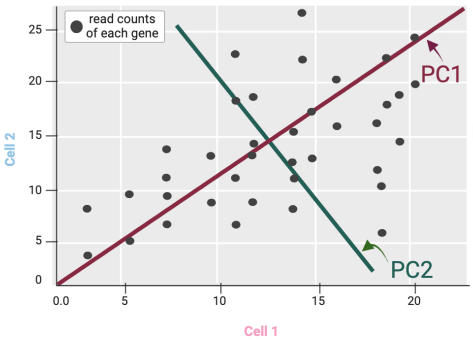
Principal Component Analysis (PCA)

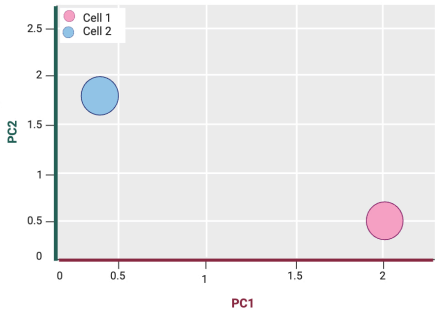
① Find principal component 1 (PC1) and principal component 2 (PC2)

PC1 is the fit line that accounts for the most variation in data. PC2 is perpendicular to PC1. The example below is highly simplified, comparing 2 cells.



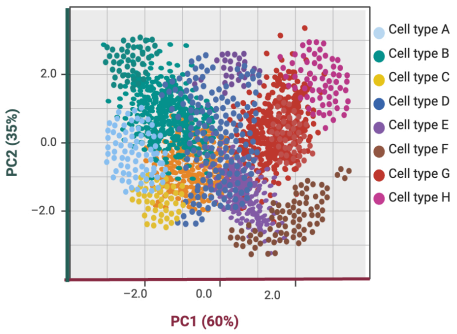
② Rotate and find coordinates of points for each cell using loading scores

- Loadings describe how much influence each gene has on the PC.
- Cell coordinate = Σ (read count * loading) for each gene in the cell, for PC 1 (x-coordinate) and PC2 (y-coordinate). This is completed for each cell.



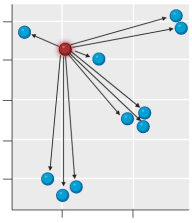
③ Observe Output

In the end, the resulting graph exhibits cells as points, and can be coloured and annotated by shared characteristics such as cell type. The percentages by each PC corresponds to the amount of variation in the data each PC accounts for.



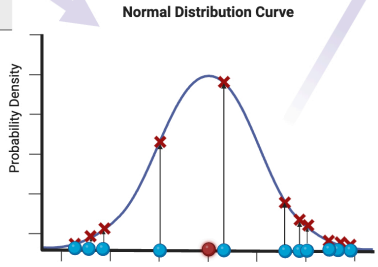
t-Distributed Stochastic Neighbour Embedding (t-SNE)

① Find likelihood that points will be neighbours in high dimensional space

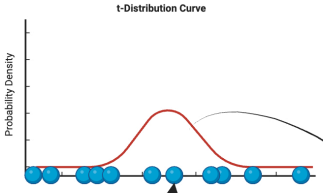


Calculate raw distances between points

Plot distances on gaussian curve to find scaled similarity scores

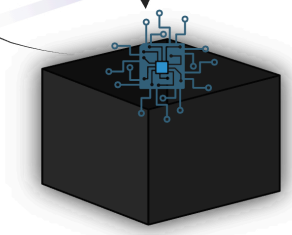


② Initialization and optimization cycle in low dimensional space

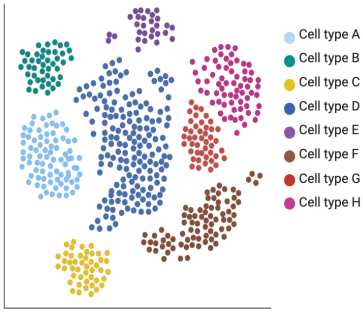


Similarly to step 1, points are plotted on a curve. Points are initially randomized .

Through many iterations, using gradient descent, final point positions are identified when they are similarly positioned to those in the higher dimension.



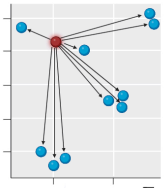
③ Observe output



The algorithm produces a graph that shows well-separated clusters with each point being a cell. The points can be coloured and annotated by shared characteristics such as cell type.

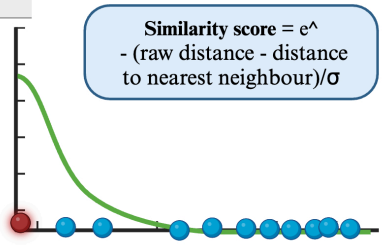
Uniform Manifold Approximation and Projection (UMAP)

① Find likelihood that points will be neighbours in high dimensional space

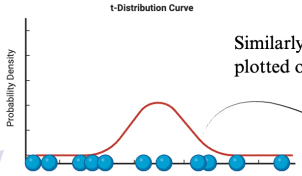


Like t-SNE, UMAP starts by calculating similarity scores in the high dimensional space. In lieu of using a gaussian curve, UMAP calculates a unique curve for each iteration of comparisons.

Similarity score = e^ - (raw distance - distance to nearest neighbour)/σ

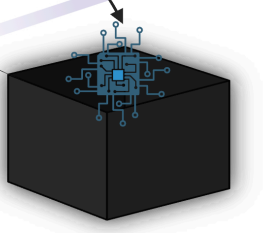


② Initialization and optimization cycle in low dimensional space

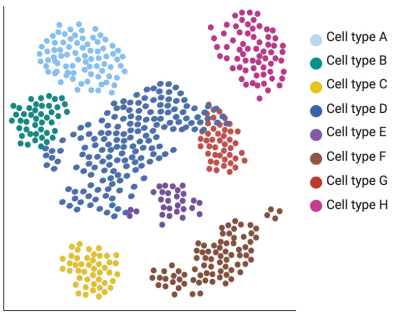


Similarly to step 1, points are plotted on a curve.

Through many iterations, using stochastic gradient descent, final point positions are identified when points that are close together in the high dimensional space, are similarly close to each other in the low dimensional space.



③ Observe output



The algorithm produces a graph that shows well-separated clusters with each point being a cell. The points can be coloured and annotated by shared characteristics such as cell type. UMAP is effective at maintaining global and local structure.