Metadata Exploratory Report

Code

Project: Differential Gene Expression and Pulmonary Function in Vaping LatinX Adolescents

Analyst: Trent Hawkins

Investigator(s): Sunita Sharma, MD

Report generated: April 01, 2022

Center for Innovative Design & Analysis

colorado school of **public health**

Background

The data presented in this report are part of a study aimed to assess differential gene expression and methylation in vaping versus non-vaping LatinX youths in Pueblo and Denver, CO. Pulmonary function data were also obtained in order to better understand the impacts of vape use on pulmonary function. To assess differential gene expression and methylation, naso-epithelial swabs were obtained from each participating subject. Pulmonary function is assessed using PFTs (Pulmonary Function Tests) and Impulse Oscillometry (IOS).

Study Population

This data set consists of samples taken from 51 people ages 12-17 yrs from the Pueblo, Denver, and Aurora, CO areas. Vape Status (did you vape in the last 6 months?) and ethnicity are self-reported.

Methods

All analyses performed using R version 4.1.2 (2021-11-01)

Clinical Data Processing

Vape Status

Subjects are dichotomized to those that used a vaping device in the last 6 months and those who have not based on the variables *ever_vape*, *vape_days*, and *last_vape*. This variable will be referred to as *Vape Status* throughout this report. One participant (SID = 111) reported that they had used a vaping device 5 out of the last 30 days, but did not respond to *last_vape*. They were labeled as "NA" in previous analyses.

Sex

Biological sex will need to be identified using available genomic data.

Geographic Location

Subjects' geographic location, *city*, was grouped into the new broader variable *recruiting_center* which encompasses the broader geographic region where they live.

Lung Function and IOS

Measures of lung function and IOS were visually inspected for normality using histograms.

Gene-Count Processing

Annotation: Ensembl annotation for GrCH 39 ver. 37
Differential Expression Analysis: DESeq2 1.34.0
Removal of Unwanted Variance: RUVSeq 1.28.0

Gene Filtering Parameters
This analysis will conduct a comparison of various gene-filtering parameters presented in previous analyses and in the current literature to select parameters best-suited for this study.

Normalization
The following analyses used the function RUVr from the R package RUVSeq. RUVr uses the deviance residuals from a first pass negative binomial GLM to perform a factor analysis which corrects for unwanted technical effects. The first-pass model formula is presented below :

$$raw\ read\ count \sim \beta_0 * vape\ status\ + \beta_1 * male\ + \beta_2 * latinx$$

RUVr will be performed with k = 1 through k = 5 factors and the best cutoff for factor analysis will be determined visually using RLE plots and dendrograms for each level k. Previous analyses used R package DESeq2 to fit the first-pass GLM. This analysis will use edgeR due to its reference in the literature for the RUVr procedure mentioned above.

Transformations
A variance stabilizing transformation (VST) was applied in previous analyses. This analysis will use the same transformation, but will apply the transformation only after normalization.

Results

After removing participants with missing values for *vape status*, we are left with **n = 50** subjects. Previous analyses showed **n = 12** participants had vaped in the last 6 months. This analysis will use **n = 13** participants who had vaped in the last 6 months. The lung function variable *fev1* and *fev1_fvc* reported 22 missing values. IOS measures *r5* and *x20* reported 1 and 6 missing values, respectively.

Table 1: Clinical Data

Code

	Did Not Vape in Last 6 Months (N=37)	Vaped in Last 6 Months (N=13)	Total (N=50)
Sex			
Female	21 (56.8%)	5 (38.5%)	26 (52.0%)
Male	16 (43.2%)	8 (61.5%)	24 (48.0%)
Age (yrs)			
Mean (SD)	14.6 (1.4)	14.8 (1.4)	14.6 (1.4)
Range	12.0 - 17.0	13.0 - 17.0	12.0 - 17.0
Recruitment Center			
Aurora	15 (40.5%)	0 (0.0%)	15 (30.0%)
CommCity/Denver	13 (35.1%)	1 (7.7%)	14 (28.0%)
Pueblo	9 (24.3%)	12 (92.3%)	21 (42.0%)
Ethnicity			
LatinX	23 (62.2%)	11 (84.6%)	34 (68.0%)
Non-LatinX	14 (37.8%)	2 (15.4%)	16 (32.0%)
FEV1			
N-Miss	10	12	22
Mean (SD)	2.6 (0.7)	3.9 (NA)	2.6 (0.7)
Range	1.2 - 3.9	3.9 - 3.9	1.2 - 3.9

	Did Not Vape in Last 6 Months (N=37)	Vaped in Last 6 Months (N=13)	Total (N=50)
FEV1/FVC (%)			
N-Miss	10	12	22
Mean (SD)	0.8 (0.1)	0.7 (NA)	0.8 (0.1)
Range	0.5 - 1.0	0.7 - 0.7	0.5 - 1.0
R5			
N-Miss	1	0	1
Mean (SD)	4.0 (0.9)	5.0 (1.3)	4.3 (1.1)
Range	2.0 - 6.1	3.7 - 7.6	2.0 - 7.6
X20			
N-Miss	4	2	6
Mean (SD)	0.1 (0.6)	0.7 (0.9)	0.2 (0.7)
Range	-1.1 - 2.4	-1.0 - 2.3	-1.1 - 2.4

Code

Figure 1: Population Demographics

Figure 1 visualizes the demographic information presented in Table 1. The majority of vaping subjects were recruited in Pueblo (92%) and identified as LatinX (85%). 62% of vapers identified as male.

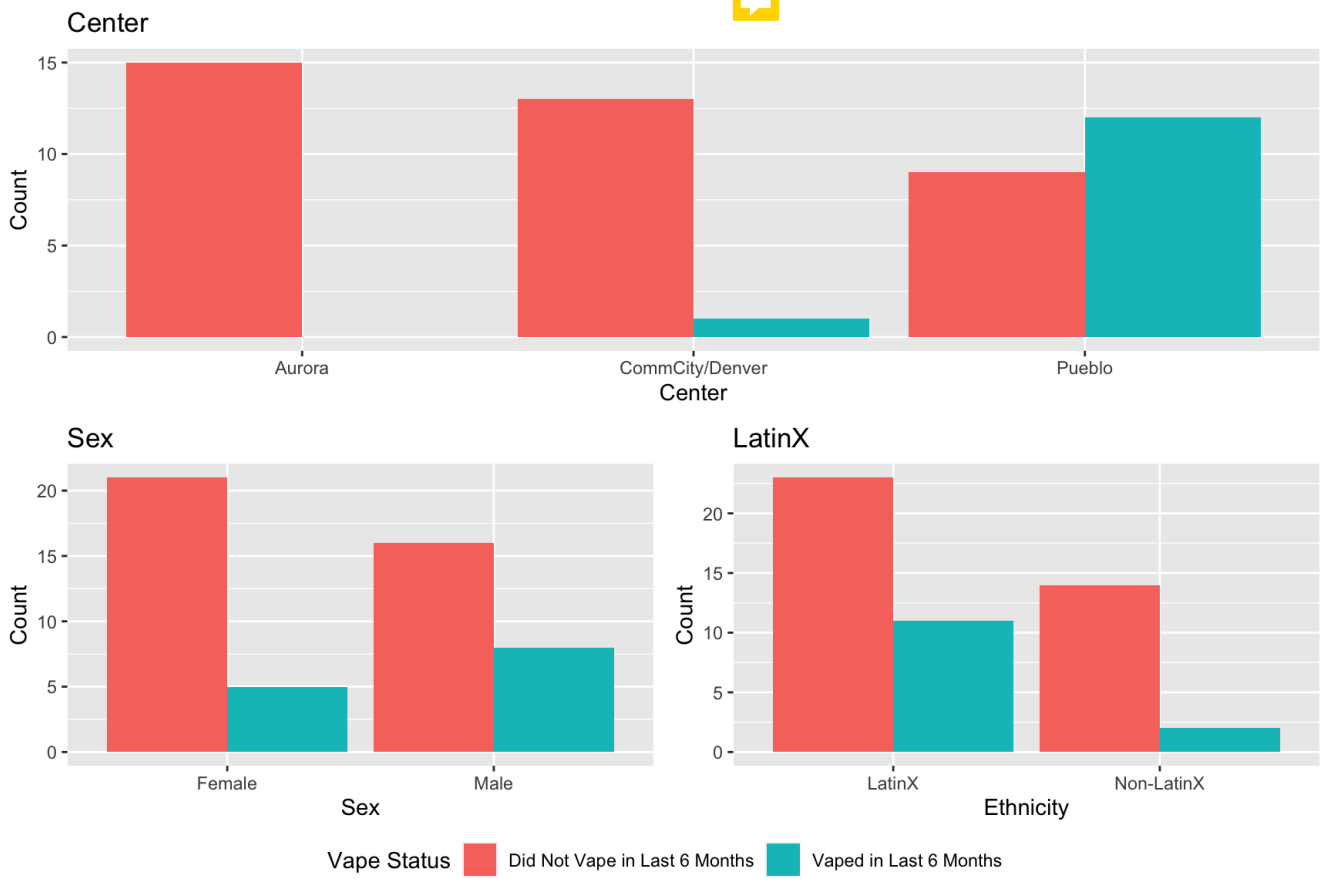


Figure 2: Pulmonary Function

Figure 2 is a visualization of the pulmonary function ($\frac{FEV1}{FVC}$) and IOS (*R5* and *X20*) variables. $\frac{FEV1}{FVC}$ was only completed by **n = 22** individuals from the study population. *R5* and *X20* represent **n = 49** and **n = 44** individuals, respectively.

Code

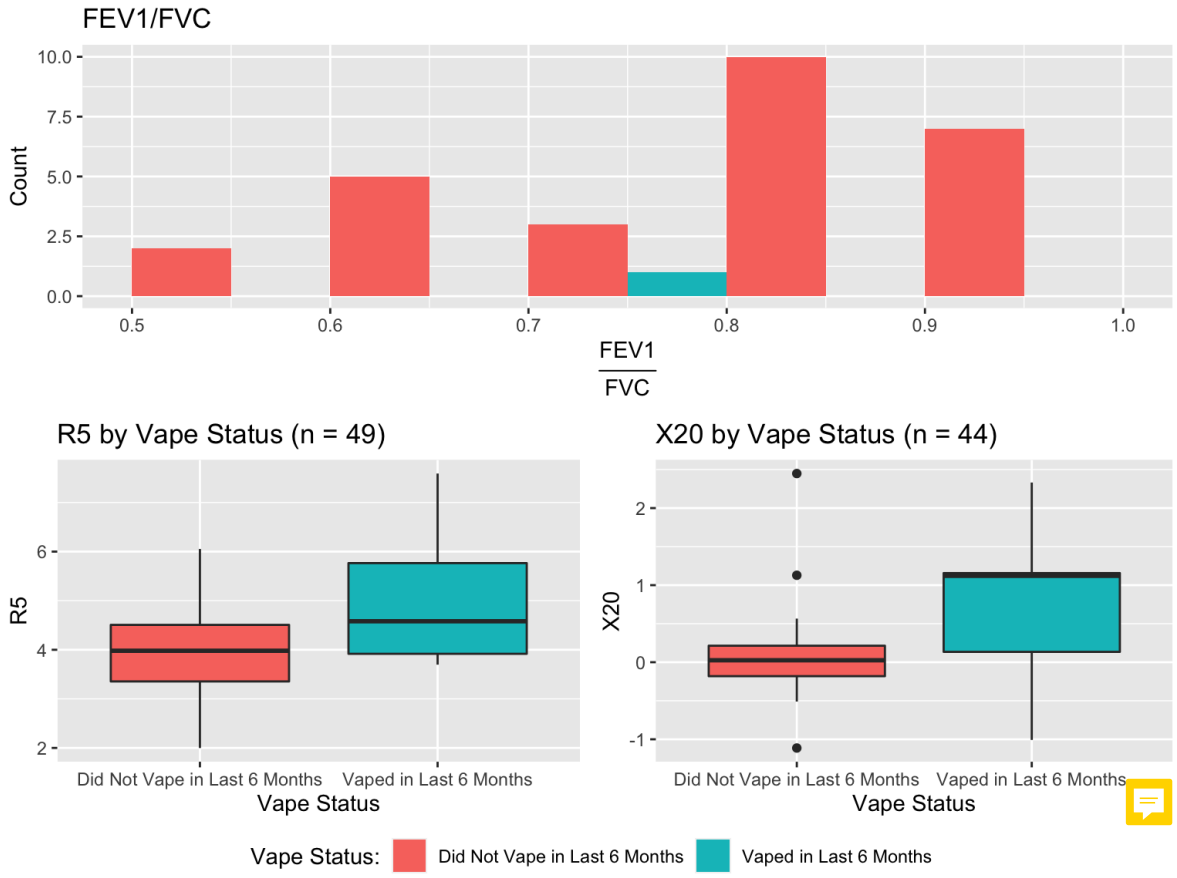


Table 2: Comparison of Gene Filtering Parameters

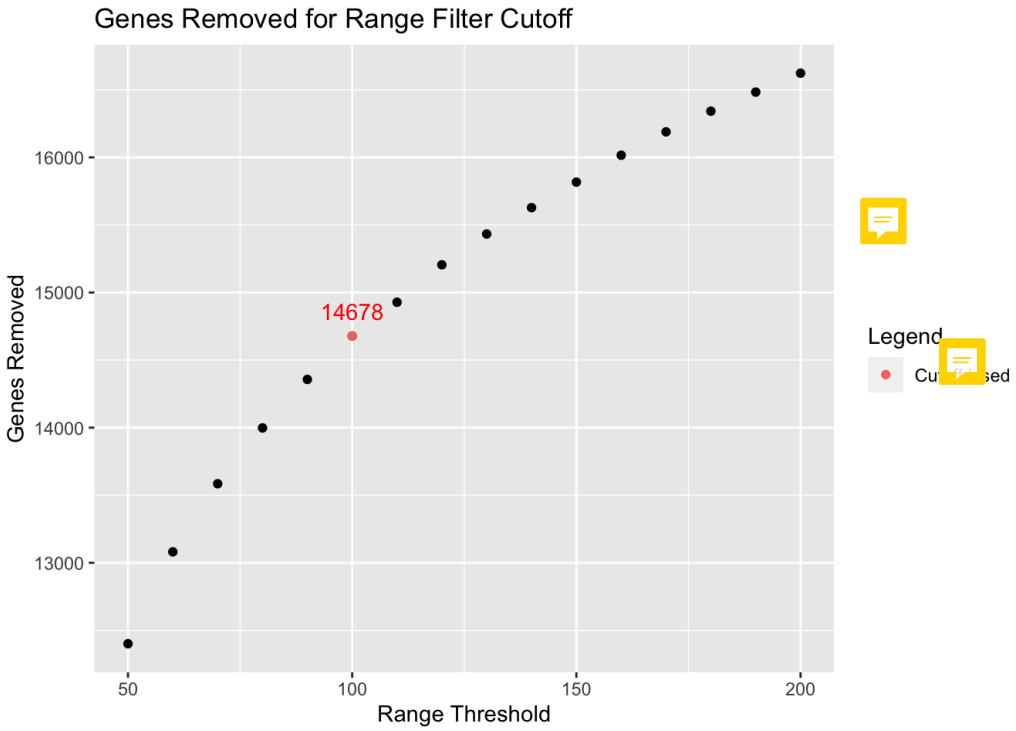
The table below compares gene-filtering parameters from previous analyses to a parameter presented by the creators of the RUVseq package.

Code

Filter	Analysis Used	Inclusion Criteria	Read Count Before	Read Count After	Reads Removed
1	Previous	At least 25% of the samples have > 0 reads	60651	31505	29146
2	Previous	The range of reads across all samples < 100	31505	14645	16860
3	Current	>5 reads in at least 2 samples (Bioconductor)	60651	29141	31510

After reviewing the comparison of filters, it appears that filtering parameters presented by the creators of RUVSeq may be overly conservative for this application. Analyses will proceed using the same filters as previous analyses. The first filter will remove genes with 0 reads in more than 75% of samples, and the second will remove genes with low variation (range of reads < 100) that may be considered “house-keeping” genes.

Figure 3: Genes removed for read-count cutoff values The figure below is used to visually assess how many genes are removed across cutoff values for the range of reads for each gene across the 49 samples after filtering out genes that have 0 read counts in 75% or more of the samples. The red line represents the number of genes (14678) removed at the cutoff range of 100 (the value used in previous analyses).



After reviewing the comparison of filters, it appears that filtering parameters presented by the creators of RUVSeq may be overly conservative for this application. Analyses will proceed using the same filters as previous analyses. The first filter will remove genes with 0 reads in more than 75% of samples, and the second will remove genes with low variation (range of reads < 100) that may be considered “house-keeping” genes. This analysis will include a total of **n = 14645** genes.

Figure 4: Relative Loge Expression and Principal Component Analysis Prior to RUV

The following figure shows the Relative Log Expression (RLE) and Principal Component Analysis (PCA) of read counts for each sample without the use of RUVr or any other transformation technique. Sample 23 (SID = 144) has been removed due to missing vape status.

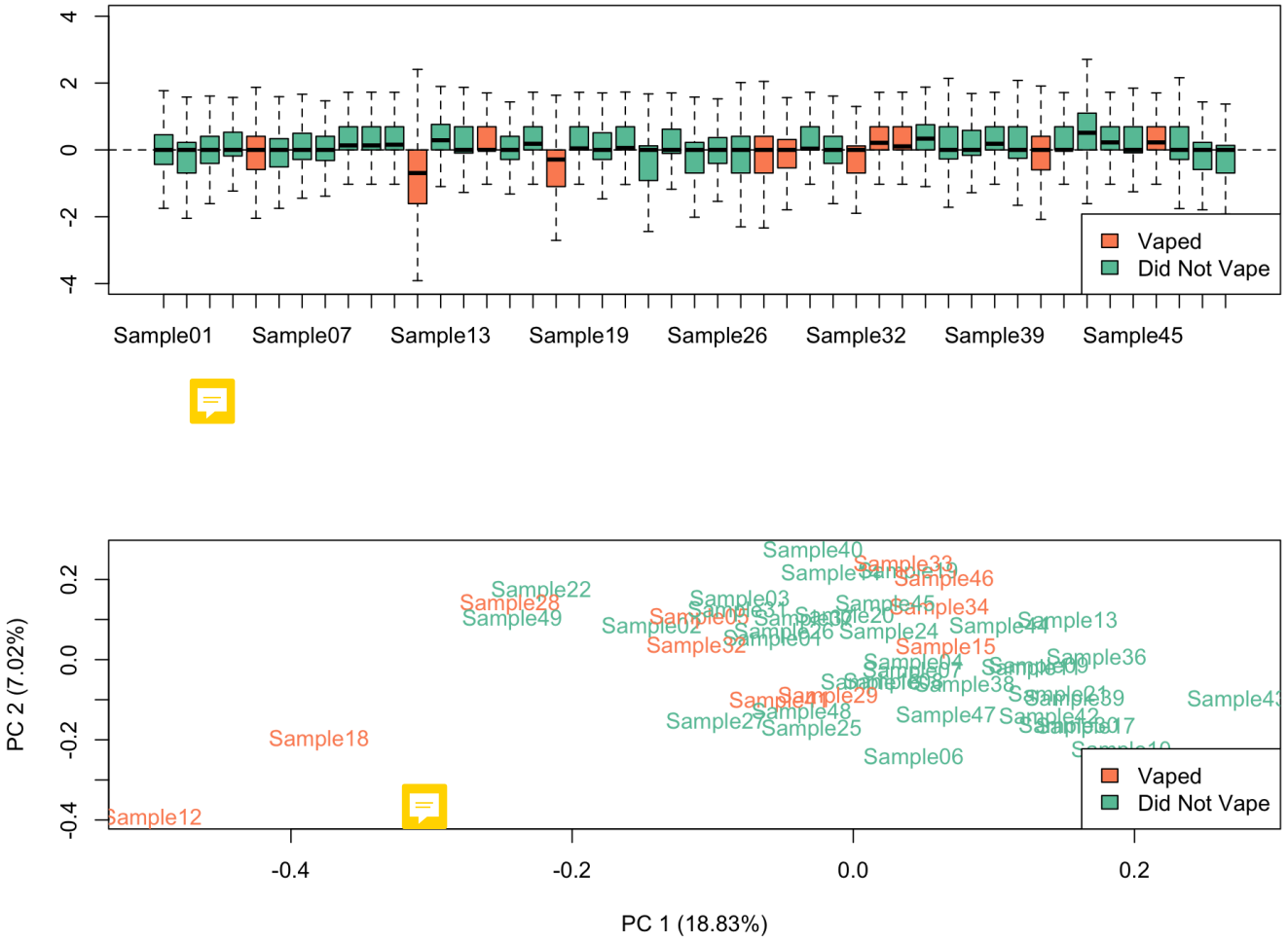
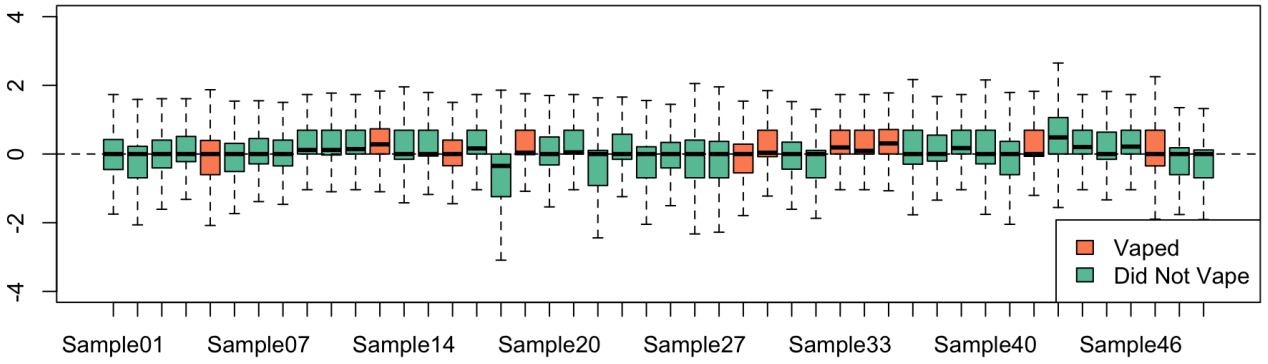
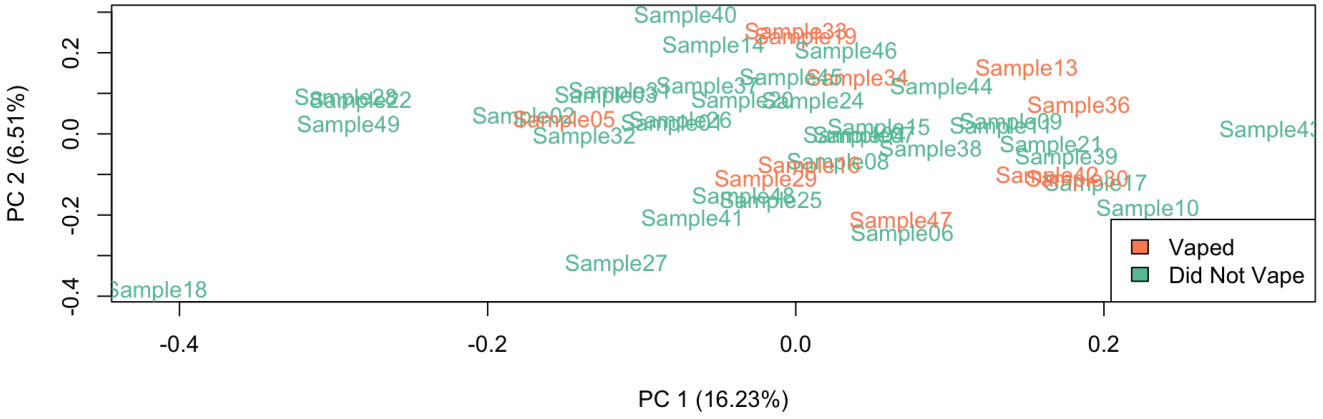


Figure 5: RLE and PCA Excluding Outlier Samples It appears that Sample 12 (SID = 102) may be an outlier. Below are the RLE and PCA plots with the sample removed.





Both the RLE and PCA plots appear to improve when Sample 12 is excluded. All following analyses will exclude Sample 12 (SID = 102).

Figure 6: Elbow Plot Comparison of RUVr Factor Inclusion

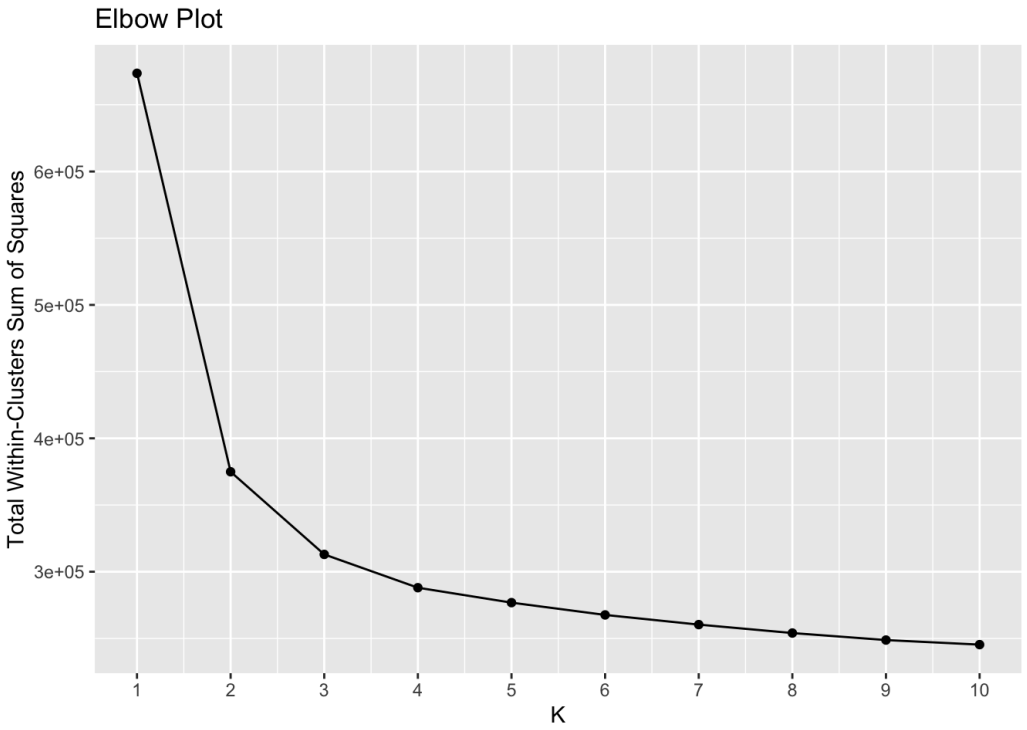
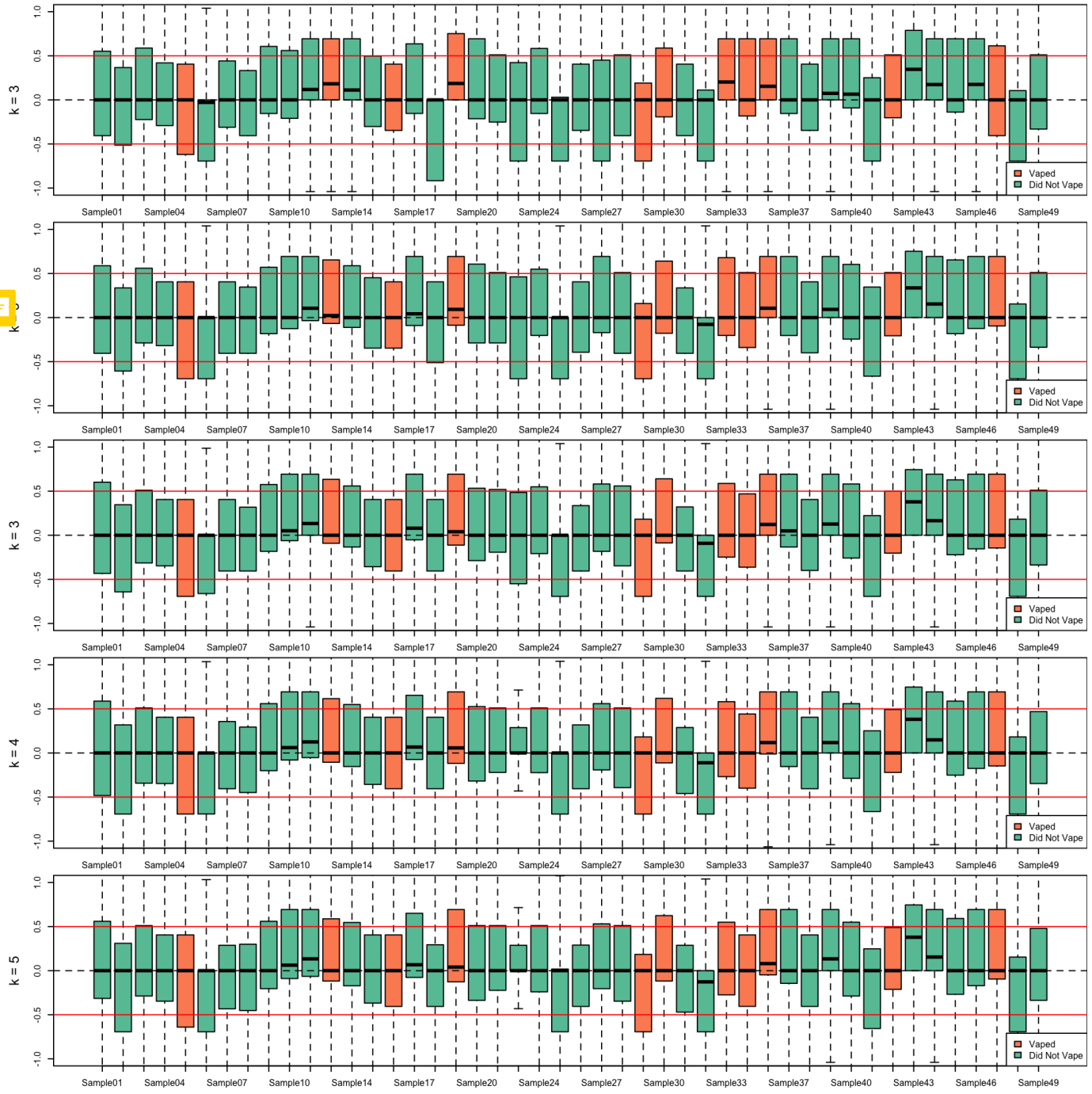
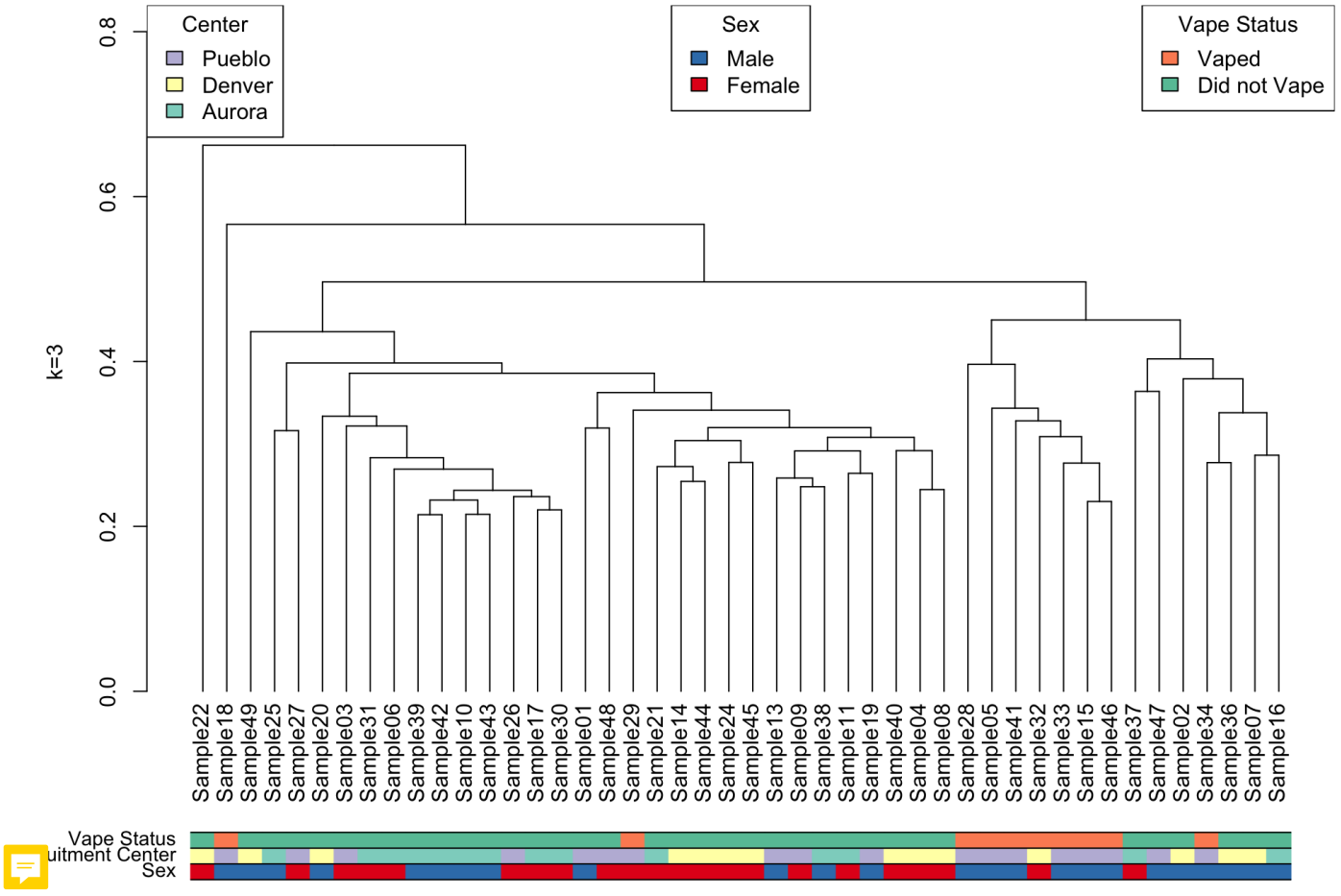
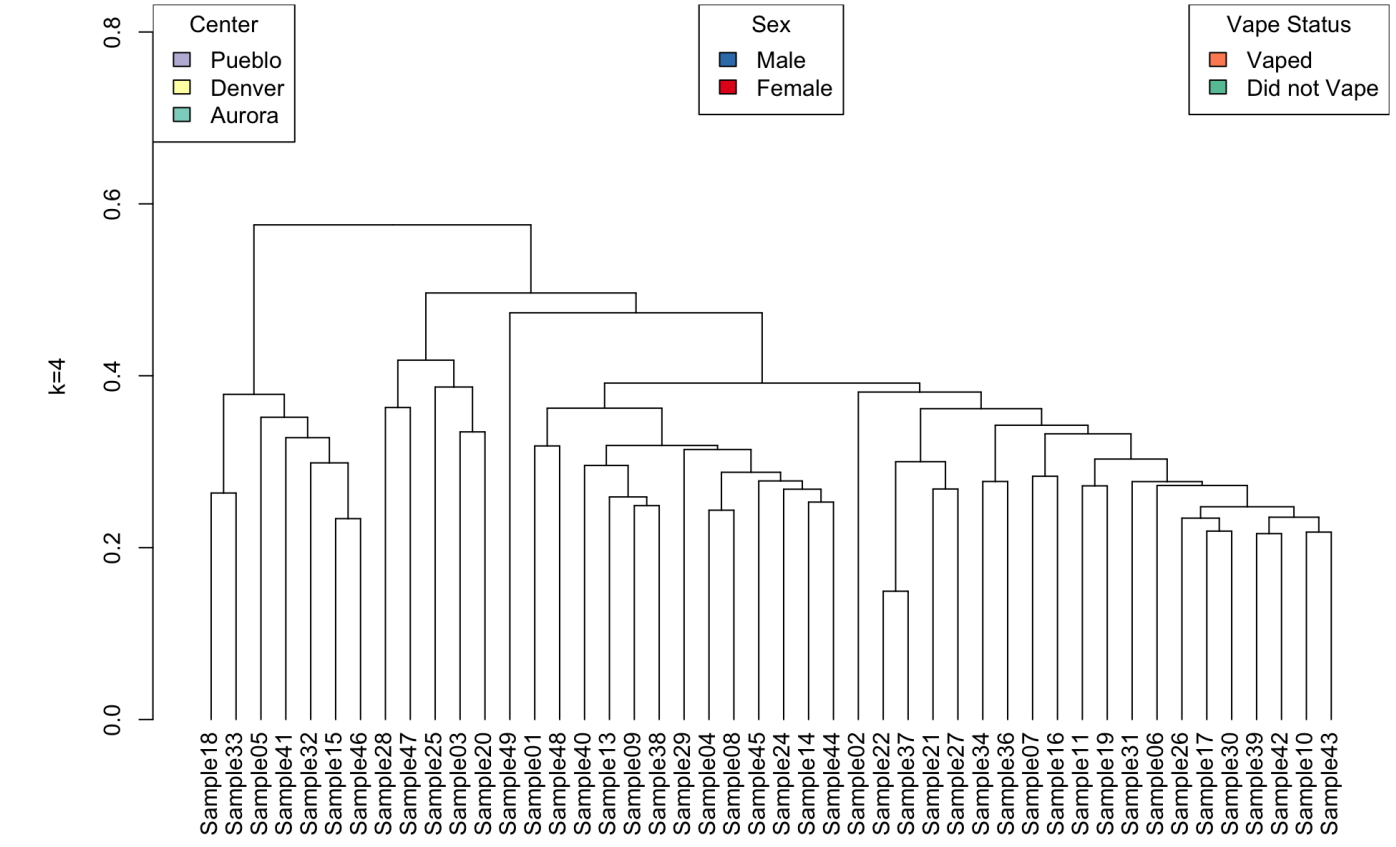


Figure 7: RLE Comparison of RUVr Factor Inclusion To visually inspect for the best cutoff for factor analysis, RUVr was run for k = 1 through k = 5 factors. The RLE plots for each are presented below.



The dendrograms for k = 3 and k = 4 are compared below


Center
Pueblo
Denver
Aurora
Sex
Male
Female
Vape Status
Vaped
Did not Vape
k=3
0.0
0.2
0.4
0.6
0.8
Sample22
Sample18
Sample49
Sample25
Sample27
Sample20
Sample03
Sample31
Sample06
Sample39
Sample42
Sample10
Sample43
Sample26
Sample17
Sample30
Sample01
Sample48
Sample29
Sample21
Sample14
Sample44
Sample24
Sample45
Sample13
Sample09
Sample38
Sample11
Sample19
Sample40
Sample04
Sample08
Sample28
Sample05
Sample41
Sample32
Sample33
Sample15
Sample46
Sample37
Sample47
Sample02
Sample34
Sample36
Sample07
Sample16
Vape Status
uitment Center
Sex


Center
Pueblo
Denver
Aurora
Sex
Male
Female
Vape Status
Vaped
Did not Vape
k=4
0.0
0.2
0.4
0.6
0.8
Sample18
Sample33
Sample05
Sample41
Sample32
Sample15
Sample46
Sample28
Sample47
Sample25
Sample03
Sample20
Sample49
Sample01
Sample48
Sample40
Sample13
Sample09
Sample38
Sample29
Sample04
Sample08
Sample45
Sample24
Sample14
Sample44
Sample02
Sample22
Sample37
Sample21
Sample27
Sample34
Sample36
Sample07
Sample16
Sample11
Sample19
Sample31
Sample06
Sample26
Sample17
Sample30
Sample39
Sample42
Sample10
Sample43

Based on the elbow plot, dendrograms, and RLE plots above, it appears that there is not significant improvement between K = 3 and K = 4. Analyses will proceed with K = 3 for RUVr.
