BASiCS: Bayesian Analysis of Single Cell Sequencing data (+scran)

Catalina Vallejos
Chancellor's Fellow, MRC Human Genetics Unit
Fellow, The Alan Turing Institute

@CataVallejosM

Questions

Q1. How do you define normalization? Why is it important?

Q2. How do you normalize? (This can be a discussion of your method)

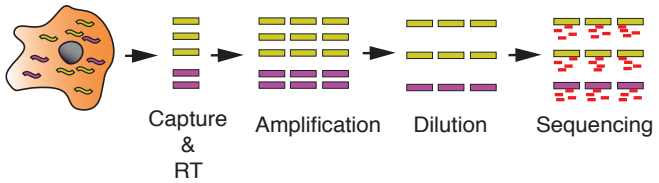
Q3. How do you demonstrate success? How do you know if you've normalized properly?

Q4. Where does your method break? Or where do you see challenges with your method or others?

Q5. What is your suggestion for how we spend the second day of the workshop to collaborate on computational problems? Are there particular tasks or challenges that you think would be good to address?

Normalising scRNAseq data — global scaling

$$\mathrm{E}(X_{ij}) = \underbrace{s_j}_{\text{cell effect}} \times \overbrace{\mu_i}^{\text{gene effect}}$$



Normalised data

$$\tilde{X}_{ij} = X_{ij} / \hat{s}_j$$

Adjusted model

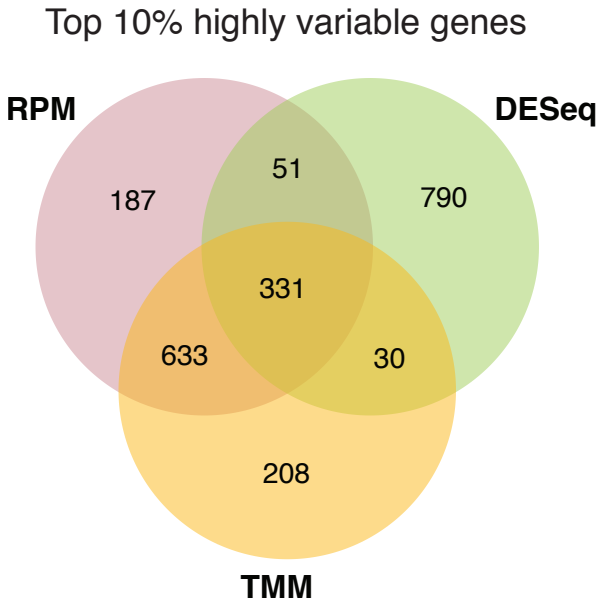
$$X_{ij} \sim \mathrm{NB}(s_j \mu_i, \delta_i)$$

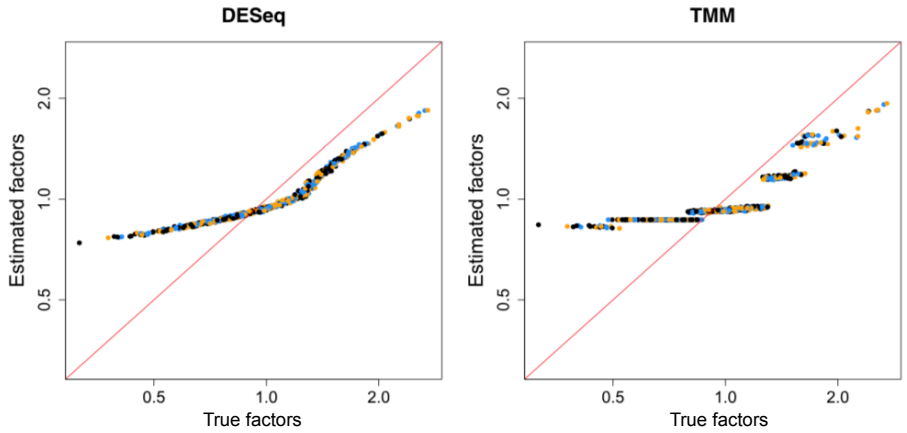
Fixed offsets

Joint inference

Note: some of these are removed when using UMIs

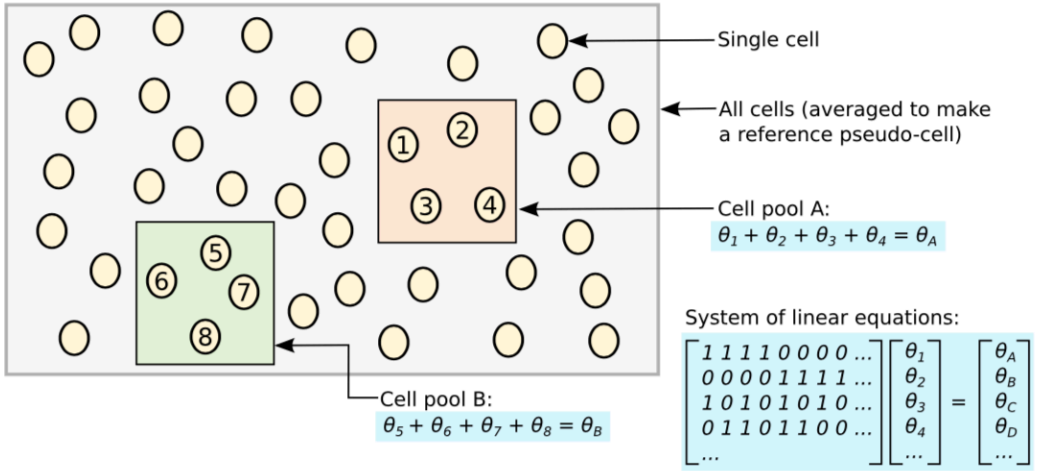
Normalising scRNAseq data — bulk methods?

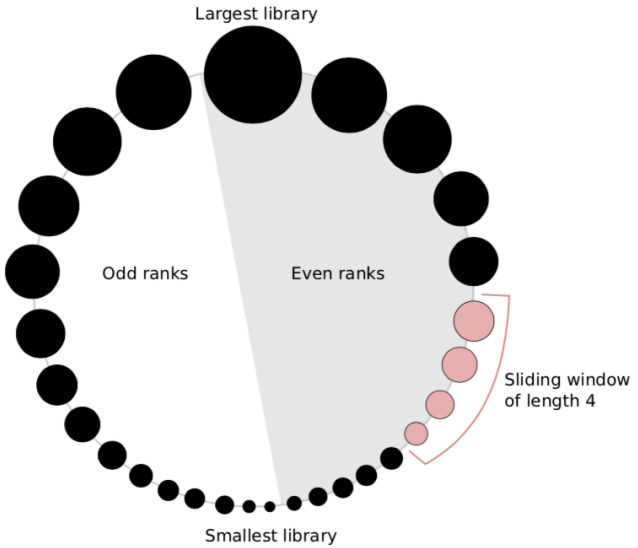




Vallejos, Risso, Scialdone *et al* (2017) *Nature Methods*
Lun *et al* (2016) *Genome Biology*

scran: pooling across cells to normalise scRNAseq data





Lun *et al* (2016) *Genome Biology*

Q2

BASiCS: Bayesian Analysis of Single Cell Sequencing data

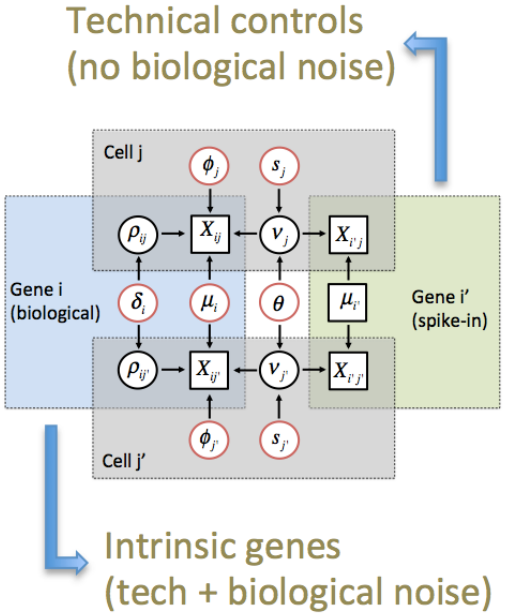
BASiCS is an integrated Bayesian statistical framework that **simultaneously** performs:

- Data normalisation
- Technical noise quantification
- **Supervised** downstream analyses
  - Detection of highly/lowly variable genes **within** a population
  - Differential mean and variability testing **between** populations

Integrated approach to propagate statistical uncertainty

Vallejos, Marioni and Richardson (2015) *PLoS Computational Biology*
Vallejos, Richardson and Marioni (2016) *Genome Biology*
Eling *et al* (2018) *Cell Systems*

BASiCS: Bayesian Analysis of Single Cell Sequencing data



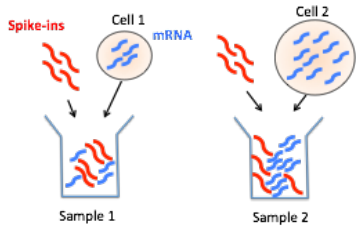
Hierarchical Bayesian model to capture **technical** and **biological** over-dispersion

$$X_{ij} \mid \mu_i, \phi_j, \nu_j, \rho_{ij} \sim \text{Poisson}(\phi_j \nu_j \mu_i \rho_{ij})$$

$$\nu_j \mid s_j, \theta \sim \text{Gamma}(s_j^{-1}, (s_j\theta)^{-1})$$

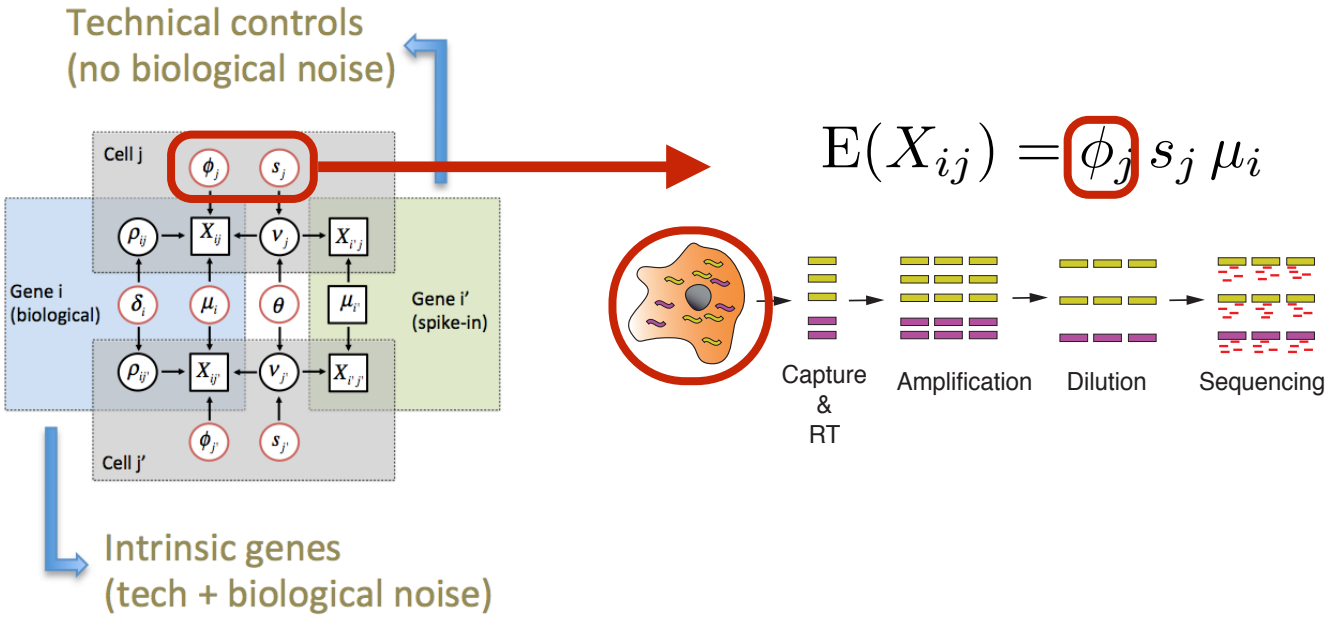
$$\rho_{ij} \mid \delta_i \sim \text{Gamma}(\delta_i^{-1}, \delta_i^{-1})$$

Requires pre-filtered data

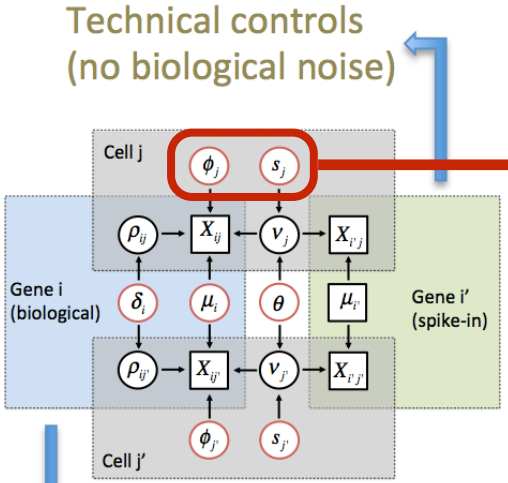


spike-in genes

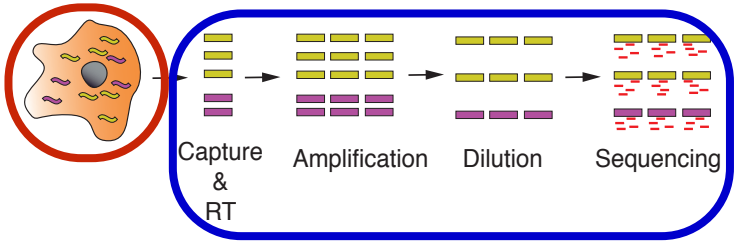
BASiCS: Bayesian Analysis of Single Cell Sequencing data



BASiCS: Bayesian Analysis of Single Cell Sequencing data

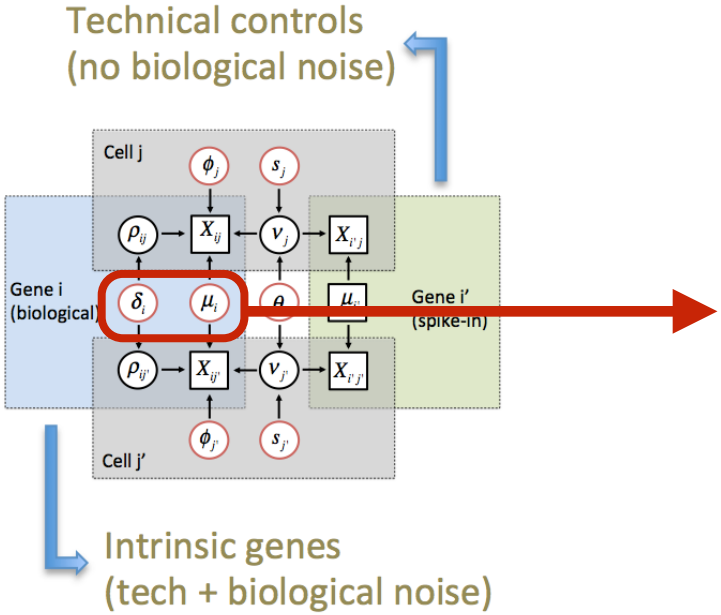


$$\mathrm{E}(X_{ij}) = \phi_j s_j \mu_i$$



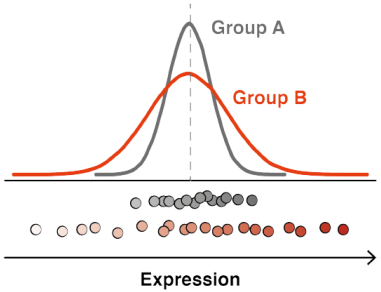
One set of scaling factors if spike-ins are not available

BASiCS: Bayesian Analysis of Single Cell Sequencing data

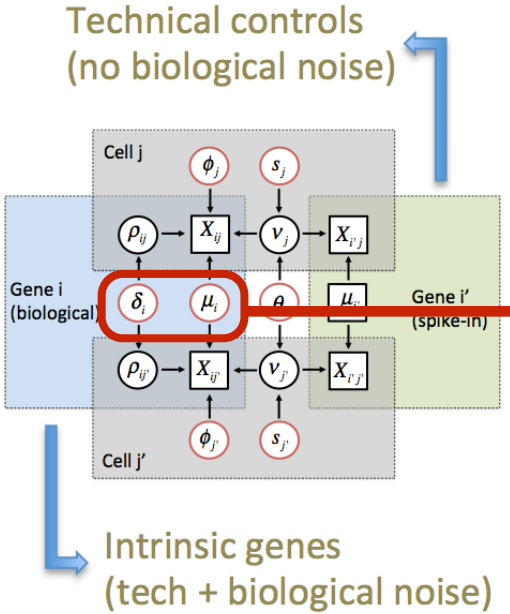


Gene-specific parameters:

- mean expression μ_i
- biological over-dispersion δ_i

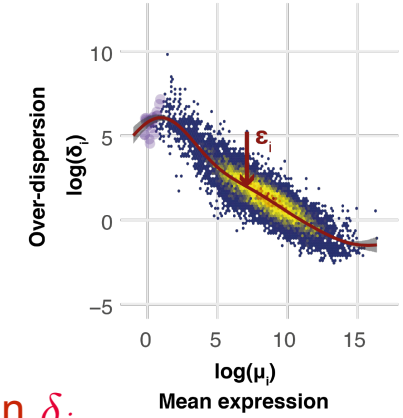


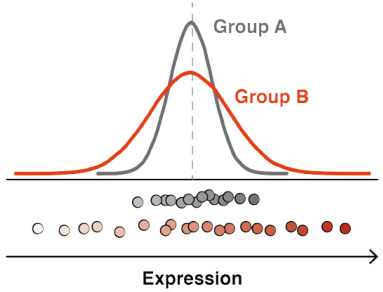
BASiCS: Bayesian Analysis of Single Cell Sequencing data



Gene-specific parameters:

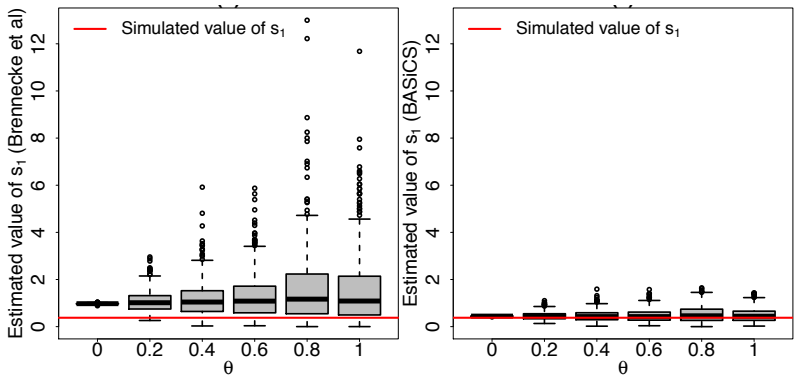
- mean expression μ_i
- biological over-dispersion δ_i



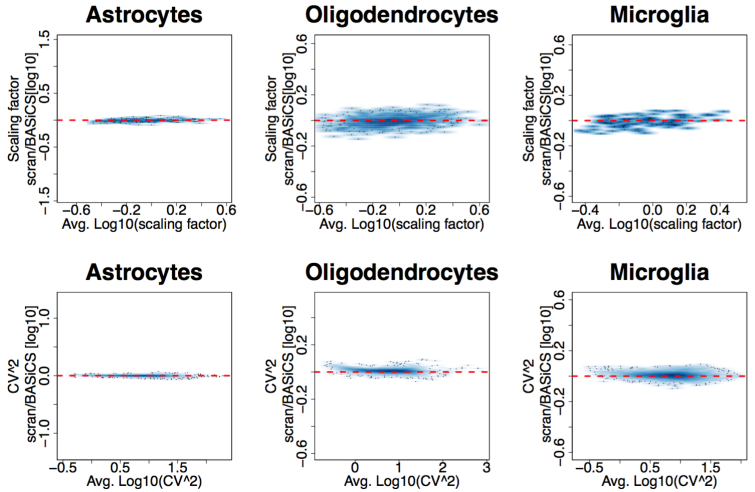


Does it work?

BASiCS and DESeq[1] — synthetic data



BASiCS and scran[2] — Zeisel *et al* data[3]

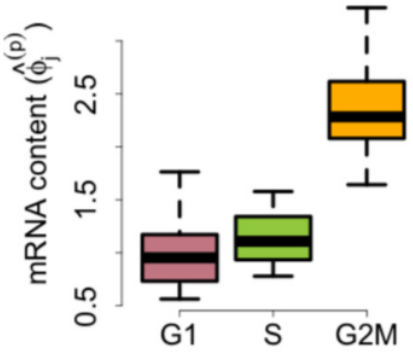


1. Anders and Huber (2010) Genome Biology

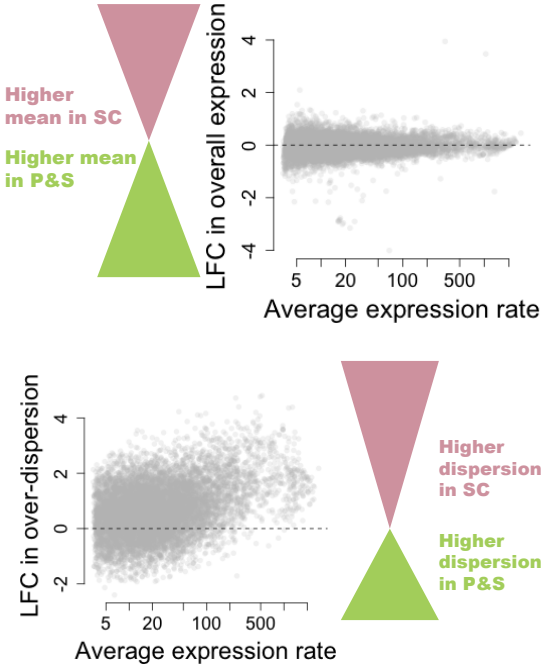
2. Lun *et al* (2016) *Genome Biology* 3. Zeisel *et al* (2015) *Science*

Does it work?

mRNA content across cell cycle



A pool and split experiment



Does it work?

Predictive checks

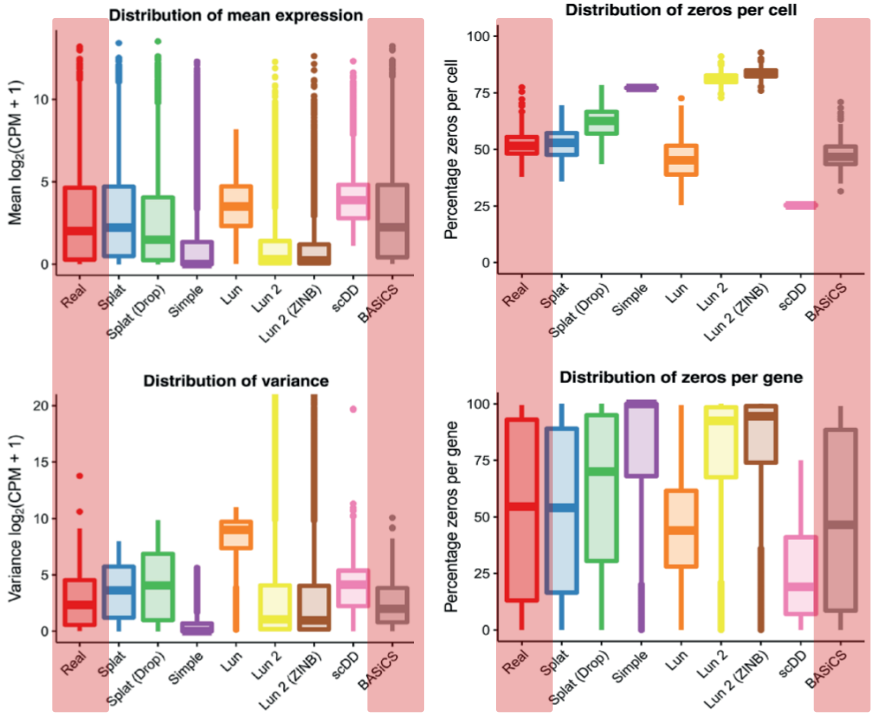
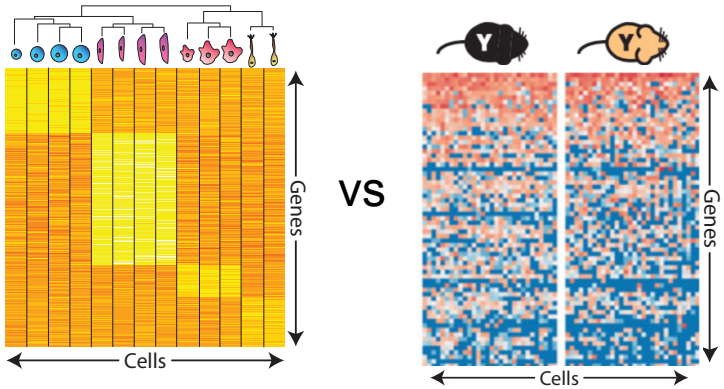


Image adapted from Zappia *et al* (2017) *Genome Biology*

Dealing with structural heterogeneity — scran & BASiCS



From Martinez-Jimenez, Eling *et al* (2017) Science

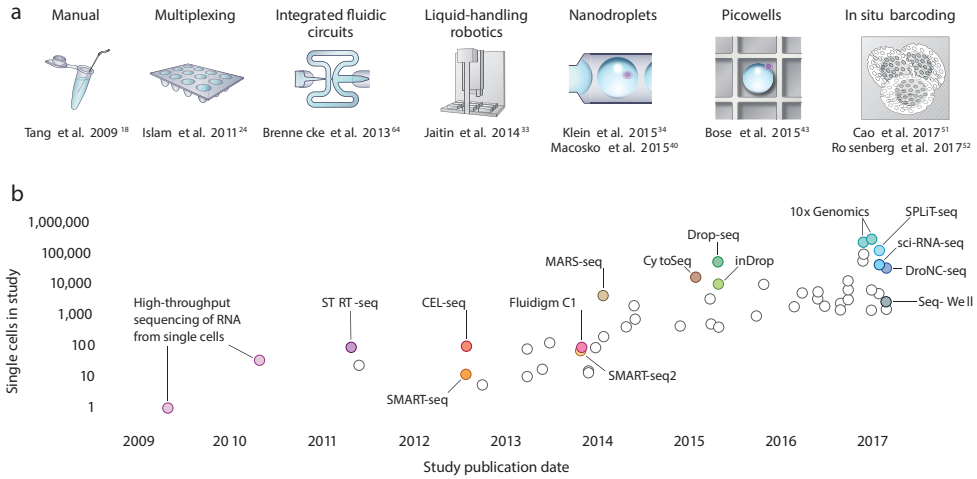
If structural heterogeneity (strong DE) exists:

- *Pre-cluster* cells
- Normalise cells within each cluster
- Normalise between clusters

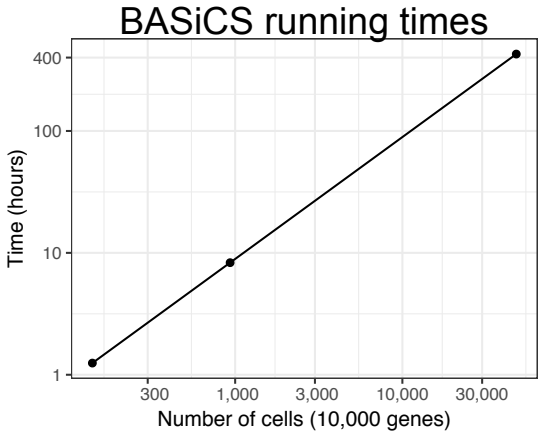
From Buettner *et al* (2015) *Nat Biotech*

?

Scalability - ongoing work



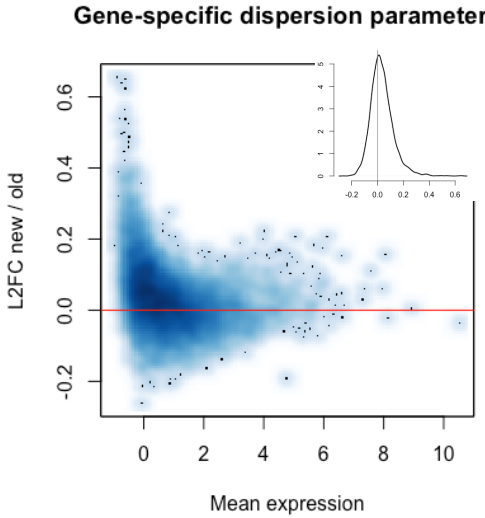
From Svensson et al (2018) Nature Protocols

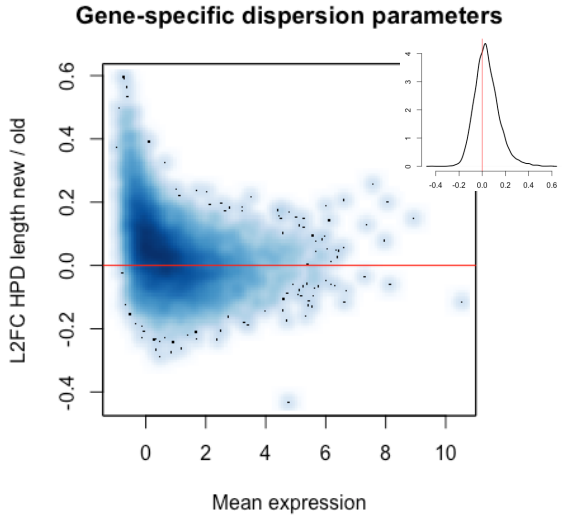


@AlanBOCallaghan

Integrated approach to propagate statistical uncertainty …

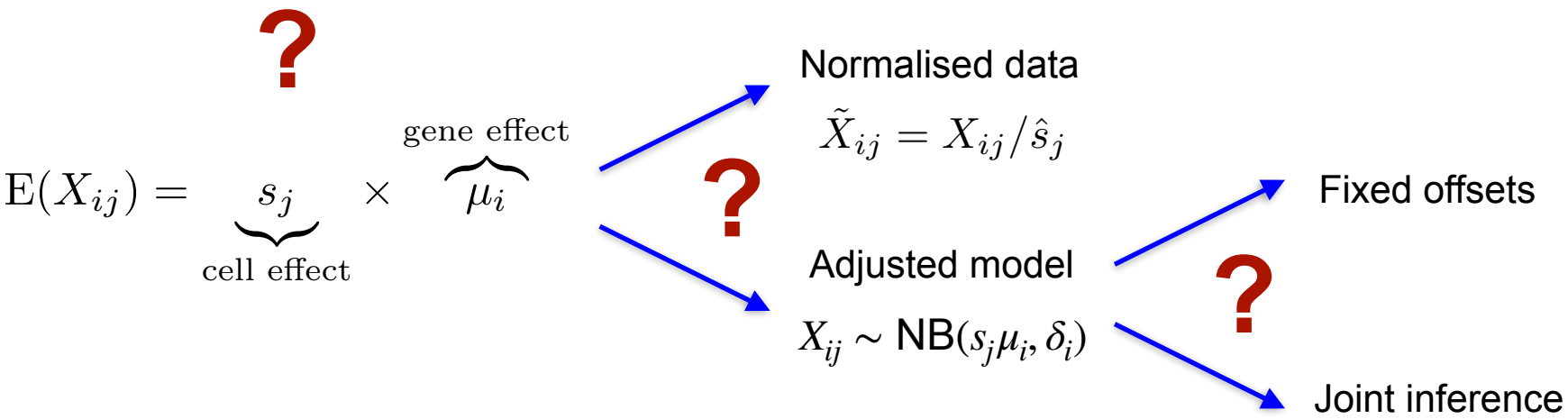
… what if we run BASiCS with scran size factors as fixed offsets?





Note: posterior uncertainty about the normalisation parameters was low

For tomorrow?



Acknowledgements

vallejosgroup.github.io
@CataVallejosM

MRC Human Genetics Unit

Vallejos Group
Alan O'Callaghan
Andreas Kapourani
Andrew Papanastasiou
Ava Khamseh
Christos Maniatis
Evgenii Lobzaeb
Fraser McPhie
Nathan Constantine-Cooke

MRC Biostatistics Unit

Sylvia Richardson

EMBL-EBI

John Marioni
Nils Eling

Cancer Research UK Cambridge Institute

Arianne Richard
Aaron Lun

HelmholtzZentrum münchen
German Research Center for Environmental Health

Antonio Scialdone

Berkeley
University of California

Sandrine Dudoit

Davide Risso

British Heart Foundation

The Alan Turing Institute

MRC Medical Research Council

The University of Edinburgh