1. Suppose you perform single cell RNA sequencing on 2 cells each from two different stages of a cancer. You want to compare gene expression between these two stages of cancer in order to identify genes that contribute to varying disease severity. You determine the expression read counts of 2000 genes per cell (See table below for example). Some of these genes have extreme expression differences between cells regardless of whether the cells are from the same or different stages of the cancer. For example, they are nearly absent or highly expressed. Before making inferences about the biological differences between the two stages of the cancer based on the varying gene expression of the cells, you would like to first assess some pairwise relationships to build a framework for general patterns. You decide to compute pairwise measures of relationships between (1) each cell; (2) between cell types; and (3) between each gene.

A. Suggest a measure of relatedness (or more than one)  
B. Explain your goals—that is, what kind of relatedness that you are trying to characterize with your measure(s)  
C. Explain your choice of measure(s) in terms of your goals.  
D. Suggest possible problems with your chosen measure (e.g., when some value is zero, the measure goes to infinity).

In the following, I will treat the 2000 genes as features.

If the kind of relatedness I’m interested in is “similarity of gene expression magnitude”, then one way to capture this would be to calculate the Euclidean distance between each pair of cells based on their gene expression. Since Euclidean distance basically sums up the squared differences between a pair of objects for each coordinate (in this case, gene expression values), this method will return lower distances for pairs of cells that have many genes with similar expression values. There are some potential problems with this method, though. For one thing, if there are many genes with zero reads shared between a pair of cells, then this will actually make the cells look quite similar (since the difference between 0 and 0 is 0). We could avoid this problem by removing any genes that are zero in all or most of the cells. Another problem is related to the wide range of gene expression values that can be observed. Some genes only vary from e.g. 10 to 100 reads, whereas other genes may vary from e.g. 1,000 to 10,000 reads. Based on Euclidean distance, the difference between 1,000 and 10,000 looks much larger than the difference between 10 and 100. But if we think about this in terms of fold change, then they are both a 10x change. So is 1000 vs 10000 really “worse” than 10 vs 100? Maybe, or maybe not... In any case, one way we could reduce this effect is to do something called “standardization”/”z-normalization”, which basically centers the expression values of each gene around the mean value and then scales it by the standard deviation (look up more on this if you are interested). Although not a perfect solution, this at least helps bring all the genes on to the same scale.

Another way to try to measure relatedness of cells based on gene expression is to use correlation. The most common correlation coefficients are Pearson and Spearman, and each has its own drawbacks when it comes to gene expression measurements (for example, Pearson correlation can be greatly affected by outliers with large magnitude, and Spearman correlation can be affected by highly variable ranks at low expression values).

Alternatively, maybe we decide that we don’t care about the magnitude of gene expression, but simply want to compare cells based on how many genes are “on” or “off” in both. In this case, we could set all the gene expression values that are greater than 0 to 1 (leaving the 0 expression as 0). Then we can count up how many genes are 1 in both and use this as our measure (i.e. take the intersection). One problem with this is that if one cell happens to express almost all the genes, it’s going to look very similar to all of the other cells. To prevent this, one idea is to divide the intersection count by the union count. In other words, we express the number of shared expressed genes as a fraction of the number of total expressed genes between the two cells. This is also known as the Jaccard similarity measure.

Before calculating the distance/similarity measure, you could also try to reduce the feature space to a more meaningful set using either prior knowledge/literature search, or a method like PCA.

(2) Between cell types

We only have two examples of each cell type here, so we really don’t have much information to work with in terms of figuring out what the prototypical cell for each type would look like. One idea is to just average the expression values for each cell type and then compare the two average profiles using one of the distance measures mentioned above. Again, we could also try to do various things to try to find a more “meaningful” subset of genes for each cell type, which may or may not work depending on how variable the cell type is and how much prior info we have.

(3) Between genes

Now we switch perspectives and have the genes be examples and the cells be features. If we’re interested in finding genes with similar expression patterns across cells, we can again use something like Euclidean distance (with pretty much the same caveats as mentioned above).

Example read counts:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Gene** | Cell Type 1 | | Cell Type 2 | |
| **Stage 1A** | **Stage1B** | **Stage 4A** | **Stage4B** |
| Upp2 | 1 | 20 | 0 | 0 |
| Dusp15 | 14 | 33 | 30 | 71 |
| Dpy19l1 | 111 | 655 | 343 | 108 |
| Mtx2 | 1770 | 995 | 1496 | 475 |
| Alkbh1 | 9 | 581 | 1 | 29 |
| Man1c1 | 3 | 699 | 848 | 112 |
| AU023762 | 2 | 0 | 0 | 56 |
| Plscr3 | 0 | 6 | 0 | 0 |
| Tas1r3 | 3 | 0 | 0 | 0 |