Identifying and Preventing Artifacts in High Dimensional Data: Computational Science in Immuno-Oncology

Hongkai Ji
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Email: hji@jhu.edu
• Artifacts are common in data generated by high-throughput technologies

• Why do we care?
  • Identifying and preventing artifacts will help better discover true signals
It is not our objective to provide an exhaustive list of all possible artifacts. Instead, we will:

- Discuss common sources of artifacts
- Discuss some general principles and methods for identifying and preventing artifacts
Common sources of artifacts

Study design
- Lack of proper control or randomization

Data generation
- Bias and noise in technology
- Bias in experimental procedure

Data analysis
- Improper normalization
- Failure to control confounders
- Wrong models, assumptions or methods

Artifacts due to study design

Study design
- Lack of proper control or randomization

Data generation

Data analysis
Example: Batch effects
Example: Batch effects
Batch effects: differential expression

A differential gene

Separate by batch

Ignore batch

Lose power
Batch effects: differential expression

A non-differential gene

Separate by batch

Treatment confounded with batch

Artifact
Reanalysis of the raw data demonstrated a perfect confound between read length and cancer status (50 base pairs [bp] for both cancer cohorts, 75 bp for normal). Raw expression principal components PC1 and PC2, which separate cancer from normal samples, highly correlate to alignment metrics (Fig. 1A and B). Following in silico read-length trimming, normal samples still exhibited perfect or near-perfect separation along a number of purely technical variables: mismatch rate, intronic rate, exonic rate, ribosomal RNA (rRNA) rate, and others (Fig. 1C and D). Based on these observations, it seems that serum from individuals with cancer was processed separately from serum from individuals without cancer, creating a perfect confound between library batch, sequencing batch, and status. Since many standard RNA sequencing
How to prevent artifacts due to batches?

- Proper control and randomization

<table>
<thead>
<tr>
<th></th>
<th>T1</th>
<th>T2</th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Batch1</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Batch2</td>
<td></td>
<td>√</td>
<td>√</td>
<td></td>
</tr>
</tbody>
</table>

- For those who run experiments
  Team up with a statistician or experiment design expert before your study
  Make sure everyone is on the same page

- For those who analyze data
  Talk to your wet lab collaborators before they generate data
How to identify batch artifacts?

- Review the study design
- Exploratory plots
- Compare results from orthogonal datasets
How to correct for batch effects?

- With proper design: regress out confounders
How to correct for batch effects?

- With proper design: regress out confounders
  - ComBat (Johnson et al. Biostatistics, 8:118-127, 2007)
  - Surrogate variable analysis (Leek & Storey, PLoS Genet. 3:e161, 2007)
  - Remove unwanted variation (Gagnon-Bartsch & Speed, Biostatistics, 13:539-52, 2012)
  - A good review (Leek et al. Nat Rev Genet. 11: 733-739, 2010)

- With perfect confounding: profile new samples
  - Include samples to be compared in the same batch
  - Generate multiple batches to estimate batch variance
Remarks

- Batch effect is just one example of unwanted variation that may cause artifacts
- Other confounders may also create artifacts
- They often can be dealt with by following the same principles
Artifacts created during data generation

Study design

Data generation
- Bias and noise in technology
- Bias in experimental procedure

Data analysis
**Example 1: Doublets in single-cell RNA-seq**

DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors

**Authors**
Christopher S. McGinnis, Lyndsay M. Murrow, Zev J. Gartner

**Correspondence**
zeug.gartner@ucsf.edu

**In Brief**
scRNA-seq data interpretation is confounded by technical artifacts known as doublets—single-cell transcriptome data representing more than one cell. Moreover, scRNA-seq cellular throughput is purposefully limited to minimize doublet formation rates. By identifying cells sharing expression features with simulated doublets, DoubletFinder detects many real doublets and mitigates these two limitations.
Example 1: Doublets in single-cell RNA-seq

Cell Systems
Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data

Graphical Abstract

Authors
Nan Miles Xi, Jingyi Jessica Li

Correspondence
jili@stat.ucla.edu

In Brief
We conduct a systematic benchmark study of nine cutting-edge computational doublet-detection methods. We evaluate the methods’ detection accuracy, impacts on downstream analyses, and computational efficiency, using a comprehensive set of real and synthetic data. Although no method dominates in all aspects, the DoubletFinder and cxds methods have the best detection accuracy and computational efficiency, respectively.
Example 2: Artifacts due to tissue dissociation
Example 2: Artifacts due to tissue dissociation

Dissociation of solid tumor tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses

Ciara H. O’Flanagan1, Kieran R. Campbell1,2,3, Allen W. Zhang1,4,5, Farihia Kabber4,6,7, Jamie L. P. Lim7, Justina Biele1, Peter Eirew1, Daniel Lai1, Andrew McPherson1,7, Esther Kong1, Cherie Bates1, Kelly Borkowski1, Matt Wiens1, Brittany Hewitson1, James Hopkins1, Jenifer Pham1, Nicholas Ceglia8, Richard Moore6, Andrew J. Mungall6, Jessica N. McAlpine6, The CRUK IMA XT Grand Challenge Team1, Sohrab P. Shah1,7,8 and Samuel Aparicio1,3,9.

Dissociation using collagenase at 37°C results in a stress response as compared to dissociation using a cold active protease at 6°C.
Example 2: Artifacts due to tissue dissociation

HSPA1A expression

The correlation is absent in immunohistochemistry staining

Zhang, Caushi, Pardoll, Smith, et al.
Example 2: Artifacts due to tissue dissociation

HSPA1A only expresses in solid tissue samples but not in PBMC (PBMC is handled without using dissociation enzyme)

Zhang, Caushi, Pardoll, Smith, et al.
How to deal with artifacts due to data generation?

Build knowledge
- Understand the data generation process
- Compare results from orthogonal data
- Analyze many datasets and find recurring patterns

Develop solution
- Improve experimental technologies
- Develop computational algorithms for artifact detection and removal
- Spike-in experiments for benchmark

Importantly, wet lab and dry lab investigators need to closely work together!
Artifacts due to data analysis

- Improper normalization
- Failure to control confounders
- Wrong models, assumptions or methods
Example 1: scRNA-seq differential expression

Wilcoxon test ignores sample variability. When there are multiple samples, it will create false discoveries.
Example 1: scRNA-seq differential expression

<table>
<thead>
<tr>
<th>Method</th>
<th>8 samples 410 cells per cluster</th>
<th>8 samples 41 cells per cluster</th>
<th>4 samples 168 cells per cluster</th>
<th>4 samples 21 cells per cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAISIN(FAVOR)</td>
<td>0.85</td>
<td>0.14</td>
<td>1.93</td>
<td>1.3</td>
</tr>
<tr>
<td>limmapb(FAVOR)</td>
<td>0.03</td>
<td>0.05</td>
<td>0.32</td>
<td>0.41</td>
</tr>
<tr>
<td>edgeRpb(count)</td>
<td>8.16</td>
<td>7.01</td>
<td>9.75</td>
<td>6.75</td>
</tr>
<tr>
<td>DESeq2pb(count)</td>
<td>0.61</td>
<td>0.17</td>
<td>2.6</td>
<td>0.41</td>
</tr>
<tr>
<td>RAISIN(norm)</td>
<td>0.24</td>
<td>0.22</td>
<td>22.15</td>
<td>102.14</td>
</tr>
<tr>
<td>RAISIN-LMM(norm)</td>
<td>4.94</td>
<td>0.32</td>
<td>22.15</td>
<td>102.14</td>
</tr>
<tr>
<td>IDEAS(count)</td>
<td>0</td>
<td>0.32</td>
<td>22.15</td>
<td>102.14</td>
</tr>
<tr>
<td>Wilcoxon(FAVOR)</td>
<td>9.2e3</td>
<td>3e3</td>
<td>8.7e3</td>
<td>3.1e3</td>
</tr>
<tr>
<td>MAST(FAVOR)</td>
<td>6e3</td>
<td>0.14</td>
<td>0.32</td>
<td>0.41</td>
</tr>
<tr>
<td>t-test(FAVOR)</td>
<td>7.3e3</td>
<td>7.01</td>
<td>9.75</td>
<td>6.75</td>
</tr>
<tr>
<td>scDD(FAVOR)</td>
<td>1.1e4</td>
<td>0.32</td>
<td>22.15</td>
<td>102.14</td>
</tr>
<tr>
<td>limmapb(count)</td>
<td>1.12</td>
<td>0.32</td>
<td>22.15</td>
<td>102.14</td>
</tr>
<tr>
<td>RAISIN-LMM(FAVOR)</td>
<td>9.2e3</td>
<td>3e3</td>
<td>8.7e3</td>
<td>3.1e3</td>
</tr>
<tr>
<td>t-test(norm)</td>
<td>7.45</td>
<td>7.01</td>
<td>9.75</td>
<td>6.75</td>
</tr>
<tr>
<td>edgeR(zinebwave)</td>
<td>439.85</td>
<td>327.71</td>
<td>307.11</td>
<td>4.91</td>
</tr>
<tr>
<td>limmapb(norm)</td>
<td>445.16</td>
<td>327.71</td>
<td>307.11</td>
<td>4.91</td>
</tr>
<tr>
<td>MAST(norm)</td>
<td>0.06</td>
<td>0.32</td>
<td>22.15</td>
<td>102.14</td>
</tr>
<tr>
<td>Wilcoxon(norm)</td>
<td>0.06</td>
<td>0.32</td>
<td>22.15</td>
<td>102.14</td>
</tr>
<tr>
<td>DESeq2(zinebwave)</td>
<td>476.3</td>
<td>123.71</td>
<td>102.91</td>
<td>102.91</td>
</tr>
<tr>
<td>scDD(norm)</td>
<td>844.21</td>
<td>327.71</td>
<td>307.11</td>
<td>4.91</td>
</tr>
</tbody>
</table>
Example 1: scRNA-seq differential pseudotemporal expression

Example 1: scRNA-seq differential pseudotemporal expression

<table>
<thead>
<tr>
<th>Method</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>tradeSeq_startVsEndTest</td>
<td>$1.88 \times 10^{-4}$</td>
</tr>
<tr>
<td>tradeSeq_associationTest</td>
<td>$6.09 \times 10^{-6}$</td>
</tr>
<tr>
<td>Monocle2</td>
<td>$4.13 \times 10^{-10}$</td>
</tr>
<tr>
<td>Monocle3</td>
<td>$1.70 \times 10^{-3}$</td>
</tr>
<tr>
<td>TSCAN</td>
<td>$1.68 \times 10^{-2}$</td>
</tr>
<tr>
<td>Lamian</td>
<td>0.999</td>
</tr>
</tbody>
</table>

Example 2: Chromatin accessibility locus effects

GM12878 scATAC-seq vs. GM12878 bulk DNase-seq

Correlation: 0.678

GM12878 scATAC-seq vs. ENCODE average bulk DNase-seq

Correlation: 0.725

Ji et al. Genome Biology. 21: 161, 2020
Example 2: Chromatin accessibility locus effects

Zhou et al. Nature Communications. 8: 1038, 2017
Zhou et al. Nucleic Acids Res. 47: e121, 2019
How to prevent and identify artifacts due to data analysis?

- Build good understanding of your data
- Choose appropriate models and assumptions
- Use proper controls
- Learn from analyzing many datasets
- Compare results from orthogonal data
- Benchmark using spike-in experiments
- Again, wet lab and dry lab investigators should work closely together
Summary: common sources of artifacts

Study design
- Lack of proper control or randomization

Data generation
- Bias and noise in technology
- Bias in experimental procedure

Data analysis
- Improper normalization
- Failure to control confounders
- Wrong models, assumptions or methods
General principles and common methods for preventing, identifying and removing artifacts

- Wet lab and dry lab investigators work closely together from day one
- Use proper control and randomization
- Validate findings using orthogonal data
- Learn from analyzing many datasets
- Use appropriate models, assumptions, and analysis methods
- Benchmark using spike-ins
Acknowledgment

The Johns Hopkins Bloomberg School of Public Health

Boyang Zhang
Weiqiang Zhou
Ruzhang Zhao
Wenpin Hou
Stephanie Hicks

The Johns Hopkins School of Medicine

Jiajia Zhang
Justina X. Caushi
Arbor G. Dykema
Srinivasan Yegnasubramanian
Drew M. Pardoll
Kellie N. Smith

Duke University School of Medicine

Zhicheng Ji

Funding
NIH R01HG009518, R01HG010889
Johns Hopkins IDIES Seed Fund
How to Submit Questions

• Click the “Q&A” icon located on at the bottom of your Zoom control panel

• Type your question in the Q&A box, then click “Send”

• Questions will be answered in the Question & Answer session at the end of the webinar (as time permits)