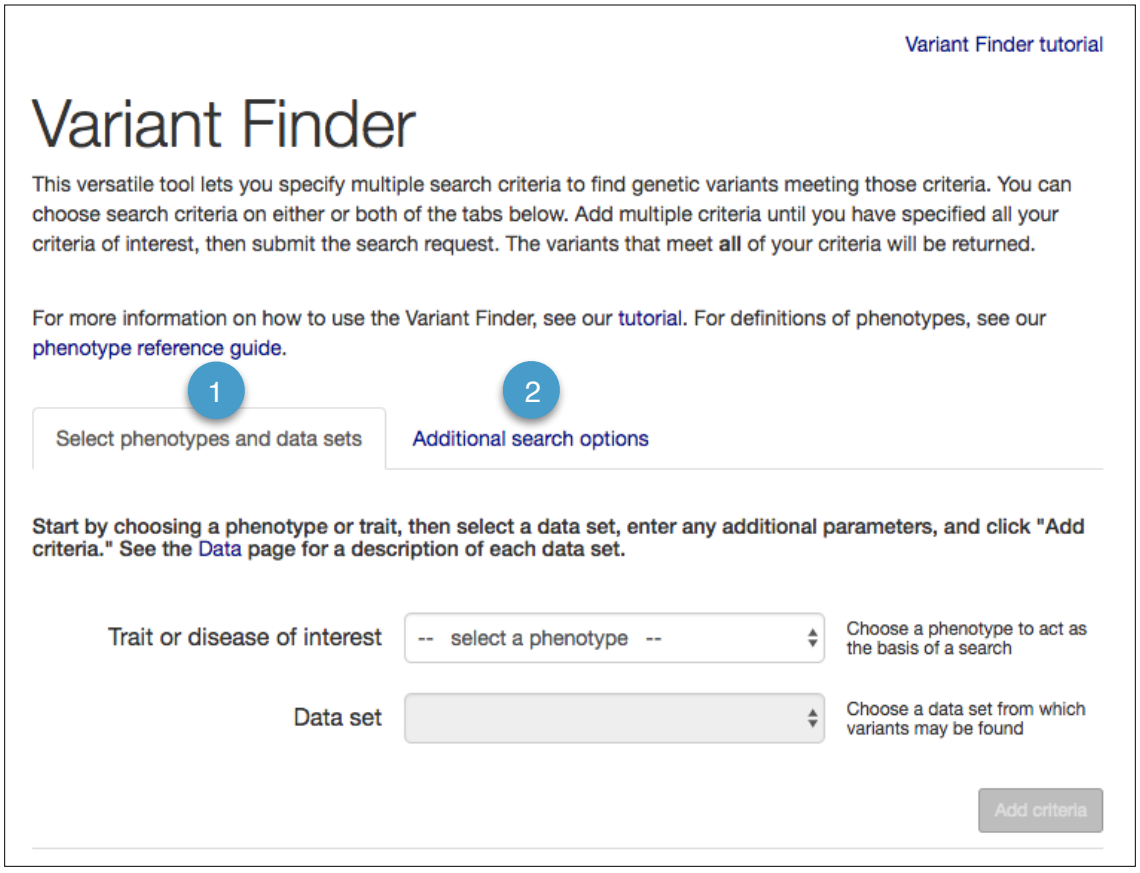
TYPE 2 DIABETES KNOWLEDGE PORTAL
type2diabetesgenetics.org

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

Variant Finder Tutorial

The Variant Finder is a versatile search builder that retrieves sets of variants that meet specific criteria. Using the Variant Finder, you can:

- find variants that are associated with multiple distinct phenotypes or traits
- set thresholds for the significance of associations retrieved
- specify the direction and magnitude of the effect of variants on a disease or trait
- specify the genomic location of variants retrieved, by gene or by region
- find variants with specific effects on proteins, such as missense mutations predicted to be deleterious
- search within individual cohorts and other subsets of datasets



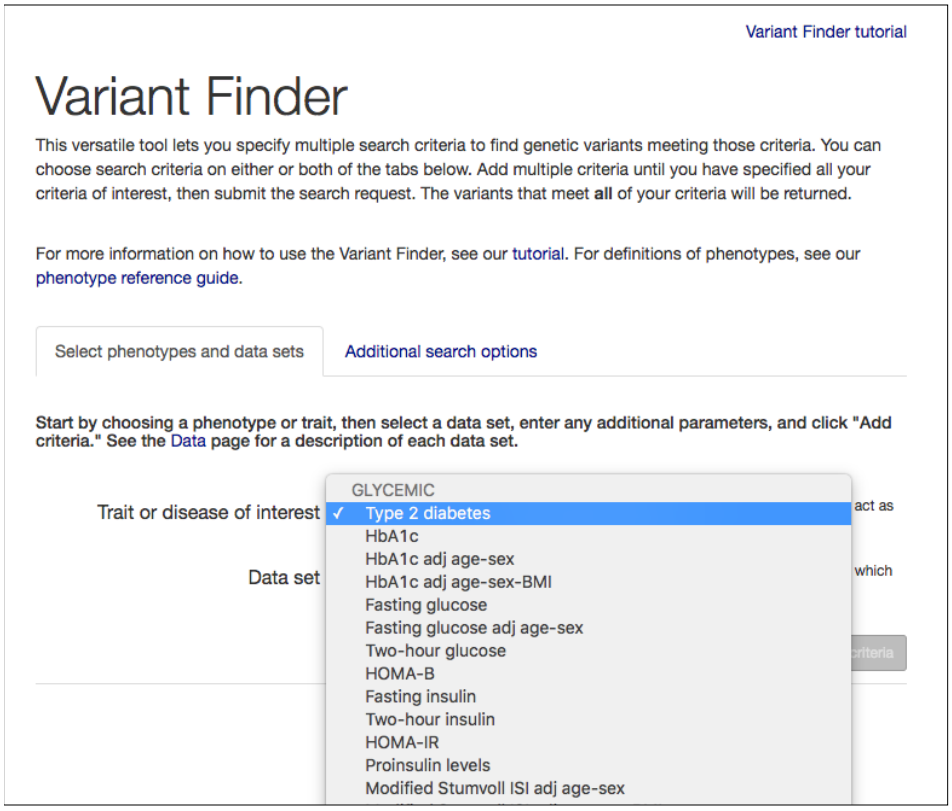
Two tabs on the Variant Finder interface allow you to set two types of criteria:

1 On the “Select phenotypes and data sets” tab, start by choosing a phenotype, then choose datasets and parameters for associations with that phenotype.

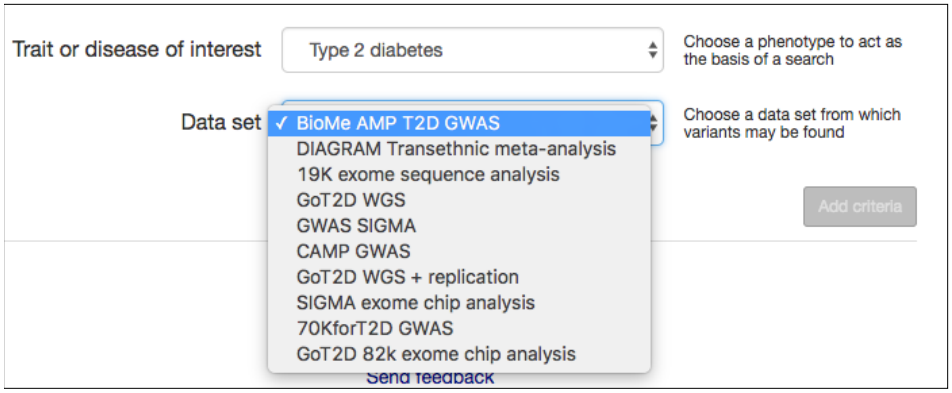
2 On the “Additional search options tab, you may choose to find variants with specific allele frequencies in datasets without associated phenotypes (1000 Genomes and gnomAD); select variants by genomic location (proximity to a gene, or chromosomal coordinate range); or select variants by their predicted effects on encoded proteins.

1 **Set criteria using the “Select phenotypes and data sets” tab**

Start by selecting a phenotype (trait or disease) from the pull-down menu. For definitions of phenotypes, see our Phenotype Reference Guide.



Next, select a dataset from the pull-down menu. The menu shows only the datasets that contain associations for the phenotype you have chosen.



The T2DKP has so many datasets, how do I know which one to pick?

GWAS datasets have the largest number of samples and contain common variants (minor allele frequency (MAF) > 5%) that are located in both coding and non-coding regions of the genome. Choose a GWAS set for the broadest and most unbiased search. Note that GWAS signals are unlikely to pinpoint the causal variant for an effect; rather, they identify a region of the genome containing a causal variant. For type 2 diabetes, the most comprehensive GWAS dataset is the **DIAGRAM transethnic meta-analysis** set.

Exome chip datasets contain variants located in protein-coding regions. Exome chips detect about 80% of common and low-frequency (5% > MAF > .5%) coding variants. Choose an exome chip set to focus on variants that may directly affect protein structure. The **GoT2D 82K exome chip analysis** set is currently the largest set.

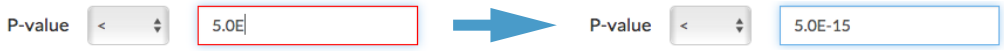
Exome sequencing datasets have smaller numbers of samples than array-based datasets, but include all variants in protein-coding regions, both common and rare. Choose the **19K exome sequence analysis** dataset to search a comprehensive set of variants that may directly affect protein structure.

Whole-genome sequencing datasets typically have the smallest number of samples but include every variant, whether common or rare, coding or non-coding. Whole genome sequences are included in our **GoT2D WGS** dataset, which has about 2,600 samples.

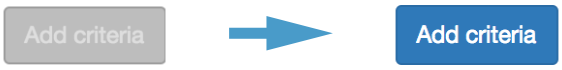
Explore all datasets by data type and phenotype on our Data page.

After you select a dataset, you will be able to specify the additional parameters that are available for that dataset.

As you type in parameters, the entry box is outlined in red until you have entered a value in the correct format; then it changes to blue.



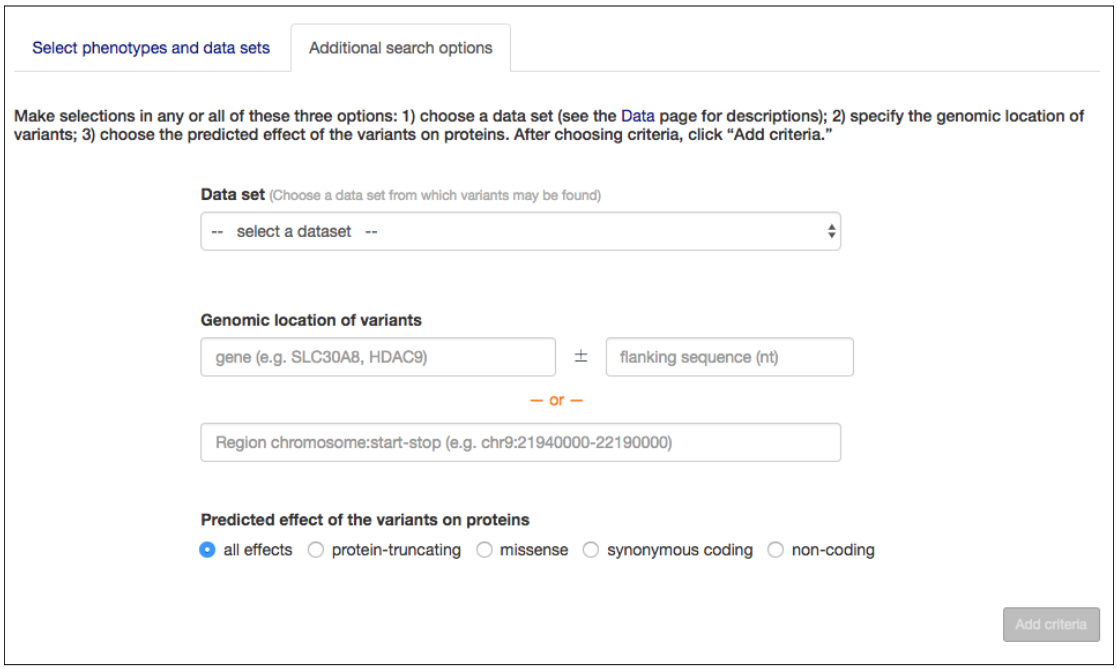
See the table at the end of this tutorial for a list of the dataset parameters that are available in the Variant Finder and their definitions. The “Add criteria” button will turn blue when you have selected all of the parameters that are required for your chosen dataset.



When finished selecting parameters, click the “Add criteria” button to save those criteria.

Set criteria using the "Additional search options" tab

In addition to, or instead of, setting search criteria by using the options on the **"Select phenotypes and data sets"** tab, you can also set criteria on the **"Additional search options"** tab.



The first option on this tab allows you to search for variant information in datasets from the 1000 Genomes and gnomAD projects. These datasets do not include phenotype associations, but include allele frequencies in cohorts of different ancestries. To query one of these datasets, you must specify a minor or effect allele frequency.

You may also search for variants by their genomic location, either by specifying a gene and an optional extent of up- and downstream flanking sequence, or by entering a chromosomal coordinate range in the format "chromosome number:start coordinate-end coordinate," *e.g.*, "9:21940000-22190000."

You can also limit the search by the predicted effect of variants on the encoded protein. For missense mutations, you can specify the severity of those effects, as predicted by three different algorithms: PolyPhen-2, SIFT, and CONDEL, provided by Ensembl's Variant Effect Predictor.

If you specify the variant effect without specifying a gene, the search will look for variants causing effects of that type in any gene.

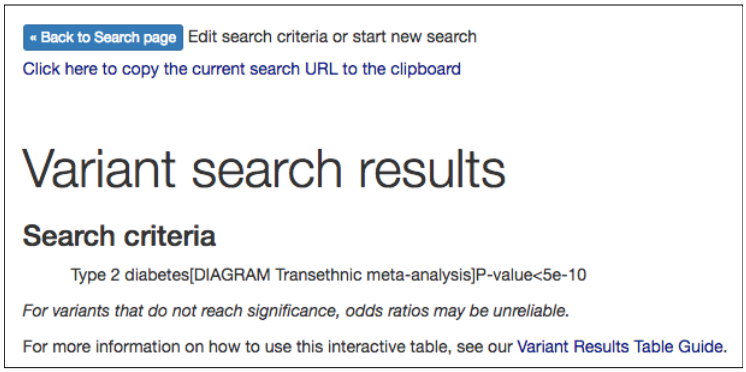
Run the search
After clicking “Add criteria,” your search parameters are stored in the “Search detail” section. You can now submit that search, or add more search criteria.

You can repeat the process of building searches as many times as you want. All your searches are listed in the Search detail section. Submitting the search request will find the set of variants that answer **all** of the search criteria.

From the search results table, click on a gene or variant name to go to its page and learn more about it. The table can be sorted by each column and may also be exported in various formats. Find complete instructions on how to use the interactive Variant Results table in our Variant Results Table Guide.

Edit the search
At the top of the Results page, the “Back to Search page” link takes you to your list of search criteria, where you can edit or delete individual criteria.

Save or share the search
The link “Click here to copy the current search URL to the clipboard” lets you save a URL that will regenerate your search, for future use.



Available parameters
The table below defines the parameters that may be available for datasets in the Variant Finder. For more information about the concepts, see our Introduction to genetic association analysis.

Some parameters are required, while others are optional. After you have entered values for all required parameters, the “Add criteria” button turns blue.

Parameter	Definition
Case minor allele count	Case minor allele count (MAC) count refers to the number of chromosomes that carry the minor allele in the group of cases (individuals who have the disease or binary trait of interest) in sampled population, and so reflects both allele frequency and sample size.
Cohort	Any available subsets of a dataset are listed. You may choose one subset, or include all.

Parameter	Definition
Control minor allele count	Control minor allele count (MAC) count refers to the number of chromosomes that carry the minor allele in the group of controls (individuals who do not have the disease or binary trait of interest) in the sampled population, and so reflects both allele frequency and sample size.
Cumulative posterior probability	Cumulative posterior probability represents the likelihood that a credible set of variants is causal for its phenotypic effect.
Effect allele frequency	Effect allele frequency (EAF) denotes the frequency of the effect allele in a given population. EAF is equivalent to the minor allele frequency when the minor and effect alleles are the same.
Effective sample size	The effective sample size is the number of samples in a study with equal number of cases/controls such that the balanced study would have equivalent power to the actual study. Effective sample size equals sample size in a balanced study, and decreases as cases/controls become more unbalanced.
Effect size (beta)	Effect size represents the magnitude and direction of association between a variant and a trait. It is analogous to the odds ratio, but it can be applied to continuous traits such BMI, fasting plasma glucose level, or cholesterol level.
Minor allele count	Minor allele count (MAC) count refers to the number of chromosomes in the sampled population that carry the minor allele, and so reflects both allele frequency and sample size.
Minor allele frequency	Minor allele frequency (MAF) denotes the frequency of the minor allele in a given population. MAF is equivalent to the effect allele frequency when the minor and effect alleles are the same. Variants with MAF > 5% are considered to be common; variants with MAF between 0.05% and 5% are termed low-frequency; and variants with MAF < 0.05% are termed rare.
Odds ratio	The odds ratio (OR) represents the magnitude of the association between an allele and a binary, or dichotomous, disease or trait (e.g., T2D, chronic kidney disease, bipolar disorder) and the direction of its effect. An odds ratio near 1 means that the allele has little or no effect on a disease; odds ratios greater than 1 mean that the effect allele is more likely to be carried by people with the disease, and odds ratios less than 1 mean that the effect allele is more likely to be carried by people without the disease, suggesting that it could be protective. • Note that the numbers of cases and controls in a data set must be roughly equivalent in order to generate a meaningful OR. • OR may not be accurate for variants with very low allele count in a particular data set. • OR for the non-effect allele is equal to 1/OR of the effect allele.
Posterior probability	Posterior probability represents the likelihood that a variant is causal for its phenotypic effect.

Parameter	Definition
P-value	The p-value calculated for genetic associations represents the probability that the observed case/control frequency difference would occur by chance if there were no association between the SNP and disease: the lower the p-value, the greater the statistical significance of the association. A p-value of 5×10^{-8} or lower represents genome-wide significance; a p-value of 5×10^{-4} or lower represents locus-wide significance; and a p-value of 0.05 or lower represents nominal significance. Acceptable formats for entering p-value in the Variant Finder include: 0.005; 5e-3; 5.0e-3; 5E-3; 5.0E-3.
Sample size	Sample size denotes the number of individuals with a particular variant for which there are data about the specified phenotype. Check the Data page to see how many total samples are present in a dataset; for any phenotype, the number of individuals with that phenotype may be the same as the total number in the dataset, or it may be smaller. The sample size for a phenotype-variant association is meaningful because larger sample numbers lead to more significant p-values.
Z-score	The Z-score represents both the p-value and direction of effect of a variant-phenotype association. Z-score must be entered as a positive number, e.g.: 3; 4.0; 4.6.