Scoring transcript variation in single cell RNA-seq data

Xiuwei Zhang
Single cell RNA-seq provides data at cellular resolution

- **Single-cell RNA-Seq**:
  - ~100,000 cells

- **Bulk RNA-Seq**:
  - Gene expression level
Single cell RNA-seq provides data at cellular resolution

Single cell RNA-seq also shows variation in read coverage profiles
Background

Traditional bulk RNA-seq tools for single cell data

- Calculating exon/intron inclusion scores:
  - MISO (Katz et al. 2010)
    used in Shalek et al. 2013
    high isoform variation, bimodal distribution of PSI scores
  - Bam2ssj (Pervouchine et al. 2013)
    used in Marinov et al, 2014
    number of isoforms for one gene in a cell

- Find novel splice junctions

Global analysis of profile variation across single cells
Sources? patterns? sub-populations?
Outline

• Method
  – Profile Variation (PV) score

• Benchmarking and thresholding
  – Various data sets
  – Various gene categories and exons
  – Compare with bulk RNA-seq

• Applications
  – Genes with high isoform variation
  – Patterns in isoform usage
  – Genes which switch isoforms
Profile variation (PV) score

Gene $g$ in Cell $s$

Gene $g$ in Cell $t$
Profile variation (PV) score

Probability distribution $P_1$

Probability distribution $P_2$

Jensen-Shannon Divergence (JSD)
Profile variation (PV) score

\[
\text{JSD}(P_1, P_2, \ldots, P_n) = H \left( \sum_{i=1}^{n} \pi_i P_i \right) - \sum_{i=1}^{n} \pi_i H(P_i)
\]

\(H(\cdot)\): entropy

increases with the number of categories in a discrete probability distribution.
Profile variation (PV) score

\[ JSD(P_1, P_2, \ldots, P_n) = H\left( \sum_{i=1}^{n} \pi_i P_i \right) - \sum_{i=1}^{n} \pi_i H(P_i) \]

\( H(\cdot) \): entropy

increases with the number of categories in a discrete probability distribution.

\[ PV = \frac{JSD}{\log_2(L)} \]
PV with gene length regressed out

\[ Y_{PV} = \beta_0 + \beta_1 X_{length} \]

\[ \hat{Y}_{PV} = \hat{\beta}_0 + \hat{\beta}_1 X_{length} \]

\[ Y_{PV} - \hat{Y}_{PV} \] is the length regressed PV scores \[ PV_{\text{length}} \]
Outline

• **Method**
  - Profile Variation (PV) score — *two versions of PV score*

• **Benchmarking and thresholding**
  - Various data sets
  - Various gene categories and exons
  - Compare with bulk RNA-seq

• **Applications**
  - Genes with high isoform variation
  - Patterns in isoform usage
  - Genes which switch isoforms
Compare between data sets

Differentiating T helper cells

Embryonic stem (ES) cells: different culture conditions
different cell cycle phases

Mahata et al 2014
Kołodziejczyk et al 2015
Buttener et al 2015

Marinov et al 2014
Compare between gene categories

![Graph comparing gene categories]

- **ERCCs**
- genes with 1 transcript
- genes with multiple transcripts

**x-axis:** length regressed out PV score

**y-axis:** density

The graph shows the distribution of PV scores for different gene categories.
Thresholding PV scores

PV of AS genes

☆ technical noise
☆ biological noise
☆ AS events

<table>
<thead>
<tr>
<th></th>
<th>cell 1</th>
<th>cell 2</th>
<th>cell 3</th>
<th>cell 4</th>
<th>cell 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>PV</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>AS</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>TN</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>BN</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
</tbody>
</table>
Thresholding PV scores

PV of AS genes
☆ technical noise
☆ biological noise
☆ AS events

PV of exons
☆ technical noise
☆ biological noise

Cell 1
Cell 2
Cell 3
Cell 4
Cell 5
Thresholding PV scores with exons

300~600 genes were found to have highly variable isoform usage
Compare with bulk RNA-seq data

single cell data

bulk data

ES cells

NPC cells

PV score

Cuffdiff
Compare with bulk RNA-seq data

What is consistent

What is different

- Genes with high PV but not detected by Cuffdiff
- Enriched in cell cycle genes
- Biological variation within one cell type
Outline

• Method
  - Profile Variation (PV) score -- two versions of PV score

• Benchmarking and thresholding
  - Various data sets -- conforms with biological heterogeneity
  - Various gene categories and exons -- significant variation
  - Compare with bulk RNA-seq -- consistent and more than bulk

• Applications
  - Genes with high isoform variation
  - Patterns in isoform usage
  - Genes which switch isoforms
Genes with highly variable isoforms

Isoform variation at two levels

$PV$

expression regulation
chromatin modification

$PV_{\text{length}}$

immunology
T helper cells

cell cycle
ES cells
Find representative read coverage patterns

<table>
<thead>
<tr>
<th>cell 1</th>
<th>cell 2</th>
<th>...</th>
<th>cell n</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\sqrt{PV}$</td>
</tr>
<tr>
<td>cell 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cell 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cell n</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Pairwise $\sqrt{PV}$ is a metric

Cluster Dendrogram
Find representative read coverage patterns

cell 1

cell 2

... cell n

Cluster Dendrogram
Find representative read coverage patterns

Example: Nsf in T cells
Find correlated genes in isoform usage

<table>
<thead>
<tr>
<th></th>
<th>Gene 1</th>
<th>Gene 2</th>
<th>Gene 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell 1</td>
<td>Pattern A</td>
<td>Pattern A</td>
<td>Pattern A</td>
</tr>
<tr>
<td>Cell 2</td>
<td>Pattern A</td>
<td></td>
<td>Pattern A</td>
</tr>
<tr>
<td>Cell 3</td>
<td></td>
<td></td>
<td>Pattern C</td>
</tr>
<tr>
<td>Cell 4</td>
<td>Pattern B</td>
<td>Pattern B</td>
<td>Pattern B</td>
</tr>
<tr>
<td>Cell 5</td>
<td>Pattern B</td>
<td>Pattern B</td>
<td></td>
</tr>
</tbody>
</table>

Difficulty: genes are expressed in a small number of cells.
Find correlated genes in isoform usage

Cell vs “gene pattern” binary matrix

<table>
<thead>
<tr>
<th></th>
<th>Gene1 PatternA</th>
<th>Gene1 PatternB</th>
<th>Gene2 PatternA</th>
<th>Gene2 PatternB</th>
<th>Gene3 PatternA</th>
<th>Gene3 PatternB</th>
<th>Gene3 PatternC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell 1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cell 2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cell 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Cell 4</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Cell 5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Find clusters of genes with Jaccard distance $< h$

Compare with random binary matrices:

Isoform usage across single cells has high stochasticity
Genes which switch isoforms between cell types

1. high ratio of:
\[
\frac{\text{Average}(\text{Inter-group distances})}{\text{Average}(\text{Intra-group distances})}
\]

2. high PV(all cells)
NPC cells

12 ES cells
5 NPC cells

Lig1
Lig1 protein domain of the long transcript

Lig1 protein domain of the short transcript

- Lig1 is a cell cycle gene
- Cells slow down cycling ES $\rightarrow$ NPC
- Not detected in bulk data
Outline

• Method
  – Profile Variation (PV) score -- two versions of PV score

• Benchmarking and thresholding
  – Various data sets -- conforms with biological heterogeneity
  – Various gene categories and exons -- significant variation
  – Compare with bulk RNA-seq -- consistent and more than bulk

• Applications
  – Genes with high isoform variation -- sources of isoform variation
  – Patterns in isoform usage -- high stochasticity
  – Genes which switch isoforms -- function change during differentiation
Acknowledgements

Sarah Teichmann

Valentine Svensson

Jong Kyoung Kim

Oliver Stegle (EBI)
John Marioni (EBI)
Nick Owens (MRC NIMR)
Kaur Alasoo (Sanger Institute)