Script Name: PC_covariates_standalone.R

Description: Expected output is a printed to standard out (or within the console of an Rstudio session, if running interactively) with a list of which PCs should be included as covariates in the association analysis mixed effects model due to statistical significance with your phenotype of interest.

General Use Instructions

- for use in interactive Rstudio Session or after all variables have been manually set in step 2, then can run script on command line
- Please do no change anything in the code except for the variable values in STEP 2, and do not change the variable names in STEP 2. Changing anything else in the code may cause errors or inaccuracies in the logic that was built

Step 1:

- Make sure you have already run the code **GENESIS_setup_ANALYSIS_standalone.R**
- PC_covariates_standalone.R is dependent on the generated data object with the following file suffix: *_GENESIS_PC_pheno_covar_data.Rdata*, as generated by the script GENESIS_setup_ANALYSIS_standalone.R
- Make sure the following R libraries are already installed:
    - BiocGenerics
    - Biobase
    - GENESIS
    - GWASTools

Step 2: Variable Assignments

- Open the script (PC_covariates_standalone.R) in Rstudio and fill in the variable section at the top called “VARIABLES THAT NEED TO BE UPDATED”. Change the values of the variable names within the double quotation marks. All the variable values should be strings, therefore, between quotation marks.

```
1  library("BiocGenerics")
2  library("Biobase")
3  library("GENESIS")
4  library("GWASTools")
5
6
7  #STEP 2: Variable Assignments
8  #---------------------<start> VARIABLES THAT NEED TO BE UPDATED <start>-----------------#
9  dataObj <- "/full/path/to/my/R/data/object/containing/scanAnnot/and/coVarMatrix/and/GRM"
10 phenotype <- "myPhenotype"
11 varType <- "logisitc or gaussian"
12 #---------------------<end> VARIABLES THAT NEED TO BE UPDATED <end>---------------------#
13
14
```

Variable Name	Type	Definition
dataObj	string	Full path to the saved R data object generated from the script GENESIS_setup_ANALYSIS_standalone.R This data object should have the suffix: *_GENESIS_PC_pheno_covar_data.Rdata* and contain the scanAnnot dataframe with your PCs, phenotypes, and covariates of interest; Additionally it should contain the following data objects: covMatList and mypcair
phenotype	string	A string the matches the phenotype of interest you would like to test for your downstream association analysis. **This string MUST match a header in the scanAnnot dataframe exactly!**
varType	string	Pick **one** of the following strings: logistic OR Gaussian. If the phenotype you are testing is binary, choose logistics. If the phenotype you are testing is a continuous variable choose Gaussian.

Step 3: Confirm Mixed Model with Random Effects Selection Type

- No changes need to be made to the code here

  Explanation of code:

  This code is literally just two print statements that print to your console or standard out if you are using the command line. Its purpose is to serve as an audit trail if you save the output to a log file to confirm your variable selections in STEP2.

Step 4: Check PC Significance

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

  Explanation of code:

  The R data object is loaded in this step, which will make the scanAnnot, covMatList, and mypcair data objects available in your environment.

  There is some basic logic built here that look at the varType variable set in STEP 2 and builds the appropriate null mixed model depending on whether the varType is set to “logistic” or “Gaussian”.

  Once the appropriate model is identified, it will loop through the each of the first 20 PCs and build an independent null model where the scanAnnot data object will be used to extract the eigenvalues for the PC of interest and set it as a covariate in the model. Additionally, the code will extract the phenotype for every individual and use it as a

predictor/outcome/dependent variable. As this point, the covMatList will also be used to accurately build the model and account for genetic relationships between individuals.

Once convergence of the null model is reached, the p-value of the model is extracted. If the p-value is < 0.05, the pc is stored in a list, which is called pc.list. Therefore pc.list will have all the pcs that are statistically significant with your phenotype of interest (p-value < 0.05) for the first 20 PCs.

Step 5: Report PC Significance

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

Explanation of code:

This will print which PCs (if any) need to be included as covariates to your mixed model for association analysis due to statistical significance with your phenotype of interest.