Integration of data from the human GWAS Catalog into the Rat Genome Database

Jennifer R Smith[1], Logan Lamers[1], Stanley JF Laulederkind[1], G Thomas Hayman[1], Shur-Jen Wang[1], Monika Tutaj[1], Mary L Kaldunski[1], Mahima Vedi[1], Wendy M Demos[1], Marek A Tutaj[1], Jyothi Thota[1], Adam C Gibson[1], Akhilanand Kundurthi[1], Varun Reddy Gollapally[1], Kent C Brodie[2], Stacy Zacher[3], Jeffrey L De Pons[1], Melinda R Dwinell[1], Anne E Kwitek[1]

[1]Rat Genome Database, Department of Physiology, [2]Clinical and Translational Science Institute, [3]Finance and Administration, Medical College of Wisconsin, Milwaukee, WI, USA.

https://rgd.mcw.edu

V = Variant rs12720356
Q = QTL GWAS681336_H

Abstract:

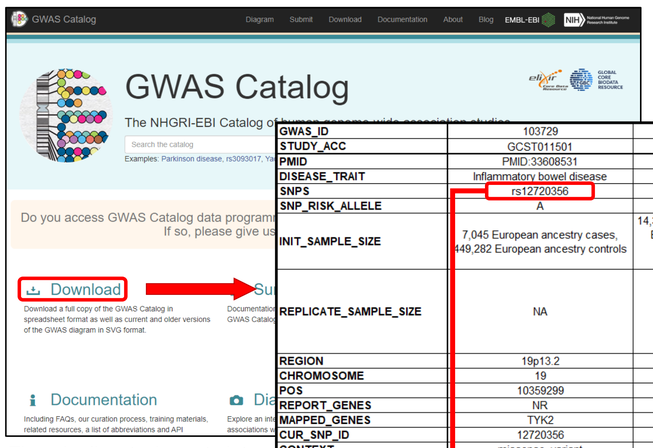
The Rat Genome Database (RGD, https://rgd.mcw.edu), the premiere online resource for genetic, genomic, disease and phenotype data for rat, also integrates a substantial corpus of data for human and other mammalian species to facilitate comparative studies. For researchers interested in cross-species studies of disease- and phenotype-related genomic regions, RGD has numerous resources including standardized disease annotations across multiple species; rat, mouse and human quantitative trait loci (QTL) associated with disease, phenotype, trait and measurement term; and human disease- and phenotype-associated variants imported from NCBI's ClinVar database. Now RGD has incorporated human variants and the associated phenotypes from the NHGRI-EBI GWAS Catalog.

All data available at GWAS Catalog for over 676,000 human GWAS results were loaded as is into a "holding" table in the RGD database. These data were filtered based on whether a risk allele was specified. SNPs with a risk allele different than reference at that location were loaded as variant records at RGD. In addition, GWAS QTL records with associated p-values for the listed trait(s) were created for all SNPs regardless of risk allele status.

GWAS Catalog records are associated with terms from the Experimental Factor Ontology (EFO) to denote the phenotypic trait that was studied. To make these data consistent with RGD's existing annotations for genes, QTL, strains, and human variants, EFO terms were matched to terms in the disease, phenotype, trait and measurement vocabularies already in use. Based on the associated EFO terms for each result, annotations to these ontologies were created. Both the original EFO term associations and the derived annotations for other ontologies are displayed on the corresponding variant and QTL report pages at RGD.

These data are discoverable at RGD by searching for terms in any of the mentioned ontologies, using "search by position" for QTLs, and using RGD's Variant Visualizer tool to search for GWAS Catalog SNPs. In addition, variants and QTL that overlap a gene are listed on that RGD gene page. Likewise, the variant and QTL report pages provide information about overlapping gene(s), GWAS QTL defined by the same peak SNP, and any phenotypic or GWAS QTL that overlap the focus variant or QTL.

The incorporation of GWAS Catalog variants has substantially expanded the corpus of human genotype-to-phenotype data at RGD. Future work will include improvements to RGD's website and tools to make the data more findable and usable. Also, to support comparative genomics analyses, RGD is actively looking to integrate comparable GWAS datasets for rat.



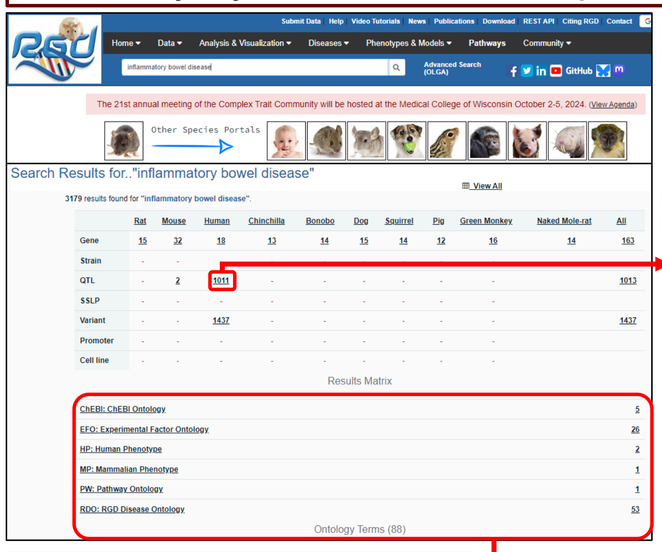
The pipeline to import data from GWAS Catalog is run on a regular basis to keep the information in RGD up-to-date

More than 676,000 human GWAS results were downloaded from the NHGRI-EBI GWAS Catalog and inserted as is into a table in the RGD database, including the trait(s) analyzed, population studied, associated position and SNP ID(s), and any information given about the risk allele, its allele frequency and the p-value. When variant and GWAS QTL records are created at RGD those IDs are entered into the corresponding rows in the GWAS_Catalog table.

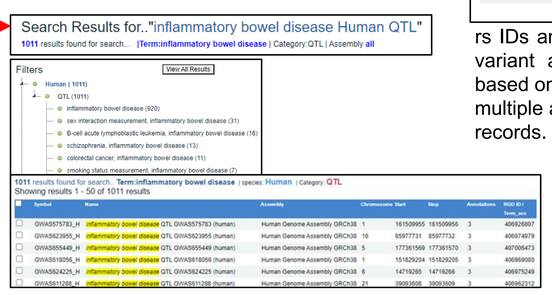
GWAS_ID	103729	1269241	170104	170029	144444
STUDY_ACC	GCST011501	GCST90292538	GCST003043	GCST003044	GCST000879
PMID	PMID:33608531	PMID:37156999	PMID:26192919	PMID:26192919	PMID:21102463
DISEASE_TRAIT	Inflammatory bowel disease	Inflammatory bowel disease	Inflammatory bowel disease	Crohn's disease	Crohn's disease
SNPS	rs12720356	rs12720356	rs12720356	rs12720356	rs12720356
SNP_RISK_ALLELE	A	?	C	C	G
INIT_SAMPLE_SIZE	7,045 European ancestry cases, 449,282 European ancestry controls	14,393 East Asian ancestry cases, 15,456 East Asian ancestry controls, 30,713 European ancestry cases, 338,106 European ancestry controls	12,882 European ancestry cases, 21,770 European ancestry controls	5,956 European ancestry cases, 14,927 European ancestry controls	6,333 European ancestry cases, 15,056 European ancestry controls
REPLICATE_SAMPLE_SIZE	NA	NA	25,273 European ancestry cases, 26,715 European ancestry controls, 548 Iranian ancestry cases, 342 Iranian ancestry control, 1,423 Indian ancestry cases, 990 Indian ancestry controls, 2,824 East Asian ancestry cases, 3,719 East Asian ancestry controls	14,594 European ancestry cases, 26,715 European ancestry controls, 151 Iranian ancestry cases, 342 Iranian ancestry controls, 184 Indian ancestry cases, 990 Indian ancestry controls, 1,690 East Asian ancestry cases, 3,719 East Asian ancestry controls	15,694 European ancestry cases, 14,026 European ancestry controls, 414 European ancestry trios
REGION	19p13.2	19p13.2	19p13.2	19p13.2	19p13.2
CHROMOSOME	19	19	19	19	19
POS	10359299	10359299	10359299	10359299	10359299
REPORT_GENES	NR	(null)	NR	NR	ICAM3, TYK2, ICAM1
MAPPED_GENES	TYK2	TYK2	TYK2	TYK2	TYK2
CUR_SNP_ID	12720356	12720356	12720356	12720356	12720356
CONTEXT	missense_variant	missense_variant	missense_variant	missense_variant	missense_variant
RISK_ALLELE_FREQ	0.9	NR	0.08572	0.08572	0.084
P_VALUE	2.00E-08	3.00E-16	4.00E-16	3.00E-11	1.00E-12
P_VALUE_MLOG	7.698970004	15.52287875	15.39794001	10.52287875	12
SNP_PASSING_QC	NR [8546065] (imputed)	Illumina [Illumina ImmunoChip| Japonica Array V1] (imputed)	Affymetrix, Illumina [~ 9000000] (imputed)	Affymetrix, Illumina [~ 9000000] (imputed)	Affymetrix, Illumina [953241] (imputed)
MAPPED_TRAIT	inflammatory bowel disease	inflammatory bowel disease	inflammatory bowel disease	Crohn's disease	Crohn's disease
EFO_IDS	EFO_0003767	EFO_0003767	EFO_0003767	EFO_0000384	EFO_0000384
OR_OR_BETA	0.86	0.1472	1.1598983	1.154103	1.12
VARIANT_RGD_ID	(null)	(null)	11543250	11543250 V	151295804
QTL_RGD_ID	406987872	407255091	407032360 Q	407030681	407019813

RGD does not create variant records when the risk allele is the same as the reference nucleotide or when the risk allele is not reported [see "(null)" for VARIANT_RGD_ID in the table above]. However, QTL records are created regardless of the risk allele reported [QTL_RGD_ID is filled]. This means that some QTL records will show the reference allele or a question mark for the associated risk allele.

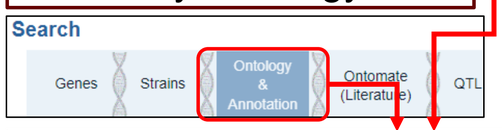
Search by Symbol/Name/Description



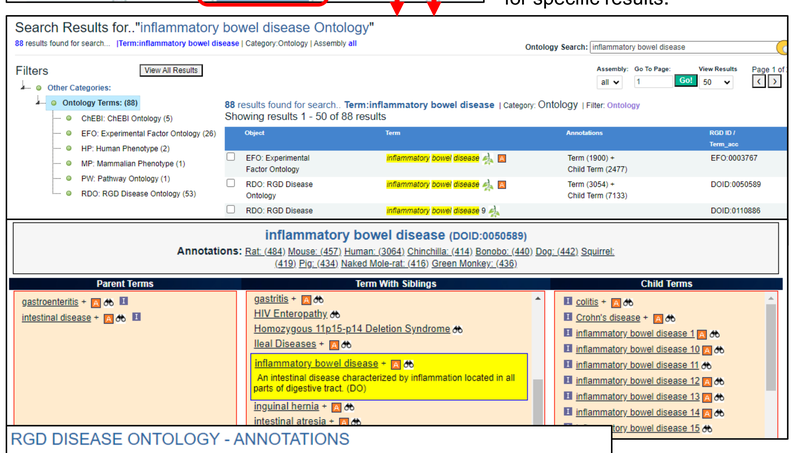
Use the keyword search at the top of most RGD pages to query across data types and species. For genes, QTL, etc. the symbol, name, description. synonyms and internal and external database IDs are searched. For ontologies, the terms, descriptions and synonyms are used. Click on a number in the upper table to see all of the results for that data type and species.

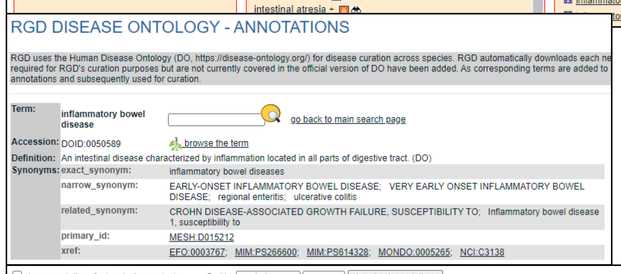


Search by Ontology Term

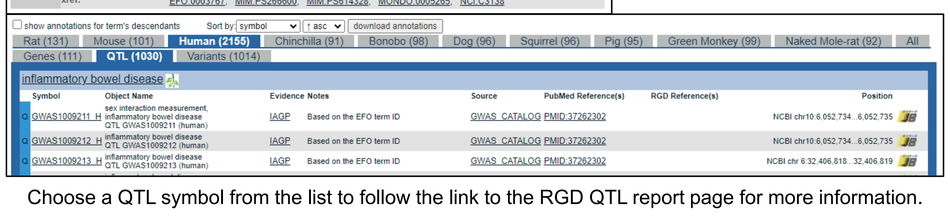


Search RGD's manual and imported functional annotations using either the keyword search or the ontology search. Results for all vocabularies are returned. Choose an ontology for specific results.



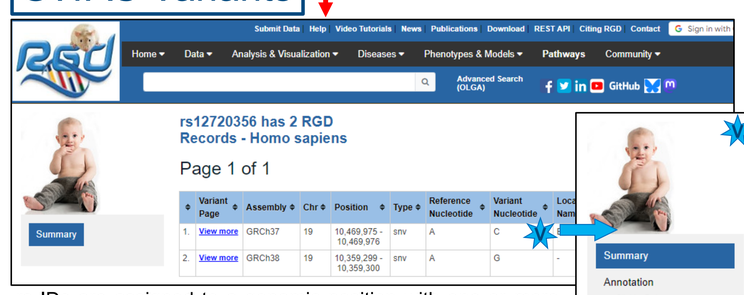


The leaf icon links to RGD's ontology browser showing the term of interest with the related terms that are more general to the left and more specific to the right. The "A" icon links to the ontology report page with the list of all, in this case, genes, QTL and variants annotated to the term.



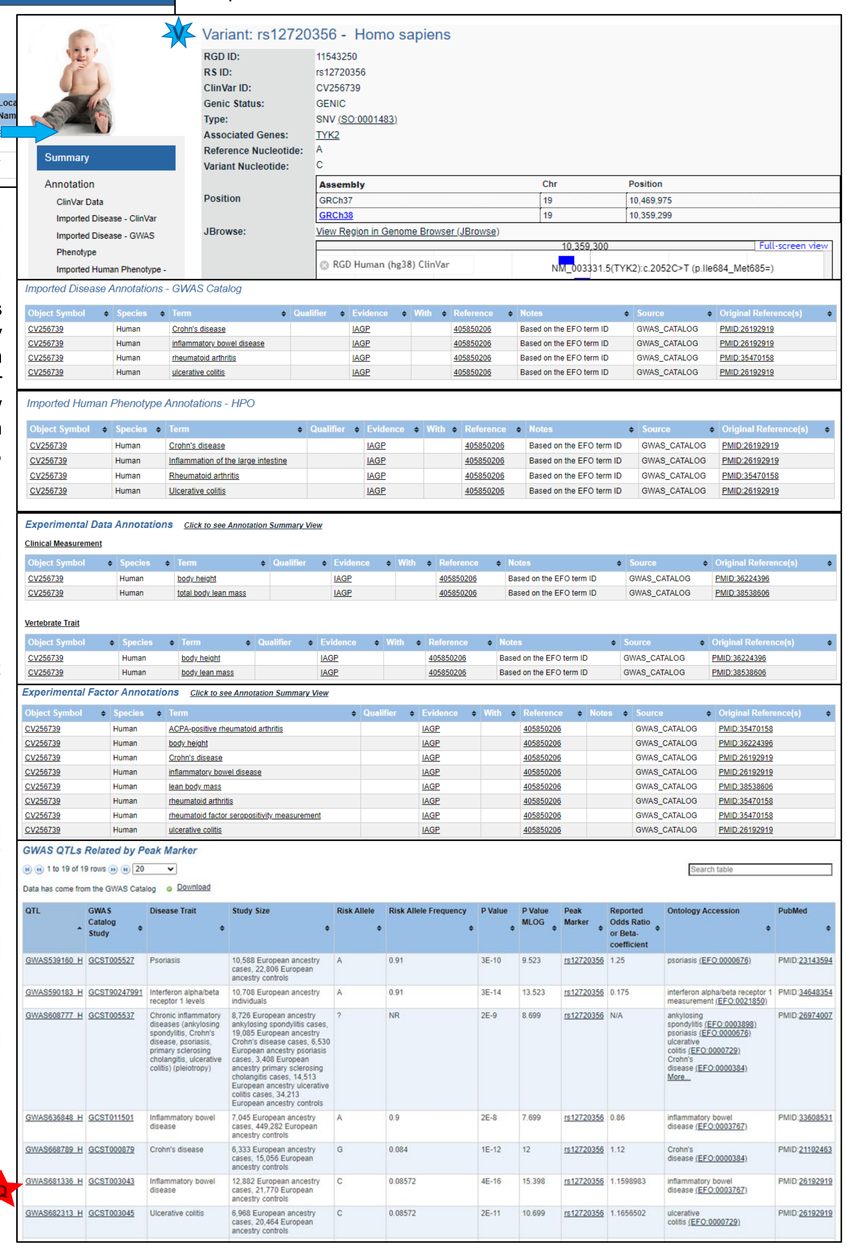
Choose a QTL symbol from the list to follow the link to the RGD QTL report page for more information.

GWAS Variants



rs IDs are assigned to a genomic position with one or more variant alleles associated. RGD creates variant records based on a position and a single variant allele. If an rs ID has multiple alleles, a search for the rs ID will return multiple RGD records.

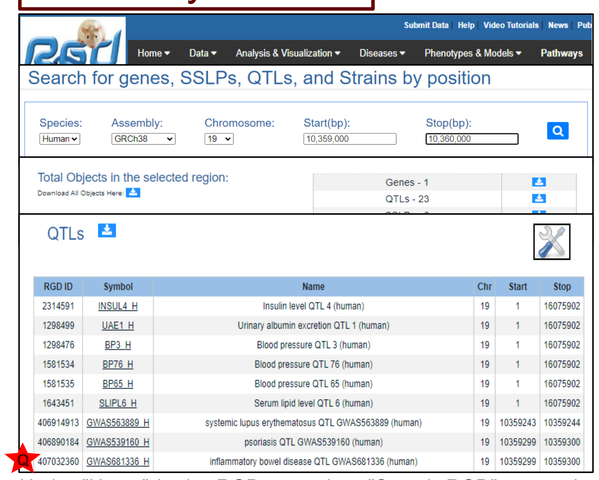
GWAS Catalog links variants to phenotypes using the Experimental Factor Ontology (EFO). To integrate these with other data in RGD, mappings between EFO and other ontologies used at RGD were created to allow GWAS variants and QTL to be annotated with RGD's vocabularies. For more information, see the poster by Laulederkind et al.



RGD assigns all applicable disease, phenotype, trait and measurement terms to each variant based on all GWAS Catalog results naming that variant. The original EFO terms provided are used to create EFO annotations in RGD.

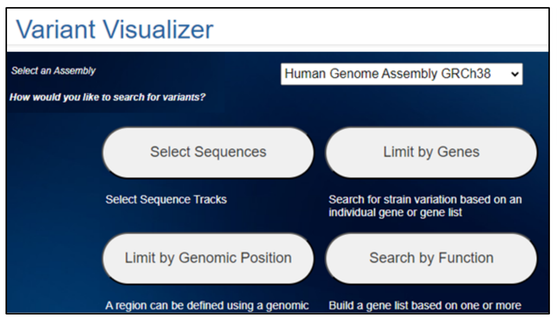
The variant page lists all RGD QTL records for which that rs ID is listed as the SNP. The rs ID is listed as the peak marker for the QTL and information about the trait(s), population(s) studied, risk allele, p-value and original reference are shown. Links are provided to the RGD QTL record, the study at GWAS Catalog, and more.

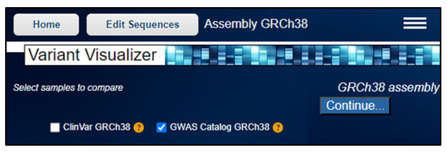
Search by Position

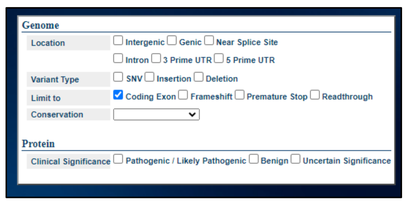


Under "Home" in the RGD menu bar, "Search RGD" opens the search by position menu. Search for specific data types by position using the options in the "Data" menu.

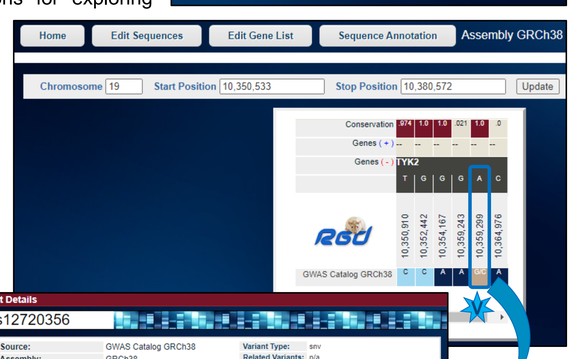
Variant Visualizer Search



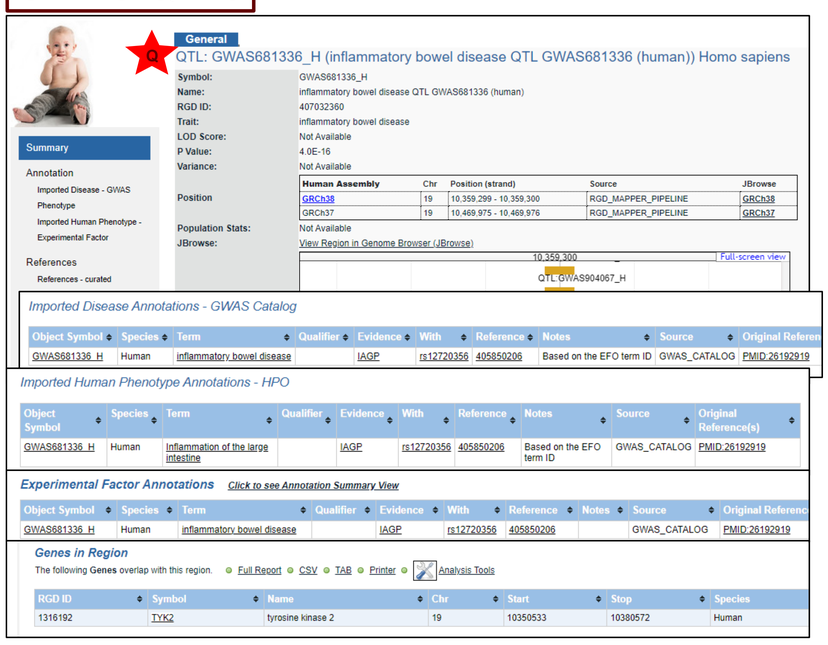


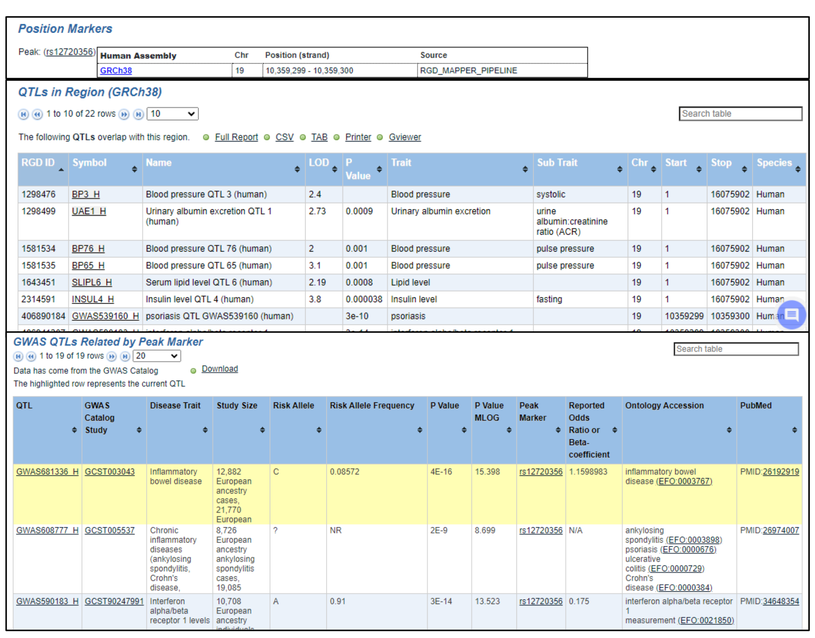


Variant Visualizer, RGD's "advanced search" for variants provides many options for exploring variant data in RGD. Begin by selecting the species and assembly. Under "Select sequences" choose the GWAS Catalog then filter using a genomic position or gene list. If desired, further filter by the type of variant. All variants matching the search criteria will be shown. Click a square in the variant display to see more details about that variant, including associated amino acid changes for coding variants.



GWAS QTLs





Each QTL report page displays all of the data in RGD associated with that QTL. The position is assigned based on the position of the GWAS variant marker. Since the QTL is defined as the association between a position and a trait, only annotations derived from the EFO associations for the corresponding GWAS result are made to the QTL. The "QTLs in Region" section contains both overlapping pQTLs and GWAS QTLs. "GWAS QTLs Related by Peak Marker" contains all GWAS QTLs that share the same rs ID.

RGD is funded by the National Heart, Lung, and Blood Institute (NHLBI; R01HL064541), and the National Human Genome Research Institute (NHGRI) as part of the Alliance of Genome Resources (U24HG010859).

ALLIANCE of GENOME RESOURCES FOUNDING MEMBER
GLOBAL CORE BIODATA RESOURCE