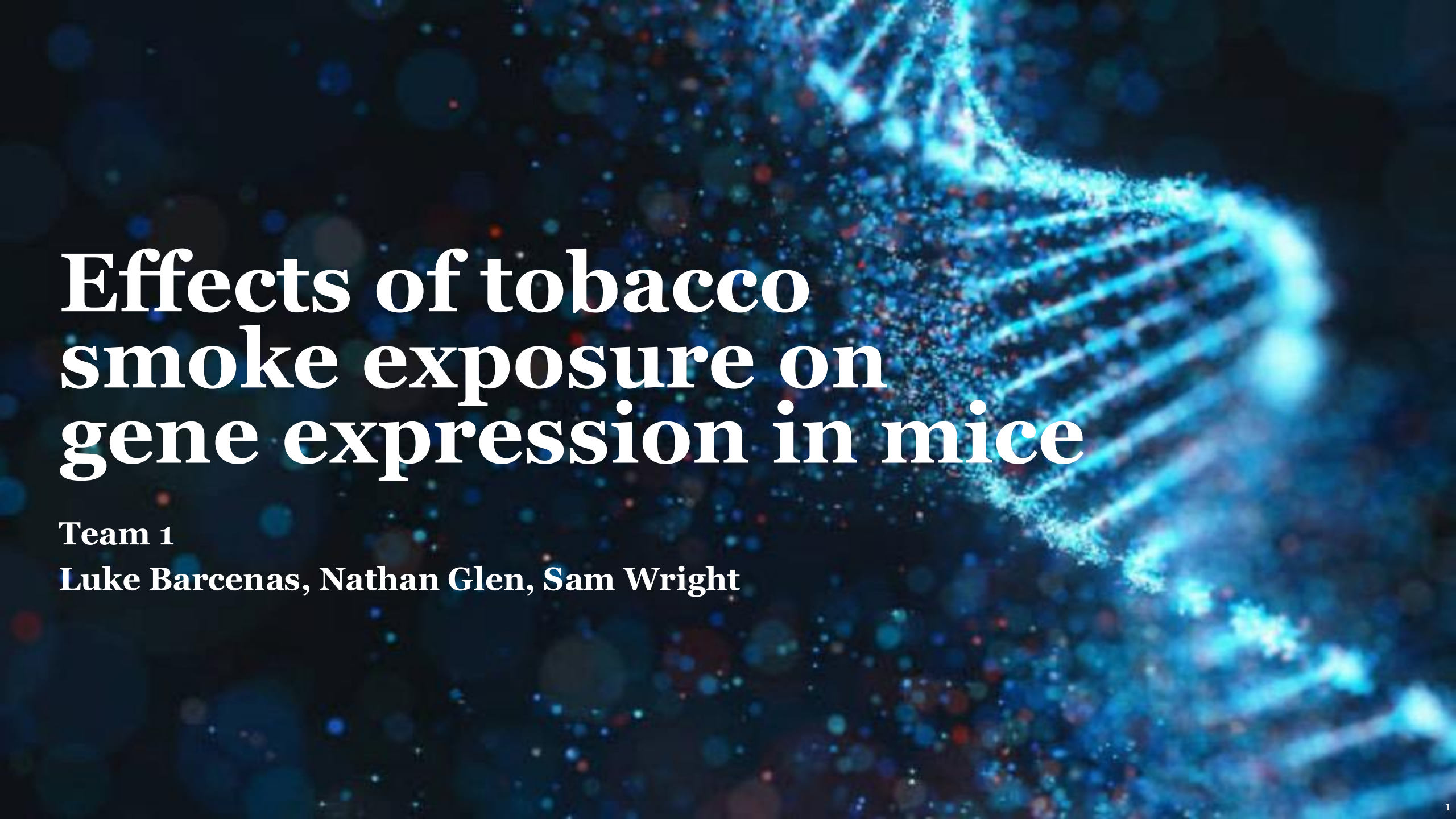


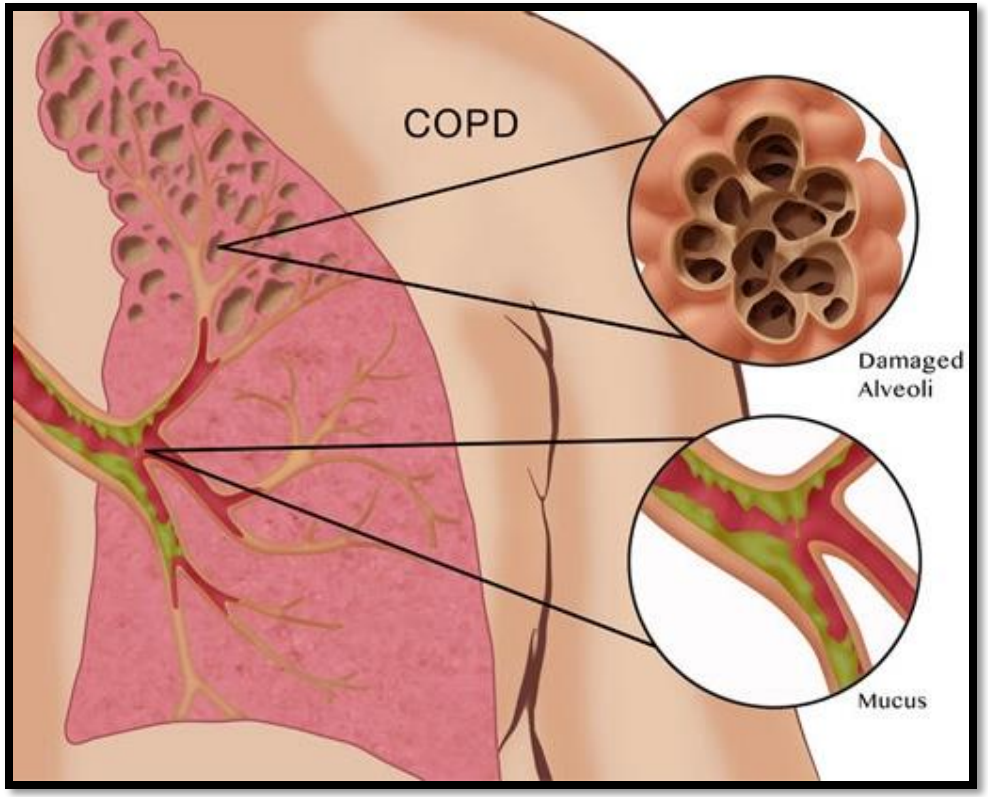
Effects of tobacco smoke exposure on gene expression in mice

Team 1

Luke Barcenas, Nathan Glen, Sam Wright

Introduction

- Chronic obstructive pulmonary disease (COPD) is a type of lung disease that is a leading cause of mortality in humans.
  - COPD is primarily caused by prolonged exposures to cigarette smoke (CS) it may persist even after stopping.
  - Our dataset tested the effect of cigarette smoke on the gene expression of mice.
- Does exposure to cigarette smoke have a significant impact on gene expression in mice?
- The experiment tested different groups to analyze the effects of cigarette smoke over varying periods of time.
  - We Generalized the data to 2 different groups: CS (Cigarette Smoke) and AC (Control Group)



Hypothesis

- Will cigarette smoke have an impact on the gene expression of mice? Of course!
- More specifically, there must be some impact on the gene expression.
- This could either be through effects on immune response, inflammatory response, and/or genes that contribute to the overall function of the cardiovascular and respiratory systems
- Could there be any other genes that may be impacted from cigarette smoke exposure?

Differential Expression & Enrichment

Differential Expression

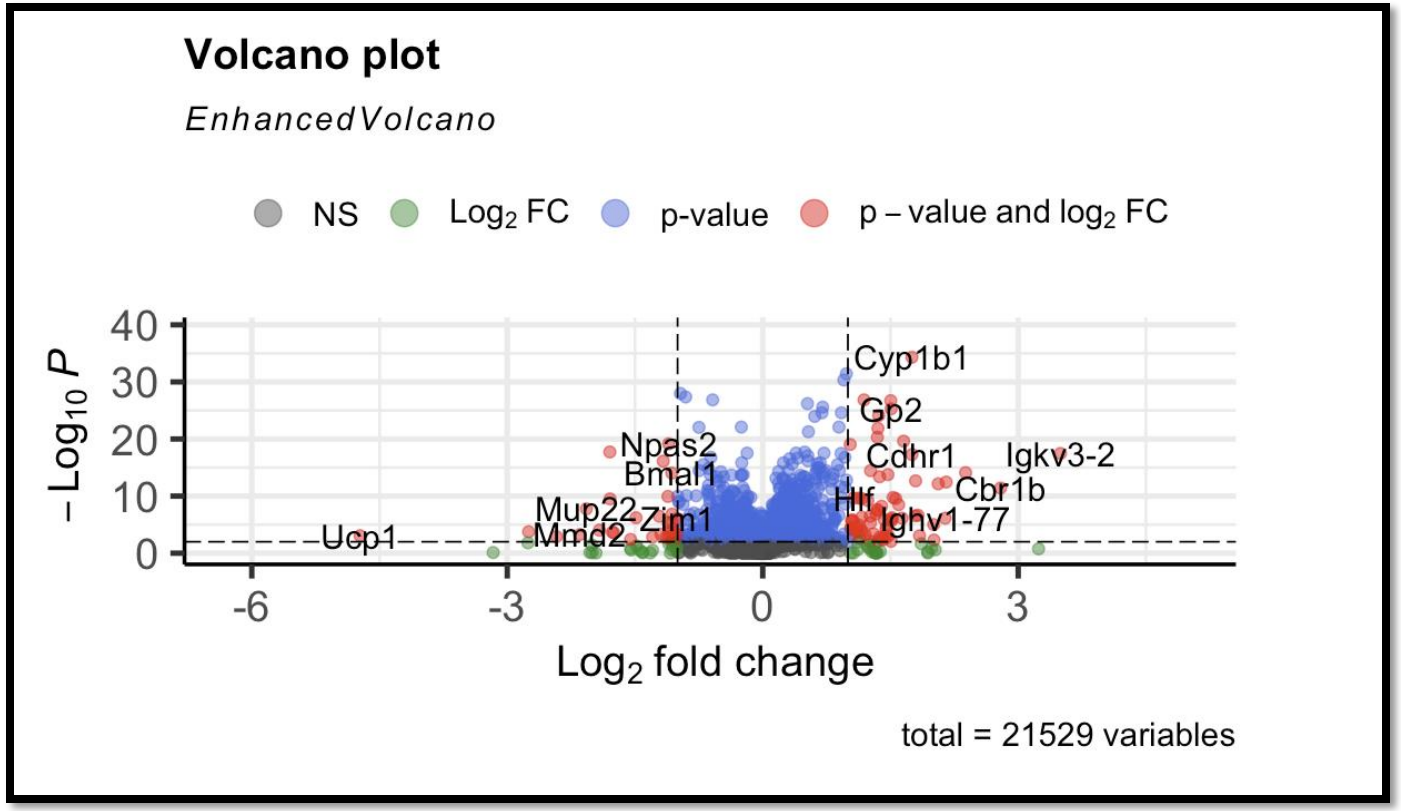
Methods:

- Used **deseq** package in R
- 0.01 threshold for p
- **ggplot()** for visualizations

Results:

- Small # of points of interest
- Log fold change between –3 and 3
- Most of the genes are upregulated

Volcano Plot:



Differential Expression & Enrichment

GSEA

Methods:	Ontology:	How:	Interesting Results
• Cluster Profiler • Gprofiler2 • TopGO	• GO("BP") • GO("CC") • GO("BP")	• Compared **p-values**	• GO:0042571 – disease and infection • Mostly Biological Processes! • Makes sense with our hypothesis!

Combined Results:

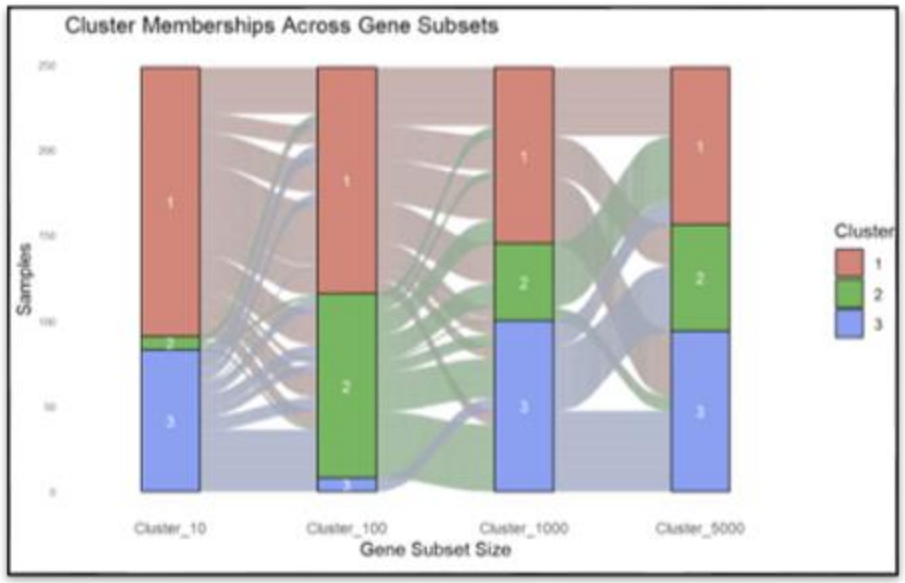
ID <chr>	Combined_Info <chr>	p_value <dbl>	Num_Methods <dbl>
GO:0042571	immunoglobulin complex, circulating	6.593228e-05	1
GO:0006805	xenobiotic metabolic process	7.319964e-05	1
GO:0071466	cellular response to xenobiotic stimulus	1.049894e-04	1
GO:0005615	extracellular space	1.094173e-04	1
GO:0019748	secondary metabolic process	2.518547e-04	2
GO:0019814	immunoglobulin complex	2.532261e-04	1
GO:0071756	pentameric IgM immunoglobulin complex	8.065625e-04	1
GO:0071754	IgM immunoglobulin complex, circulating	8.065625e-04	1
GO:0071751	secretory IgA immunoglobulin complex	8.065625e-04	1
GO:0071749	polymeric IgA immunoglobulin complex	8.065625e-04	1

1–10 of 10 rows

Clustering & Enrichment

Example Alluvial Plot:

PAM Clustering k=3



We used a total of 3 clustering methods:

- K-means, PAM, and Hierarchical Clustering

Significance of Our Clustering Methods

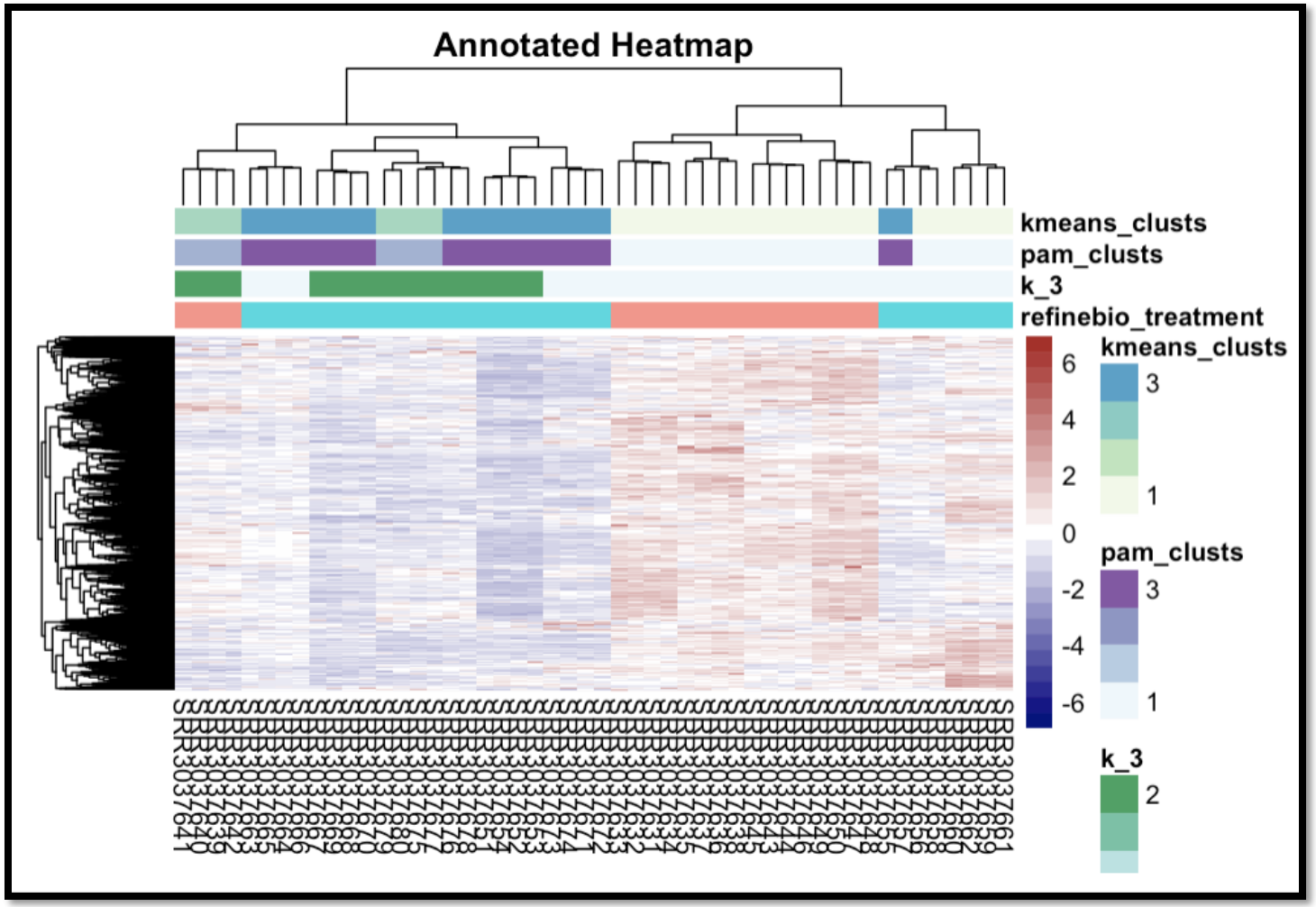
Method <chr>	Cluster <chr>	P_Value <dbl>	Adjusted_P_Value <dbl>
Kmeans	10 vs treatment	4.237103e-02	5.296379e-02
Kmeans	100 vs treatment	1.596828e-26	7.984138e-26
Kmeans	1000 vs treatment	1.037645e-04	2.594111e-04
Kmeans	5000 vs treatment	1.049751e-01	1.049751e-01
Kmeans	10000 vs treatment	2.488355e-02	4.147258e-02
PAM	Cluster_10	8.986672e-01	8.986672e-01
PAM	Cluster_100	3.265912e-05	6.531825e-05
PAM	Cluster_1000	1.306747e-05	5.226988e-05
PAM	Cluster_5000	1.135428e-03	1.513904e-03
HC	10 vs refinebio	6.456613e-04	3.228307e-03
HC	100 vs refinebio	9.223356e-01	9.223356e-01
HC	1000 vs refinebio	1.822988e-01	4.557471e-01
HC	5000 vs refinebio	4.666312e-01	5.832890e-01
HC	10000 vs refinebio	4.666312e-01	5.832890e-01

Clustering & Enrichment

Interesting Results

- Compare clustering methods vs. OG treatment
- K-means way different than other methods
- May have captured combined earlier groups in clustering!

Combined Heatmap



Predictive Modeling

Methods

- Supervised Learning
- Used **tidymodels()**
- Split into train, test
- Models for OG group and our clusters

Interesting Results

- Random Forest and SVM had high accuracy (overfitting?)
- Logistic performed better for arbitrary clusters
- We think there is something that distinguishes our two groups based on our model's performance!

Original Treatments:

Random Forest

```
Prediction ac cs
        ac 22  0
        cs  1 28
Accuracy: 98.04 %
```

Logistic Regression

```
Prediction ac cs
        ac 21 11
        cs  7 12
```

Accuracy: 64.71%

SVM

```
Predicted ac cs
        ac 38  0
        cs  0 62
Accuracy: 100 %
```

Our Clusters:

Random Forest

```
Prediction  1  2  3
         1 15  3  1
         2  0 11  0
         3  4  0 17
Accuracy: 84.31 %
```

Logistic Regression

(No table available)

Accuracy: 72.83%

SVM

```
Predicted  1  2  3
        1 27  0  0
        2  0 18  1
        3  1  0 28
Accuracy: 97.33333 %
```

Conclusions & Future Work

Our Conclusions

What we learned:

- We learned how to conduct an EDA of genomics data and interpret/visualize those results
- We learned how to use Machine Learning algorithms in general, and applied them to our situation

What we would do differently:

- We would vet our choice of dataset a little more before jumping into a project

Potential Future Work

New question: How does smoke exposure affect genes over time?

- Incorporate all the original classes from our dataset
- Look at data with different organisms