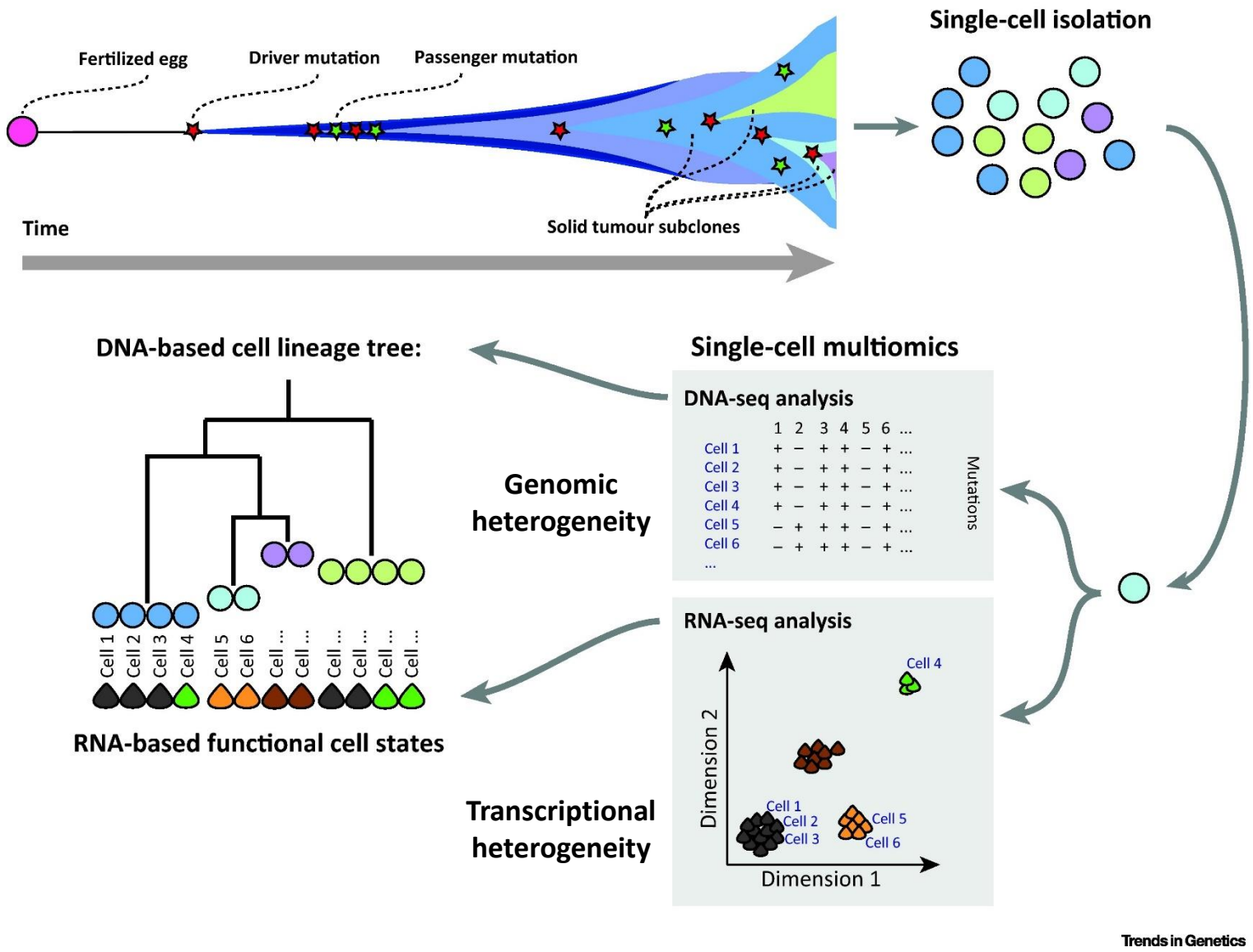
clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers

03/24/2020

CS598MEB Course Presentation

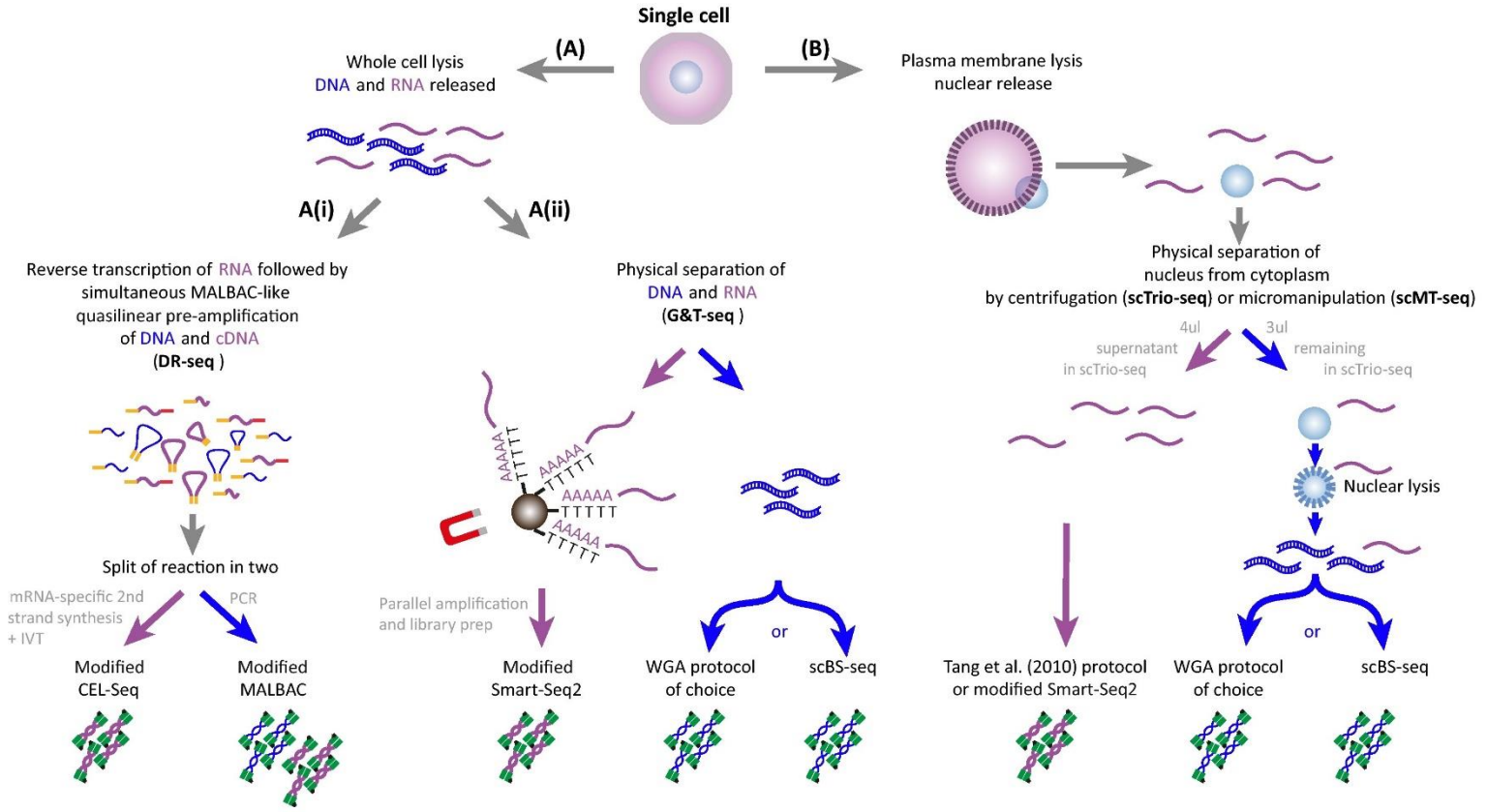
TARUN MAHAJAN

Motivation: Genotype-to-phenotype mapping



Macaulay, Iain C., Chris P. Ponting, and Thierry Voet. "Single-cell multiomics: multiple measurements from single cells." Trends in Genetics 33.2 (2017): 155-168.

Motivation: Single-cell multiomics for genotype-to-phenotype mapping



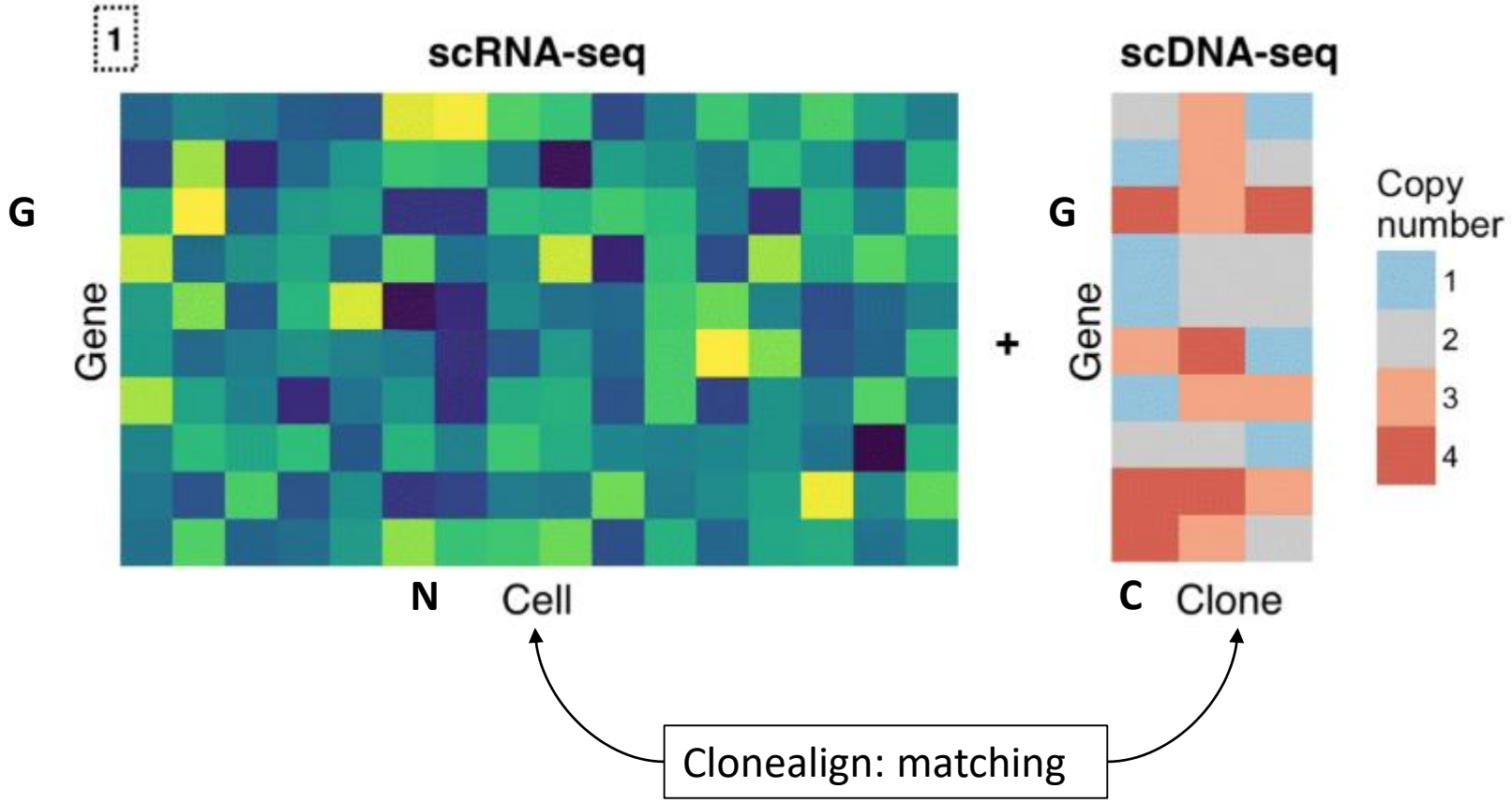
Issues: low throughput, and low quality

(C)	Loss of nucleic acids	Nature of RNA-seq	Nature of gDNA-seq	Shown amenable to bisulphite-sequencing
DR-seq	Minimal risk of loss	3'end tag transcript seq	MALBAC-like amplified gDNA, contaminated with co-amplified cDNA	no
G&T-seq (like)	Potential loss of mRNA and DNA molecules	Full-length transcript seq	In line with chosen WGA	yes
scTrio-seq	Loss of nearly half of cytoplasmic and all nuclear mRNA-molecules	Full-length transcript seq	reduced representation bisulphite-seq	yes
scMT-seq	Loss of some cytoplasmic and all nuclear mRNA-molecules during micromanipulation	Full-length transcript seq	reduced representation bisulphite-seq	yes

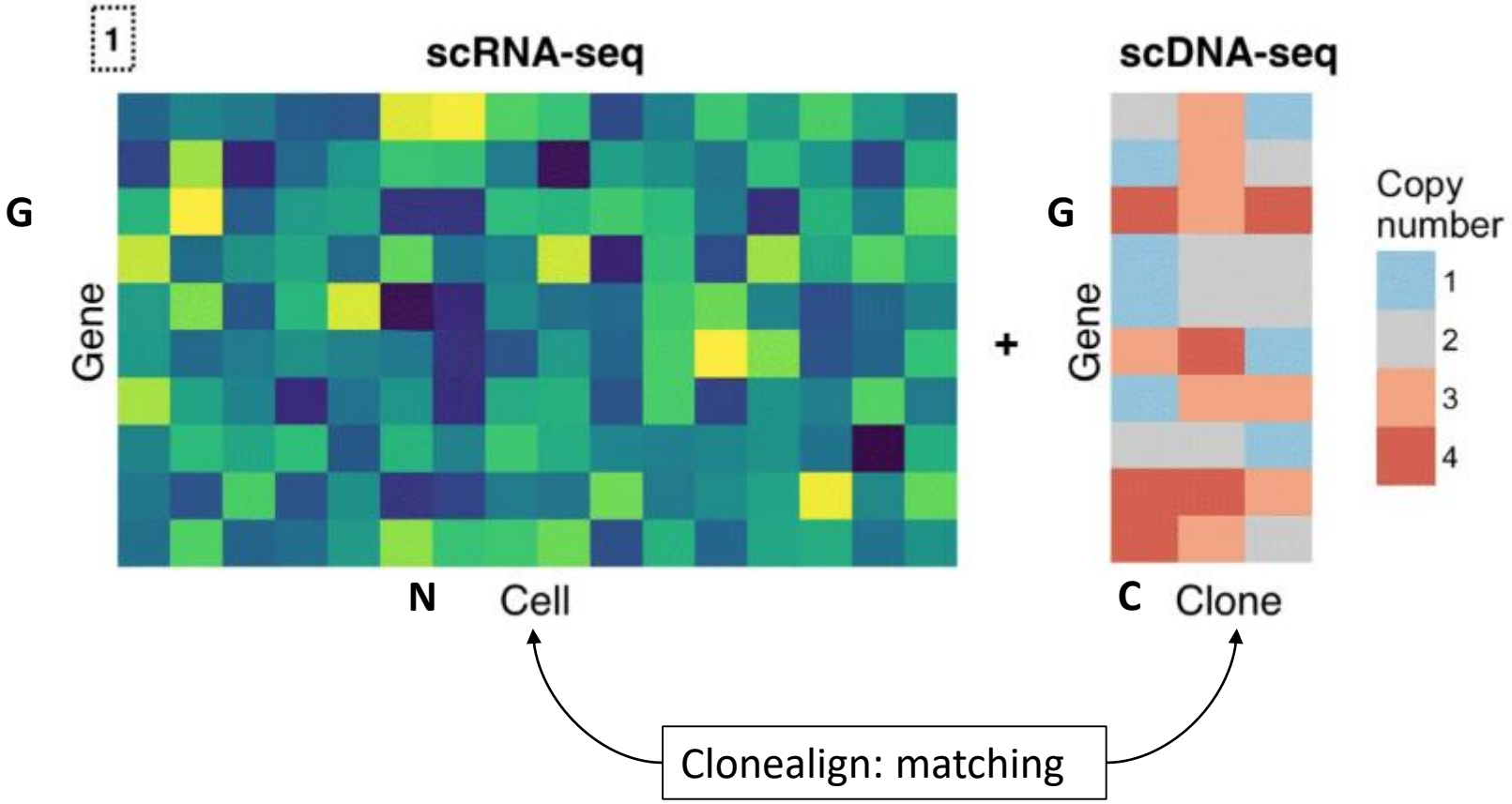
Macaulay, Iain C., Chris P. Ponting, and Thierry Voet. "Single-cell multiomics: multiple measurements from single cells." Trends in Genetics 33.2 (2017): 155-168.

Trends in Genetics

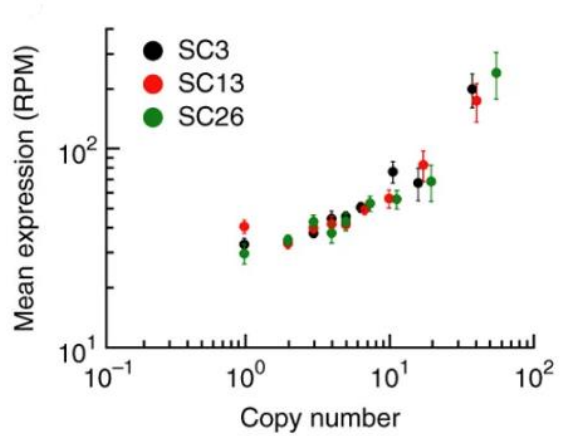
Motivation: Matching independent single-cell RNA and DNA sequencing data



Motivation: Matching independent single-cell RNA and DNA sequencing data



Copy number dosage effect



Dey, Siddharth S., et al. "Integrated genome and transcriptome sequencing of the same cell." Nature biotechnology 33.3 (2015): 285.

Problem formulation: Genome-to-Transcriptome Matching

Genome-to-Transcriptome Matching Problem:

Given $N \times G$ matrix of expression raw read counts $\boldsymbol{Y}$ for N cells and G genes, and a $G \times C$ matrix $\boldsymbol{\Lambda} = (\lambda_{gc})$ of clone specific copy numbers for C clones and G genes, find a mapping $z: [N] \rightarrow [C]$ that matches the N RNA-seq cells to the C DNA-seq clones such that expression likelihood is maximized.

Solution: Negative binomial distribution

$$f_{\mathrm{NB}}(y;\mu,\theta) = \frac{\Gamma(y+\theta)}{\Gamma(y+1)\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\mu+\theta}\right)^{y}$$
$$\forall y \in \mathbb{N}.$$

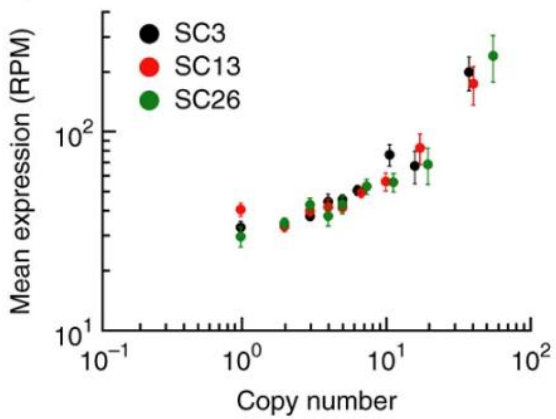
$$\underbrace{\sigma^2}_{\text{Variance}} = \underbrace{\mu}_{\text{Mean}} + \underbrace{\frac{\mu^2}{\theta}}_{\text{Inverse Over-dispersion}} = \mu + \underbrace{\phi}_{\text{Over-dispersion}}\mu^2$$

Solution: expression mean

$$\mathbb{E}[y_{ng}|z_n = c] = s_n \frac{\overbrace{\mu_g}^{\text{Per-copy expression}} \times \;f(\overbrace{\lambda_{gc}}^{\text{Copy number}}) \times e^{\overbrace{\boldsymbol{x}_n \cdot \boldsymbol{\beta}_g^T}^{\text{Known covariates}} + \overbrace{\boldsymbol{\psi}_n \cdot \boldsymbol{w}_g^T}^{\text{Residual expression}}}}{\underbrace{\sum_{g'=1}^{G} \mu_{g'} f(\lambda_{g'c}) e^{\boldsymbol{x}_n \cdot \boldsymbol{\beta}_{g'}^T + \boldsymbol{\psi}_n \cdot \boldsymbol{w}_{g'}^T}}_{\text{Total RNA normalization}}}$$

$\underbrace{s_n}_{\text{Cell read depth}}$

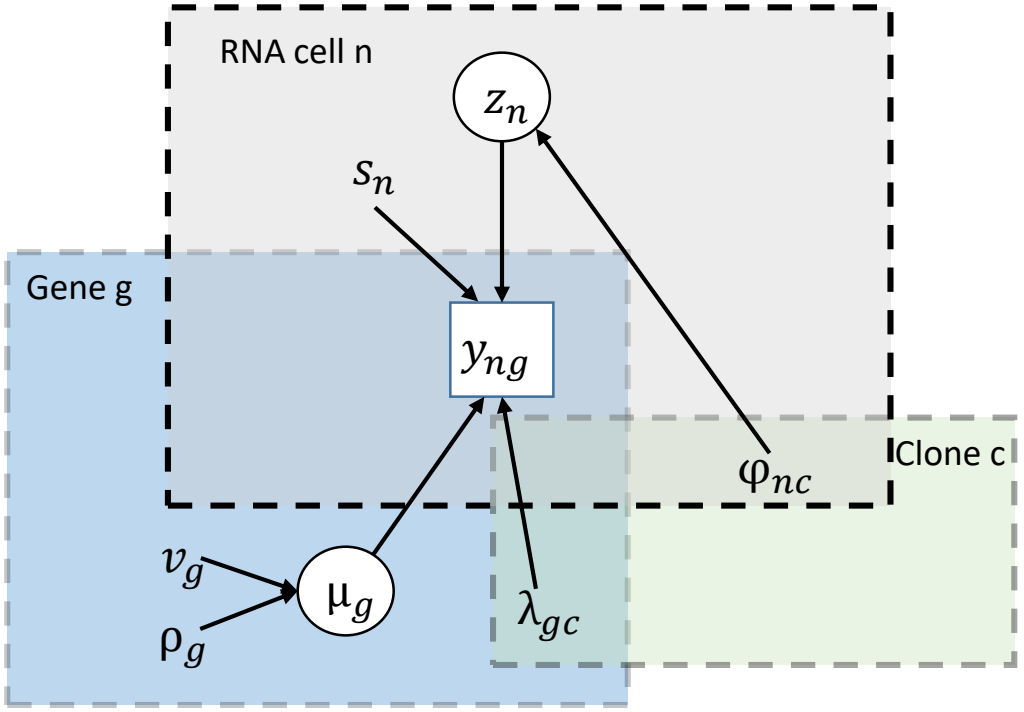
$$f(\lambda) = \begin{cases} \lambda & \text{if } \lambda < \zeta \\ \zeta & \text{if } \lambda \geq \zeta, \end{cases}$$



Solution: over-dispersion

$$\phi(\mu) = \sum_{i=1}^{M} a_i \exp(-b(\mu - c_i)^2)$$

Solution: graphical model



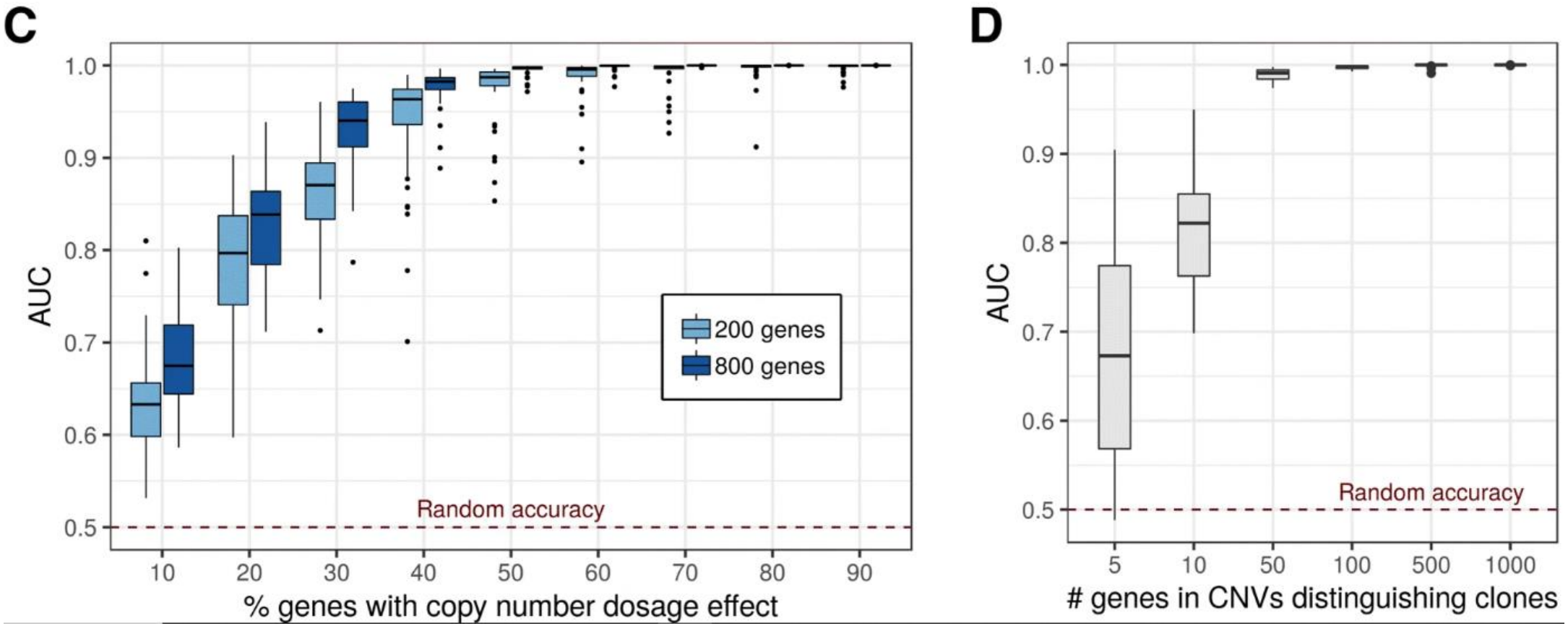
Solution: estimate posterior distribution for mapping *z:* [*N*] → [*C*] and mean parameter *μ*

$$p(z, \mu|y) = \frac{p(y|z, \mu)}{p(y)}$$

$$p(y) = \iint p(y|z, \mu)\, dz\, d\mu$$

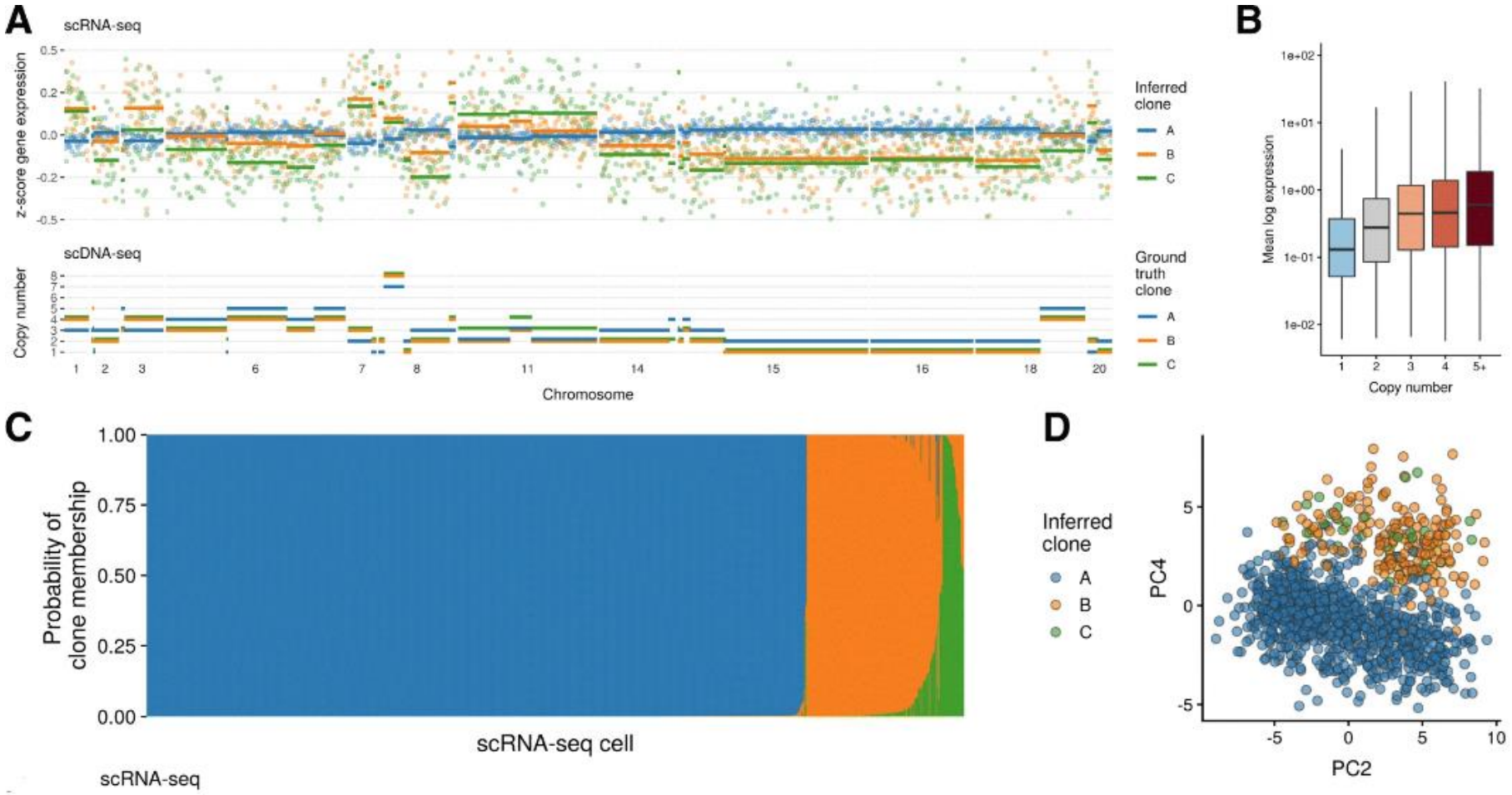
- Data likelihood, p(y), is intractable and can only be computed numerically in exponential time.
- MCMC methods are computationally expensive and do not scale well for large data.
- Using mean-field variational bayes to estimate posterior distribution by solving an optimization problem.

Validation: Simulation



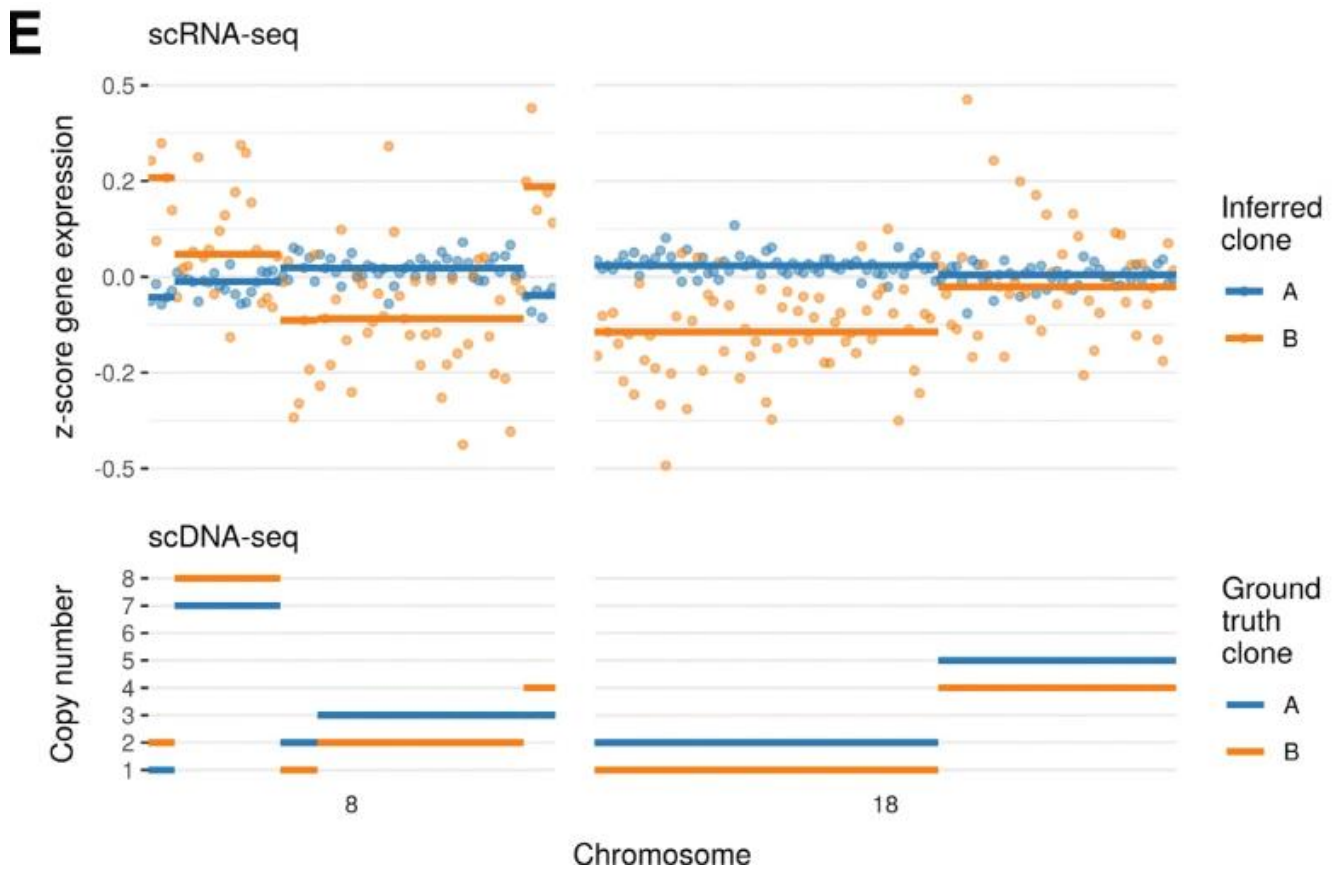
C Simulations demonstrate the robustness of clonealign to the underlying proportion of genes exhibiting a copy number dosage effect. Even if only 30% of genes have a clone-specific copy number effect on expression, clones can still be accurately assigned with an average AUC >0.8. **D** Simulations demonstrate clonal assignment is accurate even when as few as 10–50 genes lie in regions of differing copy number between clones, allowing clonal assignment from only small-scale genomic rearrangements

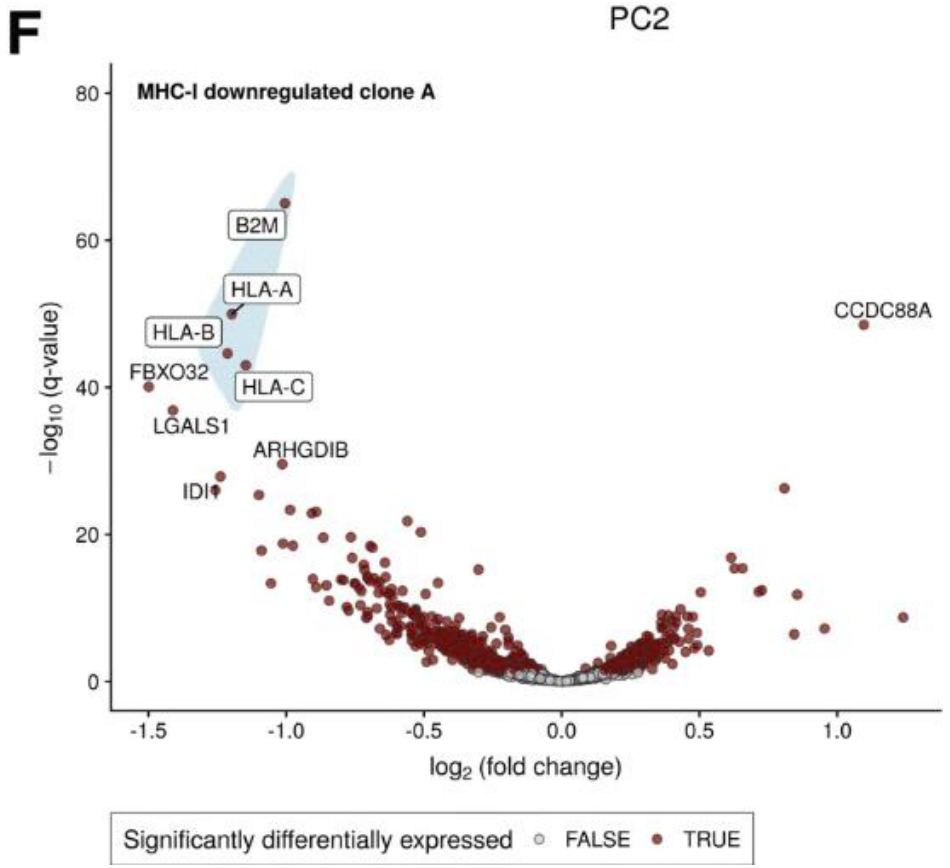
Validation: Breast cancer data



A Clone-specific copy number for ground truth clones in scDNA-seq (bottom) and clone-specific z-score expression for clonealign inferred clones in scRNA-seq (top) for regions exhibiting inter-clone copy number aberrations **B** The mean log expression as a function of copy number across all clones. **C** Clone assignment probabilities for 1152 single-cell RNA-seq profiles across three clones. **D** A PCA projection using only genes residing in copy number regions shows the cells clustering by clone along components 2 and 4.

Validation: Breast cancer data





E z-score normalized gene expression and copy number profiles for held-out data on chromosomes 8 and 18 as a function of genomic position (gene index along chromosome). In all but one copy number segment, when the copy number profile of a clone is higher, the normalized gene expression in that chromosome is also higher on average. **F** Differential expression analysis for genes residing in regions whose copy number is identical between clones highlights downregulation of MHC class I proteins

Validation: Breast cancer data

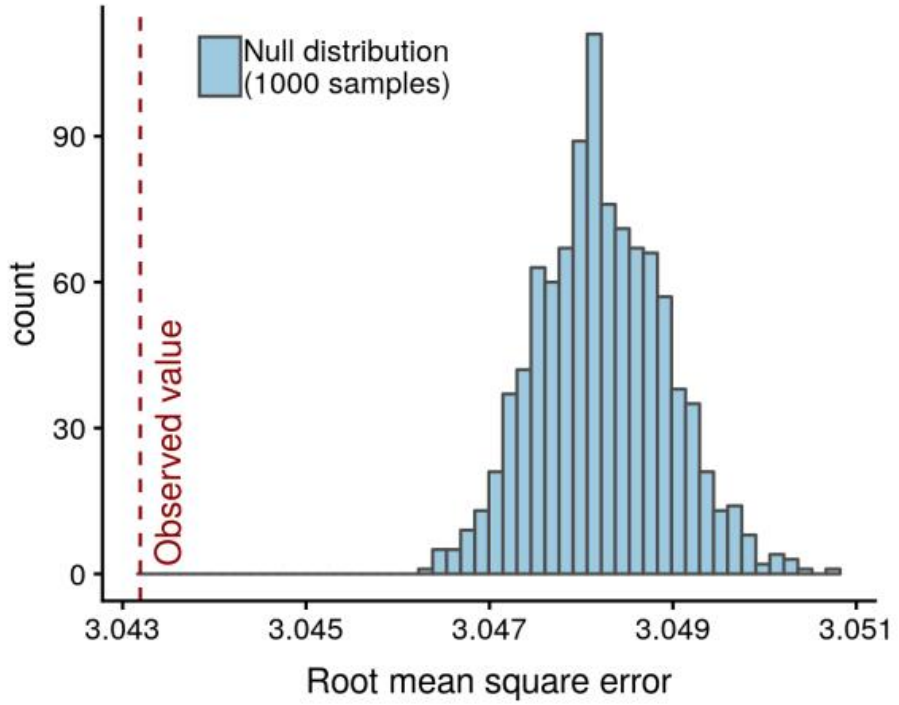
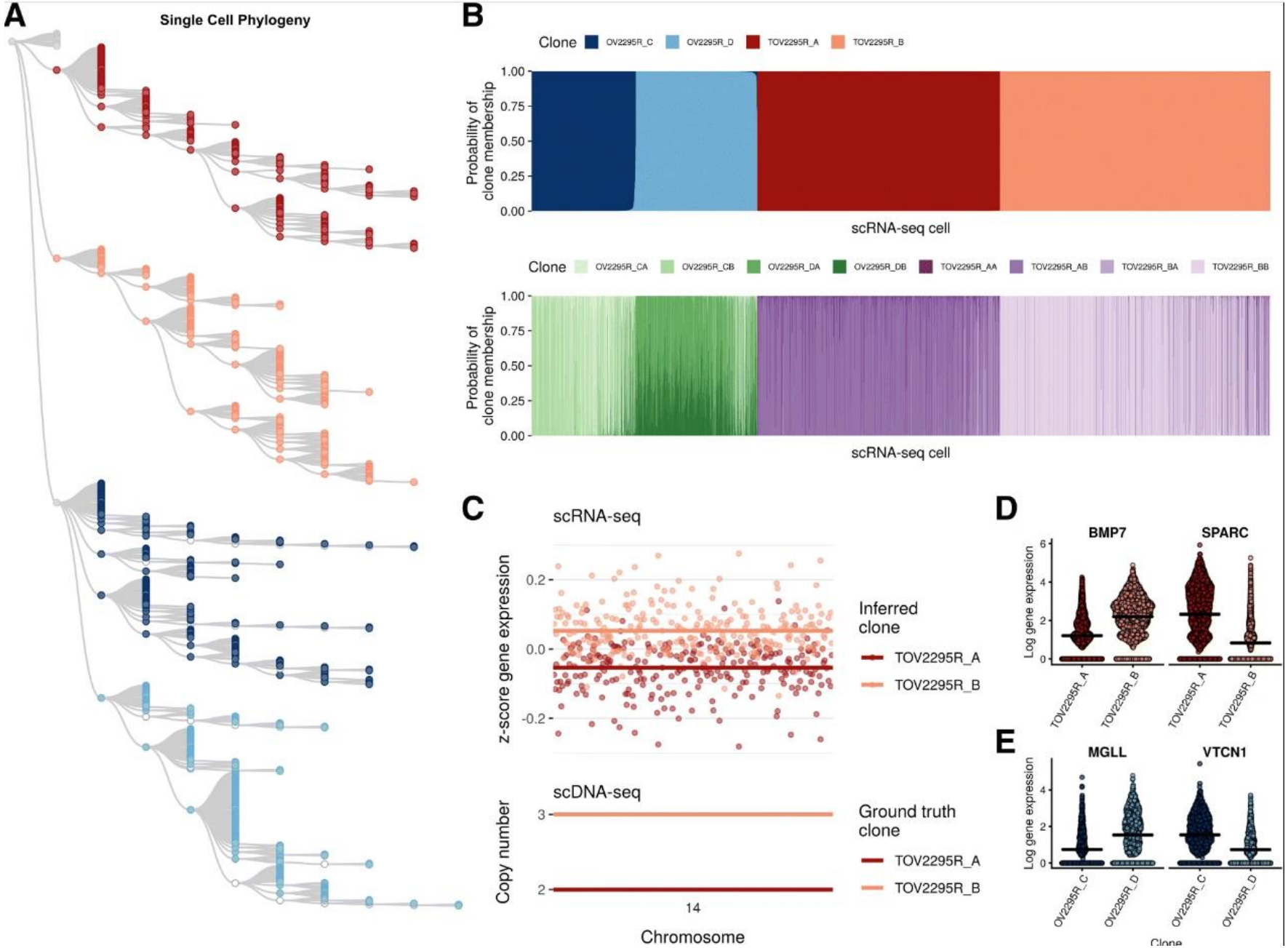


Figure S1: Distribution of root mean square error in predicting the expression of genes on held out chromosomes (8 & 18) for SA501 under random repeated permutation of clone assignments (light blue) compared to the observed error under clonealign assignments (red dashed arrow). This demonstrates the observed error is significantly less than is observed at random ($p < 10^{-3}$).

Validation: Ovarian cancer data



Validation: Ovarian cancer data

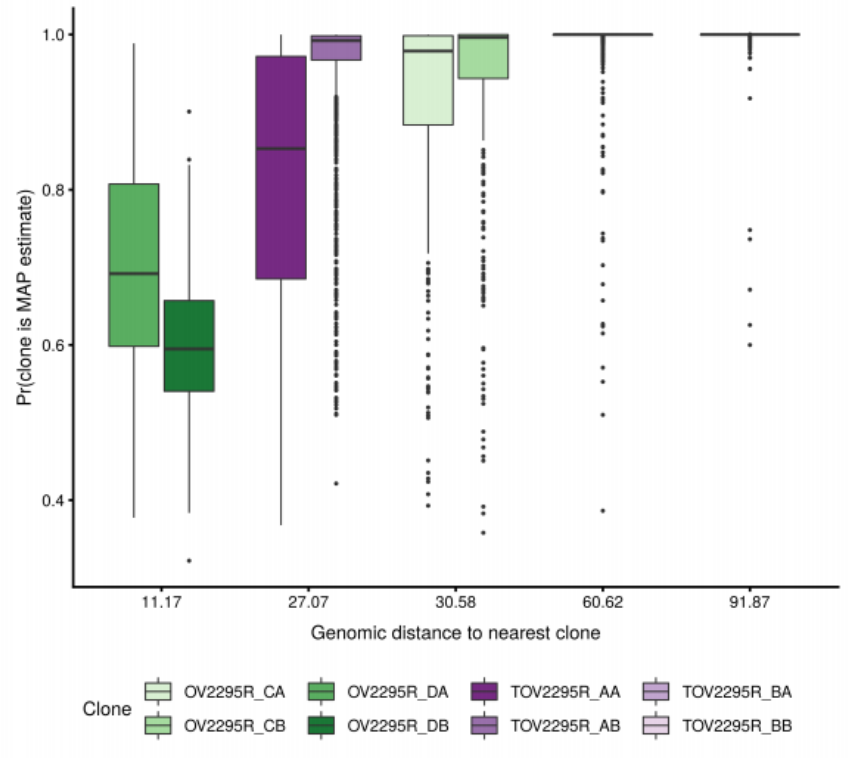
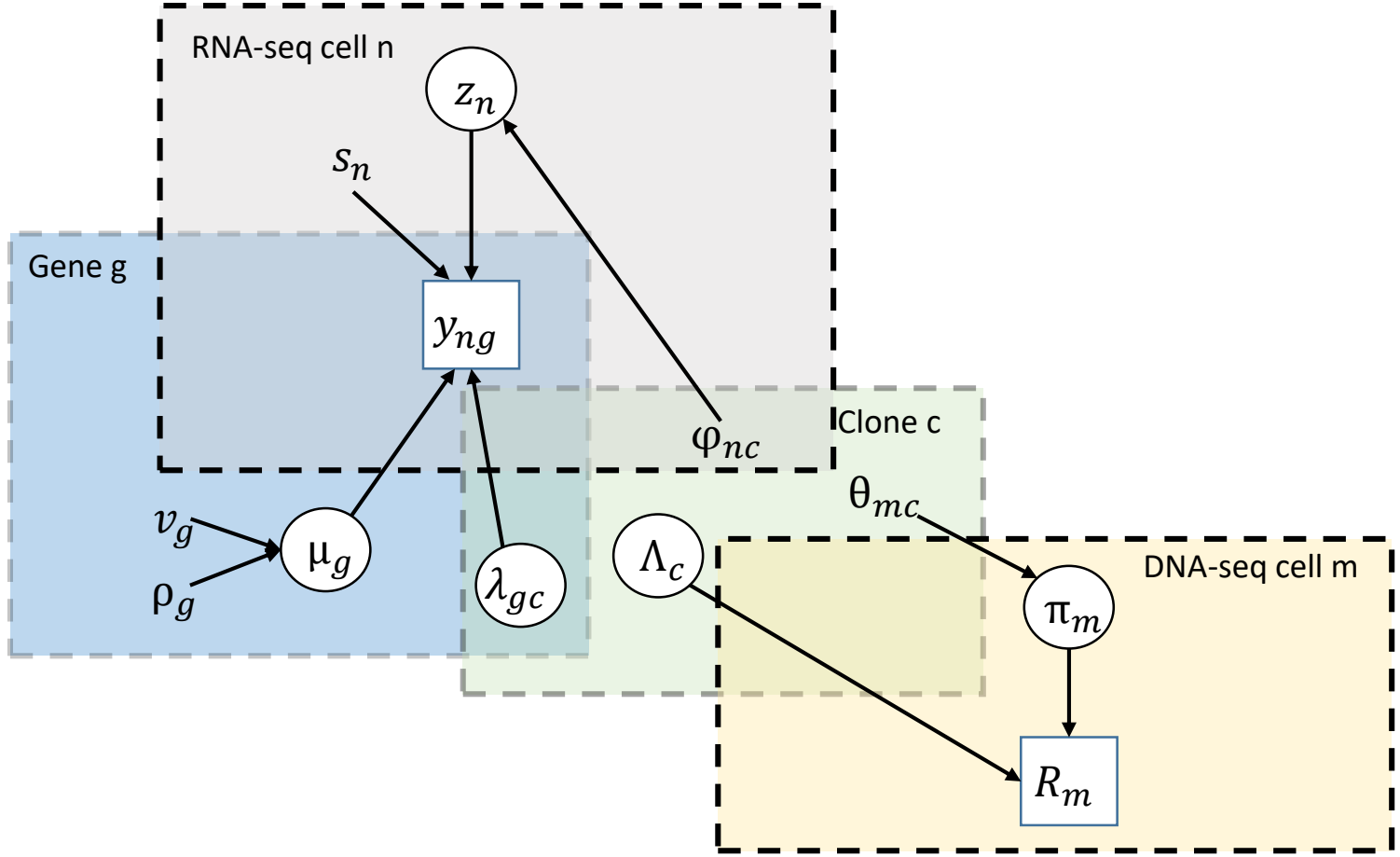


Figure S20: The maximum likelihood probability for a cell to be assigned to a clone as a function of the genomic distance (euclidean distance in copy number space) to the nearest clone. The more distinct clones are, the more certainty in clonal assignment, while for clones that are very close in copy number space the model assigns uncertainty to the assignment in RNA-space.

CS598MEB Project proposal : Inference of clone-specific expression and copy number profiles using multi-omics single-cell data



Given expression matrix ***Y*** for *N* cells and *G* genes, and an *M*×*L* matrix ***R*** of DNA-seq read counts for *M* DNA-seq cells and *L* genomic bins, find a mapping θ: $[M] \rightarrow [C]$ that matches the M DNA-seq cells to *C* clonal clusters, matrix $\boldsymbol{\Lambda} = (\lambda_{gc})$ of copy numbers for *C* clones and *G* genes, and a mapping z: $[N] \rightarrow [C]$ that matches the *N* RNA-seq cells to the *C* DNA-seq clones such that expression likelihood is maximized.

Solution: A quick introduction to variational bayes

Data

$$p(\underbrace{\mathbf{z}}_{\text{Parameters}} \mid \overbrace{\mathbf{x}}^{\text{Data}}) = \frac{p(\mathbf{z}, \mathbf{x})}{p(\mathbf{x})}.$$

Parameters

Solution: A quick introduction to variational bayes

Data

$$p(\underbrace{\mathbf{z}}_{\text{Parameters}} \mid \overbrace{\mathbf{x}}^{\text{Data}}) = \frac{p(\mathbf{z}, \mathbf{x})}{p(\mathbf{x})}.$$

- Approximate the posterior with a variational distribution q(z) such that the KL divergence is minimized

$$q^*(\mathbf{z}) = \underset{q(\mathbf{z}) \in \mathcal{Q}}{\arg\min}\ kl\left(q(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x})\right).$$

Optimization problem

Solution: A quick introduction to variational bayes

Data

$$p(\underbrace{\mathbf{z}}_{\text{Parameters}} \mid \overbrace{\mathbf{x}}^{\text{Data}}) = \frac{p(\mathbf{z}, \mathbf{x})}{p(\mathbf{x})}.$$

- Approximate the posterior with a variational distribution q(z) such that the KL divergence is minimized

$$q^*(\mathbf{z}) = \underset{q(\mathbf{z}) \in \mathcal{Q}}{\arg\min}\ kl\left(q(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x})\right).$$

Optimization problem

$$kl\left(q(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x})\right) = \mathbb{E}\left[\log q(\mathbf{z})\right] - \mathbb{E}\left[\log p(\mathbf{z}, \mathbf{x})\right] + \log p(\mathbf{x}).$$

KL divergence

Solution: A quick introduction to variational bayes

Data

$$p(\underbrace{\mathbf{z}}_{\text{Parameters}} \mid \overbrace{\mathbf{x}}^{\text{Data}}) = \frac{p(\mathbf{z}, \mathbf{x})}{p(\mathbf{x})}.$$

- Approximate the posterior with a variational distribution q(z) such that the KL divergence is minimized

$$q^*(\mathbf{z}) = \underset{q(\mathbf{z}) \in \mathcal{Q}}{\arg\min}\ kl\left(q(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x})\right).$$

Optimization problem

$$kl\left(q(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x})\right) = \mathbb{E}\left[\log q(\mathbf{z})\right] - \mathbb{E}\left[\log p(\mathbf{z}, \mathbf{x})\right] + \log p(\mathbf{x}).$$

KL divergence

$$elbo(q) = \mathbb{E}\left[\log p(\mathbf{z}, \mathbf{x})\right] - \mathbb{E}\left[\log q(\mathbf{z})\right].$$

Evidence Lower Bound (ELBO)

Solution: A quick introduction to variational bayes

Data

$$p(\underbrace{\mathbf{z}}_{\text{Parameters}} \mid \overbrace{\mathbf{x}}^{\text{Data}}) = \frac{p(\mathbf{z}, \mathbf{x})}{p(\mathbf{x})}.$$

- Approximate the posterior with a variational distribution q(z) such that the KL divergence is minimized

$$q^*(\mathbf{z}) = \underset{q(\mathbf{z}) \in \mathcal{Q}}{\arg\min}\ kl\left(q(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x})\right).$$

Optimization problem

$$kl\left(q(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x})\right) = \mathbb{E}\left[\log q(\mathbf{z})\right] - \mathbb{E}\left[\log p(\mathbf{z}, \mathbf{x})\right] + \log p(\mathbf{x}).$$

KL divergence

$$elbo(q) = \mathbb{E}\left[\log p(\mathbf{z}, \mathbf{x})\right] - \mathbb{E}\left[\log q(\mathbf{z})\right].$$

Evidence Lower Bound (ELBO)

$$\log p(\mathbf{x}) = kl\left(q(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x})\right) + elbo(q).$$

Maximize ELBO to minimize KL divergence

Solution: Mean-field variational bayes for clonealign

$$q(\boldsymbol{z}, \boldsymbol{\mu}) = \prod_n q(z_n) \prod_g q(\mu_g)$$

Latent variables are mutually independent

$$q(z_n{=}c){=}\varphi_{nc}$$

Variational distribution for clone assignment

$$\mu_g{=}\ \exp(\nu_g{+}\rho_g\varepsilon) \qquad \epsilon \sim \mathcal{N}(0,1)$$

Variational distribution for mean expression parameter

Solution: Mean-field variational bayes for clonealign

$$q(\boldsymbol{z}, \boldsymbol{\mu}) = \prod_n q(z_n) \prod_g q(\mu_g)$$

Latent variables are mutually independent

$$q(z_n{=}c){=}\varphi_{nc}$$

Variational distribution for clone assignment

$$\mu_g{=}\ \exp(\nu_g{+}\rho_g\varepsilon) \qquad \epsilon \sim \mathcal{N}(0,1)$$

Variational distribution for mean expression parameter

$$\boldsymbol{E}_{q(\mu)}[f(z)] = \boldsymbol{E}_{\boldsymbol{p}(\varepsilon)}[f(\nu_g + \rho_g\varepsilon)] \approx 1/L\sum_{l=1}^{L} f\left(\nu_g + \rho_g\varepsilon^l\right), \varepsilon^l \sim p(\varepsilon)$$

Solution: Mean-field variational bayes for clonealign

$$q(\boldsymbol{z}, \boldsymbol{\mu}) = \prod_n q(z_n) \prod_g q(\mu_g)$$

Latent variables are mutually independent

$$q(z_n{=}c){=}\varphi_{nc}$$

Variational distribution for clone assignment

$$\mu_g{=}\exp(\nu_g{+}\rho_g\varepsilon) \qquad \epsilon \sim \mathcal{N}(0,1)$$

Variational distribution for mean expression parameter

$$\boldsymbol{E}_{q(\mu)}[f(z)] = \boldsymbol{E}_{\boldsymbol{p}(\varepsilon)}[f(\nu_g + \rho_g\varepsilon)] \approx 1/L\sum_{l=1}^{L} f\left(\nu_g + \rho_g\varepsilon^l\right), \varepsilon^l \sim p(\varepsilon)$$

$$elbo(q) = E_{q(z,\mu)}[p_\theta(y|z,\mu)] - kl(q(z,\mu)||p_\theta(z,\mu))$$

Maximize