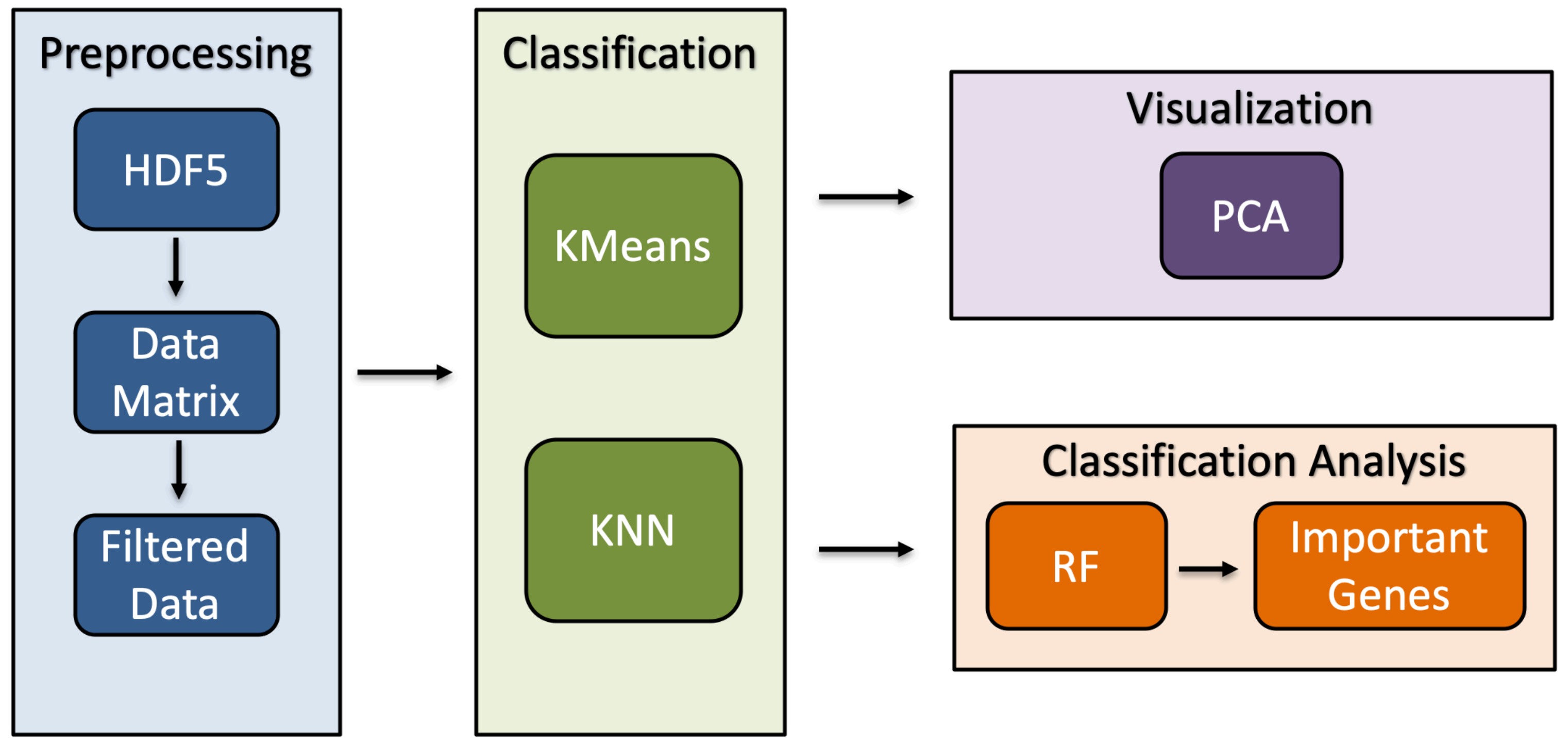
Utilizing Different Machine Learning Methods on scRNA-Seq Data to Determine Genes of High Importance

Angelica M. Walker

Abstract

Single Cell RNA-Sequencing (scRNA-Seq) has provided a break-through in biological research, allowing individual cells to be studied to determine cellular functions. ScRNA-Seq measures the gene expression count of different gene barcodes to discover which genes are expressed for different cells. Due to the number of possible single nucleotide polymorphisms (SPNs) and copy number variations (CNVs) in genomes, some gene expression could be missed in analysis. Additionally, there are inherent issues with Illumina sequencing as the strand gets longer, missing some base pairs in analysis. Since scRNA-Seq is a relatively new data format, new data analysis methods are coming about in the field. It is important to verify that this data is adequate enough for analysis before the analysis can be conducted. Can K-Means and K Nearest Neighbors (KNN) effectively classify mouse and human cells in a mixed scRNA-Seq dataset? Can we determine which genes are the most significant in determining these clusters?

Data Analysis Workflow



Preprocessing: Transform HDF5 file into a useable Pandas DataFrame, then filter out the samples that contain zeros.
Classification: Use KMeans and KNN to classify the samples
Visualization: Use Principal Component Analysis to visualize the clusters in a 2-D space
Classification Analysis: Use Random Forest to determine which genes of high importance are used to determine classes in classification.

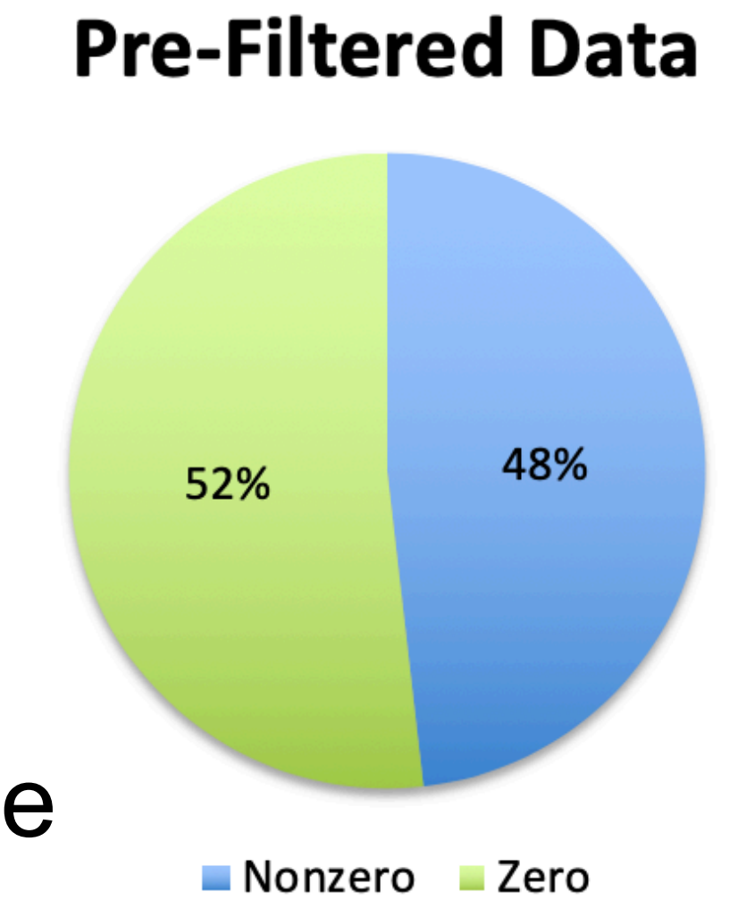


Figure 1: Filtering of Data shows that 48% of the samples contained no data.

KMeans

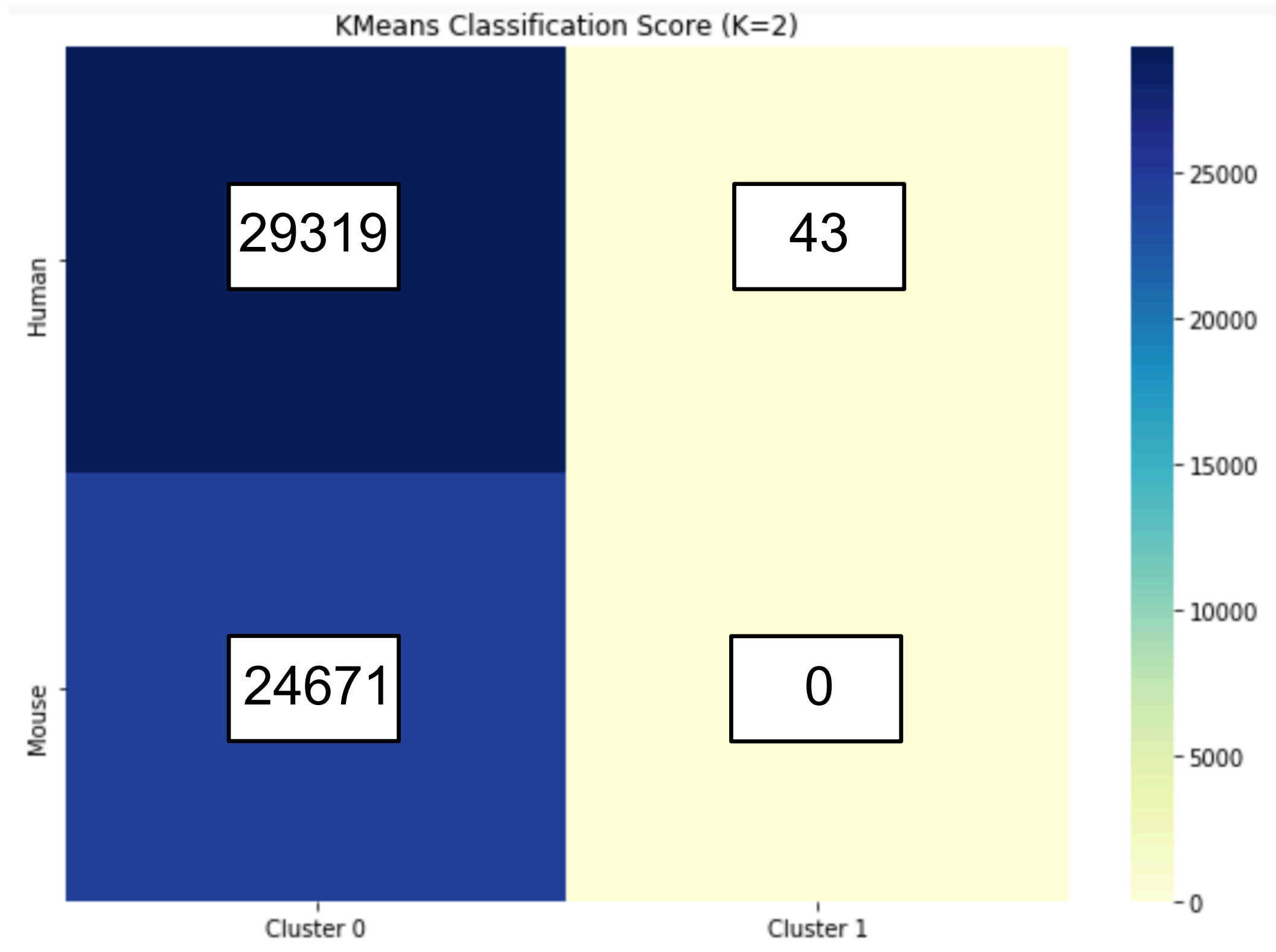


Figure 2: Confusion Matrix for KMeans Classification where K=2. Most of the samples are clustered in one cluster, which is a mix of both human and mouse cells. The second cluster contains only 43 human cells.

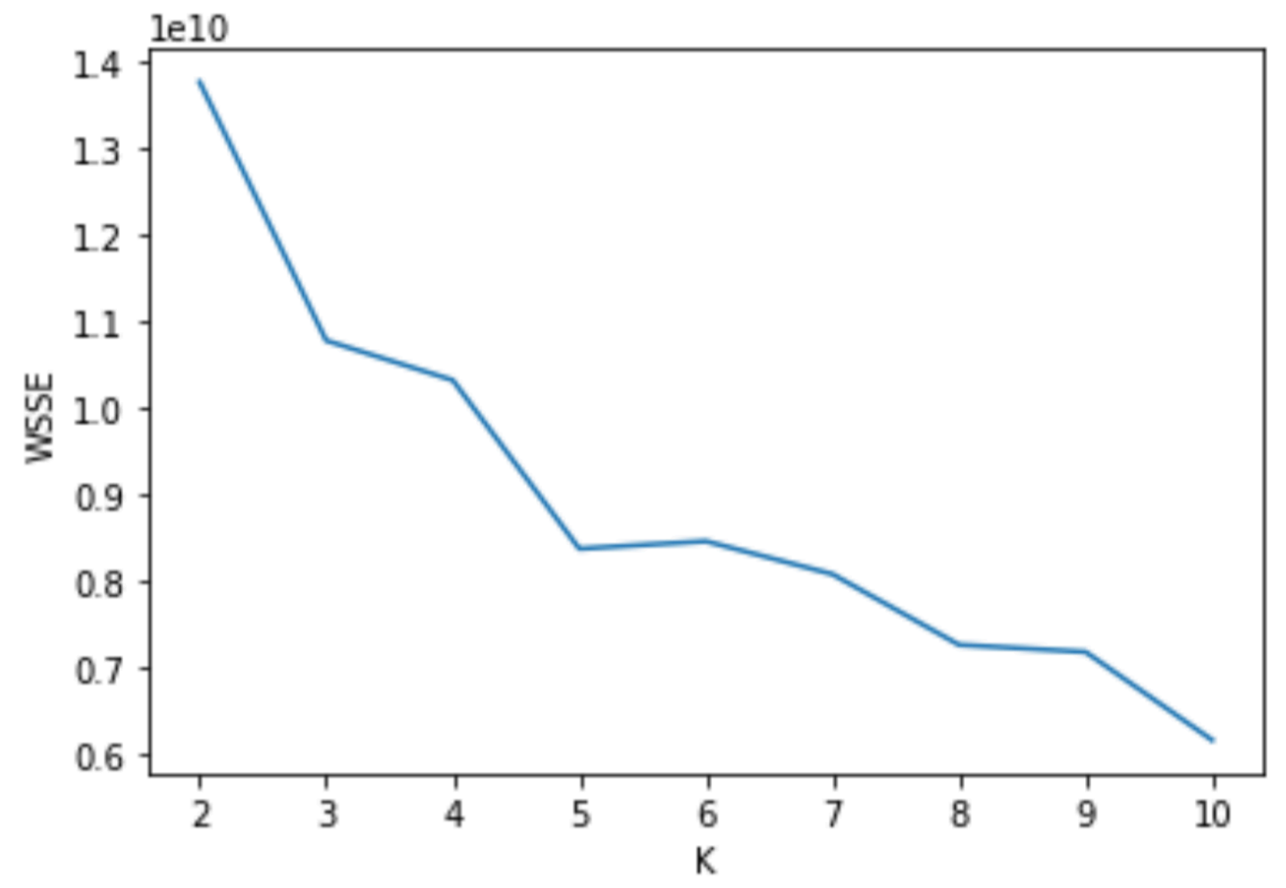


Figure 3: Weighted Sum Squared Error for KMeans are varying Ks. Elbow Method shows that optimal K = 5.

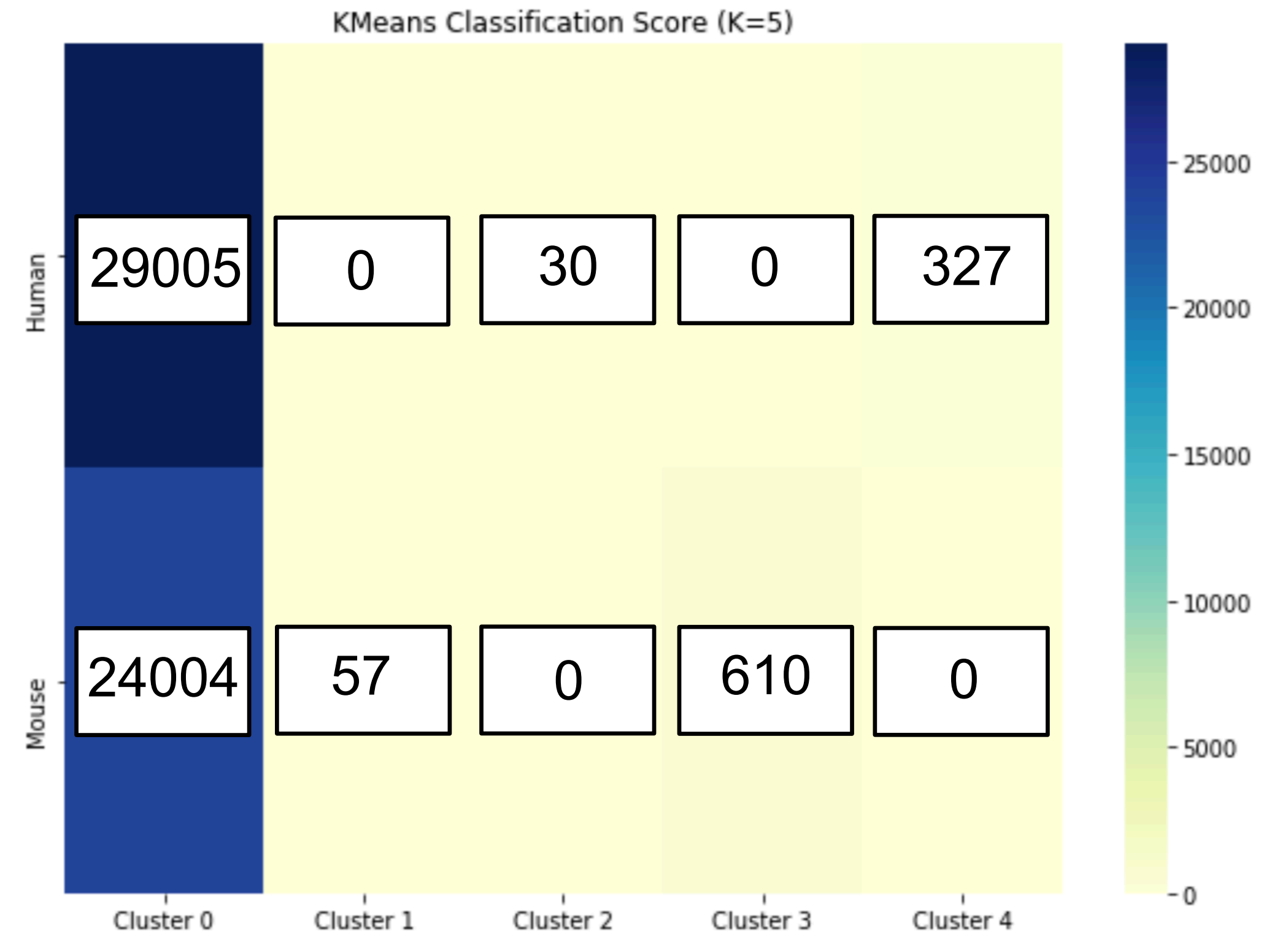


Figure 4: Confusion Matrix for KMeans Classification where K=5. Much like for K=5, one cluster contains the majority of the samples and is both human and mouse cells. The remaining clusters are pure human or pure mouse, but are small clusters.

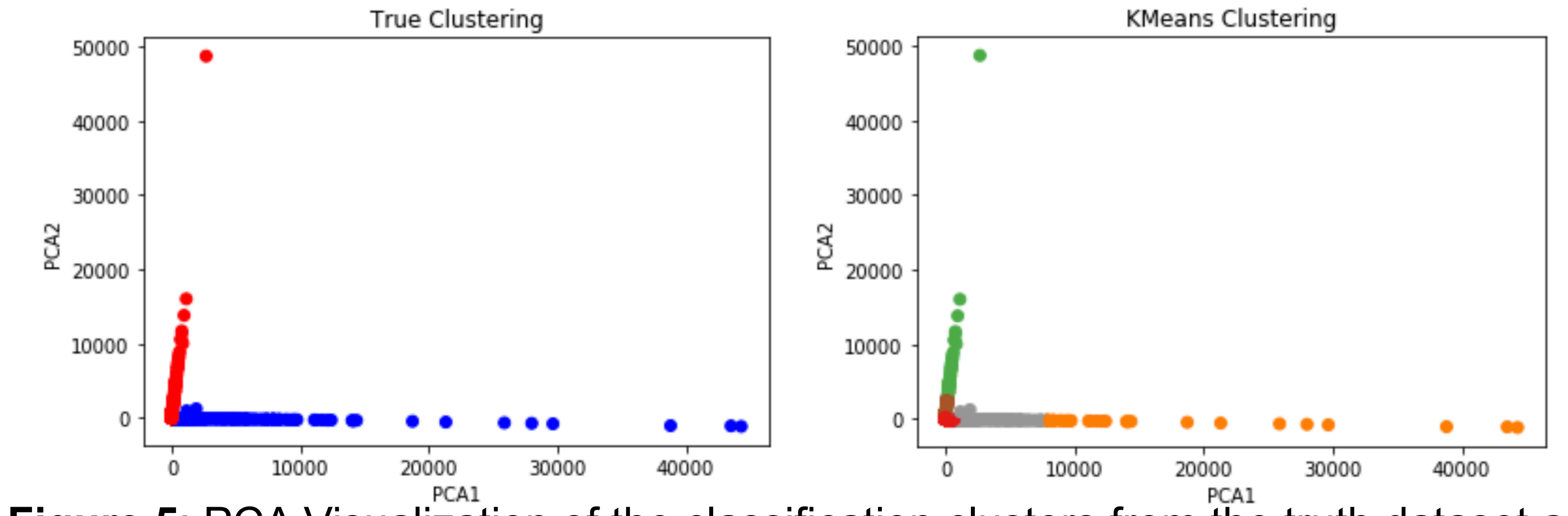


Figure 5: PCA Visualization of the classification clusters from the truth dataset and the KMeans (K=5) predictions.

K Nearest Neighbors

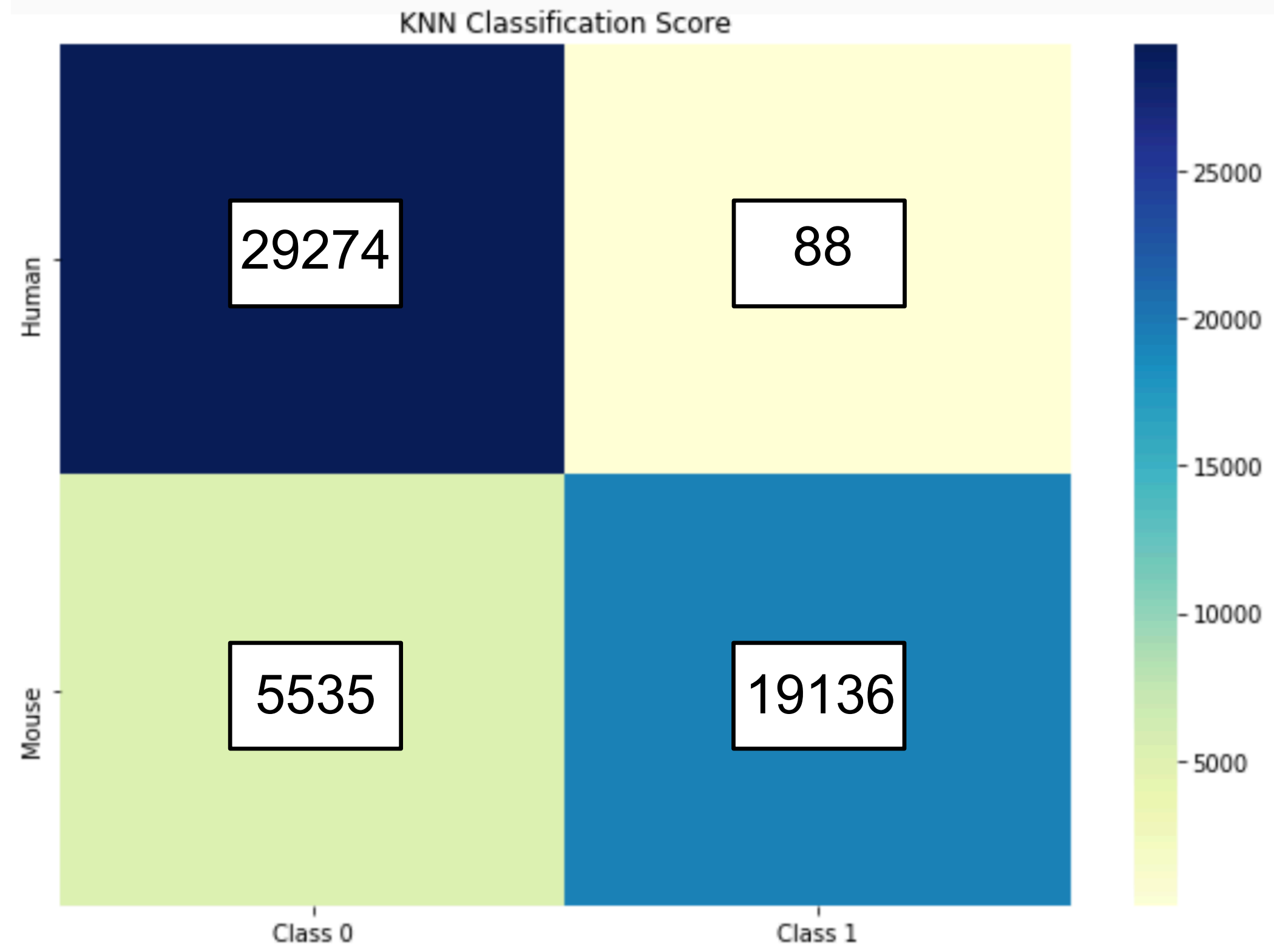


Figure 6: Confusion Matrix for KNN Classification. KNN outperforms KMeans, although neither cluster are pure, they contain predominantly one cell type.

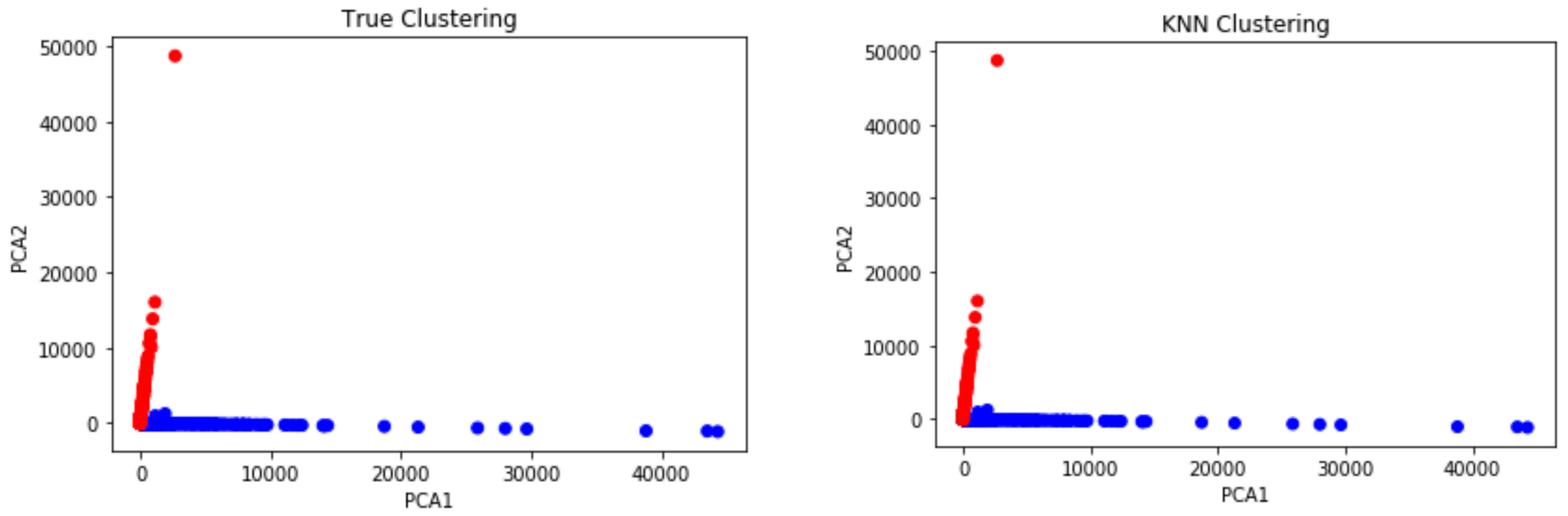


Figure 7: PCA Visualization of the classification clusters from the truth dataset and the KNN predictions.

Conclusion

Gene (Kmeans)	Importance Score (Kmeans)	Gene (KNN)	Importance Score (KNN)
CATTCCGAGTGGAATT-1	0.1707	GCACGGTTCCCGATCT-1	34.95
TGGAGGATCCGCCTAT-1	0.1665	GCAGCCAAGCACCTGC-1	32.40
ATGCCTCCACGGGCTT-1	0.0549	AGGGCCTCACAAATAG-1	32.23
CACGGGTCATTGCCTC-1	0.0295	AAGACAAAGTGGCAGT-1	32.13
AAGACAAAGAGATGCC-1	0.0259	CATCAAGAGGCTATCT-1	30.05

Figure 8: Top 5 scoring genes from each machine learning method, acquired using Random Forest with 500 and the predicted class values as the Y variable. KMeans contains very low important scores, while KNN contains higher scores. This is expected due to the poor cluster results with KMeans.

References

"5k 1:1 mixture of fresh frozen human (HEK293T) and mouse (NIH3T3) cells (Next GEM)." 10X Genomics. *Single Cell Gene Expression Datasets.* https://support.10xgenomics.com/single-cell-gene-expression/datasets

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE