KSBi-BIML 2023

Bioinformatics & Machine Learning(BIML) Workshop for Life Scientists, Data Scientists, and Bioinformatians

생물정보학&머신러닝 워크샵(온라인)

Single-cell multi-omics analysis to study tumor subclones

정효빈 _ 한양대학교

SBi 한국생명정보학회 Korean Society for Bioinformatics

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2023 워크샵 온라인 수업을 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 전적으로 불법 행위자 본인에게 있음을 경고**합니다.

KSBi-BIML 2023

Bioinformatics & Machine Learning (BIML)
Workshop for Life Scientists, Data Scientists, and Bioinformatians

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크샵인 BIML-2023에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하여 올해로 9차를 맞이하게 되었습니다. 지난 2년간은 심각한 코로나 대유행으로 인해 아쉽게도 모든 강의가 온라인으로 진행되어 현장 강의에서만 가능한 강의자와 수강생 사이에 다양한 소통의 기회가 없음에 대한 아쉬움이 있었습니다. 다행히도 최근 사회적 거리두기 완화로 현장 강의가 가능해져 올해는 현장 강의를 재개함으로써 온라인과 현장 강의의 장점을 모두 갖춘 프로그램을 구성할 수 있게 되었습니다.

BIML 워크샵은 전통적으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되었습니다. 올해 AI 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 심층학습(Deep learning) 기법들에 대한 현장 강의가 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체 분석, 신약개발에 대한 이론과 실습 강의가 함께 제공될 예정입니다. 또한 싱글셀오믹스 분석과 메타유전체분석 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다. 이외에 다양한 생명정보학 분야에 대하여 30개 이상의 온라인 강좌가 개설되어 제공되며 온라인 강의의 한계를 극복하기 위해서 실시간 Q&A 세션 또한 마련했습니다. 특히 BIML은 각 분야 국내 최고 전문가들의 강의로 구성되어 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의가 될 것입니다.

이번 BIML-2023을 준비하기까지 너무나 많은 수고를 해주신 BIML-2023 운영위원회의 남진우, 우현구, 백대현, 정성원, 정인경, 장혜식, 박종은 교수님과 KOBIC 이병욱 박사님께 커다란 감사를 드립니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 현장 강의와 온라인 강의를 준비하시는데 노고를 아끼지 않으신 모든 연사분께 깊은 감사를 드립니다.

2023년 2월

한국생명정보학회장 이 인 석

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)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
- scRNA-seq 으로부터 서브클론을 유추하기 위한 데이터 분석 방법 소개
- 서브클론을 동정한 후 functional analysis를 위한 싱글셀 멀티오믹스 접근법과 생명정보학 도구 소개

* 교육생준비물: 노트북, R (또는 R studio)

* 강의 난이도: 초급

* 강의: 정효빈 교수 (한양대학교 생명과학과 | 한양생명과학기술원)

Curriculum Vitae

Speaker Name: Hyobin Jeong, Ph.D.

▶ **Personal Info**

Name	Hyobin Jeong
Title	Research Professor (연구전임교원)
Affiliation	Hanyang University

▶ **Contact Information**

Address	222 Wangsimni-ro, Seongdong-gu, Seoul 04763, Korea
Email	hyobinjeong@hanyang.ac.kr
Phone Number	010-4365-9054

Research Interest

Systems Biology of somatic mosaicism in aging and cancer
Computational tool development for Single-cell multi-omics
Disease marker discovery using multi-omics integration

Educational Experience

2007-2011.02	B.S. in Chemical Engineering, POSTECH, Korea
2011-2015.02	Ph.D. in Systems Biology, School of Interdisciplinary Bioscience & Bioengineering, POSTECH, Korea

Professional Experience

2015	Post-doc fellow, Institute of Basic Science, Korea
2016-2017	Post-doc fellow, Institute of Molecular Biology, Germany
2018-2022.08	Post-doc fellow, European Molecular Biology Laboratory (EMBL), Germany
2022.09-present	Research Professor, Hanyang University (Dept. of Life Science, College of Natural Science | Hanyang Institute of Bioscience and Biotechnology)

Selected Publications (5 maximum)

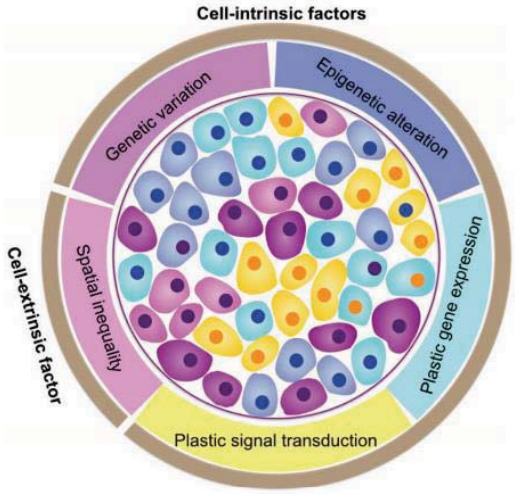
1. **Hyobin Jeong***, Karen Grimes*, Kerstin K. Rauwolf, Peter-Martin Bruch, Tobias Rausch, Patrick Hasenfeld, Eva Benito Garagorri, Tobias Roider, Radhakrishnan Sabarinathan, David Porubsky, Sophie A. Herbst, Büşra Erarslan-Uysal, Johann-Christoph Jann, Tobias Marschall, Daniel Nowak, Jean-Pierre Bourquin, Andreas E. Kulozik, Sascha Dietrich, Beat Bornhauser, Ashley D. Sanders#, Jan O. Korbel#, (2022.11) "Functional analysis of structural variants in single cells using Strand-seq", ***Nature Biotechnology*** **(*: equally contributed**).
2. Jung Yeon Kim, Juhyeon Lee, Myeong Hoon Kang, Tran Thi My Trang, Jusung Lee, Heeho Lee, **Hyobin Jeong**#, and Pyung Ok Lim#, (2022.11) "Dynamic Landscape of Long Noncoding RNAs during Leaf Aging in Arabidopsis", Accepted for publication in ***Frontiers in Plant Science***, **(#: co-corresponding)**

3. Jong-Chan Park*, Sun-Ho Han*, Hangyeore Lee*, **Hyobin Jeong***, Min Soo Byun, Jingi Bae, Hokeun Kim, Dong Young Lee, Dahyun Yi, Seong A Shin, Yu Kyeong Kim, Daehee Hwang, Sang-Won Lee, Inhee Mook-Jung (2019.12) "Prognostic plasma protein panel for brain Aβ deposition in Alzheimer's disease", ***Progress in Neurobiology***, 183:101690. **(*: equally contributed)**.
4. Hye Kyeong Kwon*, **Hyobin Jeong***, Daehee Hwang, Zee-Yong Park (2018.07) "Comparative Proteomic Analysis of Mouse Models of Pathological and Physiological Cardiac Hypertrophy, with Selection of Biomarkers of Pathological Hypertrophy by Integrative Proteogenomics", ***BBA - Proteins and Proteomics***, S1570-9639(18)30118-3. **(*: equally contributed)**.
5. **Hyobin Jeong***, Vijay K Tiwari# (2018.01) "Exploring the complexity of cortical development using single-cell transcriptomics" Mini Review, Fron. Neurosci - Neurogenesis. 2018 Jan 15; (*first)

KSBi-BIML 2023

Single-cell multi-omics analysis to
study tumor subclones

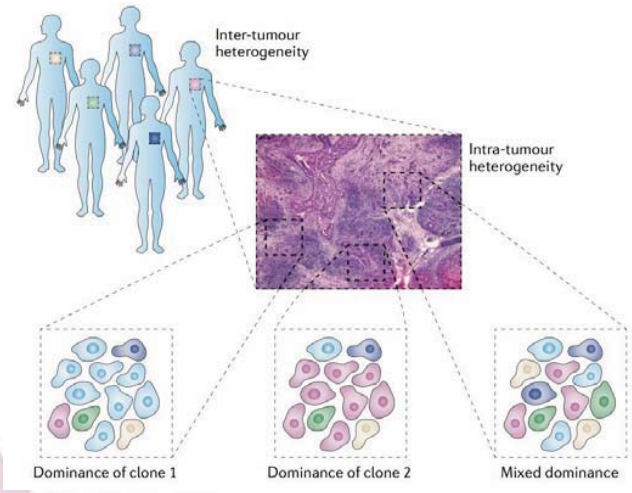
Tumor is composed of multiple subclones that makes intra-tumor heterogeneity



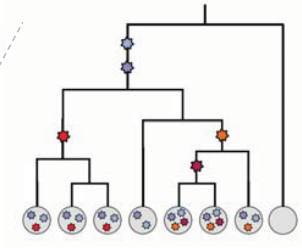
Acta Pharmacologica Sinca (2015)

2

Multi-layered heterogeneity contributes to therapy failure and cancer progression

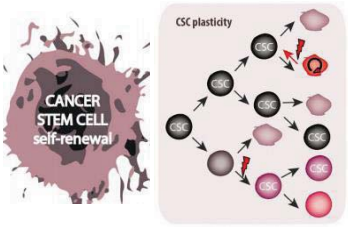


Nature review cancer, 2012



Genome Biology, 2016

Genetic variation



Molecular cancer, 2017

Epigenetic (functional) alteration

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

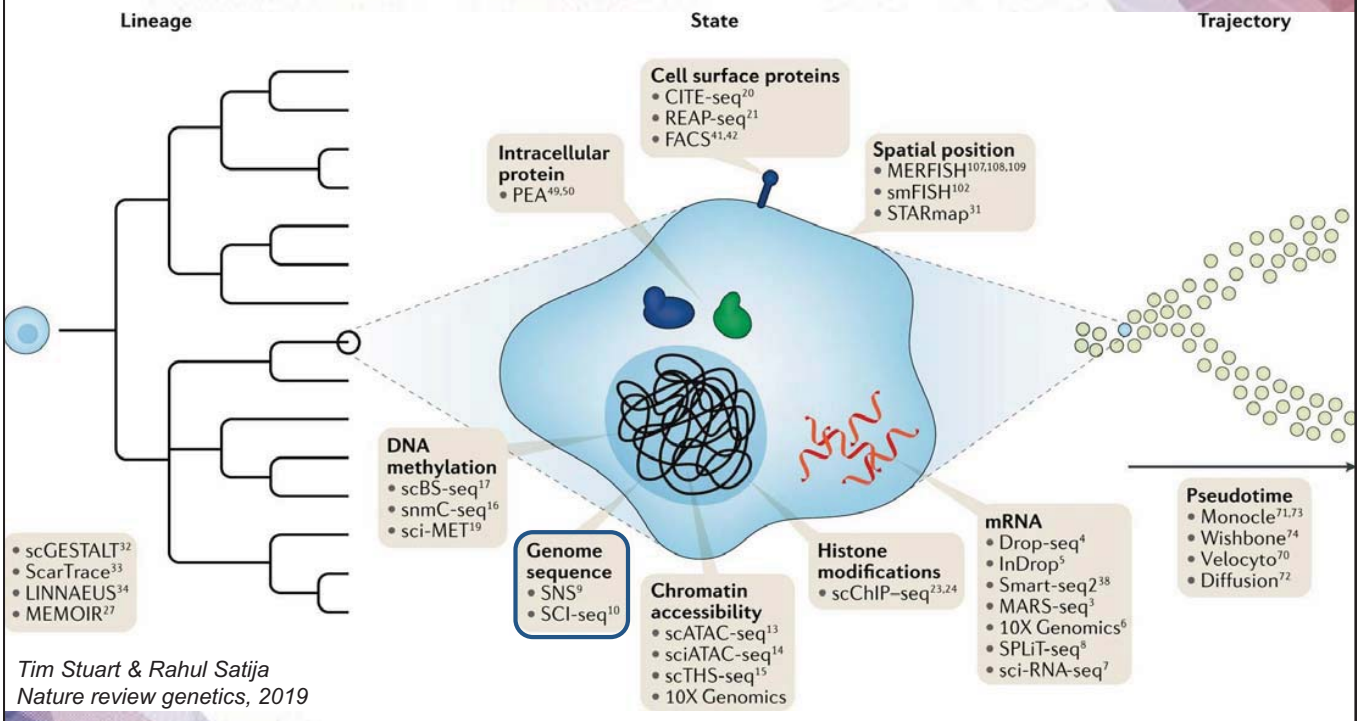
- 암에서 이러한 서브클론들을 동정하기 위해 어떤 싱글셀 오믹스 기법들이 개발되어 있을까?
- 이러한 싱글셀 오믹스 데이터를 분석하기 위해 어떤 생명 정보학적인 도구들을 사용할 수 있을까?
- 서브클론의 동정 뿐 아니라 그 기능적 특성을 파악하기 위해서는 유전체와 전사체 또는 후성유전체 데이터를 함께 분석하는 싱글셀 멀티 오믹스 분석이 필요하다. 이를 구현하기 위한 생명 정보학적인 방법에는 어떤 것들이 있을까?

SBI 한국생명정보학회 Korean Society for Bioinformatics

Part1. 암에서 서브클론을 동정하기 위한 싱글셀 오믹스 기법들에 대한 소개

Single-cell multi-omics analysis to study tumor subclones

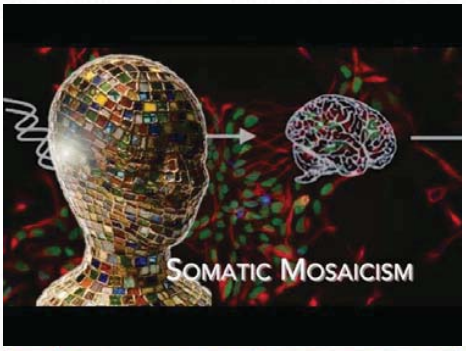
Single-cell technologies to explore cellular heterogeneity

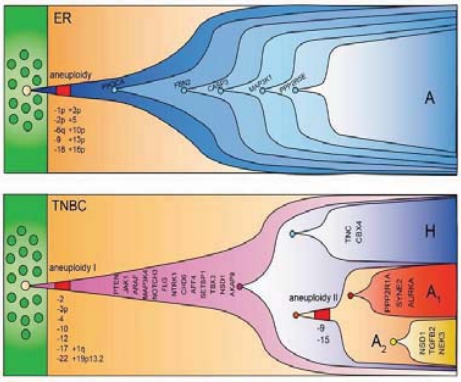


Tim Stuart & Rahul Satija
Nature review genetics, 2019

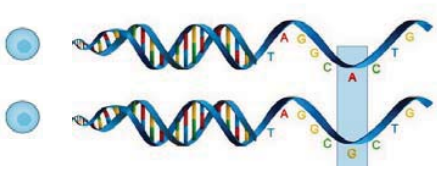
6

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

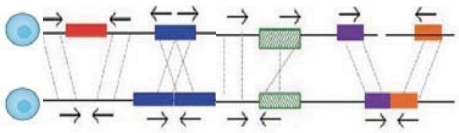




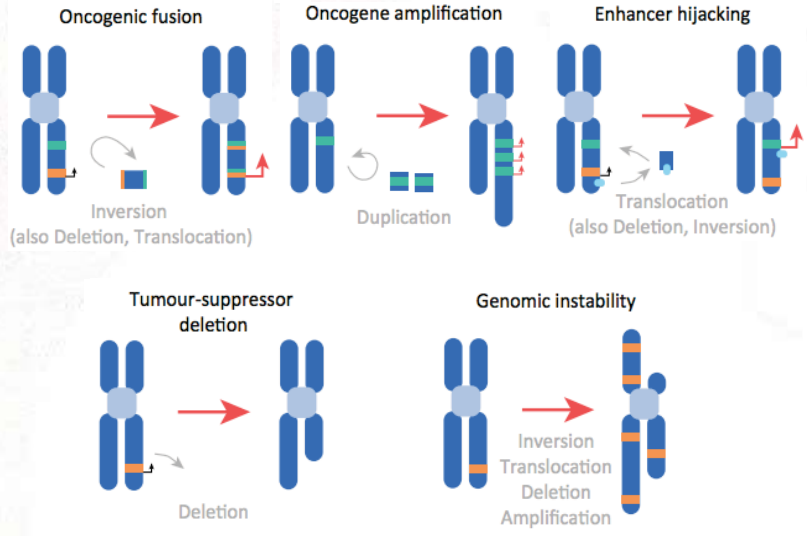
Single nucleotide variation



Structural Variation (SVs) – Strand-seq



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



Macintyre et al. 2016

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

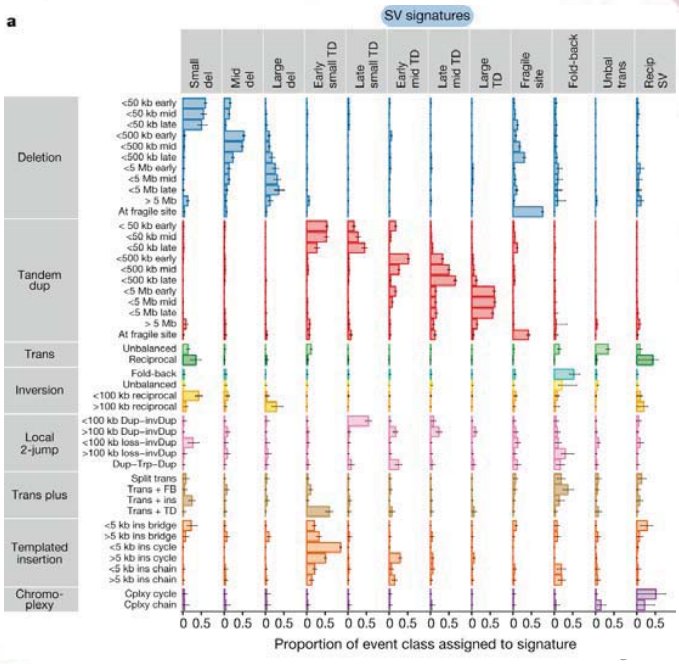
Article | Open Access | Published: 05 February 2020

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

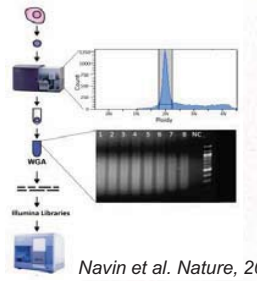
Nature **578**, 112–121 (2020) | Cite this article

79k Accesses | **267** Citations | **175** Altmetric | Metrics



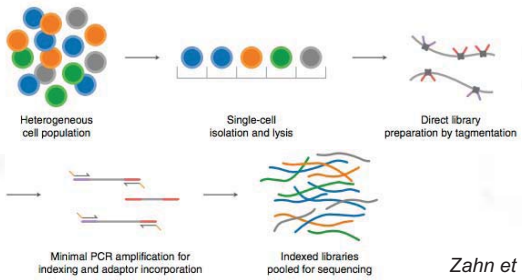
Single-cell technologies to explore *genetic* heterogeneity

Single-nucleus sequencing (SNS)



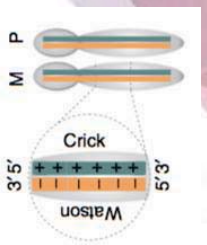
Navin et al. Nature, 2011

Direct library



Zahn et al. Nat Methods, 2017

Strand-seq



Sanders et al. Nat protocol, 2017

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

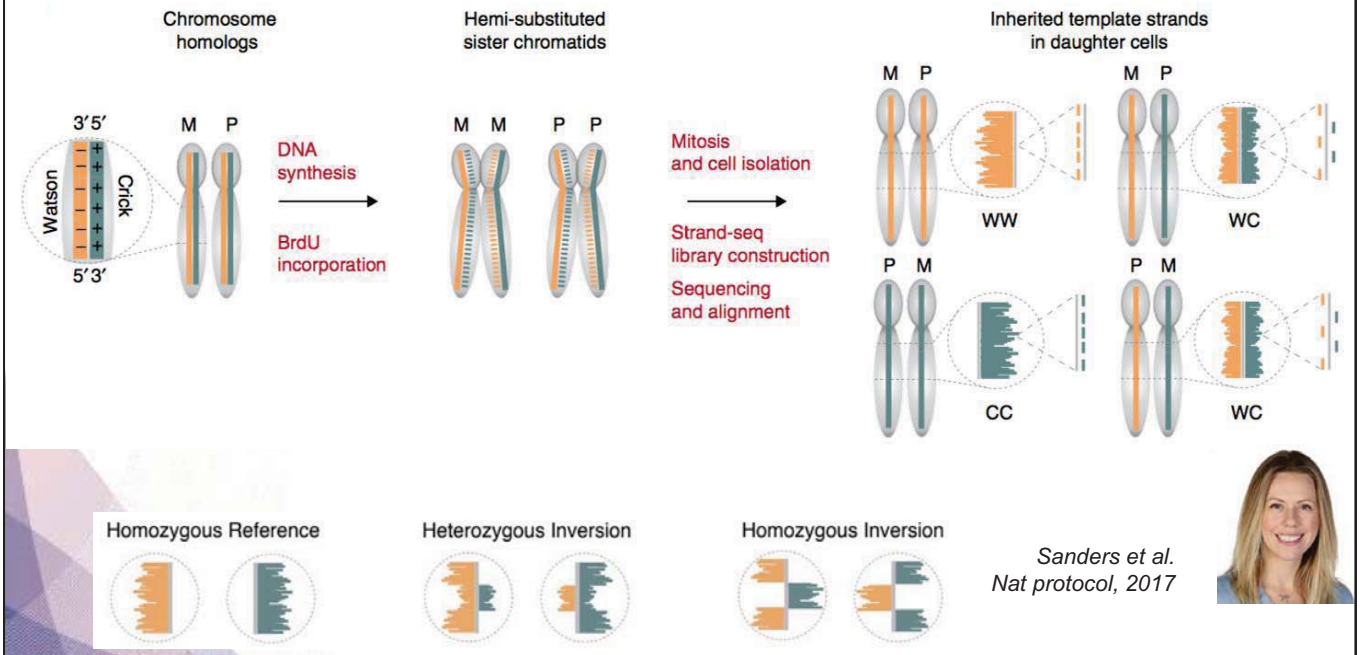
Step2. Remove PCR duplicate: Picard mark duplicate, Biobambam

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



Sanders et al. Nat protocol, 2017

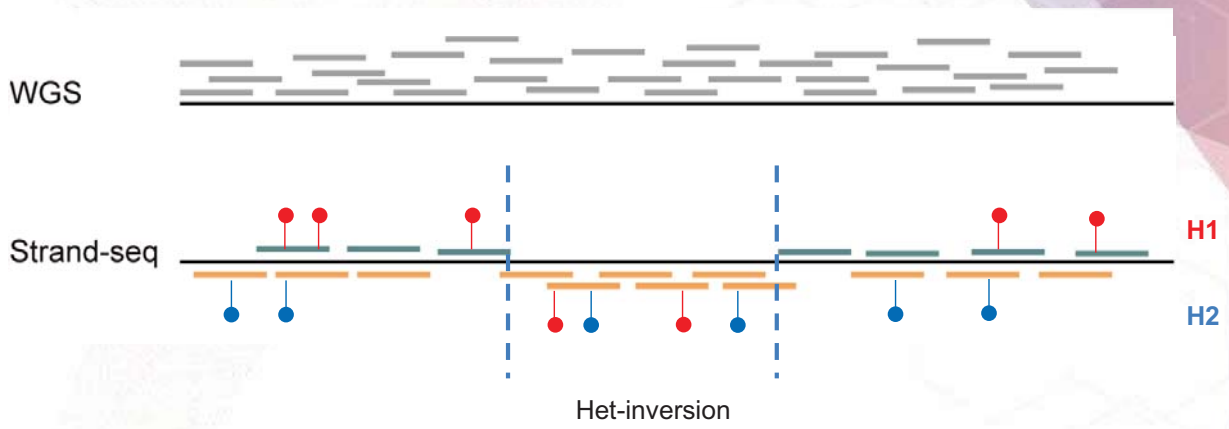
11

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

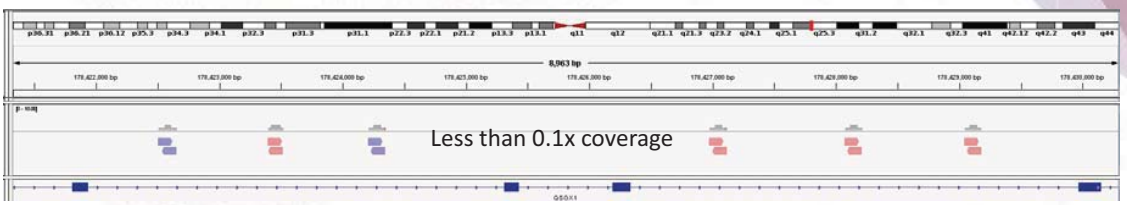


- Sequence orientation is important (Crick or Watson)
- Breakpoint needs to be detected
- Strand state and haplotypes can be assigned
- Multiple types of structural variations need to be classified

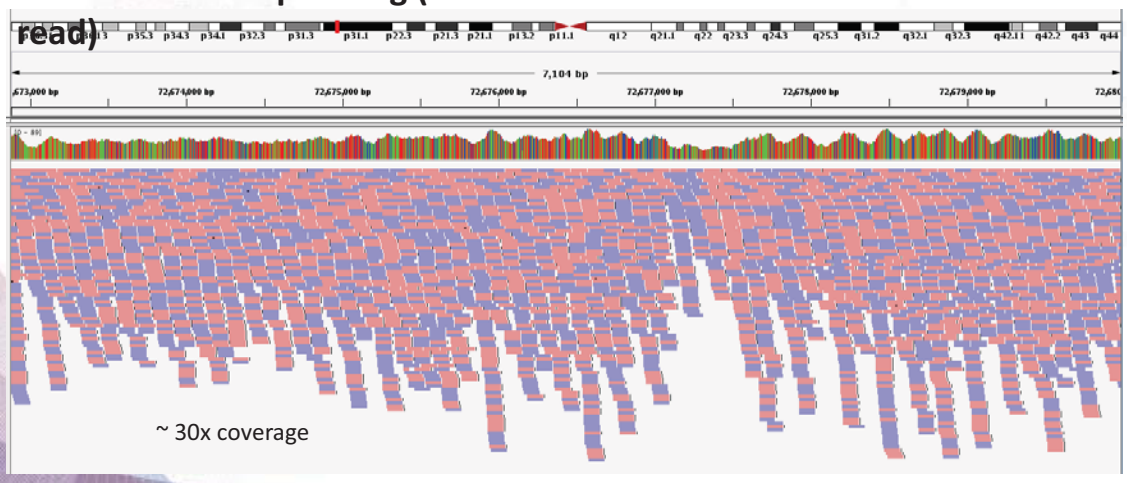
13

Challenges of the Strand-seq data analysis

Strand sequencing



Conventional sequencing (short-read)



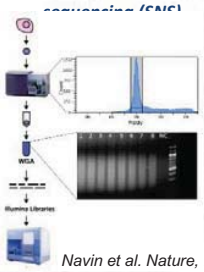
14

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

Single nucleotide variation (SNVs) - scWGS

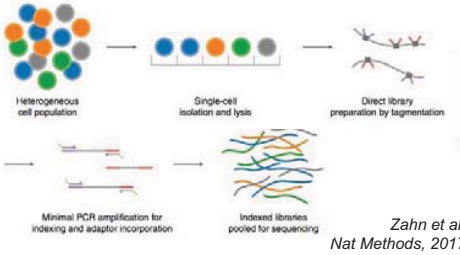
Structural Variation

Single-nucleus



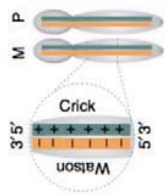
Navin et al. Nature, 2011

Direct library



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!**

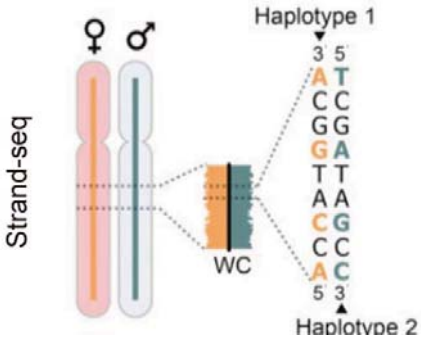
Step3. Genotyping, Haplotyping, Segmentation: StrandPhaseR, breakpointR

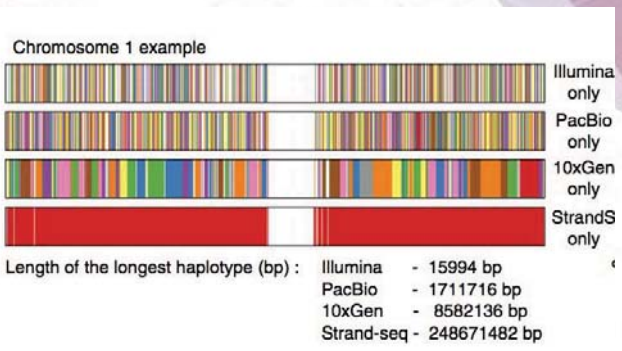
Step4. Structural variation calling: MosaiCatcher

Step5. Single-cell clustering and Phylogenetics

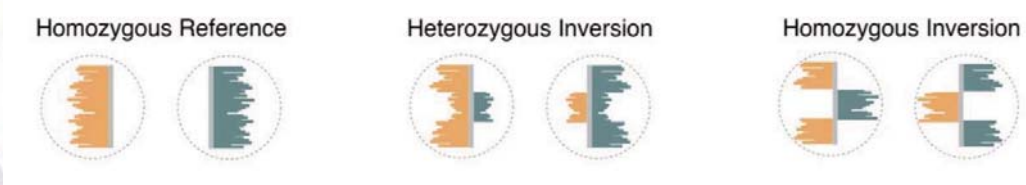
15

Why the orientation of the reads are important?





Porubsky et al. Nat comm, 2017

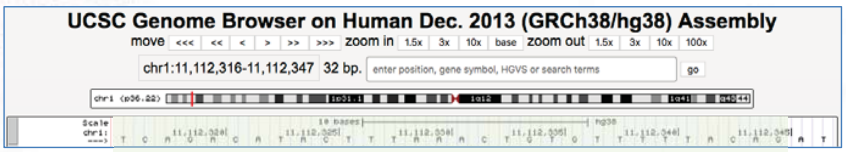


Sanders et al. Genome Res, 2016

16

How can we assign sequencing reads into Crick and Watson?

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



ATACTTT

AAAGTAT

ATACTTT

TC AGACATACTTT A A ACTGTGT T TTT ACAG

AAAGTAT

ATACTTT Forward (+) → Crick (SAMFLAG 0)

AAAGTAT Reverse (-) → Watson (SAMFLAG 16)

17

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:

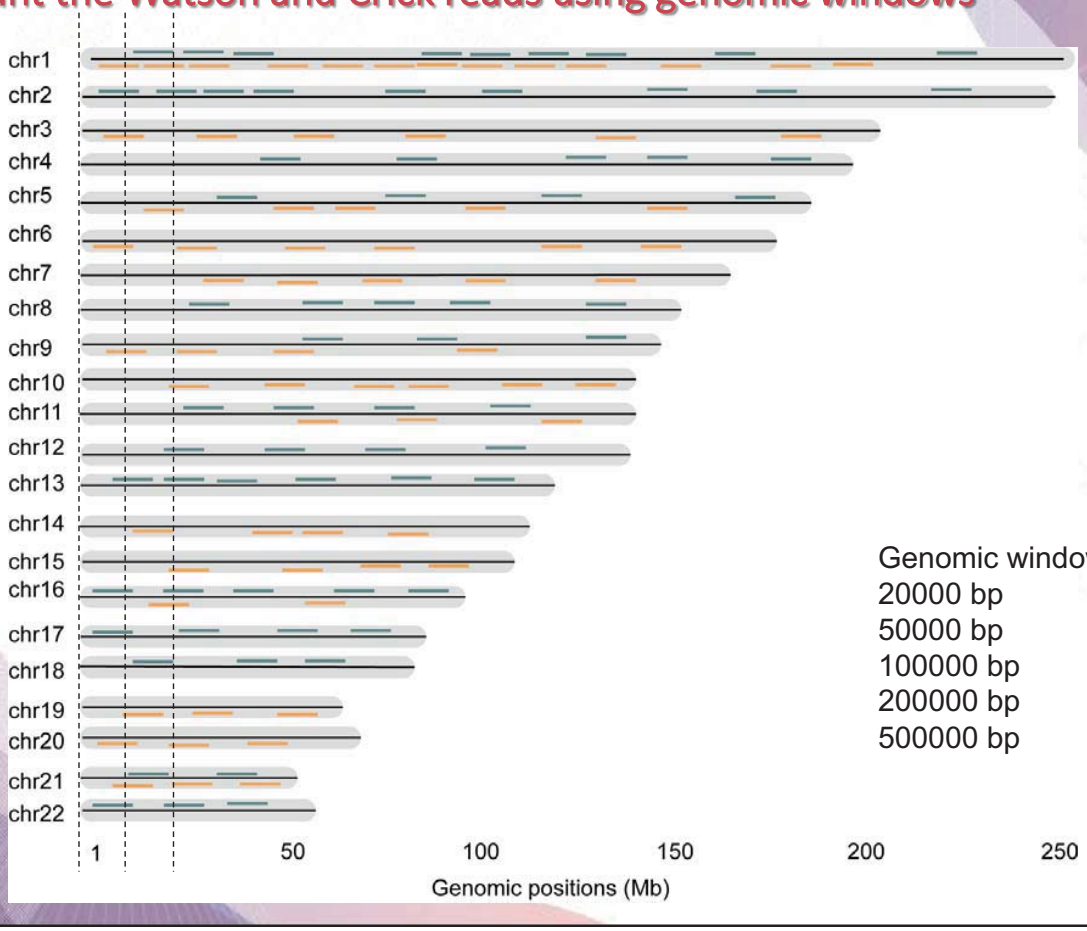
To find out what the SAM flag value would be for a given combination of properties, tick the boxes for those that you'd like to include. The flag value will be shown in the SAM Flag field above.

- [x] read paired
- [x] read mapped in proper pair
- [ ] read unmapped
- [ ] mate unmapped
- [ ] read reverse strand
- [x] mate reverse strand
- [x] 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

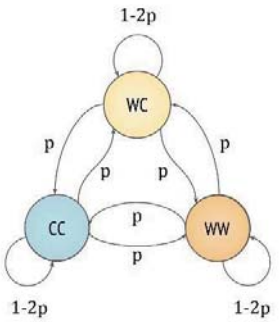
18

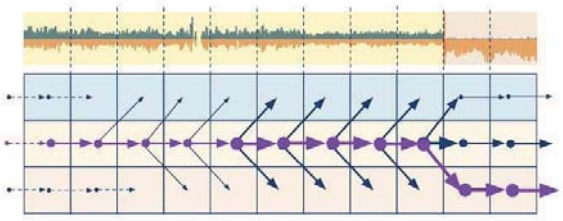
Count the Watson and Crick reads using genomic windows



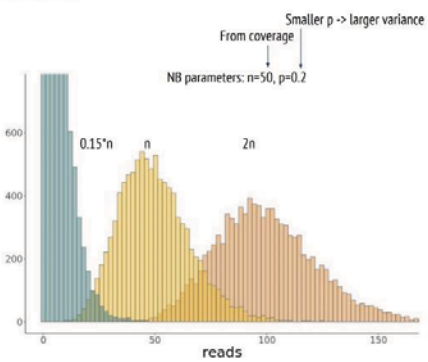
Genomic windows
20000 bp
50000 bp
100000 bp
200000 bp
500000 bp

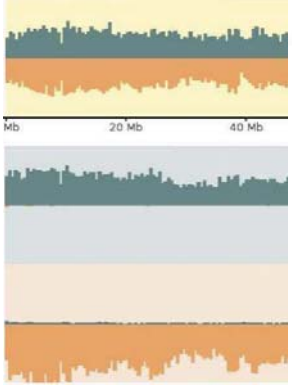
Strand states are called using a Hidden Markov Model





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



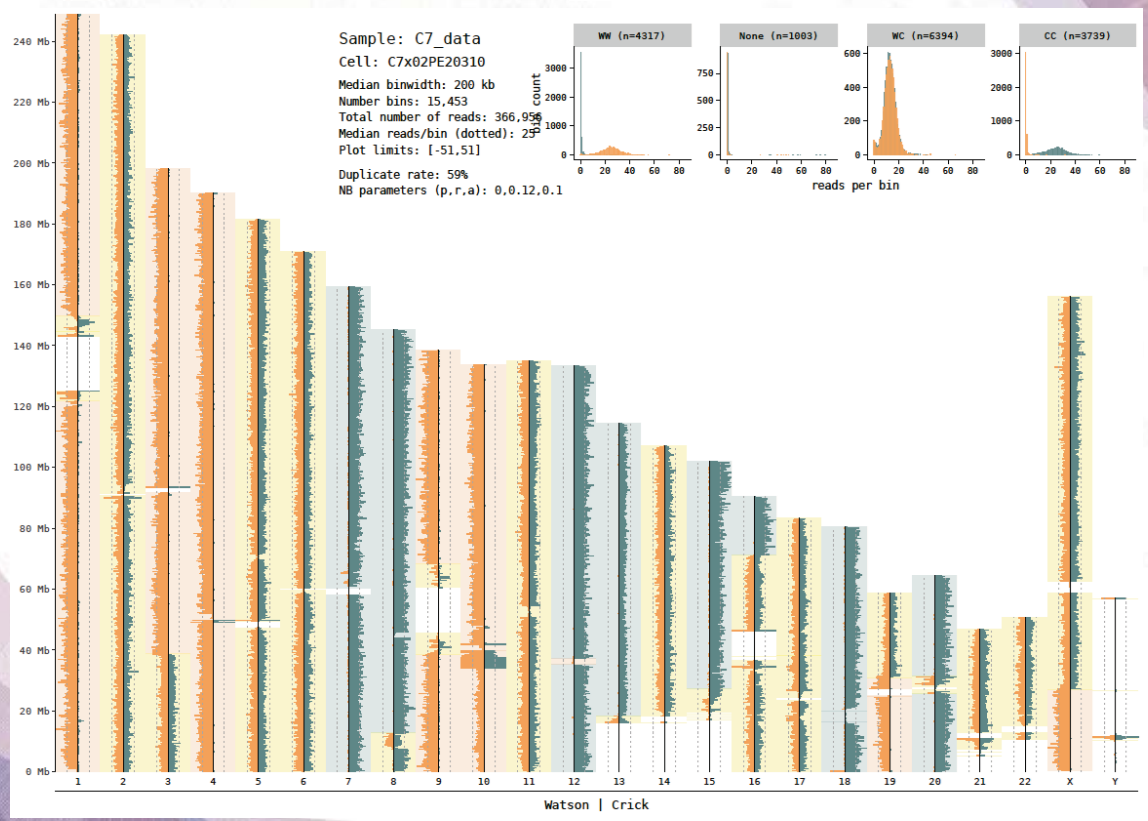


WC

CC

WW

Strand-seq result of example single-cell



21

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

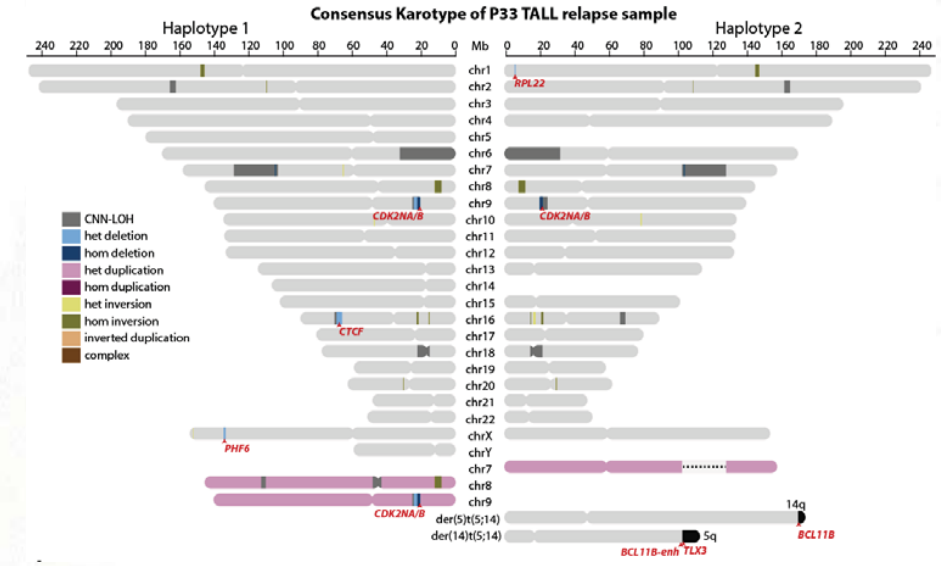


Figure from scTRIP manuscript, Sanders et al. 2020

22

Mosaicatcher towards the automatic single-cell SV calling and clustering

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

README.md

MOSAIC CATCHER

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)

Korbel group, EMBL ***Marschall group, MPI informatics***

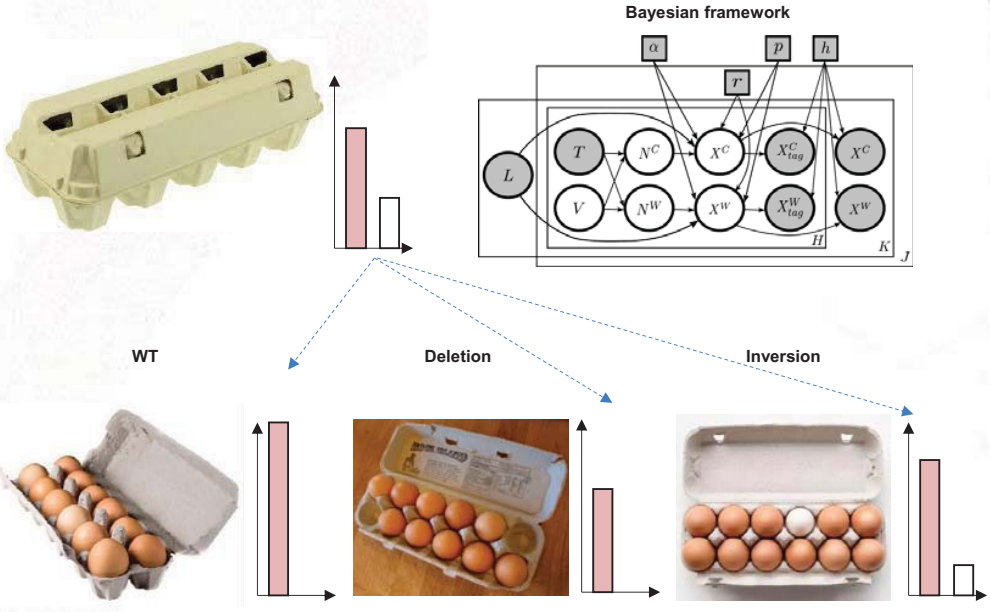
v 1

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

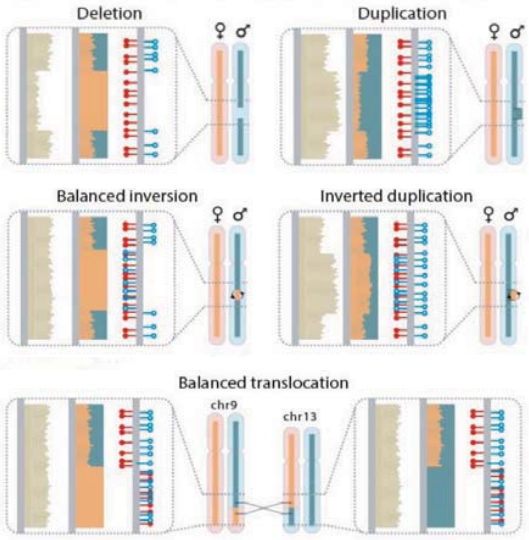
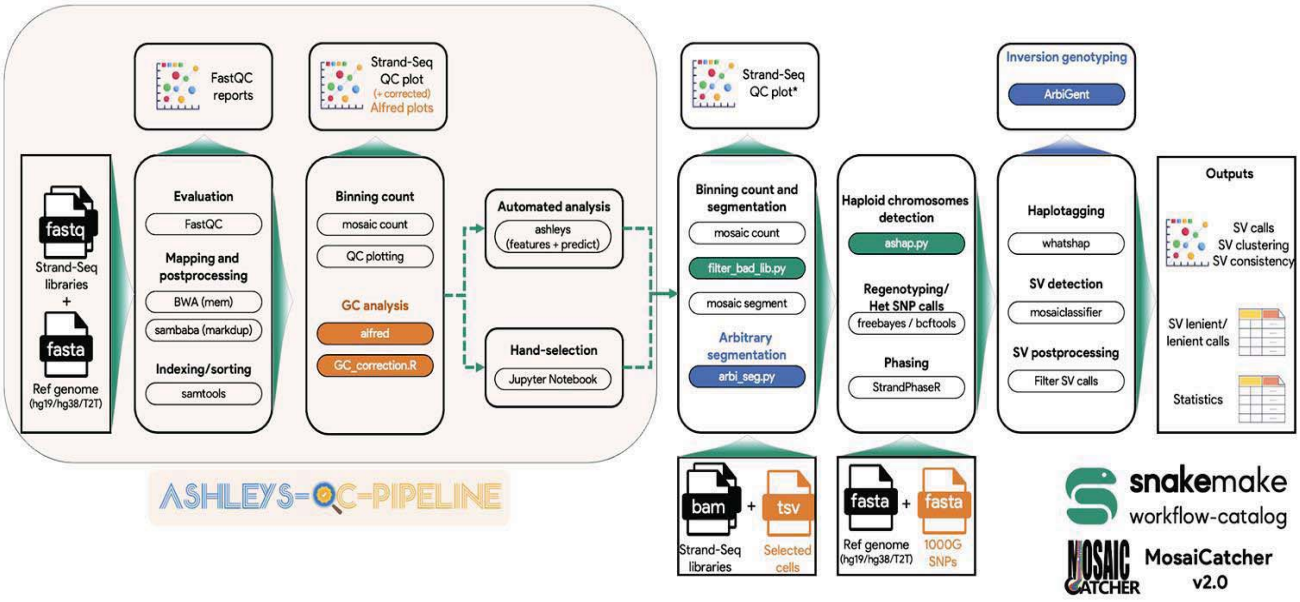


Figure from scTRIP manuscript, Sanders et al.

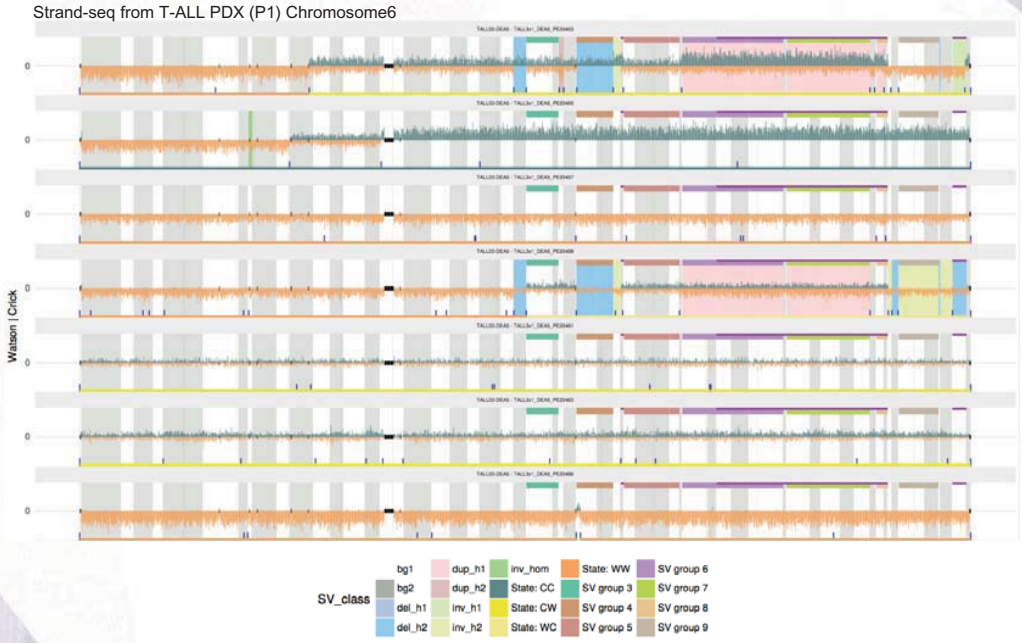
- Input: single-cell BAM files
- Workflow management: Snakemake

- 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

Mosaicatcher calls single-cell SV using Bayesian framework



Chromosome plot with SVs called by MosaiCatcher framework



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

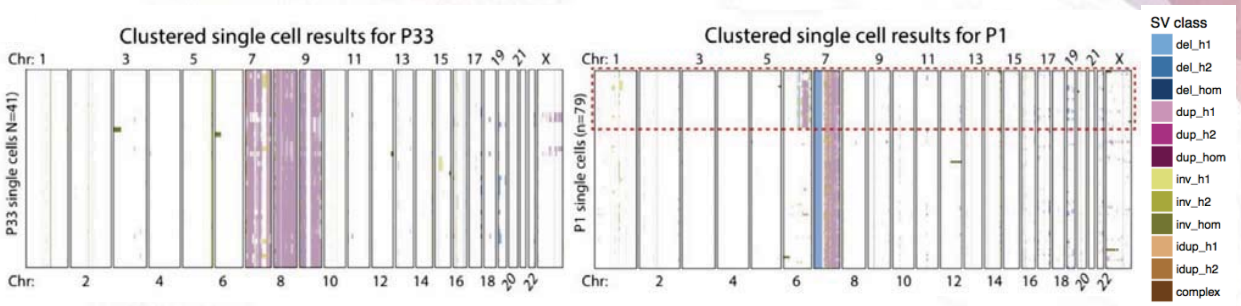
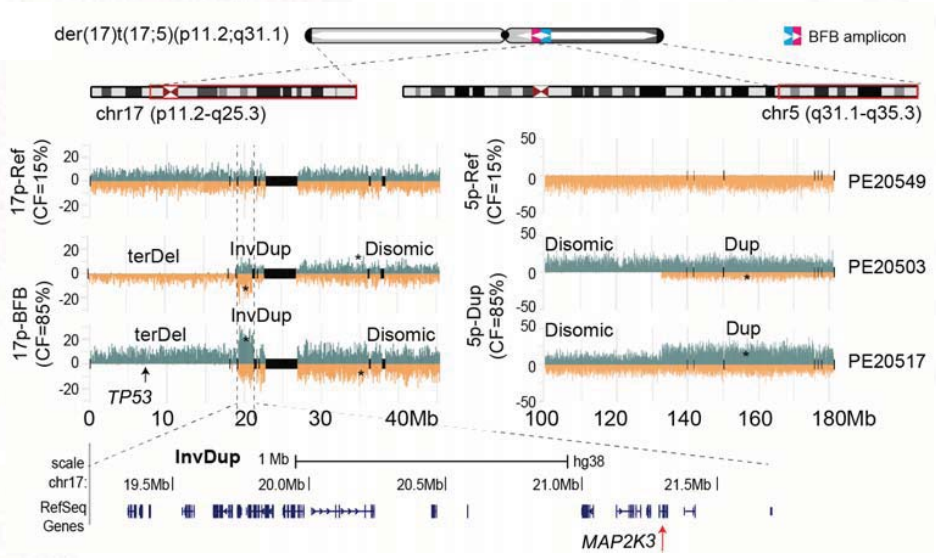


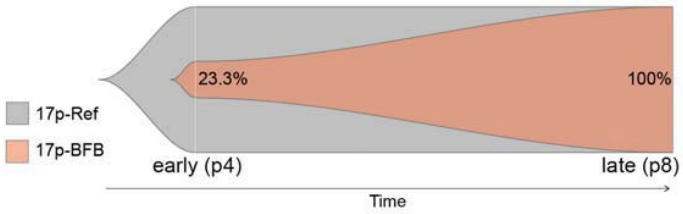
Figure from scTRIP manuscript, Sanders et al. 2019

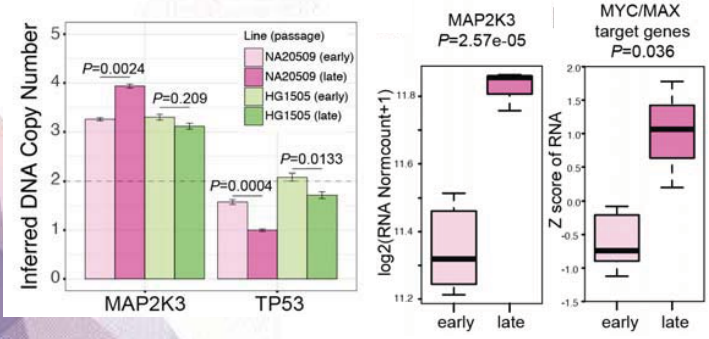
- This heatmap was arranged using Ward's method for hierarchical clustering of SVs genotype likelihoods in two PDX samples
- P33 shows single dominant clone but P1 shows subclonal population in the sample represented by 23 cells

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



Subclonal evolution can be analyzed using Strand-seq





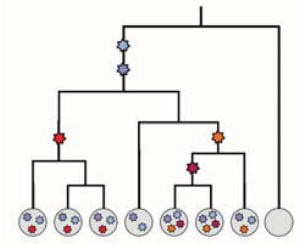
- NA20509 (=GM20509) cell line was in culture for passage 4 (early) and passage 8 (late)
- MAP2K3, and MYC/MAX target genes were increased in late passage
- MYC expression was not changed

SBI 한국생명정보학회 Korean Society for Bioinformatics

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

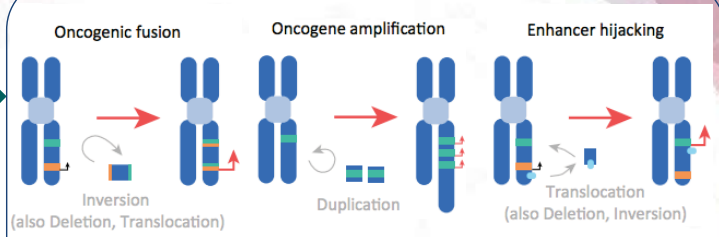
Single-cell multi-omics analysis to study tumor subclones

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



Genome Biology, 2016

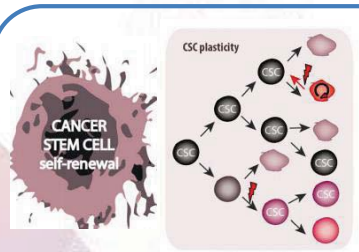
Genetic variation



Macintyre et al. 2016

?

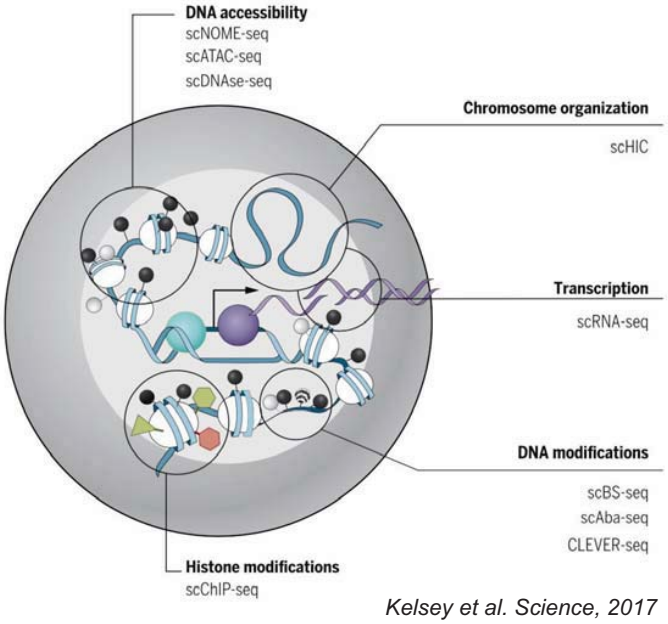
Epigenetic (functional) alteration



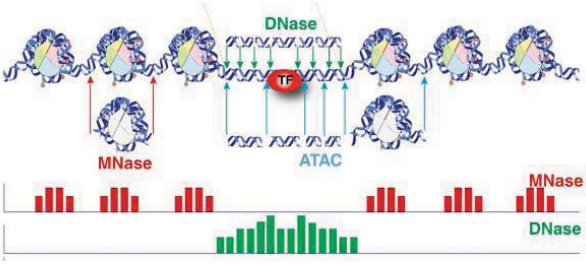
Molecular cancer, 2017

32

Single-cell technologies to explore *functional* heterogeneity



Kelsey et al. Science, 2017



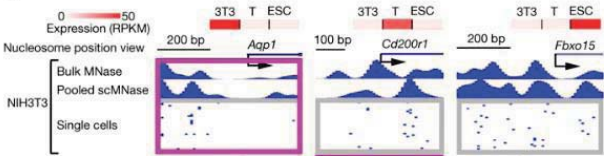
LETTER

scMNase-seq, Lai et al. 2018

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

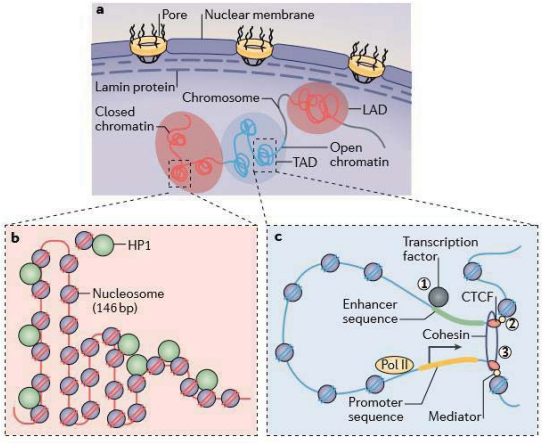
Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing

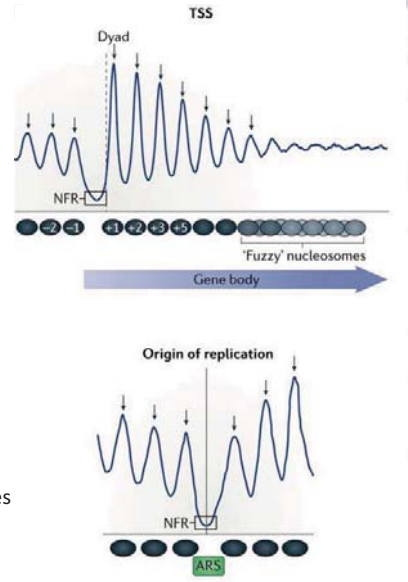
Binbin Lai[1], Weiwu Gao[1,2], Kairong Cui[1], Wanli Xie[1,3], Qingsong Tang[1], Wenfei Jin[4], Gangqing Hu[1], Bing Ni[2] & Keji Zhao[1]*



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

Nucleosomes are the basic unit of chromatin which slide along DNA





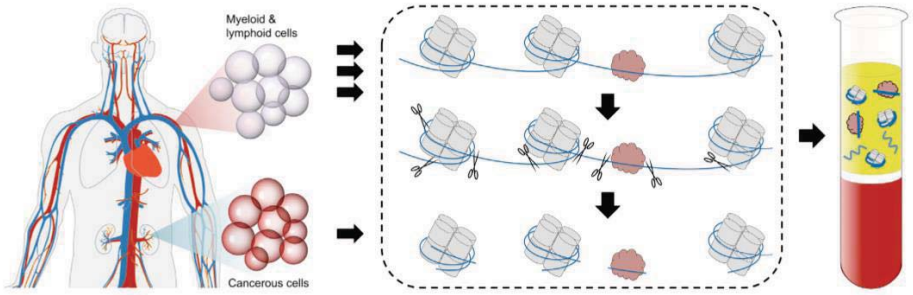
- Nucleosome is composed of two copies of four core histones together with 146~147bp of DNA
- Human diploid genomes have 30 million nucleosomes
- Transcriptionally active gene promoters exhibit a prominent nucleosome-depleted region (NDR) directly upstream of the TSS

Nat Rev Mol Cell Biol, 2017

34

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

Cell free DNA protected by nucleosome is secreted to the blood



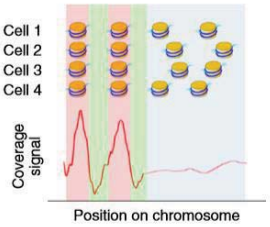
Nucleosomes pattern is informative for the gene expression and cell type of origin

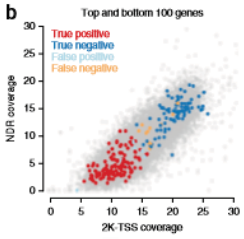
LETTERS
nature genetics

Inferring expressed genes by whole-genome sequencing of plasma DNA

Peter Ulz[1], Gerhard G Thallinger[1,3], 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





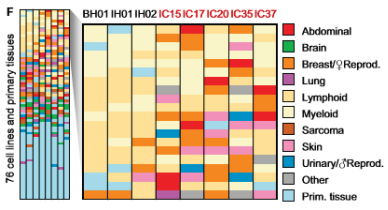
Article
Cell

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

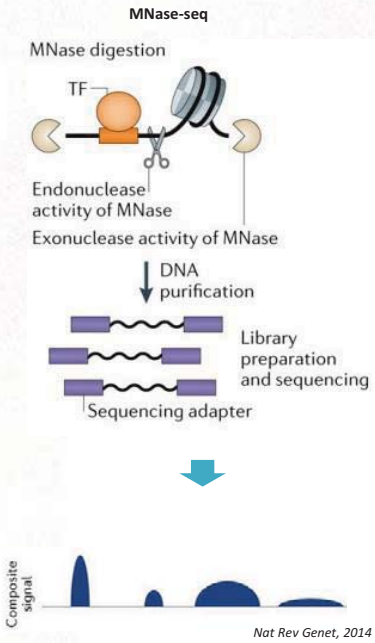
Cell, 2016

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

Correspondence
shendure@uw.edu

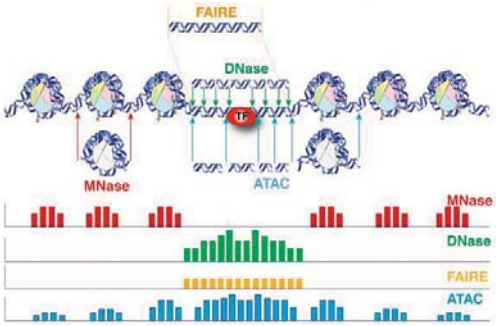


Nucleosome dynamics can be measured by genomic assays



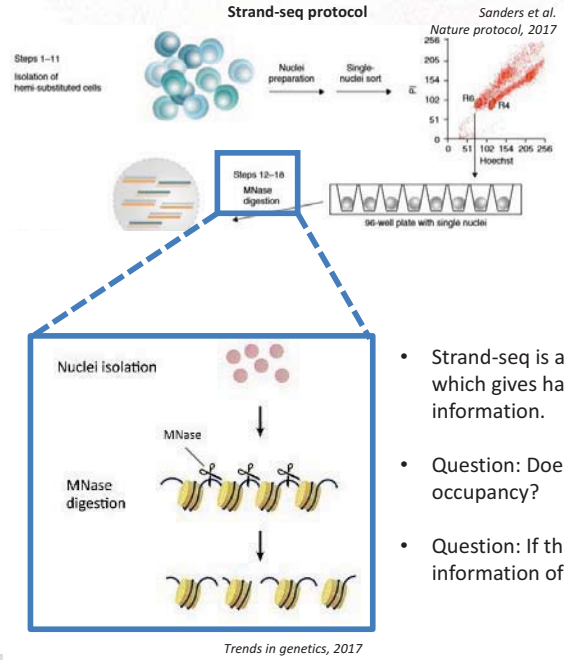
Nat Rev Genet, 2014

- 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

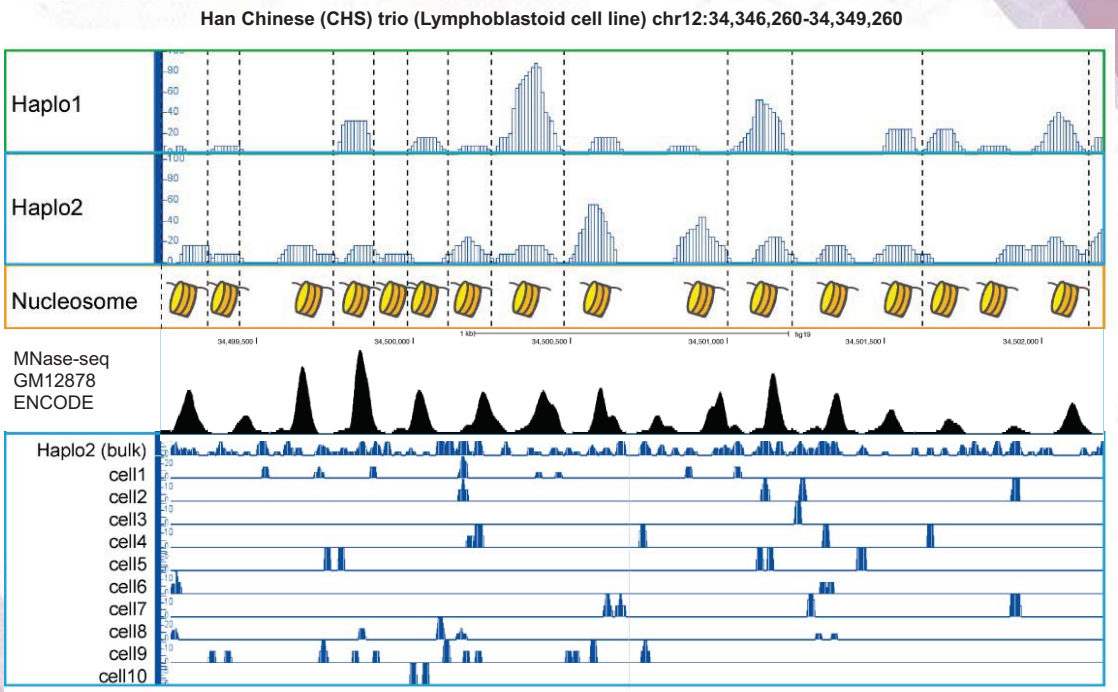
Strand-seq protocol involves MNase treatment



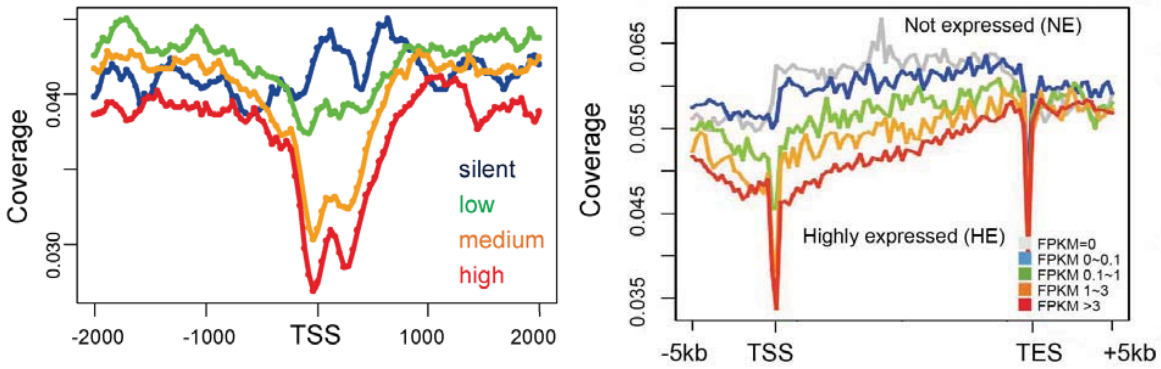
Trends in genetics, 2017

- 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



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

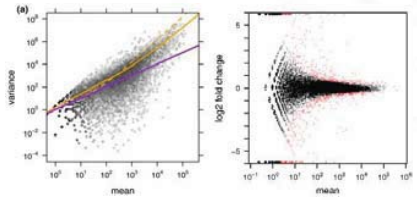
19770 genes

LCL (224 cells)

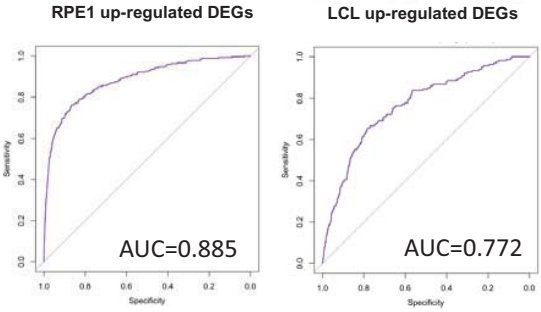
	cell1	cell2	...	cell N
Gene1	1	2	...	1
Gene2	8	4	...	5
...	...	...	...	...
Gene N	14	25	...	10

19770 genes

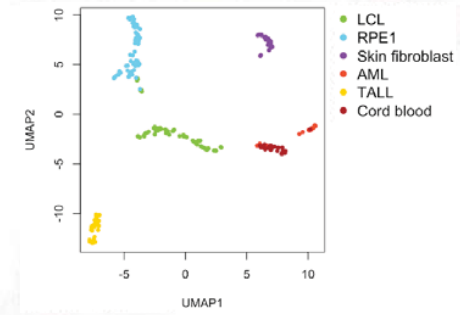
Approach (DESeq of nucleosome occupancy)

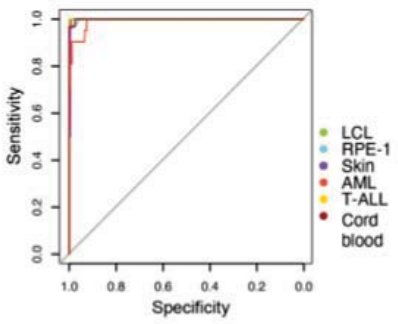


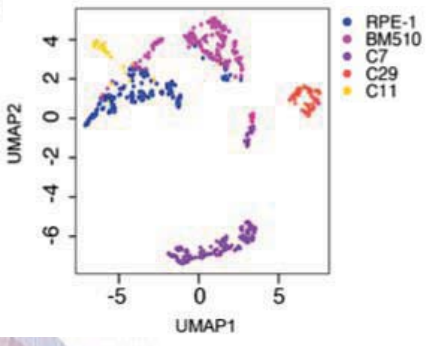
Anders et al. 2010, Love et al. 2014

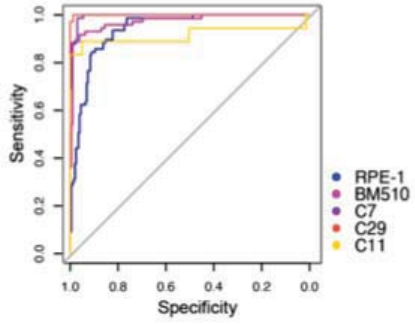


Nucleosome occupancy can be used to classify cell-type

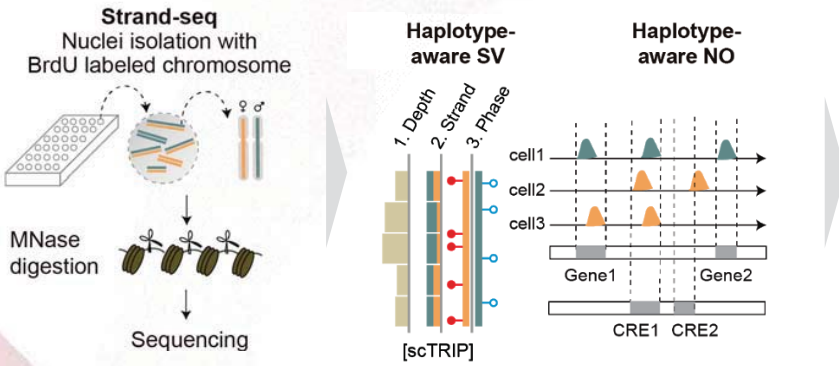






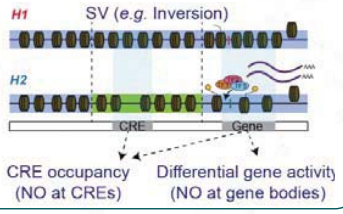


scNOVA : Coupling genome-epigenome using Strand-seq technology

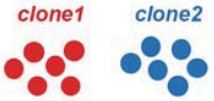


Jeong and Grimes* et al…. Sanders and Korbel Nature Biotech, 2022*

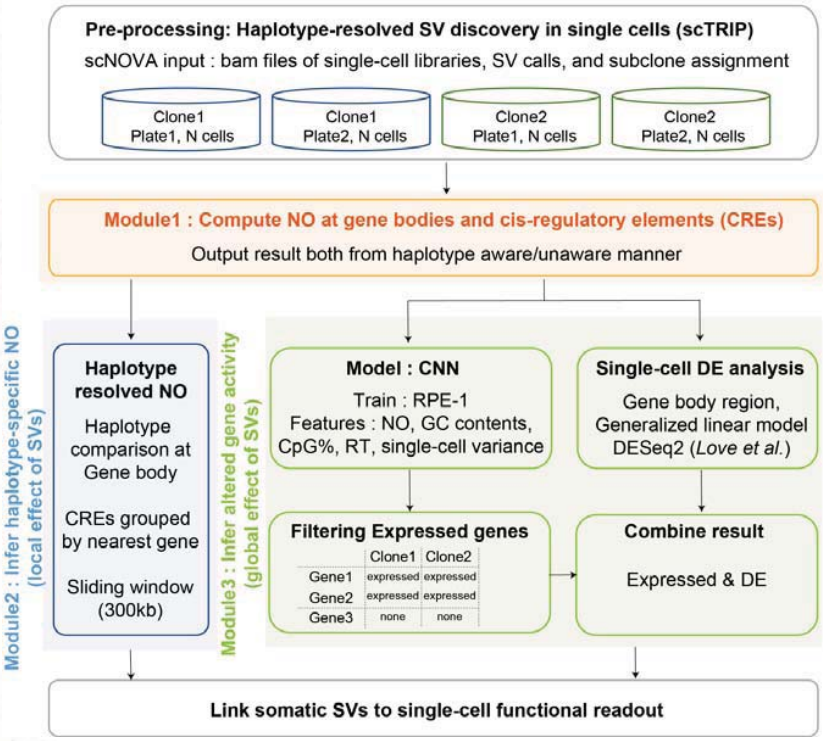
Haplotype specific NO (local/cis-effect of SVs)



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



Computational pipeline of scNOVA



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

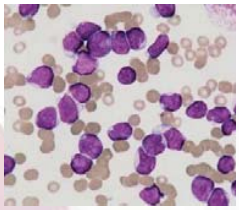
https://github.com/jeongdo801/scNOVA

How subclonal SVs alter the epigenome and phenotype?

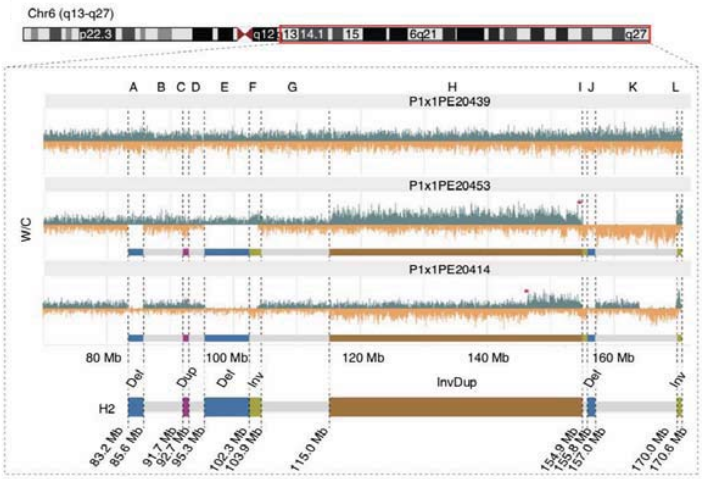
System

T-ALL P1

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



SVs



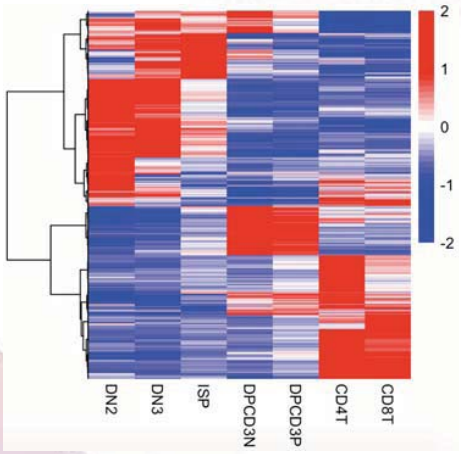
Major clone (68%)

Subclone (32%) (only appeared in relapse)

Sanders et al. 2020

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

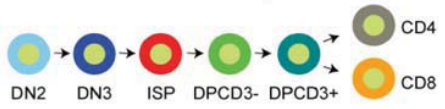
ATAC-seq signature matrix (2020 peaks)

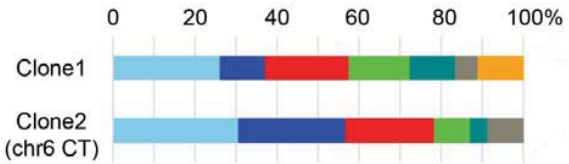


ATAC-seq and signature matrix from Erarslan-Uysal et al. EMBO Mol Med, 2020

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

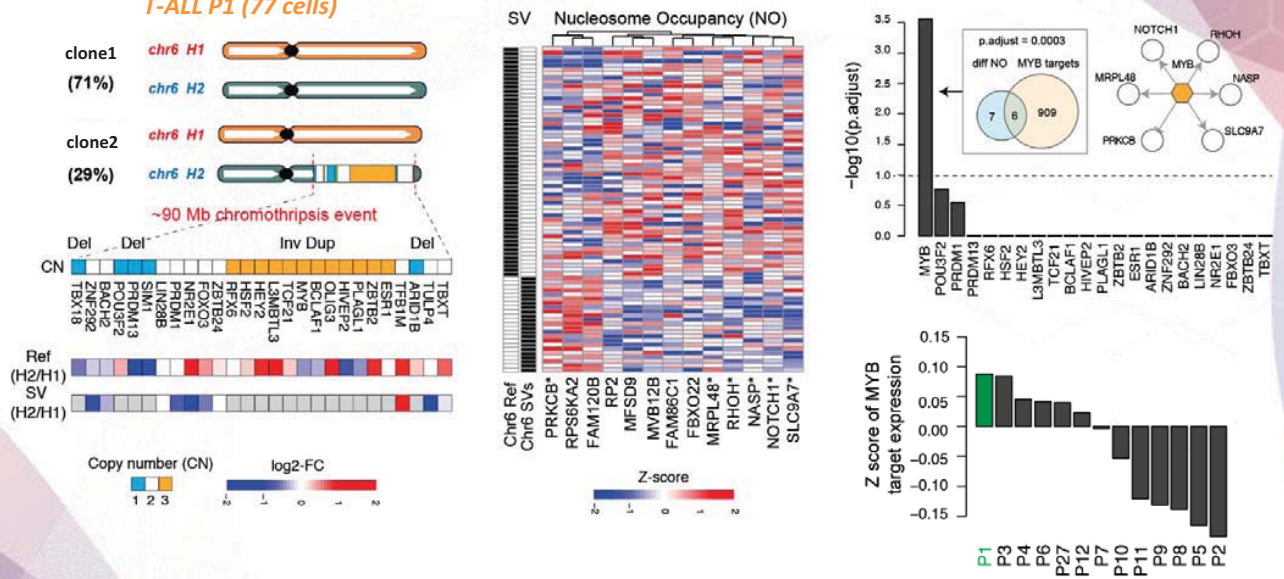
T cell differentiation stages





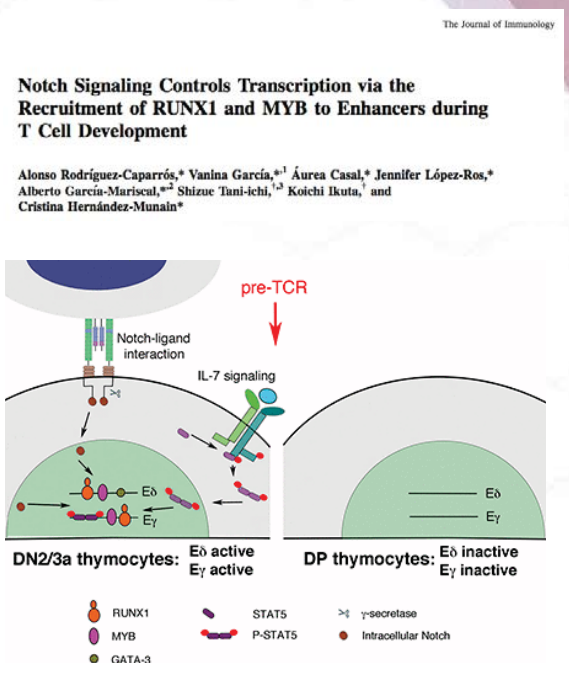
DN3 : 11% -> 26% CD8 : 11% -> 0%

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

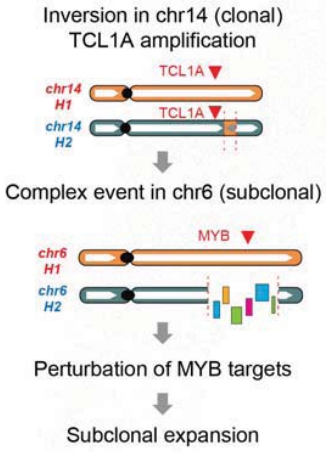


How the cell type (state) composition different in clone1 and clone2?

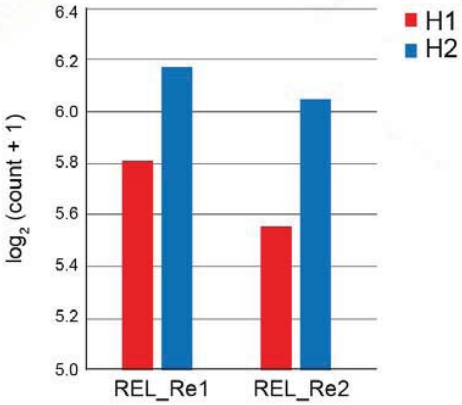
Notch signaling and MYB has been reported in T-ALL oncogenesis



Validation of increased dosage of MYB expression in rearranged haplotype



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

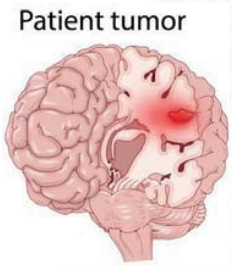
SBI 한국생명정보학회

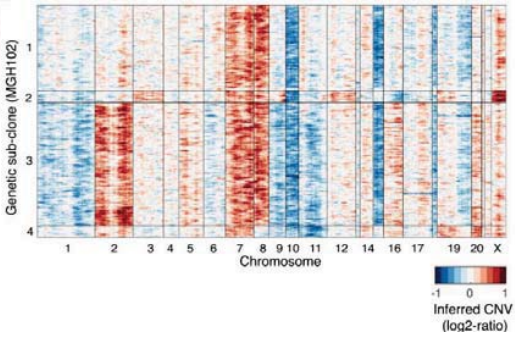
Part4. scRNA-seq에서 서브클론을 유추하고 기능적으로 분석하는 멀티오믹스 기법 소개

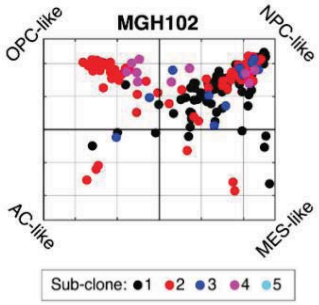
Single-cell multi-omics analysis to study tumor subclones

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

Infer CNV from single-cell RNA-seq



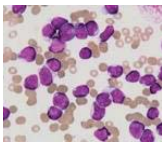




Neftel et al. Cell, 2019

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

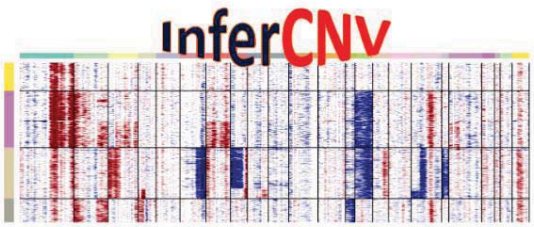
SCNA inference methods based on transcriptome



	Method	Method detail: 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
Discovery	HoneyBADGER (Genome Res, 2018)	CNV only	N	10Mb	N
Discovery	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

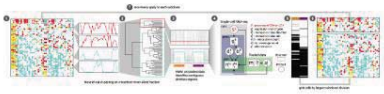
build passing

HoneyBADGER (**h**idden Markov model integrated **B**ayesian **a**pproach for **d**etecting CNV and LOH events from sin**g**le-ce**l**l **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.

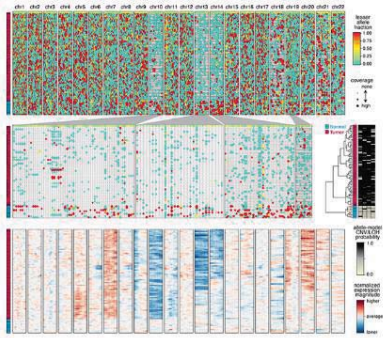
The overall approach is detailed in the following publication:
Fan J*, Lee HO*, Lee S, et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. Genome Res. 2018;

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



https://jef.works/HoneyBADGER/

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

README.md

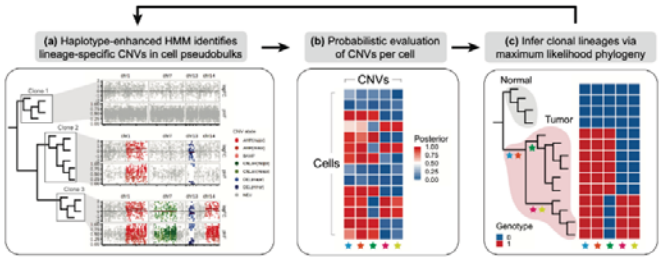
Numbat

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 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 55

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

CONICS

CONICS: COpy-Number analysis In single-Cell RNA-Sequencing

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 !

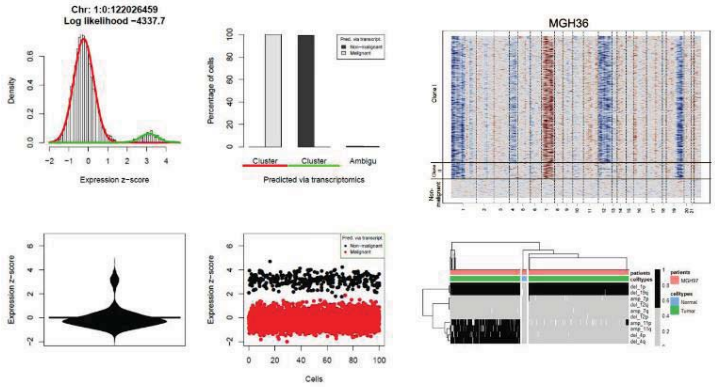
Table of contents

- 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

https://github.com/diazlab/CONICS

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].

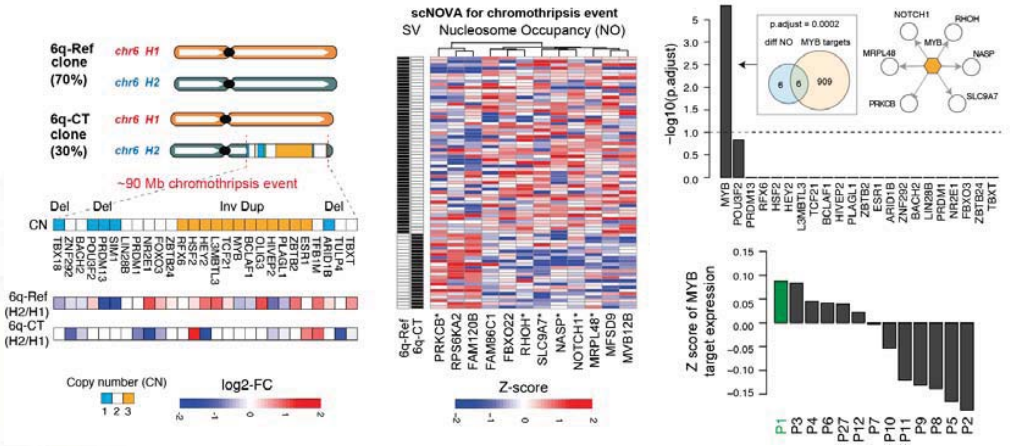


Visualizations of scRNA-seq data from Oligodendroglioma (Tirosh et al., 2016) generated with CONICSmat.

56

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

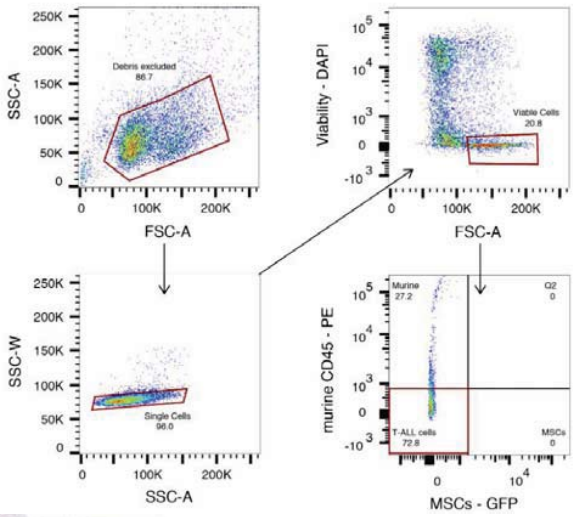
Hypothesis : 6q-CT cells have MYB-Notch activation compared to 6-Ref cells



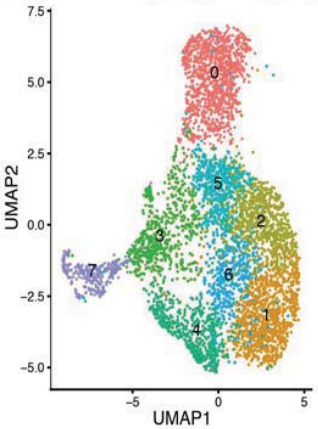
Single-cell experiment is needed to confirm subclonal level transcriptome changes

Applying CONICSmat to the T-ALL case study

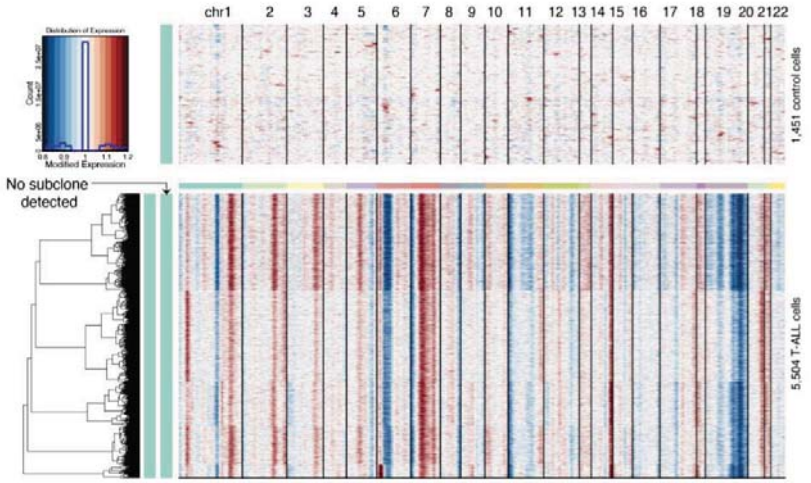
Gating strategy for single, viable T-ALL cell isolation from T-ALL sample T-ALL_P1 for scRNA-seq.



Unsupervised analysis of transcriptome



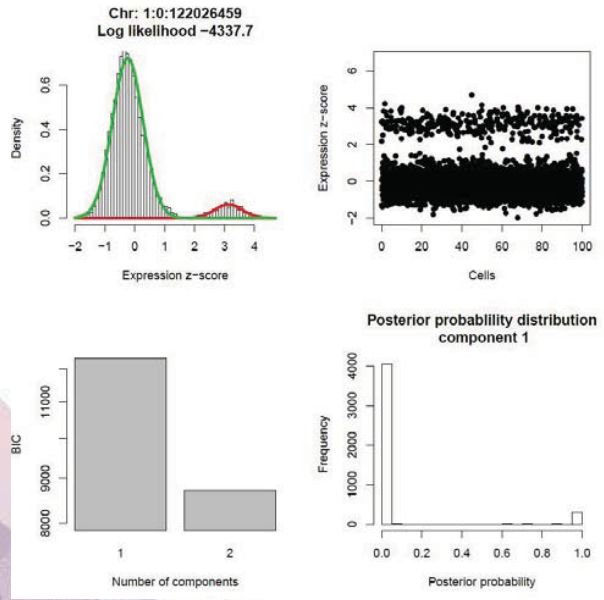
Applying InferCNV to the T-ALL case study



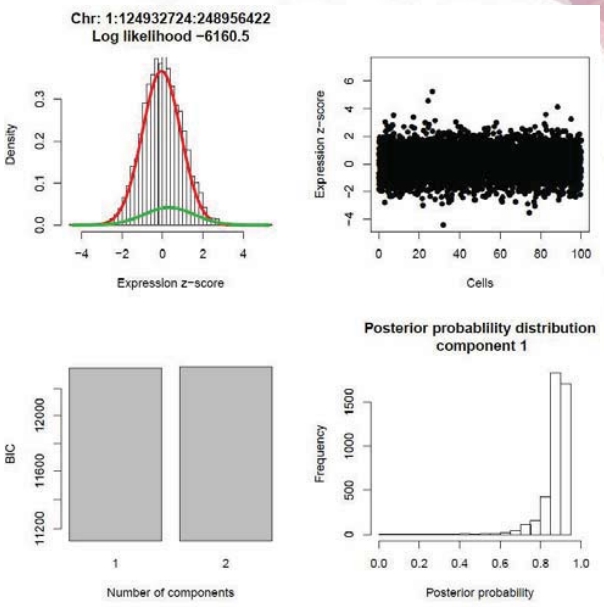
InferCNV analysis of 5,504 high quality T-ALL_P1 cells, and 1,451 control cells. Control cells were downloaded from PBMC data provided by 10X Genomics. This analysis did not discover subclones in 5,504 T-ALL cells.

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

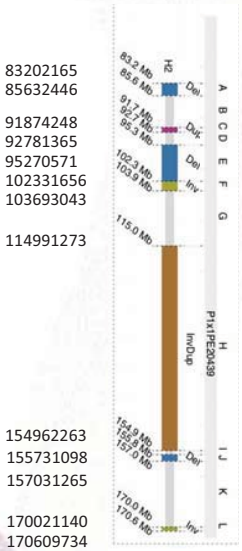
Presence of Subclonal CNA



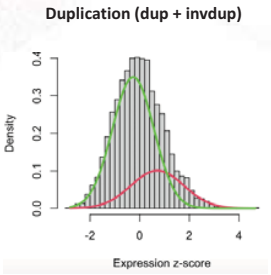
Absence of Subclonal CNA

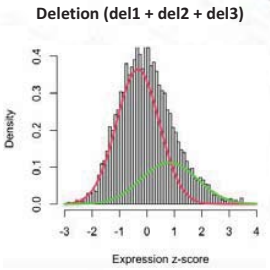


CONICSmat analysis supports the presence of chr6 deletions and duplications



Sanders et al. 2020





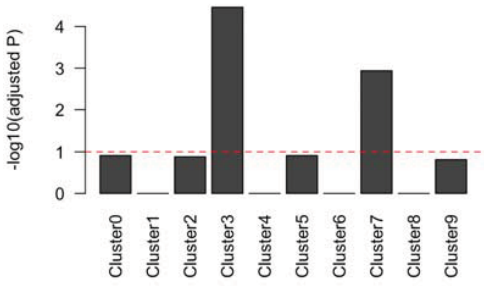
Genomic range	BIC 1 component	BIC 2 components	BIC difference	LRT adj. p-val	CNV call (CF%)
chr6_Del 35 genes	15635.9018	15553.1101	82.7917307	0	729 cells (13.2%)
chr6_Dup 192 genes	15635.9018	15481.839	154.062863	0	265 cells (4.8%)

10X transcriptome experiment from Karen Grimes

61

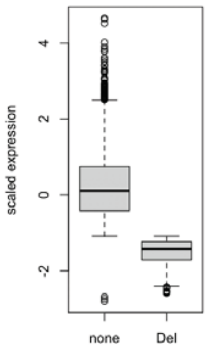
Cluster3 and Cluster7 cells are highly enriched with deletion calls

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

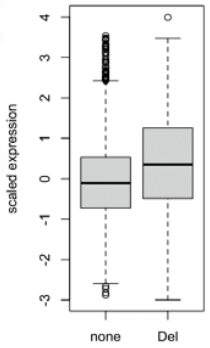


Row Labels	Del(0.9)	%Del(all)	p-value	adjustedP
Cluster0	166	14.901	0.039	0.122
Cluster1	97	9.454	1.000	1.000
Cluster2	118	15.013	0.066	0.131
Cluster3	118	19.440	0.000	0.000
Cluster4	40	6.981	1.000	1.000
Cluster5	83	15.690	0.049	0.122
Cluster6	32	6.695	1.000	1.000
Cluster7	62	20.395	0.000	0.001
Cluster8	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



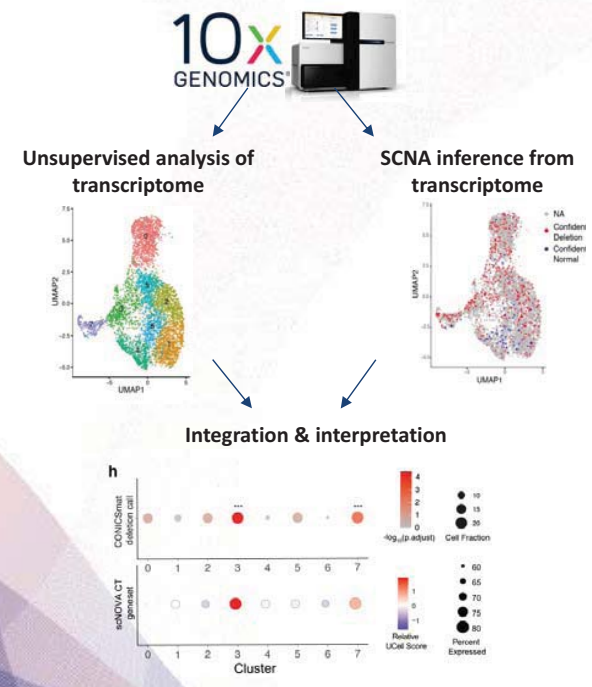
Type	Dup	none
Del	96	633
none	169	4606

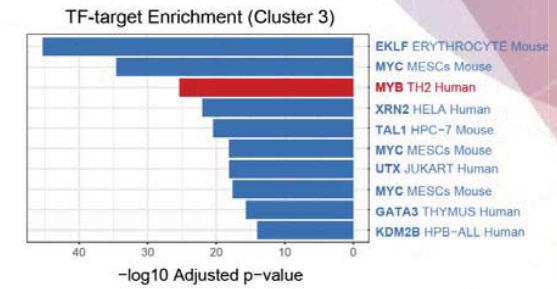
Fisher exact test

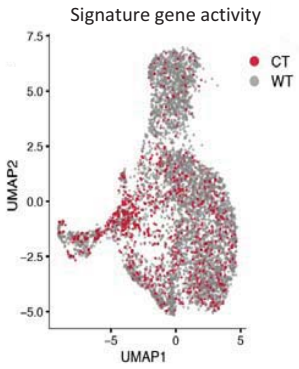
p-value < 2.2e-16

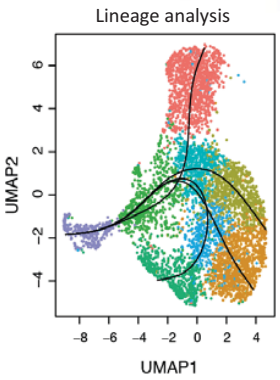
62

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









63

SBI 한국생명정보학회 Korean Society for Bioinformatics

Part5. 실제 암 샘플 분석에의 적용

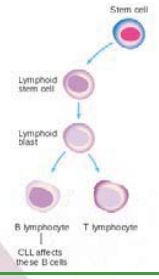
Single-cell multi-omics analysis to study tumor subclones

Case study in Chronic lymphocytic leukemia

System

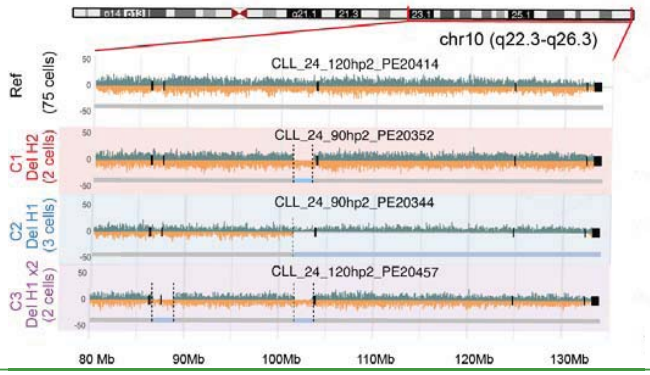
B-CLL_24

Peter-Martin Bruch
Sascha Dietrich group

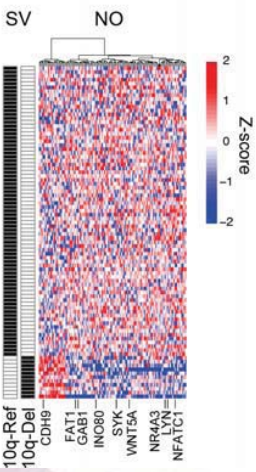


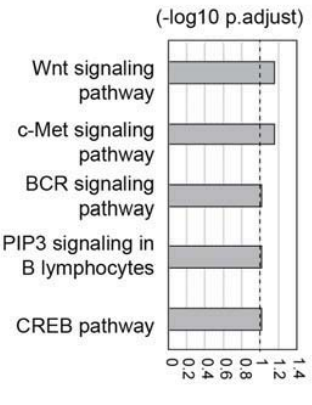
SVs

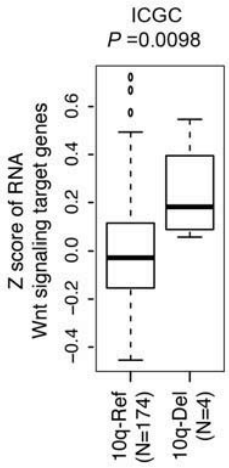
Subclonal deletions in chr10q

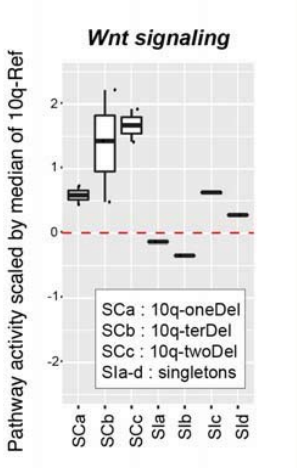


scNOVA inferred that 10q-Del clones have aberrant Wnt signaling activity

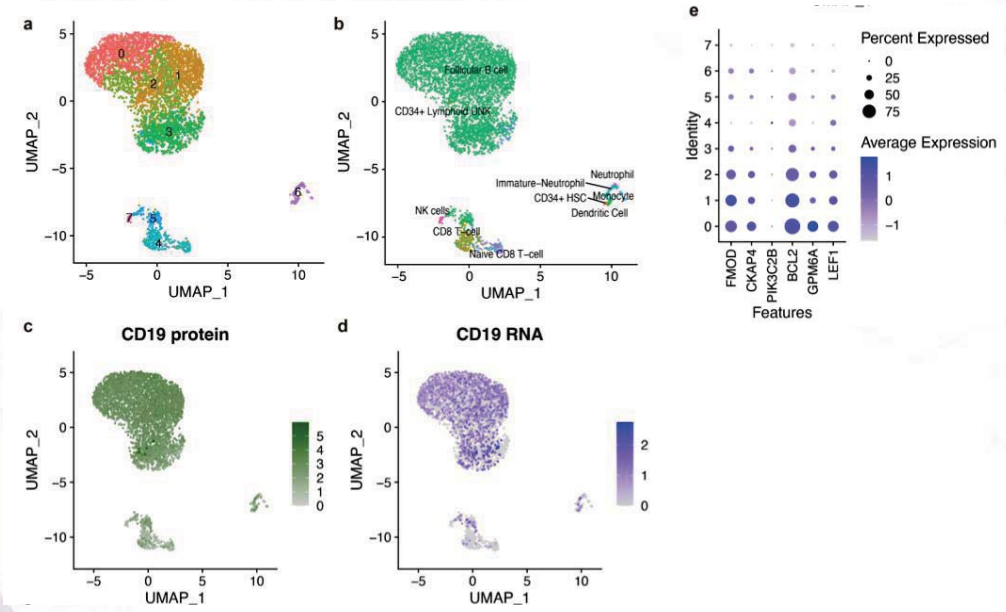




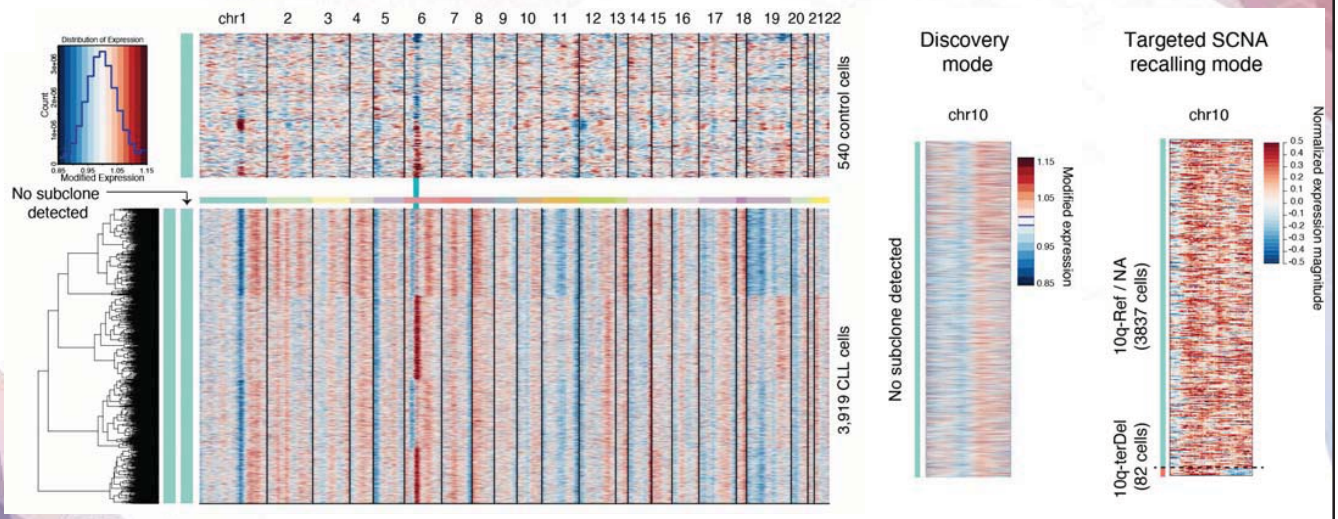




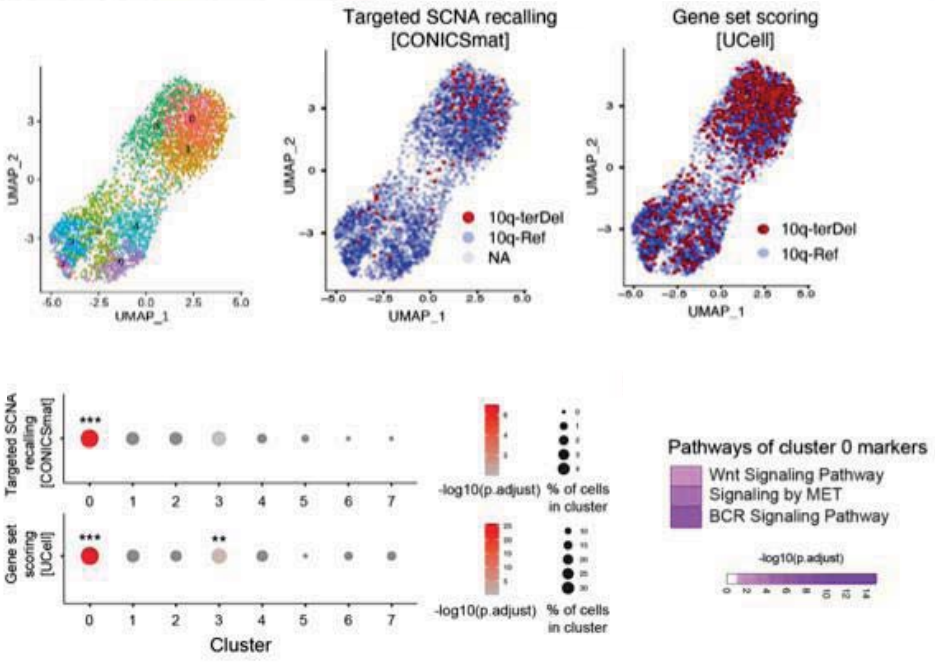
Single-cell transcriptome analysis of CLL_24 (CITE-seq)



Inference of 10q-terDel cells in CITE-seq of CLL_24

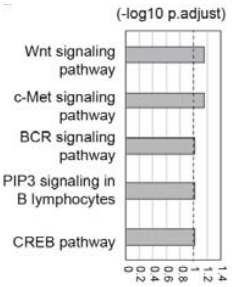


Inference of 10q-terDel cells in CITE-seq of CLL_24

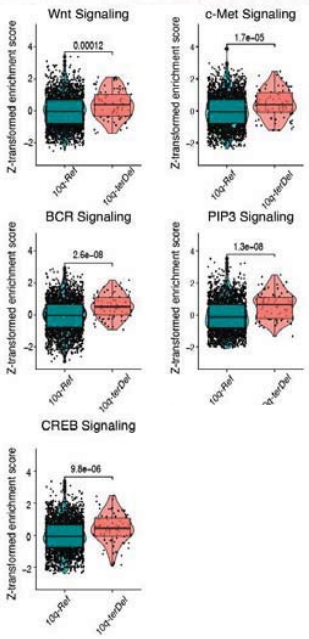


10q-terDel cells from CITE-seq shows Wnt activation and PD-1 overexpression

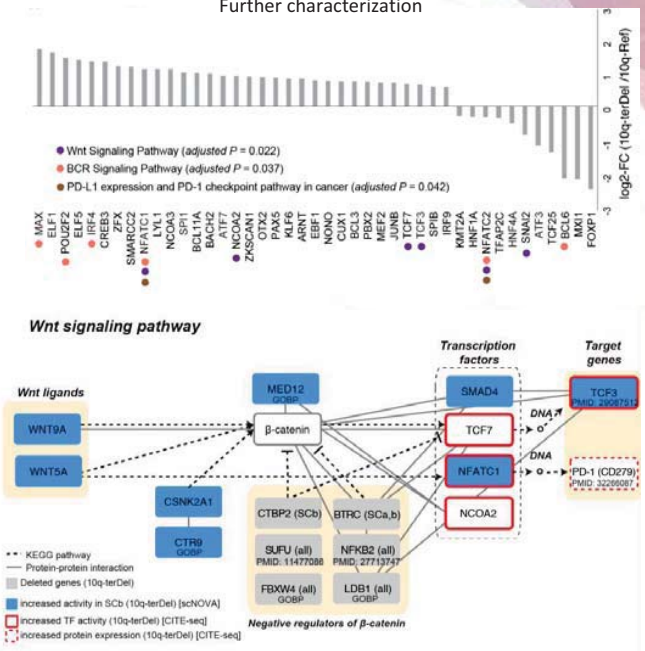
scNOVA



validation



Further characterization



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

Summary

- 암에서 이러한 서브클론들을 동정하기 위해 어떤 싱글셀 오믹스 기법들이 개발되어 있을까? → *single-cell WGS, 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.)*

71

SBI 한국생명정보학회 Korean Society for Bioinformatics

감사합니다.