Statistical challenges in the analysis of single-cell transcriptomics data

Catalina Vallejos

joint work with Sylvia Richardson and John Marioni

MRC Biostatistics Unit

EMBL-EBI

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Overview

Single-cell transcriptomics

Notation

Statistical challenges in the analysis of scRNA-seq data

BASiCS: Bayesian Analysis of Single-Cell Sequencing data

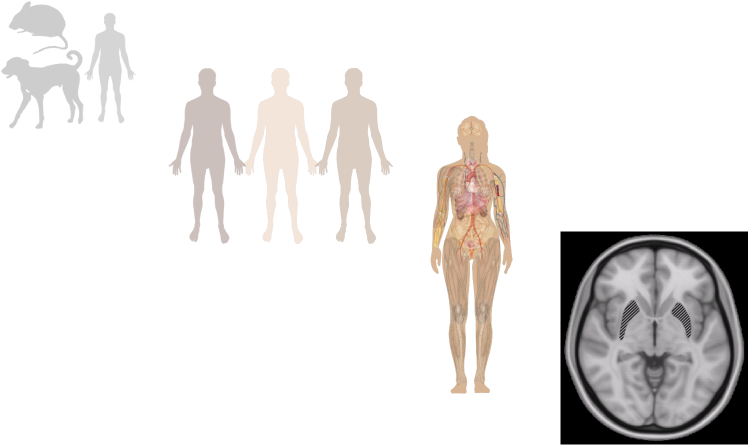
BASiCS: Posterior inference

Final remarks

Single-cell transcriptomics

Biological heterogeneity

There are multiple levels of biological heterogeneity



All images are public domain (source: https://commons.wikimedia.org/)

Understanding heterogeneity at the single-cell level

Most transcriptomic studies have focused on examining expression in large populations of cells[1,2]

Some biological processes, however, require the study of variation in gene expression at the single-cell level[3,4]

Single-cell RNA-sequencing (scRNA-seq) quantifies gene expression profiles of individuals cells

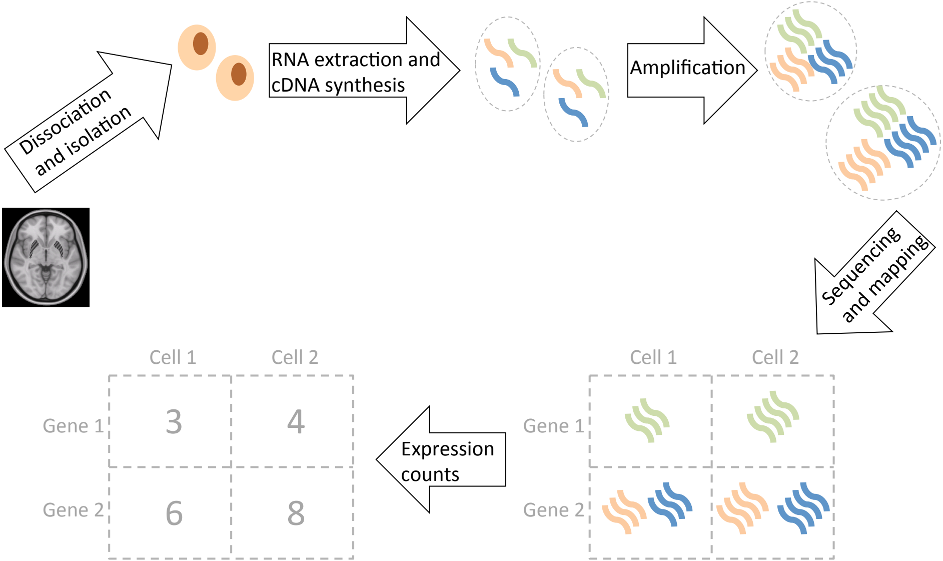
[1] Marioni et al., Genome Res (2008)

[2] Pickrell et al. Nature, (2010)

[3] Hayashi et al., Science (2007)

[4] Diez-Roux et al., PLoS Biol (2011)

scRNA-seq workflow



Dissociation and isolation

RNA extraction and cDNA synthesis

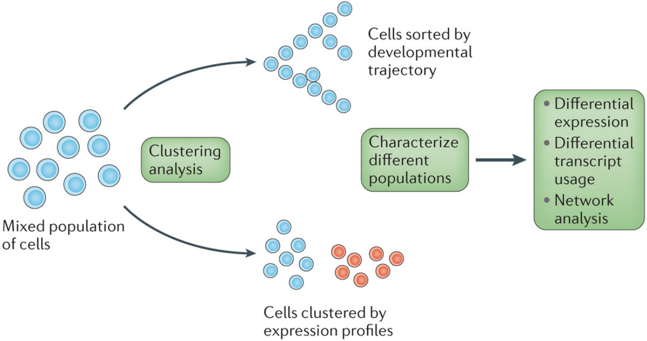
Amplification

Sequencing and mapping

Expression counts

	Cell 1	Cell 2
Gene 1	3	4
Gene 2	6	8

The power of scRNA-seq


Cells sorted by developmental trajectory
Mixed population of cells
Clustering analysis
Characterize different populations
• Differential expression
• Differential transcript usage
• Network analysis
Cells clustered by expression profiles

Stegle et al., Nature Reviews (2015)

Already this has led to identification of novel:

- Neuronal populations[1]
- Immune cell populations[2]
- Sub-populations of tumour cells[3]

[1] Zeisel et al., Science, (2015) [2] Jaitin et al., Science, (2014) [3] Patel et al., Science, (2014)

Notation

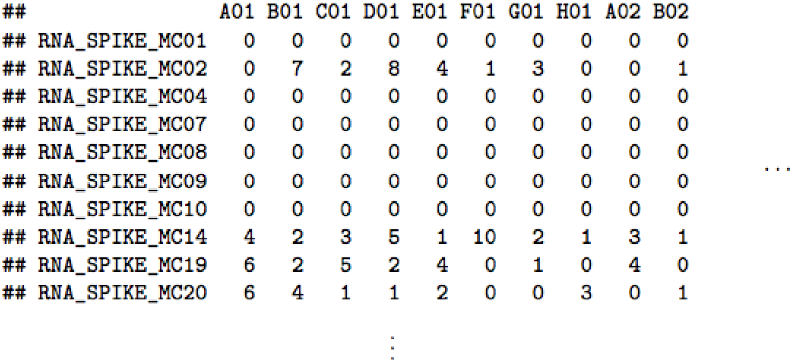
scRNA-seq data

scRNA-seq data can be represented as

$$\begin{array}{cc} \begin{array}{cccc} \text{cell } 1 & \text{cell } 2 & \cdots & \text{cell } n \end{array} & \\ \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \ldots \\ x_{q,1} & x_{q,2} & \cdots & x_{q,n} \end{bmatrix} & \begin{array}{c} \text{gene } 1 \\ \text{gene } 2 \\ \vdots \\ \text{gene } q \end{array} \end{array}$$

$x_{i,j}$: number of mRNA molecules mapped to gene i in cell j.

scRNA-seq data



```
##                 A01 B01 C01 D01 E01 F01 G01 H01 A02 B02
## RNA_SPIKE_MC01    0   0   0   0   0   0   0   0   0   0
## RNA_SPIKE_MC02    0   7   2   8   4   1   3   0   0   1
## RNA_SPIKE_MC04    0   0   0   0   0   0   0   0   0   0
## RNA_SPIKE_MC07    0   0   0   0   0   0   0   0   0   0
## RNA_SPIKE_MC08    0   0   0   0   0   0   0   0   0   0
## RNA_SPIKE_MC09    0   0   0   0   0   0   0   0   0   0        ...
## RNA_SPIKE_MC10    0   0   0   0   0   0   0   0   0   0
## RNA_SPIKE_MC14    4   2   3   5   1  10   2   1   3   1
## RNA_SPIKE_MC19    6   2   5   2   4   0   1   0   4   0
## RNA_SPIKE_MC20    6   4   1   1   2   0   0   3   0   1
```

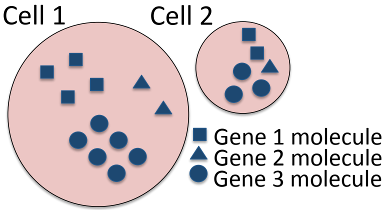
⋮

Statistical challenges in the analysis of scRNA-seq data

Normalisation

Cell-specific measurements can vary due to differences in

- total cellular mRNA content,
- sequencing depth and other amplification biases,
- capture efficiency.



Normalisation

scRNA-seq data is typically **pre-normalised** using the same strategies as for bulk RNA-seq datasets

$\Rightarrow$ to adjust the expression counts using $\tilde{x}_{ij} = x_{ij}/\hat{s}_j$ with e.g.

- Reads Per Million (RPM) $\hat{s}_j = (\sum_{i=1}^{q} x_{i,j})/1000000.$

Normalisation

scRNA-seq data is typically **pre-normalised** using the same strategies as for bulk RNA-seq datasets

$\Rightarrow$ to adjust the expression counts using $\tilde{x}_{ij} = x_{ij}/\hat{s}_j$ with e.g.

- Reads Per Million (RPM) $\hat{s}_j = (\sum_{i=1}^{q} x_{i,j})/1000000$.
- DESeq factors[1]

$$\hat{s}_j = \text{median}_{i=1,\ldots,q} \left\{ \frac{x_{ij}}{\left(\prod_{j=1}^{n} x_{ij} \right)^{1/n}} \right\}$$

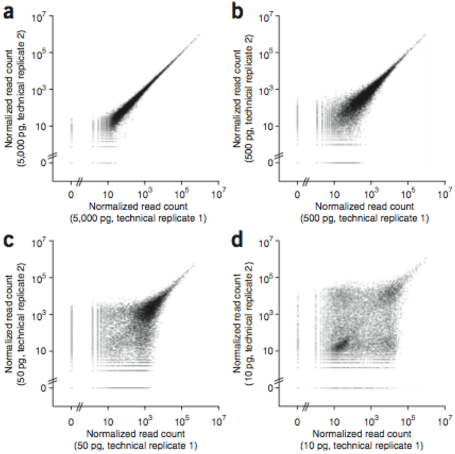
[1] Anders and Huber, Genome Biology (2010)

Normalisation

Although these strategies perform well for bulk experiments, they can lead to unstable results for scRNA-seq datasets

Technical noise

Sequencing small quantities of RNA leads to strong levels of **technical variability**



Brennecke et al., Nature Methods (2013)

Using spike-in genes to quantify technical variability

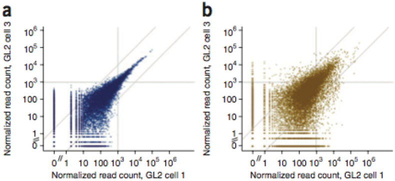
To quantify the amount of technical (non-biological) variability, non biological spike-in genes can be used[1]

⇒ e.g. the set of 92 extrinsic molecules derived by the External RNA Controls Consortium (ERCC)[2]

- are present at the same level in each cell
- spike-in empirical measurements can be compared to their known values: use as a ‘gold standard’

[1] Brennecke et al, Nat Methods (2013) [2] Jiang et al, Genome Research (2011)

Using spike-in genes to quantify technical variability



a: spike-in genes; **b**: plant genes

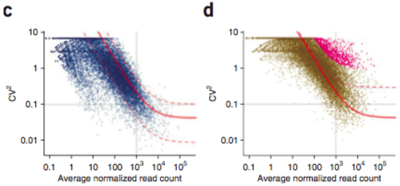
Source: Brennecke et al (2013)

- In (a), spike-in genes are compared between 2 cells: high level of technical noise (specially for genes with low read count)
- In (b), intrinsic genes are compared between 2 cells: use spikes to tease out biological variability from technical one

Using spike-in genes to quantify technical variability

Brennecke et al (2013) suggested

- to use spike-in genes to estimate relationship between technical variability and read count
- to 'plug-in' this fit to identify true cell-to-cell variability



a: spike-in genes; **b**: plant genes

Figure taken from Brennecke et al (2013)

- In (a), technical noise fit on spike-in genes (CV^2 versus means read counts)
- In (b), technical noise fit superimposed on biological genes to highlight significantly variable biological genes

Using spike-in genes to quantify technical variability

This 2-step approach ignores uncertainty in technical noise fit ⇒ development of a joint model of spike-ins and biological genes

Quality control: removing poor quality cells

Lastly (but not least!) it is important to assess how well RNA was captured and amplified from each cell[1]

⇒ e.g. some cells may contain degraded RNA (due to stress)

Some important indicators are:

- The fraction of mapped reads
- The fraction of reads mapped to the spikes
- The fraction of reads to mitochondrial genes

[1] Stegle et al, Nat Reviews (2015)

Quality control: removing poor quality cells

We might also use other experimental information (e.g. microscopy to detect multiple cells in a well)

WARNING: Be careful about removing biologically relevant cells

BASiCS: Bayesian Analysis of Single-Cell Sequencing data

BASiCS: Bayesian Analysis of Single-Cell Sequencing data

BASiCS is an integrated Bayesian hierarchical model where

- cell-specific normalising constants are treated as model parameters,

  as opposed to former pre-normalisation strategies

- unexplained technical variability is calibrated using spike-in genes,

  combining information from endogenous genes in a single step

- highly/lowly variable genes are identified via an intuitive approach

  decomposing total variability into technical and biological components

BASiCS: Bayesian Analysis of Single-Cell Sequencing data

BASiCS is an integrated Bayesian hierarchical model where

- cell-specific normalising constants are treated as model parameters,

  as opposed to former pre-normalisation strategies

- unexplained technical variability is calibrated using spike-in genes,

  combining information from endogenous genes in a single step

- highly/lowly variable genes are identified via an intuitive approach

  decomposing total variability into technical and biological components

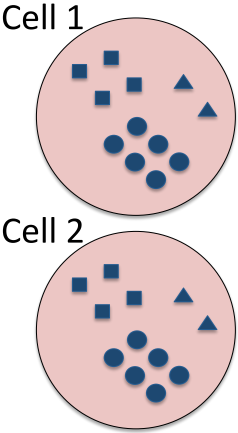
Integrative method rather than former 3-stage approaches

BASiCS: Bayesian Analysis of Single-Cell Sequencing data

Modelling expression counts of spike-in genes

If cells are identical and there is no technical variability (e.g. seq. depth, capture efficiency, etc):

$$X_{i,j}|\mu_i \overset{iid}{\sim} \text{Poisson}(\mu_i)$$



* BLUE quantities denote known parameters

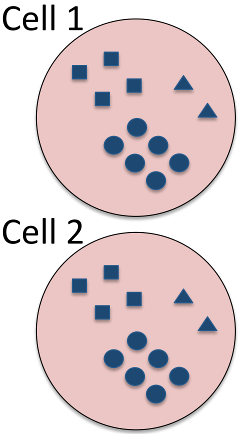
BASiCS: Bayesian Analysis of Single-Cell Sequencing data

Modelling expression counts of spike-in genes

If cells are identical and there is no technical variability (e.g. seq. depth, capture efficiency, etc):

$$X_{i,j} | \mu_i \overset{iid}{\sim} \text{Poisson}(\mu_i)$$

$$\Rightarrow \text{E}\left(X_{i,j} | \mu_i\right) = \mu_i, \text{Var}\left(X_{i,j} | \mu_i\right) = \mu_i.$$



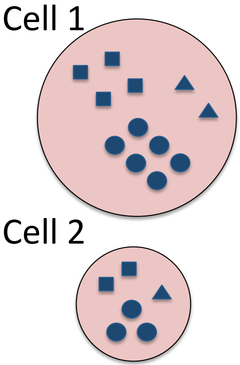
* BLUE quantities denote known parameters

BASiCS: Bayesian Analysis of Single-Cell Sequencing data

Modelling expression counts of spike-in genes

Differences in scale (e.g. seq. depth, capture efficiency, etc) can be captured by cell-specific normalising terms

$$X_{i,j} | \mu_i, s_j \overset{iid}{\sim} \text{Poisson}(s_j \mu_i)$$



* BLUE quantities denote known parameters
* RED quantities denote unknown parameters

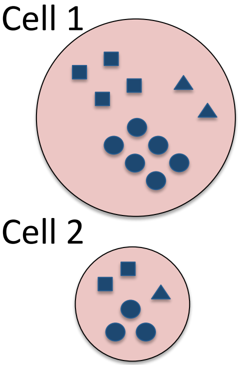
BASiCS: Bayesian Analysis of Single-Cell Sequencing data

Modelling expression counts of spike-in genes

Differences in scale (e.g. seq. depth, capture efficiency, etc) can be captured by cell-specific normalising terms

$$X_{i,j}|\mu_i, s_j \overset{iid}{\sim} \text{Poisson}(s_j \mu_i)$$

$$\Rightarrow \mathrm{E}\left(X_{i,j}|\mu_i, s_j\right) = s_j \mu_i, \mathrm{Var}\left(X_{i,j}|\mu_i, s_j\right) = s_j \mu_i.$$



* BLUE quantities denote known parameters

* RED quantities denote unknown parameters

BASiCS: Bayesian Analysis of Single-Cell Sequencing data

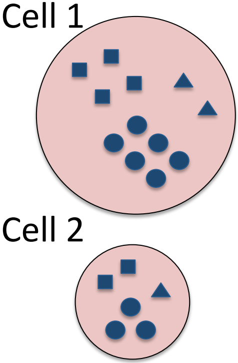
Modelling expression counts of spike-in genes

Unexplained technical variability is incorporated through random effects in a hierarchical structure

$$X_{i,j} | \mu_i, \nu_j \overset{iid}{\sim} \text{Poisson}(\nu_j \mu_i)$$

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

$$\text{i.e. } E(\nu_j | s_j, \theta) = s_j, \text{Var}(\nu_j | s_j, \theta) = s_j^2 \theta.$$



* BLUE quantities denote known parameters
* RED quantities denote unknown parameters
* GREEN quantities denote latent intermediate parameters

BASiCS: Bayesian Analysis of Single-Cell Sequencing data

Modelling expression counts of spike-in genes

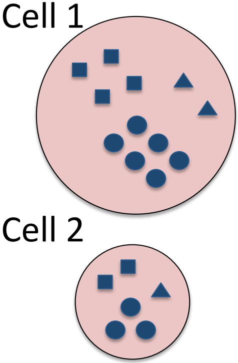
Unexplained technical variability is incorporated through random effects in a hierarchical structure

$$X_{i,j}|\mu_i,\nu_j \overset{iid}{\sim} \text{Poisson}(\nu_j\mu_i)$$

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

$$\text{i.e. } \text{E}(\nu_j|s_j,\theta)=s_j, \text{Var}(\nu_j|s_j,\theta)=s_j^2\theta.$$

$$\Rightarrow \text{E}(X_{ij}|\mu_i,s_j,\theta)=s_j\mu_i, \text{Var}(X_{ij}|\mu_i,s_j,\theta)=s_j\mu_i+\theta\left(s_j\mu_i\right)^2.$$



* BLUE quantities denote known parameters

* RED quantities denote unknown parameters

* GREEN quantities denote latent intermediate parameters

BASiCS: Bayesian Analysis of Single-Cell Sequencing data

Modelling expression counts of biological genes

$$X_{i,j} | \quad , \mu_i, \nu_j, \quad \overset{ind}{\sim} \text{Poisson}(\quad \nu_j \mu_i \quad),$$

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

* RED quantities denote unknown parameters

* GREEN quantities denote latent intermediate parameters

BASiCS: Bayesian Analysis of Single-Cell Sequencing data

Modelling expression counts of biological genes

$$X_{i,j}|\phi_j,\mu_i,\nu_j, \quad \overset{ind}{\sim} \text{Poisson}(\phi_j\nu_j\mu_i \quad),$$

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

* RED quantities denote unknown parameters

* GREEN quantities denote latent intermediate parameters

BASiCS: Bayesian Analysis of Single-Cell Sequencing data

Modelling expression counts of biological genes

$$X_{i,j}|\phi_j, \mu_i, \nu_j, \rho_{i,j} \overset{ind}{\sim} \text{Poisson}(\phi_j \nu_j \mu_i \rho_{i,j}),$$

$$\nu_j | s_j, \theta \overset{iid}{\sim} \text{Gamma}(\theta^{-1}, (s_j\theta)^{-1}), \quad \rho_{i,j}|\delta_i \overset{ind}{\sim} \text{Gamma}(\delta_i^{-1}, \delta_i^{-1})$$

Here, the ν_j's are **shared** with the technical model component and the ρ_{ij}'s are such that

$$\text{E}(\rho_{i,j}|\delta_i) = 1 \text{ and } \text{Var}(\rho_{i,j}|\delta_i) = \delta_i.$$

* RED quantities denote unknown parameters

* GREEN quantities denote latent intermediate parameters

BASiCS: Bayesian Analysis of Single-Cell Sequencing data

Modelling expression counts of biological genes

$$X_{i,j}|\phi_j,\mu_i,\nu_j,\rho_{i,j} \overset{ind}{\sim} \text{Poisson}(\phi_j\nu_j\mu_i\rho_{i,j}),$$

$$\nu_j|s_j,\theta \overset{iid}{\sim} \text{Gamma}(\theta^{-1},(s_j\theta)^{-1}), \quad \rho_{i,j}|\delta_i \overset{ind}{\sim} \text{Gamma}(\delta_i^{-1},\delta_i^{-1})$$

Here, the ν_j's are **shared** with the technical model component and the ρ_{ij}'s are such that

$$\text{E}(\rho_{i,j}|\delta_i) = 1 \text{ and } \text{Var}(\rho_{i,j}|\delta_i) = \delta_i.$$

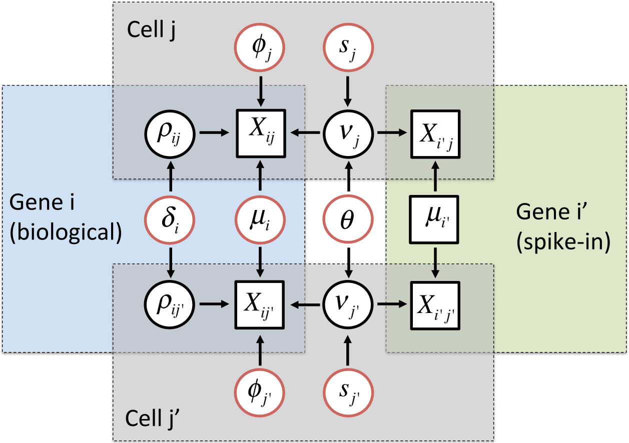
$$\Rightarrow \text{E}(X_{i,j}|\mu_i,\delta_i,s_j,\phi_j,\theta) = \phi_j s_j\mu_i,$$

$$\text{Var}(X_{ij}|\mu_i,\delta_i,s_j,\phi_j,\theta) = \phi_j s_j\mu_i + \theta(\phi_j s_j\mu_i)^2 + \delta_i(\theta+1)(\phi_j s_j\mu_i)^2$$

* RED quantities denote unknown parameters

* GREEN quantities denote latent intermediate parameters

BASiCS: Bayesian Analysis of Single-Cell Sequencing data



BASiCS: Identifiability

Definition (Identifiability)

A model for X is identifiable if and only if different parameter values lead to different probability distributions for X.

For example, if

$$X \sim \mathsf{N}(\alpha + \beta, \sigma^2),$$

α and β are not identifiable.

In fact, the distribution of X is unchanged if α and β are replaced by $\alpha^* = \alpha - \gamma$ and $\beta^* = \beta + \gamma$, respectively (for an arbitrary γ).

BASiCS: Identifiability

Using the spike-in genes, where $\mu_{q_0+1}, \ldots, \mu_q$ are known

$\Rightarrow$ We can identify s_j's and θ.

Recall:

$$\mathrm{E}(X_{i,j}|\mu_i, s_j, \theta) = s_j\mu_i, \quad \mathrm{Var}(X_{ij}|\mu_i, s_j, \theta) = s_j\mu_i + \theta\,(s_j\mu_i)^2$$

BASiCS: Identifiability

Using the biological genes, where $\mu_1, \ldots, \mu_{q_0}$ are unknown

$\Rightarrow$ We can identify δ_i's

$\Rightarrow$ But, we can't separately identify μ_i's and ϕ_j's

Recall:

$$\mathrm{E}(X_{i,j}|\mu_i, \delta_i, s_j, \phi_j, \theta) = \phi_j s_j \mu_i,$$

$$\mathrm{Var}(X_{ij}|\mu_i, \delta_i, s_j, \phi_j, \theta) = \phi_j s_j \mu_i + \theta(\phi_j s_j \mu_i)^2 + \delta_i(\theta + 1)(\phi_j s_j \mu_i)^2$$

BASiCS: Identifiability

Using the biological genes, where $\mu_1, \ldots, \mu_{q_0}$ are unknown

$\Rightarrow$ We can identify δ_i's

$\Rightarrow$ But, we can't separately identify μ_i's and ϕ_j's

Recall:

$$\mathrm{E}(X_{i,j}|\mu_i, \delta_i, s_j, \phi_j, \theta) = \phi_j s_j \mu_i,$$

$$\mathrm{Var}(X_{ij}|\mu_i, \delta_i, s_j, \phi_j, \theta) = \phi_j s_j \mu_i + \theta(\phi_j s_j \mu_i)^2 + \delta_i(\theta+1)(\phi_j s_j \mu_i)^2$$

Identifiability restriction: $n^{-1}\sum_{j=1}^{n} \phi_j = \phi_0$, for some **known** ϕ_0.

We use $\phi_0 = 1$.

BASiCS: Variance decomposition

After integrating out all random effects (intermediate parameters), our model induces:

$$\begin{aligned}
\mathrm{E}(X_{i,j}|\phi_j, s_j, \mu_i, \theta, \delta_i) &= \phi_j s_j \mu_i, \text{ and} \\
\mathrm{Var}(X_{i,j}|\phi_j, s_j, \mu_i, \theta, \delta_i) &= \underbrace{\phi_j s_j \mu_i}_{\text{Baseline}} + \underbrace{\theta(\phi_j s_j \mu_i)^2}_{\text{Technical}} + \underbrace{\delta_i(\theta + 1)(\phi_j s_j \mu_i)^2}_{\text{Biological heterogeneity}}
\end{aligned}$$

BASiCS: Variance decomposition

After integrating out all random effects (intermediate parameters), our model induces:

$$\begin{aligned} \mathrm{E}(X_{i,j}|\phi_j, s_j, \mu_i, \theta, \delta_i) &= \phi_j s_j \mu_i, \text{ and} \\ \mathrm{Var}(X_{i,j}|\phi_j, s_j, \mu_i, \theta, \delta_i) &= \underbrace{\phi_j s_j \mu_i}_{\text{Baseline}} + \underbrace{\theta(\phi_j s_j \mu_i)^2}_{\text{Technical}} + \underbrace{\delta_i(\theta+1)(\phi_j s_j \mu_i)^2}_{\text{Biological heterogeneity}} \end{aligned}$$

Using this variance decomposition we can

- Quantify the strength of technical noise (overall and per gene)
- Generate a ranking of the genes based on biological cell-to-cell heterogeneity

BASiCS: Highly and lowly variable genes

Highly Variable Genes (HVG)

- Key drivers of cell-to-cell heterogeneity
- Potential markers of novel cell sub-populations

BASiCS: Highly and lowly variable genes

Highly Variable Genes (HVG)

- Key drivers of cell-to-cell heterogeneity
- Potential markers of novel cell sub-populations

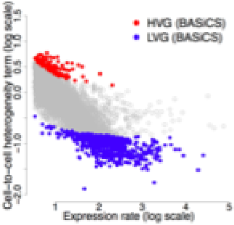
Lowly Variable Genes (LVG)

- Related to core processes of the cell
- Can help to reduce dimensionality in downstream analysis

BASiCS: Highly and lowly variable genes

Enrichment of genes related to cell differentiation →

Enrichment of genes related to translation →



BASiCS: Detecting highly variable genes

We identify HVG using tail posterior probabilities associated to a HIGH biological cell-to-cell heterogeneity component

BASiCS: Detecting highly variable genes

We identify HVG using tail posterior probabilities associated to a HIGH biological cell-to-cell heterogeneity component

For a given variance threshold $\gamma_{_H}$, and evidence threshold $\alpha_{_H}$, BASiCS labels a gene as HVG if:

$$\pi_i^H(\gamma_{_H}) = \mathrm{P}\left(\sigma_i > \gamma_{_H} \middle|\ \{\text{Data}\}\right) > \alpha_{_H}$$

$\sigma_i \Rightarrow$ proportion of total variability explained by cell-to-cell biological heterogeneity (in a typical cell)

$$\sigma_i \equiv \frac{\delta_i(\theta+1)}{[(\phi s)^* \mu_i]^{-1} + \theta + \delta_i(\theta+1)}, \quad \text{where} \quad (\phi s)^* = \underset{j \in \{1,\ldots,n\}}{\mathrm{median}} \{\phi_j s_j\},$$

BASiCS: Detecting lowly variable genes

Similarly, we identify LVG using tail posterior probabilities associated to a LOW biological cell-to-cell heterogeneity component

BASiCS: Detecting lowly variable genes

Similarly, we identify LVG using tail posterior probabilities associated to a LOW biological cell-to-cell heterogeneity component

For a given variance threshold γ_{L}, and evidence threshold α_{L}, we classify as LVG those genes for which:

$$\pi_i^H(\gamma_L) = \mathrm{P}\left(\sigma_i < \gamma_L \middle| \; \{\text{Data}\}\right) > \alpha_L$$

BASiCS: Control of error rates for HVG and LVG detection

The variance thresholds γ_H and γ_L are biologically meaningful quantities and can be fixed prior to the analysis

BASiCS: Control of error rates for HVG and LVG detection

The variance thresholds $\gamma_{_H}$ and $\gamma_{_L}$ are biologically meaningful quantities and can be fixed prior to the analysis

For fixed $\gamma_{_H}$ and $\gamma_{_L}$, evidence thresholds $\alpha_{_H}$ and $\alpha_{_L}$ can be chosen by controlling the trade-off between

- **Expected False Discovery Rate (EFDR)**

$$\text{EFDR}_\alpha = \frac{\sum_{i=1}^{q_0} (1 - \pi_i(\gamma)) I(\pi_i(\gamma) > \alpha)}{\sum_{i=1}^{q_0} I(\pi_i(\gamma) > \alpha)}$$

- **Expected False Negative Rate (EFNR)**

$$\text{EFNR}_\alpha = \frac{\sum_{i=1}^{q_0} \pi_i(\gamma) I(\pi_i(\gamma) \leq \alpha)}{\sum_{i=1}^{q_0} I(\pi_i(\gamma) \leq \alpha)}$$

BASiCS: Posterior inference

Bayesian Inference

Definition (Bayes Theorem)

$$\pi(\gamma|X) = \frac{f(X|\gamma)\pi(\gamma)}{\int f(X|\gamma)\pi(\gamma)\, d\gamma}$$

Bayesian Inference

Definition (Bayes Theorem)

$$\pi(\gamma|X) = \frac{f(X|\gamma)\pi(\gamma)}{\int f(X|\gamma)\pi(\gamma)\, d\gamma}$$

- $f(X|\gamma)$ is the likelihood function of X for a given value of γ

Bayesian Inference

Definition (Bayes Theorem)

$$\pi(\gamma|X) = \frac{f(X|\gamma)\pi(\gamma)}{\int f(X|\gamma)\pi(\gamma)\, d\gamma}$$

- $f(X|\gamma)$ is the likelihood function of X for a given value of γ
- $\pi(\gamma)$ is the prior density assigned to γ

Bayesian Inference

Definition (Bayes Theorem)

$$\pi(\gamma|X) = \frac{f(X|\gamma)\pi(\gamma)}{\int f(X|\gamma)\pi(\gamma)\,d\gamma}$$

- $f(X|\gamma)$ is the likelihood function of X for a given value of γ
- $\pi(\gamma)$ is the prior density assigned to γ
- $\pi(\gamma|X)$ is the posterior density of γ after observing X

BASiCS: The prior

The Bayesian model is completed using the following priors:

- $\mu_i \sim \text{log-Normal}(0, s_\mu^2)$ for $i = 1, \ldots, q_0$,
- $n^{-1}(\phi_1, \ldots, \phi_n)' \sim \text{Dirichlet}(p_1, \ldots, p_n)$
- $s_j \overset{iid}{\sim} \text{Gamma}(a_s, b_s)$ for $j = 1, \ldots, n$,
- $\theta \sim \text{Gamma}(a_\theta, b_\theta)$
- $\delta_i \overset{iid}{\sim} \text{Gamma}(a_\delta, b_\delta)$ for $i = 1, \ldots, q_0$,

Results are robust to changes on hyper-parameter values

BASiCS: Posterior inference

BASiCS involves a large number of parameters and exact posterior inference not possible

Instead, we use Markov Chain Monte Carlo (MCMC) methods to generate samples from the posterior distribution

BASiCS: Posterior inference

The sampler is based on the model

$$X_{ij}|\phi_j, s_j, \mu_i, \nu_j, \theta, \delta_i \overset{ind}{\sim} \begin{cases} \text{NB}\left(\delta_i^{-1}, \frac{\phi_j \nu_j \mu_i}{\phi_j \nu_j \mu_i + \delta_i^{-1}}\right), & i = 1, \ldots, q_0; \\ \text{Poisson}(\nu_j \mu_i), & i = q_0 + 1, \ldots, q. \end{cases}$$

for which the $\rho_{i,j}$'s are integrated out.

We use an Adaptive Metropolis Hastings within Gibbs algorithm

BASiCS: Posterior inference

Definition (Gibbs Sampler[1])

Let $\gamma = (\gamma_1, \ldots, \gamma_P)'$ be a P-dimensional vector of parameters. Given an initial guess $\gamma^{(0)} = (\gamma_1^{(0)}, \ldots, \gamma_P^{(0)})'$, at each iteration m

$$\begin{array}{lll} \text{sample} & \gamma_1^{(m+1)} & \text{from } \pi(\gamma_1 | \gamma_2^{(m)}, \ldots, \gamma_P^{(m)}, X), \\ \text{sample} & \gamma_2^{(m+1)} & \text{from } \pi(\gamma_2 | \gamma_1^{(m+1)}, \gamma_3^{(m)}, \ldots, \gamma_P^{(m)}, X), \\ & \vdots & \\ \text{sample} & \gamma_P^{(m+1)} & \text{from } \pi(\gamma_P | \gamma_1^{(m+1)}, \ldots, \gamma_{P-1}^{(m+1)}, X). \end{array}$$

For large m, the distribution of $\gamma^{(m)}$ converges to $\pi(\gamma | X)$

These distributions are referred to as **full conditionals**

[1] Geman and Geman, IEEE Transactions on Pattern Analysis and Machine Intelligence (1984)

BASiCS: Posterior inference

In our case, the full conditionals of parameters of the "same type" factorise due to conditional independences.

Therefore, computational complexity is simplified
$\Rightarrow$ e.g. simultaneous updates for $\mu_1, \ldots, \mu_{q_0}$

BASiCS: Posterior inference

In our case, the full conditionals of parameters of the “same type” factorise due to conditional independences.

Therefore, computational complexity is simplified
⇒ e.g. simultaneous updates for $\mu_1, \ldots, \mu_{q_0}$

However, most of the required full conditionals do not have a known form
⇒ direct samplers are not available
⇒ we need to implement specialised samplers

BASiCS: Posterior inference

Definition (Metropolis-Hastings[1,2])

Given a starting value $\gamma^{(0)}$, at each iteration m

1. Sample $\upsilon \sim \text{Unif}(0,1)$ and $\gamma^* \sim q(\gamma^*|\gamma^{(m)})$.
2. Define

$$a(\gamma^{(m)}, \gamma^*|X) = \min\left\{1, \frac{\pi(\gamma^*|X)}{\pi(\gamma^{(m)}|X)} \frac{q(\gamma^{(m)}|\gamma^*)}{q(\gamma^*|\gamma^{(m)})}\right\}.$$

3. If $\upsilon \leq a(\gamma^{(m)}, \gamma^*|X)$, return γ^*. Otherwise, return $\gamma^{(m)}$.

These steps generate samples from $\pi(\gamma|X)$.

[1] Metropolis et al., The Journal of Chemical Physics (1953)

[2] Hastings, Biometrika (1970)

BASiCS: Posterior inference

A common choice for $q(\gamma^*|\gamma^{(m)})$ is a Normal$(\gamma^{(m)}, \omega^2)$ distribution

Where the value of ω^2 is tuned to control the acceptance rate (i.e. the proportion of times that draws are accepted)

BASiCS: Posterior inference

A common choice for $q(\gamma^*|\gamma^{(m)})$ is a Normal$(\gamma^{(m)}, \omega^2)$ distribution

Where the value of ω^2 is tuned to control the acceptance rate (i.e. the proportion of times that draws are accepted)

A solution is to use an Adaptive Metropolis-Hastings[1] algorithm

Every 50 iterations

- Calculate the current acceptance rate
- If it is too high, increase ω^2
- If it is too small, decrease ω^2

Diminishing increments $\Rightarrow$ ω^2 will stabilise

[1] Roberts and Rosenthal, Journal of Computational and Graphical Statistics (2003)

Final remarks

Final remarks

- scRNA-seq can reveal novel insights about transcriptional regulation
- However, analysing scRNA-seq is not a trivial task due to
  - Quality control
  - Normalisation
  - Technical variability
- Methods used for bulk RNA-seq datasets cannot be directly applied
- Our approach borrows information from intrinsic genes and technical spike-in genes, simultaneously $\Rightarrow$ avoid stepwise procedures

Extensions

BASiCS will soon incorporate 2 of the most widely applied downstream analyses

- Differential expression
- Clustering

Extensions

BASiCS will soon incorporate 2 of the most widely applied downstream analyses

- Differential expression
- Clustering

Another extension relates to scalability

⇒ New technologies allow sequencing of huge numbers of cells e.g. Drop-seq[1] $\sim$ 40000 cells (and no spikes!)

[1] Macosko et al., Cell (2015)

Before the lab session ...

After lunch we will have a practical session

Before we start, please visit:

https://github.com/catavallejos/TutorialBASiCS

Acknowledgments

EMBL-EBI CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

John Marioni's lab

Aaron Lun
Antonio Scialdone
Christopher Laumer
Jong Kyoung Kim
Konrad Rudolph
Liora Vilmovsky
Luis Saraiva
Nils Eling (here!)
Tim Hu

MRC | Biostatistics Unit

Sylvia Richardson's group

Daniel Ahfock
Daniel Greene
Gwenael Leday
Harry Gray
James Peters
Paul Newcombe
Paul Kirk