSG000001	6.46002503	-0.3960438	0.21062074	-1.8803649	0.06005837	0.34450411	0	0.43698305	0		
6	5	chr1	944204	959309	ENSG000001	5.88780816	-0.2576724	0.22619832	-1.1391436	0.25464325	0.63533931	0	0.64471385	0		
7	6	chr1	960587	965715	ENSG000001	1.46755257	-0.1181165	0.45558389	-0.259264	0.79543155	NA	0	0.44806094	0		
8	7	chr1	966497	975865	ENSG000001	3.06086433	-0.9253158	0.3575531	-2.5879114	0.00965598	0.13142863	0	0.35206678	0		
9	8	chr1	975204	982093	ENSG000001	2.00238629	-0.2246833	0.39632354	-0.566919	0.57076923	NA	0	0.14545784	0		
10	9	chr1	998962	1000172	ENSG000001	0.23626821	-0.9713539	1.05248606	-0.9229138	0.35605213	NA	0	0.50299837	0		
11	10	chr1	1001138	1014541	ENSG000001	3.64719679	-0.1820437	0.29625061	-0.6144921	0.5388902	0.84138133	0	0.73696279	0		
12	11	chr1	1020123	1056118	ENSG000001	8.71175255	-0.4118836	0.19404028	-2.1226704	0.03378148	0.25755599	0	0.67562452	0		
13	12	chr1	1070966	1074307	ENSG000002	1	1	1	1	1	1	0	0.08617284	0		
14	13	chr1	1081818	1116361	ENSG000001	13.3535659	-0.0323651	0.14870756	-0.2176423	0.8277078	0.94896767	0	0.41408414	0		
15	14	chr1	1173884	1197935	ENSG000001	8.35775996	-0.2039737	0.18268609	-1.1165255	0.26419727	0.64221965	0	0.08320931	0		
16	15	chr1	1203508	1206691	ENSG000001	1.34356182	0.30893764	0.46867774	0.65916857	0.50978753	NA	0	0.46359975	0		
17	16	chr1	1211326	1214138	ENSG000001	0.99219791	0.06526064	0.57500097	0.11349657	0.90963689	NA	0	0.06677801	0		
18	17	chr1	1216908	1232031	ENSG00000C	4.420115	-0.1837014	0.25922929	-0.7086446	0.47854503	0.80718021	0	0.84957508	0		
19	18	chr1	1232265	1235041	ENSG000001	0.25809378	0.2996512	1.00827724	0.29719127	0.76632048	NA	0	0.84404128	0		
20	19	chr1	1242446	1246722	ENSG000001	1.17317178	-1.1792686	0.53539976	-2.2025945	0.02762334	NA	0	0.40187508	0		
21	20	chr1	1253909	1273885	ENSG000001	4.94192657	-0.6107308	0.24454275	-2.49744	0.01250936	0.15246679	0	0.7524455	0		
22	21	chr1	1280436	1292029	ENSG000001	3.17211492	-0.2269471	0.29652114	-0.7653657	0.4440538	0.78398215	0	0.14812104	0		
23	22	chr1	1308567	1311677	ENSG000001	1.05113951	-0.7560857	0.57481455	-1.3153559	0.18839032	NA	0	0.3731774	0		
24	23	chr1	1292376	1309609	ENSG000001	4.83928988	-0.7271657	0.25238422	-2.8811855	0.00396182	0.0748867	0	0.64854132	1		
25	24	chr1	1311585	1324691	ENSG000001	3.99016865	-0.501079	0.25504467	-1.9646716	0.04945226	0.31190587	0	0.8577289	0		
26	25	chr1	1324756	1328897	ENSG000002	1.37990224	-0.2838296	0.47296978	-0.6001009	0.54843897	NA	0	0.78980667	0		
27	26	chr1	1331314	1335306	ENSG000001	1.25906914	-0.9786393	0.57140569	-1.7126874	0.08677006	NA	0	0.17447004	0		

Output – altered gene activity between clones (DESeq2) plots

Result of infer altered gene activity between clones (DESeq2) - plots

data/RPE_mixture/scNOVA_result_plots/Result_scNOVA_plots_RPE_mixture.pdf

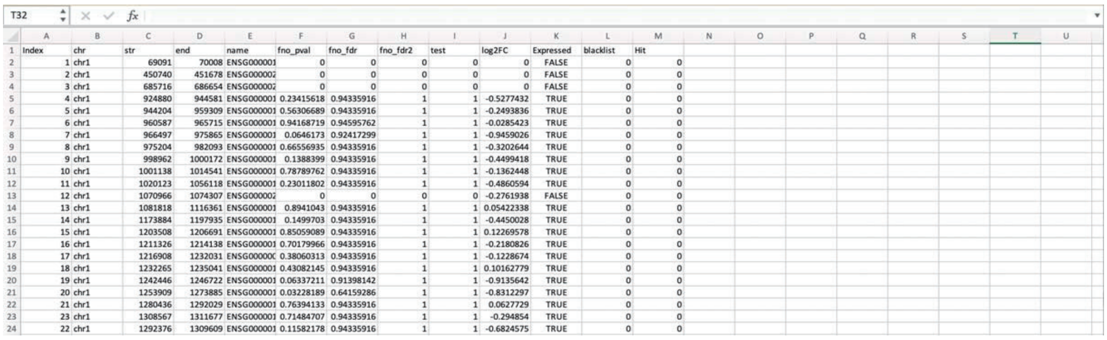


🟩 Clone1 (RPE1_WT)

🟪 Clone2 (RPE1_BM510)

Output – altered gene activity between clones (PLSDA)

data/RPE_mixture/scNOVA_result/result_PLSDA_RPE_mixture.txt

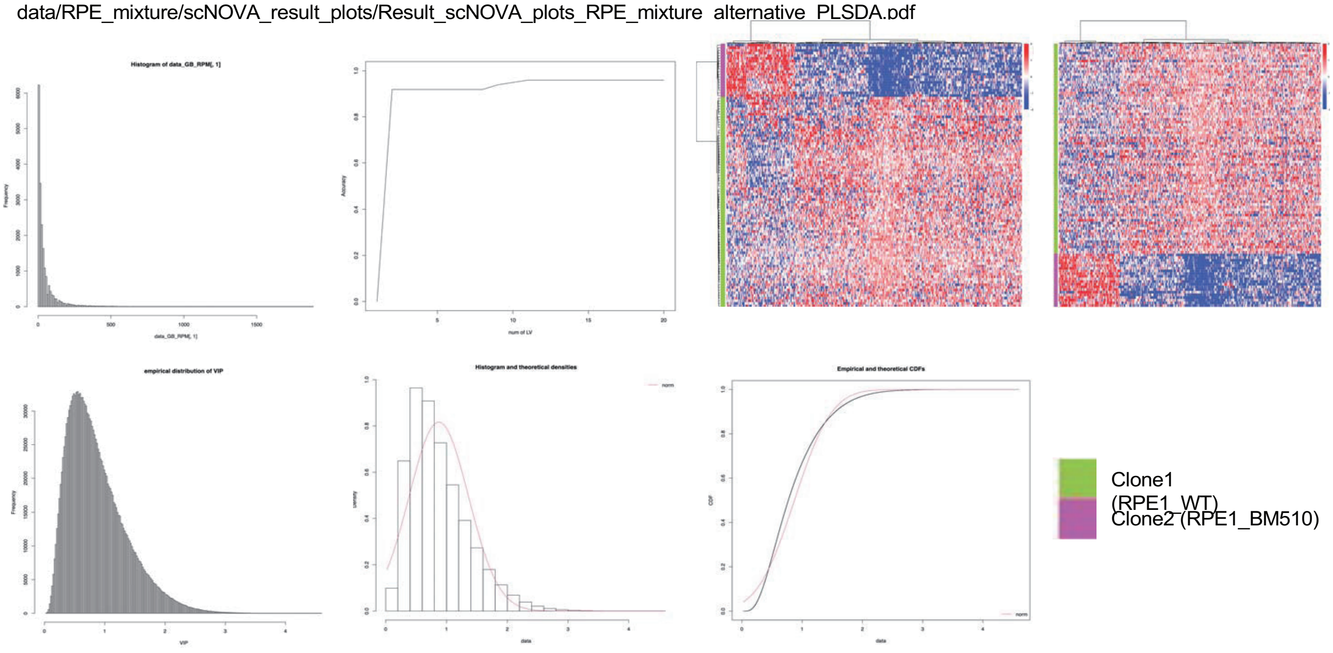


T32

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Index	chr	str	end	name	fno_pval	fno_fdr	fno_fdr2	test	log2FC	Expressed	blacklist	Hit
2	1	chr1	69091	70008	ENSG000001	0	0	0	0	0	FALSE	0	0
3	2	chr1	450740	451678	ENSG000002	0	0	0	0	0	FALSE	0	0
4	3	chr1	685716	686654	ENSG000002	0	0	0	0	0	FALSE	0	0
5	4	chr1	924880	944581	ENSG000001	0.23415618	0.94335916	1	1	-0.5277432	TRUE	0	0
6	5	chr1	944204	959309	ENSG000001	0.56306689	0.94335916	1	1	-0.2493836	TRUE	0	0
7	6	chr1	960587	965715	ENSG000001	0.94168719	0.94595762	1	1	-0.0285423	TRUE	0	0
8	7	chr1	966497	975865	ENSG000001	0.0646173	0.92417299	1	1	-0.9459026	TRUE	0	0
9	8	chr1	975204	982093	ENSG000001	0.66556935	0.94335916	1	1	-0.3202644	TRUE	0	0
10	9	chr1	998962	1000172	ENSG000001	0.1388399	0.94335916	1	1	-0.4499418	TRUE	0	0
11	10	chr1	1001138	1014541	ENSG000001	0.78789762	0.94335916	1	1	-0.1362448	TRUE	0	0
12	11	chr1	1020123	1056118	ENSG000001	0.23011802	0.94335916	1	1	-0.4860594	TRUE	0	0
13	12	chr1	1070966	1074307	ENSG000002	0	0	0	0	-0.2761938	FALSE	0	0
14	13	chr1	1081818	1116361	ENSG000001	0.8941043	0.94335916	1	1	0.05422338	TRUE	0	0
15	14	chr1	1173884	1197935	ENSG000001	0.1499703	0.94335916	1	1	-0.4450028	TRUE	0	0
16	15	chr1	1203508	1206691	ENSG000001	0.85059089	0.94335916	1	1	0.12269578	TRUE	0	0
17	16	chr1	1211326	1214138	ENSG000001	0.70179966	0.94335916	1	1	-0.2180826	TRUE	0	0
18	17	chr1	1216908	1232031	ENSG00000C	0.38060313	0.94335916	1	1	-0.1228674	TRUE	0	0
19	18	chr1	1232265	1235041	ENSG000001	0.43082145	0.94335916	1	1	0.10162779	TRUE	0	0
20	19	chr1	1242446	1246722	ENSG000001	0.06337211	0.91398142	1	1	-0.9135642	TRUE	0	0
21	20	chr1	1253909	1273885	ENSG000001	0.03228189	0.64159286	1	1	-0.8312297	TRUE	0	0
22	21	chr1	1280436	1292029	ENSG000001	0.76394133	0.94335916	1	1	0.0627729	TRUE	0	0
23	23	chr1	1308567	1311677	ENSG000001	0.71484707	0.94335916	1	1	-0.294854	TRUE	0	0
24	22	chr1	1292376	1309609	ENSG000001	0.11582178	0.94335916	1	1	-0.6824575	TRUE	0	0

Output – altered gene activity between clones (PLSDA) plots

data/RPE_mixture/scNOVA_result_plots/Result_scNOVA_plots_RPE_mixture_alternative_PLSDA.pdf



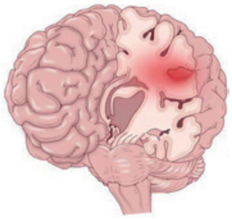
KSBi 한국생명정보학회 Korean Society for Bioinformatics

Part4. scRNA-seq에서 서브클론을 유추하고 기능적으로 분석하는 멀티오믹스 기법 소개

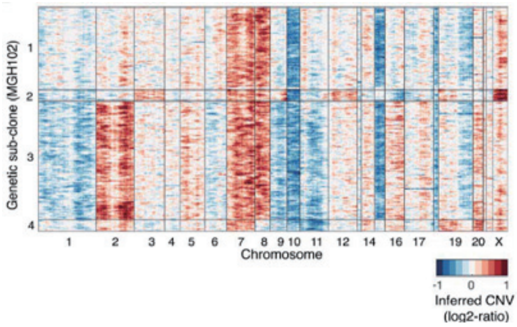
Single-cell multi-omics analysis to study tumor subclones

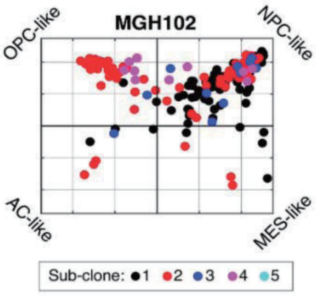
Recent strategy to study genome and functional readout from single-cell RNA-seq

Patient tumor



Infer CNV from single-cell RNA-seq

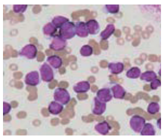




Neftel et al. Cell, 2019

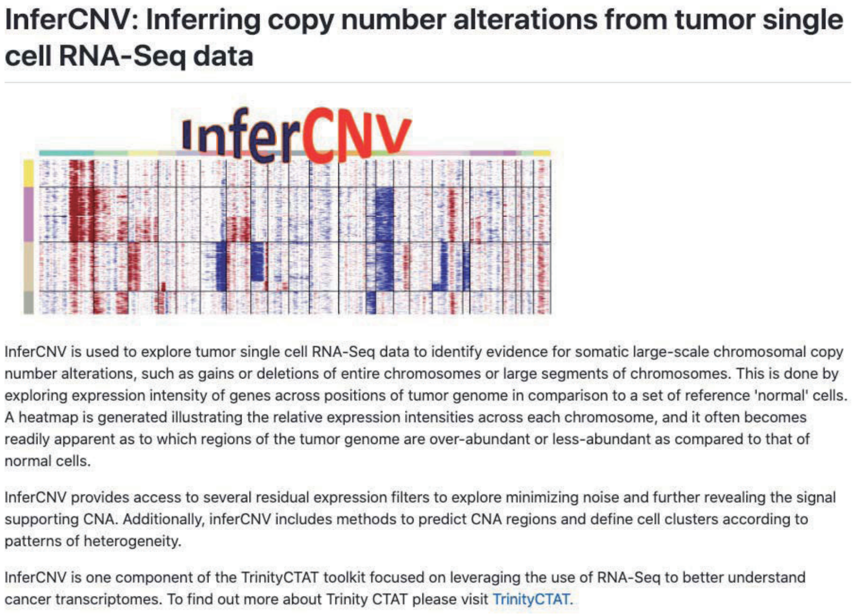
Recent strategy to study genome and functional readout from single-cell RNA-seq

SCNA inference methods based on transcriptome



		Method detail			
	Method	SV class	Require pre-defined SV breakpoint	Size resolution in the paper	Chr6 SV detection
Discovery	InferCNV (Science, 2014)	CNV only	N	entire chromosomes or large segments of chromosomes	N
	HoneyBADGER (Genome Res, 2018)	CNV only	N	10Mb	N
	CONICSmat 'discovery mode' (Bioinformatics, 2018)	CNV only	N	100 expressed genes (by default)	N
Genotyping	CONICSmat 'genotype mode' (Bioinformatics, 2018) User provide candidate SCNA	CNV only	Y	100 expressed genes (by default)	Y

Recent strategy to study genome and functional readout from single-cell RNA-seq (InferCNV)



InferCNV: Inferring copy number alterations from tumor single cell RNA-Seq data

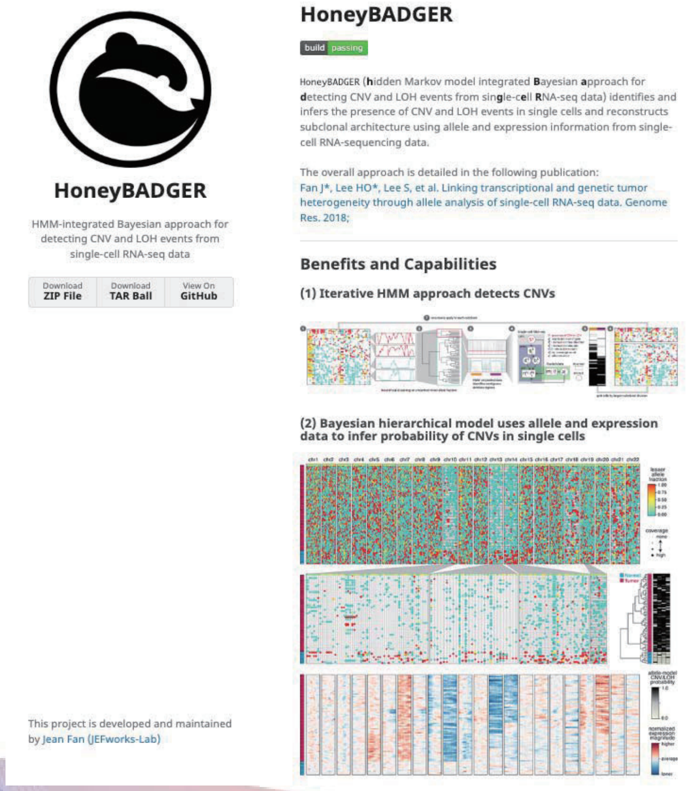
InferCNV is used to explore tumor single cell RNA-Seq data to identify evidence for somatic large-scale chromosomal copy number alterations, such as gains or deletions of entire chromosomes or large segments of chromosomes. This is done by exploring expression intensity of genes across positions of tumor genome in comparison to a set of reference 'normal' cells. A heatmap is generated illustrating the relative expression intensities across each chromosome, and it often becomes readily apparent as to which regions of the tumor genome are over-abundant or less-abundant as compared to that of normal cells.

InferCNV provides access to several residual expression filters to explore minimizing noise and further revealing the signal supporting CNA. Additionally, inferCNV includes methods to predict CNA regions and define cell clusters according to patterns of heterogeneity.

InferCNV is one component of the TrinityCTAT toolkit focused on leveraging the use of RNA-Seq to better understand cancer transcriptomes. To find out more about Trinity CTAT please visit TrinityCTAT.

https://github.com/broadinstitute/inferCNV/wiki

Recent strategy to study genome and functional readout from single-cell RNA-seq (HoneyBADGER)



HoneyBADGER

HMM-integrated Bayesian approach for detecting CNV and LOH events from single-cell RNA-seq data

Download ZIP File | Download TAR Ball | View On GitHub

This project is developed and maintained by Jean Fan (JEFworks-Lab)

HoneyBADGER

build passing

HoneyBADGER (**h**idden Markov model integrated **B**ayesian **a**pproach for **d**etecting CNV and LOH events from sin**g**le-c**e**ll **R**NA-seq data) identifies and infers the presence of CNV and LOH events in single cells and reconstructs subclonal architecture using allele and expression information from single-cell RNA-sequencing data.

The overall approach is detailed in the following publication:
Fan J*, Lee HO*, Lee S, et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. Genome Res. 2018;

Benefits and Capabilities

(1) Iterative HMM approach detects CNVs

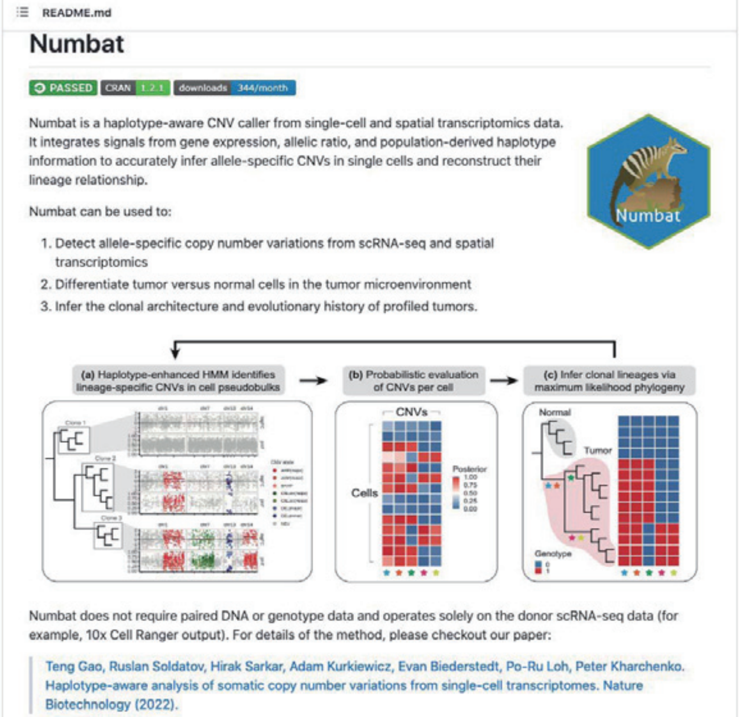
(2) Bayesian hierarchical model uses allele and expression data to infer probability of CNVs in single cells

chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22
lesser allele fraction
coverage
Normal
Tumor
allele-model CNV/LOH probability
normalized expression magnitude

89

https://jef.works/HoneyBADGER/

Recent strategy to study genome and functional readout from single-cell RNA-seq (NumBat)



README.md

Numbat

PASSED | CRAN 1.2.1 | downloads 344/month

Numbat is a haplotype-aware CNV caller from single-cell and spatial transcriptomics data. It integrates signals from gene expression, allelic ratio, and population-derived haplotype information to accurately infer allele-specific CNVs in single cells and reconstruct their lineage relationship.

Numbat can be used to:

1. Detect allele-specific copy number variations from scRNA-seq and spatial transcriptomics
2. Differentiate tumor versus normal cells in the tumor microenvironment
3. Infer the clonal architecture and evolutionary history of profiled tumors.

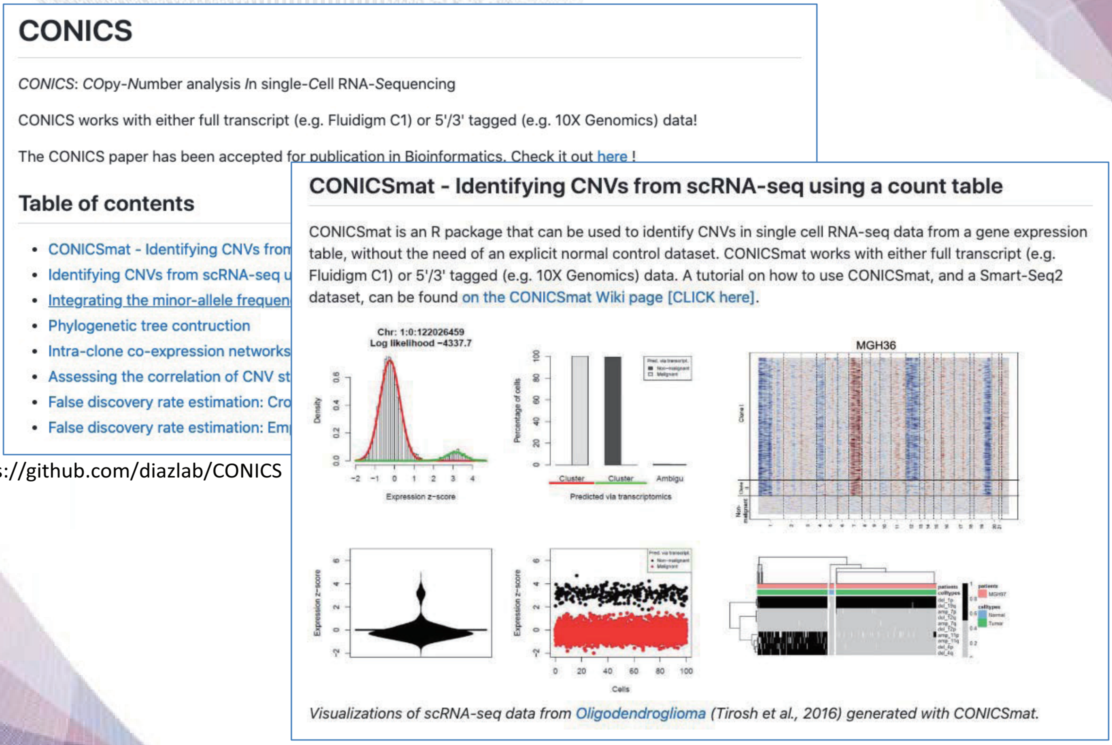
Numbat
(a) Haplotype-enhanced HMM identifies lineage-specific CNVs in cell pseudobulks
(b) Probabilistic evaluation of CNVs per cell
(c) Infer clonal lineages via maximum likelihood phylogeny
Clone 1
Clone 2
Clone 3
CNVs
Cells
Posterior
1.00
0.75
0.50
0.25
0.00
Normal
Tumor
Genotype
0
1

Numbat does not require paired DNA or genotype data and operates solely on the donor scRNA-seq data (for example, 10x Cell Ranger output). For details of the method, please checkout our paper:

> Teng Gao, Ruslan Soldatov, Hirak Sarkar, Adam Kurkiewicz, Evan Biederstedt, Po-Ru Loh, Peter Kharchenko. Haplotype-aware analysis of somatic copy number variations from single-cell transcriptomes. Nature Biotechnology (2022).

https://github.com/kharchenkolab/numbat 90

Recent strategy to study genome and functional readout from single-cell RNA-seq (CONICS)



CONICS

CONICS: *CO*py-*N*umber analysis *I*n single-*C*ell RNA-*S*equencing

CONICS works with either full transcript (e.g. Fluidigm C1) or 5'/3' tagged (e.g. 10X Genomics) data!

The CONICS paper has been accepted for publication in Bioinformatics. Check it out here !

Table of contents

- CONICSmat - Identifying CNVs from
- Identifying CNVs from scRNA-seq u
- Integrating the minor-allele frequen
- Phylogenetic tree contruction
- Intra-clone co-expression networks
- Assessing the correlation of CNV st
- False discovery rate estimation: Cro
- False discovery rate estimation: Em

CONICSmat - Identifying CNVs from scRNA-seq using a count table

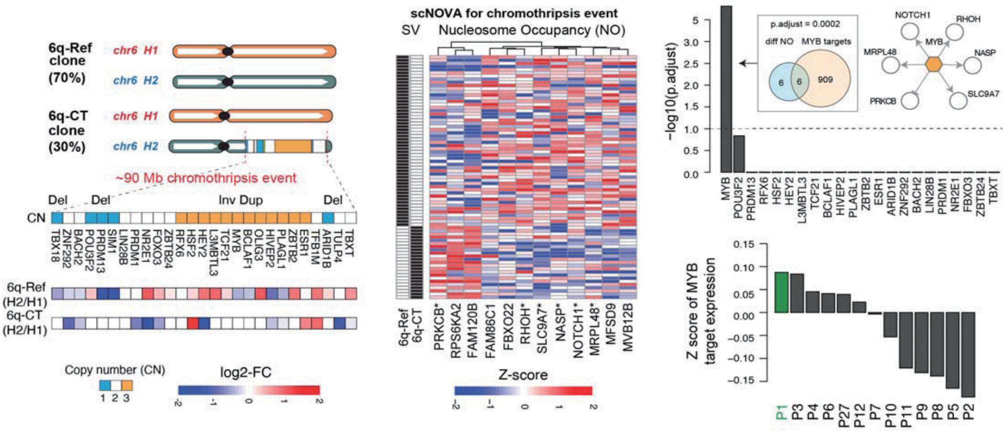
CONICSmat is an R package that can be used to identify CNVs in single cell RNA-seq data from a gene expression table, without the need of an explicit normal control dataset. CONICSmat works with either full transcript (e.g. Fluidigm C1) or 5'/3' tagged (e.g. 10X Genomics) data. A tutorial on how to use CONICSmat, and a Smart-Seq2 dataset, can be found on the CONICSmat Wiki page [CLICK here].

Visualizations of scRNA-seq data from Oligodendroglioma (Tirosh et al., 2016) generated with CONICSmat.

https://github.com/diazlab/CONICS

Applying CNV inference of scRNA-seq to the T-ALL case study

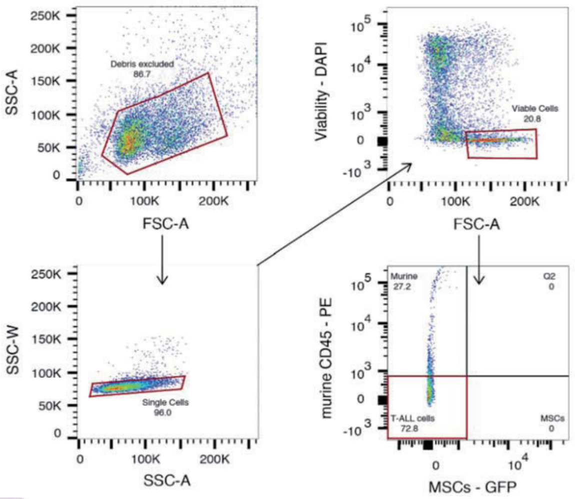
Hypothesis : 6q-CT cells have MYB-Notch activation compared to 6-Ref cells



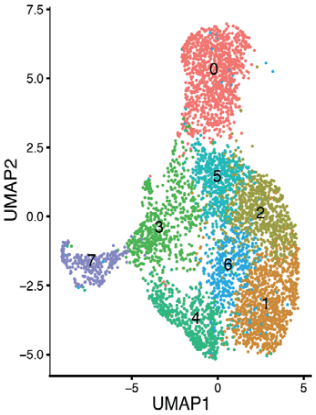
Single-cell experiment is needed to confirm subclonal level transcriptome changes

Applying CONICSmat to the T-ALL case study

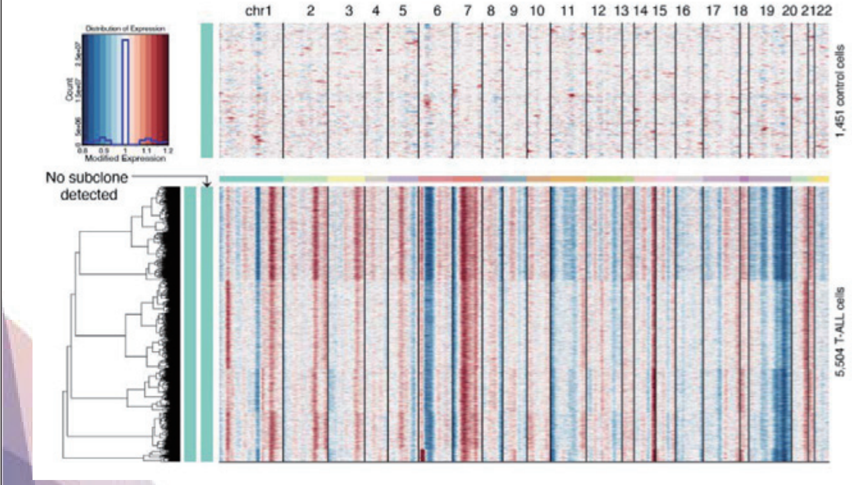
Gating strategy for single, viable T-ALL cell isolation from T-ALL sample T-ALL_P1 for scRNA-seq.



Unsupervised analysis of transcriptome



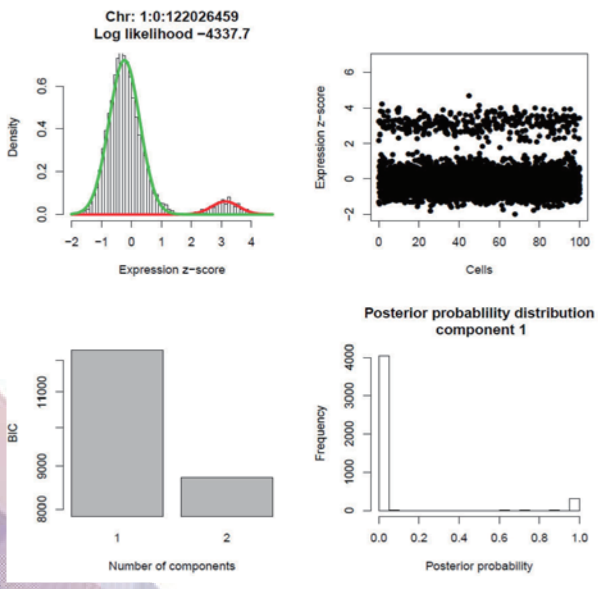
Applying InferCNV to the T-ALL case study



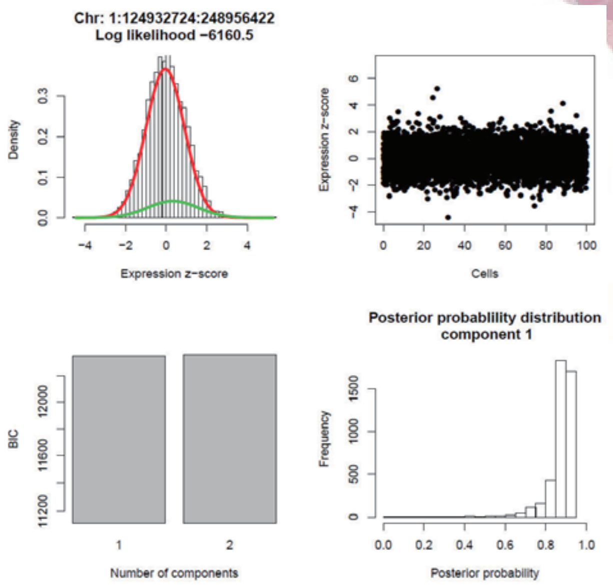
InferCNV analysis of 5,504 high quality T-ALL_P1 cells, and 1,451 control cells. Control cells were downloaded from PBMC data provided by 10X Genomics. This analysis did not discover subclones in 5,504 T-ALL cells.

Applying CONICSmat to the T-ALL scRNA-seq (Genotyping mode)

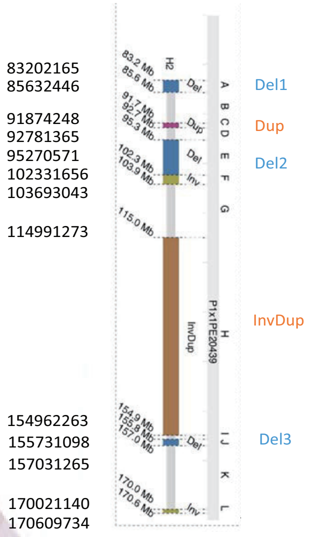
Presence of Subclonal CNA



Absence of Subclonal CNA

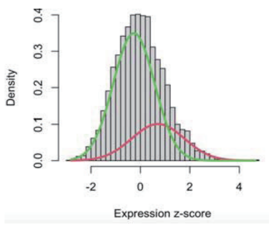


CONICSmat analysis supports the presence of chr6 deletions and duplications

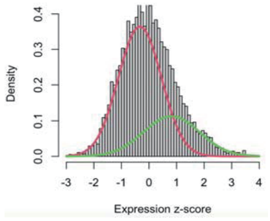


Sanders et al. 2020

Duplication (dup + invdup)



Deletion (del1 + del2 + del3)

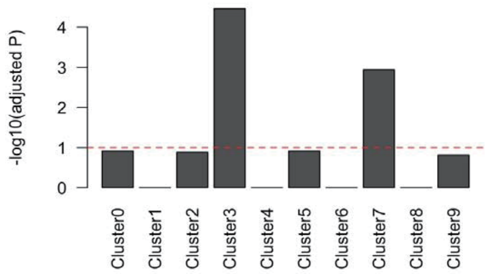


Genomic range	BIC 1 component	BIC 2 components	BIC difference	LRT adj. p-val	CNV call (CF%)
chr6_Del 35 genes	15635.9018	15553.1101	82.7917307	0	729 cells (13.2%)
chr6_Dup 192 genes	15635.9018	15481.839	154.062863	0	265 cells (4.8%)

10X transcriptome experiment from Karen Grimes

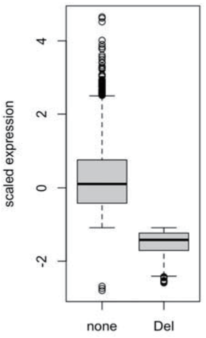
Cluster3 and Cluster7 cells are highly enriched with deletion calls

Deletion call (del1 + del2 + del3) probability>0.9

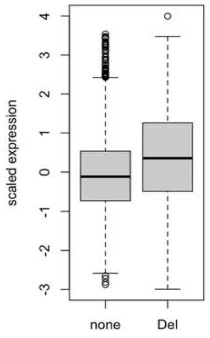


Row Labels	Del(0.9)	%Del(all)	p-value	adjustedP
Cluster0	166	14.901	0.039	0.122
Cluster1	97	9.454	1.000	1.000
Cluster2	118	15.013	0.066	0.131
Cluster3	118	19.440	0.000	0.000
Cluster4	40	6.981	1.000	1.000
Cluster5	83	15.690	0.049	0.122
Cluster6	32	6.695	1.000	1.000
Cluster7	62	20.395	0.000	0.001
Cluster8	6	10.526	0.785	1.000
Cluster9	7	23.333	0.092	0.154
Grand Total	729	13.245	-	-

Deleted region
Ranksum test, t.test
p-value < 2.2e-16



Duplicated region
Ranksum test, t.test
p-value < 2.2e-16

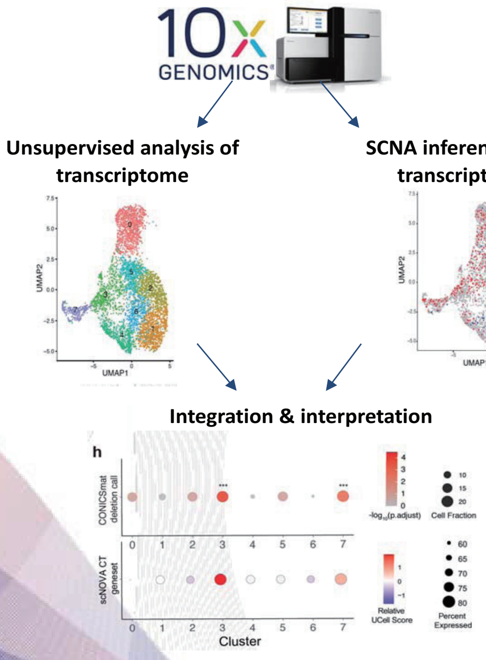


Type	Dup	none
Del	96	633
none	169	4606

Fisher exact test

p-value < 2.2e-16

SV subclone in P1 shows increase of MYB target expression and cells with premature stages in the cellular hierarchy

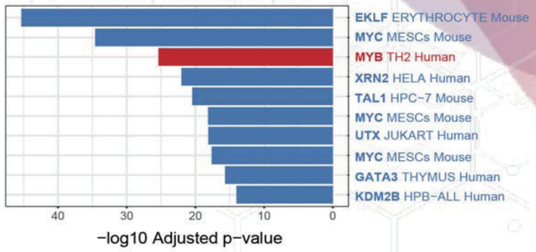


Unsupervised analysis of transcriptome

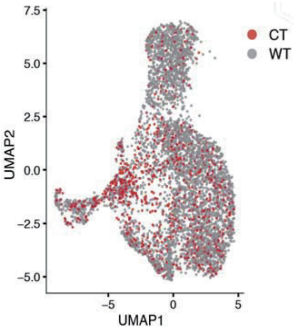
SCNA inference from transcriptome

Integration & interpretation

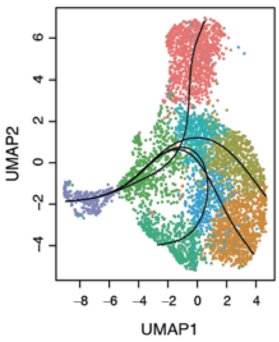
TF-target Enrichment (Cluster 3)



Signature gene activity



Lineage analysis



Summary

- 암에서 이러한 서브클론들을 동정하기 위해 어떤 싱글셀 오믹스 기법들이 개발되어 있을까? → *single-cell WGS, SDR-seq, Strand-seq, etc*
- 이러한 싱글셀 오믹스 데이터를 분석하기 위해 어떤 생명 정보학적인 도구들을 사용할 수 있을까? → *MosaiCatcher for Strand-seq analysis*
- 서브클론의 동정 뿐 아니라 그 기능적 특성을 파악하기 위해서는 유전체와 전사체 또는 후성유전체 데이터를 함께 분석하는 싱글셀 멀티 오믹스 분석이 필요하다. 이를 구현하기 위한 생명 정보학적인 방법에는 어떤 것들이 있을까?

  → *scNOVA for Strand-seq analysis*

  → *Infer copy number alteration from scRNA-seq (InferCNV, HoneyBADGER, Numbat, CONICS etc.)*

감사합니다.