Statistical challenges in the analysis of single-cell transcriptomics data

Catalina Vallejos
joint work with Sylvia Richardson and John Marioni
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
Understanding heterogeneity at the single-cell level

Most transcriptomic studies have focused on examining expression in large populations of cells\textsuperscript{1,2}.

Some biological processes, however, require the study of variation in gene expression at the single-cell level\textsuperscript{3,4}.

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

\textsuperscript{1} Marioni et al., Genome Res (2008)
\textsuperscript{3} Hayashi et al., Science (2007)
\textsuperscript{4} Diez-Roux et al., PLoS Biol (2011)
scRNA-seq workflow

Dissociation and isolation → RNA extraction and cDNA synthesis → Amplification → Sequencing and mapping

Gene 1

<table>
<thead>
<tr>
<th>Cell 1</th>
<th>Cell 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
</tr>
</tbody>
</table>

Gene 2

Expression counts

Gene 1

Gene 2
The power of scRNA-seq

Already this has led to identification of novel:

- Neuronal populations\(^1\)
- Immune cell populations\(^2\)
- Sub-populations of tumour cells\(^3\)

---

Notation
scRNA-seq data

scRNA-seq data can be represented as

\[
\begin{bmatrix}
X_{1,1} & X_{1,2} & \cdots & X_{1,n} \\
X_{2,1} & X_{2,2} & \cdots & X_{2,n} \\
\vdots & \vdots & \ddots & \vdots \\
X_{q,1} & X_{q,2} & \cdots & X_{q,n}
\end{bmatrix}
\]

gene 1

gene 2

\[x_{i,j}: \text{number of mRNA molecules mapped to gene } i \text{ in cell } j.\]
### scRNA-seq data

<table>
<thead>
<tr>
<th></th>
<th>A01</th>
<th>B01</th>
<th>C01</th>
<th>D01</th>
<th>E01</th>
<th>F01</th>
<th>G01</th>
<th>H01</th>
<th>A02</th>
<th>B02</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA_SPIKE_MC01</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RNA_SPIKE_MC02</td>
<td>0</td>
<td>7</td>
<td>2</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>RNA_SPIKE_MC04</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RNA_SPIKE_MC07</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RNA_SPIKE_MC08</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RNA_SPIKE_MC09</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RNA_SPIKE_MC10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RNA_SPIKE_MC14</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>RNA_SPIKE_MC19</td>
<td>6</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>RNA_SPIKE_MC20</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

...
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

⇒ 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

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

\(\Rightarrow\) 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’

Using spike-in genes to quantify technical variability

- 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

a: spike-in genes; b: plant genes

Source: Brennecke et al (2013)
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

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

**a:** spike-in genes; **b:** plant genes
Using spike-in genes to quantify technical variability

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

---

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
**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
Modelling expression counts of spike-in genes

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

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

\[ \Rightarrow E(X_{i,j} | \mu_i) = \mu_i, \text{Var}(X_{i,j} | \mu_i) = \mu_i. \]

* **BLUE** quantities denote known parameters
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
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

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

\[ \Rightarrow E(X_{i,j} | \mu_i, s_j) = s_j \mu_i, \text{Var}(X_{i,j} | \mu_i, s_j) = s_j \mu_i. \]

* **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

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

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

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

\[
X_{i,j} | \mu_i, \nu_j \sim \text{Poisson}(\nu_j \mu_i)
\]

\[
\nu_j | s_j, \theta \sim \text{Gamma}(\theta^{-1}, (s_j \theta)^{-1})
\]

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

\[
\Rightarrow E(X_{ij} | \mu_i, s_j, \theta) = s_j \mu_i, \ Var(X_{ij} | \mu_i, s_j, \theta) = s_j \mu_i + \theta (s_j \mu_i)^2.
\]

* **BLUE** quantities denote known parameters
* **RED** quantities denote unknown parameters
* **GREEN** quantities denote latent intermediate parameters
Modelling expression counts of biological genes

\[ X_{i,j} \mid \mu_i, \nu_j, \overset{\text{ind}}{\sim} \text{Poisson}(\nu_j \mu_i), \]
\[ \nu_j \mid s_j, \theta \overset{\text{iid}}{\sim} \text{Gamma}(\theta^{-1}, (s_j \theta)^{-1}), \]

* RED quantities denote unknown parameters
* GREEN quantities denote latent intermediate parameters
Modelling expression counts of biological genes

\[ X_{i,j} \mid \phi_j, \mu_i, \nu_j, \overset{ind}{\sim} \text{Poisson}(\phi_j \nu_j \mu_i), \]
\[ \nu_j \mid s_j, \theta \overset{iid}{\sim} \text{Gamma}(\theta^{-1}, (s_j \theta)^{-1}), \]

* **RED** quantities denote unknown parameters
* **GREEN** quantities denote latent intermediate parameters
Modelling expression counts of biological genes

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

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

Here, the \( \nu_j \)'s are **shared** with the technical model component and the \( \rho_{ij} \)'s are such that

\[
\mathbb{E}(\rho_{i,j} | \delta_i) = 1 \quad \text{and} \quad \text{Var}(\rho_{i,j} | \delta_i) = \delta_i.
\]

* **RED** quantities denote unknown parameters
* **GREEN** quantities denote latent intermediate parameters
Modelling expression counts of biological genes

\[ X_{i,j} | \phi_j, \mu_i, \nu_j, \rho_{i,j} \overset{\text{ind}}{\sim} \text{Poisson}(\phi_j \nu_j \mu_i \rho_{i,j}), \]
\[ \nu_j | s_j, \theta \overset{\text{iid}}{\sim} \text{Gamma}(\theta^{-1}, (s_j \theta)^{-1}), \quad \rho_{i,j} | \delta_i \overset{\text{ind}}{\sim} \text{Gamma}(\delta_i^{-1}, \delta_i^{-1}) \]

Here, the \( \nu_j \)'s are \textbf{shared} with the technical model component and the \( \rho_{ij} \)'s are such that

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

\[ \Rightarrow E(X_{i,j} | \mu_i, \delta_i, s_j, \phi_j, \theta) = \phi_j s_j \mu_i, \]
\[ \text{Var}(X_{ij} | \mu_i, \delta_i, s_j, \phi_j, \theta) = \phi_j s_j \mu_i + \theta(\phi_j s_j \mu_i)^2 + \delta_i(\theta + 1)(\phi_j s_j \mu_i)^2 \]

* \textbf{RED} quantities denote unknown parameters
* \textbf{GREEN} quantities denote latent intermediate parameters
BASiCS: Bayesian Analysis of Single-Cell Sequencing data

Cell $j$

Gene $i$ (biological)

Gene $i'$ (spike-in)

Cell $j'$

$\phi_j$

$s_j$

$\rho_{ij}$

$X_{ij}$

$\delta_i$

$\mu_i$

$X_{ij'}$

$\phi_{j'}$

$s_{j'}$

$\rho_{ij'}$

$X_{ij'}$

$\theta$

$\mu_{i'}$

$X_{i'j'}$

$\mu_{i'}$

$X_{i'j'}$
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

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

$\alpha$ and $\beta$ are not identifiable.

In fact, the distribution of $X$ is unchanged if $\alpha$ and $\beta$ are replaced by $\alpha^* = \alpha - \gamma$ and $\beta^* = \beta + \gamma$, respectively (for an arbitrary $\gamma$).
Using the spike-in genes, where $\mu_{q_0+1}, \ldots, \mu_q$ are known

$\Rightarrow$ We can identify $s_j$’s and $\theta$.

Recall:

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

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

$\Rightarrow$ We can identify $\delta_i$’s

$\Rightarrow$ But, we can’t separately identify $\mu_i$’s and $\phi_j$’s

Recall:

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

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

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

$\Rightarrow$ We can identify $\delta_i$’s

$\Rightarrow$ But, we can’t separately identify $\mu_i$’s and $\phi_j$’s

Recall:

\[
E(X_{i,j}|\mu_i, \delta_i, s_j, \phi_j, \theta) = \phi_j s_j \mu_i,
\]
\[
\text{Var}(X_{ij}|\mu_i, \delta_i, s_j, \phi_j, \theta) = \phi_j s_j \mu_i + \theta (\phi_j s_j \mu_i)^2 + \delta_i (\theta + 1) (\phi_j s_j \mu_i)^2
\]

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

We use $\phi_0 = 1$. 

BASiCS: Variance decomposition

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

\[
\begin{align*}
E(X_{i,j} | \phi_j, s_j, \mu_i, \theta, \delta_i) &= \phi_j s_j \mu_i, \text{ and} \\
\text{Var}(X_{i,j} | \phi_j, s_j, \mu_i, \theta, \delta_i) &= \phi_j s_j \mu_i + \theta(\phi_j s_j \mu_i)^2 + \delta_i (\theta + 1)(\phi_j s_j \mu_i)^2 \\
& \quad \text{Baseline} \quad \text{Technical} \quad \text{Biological heterogeneity}
\end{align*}
\]
After integrating out all random effects (intermediate parameters), our model induces:

\[
E(X_{i,j} | \phi_j, s_j, \mu_i, \theta, \delta_i) = \phi_j s_j \mu_i, \quad \text{and} \\
\text{Var}(X_{i,j} | \phi_j, s_j, \mu_i, \theta, \delta_i) = \phi_j s_j \mu_i + \theta (\phi_j s_j \mu_i)^2 + \delta_i (\theta + 1)(\phi_j s_j \mu_i)^2
\]

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
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
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:

$$\pi_i^H(\gamma_H) = P(\sigma_i > \gamma_H \mid \{\text{Data}\}) > \alpha_H$$

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

$$\sigma_i = \frac{\delta_i(\theta + 1)}{[(\phi s)^* \mu_i]^{-1} + \theta + \delta_i(\theta + 1)}, \text{ where } (\phi s)^* = \text{median}_{j\in\{1,\ldots,n\}} \{\phi_j s_j\},$$
Similarly, we identify LVG using tail posterior probabilities associated to a LOW biological cell-to-cell heterogeneity component.
Similarly, we identify LVG using tail posterior probabilities associated to a low biological cell-to-cell heterogeneity component.

For a given variance threshold $\gamma_L$, and evidence threshold $\alpha_L$, we classify as LVG those genes for which:

$$\pi_i^H(\gamma_L) = P(\sigma_i < \gamma_L \mid \text{Data}) > \alpha_L$$
The variance thresholds $\gamma_H$ and $\gamma_L$ are biologically meaningful quantities and can be fixed prior to the analysis.
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 \(\gamma\)

Catalina Vallejos

MRC Biostatistics Unit - EMBL European Bioinformatics Institute
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 $\gamma$
- $\pi(\gamma)$ is the **prior** density assigned to $\gamma$
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 \( \gamma \)
- \( \pi(\gamma) \) is the **prior** density assigned to \( \gamma \)
- \( \pi(\gamma|X) \) is the **posterior** density of \( \gamma \) after observing \( X \)
The Bayesian model is completed using the following priors:

- $\mu_i \sim \text{log-Normal}(0, s^2_\mu)$ 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
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
Definition (Gibbs Sampler$^1$)

Let $\gamma = (\gamma_1, \ldots, \gamma_P)'$ be a $P$-dimensional vector of parameters. Given an initial guess $\gamma^{(0)} = (\gamma^{(0)}_1, \ldots, \gamma^{(0)}_P)'$, at each iteration $m$

\begin{align*}
&\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), \\
&\quad \vdots \\
&\text{sample } \gamma_P^{(m+1)} \text{ from } \pi(\gamma_P | \gamma_1^{(m+1)}, \ldots, \gamma_{P-1}^{(m+1)}, X).
\end{align*}

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

These distributions are referred to as full conditionals.

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 $\mu_1, \ldots, \mu_{q_0}$
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
**Definition (Metropolis-Hastings\textsuperscript{1,2})**

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

1. Sample $\nu \sim \text{Unif}(0, 1)$ 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 $\nu \leq a(\gamma^{(m)}, \gamma^*|X)$, return $\gamma^*$. Otherwise, return $\gamma^{(m)}$.

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

---

1. Metropolis et al., The Journal of Chemical Physics (1953)
A common choice for $q(\gamma^* | \gamma^{(m)})$ is a $\text{Normal}(\gamma^{(m)}, \omega^2)$ distribution.

Where the value of $\omega^2$ is tuned to control the acceptance rate (i.e. the proportion of times that draws are accepted).
A common choice for \( q(\gamma^*|\gamma^{(m)}) \) is a Normal(\(\gamma^{(m)}, \omega^2\)) distribution

Where the value of \(\omega^2\) 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 \(\omega^2\)
- If it is too small, decrease \(\omega^2\)

Diminishing increments \(\Rightarrow\) \(\omega^2\) will stabilise

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 ⇒ avoid stepwise procedures
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 $\sim 40000$ cells (and no spikes!)

1. Macosko et al., Cell (2015)
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

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

Sylvia Richardson’s group
Daniel Ahfock
Daniel Greene
Gwenael Leday
Harry Gray
James Peters
Paul Newcombe
Paul Kirk