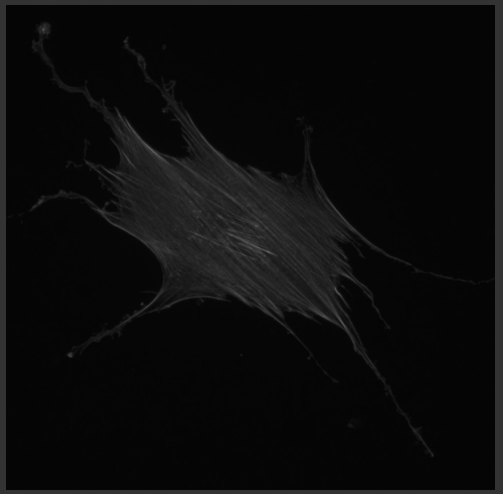
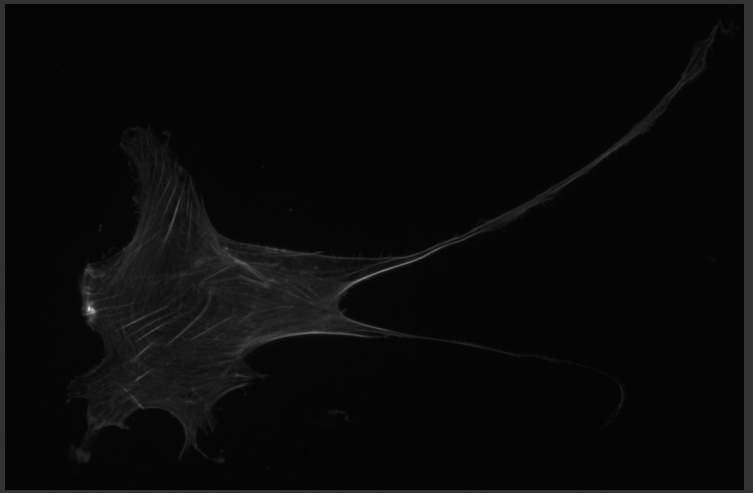
Stem Cell Classification using TDA*

Yossi Bokor Bleile, Aalborg University

Joint work with Florian Rehfeldt (Bayreuth) and Patrice Koehl (UC Davis)





*it's it all persistent homology black magic and clustering, but maybe we need slightly more geometry and topology

Biology Crash Course

Human mesenchymal stem cells (hMSC) are multi potent stem cells, which can differentiate into various *lineages* of cells, including:
* osteocytes
* adipocytes
* chondrocytes
* neurocytes
* hepatocytes

Differentiation into lineages can be triggered by environmental stimuli such as the rigidity of the extracellular matrix they are cultured on.

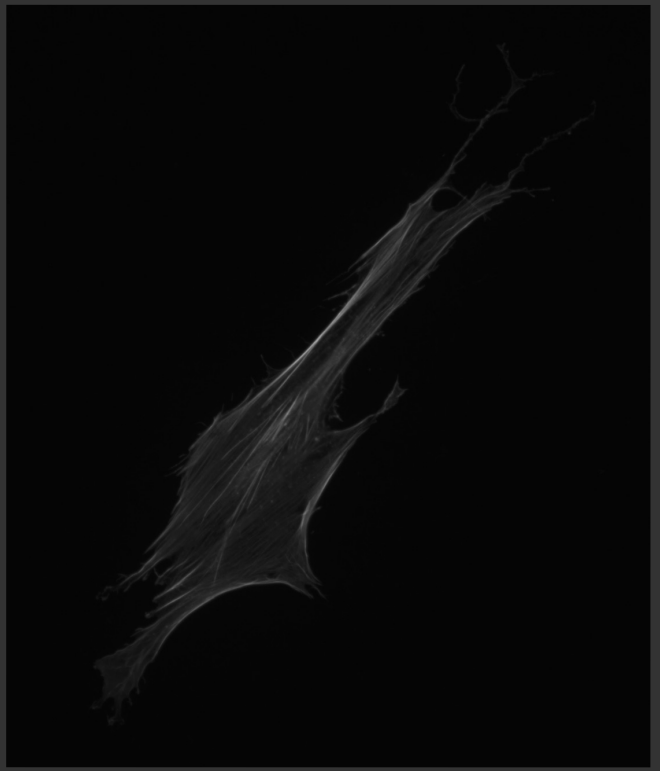
Question: can we distinguish between these lineages based on morphological features of a cell?

Morphological Features

Given a stem cell, what information can we consider to examine the morphological features?

There are many options, including:
* inscribed ellipse
* inscribing ellipse
* least square ellipse
* Willmore energy
* persistent homology

Given 2 cells, we can compare morphological features using:
* distances between ? ellipse
* Willmore distance
* Frechet distance
* compare persistent homology



Experimental Set Up and Data Pipeline

* Sets of cells from same donor

* Cultured on cellular matrix for 24 hours

* Fixed, stained, imaged

* Contour extracted using FilamentSensor

* Compute dissimilarity matrix

* k++-means clustering to identify/validate lineages

Oh no, problem!

The preferred method for collecting hMSCs is from bone marrow, and unfortunately, it is not possible to avoid the presence of other cells, such as bone barrow fibroblasts, in the sample.

We anticipate about 6% of non-hMSCs to be present in a given sample.

Hence, we need to *clean* our population of cells, to ensure we are only learning information about the morphology of hMSCs.

Due to issues in the culturing and imagining process, we expect 6% of cells to display *abnormal* growth patterns.

New Aim

Due to the the presence of the ~12% of cells which are not hMSCs/display abnormal morphology, we can not immediate identify morphological features to characterise lineages. So, we have a new problem.

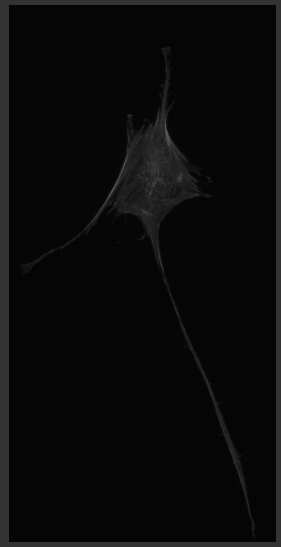
Question: can we identify the 2 sub-populations present in a sample?

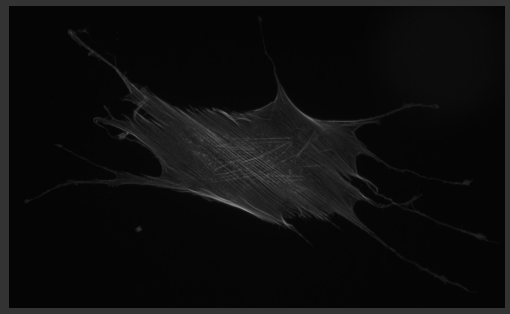
Spoiler

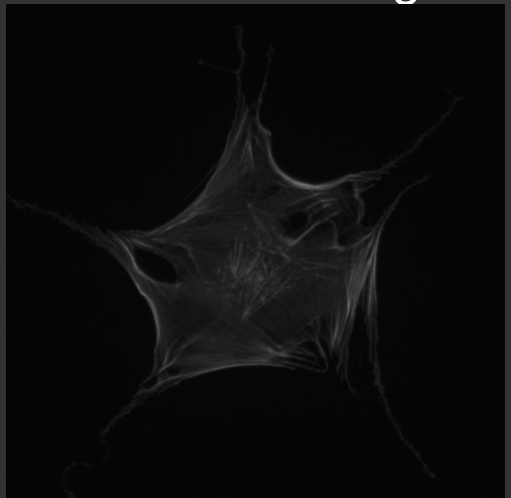
YES

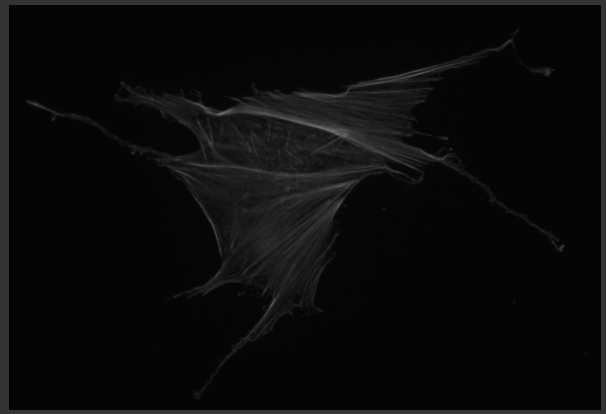
Identifying sub-populations

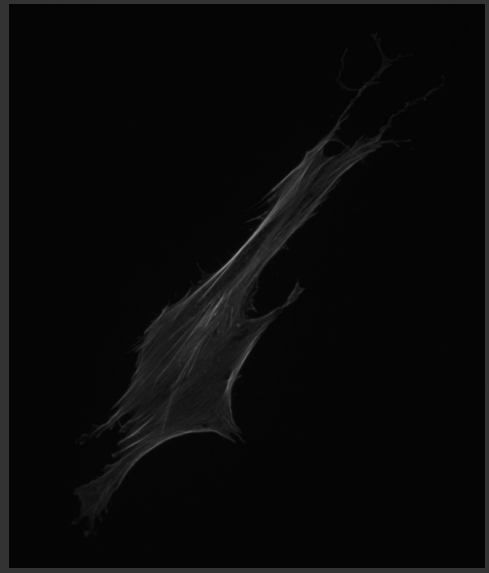
Problem set up:
Given a population of cells cultured in the same environmental conditions (extracellular matrix, treatment, time cultured, staining), can we identify the ~12% of cells we expect to exhibit abnormal growth patterns?











New Experimental Set Up and Data Pipeline

* Sets of cells from same donor

* Cultured on the same cellular matrix for 24 hours

* Fixed, stained, imaged

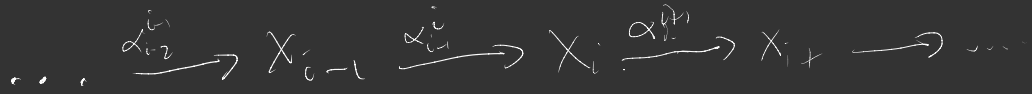
* Contour extracted using FilamentSensor

* Compute dissimilarity matrix

* k++-means clustering to identify/validate lineages

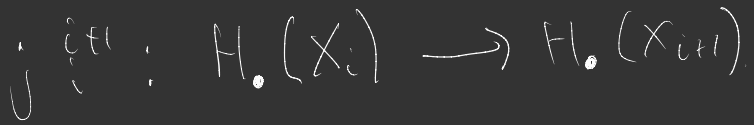
Persistent Homology Refresher

Given a sequence of topological spaces with maps between them



we can look at the homology at each *i*.

Then, a map between spaces induces a map on homology:



We can then track how the homology changes as we increase the parameter.

Persistent Homology Refresher

We now have

$$\cdots \longrightarrow H_0(X_{i-1}) \longrightarrow H_0(X_i) \longrightarrow H_0(X_{i+1}) \longrightarrow \cdots$$

At each time *i* we can ask the question: how does the homology change, if it changes?

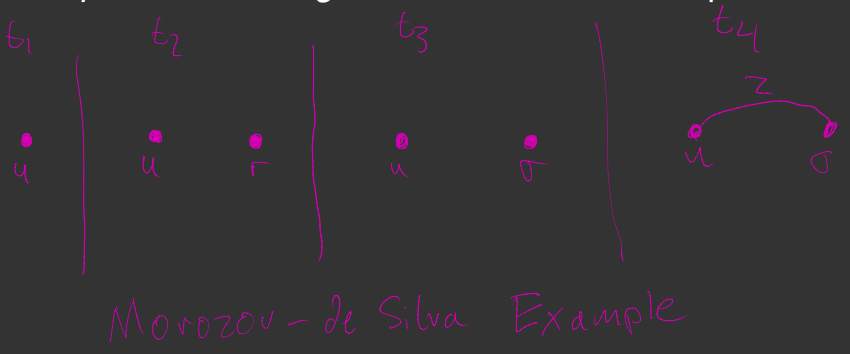
Any of the following three things can occur:
* nothing
* a homology class *disappears*
* a homology class *appears*

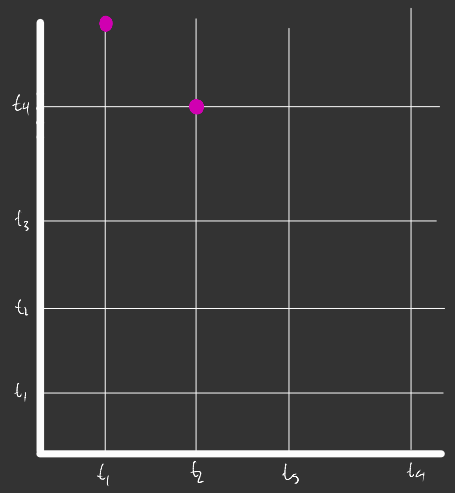
Persistent Homology Refresher

We can summarise these events in a *persistence diagram*.

Definition:

A *persistence diagram PD* is a multi-set of points in the plane.





Persistent Homology Refresher

Now we have a way for obtaining a summary of the morphology of a cell, so one thing left to do: have (dis)similarity measure on this summary.

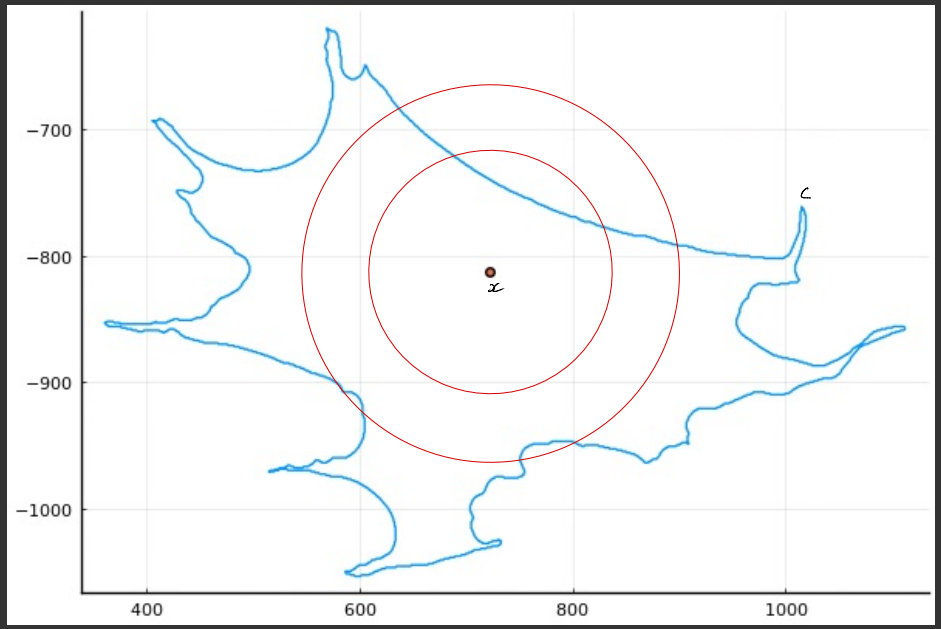
Definition:

Given two persistence diagrams *C* and *D*, the *p-Wasserstein distance* between them is

$$W_p(C,D) = \min_{\varphi \in M} \left(\sum_{x \in C} \|x - \varphi(x)\|^p \right)^{1/p}$$

where *M* is the space of matchings between *C* and *D*.

Identifying subpopulations: how do?



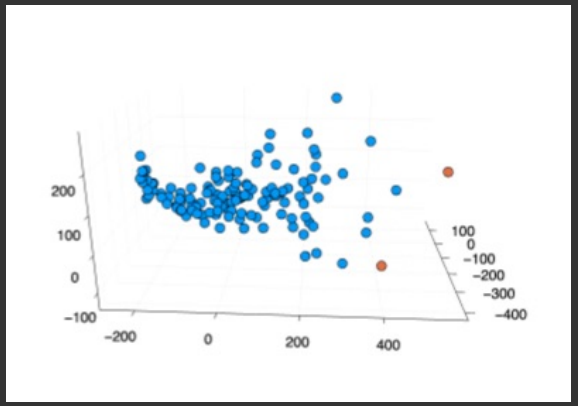
Let *x* be the center of the nucleus and *f* the distance to *x* function.

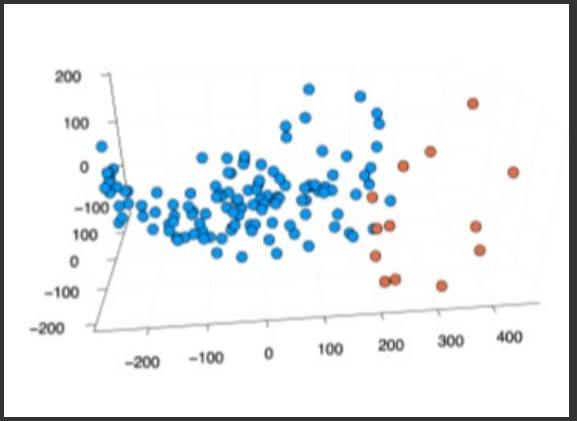
We then take the persistent homology of *C* using *f*, and pair the essential 0- and essential 1-cycles to obtain the 0th persistence diagram.

Identifying subpopulations: how do?

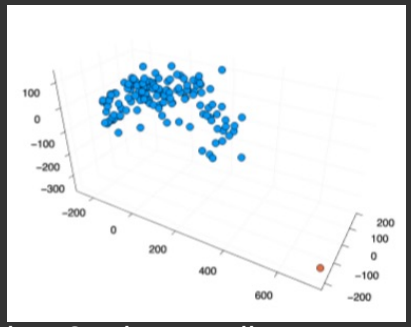
Comparing the contours of the cells in the population, we construct a dissimilarity matrix, which allows us to do *k*-means clustering (*k*=2 or 3), which we can the visualise using multi-dimensional scaling.

We use the 2-Wasserstein distance on persistence diagrams as our dissimilarity score between two cells.

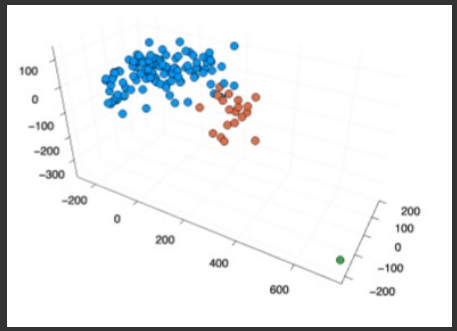




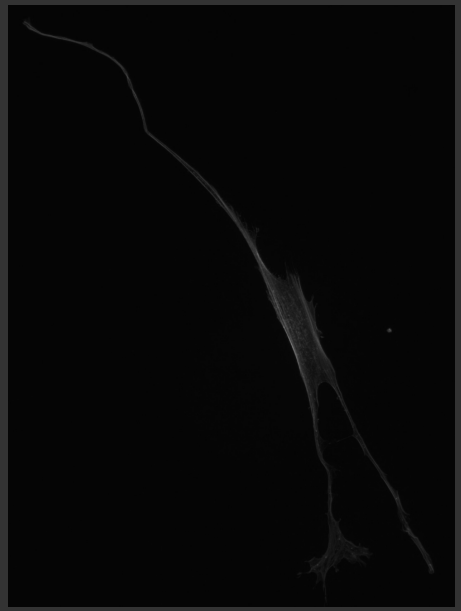
Results - set 1

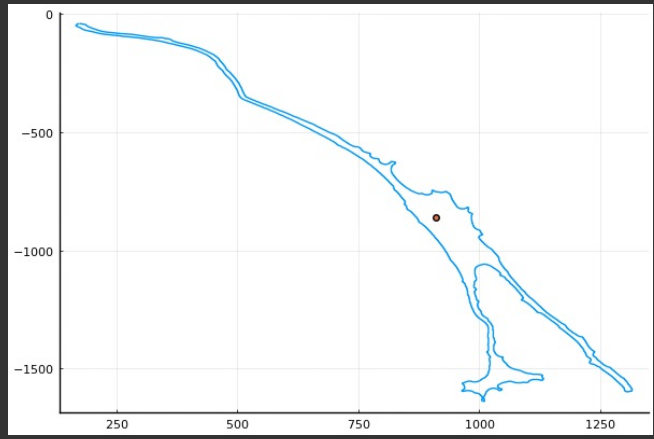


k = 2, clear outlier

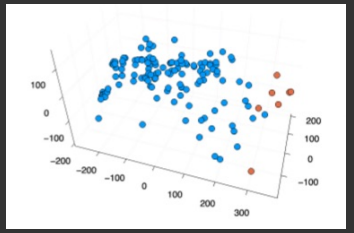


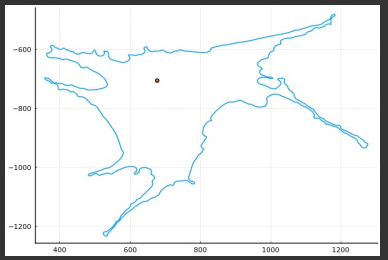
k = 3, clear outlier

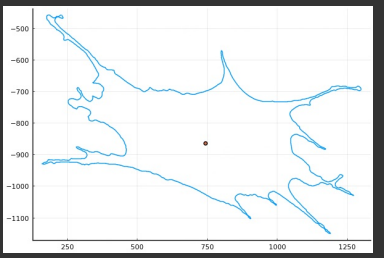


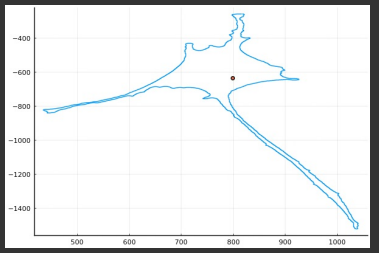


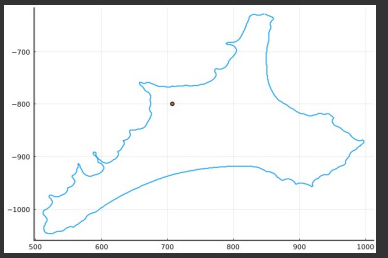
Results - set 1

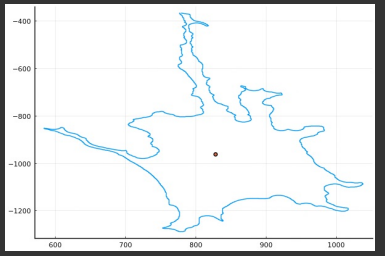


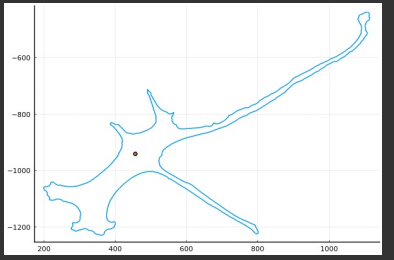




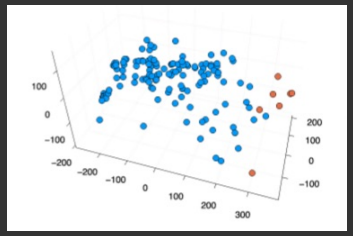


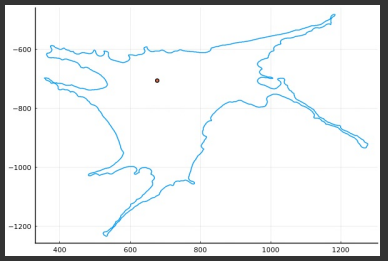


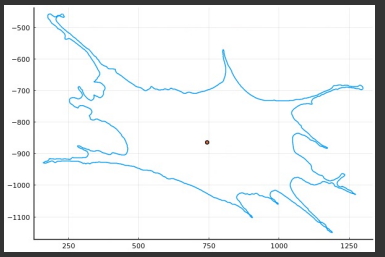


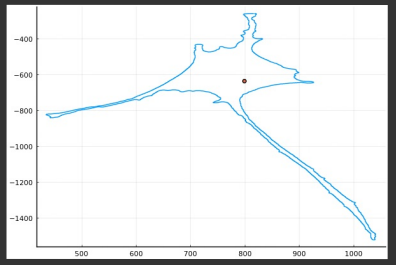


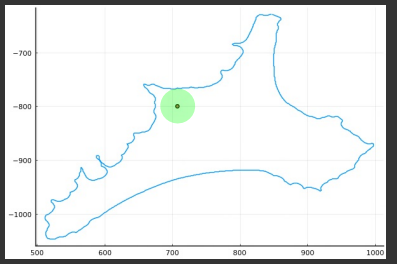
Results - set 1

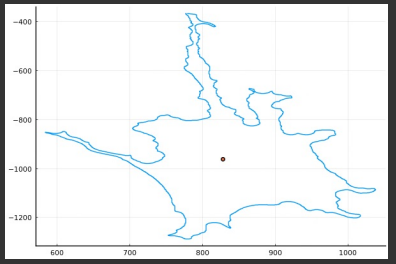


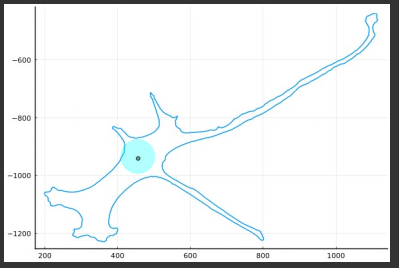




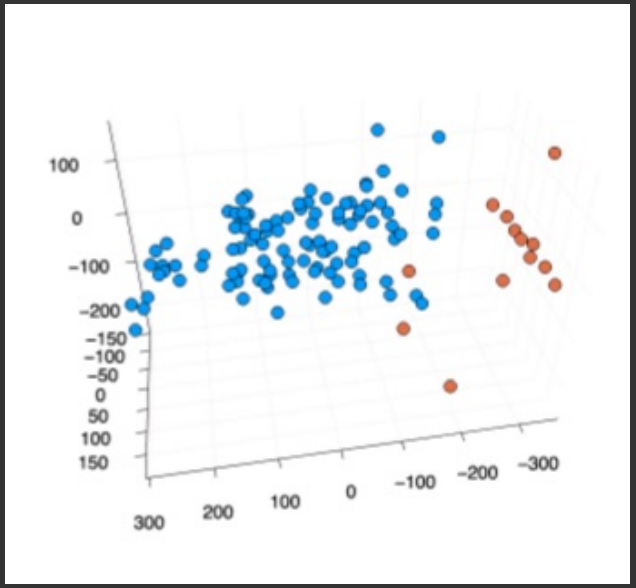


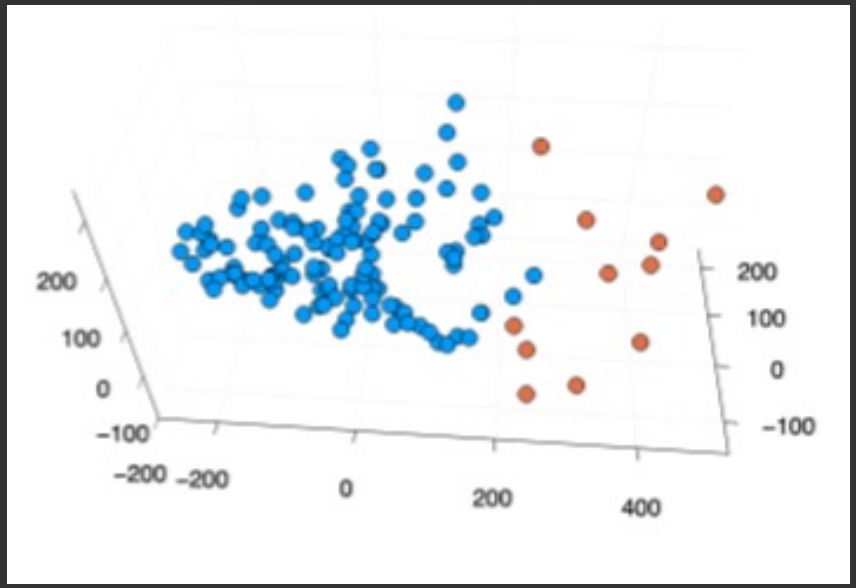




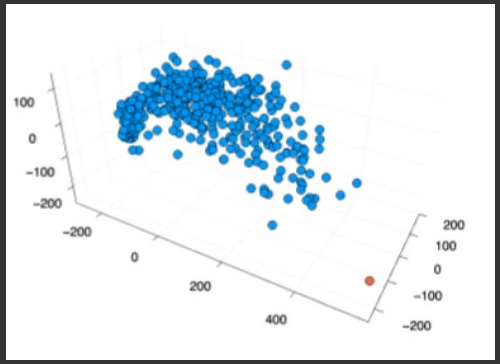


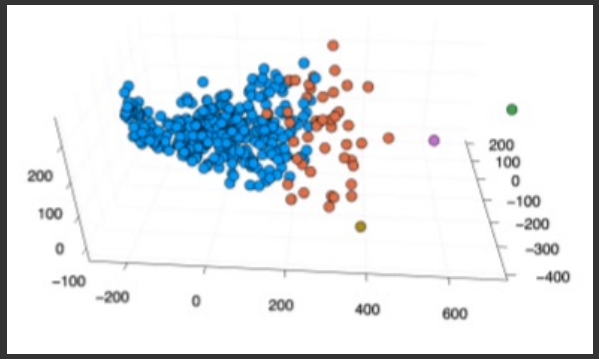
Results - set 2 and set 3

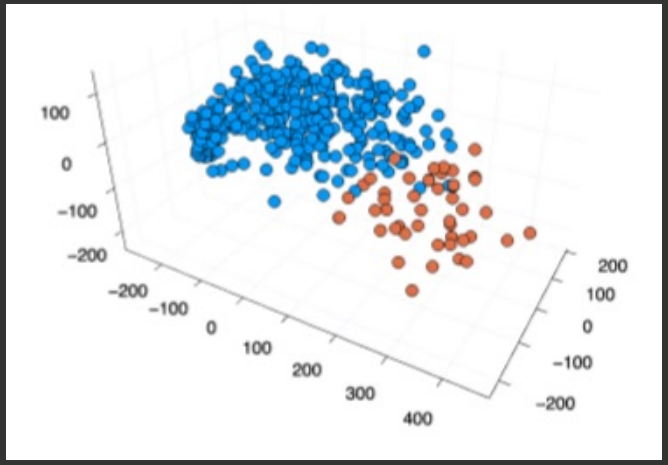




Results - combined set

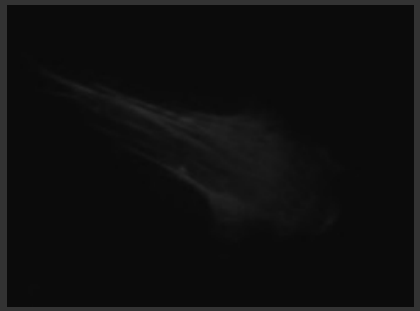
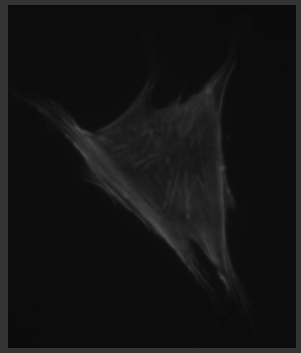
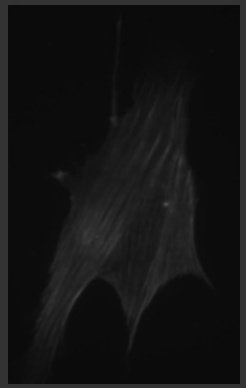
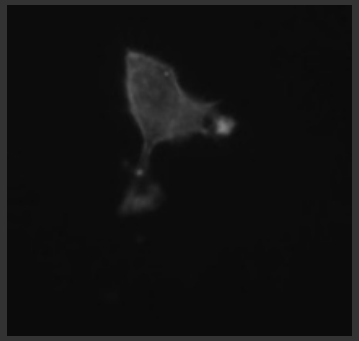
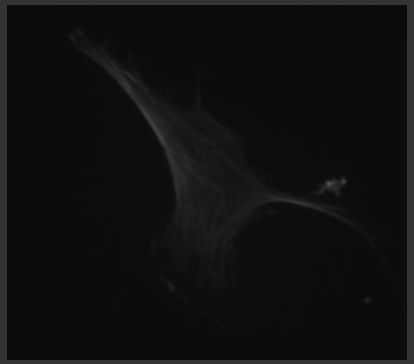
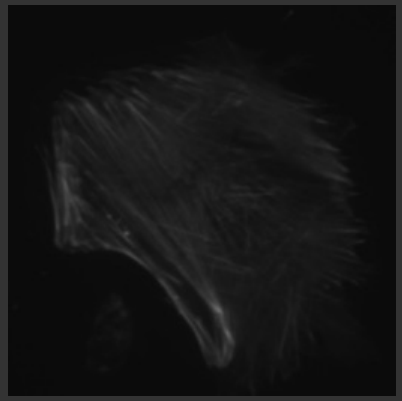






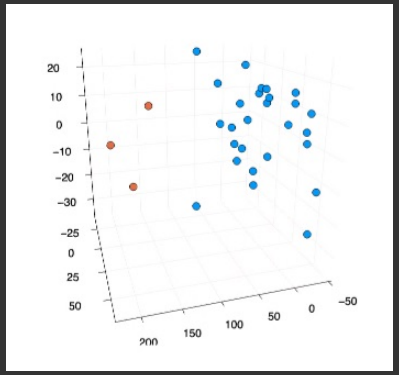
Back to initial problem

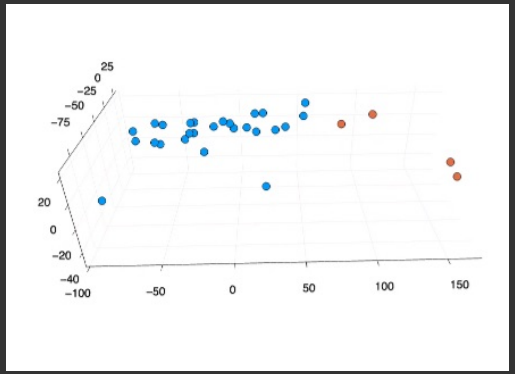
Given populations of human mesenchymal cells cultured in the different environmental conditions (extracellular matrix, treatment, time cultured, staining), can we classify their lineages based on topological and geometric information?

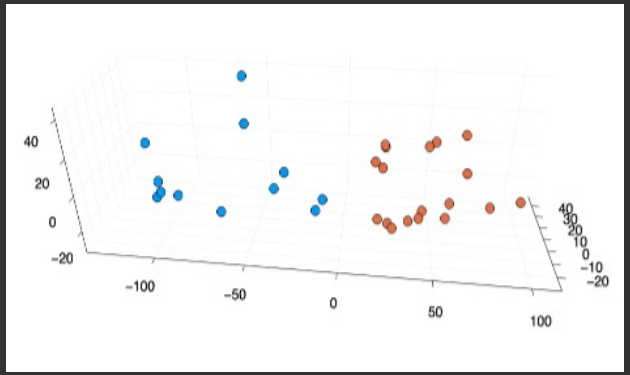







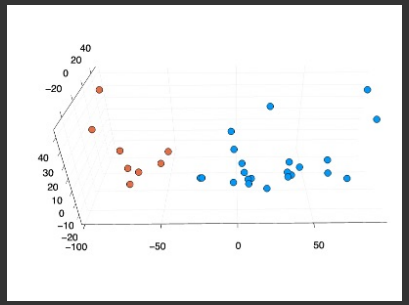
Back to initial problem

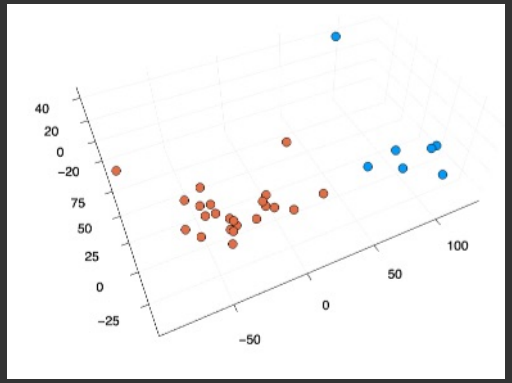
First we need to clean the data, removing any cells which exhibit abnormal morphology.

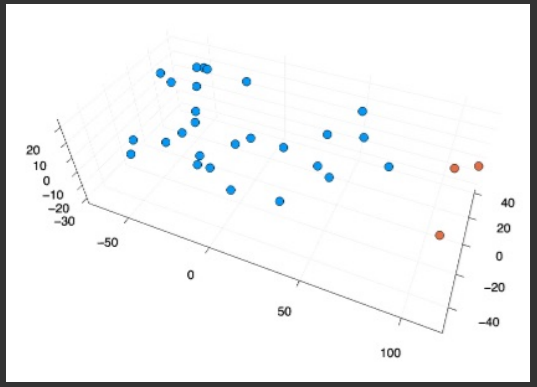












Back to initial problem

The jury is still out, about the success of persistent homology to seperate these lineages.

Next steps are to use summaries including

* inscribed ellipse
* inscribing ellipse
* least square ellipse
* Willmore energy
* persistent homology (including rank functions and APF)
* conformal map classes

and the distances between

* distances between ? ellipse
* Willmore distance
* Frechet distance
* Wasserstein distance between persistence diagrams
* PCA on rank functions
* PCA on APF

to understand the space of lineages.

Persistence of radial function not as discerning as we would have hoped.

Potentially due to ‘global’ outliers.