Genome Characterization from **Bulk Tissue** to **Single Cells**

Statistics in the Big Data Era Conference in Celebration of Peter Bickel's 80th Birthday

Nancy R. Zhang

Department of Statistics

The Wharton School, University of Pennsylvania

ENCODE

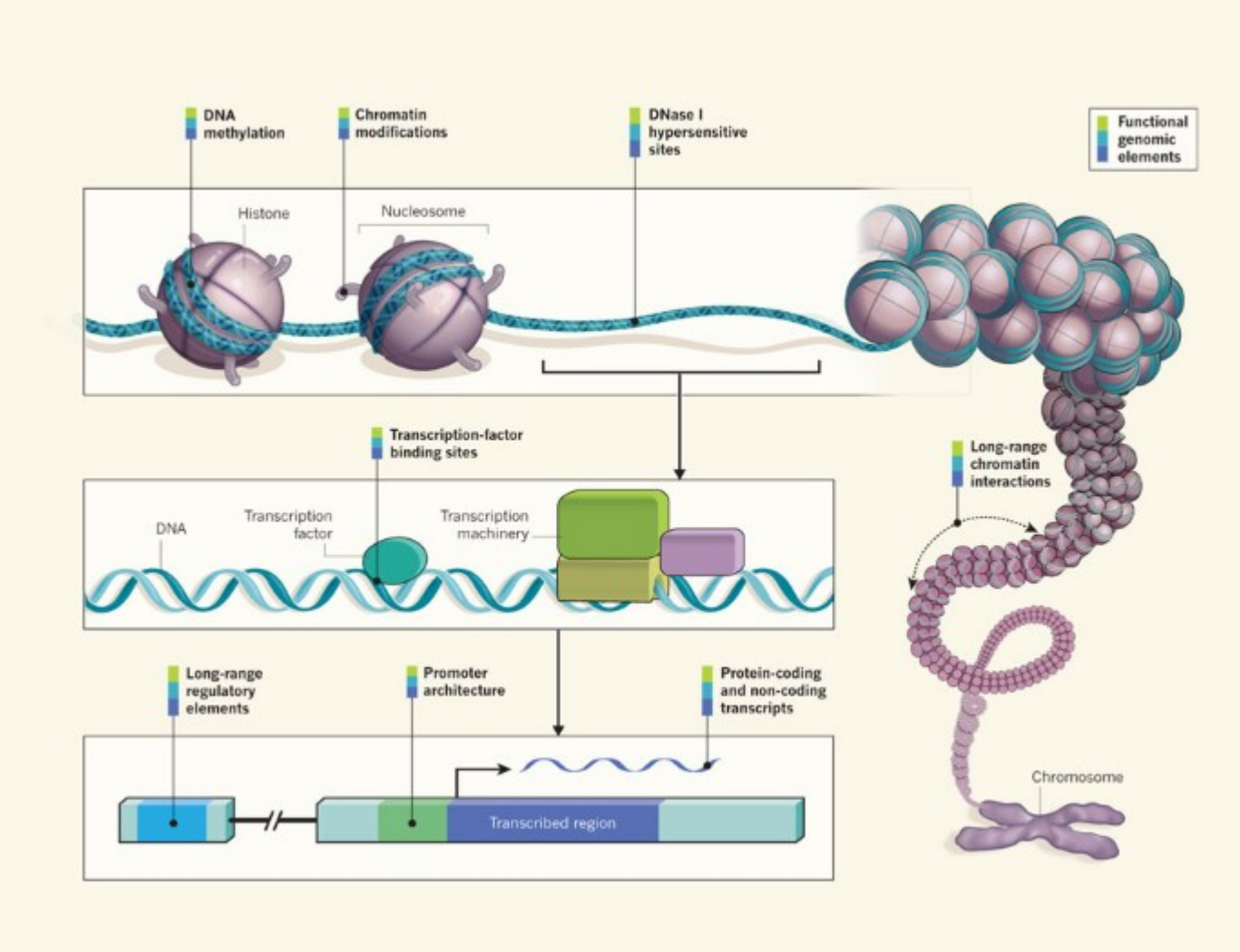
The Encyclopedia of DNA Elements (ENCODE)

ENCODE is a public research consortium aimed at identifying all functional elements in the human and mouse genomes.

2000 Rough draft of human genome revealed.

2003 ENCODE project launched

2005 Peter gave a talk on ENCODE at Stanford



2000 — Rough draft of human genome revealed.

2003 — ENCODE project launched

2005 — Peter gave a talk on ENCODE at Stanford

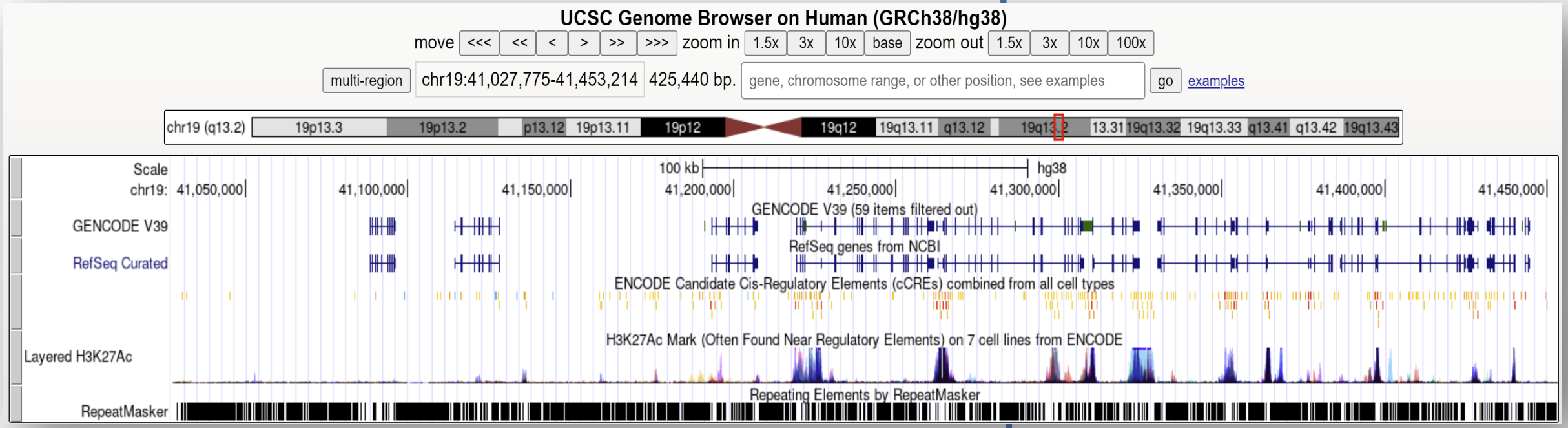
ENCODE Pilot phase:

- Sample 1% of the genome (30 Megabases)
- 35 laboratories, doing different assays
  - How accessible is the DNA?
  - Which parts are being made into RNA?
  - What proteins are binding and where?
- In the end: ~ 200 “features were measured

2000 — Rough draft of human genome revealed.

2003 — ENCODE project launched

2005 — Peter gave a talk on ENCODE at Stanford



ENCODE Pilot phase:

- Sample 1% of the genome (30 Megabases)
- 35 laboratories, doing different assays
  - How accessible is the DNA?
  - Which parts are being made into RNA?
  - What proteins are binding and where?
- In the end: ~ 200 features were measured

Pervasive question: How do the features relate to each other?

What were people doing?

Random reshuffling.

Two "point processes": $\{X_1, X_2, \ldots, X_n\}, \quad \{Y_1, Y_2, \ldots, Y_m\}$

Randomly sample $\boldsymbol{Y}$, compute overlap, do this many times.

What is wrong with this?

Genome features are highly non-uniform, clumpy!

"Two features overlap more than random chance."

-- What do we mean by "random"?

2000 Rough draft of human genome revealed.

2003 ENCODE project launched

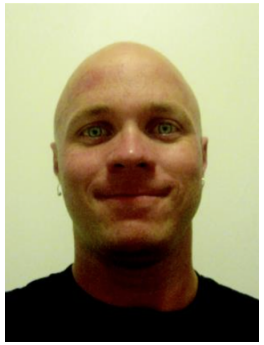
2005 Peter gave a talk on ENCODE at Stanford

"The essential challenge in the statistical formulation of this problem is the appropriate modeling of randomness of the genome, since we observe only one of the multitudes of possible genomes that evolution might have produced for our and other species."

- Bickel et al. AoAS 2010

Ben Brown


Haiyan Huang

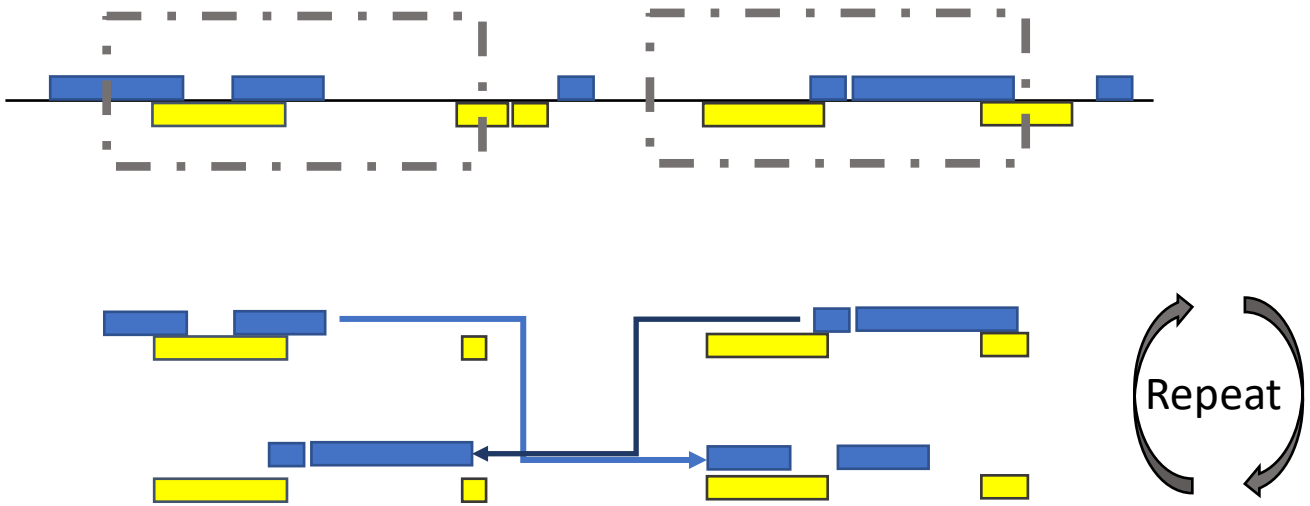

Nathan Boley

2000 Rough draft of human genome revealed.

2003 ENCODE project launched

2005 Peter gave a talk on ENCODE at Stanford

Peter guided us towards nonparametric approaches.



Bickel, Boley, Brown, Huang, and Zhang (AOAS, 2010) Subsampling Methods for Genomic Inference:

- Piecewise stationary model for the genome
- Stationary block sampling to get standard errors
- Normal approximations to the null distributions of overlap test statistics

2000 — Rough draft of human genome revealed.

2003 — ENCODE project launched

2005 — Peter gave a talk on ENCODE at Stanford

2000 Rough draft of human genome revealed.

2003 ENCODE project launched

2005 Peter gave a talk on ENCODE at Stanford

2007

"Genome Structure Correction" provided p-values for tests of overlap.

Where are we now after 15 years?

April 1 2022: Human genome sequence is ***finally*** complete.

Rest of this talk: **single cell, multiomic profiling**

2000 — Rough draft of human genome revealed.

2003 — ENCODE project launched

2005 — Peter gave a talk on ENCODE at Stanford

2007 — First publication of ENCODE

2022

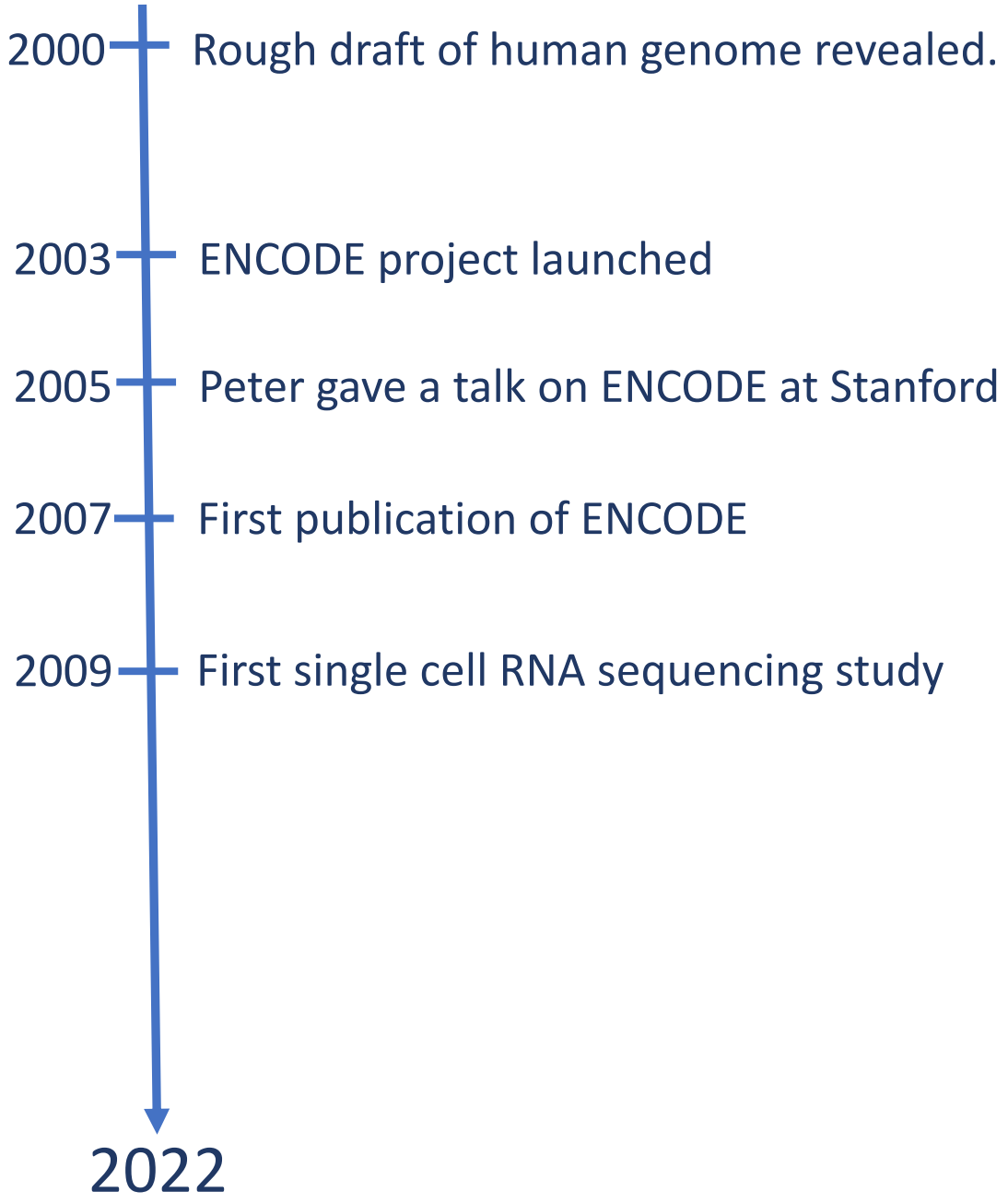
The rest of this talk:

- Single cell allele-specific copy number estimation

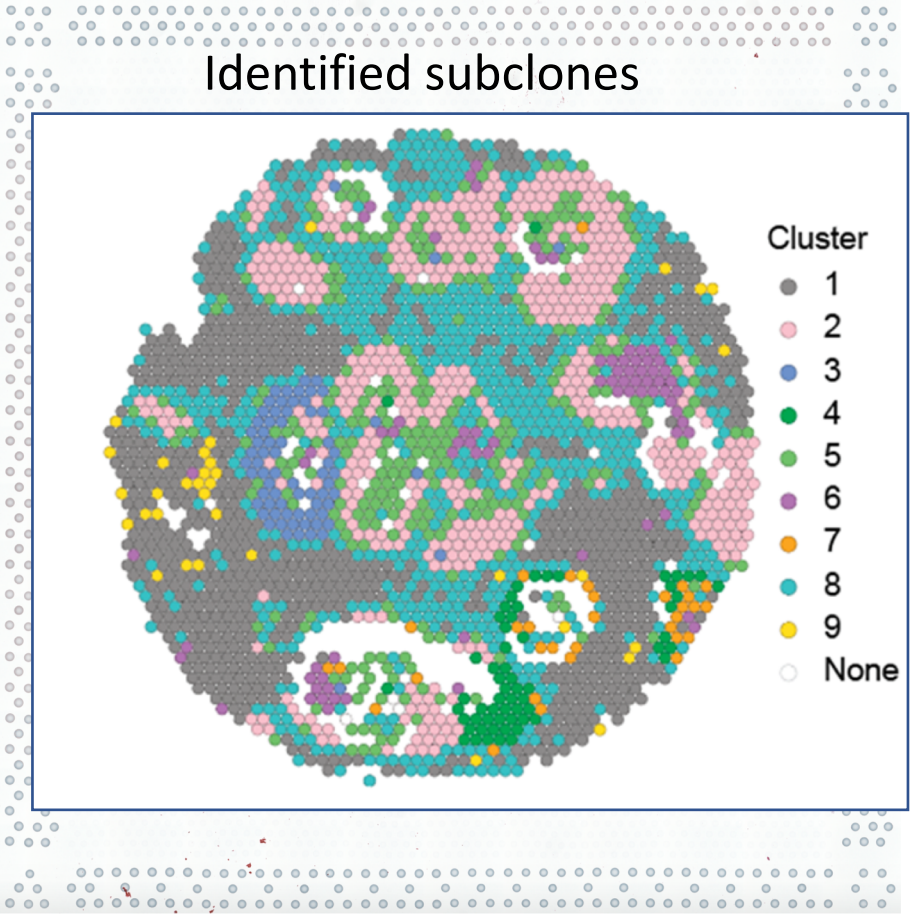
  Chi-Yun Wu et al. Integrative single-cell analysis of allele-specific copy number alterations and chromatin accessibility in cancer. ***Nature Biotechnology 2021***.

- Cancer subclone detection in spatial transcriptomic data

  Chi-Yun Wu et al. Subclone detection on copy number profiles in single cell and spatial tumor sequencing data. ***Under preparation 2022***

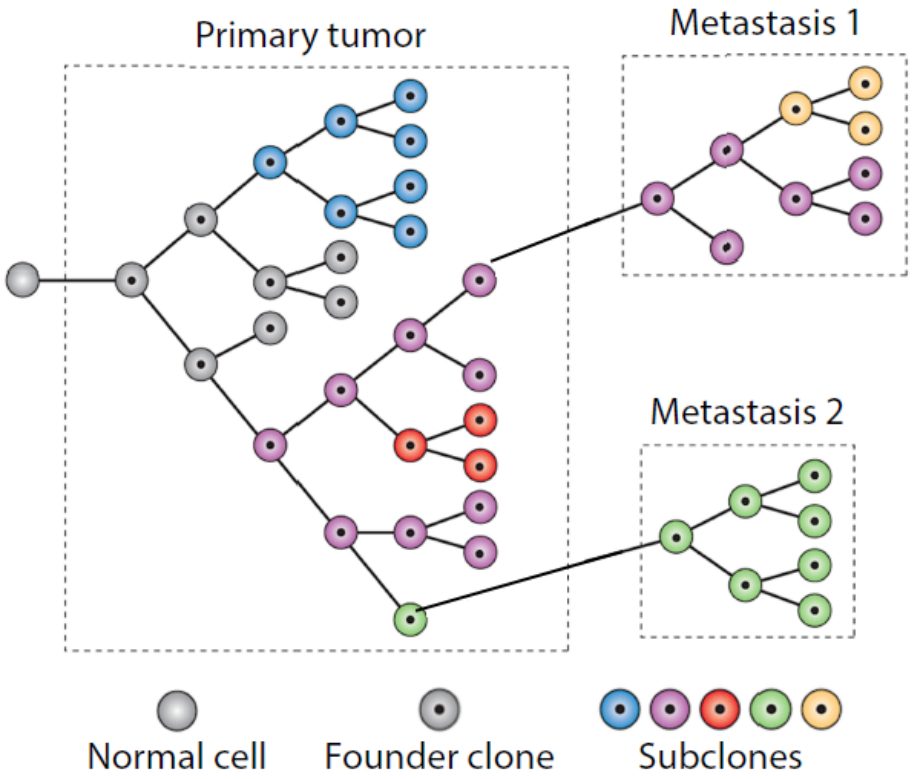


Why is cancer so interesting to me?

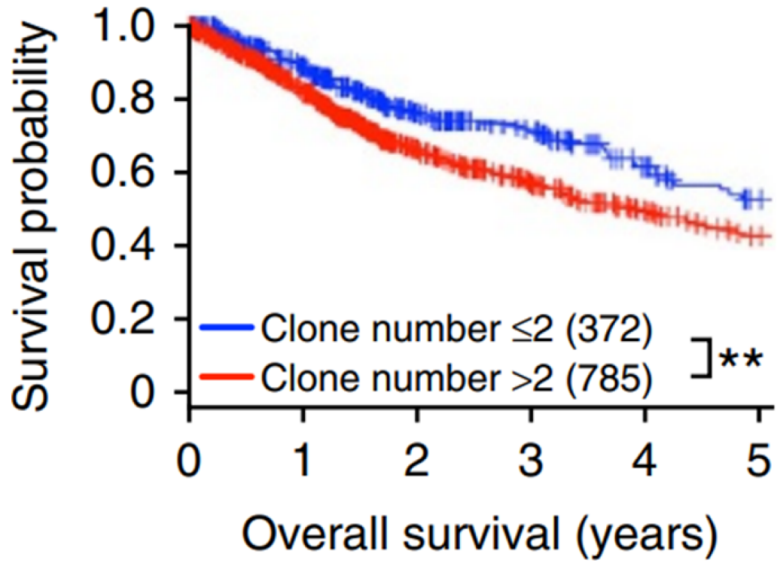


Cancer is a system of active intra-tissue competition and cooperation, selection and adaptation, when cells of a multicellular organism undergo Darwinian evolution.

Cancer cells follow Darwinian evolution

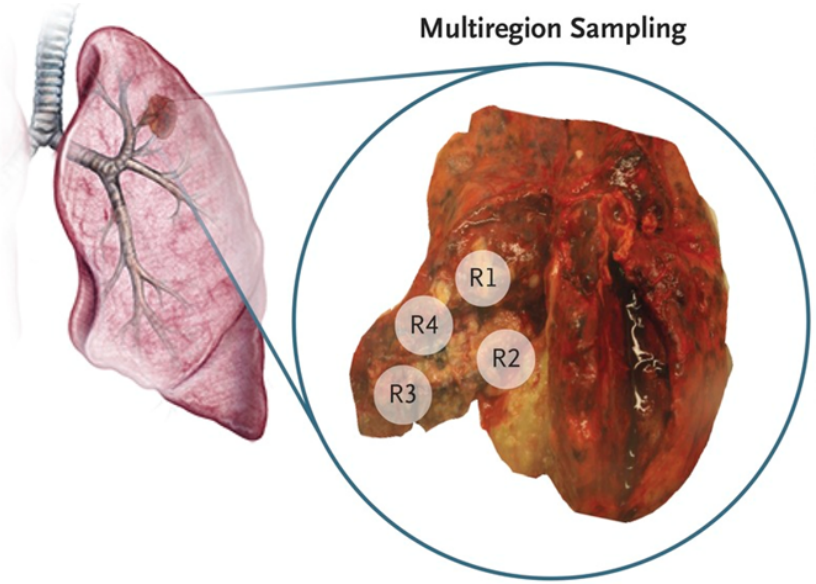


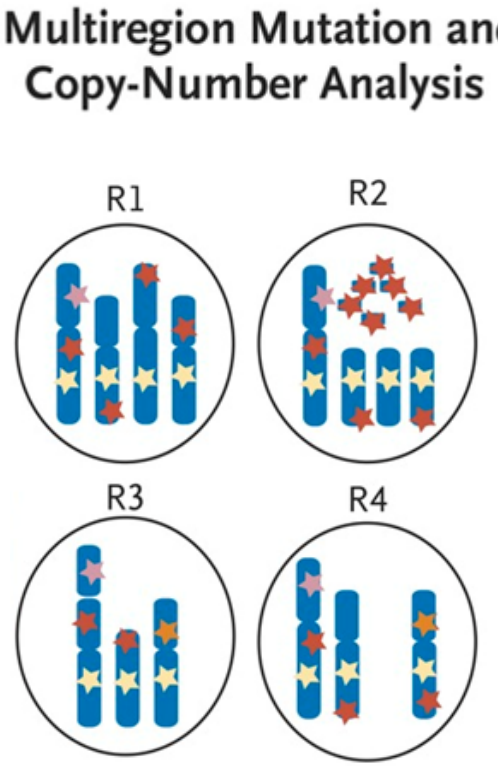
The subclone diversity of a tumor is directly linked to clinical outcome of the patient.



From: Andor et al. (2016) ***Pan-cancer analysis of the extent and consequences of intratumor heterogeneity,*** Nature Medicine 22, 105.

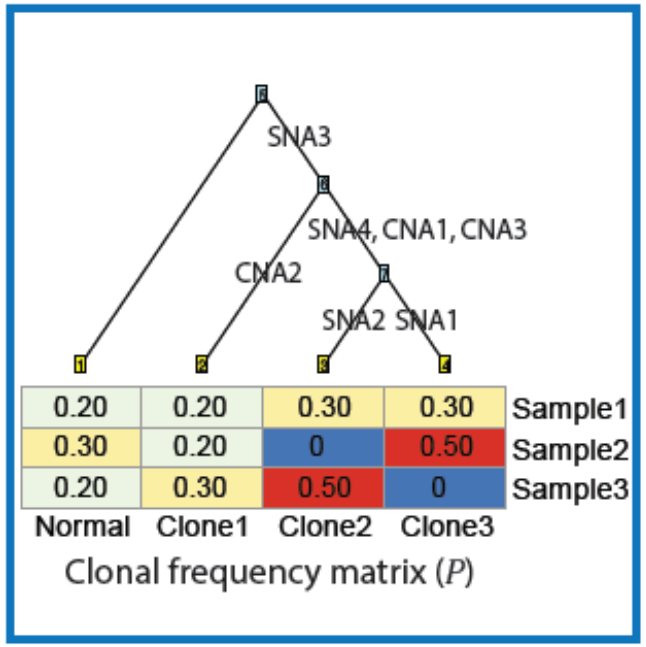
Multi-region sampling





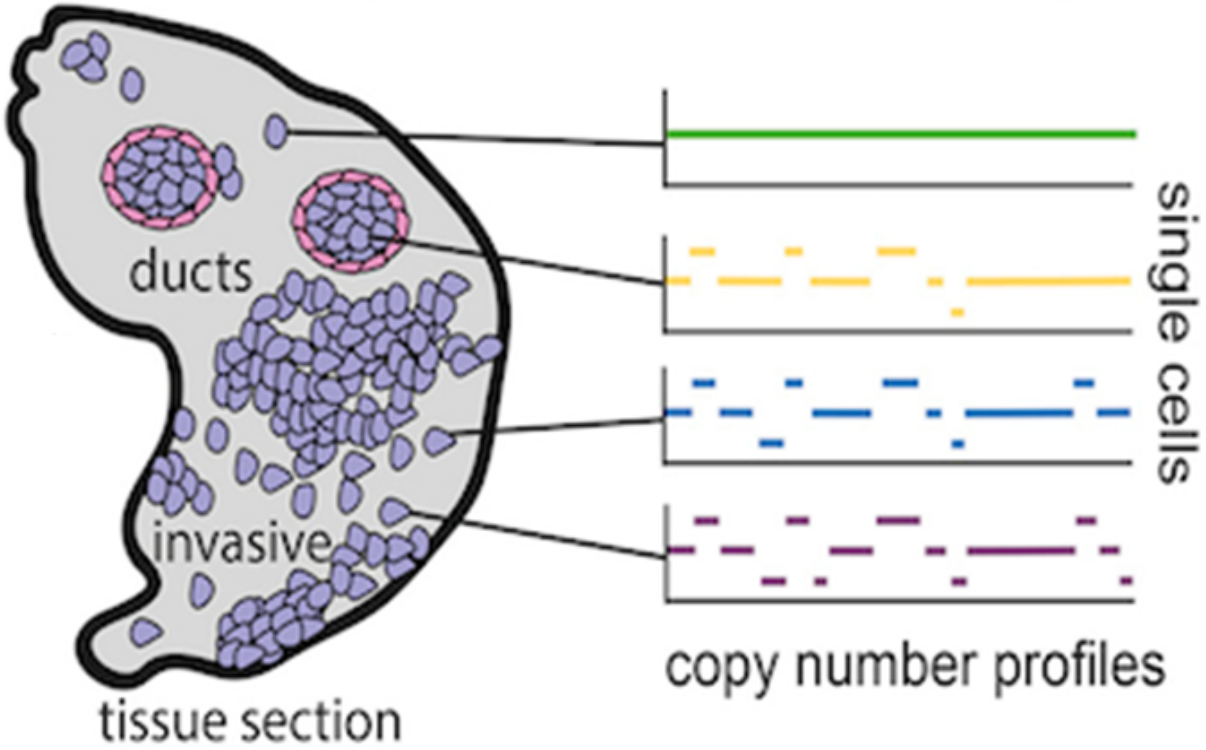
Jamal-Hanjani et al. Tracking the Evolution of Non–Small-Cell Lung Cancer. New England Journal of Medicine (2017)

Infer underlying phylogeny and subclone proportions in each sample:



Jiang et al. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. PNAS (2016)

Single cell DNA sequencing



Navin et al. (**2011**) Tumor evolution inferred by single-cell sequencing, Nature 472, p90.

At least 10 protocols to date for scDNA-seq.

Methods:

Ginkgo
2015

AneuFinder
2016

SCOPE
2020

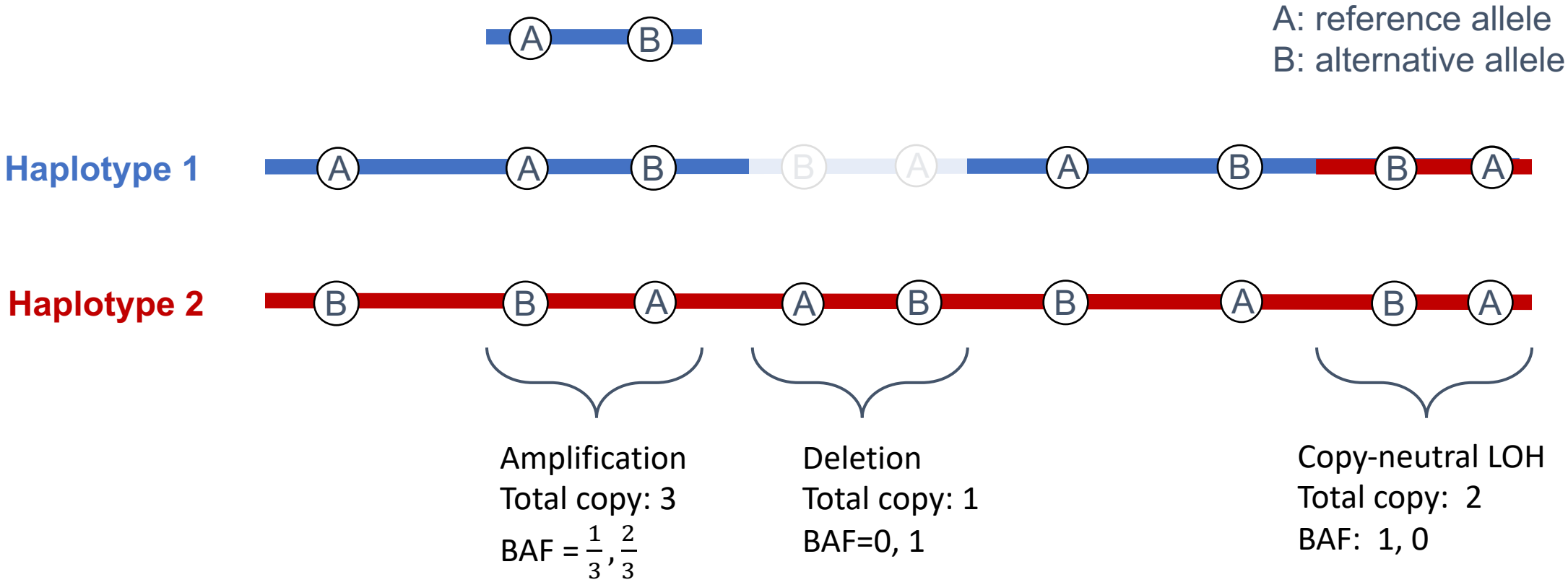
CHISEL
2021

HMMcopy
2021

Alleloscope
2021

Allele-specific copy number analysis in single cells

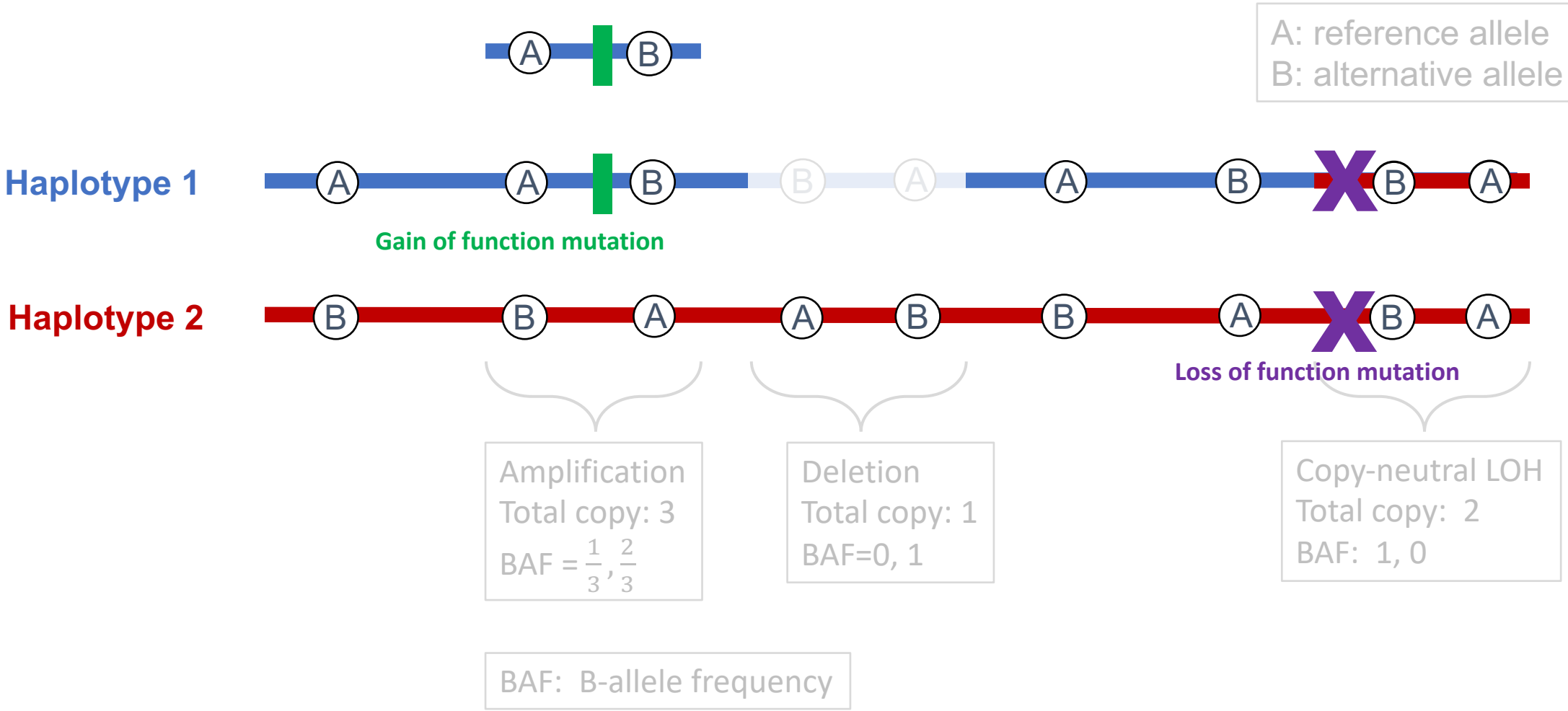
Allele-specific signals



BAF: B-allele frequency

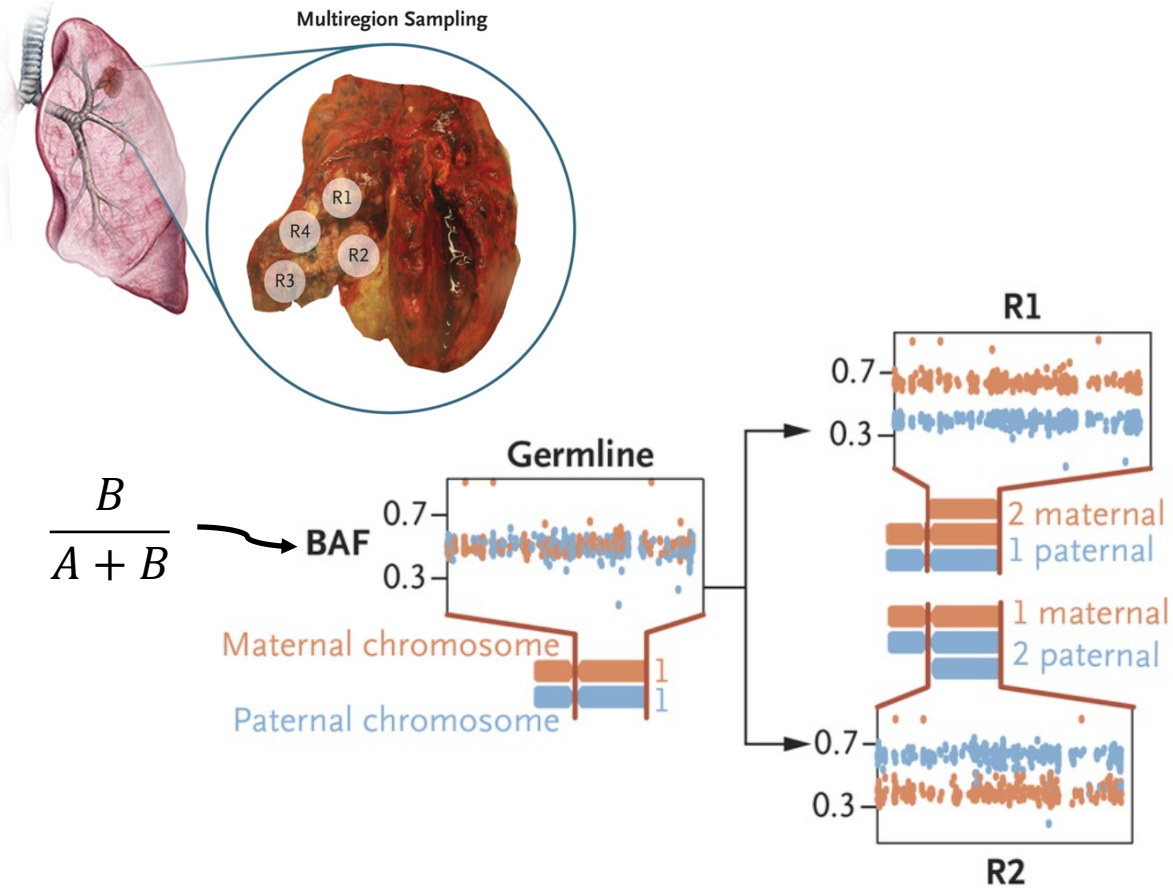
Allele-specific methods: jointly model the copy numbers of the two alleles.

What is **allele-specific** copy number



“Mirrored subclones”, a hidden variation

Jamal-Hanjani, M. et al. NEJM 2017



nature biotechnology
ARTICLES
https://doi.org/10.1038/s41587-020-0661-6
Check for updates

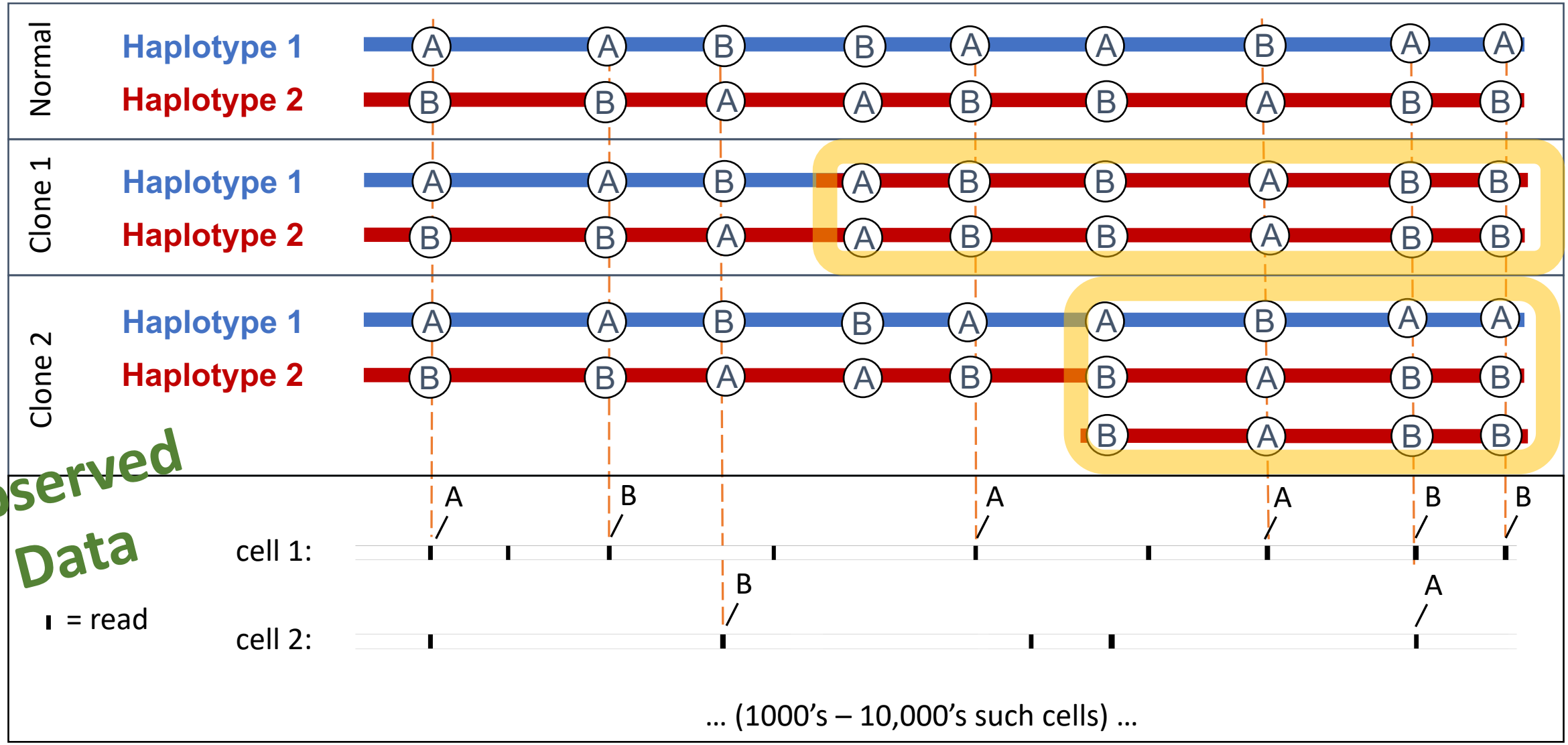
Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL

Simone Zaccaria[1] and Benjamin J. Raphael[1,2]

February 2021

- **Found 2 instances of such mirrored subclones in a high coverge breast cancer sample**
- **Applicable to high coverage single cell DNA sequencing data.**
- **Relies on external phasing of heterozygous sites.**
- **Not applicable to single cell ATAC sequencing data (more on this later)**

Underlying structure of single cell allele-specific data



Challenges: (1) Reads per cell are **sparse**. (2) Don't know **where the "breakpoints" are**. (3) Don't know the **underlying phase**.

Model (in formula)

For each cell i, let the reads be sampled according to

$$Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{iN_i}) \sim Poisson(\mu_t \delta_{it}), \qquad (\boldsymbol{\delta_{it}}, \boldsymbol{\theta_{it}}) \sim$$

where δ_{it} piecewise continuous rate function with change-points

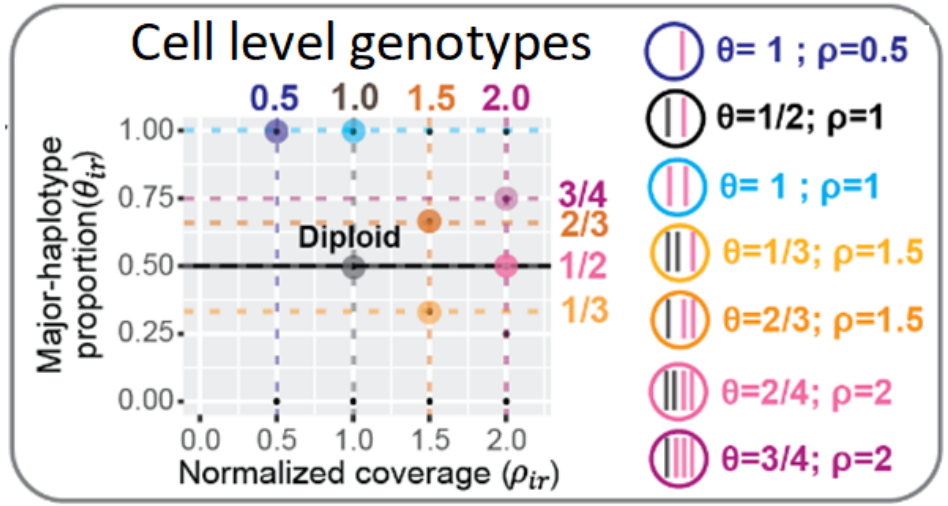
$$\boldsymbol{\tau}_i \subseteq \{\tau^{(1)}, \dots, \tau^{(K)}\},$$

and μ_t is the background rate.

If a read Y_{ij} overlaps with a heterozygous site (known a priori), let Z_{ij} be indicator of whether it carries the alternative ("B") allele,

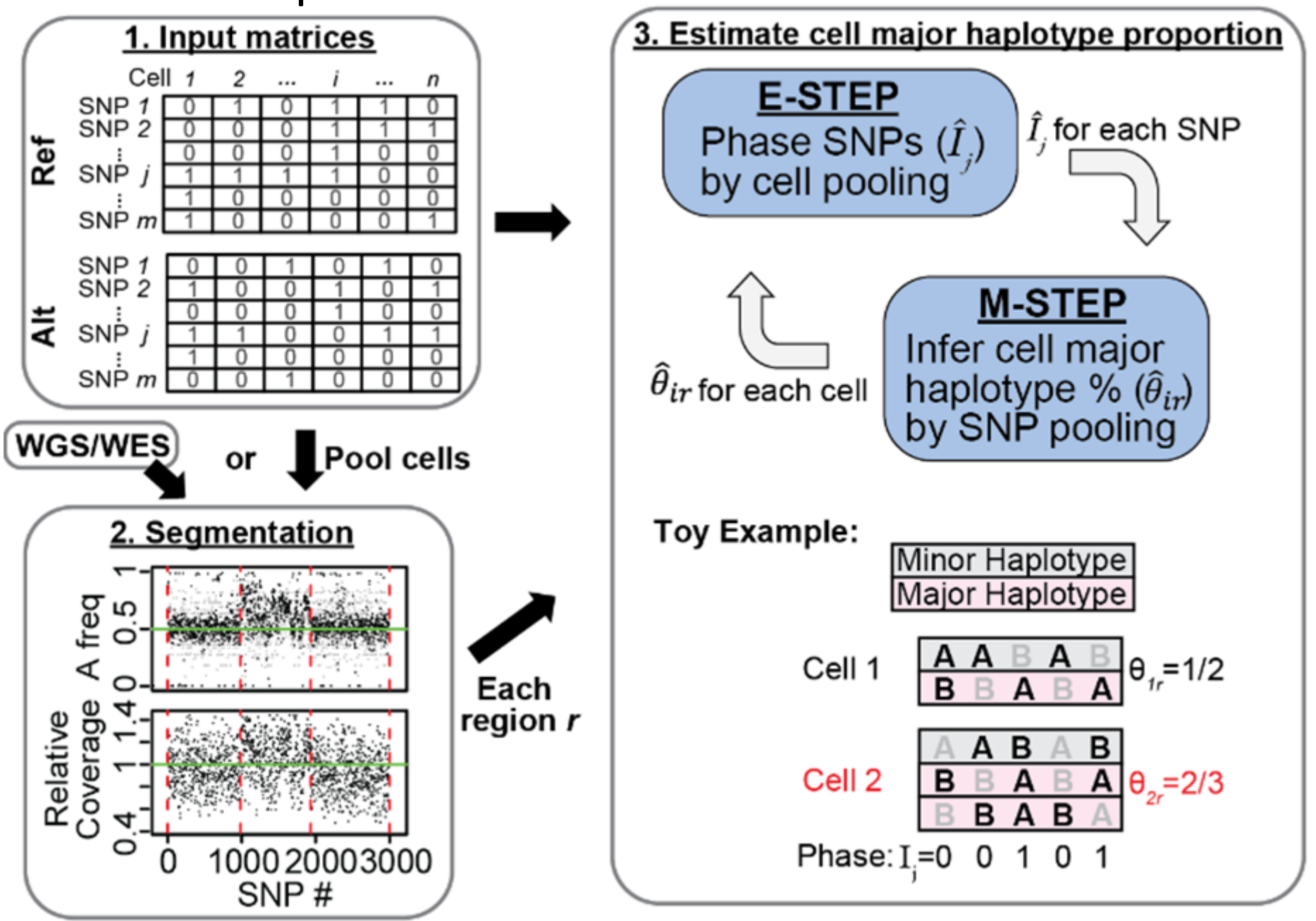
$$Z_{ij} \sim Bernoulli\left(\theta_{iY_{ij}} I_{Y_{ij}} + (1 - \theta_{iY_{ij}})(1 - I_{Y_{ij}})\right),$$

Where for any position t, θ_{it} is the major haplotype proportion of cell i, and I_t is indicator of whether allele B is on the major haplotype ("phase").

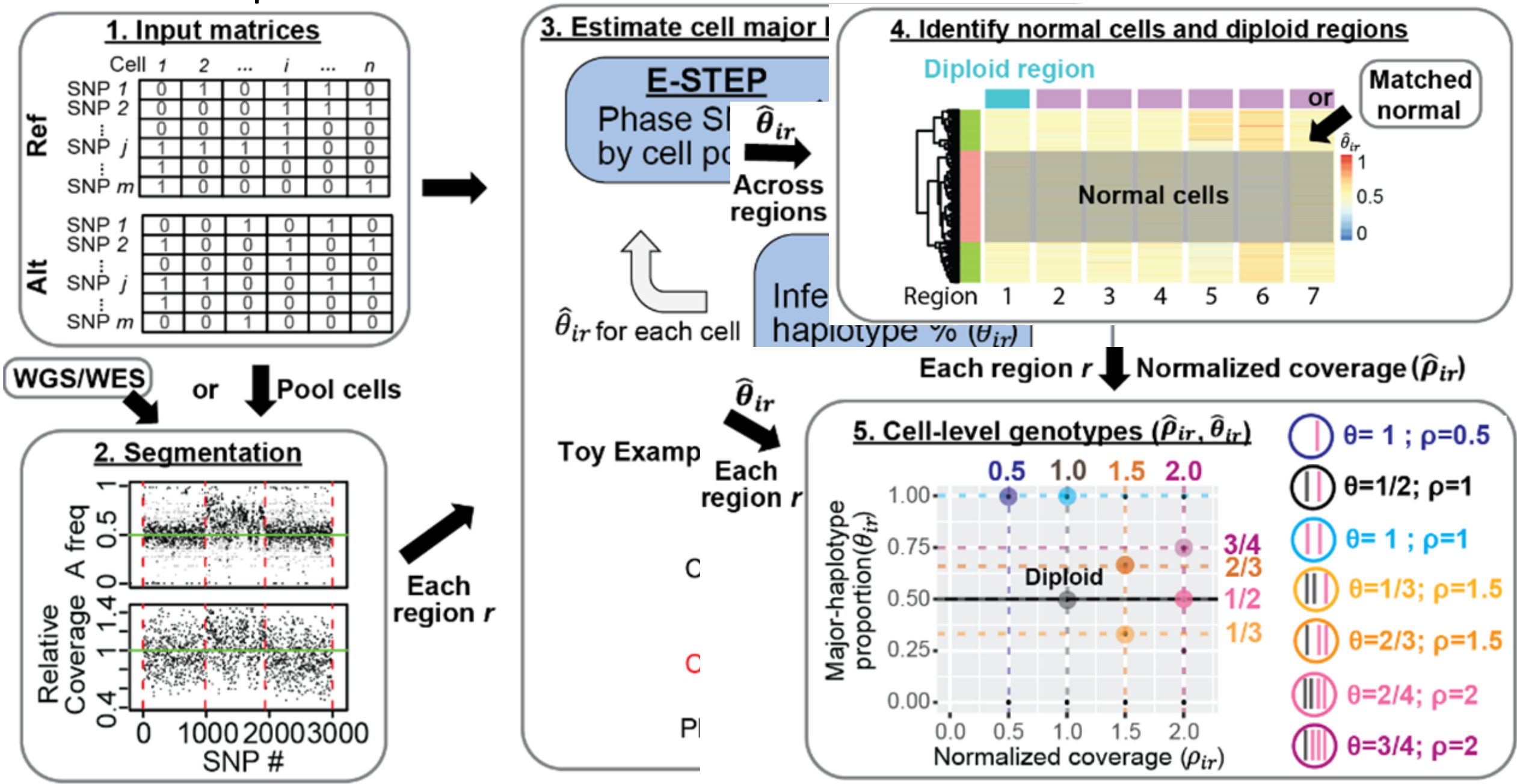


Challenges: (1) Reads per cell are **sparse**. (2) Don't know **where the "breakpoints" are**. (3) Don't know the **underlying phase**.

Alleloscope



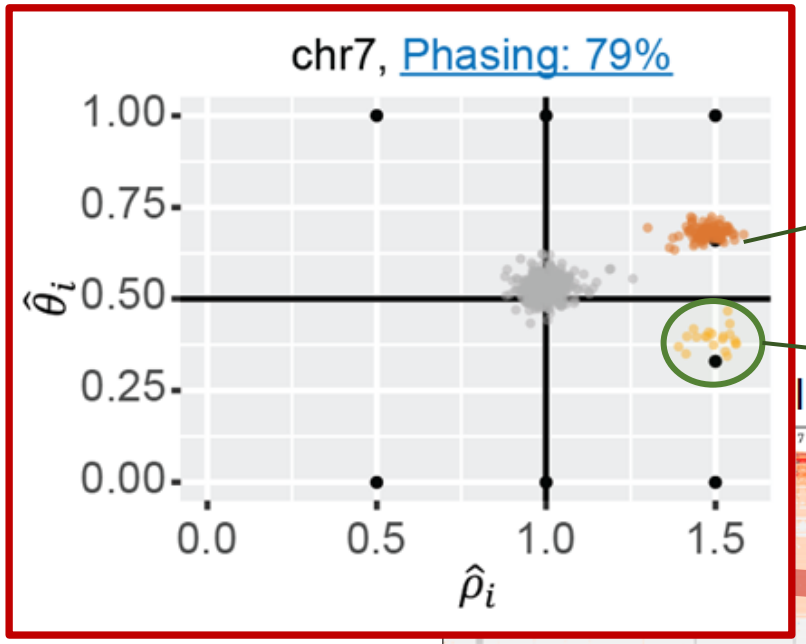
Alleloscope

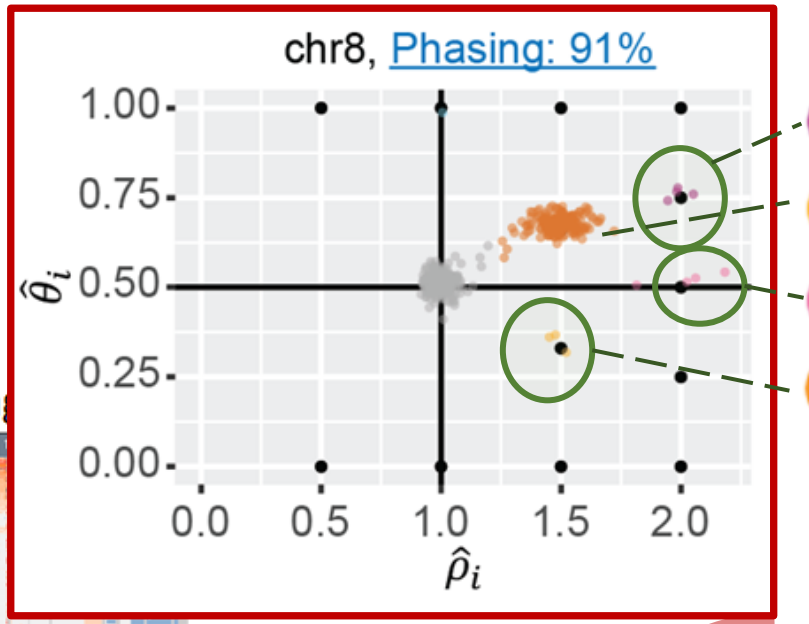


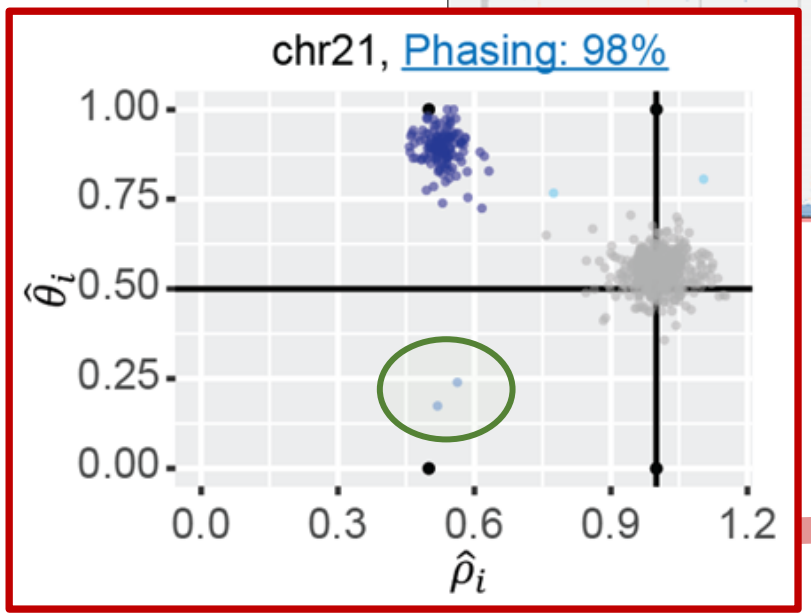
What can we see using Alleloscope?

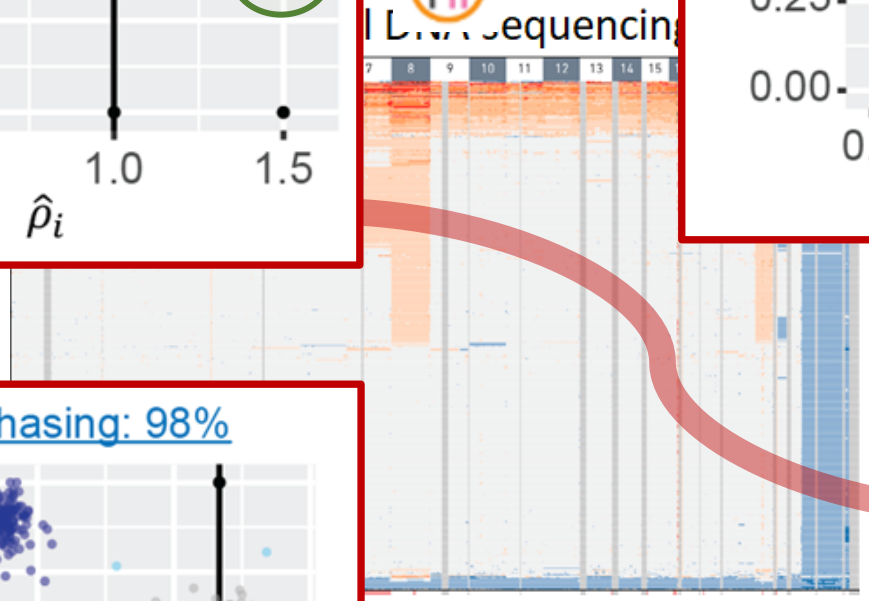
Gastric tumor
MSI subtype

800 cells
~700K reads/cell





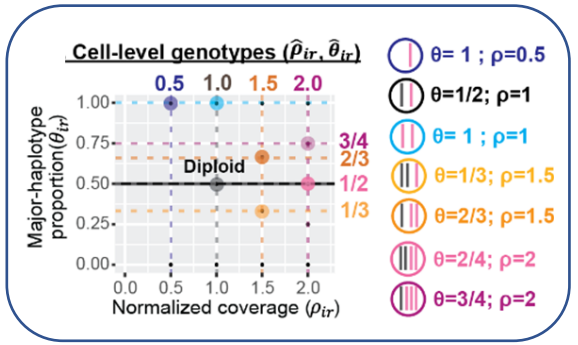




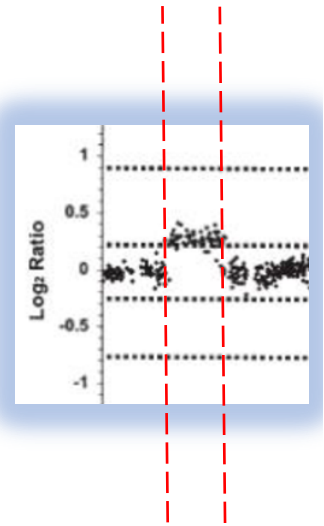
Each dot is a cell.

Each plot is for a region.

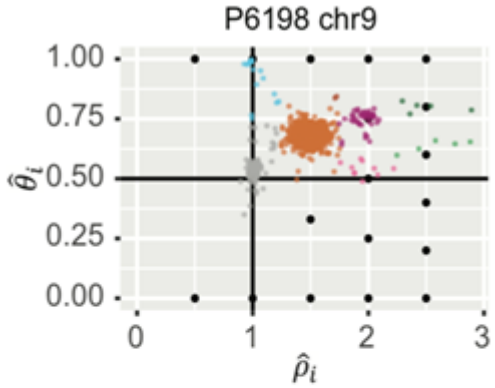
What can we see using Alleloscope?

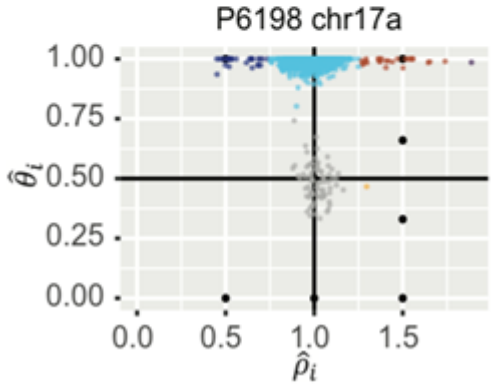


Each dot is a cell.
Each plot is for a region.

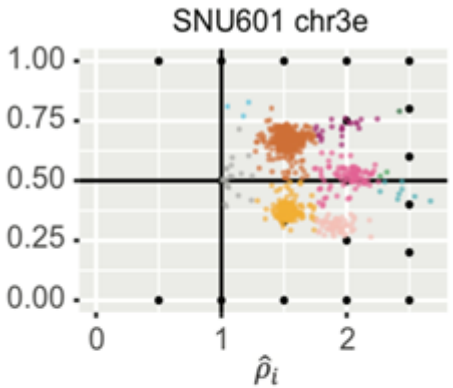


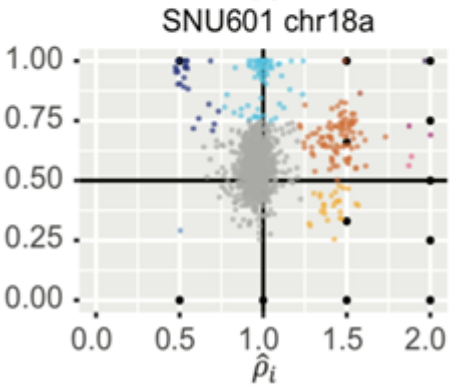
Colon cancer



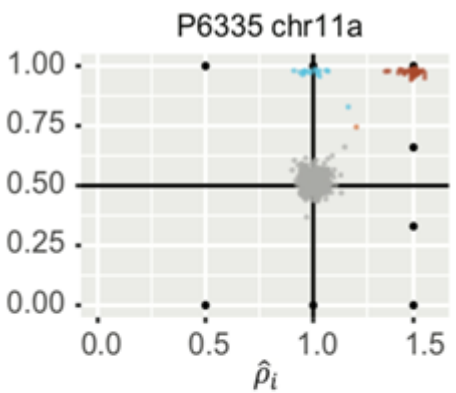


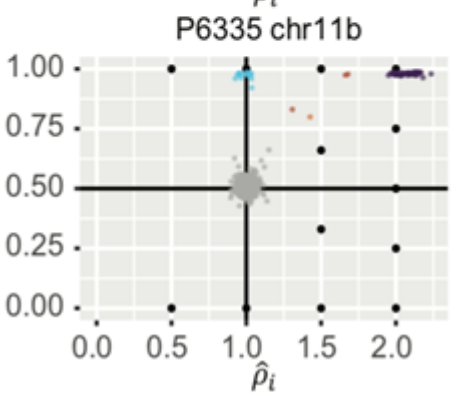
Gastric cancer Cell line



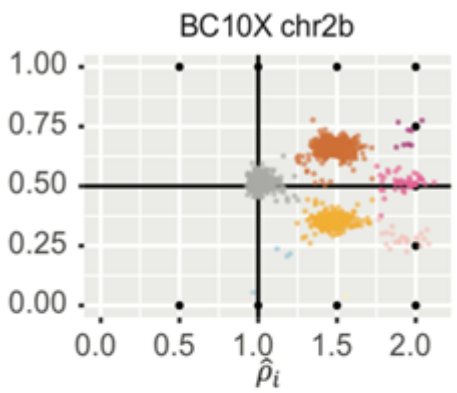


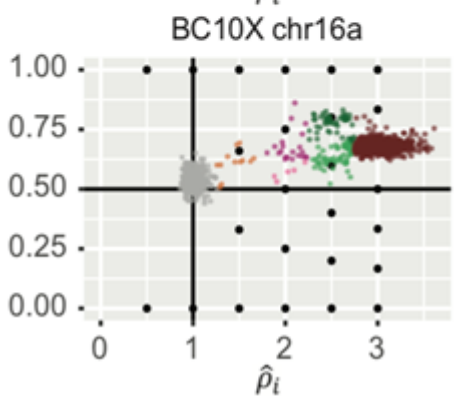
Colon cancer





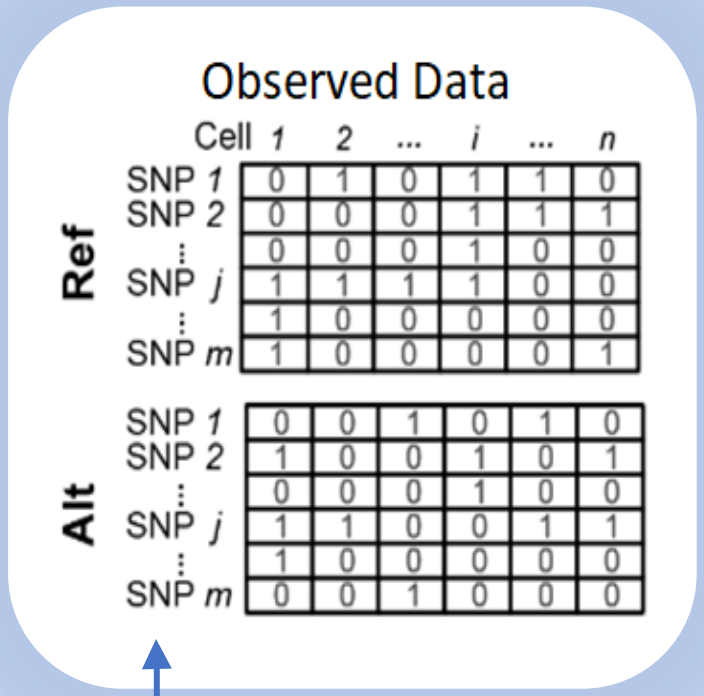
Breast cancer

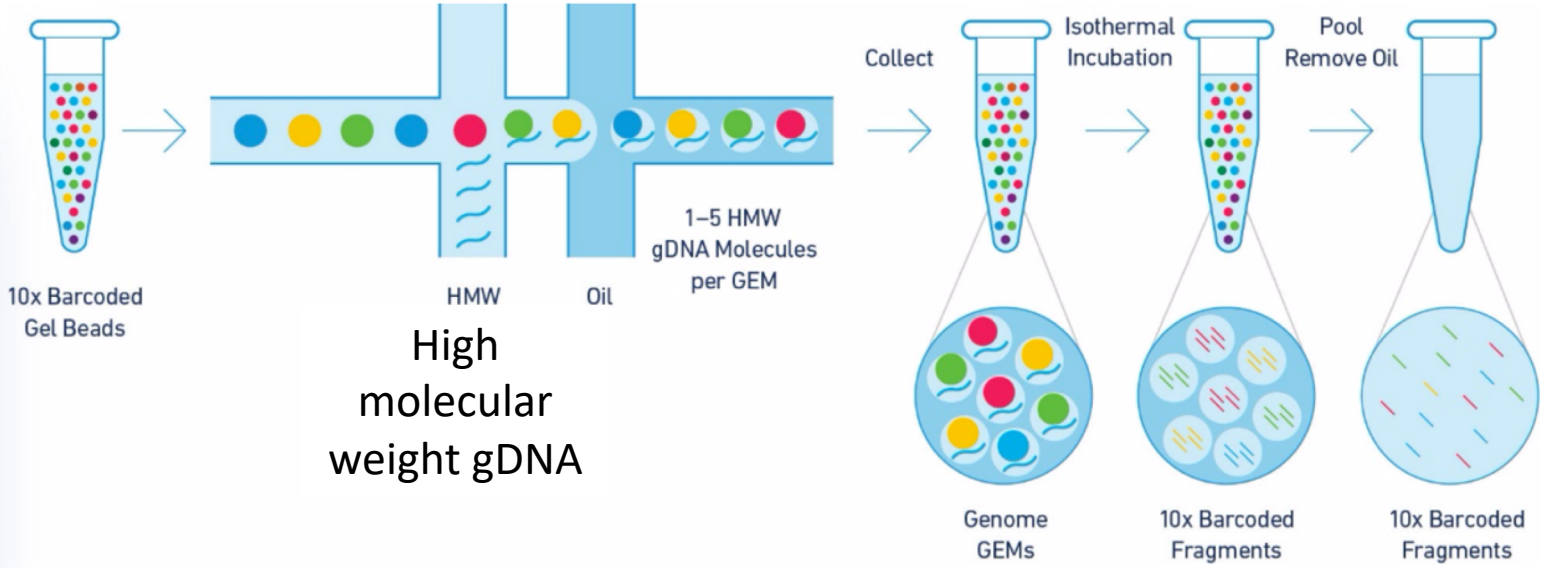


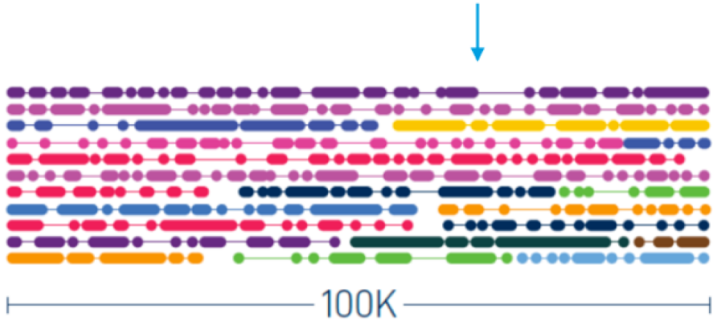


Are these clusters real or artifacts?

Validation by matched bulk linked-reads sequencing (a.k.a. “haplotype sequencing”)



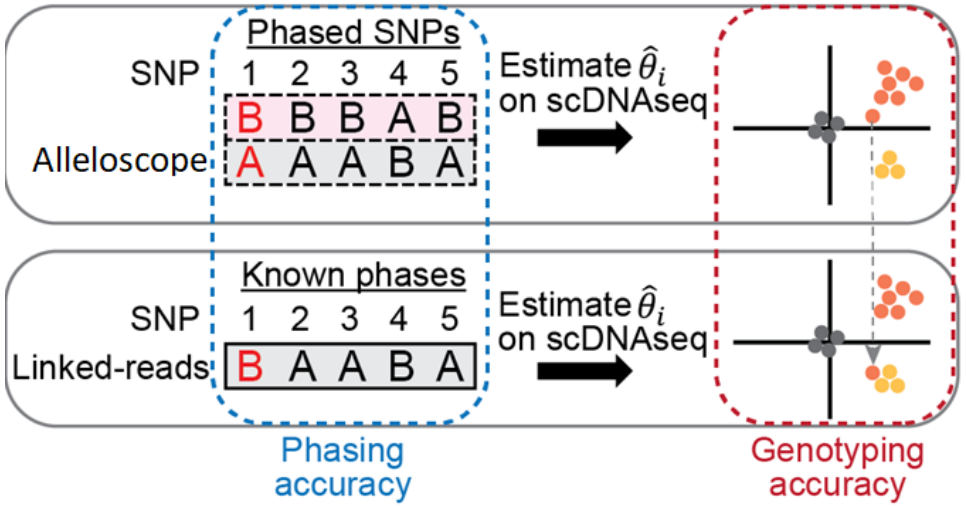




Linked-Reads

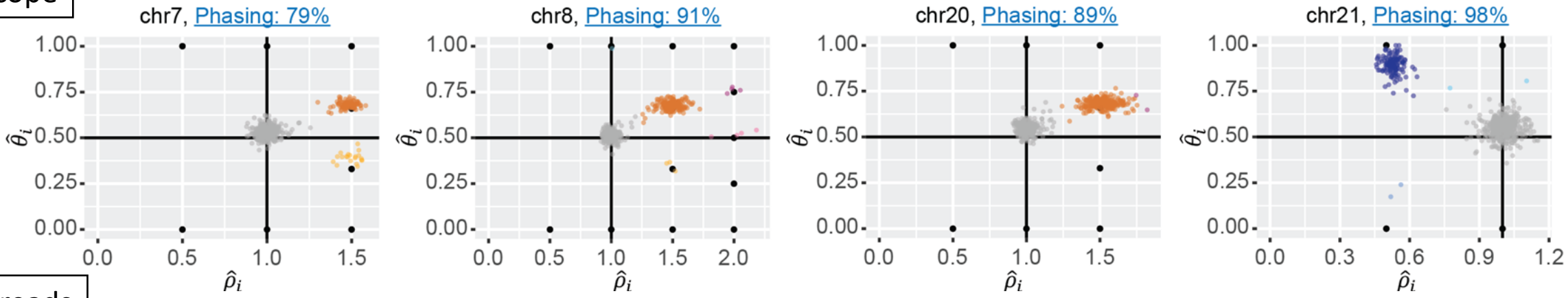
Phase SNPs

Validation by Linked-reads sequencing

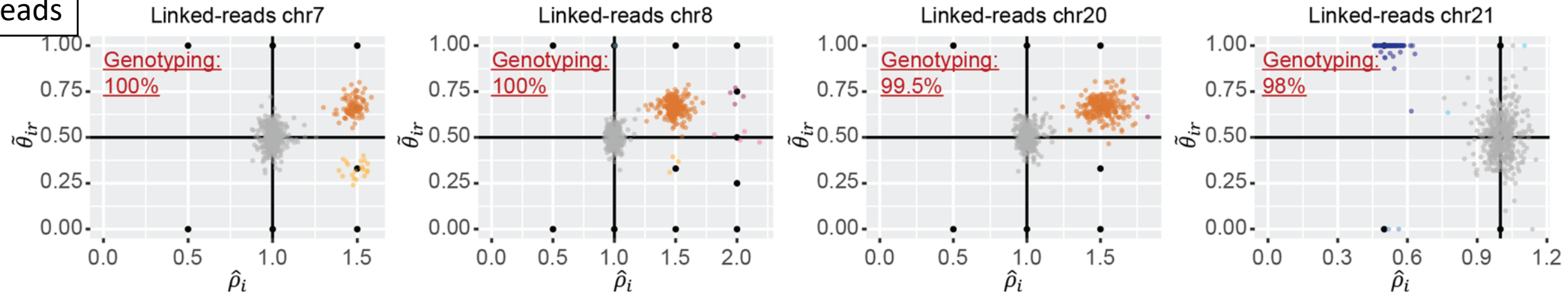


Each dot is a cell.
Each plot is for a region.

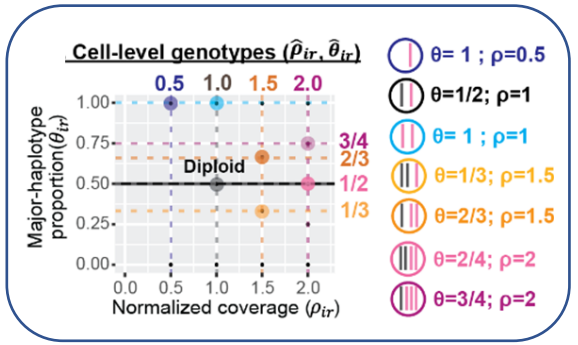
Alleloscope



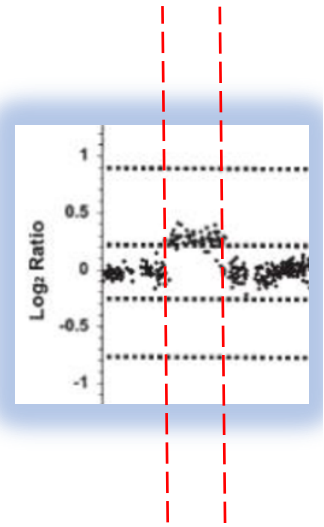
Linked reads



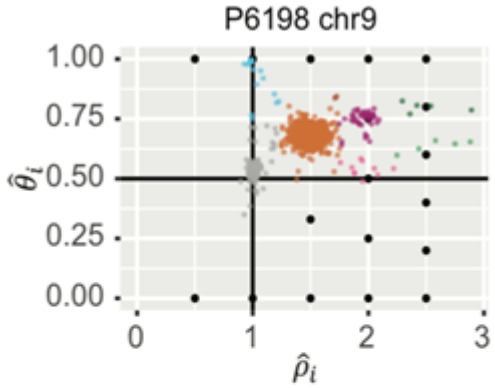
What can we see using Alleloscope?

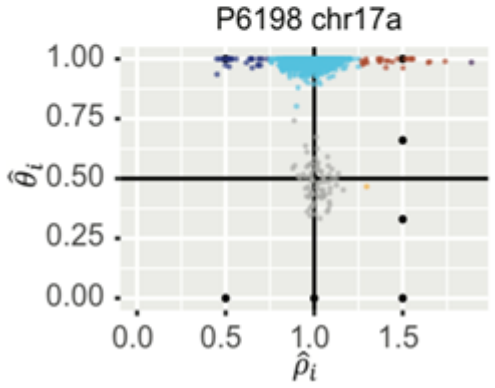


Each dot is a cell.
Each plot is for a region.

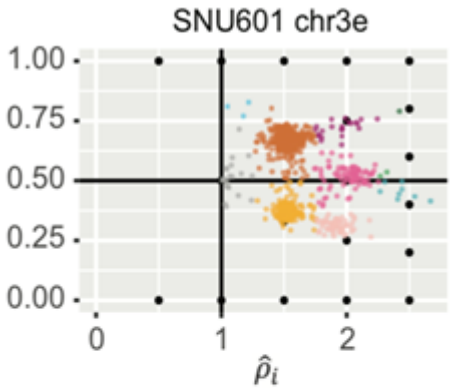


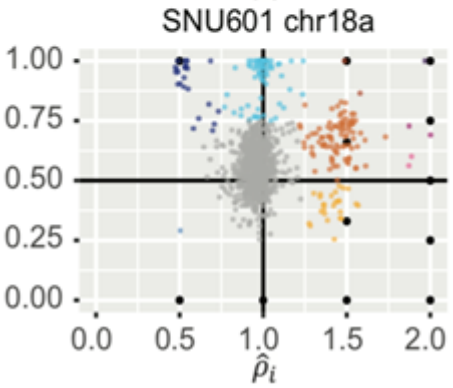
Colon cancer



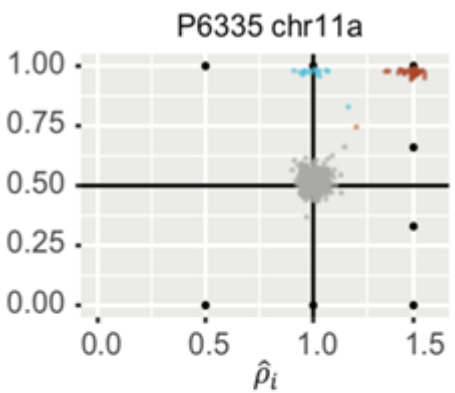


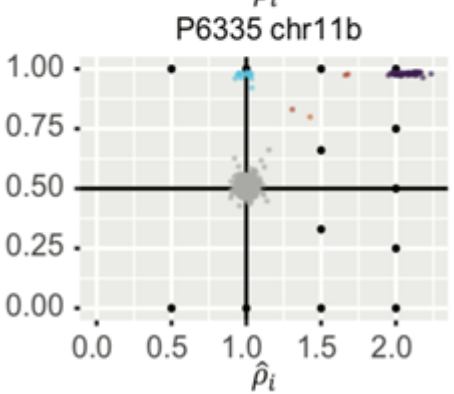
Gastric cancer Cell line



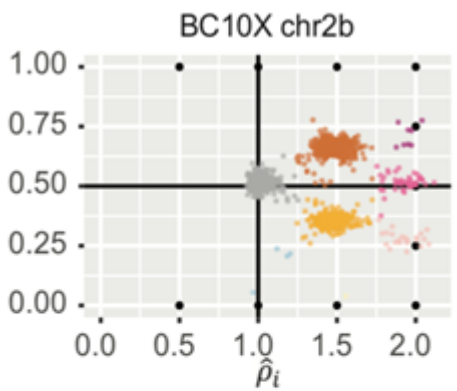


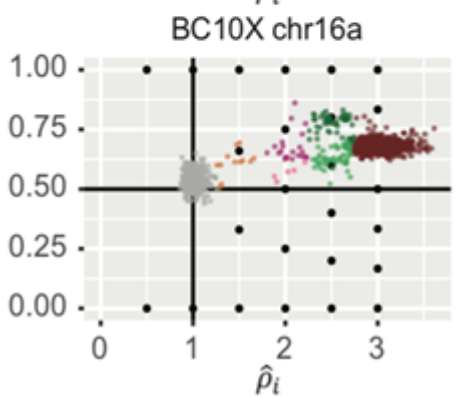
Colon cancer





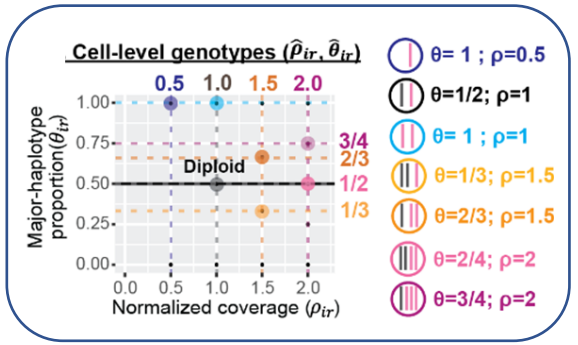
Breast cancer



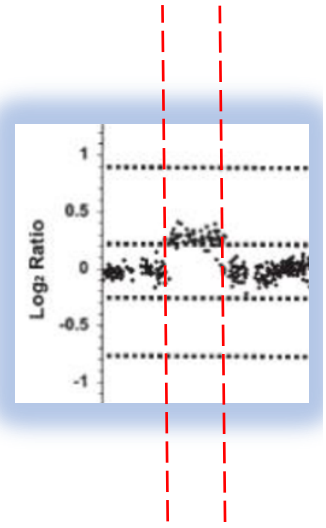


Clusters validated by linked-reads sequencing.

What can we see using Alleloscope?



Each dot is a cell.
Each plot is for a region.

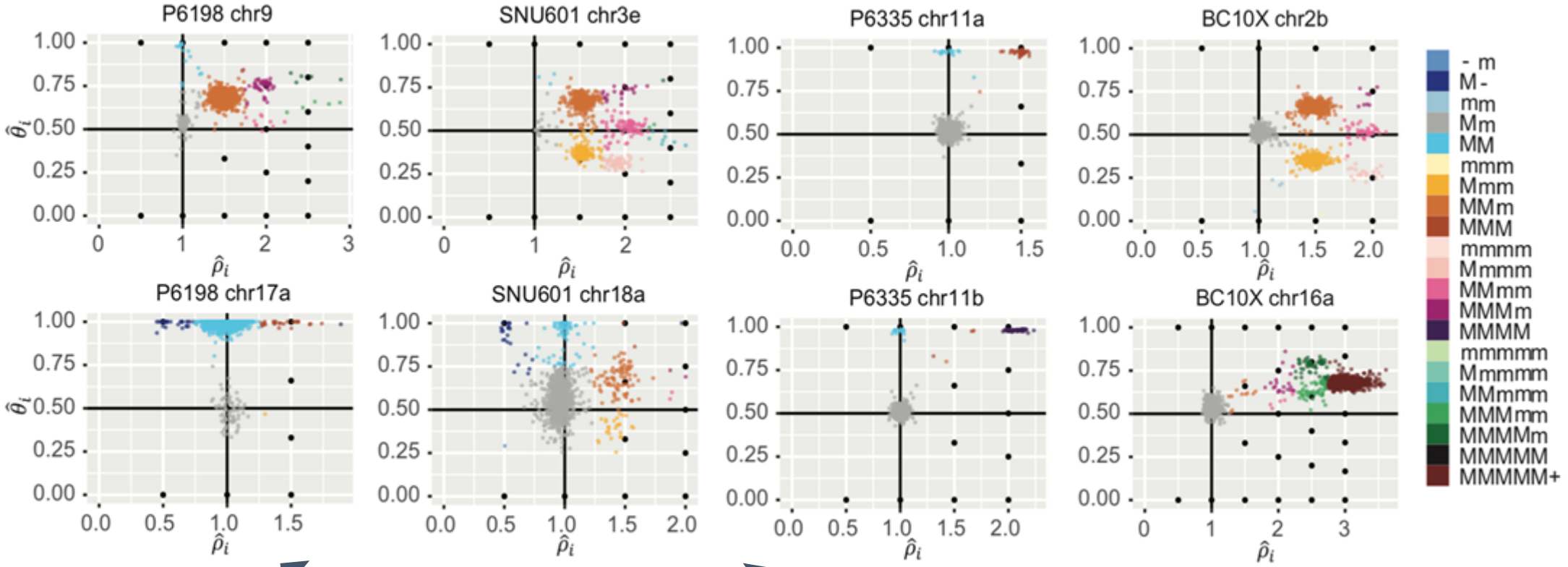


Colon cancer

Gastric cancer Cell line

Colon cancer

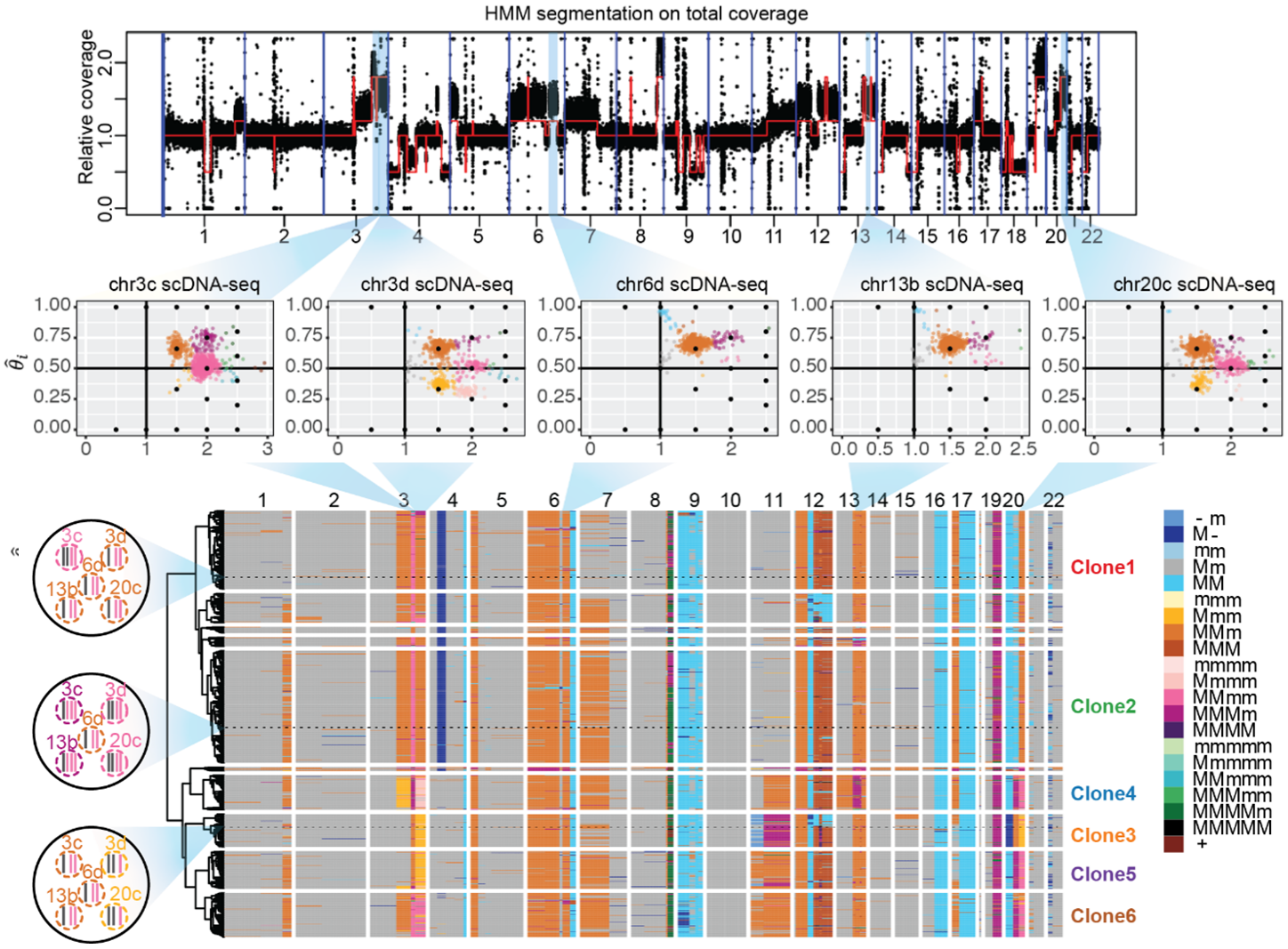
Breast cancer

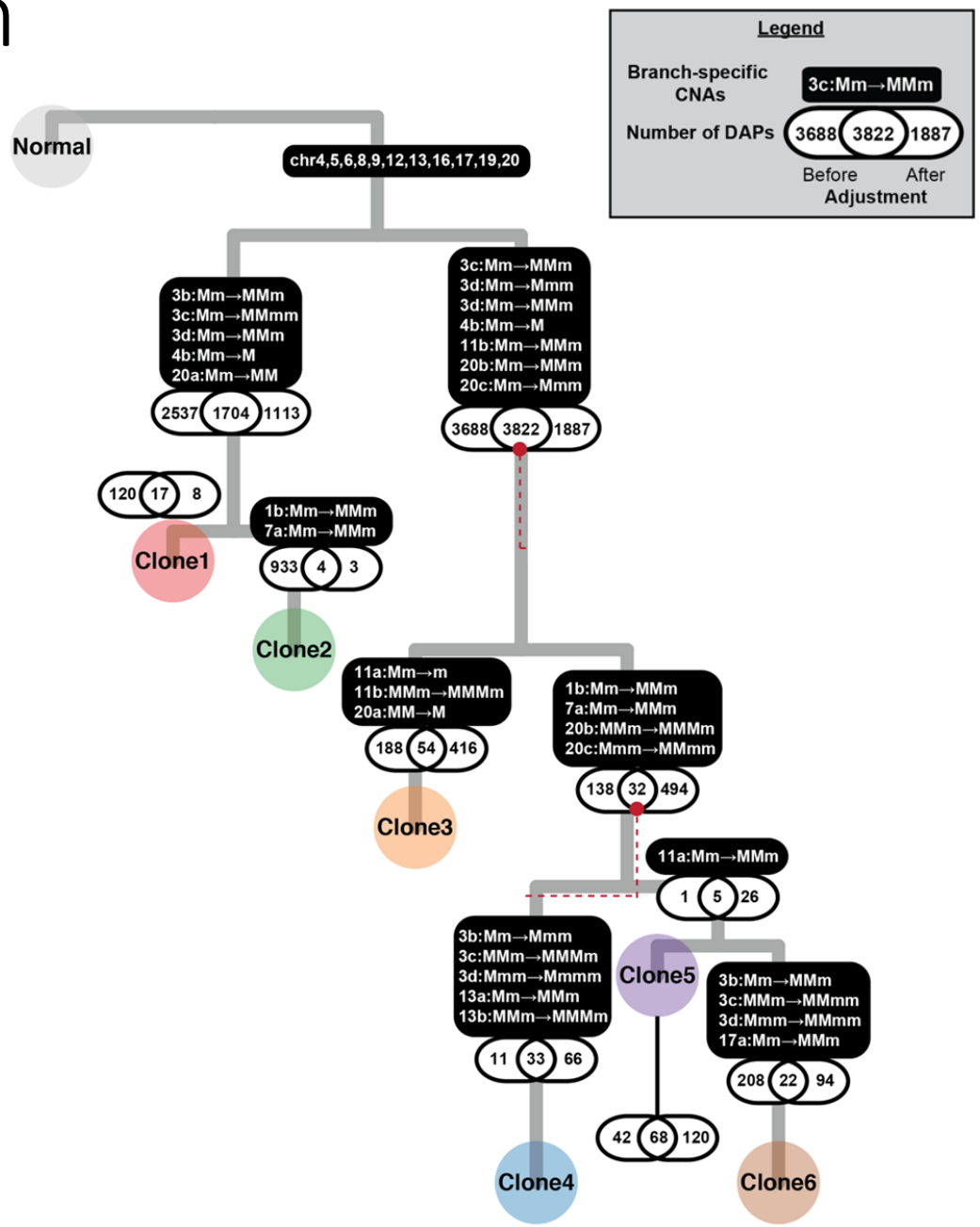


metastasis

Watkins et al. (2020) ***Pervasive chromosomal instability and karyotype order in tumour evolution***, Nature 587, 126.

Detailed study of tumor evolution





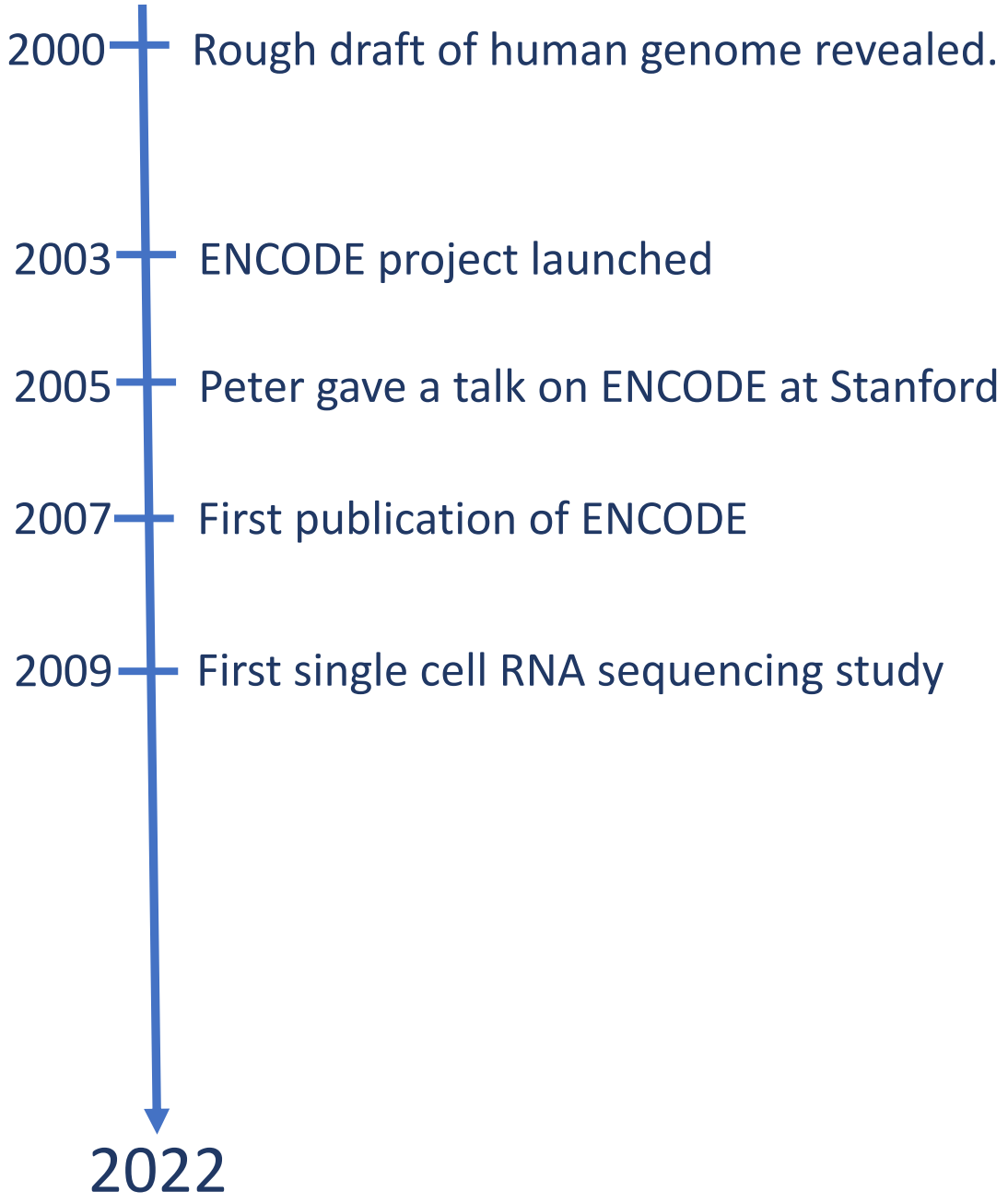
The rest of this talk:

- Single cell allele-specific copy number estimation

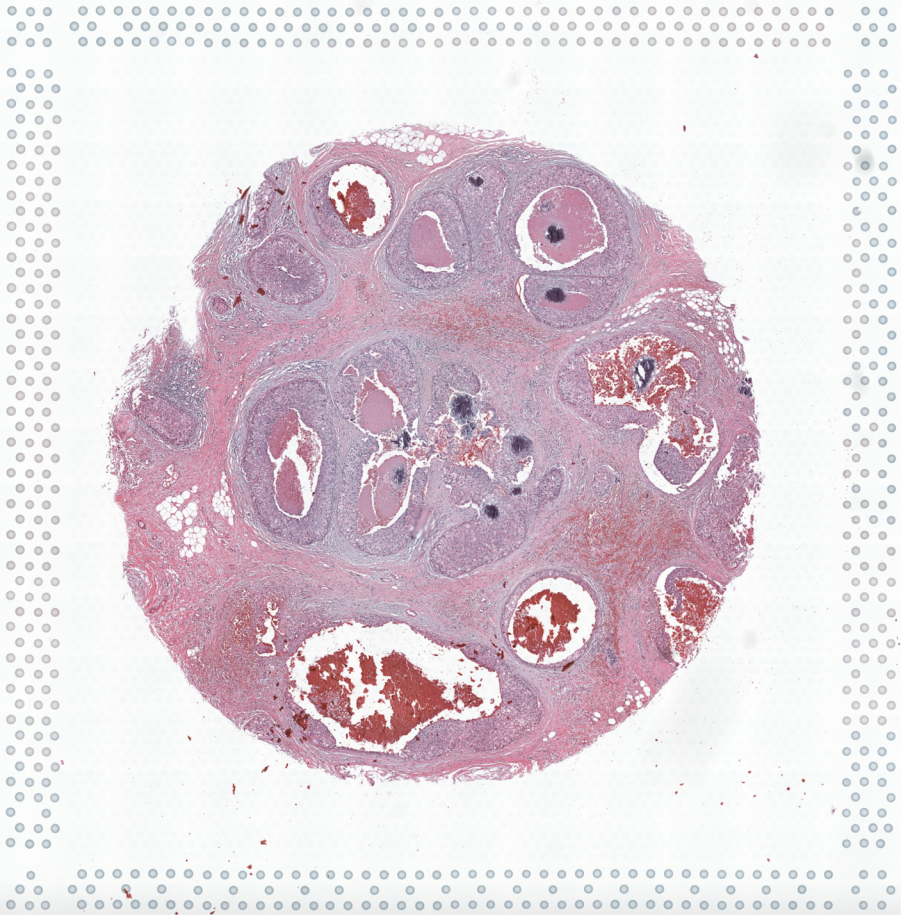
  Chi-Yun Wu et al. Integrative single-cell analysis of allele-specific copy number alterations and chromatin accessibility in cancer. ***Nature Biotechnology 2021***.

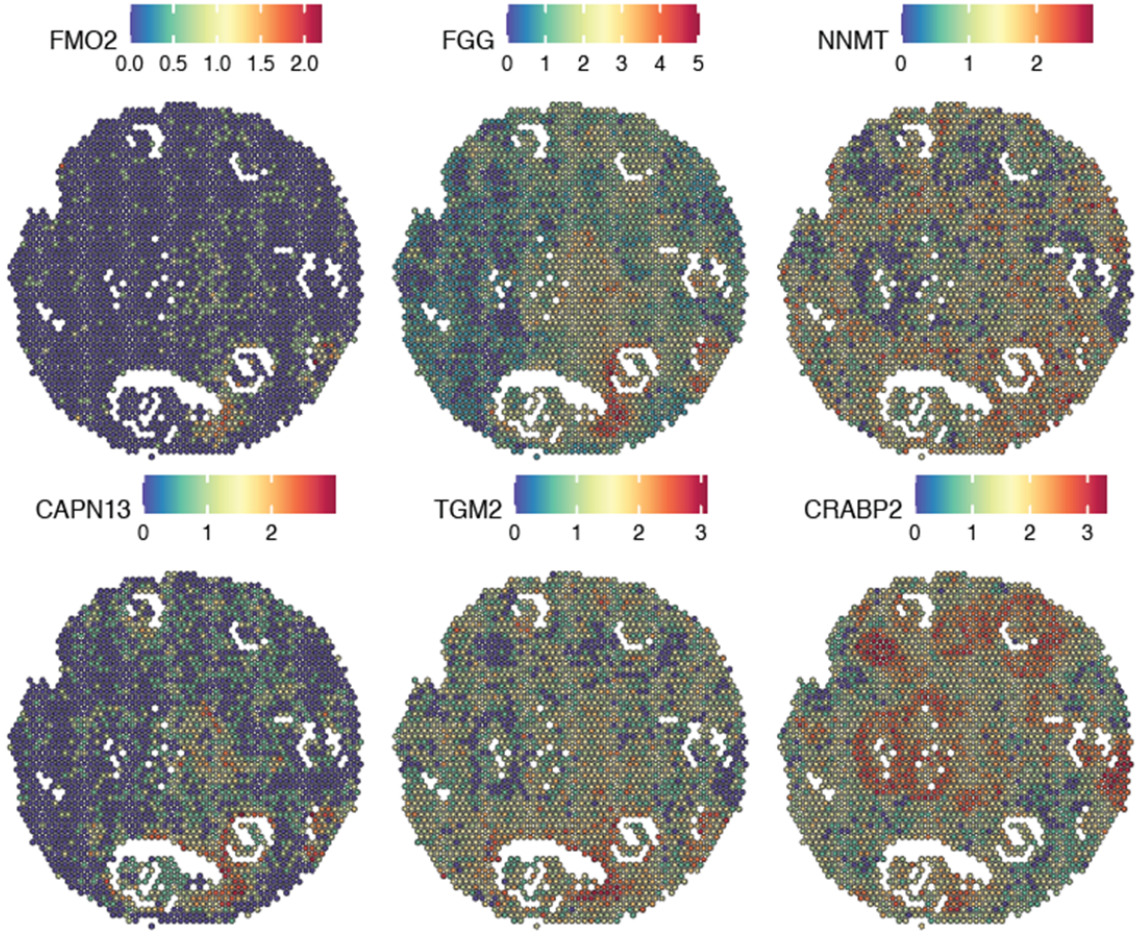
- Cancer subclone detection in spatial transcriptomic data

  Chi-Yun Wu et al. Subclone detection on copy number profiles in single cell and spatial tumor sequencing data. ***Under preparation 2022***



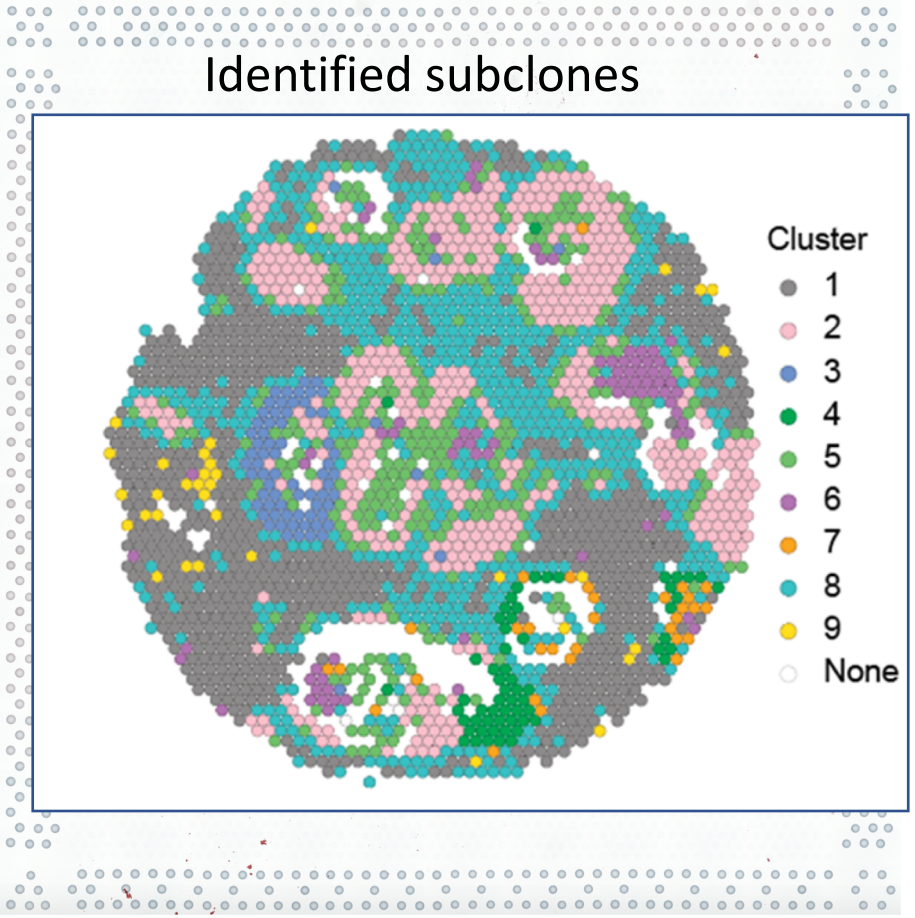
Spatial transcriptomic data

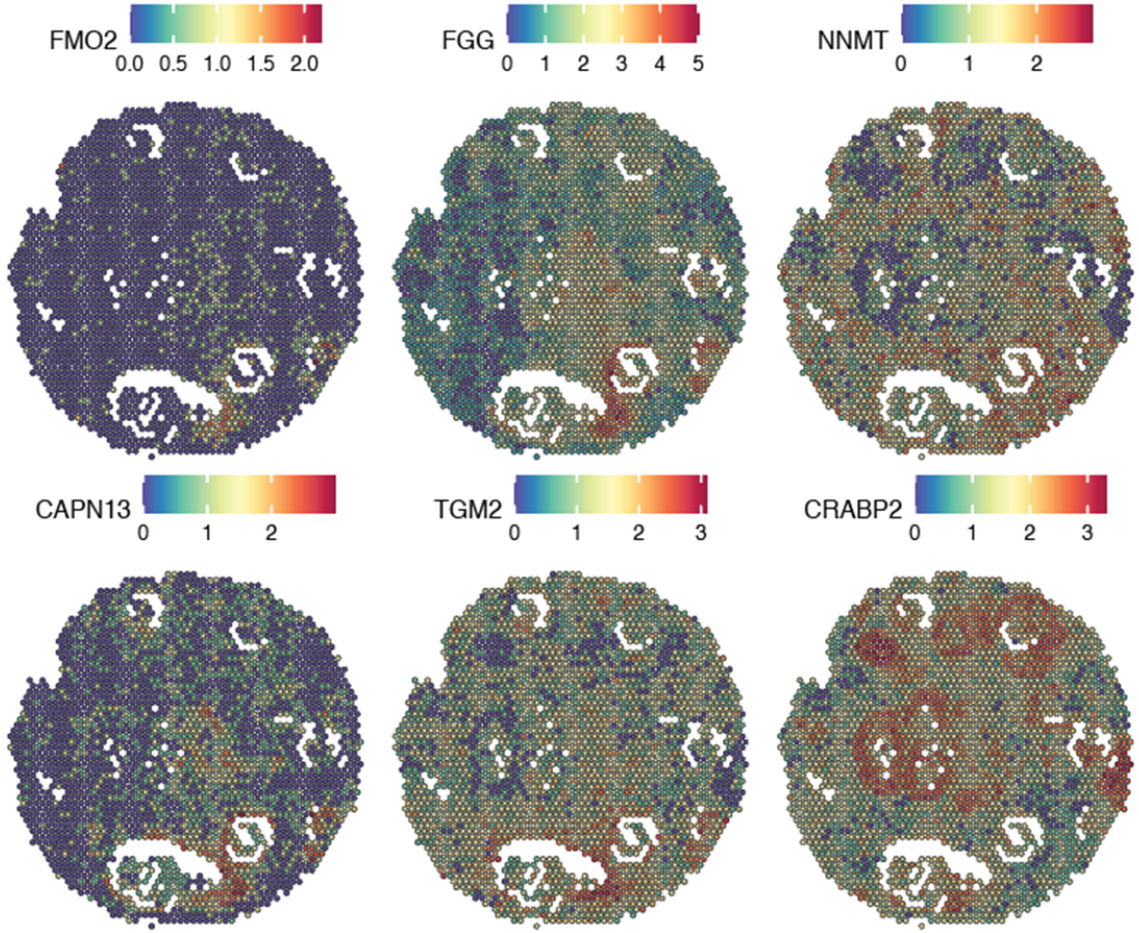




Full transcriptome = ~ 20,000 measurements per spot

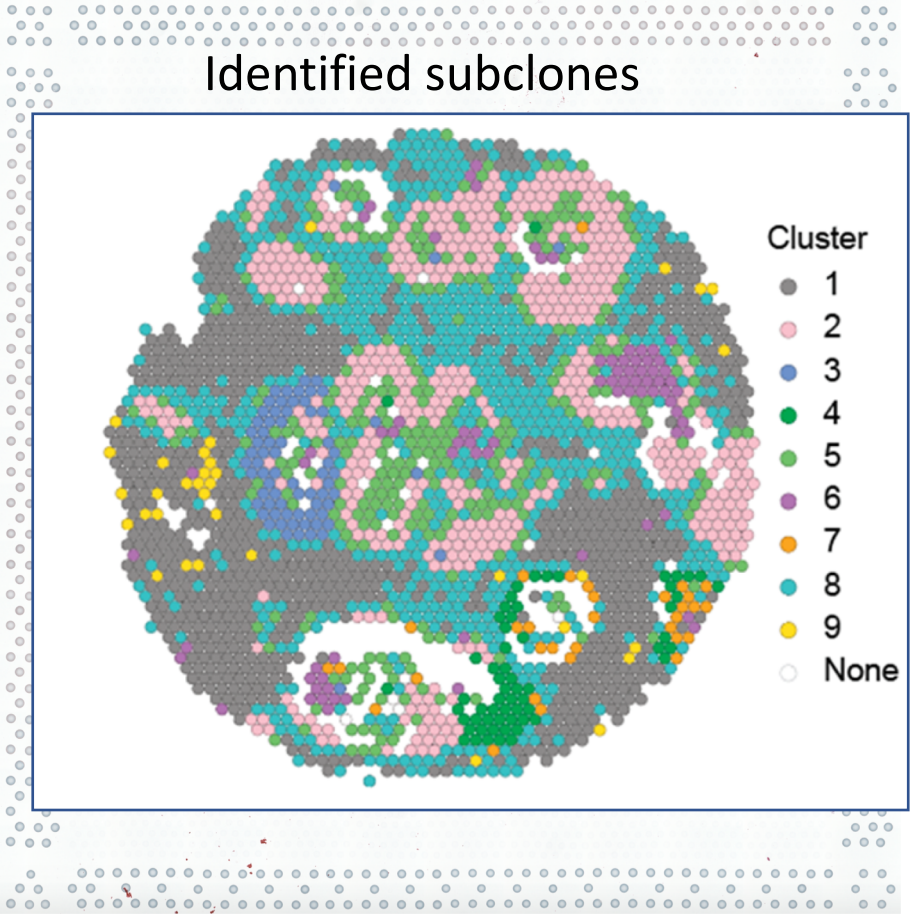
Spatial transcriptomic data





Full transcriptome = ~ 20,000 measurements per spot

Spatial transcriptomic data



Questions:

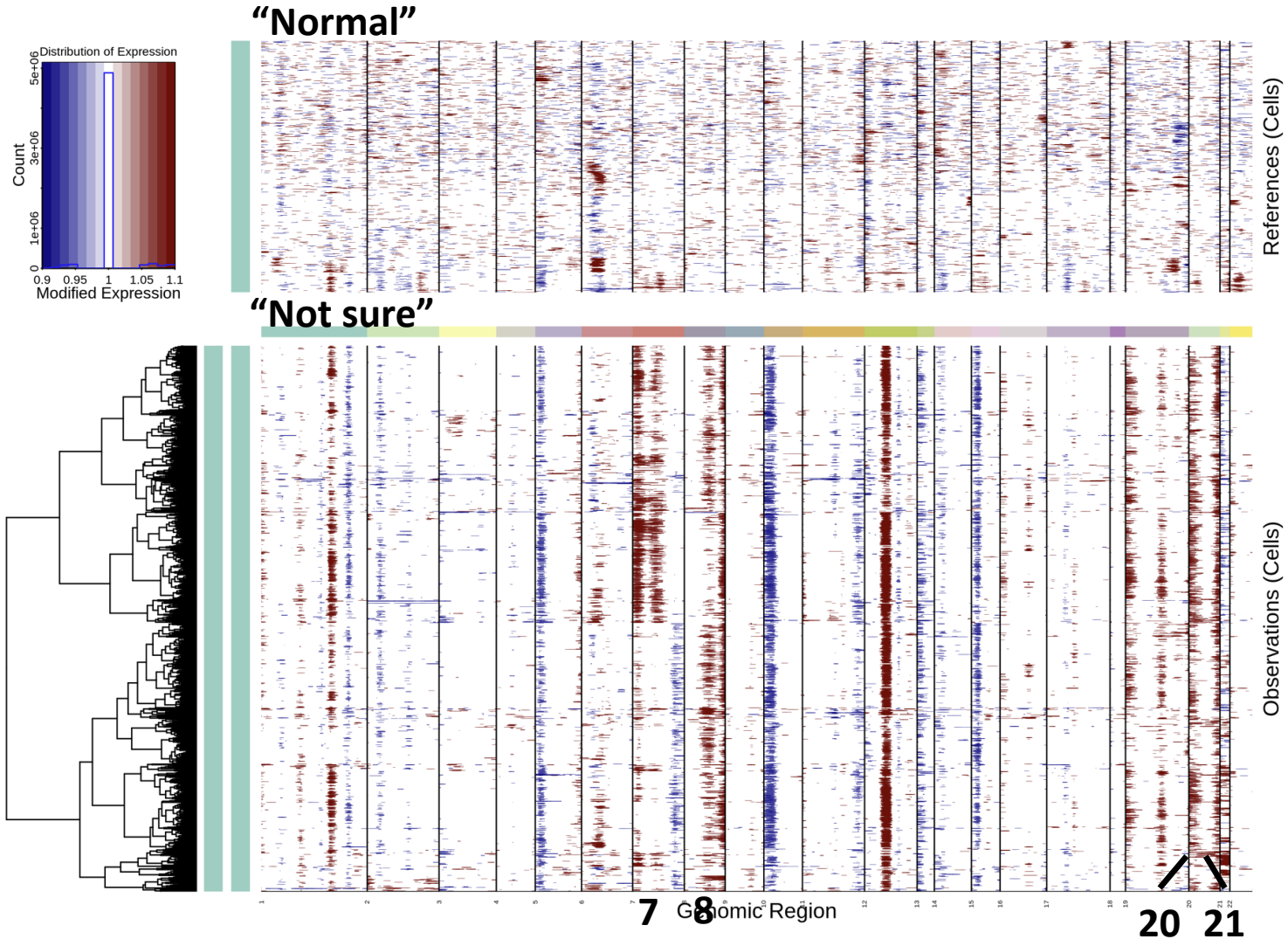
1. Where are the cancer cells?
2. Where are the geneticially and epigenetically distinct subclones of cancer cells?
3. How do these subclones differ?
4. Are subclones mixing in space?
5. How are the subclones interacting with their immune and stromal microenvironment?
6. Is there competition or cooperation between subclones?

Why is copy number estimation in scRNA-seq hard?

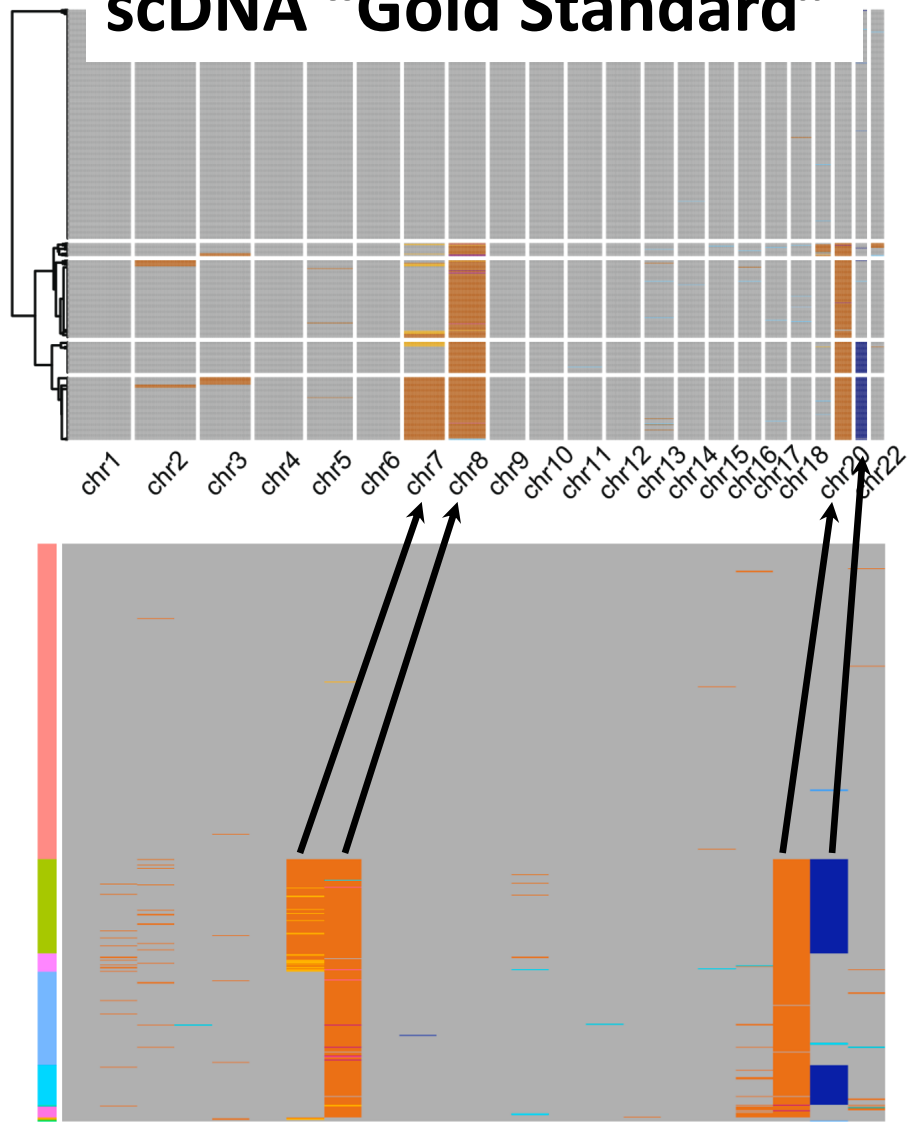
- Gene expression is only a proxy for underlying DNA copy number
- Have to deal with transcriptional stochasticity
- Less heterozygous sites in coding regions, thus, less reads contain allelic information

The problem with exiting smoothing-based techniques

inferCNV on single cell RNAseq



scDNA “Gold Standard”

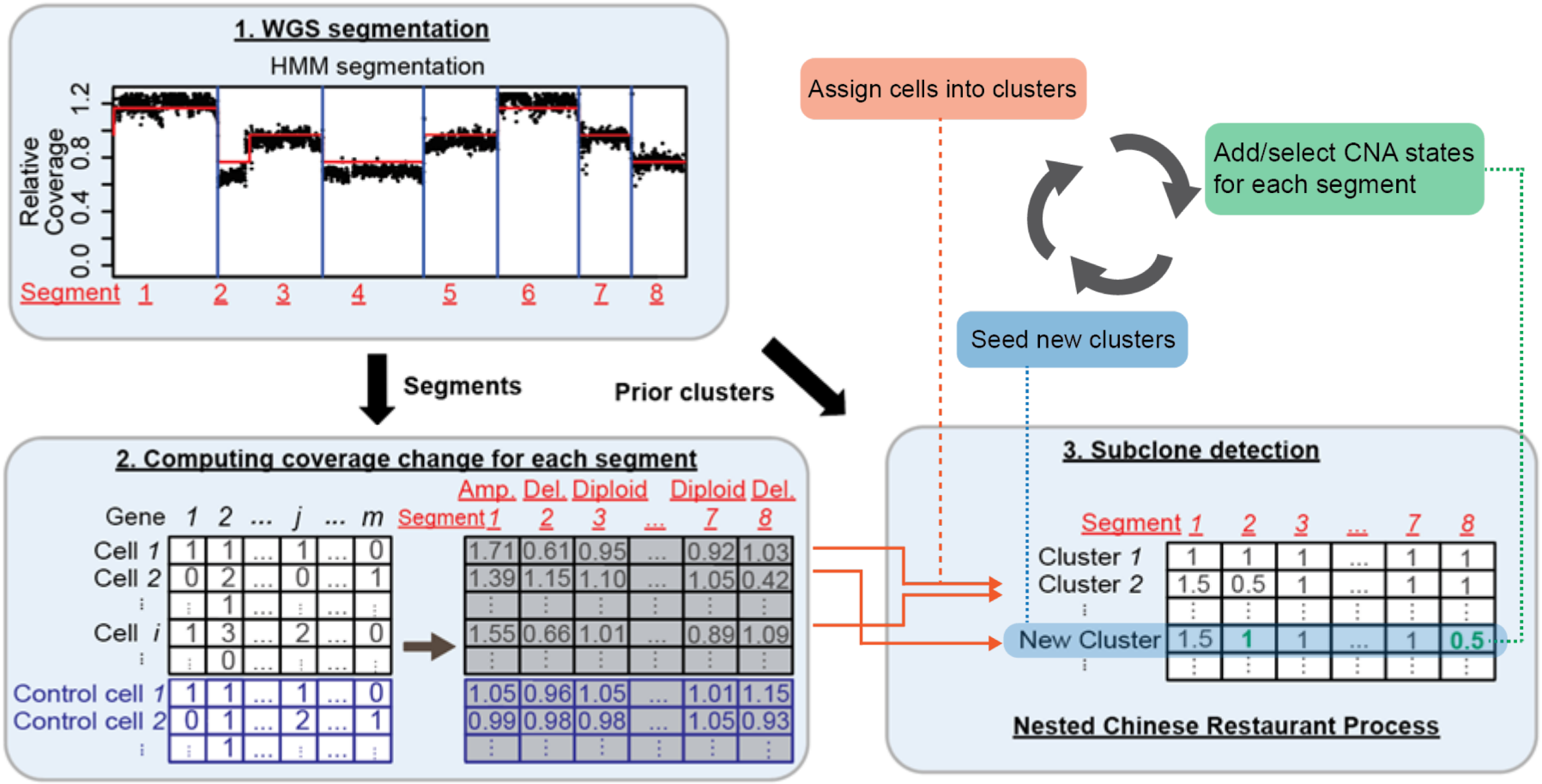


New approach: Clonoscope

Bayesian non-parametric clustering

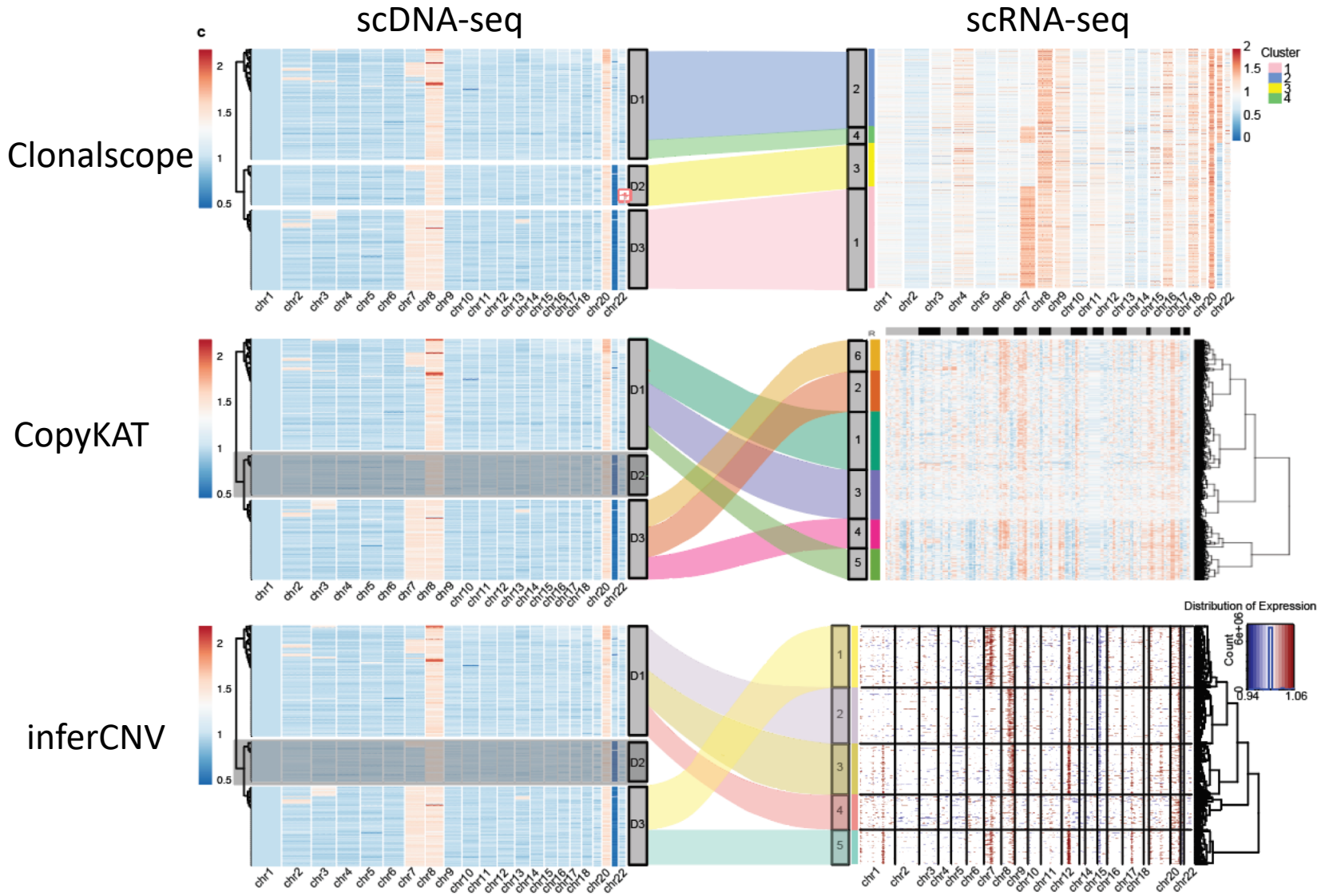
Bulk DNA Sequencing

Single-cell /Spatial Transcriptomic Data



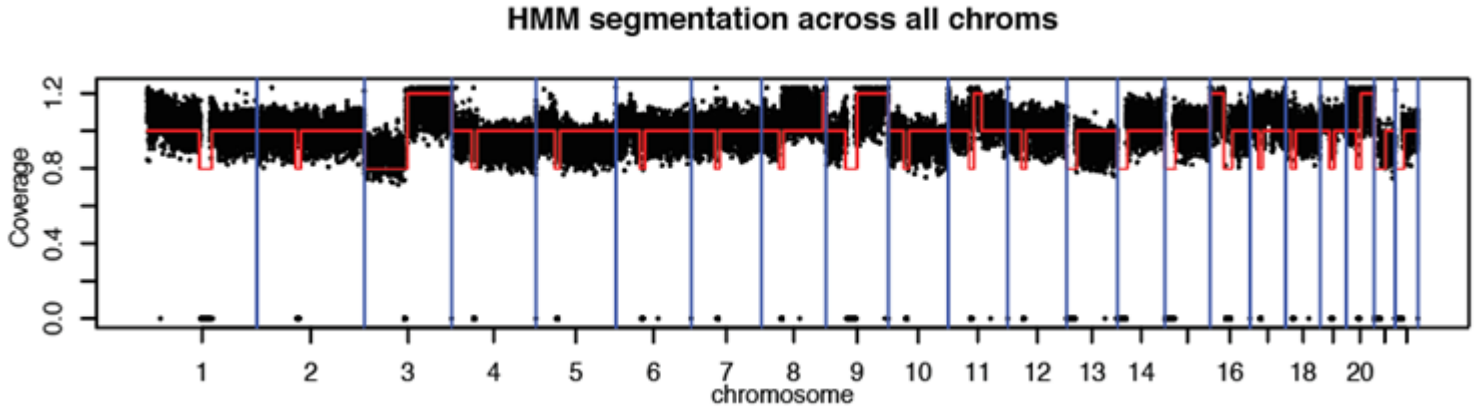
There is a more complicated version of this algorithm that makes use of allele-specific reads.

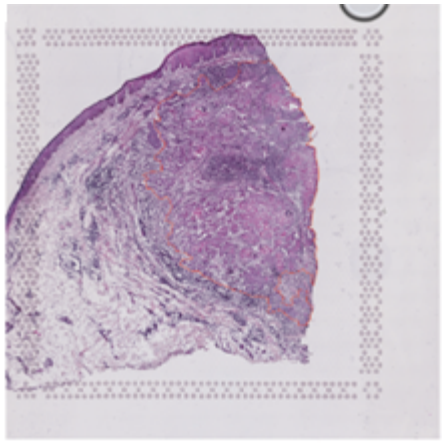
Comparison to scDNA-seq “gold standard

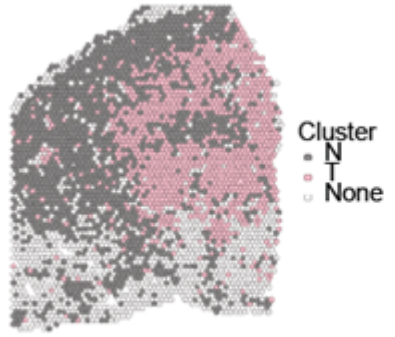


P5931 gastric cancer sample

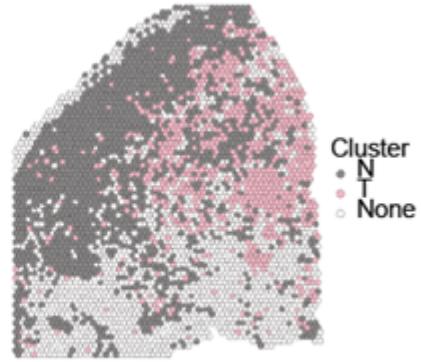
Malignant spot labeling in two adjacent slices of a squamous cell carcinoma sample



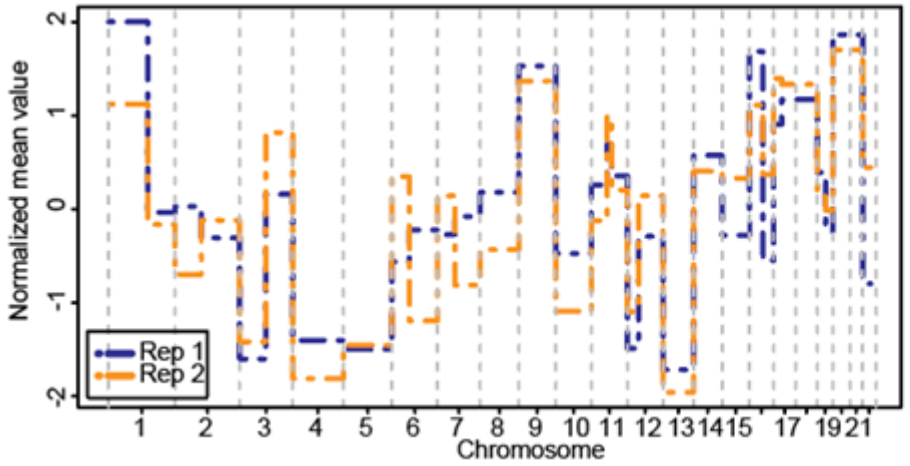


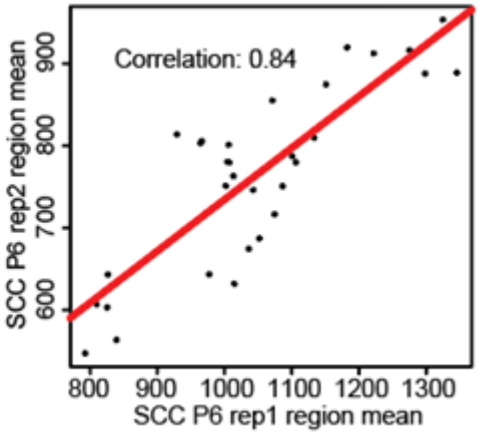


Slice 1

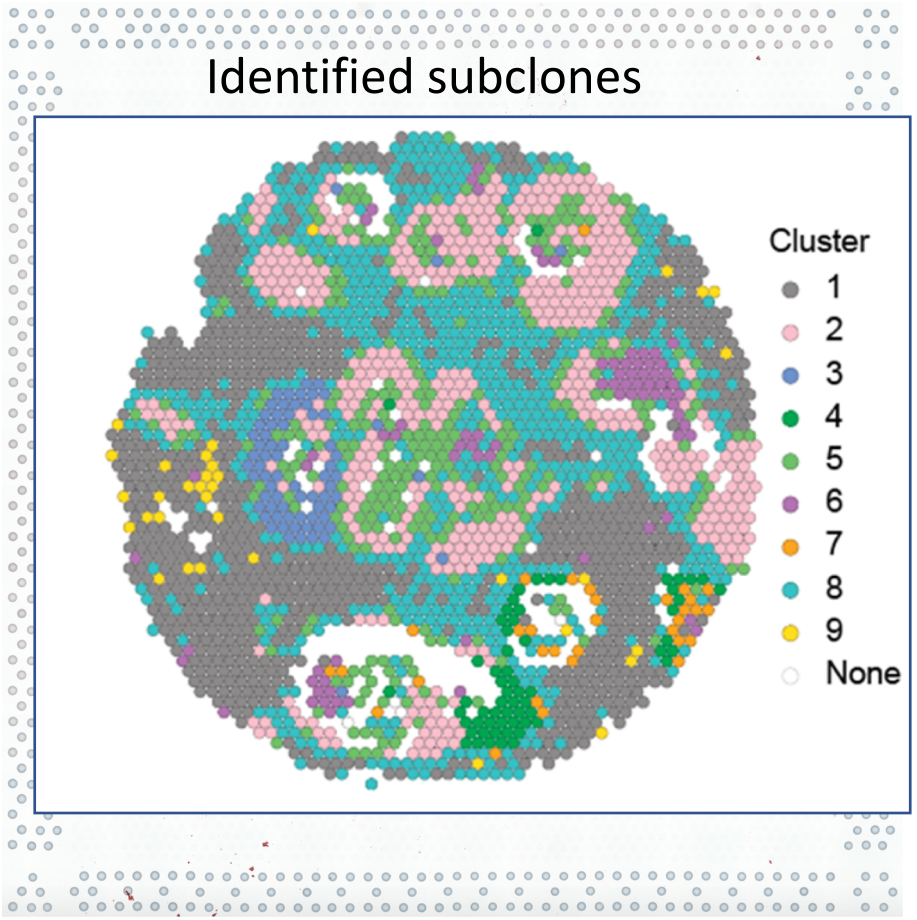


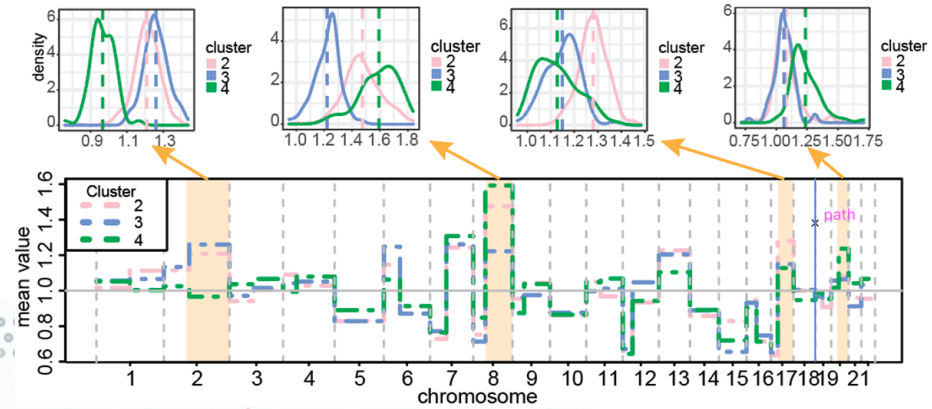
Slice 2

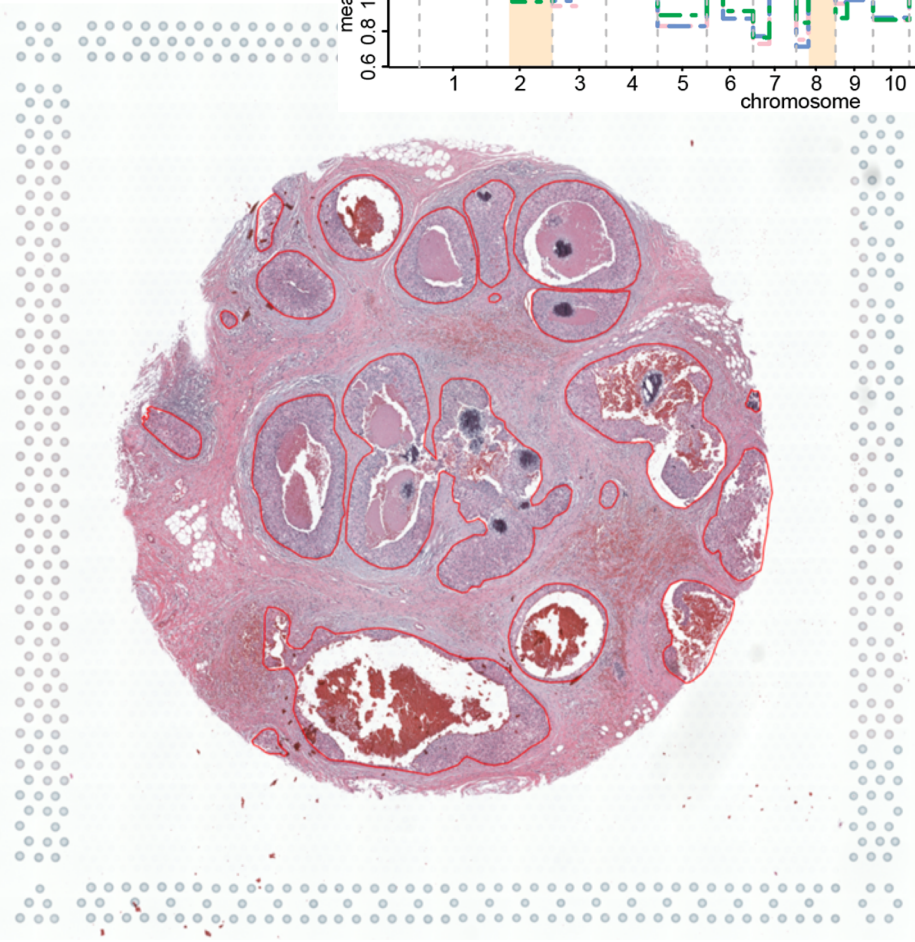




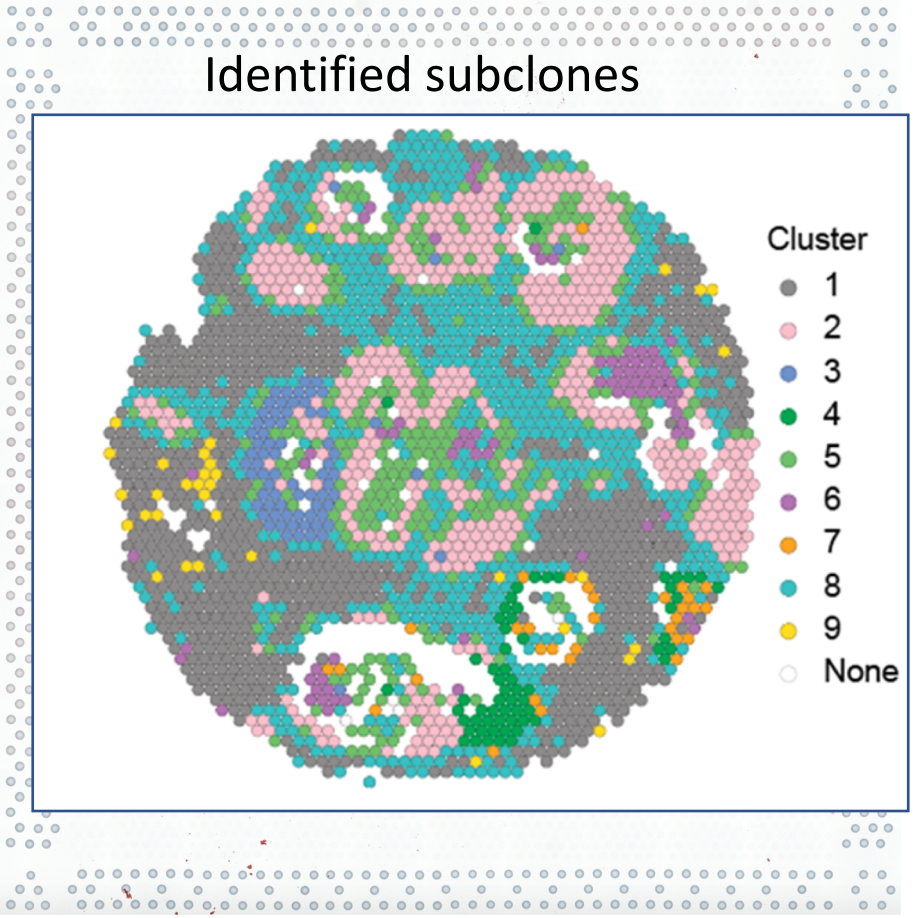
Spatial transcriptomic data

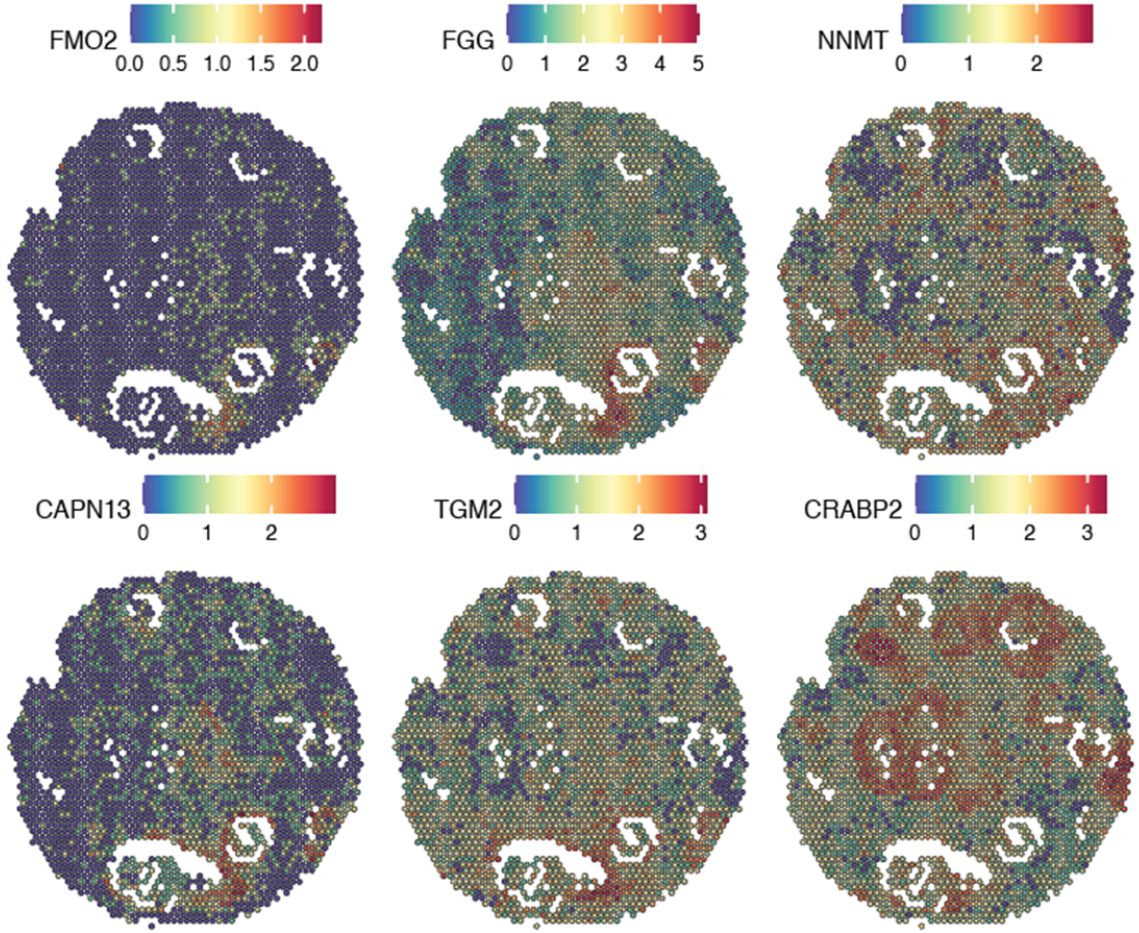




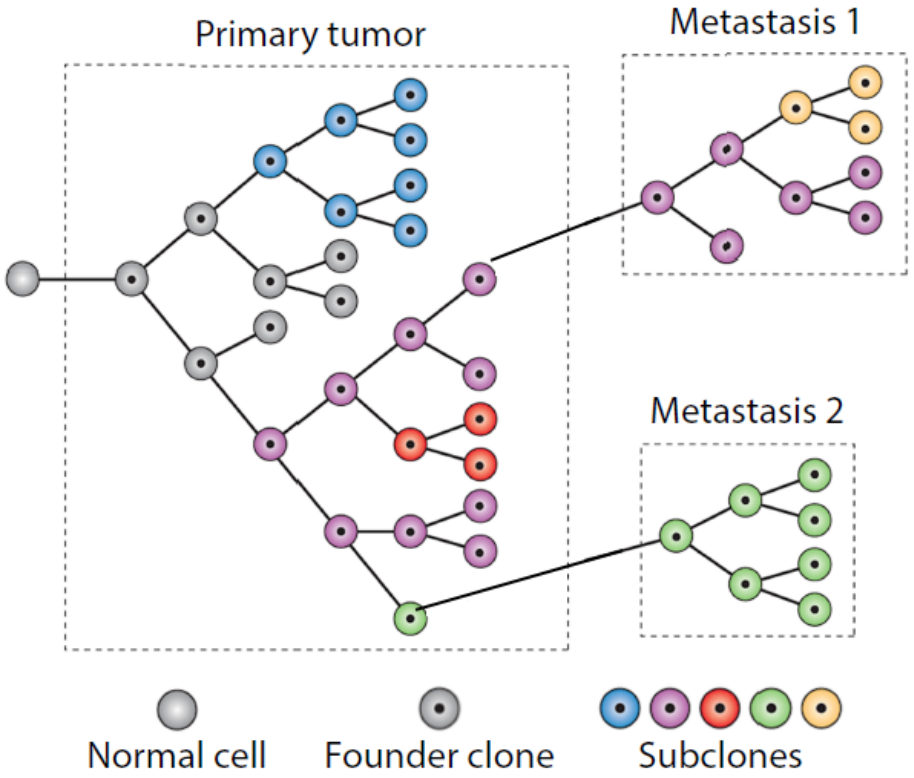


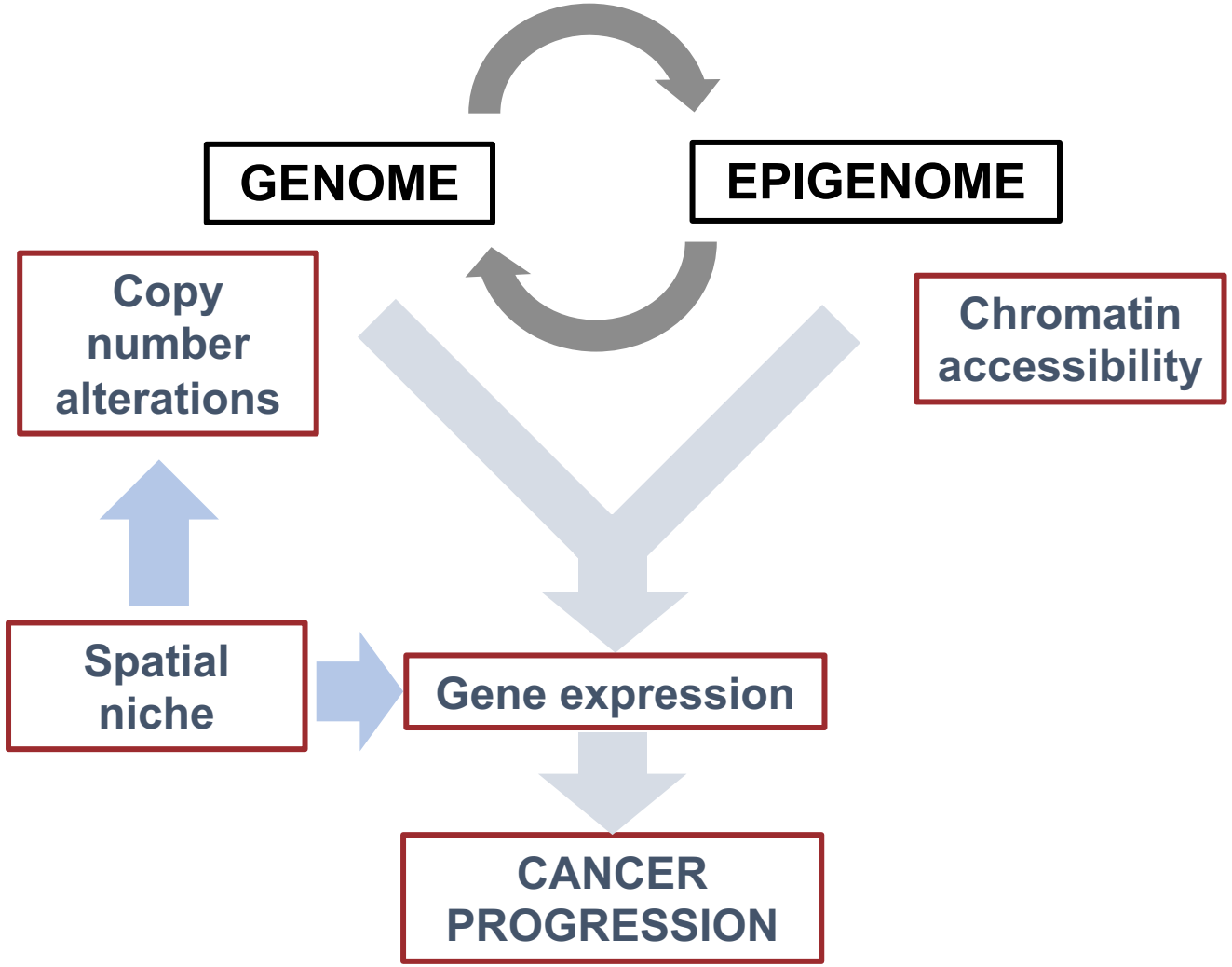
Spatial transcriptomic data





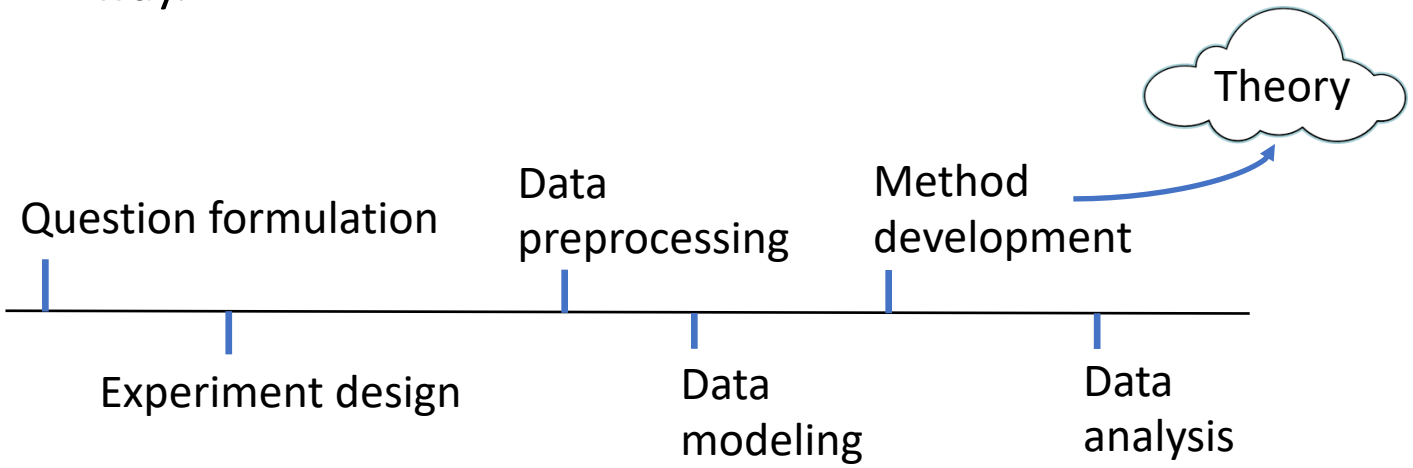
Genome instability and **epigenetic plasticity** shape cancer evolution





Lessons from 15 years ago

- The most important, and perhaps difficult step is formulating the null hypothesis (Biologists sometimes call it the “control”)
- Biological data does not conform to clean models. When doing at the genome level, the most significant signals are often model violations.
- Work closely with domain experts, every step of the way.



2000 Rough draft of human genome revealed.

2003 ENCODE project launched

2005 Peter gave a talk on ENCODE at Stanford

Acknowledgements

U. Penn

Chi-Yun Wu, Genomics and Computational Biology Program

Paul Hess

Kaishu Mason

Zilu Zhou, Google Inc

University of North Carolina

Yuchao Jiang, Departments of Biostatistics and Genetics

Stanford (Hanlee Ji Lab)

Billy T. Lau

Heonseok Kim

Anuja Sathe

Sue Grimes

Hanlee Ji

Thank you !!