CCPM Biobank GWAS Pipeline

Release 0.1.0

Tonya Brunetti

Jan 05, 2021

Contents

CHAPTER 1

Installation

Good news! There is no installation required to use this pipeline and it should be OS agnostic. There are only 2 true system dependencies and most HPCs and shared resources already have these dependencies installed:

- Singularity version >= 3.0
- Golang version >= go1.13.5 (has been tested on versions go1.13.5 and go1.15.2 on a Linux OS).

If not, you can install Singularity and install Golang using the links on your local computer. Both should be available across OS platforms for Windows, MacOS, and Linux.

CHAPTER 2

Getting Started

2.1 Deciphering the config file

The config file is a text file that contains key-value mappings to all input parameters and pipeline logic. The file can be broken down into three main parts: the pipeline logic, environment setup, and the user data input options.

2.1.1 Pipeline Logic

There are five keywords that control the logic of the pipeline and all five accept only boolean arguments (true/false):

```
GenerateGRM:true
GenerateNull:true
GenerateAssociations:true
GenerateResults:true
SkipChunking:false
```

2.1.2 Environment Setup

There are three keywords that control the environment setup of the pipeline. It is a requirement to use the container provided as an LSF object on github. This ensures all software is properly versioned and it reduces the complexity of the pipeline by using predefining software paths as well as reducing installation issues that arise with many softawre dependencies.

```
BindPoint:/path/to/bind/container
BindPointTemp:/path/to/tmp/
Container:/path/to/SAIGE_v0.39_CCPM_biobank_singularity_recipe_file_11162020.simg
```

`BindPoint` and `BindPointTemp` are directories to where you want the container to be mounted. `BindPointTemp` allows you to give the container a secondary binding point where all temp files and calculations will be performed, and once finished will move the final files to the `BindPoint`. The files generated in `BindPointTemp` will be deleted after the run completes.

Note: If you do not want to use or do not have a temp directory, please set this parameter to be the same as `BindPoint`.

Warning: **IMPORTANT PLEASE READ!** The container follows the same rules of inheritance as Singularity specifies. This means the `BindPoint` and `BindPointTemp` become the highest point in your directory tree. Thus, you can only access paths, directories, and files if you are able to traverse below the tree from these starting points but not above these entry points from your host system. Therefore, be sure all the paths, directories, and files specified in the config file are contained within the scope of at least one of these two entry points.

2.1.3 User Data Input

The remainder of the keywords are parameters offered to the user and the user can specify paths and options for all or some of these keywords:

```
ChromosomeLengthFile:/path/to/chromosomeLengths.txt
Build:hg38
Chromosomes:1-22
ImputeSuffix:_rsq70_merged_renamed.vcf.gz
ImputeDir:/path/to/imputed/data/directory/
OutDir:/path/to/output/final/results
OutPrefix:myGWAS
PhenoFile:/path/to/phenotype/covariate/file
Plink:/path/to/LDpruned/plink/file/prefix
Trait:binary
Pheno:myPhenotype
InvNorm:FALSE
Covars:PC1,PC2,PC3,PC4,PC5,age,sex
SampleID:SampleIDcol
NThreads:
SparseKin:True
Markers:30
Rel:0.0625
Loco:TRUE
CovTransform:True
VcfField:DS
MAF:0.05
MAC:10
IsDropMissingDosages:FALSE
InfoFile:/path/to/info/file
SaveChunks:false
GrmMAF:0.01
ChunkVariants:1000000
SaveAsTar:false
ImputationFileList:/path/to/list/of/chunked/chromosomes/file.txt
SparseGRM:/path/to/grm/file.mtx
SampleIDFile:/path/to/grm/sample/id/file.mtx.SampleID.txt
NullModelFile:/path/to/null/model/file.rda
VarianceRatioFile:/path/to/variance/ratio/file.txt
AssociationFile:/path/to/concatenated/association/results/file.txt
```

2.2 Parameters

This explains in more detail how and when to use each keyword in the config file. For readability, the minimum required parameters are listed for each step. If more than one step will be run (>1 out of 5 possible pipeline logic keywords is set to true), the minimum parameters required is the sum of the minimum required parameters for all true keywords.

2.2.1 Minimum Required Parameters for Each Step

Minimum Requirements to Run Genetic Relatedness Matrix

To calculate the GRM, `GenerateGRM:true` needs to be specified in the config file. If it is set to `false`, the pipeline assumes one of two things:

1. *GRM is not needed because one is provided in the config file from from a previous calculation*
2. *GRM is not needed at all for the scope of the pipeline logic*

Minimum Requirements to Run Null Model

To calculate the null model, `GenerateNull:true` needs to be specified in the config file. If it is set to `false`, the pipeline assumes one of two things:

1. *The null model files are not needed because one is provided in the config file from from a previous calculation*
2. *The null model files are not needed at all for the scope of the pipeline logic*

Minimum Requirements to Run Association Analysis

To calculate the association analyis, `GenerateAssociations:true` needs to be specified in the config file. If it is set to `false`, the pipeline assumes one of two things:

1. *The association file results are not needed because one is provided in the config file from from a previous calculation*
2. *The association analysis files are not needed at all for the pipeline*

Minimum Requirements to Run Results

To calculate the association analyis, `GenerateResults:true` needs to be specified in the config file. If it is set to `false`, the pipeline assumes that no data clean up, data merges, and graphical summaries are needed.

2.2.2 Full List of User Input Data Parameters

Table 1: Definitions and Use

Parameter	Type	Default	Description	Relevance
ChromosomeLengthFile	string		This is a tab-delimited text file that contains the same names as the chromosomes in your dataset followed by the chromosome length. This file can generally be pulled down from the NCBI website under the build you are using (hg19 or hg38). This is used for chunking the imputed data.	if `skipChunking:false`
Build	option: hg19 or hg38	hg38	This determines whether the software needs to be searching for a "chr" before the chomosome name or not	
Chromosomes	string	1-22	This lets the software know which chromosomes you want to use for association analysis. It must be a range. If you want to only run analysis on a single chromsome, the start and end will be the same value. For example: running chromome 2 only will look like 2-2.	if `GenerateAssociation`
ImputeSuffix	string		The full suffix for the software to determine which files in the directory are impuation files. Important, it is assuming the prefix is either a chromosome number (hg19) or the string chr followed by the chromosome number (hg38)	if `GenerateAssociation` and `skipChunking:false`
ImputeDir	string		Full path to directory where imputed results are located	if `GenerateAssociation`
OutDir	string		Full path to directory where final results should be transferred	all
OutPrefix	string		string (no whitespace or special characters) to prefix to the output files generated	all
PhenoFile	string		Full path to tab-delimted phenotype file containing sample IDs, phenotypes, and covariates, with whatever string of headers you choose. NO WHITESPACES in header names.	If `:GenerateNull:true`
Plink	string		Full path to the directory and plink file prefix (dropping the suffix .bed,.bim,.fam) to an LD-pruned set of data to be used to generate GRM relatedness and to select random markers from for the variance ratio value	If `GenerateGRM:true` or `GenerateNull:true`
Trait	option: binary or quantitative		Based upon your association phenotype. If binary, all values will be 0/1/NA, if quantitivate all phenotype traits to be analyzed will be continuous or have numeric quantitative meaning	if `GenerateNull:true` or `GenerateResults:tru`
Pheno	string		The exact name (case-sensitive) of the phenotype to be analyzed in your PhenoFile. Must be present in PhenoFile.	if `GenerateNull:true` or `GenerateResults:tru`

continues on next page

Table 1 – continued from previous page

Parameter	Type	Default	Description	Relevance
InvNorm	boolean	FALSE	This applies to the phenotype of interest to be analzyed and whether to perform an inverse normalization. For binary traits, this should be set to FALSE and for quantitative traits, set this to TRUE.	if `GenerateNull:true`
Covars	comma-separated list		A comma-separated list (no whitespaces) of all the covariate names to regress out in the model. These variables need to be in your PhenoFile.	if `GenerateNull:true` or `GenerateResults:tru`
SampleID	string		A string (no whitespaces) that is contained in the header of your PhenoFile. This is the sampleID names and they must be the same names as listed in the PhenoFile, Imputation Files, and Plink Files.	if `GenerateNull:true`
Nthreads	int		Strongly Recommned to leave this blank! If left blank, it will auto-decect available resources and scale steps automatically on the back-end. By specifiying the threads it tells the program to use max Nthreads for parallelization and concurrency.	all
SparseKin	boolean	TRUE	If set to true, takes advantage of the sparsity of the GRM, otherwise will not use the sparsity to make assessments	if `GenerateNull:true`
Markers	int	30	The number of random markers selected from the LD-pruned plink file to estimate the variance ratio component in the null model. Warning! This number increases time linearly	if `GenerateNull:true`
Rel	float	0.0625	A float between 0.0-1.0. This is the threshold in kinship estimate to consider someone related. Anything below this value will be treated as an unrelated individual in the pairwise comparison and calculation for the sparse GRM.	if `GenerateGRM:true`
Loco	boolean	TRUE	Leave-One-Chromosome-Out method. Warning – Setting this to true, increases the time complexity of the algorithm.	if `GenerateNull:true`a :code:`GenerateAsso
CovTransform	boolean	TRUE	Recommended to set to true. It is a QR decomposition that aids in the covergence of the null model.	if `GenerateNull:true`
VcfField	option: DS or GT	DS	This determines what metric to base association upon. DS = dosages and GT = genotypes. If you have genotypes only, i.e. chip data withouth dosage calculations, DS cannot be used!	if `GenerateAssociation`

continues on next page

Table 1 – continued from previous page

Parameter	Type	Default	Description	Relevance
MAF	float	0.05	Float between 0.0-0.50 that specifies the cutoff to be considered a common snp or a rare snp. For example, keeping this to the default of 0.05 will assume common snps are defined as those with a minor allele frequency >5% and that rare snps are defined as those with a minor allele frequency 5%. THIS IS NOT A FILTER!	if `GenerateResults:tru`
MAC	int	10	A filter applied to the cleaned association results to remove snps that have low minor allele counts. Default recommendation is to set this to 10.	if `GenerateResults:tru`
IsDropMissingDosages	boolean	FALSE		if `GenerateAssociation`
InfoFile	string		Path to the info file. This file contains snps information pertaining to chromosome, positions, genotype/imputation status, R2, ER2 values. For formatting of this file please refer to <----------→	if `GenerateResults:tru`
SaveChunks	boolean	TRUE	Specifies whether to save the chunked files and the queue list for future use.	if `GenerateAssociation` and `skipChunking:false`
GrmMAF	float	0.01	The minor allele frequency threshold for a snp to be included in the GRM calculation based on the LD-pruned plink file. For example, if set to 0.01 this means any snp with a MAF > 0.01 wil be used to calculate relatedness in the GRM.	if `GenerateGRM:true`
ChunkVariants	int	1000000		if `GenerateNull:true` and `SkipChunking:false`
SaveAsTar	boolean	FALSE		all
ImputationFileList	string		Ends in _chunkedImputationQueue.txt	if `GenerateAssociation` and `skipChunking:true`
SparseGRM	string		Ends in .sparseGRM.mtx	if `GenerateGRM:false` and `GenerateNull:true`
SampleIDFile	string		Ends in sparseGRM.mtx.sampleIDs.txt	if `GenerateGRM:false` and `GenerateNull:true`
NullModelFile	string		Ends in .rda	if `GenerateNull:false` and `GenerateAssociation`

continues on next page

Table 1 – continued from previous page

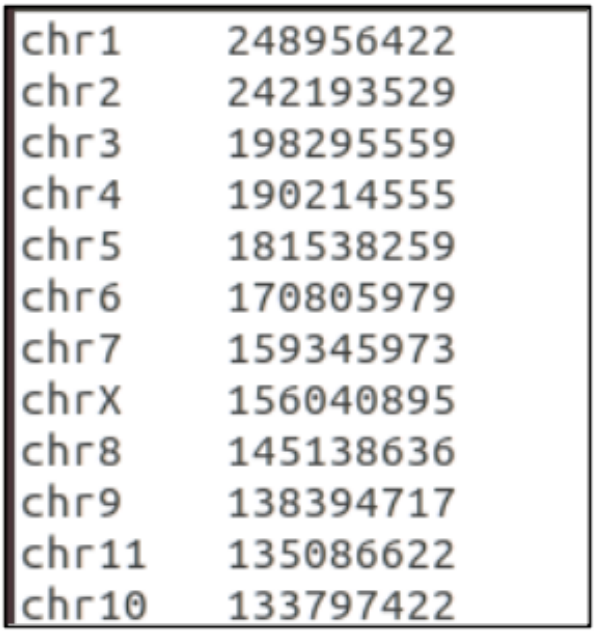
Parameter	Type	Default	Description	Relevance
VarianceRatioFile	string		Ends in .varianceRatio.txt	if `GenerateNull:false` and `GenerateAssociation`
AssociationFile	string		Ends in _SNPassociationAnalysis.txt	If `GenerateAssociation` and `GenerateResults:tru`

2.3 Formatting the Required Files

The pipeline does require some formatting for file names and contents within files to be present. This section explains the file format expectations, as well as file name expectations.

2.3.1 Parameter: ChromosomeLengthFile

The `ChromsomeLengthFile` parameter is a file that can be downloaded and modified from sources such as NCBI and UCSC. It should be a tab-delimted file with chromosomes 1-22 listed in the first column and the length of the chromosome in the second column. This file contains **no header**. The build you select should be based upon the build used for your imputation file. If only using genotyping file, base this file on the genotype build.



Example snippet for hg38

- 1st column is chromosome
- 2nd column is length

Warning: hg19 vs hg38 formatting Depending on the build you choose you must format your file accordingly.

- *hg19:* omit "chr" and should strictly be integer values between 1-22 for the first column.
- *hg38:* include the preceding "chr" string (no spaces) for integer values between 1-22 for the first column.

2.3.2 Parameter: ImputeSuffix

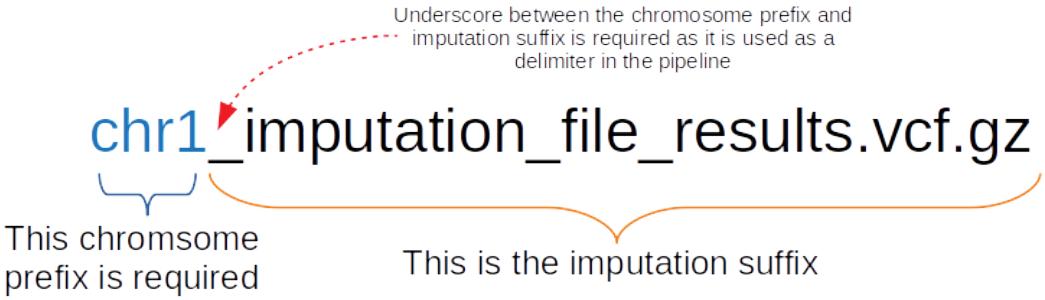
This string is the suffix of your imputation file. The imputation files require to be named and split in a particular way.

First, all imputation files follow these rules:

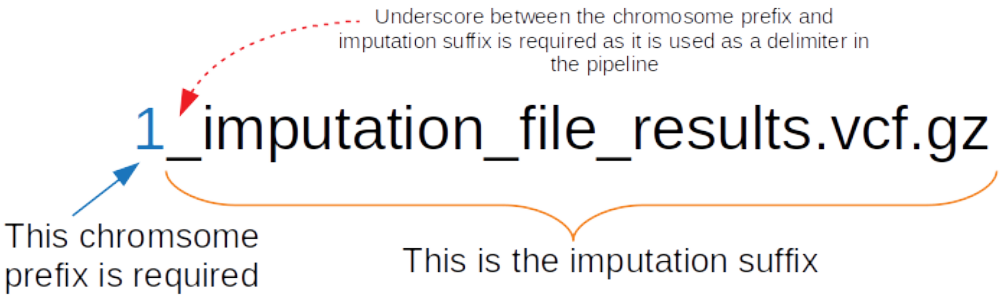
- split by chromosome
- sample names are the samples names used in `GenerateGRM` and `GenerateNull` (order doesn't need to be maintained)
- gzipped vcf files, and
- have the matching tabix index ending in the imputation file name followed by the .tbi string

This is how each imputation file needs to be named:

for hg38



for hg19



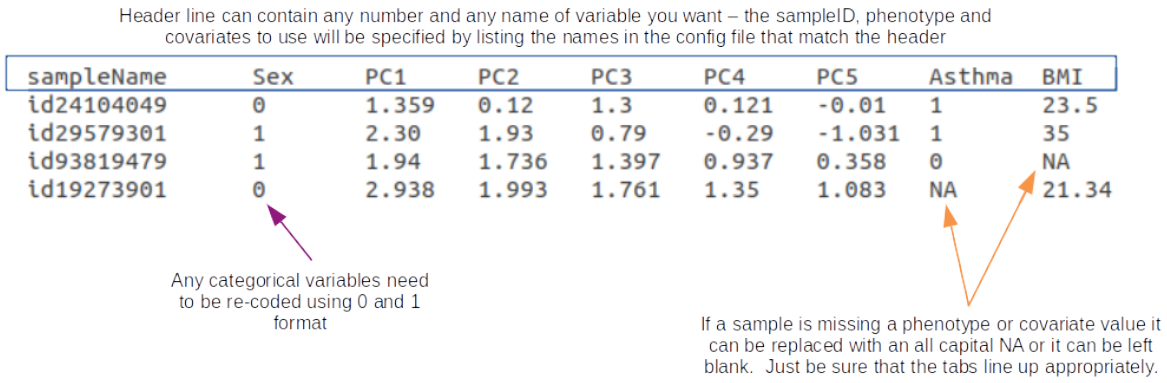
Warning: hg19 vs hg38 formatting Depending on the build you choose you must format your file name accordingly.

- *hg19:* omit "chr" and strictly be integer values between 1-22 as the prefix, followed by a required underscore, following by remaining string.
- *hg38:* include the preceding "chr" string (no spaces) for integer values between 1-22 as the prefix, followed by a required underscore, following by remaining string.

2.3.3 Parameter: PhenoFile

This is tab-delimited txt file that contains all the sample IDs (must be the same IDs used in the plink file, imputation file, GRM, and null model -- order agnostic). In addition to the sample IDs, it also contains any phenotype(s) you may want to run and any covariates you may want to use, although the user is not required to use everything listed in the header/file.

Below is an example of a tab-delimited `PhenoFile`. Again, none of these header names are required, however, there needs to be a header variable at minimum that denotes the sample ID and a phenotype to analyze.



Note: It is worthwhile to generate a single PhenoFile that contains many phenotypes and covariates you may want to analyze for the sample set. Within the config file, the user can specify which single phenotype to run and which covariates to run. That way, the user can run several jobs in parallel using the same PhenoFile but just changing the trait, phenotype, covariates, and invNorm parameters without having to change anything else.

2.3.4 Parameter: Plink

2.3.5 Parameter: Pheno

2.3.6 Parameter: Covars

2.3.7 Parameter: SampleID

2.3.8 Parameter: InfoFile

2.3.9 Parameter: ImputationFileList

2.3.10 Parameter: SparseGRM

2.3.11 Parameter: SampleIDFile

2.3.12 Parameter: NullModelFile

2.3.13 Parameter: VarianceRatioFile

2.3.14 Parameter: AssociationFile

2.4 Output Generated

2.5 Parsing Through StdErr and StdOut

When using the pipeline with the container and go binary file, all standard out messages and errors are reported in the generated log file. It is important to note, that since the pipeline automatically runs steps concurrently when possible, the order of the log files may show some steps running as dependencies become available and based on the threads available. Also, this standard out is used by SAIGE as well, so anything that does not follow the format below is a results of the software calculations generated by SAIGE, and not the pipeline.

The following shows how messages are formatted in the log output:

2.5.1 Logging Levels

There are 4 different levels or types of debugging and error handling. Some are meant to inform, while other are meant to stop the pipeline if an error occurs that cannot be resolved.

Level	Interpretation
UPDATE!	To inform the user of the status of the pipeline.
CONFIRMED!	To let user know a check or validation has been passed.
WARNING!	There may be a problem but it continues through the pipeline and skips the issue.
ERROR!	There is definitely a problem and forces the pipeline to prematurely exit.

By using the levels and function names, one can `grep` through the log file to find specific messages related to functions or error types.

2.5.2 Exit Status Codes

Although I am not a huge fan of relying on exit status codes, for those that use Gnu Parallel or other software that records exit status to determine if a job was successful or not, I have coded up a few exit codes so that if an error does occur the number can tell the user which function is responsible for the error. This make troubleshooting and errors easier to identify.

functions	Exit Status
func(main)	return 42
func(chunk)	return 10
func(smallerChunk)	None
func(processing)	None
func(nullModel)	None
func(associationAnalysis)	None
func(checkInput)	return 5
func(saveResults)	return 99
func(saveQueue)	None
func(usePrevChunks)	return 17
func(findElement)	return 20
func(parser)	return 3

CHAPTER 3

Quick Start and Examples

3.1 Example Work Flows

3.1.1 Quick Start Command

In order to run the pipeline, open a shell or bash prompt (or batch script for a job-scheduler) and type:

```
$ ./CCPM_GWAS_pipeline myConfigFile.txt
```

3.1.2 Full Pipeline

Full pipline means you want to run every component of the pipeline from beginning to end in one go, without re-using any previously calculated data from the pipeline. This is analagous to setting the pipeline logic kewords to the following:

```
GenerateGRM:true
GenerateNull:true
GenerateAssociations:true
GenerateResults:true
SkipChunking:false
```

3.1.3 GRM only

GRM only means you want to run the GRM step. This is analagous to setting the pipeline logic kewords to the following:

Note, `SkipChunking` *can be set to either* `true` *or* `false` *because it is only used if* `GenerateAssociation` *is set to* `true`.

```
GenerateGRM:true
GenerateNull:false
GenerateAssociations:false
GenerateResults:false
SkipChunking:false
```

3.1.4 Null Model only

Null Model only means you want to run the Null Model only. It makes an assumption that you already have the GRM pre-calculated and want to re-use it in this step by setting the keywords `SparseGRM` and `SampleIDFile` located in the config file. These two files are the result of running `GenerateGRM:true`.

Choosing to run just the null model generation step is analagous to setting the pipeline logic kewords to the following:

```
GenerateGRM:false
GenerateNull:true
GenerateAssociations:false
GenerateResults:false
SkipChunking:false
```

`SkipChunking` *can be set to either* `true` *or* `false` *because it is only used if* `GenerateAssociation` *is set to* `true`.

3.1.5 Association Analyses Only

Association Analyses only means you only want to run the association anlysis. It makes an assumption that you already have the null model file (.rda) pre-calculated and have a pre-calculate variance ratio file (.varianceRatio.txt) and want to re-use/use it in this step by setting the keywords `NullModelFile` and `VarianceRatioFile` located in the config file. These two files are the result of running `GenerateNull:true`.

Choosing to run just the association analysis step is analagous to setting the pipeline logic kewords to the following:

```
GenerateGRM:false
GenerateNull:false
GenerateAssociations:true
GenerateResults:false
```

When `GenerateAssociations:true`, the `SkipChunking` logic comes into play.

Note: This step produces the raw associaions results concatenated into a file. It does not clean up the data, perform the proper flips, or generate graphs/figures. If you want the raw data in addition to the previously mentioned actions, be sure to also set `GenerateResults:true`.

3.1.6 Results and Graphs Only

Results and Graphs Only cleans up raw data that was previously generated from an association analysis and generates cleaned data in addition to some figures/graphs. You can use any association analysis here as long as it meets the file formatting specifications for supplying results in `AssociationFile`.

3.1.7 Reuse Previously Indexed and Chunked Files

3.1.8 Combinations of Logic

CHAPTER 4

FAQs

CHAPTER 5

Acknowledgements

CHAPTER 6

Indices and tables

- genindex
- modindex
- search