

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

Table of Contents

Objectives	4
Overview of the GENESIS pipelines	4
Organizing a Workspace and Preparing Data	6
Project Creation	6
Accessing TOPMed WGS variant (VCF) data on BioData Catalyst	7
Adding GENESIS Apps from Public Gallery	9
VCF to GDS Conversion	10
Uploading additional files for the association test	13
Phenotype Harmonization	14
Find hosted phenotype data in Data Browser	14
RStudio on the platform	15
Formatting the phenotype file for GENESIS	17
Preparation of Other Optional Files for GENESIS	17
Ancestry and relatedness	17
Running an analysis on a subset of samples	18
Running an analysis on a subset of variants	18
Fit Null Model	18

Running Single Variant Association Tests	19
Multiple-variant Association Tests	21
Preparation of Files for Multiple Variant Associations using GENESIS	21
User provided variant groupings for aggregate tests	21
User provided variant weights for aggregate tests	21
Annotation Explorer	22
Launch Annotation Explorer	22
Authentication and Data selection	26
Running Multiple-variant association test	31
Results and Quality Control	33
Finding Platform Support	34

Objectives

The purpose of this tutorial is to guide users through the steps of running a single variant or multiple variant association test on NHLBI BioData Catalyst powered by Seven Bridges using a particular set of association pipelines: the GENESIS R package pipelines. This tutorial is designed to be a starting point for researchers who have a general understanding of genome-wide association studies. Previous use of R or the GENESIS package is not required.

Prior to beginning this tutorial, please go through the steps of the [Onboarding Tutorial](#) which provides 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
- Using the Data Browser feature

Overview of the GENESIS pipelines

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. 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 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 GMMAT¹ for association testing of binary (e.g. case/control) phenotypes. The mixed models utilize the PC-AiR² PCs and a relatedness matrix of PC-Relate³ 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

¹ (2016, March 24). Control for Population Structure and Relatedness for Binary Retrieved March 24, 2020, from <https://www.ncbi.nlm.nih.gov/pubmed/27018471>

² (2015, March 23). Robust inference of population structure for ancestry Retrieved March 24, 2020, from <https://www.ncbi.nlm.nih.gov/pubmed/25810074>

³ (n.d.). Model-free Estimation of Recent Genetic Relatedness.. Retrieved March 24, 2020, from <https://www.ncbi.nlm.nih.gov/pubmed/26748516>

variant association is subsequently tested genome-wide. Available association tests include single variant tests as well as multiple 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 Multiple variant tests can be performed using the burden, SKAT,⁵ SKAT-O,⁶ fastSKAT,⁷ and SMMAT⁸ methods. Each of these methods can be run with user-defined variant groupings (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.

This tutorial will first guide you through the steps of performing a single variant association test using TOPMed study data and then move on to a multiple variant test. The overall steps of the single variant test are as follows:

1. Create your project
2. Bring TOPMed VCF data to your project
3. Upload kinship and PCA data
4. Bring the GENESIS workflows to your project
5. Convert VCF files to GDS file format
6. Upload harmonized phenotype or create harmonized phenotype in RStudio on platform
7. Fit Null Model
8. Run the single variant association test workflow

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/

⁴ (2018, August 13). Efficiently controlling for case-control imbalance and sample Retrieved March 24, 2020, from <https://www.nature.com/articles/s41588-018-0184-y>

⁵ (n.d.). Rare-Variant Association Testing for Sequencing Data with Retrieved March 24, 2020, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3135811/>

⁶ (n.d.). Optimal tests for rare variant effects in sequencing association Retrieved March 24, 2020, from <https://www.ncbi.nlm.nih.gov/pubmed/22699862>

⁷ (2018, June 22). FastSKAT: Sequence kernel association tests for very large Retrieved March 24, 2020, from <https://www.ncbi.nlm.nih.gov/pubmed/29932245>

⁸ (2019, January 10). Efficient Variant Set Mixed Model Association Tests for Retrieved March 24, 2020, from <https://www.ncbi.nlm.nih.gov/pubmed/30639324>

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:

- Create a project for the test GWAS
- 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
- Fit Null Model

Project Creation

The first step 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.

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.

Frequency	Percentage
Daily	75%
Weekly	15%
Monthly	10%

Controlled data: Yes

Users can find data from the hosted TOPMed studies in the Data Browser feature. This includes multi-sample VCF files and phenotype data on a per study and consent code basis, as well as single-sample VCF files and CRAM files. In the Data Browser feature, users can search over open metadata for the hosted TOPMed studies. However, users can only bring files to a project for analysis if they have dbGaP approval to work with the selected files.

For this tutorial, please select one or more TOPMed studies that you are approved to access. Choose “Explore Selected” and then search for the consent codes you are approved for and the multi-sample VCFs. The following example shows how to do this for the “HMB-IRB” consent group of the Jackson Heart Study. The instructions are written for this specific example, but please note that you will not be able to bring these data to a Project unless you are approved for this study and consent group. Therefore, please follow the flow for a study and consent group that you have approval for in dbGaP.

After selecting “Explore 1 selected,” the visual graph query builder pops up where you can search over open metadata for the TOPMed studies. Select to search over the File entity and then select the “Data Type” property within the File entity. Listed are Aligned Reads (CRAMs), Simple Germline Variation (single-sample VCF files), the Unharmonized Clinical data (clinical data from dbGaP), and Variant Call (the multi-sample VCFs). Filter by the Variant Calls and refresh in the lower left-hand corner.

Next, search for the consent group(s) by selecting the “Consent” property under the File entity. Choose your consent groups, filter by this property, and refresh results. For Jackson Heart, we will select the “HMB-IRB” consent group. This search identifies three files in the tar.gz format (same as dbGaP). Now select “Copy files to project.” This links these files to your project (as opposed to duplicating the files on storage infrastructure). Since the identified files are controlled access, they must be copied to a project marked as controlled (selected during project creation).

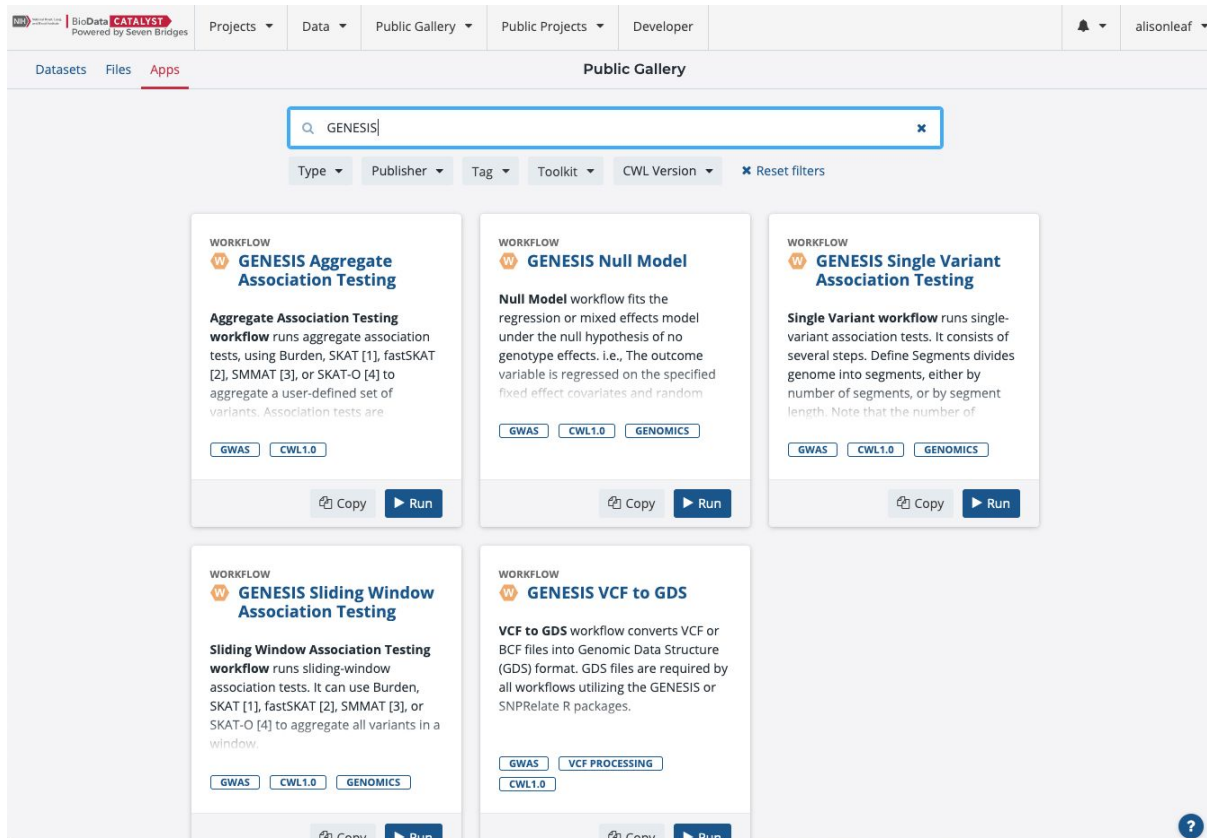
The screenshot shows the JHS interface with a query builder and a list of files. The query builder is on the left, showing a search for 'File' with 3 results. The 'Data Type' is set to 'Variant Call' and the 'Consent' is set to 'HMB-IRB'. The 'Filter by' section is empty. The 'File name' field is also empty. The 'Export' button is visible. The 'Details' and 'Analytics' buttons are also visible.

File	Details for phg000890.v1.TOPMed_WGS_JHS_v2.genotype-calls-vc...	Connections
phg000890.v1.TOPMed_WGS_JHS_v2.genotype-calls-vc...	JHS	Inbound:
phg001028.v1.TOPMed_WGS_JHS_v3.genotype-calls-vc...	Access level ⓘ Controlled	No inbound connections
phg000802.v1.TOPMed_WGS_JHS.genotype-calls-vc...	Consent ⓘ HMB-IRB	Outbound:
	Data Format ⓘ VCF	No outbound connections

Verify that the files are in your designated project by opening the “test GWAS” project and going to the “Files” tab.

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 pipelines to BioData Catalyst. These workflows were optimized for the cloud and can be found in the Public Gallery as shown below:















To find the pipelines click on **Public Gallery - Apps** and search for “GENESIS.” You will find the VCF to GDS workflow, the null model workflow, and 3 options for association testing (single variant, aggregate, sliding window). For the needed workflows, 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

Because the TOPMed Freeze5 multi-sample VCFs are in a tar file format, you will also need an App to decompress the file. Please also search for and copy the “SBG Decompressor” App to your project.

On the right side of the screen click on “copy” and add this tool to your test GWAS project. You should now see all four workflows in the Apps tab of your “test GWAS” project.

^ Name	Type	Actions
 COPY GENESIS Null Model **Null Model** workflow fits the regression or mixed effects model under the null hypothesis of ...	Workflow	 
 COPY GENESIS Single Variant Association Testing **Single Variant workflow** runs single-variant association tests. It con...	Workflow	 
 COPY GENESIS VCF to GDS **VCF to GDS** workflow converts VCF or BCF files into Genomic Data Structure (GDS) format. G...	Workflow	 
 COPY SBG Decompressor CWL1.0 The SBG Decompressor tool performs extraction of files from an input archive file. The s...	CommandLineTool	 

VCF to GDS Conversion


The next key step for running the GENESIS pipelines is to convert the VCF files into the GDS format. Since we are working with the hosted TOPMed Freeze5 VCF data in the tar file format, these tar bundles need to be decompressed first. To do this we will use a CWL app called **SBG Decompressor** which simply opens up the tar bundles so we can access the per chromosome multi-sample vcfs inside.

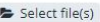
On the Apps tab, click the blue button to the right of the SBG Decompressor tool to bring up the draft task screen (shown below).

Task Inputs

Execution Settings

Inputs

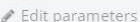
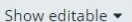
Batching ⓘ Off 

▼ Input archive file * ⓘ 

No files selected

This field is required and cannot be empty.

App Settings

Flatten outputs ⓘ

No value ▼

Outputs

Output files ⓘ

No value

Click on “Select Files” under the Inputs section of the draft Task setup. This will open up the file picker.

Files

Search

Type: 8 selected Sample ID: All Task ID: All Tags: All + Save selection

Copy

Name	Size
<input checked="" type="checkbox"/> JHS HMB-IRB phg001028.v1.TOPMed_WGS_JHS_v3.genotype-calls.vcf.WGS_markerset_grc38.c3.HMB-IRB.tar.gz	15.4 GiB
<input type="checkbox"/> JHS HMB-IRB phg000890.v1.TOPMed_WGS_JHS_v2.genotype-calls.vcf.WGS_markerset_grc37.c3.HMB-IRB.tar.gz	37.9 GiB
<input type="checkbox"/> JHS HMB-IRB phg000802.v1.TOPMed_WGS_JHS.genotype-calls.vcf.WGS_markerset_grc37.c3.HMB-IRB.tar.gz	23.6 GiB

As the example for this tutorial, we will select JHS accession number phg001028.v1 which is using human reference grc38. If you are working with another study for this tutorial, find a single tar bundle that uses reference 38. Click the save selection button which will send you back to the draft task screen. Next, start the task execution by clicking “Run” in the upper right hand corner. The completed task will look like this:

COMPLETED SBG Decompressor CWL1.0 run - 04-... [Get support](#) [View stats & logs](#) [Publish](#) [Edit and rerun](#)

Executed on Mar. 31, 2020 23:35 by dave

Spot Instances: **On** | Memorization: **Off** | Price: **\$0.08** | [Refund](#) | [View refunds](#) | Duration: **16 minutes**

App: SBG Decompressor CWL1.0 - Revision: 0

Inputs

Input archive file

JHS HMB-IRB phg001028.v1.TOPMed_WGS_JHS_v3....

App Settings

Flatten outputs **true**

Outputs

Output files

cg3_phs000964_TOPMed_WGS_JHS_freeze.5b...

cg3_phs000964_TOPMed_WGS_JHS_freeze.5b...

cg3_phs000964_TOPMed_WGS_JHS_freeze.5b...

cg3_phs000964_TOPMed_WGS_JHS_freeze.5b...

cg3_phs000964_TOPMed_WGS_JHS_freeze.5b...

...and 18 more items

As with all tasks that you run on the BDC platform, your output files will be available for you in the project files tab and available for you to use in the next step of the pipeline.

TOPMed variant data is in the VCF file format. The Freeze 5 data will have 1 VCF file per chromosome. 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. 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 uses 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:

DRAFT GENESIS VCF to GDS run [Get support] [Discard] [Run]

Last update by dave on Apr. 1, 2020 08:16
 App: GENESIS VCF to GDS - Revision: 0

Task Inputs | Execution Settings

Inputs

Batching ☐ Off

▼ **Variants Files** *

No files selected

This field is required and cannot be empty.

App Settings

▼ **vcf2gds** (#vcf2gds)

▼ **Format**

This input is set to null.

Number of CPUs

memory GB

Outputs

Unique variant ID corrected GDS files per chromosome

No value

Select 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:

QUEUED GENESIS VCF to GDS run [Get support] [View stats & logs] [Abort] [Edit and rerun]

Queued on Apr. 1, 2020 08:40 by dave

Spot Instances: **On** | Memoization: **Off** | Progress: The task has been submitted and is awaiting execution. | Duration: Less than a minute

App: GENESIS VCF to GDS - Revision: 0

Inputs

▼ **Variants Files**

cg3_phs000964_TOPMed_WGS_JHS_freeze.5b.chr9....
 cg3_phs000964_TOPMed_WGS_JHS_freeze.5b.chr8....
 cg3_phs000964_TOPMed_WGS_JHS_freeze.5b.chr7....
 cg3_phs000964_TOPMed_WGS_JHS_freeze.5b.chr6....
 cg3_phs000964_TOPMed_WGS_JHS_freeze.5b.chr5....
 ...and 17 more items

App Settings

▼ **vcf2gds** (#vcf2gds)

▼ **Format**

GDS File No value

Number of CPUs No value

memory GB No value

Outputs

Unique variant ID corrected GDS files per chromosome

No value

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. 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.

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-CPU” 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.

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 TOPMed studies or consent groups, we recommend first merging VCFs from the separate consent groups 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 a single VCF file per chromosome. Each VCF should include all samples in your analysis cohort.

Uploading additional files for the association test

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

- Principal components
- Kinship matrix
- Phenotype file

BioData Catalyst powered by Seven Bridges does not currently host the principal components files or kinship matrix files. Please upload them to your project using either the command line uploader (linux) or the GUI-based Desktop Uploader (Windows, Mac OS, Linux), both of which can be found under “Data Tools” in the Data tab of the top navigation bar. Please refer to the [Documentation pages on “Bring your own data”](#) to read additional details on how to use these features. In addition, for



Many researchers need to perform phenotype harmonization for 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 interest. We recommend using the platform implementation of RStudio for these steps (Jupyterlab notebooks are also available). The first section below outlines how to find the hosted raw phenotype files on the platform. The second section describes how to use RStudio.

BioData Catalyst hosts phenotype files for each TOPMed study and consent code. Phenotype data is hosted in the same format as it is found on dbGaP, in tar files that can be decompressed. In addition, the system hosts the raw phenotype data in the avro file format. The phenotype data can be found in the Data Browser feature. The TOPMed parent studies are listed on the first page of the Data Browser. Select your study of interest and then proceed to the visual interface. Search over the **File** entity, choose the **Data Type** property and select the “unharmonized clinical data.” In addition, be sure to search for the consent group(s) you have access to. After refreshing your results, you will find both the tar format files and the avro format files. Select **Copy files to project** to link to your “test GWAS” project.

RStudio on the platform

RStudio can be opened from within projects. 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. The default instance (aws c3.2xlarge) will be sufficient for this tutorial.

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 -AND- there are no running jobs. When the suspend time is reached, the instance and analysis are stopped. All files are saved that meet the criteria for automatic saving or have been manually selected to be saved as project files. All other files will be lost. By setting suspend time, you don’t have to worry about stopping the RStudio instance to prevent it from continually running in the background.

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

Create new analysis

✓Basic information

✓Compute requirements

Analysis name

My Phenotype Harmonization

Environment

JupyterLab

Web-based UI for Project Jupyter

RStudio BETA

IDE for R

Skip wizard

Previous

Next

Create new analysis

✓Basic information

✓Compute requirements

Select an Instance type with adequate CPU, memory and storage allocation for your analysis. This can be changed between analysis runs, but not while the analysis is running.

Instance type

c3.2xlarge (160GB SSD, 8vCPUs, 15GB RA...

Suspend time ⓘ On

30

Minutes

Price: \$0.44 per hour

Skip wizard

Previous

Start the analysis

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. 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. The sample.ids 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.

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. This RData object should take the form of the output from the pcair() function in the GENESIS package; specifically, as 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. This RData object should 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 specified file format.

Running an analysis on a subset of samples

Researchers may want to run an analysis with only a subset of the samples in their project. Rather than requiring subsetting of the GDS file, Phenotype file, PCA file, and Relatedness Matrix file, a subset of samples can be specified 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 analyzed.

Running an analysis on a subset of variants

Researchers may want 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 analyzed.

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

Both continuous (**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

⁹ (n.d.). GCTA: A Tool for Genome-wide Complex Trait Analysis. Retrieved March 24, 2020, from [https://www.cell.com/ajhg/pdf/S0002-9297\(10\)00598-7.pdf](https://www.cell.com/ajhg/pdf/S0002-9297(10)00598-7.pdf)

covariates using the PCA file input, and genetic relatedness can be accounted for as a random effect using the **Relatedness matrix file** input. See section on “Ancestry and Relatedness” above.

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

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. See section on “Ancestry and Relatedness.” 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 **relatedness_matrix_file** via the R Matrix package, and this can substantially reduce memory usage and CPU time.¹⁰

Running Single Variant Association Tests

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” and 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^{11, 12} can be used by specifying **test_type = ‘score.spa’**; this is generally recommended. 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 imputed to the mean alternate allele count 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).

¹⁰ (2016, March 24). Control for Population Structure and Relatedness for Binary Retrieved March 24, 2020, from <https://www.ncbi.nlm.nih.gov/pubmed/27018471>

¹¹ (2018, August 13). Efficiently controlling for case-control imbalance and sample Retrieved March 24, 2020, from <https://www.nature.com/articles/s41588-018-0184-y>

¹² (n.d.). A Fast and Accurate Algorithm to Test for Binary Phenotypes Retrieved March 24, 2020, from <https://www.ncbi.nlm.nih.gov/pubmed/28602423>

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

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 the null model. If a run fails with **error 137**, and with the message “killed” displayed, the most likely cause is a lack of memory.

Your completed task for the GENESIS Single Variant Association test should be similar to the image below:

COMPLETED **GENESIS Single Variant Association ...**

[Get support](#) [View stats & logs](#) [Publish](#) [Edit and rerun](#)

Executed on Apr. 6, 2020 10:19 by dave

Spot Instances: **On** | Memoization: **Off** | Price: **\$16.58** | [Refund](#) | [View refunds](#) | Duration: **3 hours, 8 minutes**

App: GENESIS Single Variant Association Testing - Revision: 0

Inputs

GDS files

cg3_phs000964_TOPMed_WGS_JHS_freeze.5b...
cg3_phs000964_TOPMed_WGS_JHS_freeze.5b...
cg3_phs000964_TOPMed_WGS_JHS_freeze.5b...
cg3_phs000964_TOPMed_WGS_JHS_freeze.5b...
cg3_phs000964_TOPMed_WGS_JHS_freeze.5b...
...and 17 more items

Known hits file

No files selected

Null model file

null_model_invnorm.RData

Phenotype file

phenotypes.RData

Variant Include Files

No files selected

App Settings

Show non-default

define_segments.R (#define_segments_r)

Genome build hg38

assoc_single.R (#assoc_single_r)

Output prefix single_variant

Test type score

Outputs

Association testing results

single_variant_chr1.RData
single_variant_chr2.RData
single_variant_chr3.RData
single_variant_chr4.RData
single_variant_chr5.RData
...and 17 more items

QQ and Manhattan plots

single_variant_manh.png
 single_variant_qq.png

20

Multiple-variant Association Tests

Multiple-variant association tests are often used when analyzing variants from WGS or WES studies since they jointly test a group of variants and can increase power when using rare variants. These tests are also referred to as aggregate tests and each grouped variant set can be referred to as an aggregation unit. This section of the tutorial describes multiple-variant association tests where the user specifies sets of variants to be jointly tested using genomic annotations. BioData Catalyst powered by Seven Bridges also has the **GENESIS Sliding Window Association Testing** app which aggregates variants by genomic positions. The sliding window test will not be covered in this tutorial and more information can be found in the [app description](#).

Preparation of Files for Multiple Variant Associations using GENESIS

When performing multiple-variant association where the variant sets or groups are specified by the user, the user must provide a variant grouping file and, optionally, a variant weight file. The format of these files is described below. Both these files can be generated using the Annotation Explorer feature.

User provided variant groupings for aggregate tests

When performing multiple variant aggregate association tests, you can specify the groups and the variants that should be grouped under them using a variant grouping file. This is an RData file with a 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 (e.g. 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** feature on BioData Catalyst powered by Seven Bridges, which allows the user to filter and group variants based on 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.

User provided variant weights for aggregate tests

When performing multiple variant aggregate association tests, a researcher may want to weigh each variant included in the aggregation unit. You can 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 previous section), "Weight user" may also refer to a column in that file and a separate variant weight file is not needed.

Annotation Explorer

The Annotation Explorer is an application developed by Seven Bridges in collaboration with the TOPMed Data Coordinating Center which enables users to interactively explore, query, and study characteristics of an inventory of annotations for the variants called in TOPMed studies. This application can be used pre-association testing to interactively explore aggregation and filtering strategies for variants based on annotations and generate input files for multiple-variant association testing. It can also be used post-association testing to explore annotations associated with a set of variants, like variant sets found significant during association testing.

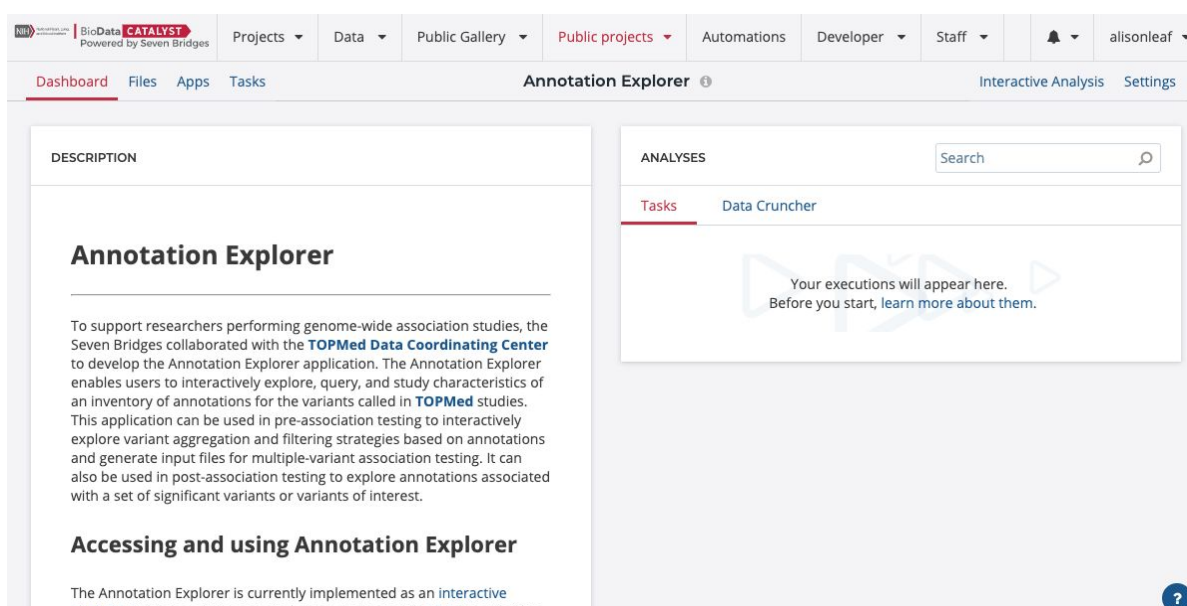
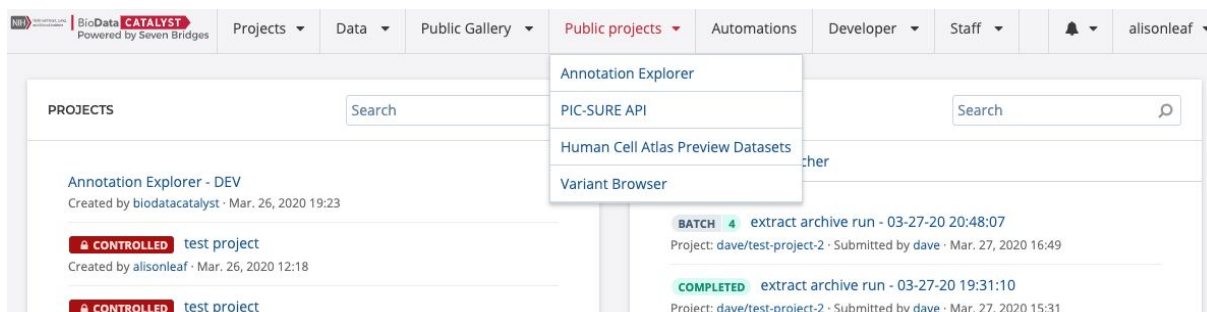
The Annotation Explorer currently hosts a subset of genomic annotations obtained using Whole Genome Sequence Annotator¹³ software for TOPMed variants. Currently, annotations for TOPMed Freeze5 variants and TOPMed Freeze8 variants are integrated with the Annotation Explorer. Researchers who are approved to access one or more of the TOPMed Freeze5 studies will have the ability to access the Freeze5 data on Annotation Explorer. Access to annotations for Freeze8 data will be made available after the Freeze8 data is released on dbGaP.

For the purposes of this tutorial, we will use the Annotation Explorer to generate a variant grouping file by applying simple variant filtering and aggregation strategy using annotations from TOPMed Freeze5.

Launch Annotation Explorer

The Annotation Explorer is currently implemented as an interactive analysis which has all the code to launch an interactive R shiny application. To launch the feature, choose **Public Projects** from the top navigation menu and select the **Annotation Explorer** option.

¹³ (n.d.). WGS: an annotation pipeline for human genome sequencing Retrieved March 24, 2020, from <https://img.bmj.com/content/53/2/111>



Select **Interactive Analysis** in the upper-right hand corner of the dashboard and then the **Data Cruncher** feature. Select the analysis for the Annotation Explorer on the next page and that will bring you to the page below:

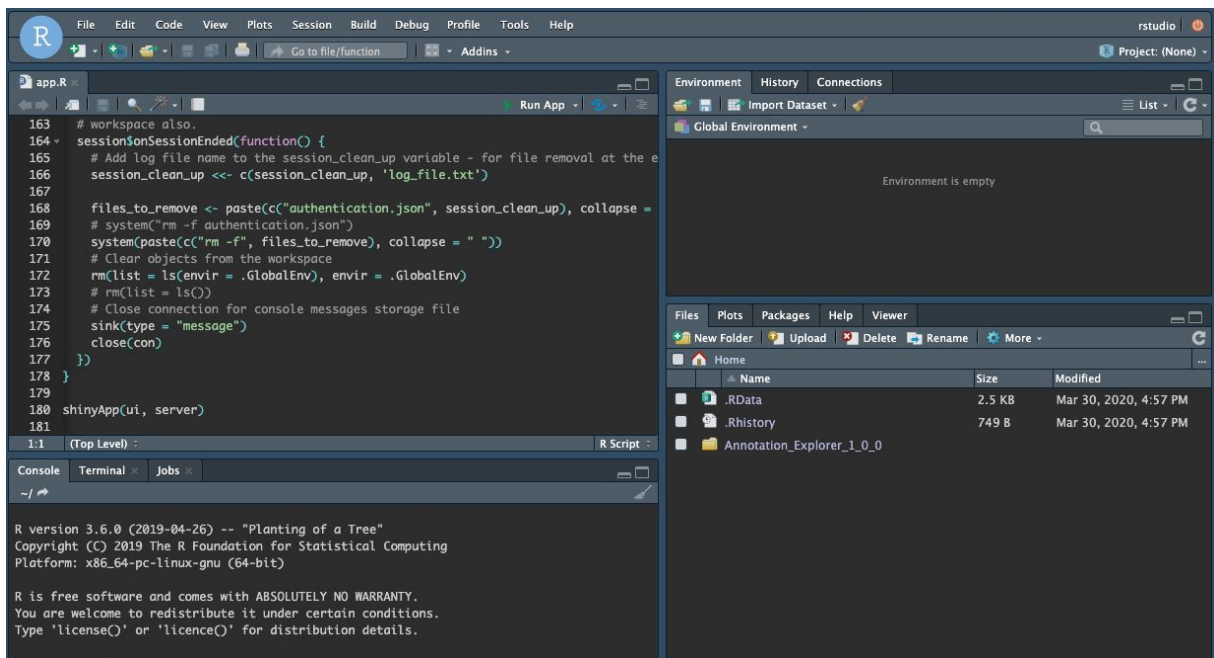
Copy this analysis to your “test gwas” project using the **Copy** button in the upper right hand corner and then go back to your “test gwas” project. Select **Interactive Analysis** and open **Data Cruncher** (middle tile).

Analysis name	Created by	Environment	Created on	Status	Action
Copy of Annotation Explorer	dave	RStudio (Seven Bridges default)	Apr. 2, 2020 14:49	SAVED	

Start the Annotation Explorer analysis by pressing the blue “play” button. Once the RStudio instance has been initialized, which takes a few minutes, click the “open editor” button which has replaced the play button in the image below. For more information about RStudio cruncher please see the knowledge center RStudio documentation at <https://sb-biodatacatalyst.readme.io/docs/rstudio-quick-reference>.

Analysis name	Created by	Environment	Created on	Status	Actions
Copy of Annotation Explorer	dave	RStudio (Seven Bridges default)	Apr. 2, 2020 14:49	RUNNING	

The RStudio interface will open up (shown below). Click “Run App” at the top right hand corner of the app.R script to launch the interactive R shiny app in a new window.



The Annotation Explorer will pop up and look like this:

Instructions

Authentication & Dataset Selection

[Reset Session](#)

[Close Window / Stop App](#)

© 2020 Seven Bridges
Terms Privacy Copyright

TOPMed Annotation Explorer

Powered by Seven Bridges

The TOPMed annotation explorer provides you with the functionality to interactively explore an inventory of annotations from the annotation databases hosted by Google BigQuery.

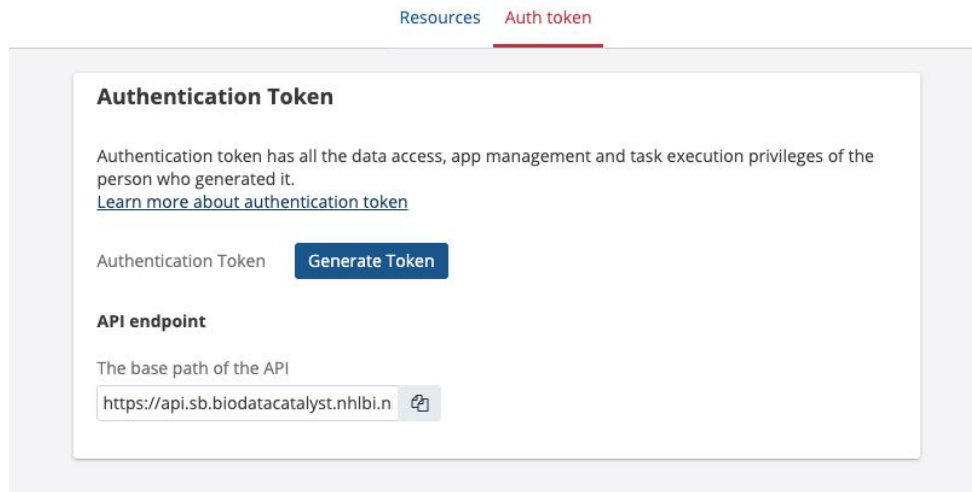
To use it, one must first provide a JSON authentication file to gain access to BigQuery. This can be done from the *Authentication and Dataset Selection* tab. If the query results are to be uploaded to a project on the Seven Bridges powered NHLBI BioData Catalyst platform — then a valid API token is also needed.

The *Dataset Info* tab lets you preview the chosen dataset and explore different annotation fields by looking at their metadata or graphical summaries.

There are two options for querying the data — one is by writing standard SQL in the *SQL Query* tab, or if you prefer composing the query interactively through GUI — use the *Query Composer*.

Authentication and Data selection

After starting the Annotation Explorer you will need to enter your private authentication token. Follow the provided link to the platform page where you can generate your API token (see below) and select **Generate Token**. Copy the token to your clipboard and go back to the Annotation Explorer.



The screenshot shows the 'Auth token' tab selected in the top navigation bar, with 'Resources' also visible. The main content area is titled 'Authentication Token' and contains the following text: 'Authentication token has all the data access, app management and task execution privileges of the person who generated it.' Below this is a link: '[Learn more about authentication token](#)'. There is a label 'Authentication Token' followed by a blue button labeled 'Generate Token'. Below this is the section 'API endpoint' with the text 'The base path of the API' and a text input field containing 'https://api.sb.biodatacatalyst.nhlbi.n'. To the right of the input field is a small icon of a document with a plus sign.

Paste your API token into the space provided and **Authenticate**. A message in a green bar will appear to let you know if the authentication was successful.

Next, under the “Project and Data Selection” tab at the top, select your Destination project as “test gwas”, this is the project where all the output files will be stored. Under “Dataset and Table selection,” choose “freeze-5” and then click “**Send request to BigQuery.**”

Authentication

Project and Dataset Selection

NHLBI BioData Catalyst platform project selection

Choose the adequate NHLBI BioData Catalyst platform project, where your results will be saved.

Destination project

dave/test-gwas

Dataset and Table selection

Choose the adequate BigQuery dataset and table

Select Freeze you want to analyze

freeze_5

Send request to BigQuery

This should take about 20 seconds and will result in a preview (1,000 rows) of the unfiltered annotations. The Dataset Info page (see image below) is a great place to learn about the annotations available for your research in real time. The Metadata Explorer tab gives text descriptions of the hundreds of annotations available. The Data Explorer gives distribution and summary statistics for the 100's of annotation fields.

Instructions

Authentication & Dataset Selection

Dataset Info

Query Composer

Query Billing

Reset Filters

Reset Session

Close Window / Stop App

© 2020 Seven Bridges
Terms Privacy Copyright

Table preview

Metadata explorer

Data explorer

Dataset Info

Preview for dataset: freeze_5

Show 10 entries

Search:

CHROM	POS	REF	ALT	VEP_ensembl_Amino_Acid_Change	VEP_ensembl_Gene_ID	ALSPAC_AC	ALSPAC_
X	119851748	CT	C	.	ENSG000000125351	2528	0.8750432675
X	136033510	GA	G	.	ENSG000000198689		
14	90899302	GA	G	.	ENSG000000100784		
2	37241551	AT	A	.	ENSG000000003509		
1	222650745	CAAAAGT	C	.	ENSG000000154305		
16	20813344	CT	C	.	ENSG000000005189		
5	138171173	GA	G	.	ENSG000000112983		
5	178270372	CGA	C	.	ENSG000000050767		
5	79990547	AG	A	.	ENSG000000177034		
4	65420395	GA	G	.	ENSG000000145242		

Showing 1 to 10 of 1,000 entries

Previous

1

2

3

4

5

...

100

Next

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.

Filter manually

Select filtering query logic

AND

Annotation fields

CHROM POS REF ALT VEP_ensembl_Gene_ID
VEP_ensembl_Consequence CADD_phred

Chromosomes

chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9
chr10 chr11 chr12 chr13 chr14 chr15 chr16
chr17 chr18 chr19 chr20 chr21 chr22

Annotation Field

CADD_phred

Annotation Type

Quantitative

Filter

>

Value

20

Remove filter

Annotation Field

VEP_ensembl_Consequence

Annotation Type

Qualitative

Filter

Contains

Value

missense

Remove filter

Add filter

Filter

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.

Filter annotation fields and chromosomes

2,231,478

Number of variants

Save filter results to a file
+

Show 5 entries
Search:

Table preview

CHROM	POS	REF	ALT	VEP_ensembl_Gene_ID	VEP_ensembl_Consequence
1	169310010	C	T	ENSG00000143156	missense_variant,NMD_transcript_variant
1	201989474	C	G	ENSG00000176393	missense_variant
1	202301229	A	G	ENSG00000133067	missense_variant
1	2785854	G	T	ENSG00000215912	missense_variant
1	246621546	C	T	ENSG00000162852	missense_variant

Showing 1 to 5 of 1,000 entries

Previous
1
2
3
4
5
...
200
Next

Filter using file

Upload filter file

Browse... No file selected

Filter manually

Select filtering query logic

AND

Annotation fields

CHROM POS REF ALT VEP_ensembl_Gene_ID
VEP_ensembl_Consequence CADD_phred

Chromosomes

chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9
chr10 chr11 chr12 chr13 chr14 chr15 chr16
chr17 chr18 chr19 chr20 chr21 chr22

Annotation Field

CADD_phred

Annotation Type

Quantitative

Filter

>

Value

20

Remove filter

Annotation Field

VEP_ensembl_Consequence

Annotation Type

Qualitative

Filter

Contains

Value

missense

Remove filter

Add filter

Filter

Now navigate 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.

Aggregate manually

Choose annotation field for aggregation:

VEP_ensembl_Gene_ID ▼

Aggregate

Finally, go to the **Export results** tab and 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.

Save variant grouping file to project

Output file name

variant_grouping_file_test

Select folder where you want to save files

root_folder ▼

Save to project

When you are done using the Annotation Explorer, please make sure to close out by going to “Close Window/Stop App” on the side navigation bar.

The screenshot shows the Annotation Explorer interface. A confirmation dialog box is centered on the screen, asking "Are you sure you want to quit?" with "Yes" and "No" buttons. The background interface includes a sidebar on the left with navigation options: "Instructions", "Authentication & Dataset Selection", "Dataset Info", "Query Composer", "Query Billing", "Reset Filters", "Reset Session", and "Close Window / Stop App". The main area displays the "Filtering" tab, showing a table of variant data with columns: CHROM, POS, REF, ALT, VEP_ensembl_Amino_Acid_Change, and VEP_ensembl_Gene_ID. The table shows 2,064 entries. The "Export results" tab is also visible in the top navigation bar.

Running Multiple-variant association test

For users who would like to perform a multiple-variant association test by specifying aggregated sets of variants, please use the “GENESIS Aggregate Association Testing” workflow. This workflow runs multiple variant tests, also referred to as aggregate association tests, using Burden, SKAT,¹⁴ SKAT-O,¹⁵ fastSKAT,¹⁶ or SMMAT¹⁷ on each aggregated set of variants specified through the user-provided variant grouping file.

The aggregate testing workflow consists of several steps. The “**define_segments.R**” tool 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 done 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 null model workflow outputs, variant grouping files (generated using the Annotation Explorer), and the genotype data to perform the specified test for each aggregation unit. Once you have these files in your project, you can set up a draft task and run the multiple variant association test.

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 imputed to the mean alternate allele count 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.

¹⁴ (n.d.). Rare-Variant Association Testing for Sequencing Data with Retrieved March 24, 2020, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3135811/>

¹⁵ (n.d.). Optimal tests for rare variant effects in sequencing association Retrieved March 24, 2020, from <https://www.ncbi.nlm.nih.gov/pubmed/22699862>

¹⁶ (2018, June 22). FastSKAT: Sequence kernel association tests for very large Retrieved March 24, 2020, from <https://www.ncbi.nlm.nih.gov/pubmed/29932245>

¹⁷ (2019, January 10). Efficient Variant Set Mixed Model Association Tests for Retrieved March 24, 2020, from <https://www.ncbi.nlm.nih.gov/pubmed/30639324>

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.

The completed aggregate association test will look like the below image. Run time will increase with the number of subjects in your study.

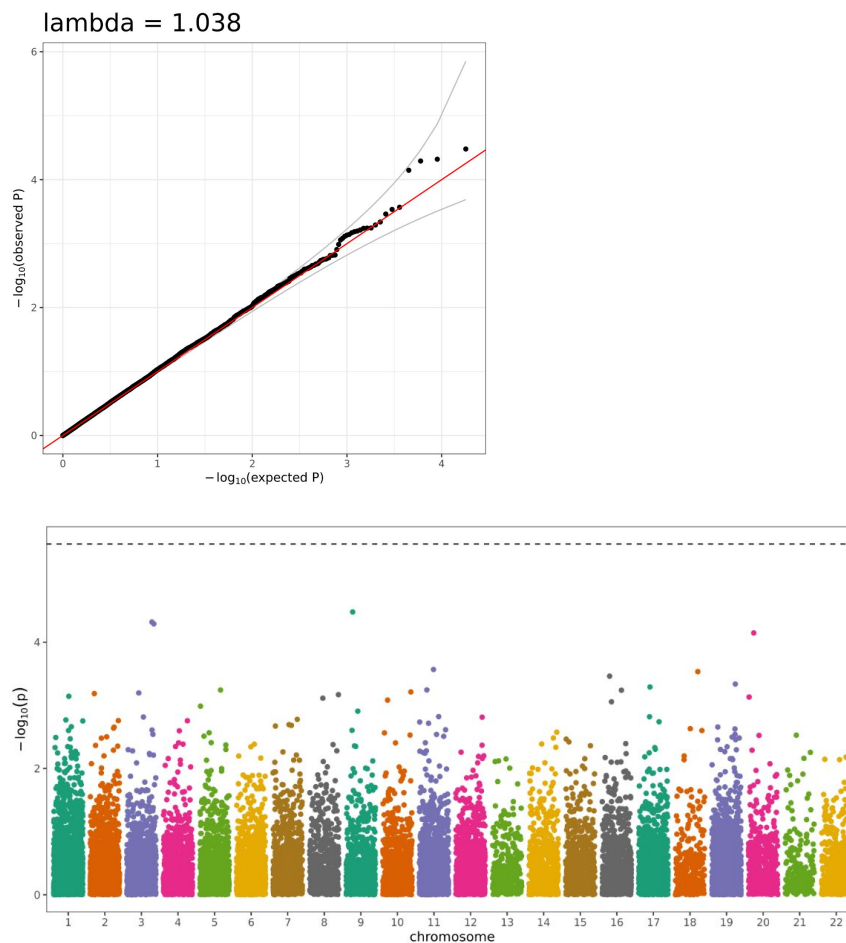
COMPLETED
GENESIS Aggregate Association Test...
Get support
View stats & logs
Publish
Edit and rerun

Executed on Apr. 8, 2020 11:10 by dave
 Spot Instances: **On** Memoization: **Off** Price: **\$3.92** Refund View refunds Duration: **1 hour, 13 minutes**
 App: GENESIS Aggregate Association Testing - Revision: 0

Inputs	App Settings	Outputs
GDS files cg3_phs000964_TOPMed_WGS_JHS_freeze.5b... cg3_phs000964_TOPMed_WGS_JHS_freeze.5b... cg3_phs000964_TOPMed_WGS_JHS_freeze.5b... cg3_phs000964_TOPMed_WGS_JHS_freeze.5b... cg3_phs000964_TOPMed_WGS_JHS_freeze.5b... ...and 17 more items	define_segments.R (#define_segments_r) Genome build hg38 assoc_aggregate (#assoc_aggregate) Aggregate type (Aggregate type) allele Output prefix aggregate Test smmat aggregate_list (#aggregate_list) Aggregate type (Aggregate type) allele	Assoc combined aggregate_chr1.RData aggregate_chr2.RData aggregate_chr3.RData aggregate_chr4.RData aggregate_chr5.RData ...and 17 more items Assoc plots aggregate_manh.png aggregate_qq.png
Known hits file Null model file null_model_invnorm.RData Phenotype file phenotypes.RData Variant Include Files Variant Weight file No files selected Variant group files variant_grouping_file_chr21.RData variant_grouping_file_chr11.RData variant_grouping_file_chr16.RData variant_grouping_file_chr1.RData variant_grouping_file_chr7.RData ...and 17 more items		

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 run. The “png” files can be viewed natively on the platform using your web browser. As you iterate over different association test experiments, these plots will remain saved in your project for future reference.





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.