Genetic Association Testing using GENESIS Workflows on *NHLBI BioData Catalyst powered by Seven Bridges*

Table of contents

Objectives

Hello and welcome to the ***NHLBI BioData Catalyst powered by Seven Bridges*** analysis platform. The purpose of this tutorial is to guide users through the steps of running a single variant or multi-variant association test using the GENESIS workflows. After reading the manual users will be able to:

- Create and manage projects for genome-wide association studies
- Find VCF files using the Data Browser
- Use the TOPMed Annotation Explorer to aggregate and filter variants from TOPMed studies for use in multi-variant association tests
- Find appropriate GENESIS workflows in the Public Gallery and setup and run these workflows as Tasks.
- QC and interpret results
- Find support from Seven Bridges if needed

Prerequisites

- Please begin with the **Onboarding Tutorial** which provides information and a step-by-step guide on:
    - Accessing hosted TOPMed datasets on BioData Catalyst
    - Creating an account with eRA Commons credentials
    - Creating projects
    - Running a practice analysis
- A general understanding of running a genome-wide association study
    - Previous use of R or the GENESIS package is not required

Overview of the GENESIS Pipeline

The GENetic EStimation and Inference in Structured samples (GENESIS) R package provides efficient methods of working with genotypes measured in sequencing and microarrays. These tools were developed by the TOPMed Data Coordinating Center (DCC) at the University of Washington, and the Seven Bridges team worked with the TOPMed DCC to create Common Workflow Language (CWL) tools for the GENESIS R functions and arranged these tools into computationally efficient workflows.

The goal is to allow a user with the appropriate dbGaP credentials to easily and reproducibly execute these workflows on TOPMed study data.

The GENESIS workflows are a good fit for performing association studies on TOPMed data because of their robust ability to estimate and account for population and pedigree structure. GENESIS implements linear mixed models for association testing of quantitative phenotypes and logistic mixed models via the penalized quasi-likelihood approach of GMMAT[1] for association testing of binary (e.g. case/control) phenotypes. The mixed models utilize the PC-AiR[2] PCs and a relatedness matrix of PC-Relate[3] kinship coefficient estimates to accurately and efficiently adjust for ancestry and relatedness in the sample. When no relatedness matrix is provided, simple linear or logistic regression models are used for quantitative or binary traits, respectively. Heterogeneous residual variances can be used to account for differences in quantitative phenotype variability by user specified group. For computational efficiency, a “null model” is fit once under the null hypothesis of no genotype effect, and variant association is subsequently tested genome-wide. Available association tests include single variant tests as well as multi-variant aggregate tests. Single variant tests are performed with score tests, and approximations of effect sizes are provided. The saddlepoint approximation (SPA) of p-values is available when testing binary phenotypes[4]. Multi-variant tests can be performed using the burden, SKAT[5], SKAT-O[6], fastSKAT[7], and SMMAT[8] methods. Each of these methods can be run with user-defined aggregation units (e.g. by gene) or with a sliding window approach, and they can incorporate either allele frequency-based or user-defined (e.g. utilizing variant annotation) variant weighting. GENESIS can utilize sparse matrix representation of relatedness and genotype matrices for a substantial reduction in computational demand in large samples such as TOPMed.

Running association tests on BioData Catalyst requires three things: **a collection of bioinformatic workflows, data, and a project to keep everything organized**. Project creation was covered in the ***Onboarding Tutorial*** so please reference that if needed. The section “**Organizing a Workspace and Preparing Data”** will walk you through how to identify a cohort you have permission to analyze and convert the VCF data to the GDS file format. The next section, **“Running Single Variant Association Tests”** will explain how to set up the single variant association workflow using the input files created in the “Preparing Data” section. After running the single variant association test, you can move onto the final section on **“Running Aggregate Associations.”**

For more information on association testing with the GENESIS R package please see:

Gogarten SM, Sofer T, Chen H, Yu C, Brody JA, Thornton TA, Rice KM, Conomos MP. 2019. Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics*. Btz567. https://doi.org/10.1093/bioinformatics/btz567

Course materials for “Computational Pipeline for WGS Data” at the Summer Institute for Statistical Genetics: https://uw-gac.github.io/SISG_2019/

GENESIS software can be found on Bioconductor and Github:
http://bioconductor.org/packages/release/bioc/html/GENESIS.html
https://github.com/UW-GAC/GENESIS

Organizing a Workspace and Preparing Data

Prior to running the association test, we need to do some preparation work. The following steps take you through how to:

- Add variant data to a secure project on the platform
- Find the GENESIS workflows on the platform and add them to a project
- Run the VCF to GDS tool to get the variant data in the appropriate format for the GENESIS workflows
- Access RStudio for use in phenotype harmonization

Project Creation

The first step in preparing for the association analysis is to create a secure project for working with the WGS data. Projects are workspaces that serve as the core building blocks of *BioData Catalyst powered by Seven Bridges*. Each project corresponds to a distinct scientific investigation and serves as a container for data, analysis workflows, and results. Multiple analyses can be carried out within a project. Projects are an important way to keep your analysis organized.

Please refer to the **Onboarding Tutorial** for more information about project creation. The terms below will be used throughout the tutorial:

Project: Workspace for organizing files and analyses.
App: Bioinformatics tools and workflows. There are hundreds of Apps hosted on the platform and users can also bring their own.
CWL: Common Workflow Language. All hosted tools and workflows are described in CWL which is both human and machine-readable and has all the necessary information to run the tool in a reproducible way.
Task: Single execution of a bioinformatics tool or workflow.

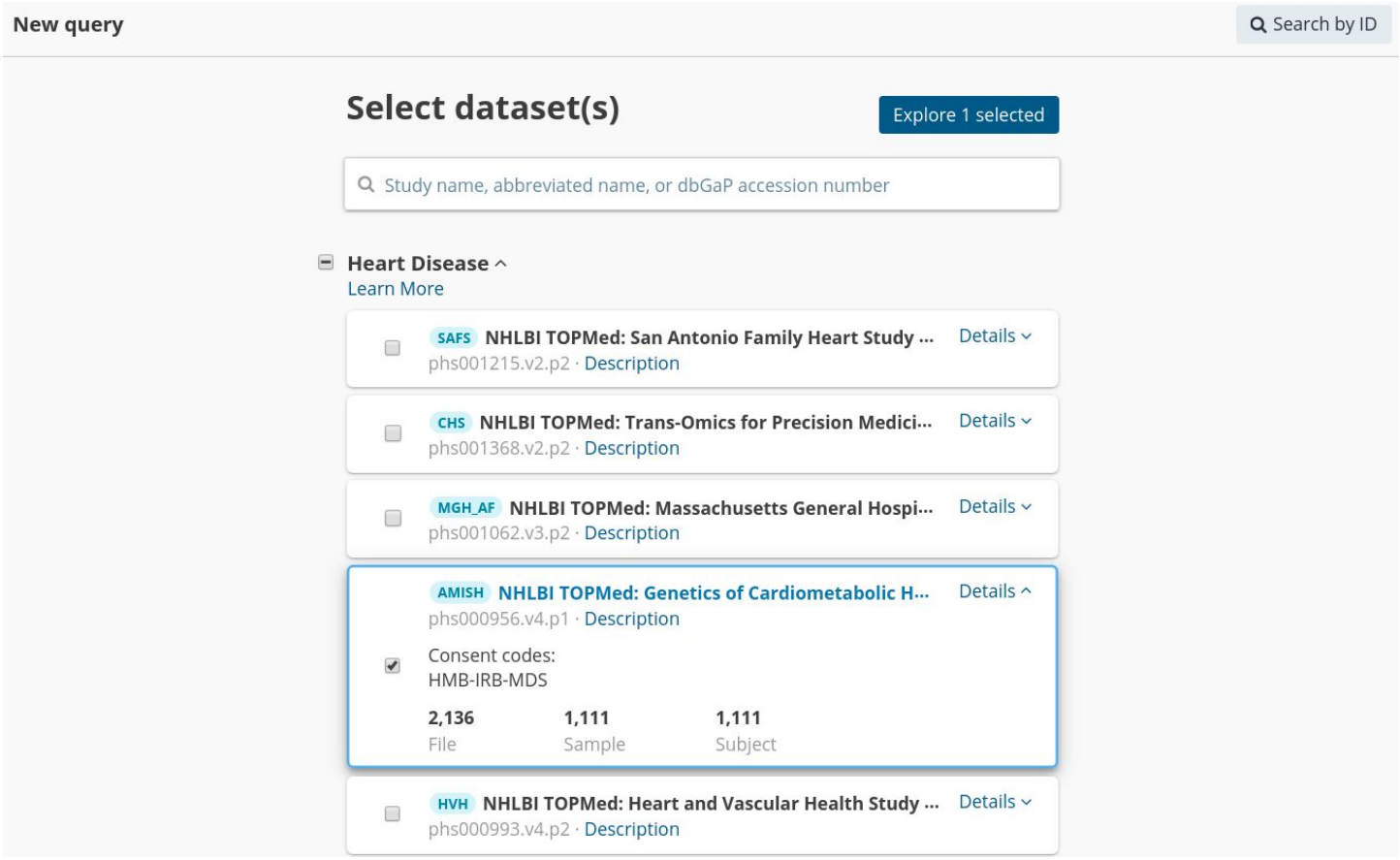
To complete the steps in this tutorial, please create a new project with the title “test GWAS.” Please choose the following settings for the project:

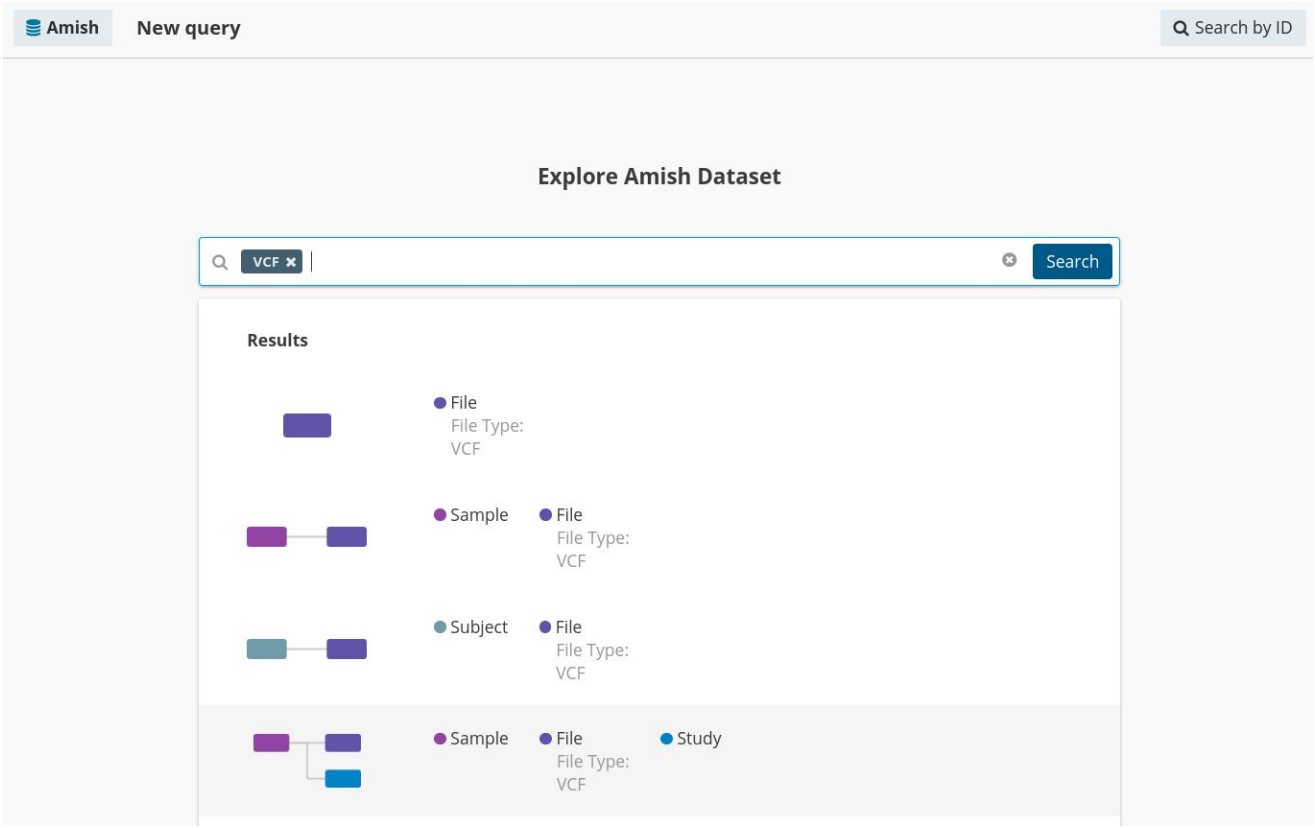
Billing Group: Pilot Funds
Location: aws-us-east-1
Spot instances: On
Memoization: Off
Controlled data: Yes

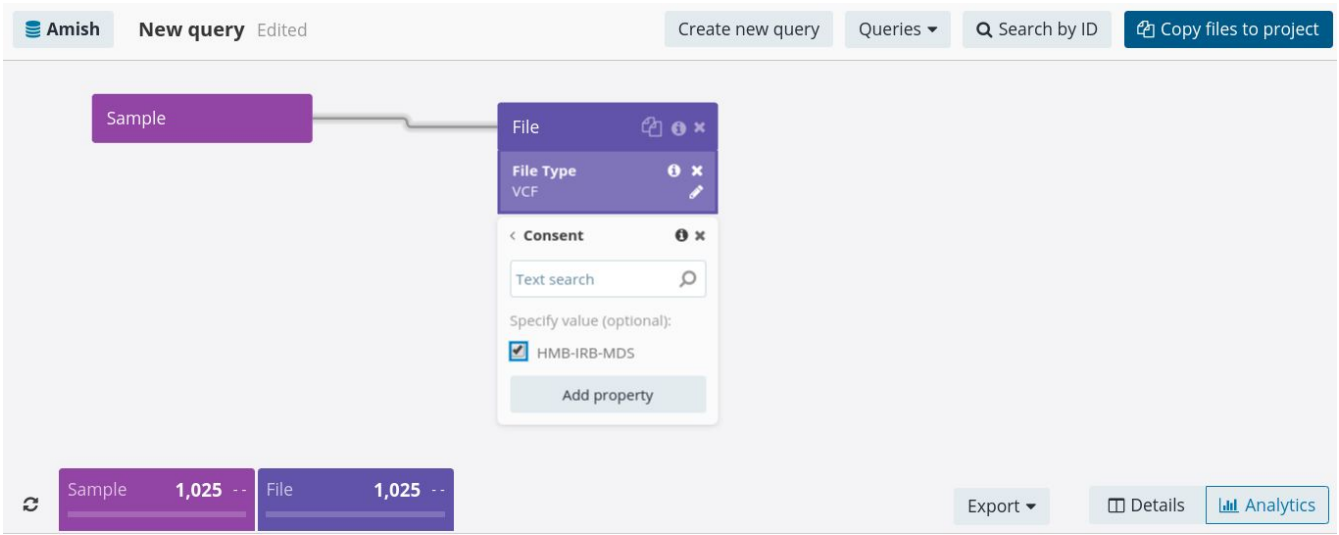
Accessing TOPMed WGS variant (VCF) data on BioData Catalyst

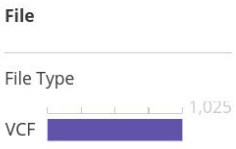
Users can find the hosted TOPMed studies in the Data Browser feature. This includes multi-sample VCF files on a per study and consent code basis, as well as single-sample VCF files and CRAM files. The Data Browser enables users to select files from TOPMed studies and consent codes that they have access to and bring these to a project for analysis. Users can form cohorts of multiple TOPMed studies.

For the purposes of this tutorial, please select one or more TOPMed studies that you know you are approved to access. Choose “Explore Selected” and then search for the consent codes you are approved for and the multi-sample VCFs.









Now that we have selected our files, we will copy them to the project we created. In the top right corner, select "Copy files to project" and select the "test GWAS" project. Since the identified files are controlled access, the

platform will only allow them to be copied to a controlled project. Importantly, when files are “copied” to a project, they are not duplicated on the storage infrastructure. Instead, a soft link is created to the original file.

Open the “test GWAS” project and go to the “Files” tab to see all of the files available in the project.

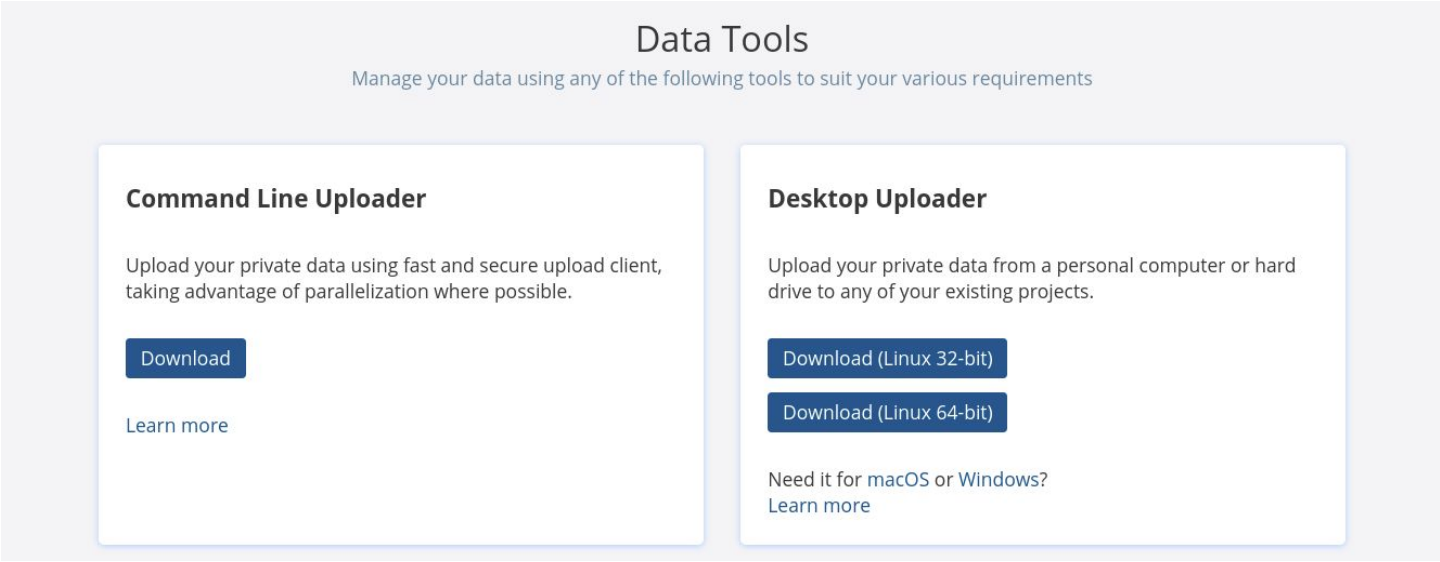
Uploading TOPMed Exchange Area files

In addition to the VCF files, you will need the following additional types of files to perform the GWAS:

1. Principal components
2. Kinship Matrix
3. Phenotype files

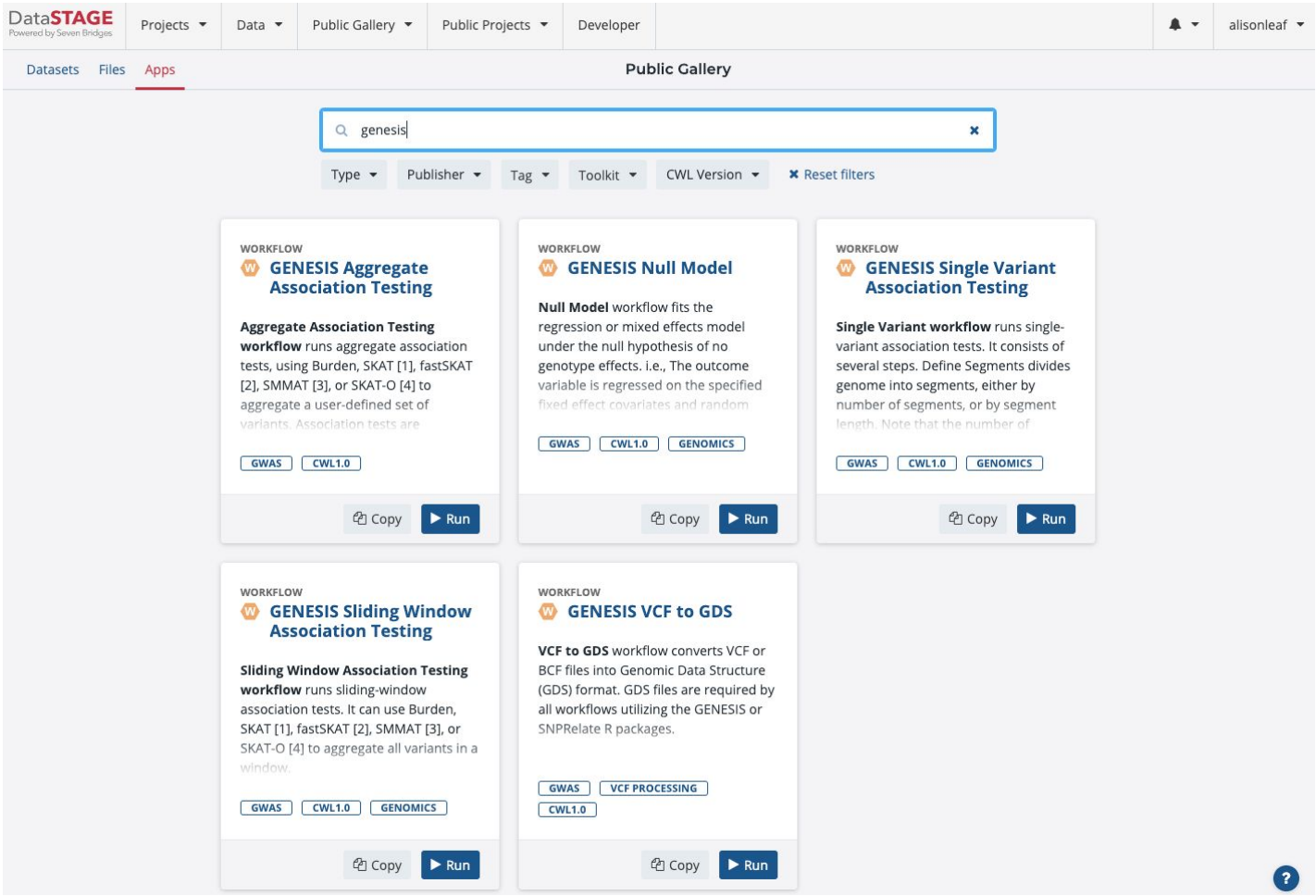
BioData Catalyst powered by Seven Bridges does not currently host these files from the TOPMed Exchange Area. To get these files on the platform, please upload them to your project using either the command line uploader (linux) or the GUI-based Desktop Uploader (Windows, Mac OS, Linux).

The data tools and others can be accessed on the platform by going to “Data” on the top navigation bar and then selecting “Data Tools.” Please refer to the Documentation pages on “Bring your own data” to read additional details on how to use these features (https://f4c.readme.io/docs/upload-to-f4c).



Adding GENESIS Apps from Public Gallery

The Seven Bridges bioinformatics team worked with the TOPMed Data Coordinating Center to bring CWL versions of the GENESIS tools to *BioData Catalyst*. These tools have been optimized for the cloud and configured into workflows (combinations of multiple tools) to make running the analyses more efficient. The GENESIS Apps can easily be found in the Public Gallery as shown below:



To find the pipelines click on ***Public Gallery*** - ***Apps***. Search for “GENESIS” and the workflows will pop up. This will include 5 apps: VCF to GDS, null model, and 3 options for association apps (single variant, aggregate, sliding window). For each App you plan to use, click **Copy** and select your working project as a destination.

For this tutorial, please copy the following workflows to the “test GWAS” project:

- VCF to GDS
- Null model
- Single-variant association testing
- Aggregate association testing

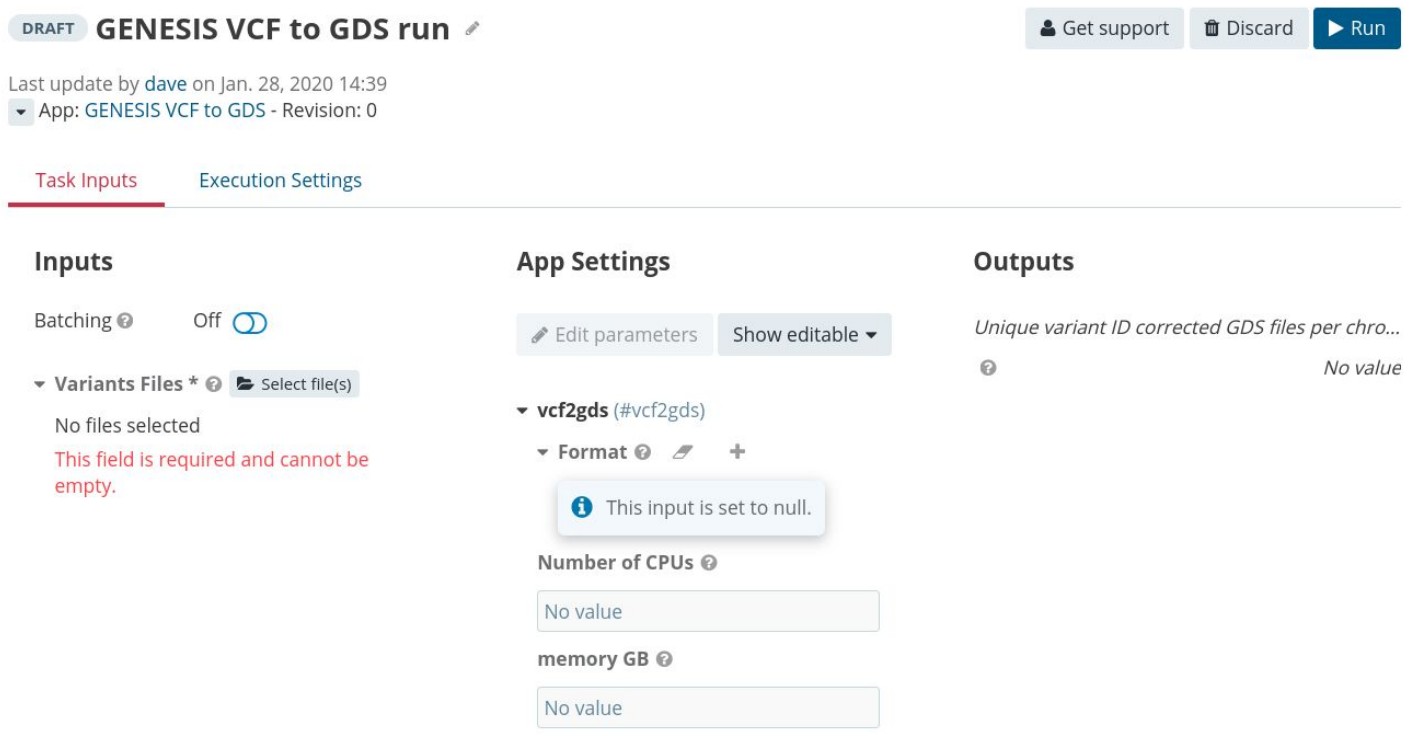
You should now see these workflows in the **Apps** tab of your “test GWAS” project.

VCF to GDS Conversion

The hosted TOPMed data is in the VCF file format. However the GENESIS packages require the inputs to be in the Genomic Data Structure (GDS) format. Therefore, now that the data files have been added to the project, the next step is to convert them to GDS files using the "GENESIS VCF to GDS" workflow.

This workflow is composed of three tools. Step 1 (**vcf2gds**) converts VCF or BCF files (one per chromosome) into GDS files, with option to keep a subset of FORMAT fields, by default only GT field. (BCF files may be used instead of VCF.) Step 2 (**Unique Variant IDs**) ensures that each variant has a unique integer ID across the genome. Step 3 (**Check GDS**) ensures that no important information is lost during conversion. If **Check GDS** fails, it is likely that there was an issue during the conversion. The workflow will run **vcf2gds** once per chromosome and check that each use the **unique_variant_id** app to make sure each variant in the new GDS has a unique integer id.

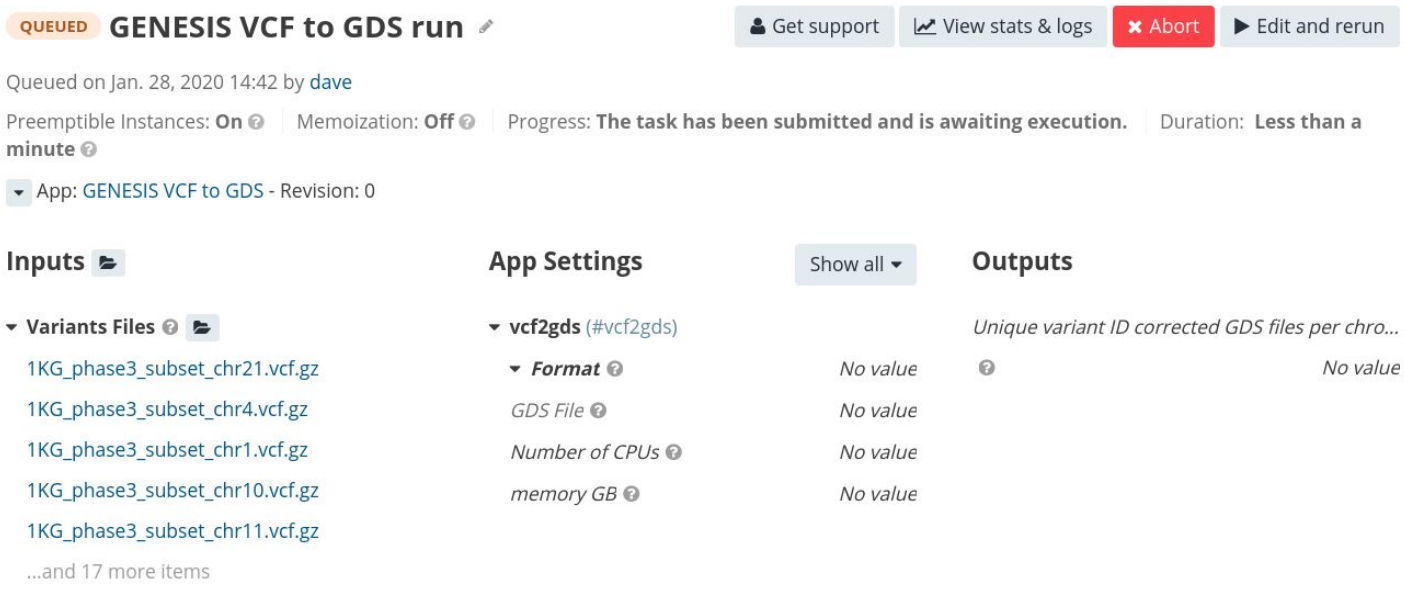
When you are ready to run the workflow, click on the **Apps** tab in your project and select "Run" on the VCF to GDS conversion app. Your next screen will show a **DRAFT** Task setup page and look like this:



Inputs / Settings / Expected Outputs

Select all the VCF files in your project and add them as the **Variant Files** input. The app will process them all in parallel. This is an example of a form of cloud platform parallelization called **scattering**.

Click **RUN** and your next screen will look like:



The task will go through 3 status labels: **Queued, Running** and **Complete.** Once your task is complete you will be able to continue with the next section.

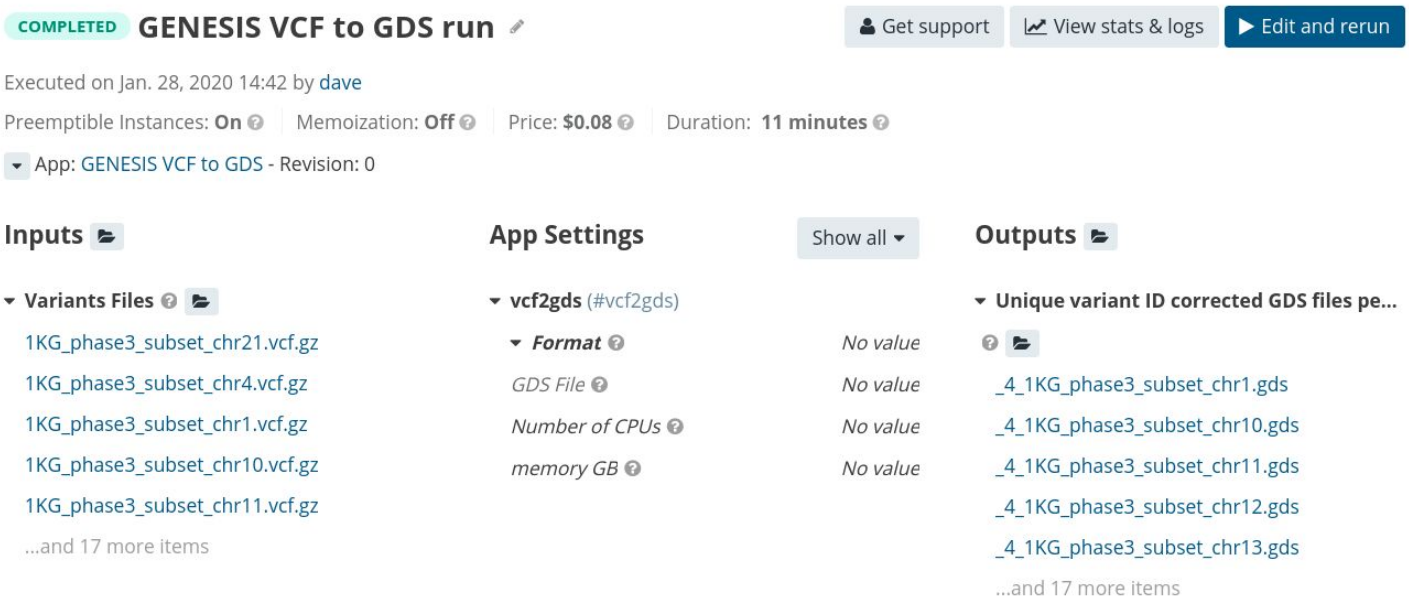
Using the 1000G subset data will allow for quick testing of these pipelines. If you are using TOPMed study data the tasks will take longer. Once completed, your GDS file outputs will be listed under the “Outputs” section of the completed Task page.

Please note that you will only need to generate the GDS files once for a particular set of samples. These GDS files can be used in multiple association tests with different phenotypes.

Common issues and important notes:
This pipeline expects that input VCF files are separated per chromosome and that file is properly named. The expected format of the file name is: **basename_chr##.[bcf, bcf.gz, vcf, vcf.gz]**, where the “##” is the name of the chromosome (1-22, X, Y).

The parameter for “number-of-CPUs” should only be used when working with VCF files. The workflow is unable to utilize more than one thread when working with BCF files, and will fail if the number of threads is set for BCF conversion.

Note: Variant IDs in output workflow might be different than expected. Unique variants are assigned for one chromosome at a time, in ascending order, and chromosomes are sorted in natural order (1,2,3,..,22,X,Y). Variant IDs are integer IDs unique to your data and do not map to rsIDs or any other standard identifier. Be sure to use **variant_id** file for down-the-line workflows generated based on GDS files created by this workflow.



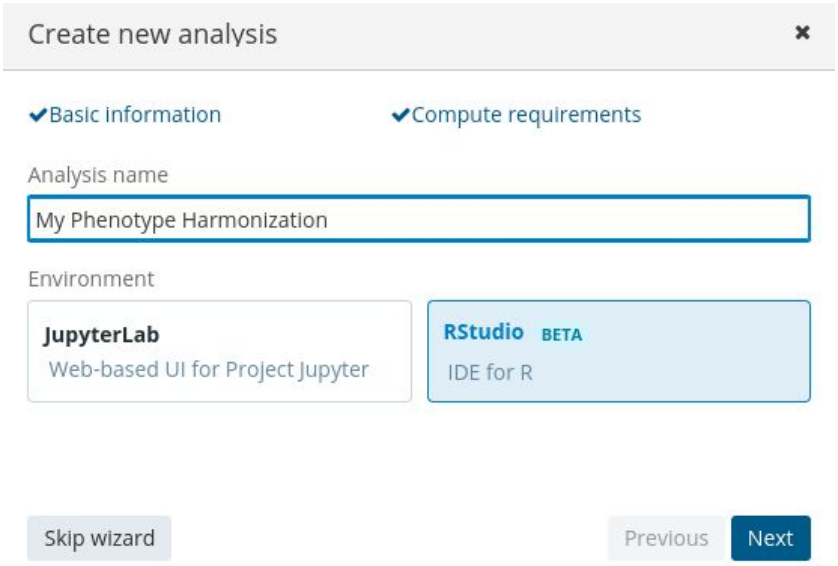
Please note that if you are working with multiple VCF files, we recommend first merging them into a single VCF and then converting that single VCF file to the GDS format. This was found to be faster than converting multiple VCFs to the GDS format and then merging the GDS files. We recommend using Bcftools Merge to combine the VCF files. This tool can be found in the Public Apps Gallery. It is not recommended to merge across chromosomes. The end results should be 1 vcf files per chromosome. Each VCF should include all your study samples.

Phenotype Harmonization

Many researchers will need to perform a phenotype harmonization step in their association studies. This step can include renaming, transforming, or otherwise processing measured clinical variables to create a comparable variable across studies for one or more phenotypes of interfest. We recommend using the platform implementation of RStudio for these steps, although Jupyterlab notebooks are also available. Users can use either the raw phenotype files (txt) that are hosted on the platform or they can upload their own phenotype files to the platform using the same steps outlined in “Uploading TOPMed Exchange Area Files.” The first section below outlines how to find the hosted raw phenotype files on the platform. The second section describes how to use RStudio on the platform.

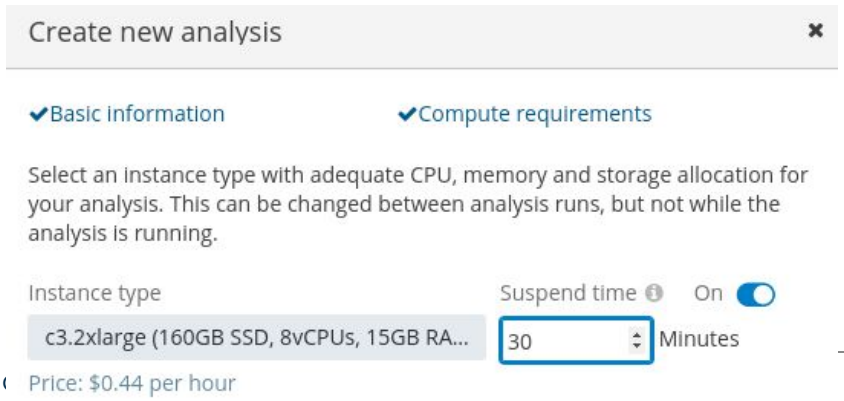
Find hosted phenotype data in Data Browser

BioData Catalyst hosts phenotype files for each TOPMed study and consent code. These files can be found in the Data Browser in the same way the multi-sample VCF files were found. Go to “Data” on the top navigation bar and then select the “Data Browser.” Choose the TOPMed studies and consent codes that were used for the VCF data by doing XXXX



RStudio on the Cloud

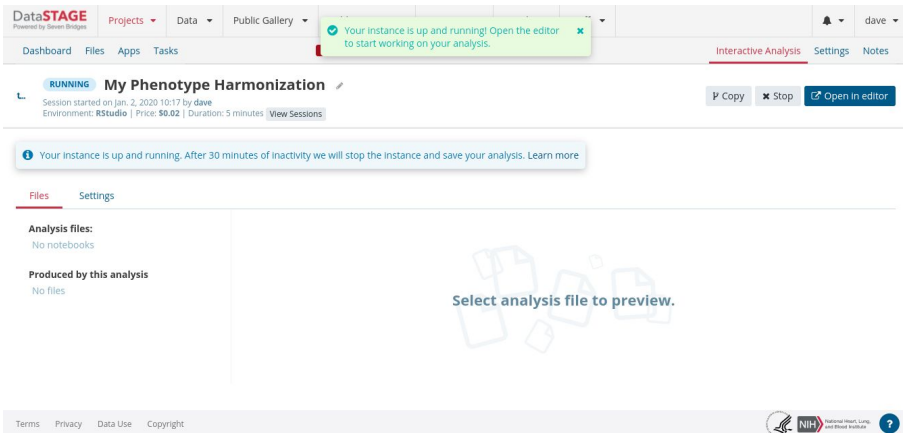
RStudio can be opened from within a project. From the “test GWAS” project, go to the upper right-hand corner and select “Interactive Analysis.” Then select the “Data Cruncher” feature and “Create your first analysis.” After naming your analysis, select to open “RStudio.”



You will have the option to select an instance type. For this tutorial the default instance (aws c3.2xlarge) will be sufficient . Users can also set the **Suspend time** which is the period of analysis inactivity after which the instance is stopped automatically. Inactivity implies that:

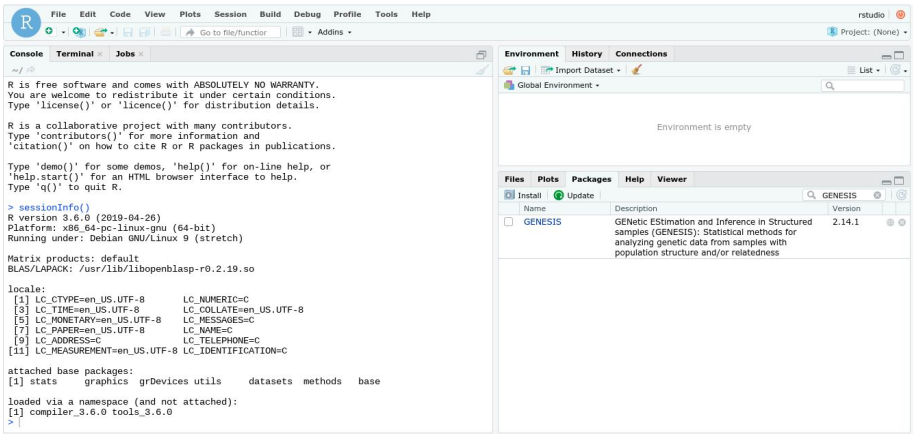
- No files have been modified or created under the **Files** tab (in the /sbgenomics/workspace directory if you are using the Terminal).
- There are no running jobs.

When the suspend time is reached, the instance is stopped as well as the analysis. All files are saved that meet the criteria for automatic saving or have been selected to be saved as project files. Files that do not meet the criteria and are not manually saved to the project will be lost. By setting suspend time, users don't have to worry about stopping the RStudio instance to prevent it from continually running in the background if the user moves on to other things.



Make sure to click "Open in editor" once the RStudio instance has been initialized. This will open a new window for you which contains the RStudio IDE.

The cloud version of RStudio is the same software that you would run locally with the exception that Seven Bridges has pre-installed many relevant R packages. These include the tidyverse packages (dplyr, ggplot2, readr, tidyr..etc) and other useful data science packages.



For more information on launching RStudio in the cloud please see the documentation at https://f4c.readme.io/docs/rstudio-quick-reference

Formatting the Phenotype File for GENESIS

The GENESIS workflows expect the data in the phenotype file to be provided as an R data.frame or AnnotatedDataFrame, which can be constructed from a standard R data.frame using the AnnotatedDataFrame()

function in the Biobase package [REF]. All phenotype data to be used in the association testing models should be present in the phenotype file, including the outcome, covariates, and group variate values (principal components of ancestry do not need to be included, as they can be specified via a special PCA File). The only other requirement is in regard to how the unique sample identifiers (i.e. sample.id) are specified. These sample.id must take the same values as the sample.id in the GDS genotype files (which will be the same sample identifiers from the VCF files if constructed using the VCF to GDS conversion workflow). If using a data.frame, the sample.id must be the rownames; if using an AnnotatedDataFrame, the sample.id must be provided in a column named 'sample.id'.

Preparation of Other Optional Files for GENESIS

Ancestry and Relatedness

GENESIS implements PC-AiR, which unlike standard principal components analysis (PCA), accounts for relatedness in the sample to provide accurate ancestry inference that is not confounded by family structure. In addition, GENESIS implements PC-Relate, which uses ancestry representative principal components (PCs), such as those from PC-AiR, to adjust for population structure and provide accurate estimates of kinship coefficients. An ancestry and relatedness inference pipeline utilizing these methods is being developed on the Seven Bridges platform; until that pipeline is released, these analyses can be performed in R using the GENESIS library, or pre-computed PCs and kinship coefficients for TOPMed freeze 5b are available <where?>.

Inclusion of ancestry PCs and/or a kinship coefficient matrix is not required to run the GENESIS association testing workflow, but it is generally recommended in TOPMed, as most studies have complex population and/or pedigree structure.

The GENESIS association testing workflow expects ancestry PCs to be provided as an RData object in a separate PCA file. It expects this RData object to take the form of the output from the pcair() function in the GENESIS package; specifically, to be a list containing a data.frame named 'vectors' of PC values, with the sample.id as the rownames of this 'vectors' data.frame.

The GENESIS association testing workflow expects the kinship matrix to be provided as an RData object in the Relatedness Matrix file. It expects this RData object to be of class matrix or Matrix, with pairwise kinship coefficients as the matrix entries, and sample.id as both the rownames and colnames. Using the GENESIS package, the output of the pcrelate() function can be transformed to this format using the pcrelateToMatrix() function. Sparse matrices are supported for significantly improved computational efficiency in large samples, and can be constructed using the `thresh` parameter of the pcrelateToMatrix() or makeSparseMatrix() functions in GENESIS. Other types of relatedness matrices, such as a genetic relatedness matrix (GRM) calculated with GCTA[9], can be used in the GENESIS association testing workflow, as long as they are provided in the file format specified above.

Running an Analysis on a Subset of Samples

An analyst may wish to run an analysis with only a subset of the samples in their project. Rather than requiring them to subset their Phenotype file, PCA file, and Relatedness Matrix file, a subset of samples can be specified

for analysis by using the Sample include file. This Sample include file should be an RData file with a vector of sample.id to include in the analysis. When not specified, all samples in the provided data are included in the analysis.

Running an Analysis on a Subset of Variants

An analyst may wish to run an analysis with only a subset of variants included in their GDS genotype files. Rather than requiring them to create a new GDS file, a subset of variants can be specified for analysis by using Variant Include files. There should be one Variant Include file per chromosome being tested; the expected format of the file name is **basename_chr##.RData**. Each of these RData files should be a vector of variant.id (matching the 'variant.id' field in the GDS file that was assigned by the VCF to GDS conversion Unique Variant ID step) to include in the analysis. When combined with other variant filters (e.g. PASS only, MAC threshold, MAF threshold), only variants specified in the Variant Include files and passing the filters will be included in the analysis.

User Provided Variant Weights for Aggregate Tests

When performing multi-variant aggregate association tests, it is common practice to weight each variant included in the aggregation unit. An analyst has the option to specify weights, perhaps based on variant annotation information, using a Variant Weight file. This RData file should be a data.frame with columns 'variant.id' (matching the 'variant.id' field in the GDS file that was assigned by the VCF to GDS conversion Unique Variant ID step), or 'chr', 'pos', 'ref', 'alt' to identify the variant, and a column of weights; the name of the column of weights can be anything and is provided via the Weight user parameter. If a variant grouping file is used (see next section), "Weight user" may also refer to a column in that file, and a separate variant weight file is not needed.

User provided Variant Groupings for Aggregate Tests

When performing multi-variant aggregate association tests, a user can specify the groups and the variants that should be grouped under them using a variant grouping file. This is an RData file with data frame defining aggregate groups. If the value for the config parameter 'aggregate_type' is 'allele', columns should be 'group_id' , 'chr', 'pos', 'ref', 'alt'. Value of 'group_id" is a group identifier (example gene name) under which the variant specified by 'chr', 'pos', 'ref', 'alt' should be grouped. A given variant can be assigned to more than one group by including multiple row entries for that variant, each with a different 'group_id'. Variant grouping files of this format can be generated using the `Annotation Explorer` application on BioData Catalyst platform which allows the user to perform variant filtering and groupings based on variant annotations. If the value for the config parameter 'aggregate_type' is 'position', columns should be 'group_id', 'chr', 'start', 'end'. The latter format is used to group variants in the contiguous genomic regions defined by 'chr', 'start', 'end' . Only variants that pass the specified variant filters (e.g. PASS only, MAC threshold, MAF threshold) will be included in the aggregate analysis even if they are included in the variant grouping file.

Fit Null Model

The next step is to fit a **Null Model**. The **Null Model** workflow fits the regression or mixed effects model under the null hypothesis of no genotype effects; i.e., the outcome variable is regressed on only the specified fixed effect covariates and random effects. The single variant genotypes or multi-variant aggregation units are not included in this model. The output of this **Null Model** is then used in the association test workflow of choice.

This workflow consists of two steps: The first step fits the **Null Model**, and the second one generates reports based on the fitted model and data. Reports are available both in Rmd and html format.

Both quantitative (**binary** = FALSE) and binary, e.g. case/control, (**binary** = TRUE) phenotypes are supported. The outcome and covariate values must be included in the **Phenotype file** and specified with the **Outcome** and **Covariates** parameters. Principal components of ancestry can be included as covariates using the **PCA file** input, and genetic relatedness can be accounted for as a random effect using the **Relatedness matrix file** input.

When working with a quantitative phenotype, heterogeneous residual variances by group can be specified using the **Group variate** parameter.

This workflow utilizes the `fitNullModel` function from the GENESIS software.

Common issues and important notes

If **pca_file** is not provided, the **n_pcs parameter** must be set to 0.
pca_file must be an RData object output from the pcair function in the GENESIS package.
The **Null Model** job can be very computationally demanding when fitting mixed models in large samples (e.g. > 20K). GENESIS supports using sparse representations of matrices in the r**elatedness_matrix_file** via the R Matrix package, and this can substantially reduce memory usage and CPU time[11].

Running Single Variant Association Tests

For users who would like to perform a single-variant association test, please use the “GENESIS Single Variant Association Testing” workflow. The single-variant test is similar to a traditional “GWAS”.

The Single Variant Workflow consists of several steps. First, the “**define_segments.R**” tool divides the genome into segments, either by a set number of segments, or by segment length. Note that number of segments refers to the whole genome, not a number of segments per chromosome. Next, the association test is performed for each segment in parallel, before combining results on the chromosome level. The final step in the workflow produces QQ and Manhattan plots of the association p-values.

This workflow uses the output from the null model workflow and the genotype data to perform score tests for all variants individually. The reported effect estimate is for the alternate allele, and multiple alternate alleles for a single variant are tested separately.

When testing a binary outcome, the saddlepoint approximation (SPA) for p-values [REFS] can be used by specifying **test_type = ‘score.spa’**; this is generally recommended[4, 12]. SPA will provide better calibrated p-values, particularly for rarer variants in samples with case-control imbalance.

If your genotype data has sporadic missing values, they are mean imputed using the allele frequency observed in the sample.

On the X chromosome, males have genotype values coded as 0/2 (females as 0/1/2).

This workflow utilizes the `assocTestSingle` function from the GENESIS software.

Common issues and important notes

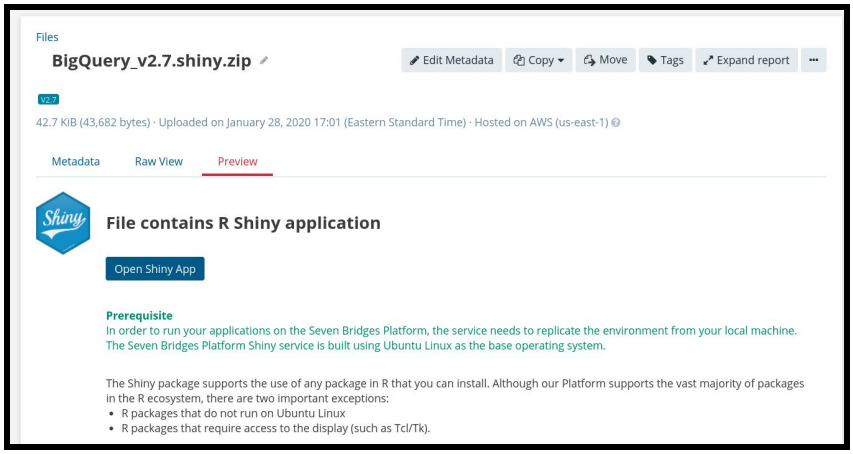
Assoc_single job can be very memory demanding, depending on the number of samples and null model used. We suggest running with at least 5GB of memory allocated for small studies, and to use approximation of 0.5GB per thousand samples for larger studies (with more than 10k samples), but this again depends on complexity of null model. If a run fails with **error 137**, and with the message “killed” displayed, the most likely cause is a lack of memory.

Running Aggregate Associations

This section is specifically for those doing an aggregate unit association which is a specific type multiple-variant association test. The BioData Catalyst platform also has *GENESIS Sliding Window Association Testing* app which aggregates by genomic positions. The sliding window test will not be covered in this tutorial but please see the app description for more information.

TOPMed Annotation Explorer

The TOPMed Annotation Explorer is an application developed by Seven Bridges in collaboration with the TOPMed Data Coordinating Center which enables users to interactively explore an inventory of annotations for the TOPMed studies. This feature can be used to aggregate and filter variants based on annotations for use in a multi-variant association study.

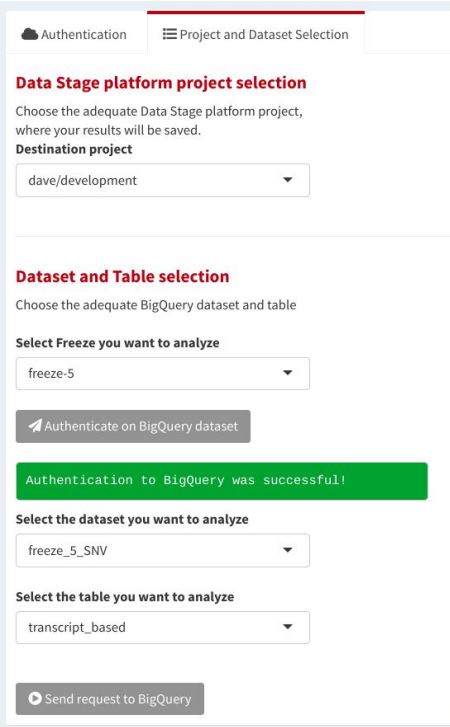


The Annotation Explorer currently hosts annotations for TOPMed Freeze5 variants as well as TOPMed Freeze8 variants. Researchers who are approved to access one or more of the TOPMed Freeze5 studies will have the ability to access the Annotation Explorer Freeze5 data.

At this time, only the TOPMed Data Coordinating Center has access to the Freeze8 annotations; however, as Freeze8 is being released on dbGaP, the BioData Catalyst consortium will work to make this data available on the platform as well with associated permissions. After this work is completed, Freeze8 annotations will also be available to researchers.

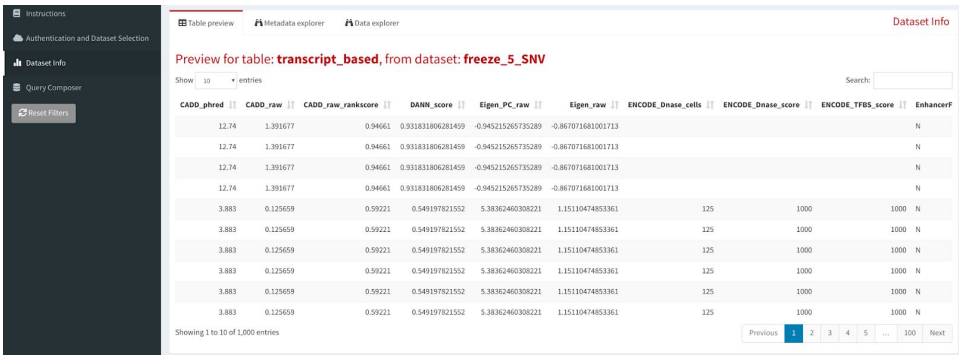
For the purposes of this tutorial, we will show a simple query of the Freeze5 annotations. Please refer to XXX tutorial written by the TOPMed Data Coordinating Center for more detail on how to take advantage of the Annotation Explorer feature.

To start the Annotation Explorer, go to your files table in the “test gwas” project. The Annotation Explorer is loaded from a file that can be added to your project (please contact Seven Bridges for this file). Once added to your project click on file to view the description. Click “**Open Shiny App**”. The app will load quickly.



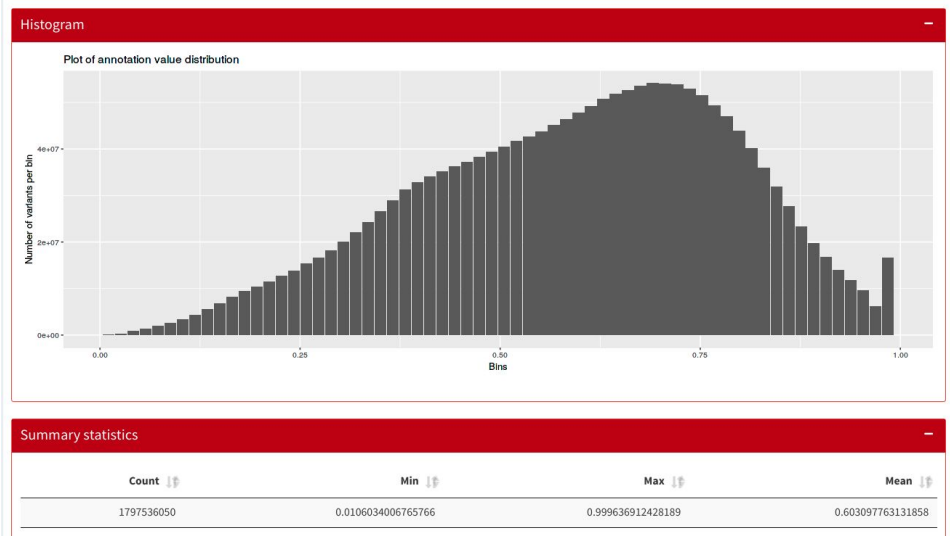
After starting the Annotation Explorer you will need to enter your private authentication token found in this link. Next select your Destination project for the aggregate unit file. For this tutorial, we will use freeze-5 as the data to analyze so select "freeze-5" in the **Dataset and Table selection** and click **Authenticate on BigQuery Dataset**. You will see a message that indicates "Authentication to BigQuery was successful"

Next select freeze_5_SNV and transcript_based for the specific annotation table we will use in this tutorial. Select send request to BigQuery. What this will do is send back a preview (1,000 rows) of the unfiltered annotations.



The dataset info tab is a great place to learn about the annotations available for your research in real time. Of important note is the metadata explorer tab which will give text descriptions of the hundreds of annotations available.

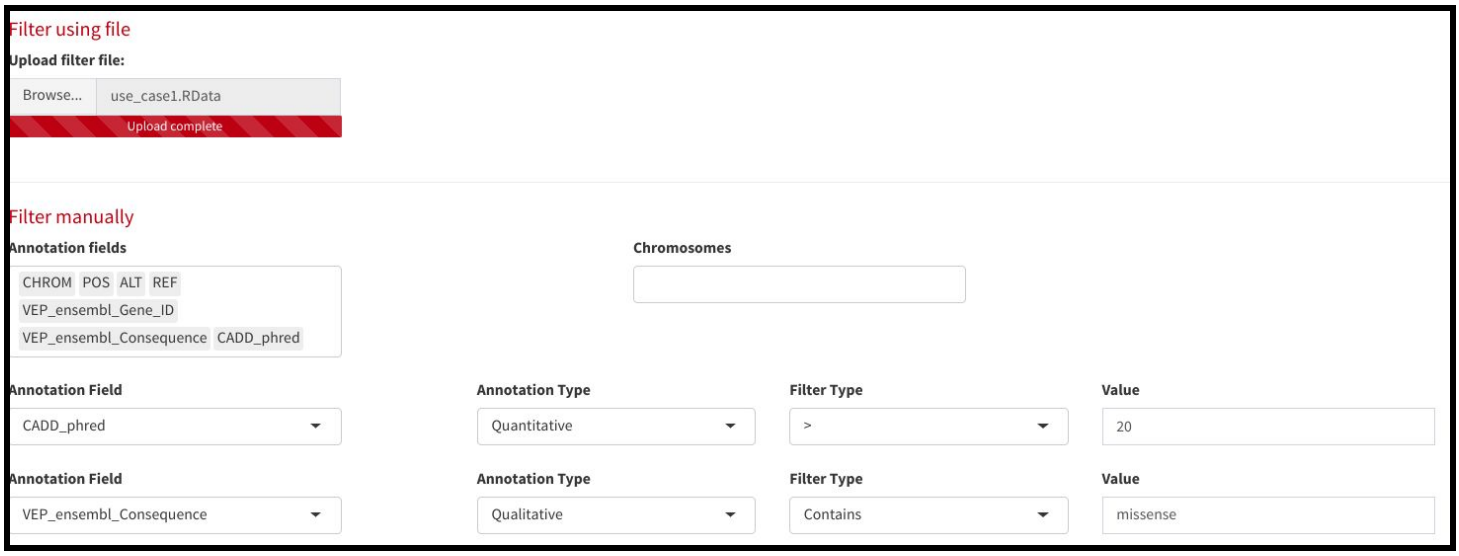
The data explorer will give distribution and summary statistics for the 100's of annotation fields.



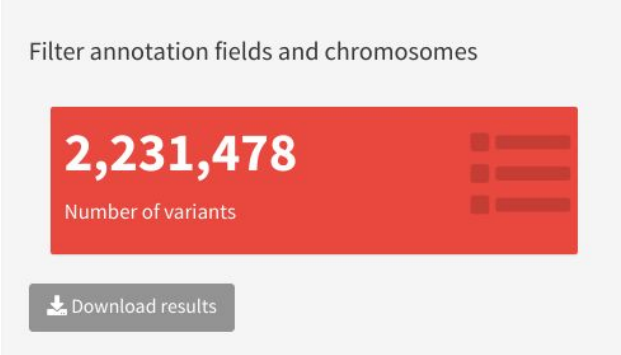
The final section of the Annotation Explorer application is the query composer where you can interactively create the aggregation units for the subsequent association tests.

For this example, we will use a quantitative and a qualitative field type. We will filter for variants that have a CADD_phred score of greater than 20 and contain the term “missense” in the VEP_Ensembl_Consequence field.

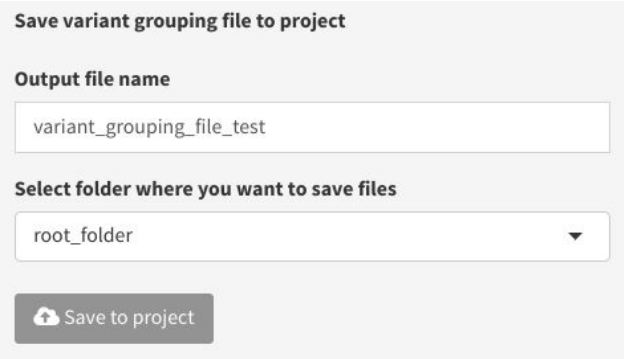
.



Click “Filter” towards the bottom of the app screen. This filters through the 100’s of millions of variants by using the parameters we loaded in. The process takes only a few seconds and it will create a new table with only the filtered variants.



Next switch to the aggregation tab, select VEP_ensembl_Gene_ID for the aggregation field and click on “Aggregate”. The “Interactive plots” tab is used to QC the total number of aggregate units and the number of variants within each aggregation.



Next switch to the “Export results” tab, fill out the section titled “Save variant grouping file to project”

The aggregate files will be saved to the root directory of your project. There will be 1 file for each chromosome. This completes the section on the Annation Explorer. For more in depth information and additional user cases please look for the stand alone Annotation Explorer reference materials.

Aggregate association test

Once we have the aggregate variant files in our project we can setup our draft task and run the multi-variant association test.

This workflow runs aggregate association tests, using Burden, SKAT[5], fastSKAT[7], SMMAT[8], or SKAT-O[6] on a user-defined set of variants. Association tests are parallelized by segments within chromosomes.

Define segments splits the genome into segments and assigns each aggregate unit to a segment based on the position of its first variant. Note that the number of segments refers to the whole genome, not a number of segments per chromosome. Association testing is then for each segment in parallel, before combining results on the chromosomal level. Finally, the last step creates QQ and Manhattan plots of the association p-values.

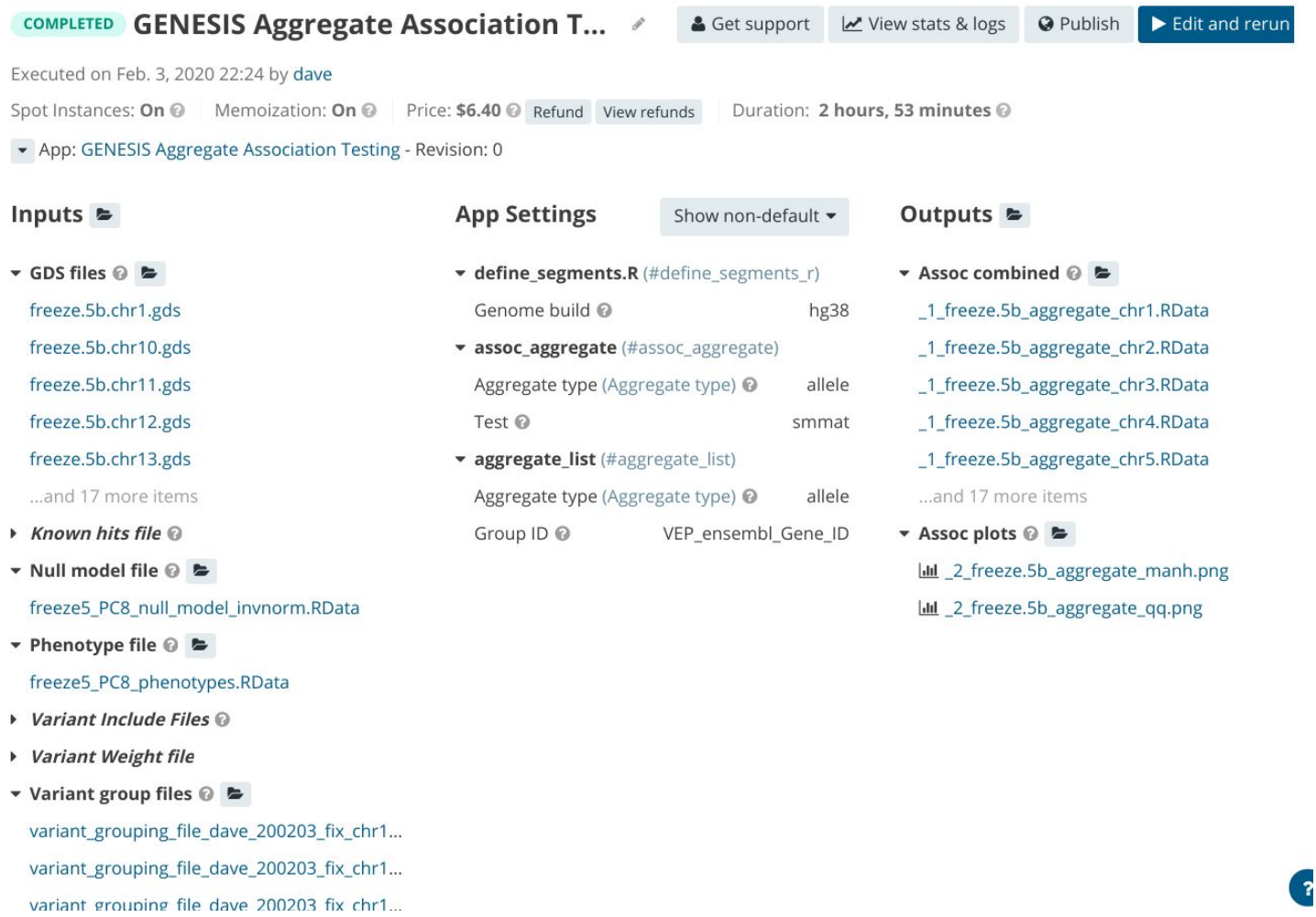
This workflow uses the output from the null model workflow and the genotype data to perform the specified test for each aggregation unit.

Aggregate tests are typically used to jointly test rare variants. The **alt_freq_max** parameter allows specification of the maximum alternate allele frequency allowable for inclusion in the test. Included variants are usually weighted using either a function of allele frequency (specified via the **weight_beta** parameter) or some other annotation information (specified via the **variant_weight_file** and **weight_user** parameters). Multiple alternate alleles for a single variant are treated separately.

If your genotype data has sporadic missing values, they will be mean imputed using the allele frequency observed in the sample.

When running a burden test, the effect estimate is for each additional unit of burden; there are no effect size estimates for the other tests.

This workflow utilizes the `**assocTestAggregate**` function from the GENESIS software.

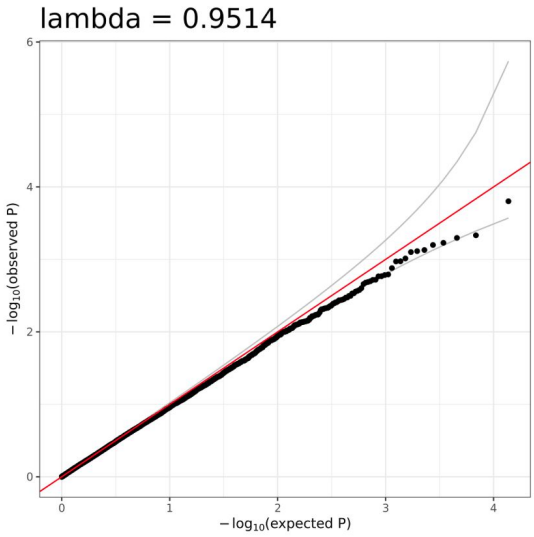


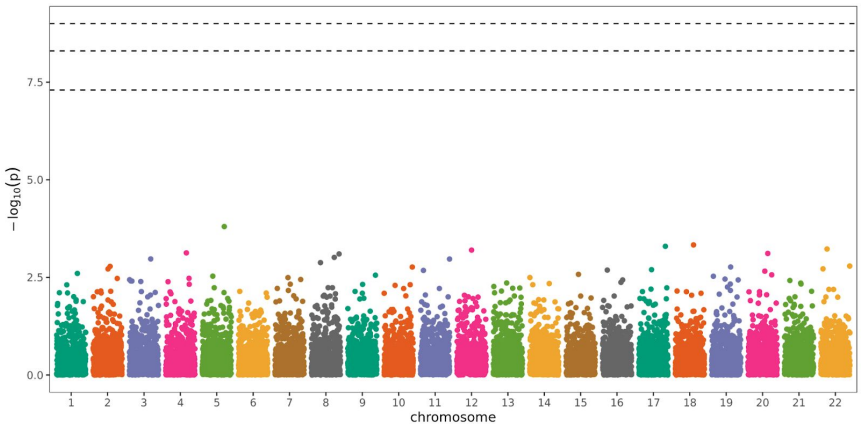
The **Number of Segments** parameter, if provided, needs to be equal or higher than the number of chromosomes.

Testing showed that default parameters for CPU and Memory (8GB) are sufficient for testing studies (up to 50k samples), however different null models might increase the requirements.

Results and Quality Control

The typical quality control plots are QQ plot and Manhattan plot. These are both available to you regardless of the association test you will choose to run. The “png” files can be viewed natively on the platform using your web browser. As you iterate over your different association test experiments these plots will remain saved in your project to refer back to at a later time.





Finding Platform Support

Please reach out to the Seven Bridges team if you need assistance with performing an association study on *BioData Catalyst powered by Seven Bridges*. We are happy to schedule a tailored training with you and your research group to help you get started with analyses quickly.

If you test out the association pipelines and encounter any difficulties with completing the analyses, you can also click the “?” on the lower right-hand part of the page and reach out to our Support Team.

References

1. Chen, H., Wang, C., Conomos, M. P., Stilp, A. M., Li, Z., Sofer, T., & Redline, S. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. The American Journal of Human Genetics. 2016;98(4), 653-666.

2. Conomos, M. P., Miller, M. B., & Thornton, T. A. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genetic Epidemiology. 2015;39(4), 276-293.

3. Conomos, M. P., Reiner, A. P., Weir, B. S., & Thornton, T. A. Model-free estimation of recent genetic relatedness. The American Journal of Human Genetics, 2016;98(1), 127-148.

4. Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., & Bastarache, L. A. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nature Genetics. 2018;50(9), 1335-1341.

5. Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. Rare-variant association testing for sequencing data with the sequence kernel association test. The American Journal of Human Genetics. 2011;89(1), 82-93.

6. Lee, S., Wu, M. C., & Lin, X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012;13(4), 762-775.

7. Lumley, T., Brody, J., Peloso, G., Morrison, A., & Rice, K. FastSKAT: Sequence kernel association tests for very large sets of markers. Genetic Epidemiology, 2018;42(6), 516-527.

8. Chen, H., Huffman, J. E., Brody, J. A., Wang, C., Lee, S., Li, Z., ... & Blangero, J. Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. The American Journal of Human Genetics. 2019;104(2), 260-274.

9. Yang et al. GCTA: a tool for Genome-wide Complex Trait Analysis. American Journal of Human Genetics. 2011;88(1): 76-82.

10. Sofer, T., Zheng, X., Gogarten, S. M., Laurie, C. A., Grinde, K., Shaffer, J. R., ... & Lange, L. A fully adjusted two-stage procedure for rank-normalization in genetic association studies. Genetic Epidemiology. 2019;43(3), 263-275.

11. MISSING INFO

12. Dey, R., Schmidt, E. M., Abecasis, G. R., & Lee, S. A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. The American Journal of Human Genetics. 2017;101(1), 37-49.