Script Name: PCA_indi_standalone.R

Description: Expected output is a PDF of scree plots, PCA plots, and boxplots loaded from .Rdata object generated from script GENESIS_setup_ANALYSIS_standalone.R

General Use Instructions

- for use in interactive Rstudio Session or after all variables have been manually set

Step 1:

- Run GENESIS_setup_ANALYSIS_standalone.R script

Step 2: Variable assignments

- Open script in Rstudio and fill in variable section at the top called: "VARIABLES THAT NEED TO BE UPDATED" Change the items in quotes for each variable. This is the only step where anything needs to be updated or changed within the script.

```
#source("https://bioconductor.org/biocLite.R")
#biocLite("GENESIS")

library("BiocGenerics", lib.loc = args[3])
library("Biobase", lib.loc = args[3])
library("GENESIS", lib.loc = args[3])
library("GWASTools", lib.loc = args[3])

#STEP 2: Variable Assignments
#----------------------<start> VARIABLES THAT NEED TO BE UPDATED <start>------------------#
load("full/path/to/my_GENESIS_PC_pheno_covar_data.Rdata")
cohort <- "myPopulationName" # only used for labeling plots and output files
outDir_path <- "/my/output/directory/"
#----------------------<end> VARIABLES THAT NEED TO BE UPDATED <end>----------------------#
```

Variable Name/Function	Type	Definition
load()	function	Full path and name of the .Rdata binary object generated from GENESIS_setup_ANALYSIS_standalone.R script
cohort	string	Name of population; only used to label graphs and embed final file names
outDir_path	string ending in "/" of an already existing directory on local computer	Full path to your output directory where files should be saved. Must already exist and end in "/"

Step 3: Extract PC data frame from loaded R data object and generate PDF

- No changes to the code need to be made here.

Explanation of code:

This part of the code will extract the PC data frame generated by pcair from the previous script and store it in the variable *get_dataframe*. A new empty PDF with specific figure dimensions is created in the output directory that was specified in step 2 with the following naming convention:

```
yourStringCohort_individual_PCA_plots.pdf
```

Step 4: Generate scree plots

- No changes to the code need to be made here.

Explanation of code:

Two scree plots are generated and printed to the newly created PDF. The first scree plot is of all 200 PCs (if available) from the pcair dataframe saved in the R data object. The proportion of variance is calculated by taking each PC and dividing is over the sum of the 200 PC eigenvalues.

The second scree plot is the same as above except it subsets it to the first 20 PCs instead of all 200. The sum of the eigenvalues is also subset to only the first 20 PCs, therefore, the proportion of variance is only across the first 20 PCs.

Step 5: Create variables that make plots look nice

- No changes to the code need to be made here.

Explanation of code:

This just changes colors of the points by population. In this script, it is only one population so only one color has been added.

Step 6: Generate scatter plots of pairwise PCs

- No changes to the code need to be made here.

Explanation of code:

This generates pairwise combinations of PCs across multiple scatter plots. Each horizontal and veritcal line (i.e. abline() function calls) indicates n number of standard deviations away from the mean for each PC being compared. These boxplots can easily

be condensed into one for loop, however, for troubleshooting purposes, we have written each chunk of code out separately.

The PC scatter plots that are generated here are the following:

- PC1 vs PC2
- PC1 vs PC3
- PC1 vs PC4
- PC1 vs PC5
- PC2 vs PC3
- PC2 vs PC4
- PC2 vs PC5
- PC3 vs PC4
- PC3 vs PC5

Step 7: Generate boxplots of PCs

- No changes to the code need to be made here.

Explanation of code:

This will generate a boxplot with a scatter plot overlay for each individual. A mean and standard deviation is calculated and horizontal lines of the mean and 6 standard deviations away from the mean are also plotted on the graph. The first 5 PC plots are generated here. Again, this can easily be condensed into one for loop, but easier to troubleshoot and modify if each boxplot is generated individually.

Step 8: Save and close PDF

- No changes to the code need to be made here.

Explanation of code:

This one-liner will just save and close the PDF and make sure all the figures in the buffer have been pushed and saved to the PDF.