Optimization of Biopsies Using Tumor Modeling and Clustering Methods

Richa Sehgal

Abstract

Currently, tumor biopsies do not provide doctors with all the possible information that could be used in determining a treatment plan for patients with cancer. The reason for this is that biopsies only remove a very small part of the tumor. However, there are certain “rare” cells (those that have a very large impact on determining key facts like how aggressive the cancer is and how fast it will spread and grow) that exist scattered around the tumor, but because of its size and lack of direction, the biopsy often does not pick these up. This project focused on three types: cancer stem cells (give rise to all cell types in a particular cancer due to self-renewal, lead to metastases, and not killed by common therapeutic treatments), transient amplifying cells (cells in transition between stem cells and differentiated cells), and terminally differentiated cells (“dead” cells, can no longer proliferate). Using a tumor growth model, a dataset consisting of data points (each representing different cells within the tumor) was obtained. After removing transient amplifying cells from the data set (because they are the most in number and therefore ensured a presence in every cluster), a combination matrix was created that was the sum of a scaled Euclidian distance matrix (0-1) and dissimilarity matrix (1 if similar, 0 if dissimilar). Then, using the PROC CLUSTER and TREE procedures in SAS®, clusters of dissimilar cells were found. This was then plotted 3-dimensionally to visualize the location and size of each cluster, seeing which would be most accessible. This is where a surgeon would want to aim during a biopsy in order to get the most accurate representation of the tumor and therefore create the most accurate prognosis and treatment plan. This model can then be applied to tumors of different cancers, sizes, and stages. It’s currently theorized that tumors of the same cancer and stage have the same cell types existing in similar positions relative to tumor size; so, a

comprehensive database of the areas of greatest variability in all tumor types could be compiled if such training data was available.

Background

Malignant tumors contain various types of cells. Knowing what types of cells exist in a tumor and how many of each are there holds great prognostic value. Cancer stem cells (CSC), for example, can therefore give rise to all cell types in a particular cancer due to self-renewal. These CSCs are hypothesized to lead to persisting relapses and metastases. Other rare cells of varying types also exist, and make up approximately 5% of a tumor.

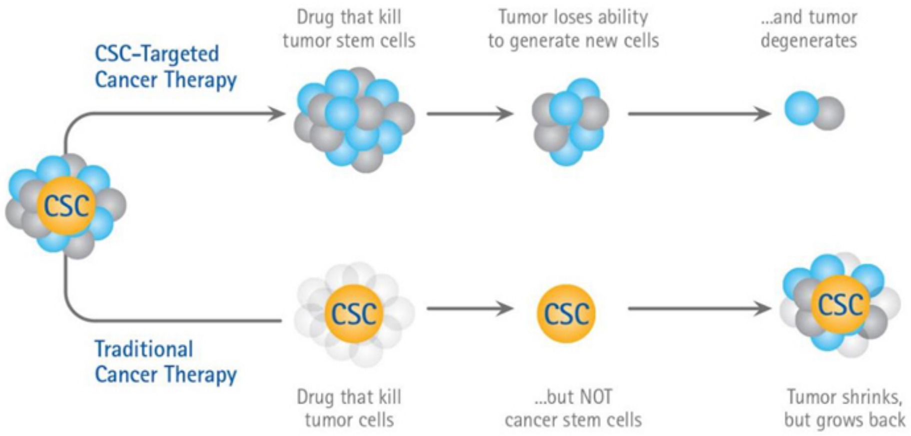
CSCs exist in small amounts in tumors and are often not killed by common chemotherapeutic treatments. With even a small amount of these CSCs remaining after treatment, relapse is highly possible. Therefore, CSCs must be detected early on with a biopsy, so that there is enough time for specific stem cell therapies to be used. Because of their minute existence, these CSCs may not always be detected in biopsies. This is also the case for other rare cells, such as circulating tumor cells, that have an effect on prognosis and treatment.

Transient amplifying cells (TAC) make up the bulk of a tumor and regulate the balance of stem cell usage and tissue generation. They reactivate dormant stem cells to begin self-renewal, which fuels the growth of TACs. TACs are in transition between stem cells and differentiated cells; they arise from stem cells and divide a number of times until they become differentiated.

Terminally differentiated cells (TDC) are cells that have lost the ability to proliferate. They no longer divide, and therefore don’t need to be treated with therapy because they will die off eventually.

A biopsy is an extraction of cells or tissue for examination to discover the presence or extent of a disease. For tumors, a biopsy is most often taken to determine whether or not the tumor is cancerous. Biopsies are usually taken with a needle, extracting a thin cylindrical shape.

Currently, biopsies don't really have any specific means of where exactly to extract cells from the tumor. There are no guidelines for how a tumor should be cut into, or what section and how much of a tumor should be cut into for the greatest variation in cell type. Additionally, because biopsies extract such a small portion of a tumor, it is likely that certain cells will not be extracted, so doctors won't be able to make the most educated decisions about treatment and prognosis. Additionally, it is currently theorized that different types of cells exist in the same relative positions in cancerous tumors.



Differences in growth and treatment of CSCs and non-CSC

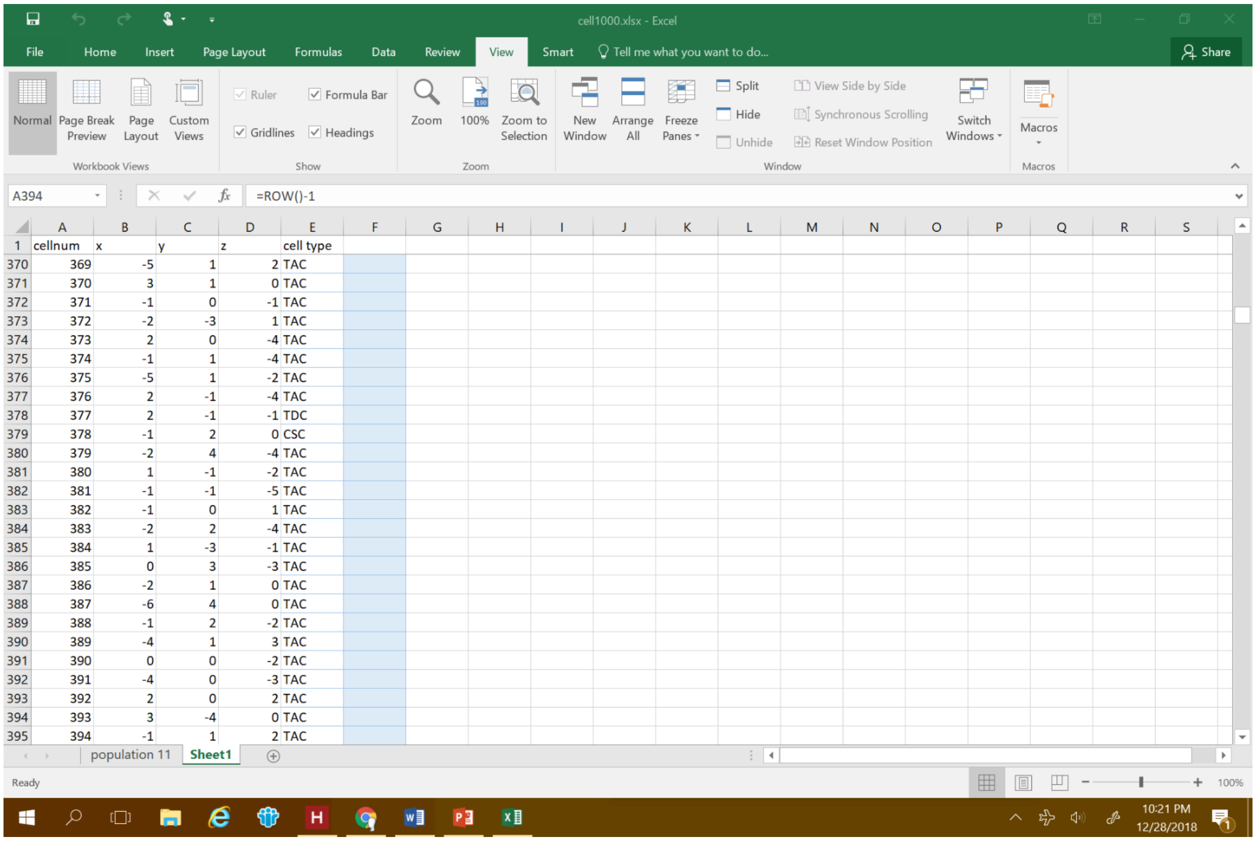
Experiment/Method

Using a tumor growth model called *Tumopp* (created by Iwasaki and Hinnan), I obtained a dataset with with x, y, and z coordinates of each cell, along with the cell's type (CSC, TAC, TDC). The growth model assumes that:

1. a cell occupies a single node in the lattice
2. normal (noncancer) cells are not simulated
3. extracellular matrix surrounding the tumor is ignored
4. the environment is not affected by changes in the configuration of the tumor

The original dataset included about 4.5% cancer stem cells (CSCs), 81.69% transient amplifying cells (TACs), and 13.81% terminally differentiated cells (TDCs).

A sample of the data:



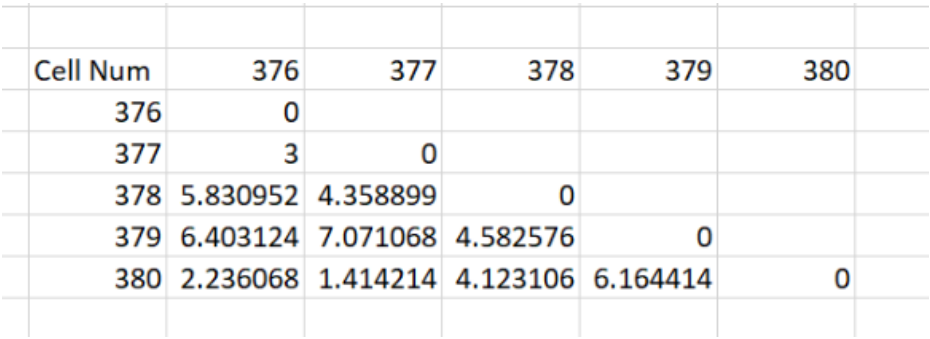
A394 =ROW()-1

	A	B	C	D	E
1	cellnum	x	y	z	cell type
370	369	-5	1	2	TAC
371	370	3	1	0	TAC
372	371	-1	0	-1	TAC
373	372	-2	-3	1	TAC
374	373	2	0	-4	TAC
375	374	-1	1	-4	TAC
376	375	-5	1	-2	TAC
377	376	2	-1	-4	TAC
378	377	2	-1	-1	TDC
379	378	-1	2	0	CSC
380	379	-2	4	-4	TAC
381	380	1	-1	-2	TAC
382	381	-1	-1	-5	TAC
383	382	-1	0	1	TAC
384	383	-2	2	-4	TAC
385	384	1	-3	-1	TAC
386	385	0	3	-3	TAC
387	386	-2	1	0	TAC
388	387	-6	4	0	TAC
389	388	-1	2	-2	TAC
390	389	-4	1	3	TAC
391	390	0	0	-2	TAC
392	391	-4	0	-3	TAC
393	392	2	0	2	TAC
394	393	3	-4	0	TAC
395	394	-1	1	2	TAC

The data above was then converted into a distance matrix by taking the Euclidian distance between each pair.

$$Distance = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

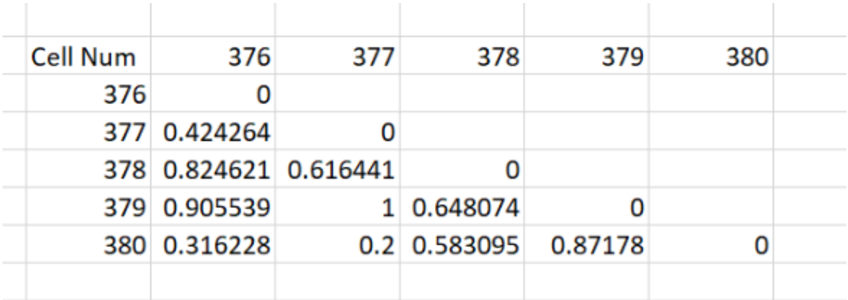
A sample of the distance matrix:



Cell Num	376	377	378	379	380
376	0				
377	3	0			
378	5.830952	4.358899	0		
379	6.403124	7.071068	4.582576	0	
380	2.236068	1.414214	4.123106	6.164414	0

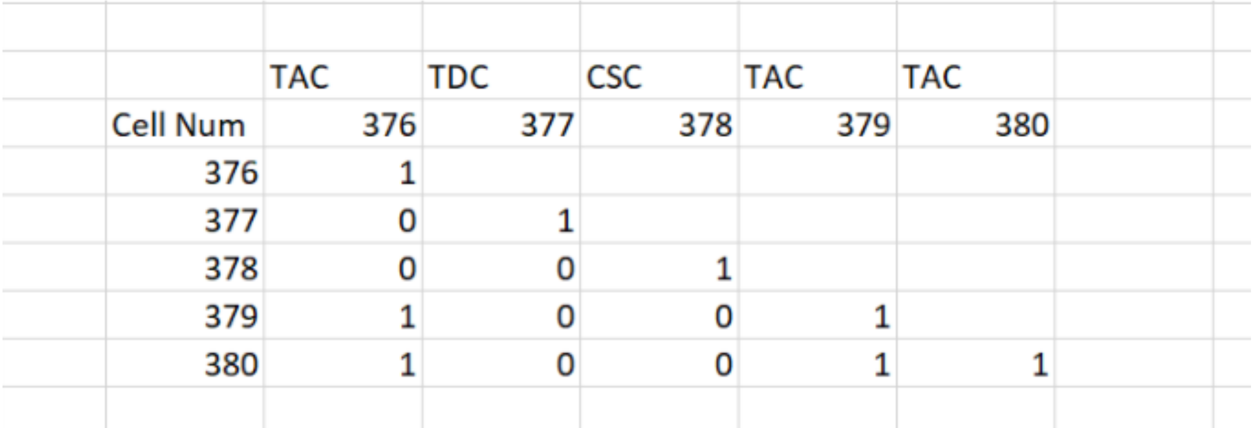
The distance was then scaled so that it would range from 0 and 1.

$$Scaled\ distance = \frac{distance}{distance_{MAX}}$$



Cell Num	376	377	378	379	380
376	0				
377	0.424264	0			
378	0.824621	0.616441	0		
379	0.905539	1	0.648074	0	
380	0.316228	0.2	0.583095	0.87178	0

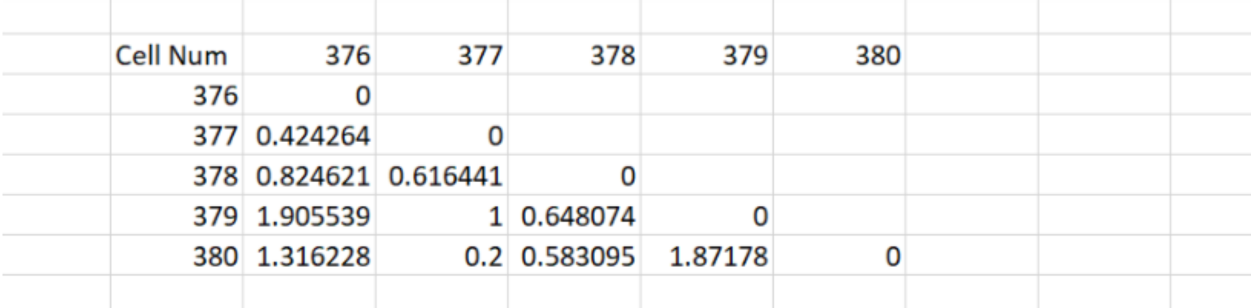
A dissimilarity matrix was then generated by assigning a "1" if the cells were similar and a "0" if they were dissimilar.



	TAC	TDC	CSC	TAC	TAC
Cell Num	376	377	378	379	380
376	1				
377	0	1			
378	0	0	1		
379	1	0	0	1	
380	1	0	0	1	1

Dissimilar cells that are close together are of interest. So, a combination matrix was generated by adding the scaled distance matrix and the dissimilarity matrix.

The close dissimilar cells will have the lowest scores/values.

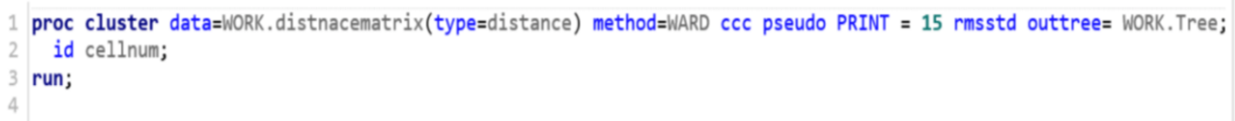


Cell Num	376	377	378	379	380
376	0				
377	0.424264	0			
378	0.824621	0.616441	0		
379	1.905539	1	0.648074	0	
380	1.316228	0.2	0.583095	1.87178	0

TACs were then filtered out from the data for the following reasons:

1. They make up bulk of the tumor, and are therefore essentially guaranteed to exist in each cluster.
2. SAS university edition was being used, so there were limitations on data size.

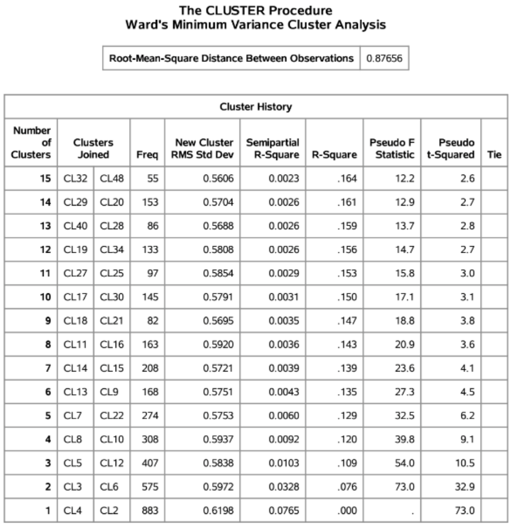
The "proc cluster" function in SAS was then used to create the clusters.



```
proc cluster data=WORK.distnacematrix(type=distance) method=WARD ccc pseudo PRINT = 15 rmsstd outtree= WORK.Tree;
  id cellnum;
run;
```

The WARD method was used for this function because it is biased towards producing clusters with about same number of observations.

Proc cluster displays the table of eigenvalues of the covariance matrix, and these eigenvalues are used in the computation of the cubic clustering criterion. The output of the proc-cluster function gives insight on the number of clusters.



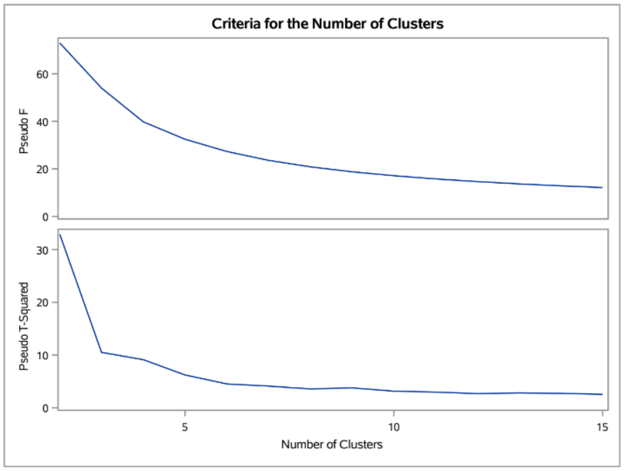
The CLUSTER Procedure
Ward's Minimum Variance Cluster Analysis

Root-Mean-Square Distance Between Observations	0.87656

Cluster History

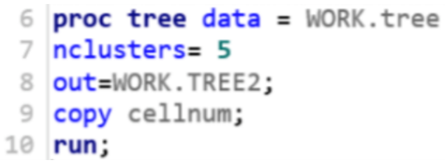
Number of Clusters	Clusters Joined		Freq	New Cluster RMS Std Dev	Semipartial R-Square	R-Square	Pseudo F Statistic	Pseudo t-Squared	Tie
15	CL32	CL48	55	0.5606	0.0023	.164	12.2	2.6	
14	CL29	CL20	153	0.5704	0.0026	.161	12.9	2.7	
13	CL40	CL28	86	0.5688	0.0026	.159	13.7	2.8	
12	CL19	CL34	133	0.5808	0.0026	.156	14.7	2.7	
11	CL27	CL25	97	0.5854	0.0029	.153	15.8	3.0	
10	CL17	CL30	145	0.5791	0.0031	.150	17.1	3.1	
9	CL18	CL21	82	0.5695	0.0035	.147	18.8	3.8	
8	CL11	CL16	163	0.5920	0.0036	.143	20.9	3.6	
7	CL14	CL15	208	0.5721	0.0039	.139	23.6	4.1	
6	CL13	CL9	168	0.5751	0.0043	.135	27.3	4.5	
5	CL7	CL22	274	0.5753	0.0060	.129	32.5	6.2	
4	CL8	CL10	308	0.5937	0.0092	.120	39.8	9.1	
3	CL5	CL12	407	0.5838	0.0103	.109	54.0	10.5	
2	CL3	CL6	575	0.5972	0.0328	.076	73.0	32.9	
1	CL4	CL2	883	0.6198	0.0765	.000	.	73.0	

The CLUSTER Procedure
Ward's Minimum Variance Cluster Analysis



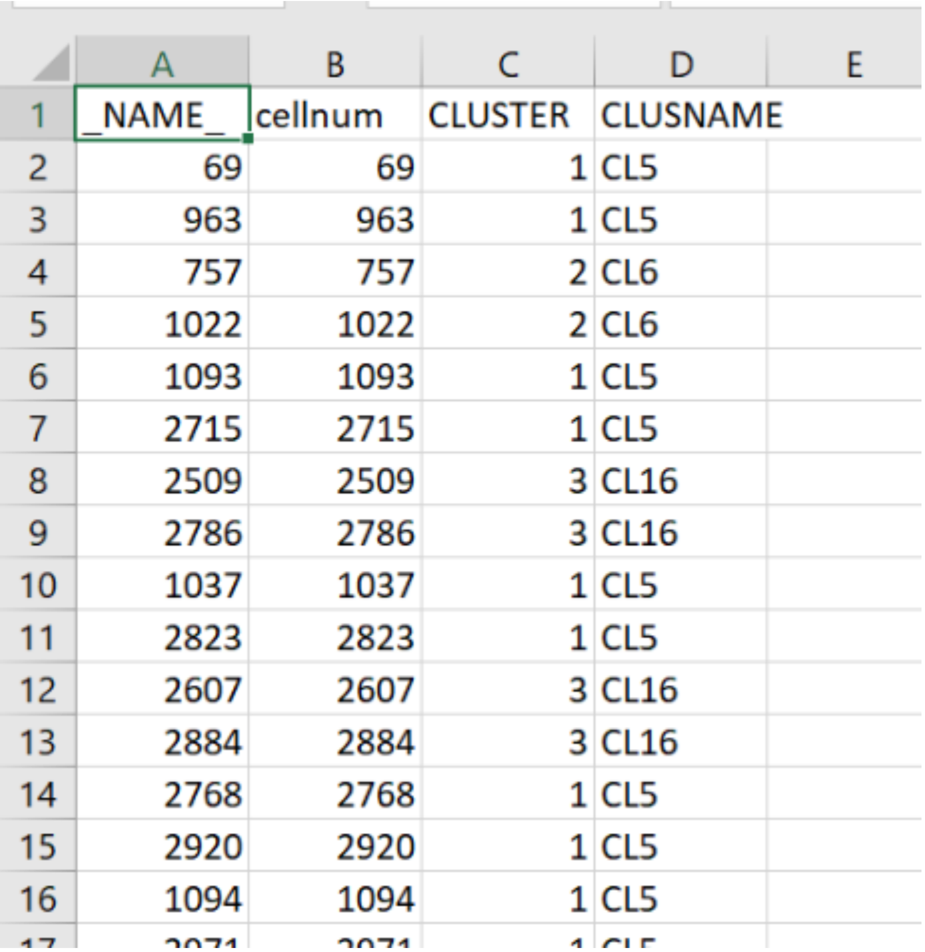
The greatest change in Pseudo T-Squared corresponds to the optimum number of clusters; in this case, that is 5 clusters.

The “tree” procedure was then used to produce a tree diagram of the clusters. Here, the number of clusters inputted was 5.



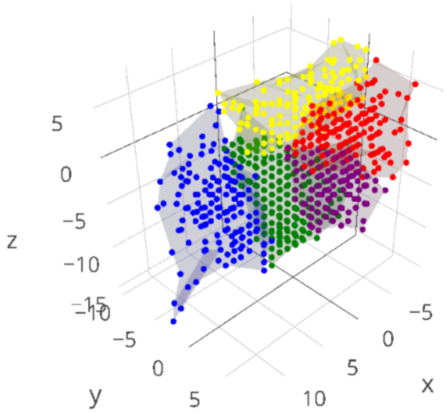
```
proc tree data = WORK.tree
nclusters= 5
out=WORK.TREE2;
copy cellnum;
run;
```

The tree was exported into a csv file that looks like this:



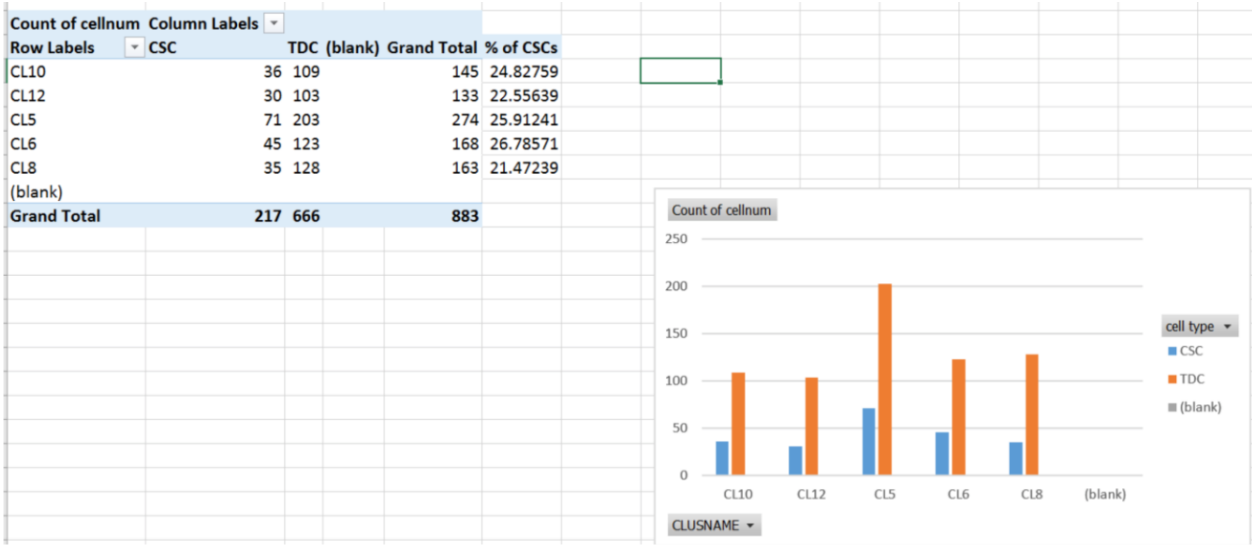
	A	B	C	D	E
1	_NAME_	cellnum	CLUSTER	CLUSNAME	
2	69	69	1	CL5	
3	963	963	1	CL5	
4	757	757	2	CL6	
5	1022	1022	2	CL6	
6	1093	1093	1	CL5	
7	2715	2715	1	CL5	
8	2509	2509	3	CL16	
9	2786	2786	3	CL16	
10	1037	1037	1	CL5	
11	2823	2823	1	CL5	
12	2607	2607	3	CL16	
13	2884	2884	3	CL16	
14	2768	2768	1	CL5	
15	2920	2920	1	CL5	
16	1094	1094	1	CL5	
17	2071	2071	1	CL5	

Using Plotly in Python, the (x, y, z) coordinates of every cell along with its cluster were plotted, giving this output:



Analyzing the Clusters:

The TREE output of Proc Cluster was also used to analyze each cluster. As mentioned previously, the WARD method is inclined towards producing balanced clusters. In this sample, the ratio of CSCs to TDCs was roughly 1:4. Thus, cluster 8 is most representative of the tumor, but cluster 6 contains the greatest variability.



Count of cellnum	Column Labels ▾				
Row Labels ▾	CSC	TDC	(blank)	Grand Total	% of CSCs
CL10	36	109		145	24.82759
CL12	30	103		133	22.55639
CL5	71	203		274	25.91241
CL6	45	123		168	26.78571
CL8	35	128		163	21.47239
(blank)					
Grand Total	**217**	**666**		**883**	

Conclusion

The project showed that clustering by dissimilarity can be used to identify areas of greatest variability. The WARD method of the “proc cluster” function in SAS was used effectively to create and visualize the clusters. The clustering was done using combination of Euclidian distance and variability. Simple addition of scaled distance and dissimilarity (0-similar, 1-dissimilar) was used.

The project can be performed on different tumor models that are already being studied. Different weights can also be given to distance and dissimilarity. This analysis was done on 3 types of cells and a Boolean (0 or 1) measure of dissimilarity. The approach can be expanded to cells in different stages of their life cycle with different levels of similarity (e.g. – 0.03 is more similar that 0.004).

Also different clustering algorithms can be experimented with to see the effectiveness (distance between clusters and mean distance between dissimilar cells). As tumor models become more and more popular, data from different stages of a type of tumor can be analyzed to see if patterns can be produced and if machine learning can be used to predict cross sections of greatest variability based on the type and stage of the tumor.

Bibliography

Hsu, Ya-Chieh, et al. "Transit-Amplifying Cells Orchestrate Stem Cell Activity and Tissue Regeneration." Cell, vol. 157, no. 4, 8 May 2014, pp. 935–949. https://doi.org/10.1016/j.cell.2014.02.057.

Iwasaki WM, Innan H (2017) Simulation framework for generating intratumor heterogeneity patterns in a cancer cell population. PLoS ONE 12(9): e0184229. https://doi.org/10.1371/journal.pone.0184229

Ludwig Center, Ludwig Center. "The Stem Cell Theory of Cancer." Stanford Medicine, Stanford University, 2016, med.stanford.edu/ludwigcenter/overview/theory.html.

Neeley, Cindy. "The Use of Spheroids in Cancer Research." Thermo Fisher Scientific, Thermo Fisher Scientific, 25 Apr. 2016, www.thermofisher.com/blog/cellculture/the-use-of-spheroids-in-cancer-research/.

Sacco A, Pajalunga D, Latella L, et al. Cell Cycle Reactivation in Skeletal Muscle and Other Terminally Differentiated Cells. In: Madame Curie Bioscience Database [Internet]. Austin (TX): Landes Bioscience; 2000-2013. Available from: https://www.ncbi.nlm.nih.gov/books/NBK6180/

Slack, Jonathan M.W. "Transit Amplifying Cell." Encyclopædia Britannica, Encyclopædia Britannica, Inc., www.britannica.com/science/transit-amplifying-cell.

Zhou, Mingyuan, et al. "Transcriptional Profiling of Enriched Populations of Stem CellsVersusTransient Amplifying Cells." Journal of Biological Chemistry, vol. 281, no. 28, 4 May 2006, pp. 19600–19609., doi:10.1074/jbc.m600777200.