Overview of the Human Genetic Evidence (HuGE) framework

The purpose of the Human Genetic Evidence (HuGE) framework is to evaluate the extent to which human genetic data support the hypothesized involvement of a gene in human disease. The goal of this document is to provide (a) an overview as to how the HuGE framework was designed (*i.e.* its methodology) and (b) a step-by-step process to use the framework – together with publicly available resources – to evaluate genetic support for a gene.

The HuGE framework estimates genetic support using a Bayesian approach. Bayesian statistics provides a series of rules to model probabilities as degrees of "belief" in an event – in our case, the event is that genetic perturbations of a gene are associated with disease. We define "genetic support" as the estimated *increase* in probability of this event based on the observed genetic data. Bayesian statistics are more naturally applied to the "odds" of events, rather than probabilities, and there is a one-to-one correspondence between odds and probabilities: if an event has probability *p*, then its odds are *p* / (*1-p*).

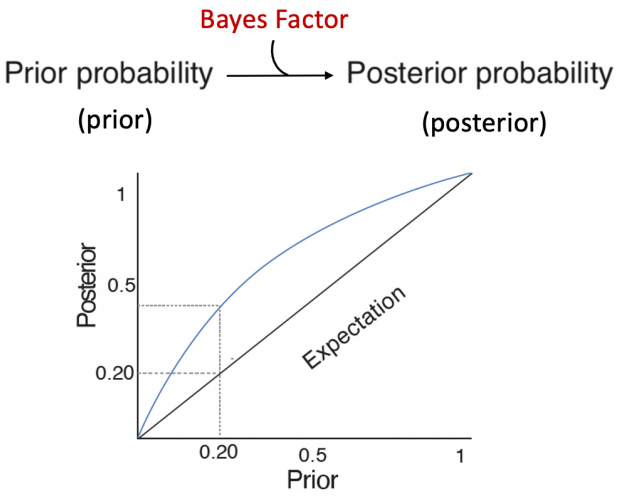
The HuGE framework defines a set of rules to incorporate data from different classes of genetic variation into a Bayesian approach. Before examining genetic data, our belief is reflected in a "prior" probability (or, equivalently, a prior odds). We then express both rare and common variant genetic data as 'Bayes Factors (BFs)', which update the prior probability into a "posterior" probability through a simple multiplicative factor of the odds:

$$PO = BF * \left(\frac{prior}{1 - prior}\right)$$

where PO indicates the posterior odds, BF indicates the Bayes Factor, and $prior$ indicates the prior probability. We can then convert the posterior odds into the posterior probability through:

$$PPA = \frac{PO}{1 + PO}$$

The *BF* value is the final estimated genetic support for the gene. BFs of 1 result in a posterior equal to the prior, and BFs greater than 1 result in a posterior greater than the prior.



Thus, calculating the degree to which genetic data provide genetic support is reduced to calculating a Bayes Factor from the observed genetic data. The final *PPA* value then follows from the above equations, together with a prior probability. Detailed derivations of the Bayes Factors used in the HuGE framework can be found below. At the end, genetic support is communicated via three independent mechanisms, corresponding to the three values shown in each cell of **Figure 1**:

Common Variation	No evidence $p \geq 0.1$	Weak $p < 0.1$	Nominal $p < 0.05$	Strong $p < 1 \times 10^{-3}$	Exome-wide $p < 2.5 \times 10^{-6}$
Causal coding variant	**Compelling** 95% | 99%	**Compelling** 95% | 99%	**Compelling** 99% | 9%	**Compelling** 99% | 99%	**Compelling** 99% | 99%
Nearest gene	**Very Strong** 70% | 90%	**Very Strong** 80% | 95%	**Extreme** 90% | 95%	**Compelling** 99% | 99%	**Compelling** 99% | 99%
Coding variant	**Strong** 50% | 85%	**Very Strong** 60% | 90%	**Very Strong** 75% | 95%	**Compelling** 95% | 99%	**Compelling** 99% | 99%
GWAS locus	**Moderate** 15% | 40%	**Moderate** 20% | 55%	**Moderate** 30% | 70%	**Very Strong** 75% | 95%	**Compelling** 99% | 99%
No evidence	**No evidence** 5% | 20%	**Anecdotal** 5% | 25%	**Moderate** 15% | 45%	**Strong** 50% | 85%	**Compelling** 95% | 99%

Rare Variation

1. The text labels refer to qualitative values of evidence, derived directly from the calculated Bayes Factors. These labels have been previously applied to Bayes Factors in other settings (Jeffreys, 1961).
2. The left numbers correspond to a conservative quantitative estimate of genetic support. These follow from applying the Bayes Factors with a prior of 5%, which corresponds to an order-of-magnitude level assumption that, for a common disease, there are likely to be ~1,000 genes (*i.e.* 5% of genes in genome) that impact susceptibility. This estimate has been used in previous Bayesian genetic studies (Satterstrom et al., 2020).
3. The right numbers correspond to an optimistic estimate of genetic support. These follow from applying the Bayes Factors with a prior of 20%. It has previously been estimated that genes with supporting evidence from a mouse model have about a 20% chance of exhibiting a disease association (Flannick et al., 2019). This optimistic model may be most appropriate for researchers to use when evaluating genes with experimental support.

Applying the HuGE calculator to a gene involves three steps: (1) evaluating evidence from common variation; (2) evaluating evidence from rare variation; (3) using the two classes of evidence to obtain a HuGE score. Below we describe how to conduct these steps using public resources, deriving within each step how Bayes Factors are calculated. All of the below steps apply to a single trait for which genetic support is being evaluated.

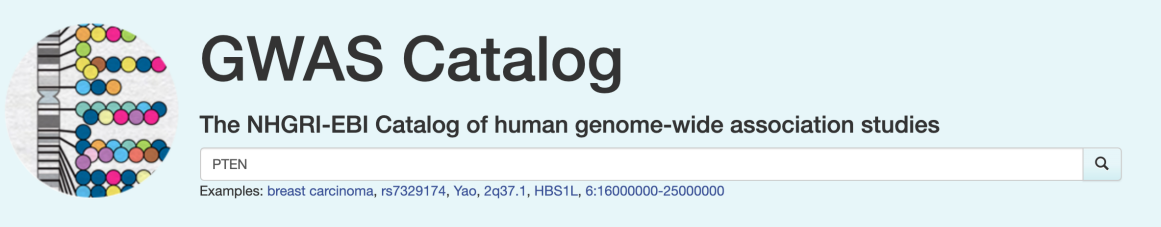
Step 1: Assess evidence from common variation

1.1 Query GWAS associations nearby the gene. Any resource that supports gene-level queries can be used for this step. Potential resources include:

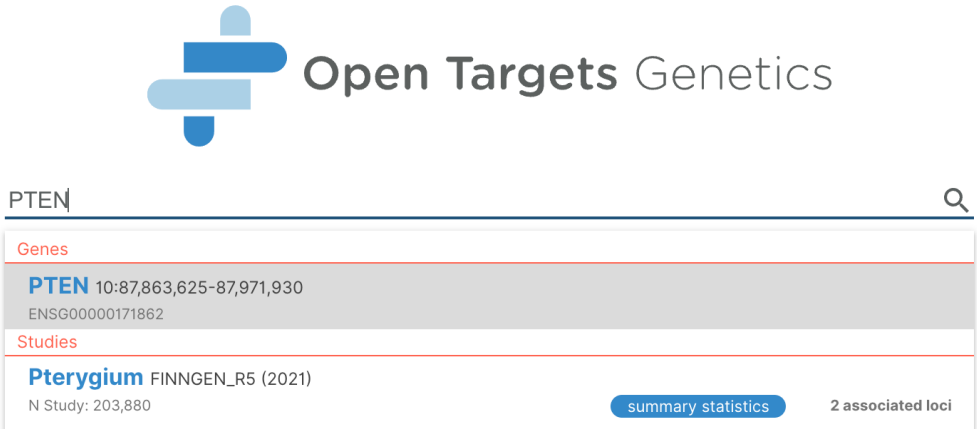
The common metabolic disease knowledge portal (CMDKP): https://hugeamp.org/
The CMDKP contains GWAS associations from >350 metabolic traits across >300 datasets and can be freely queried. To assess a gene of interest, visit the website and

enter the gene into the search bar. The subsequent page will show a list of traits for which GWAS associations are observed nearby the gene.

The GWAS catalog: https://www.ebi.ac.uk/gwas/
The GWAS catalog is a large, trait-agnostic repository for association statistics from GWAS studies and, currently, consists of data from >5500 publications. To scan for trait associations, enter a gene of interest in the search bar and the website will return all variants nearby the gene that have an association in the database.

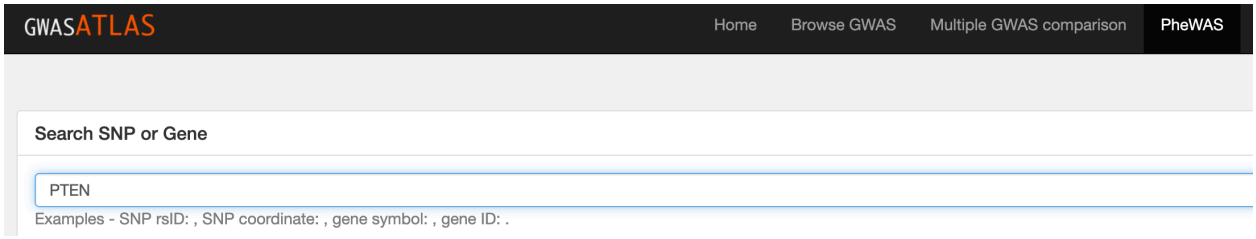


Open Targets Genetics: https://genetics.opentargets.org/
The Open Targets Genetics database consists of aggregated results based on reported GWAS summary statistics in the GWAS catalog (see above). To focus on a gene of interest, use the search bar to enter and search for a specific gene:



The GWAS Atlas: https://atlas.ctglab.nl/

The GWAS Atlas is a GWAS repository consisting of >4,700 studies. This database is focused on study-level results and, thus, requires more work from a user to understand the details behind the association. The PheWAS page can be used to query a gene of interest and assess whether there is a known trait association with a variant near the gene.



1.2 Assess the properties of variants associated with the phenotype of interest. Depending on the results of the above query, we assign the gene to one of five categories. The categories are designed to provide simple rules to overcome the major challenge of using GWAS data: faced with an association, we do not know which gene is "causal" for the association – most GWAS associations map to noncoding regions of the genome and are spread over large regions that contain many genes. The HuGE framework provides a few simple steps that can be easily conducted as a first pass to obtain evidence for the gene's causality.

The general approach in each step is to determine (based on prior literature evidence) the likelihood that a gene with the observed evidence is causal. Assuming this probability is *PPA*, we then derive a BF using the following equation:

$$BF = \left(\frac{PPA}{1 - PPA} \Big/ \frac{prior}{1 - prior} \right)$$

In all cases, we use a prior of 5%, as the literature estimates for the posteriors are provided for genes about which nothing else is known.

***No evidence*: No association within 100kb of the gene achieves $p<5\times10^{-8}$.** In this case, the common variant data provide no genetic support, and we assign a BF=1. Reasons for not assigning BF<1 are discussed in the corresponding manuscript.

***Causal coding variant:* Most significant association in the genomic locus is a nonsynonymous variant.** Based on previous literature (Mahajan et al., 2018b), a causal nonsynonymous variant association is the strongest tier of genetic support for the gene. While rigorously determining causality requires in-depth fine mapping and credible set analysis, for the purposes of our guidelines we assign genes to this tier if the variant with the lowest *p*-value in the region is a nonsynonymous variant within the gene. This search can be conducted within resources that show all variants associated with a trait and allow filters on variant annotations shown (see below for an example using the CMDKP).

If a gene falls into this tier, we assign a Bayes Factor that raises a 5% prior to a ~95% posterior probability, corresponding to a rough intuition of "near certainty" for the gene. This corresponds to a BF of ~350. Note that a BF of 350 raises a prior of 20% to an even greater posterior.

Nearest gene: **Most significant association in the genomic locus is closest to the gene of interest.** Previous reports have noted that, the gene nearest to the most significant association within a GWAS locus is the causal gene for that association ~70% of the time (Stacey et al., 2019). This search can be conducted on resources that either annotate the gene nearest to the associations, or that display graphical LocusZoom plots of the associations and genes within a region.

If a gene does not fall into the "causal coding variant" tier, but it is the gene nearest to the strongest association, then the BF becomes the value required to raise a prior of 5% to ~70%. This corresponds to a BF of ~45.

Coding variant: **A nonsynonymous variant in the gene of interest is associated at $p<5\times10^{-8}$.** The next level of evidence is based on whether there is a nonsynonymous variant associated with the phenotype of interest. Previous studies have shown that, if a gene harbors a nonsynonymous variant association – even if there might be stronger associations in the region – then it is the causal gene ~50% of the time (Mahajan et al., 2018b).

If the gene is not in the "causal coding variant" or "nearest gene" tiers, but it does harbor an associated coding variant, then we assign it a BF that raises a prior of 5% to a posterior probability of ~50%. This corresponds to a BF of ~20.

GWAS locus: **An association is detected within 100kb of the gene of interest**. If a gene is not in any of the above tiers, than we assume that it is equally likely as any gene in the region to be causal for the association. On average, there are ~7 genes nearby (*i.e.* within 100kb) of a GWAS association. Thus, we assign the gene a BF sufficient to raise a prior of 5% to a posterior of ~14%. This corresponds to a BF of ~3.

If we wish to obtain a more accurate BF specific to a GWAS locus, we can redefine this calculation according to the actual number of genes nearby the association. For example, if the locus of interest contains 6 genes (as opposed to the average of 7), the posterior probability can be assumed to 1/6, or ~17%, which would result in a BF to ~3.9.

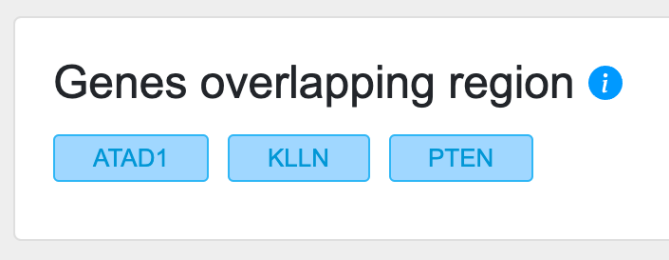
Step 2: Assess evidence from rare variation

2.1 *Finding rare variant association statistics.* As with common variation, there are several freely available databases of rare coding variation association statistics. These databases allow queries of gene-level "aggregate" associations, since individual rare variants rarely have enough statistical power to detect associations and thus are most often analyzed in "groups" to boost statistical power. Two recommended resources are:

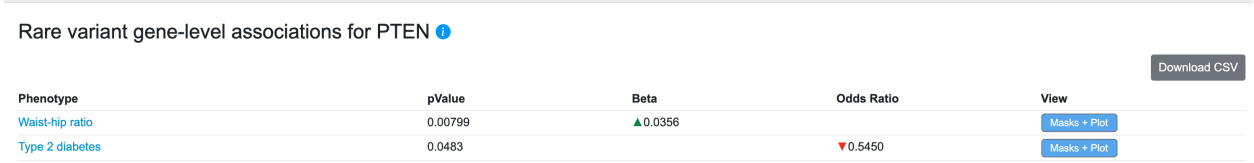
The common metabolic disease knowledge portal (CMDKP): https://hugeamp.org/

Along with GWAS data, the CMDKP contains association statistics from rare variant gene-level tests for up to 23 metabolic traits across nearly 50,000 individuals. To evaluate a gene of interest, search the homepage for the gene:

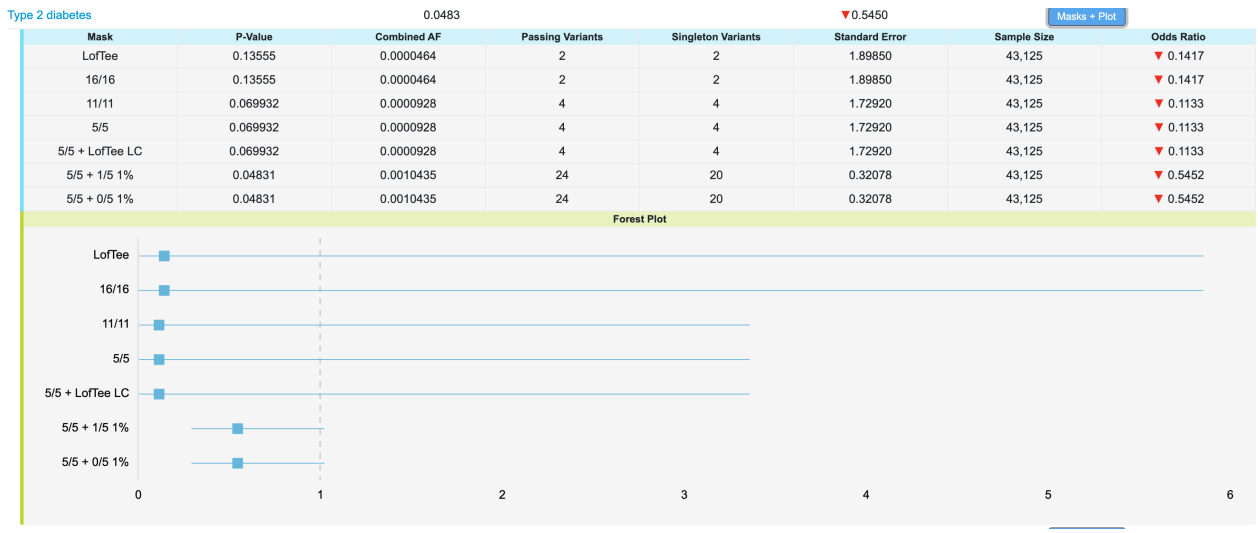
On the subsequent page, click on the gene name in the “Genes overlapping region” section:



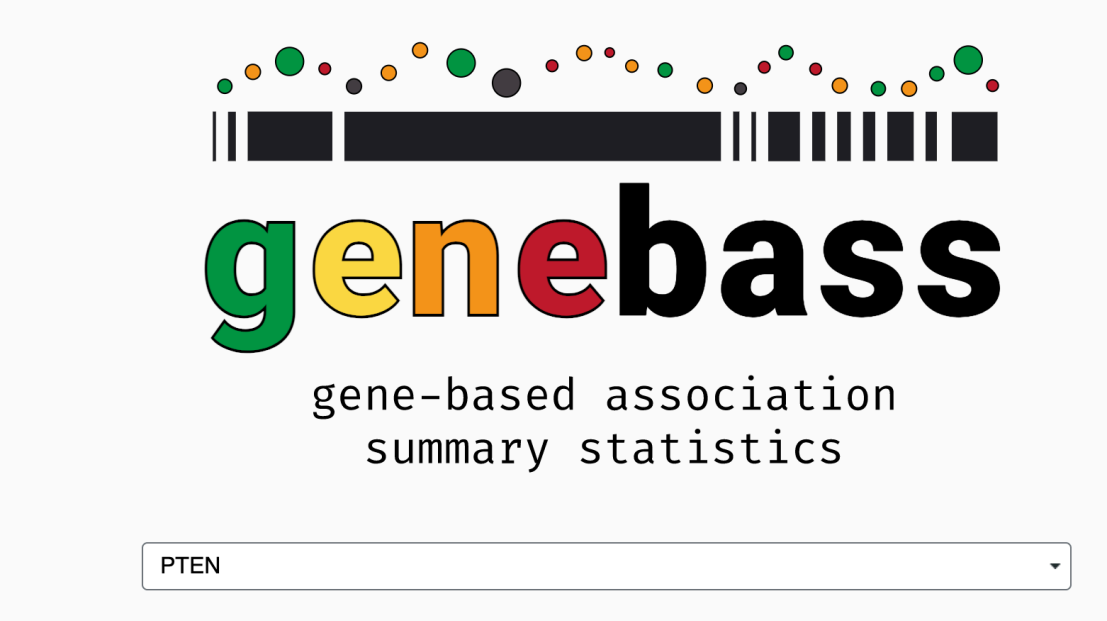
On the subsequent page, scroll down to the rare variant gene-level association statistics:



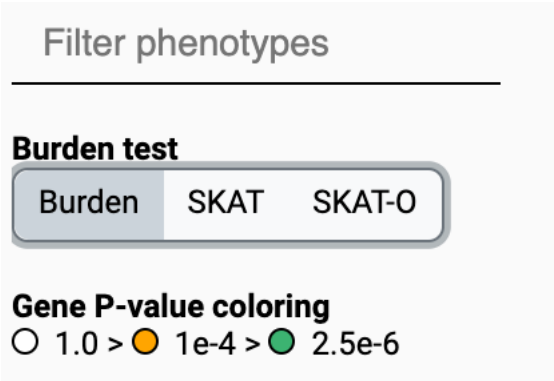
The *p*-values shown in the table are corrected for the number of variant groupings tested and are what should be used in the HuGE framework. If desired, a researcher can inspect the group-level statistics by clicking on “Masks + Plot”:



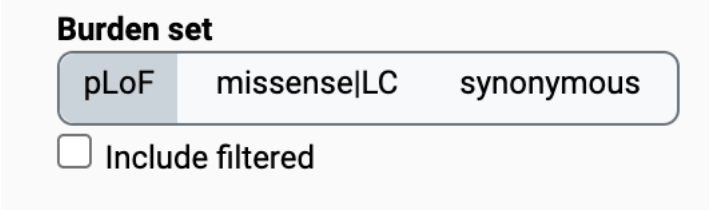
Gene-based association summary statistics database (genebass): https://genebass.org/
Genebass is a database containing gene-level association statistics from over 280,000 individuals from the UK Biobank across >3,800 traits. Note that, while most variants in the genome are rare, this database contains gene-level results that were not filtered on minor allele frequency (MAF) and, thus, may contain common variants in the gene-level test results. To analyze results for a specific gene, either (a) visit the detailed walkthrough page (https://genebass.org/walkthrough) or (b) follow several steps, which begin with a query for the gene of interest:



Next, select the “burden test” tab



Next, select variant groupings of interest (*i.e.* only loss-of-function variants or loss of function + missense variants):



These queries will produce up to two *p*-values for the gene. The *p*-value used in the HuGE framework should be the minimum *p*-value, corrected for two tests (2 * p is an acceptable approximation to use).

2.2 Assessing the strength of the rare variant association. Unlike common variants, rare variant gene-level based associations have less statistical power, but, when detected, point to the causal gene. Thus, the framework for assessing rare variant-related evidence focuses on evaluating association *strength* rather than evaluating gene causality (as is done in the common variant step). To convert the strength of the association to a Bayes Factor, we employ a previously described method (Wakefield, 2008) where the estimated beta and standard error from the association test is used to calculate an "approximate bayes factor":

$$ABF = \sqrt{\frac{V}{V + \omega}} * exp\left(\frac{\omega\beta^2}{2V(V + \omega)}\right)$$

where ABF indicates "approximate bayes factor", β indicates the effect size of the gene-level association, V indicates the square of the standard error of the association effect size, and ω indicates the estimated prior allelic variance. We recommend a value of 0.3696 for the prior allelic variance, reflecting a distribution of true effect sizes ~9 times more variable than that previously used for common variants (Mahajan et al., 2018a)). If precise BFs are important for the analysis, we recommend that researchers perform a sensitivity analysis to determine the degree to which their conclusions depend on this parameter.

Formally, the ABFs depend on the effect sizes and standard errors of the association, rather than the *p*-values. Both effect sizes and standard errors can be obtained through the CMDKP and genebass, and we recommend that the exact ABF calculation be performed

as a matter of best practice. However, a rough estimate of genetic support will be adequate for many researchers, and we therefore provide a simple approximate mapping between categories of *p*-values and ABFs. We obtained the values below by calculating ABFs for a sample of genes with *p*-values in the specified range and then taking the average value.

***No evidence*:** if the rare variant gene-level association has $p \geq 0.1$, the BF is set as 1 (equivalent to no evidence provided by rare variants). Reasons for not assigning a BF<1 are discussed at length in the corresponding manuscript.

***Exome-wide*:** if the rare variant gene-level association has $p \leq 2.5 \times 10^{-6}$, the BF is set at 350, corresponding to a maximum posterior of ~95% for a prior of 5%. We cap the posterior to be conservative, although a case could be made for higher values.

***Strong*:** if the rare variant gene-level association has $2.5 \times 10^{-6} \leq p \leq 1 \times 10^{-3}$, the BF is set as 20.

***Nominal*:** if the rare variant gene-level association has $1 \times 10^{-3} \leq p \leq 0.05$, the BF is set as 3.

***Weak*:** if the rare variant gene-level association has $0.05 < p < 0.1$, the BF is set as 1.5.

Step 3: Combine evidence for common and rare variation to calculate the HuGE score
In the final step, we assume that the common variant and rare variant genetic data are independent. This is a reasonable approximation for most genes, although researchers should be aware that there are cases were common variants affect rare variant associations through linkage disequilibrium or through inclusion in gene-level tests (*e.g.* within genebass). Under the independence assumption, the common variant and rare variant BFs simply multiply to yield a combined BF. This BF is then assigned to one of seven tiers based on previous terminology (Jeffreys, 1961).

HuGE score value	Evidence level
1	No evidence
>1	Anecdotal
≥3	Moderate
≥10	Strong
≥30	Very strong
≥100	Extreme
≥350	Compelling

These tiers are displayed as the qualitative categories in **Figure 1**. To map them to PPA values, we then apply the equations:

$$PO = HuGE * \left(\frac{prior}{1 - prior}\right)$$

$$PPA = \frac{PO}{1 + PO}$$

where PO indicates the posterior odds, $HuGE$ indicates the combined BF, $prior$ indicates the prior probability of association, and PPA indicates posterior probability of association. The PPA is displayed in **Figure 1** under conservative (prior=5%) and optimistic (prior=20%) scenarios. If researchers are so inclined, they can calculate different PPA values for different priors, if they have additional data supporting the gene that justifies a prior other than 5% or 20%.

To better visualize the impact of combining different levels of common and rare variant evidence and the dependence on priors brought by a researcher, **Figure 1** in the manuscript summarizes the above calculations in a matrix:

Common Variation	No evidence $p \geq 0.1$	Weak $p < 0.1$	Nominal $p < 0.05$	Strong $p < 1 \times 10^{-3}$	Exome-wide $p < 2.5 \times 10^{-6}$
Causal coding variant	**Compelling** 95% | 99%	**Compelling** 95% | 99%	**Compelling** 99% | 9%	**Compelling** 99% | 99%	**Compelling** 99% | 99%
Nearest gene	**Very Strong** 70% | 90%	**Very Strong** 80% | 95%	**Extreme** 90% | 95%	**Compelling** 99% | 99%	**Compelling** 99% | 99%
Coding variant	**Strong** 50% | 85%	**Very Strong** 60% | 90%	**Very Strong** 75% | 95%	**Compelling** 95% | 99%	**Compelling** 99% | 99%
GWAS locus	**Moderate** 15% | 40%	**Moderate** 20% | 55%	**Moderate** 30% | 70%	**Very Strong** 75% | 95%	**Compelling** 99% | 99%
No evidence	**No evidence** 5% | 20%	**Anecdotal** 5% | 25%	**Moderate** 15% | 45%	**Strong** 50% | 85%	**Compelling** 95% | 99%

Rare Variation

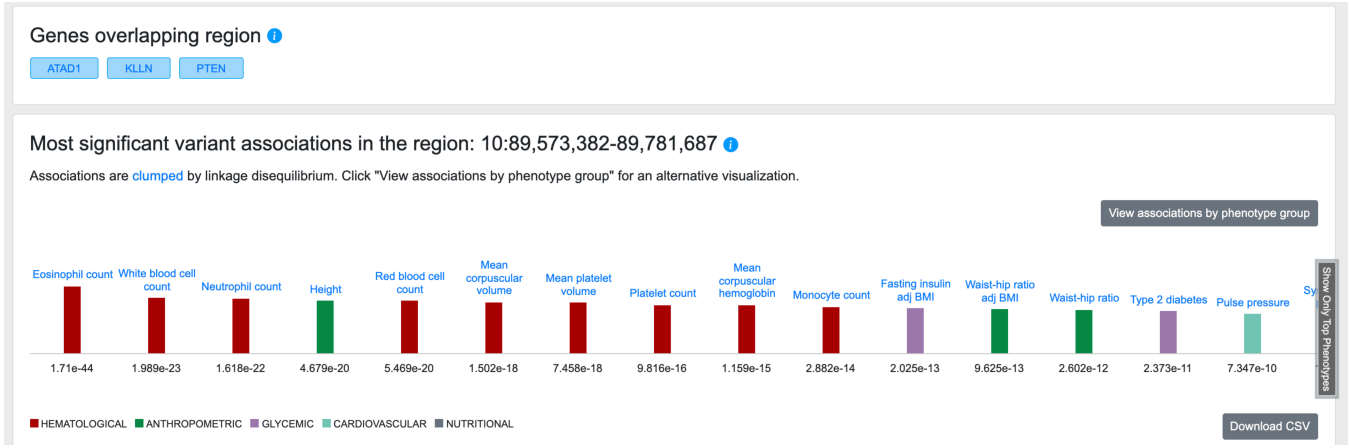
where the percentage values indicate posterior probabilities calculated using a 5% prior (left of vertical bar) and a 20% prior (right of vertical bar).

Real-world demonstration of the HuGE framework

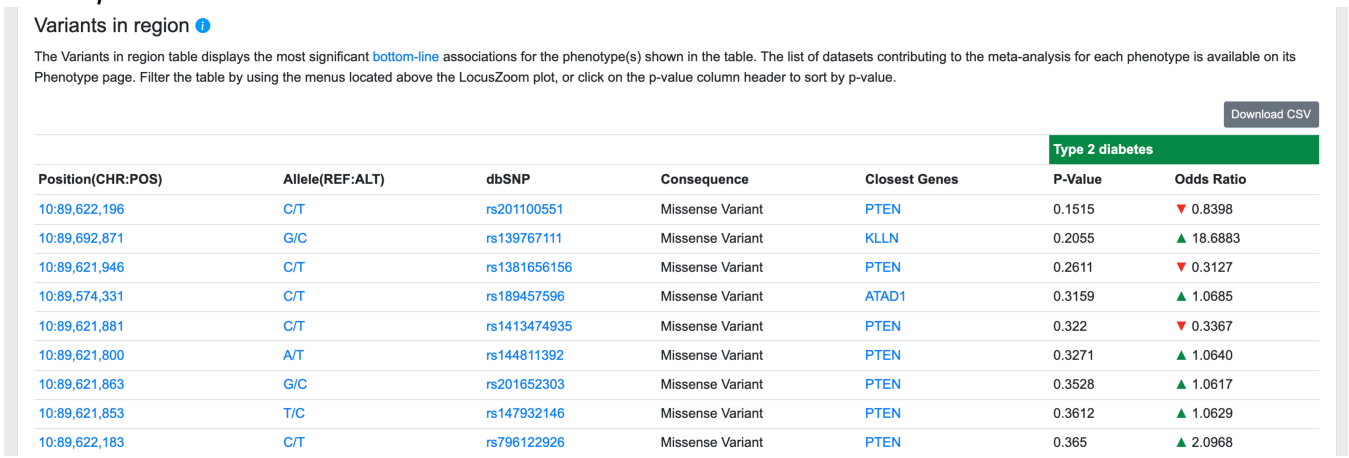
Here, we outline how the HuGE framework is implemented for *PTEN* and type 2 diabetes. We use the CMDKP for the analysis, although other resources could be used as well.

Step 1: Common variant association results for PTEN indicate 'nearest gene' evidence

A query of *PTEN* in the CMDKP shows a nearby genome-wide significant association of T2D:

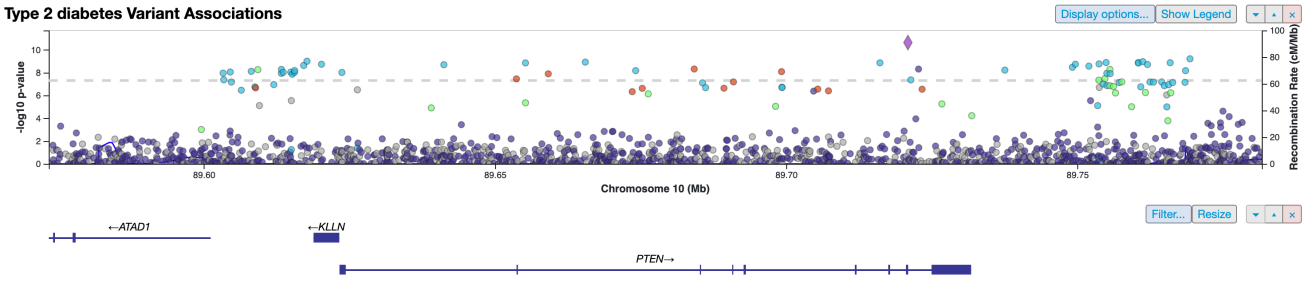


Filtering the associations to coding variants, no nonsynonymous variant is associated at $p \leq 5 \times 10^{-8}$:



					Type 2 diabetes	
Position(CHR:POS)	**Allele(REF:ALT)**	**dbSNP**	**Consequence**	**Closest Genes**	**P-Value**	**Odds Ratio**
10:89,622,196	C/T	rs201100551	Missense Variant	PTEN	0.1515	▼ 0.8398
10:89,692,871	G/C	rs139767111	Missense Variant	KLLN	0.2055	▲ 18.6883
10:89,621,946	C/T	rs1381656156	Missense Variant	PTEN	0.2611	▼ 0.3127
10:89,574,331	C/T	rs189457596	Missense Variant	ATAD1	0.3159	▲ 1.0685
10:89,621,881	C/T	rs1413474935	Missense Variant	PTEN	0.322	▼ 0.3367
10:89,621,800	A/T	rs144811392	Missense Variant	PTEN	0.3271	▲ 1.0640
10:89,621,863	G/C	rs201652303	Missense Variant	PTEN	0.3528	▲ 1.0617
10:89,621,853	T/C	rs147932146	Missense Variant	PTEN	0.3612	▲ 1.0629
10:89,622,183	C/T	rs796122926	Missense Variant	PTEN	0.365	▲ 2.0968

However, the most significant variant in the genomic locus is closest to *PTEN* as compared to other genes in the locus:



We thus assign *PTEN* to the “nearest gene” category.

Step 2: Rare variant association results for PTEN indicate ‘nominal’ evidence
Clicking on the *PTEN* gene to navigate to a new page in the CMDKP, and then scrolling down, shows a rare variant T2D association of p=0.0483.

Rare variant gene-level associations for PTEN

Phenotype	pValue	Beta	Odds Ratio	View
Waist-hip ratio	0.00799	▲0.0356		Masks + Plot
Type 2 diabetes	0.0483		▼0.5450	Masks + Plot

We thus assign *PTEN* to the “nominal” category.

Step 3: Combination of common and rare variation BF values indicate ‘extreme’ evidence
We next use **Figure 1** to map these two categories to the final HuGE score. This yields the “extreme” category of genetic support for *PTEN*, corresponding to a posterior of 90% (under a prior of 5%) or 95% (under a prior of 20%).

Limitations

Users of the HuGE framework must foremost understand that the guidelines above are simply a first step toward using human genetic data to evaluate experimentally identified genes. The framework is *based on estimates*, *not exact values*, and in the interest of accessibility it omits many sources of data that expert geneticists use to evaluate genes. It is *not* designed to result in precise posterior probability estimates.

As discussed in the manuscript, human genetics – like all scientific models – has caveats. For example, variants are inherited at birth and may not be a good proxy for perturbations introduced after the onset of diseases. Many disease genes may not exhibit associations, either because (by chance) no genetic variants that perturb the gene are observed in a population, or because statistical power is too low to detect associations.

Finally, as mentioned several times above, the current framework only provides support *for a gene, not against*. The primary reason for this limitation is based on the current state of human genetics – it is hard to identify variants that (with high confidence) perturb a gene such that the lack of an association implies a lack of disease relevance. As sample sizes increase, we will be better able to identify “human knockouts” which may enable us to determine true “evidence from absence” of a gene’s involvement in disease.

References

Flannick, J., Mercader, J.M., Fuchsberger, C., Udler, M.S., Mahajan, A., Wessel, J., Teslovich, T.M., Caulkins, L., Koesterer, R., Barajas-Olmos, F., et al. (2019). Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. Nature *570*, 71-76.

Jeffreys, H. (1961). Theory of probability Clarendon Press. Oxford.

Mahajan, A., Taliun, D., Thurner, M., Robertson, N.R., Torres, J.M., Rayner, N.W., Payne, A.J., Steinthorsdottir, V., Scott, R.A., Grarup, N., et al. (2018a). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nature genetics *50*, 1505-1513.

Mahajan, A., Wessel, J., Willems, S.M., Zhao, W., Robertson, N.R., Chu, A.Y., Gan, W., Kitajima, H., Taliun, D., Rayner, N.W., et al. (2018b). Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. Nature genetics *50*, 559-571.

Satterstrom, F.K., Kosmicki, J.A., Wang, J., Breen, M.S., De Rubeis, S., An, J.Y., Peng, M., Collins, R., Grove, J., Klei, L., et al. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell *180*, 568-584.e523.

Stacey, D., Fauman, E.B., Ziemek, D., Sun, B.B., Harshfield, E.L., Wood, A.M., Butterworth, A.S., Suhre, K., and Paul, D.S. (2019). ProGeM: a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. Nucleic acids research *47*, e3.

Wakefield, J. (2008). Reporting and interpretation in genome-wide association studies. International journal of epidemiology *37*, 641-653.