Understanding drop-outs in single-cell UMI: two papers with different approaches

Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics
K Choi, Y Chen, DA Skelly, GA Churchill

Demystifying "drop-outs" in single-cell UMI data
TH Kim, X Zhou, M Chen

CSE 590C Fall 2020
October 19th, 2020
Ayse Dincer & Walter L. Ruzzo
Single-cell RNA sequencing (scRNA-seq)

Genotype ↔ Phenotype

A challenge in biology and medicine

Transcriptomes can be informative

- Bulk population sequencing can provide only the average expression signal for an ensemble of cells
- However, diverse cell types in our body each express a unique transcriptome

Single-cell RNA sequencing (scRNA-seq)

We need a more precise understanding of the transcriptome in individual cells

Single-cell RNA sequencing (scRNA-seq)

- Pioneered by James Eberwine et al. and Iscove et al.
- First analysis in 2009 by Tang et al.
  - characterization of cells from early developmental stages
- Many studies followed:
  - Identify rare cell populations
  - Characterize outlier cells to understand drug resistance and relapse in cancer treatment
  - Detect diverse immune cell populations
  - Understand cell lineage relationships in early development

scRNA-seq Technology

First step: single-cell isolation

Second step: generation of scRNA-seq libraries

Many techniques exist to isolate cells

example of droplet-based library generation

scRNA-seq Technology: What is UMI?

“Unique molecular identifiers (UMI) are molecular tags that are used to detect and quantify unique mRNA transcripts”

Drop-Seq workflow

Paired-end reads

Illumina, Data Science Sequencing Lecture 16
scRNA-seq: Computational pipeline

**Technical variation**
- Batch effect
- Cell-specific capture efficiency
- Amplification bias
- Dropout

**Biological variation**
- Stochastic gene expression
- Environmental niche
- Cell cycle

**Phase 1: Data acquisition**
- Alignment
- De-duplication
- Quantification

**Observed data**
- Genes
- Cells

scRNA-seq: Computational pipeline

scRNA-seq Applications

scRNA-seq Applications

a. Drug resistance clone identification

Single-cell RNA sequencing (scRNA-seq)

- Single-cell RNA sequencing is a very promising technology
- It can allow new biological insights

- Yet it also presents many technical and computation challenges
- One problem we will focus on today is drop-out or zero-inflation
What is dropout in single cell?

A gene is observed at a moderate or high expression level in one cell but is not detected in another cell.

There are many many different approaches

scDoc: correcting dropout events in RNA-seq data
Di Ran, Shanshan Zhang, Nicholas Lytal, Li

Bioinformatics, Volume 36, Issue 15, 1 August 2020

Droplet scRNA-seq is not zero-inflated
Gong et al. BMC Bioinformatics (2018) 19:220
https://doi.org/10.1186/s12859-018-2226-y

DrlImpute: imputing dropout events in single cell RNA sequencing data
Wuming Gong, Il-Youp Kwak, Pruthvi Pota, Naoko Koyano-Nakagawa and Daniel J. Garry

An accurate and robust imputation method sclImpute for single-cell RNA-seq data
Wei Vivian Li & Jingyi Jessica Li

BMC Bioinformatics

2020, Pages 4021–4029, iaaa278

January 11, 2021
Why do dropouts occur in single cell?

There are different views

<table>
<thead>
<tr>
<th>Why do we observe dropouts?</th>
<th>What should we do about them?</th>
</tr>
</thead>
<tbody>
<tr>
<td>• technical artifacts</td>
<td>• impute before learning</td>
</tr>
<tr>
<td>• statistical sampling</td>
<td>• preprocess/cluster/reduce dimensions</td>
</tr>
<tr>
<td>• cell type differences</td>
<td>• incorporate technical variates</td>
</tr>
<tr>
<td>• biological factors</td>
<td>• incorporate biological variates</td>
</tr>
<tr>
<td></td>
<td>• model zero inflation</td>
</tr>
<tr>
<td></td>
<td>• ignore zero inflation</td>
</tr>
</tbody>
</table>
Today we are going to examine 2 papers

There are two main views

**Drop-outs are technical artefacts**

**Drop-outs are related to biological signals**

---

To solve drop-outs ->
Take cell type heterogeneity and biological covariates into account

---

To detect cell type heterogeneity ->
Use drop-out rates
Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics

Paper 1
Short summary of paper 1

• They apply a **Bayesian model selection approach** to demonstrate zero inflation in multiple biologically realistic scRNA-seq datasets

• They show that the primary causes of zero inflation are **not technical but rather biological in nature**

• They recommend the **negative binomial count distribution, not zero-inflated**, as a suitable reference model for scRNA-seq analysis
Outline for paper 1

Problem: Potential reasons for zero inflation/dropout
Method: Bayesian model selection approach to identify genes with zero inflation
Results #1: scRATE can identify genes with zero inflation
Results #2: Zero-inflation of genes is highly associated with cell types
Problem: Why are there so many zeros?

1. Sequencing Depth
2. Per-gene average rate of expression

Sequencing depth explains 95% of variation in the number of zeros per cell.
Background: Statistical Models

1. Poisson (P)

2. Negative Binomial (NB)

3. Zero-inflated Poisson (ZIP)

4. Zero-inflated Negative Binomial (ZINB)
Method: Bayesian model selection to identify genes exhibiting zero inflation

What is Bayesian model selection?

• The goal is to select the model that maximizes the likelihood of the observed data
• The probability of the data given the model is computed by integrating over the unknown parameter values in that model:

\[ p(D|M) = \int_\theta p(D|\theta)p(\theta|M)d\theta \]
Method: Bayesian model selection to identify genes exhibiting zero inflation

- Is based on generalized linear models (GLMs)
- Implemented a Bayesian model selection criterion the expected log predictive density (ELPD)
  \[
  \text{ELPD} = \sum_{c=1}^{C} \log p(y_c | y_{-c}).
  \]
  denotes LOOCV value for each cell vs. all the other cells

- ELPD score is calculated for four statistical models (P, ZIP, NB, or ZINB)
- scRATE examines all the data, including non-zero counts
- Uses leave-one-out cross-validation, which provides a standard error (SE) to quantify uncertainty in the estimated ELPD scores
- Penalizes both underfitting and overfitting models, a more complex model is selected only when the ELPD is substantially better
Results #1: Model selection can identify genes exhibiting zero inflation

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Error rates and power of scRATE classification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sequencing depth</strong></td>
<td><strong>Threshold</strong></td>
</tr>
<tr>
<td>(a)</td>
<td>10k</td>
</tr>
<tr>
<td>50k</td>
<td>0.1837±0.0557</td>
</tr>
<tr>
<td>(b)</td>
<td>10k</td>
</tr>
<tr>
<td>50k</td>
<td>0.8955±0.0158</td>
</tr>
</tbody>
</table>
Results #2: Most zero-inflated genes are due to variable expression rates across cell types

After accounting for cell type, the number of zero-inflated genes drops

Genes that are no longer ZI vary across cell types

Examples: Col1a2 -> fibroblasts, Ptpn18 -> immune cells

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Selected model</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P</td>
<td>NB</td>
<td>ZIP</td>
</tr>
<tr>
<td>(a)</td>
<td></td>
<td>1111</td>
<td>2930</td>
<td>525</td>
</tr>
<tr>
<td>0 SE</td>
<td></td>
<td>2112</td>
<td>3183</td>
<td>81</td>
</tr>
<tr>
<td>1 SE</td>
<td></td>
<td>2930</td>
<td>2509</td>
<td>5</td>
</tr>
<tr>
<td>2 SE</td>
<td></td>
<td>3445</td>
<td>2035</td>
<td>1</td>
</tr>
</tbody>
</table>
Results #2: Most zero-inflated genes are due to variable expression rates across cell types

Most zero-inflated genes are due to variable expression rates across cell types.

Majority of genes were originally classified as ZI are no longer ZI after accounting for cell type.

A few of genes remain or become ZI: female-specific \textit{Xist} Y-chromosome gene \textit{Ddx3y}.

After accounting for sex as an explanatory variable, these genes are no longer ZI.
Paper 1

Their conclusions:

• High frequency of zeros does not necessarily imply technical dropout
• Instead, zero inflation is largely explained by biological factors, such as cell type and sex
• Recommend against the practice of replacing zeros in data with imputed non-zero values, could mask biological signals
• Recommend the generalized linear model with negative binomial error, and taking cell types and biological factors as explanatory variables
Paper 1

• Do you think simulation tests make sense?
  • What other simulation experiments can be carried?
  • Do you think simulated data can reflect true patterns?

• Do you prefer to see more real-data experiments and biological covariate examples?

• What are the advantages/disadvantages of this model?
  • Does it make sense that cell type is a determinant of zero-inflation?
Demystifying “drop-outs” in single-cell UMI data

Paper 2
Short summary of paper 2

• Proposed a novel framework HIPPO (Heterogeneity-Inspired Pre-Processing tOol) that leverages zero proportions to explain cellular heterogeneity and integrates feature selection with iterative clustering
• Showed that clustering should be the foremost step of the workflow
• Showed that cell-type heterogeneity can resolve drop-outs, while imputing or normalizing heterogeneous data can introduce unwanted noise
Outline for paper 2

**Problem:** Potential reasons for zero inflation/dropout

**Method:** Zero inflation test to detect cellular heterogeneity and HIPPO

**Results #1:** Zero inflation test is successful at detecting cellular heterogeneity

**Results #2:** Appropriate pre-processing introduces unwanted noise in the downstream analysis

**Results #3:** HIPPO can identify cell types
Problem: Demystifying drop-outs

1. For a homogeneous cell population, zero proportions in most genes can be modeled by the Poisson distribution (more than 95% of absolute z values are below 2)

2. For mixed cell types, zero proportions considerably deviate from expected values under the Poisson model (less than 30% of the genes have z values below 2)

Conclusion: Zero-inflation test is an effective way to find genes that contribute to cellular heterogeneity
Problem: Demystifying drop-outs

Conclusion: Zero proportions can be a metric to evaluate cellular heterogeneity and can discern cell types
Method: Zero inflation test for cellular heterogeneity

They developed a new feature selection strategy that uses detected zero proportion of a given gene as the statistic to test for cellular heterogeneity.

Framework:
- **Null hypothesis** = assumes complete cellular homogeneity = the proportion of zeros is equal to the expected zero proportion under Poisson distribution
- **Alternative hypothesis** = zero proportion is inflated, as if the count data follows mixture of Poisson distributions

\[
H_0 : \ p_g = e^{-\lambda_g}, \quad p_g = \text{true zero proportion}
\]

\[
H_A : \ p_g > e^{-\lambda_g}
\]

Advantages of the framework:
1. Only the proportion of zeros is used
2. Allows each gene to have different grouping structure across cells
3. No complicated modeling
Results #1: Zero inflation test is successful at detecting cellular heterogeneity

- PPBP was identified with a high zero proportion of 26% within CD34+ cells, indicating very high zero inflation.
- After they separated CD34+ cells into three subtypes, the test within each subtype is no longer statistically significant.

Conclusion: cellular heterogeneity can drive excessive zeros and zero proportions can be used to discern cell types.
Results #2: Inappropriate pre-processing introduces unwanted noise in the downstream analysis

A popular pre-processing step is to apply deep learning based de-noising tools (e.g. Deep Count Autoencoder (DCA)) which de-convolute the technical effects from biological effects and impute zero accounts due to drop-outs.

Conclusion: imputing the UMI data without resolving cell heterogeneity can lead to loss of important biological information.
Method: HIPPO: Heterogeneity-Inspired Pre-Processing Tool

HIPPO integrates the proposed zero inflation test into a hierarchical clustering framework

**Step 1 Feature Selection:**
- Select genes with strong indication for cellular heterogeneity (cutoff of 2 on z score)

**Step 2 Cluster:**
- With the selected features, cluster the cells into 2 groups using PCA + K-means
- Each cluster is evaluated with their intra-variability using the mean Euclidean distance from the centers of K-mean algorithm. The group with the highest intra-variability is selected and assigned for next round of clustering.

*Computationally cheap because fewer and fewer features will be left for the next round of clustering*
Results #3: HIPPO can successfully identify cell types
Results #3: HIPPO can identify cell types

Seurat and Sctransform fails to separate the memory T cells, regulatory T cells, and helper T cells, grouping them as one cluster.
Paper 2

Their conclusions:

• Cell-type heterogeneity must be tackled as the first step of analysis for more reliable downstream analysis

• They introduced computationally and mathematically simple analysis tool for feature selection with great interpretability

• This pre-processing tool can resolve cellular heterogeneity and help avoid unnecessary normalizing steps that can introduce unwanted bias and noise
• What are the advantages of this model?
  • Do you think having a simple model can be helpful?
  • What are the advantages/disadvantages of not taking non-zero counts into account?

• What can be potential limitations of predicting cell type heterogeneity from drop-out rates?
  • Do you think more datasets are required to support conclusions?

• What are the advantages/disadvantages of inferring cell types from zero-inflation?
  • How can we solve the circular dependence of cell type heterogeneity and dropout?
Summary and comparison of 2 papers

PAPER 1: Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics

PAPER 2: Demystifying “drop-outs” in single-cell UMI data

Common Points

• Drop-out rates in scRNA-Seq is determined by cell types
• Drop-out rates are not technical problems that should be eliminated but provide important biological information
• Zero-inflated distributions are not good fits for scRNA-Seq especially after taking cell type into account
Summary and comparison of 2 papers

PAPER 1: Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics

PAPER 2: Demystifying “drop-outs” in single-cell UMI data

Differences

• To solve drop-outs -> uses cell type heterogeneity and biological covariates
• The goal is to select the best distribution for each gene
• Negative binomial distribution should be used to model scRNA-Seq

• To detect cell type heterogeneity -> uses drop-out rates
• The goal is to cluster the cells using drop-out rates
• Poisson distribution should be used to model scRNA-Seq