**Applying dimensional reduction to HiPR-FISH data**

Software companies are using Computer vision to recognize patterns in photos etc.

Can we recognize patterns in microbiome structure?

What are the data like?

* Taxon ID
* Location
  + Relative to objects (epithelium, food particles)
  + Relative to all other cells
* Shape
* Volume
* Brightness

What is the scale of measurement?

* Within one image – 150μm square
* Between images – along the length of the gut?

What is size of the data?

* About 1,000-10,000 cells per FOV

What information are we trying to pull out?

* Single cell is all about separating cell types
* Spatial patterns in a single FOV?
* Spatial patterns in a taxon?
* Spatial patterns for single cells?

Can we have an array

* Rows for each image
* Columns for each taxon
* 3rd dimension with various data parameters
  + Radial distance to other cell types
  + Distance from objects
  + Radial distance from each other (clustering, dist betw clusters, dist betw single cells)
  + Fourier stuff
  + Abundance

Look at dimensional reduction

* Images within the same sample vs betw/ samples
* Clustering between treatments

**Understanding UMAP**

<https://pair-code.github.io/understanding-umap/>

The biggest difference between the the output of UMAP when compared with t-SNE is this balance between local and global structure - UMAP is often better at preserving global structure in the final projection.

UMAP constructs a high dimensional graph representation of the data then optimizes a low-dimensional graph to be as structurally similar as possible.

To determine connectedness, UMAP extends a radius outwards from each point, connecting points when those radii overlap.

Choosing this radius is critical - too small a choice will lead to small, isolated clusters, while too large a choice will connect everything together. UMAP overcomes this challenge by choosing a radius locally, based on the distance to each point's nth nearest neighbor.

UMAP then makes the graph "fuzzy" by decreasing the likelihood of connection as the radius grows. Finally, by stipulating that each point must be connected to at least its closest neighbor, UMAP ensures that local structure is preserved in balance with global structure.

**PHATE**

<https://doi-org.proxy.library.cornell.edu/10.1038/s41587-019-0336-3>

in biological systems, where structure exists at many different scales and a faithful visualization can lead to hypothesis generation.

methods tend to be sensitive to noise. Biomedical data is generally very noisy, and methods like PCA and Isomap4 fail to explicitly remove this noise for visualization, rendering fine-grained local structure impossible to recognize.

nonlinear visualization methods such as t-SNE often scramble the global structure in data.

many dimensionality-reduction methods (for example, PCA and diffusion maps) fail to optimize for two-dimensional (2D) visualization as they are not specifically designed for visualization

most available packages still follow the original implementation and thus cannot run on big data, which severely limits the usability of these methods in the medium-to-long term.

any data will be transformed to fit a tree with Monocle212 or clusters with t-SNE. While such methods are useful for data that fit their prior assumptions, they can generate misleading results otherwise, and are often ill suited for hypothesis generation or data exploration

PHATE generates a low-dimensional embedding specific for visualization, which provides an accurate, denoised representation of both local and global structure of a dataset in the required number of dimensions without imposing any strong assumptions on the structure of the data, and is highly scalable both in memory and runtime.