Guide to the Genetic Association Interactive Tool (GAIT)

GAIT is a powerful tool that allows you to design custom analyses of genetic associations for the variant or gene of your choice. It offers the ability to filter samples by multiple criteria and to select specific covariates and other parameters. GAIT accesses individual-level data to compute p-values and odds ratios or effect sizes for these associations; to protect patient confidentiality, results are only displayed for sample sets that consist of more than 100 individuals.

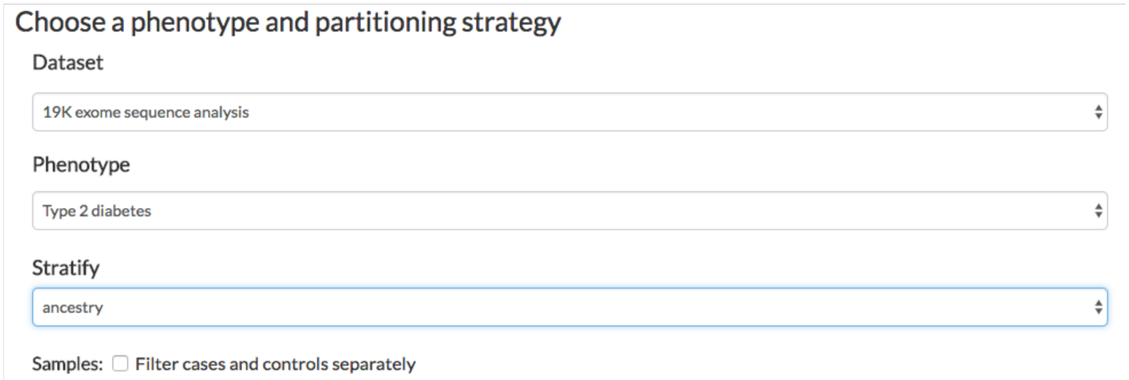
GAIT is available for two types of analysis. On Variant pages of the Type 2 Diabetes Knowledge Portal (T2DKP), the Cerebrovascular Disease Knowledge Portal (CDKP), and Cardiovascular Disease Knowledge Portal (CVDKP), GAIT allows you to compute associations for a single variant. On the “High impact variants” tab of Gene pages in the T2DKP and CVDKP, GAIT allows you to assess the disease burden for a gene using a custom set of variants.

GAIT is an exploratory tool. It is intended to produce results that are broadly concordant with those from an expert analysis, but results produced with GAIT should not be considered definitive. Rather, they may suggest hypotheses and directions for further investigation. Additionally, results may change over time, as the software and the data are under development. We are happy to provide help in evaluating the results from this tool; please contact us at the T2DKP helpdesk, CDKP helpdesk, or CVDKP helpdesk.

GAIT on Variant Pages

GAIT is accessible in the “Genetic Association Interactive Tool” section of Variant pages. It analyzes associations for the variant featured on that page.

Choose initial criteria



Choose a phenotype and partitioning strategy

Dataset: 19K exome sequence analysis

Phenotype: Type 2 diabetes

Stratify: ancestry

Samples: □ Filter cases and controls separately

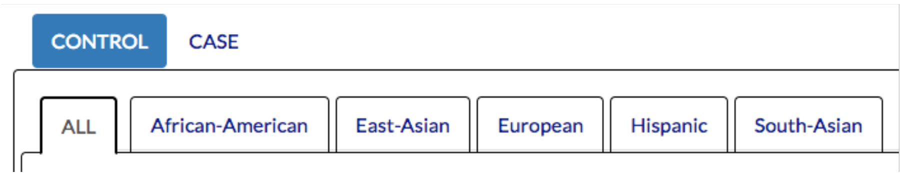
- First, choose a dataset for analysis. If multiple datasets are available, they are listed in the “Dataset” pull-down menu.
- Next, choose a phenotype to compute associations with that phenotype.

- With the “Stratify” pull-down menu, you may choose to stratify by ancestry so that results are generated separately for each ancestry group. This menu only appears if datasets with samples derived from multiple ancestries are available.
- You may choose to filter cases and controls separately.

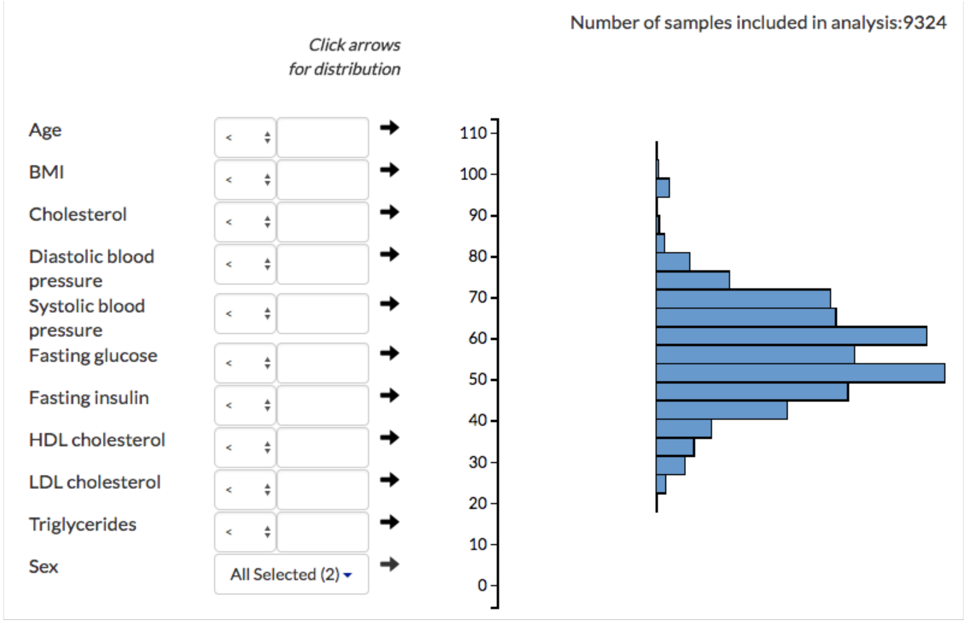
Step 1: Select a subset of samples based on phenotypic criteria

This step allows you to filter samples such that a custom subset is used for association analysis. Filtering the samples before performing association analysis may allow you to see effects that were not detectable in the larger sample set.

If you chose to stratify by case/control status or by ancestry, you may set separate filters for each group on the individual tabs.



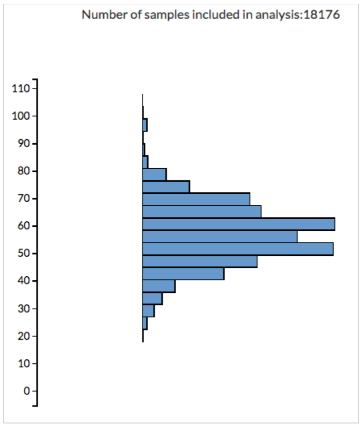
The section below the tabs displays other phenotypes that have been measured for your chosen sample set, with their ranges. The available phenotypes differ for each dataset-phenotype combination chosen in the initial steps.


Number of samples included in analysis:9324

Phenotypes available for filtering when using GAIT to perform association analysis for type 2 diabetes in variants from the 19K exome sequence analysis dataset. Different sets of phenotypes may be available for other datasets.

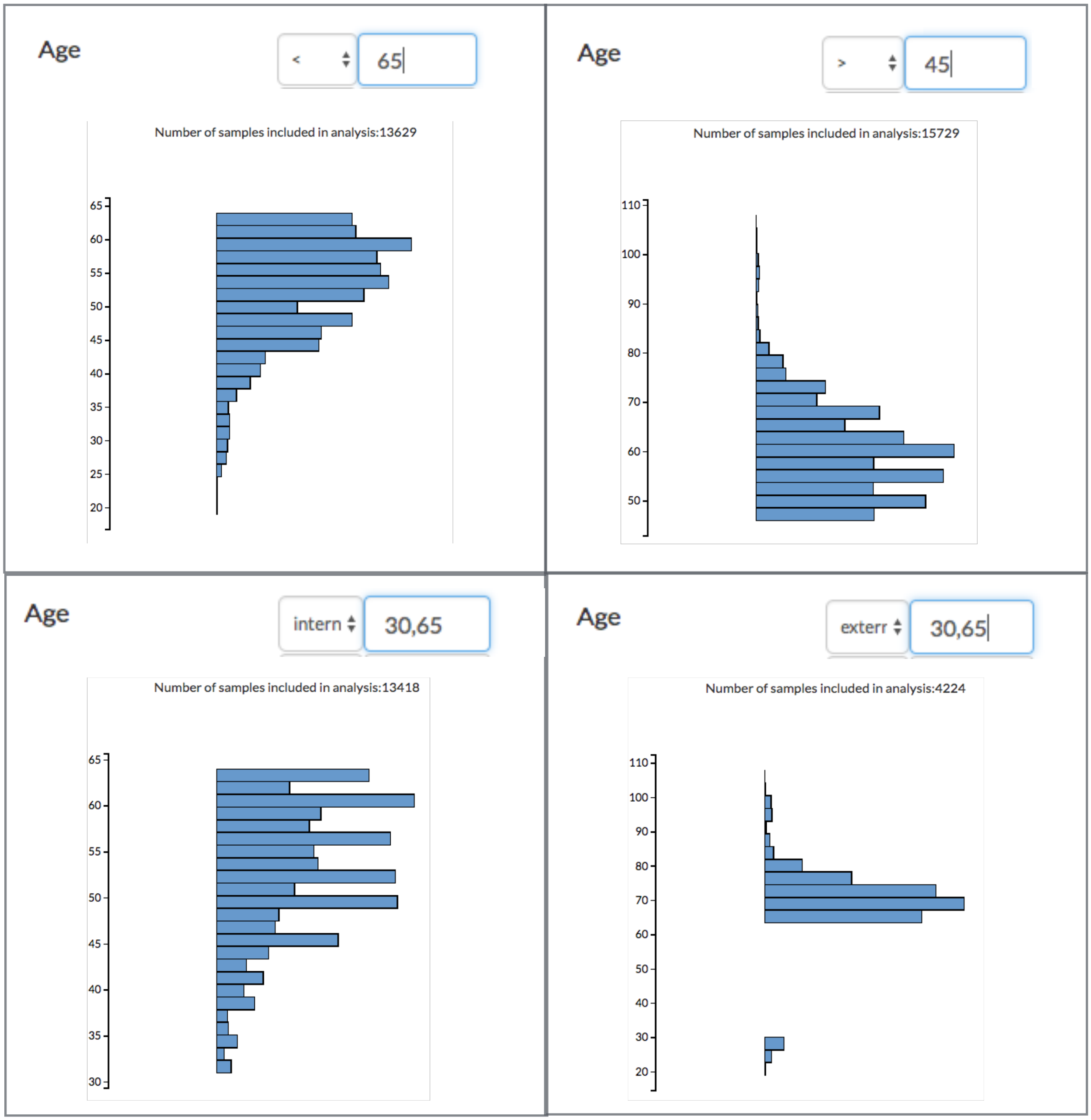
Click on an arrow to generate a bar chart showing the values present in the sample set for that phenotype. The total number of samples is shown above the bar chart.

To filter samples, enter a number or numbers in the box and use the pull-down menu to select samples greater than or less than the entered value, or an internal or external range of values, for example:



Number of samples included in analysis:18176

Age, unfiltered



Age

< 65

Number of samples included in analysis:13629

Age

> 45

Number of samples included in analysis:15729

Age

intern 30,65

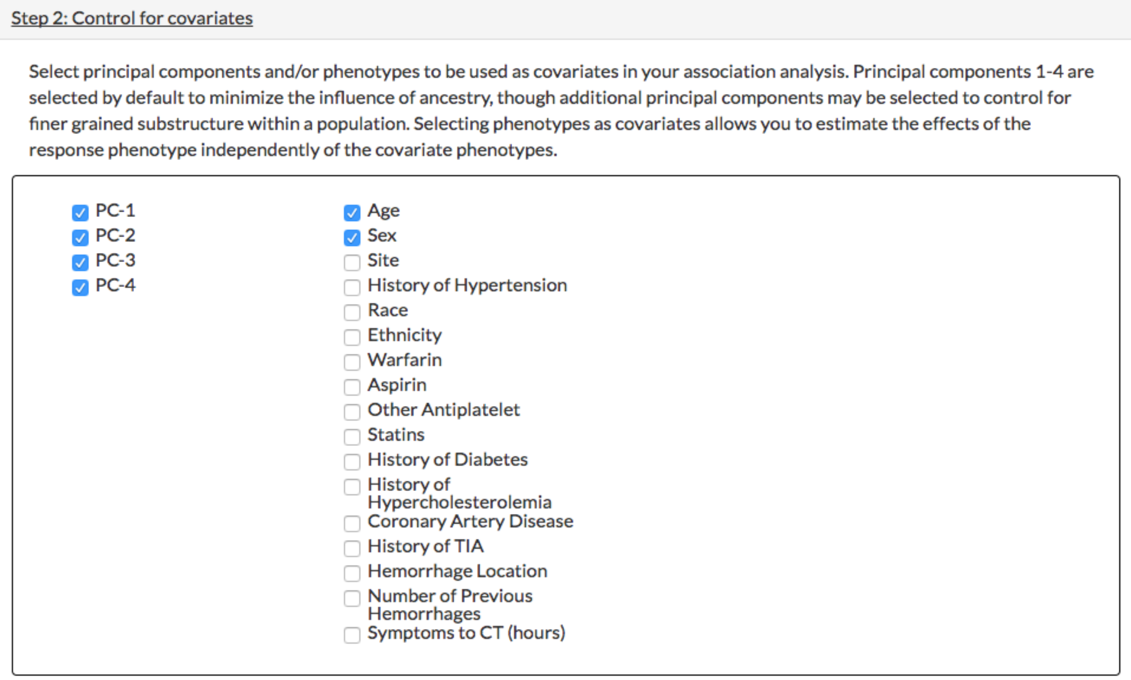
Number of samples included in analysis:13418

Age

exterr 30,65

Number of samples included in analysis:4224

Step 2: Choose covariates in order to control for population structure or test whether your chosen phenotype is dependent on other phenotypes.



Step 2: Control for covariates

Select principal components and/or phenotypes to be used as covariates in your association analysis. Principal components 1-4 are selected by default to minimize the influence of ancestry, though additional principal components may be selected to control for finer grained substructure within a population. Selecting phenotypes as covariates allows you to estimate the effects of the response phenotype independently of the covariate phenotypes.

- ☑ PC-1
- ☑ PC-2
- ☑ PC-3
- ☑ PC-4

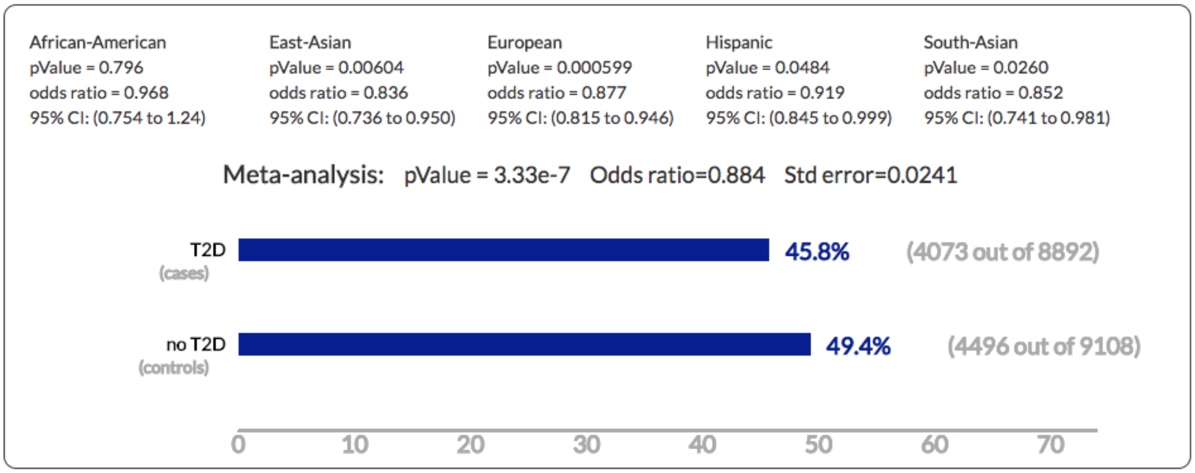
- ☑ Age
- ☑ Sex
- □ Site
- □ History of Hypertension
- □ Race
- □ Ethnicity
- □ Warfarin
- □ Aspirin
- □ Other Antiplatelet
- □ Statins
- □ History of Diabetes
- □ History of Hypercholesterolemia
- □ Coronary Artery Disease
- □ History of TIA
- □ Hemorrhage Location
- □ Number of Previous Hemorrhages
- □ Symptoms to CT (hours)

In this step, Principal Components 1-4 are checked by default to control for the relatedness within ancestral groups. Additional principal components are available for some datasets, and these may be selected to control for sub-groups within ancestries.

You may also select phenotypes to be used as covariates. If you chose to stratify by ancestry, choosing covariates on the "All" tab sets them for every ancestry. You may select different covariates for each ancestry on the respective ancestry tabs.

After setting parameters in steps 1 and 2, compute associations by clicking the "Launch analysis" button.

Results are shown in different formats, depending on your selections in the previous steps.



African-American	East-Asian	European	Hispanic	South-Asian
pValue = 0.796	pValue = 0.00604	pValue = 0.000599	pValue = 0.0484	pValue = 0.0260
odds ratio = 0.968	odds ratio = 0.836	odds ratio = 0.877	odds ratio = 0.919	odds ratio = 0.852
95% CI: (0.754 to 1.24)	95% CI: (0.736 to 0.950)	95% CI: (0.815 to 0.946)	95% CI: (0.845 to 0.999)	95% CI: (0.741 to 0.981)

Meta-analysis: pValue = 3.33e-7 Odds ratio=0.884 Std error=0.0241

Example results from a stratified analysis for association of a variant with T2D

- p-value is shown for each association
- odds ratio is shown for dichotomous (binary) traits
- beta (effect size) is shown for continuous traits
- confidence interval is shown for odds ratios and effect sizes
- if a dichotomous trait was selected initially, a bar chart is displayed showing the occurrence of the variant in cases and controls
- if stratification was selected, p-value, odds ratio or effect size, and confidence interval are shown for each ancestry
- if stratification was selected, a meta-analysis across all ancestries is performed, and p-value, odds ratio or effect size, and confidence interval are shown for the meta-analysis.

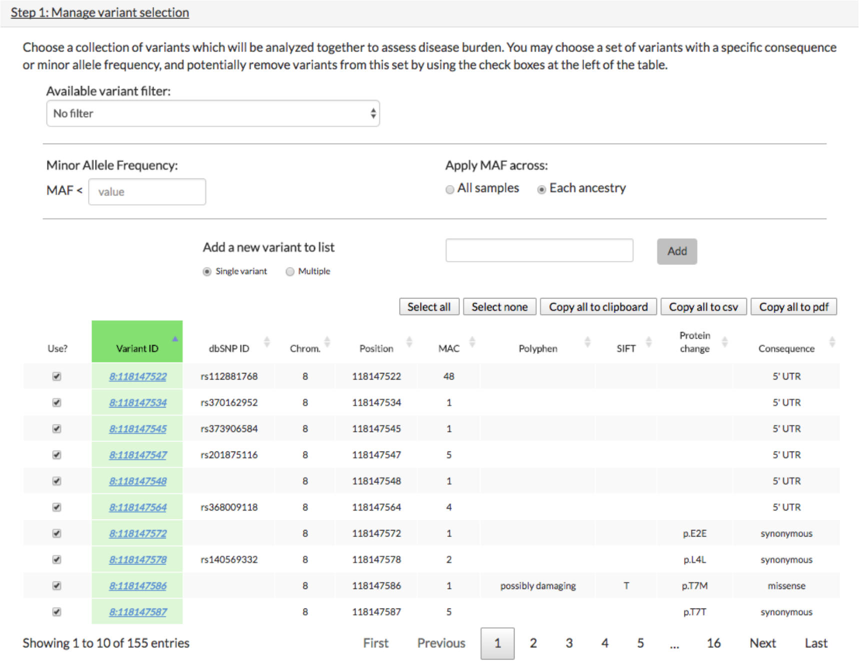
See the “Interpreting the results of GAIT analysis” section below for more information on these parameters.

GAIT on Gene Pages

GAIT is accessible in the “Run a custom burden test” section on the “High-impact variants” tab of Gene pages (currently, available in the T2DKP and CVDKP). It analyzes associations for sets of variants in or near the gene featured on that page. Rather than computing associations for one variant, the interactive burden test considers the whole gene as the unit of inquiry, and computes the aggregate disease or phenotype burden for the gene.

The interface for the GAIT-powered interactive burden test is identical to that of GAIT on Variant Pages (see above) except for an additional step that allows you to select a custom set of variants for analysis.

Step 1: Manage variant selection



Step 1: Manage variant selection

Choose a collection of variants which will be analyzed together to assess disease burden. You may choose a set of variants with a specific consequence or minor allele frequency, and potentially remove variants from this set by using the check boxes at the left of the table.

Available variant filter: No filter

Minor Allele Frequency: MAF < value

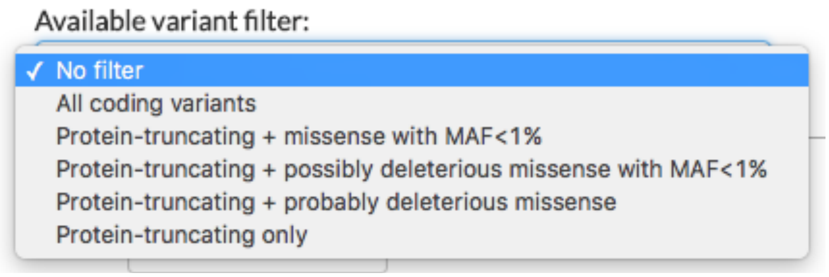
Apply MAF across: ○ All samples ● Each ancestry

Add a new variant to list ● Single variant ○ Multiple

Use?	Variant ID	dbSNP ID	Chrom.	Position	MAC	Polyphen	SIFT	Protein change	Consequence
☑	8:118147522	rs112881768	8	118147522	48				5' UTR
☑	8:118147534	rs370162952	8	118147534	1				5' UTR
☑	8:118147545	rs373906584	8	118147545	1				5' UTR
☑	8:118147547	rs201875116	8	118147547	5				5' UTR
☑	8:118147548		8	118147548	1				5' UTR
☑	8:118147564	rs368009118	8	118147564	4				5' UTR
☑	8:118147572		8	118147572	1			p.E2E	synonymous
☑	8:118147578	rs140569332	8	118147578	2			p.L4L	synonymous
☑	8:118147586		8	118147586	1	possibly damaging	T	p.T7M	missense
☑	8:118147587		8	118147587	5			p.T7T	synonymous

Showing 1 to 10 of 155 entries

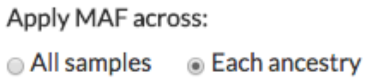
The variant filter scroll box enables you to select sets of variants. The default list ("No filter") includes all variants within the gene, including 100 kb of upstream and downstream flanking sequences.. These filters were documented in Fuchsberger, Flannick,Teslovich, Mahajan, Agarwala, Gaulton *et al.* 2016. The genetic architecture of type 2 diabetes. Nature **536**(7614): 41-7. The set of variants includes only those that are included in the dataset that you selected.


Available variant filter:
✓ No filter
All coding variants
Protein-truncating + missense with MAF<1%
Protein-truncating + possibly deleterious missense with MAF<1%
Protein-truncating + probably deleterious missense
Protein-truncating only

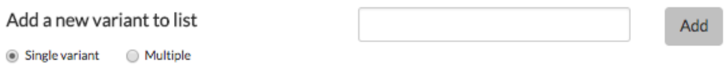
- "All coding variants" selects variants within the coding sequence.
- "Protein-truncating + missense with MAF<1%" selects variants that are protein-truncating OR are missense AND have minor allele frequency of less than 1%, since variants of deleterious effect are likely to occur at lower frequency.
- "Protein-truncating + possibly deleterious missense with MAF<1%" selects variants that are protein-truncating OR are missense AND are predicted to be deleterious by at least one of 5 algorithms (LRT, MutationTaster, PolyPhen2-HumDiv, PolyPhen2-HumVar, or SIFT) AND have minor allele frequency of less than 1%.
- "Protein-truncating + probably deleterious missense" selects variants that are protein-truncating OR are missense AND are predicted to be deleterious by all 5 algorithms (LRT, MutationTaster, PolyPhen2-HumDiv, PolyPhen2-HumVar, and SIFT).
- "Protein-truncating only" selects variants that would cause a truncated protein to be generated, either by creating a premature stop codon or by causing a frameshift.

The **Minor allele frequency (MAF)** text entry box allows you to select variants whose allele frequency is below an entered value, expressed as a fraction of chromosomes that carry the allele in the sampled population. For example, an allele with a frequency of 0.2 is present on 20% of the chromosomes in the population.

You may choose to apply the MAF cutoff across all samples or per ancestry. MAF may differ substantially between different ancestries. Applying the cutoff per ancestry means that variants with a MAF above the threshold in any ancestry will be excluded from the analysis. The option to apply the cutoff across all samples is also available.



The **Add a new variant to list** text entry box allows you to add one or more variants to the set for inclusion in the burden test. Enter either dbSNP IDs (*e.g.*, rs112881768) or variant identifiers in the format "chromosome_coordinate_reference-nucleotide_variant-nucleotide" (*e.g.*, 8_112881768_G_A). Multiple IDs should be separated by commas or returns. Variants added manually to the table are not subject to the variant filters.



Use the “Select all” and “Select none” buttons to select or un-select the entire set of variants, or use the checkboxes to select and un-select individual variants.

If you would like to run the burden test using a small subset of the available variants for a gene, the quickest way to do this is to:

- Select the most restrictive variant filter to generate a table with the smallest number of variants
- Add the IDs of your custom variant set via the “Add a new variant to list” box
- Use the checkboxes to remove variants from the final list, if desired.

In the variant table, all columns except “Use” may be sorted by clicking the up and down arrows. Columns in the table include:

- **Use?** – checking the box for a variant includes it in the analysis, while un-checking the box removes it.
- **Variant ID** – The variant ID specifies the number of the chromosome on which the variant is located and its chromosomal coordinate in the human genome build hg19, separated by a colon. Variant IDs in the table are linked to Variant pages.
- **dbSNP ID** – the reference SNP identifier of the variant in dbSNP.
- **Chrom.** – the number of the chromosome on which the variant is located.
- **Position** – the coordinate of the variant, from the human genome build hg19.
- **MAC** – minor allele count; number of chromosomes in the sample set that contain the minor allele.
- **Polyphen** – effect on protein structure and function predicted by PolyPhen-2 as calculated by the Variant Effect Predictor: benign, possibly damaging, or probably damaging.
- **SIFT** – effect on protein function predicted by SIFT (Sorting Intolerant from Tolerant) as calculated by the Variant Effect Predictor: tolerated (T) or deleterious (D).
- **Protein change** – If the variant changes the encoded protein sequence of a gene, the change is shown in this column. The format is: “p” (for protein).(identity of the amino acid in the reference allele, in single letter code)(protein sequence coordinate of the altered residue)(identity of the amino acid in the variant allele, in single letter code). For example, “p.R325W” indicates that the variant changes amino acid 325 from arginine to tryptophan.
- **Consequence** – the effect of the variant on the protein or transcript within which it lies. This is expressed using controlled vocabulary terms from the Sequence Ontology.

Interpreting the results of GAIT analysis

- The **p-value** calculated for genetic associations represents the probability that the observed case/control frequency difference would occur by chance: the lower the p-value, the greater the statistical significance of the association. A p-value of 5 x 10^{-8} or lower represents **genome-wide significance**; a p-value of 5 x 10^{-4} or lower represents **locus-wide significance**; and a p-value of 0.05 or lower represents **nominal significance**.

- The **odds ratio (OR)** is used to represent the strength of the association between a variant and a binary disease or trait (*e.g.*, T2D, ischemic stroke, chronic kidney disease). An odds ratio near 1 means that the variant has little or no effect on disease; odds ratios greater than one mean that a carrier of the variant is more likely to have disease, and odds ratios less than one mean that a carrier is less likely to have the disease, suggesting that the variant could be protective. Note that the numbers of cases and controls in a dataset must be roughly equivalent in order to generate a meaningful OR.

- **Effect size (beta)** is analogous to the odds ratio but it can be applied to continuous traits such BMI, fasting plasma glucose level, or cholesterol level. It represents the strength of association between a variant and that trait.

- The **confidence interval (CI)** represents the probability that the odds ratio or effect size falls within the given range. For example, 95% CI: (0.852 to 0.941) for an odds ratio signifies that there is a 95% chance that the OR is between 0.852 and 0.941. If the confidence interval for an OR does not include 1, that supports the possibility that there is an effect on disease risk.

- **Meta-analysis** combines the independently association statistics for ancestry-specific associations to generate a single measure of association across all ancestries. If all groups are relatively similar to one another, performing a meta-analysis should yield similar results to a single analysis across all samples. If there are major differences between groups–for example, if there is more severe population stratification within one group than the others–then analyzing them separately has benefits. Additional parameters may be specified for the association test in the problematic group to control for stratification, and then results may be combined with the others in the meta-analysis.