

Cancer genomics: one cell at a time

Yuchao Jiang

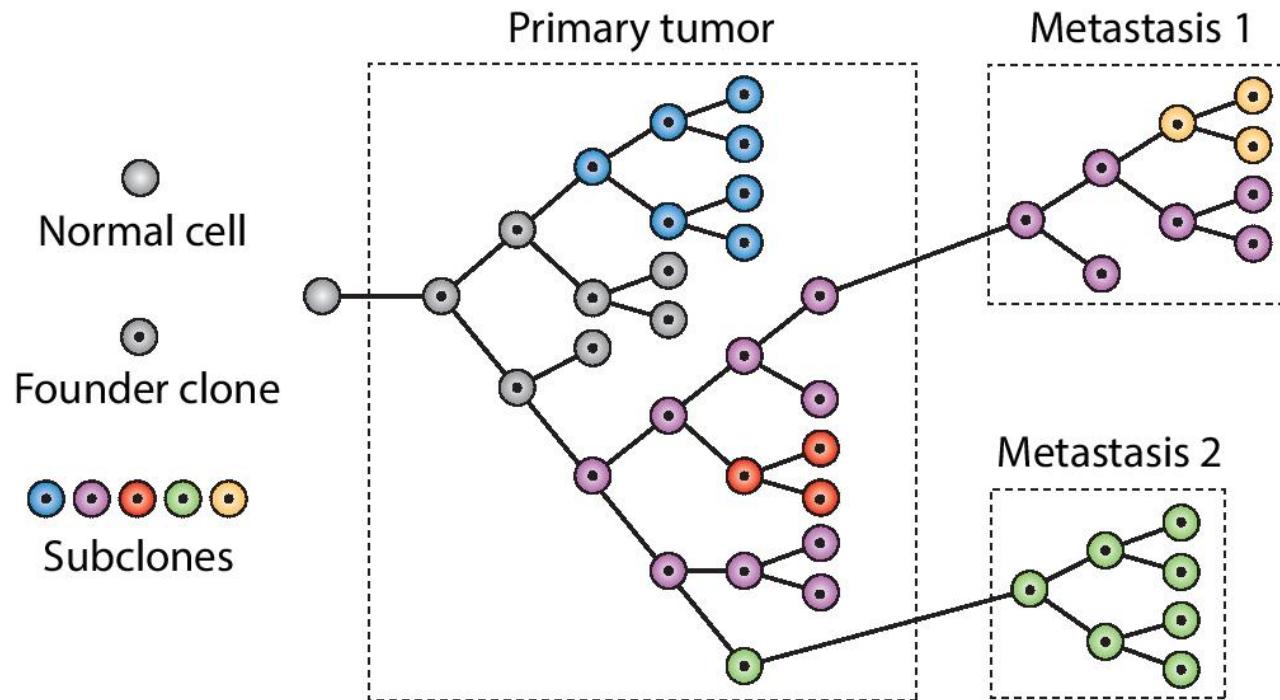
UNC Chapel Hill

ISMB 2019

July 21st, 2019

Background: tumor progression

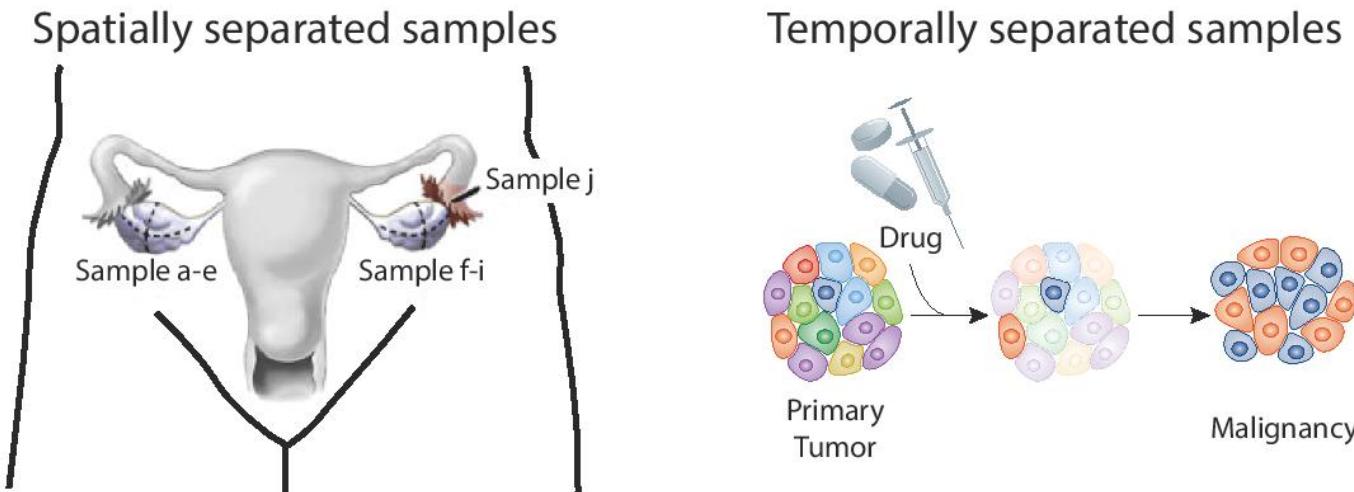
- ▶ Follows Darwinian evolution



- ▶ **Clones:** genotypically distinct cell populations
- ▶ **Longitudinal and spatial subclonal dynamics** lead to failures of targeted therapies and disease recurrence

Tracking clonal evolution by bulk sequencing

- ▶ Study design:
repeated sequencing of **temporally and/or spatially separated tumor specimens** from the same patient.



Canopy for repeated bulk DNA sequencing

<https://CRAN.R-project.org/package=Canopy>

<https://github.com/yuchaojiang/Canopy>

The figure illustrates the workflow for phylogenetic analysis of SNA-CNA data. It starts with 'Input' data, which includes observed MAF (R/X), observed major copy (W^M), observed minor copy (W^m), and SNA-CNA overlap (γ). These inputs are converted into a 'Matrix representation' (Normal and Integer matrices for SNA carrier status and clone frequencies) and a 'SNA-CNA phase' matrix (H). Finally, the data is used to generate a phylogenetic tree ('Phylogeny (output)') and a clonal frequency matrix (P).

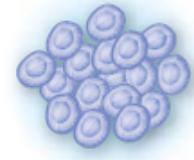
Input			Matrix representation			Phylogeny (output)		
Observed MAF (R/X)	0.15 0.16 0.52 0.34	0.26 0.00 0.48 0.24	0.00 0.30 0.52 0.22	SNA1 SNA2 SNA3 SNA4	Normal Clone1, 2, 3 SNA carrier status (Z)	[0 0 0 1] [0 0 1 0] [0 1 1 1] [0 0 1 1]	SNA1 SNA2 SNA3 SNA4	
Observed major copy (W^M)	1.64 1.01 1.60	1.49 0.98 1.54	1.51 1.04 1.48	CNA1 CNA2 CNA3	Normal Clone1, 2, 3 Integer major copy (\tilde{C}^M)	[1 1 2 2] [1 1 1 1] [1 1 2 2]	CNA1 CNA2 CNA3	
Observed minor copy (W^m)	0.37 0.78 0.97	0.50 0.82 0.99	0.47 0.67 1.00	CNA1 CNA2 CNA3	Normal Clone1, 2, 3 Integer minor copy (\tilde{C}^m)	[1 1 0 0] [1 0 1 1] [1 1 1 1]	CNA1 CNA2 CNA3	
SNA-CNA overlap (γ)	non-CNA [0 1 0 0] [0 0 1 0] [0 0 0 1] [1 0 0 0]	CNA1, 2, 3 SNA1 SNA2 SNA3 SNA4	SNA-CNA phase (H)	SNA1 SNA2 SNA3 SNA4	[0 0 1 0]			
						Clonal frequency matrix (P)	Sample 1 Sample 2 Sample 3	

(Jiang et al. 2016 PNAS)

Tumor heterogeneity: from bulk tissue to single cells

- **Repeated bulk-tissue DNA sequencing:** identify clonal mixtures and recover clonal history through deconvolution.
- **Single-cell RNA sequencing:** study cellular heterogeneity on expression level with single-cell resolution.

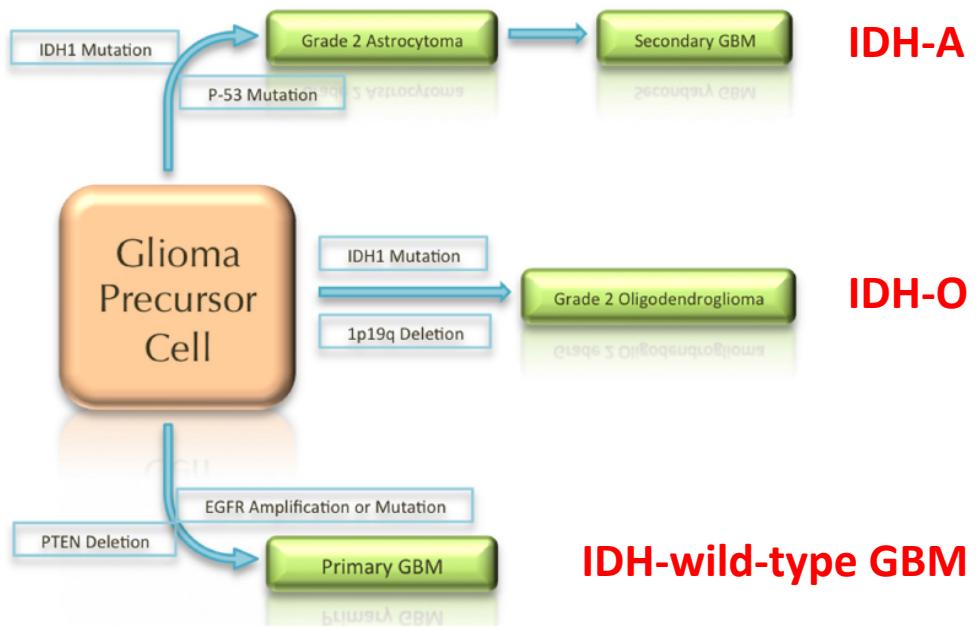
Population of cells
(bulk sequencing)



Single cell
(single-cell sequencing)



Primary glioblastoma & IDH-mutant gliomas



CANCER GENOMICS

Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq

9879 single cells from 10 IDH-A tumors
(Venteicher et al., Science, 2017)

LETTER

doi:10.1038/nature20123

Single-cell RNA-seq supports a developmental hierarchy in human oligodendrogloma

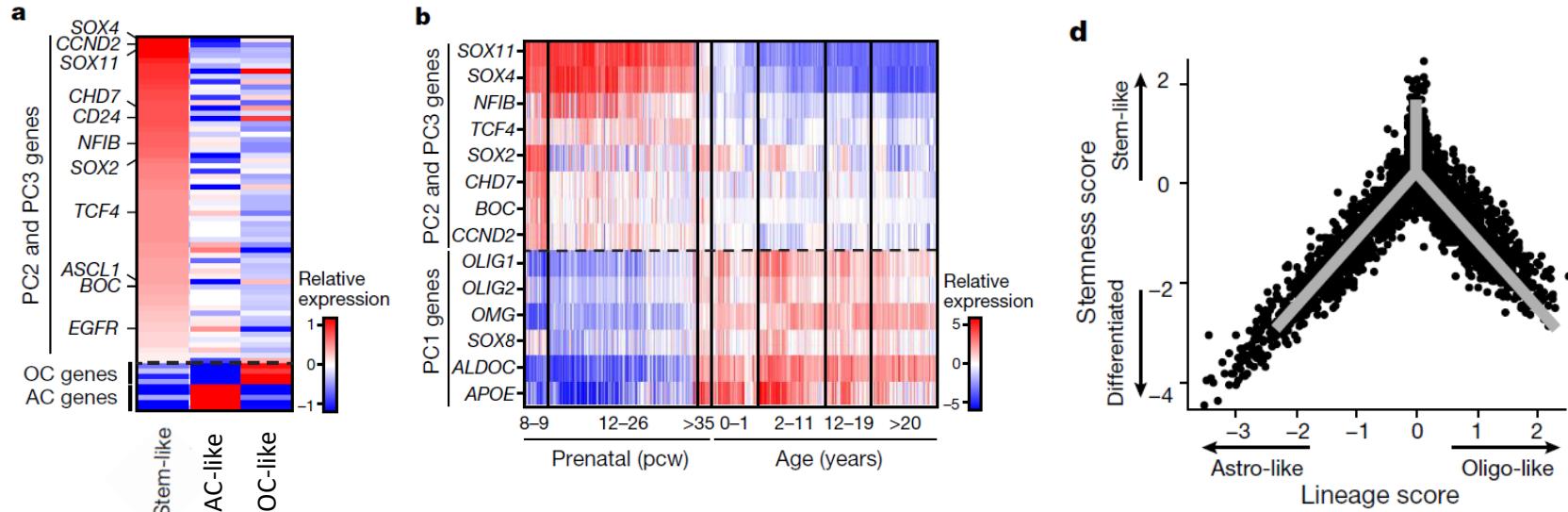
4347 single cells from 6 IDH-O tumors
(Tirosh et al., Nature, 2016)

CANCER GENOMICS

Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma

430 single cells from 5 primary GBMs
(Patel et al., Science, 2014)

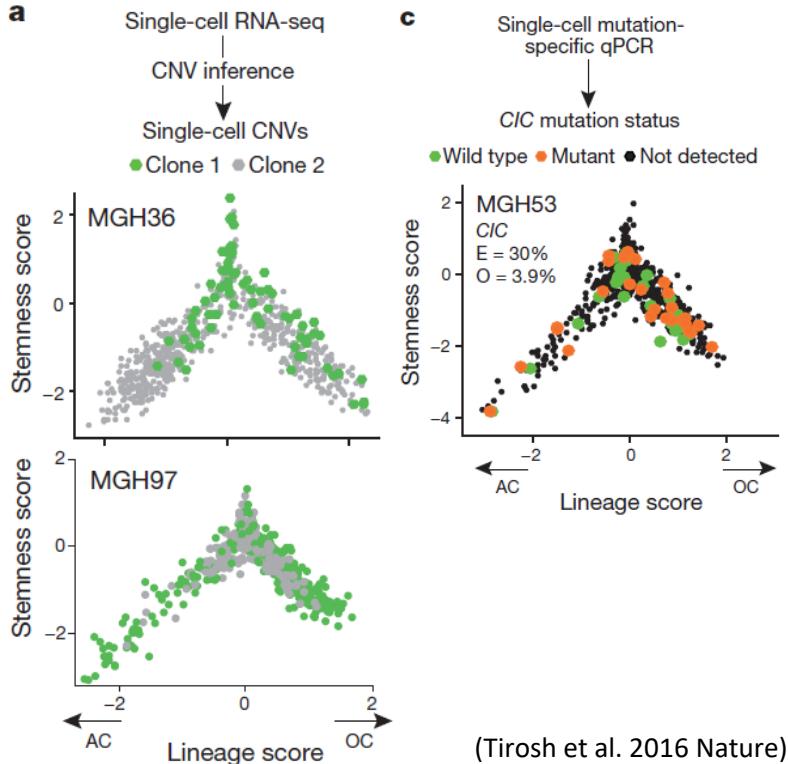
Major findings: three subpopulation of malignant cells



(Tirosh et al. 2016 Nature)

- PC1 associated genes separate two subpopulations of glial cells, with oligodendrocytic and astrocytic markers. **“Lineage Score”**
- PC2 and PC3 associated genes represent stemness signatures. **“Stemness Score”**

Role of genetics on tumor architecture?



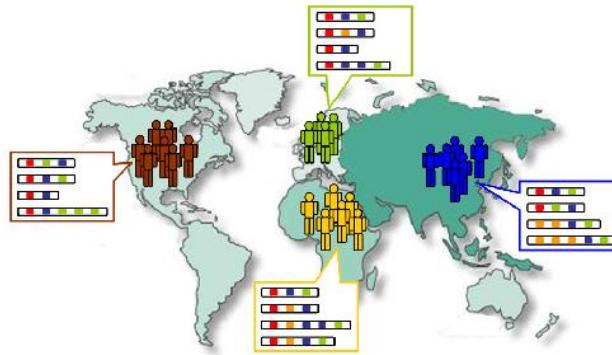
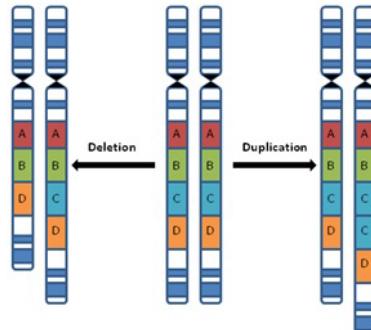
- Overall tumor architecture was preserved across clones reconstructed by CNVs and point mutations.
- Tirosh et al., Nature 2016:
“Many subclonal mutations can accrue within the hierarchy (but not drive it), although **without a comprehensive phylogenetic reconstruction, we cannot categorically rule out a genetic influence.**”
- Venteicher et al., Science 2017:
“... patterns of differentiation and proliferation can be partially modulated by genetics and subject to selection. **Future studies should further investigate the modulation of our inferred cellular architecture by genetic evolution**”

Two broad types of mutations

- Copy number aberrations (CNAs)
 - Bulk and single-cell DNA sequencing: How to remove biases and artifacts?
 - Single-cell RNA sequencing: How to infer copy number from expression?
- Single-nucleotide variants (SNVs)
 - Bulk DNA sequencing: MuTect, VarScan 2, etc., fairly standardized.
 - Single-cell RNA sequencing: Technical noise? Transcriptional bursting?

Copy number variation (CNV)

- Large deletions or duplications
- Abundant source of variations
- Associated with diseases



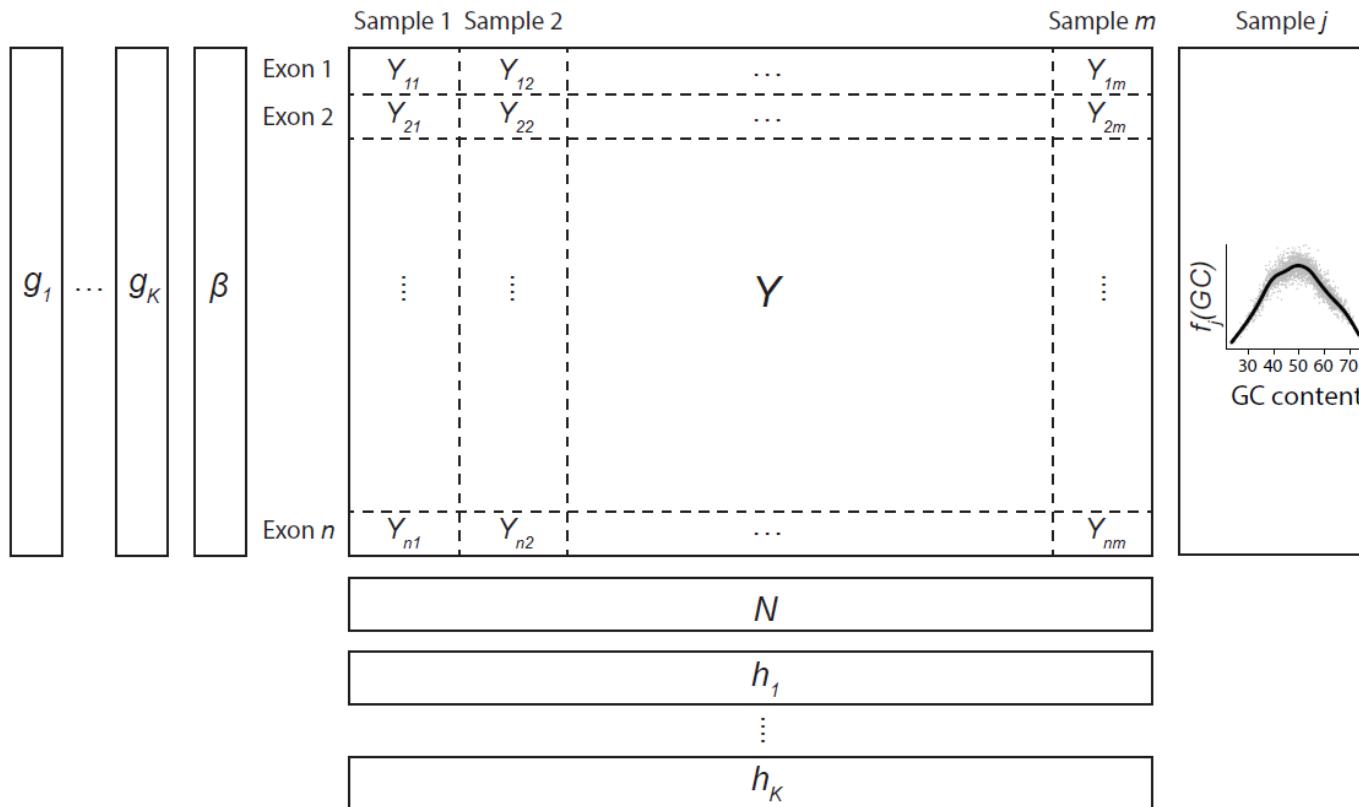
Normal



Tumor



Poisson latent factor model for normalization



Null Model:

$$Y_{ij} \sim \text{Poisson}(\lambda_{ij})$$

$$\lambda_{ij} = N_j \beta_i f_j(GC_i) \exp \left(\sum_{k=1}^K g_{ik} h_{jk} \right)$$

i : exon number; j : sample number

Y : raw coverage

λ : expected coverage (no CNV)

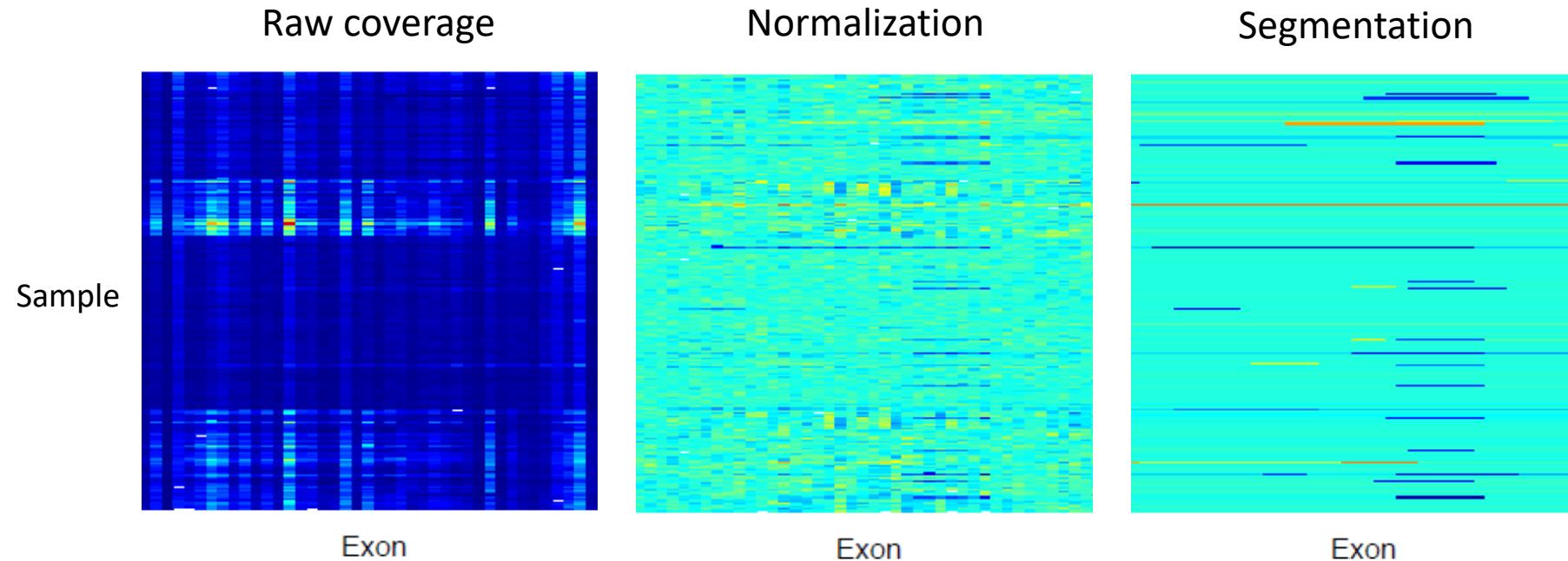
$$N_j = \sum_{i=1}^n Y_{ij}$$

β_i : exonic-specific bias for exon i

$f_j(GC_i)$: bias due to GC content

$g_{ik} h_{jk}$ ($1 \leq k \leq K$) : k th latent Poisson factors

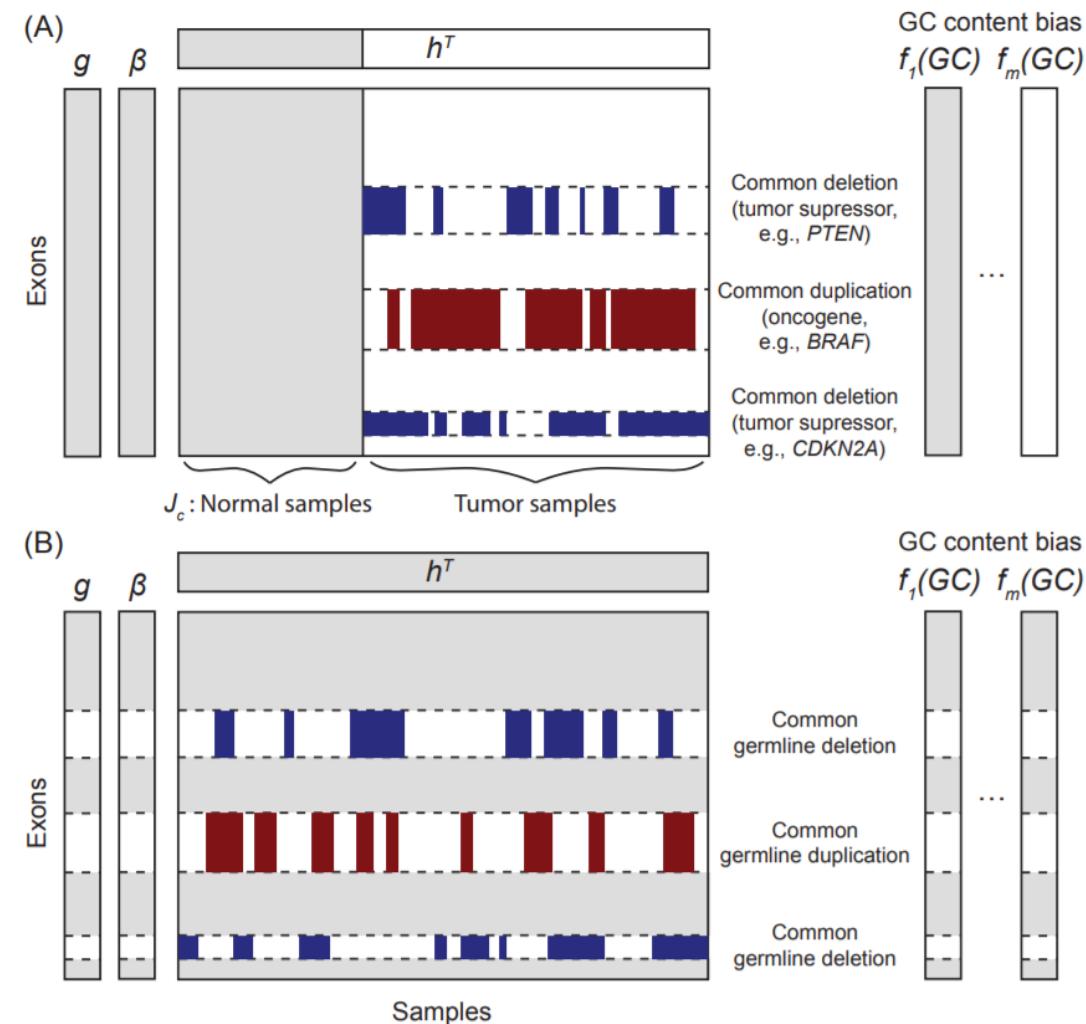
CODEX: Copy number Detection by EXome-seq



<http://bioconductor.org/packages/CODEX/>
(Jiang et al., Nucleic Acids Research, 2015)

CODEX2: full-spectrum CNV detection by NGS

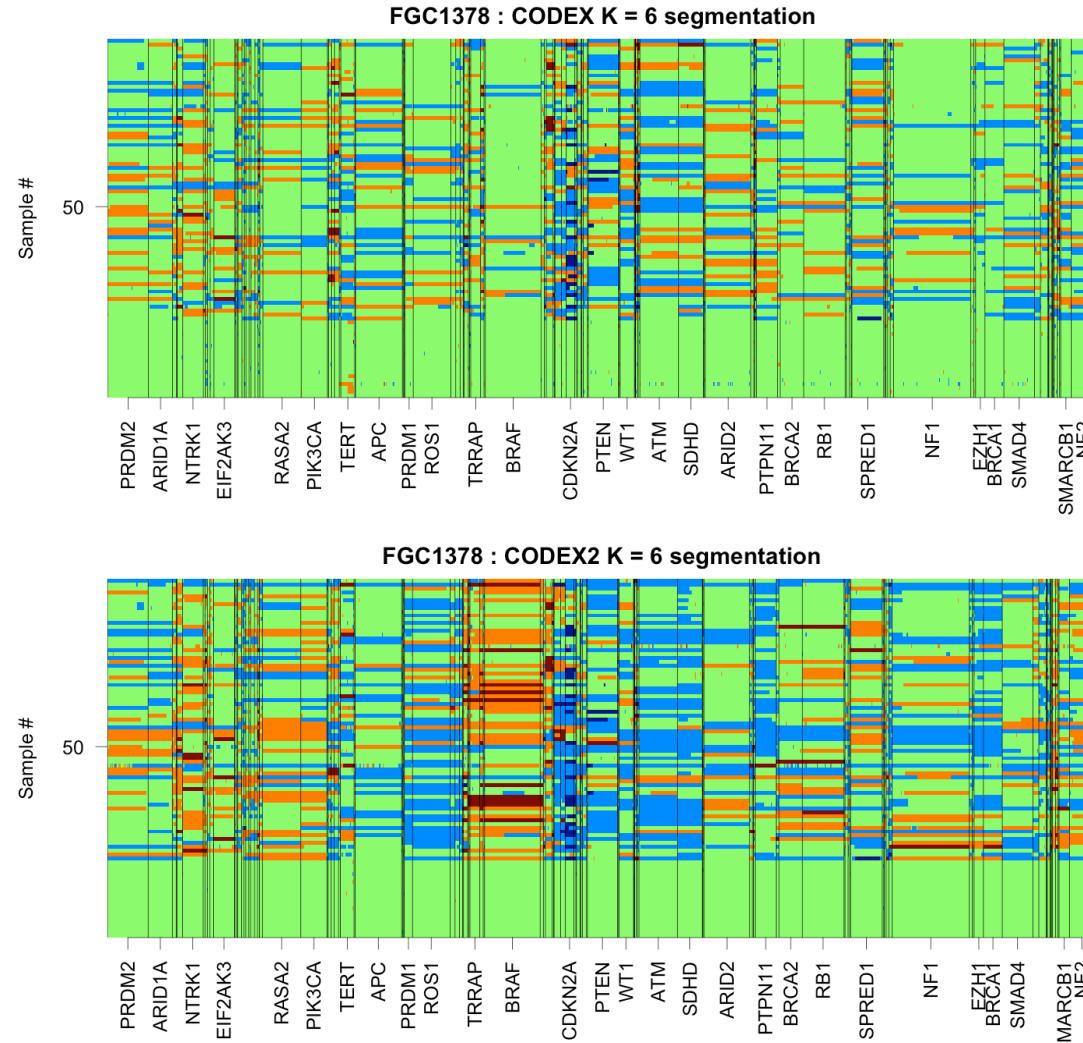
Negative control samples



<https://github.com/yuchaojiang/CODEX2>
(Jiang et al., Genome Biology, 2018)

Melanoma targeted sequencing: sanity check

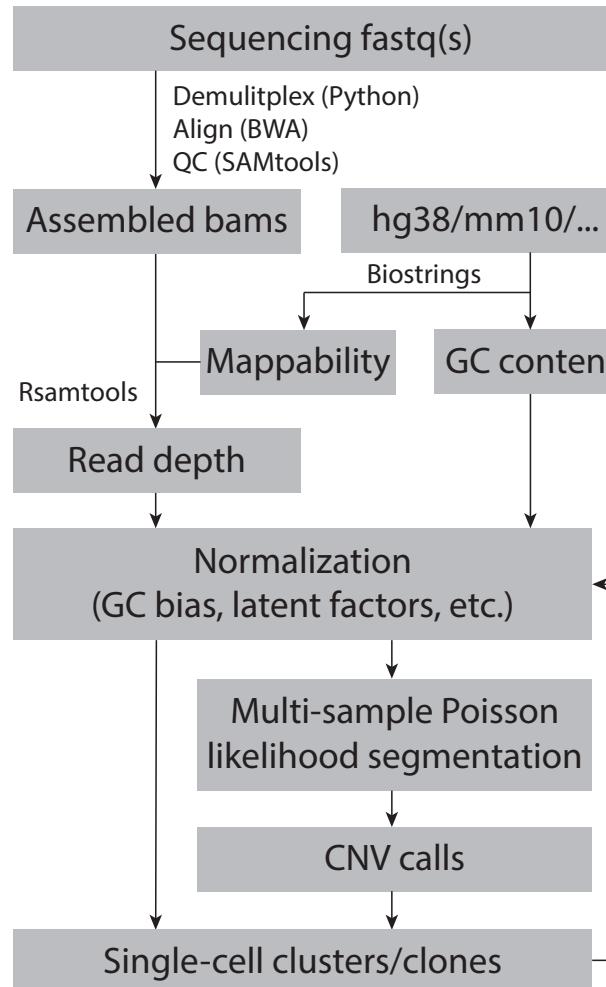
- High amp
- Gain
- Null
- Hemi. del
- Homo. del



CODEX

CODEX2

SCOPE: Single-cell COPy number Estimation

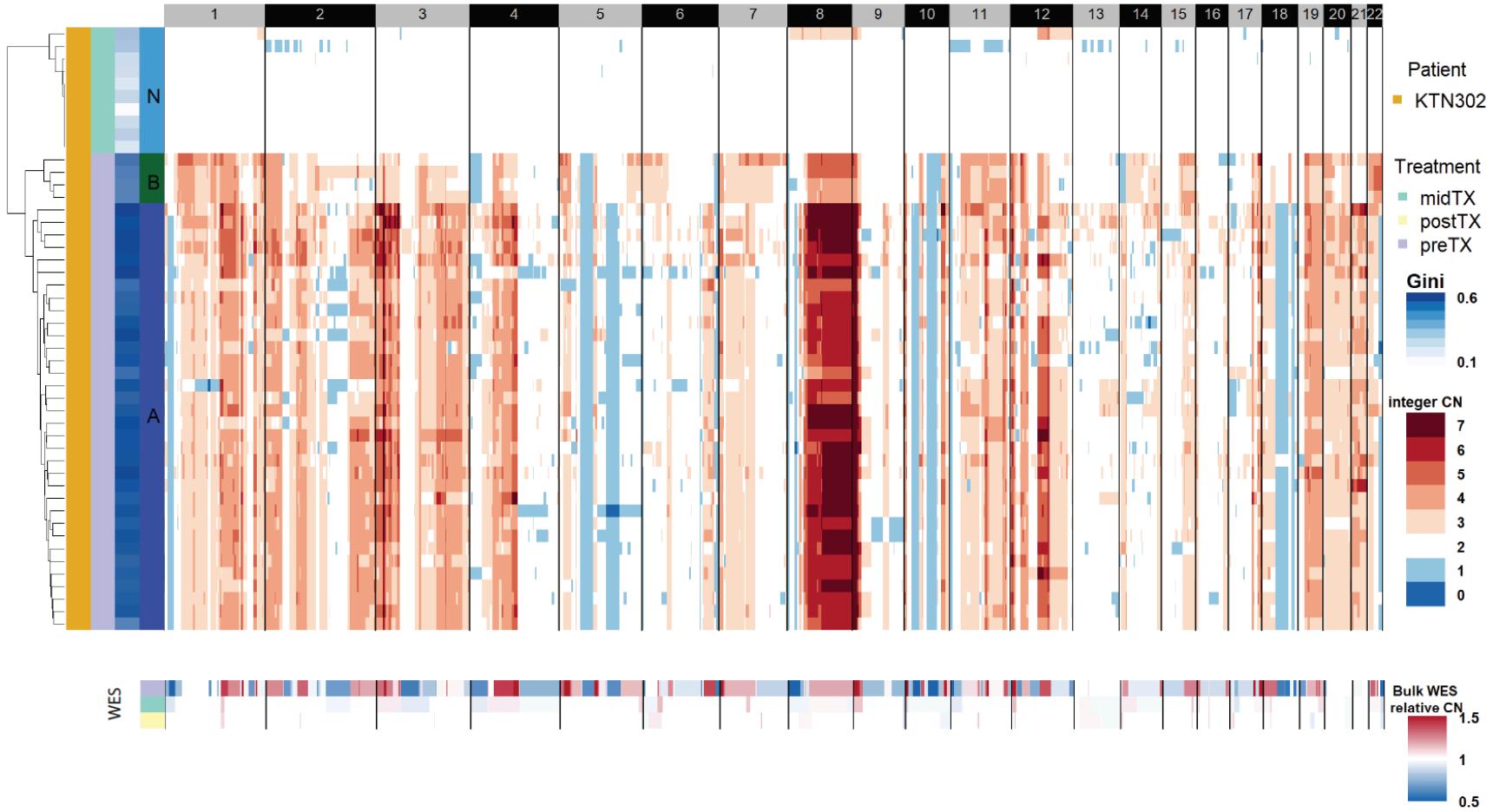


<https://github.com/rujinwang/SCOPE>

Wang et al., bioRxiv, 2019

Copy number profiles by scDNA-seq, WES, scRNA-seq

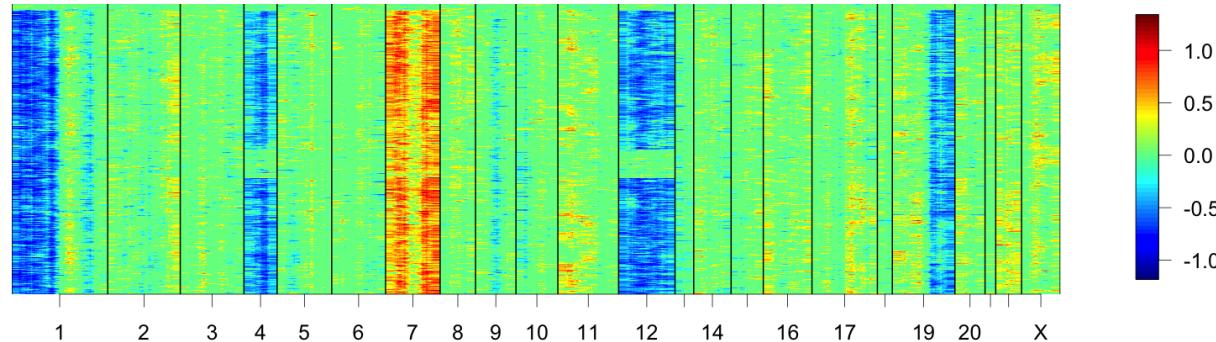
Triple negative breast cancer patient from Kim et al. (Cell, 2018)



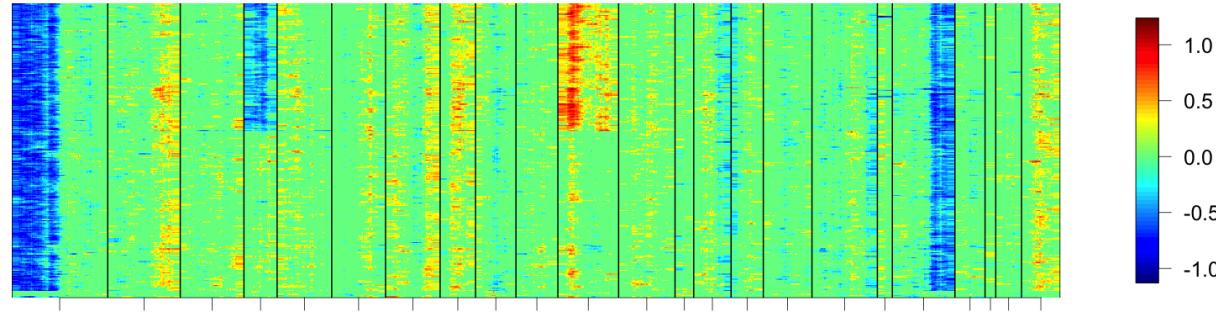
Sliding window approach for copy number profiling by scRNA-seq

Processed glioblastoma scRNA-seq dataset from Tirosh et al., Nature, 2016

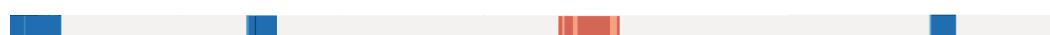
MGH 36
scRNA-seq



MGH 97
scRNA-seq

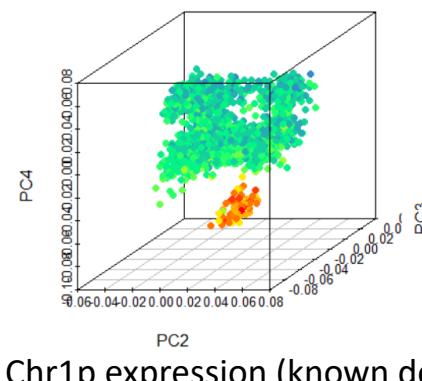
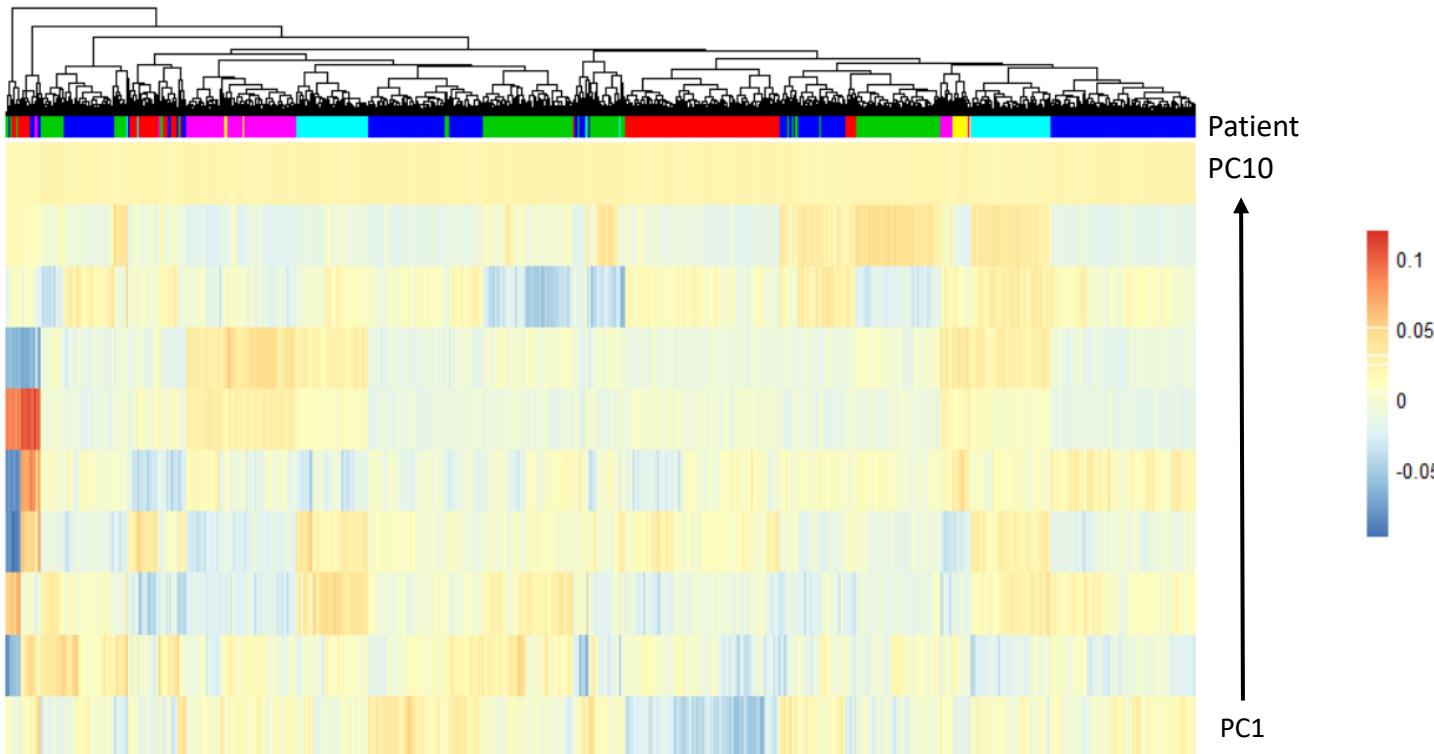


MGH 97
Bulk WES

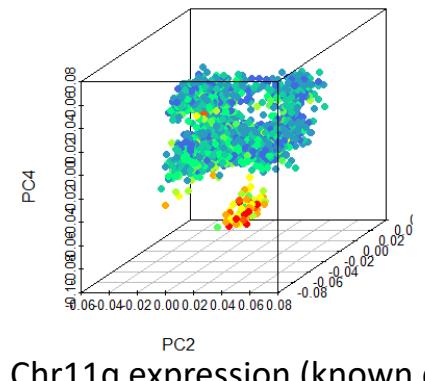


How to identify non-malignant cells?

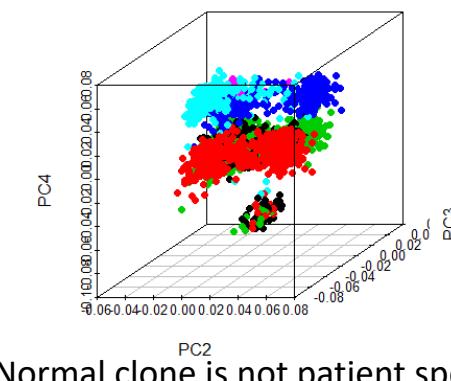
PCA on log(TPM/10+1)



Chr1p expression (known del)



Chr11q expression (known del)

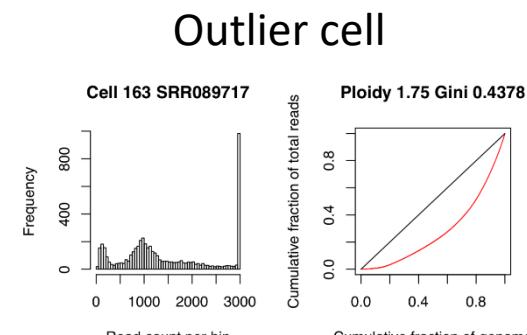
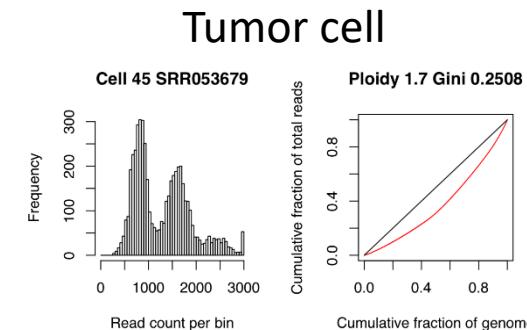
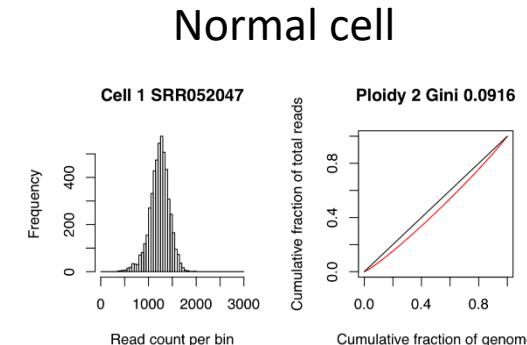
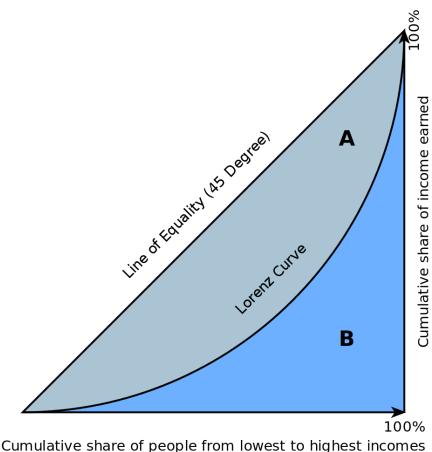


Normal clone is not patient specific!

How to identify normal cells?

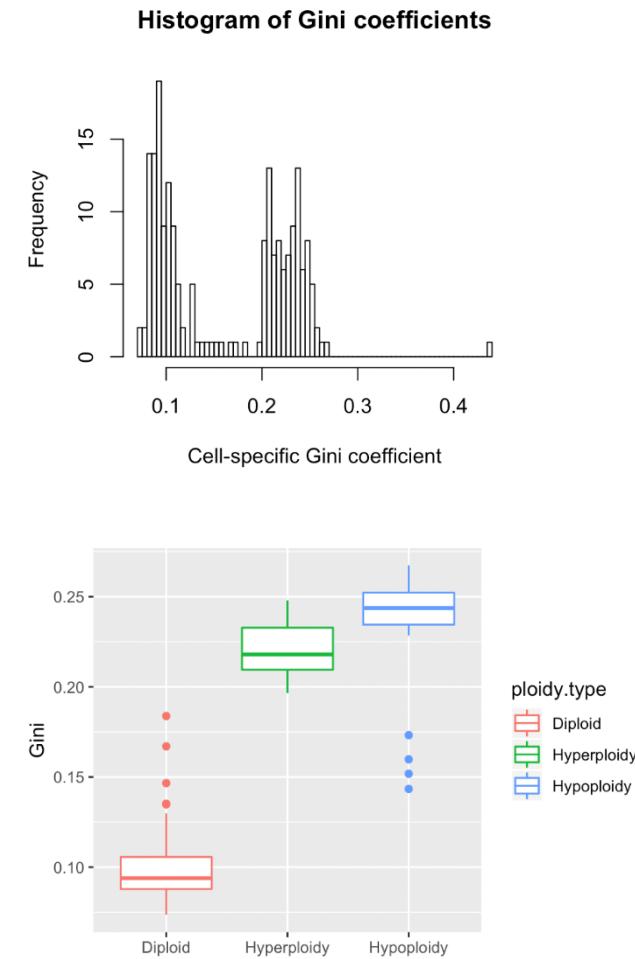
- Lorenz curve: assess coverage uniformity for each cell.
- Gini coefficient: two times the area between the Lorenz curve and the diagonal.
Equivalently,

$$Gini_j = \frac{\sum_{i=1}^m \sum_{k=1}^m |Y_{ij} - Y_{kj}|}{2m \sum_{i=1}^m Y_{ij}}.$$



Gini index can be used to identify normal cells

- Cell-specific Gini index distribution
- Ploidy measured from FACS:
 - Hyperploid: ploidy ≥ 2.1
 - Hypoploid: ploidy ≤ 1.9
 - Diploid: $1.9 < \text{ploidy} < 2.1$
- Hyperploid and hypoploid cells have higher Gini index compared to diploid cells.



Genomic and transcriptomic heterogeneity

Method

Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data

Jean Fan,^{1,5} Hae-Ock Lee,^{2,5} Soohyun Lee,¹ Da-eun Ryu,² Semin Lee,¹ Catherine Xue,¹ Seok Jin Kim,³ Kihyun Kim,³ Nikolaos Barkas,¹ Peter J. Park,¹ Woong-Yang Park,² and Peter V. Kharchenko^{1,4}

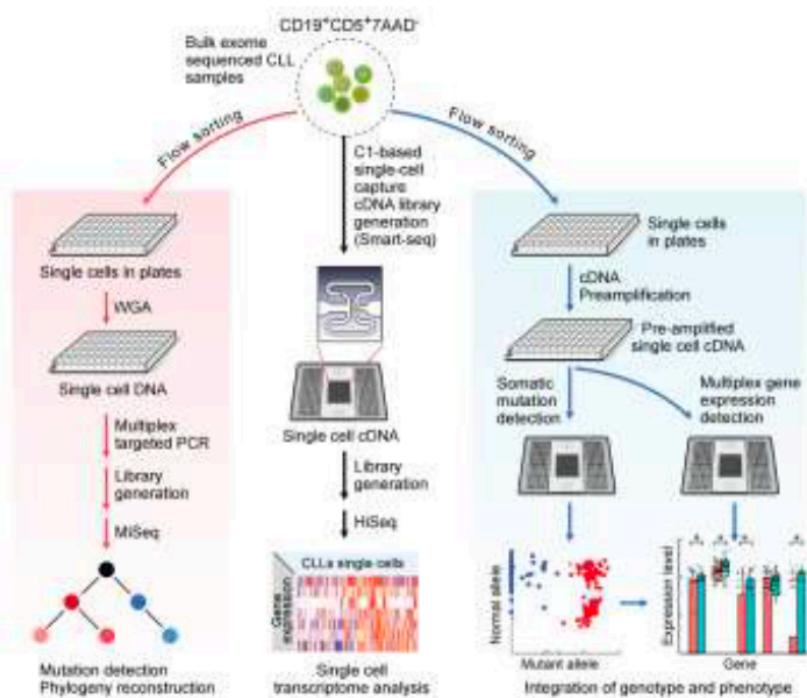
Cardelino: Integrating whole exomes and single-cell transcriptomes to reveal phenotypic impact of somatic variants

Davis J. McCarthy^{1,4,*}, Raghd Rostom^{1,2,*}, Yuanhua Huang^{1,*}, Daniel J. Kunz^{2,5,6}, Petr Danecek², Marc Jan Bonder¹, Tzachi Hagai^{1,2}, HipSci Consortium, Wenyi Wang⁸, Daniel J. Gaffney², Benjamin D. Simons^{5,6,7}, Oliver Stegle^{1,3,9,#}, Sarah A. Teichmann^{1,2,5,#}

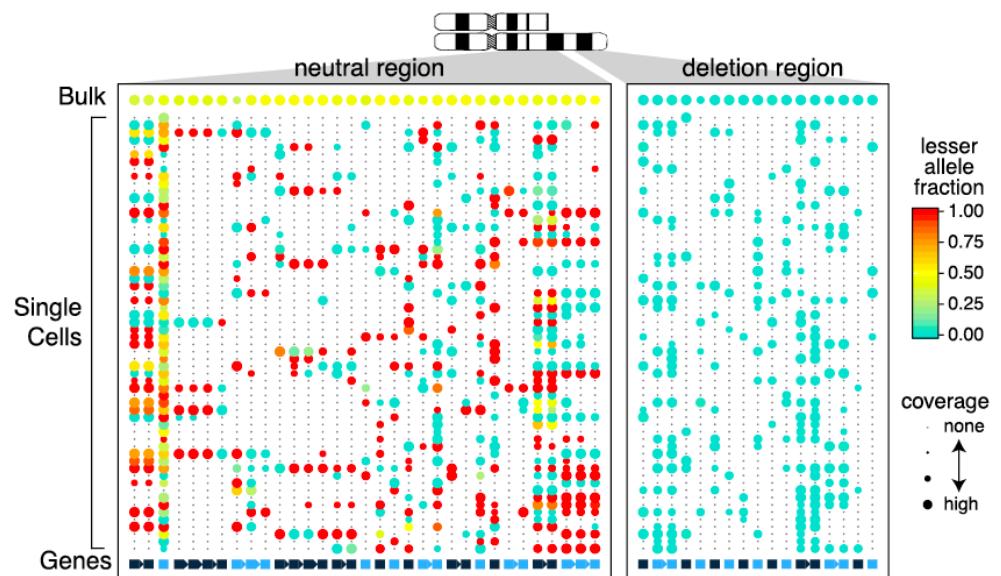
HoneyBADGER: CNAs by scRNA-seq

How to simultaneously assess transcriptional and genetic heterogeneity at the single cell level?

1. Technology Development Approach (Targeted RT-QPCR)



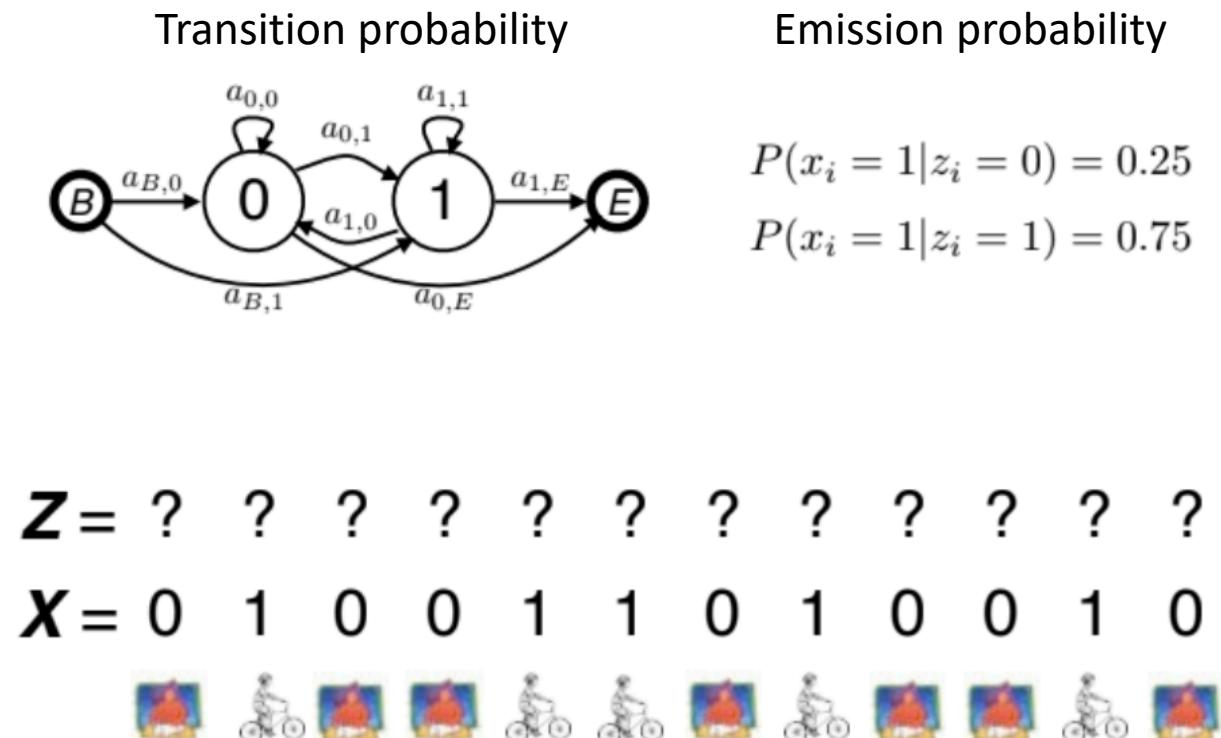
2. Computational Methods Approach (HoneyBADGER)



<https://jef.works/HoneyBADGER/>

Hidden Markov model (HMM)

- Let $\mathbf{X} = (X_1, \dots, X_L)$ indicate whether YJ bikes on day i ($X_i = 1$) or not ($X_i = 0$)
- Suppose YJ bikes on day i with probability $\theta_0 = 0.25$ if it is cloudy ($Z_i = 0$) and with probability $\theta_1 = 0.75$ if it is sunny ($Z_i = 1$)
- Further suppose the Z_i s are *hidden*; we see only $\mathbf{X} = (X_1, \dots, X_L)$
- This *hidden Markov model* is a mixture model in which the Z_i s are correlated
- We call $\mathbf{Z} = (Z_1, \dots, Z_L)$ the *path*



HMM for HoneyBADGER

HoneyBADGER implements an expression-based HMM as well as an allele-based HMM to identify regions potentially affected by CNVs. For the expression-based HMM, a transition matrix is defined on three hidden states representing deletion, neutral, and amplification:

$$\begin{pmatrix} 1 - 2t & t & t \\ t & 1 - 2t & t \\ t & t & 1 - 2t \end{pmatrix},$$

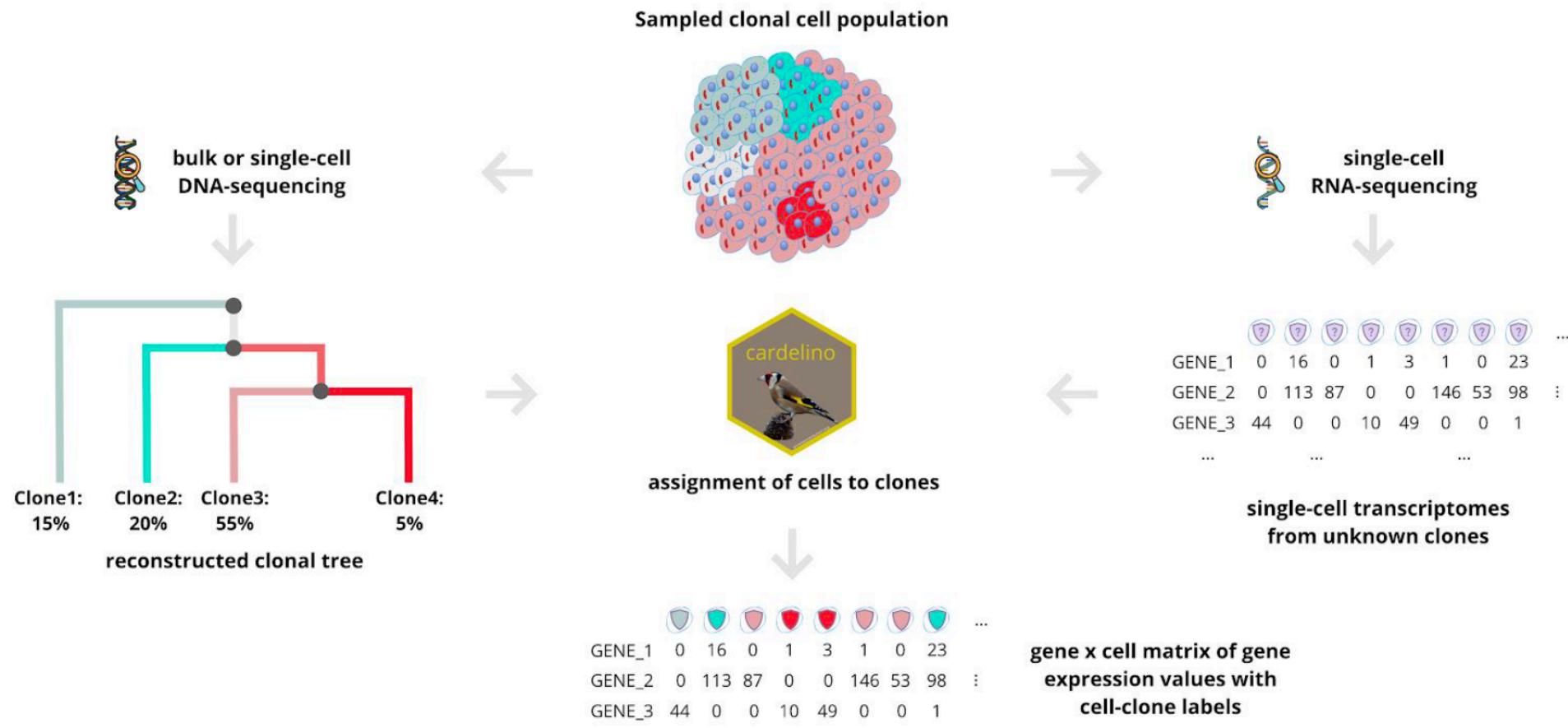
where $t = 1 \times 10^{-5}$ by default. Emission probabilities are defined by a normal distribution with means and variance estimated from the normalized expression data (see Supplemental Methods). For the allele-based HMM, a transition matrix is defined on two hidden states representing deletion or LOH, and neutral:

$$\begin{pmatrix} 1 - t & t \\ t & 1 - t \end{pmatrix},$$

where $t = 1 \times 10^{-5}$ by default. Emission probabilities are defined by a binomial distribution with the size parameter given by the pooled coverage at the SNP position and an expected $P = 0.1$ for the lesser allele in the case of deletion or LOH and $P = 0.45$ for neutral.

Cardelino: SNVs by scRNA-seq

Probabilistic mapping of single-cell transcriptomes to reconstructed clones



<https://github.com/PMBio/cardelino>

Cardelino: two-step approach

- First, a clonal tree is inferred using variant allele frequencies from bulk or single-cell DNA sequencing data.
- Subsequently, cardelino performs probabilistic assignment of individual single-cell transcriptomes to inferred clones, using variant information extracted from scRNA-seq.

References

- Jiang, Yuchao, et al. "Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing." *Proceedings of the National Academy of Sciences* 113.37 (2016): E5528-E5537.
- Tirosh, Itay, et al. "Single-cell RNA-seq supports a developmental hierarchy in human oligodendrogloma." *Nature* 539.7628 (2016): 309.
- Venteicher, Andrew S., et al. "Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq." *Science* 355.6332 (2017): eaai8478.
- Fan, Jean, et al. "Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data." *Genome research* 28.8 (2018): 1217-1227.
- McCarthy, Davis James, et al. "Cardelino: Integrating whole exomes and single-cell transcriptomes to reveal phenotypic impact of somatic variants." *bioRxiv* (2018): 413047.

References, cont'd

- Jiang, Yuchao, et al. "CODEX: a normalization and copy number variation detection method for whole exome sequencing." *Nucleic acids research* 43.6 (2015): e39-e39.
- Jiang, Yuchao, et al. "CODEX2: full-spectrum copy number variation detection by high-throughput DNA sequencing." *Genome biology* 19.1 (2018): 202.
- Urrutia, Eugene, et al. "Integrative pipeline for profiling DNA copy number and inferring tumor phylogeny." *Bioinformatics* 34.12 (2018): 2126-2128.
- Wang, Rujin, Dan-Yu Lin, and Yuchao Jiang. "SCOPE: a normalization and copy number estimation method for single-cell DNA sequencing." *bioRxiv* (2019): 594267.