# Gene Expression Data Mining of ARCHS4

**Aleksander Braksator**
Department of Statistics
University of Washington
alexioso@uw.edu

**Qinai Xu**
Department of Statistics
University of Washington
xqn07899@uw.edu

**Haobo Zhang**
Department of Electrical Engineering
University of Washington
haoboz@uw.edu

## Introduction

The ARCHS4 dataset contains gene counts for humans and mice from multiple authoritative sources including Gene Expression OmniBus (GEO) and Sequence Read Archive (SRA). The file format allows us to programmatically access the matrix entries based on column and row indices. We use Python and R to process and analyze the dataset. Considering the large size of ARCHS4 and our personal interest, we are going to focus mainly on the human gene expression data. We examine several clustering methods on the dataset.

Our work could be mainly divided into three stages. In the first stage, we present our results on one-way clustering to find which clustering methods work effectively on different subsets of the data. In the second stage, we discuss results from two-way clustering on genes and samples simultaneously. And in the third stage, we summarize what we find in stage 1 and 2 about our data, use the findings to classify samples from our database and tackle a real-world problem using supervised machine learning.

Phase 1 consists of exploring and analyzing the performance of 1-way clustering techniques on different subsets of samples by gene arrays and 1-way clustering on genes by sample arrays. For both scenarios, we attempt to develop a non-trivial means of reducing the number of genes considered to improve clustering and minimize the number of observations to analyze in the latter scenario. In addition, we have access to data from most human tissue types and significant cell lines corresponding to known diseased tissue such as cancer. We plan to assess our clustering methods based on the purity of sample labels for a multitude of different human tissue. We had access to three subsets of comparable samples: healthy ovarian tissue vs. SKOV3 and ES2 cell line (ovarian cancer cell lines), primary liver cells vs bulk liver tissue vs HEPG2 (liver cancer cell line), and pancreatic islets (alpha and beta cells) vs PANC1 cell line (pancreatic cancer). Phase 2 consists of developing a two-way clustering algorithm along with gene shaving to explore important gene patterns and perform cluster discovery on samples. We will be looking for novel clusters in an unsupervised fashion in hopes to discover novel combinations of similarly behaving genes and samples that have a high purity of sample labels. As mentioned, we have sample data (with metadata) of healthy samples vs. cell lines associated with some disease (cancer). We hope to use our results from Phase 1 and Phase 2 on each tissue type (ovarian, liver, pancreas) for a prediction scenario versus their corresponding cell lines via supervised learning on the gene expression arrays to see if we can predict healthy tissue vs cancerous tissue based on the gene expression array.

Preprint. Under review.

## Review of Prior Work

We found 5 different materials that relate to our project including three recently published papers, a bioinformatics lecture, and one paper from 2000. It is helpful to discuss (4) first to get an understanding of the ARCHS4 data background and what it can be used for. This paper was written by the authors of the ARCHS4 website which we have accessed our data from. They developed a cost-effect solution to storing data via AWS and a clever way to share this data via R scripts so that even non-programmatically inclined users can locally download the exact data they queried for without having to send over a file larger than a small R file. In addition to the ease of use, the data ARCHS4 has combined from several reputable gene expression data repositories including GEO and SRA. These are both repositories of sequencing data where GEO focuses on gene expression arrays. Both of these are public repositories, but ARCHS4 was able to merge the data from human and mouse samples and use the H5 file format to store large matrices of gene expression counts along with metadata for each of the rows and columns. Rows of this matrix correspond to a single gene, columns correspond to a sample (a functional product of gene expressions such as RNA or a protein). We decided to focus on only the human sample matrix for this project.
Several sources warn about the batch effect of analyzing these samples (4,6). This is one major disadvantage of having such a large, diverse database of gene expression data such as ARCHS4. These methods may not apply to the entire ARCHS4 dataset due to its diversity of sources, however, it may come in handy when filtered down to specific, related data samples from the same laboratory. As a result of this, outliers have been an issue when we deal with samples that come from a series (experiment) that has very few samples. These tend to cluster away from samples in series with a large number of samples even if they are from the same tissue.
Several sources also prefer using hierarchical clustering over k-means clustering (1,2,5), and in our first batch of results, we tend to see hierarchical clustering producing more pure clusters on this dataset.

## Data Collection Process

The data from the ARCHS4 repository is novel in the sense that it combines independent sample and gene expression arrays from the Gene Expression Omnibus repository into one large gene expression matrix with corresponding metadata for each row (gene) and column (sample). We have downloaded their entire 7.5 GB human matrix in H5 format and have been using their website to download R scripts for filtering down to specific tissues and cell lines of interest. From these R scripts, we are able to obtain a submatrix of gene expressions from the original human matrix corresponding to our samples of interest. Before clustering, we merge together datasets we want to analyze concurrently. The three subsets of data we focus on in this project are ovaries (healthy tissue vs. SKOV3), pancreas (healthy pancreatic islet vs PANC1), and liver (healthy tissue vs. HEPG2).

## Initial Findings and Summary Statistics for Dataset

The whole human dataset consists of 238,522 samples and 35,238 genes. Of the commonly collected samples, 5570 are fibroblast, 4912 are breast, 4032 are lung, 2271 are skin, 1884 are colon, 806 are myoblast, 207 are microglia, 1156 are adipose, 1482 are kidney, 242 are skeletal muscle, 4752 are pancreatic islet, 1757 are midbrain, 3730 are neuron, and 1810 are macrophage.

The min value for gene expression is 0 (no expression by the gene) and we have seen values as high as 308128. Average values tend to vary widely among the samples, and even the author of the ARCHS4 website warns about comparing different batches of samples with each other due to a significant batch effect present in the data (4). The matrices are very sparse. In the ovarian subsample, for example, there is a 0.9992 sparsity ratio. The sparsity is similar across all samples, and we have not encountered a sample with less than 0.5 sparsity ratio.

**1-Way Clustering of Samples on Gene Arrays**

**Mathematical Background**

For each of the three subsets of data we are analyzing, we obtain a $n \times p$ gene expression matrix X with n sample and p genes. When treating genes as features, we normalize each gene expression array so that the vector has a mean 0 and magnitude 1. Each sample (observation) has associated metadata which allows us to analyze each cluster in a supervised fashion using purity.

$$\frac{(\text{col} - \bar{\text{col}})}{\|\text{col}\|_2} \text{, for each vector col in data}$$

The Pearson Correlation coefficient is ideal for comparing similarity of genes. Here, it is shown that using Euclidean distance is equivalent to using gene correlation when our data has mean 0 and standard deviation 1 (using normalization formula above).

If you preprocess your data ($n$ observations, $p$ features) such that each feature has $\mu = 0$ and $\sigma = 1$ (which disallows constant features), then correlation reduces to cosine:

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right]}{\sigma_X \sigma_Y} = \mathbb{E}[XY] = \frac{1}{n} \langle X, Y \rangle$$

Under the same conditions, squared Euclidean distance also reduces to cosine:

$$\begin{aligned} d^2_{\text{Euclid}}(X, Y) = \sum (X_i - Y_i)^2 &= \sum X_i^2 + \sum Y_i^2 - 2 \sum X_i Y_i \\ &= 2n - 2 \langle X, Y \rangle = 2n \left[1 - \text{Corr}(X, Y)\right] \end{aligned}$$

We use the Cophenet Score as a baseline measurement for hierarchical clustering to see if the clustering is worth exploring further given a set of hyperparameters and data. It is a measure of how faithfully a dendrogram preserves the pairwise distances between the original unmodeled data points. The closer the score is to 1, the better the clustering. Often, if the score is below 0.9, we discarded the method and went on to the next set of hyperparameters. The score is defined as:

$$c = \frac{\sum_{i<j}(x(i,j) - \bar{x})(t(i,j) - \bar{t})}{\sqrt{[\sum_{i<j}(x(i,j) - \bar{x})^2][\sum_{i<j}(t(i,j) - \bar{t})^2]}}.$$

Where $x(i, j)$ is the ordinary Euclidean distance between the we and j observations and $t(i, j)$ is the vertical distance in the dendrogram between the model points $T_i$ and $T_j$.

Purity is the supervised approach we are using to assess our clustering results. For analyzation purposes, we assess the purity for three variables in our metadata: RNA molecule, series ID, and sample label. By far, the sample label is the most important purity metric, but the molecule and series ID were helpful in order to analyze why some of our clustering algorithms did not work as intended. For example, sometimes the clustering algorithm would create a separate cluster of Ovary samples with 100% label purity because they all came from the same series/experiment (100% series purity). The calculation for the sample label purity would be the maximum proportion of sample labels associated with the data points present in a given cluster.

Our final method of cluster evaluation is to use an unsupervised method of the Bootstrap algorithm process using the Jaccard distance between cluster assignments. The algorithm is outlined in the next section. The Jaccard similarity between two sets of clusters of points A and B is:

$$\frac{\|A \cap B\|}{\|A \cup B\|}$$

We use k-means and hierarchical clustering as our base clustering algorithms for this dataset. In essence we treat each sample as an observation with a feature array consisting of about 35000 genes. For hierarchical clustering, we noticed via the use of a cophenet score that the best hyperparameters to use for hierarchical clustering were always average linkage with Euclidean distance between each sample. From the resulting dendrogram, a proper value of k was chosen.

### Ovarian and SKOV3 Clustering

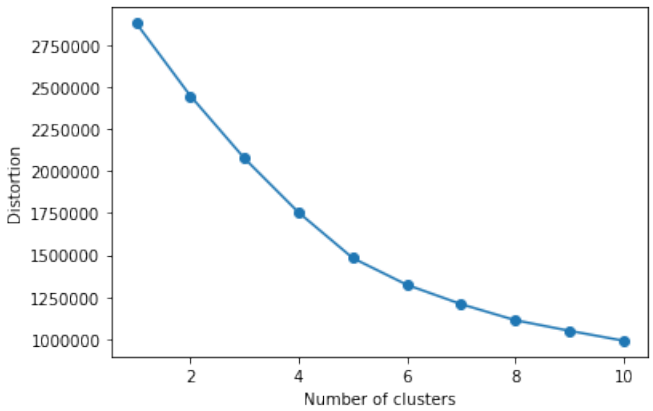
First, we chose to focus on comparing the ovarian samples with an ovarian cancer cell line we had access to, SKOV3. Once again, this matrix has a high sparsity ratio of 0.99968. There are 33 samples in SKOV3 and 99 samples in Ovarian, which makes the worst case ratio of the label purity 0.75 for Ovarian and 0.25 for SKOV3.

Using the raw data did not give us high purity in all of the clusters, however, when the data is normalized, using the average linkage method with Euclidean distance and choosing K=12 based on the dendrogram:

| Cluster ID | Cluster Size | Dominant Molecule | Molecule Purity | Dominant Series | Series Purity | Dominant Label | Label Purity |
|---|---|---|---|---|---|---|---|
| 1 | 2 | total RNA | 1 | GSE61474 | 1 | Ovary | 1 |
| 2 | 9 | total RNA | 1 | GSE55946 | 1 | SKOV3 | 1 |
| 3 | 1 | total RNA | 1 | GSE87965 | 1 | Ovary | 1 |
| 4 | 21 | total RNA | 1 | GSE126797 | 0.429 | SKOV3 | 0.857 |
| 5 | 4 | total RNA | 1 | GSE106848 | 1 | SKOV3 | 1 |
| 6 | 11 | total RNA | 1 | SE63392-GSE6339 | 0.818 | Ovary | 1 |
| 7 | 5 | total RNA | 1 | E112855-GSE1128 | 1 | Ovary | 1 |
| 8 | 3 | total RNA | 1 | GSE137237 | 1 | Ovary | 1 |
| 9 | 56 | total RNA | 0.964 | GSE134375 | 0.429 | Ovary | 0.964 |
| 10 | 18 | polyA RNA | 1 | GSE118127 | 1 | Ovary | 1 |
| 11 | 1 | polyA RNA | 1 | GSE118127 | 1 | Ovary | 1 |
| 12 | 1 | polyA RNA | 1 | GSE30017 | 1 | Ovary | 1 |

We have a nicer, but near perfect, separation between the SKOV3 and Ovary samples. The issue is we needed k=12 to separate out the outliers before we could see such separation. A solution to this would be to remove the outliers which would not be helpful in a real-life scenario where we must account for all possible samples. Otherwise, we can see a lot of the separation is due to series (which tends to happen when we normalize the data). However, even clusters 4 and 9 with low series purity have high label purity which is exactly what we are trying to see from successful clustering results. This clustering method performs exceptionally well under bootstrap validation as well with none of the clusters dissolving.

We also perform K-means clustering on this dataset. We first use the elbow method to help us find a good k. Specifically, we plot the distortion of the clustering result versus the number of k. The k corresponding to the "elbow" is likely a proper choice of k for the dataset. The plot for this case is shown below.




We can notice that the distortion is really high. It is because we have such a high dimensional feature space. For this dataset specifically, we have 35238 gene arrays as features. Even though the elbow point is not very clear in the plot, it is likely located around k=5. We then proceed to perform K-means with k = 5 and the result is as follows:

| Cluster ID | Cluster Size | Dominant Molecule | Molecule Purity | Dominant Series | Series Purity | Dominant Label | Label Purity |
|---|---|---|---|---|---|---|---|
| 0 | 18 | polyA RNA | 1 | GSE118127 | 1 | Ovary | 1 |
| 1 | 21 | total RNA | 1 | GSE126797 | 0.429 | SKOV3 | 0.857 |
| 2 | 64 | total RNA | 0.938 | GSE134375 | 0.375 | Ovary | 0.969 |
| 3 | 12 | total RNA | 1 | GSE55946 | 0.75 | SKOV3 | 0.75 |
| 4 | 17 | total RNA | 1 | GSE63392-GSE63394 | 0.529 | Ovary | 0.765 |

The result is promising. Except cluster 0 and cluster 4, all clusters have low series purity and high label purity compared to the null hypothesis. We then try to reduce the number of clusters to k = 3 to see if it still gives us a good result, and the answer is positive as shown:

| Cluster ID | Cluster Size | Dominant Molecule | Molecule Purity | Dominant Series | Series Purity | Dominant Label | Label Purity |
|---|---|---|---|---|---|---|---|
| 0 | 32 | polyA RNA | 0.594 | GSE118127 | 0.562 | Ovary | 0.719 |
| 1 | 79 | total RNA | 0.962 | GSE134375 | 0.304 | Ovary | 0.924 |
| 2 | 21 | total RNA | 1 | GSE126797 | 0.429 | SKOV3 | 0.857 |

It is cleaner than the result of k = 5. Cluster 2 and 3 have low series purity and high label purity. It combines cluster 1 in the k = 5 result with other clusters that have SKOV3 as the dominant label so we now get rid of the cluster with a series purity of 1. Cluster 1 has a lower label purity compared to the null hypothesis. It could be the result of having a relatively even distributed polyA RNA molecule and total RNA molecule. The difference between those two types of sampling depending on the molecule could influence the clustering results.

Next, we subset the gene feature space to only include genes mentioned in this paper about ovarian cancer (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4449870/) BRCA1, BRCA2, CHEK2, RAD51, BRIP1, and PALB2. Here is the resulting purity table for average euclidean k=3 on the raw data with hierarchical clustering:

| Cluster ID | Cluster Size | Dominant Molecule | Molecule Purity | Dominant Series | Series Purity | Dominant Label | Label Purity |
|---|---|---|---|---|---|---|---|
| 1 | 13 | total RNA | 0.846 | GSE104295 | 0.462 | SKOV3 | 0.769 |
| 2 | 118 | total RNA | 0.839 | GSE134375 | 0.203 | Ovary | 0.805 |
| 3 | 1 | polyA RNA | 1 | GSE118127 | 1 | Ovary | 1 |

Here we see a nice separation of the data into two main clusters and one outlier cluster. The purities are slightly above average which shows some promise that these genes are essential indicators of determining whether ovarian tissue is cancerous or not. This shows promise to explore later on in Phase 3 supervised learning.

However, K-means clustering does not perform as good as hierarchical clustering. After choosing k = 3 from the elbow method, we got the clustering result:

| Cluster ID | Cluster Size | Dominant Molecule | Molecule Purity | Dominant Series | Series Purity | Dominant Label | Label Purity |
|---|---|---|---|---|---|---|---|
| 0 | 29 | total RNA | 0.966 | GSE134375 | 0.31 | Ovary | 0.552 |
| 1 | 91 | total RNA | 0.769 | GSE118127 | 0.209 | Ovary | 0.846 |
| 2 | 12 | total RNA | 1 | GSE104295 | 0.5 | Ovary | 0.5 |

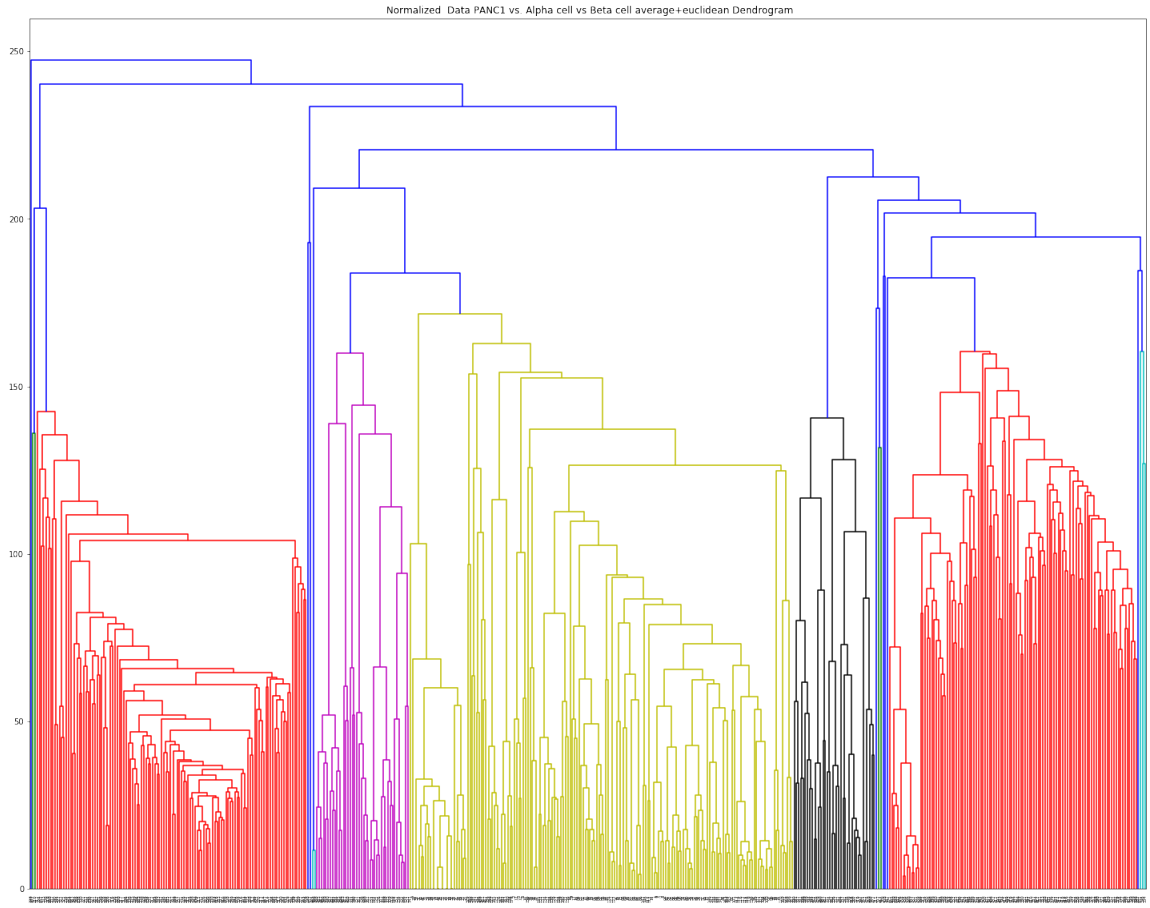
The label purity of cluster 0 and 2 falls under the null hypothesis. It is not an interesting result at all.

**Pancreatic Islet Clustering**

Here we analyze the hierarchical clustering method on three types of Pancreatic Islet cells: alpha cells, beta cells, and cells from the PANC1 cell line, a well-known pancreatic cancer cell line.

There are 142 samples in Alpha cells, 174 samples in beta cells, and 187 samples in the PANC1 cell line, which makes the worst case scenario ratios of the label purity 0.28 for alpha cells, 0.35 for beta cells, and 0.37 for the PANC1 cells.

The most successful of the initial hierarchical clustering combinations was using average linkage, Euclidean distance, and normalized data. Here is the dendrogram and purity table:




| Cluster ID | Cluster Size | Dominant Molecule | Molecule Purity | Dominant Series | Series Purity | Dominant Label | Label Purity |
|---|---|---|---|---|---|---|---|
| 1 | 124 | total RNA | 1 | GSE97655 | 0.919 | ALPHA | 0.984 |
| 2 | 2 | total RNA | 1 | GSE97655 | 1 | ALPHA | 0.5 |
| 3 | 217 | total RNA | 0.765 | GSE106290 | 0.111 | PANC1 | 0.862 |
| 4 | 159 | total RNA | 1 | GSE97655 | 0.673 | beta | 0.893 |
| 5 | 1 | total RNA | 1 | GSE67543 | 1 | ALPHA | 1 |

Aside from the cluster 2 and 5 which is an outlier cluster, the other clusters do a nice job partitioning the data into the three sample types with high purity (>0.85) and combining similar samples from different series. The two outliers in cluster 2 come from a large series of pancreatic islet cells from East Asian samples, a series which contains the majority of Alpha and Beta cells. In fact, this series makes up the majority of Clusters 1 and 4 which have high purity along with successfully combining with other series with the same label (label purity > series purity). This experiment shows great promise for supervised learning; however limited implications may come as a result. Upon further investigation, the PANC1 cell line comes from the ducts of a pancreas, whereas the alpha and beta cells come from pancreatic islets just outside these ducts. The difference in the anatomical sources of these samples may be the cause of the pure clusterings and not the fact that PANC1 cells are cancerous. We are unable to find data on the ARCHS4 repository corresponding to healthy pancreatic ductal tissue, and likewise, there is no cell line for pancreatic cancer from the pancreatic islets. However, the nice separation of alpha and beta cells shows promising results for the effective clustering of pancreatic samples.

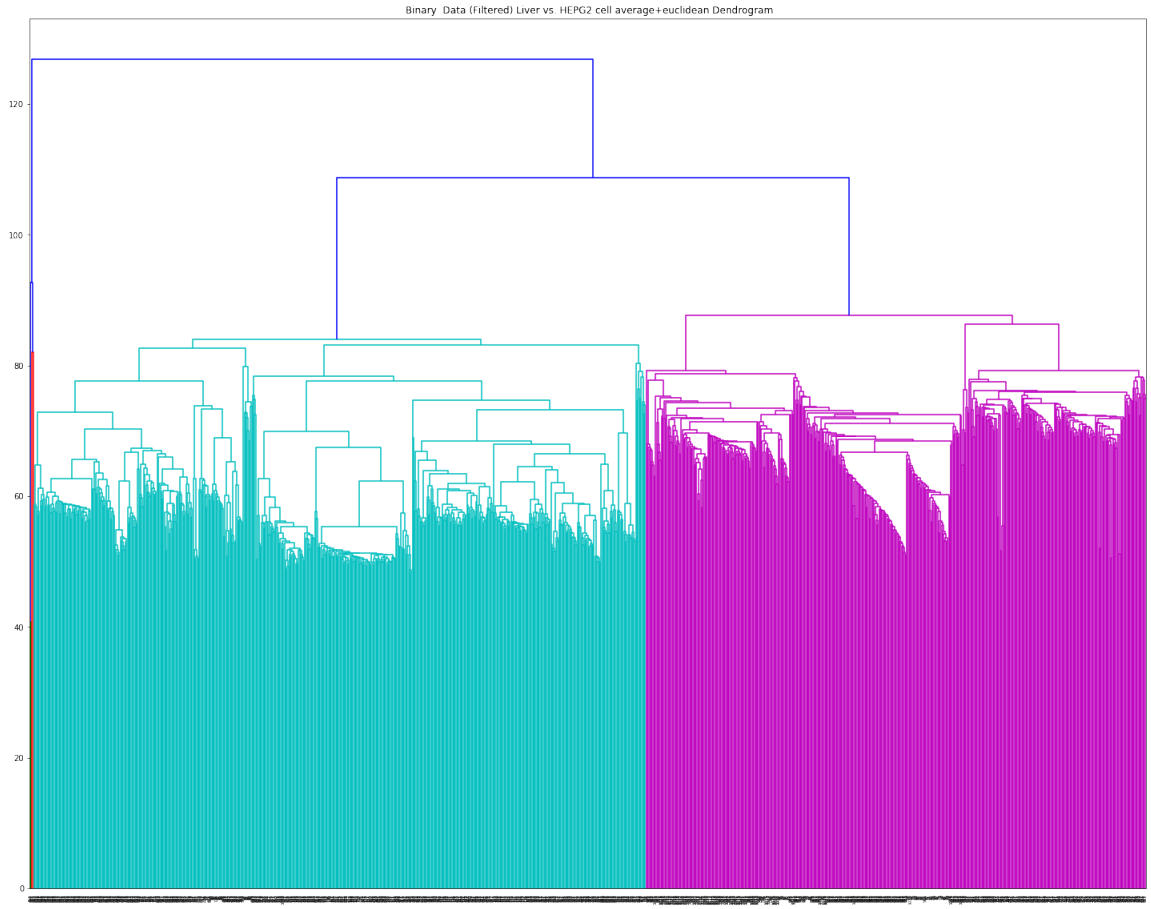
The clustering result with K-means is more promising and the result is as shown:

| Cluster ID | Cluster Size | Dominant Molecule | Molecule Purity | Dominant Series | Series Purity | Dominant Label | Label Purity |
|---|---|---|---|---|---|---|---|
| 0 | 163 | total RNA | 0.969 | GSE97655 | 0.663 | beta | 0.902 |
| 1 | 207 | total RNA | 0.778 | GSE106290 | 0.116 | PANC1 | 0.879 |
| 2 | 133 | total RNA | 1 | GSE97655 | 0.857 | ALPHA | 0.977 |

We use k = 3 according to the elbow method and the clustering did a nice job separating the 3 types of cells. Each of the clusters corresponds to 1 potential label. The series purities are moderate and the label purities are significantly higher than the worst case scenario ratios. The success on the clustering could lead to a potential good supervising model on distinguishing healthy and unhealthy Pancreatic Islet cells in phase 3 of the project

### Liver vs HEPG2 cell line

Even after removing potential outlier samples by removing all samples from series with 3 or fewer observations, the resulting dendrograms showed no promise of separating the noise from meaningful clusters. Here is an example of our best effort using the binary representation of the filtered data with average linkage, Euclidean distance:




| Cluster ID | Cluster Size | Dominant Molecule | Molecule Purity | Dominant Series | Series Purity | Dominant Label | Label Purity |
|---|---|---|---|---|---|---|---|
| 1 | 4 | total RNA | 1 | GSE107169 | 1 | Liver | 1 |
| 2 | 597 | total RNA | 0.874 | SE125570-GSE1255 | 0.134 | Liver | 0.884 |
| 3 | 488 | total RNA | 1 | GSE96981 | 0.613 | HEPG2 | 0.613 |

We can see that the purity for the cluster 3 Series ID is dominated by GSE96. The fact that clustering on the normalized data proved to be futile for k-means and hierarchical foreshadow our troubles during 2-way clustering on the liver samples.

### Bootstrap Validation

Bootstrap validation answers the important question: how can we know if a given cluster represents the actual structure in the data, or is it just an artifact of the clustering algorithm? So we write the bootstrapping algorithm to figure this out: First we cluster the data as usual. Then we resampled the original dataset with replacement to create a new dataset, and then cluster the new dataset. For every

cluster in the original clustering, find the most similar cluster in the new clustering (the one that gives the maximum Jaccard coefficient) and record that value. If this maximum Jaccard coefficient is less than 0.5, then the original cluster is considered to be "dissolved", that means it didn't show up in the new clustering. A cluster that is dissolved too often (more than half times) will be considered as an unreal cluster. Repeat steps 2-3 several times (100). In our presented results, the hierarchical clustering methods experienced no dissolved clusters with an average Jaccard similarity above 0.9. K-means also experienced no dissolved clusters and achieved Jaccard similarities no lower than 0.75.

**Two Way Clustering Results**

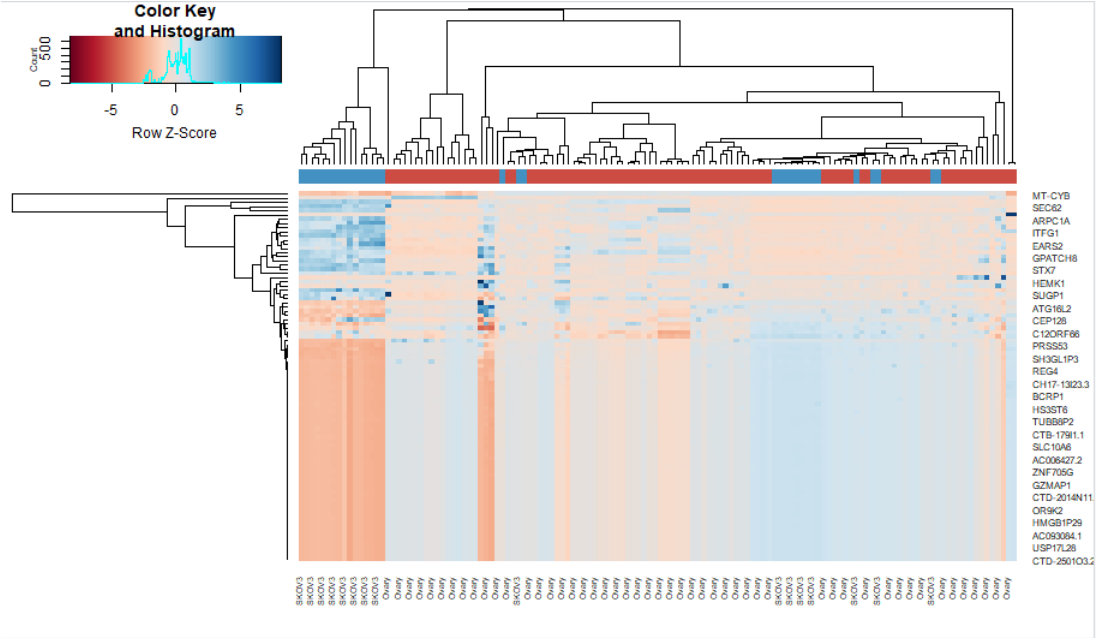
Before we can discuss the two-way clustering algorithm, we found it necessary to implement the gene shaving algorithm to reduce the large number of genes to a small subset which maximizes the variance of the average value of each gene across each sample in order to create distinct clusters. We implemented the gene shaving algorithm (10) using numpy:

1. Start with the entire expression matrix X, each row centered to have zero mean
2. Compute the leading principal component of the rows of X
3. Shave of the proportion (10%) of the genes having the smallest inner product with the leading principal component.
4. Repeat steps 2 and 3 until only one gene remains.
5. This produces a nested sequence of gene clusters $S_k$ where k denotes a cluster of k genes. We estimate k according to the $R^2$, variance, and Gap formulas given in (10).

For our purposes, we did not complete steps 6 - 7 as outlined in (10) due to time limitations, so we stuck primarily to the clustering result after step 5. This algorithm was fairly successful in reducing our genespace from 35000 to below 100 which made it possible for us to visualize our two-way clustering results with a heatmap. Essentially, two way clustering independently clusters the rows of our matrix on the columns and vice versa so that our matrix can be resorted with respect to both axes to group like gene/sample values together whose values on the heatmap ideally map to different shades of colors depending on the sample label. For each subset, we use the clustering method that was most successful during phase 1. That would be hierarchical for the ovaries and liver and k-means for the pancreas.

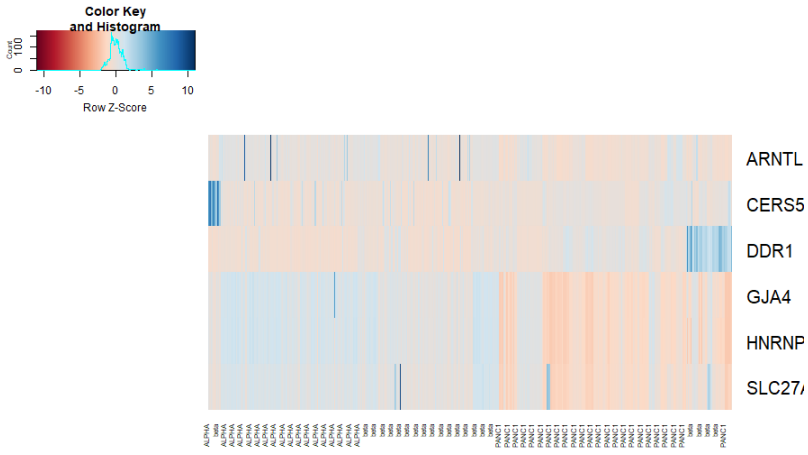
**Ovary vs SKOV3 Two Way Results**

After gene shaving, we saw it appropriate to work with a cluster containing k=80 genes for our two-way visualization. Since we use hierarchical clustering, we are able to visualize both the sample and gene dendrograms simultaneously in our heatmap in addition to sorting each axis.




Here we see that the bottom 3 quarters of the genes do a fairly good job separating a majority of the ovary samples with non-blue chunks corresponding to healthy ovary tissues in that region. These correctly correspond to a distinct region of color corresponding to z-scores near 0. We were also able to separate a significant blue chunk of SKOV3 samples on the left part of the heatmap, but the color patterns observed here do not extrapolate to any of the other clusters on the map corresponding to SKOV3. This could be due to a batch effect on the specific series of SKOV3 that we saw in 1-way clustering. It is evident that clustering on the ARCHS4 data is acceptable to separating out the majority healthy ovarian tissue and a specific batch of SKOV3 samples.

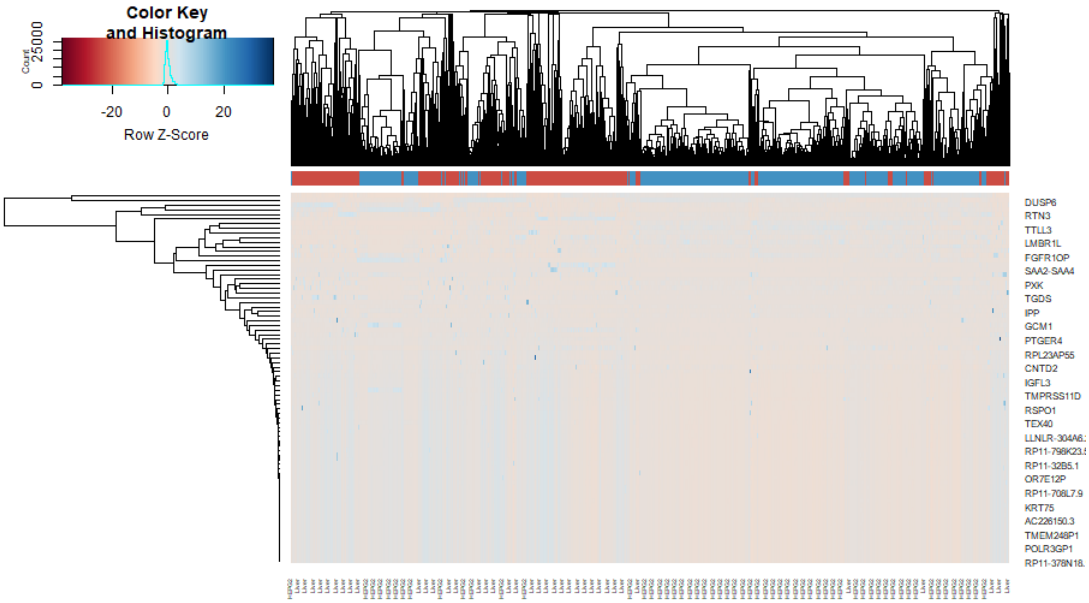
### PANC1 vs. Alpha vs Beta Two Way Results




Since k-means performed better with this subset of data, we do not have the luxury of seeing a dendrogram for our data. Instead, two-way k-means simply ordered our data in such a fashion so that similar values sample/gene combos group together. Here we see that the GJA4, HNRNP and SLC27A genes express higher than average for the alpha and beta pancreatic islets, but fail to express high at all for the PANC1 samples. It is also interesting to note the blocks of blue corresponding to extremely high expression values of DDR1 for a subsample of beta's and PANC1, as well as CERS5 on a few alpha/beta samples. We did not have time to run by these findings with a geneticist.

**Liver vs HEPG2 Two Way Results**

Unfortunately, two-way clustering with gene-shaving down to 80 genes did not give us any meaningful heatmap for the liver subset. This can be seen by the lack of distinct color regions in the following heatmap.




It is interesting to note, however, that the reordering of samples in the one-way hierarchical clustering of the columns in the heatmap create two moderately pure partitions of red and blue (corresponding to healthy Liver and HEPG2). This indicates that even though we were unable to get pure clusters with k=2 in our original hierarchical clustering, the way that two-way clustering orders the samples does a moderately good job at grouping together like sample labels, even though the gene expression color heatmap does not have a strong correlation to these differences.

## Phase 3: Machine Learning

The results we are about to present for the supervised machine learning cross-validation show a significant performance enhancement when it comes to classifying by sample labels, yet do not serve the purpose of discovering the actual genetic differences between these samples based on their gene expression profiles due to the black box nature of machine learning. The model we have decided to use is the current kingpin of machine learning classification: XGBoost. XGBoost is a powerful gradient boosting classification algorithm known for its superior accuracy scores compared to other algorithms on the same classification tasks. Lasso and Ridge Regression Regularization are built into the model to prevent overfitting. It is also faster than most gradient boosting models with its parallel processing abilities, and it has more effective tree pruning algorithms that can prune the tree backward and remove splits beyond which there is no positive gain (9). We used the cross validation algorithm (6 folds) with accuracy, f1 micro, and f1 macro scores to determine whether our prediction scores were significantly better than our sample proportion ratios.

Using max depth of 3, XGBoost achieved over 99% accuracy when it came to classifying 99 Ovary vs 33 SKOV3 samples on 80 gene shaved features and 97% f1 macro and micro scores. As a sanity check for the gene shaving algorithm, the cross validation score was near 80% when we chose a random sample of 80 genes which is too close for comfort to the sample proportion of ovaries. For the 994 HEPG2 Liver cancer cell line samples vs the 721 healthy liver samples, XGBoost scored over 99% in all three scoring categories also using max depth of 3. Finally, for the PANC1 vs. alpha vs. beta samples, using a max-depth of 5 we were able to obtain 94% accuracy with 94% f1 micro and 93% f1 macro scores. Overall, machine learning was able to perform exceptionally well overall all three sample subsets.

## Discussion

Despite the wonderful organization and abundance of data in the ARCHS4 gene expression repository, there are some drawbacks that we did not have time to tackle in Phases 1 and 2. Our approaches using standard clustering methods, gene shaving, and data normalization with Euclidean distance were not enough to distinguish samples into their respective labels. Instead, we were only able to see success when it came to separating large chunks of samples which come from the same series (experiment). Gene expression is a tricky process to standardize across experiments due to the fact that gene expressions vary with time in unpredictable fashion across all possible genes. That is two samples from the same tissue can have two different gene expression readings due to environmental factors such as time, subject type, temperature or state of the subject when the samples were taken. This led us to not so clear distinctions in two-way clustering despite our attempts to reduce the magnitude differences via normalization of the data.

One problem that kept coming up is the effect of different batches producing outliers which makes it difficult to group together observations in hierarchical clustering. Often, we would look at a dendrogram where the first few separations involve clusters of individual samples before the rest of the data separates into meaningful clusters. This means we would have to choose k=5 to view clustering results where we would expect 2 groupings, for example, because 3 of the points were outliers that became their own clusters early on. Otherwise, if we had picked k=3 there would be two clusters each with one outlier point and one cluster with the rest of the points which is not desired. One possible solution could be to filter out some of the outlier points by identifying samples that come from series (experiments) with only one or two samples. These samples tend to be different from the rest due to environmental conditions, but they are not exhaustive of the whole set of outliers in a given dataset.

Another problem that we encountered was having too many genes, not enough metadata on genes. We have not yet put our full efforts into one-way clustering genes based on their sample arrays, but our preliminary efforts have shown that dendrograms are impossible to generate. This means analyzing these clusters may be difficult. One option would be to filter down our list of 35,000 genes down to those which have significant counts in the samples of interest and have known biological relevance to these samples (such as the BRCA1 and BRCA2 genes for ovarian cancer). Another possibility would be to perform PCA on our gene space. This would make our clusters easy to visualize, but hard to interpret.

Differences among samples may be due to batch effect and it is hard to distinguish the sample effect from the batch effect. This leads to frustrating results in our clustering algorithms that have few clusters with high series ID purities (low label purities) while the rest of the clusters have low sample label purities.

The last huge problem is that analyzing clusters is a time-consuming, objective process. Even though we are not biologists, we have written a function to export our clustering results to excel tables to analyze the purity and individual sample metadata contained in each cluster. The ability to visualize our findings with two-way clustering is crucial, and we did have some interesting observations with regards to the ovary and pancreas analyses. We saw huge variation from different samples with the same labels in gene expression. Some SKOV3 (and likewise ovary) cells were highly expressed by the bottom half of the genes, whereas others were very low in their expression which shows the extent to which the environmental conditions leading up to a gene expression microarray analysis play a huge role for a batch of samples. On the other hand, we saw some consistency among the pancreas samples for genes such as GJA4, HNRNP and SLC27A which expressed higher than average for healthy pancreatic islet tissue but not so much for the PANC1 cell line.

Machine learning was able to resolve the issue of supervised classification between sample labels quite well. We were delighted to see accuracy and f1 scores over 93

## Contributions

Aleksander Braksator: Review of prior work, coming up with the 3 phases, data transformation pipeline, 1-way hierarchical clustering, purity table pipeline, bootstrap validation, two-way clustering, 2-way heatmaps, gene shaving

Haobo Zhang: K-Means clustering pipeline (1 way and 2 way)

Qinai Xu: Bootstrap validation, SCP clustering (did not make it into final paper due to bug and lack of time)

## References


[1] Getz, G., Levine E., & Domany, E. (2000) Coupled two-way clustering analysis of gene microarray data. *Proc. Natl. Acad. Sci. USA*, 10.1073

[2] Hastie, T., Tibshirani, R., Eisen, M. B., Alizadeh, A., Levy, R., Staudt, L., Chan, W. C., Botstein, D., & Brown, P. (2000). 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. Genome biology, 1(2), RESEARCH0003. https://doi.org/10.1186/gb-2000-1-2-research0003

[3] Koch, C. M., Chiu, S. F., Akbarpour, M., Bharat, A., Ridge, K. M., Bartom, E. T., & Winter, D. R. (2018). A Beginner's Guide to Analysis of RNA Sequencing Data. American journal of respiratory cell and molecular biology, 59(2), 145–157. https://doi.org/10.1165/rcmb.2017-0430TR

[4] Lachmann, A., Torre, D., Keenan, A.B. et al. (2018) Massive mining of publicly available RNA-seq data from human and mouse. Nat Commun 9, 1366 . https://doi.org/10.1038/s41467-018-03751-6

[5] Sindhu, S. & Sindhu, D. (2017) Data Mining and Gene Expression Analysis in Bioinformatics. IJCSMC, Vol. 6, Issue. 5, ISSN:2320–088X

[6] Stevens, R. (2014) "Clustering with Gene Expression Data." Utah State University Bioinformatics Lecture. Utah State University.

[7] Tibshirani, R., Et al. (1999) Clustering methods for the analysis of DNA microarray data. Department of Health Research and Policy, Department of Statistics, Department of Genetics and Department of Biochemistry, Stanford University

[8] Yeung, K.Y., Medvedovic, M., & Bumgarner, R.E. (2003) *Clustering gene-expression data with repeated measurements*. Genome Biol 4, R34. https://doi.org/10.1186/gb-2003-4-5-r34

[9] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. DOI:https://doi.org/10.1145/2939672.2939785

[10] Hastie, Trevor Tibshirani, Rob Eisen, Michael Alizadeh, Ash Levy, Ronald Staudt, Louis Chan, Wing Botstein, David. (2000). 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. Genome biology. 1. RESEARCH0003. 10.1186/gb-2000-1-2-research0003.