

# DNA copy number profiling: from bulk tissue to single cells

Yuchao Jiang

Department of Statistics

Department of Biology

Department of Biomedical Engineering

Texas A&M University

<https://yuchaojiang.github.io>

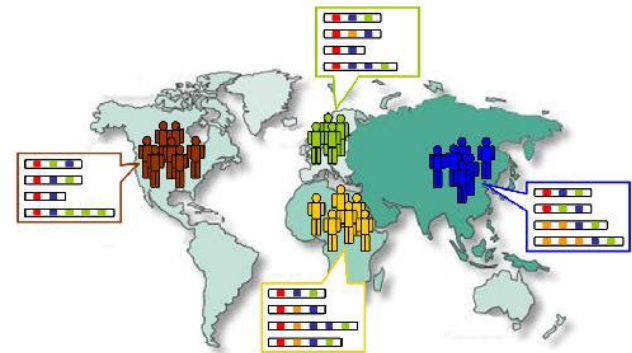
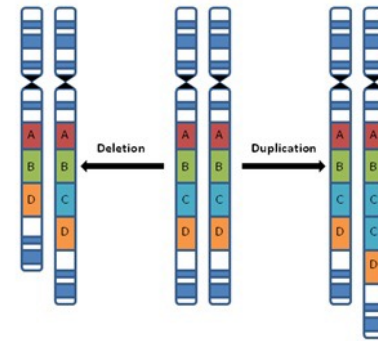
<https://twitter.com/yuchaojiang>



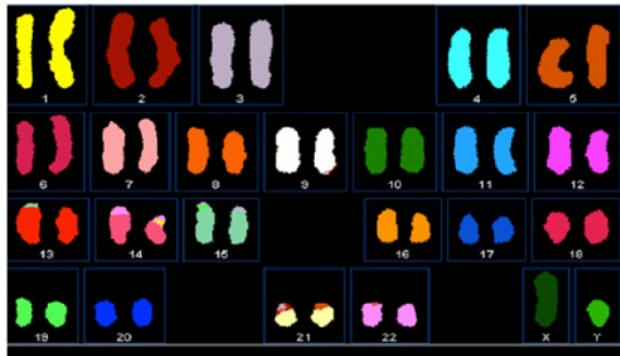
TEXAS A&M  
UNIVERSITY

# Copy number variation (CNV)

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



Normal



Tumor



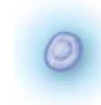
# CNV profiling: from bulk tissue to single cells

- Bulk DNA-seq
  - Whole-genome sequencing (WGS)
  - Whole-exome sequencing (WES) & targeted sequencing
- Single-cell DNA-seq
  - Conventional whole-genome amplification
  - 10X Genomics Chromium Single Cell CNV Solution / DLP+

**Population of cells  
(bulk sequencing)**

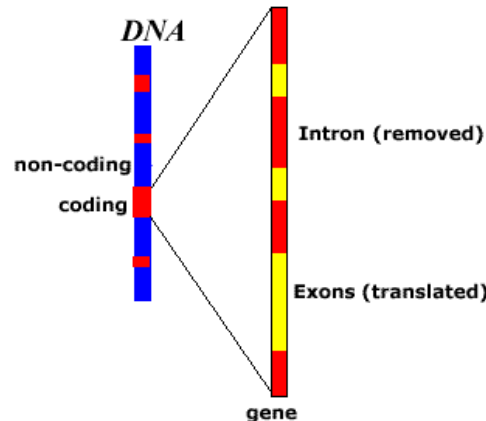


**Single cell  
(single-cell sequencing)**

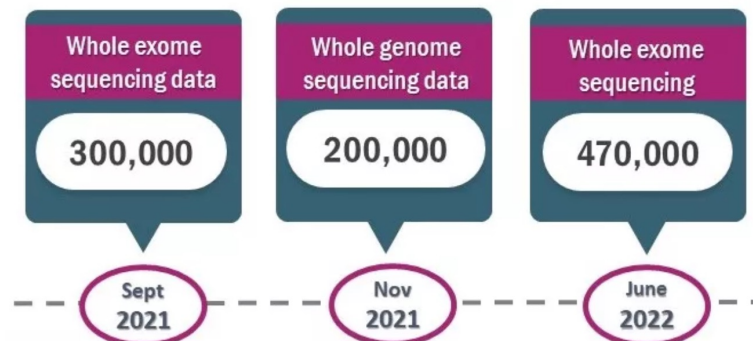


# Whole-exome sequencing

- Exome/exons: “functional” protein coding regions (1%) of the genome

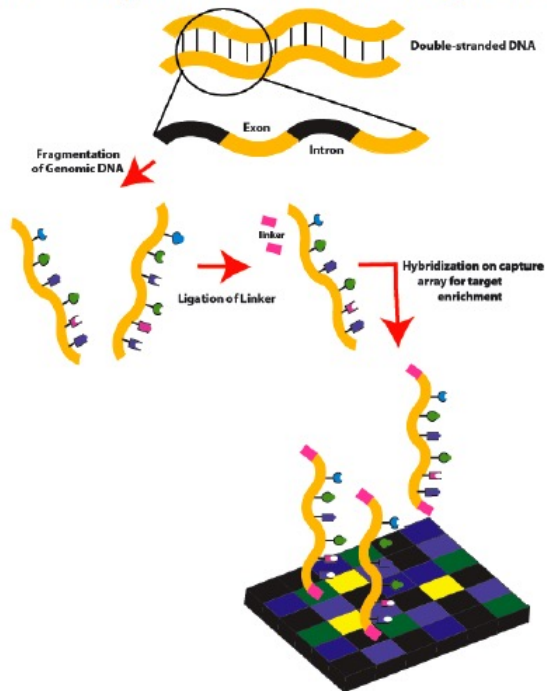


- Whole-exome sequencing
  - A cheaper, faster, but still effective alternative to whole-genome sequencing
  - Method of choice for large-scale studies

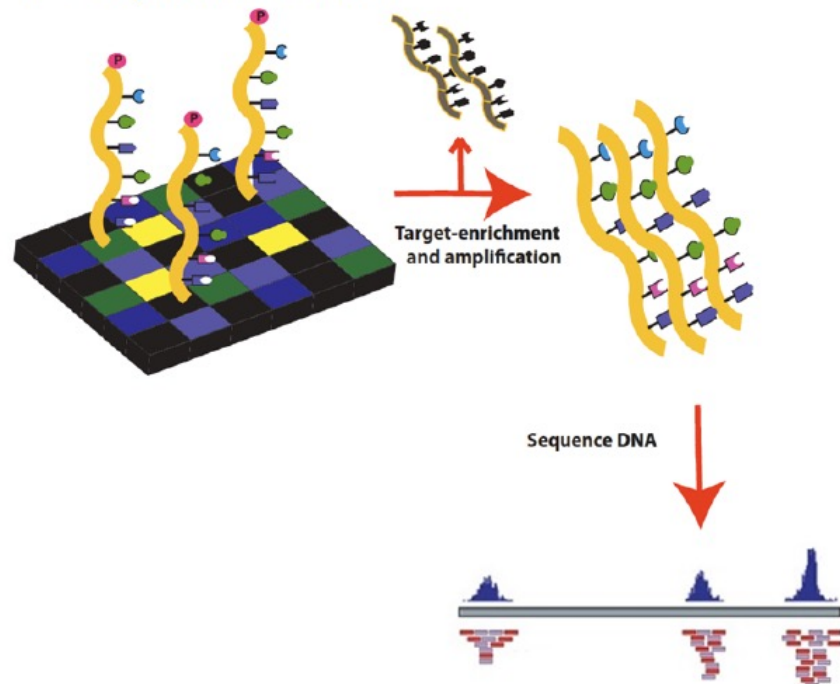


# Whole-exome sequencing protocols

1. Fragment DNA.
2. Capture exonic regions.



3. Amplify captured/exonic regions.
4. Sequence DNA.

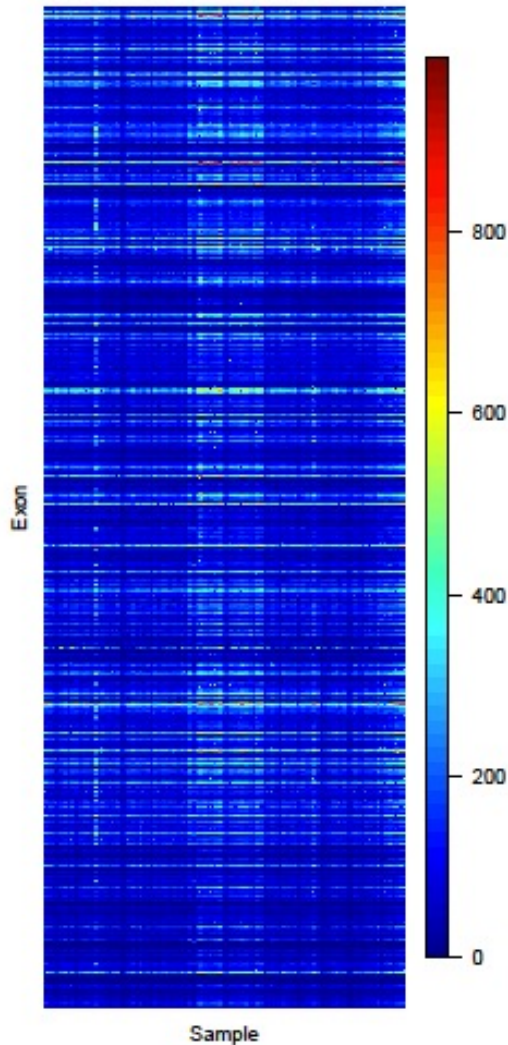


([https://en.wikipedia.org/wiki/Exome\\_sequencing](https://en.wikipedia.org/wiki/Exome_sequencing))

# Our goal

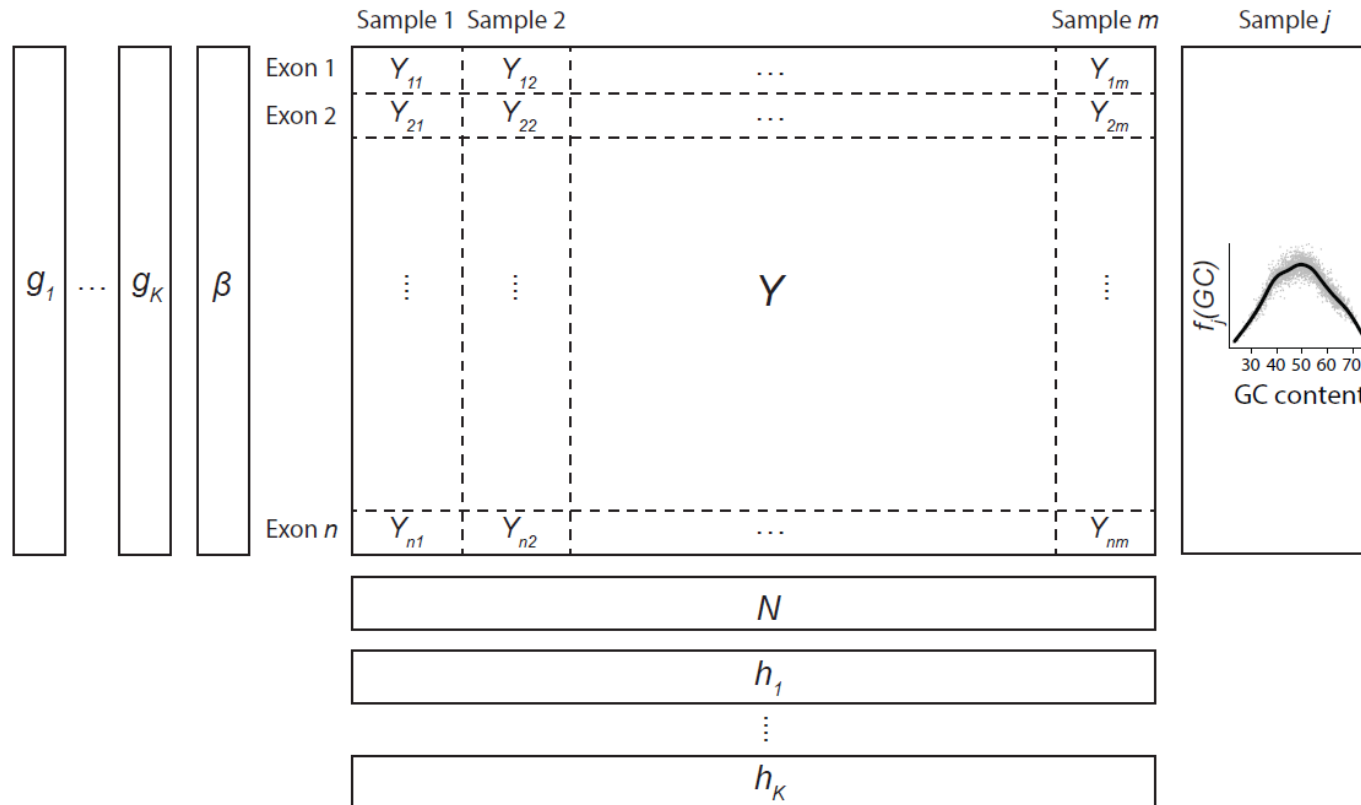
- Use bulk sequencing to accurately detect CNV
  - Based on depth of coverage  
(i.e., number of times a genomic region is “read”)
- What are the biases?
  - GC content (ease of segmentation)
  - Exon capture and amplification efficiency
  - Batch effect, population stratification
  - Latent factors
  - ...

# Cross-sample normalization model



- What has been done: PCA
  - Continuous measurements
  - Assume Gaussian noise structure
  - Ignores known quantifiable biases, such as GC content, exon lengths, etc.
  - GC content bias cannot be captured by a linear PC

# Poisson latent factor model



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

$\lambda$  : expected coverage (no CNV)

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

$\beta_i$  : 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



# Poisson latent factor model

- CNV detection:

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

(Jiang et al., Nucleic Acids Research, 2015)

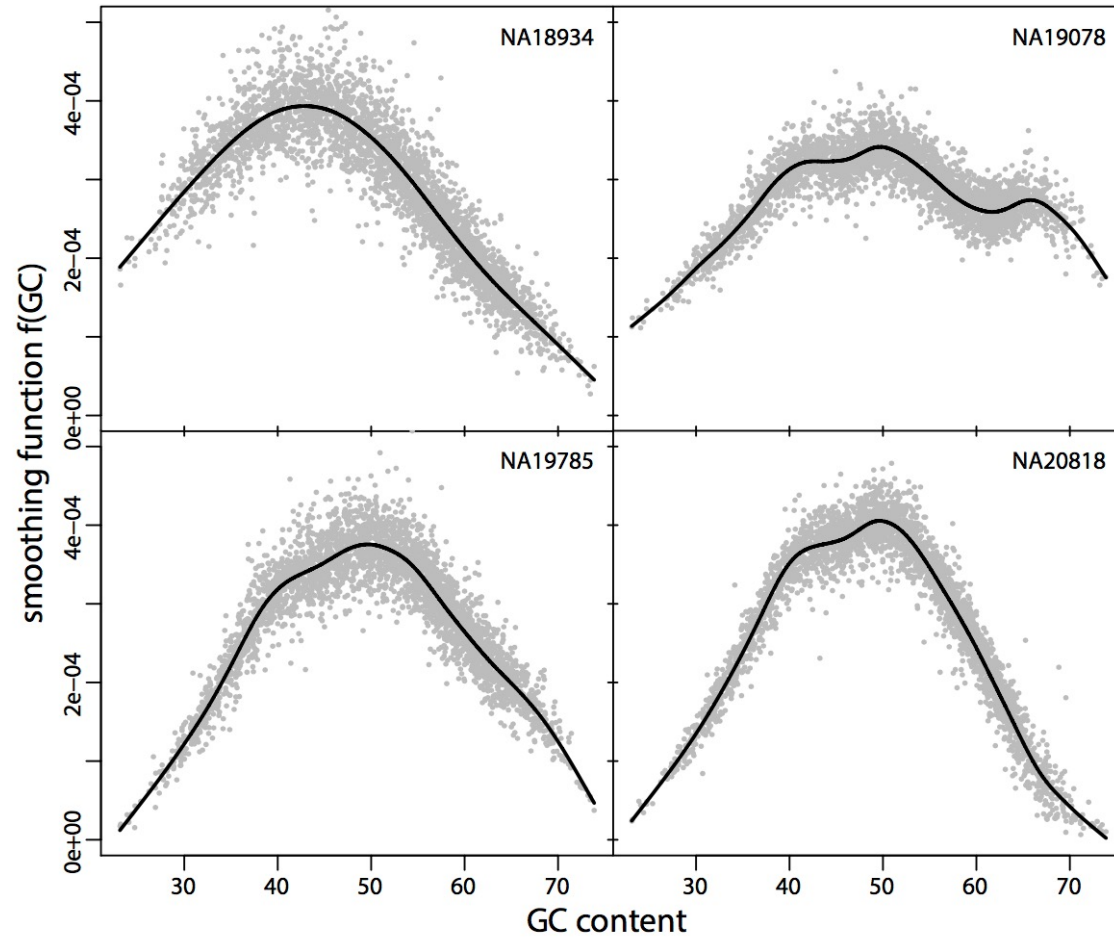
- Spatial transcriptomics:

$$Y_{i,j} | \lambda_{i,j} \sim \text{Poisson}(N_i \lambda_{i,j})$$
$$\log(\lambda_{i,j}) = \alpha_i + \log \left( \sum_{k=1}^K \beta_{i,k} \mu_{k,j} \right) + \gamma_j + \varepsilon_{i,j}$$

(Cable et al., Nature Biotechnology, 2022)

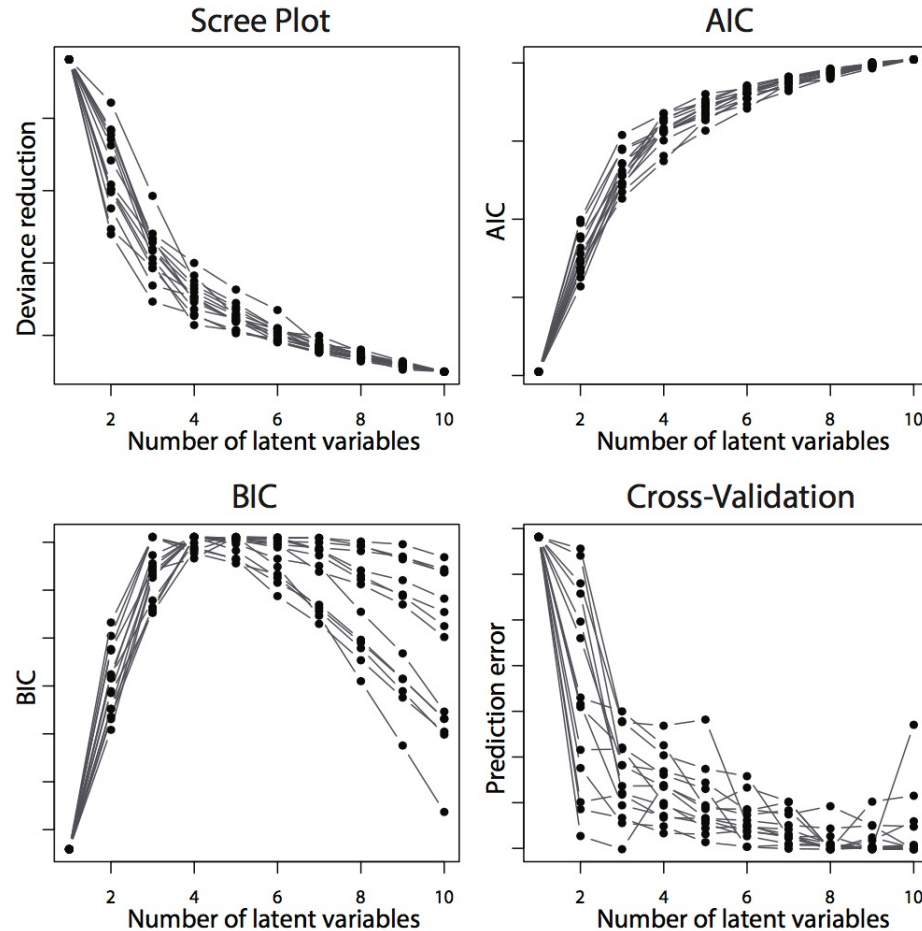
# GC content bias

Whole-exome sequencing data from the 1000 Genomes Project

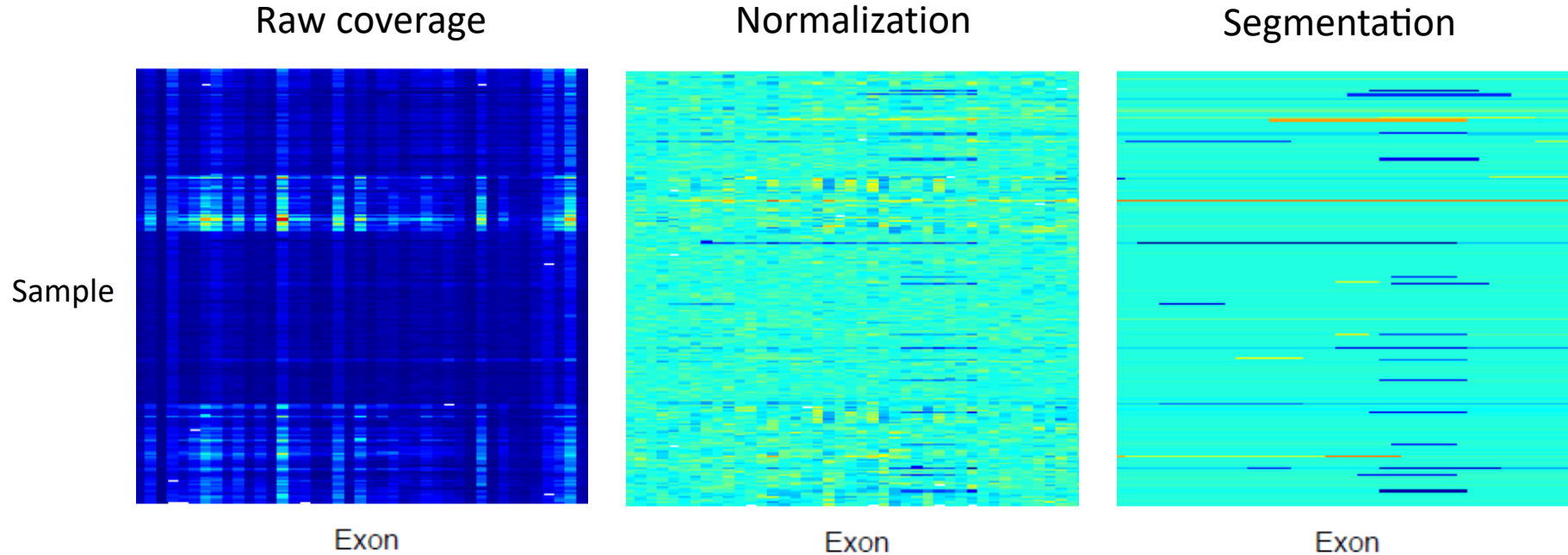


# How to choose the number of latent factors?

Whole-exome sequencing data from the TARGET initiative

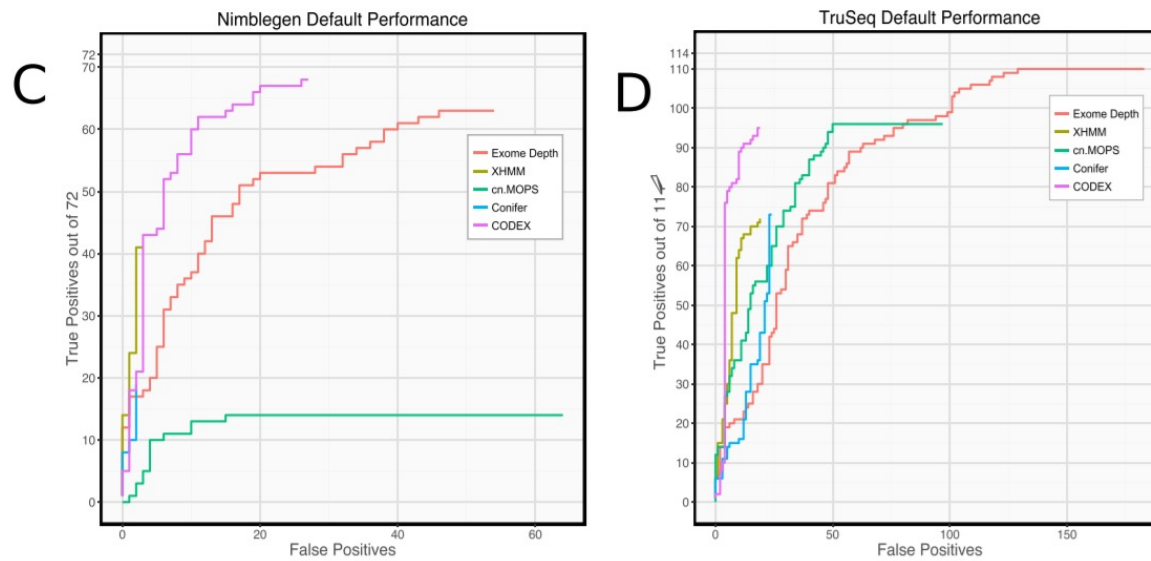


# CODEX: COpy number Detection by EXome-seq



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

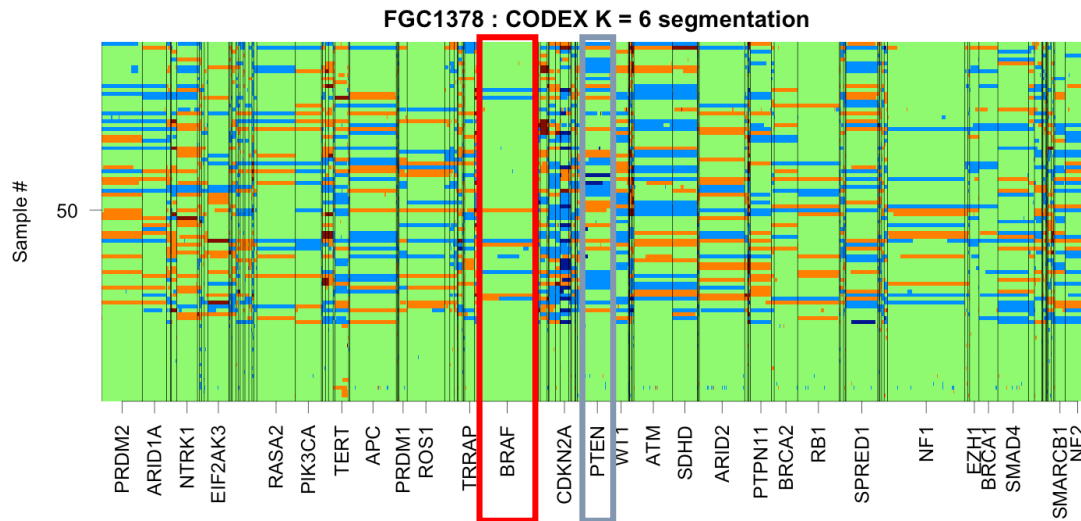
# Independent benchmark results



(Sadedin et al., GigaScience, 2018)

# Melanoma targeted sequencing

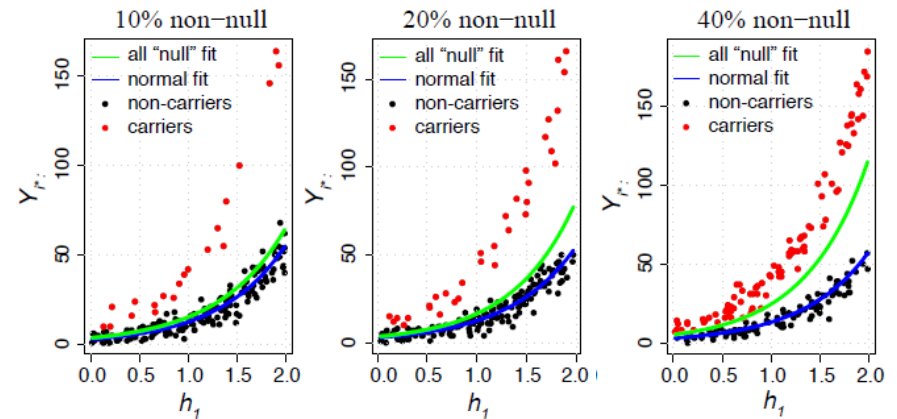
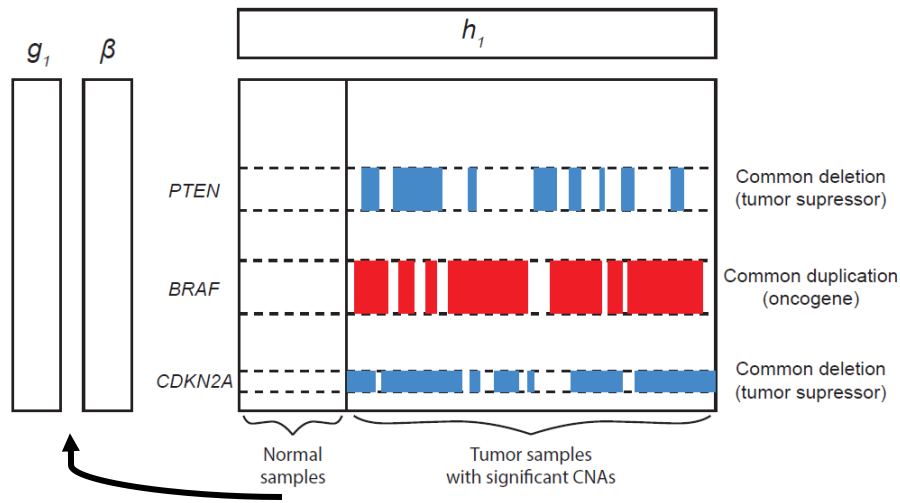
- High amp
- Amp
- Null
- Hemi. del
- Homo. del



**CODEX**

**Common CNV signals are attenuated by  
the Poisson latent factors!**

# Low sensitivity for common CNVs



Null Model:

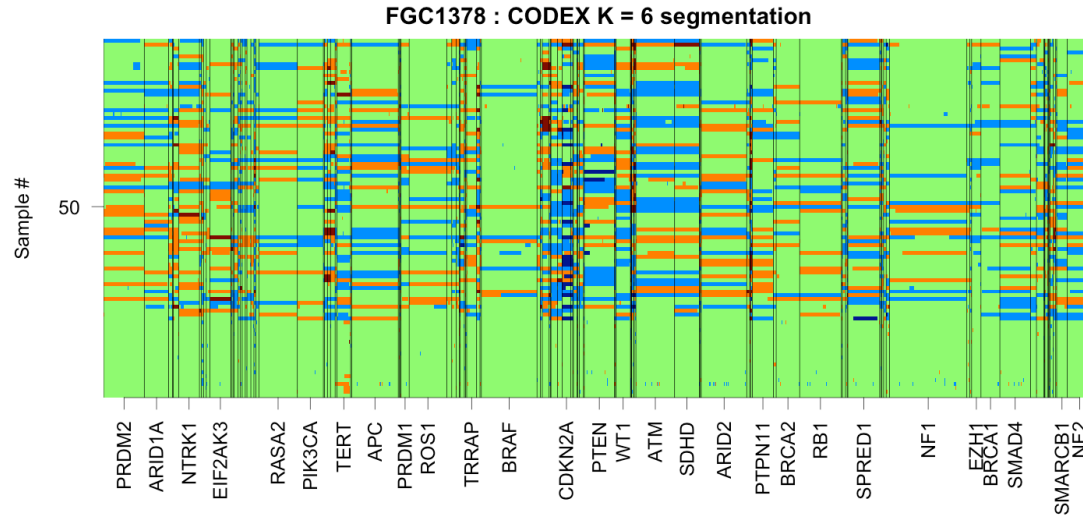
$$Y_{ij} \sim \text{Poisson}(\lambda_{ij})$$

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

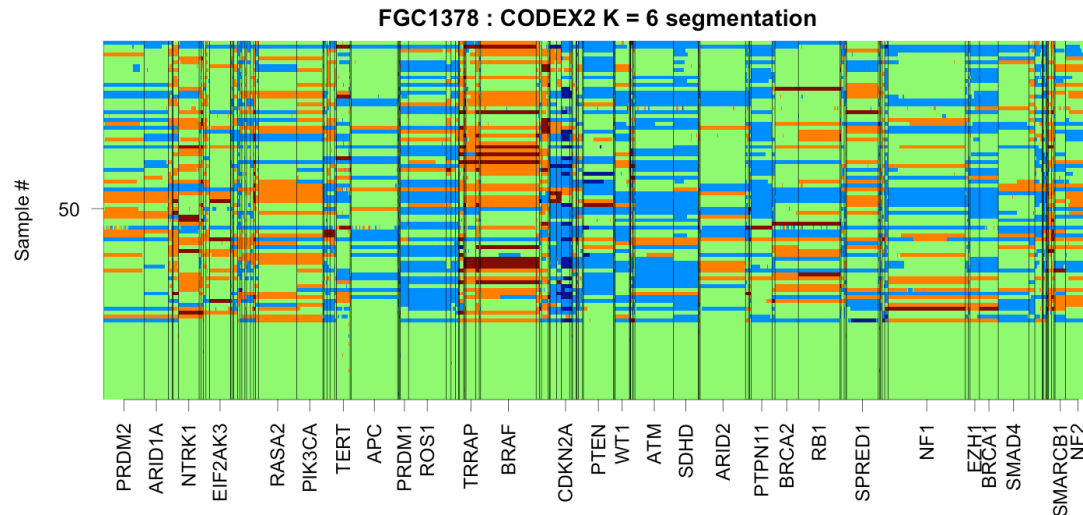
**Solution:**

**ONLY use normal samples to estimate exon-specific bias and latent factors!**

# Melanoma targeted sequencing: sanity check



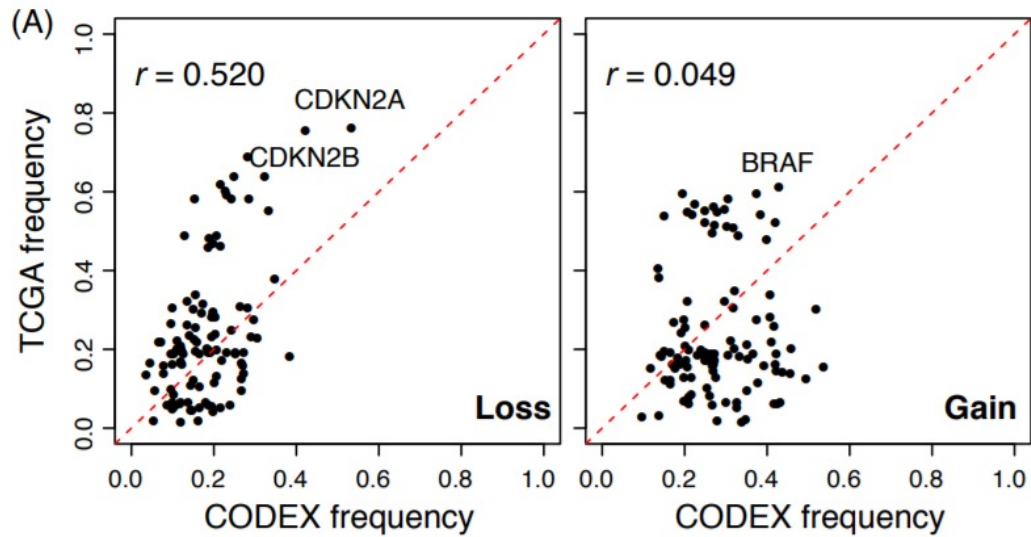
**CODEX**



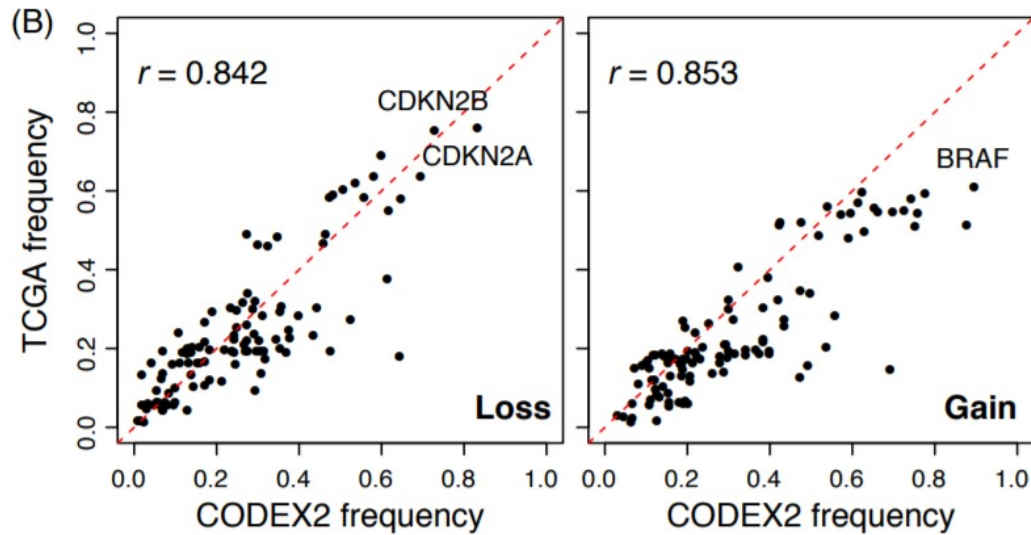
**CODEX2**



# TCGA validation

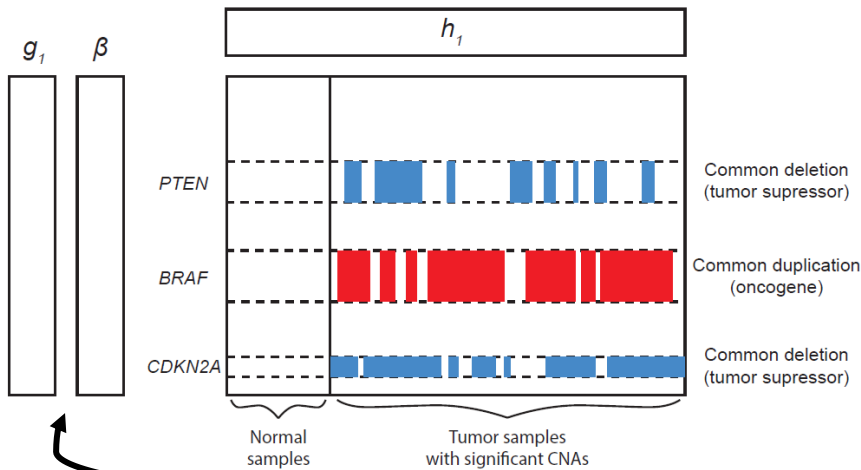


**CODEX**

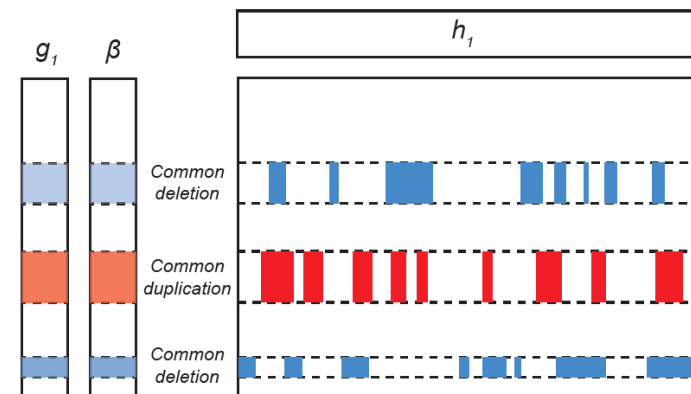


**CODEX2**

# Recovering common CNV signals

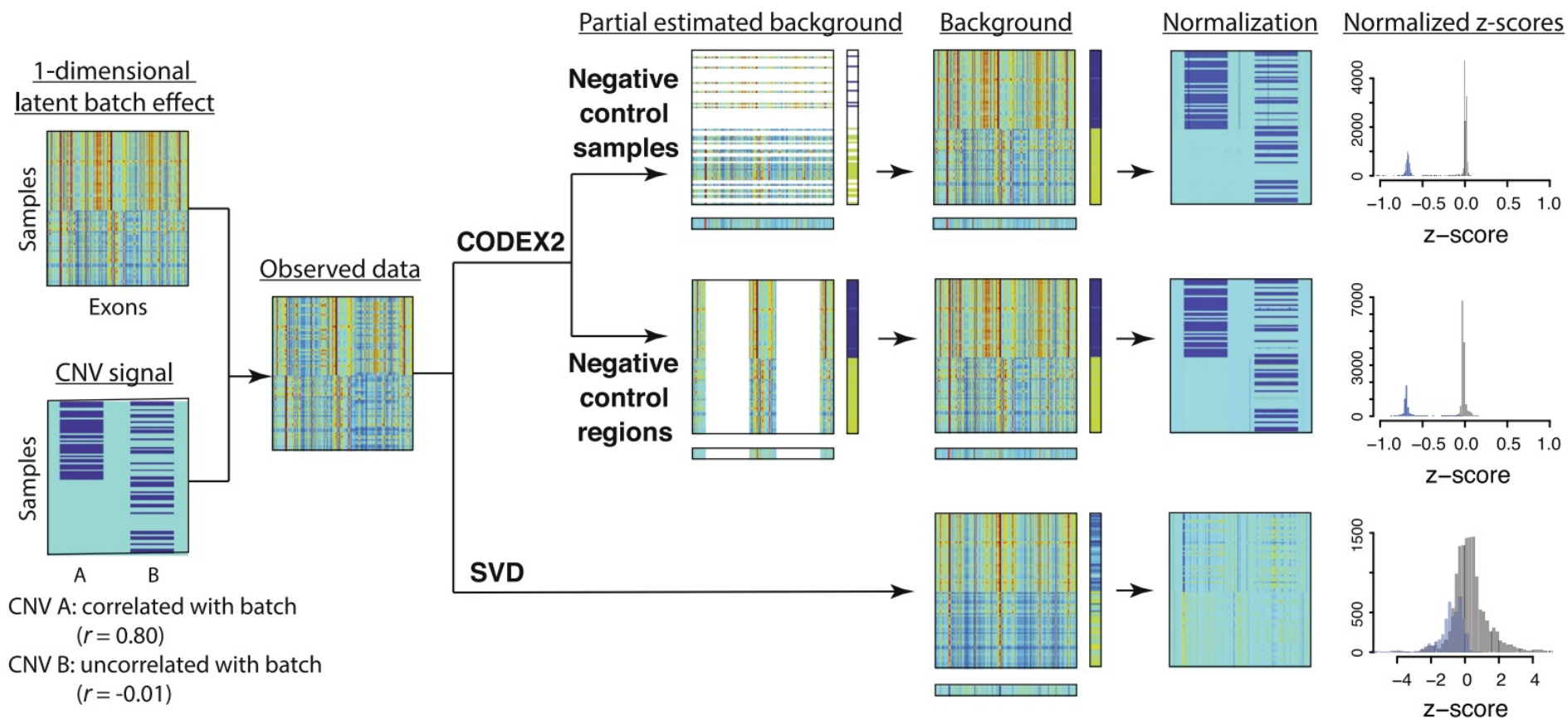


**ONLY use normal samples to estimate exon-specific bias and latent factors.**



**What if we don't know which samples are "normal" (common germline CNV detection)?**

# CODEX2: full-spectrum CNV detection by NGS

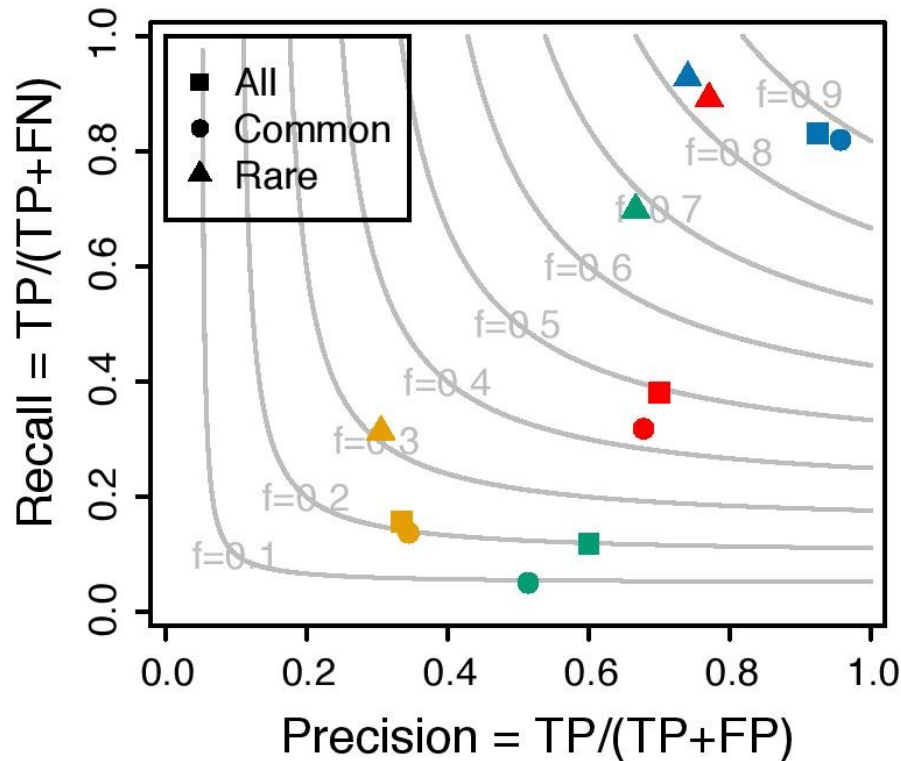


<https://github.com/yuchaojiang/CODEX2>

(Jiang et al., Genome Biology, 2018)

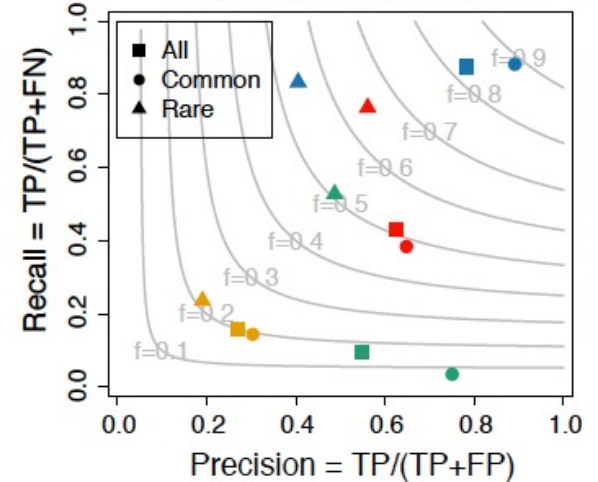
# HapMap samples: SNP-array validation

McCarroll *et al.*

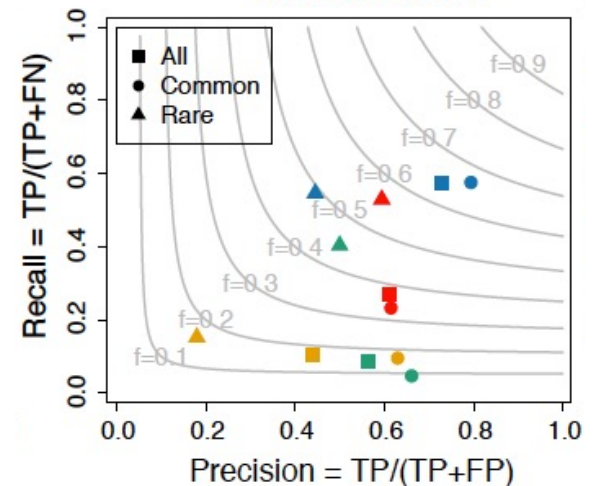


- CODEX2
- CODEX
- EXCAVATOR
- XHMM

HapMap3 Consortium



Conrad *et al.*



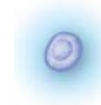
# CNV profiling: from bulk tissue to single cells

- Bulk DNA-seq
  - Whole-genome sequencing (WGS)
  - Whole-exome sequencing (WES) & targeted sequencing
- Single-cell DNA-seq
  - Conventional whole-genome amplification
  - 10X Genomics Chromium Single Cell CNV Solution / DLP+

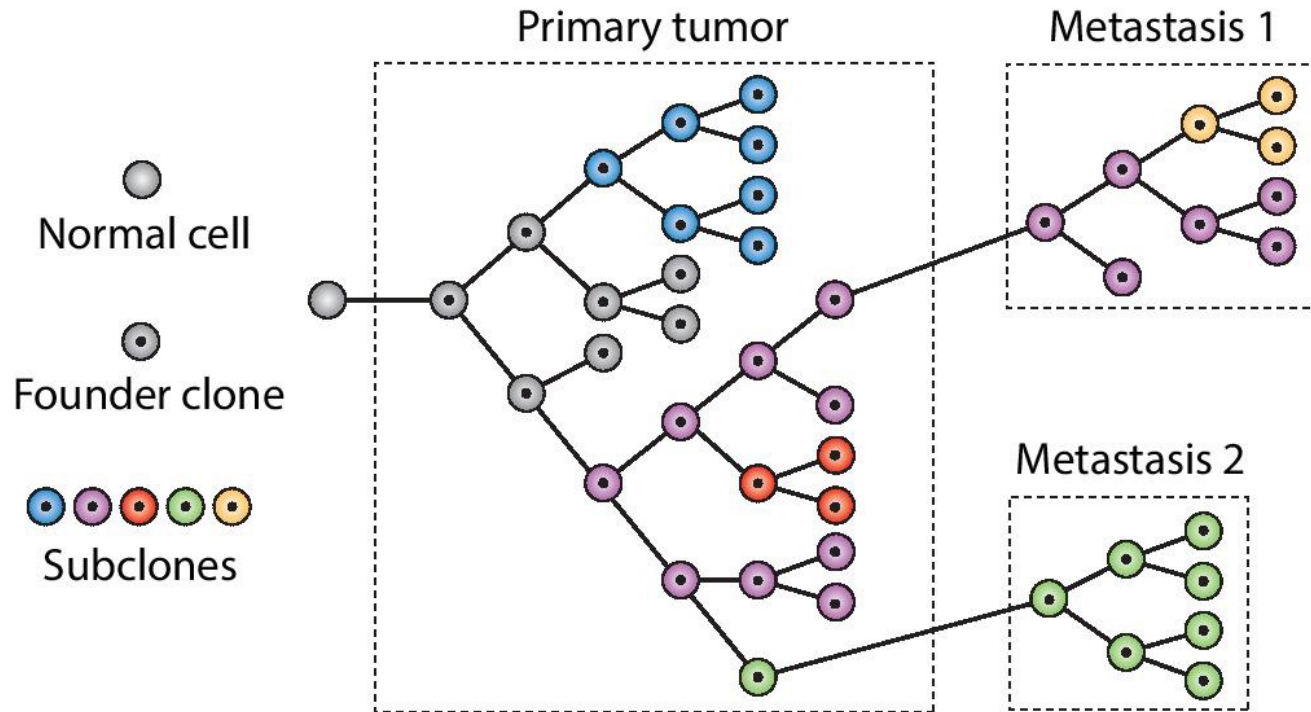
**Population of cells  
(bulk sequencing)**



**Single cell  
(single-cell sequencing)**



# Profiling somatic copy number aberrations by scDNA-seq

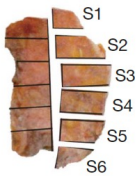


(Jiang et al., *PNAS*, 2016)

- Goal: Use scDNA-seq to assess intra-tumor heterogeneity by profiling somatic copy number aberrations with single-cell resolution.

# scDNA-seq data breast cancer patients

Patient T10

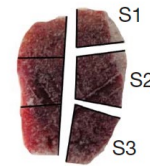


Polygenomic tumor

Patient T16 Primary Tumor



Patient T16 Metastasis



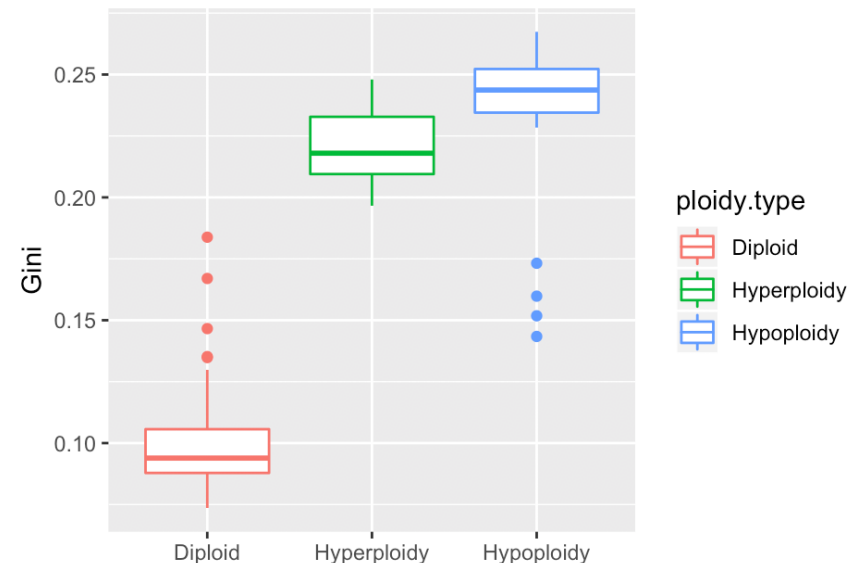
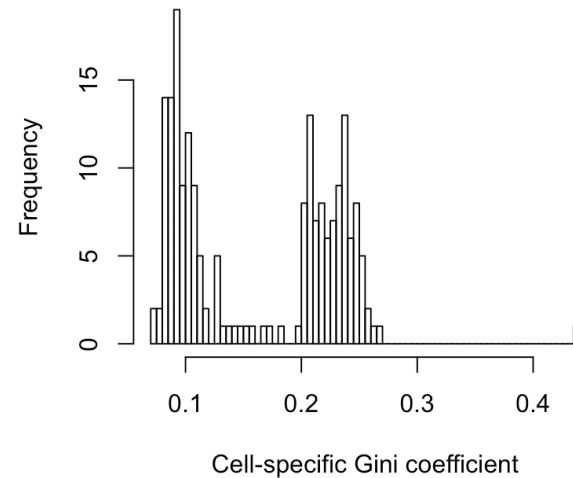
Monogenomic tumor

FACS sorting followed by copy number profiling by scDNA-seq revealed three distinct subclones in patient T10 and a single clonal expansion in patient T16.

# Using Gini index to identify normal cells

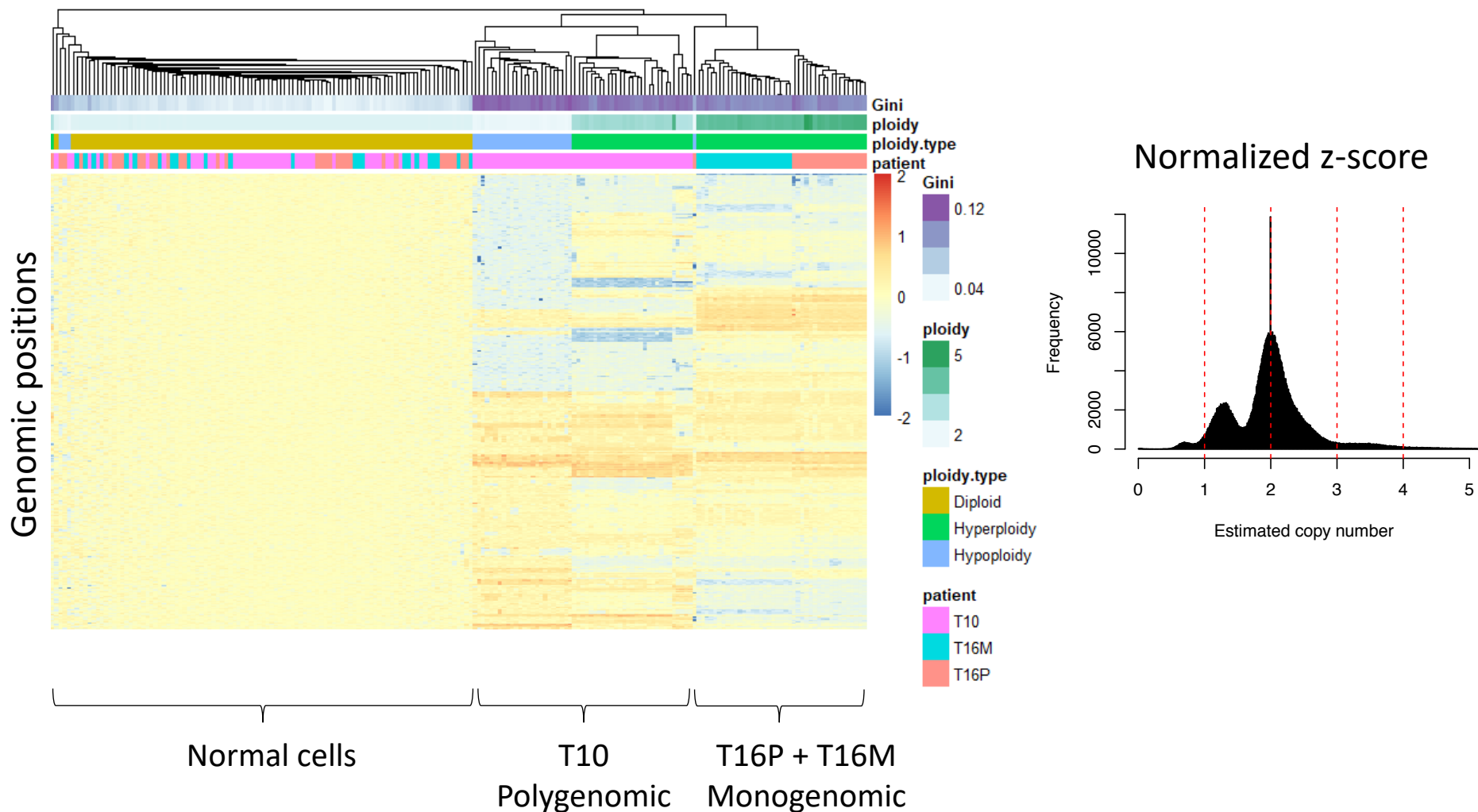
- Cell-specific Gini index distribution.
- Cancer cells have higher Gini index compared to normal cells by FACS sorting.

Histogram of Gini coefficients



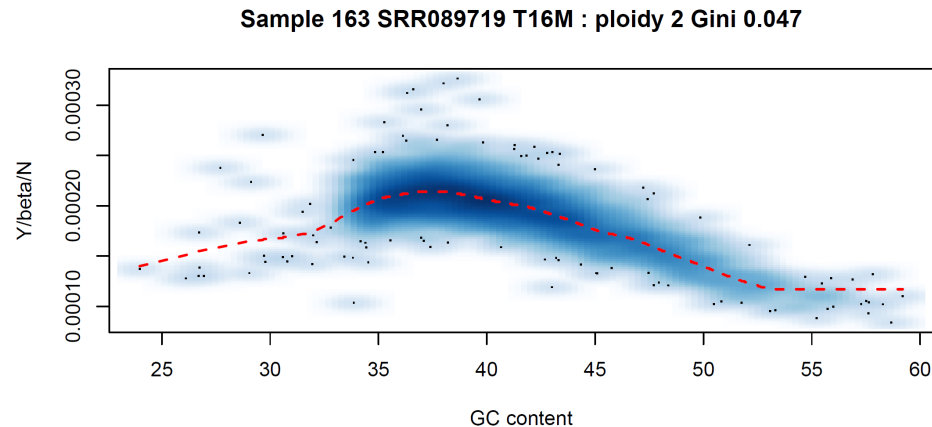


# CODEX2 with negative control samples (i.e., normal cells)



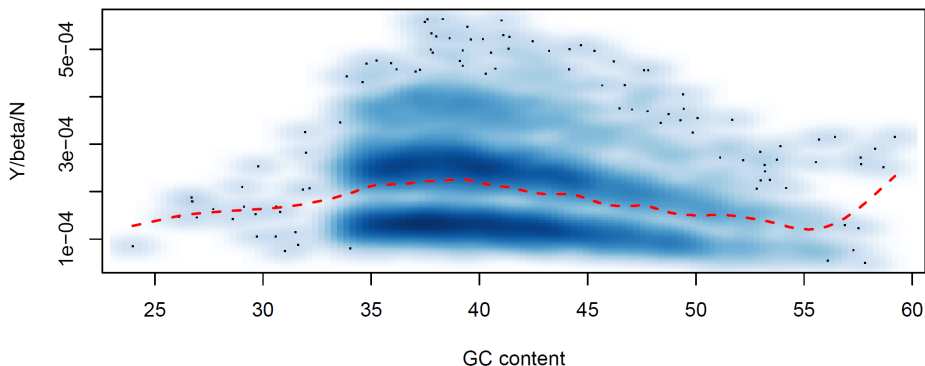
# Sanity check: GC content bias $f(GC)$ by CODEX2

- Good fit:

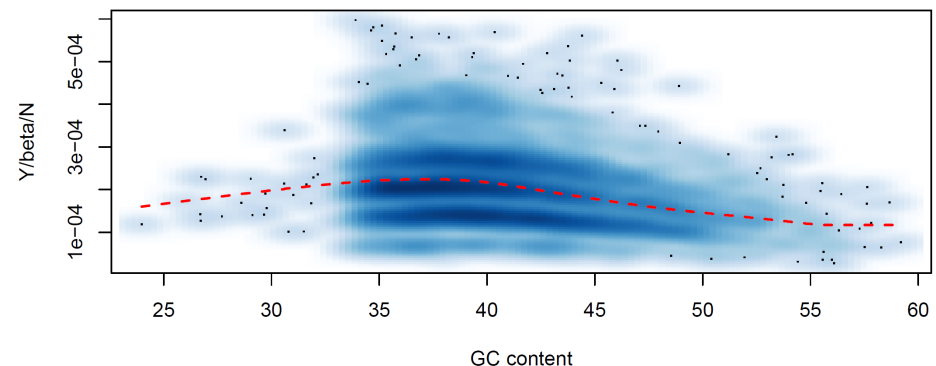


- Poor fit: multiple different copy number states contribute to multiple increments in  $f(GC)$

Sample 45 SRR053679 T10 : ploidy 1.7 Gini 0.1254

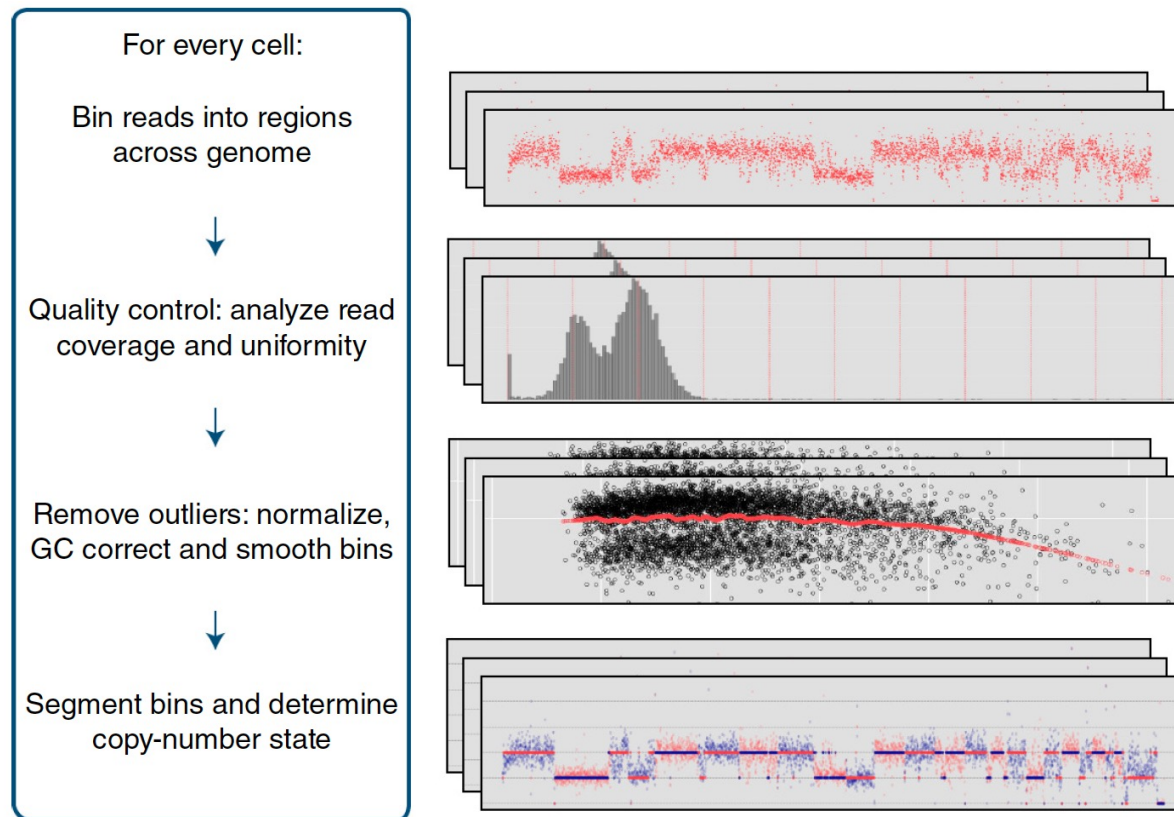


Sample 70 SRR054602 T10 : ploidy 3.15 Gini 0.1176



# Benchmark against existing method

- Ginkgo (Garvin et al., Nature Methods, 2015)
  - Normalization: Lowess fit to correct for GC bias
  - Segmentation: Cell-specific circular binary segmentation



(Figure 1: Gavin et al., *Nature Methods*, 2015)

# EM algorithm for GC bias

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

- If there are significant genome-wide copy number changes,

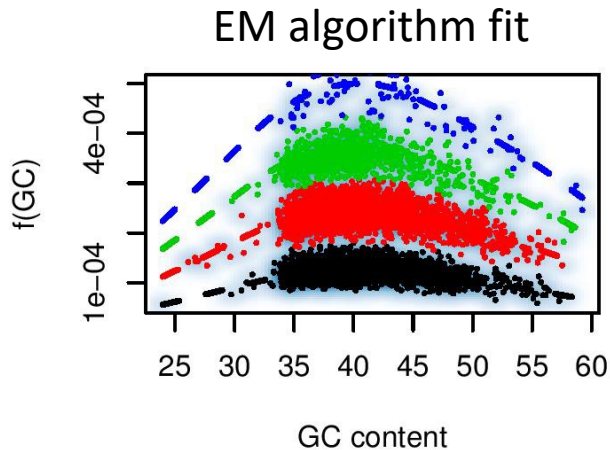
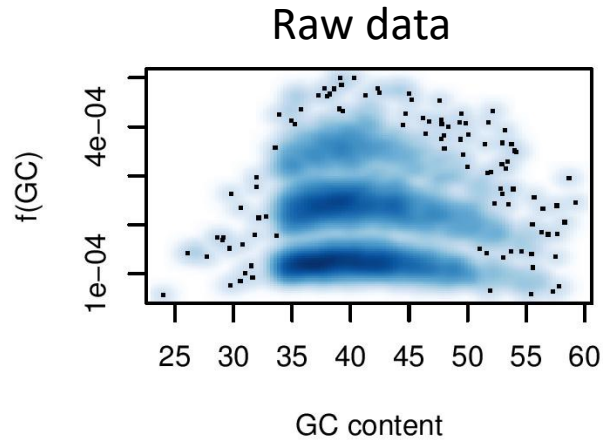
$$\alpha_{ij} = \begin{cases} 1/2 & \text{with probability } \pi_1^{(j)} \\ 2/2 & \text{with probability } \pi_2^{(j)} \\ \vdots & \vdots \\ T_j/2 & \text{with probability } \pi_{T_j}^{(j)} \end{cases},$$

where  $T_j$  is the number of copy number groups and  $\sum_{t=1}^{T_j} \pi_t^{(j)} = 1$ .

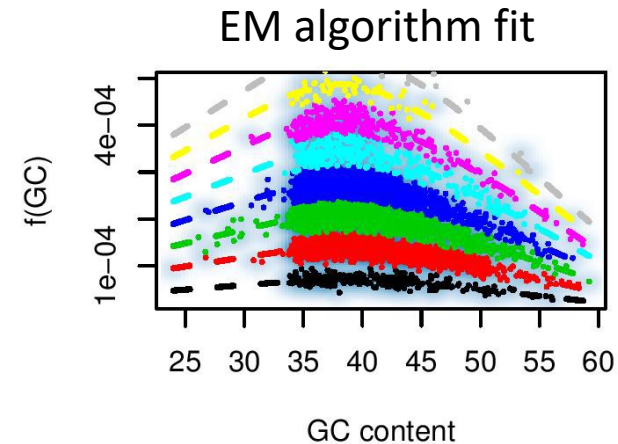
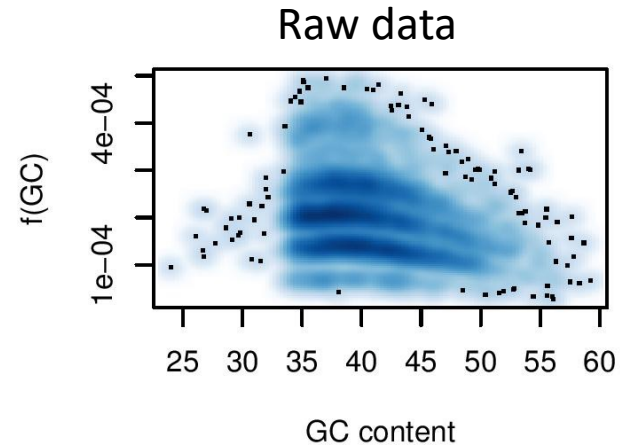
- $\phi(Y_{ij}) = \sum_{t=1}^{T_j} \pi_t^{(j)} \phi_t(Y_{ij}) =$   
 $\sum_{t=1}^{T_j} \pi_t^{(j)} \text{pPoisson}\left(Y_{ij}; N_j \beta_i f_j(GC_i) \frac{t}{2} \exp\left(\sum_{k=1}^K g_{ik} h_{jk}\right)\right).$

# EM algorithm for GC bias: empirical fit

Hypodiploid cell (cell ploidy < 2)

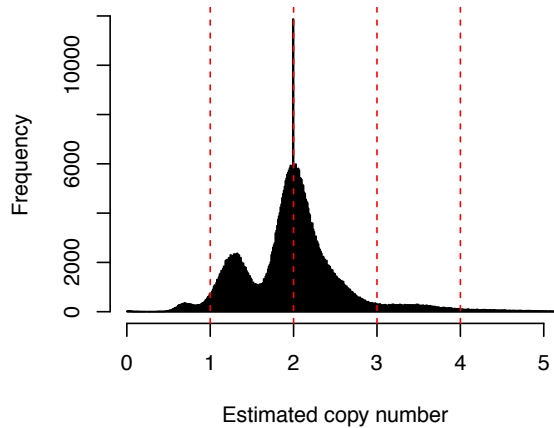


Hyperdiploid cell (cell ploidy > 2)

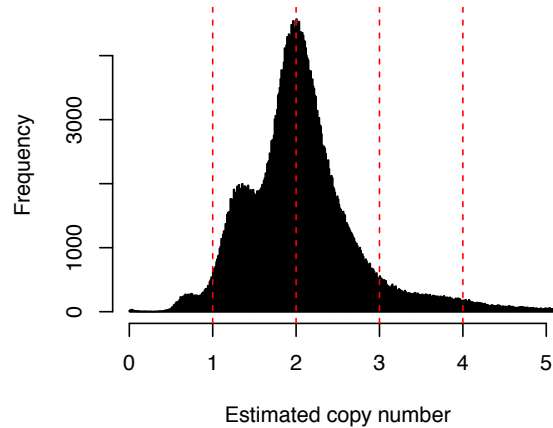


# Normalization results for patient T10

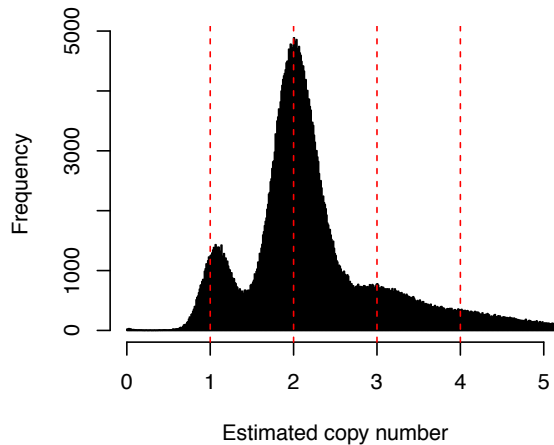
**CODEX2 normalization**



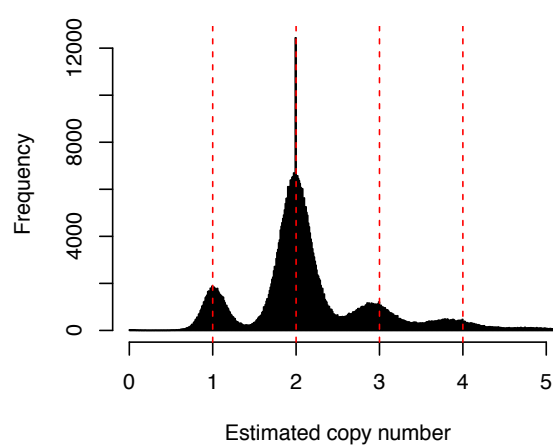
**Ginkgo normalization**



**Ginkgo normalization + ploidy adjustment**

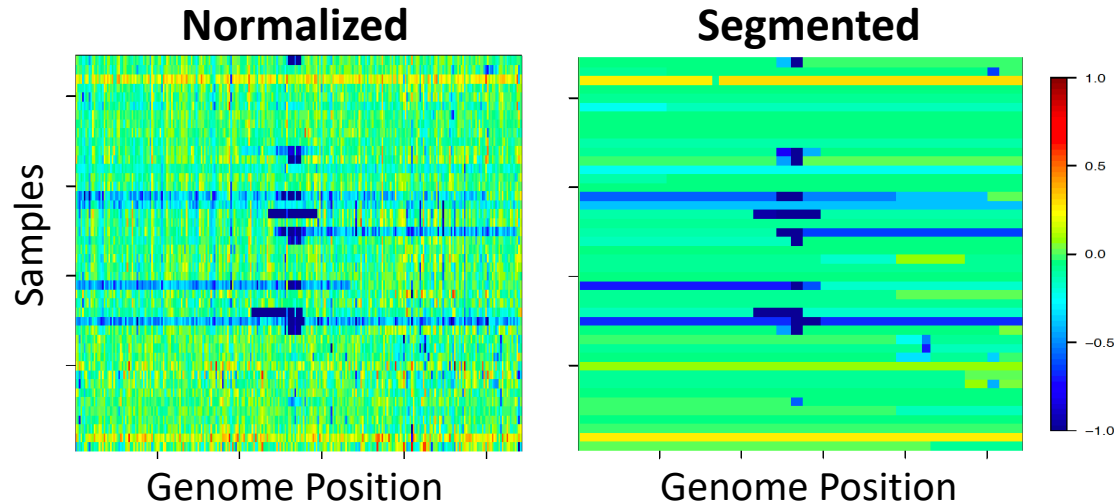


**SCOPE normalization**



# Cross-sample segmentation

- Why across samples? Breakpoints are shared across cells from the same clone.



- The scan statistic across all cells is  $\max_{s,t} Z_{s,t}$ :

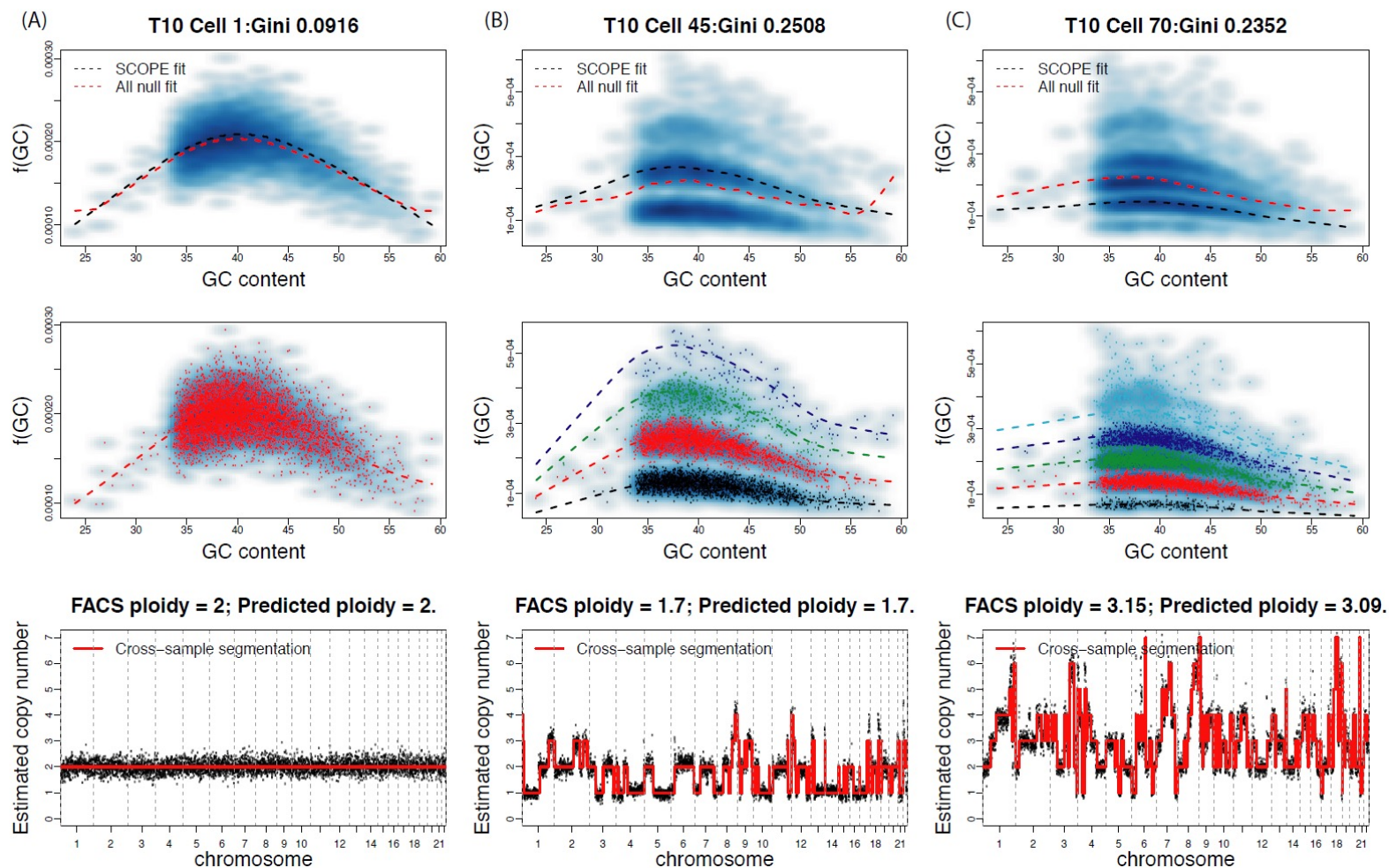
$$Z_{s,t} = \sum_j U_j(s, t) = \sum_j Y_{s:t,j} \log \left( \frac{Y_{s:t,j}}{\lambda_{s:t,j}} \right) - (Y_{s:t,j} - \lambda_{s:t,j}).$$

- Recursively find the maximizing cut in the sub regions, using a cross-sample mBIC as stopping rule:

$$\begin{aligned} \text{mBIC}(p) = & \log \left( \frac{L_\tau}{L_0} \right) - \frac{P}{2} \log \frac{2 \log(L_\tau/L_0)}{P} - \frac{P}{2} - \log \binom{m}{p} - p(\kappa_1 - \kappa_2) \\ & - \sum_{\rho=1}^p \log \left( \sum_{j \text{th carrier}} \hat{\delta}_{\rho,j}^2 \right) + P \log \pi + (np - P) \log(1 - \pi). \end{aligned}$$



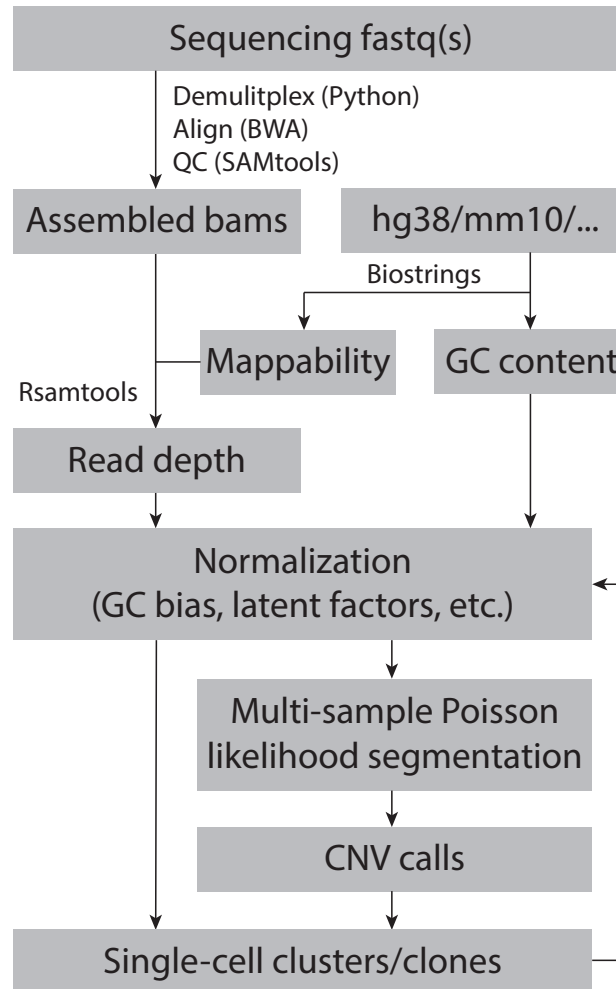
# Embed EM for fitting GC bias in Poisson latent factor model with negative control samples



Estimated ploidy is taken as the mean estimated copy number across all bins. Close to previous reports.



# SCOPE: Single-cell COPY number Estimation



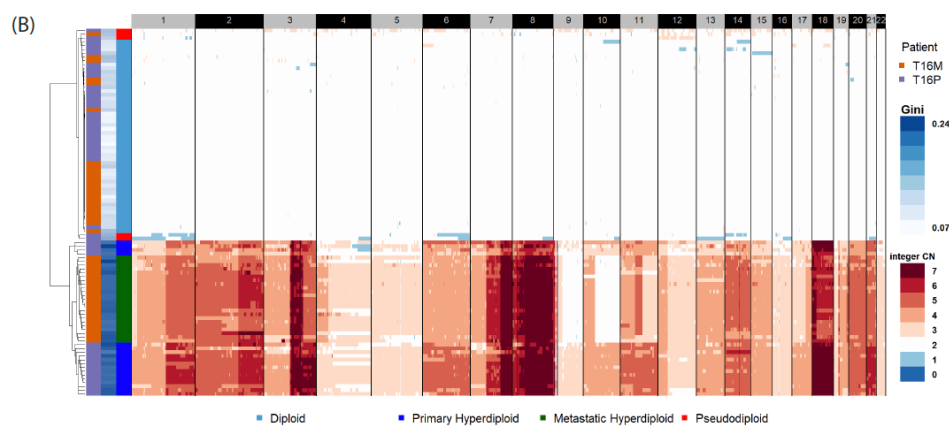
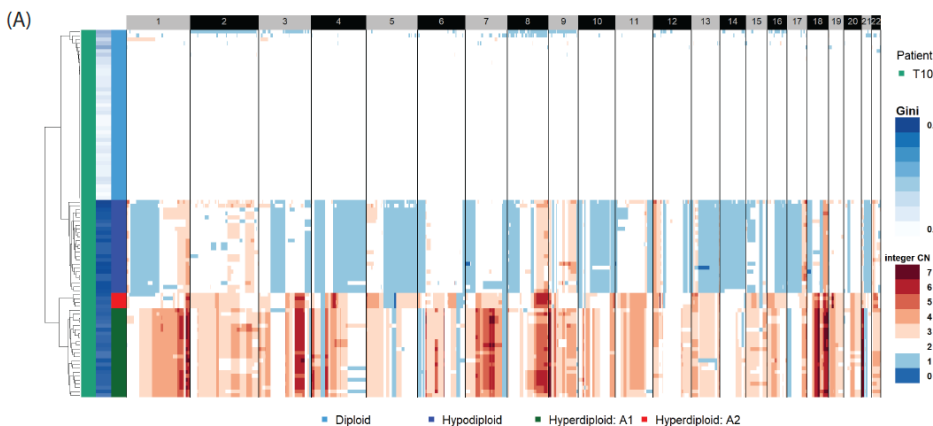
<https://github.com/rujinwang/SCOPE>

(Wang et al., *Cell Systems*, 2020)

# CNV profiles of breast cancer patients T10, T16

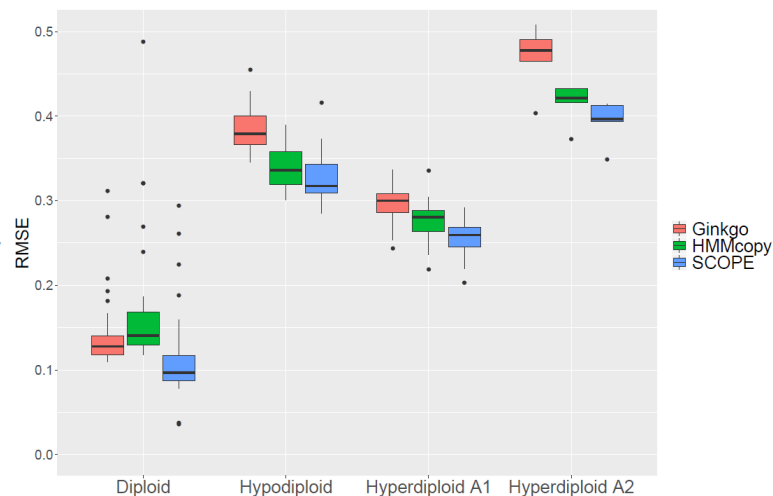
## Polygenomic tumor

## Monogenomic tumor

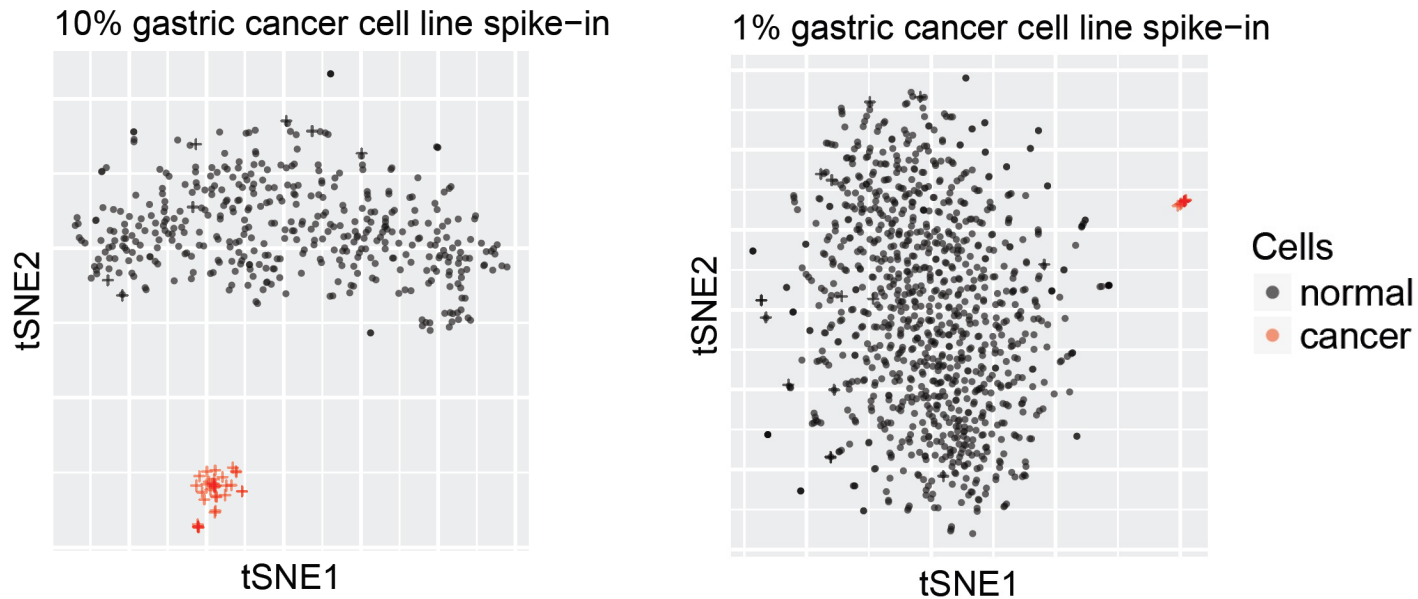


## Gold standard for orthogonal validation:

CNV calls by aCGH of purified bulk samples (Navin et al., *Genome Research*, 2010)



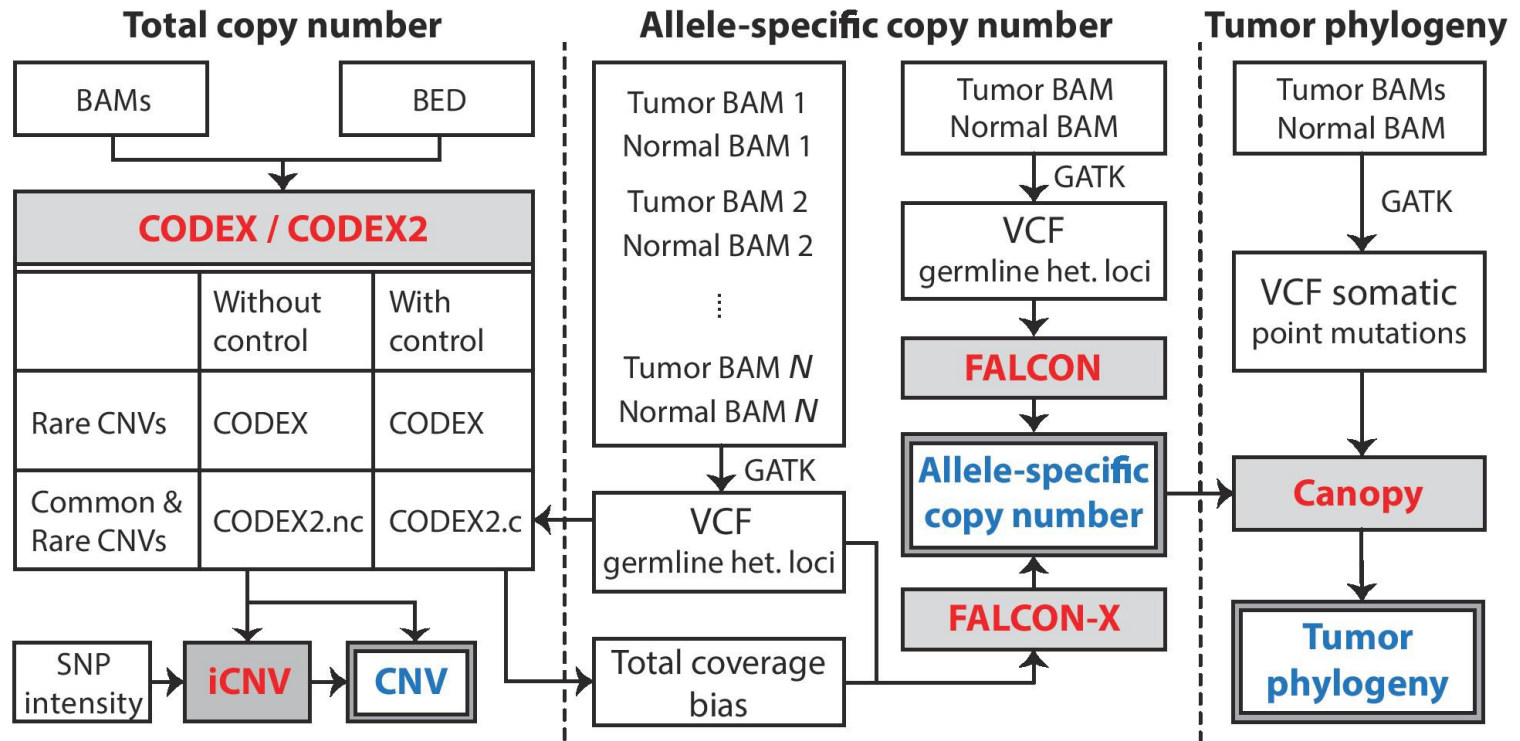
# 10X Genomics: gastric cancer cell line spike-in



Estimated % cancer cells:  $36/462 \approx 8\%$ .

Estimated % cancer cells:  $11/1055 \approx 1\%$ .

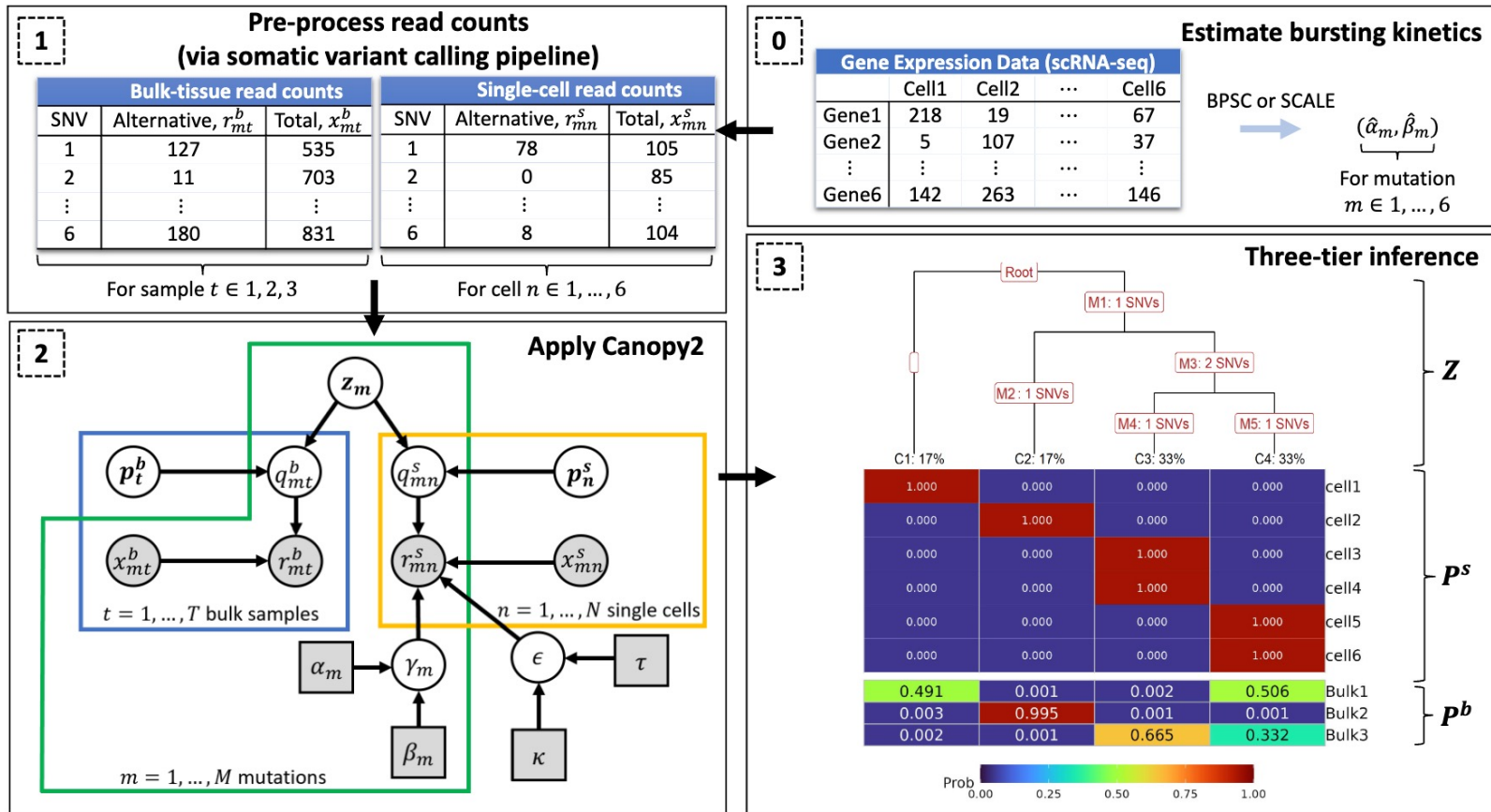
# Intratumor heterogeneity by bulk-tissue DNA-seq



Canopy: <https://CRAN.R-project.org/package=Canopy>  
(Jiang et al., PNAS, 2016)

MARATHON: <https://github.com/yuchaojiang/MARATHON>  
(Urrutia et al., Bioinformatics, 2018)

# Intratumor heterogeneity by bulk-tissue DNA-seq and single-cell RNA-seq



Canopy2: tumor phylogeny inference by bulk DNA and single-cell RNA sequencing

<https://github.com/annweideman/canopy2>

TAMU Biomedical Engineering Seminar: Nov 16<sup>th</sup> 2023, 2:20pm-3:35pm

**Thanks! Questions?**