Analyzing High-Dimensional RNA Data with Principal Component Analysis

Michael Li

08/27/2020

Often when analyzing genetic data, either DNA or RNA, there are an extremely large number of feature genes. This makes it difficult to correlate features to other variables (e.g. age, gender, disease status) and identify important distinguishing features as the number of independent variables (feature genes) is often much larger than the number of samples. To tackle this statistical challenge, Principal Component Analysis is often used to distill data down to lower dimensions prior to further investigation.

1. Objectives

The purpose of this analysis was to project the samples of data from tens of thousands of features into only 2 dimensions through Principal Component Analysis (PCA) in an effort to identify trends across the samples. Through this reduction of data dimension, this statistical analysis was intended to identify features that distinguish samples by Age groups and/or by Sex.

First the dataset used will be described in Section 2 and methods utilized to explore and analyze the data will be outlined in Section 3. Plots and generated test results will be provided in Section 4 of this report, and then discussion and conclusions will be covered in Section 5.

2. Description of the Dataset

The dataset consists of a data matrix (data_matrix.csv) with 48440 genetic features (large p=48440) for 16 RNA/RNA group sample observations (n=16). Another table (sample_annotation.xlsx) is included with information about each of the 16 samples: Age group (Young/Aged) and Sex (Female/Male).

3. Methods

This analysis was conducted using the Python programming language on the JupyterLab platform. The packages utilized include: Matplotlib (pyplot and patches), pandas, NumPy, seaborn, scikit-learn (StandardScaler from preprocessing and PCA from decomposition), SciPy (ttest_ind and f_oneway from stats), statsmodels (ols).

The data_matrix.csv was imported into JupyterLab as a dataframe, which was then cleaned to remove its default indexing and transposed. Sample annotations were imported as a separate dataframe. It is noted here that its Age field for RNA 4 included an additional space that required modification later on for proper analysis. Preliminary analysis in the form of checking dimensions, counts, mean, standard deviations, quartile identifications was conducted for each of the 48440 features. Null values were also

checked for.

The function StandardScaler within the scikit-learn package was subsequently used to fit and transform the data matrix. An initialized PCA algorithm was then fitted to and used to transform the scaled data. The data matrix was reduced to an array/dataframe of 16 samples and 2 principal components (PC1 and PC2). The data was explored with a combination of scatter plots, box plots, and heat maps. Boxplots of each principal component were generated, first against Age groups then against Sex. Grouped boxplots including both variables for both principal components were also created. Following extraction of PCA component data, a heatmap was generated with each principal component mapped against each gene feature.

Because PCA component data represents the correlation (with both positive and negative loadings) between each principal component and each gene feature, the 5 genes with the greatest magnitude of correlation to each principal component were identified. These are helpful in identifying the features that best distinguish Age and Sex based on the principal component most associated with each variable.

With regards to statistical analyses, t-tests, one-way ANOVA, and two-way ANOVA were conducted for both principal components obtained for each sample annotation variable (Age and Sex) to identify using p-value which variables are statistically significant/impactful to each principal component.

4. Results

4.1 Principal Components

In this analysis, the focus will be on the first two principal components, which turn out to account for 31.4% and 8.3% respectively, of the total variation in the dataset. The next 8 principal components represent 7.5%, 6.4%, 5.9%, 5.3%, 4.8%, 4.5%, 4.3%, and 3.9% of total variation respectively, or approximately 80% in total with all first 10 principal components.

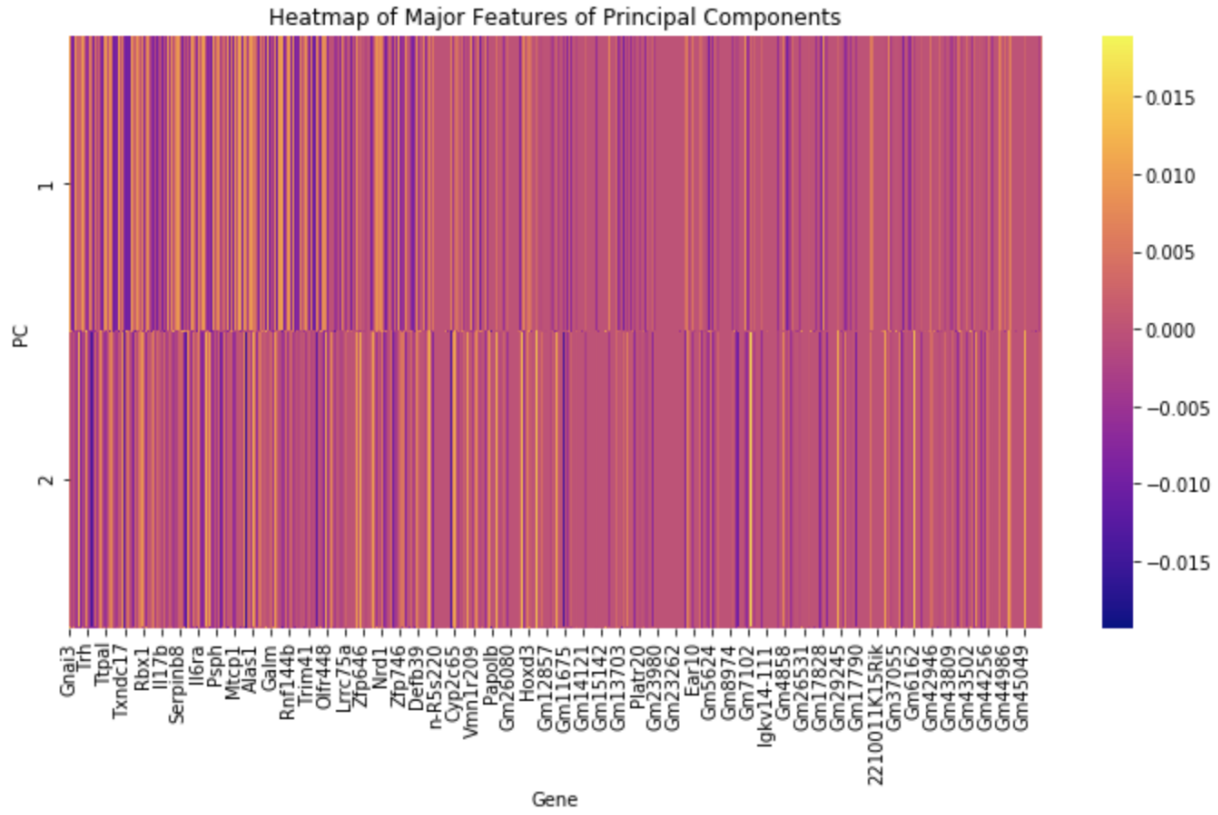
Table 1 shows a table of the features most correlated to each principal component, defined by magnitude of loadings.

Figure 1 is a heatmap of correlation between each principal component and each gene feature.

Table 1: Most Correlated Gene Features To Each Principal Component

Gene	PC1	Gene	PC2
Elf4	0.010536	Dnajc8	0.019294
Ankrd44	0.010524	Bms1	0.018986
Elmo1	0.010520	Fars2	0.018983
Wipf1	0.010513	Cox5a	0.018977
Ptpre	0.010504	Nop14	0.018849

Figure 1: Heatmap of Dominant Features of Each Principal Component



4.2 Exploratory Data Analysis

Figure 2 shows scatter plots of the second principal component (PC2) vs. first principal component (PC1). Samples are color-labelled by Age group in A and by Sex in B.

As displayed in Figure 2A, there is a very clear and striking separation along the first principal component between the two Age groups, with all the Aged group samples below -50 while all the Young group samples are above 50. No such clear separation by Age group can be observed for the second principal component in Figure 2A.

In Figure 2B, Sex does not appear to have any significant impact on either principal component, as both Female and Male samples were interspersed across the chart without any clear separation. In terms of each principal component, Female samples are generally located at lower values of the first component compared to Male samples, while Male samples are generally located at lower values of the second component compared to the Female samples.

Figure 2: Scatter Plots of PC2 vs PC1 by Age Group and by Sex

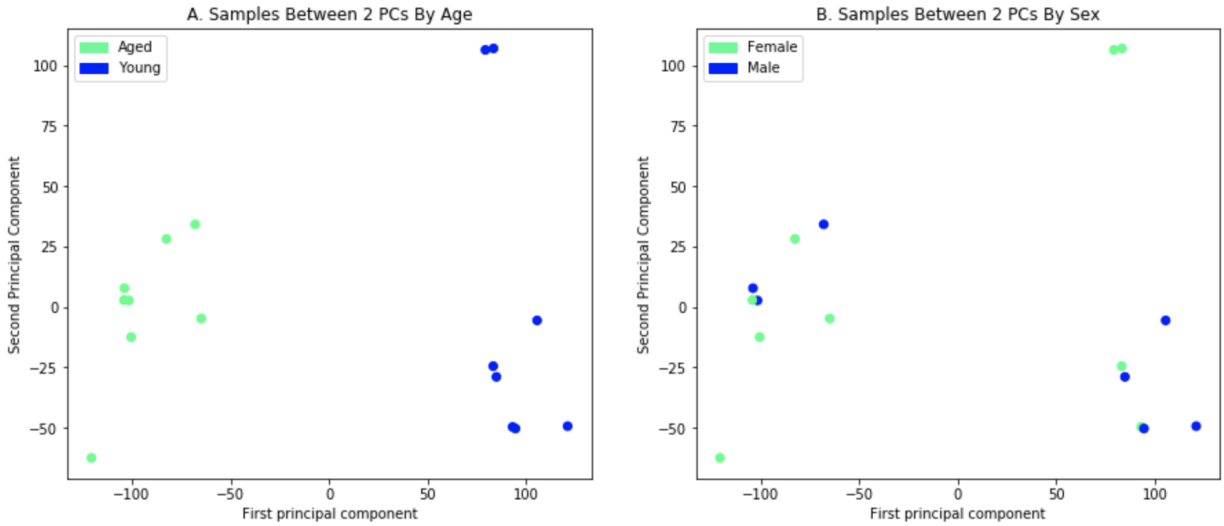


Figure 3 shows box plots of PC1 or PC2, by Age (3A and 3B) or Sex (3C and 3D).

In Figure 3A, group mean are clearly different between the Young group and Aged group on PC1, suggesting that Age is a significant factor to PC1. However, there is no clear difference between Age groups in PC2 as shown in Figure 3B.

In both Figures 3C and 3D, the box plots overlap largely and confirm that Female samples are generally located at lower values of the first component compared to Male samples, while Male samples are generally located at lower values of the second component compared to Female samples.

Figure 3: Box Plots of PC1 vs Age, PC2 vs. Age, PC1 vs Sex, PC2 vs. Sex

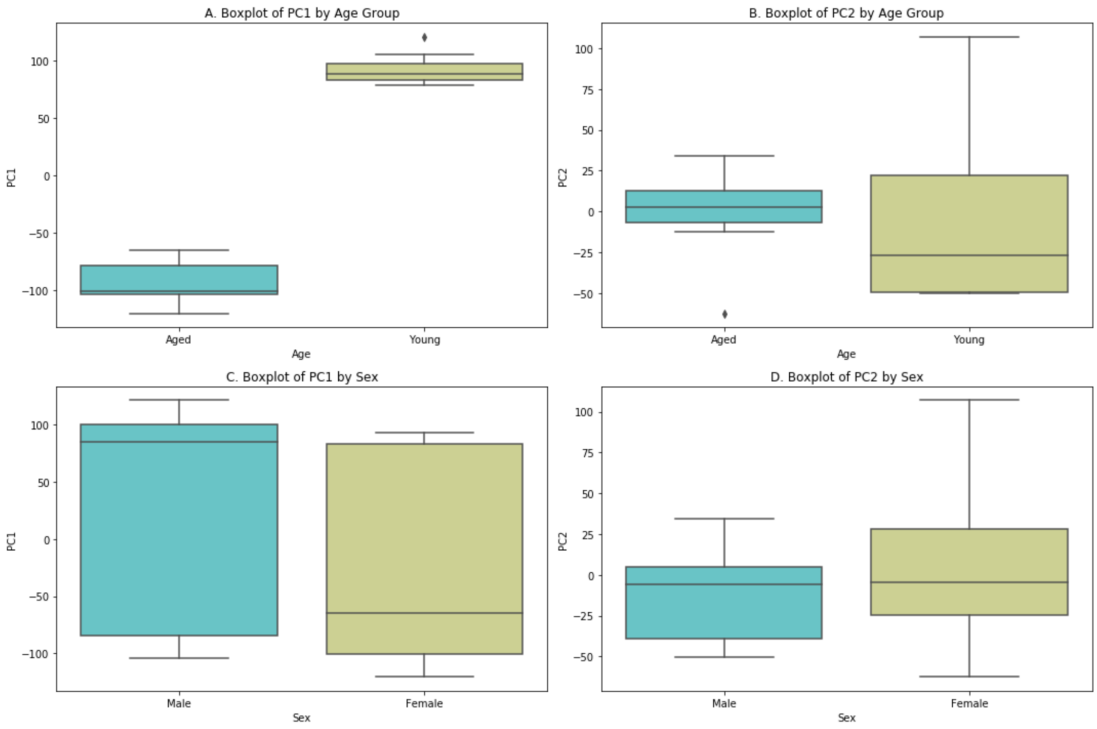


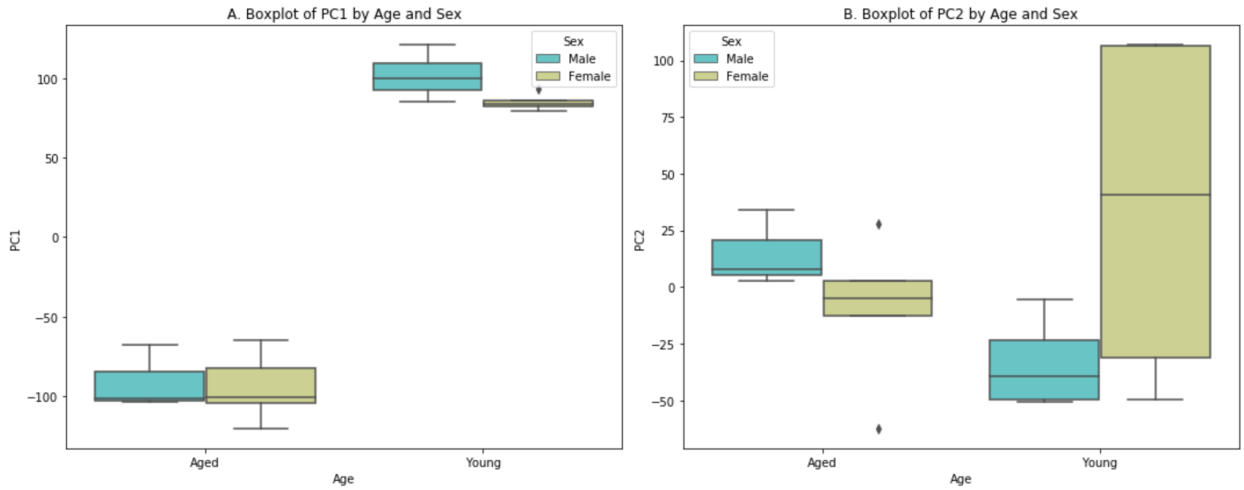
Figure 4 consists of box plots of PC1 (4A) and PC2 (4B), by Sex and Age group.

As indicated by Figure 4A, Age is highly significant for PC1 even after accounting for Sex. Between the two Age groups for Males, there exists a large difference in group mean. There is a similar and clear difference in group mean between the two Age groups amongst the Females.

Figure 4B suggests that there may be an interaction between Age and Sex in term of PC2 because the directions of group mean differences between Aged/Young samples is the opposite of that between Male and Female groups.

Figure 4B also displays a separation between Aged Male and Aged Female samples as opposed to partially overlapping box plots between Young Male and Young female samples. Within the Aged group, the group mean of PC2 for Females is lower than that of Males, while the group mean of PC2 among Males is lower than that of Females in the Young group.

Figure 4: Box Plots of PC1 vs. Age Grouped by Sex and PC2 vs. Age Grouped by Sex



4.3 Statistical Tests

As suggested by the scatter plot (Figure 2A) and the box plots (Figures 3A and 4A), there is a clear separation in PC1 between the two Age groups. This is confirmed by the t-test in Table 2, where the null hypothesis is that PC1 among Young samples and PC1 among Aged samples is equal. The t-test produced a p-value of 3.0640e-12 (<0.05), this results in the rejection of the above null hypothesis. Similar results are observed in the following one-way ANOVA.

As expected based on previous plots, no other p-values are statistically significant.

Table 2: Results of t-test and One-way ANOVA of Principal Components by Age or Sex

t-test			
		t-statistic	p-value
PC1	Age	21.94	3.0640e-12
	Sex	0.67	0.5114
PC2	Age	0.04	0.9653
	Sex	-0.90	0.3848
One-way ANOVA			
		F statistic	p-value
PC1	Age	481.15	3.0640e-12

	Sex	0.45	0.5114
PC2	Age	0.002	0.9653
	Sex	0.80	0.3848

In the two-way ANOVA shown in Table 3, even after account for Sex, Age group remains highly significant to PC1 with a p-value of 1.16e-11.

Neither Age or Sex variables are significant for PC2. As a result of a two-way ANOVA, the p-values of 0.88 and 0.40 for Age and Sex respectively do not make it clear that either variable is significant for PC2.

The two-way ANOVA model with interaction between Sex and Age shows that the interaction between Age and Sex for PC2 is close to being significant with a p-value of 0.0792.

Table 3: Results of Two-way ANOVA for Principal Components by Age and Sex

	sum_sq	df	F	p-value
PC1				
C(Age)	134839.07	1	482.41	1.1614e-11
C(Sex)	408.52	1	1.46	0.2482
PC2				
C(Age)	65.21	1	0.02	0.8799
C(Sex)	2115.40	1	0.77	0.3960
PC 1 With Interaction				
C(Age)	134839.07	1	467.56	5.5984e-11
C(Sex)	408.52	1	1.42	0.2570
C(Age):C(Sex)	172.93	1	0.60	0.4537
PC2 With Interaction				

C(Age)	65.214650	1	0.028646	0.8684
C(Sex)	2115.40	1	0.93	0.3541
C(Age):C(Sex)	8373.11	1	3.68	0.0792

5. Discussion and Conclusion

The first two principal components identified by the PCA, PC1 and PC2, account for roughly 40% of the overall variation in the dataset. In PC1, the top 5 dominating features are Elf4, Ankrd44, Elmo1, Wipf1, and Ptpre. In PC2, the top 5 dominating features are Dnajc8, Bms1, Fars2, Cox5a, and Nop14.

As shown in Section 4, Age is a significant factor to PC1, even when accounting for Sex. Hence, the most impactful features of Age are the dominating features included in PC1. Sex does not have a significant impact on PC1.

Neither Age or Sex are significant for PC2. However, within the Aged group, the group mean of PC2 for Females is lower than that of Males, while the group mean of PC2 among Males is lower than that of Females in the Young group. Based on this indication, the most impactful features of Sex are the dominating features included in PC2.

There appears to have an interaction effect between Age group and Sex on PC2, with a close to being significant p-value of 0.0792. This is worthy of further investigation, especially if a larger sample size is available.