Bioinformatics Bite #3: Intro to PCA Plots

Worksheet
Teacher Materials

OVERVIEW

This activity introduces students to data visualizations relevant to omics studies. A Principal Component Analysis (PCA) plot is a common method of processing large data sets.

KEY CONCEPTS

- Traditional graphs are an example of one method for visualizing data.
- Principal Component Analysis (PCA) is another method of visualizing data, especially when pertinent to large data sets.

OBJECTIVES

- Students will be able to define what a PCA plot is and justify its role as a data analysis tool.
- Students will be able to read a PCA plot and evaluate the utility of information from it.
- Students will be able to navigate through the NASA GeneLab database and use one tool from the Galaxy platform.

TEACHING TIPS

- **Refresh students on the purpose of graphs, and their importance in visualizing data.**
- **Communicate with NASA GeneLab if you will have your class participate in generating the PCA plot using the Galaxy platform (https://genelab.nasa.gov/help/contact) in order to ensure that the class can be accommodated for data processing power.**
- **This lesson can be used to promote effective-notetaking skills. A teacher can encourage students to take short sets of notes, rather than writing everything they see/hear.**
- **Additional notes for each step are annotated in red text throughout this teacher guide.**

BIOINFORMATICS BITE #3: PCA PLOTS

Part 1: Brainstorm

Brainstorm: "If you measure the expression of 15 genes from 60 mice, and the data come back as a 15×60 table, how do you make sense of all that?" (Ngo, 2018).

Student answers will vary!

Encourage students to brainstorm and talk in partnerships about their answer to this question.

Sample student answers:
- I would try to group mice that had similar gene expression overall (or by differences)
-I would try to group the genes by putting them in categories of what similarities they are responsible for.

What would be the challenges of trying to make sense of that?

Student answers will vary!

Encourage students to brainstorm and talk in partnerships about their answer to this question.

Sample student answers:
-Dataset is large
-Dataset is complex
-How do you know what mice are similar or different?
-How do you know what genes are responsible for being similar/different?
-Lots of potential different options

Part 2: PCA Plots--Video Introduction

Watch StatQuest: PCA main ideas in only 5 minutes!!! and take notes on the main ideas related to PCA plots.

Be sure to sketch a PCA plot when he starts explaining them, your teacher will look for that as part of your notes!

Video Notes (including PCA plot):

Student answers will vary and include sketches

Sample student notes:

-We can't observe the differences from the outside so we sequence the mRNA in each cell to identify which genes are active.

-Ex: Each column shows how much each gene is transcribed in each cell

-2 sample example
- **-plot it on a graph**
- **-Cell 1 and 2 have an inverse correlation which means they are probably two different cells since they are using different genes**

-Three sample example
- **-plot it on multiple graphs (each comparing two cells)**
- **-Cell 1 and 3 have positive correlation, suggesting they are doing something similar**
- **-cell 2 and 3 have a negative correlation, suggesting cell 2 is doing something different than cell 3**
- **-Could try plotting it on a three dimensional graph**

-What do we do when we have four or more samples?
- **-We draw a Principal Component Analysis Plot (PCA Plot)**

-A PCA plot converts the correlations (or lack thereof) among all the cells into a 2-D graph

-Cells that are highly correlated cluster together

-To make clusters easier to see, we can color code them
    -Once we've identified the clusters in the PCA plot, we can go back to the original cells, and see that they represent three different types of cells doing three different things with their genes

-IMPORTANT for interpreting PCA plots--The axes are ranked in order of importance
    -Differences along the first principal component axis (PC1) are more important than differences along the second principal component analysis (PC2)
    -So even if there is equal distance between two clusters on PC1 as there is between two clusters on PC2, the two clusters on PC1 are more different

-PCA is just one way to make sense of this type of data, many methods for dimension reduction
    -I.e. Heatmaps, t-SNE, MDS

Part 3: PCA Plots--Reading & Notes

Your teacher will assign you a section of the chart to take notes in the highlighted portion of the table below (unless you are assigned *Introduction/1, which will use the article). Don't be afraid to sketch a PCA plot in your notes if it is relevant, it is encouraged! When you are done, you will collaborate with your classmates to fill in the rest of the table.

Walk around and facilitate student collaboration as they work! Give students 5-7 minutes to work on their individual section and then put them in groups to teach each other.

The notes here contain a lot of extra detail/copied phrases from the passages! You may want to encourage students to summarize more and put it in their own words--as well as add their own sketches/snips of the PCA plots (not shown here, but if you'd like to see pictures, click the article link).

Section	Notes-Sample Notes
Introduction Principal component analysis explained simply	-We are entering an era of Big Data -PCA is short for principal component analysis -PCA is a way to bring out strong patterns from large and complex datasets -Ex: How to measure the expression of 15 genes from 60 mice that results in data coming back as a 15*60 table -PCA takes the expression data of 15 genes from each mouse and smooshes them down to one single dot that represents the expression profile of that mouse. One dot for one mouse-sixty in total. -This makes it much easier to compare mice. Those with similar expression profiles will cluster together.
1. Principal components capture the most variation in a dataset	-Main focus: we are trying to compare the mice

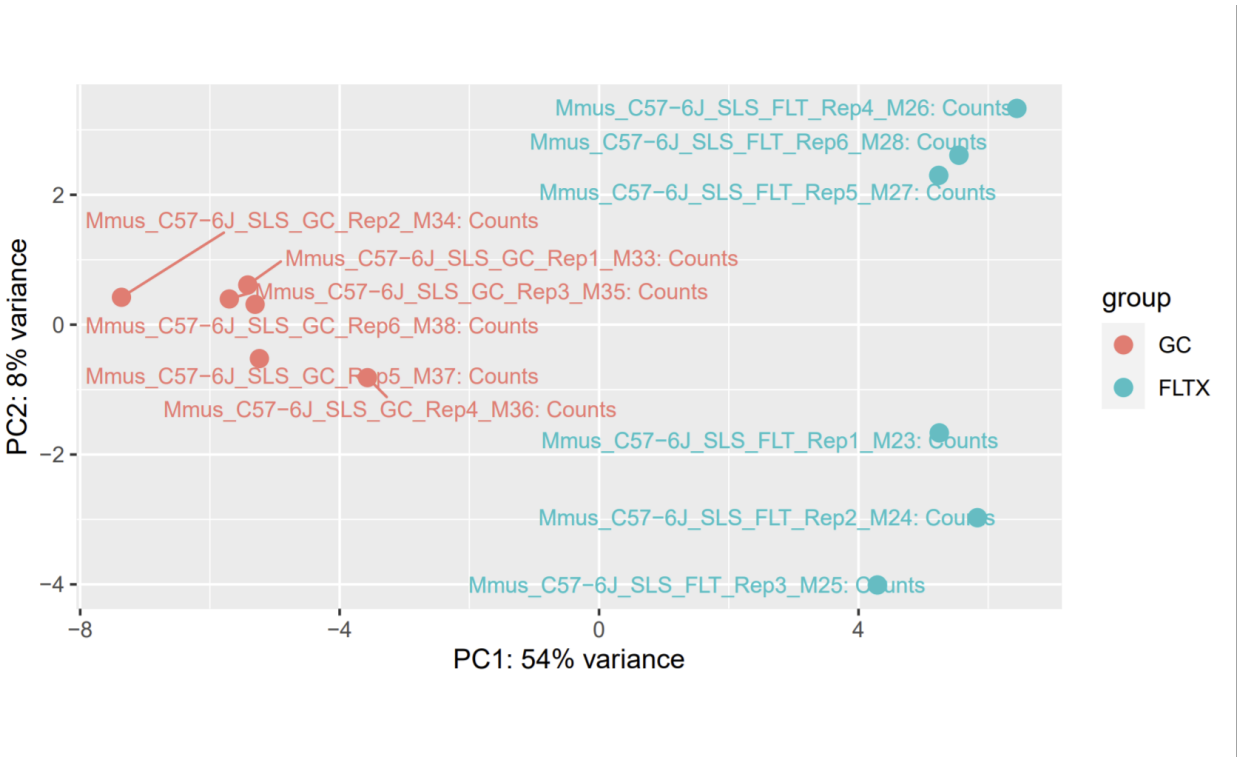
Principal component analysis explained simply	-each dot carries read counts of 2 genes from one mouse, and together they form a flat "cloud." Principal component 1 (PC1) is a line that goes through the center of that cloud and describes it best. -The total distance among the projected points is maximum...so they can be distinguished from one another as clearly as possible. -The total distance from the original points to their corresponding projected points is minimum. This means we have a representation that is as close to the original data as possible. -The best line — our PC1 — must convey the maximum variation among data points and contain minimum error.
2. PCA deals with the curse of dimensionality by capturing the essence of data into a few principal components. Principal component analysis explained simply	-The more genes you've got, the more axes (dimensions) there are when you plot their expression, so this is where PCA can be VERY helpful! -PC2 is the second line that meets PC1, perpendicularly, at the center of the cloud, and describes the second most variation in the data. -If PCA is suitable for the data, just the first 2 or 3 principal components should convey most of the information and makes it possible to see strong patterns. -Don't have to throw away any genes in doing so. -Principal components take all dimensions and data points into account. -Since PC1 and PC2 are perpendicular to each other, we can rotate them and make them straight. These are the axes of our pretty PCA plot.
3. Dimensions vary in the weights they have on each principal component. Principal component analysis explained simply	-The original dots (mouse gene expression) are pinned in place by their values on all axes (genes). These dots are projected on PC lines, which are trying to stay as close as possible to all the dots -Dimensions have different weights in defining each PC. -Finding out where they put Mouse #1 on PC1 is simple math: (gene1 read count * gene1 weight on PC1)

	+ (gene2 read count * gene2 weight on PC1) **+ …** **+ (gene15 read count * gene15 weight on PC1)** **= PC1 value of Mouse #1** **Mouse #1 has a spot on PC2, too. To calculate it, use the same formula with weights of genes on PC2 instead of PC1.** **With a value of PC1 and a value of PC2, Mouse #1 now can be graphed as a dot on the PCA plot.** **-Do this math again on all the mice, and they will each become a dot on the PCA plot**
4. How to read a PCA plot **Principal component analysis explained simply**	**-Mice that have similar expression profiles are now clustered together.** **-If 2 clusters of mice are different based on PC1, such differences are likely to be due to the genes that have heavy influences on PC1.** **-If 2 clusters are different based on PC2, like the red and blue clusters, then the genes that heavily influence PC2 are likely to be responsible.** **-PC1 reveals the most variation, while PC2 reveals the second most variation.** **-Differences among clusters along PC1 axis are actually larger than the similar-looking distances along PC2 axis.** **-Is this plot meaningful? PCA is worthy if the top 2 or 3 PCs cover most of the variation in your data. Otherwise, you should consider other dimension reduction techniques, such as t-SNE and MDS**
***Introduction/1. A PCA plot shows clusters of samples based on their similarity.** **How to read PCA biplots and scree plots**	**-Principal component analysis (PCA)** **- PCA is a tool to bring out strong patterns from complex biological datasets.** **-PCA captures the essence of the data in a few**

	principal components, which convey the most variation in the dataset. **-PCA does not discard any characteristics** **-PCA reduces the overwhelming number of dimensions by constructing principal components (PCs).** **-PCs describe variation and account for the varied influences of the original characteristics.** **-Such influences, or loadings, can be traced back from the PCA plot to find out what produces the differences among clusters.**
Principal Component Analysis **Principal Component Analysis (Excerpt)**	**-PCA is a dimension reduction method** **-PCA transforms a large set of data into smaller ones/reduces the number of variables in the dataset without reducing the amount of overall data present** **-Demonstrates the first two dimensions from the PCA run on the normalized counts of the samples.** **-It shows the samples in the 2D plane spanned by their first two principal components.** **-Each replicate is plotted as an individual data point.**

Part 4: Practice--the GLDS-104 PCA Plot

GLDS-104 PCA Plot



Go to https://genelab.nasa.gov

Then click on the Data Repository button

Then search "GLDS-104" in the provided search bar.

Then click on GLDS-104: Rodent Research-1 (RR1) NASA Validation Flight: Mouse soleus muscle transcriptomic and epigenomic data

Use the instructions above to answer the first two questions and enhance your understanding in general! Use the graph for all the questions!

For reference, questions 4, 5, and 6 and answers are pulled from the GL4HS Teacher's Manual!

1. **What is mus musculus (Mmmus)? (Go to the organisms tab).**

   Mouse

2. **What were the two differing conditions mus musculus was exposed to in this experiment? (Go to the study description tab).**

   Spaceflight and ground control

3. **Match the conditions to their color/acronym on the PCA Plot.**

   Spaceflight-blue/FLTX Ground control-orange/GC

4. **What is the first dimension (PC1) separating?**

The first dimension is separating the two factors-spaceflight vs ground control.

5. **And the second dimension (PC2)?**

Your answer: The second dimension is separating differences within the two factors e.g. there are two groupings of ground control samples within the second dimension but not in the first dimension.

6. **What can we conclude about the DESeq design (factors, levels) we choose?**

Your answer: Because variance of the first dimension is greater than the variance of the second dimension, this means that the differences between the two factors are greater than the differences within either factor. It is important to look at both the distribution on the plot and the numerical value associated with the variance of each dimension.

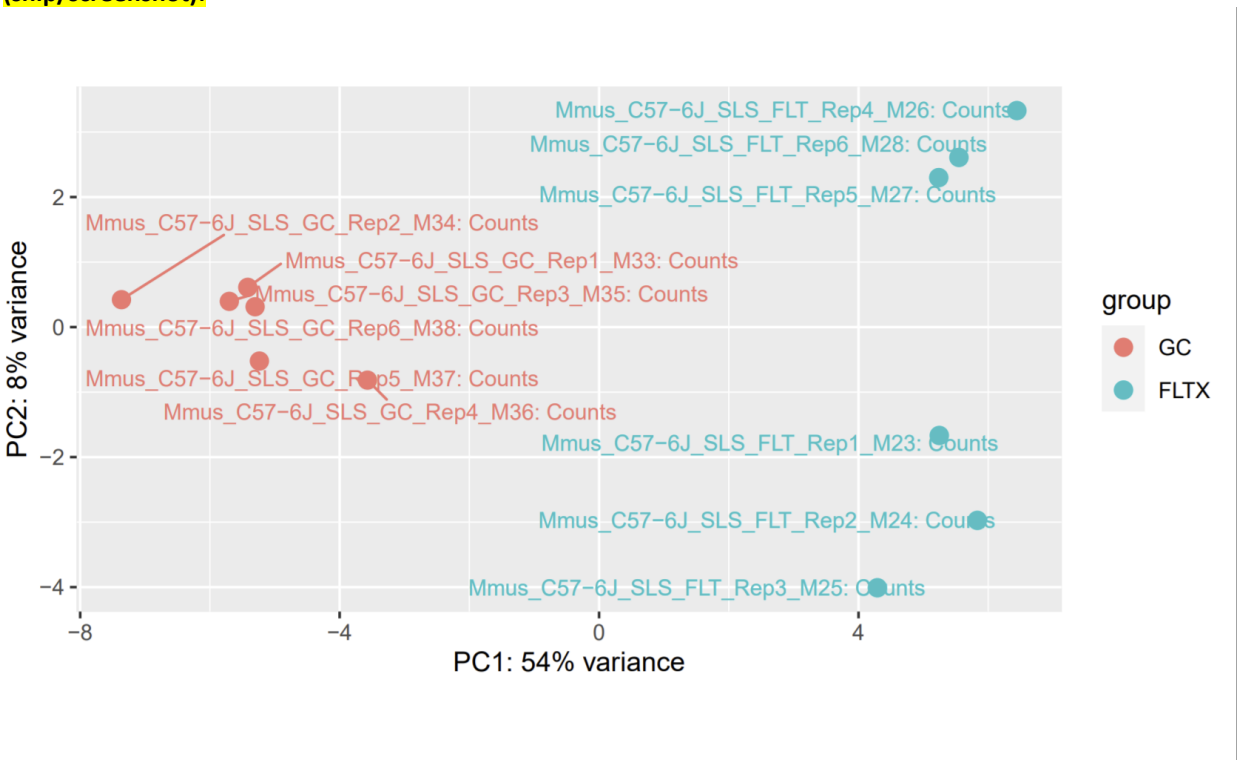
Part 5: Database Practice--Making the PCA Plot

Just for fun, let's practice briefly navigating the database to see what it looks like to take the final step towards a PCA Plot. We will check off the steps as we go and take a snip and insert a picture of the final product!

Students will check off the steps as they go! At the end I have included a picture of the PCA plot they should get, for your reference!

Completed? (Add checkmark)	
	Go to https://genelab.nasa.gov
	Click the Analyze Data button.
	Sign in with google into your account. If you do not have an account, use your school gmail to set one up.
	Navigate to Shared Data -> Histories.
	Select GLDS-104: DESeq2 DGE Files.
	Click on the + on the far-right corner to add these files to your history.
	In the pop-up, keep the name as is and click Import.
	Now the files, will appear in your current history.
	Select the file labeled 'DESeq2 plots on data…'. This file will enable us to visualize the results. Remember to select the eye icon to view the results.
	Let's look at the PCA plot first. Does it resemble the one below?

Your PCA plot (snip/screenshot):



Part 6: Learning Aims and Evaluation

Please rate where you personally are at, with regards to the learning aims, at the end of the lesson and why.

Rating Scale

1- I do not understand it at all yet.

2-I understand parts of it, but I need my teacher and/or classmates' support to answer questions.

3-I understand it and can complete an assignment by myself.

4-I understand it so well I can teach others and apply my knowledge to new situations.

Talk to students about the importance of metacognition ("thinking about your thinking") and reflecting on their learning and where they are at with what they are learning. It may feel strange to students, but it is very important for their growth and helps them see learning aims as not something to simply "glaze over" at the beginning of the assignment but tools to see what they have learned. Encourage them to be honest, letting them know that rating themselves below a 4 for any or all of these learning aims does not correlate with losing points for this section.

Learning Aim #1: Students will be able to define what a PCA plot is and why it is a useful tool for data analysis.

My Evaluation of Learning Aim #1 and Explanation:

Student answers will be personal.

Learning Aim #2: Students will be able to read a PCA plot and pull useful information from it.

My Evaluation of Learning Aim #2 and Explanation:

Student answers will be personal.

Learning Aim #3: Students will build confidence and experience in navigating the NASA GeneLab database.

My Evaluation of Learning Aim #3 and Explanation:

Student answers will be personal.

References

Ngo, L. (2018). How to read PCA biplots and scree plots. BioTuring.com. https://blog.bioturing.com/2018/06/18/how-to-read-pca-biplots-and-scree-plots/

Ngo, L. (2018). Principal component analysis explained simply. BioTuring.com. https://blog.bioturing.com/2018/06/14/principal-component-analysis-explained-simply/

GL4HS Manual: GeneLab for High School Bioinformatics Manual. Blaber, Elizabeth. 2021.

Stamer, J. (2017). StatQuest: PCA main ideas in only 5 minutes!!!!. YouTube.com. https://www.youtube.com/watch?v=HMOI_lkzW08

AUTHOR
Catherine Boileau, Bloomington High School South (Bloomington, Indiana)
Edited by GL4HS Staff