Dimension Reduction
PCA, tSNE, UMAP, Integration

v2023-02

Simon Andrews
simon.andrews@babraham.ac.uk
## Where are we heading?

<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</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>Inpp5d</td>
<td>inositol polyphosphate-5-phosphatase D</td>
<td>7.00</td>
<td>5.45</td>
<td>5.89</td>
<td>6.03</td>
<td>5.75</td>
</tr>
<tr>
<td>Aim2</td>
<td>absent in melanoma 2</td>
<td>3.01</td>
<td>4.37</td>
<td>4.59</td>
<td>4.38</td>
<td>4.18</td>
</tr>
<tr>
<td>Gldn</td>
<td>gliomedin</td>
<td>3.48</td>
<td>3.63</td>
<td>4.61</td>
<td>4.70</td>
<td>4.74</td>
</tr>
<tr>
<td>Frem2</td>
<td>Fras1 related extracellular matrix protein 2</td>
<td>4.75</td>
<td>4.66</td>
<td>3.46</td>
<td>3.74</td>
<td>3.45</td>
</tr>
<tr>
<td>Rps3a1</td>
<td>ribosomal protein S3A1</td>
<td>6.10</td>
<td>7.23</td>
<td>7.44</td>
<td>7.36</td>
<td>7.34</td>
</tr>
<tr>
<td>Slc38a3</td>
<td>solute carrier family 38, member 3</td>
<td>1.90</td>
<td>3.16</td>
<td>3.52</td>
<td>3.61</td>
<td>3.19</td>
</tr>
<tr>
<td>Mt1</td>
<td>metallothionein 1</td>
<td>5.07</td>
<td>6.49</td>
<td>6.46</td>
<td>6.04</td>
<td>6.05</td>
</tr>
<tr>
<td>C1s1</td>
<td>complement component 1, s subcomponent 1</td>
<td>2.74</td>
<td>3.02</td>
<td>3.86</td>
<td>4.10</td>
<td>4.10</td>
</tr>
<tr>
<td>Cds1</td>
<td>CDP-diacylglycerol synthase 1</td>
<td>4.55</td>
<td>4.22</td>
<td>3.80</td>
<td>3.16</td>
<td>3.12</td>
</tr>
<tr>
<td>Ifi44</td>
<td>interferon-induced protein 44</td>
<td>4.82</td>
<td>4.52</td>
<td>3.87</td>
<td>3.42</td>
<td>3.59</td>
</tr>
<tr>
<td>Lefty2</td>
<td>left-right determination factor 2</td>
<td>6.95</td>
<td>6.28</td>
<td>5.88</td>
<td>5.60</td>
<td>5.61</td>
</tr>
<tr>
<td>Fmr1nb</td>
<td>fragile X mental retardation 1 neighbor</td>
<td>4.28</td>
<td>2.78</td>
<td>3.10</td>
<td>3.25</td>
<td>2.57</td>
</tr>
<tr>
<td>Tagln</td>
<td>transgelin</td>
<td>7.93</td>
<td>7.91</td>
<td>7.20</td>
<td>7.02</td>
<td>6.68</td>
</tr>
</tbody>
</table>

Each dot is a cell

Groups of dots are similar cells

Separation of groups could be interesting biology
Too much data!

- 5000 cells and 2500 measured genes
- Realistically only 2 dimensions we can plot (x,y)
Principle Components Analysis

- Method to optimally summarise large multi-dimensional datasets
- Can find a smaller number of dimensions (ideally 2) which retain most of the useful information in the data
- Builds a recipe for converting large amounts of data into a single value, called a Principle Component (PC), eg:

  \[ PC = (\text{GeneA} \times 10) + (\text{GeneB} \times 3) + (\text{GeneC} \times -4) + (\text{GeneD} \times -20) \ldots \]
Principle Components Analysis

• Method to optimally summarise large multi-dimensional datasets
• Can find a smaller number of dimensions (ideally 2) which retain most of the useful information in the data

• Builds a recipe for converting large amounts of data into a single value, called a Principle Component (PC), eg:

\[ PC = (\text{GeneA} \times 10) + (\text{GeneB} \times 3) + (\text{GeneC} \times -4) + (\text{GeneD} \times -20) \ldots \]
How does PCA work?

Simple example using 2 genes and 10 cells
How does PCA work?

Find line of best fit, passing through the origin
Assigning Loadings to Genes

Loadings:
- Gene1 = 0.82
- Gene2 = 0.57

Higher loading equals more influence on PC
More Dimensions

• The same idea extends to larger numbers of dimensions (n)

• First PC rotates in (n-1) dimensions
  – Next PC is perpendicular to PC2, but rotated similarly (n-2)
  – Last PC is remaining perpendicular (no choice)
  – Same number of PCs as genes
Explaining Variance

• Each PC always explains some proportion of the total variance in the data. Between them they explain everything
  – PC1 always explains the most
  – PC2 is the next highest etc. etc.

• Since we only plot 2 dimensions we’d like to know that these are a good explanation

• How do we calculate this?
Explaining variance

- Project onto PC
- Calculate distance to the origin
- Calculate sum of squared differences (SSD)
  - This is a measure of variance called the ‘eigenvalue’
  - Divide by (points-1) to get actual variance
Explaining Variance – Scree Plots

Variance / Eigenvector

PC1 | PC2 | PC3 | PC4 | PC5

Variance / Eigenvector

PC1 | PC2 | PC3 | PC4 | PC5
So PCA is great then?

Kind of...

Non-linear separation of values
So PCA is great then?

Kind of...

PC2

PC1

Not optimised for 2-dimensions
tSNE to the rescue...

• T-Distributed Stochastic Neighbour Embedding

• Aims to solve the problems of PCA
  – Non-linear scaling to represent changes at different levels
  – Optimal separation in 2-dimensions
How does tSNE work?

- Based around all-vs-all table of pairwise cell to cell distances
Distance scaling and perplexity

- Perplexity = expected number of neighbours within a cluster
- Distances scaled relative to perplexity neighbours
Perplexity Robustness

Perplexity = 4

Perplexity = 2

Perplexity = 5
tSNE Projection

• Randomly scatter all points within the space (normally 2D)

• Start a simulation
  – Aim is to make the point distances match the distance matrix
  – Shuffle points based on how well they match
  – Stop after fixed number of iterations, or
  – Stop after distances have converged
tSNE Projection

- X and Y don’t mean anything (unlike PCA)
- Distance doesn’t mean anything (unlike PCA)
- Close proximity is highly informative
- Distant proximity isn’t very interesting
- Can’t rationalise distances, or add in more data
tSNE Practical Examples

Perplexity Settings Matter

Original   Perplexity = 2   Perplexity = 30   Perplexity = 100

https://distill.pub/2016/misread-tsne/
tSNE Practical Examples

Cluster Sizes are Meaningless

Original  Perplexity = 5  Perplexity = 50

https://distill.pub/2016/misread-tsne/
tSNE Practical Examples

Distances between clusters can’t be trusted

Original

Perplexity = 5

Perplexity = 30

https://distill.pub/2016/misread-tsne/
So tSNE is great then?

Kind of...

Imagine a dataset with only one super informative gene

- Now 3 genes
- Now 3,000 genes
- Everything is the same distance from everything
So everything sucks?

- **PCA**
  - Requires more than 2 dimensions
  - Thrown off by quantised data
  - Expects linear relationships

- **tSNE**
  - Can’t cope with noisy data
  - Loses the ability to cluster

**Answer: Combine the two methods, get the best of both worlds**

- **PCA**
  - Good at extracting signal from noise
  - Extracts informative dimensions

- **tSNE**
  - Can reduce to 2D well
  - Can cope with non-linear scaling

This is what many pipelines do in their default analysis
So PCA + tSNE is great then?

Kind of...

- tSNE is slow. This is probably its biggest crime
  - tSNE doesn’t scale well to large numbers of cells (10k+)

- tSNE only gives reliable information on the closest neighbours. Large distance information is almost irrelevant
UMAP to the rescue!

• UMAP is a replacement for tSNE to fulfil the same role

• Conceptually very similar to tSNE, but with a couple of relevant (and somewhat technical) changes

• Practical outcome is:
  – UMAP is quite a bit quicker than tSNE
  – UMAP can preserve more global structure than tSNE*
  – UMAP can run on raw data without PCA preprocessing*
  – UMAP can allow new data to be added to an existing projection

* In theory, but possibly not in practice
UMAP differences

• Instead of the single perplexity value in tSNE, UMAP defines
  – **Nearest neighbours**: the number of expected nearest neighbours – basically the same concept as perplexity

  – **Minimum distance**: how tightly UMAP packs points which are close together

• Nearest neighbours will affect the influence given to global vs local information. Min dist will affect how compactly packed the local parts of the plot are.
UMAP differences

- Structure preservation – mostly in the 2D projection scoring

https://towardsdatascience.com/how-exactly-umap-works-13e3040e1668
So UMAP is great then?

Kind of...
So UMAP is all hype then?

No, it really does better for some datasets...

3D mammoth skeleton projected into 2D

- tSNE: Perplexity 2000, 2h 5min
- UMAP: Nneigh 200, mindist 0.25, 3min

https://pair-code.github.io/understanding-umap/
Practical approach PCA + tSNE/UMAP

• Filter heavily before starting
  – Nicely behaving cells
  – Expressed genes
  – Variable genes

• Do PCA
  – Extract most interesting signal
  – Take top PCs. Reduce dimensionality (but not to 2)

• Do tSNE/UMAP
  – Calculate distances from PCA projections
  – Scale distances and project into 2-dimensions
So PCA + UMAP is great then?

Kind of... as long as you only have one dataset

– In 10X every library is a 'batch'
– More biases over time/distance
– Biases prevent comparisons
– Need to align the datasets
Data Integration

- Works on the basis that there are 'equivalent' collections of cells in two (or more) datasets
- Find 'anchor' points which are equivalent cells which should be aligned
- Quantitatively skew the data to optimally align the anchors
Define key 'anchor' points between equivalent cells
UMAP/tSNE integration

Skew data to align the anchors
Defining Integration Anchors

Mutual Nearest Neighbours (MNN)

For each cell in data1 find the 3 closest cells in data2
Defining Integration Anchors

Mutual Nearest Neighbours (MNN)

Do the same thing the other way around
Defining Integration Anchors

Mutual Nearest Neighbours (MNN)

Select pairs of cells which are in other's nearest neighbour groups.
Defining nearest neighbours

• Distance in original expression quantitation
  – Really noisy (different technology, normalisation, depth)
  – Slow and prone to mis-prediction

• Use a cleaner (less noisy) representation
  – Correlation (CCA)
  – Principal Components (rPCA)
Defining Nearest Neighbours

Canonical Correlation Analysis

Gene Expression Values may match poorly, but gene correlations are more robust.
Defining Integration Anchors

Reciprocal PCA

1. Define PCA Space for Data 1
2. Project cells from data 2 into the data1 PCA space
3. Find nearest neighbours
4. Repeat by projecting Data1 into Data2 PCA space
5. Find mutual nearest neighbours
Factors Affecting Integration

• Which genes are submitted to the integration
  – Expressed in all datasets
  – Variable in all datasets

• Which method is used to define nearest neighbours
  – Normalised data, Correlation, Reverse PCA

• How many nearest neighbours you consider
  – Default is around 5, some clusters require more (20ish)

• Other filters to remove artefacts