Script Name: GENESIS_Association_Analysis_standalone.R

Description: Expected output saved in a table with all SNPs on a chromosome (or chunk, depending on how you split your dosage mach files) of both imputed and typed tested for association with phenotype of interest using a mixed model with random effects with added covariates and genetic relationships. Output will be a file containing results of the association analysis following one of the following file naming conventions:

- chr<chr_num>_<phenotype>_logistic_association_results.txt
- chr<chr_num>_<phenotype>_Gaussian_association_results.txt

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

IMPORTANT!! Please go through the checklist and make sure all the following is completed before proceeding:

- ✔ Successfully run GENESIS_setup_ANALYSIS_standalone.R and have location of saved data object
- ✔ Successfully run PC_covariates_standalone.R and have a list of pcs that should be included as covariates in your association analysis model
- ✔ Imputed data on the same sample set used in the previous 2 scripts above
- ✔ Sample order of GENESIS scanAnnot is in the same order as the samples in imputed files **– order is critical! It does not use header or sample names to check this, completely positional!**
- ✔ Successfully run DosageConverter on the properly ordered imputed files and have generated MACH files **– the imputed file here must be in the same order as the GENESIS scanAnnot before running DosageConverter!**

Step 1:

- Make sure all the checks above have been successfully checked and completed.
- Make sure the following R libraries are already installed:

- gdsfmt
- SNPRelate
- GWASTools
- SeqVarTools
- GENESIS

Step 2: Variable assignments

- Open the script GENESIS_Association_Analysis_standalone.R
- Update all the variable values, however, do not change any of the variable names or you risk breaking the code.
- Variables `dataObj`, `phenotype`, and `varType` are actually the same variables from the script *PC_covariates_standalone.R*, therefore, you can just copy the values of those three into this script.
- The variables `dosefile`, `markfile`, and `posfile`, are the full path to files generated from the DosageConverter script. Please open the file: *Instructions_DosageConverter.pdf*. You will see on the front page these three variables are highlighted in yellow to show you what needs to be populated here.

```
library("gdsfmt")
library("SNPRelate")
library("GWASTools")
library("SeqVarTools")
library("GENESIS")


#STEP 2: Variable Assignments
#--------------------<start> VARIABLES THAT NEED TO BE UPDATED <start>-----------------#
dataObj <- "/full/path/to/my/R/data/object/containing/scanAnnot/and/coVarMatrix/and/GRM"
phenotype <- "myPhenotype"
varType <- "logisitc or gaussian"
chr_num <- "1" # options: string numbers 1-22,"X", "XY", "Y", "M", or "MT"
dosefile <- "/full/path/to/my/converted/chr/dose/vcf/file.txt"
markfile <- "/full/path/to/my/converted/chr/info/file.txt"
posfile <- "/full/path/to/my/positionFile.txt"
covariates <- c("myCov1", "myCov2", "...") # Note these must be available in scanAnnot with exact same header names, comma-separated
#--------------------<end> VARIABLES THAT NEED TO BE UPDATED <end>---------------------#

```

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.
chr_num	string	A string specifying chromsome. **Pick one of the following options as a string, not an integer:** 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, X, XY, Y, M, MT
dosefile	string	Full path to MACH converted dose file. File name usually looks similar to: prefex.mach.dose; please see Instructions_DosageConverter.pdf for more info
markfile	string	Full path to MACH converted info file. File name usually looks similar to: prefex.mach.info; please see Instructions_DosageConverter.pdf for more info
posfile	string	Full path to MACH SNP position file. File name usually looks similar to prefix.posFile.txt; please see Instructions_DosageConverter.pdf for more info. Essentially a tab-delimited file with column headers SNP and position of all the SNPs and their positions.
covariates	list/vector	A list or vector of strings of covariates to include in the mixed model for association analysis. All covariates must be in the scanAnnot data object and the list/vector **must be a header(s) in the scanAnnot data object**.

Here is an example to illustrate how to populate some of these variables:

For example if the following image below is my scanAnnot data object from the script: *GENESIS_setup_ANALYSIS_standalone.R*, you can see all the available headers for my dataset:

Example headers in my scanAnnot data object that can be used in the the covariates variable in step 2 or in the phenotype variable in step 2

scanID	pc1	pc2	pc3	pc4	pc5	pc6	...	pc46	pc47	pc48	pc49	pc50	pheno	testGaussian
1	-0.0415033881	-4.395812e-03	-4.068281e-03	1.857896e-03	-2.105147e-03	2.923564e-03		2.388767e-03	0.0130364040	0.0121836312	8.311109e-03	1.064784e-02	NA	0.980832307
2	-0.0128411317	2.403853e-02	-2.804255e-02	4.480715e-03	-9.961323e-03	3.173191e-02		2.859700e-02	-0.0013945356	-0.0140777233	-3.835080e-02	-1.171777e-02	0	0.540301633
3	-0.0430089041	-2.283324e-03	-2.928220e-02	1.163721e-02	-8.469856e-03	7.458786e-03		-1.066196e-02	-0.0044334429	-0.0079655944	2.445023e-03	-2.253554e-03	0	1.888623004
4	0.0325683664	7.244507e-02	-4.320467e-02	-3.002501e-02	5.402558e-02	-9.316456e-03		7.765757e-02	0.0419505752	0.0246058429	-5.321560e-02	6.442680e-02	1	1.014650527
5	-0.0347231762	-2.207474e-03	-2.845047e-04	3.883748e-04	1.289349e-02	1.368586e-02		-1.240800e-02	-0.0177225359	-0.0037370670	-7.739026e-03	-1.313527e-02	1	0.999108545
6	-0.0154609861	2.396421e-03	9.296539e-03	2.567734e-03	-4.593430e-03	3.764475e-03		3.714938e-03	-0.0076803688	-0.0081845147	-4.696618e-04	-5.440926e-03	1	0.532955746
7	-0.0495233104	-2.167840e-02	2.750656e-02	3.548701e-03	-3.428966e-04	-6.092982e-03		1.044132e-02	0.0118394113	0.0079029514	-8.778409e-04	8.828677e-03	1	0.009528681
8	0.0336110508	-8.353631e-02	1.668054e-02	-9.722934e-03	-7.725907e-03	2.443148e-02		-6.193555e-02	-0.0087694474	-0.0189794616	-4.017238e-02	1.279125e-02	0	1.365605458
9	-0.0363351328	-7.207096e-03	6.157733e-03	8.863733e-03	-7.889091e-03	5.778005e-03		-1.347135e-02	0.0009493933	0.0102418333	-2.906483e-03	1.188159e-02	0	1.176348811
10	-0.0260817331	-9.158682e-03	3.240573e-02	-1.842537e-02	-1.138193e-02	1.984601e-02		3.937523e-03	0.0115597694	-0.0036661265	1.500476e-02	1.787126e-02	0	1.475881217
11	-0.0466545301	-4.432917e-03	-1.602434e-02	-1.070095e-04	4.796071e-04	-5.105556e-03		2.282797e-03	-0.0017273671	0.0012249618	2.393627e-03	1.152666e-04	0	1.674142556
12	-0.0230859171	5.756189e-03	-3.045375e-02	1.739805e-02	-1.937121e-02	-6.286646e-03		2.392450e-02	0.0200439653	-0.0104385052	-2.517072e-02	-1.100199e-02	1	1.314864267
13	-0.0072667849	1.293976e-02	-2.317069e-03	5.169499e-03	-4.855044e-03	3.314499e-03		-2.469241e-03	0.0004842473	0.0035509422	1.340613e-02	-5.097108e-03	1	0.655132415
14	-0.0141898336	8.695803e-03	-6.717616e-03	1.879957e-02	-2.602353e-03	-2.261351e-03		-3.388077e-02	-0.0111022192	0.0003890598	6.283776e-03	-1.227074e-02	1	0.889105295
15	-0.0277832396	1.142366e-03	-1.936378e-02	2.395099e-02	-1.581723e-02	2.006447e-02		5.777856e-03	-0.0174015824	0.0086127347	-1.452201e-02	4.262558e-03	1	1.551101319
16	-0.0296955350	-5.305111e-03	3.990618e-03	3.011553e-03	6.844902e-03	1.789383e-02		1.104718e-03	0.0026626714	-0.0067646856	-2.898426e-04	-2.833395e-03	1	0.597304942
17	-0.0374578031	-8.182722e-03	-1.103914e-02	7.515494e-04	-1.122761e-02	1.059161e-02		-1.033877e-05	0.0027056537	-0.0016942066	-1.167353e-02	-1.470326e-02	1	0.409351062
18	-0.0303413556	-2.715526e-03	-3.280925e-02	-1.158681e-03	2.078840e-02	-1.032930e-02		-1.278511e-02	-0.0058026624	-0.0130676109	-7.394270e-03	4.496717e-03	0	1.855716458

Given the available headers in my scanAnnot data object above, let's say I have chromosome 22 mach files available and I want to run a logistic mixed effects model for association analysis on my

phenotype of interest called “pheno” and add pc1, pc2, pc3, and pc7 as covariates to my model. I can update the following variables in step 2 as follows:

```
1  library("gdsfmt")
2  library("SNPRelate")
3  library("GWASTools")
4  library("SeqVarTools")
5  library("GENESIS")
6
7
8  #STEP 2: Variable Assignments
9  #---------------------<start> VARIABLES THAT NEED TO BE UPDATED <start>------------------#
10 dataObj <- "/full/path/to/my/R/data/object/containing/scanAnnot/and/coVarMatrix/and/GRM"
11 phenotype <- "pheno"
12 varType <- "logistic"
13 chr_num <- "22" # options: string numbers 1-22,"X", "XY", "Y", "M", or "MT"
14 dosefile <- "/full/path/to/my/converted/chr/dose/vcf/file.txt"
15 markfile <- "/full/path/to/my/converted/chr/info/file.txt"
16 posfile <- "/full/path/to/my/positionFile.txt"
17 covariates <- c("pc1", "pc2", "pc3", "pc7") # Note these must be available in scanAnnot with e
18 #---------------------<end> VARIABLES THAT NEED TO BE UPDATED <end>---------------------#
19
```

After updating all the variables in step 2, you can go ahead and run the script, either interactively in Rstudio or on the command line for each chromosome. All other steps outlined below are just code explanations, but please do not change anything at step 3 and 4.

Step 3: File conversions and extractions

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

  Explanation of code:

  At this step the binary R data object generated from *GENESIS_setup_ANALYSIS_standalone.R* is loaded into the environment.

  The converted MACH files are used here along with the SNP position marker file all generated from dosage_converter_script.sh are read and stored in a temporary gds file. The genotypes are then extracted and stored in the variable `genoData`.

Step 4: Association analysis

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

  Explanation of code:

There is some very basic logic built here that determines if the value set in step 2 for the variable `varType` is a logistic model or a Gaussian model. The code will then choose the appropriate model with the proper configurations. It will use the covariates listed in step 2 to the model in addition to the phenotype you specified in step 2.

After the model is built and finished running, all results will be written to a file that is tab-delimited. The file output name will follow one of the two file naming conventions depending on which model was run:

- chr<chr_num>_<phenotype>_logistic_association_results.txt
- chr<chr_num>_<phenotype>_Gaussian_association_results.txt