TYPE 2 DIABETES
KNOWLEDGE PORTAL

Providing data and tools to promote understanding and treatment of type 2 diabetes and its complications

# Custom Aggregation Test Guide

The custom aggregation tests allow you to compute gene-level genetic association scores for the gene of your choice, offering the ability to set multiple parameters for the analysis. For the type 2 diabetes phenotype, the tests are powered by individual-level data from the **AMP T2D-GENES exome sequence analysis** dataset; for all other phenotypes, the tests are powered by the **19k exome sequence analysis** dataset. See the T2DKP Data page for descriptions of these datasets. Note that not all variants available in the variant selector will necessarily be available in the dataset being tested, and therefore only a subset of variants will be used in the aggregation test.To protect patient confidentiality, tests may not be run on sample sets that have been filtered to include fewer than 100 individuals.

The custom aggregation tests are exploratory tools. They are intended to produce results that are broadly concordant with those from expert analyses, but results produced with these tests 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.

## Custom aggregation test methods

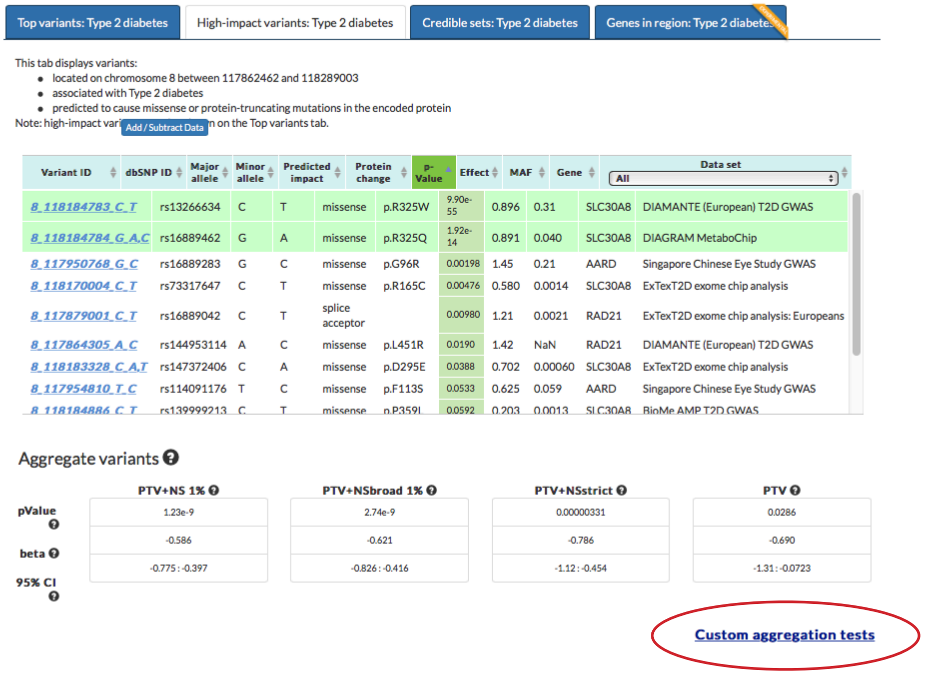
| Method | Considerations for use | References |
|---|---|---|
| Additive burden test | Counts variants to generate a significance score that is proportional to the number of variants. Most powerful when risk-associated variants in a gene have the same direction of effect and when a large fraction of variants affect risk (*e.g.*, when restricting the set of variants to protein-truncating variants). | See Lee et al, 2014 |
| Collapsing burden test (aggregate count) | Test for association between the total number of rare alleles observed per individual and a trait. Most powerful when risk-associated variants in a gene have the same direction of effect and when a large fraction of variants affect risk (*e.g.*, when restricting the set of variants to protein-truncating variants). | Morris and Zeggini 2010<br>Software |
| SKAT | Variance component test. More powerful than burden tests when combining rare variants that both increase and decrease the trait. Also more powerful when only a small fraction of included variants are causal. | Wu et al. 2011<br>Software |
| SKAT-O | Robust test that combines burden and SKAT methods via an optimal grid search. Typically more powerful than burden and SKAT separately, unless the assumptions of either test are closely matched. Computationally expensive. | Lee et al. 2013<br>Software |
| Variable threshold burden test | Runs successive collapsing burden tests across a range of minor allele frequencies, selecting the optimal statistic adaptively. P-value is calculated analytically. Most powerful when risk-associated variants in a gene have the same direction of effect and when a large fraction of variants affect risk. Computationally expensive. | Lin and Tang 2011<br>Software |

All tests use covariates to account for population structure and potential confounders. By default, the first 4 principal components along with age and sex are used as covariates.

For more information about aggregation test methods, see "Rare-Variant Association Analysis: Study Designs and Statistical Tests", Lee et al, 2014.

## Accessing the custom aggregation tests

To access the tests, click the “Custom aggregation tests” link on the **High-impact variants** tab of the Gene page.



Top variants: Type 2 diabetes
High-impact variants: Type 2 diabetes
Credible sets: Type 2 diabetes
Genes in region: Type 2 diabetes

This tab displays variants:

- located on chromosome 8 between 117862462 and 118289003
- associated with Type 2 diabetes
- predicted to cause missense or protein-truncating mutations in the encoded protein

Note: high-impact vari Add / Subtract Data n on the Top variants tab.

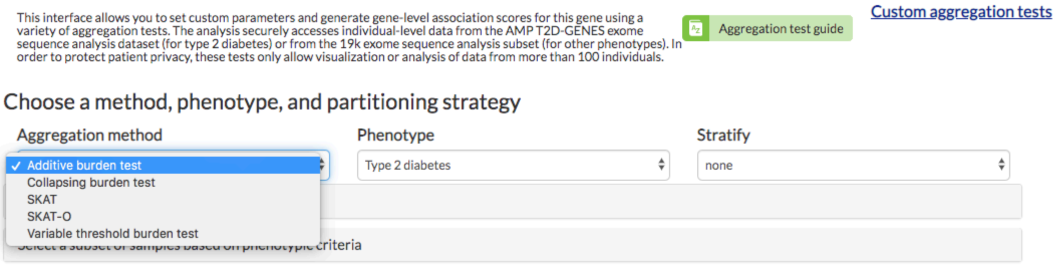
| Variant ID | dbSNP ID | Major allele | Minor allele | Predicted impact | Protein change | p-Value | Effect | MAF | Gene | Data set<br>All |
|---|---|---|---|---|---|---|---|---|---|---|
| 8_118184783_C_T | rs13266634 | C | T | missense | p.R325W | 9.90e-55 | 0.896 | 0.31 | SLC30A8 | DIAMANTE (European) T2D GWAS |
| 8_118184784_G_A,C | rs16889462 | G | A | missense | p.R325Q | 1.92e-14 | 0.891 | 0.040 | SLC30A8 | DIAGRAM MetaboChip |
| 8_117950768_G_C | rs16889283 | G | C | missense | p.G96R | 0.00198 | 1.45 | 0.21 | AARD | Singapore Chinese Eye Study GWAS |
| 8_118170004_C_T | rs73317647 | C | T | missense | p.R165C | 0.00476 | 0.580 | 0.0014 | SLC30A8 | ExTexT2D exome chip analysis |
| 8_117879001_C_T | rs16889042 | C | T | splice acceptor | | 0.00980 | 1.21 | 0.0021 | RAD21 | ExTexT2D exome chip analysis: Europeans |
| 8_117864305_A_C | rs144953114 | A | C | missense | p.L451R | 0.0190 | 1.42 | NaN | RAD21 | DIAMANTE (European) T2D GWAS |
| 8_118183328_C_A,T | rs147372406 | C | A | missense | p.D295E | 0.0388 | 0.702 | 0.00060 | SLC30A8 | ExTexT2D exome chip analysis |
| 8_117954810_T_C | rs114091176 | T | C | missense | p.F113S | 0.0533 | 0.625 | 0.059 | AARD | Singapore Chinese Eye Study GWAS |
| 8_118184886_C_T | rs139999213 | C | T | missense | p.P359L | 0.0592 | 0.203 | 0.0013 | SLC30A8 | BioMe AMP T2D GWAS |

Aggregate variants

| | PTV+NS 1% | PTV+NSbroad 1% | PTV+NSstrict | PTV |
|---|---|---|---|---|
| pValue | 1.23e-9 | 2.74e-9 | 0.00000331 | 0.0286 |
| beta | -0.586 | -0.621 | -0.786 | -0.690 |
| 95% CI | -0.775 : -0.397 | -0.826 : -0.416 | -1.12 : -0.454 | -1.31 : -0.0723 |

## Navigating the interface

- First, select an aggregation test method (see descriptions above).
- Next, select a phenotype to compute gene-level 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.
- Different options to further customize the test will be visible in the interface depending on which method you have chosen.



This interface allows you to set custom parameters and generate gene-level association scores for this gene using a variety of aggregation tests. The analysis securely accesses individual-level data from the AMP T2D-GENES exome sequence analysis dataset (for type 2 diabetes) or from the 19k exome sequence analysis subset (for other phenotypes). In order to protect patient privacy, these tests only allow visualization or analysis of data from more than 100 individuals.

Choose a method, phenotype, and partitioning strategy

Aggregation method: Additive burden test

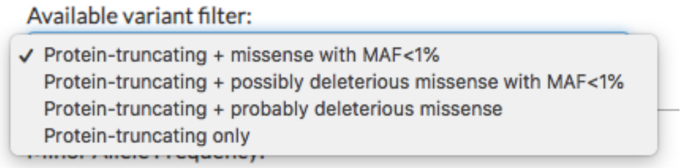
Phenotype: Type 2 diabetes

Stratify: none

Select a subset of samples based on phenotypic criteria

### Manage variant selection (available for all methods)

The variants available for analysis are those within the coding sequence of the gene whose page you are viewing, filtered according to the criteria in the “Available variant filter” pull-down menu:


Available variant filter:
✓ Protein-truncating + missense with MAF<1%
Protein-truncating + possibly deleterious missense with MAF<1%
Protein-truncating + probably deleterious missense
Protein-truncating only

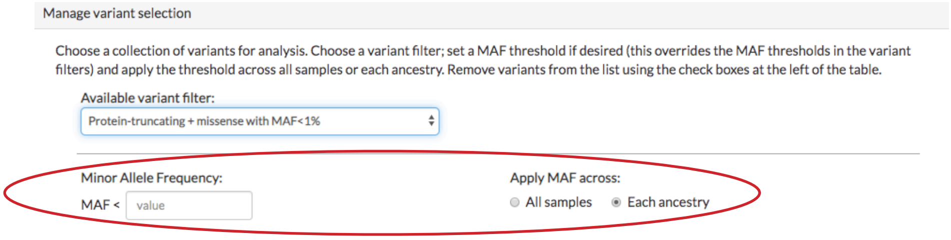
These filters were documented in Fuchsberger, Flannick,Teslovich, Mahajan, Agarwala, Gaulton *et al.* 2016. The genetic architecture of type 2 diabetes. Nature **536**(7614):41-7.

- “Protein-truncating + missense with MAF<1%” selects variants that are protein-truncating OR are missense AND have minor allele frequency of less than 1%. The MAF criterion is added because 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.

You may remove specific variants from the list by un-checking the box to the left of the variant ID.

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. Entering a MAF threshold manually overrides the MAF threshold in the chosen variant filter. For example, selecting the “Protein-truncating + missense with MAF<1%” filter and then entering MAF < 0.2 will add to the table any protein-truncating or missense variants in the gene with MAF between 1% and 20%.

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.



Choose a collection of variants for analysis. Choose a variant filter; set a MAF threshold if desired (this overrides the MAF thresholds in the variant filters) and apply the threshold across all samples or each ancestry. Remove variants from the list using the check boxes at the left of the table.

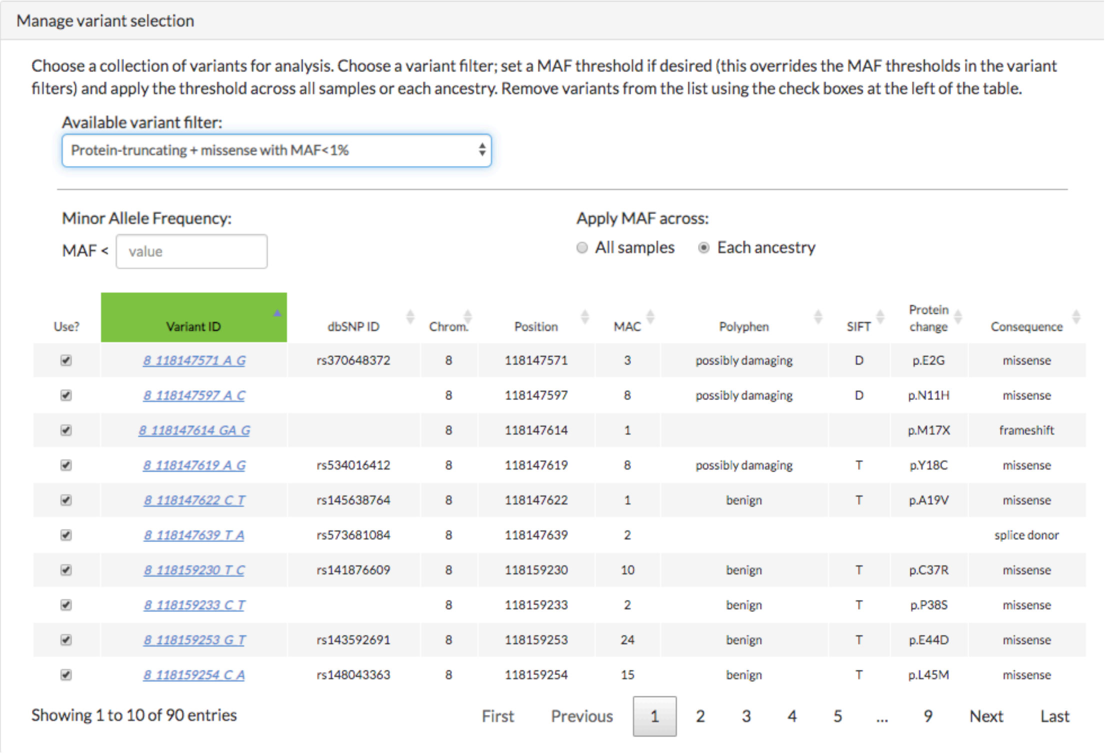
Available variant filter: Protein-truncating + missense with MAF<1%

Minor Allele Frequency: MAF < value

Apply MAF across: ○ All samples ● Each ancestry

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. An asterisk indicates a stop codon.
- **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.



Manage variant selection

Choose a collection of variants for analysis. Choose a variant filter; set a MAF threshold if desired (this overrides the MAF thresholds in the variant filters) and apply the threshold across all samples or each ancestry. Remove variants from the list using the check boxes at the left of the table.

Available variant filter: Protein-truncating + missense with MAF<1%

Minor Allele Frequency:

MAF < value

Apply MAF across:

○ All samples ● Each ancestry

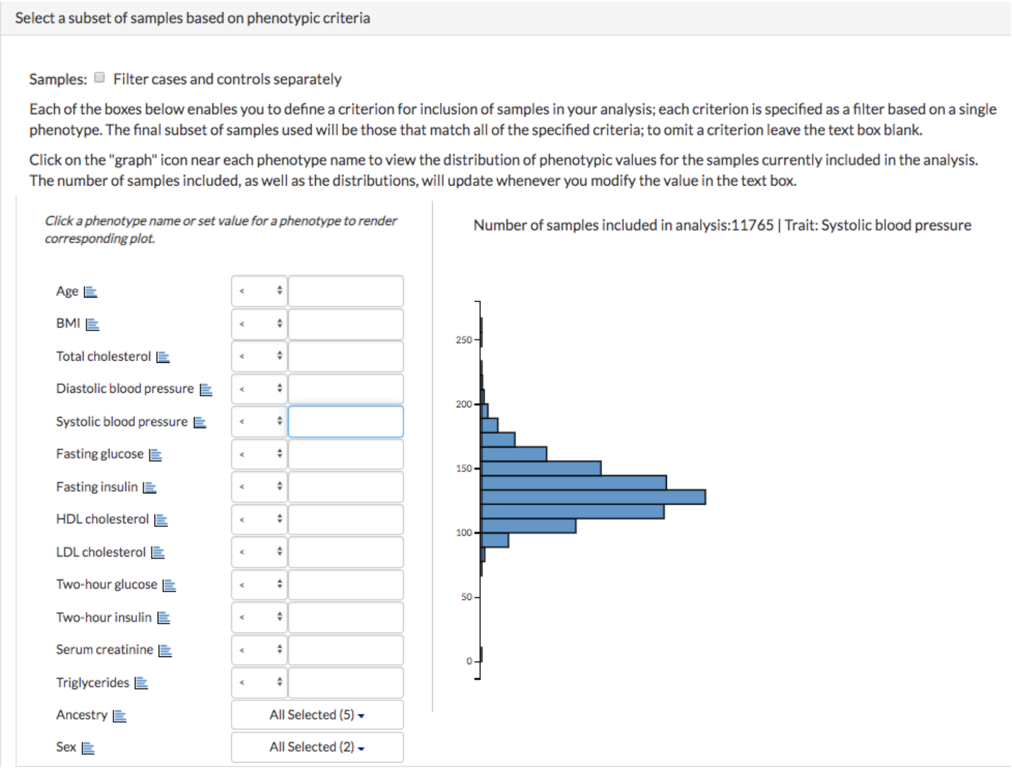
| Use? | Variant ID | dbSNP ID | Chrom. | Position | MAC | Polyphen | SIFT | Protein change | Consequence |
|---|---|---|---|---|---|---|---|---|---|
| ☑ | 8 118147571 A G | rs370648372 | 8 | 118147571 | 3 | possibly damaging | D | p.E2G | missense |
| ☑ | 8 118147597 A C |  | 8 | 118147597 | 8 | possibly damaging | D | p.N11H | missense |
| ☑ | 8 118147614 GA G |  | 8 | 118147614 | 1 |  |  | p.M17X | frameshift |
| ☑ | 8 118147619 A G | rs534016412 | 8 | 118147619 | 8 | possibly damaging | T | p.Y18C | missense |
| ☑ | 8 118147622 C T | rs145638764 | 8 | 118147622 | 1 | benign | T | p.A19V | missense |
| ☑ | 8 118147639 T A | rs573681084 | 8 | 118147639 | 2 |  |  |  | splice donor |
| ☑ | 8 118159230 T C | rs141876609 | 8 | 118159230 | 10 | benign | T | p.C37R | missense |
| ☑ | 8 118159233 C T |  | 8 | 118159233 | 2 | benign | T | p.P38S | missense |
| ☑ | 8 118159253 G T | rs143592691 | 8 | 118159253 | 24 | benign | T | p.E44D | missense |
| ☑ | 8 118159254 C A | rs148043363 | 8 | 118159254 | 15 | benign | T | p.L45M | missense |

Showing 1 to 10 of 90 entries

**Select a subset of samples based on phenotypic criteria (available for the additive and collapsing burden tests)**

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 ancestry, you may set separate filters for each group on individual tabs. A checkbox allows you to filter cases and controls separately.

The graphic in this section 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. Click on a phenotype 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.


Select a subset of samples based on phenotypic criteria

Samples: □ Filter cases and controls separately

Each of the boxes below enables you to define a criterion for inclusion of samples in your analysis; each criterion is specified as a filter based on a single phenotype. The final subset of samples used will be those that match all of the specified criteria; to omit a criterion leave the text box blank.

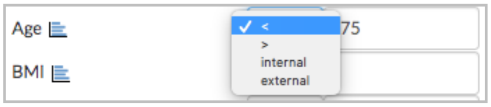
Click on the "graph" icon near each phenotype name to view the distribution of phenotypic values for the samples currently included in the analysis. The number of samples included, as well as the distributions, will update whenever you modify the value in the text box.

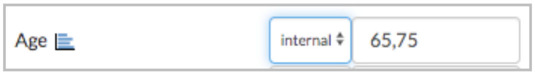
*Click a phenotype name or set value for a phenotype to render corresponding plot.*

Age <
BMI <
Total cholesterol <
Diastolic blood pressure <
Systolic blood pressure <
Fasting glucose <
Fasting insulin <
HDL cholesterol <
LDL cholesterol <
Two-hour glucose <
Two-hour insulin <
Serum creatinine <
Triglycerides <
Ancestry All Selected (5)
Sex All Selected (2)

Number of samples included in analysis:11765 | Trait: Systolic blood pressure

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. To specify a range, enter two values separated by commas. Filters may be applied to multiple phenotypes. If the number of filtered samples is fewer than 100, the bar graph does not display in order to protect patient privacy.


Age 75
BMI


Age internal 65,75

### Control for covariates (available for the additive and collapsing burden tests)

The additive and collapsing burden tests allow you to choose covariates in order to control for population structure or test whether your chosen phenotype is dependent on other phenotypes. (The covariates Principal Components 1-4, age, and sex are included by default in the other tests.)

In the additive and collapsing burden tests, 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 all desired parameters, compute associations by clicking the "Launch analysis" button.

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

- 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.

---

### Interpreting burden test results

- The **p-value** calculated for genetic associations represents the probability that the observed frequency difference would occur by chance: the lower the p-value, the greater the statistical significance of the association.

- The **odds ratio (OR)** is used to represent the strength of the gene-level association with a binary disease or trait (*e.g.*, T2D, ischemic stroke, chronic kidney disease). An odds ratio near 1 indicates little or no effect on disease; odds ratios greater than one mean increased likelihood of having disease, and odds ratios less than one mean decreased likelihood of having disease. 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.

- 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 computed 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.