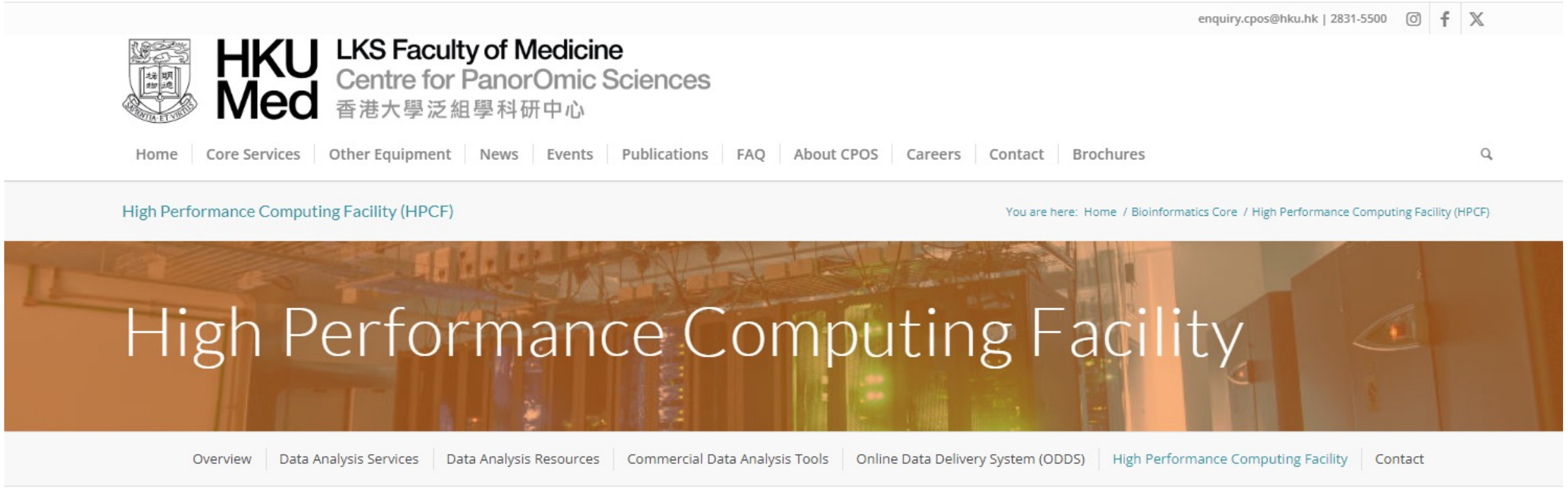
Public bioinformatics resources for genomic analysis interpretation

Clara Tang

“Public” resources in HKU : HPCF

- CPOS HPCF



“Public” resources in HKU : HPCF

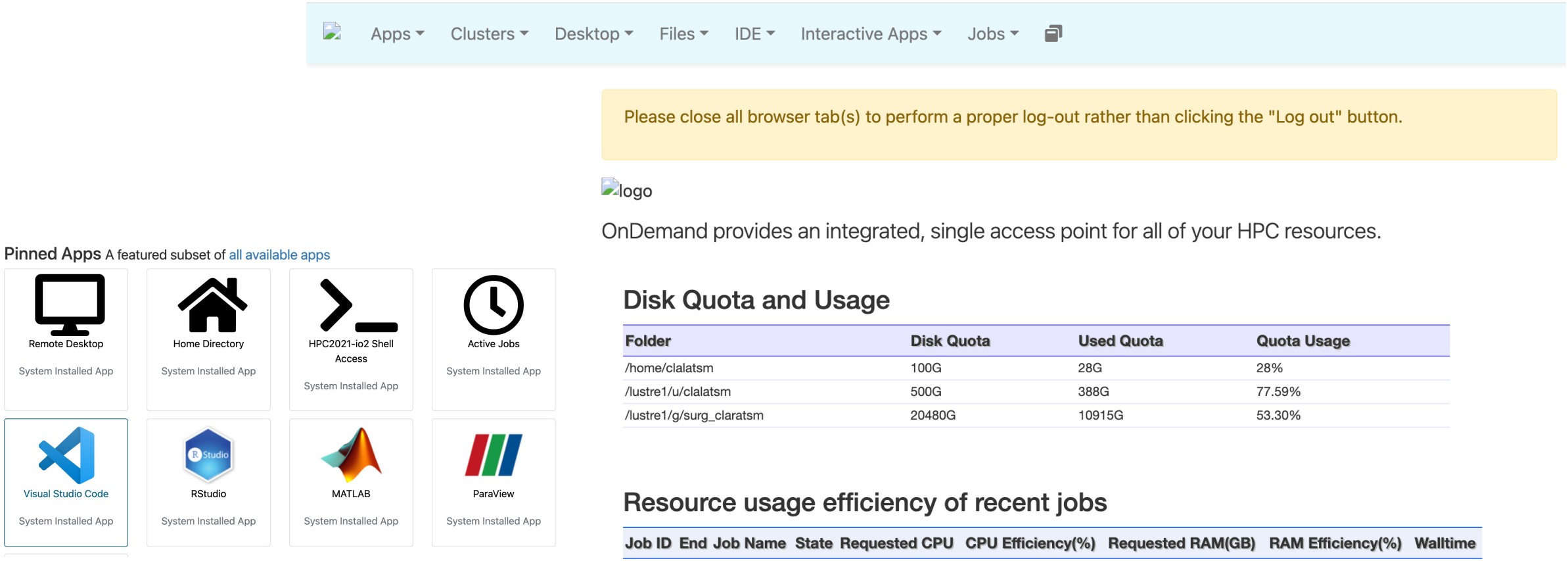
- CPOS HPCF

General Job Queues

Queue name	Max no of processors per job	Max memory (GB) per job	Max no of job running per user	Max no of job queuing per user	Max Walltime (hr)
small	2	10	18	40	6
small_ext	2	10	6	12	60
medium	12	50	12	25	24
medium_ext	12	50	6	8	60
large	12	120	3	4	84
legacy	12	45	8	16	96
test	24	190	1	1	1

“Public” resources in HKU : HPCF

- ITS HPC2021 (https://hpc.hku.hk/guide/hpc-one-web-portal/)



“Public” resources in HKU : HPCF

- ITS HPC2021 (https://hpc.hku.hk/guide/hpc-one-web-portal/)

Storage Type	$HOME	$WORK	PI Group Share	$TMP_DIR
Path	`/home/$USER`	`/lustre1/u/$USER`	`/lustre1/g/$PI_GROUP`	`/tmp`
Usage	Long term, small size	Short term, high performance	Moderate term, high performance, shared between members in a research group	Short term (for the duration a job is being executed), high performance
Availability	Available across any nodes in HPC2021	Available across any nodes in HPC2021	Available across any nodes in HPC2021	Available on the attached node only
Capacity	100GB per user	500GB per user	• 5TB (Default) per PI group • Up to 15TB at no charge • Above 15TB may be charged at standard storage rates	• ~400GB per general purpose compute node • 8.6TB per GPU node • 11TB per large memory node
Performance	Moderate – Not appropriate for workload requiring high throughput or small file operations	High throughput for large files and IO: • Good for sequential access to files of moderately large size (>1MB) • Not so good for small or random file access (e.g source code, scripts, software, temporary files)	High throughput for large files and IO: • Good for sequential access to files of moderately large size (>1MB) • Not so good for small or random file access (e.g source code, scripts, software, temporary files)	High performance, especially in terms of operation of small files
Clean-up Policy	No scheduled clean-up	Files not accessed in the past 60 days are subject to clean-up by system (Be reminded to move important data to $HOME)	No scheduled clean-up	Cleaned-up immediately upon job termination

“Public” resources in HKU : HPCF

- ITS HPC2021

Login node	Role	Usage
hpc2021.hku.hk	Head Node	Reserved for file editing, compilation and **job submission/management**
hpc2021-io1.hku.hk hpc2021-io2.hku.hk	IO Node	Reserved for **file transfer**, file management and data analysis/visualization

For HPC2021 System

QoS	Supported Partition(s)	Max Job Duration	Max Resources per job
debug	intel, amd, gpu	30min	2 nodes, 2 GPUs
normal (default)	intel, amd	1 Week	1024 cores
long	intel, amd	2 Weeks	1 node
^ special	intel, amd	1 Day	2048 cores
^ gpu	gpu	1 Week	1 node, 4 GPUs
^ hugemem	hugemem	1 Week	1 node, 2TB RAM

For HPC2021 System

Partition	Default / Max Job duration	# of nodes	cores per node	RAM(GB) per node	RAM(GB) per core	Features
intel (default)	1 Day / 1 Week	84	32	192	6	GOLD6626R
amd	1 Day / 1 Week	28	64	256	4	EPYC7542
		28	128	512	4	EPYC7742
		10	192	768	4	EPYC9654
gpu	1 Day / 1 Week	4	32	384	12	4x V100
		3	32	384	12	8x V100
		2	64	512	8	8x L40S
hugemem	1 Day / 1 Week	2	128	2048	16	EPYC7742 + 2TB RAM

Motivation

- Greatest challenge for genetic association studies
  - Assign functionality to associated SNPs and genes
  - Predict the most likely genes and variants driving the phenotype associations
- Greatest challenge for rare variant screening
  - Predict the deleterious effect and causality
- Need to integrate as much information as possible through genome annotation tools to drive the experimental validation

Publicly available bioinformatics resources

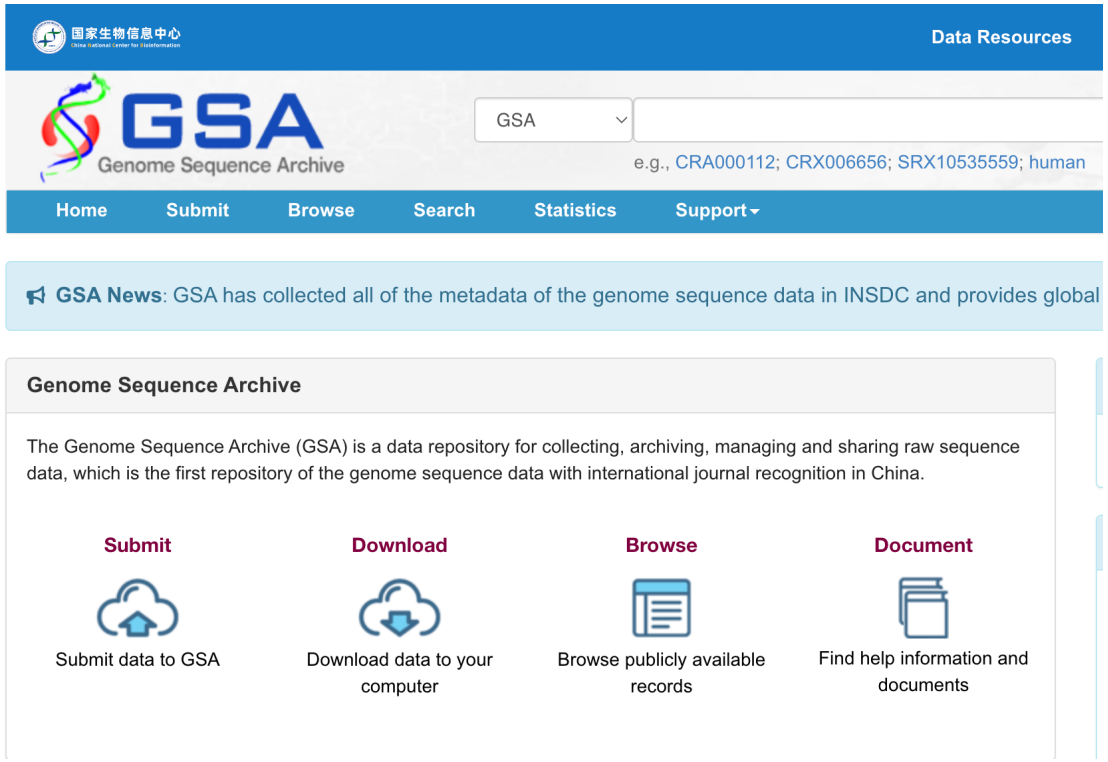
- Genome Browsers
- Genetic Variations Databases
- Human Traits & Diseases Databases
- Nucleotide Sequence Databases
- Gene Expression Databases
- Protein Sequence & domains: Databases and Search Tools
- Phylogeny & Taxonomy
- Databases of other organisms
- Gene Prediction
- Metabolic, Gene Regulatory & Signal Transduction Network Databases
- Publications Database

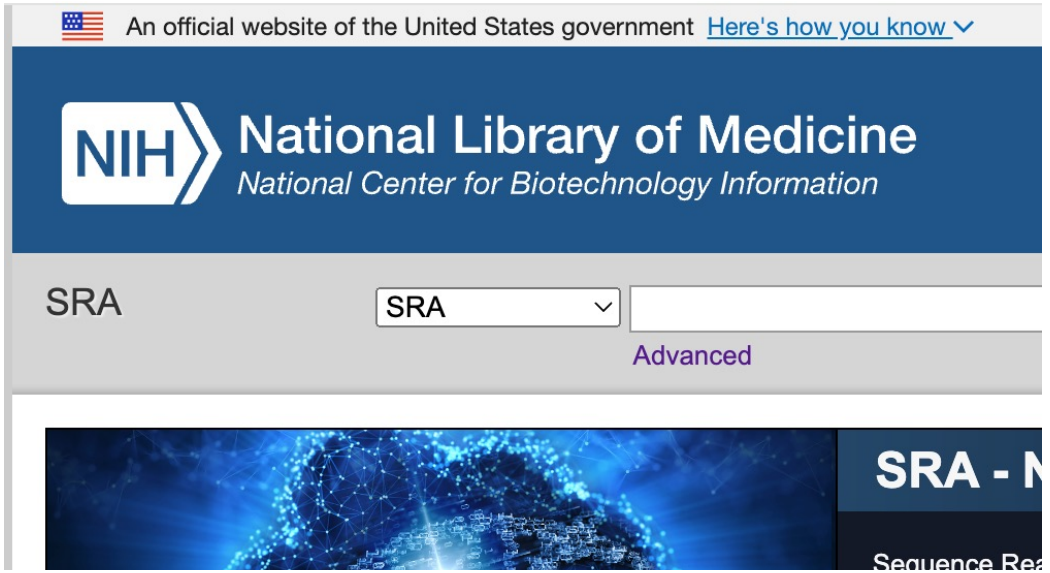
Publicly available bioinformatics resources

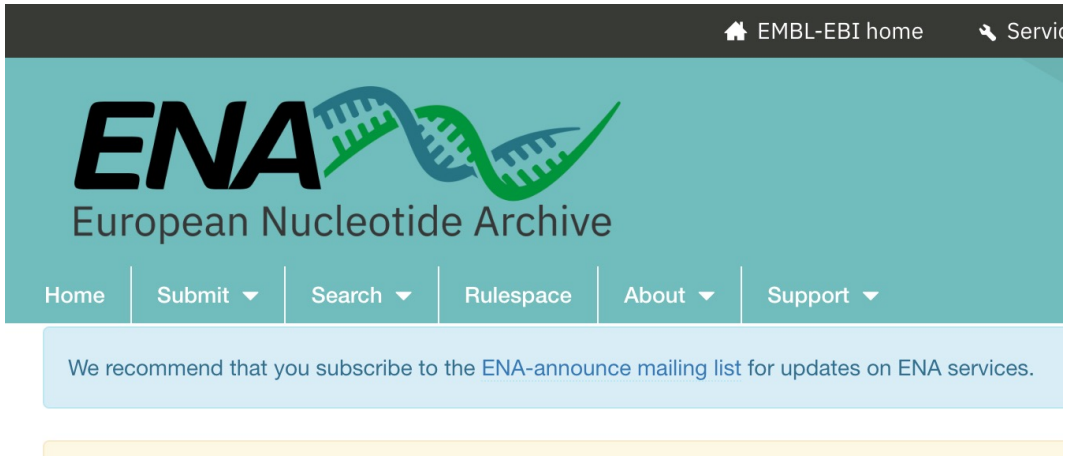
- Genome Browsers
- Genetic Variations Databases
- Human Traits & Diseases Databases
- Nucleotide Sequence Databases
- Gene Expression Databases
- Protein Sequence & domains: Databases and Search Tools
- Phylogeny & Taxonomy
- Databases of other organisms
- Gene Prediction
- Metabolic, Gene Regulatory & Signal Transduction Network Databases
- Publications Database

Nucleotide Sequence Databases

- NCBI SRA, EBI ENA, GSA, and DDBJ







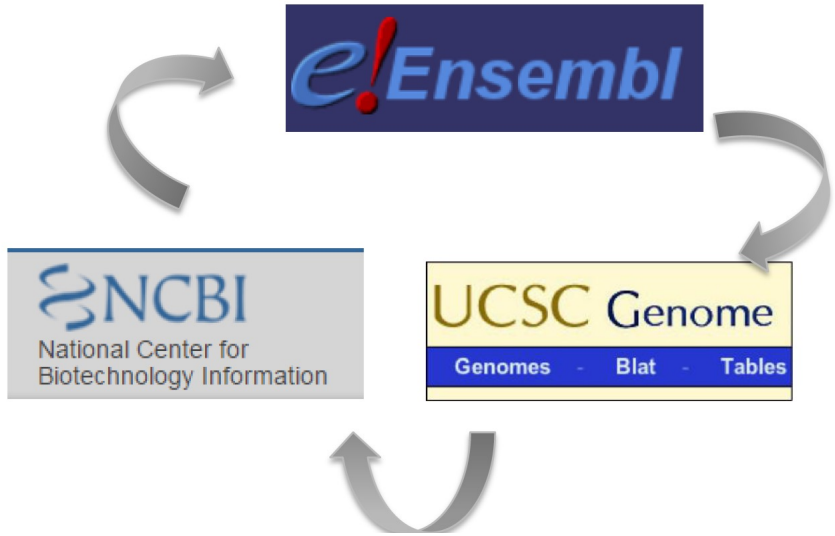
Publicly available bioinformatics resources

- Genome Browsers
- Genetic Variations Databases
- Human Traits & Diseases Databases
- Nucleotide Sequence Databases
- Gene Expression Databases
- Protein Sequence & domains: Databases and Search Tools
- Phylogeny & Taxonomy
- Databases of other organisms
- Gene Prediction
- Metabolic, Gene Regulatory & Signal Transduction Network Databases
- Publications Database

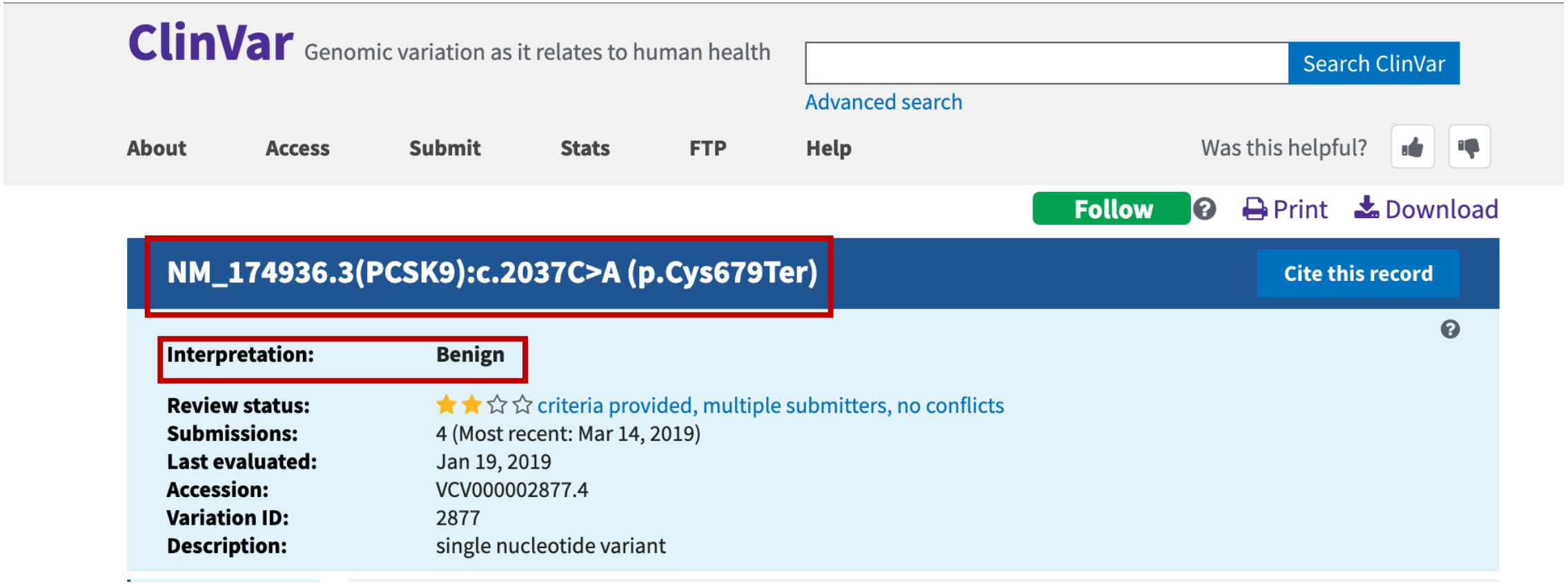
Genome browsers

More than a browser!

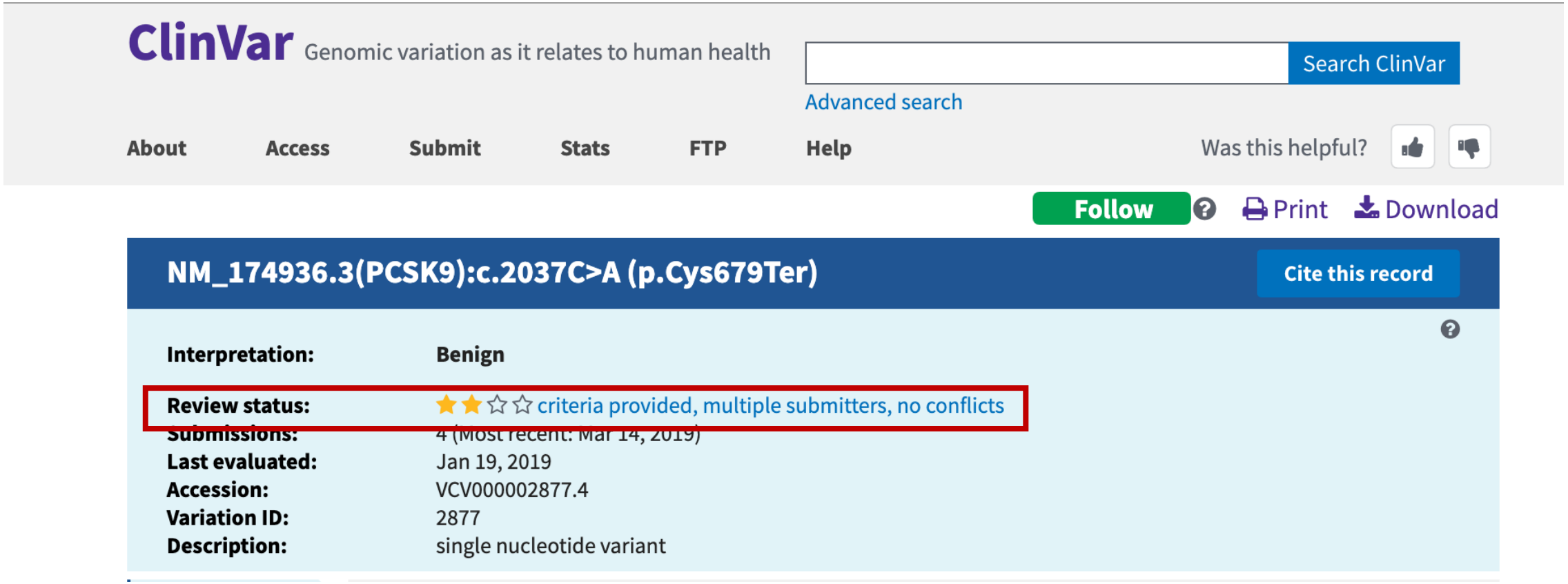
- NCBI: https://www.ncbi.nlm.nih.gov/
- EMBL-EBI Ensembl : http://www.ensembl.org
- UCSC: http://genome.ucsc.edu/



Genetic variation databases - ClinVar



Genetic variation databases - ClinVar



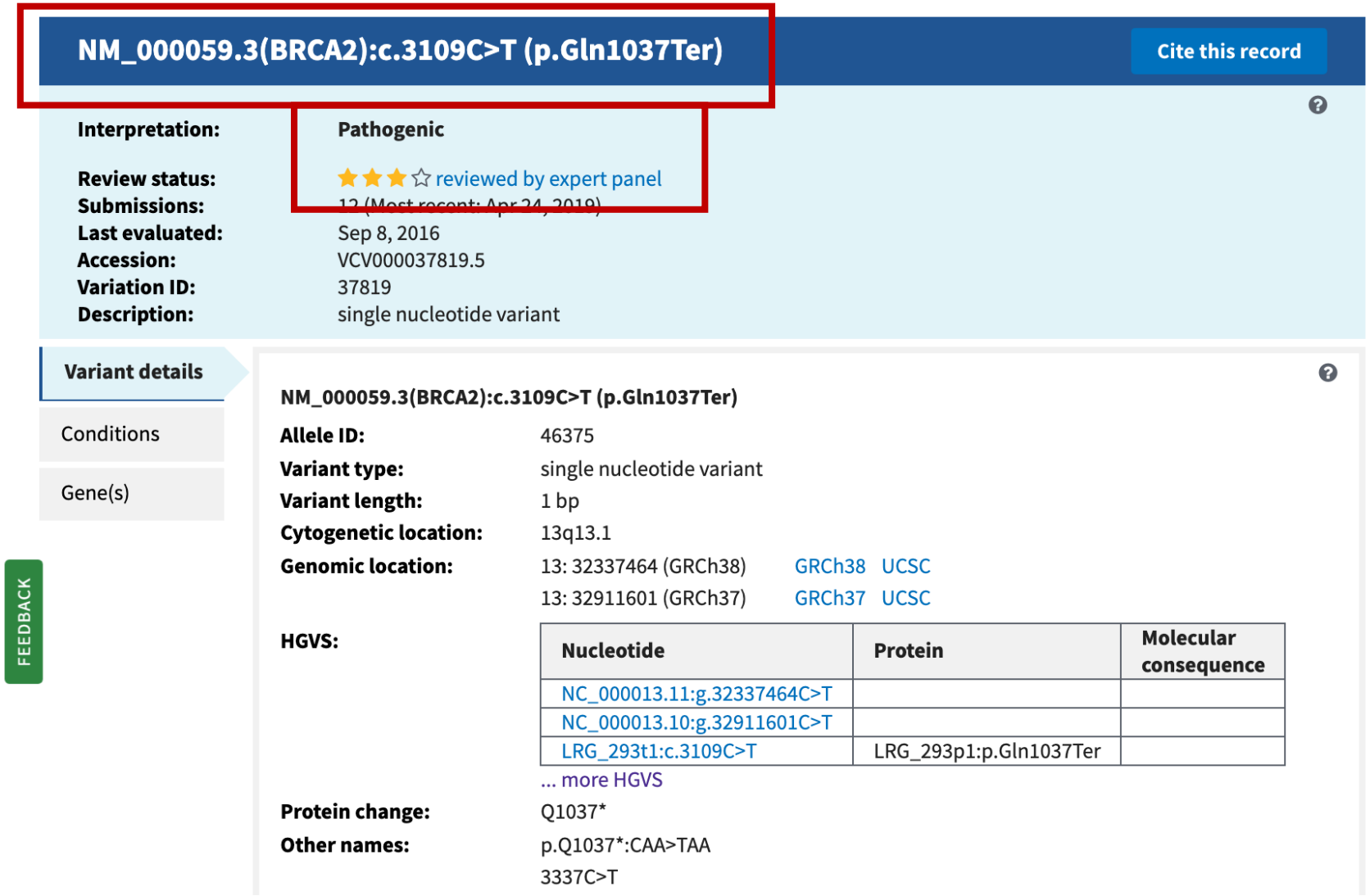
Genetic variation databases - ClinVar

Submitted interpretations and evidence

Interpretation (Last evaluated)	Review status (Assertion criteria)	Condition (Inheritance)	Submitter	Supporting information (See all)
Benign (Jan 19, 2019)	criteria provided, single submitter (Nykamp K et al. (Genet Med 2017)) Method: clinical testing	not provided Allele origin: germline	Invitae Accession: SCV000644868.3 Submitted: (Mar 14, 2019)	Evidence details
Benign (May 24, 2017)	criteria provided, single submitter (ACMG Guidelines, 2015) Method: clinical testing	Familial hypercholesterolemias Allele origin: germline	Color Accession: SCV000902915.1 Submitted: (Nov 06, 2018)	Evidence details
association (Mar 23, 2006)	no assertion criteria provided Method: literature only	LOW DENSITY LIPOPROTEIN CHOLESTEROL LEVEL QUANTITATIVE TRAIT LOCUS 1 Allele origin: germline	OMIM Accession: SCV000023169.2 Submitted: (Dec 30, 2010)	Evidence details Publications PubMed (2)

FEEDBACK

Genetic variation databases - ClinVar



NM_000059.3(BRCA2):c.3109C>T (p.Gln1037Ter)

Cite this record

Interpretation: Pathogenic

Review status: ★★★☆ reviewed by expert panel
Submissions: 12 (Most recent: Apr 24, 2019)
Last evaluated: Sep 8, 2016
Accession: VCV000037819.5
Variation ID: 37819
Description: single nucleotide variant

Variant details
Conditions
Gene(s)

FEEDBACK

NM_000059.3(BRCA2):c.3109C>T (p.Gln1037Ter)

Allele ID: 46375
Variant type: single nucleotide variant
Variant length: 1 bp
Cytogenetic location: 13q13.1
Genomic location: 13: 32337464 (GRCh38) GRCh38 UCSC
13: 32911601 (GRCh37) GRCh37 UCSC

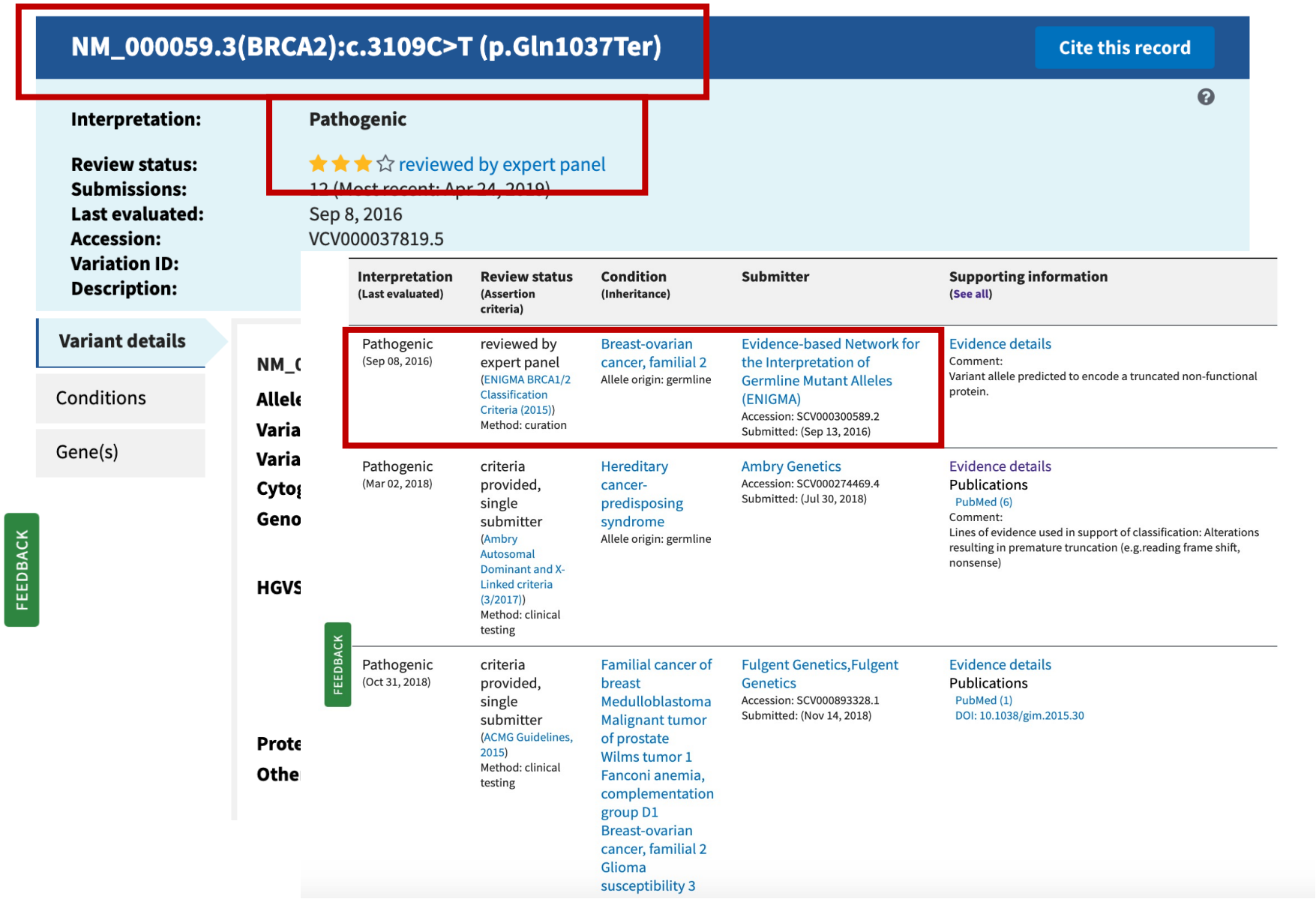
HGVS:

Nucleotide	Protein	Molecular consequence
NC_000013.11:g.32337464C>T		
NC_000013.10:g.32911601C>T		
LRG_293t1:c.3109C>T	LRG_293p1:p.Gln1037Ter	

... more HGVS

Protein change: Q1037*
Other names: p.Q1037*:CAA>TAA
3337C>T

Genetic variation databases - ClinVar



NM_000059.3(BRCA2):c.3109C>T (p.Gln1037Ter)

Cite this record

Interpretation: Pathogenic

Review status: ★★★☆ reviewed by expert panel

Submissions: 12 (Most recent: Apr 24, 2019)

Last evaluated: Sep 8, 2016

Accession: VCV000037819.5

Variation ID:

Description:

Variant details

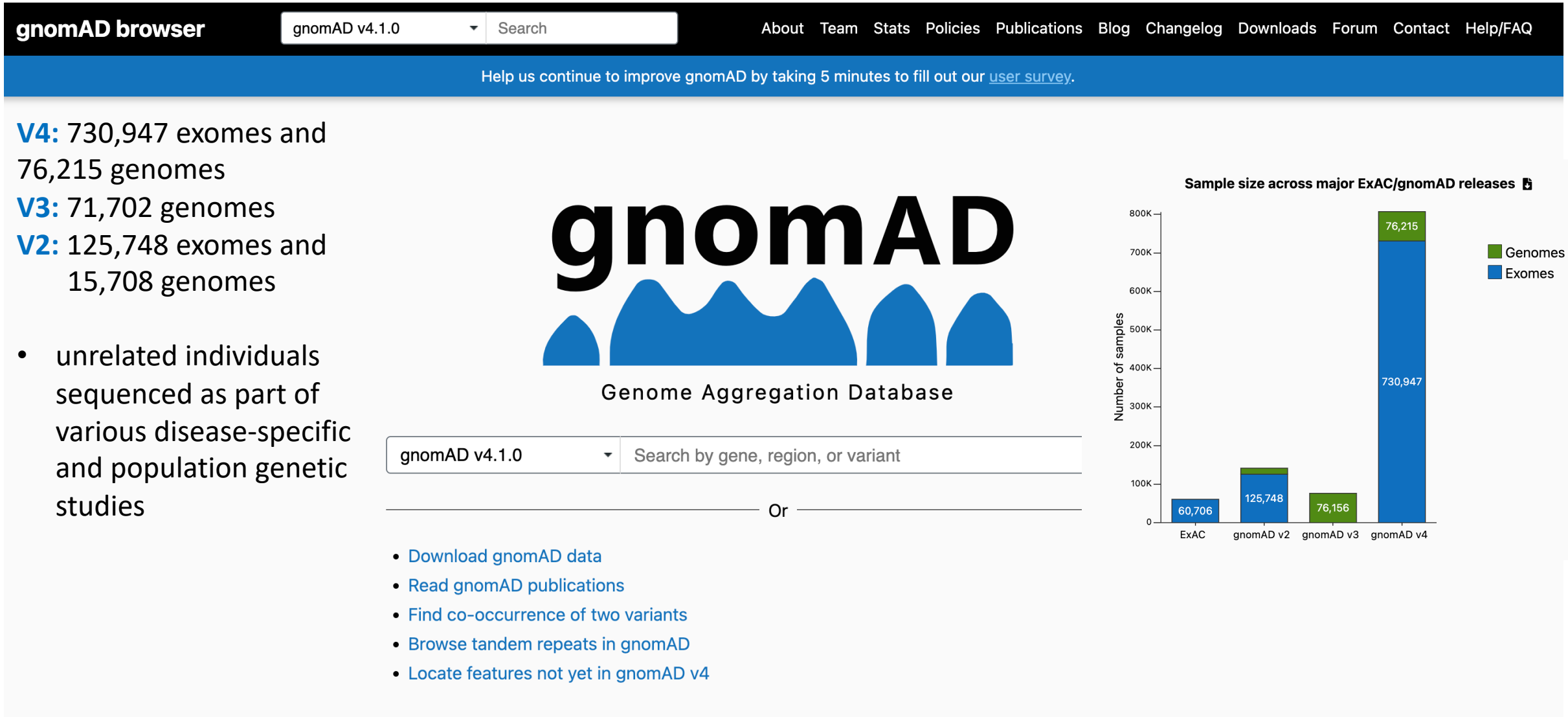
Conditions

Gene(s)

FEEDBACK

Interpretation (Last evaluated)	Review status (Assertion criteria)	Condition (Inheritance)	Submitter	Supporting information (See all)
Pathogenic (Sep 08, 2016)	reviewed by expert panel (ENIGMA BRCA1/2 Classification Criteria (2015)) Method: curation	Breast-ovarian cancer, familial 2 Allele origin: germline	Evidence-based Network for the Interpretation of Germline Mutant Alleles (ENIGMA) Accession: SCV000300589.2 Submitted: (Sep 13, 2016)	Evidence details Comment: Variant allele predicted to encode a truncated non-functional protein.
Pathogenic (Mar 02, 2018)	criteria provided, single submitter (Ambry Autosomal Dominant and X-Linked criteria (3/2017)) Method: clinical testing	Hereditary cancer-predisposing syndrome Allele origin: germline	Ambry Genetics Accession: SCV000274469.4 Submitted: (Jul 30, 2018)	Evidence details Publications PubMed (6) Comment: Lines of evidence used in support of classification: Alterations resulting in premature truncation (e.g.reading frame shift, nonsense)
Pathogenic (Oct 31, 2018)	criteria provided, single submitter (ACMG Guidelines, 2015) Method: clinical testing	Familial cancer of breast Medulloblastoma Malignant tumor of prostate Wilms tumor 1 Fanconi anemia, complementation group D1 Breast-ovarian cancer, familial 2 Glioma susceptibility 3	Fulgent Genetics,Fulgent Genetics Accession: SCV000893328.1 Submitted: (Nov 14, 2018)	Evidence details Publications PubMed (1) DOI: 10.1038/gim.2015.30

Genetic variation databases - GnomAD



V4: 730,947 exomes and 76,215 genomes
V3: 71,702 genomes
V2: 125,748 exomes and 15,708 genomes

- unrelated individuals sequenced as part of various disease-specific and population genetic studies

Genetic variation databases - GnomAD

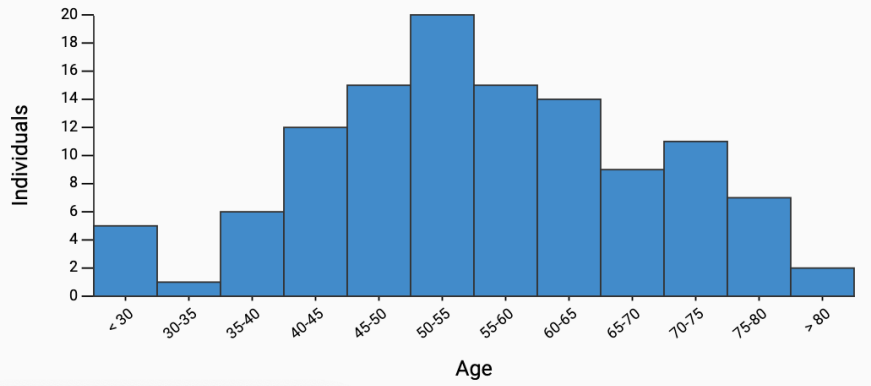
- Known / novel variants (using **dbSNP**/**gnomAD**/others)
- Allele and genotype frequencies (using **gnomAD**/others)

Population Frequencies

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
African	195	24484	1	0.007964
Latino	8	35346	0	0.0002263
Other	1	7174	0	0.0001394
South Asian	1	30590	0	0.00003269
European (non-Finnish)	2	126916	0	0.00001576
Ashkenazi Jewish	0	10288	0	0.000
East Asian	0	19880	0	0.000
European (Finnish)	0	24970	0	0.000
Female	117	127432	1	0.0009181
Male	90	152216	0	0.0005913
Total	**207**	**279648**	**1**	**0.0007402**

Include: Exomes Genomes

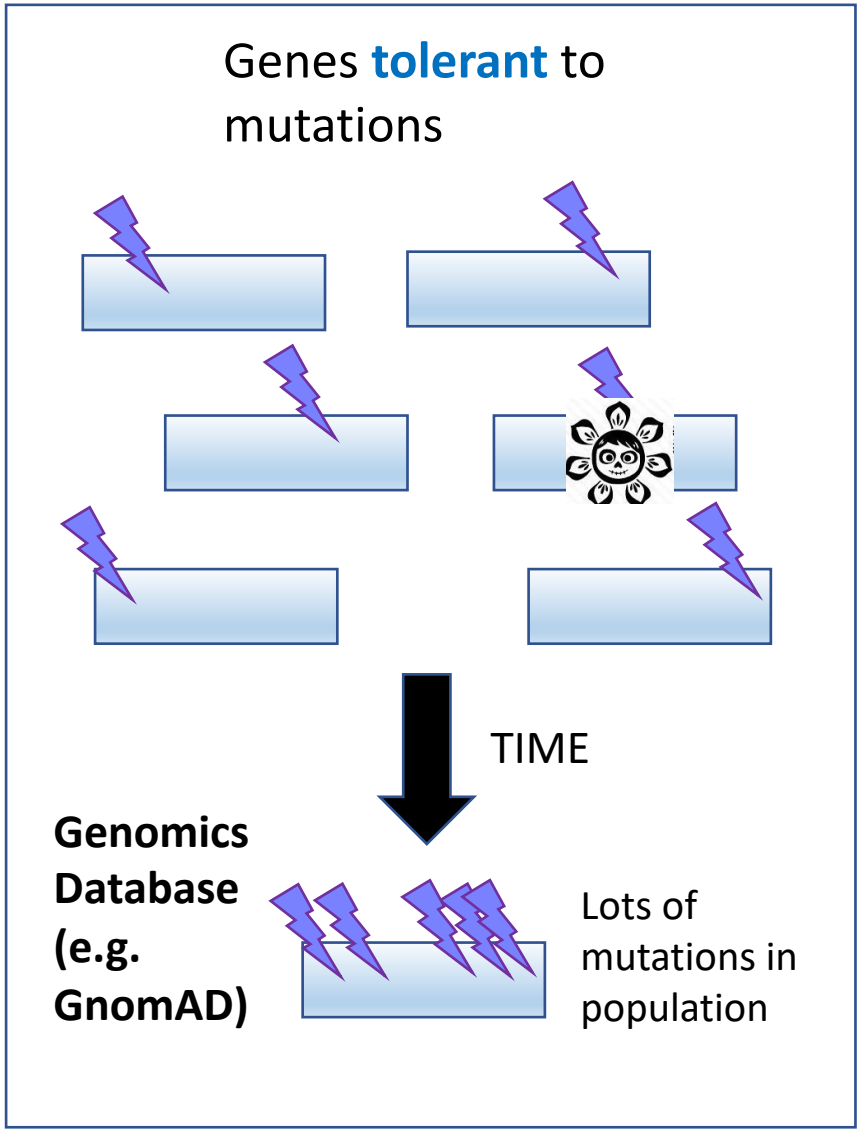
Age Distribution

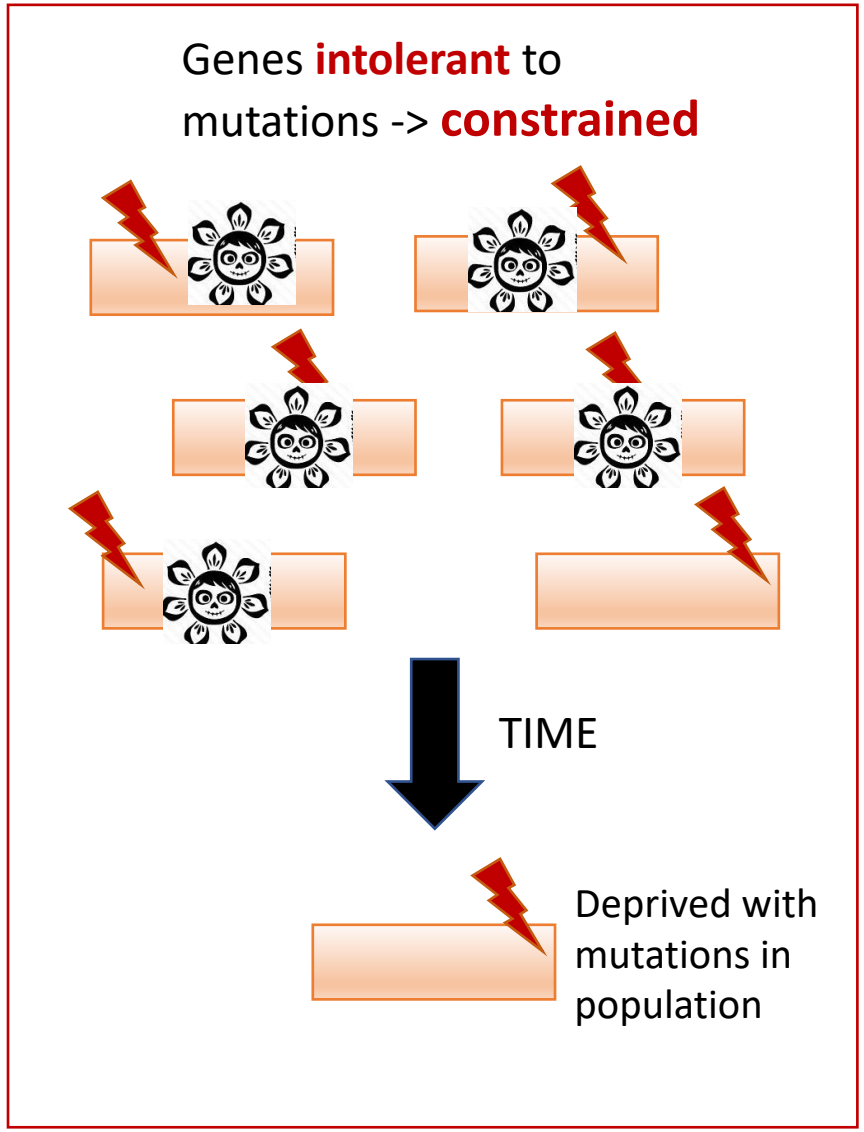


***PCSK9* C679X**

https://gnomad.broadinstitute.org/variant/1-55529215-C-A

Genetic variation databases - GnomAD





Genetic variation databases - GnomAD

CHD8 chromodomain helicase DNA binding protein 8

Dataset: gnomAD v4.1.0 | gnomAD SVs v4.1.0

Genome build GRCh38 / hg38
Ensembl gene ID ENSG00000100888.15
MANE Select transcript ENST00000646647.2 / NM_001170629.2
Ensembl canonical transcript ENST00000646647.2
Other transcripts
ENST00000555301.1, ENST00000556833.1, and 21 more
Region 14:21385194-21456126
External resources Ensembl, UCSC Browser, and more

Constraint | Variant co-occurrence

Category	Expected SNVs	Observed SNVs	Constraint metrics
Synonymous	1104	1012	Z = 1.51 o/e = 0.92 (0.87 - 0.97)
Missense	3052.9	1988	Z = 7.04 o/e = 0.65 (0.63 - 0.68)
pLoF	234.8	26	pLI = 1 o/e = 0.11 (0.08 - 0.15)

CFTR CF transmembrane conductance regulator

Dataset: gnomAD v4.1.0 | gnomAD SVs v4.1.0

Genome build GRCh38 / hg38
Ensembl gene ID ENSG00000001626.17
MANE Select transcript ENST00000003084.11 / NM_000492.4
Ensembl canonical transcript ENST00000003084.11
Other transcripts
ENST00000426809.5, ENST00000429014.1, and 23 more
Region 7:117287120-117715971
External resources Ensembl, UCSC Browser, and more

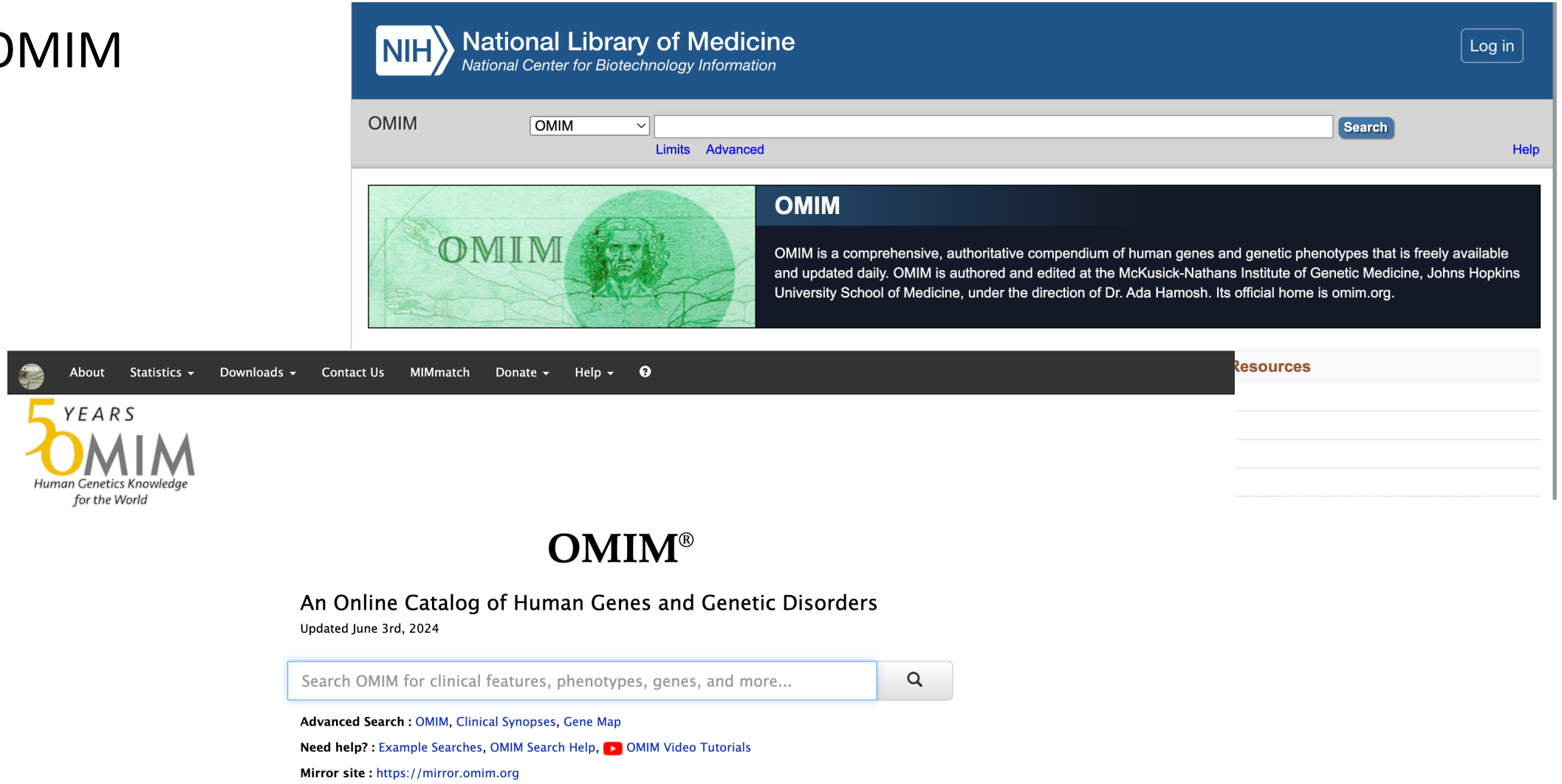
Constraint | Variant co-occurrence

Category	Expected SNVs	Observed SNVs	Constraint metrics
Synonymous	511.8	476	Z = 0.86 o/e = 0.93 (0.86 - 1)
Missense	1412.5	1523	Z = -1.07 o/e = 1.08 (1.03 - 1.13)
pLoF	117.5	116	pLI = 0 o/e = 0.99 (0.85 - 1.15)

Constraint metrics based on MANE Select transcript (ENST00000003084.11).

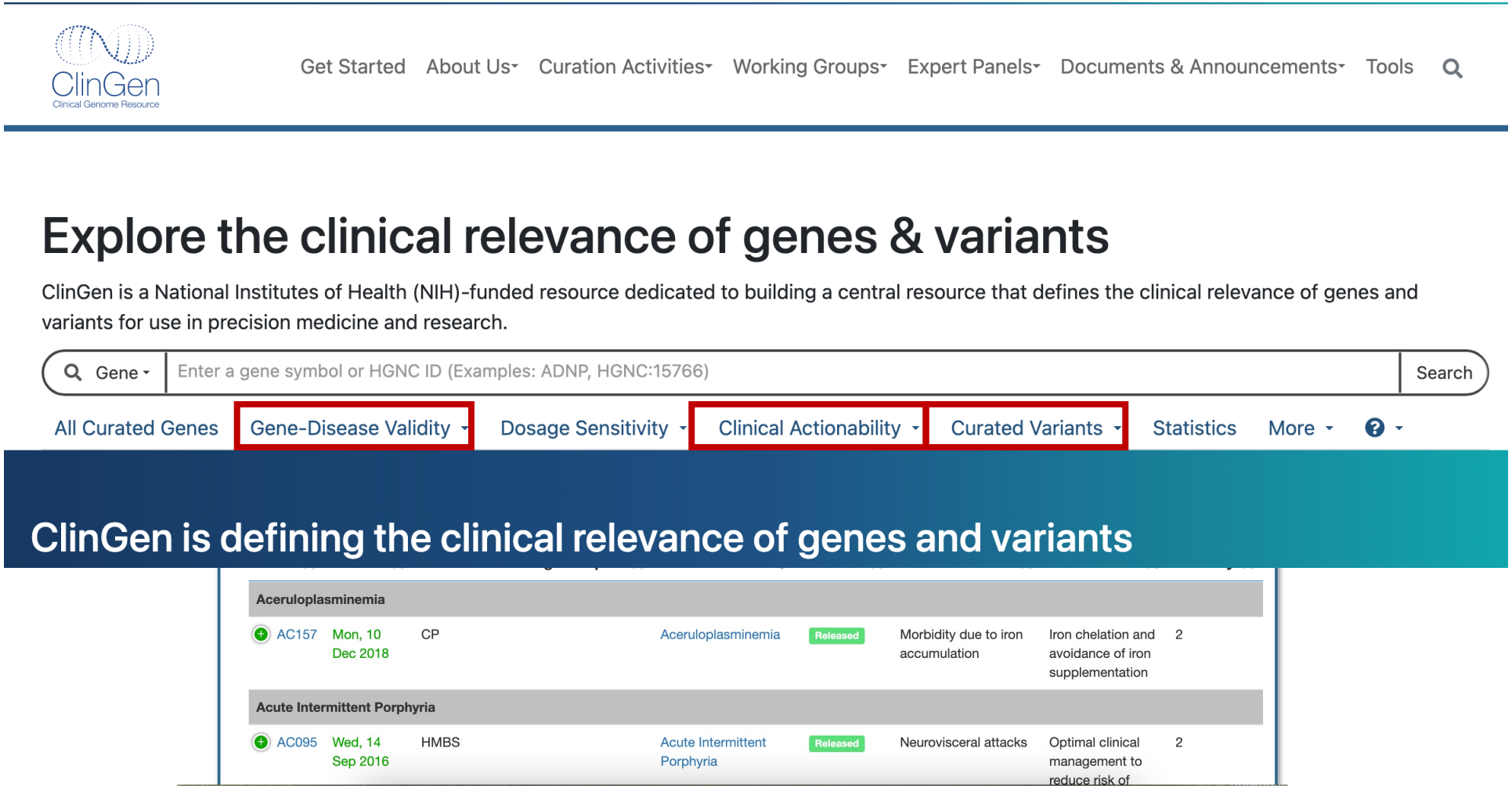
Human Traits & Diseases Databases

- OMIM



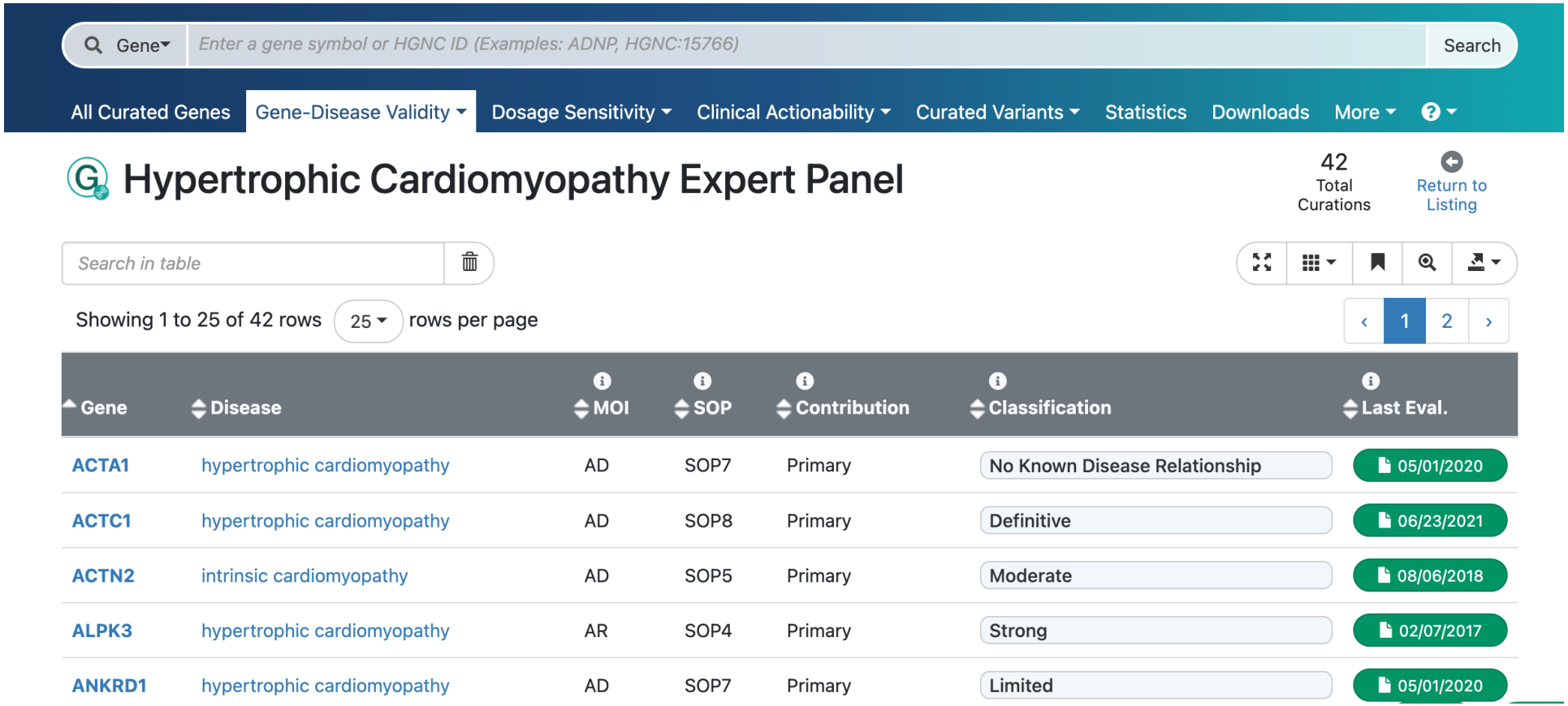
Human Traits & Diseases Databases

- https://www.clinicalgenome.org/



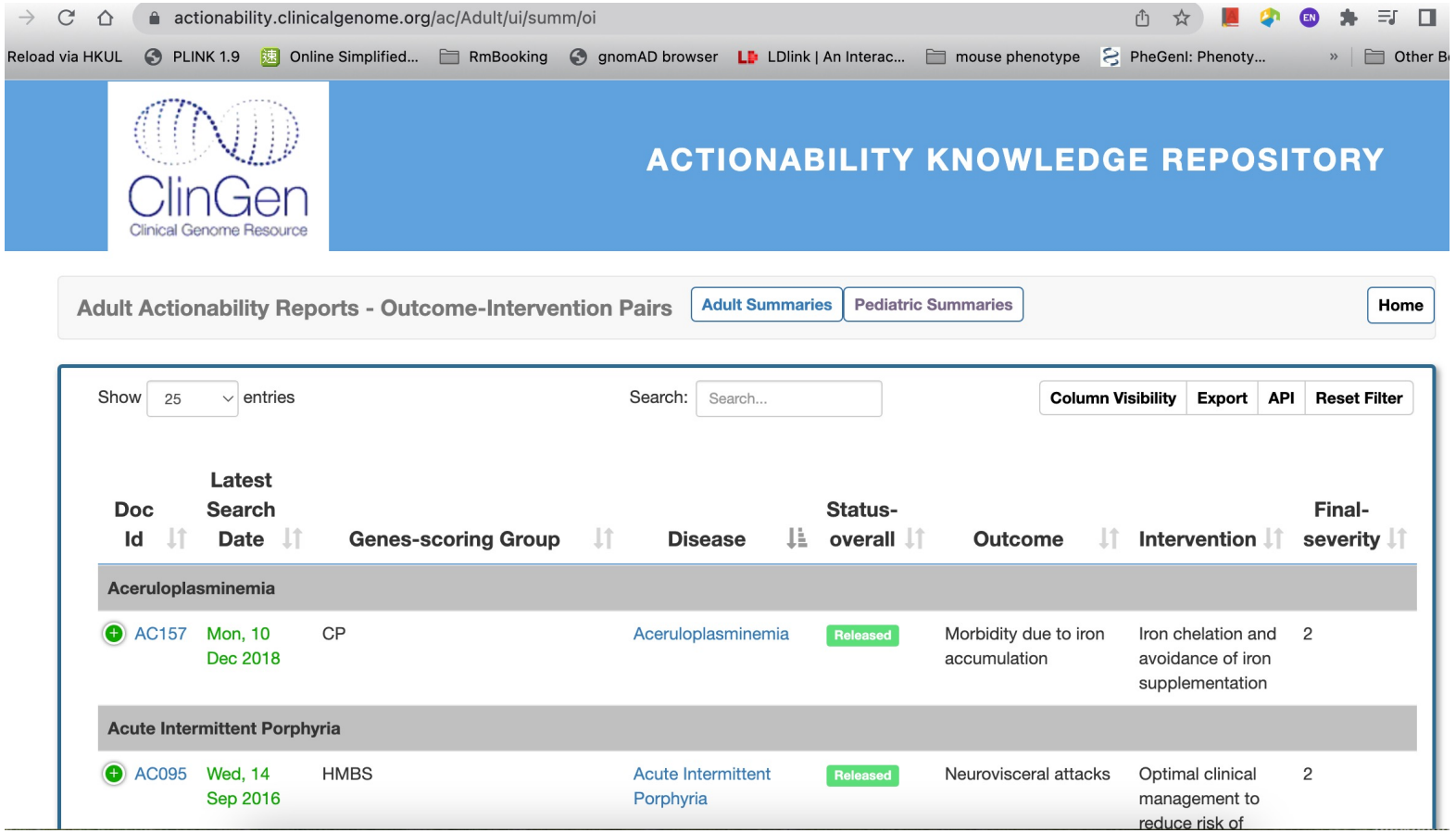
Human Traits & Diseases Databases

- Gene-disease validity



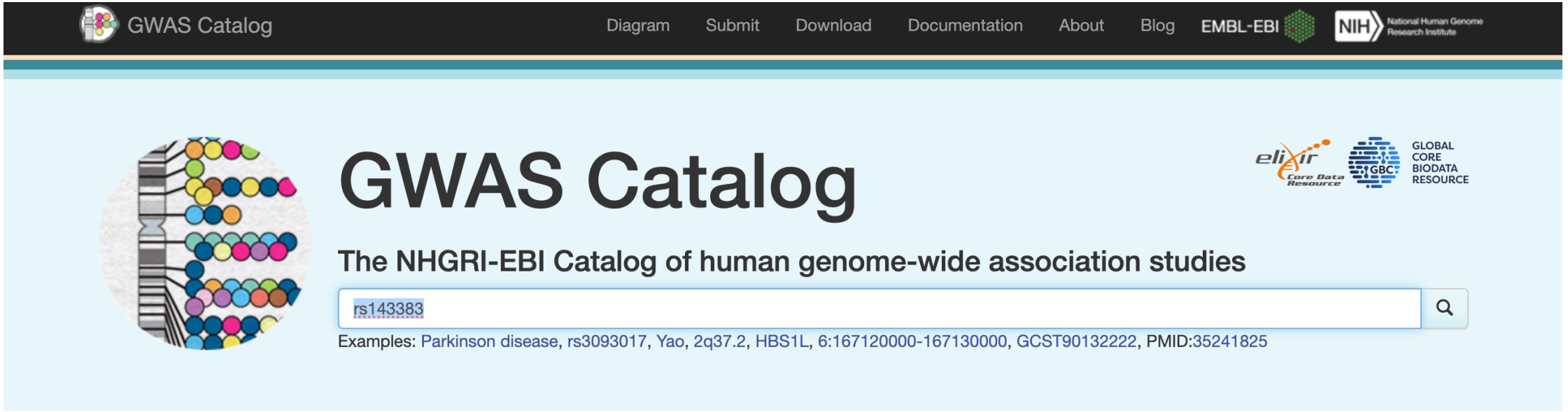
Human Traits & Diseases Databases

- Clinical actionability: the findings should be medically actionable



Human Traits & Diseases Databases

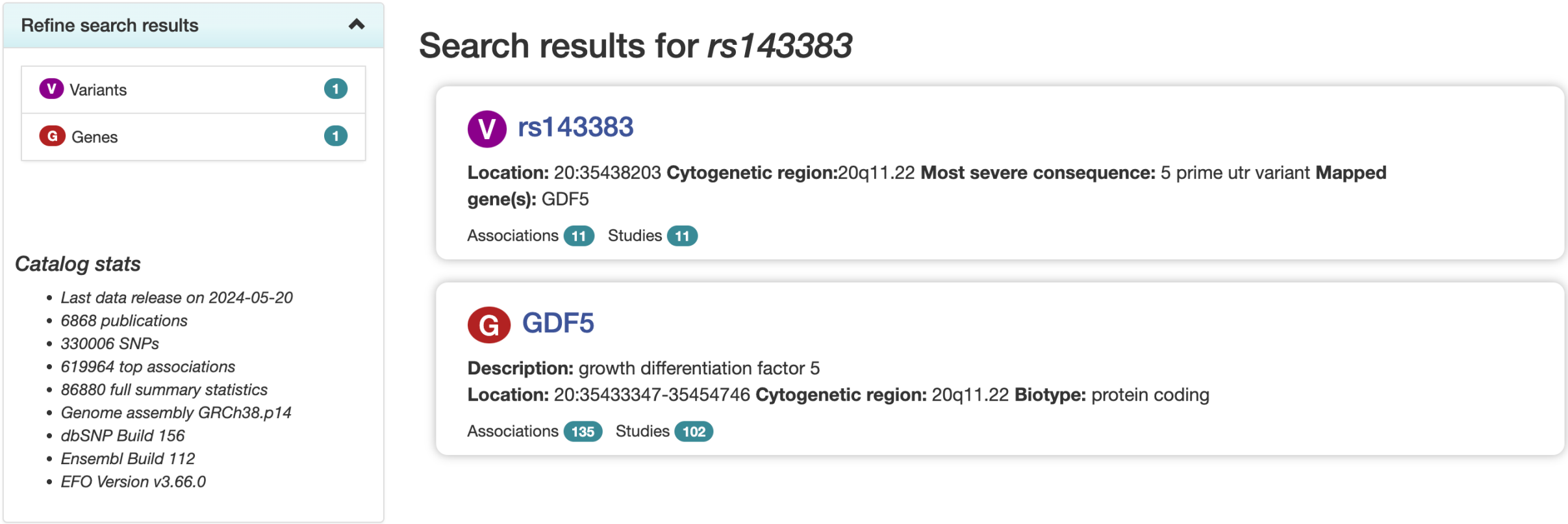
- GWAS catalogue (https://www.ebi.ac.uk/gwas/home)



Do data resources managed by EMBL-EBI and our collaborators make a difference to your work? Please take 10 minutes to fill in our annual user survey, and help us make the case for why sustaining open data resources is critical for life sciences research.

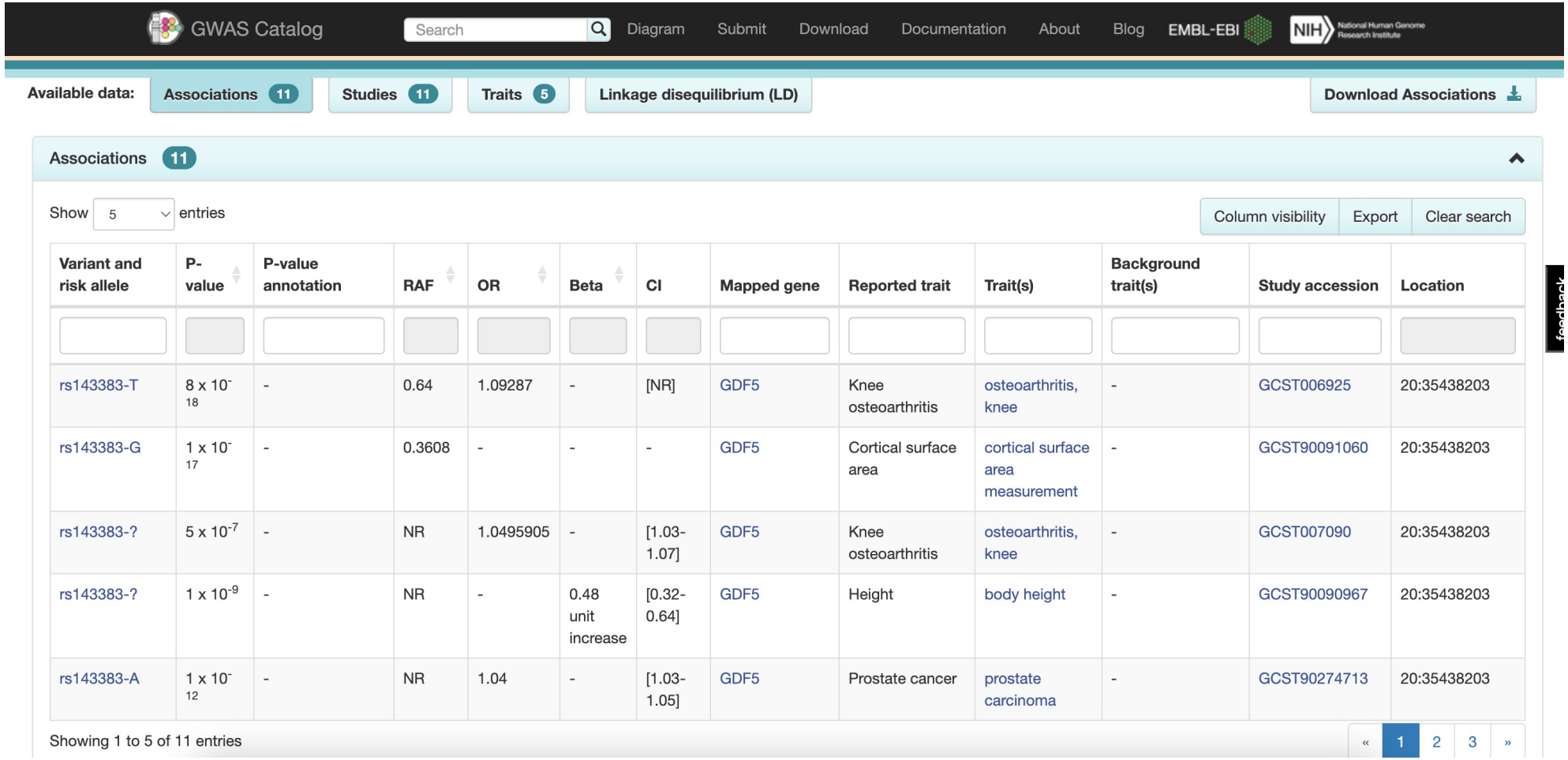
Human Traits & Diseases Databases

- GWAS catalogue (https://www.ebi.ac.uk/gwas/home)
- Let's have a look at a variant, **rs143383**



Human Traits & Diseases Databases

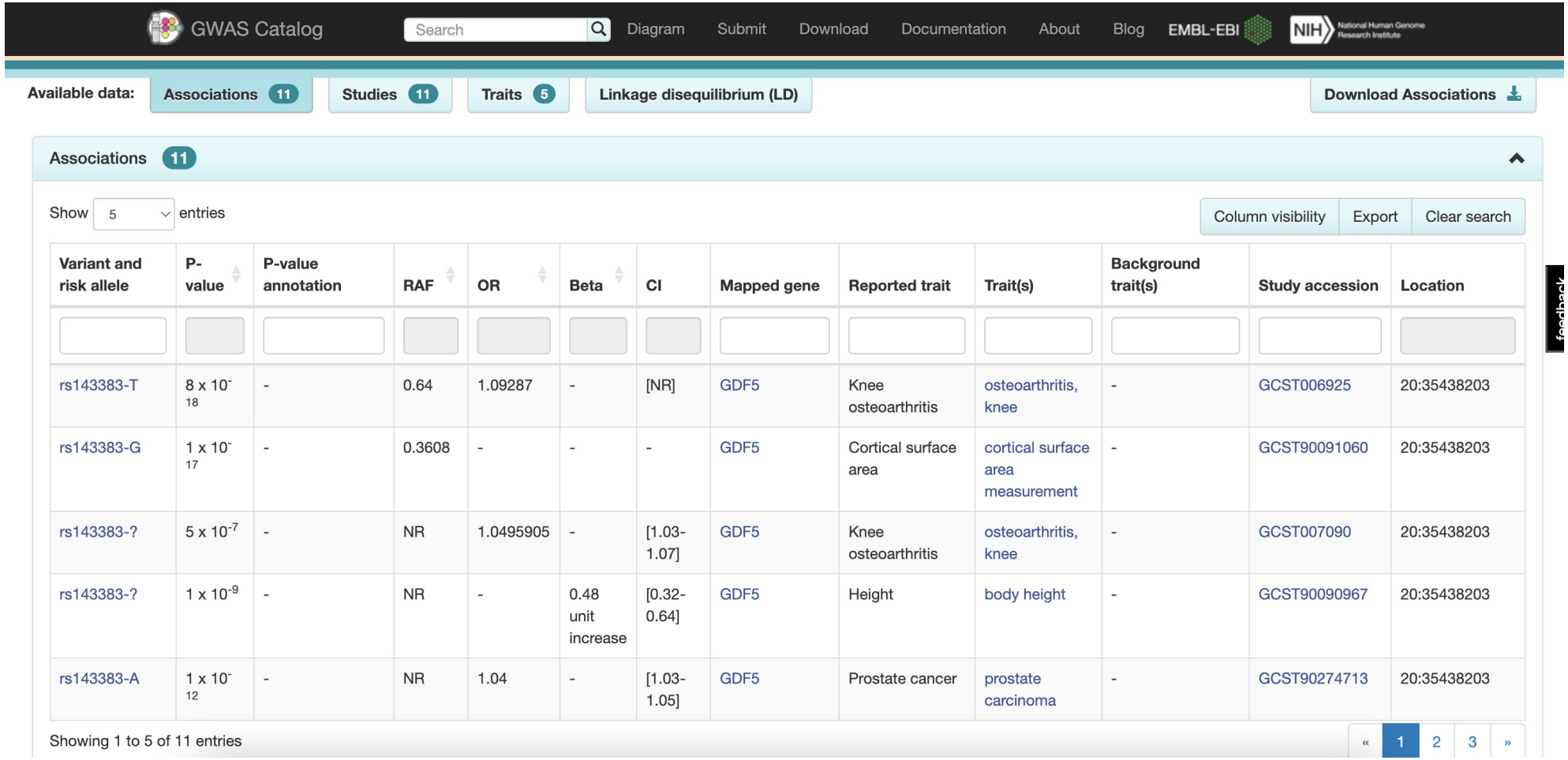
- GWAS catalogue (https://www.ebi.ac.uk/gwas/home)
- Let's have a look at a variant, **rs143383**



Variant and risk allele	P-value	P-value annotation	RAF	OR	Beta	CI	Mapped gene	Reported trait	Trait(s)	Background trait(s)	Study accession	Location
rs143383-T	8 x 10^{-18}	-	0.64	1.09287	-	[NR]	GDF5	Knee osteoarthritis	osteoarthritis, knee	-	GCST006925	20:35438203
rs143383-G	1 x 10^{-17}	-	0.3608	-	-	-	GDF5	Cortical surface area	cortical surface area measurement	-	GCST90091060	20:35438203
rs143383-?	5 x 10^{-7}	-	NR	1.0495905	-	[1.03-1.07]	GDF5	Knee osteoarthritis	osteoarthritis, knee	-	GCST007090	20:35438203
rs143383-?	1 x 10^{-9}	-	NR	-	0.48 unit increase	[0.32-0.64]	GDF5	Height	body height	-	GCST90090967	20:35438203
rs143383-A	1 x 10^{-12}	-	NR	1.04	-	[1.03-1.05]	GDF5	Prostate cancer	prostate carcinoma	-	GCST90274713	20:35438203

Human Traits & Diseases Databases

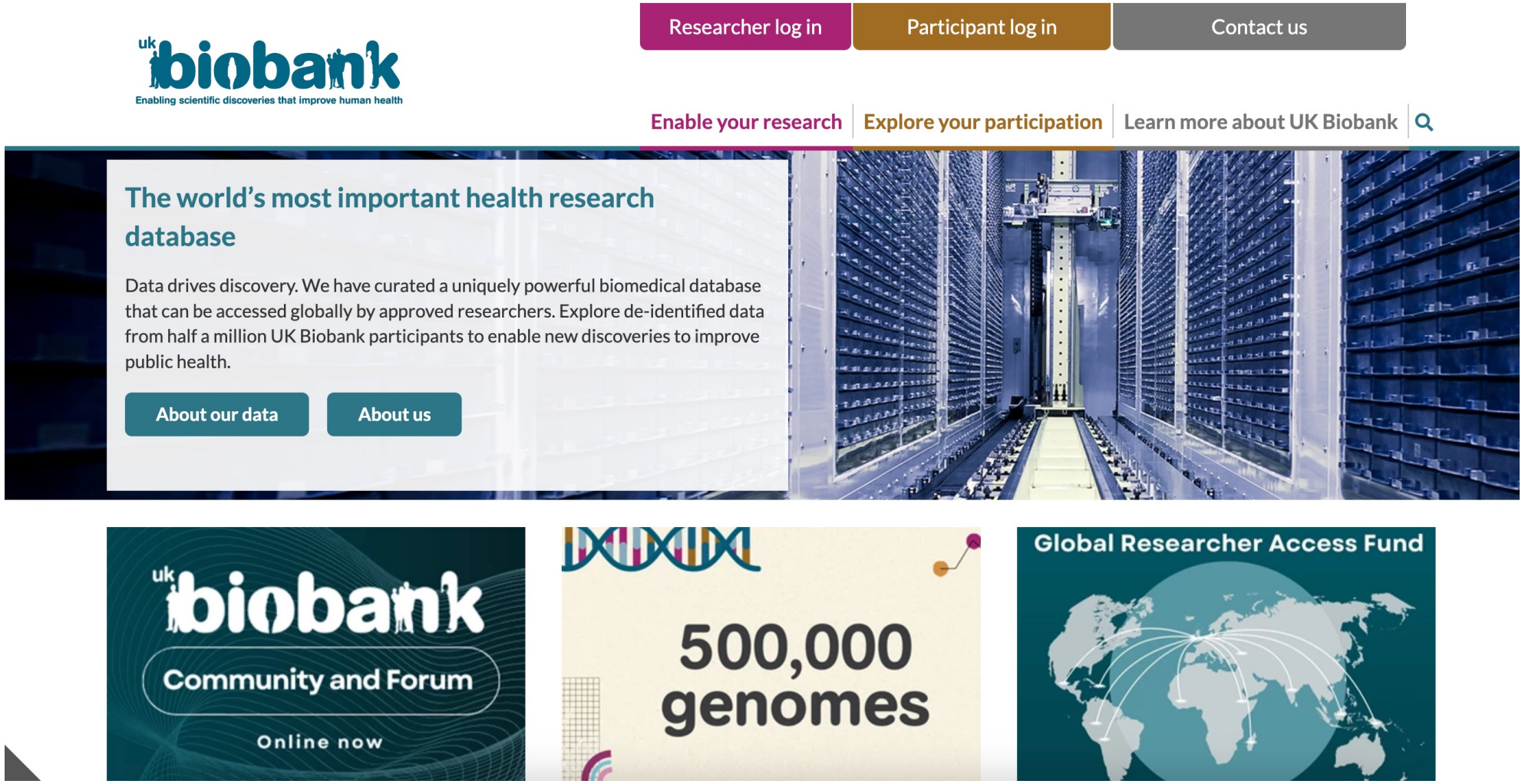
- GWAS catalogue (https://www.ebi.ac.uk/gwas/home)
- Let's have a look at a variant, **rs143383**



Variant and risk allele	P-value	P-value annotation	RAF	OR	Beta	CI	Mapped gene	Reported trait	Trait(s)	Background trait(s)	Study accession	Location
rs143383-T	8 x 10^{-18}	-	0.64	1.09287	-	[NR]	GDF5	Knee osteoarthritis	osteoarthritis, knee	-	GCST006925	20:35438203
rs143383-G	1 x 10^{-17}	-	0.3608	-	-	-	GDF5	Cortical surface area	cortical surface area measurement	-	GCST90091060	20:35438203
rs143383-?	5 x 10^{-7}	-	NR	1.0495905	-	[1.03-1.07]	GDF5	Knee osteoarthritis	osteoarthritis, knee	-	GCST007090	20:35438203
rs143383-?	1 x 10^{-9}	-	NR	-	0.48 unit increase	[0.32-0.64]	GDF5	Height	body height	-	GCST90090967	20:35438203
rs143383-A	1 x 10^{-12}	-	NR	1.04	-	[1.03-1.05]	GDF5	Prostate cancer	prostate carcinoma	-	GCST90274713	20:35438203

Human Traits & Diseases Databases - UKBiobank

- https://www.ukbiobank.ac.uk/



Human Traits & Diseases Databases - UKBiobank

- https://pheweb.org/

Datasets:

UKBiobank TOPMed-imputed: **1400** EHR-derived broad PheWAS codes for **57 million** TOPMed-imputed variants in **400,000** white British individuals.

UKBiobank HRC-imputed: **1400** EHR-derived broad PheWAS codes for **20 million** imputed variants in **400,000** white British individuals.

UKBiobank Neale v1: **2400** traits for **11 million** imputed variants in **337,000** unrelated white British individuals.

MGI BioVU Labs Meta: **70** meta-analyzed phenotypes for **800,000** variants.

FinMetSeq: **60** phenotypes for **400,000** variants in **8000-19,000** individuals.

Private Datasets:

HUNT

SardiNIA

MGI freeze3: **1542** EHR-derived broad PheWAS codes for approximately **51.8 million** imputed variants in **51,583** EUR-ancestry.

MGI freeze6: **1728** EHR-derived broad PheWAS codes for approximately **52** million imputed variants in a multi-ancestry cohort of **80,381** individuals

Human Traits & Diseases Databases - UKBiobank

- https://pheweb.org/

Datasets:

UKBiobank TOPMed-imputed: **1400** EHR-derived broad PheWAS codes for **57 million** TOPMed-imputed variants in **400,000** white British individuals.

UKBiobank HRC-imputed: **1400** EHR-derived broad PheWAS codes for **20 million** imputed variants in **400,000** white British individuals.

UKBiobank Neale v1: **2400** traits for **11 million** imputed variants in **337,000** unrelated white British individuals.

MGI BioVU Labs Meta: **70** meta-analyzed phenotypes for **800,000** variants.

FinMetSeq: **60** phenotypes for **400,000** variants in **8000-19,000** individuals.

Private Datasets:

HUNT

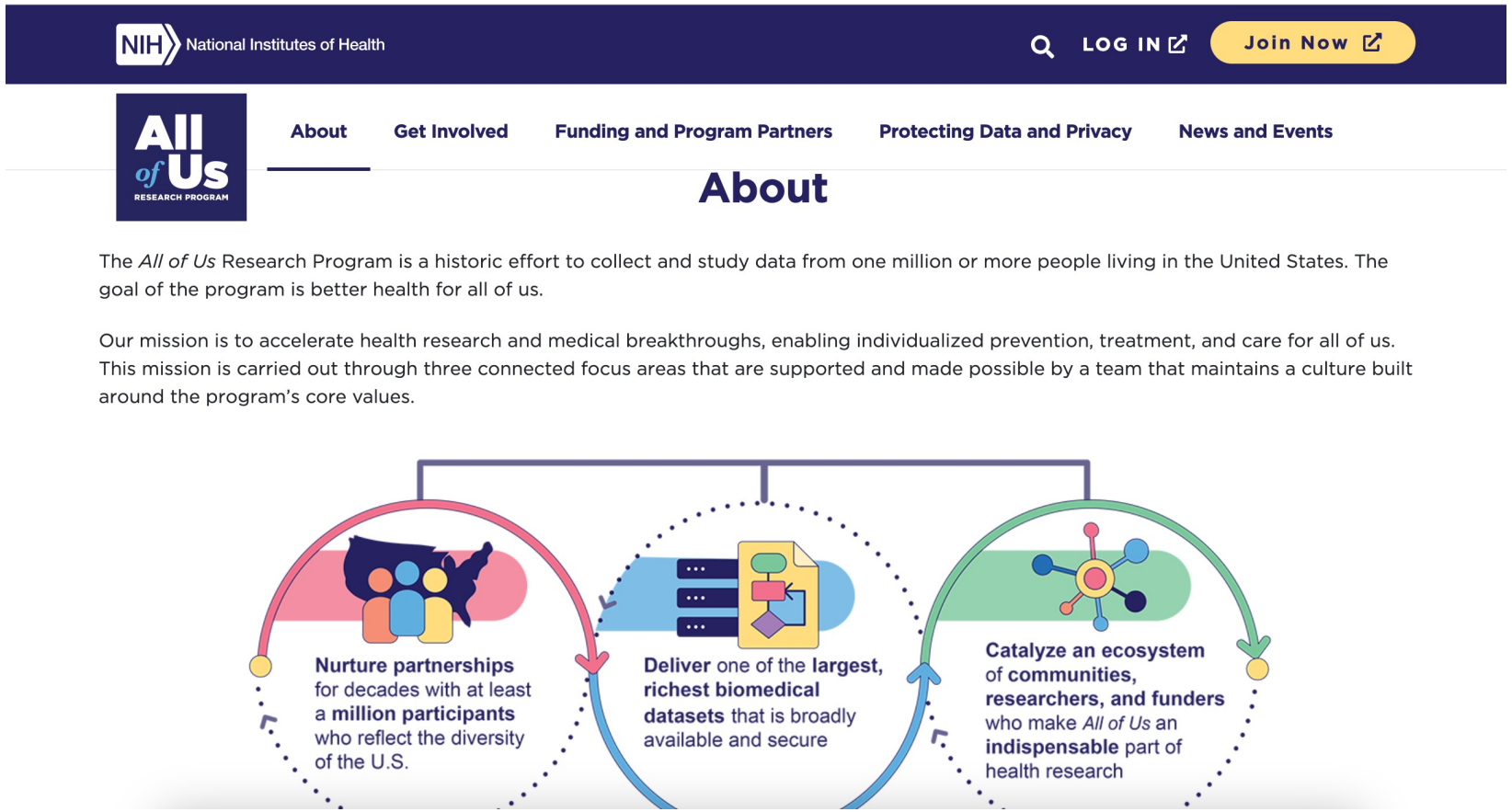
SardiNIA

MGI freeze3: **1542** EHR-derived broad PheWAS codes for approximately **51.8 million** imputed variants in **51,583** EUR-ancestry.

MGI freeze6: **1728** EHR-derived broad PheWAS codes for approximately **52** million imputed variants in a multi-ancestry cohort of **80,381** individuals

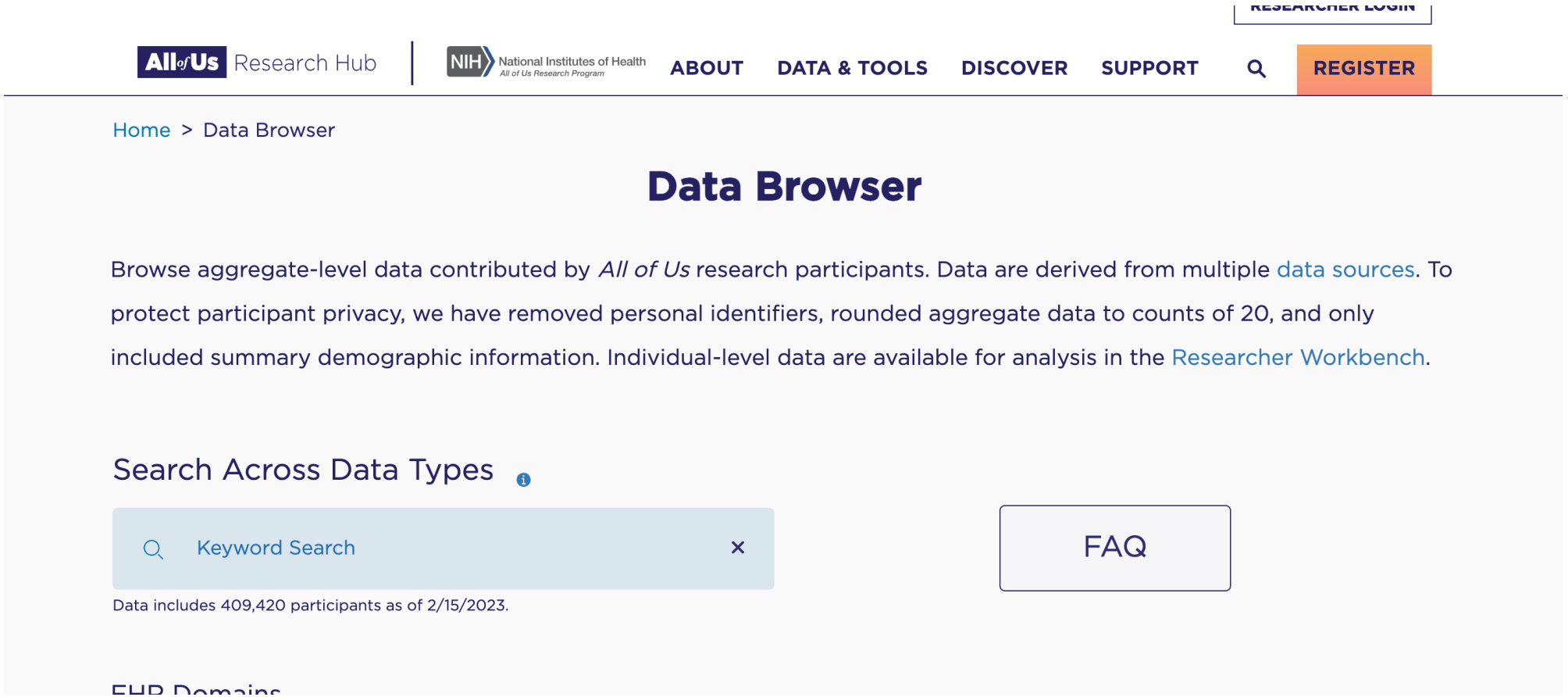
Human Traits & Diseases Databases - All of Us

- https://allofus.nih.gov



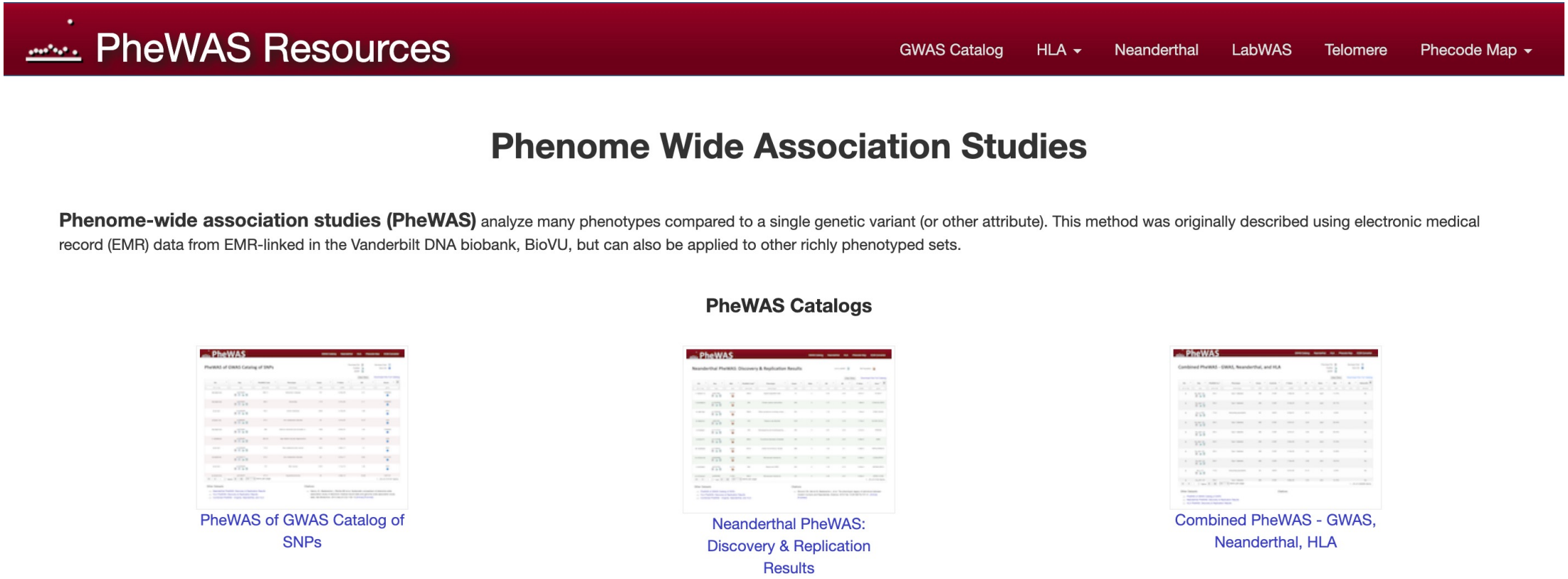
Human Traits & Diseases Databases - All of Us

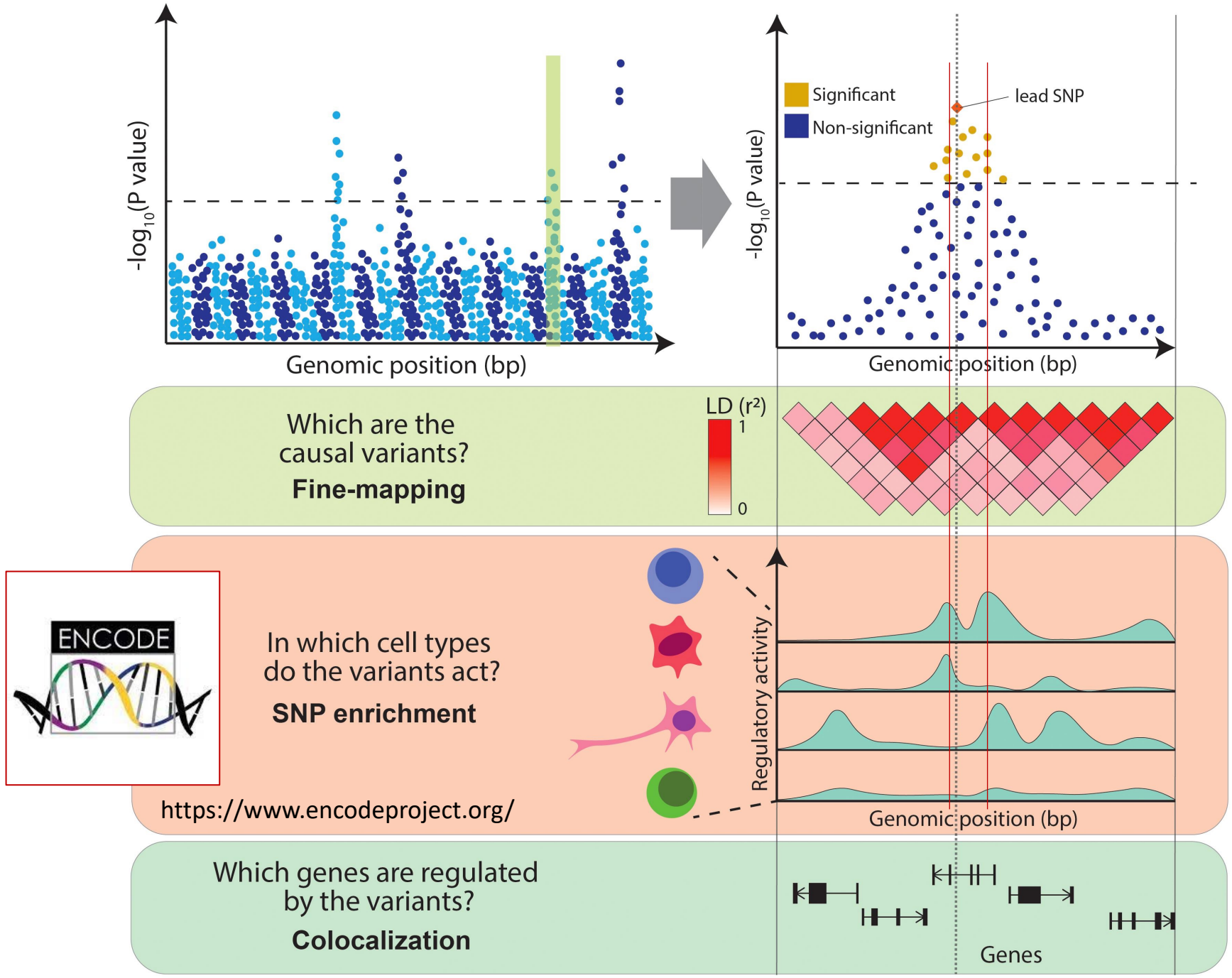
- https://allofus.nih.gov

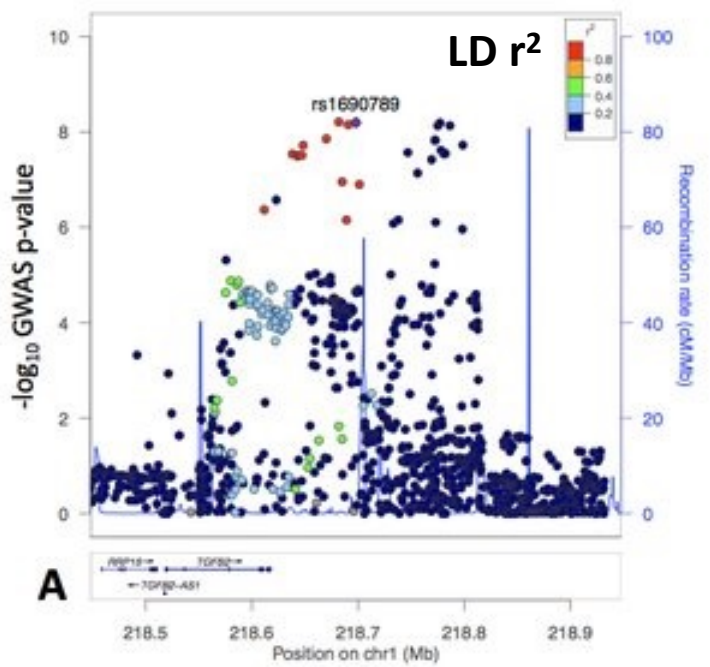


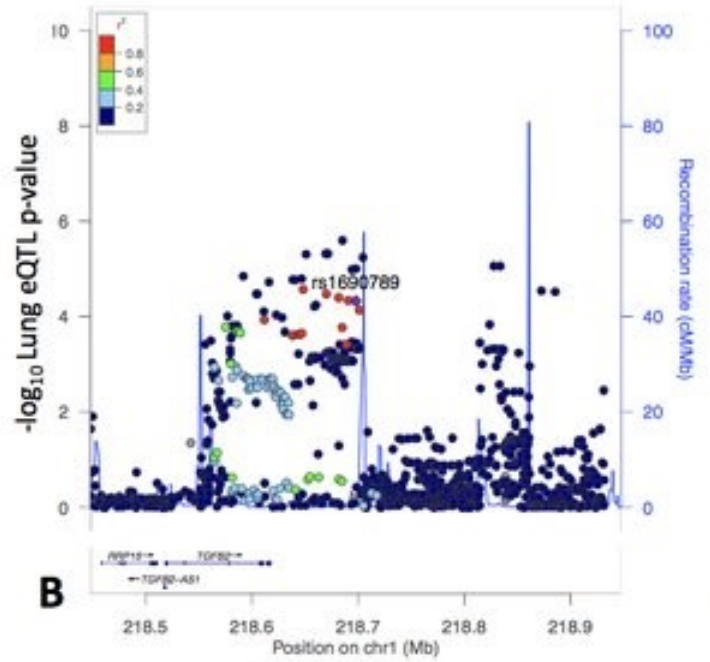
Human Traits & Diseases Databases - PheWAS

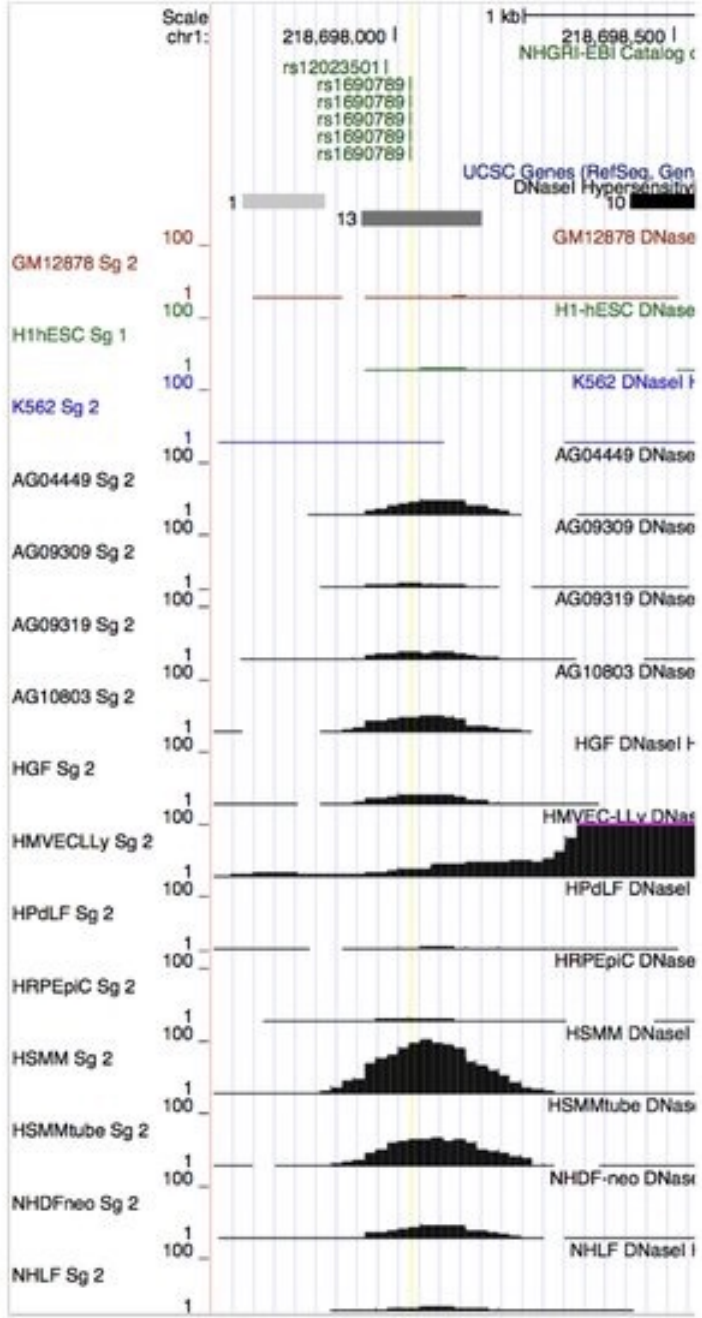
- https://www.phewascatalog.org/




-log10(P value)
Genomic position (bp)
Significant
Non-significant
lead SNP
-log10(P value)
Genomic position (bp)
Which are the causal variants?
Fine-mapping
LD (r2)
1
0
ENCODE
In which cell types do the variants act?
SNP enrichment
https://www.encodeproject.org/
Regulatory activity
Genomic position (bp)
Which genes are regulated by the variants?
Colocalization
Genes







https://pubmed.ncbi.nlm.nih.gov/31343404/

Gene Expression Databases - GTEx

https://gtexportal.org/home/

About GTEx

- GTEx
- GTEx Consortium
- Data Release and Publication Policy

The GTEx Project

Correlations between genotype and tissue-specific gene expression levels will help ider genome that influence whether and how much a gene is expressed. GTEx will help rese understand inherited susceptibility to disease and will be a resource database and tissu studies in the future.

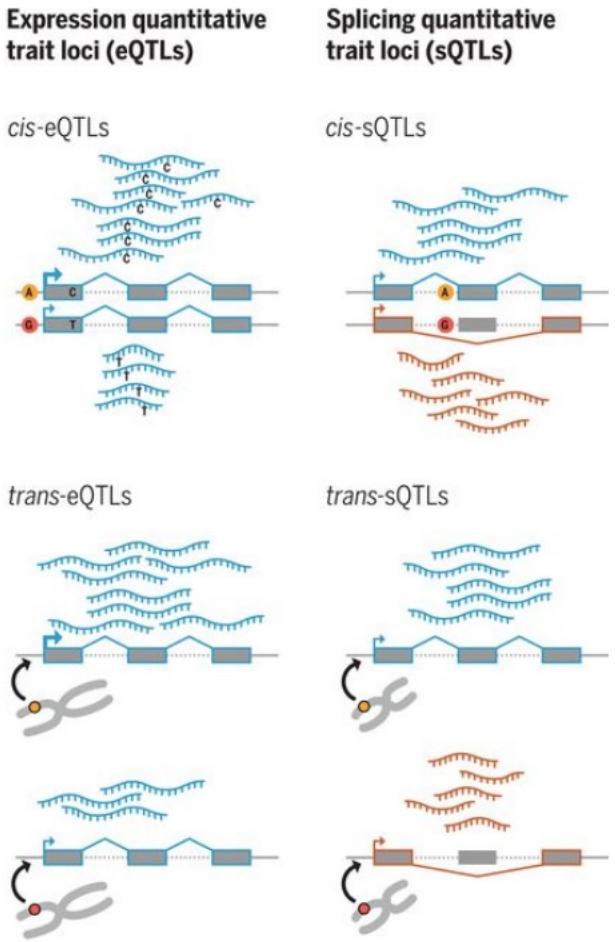
The Genotype-Tissue Expression (GTEx) project aims to provide to the scientific commu which to study human gene expression and regulation and its relationship to genetic va will collect and analyze multiple human tissues from donors who are also densely genot genetic variation within their genomes. By analyzing global RNA expression within indivi treating the expression levels of genes as quantitative traits, variations in gene expressi correlated with genetic variation can be identified as expression quantitative trait loci, o

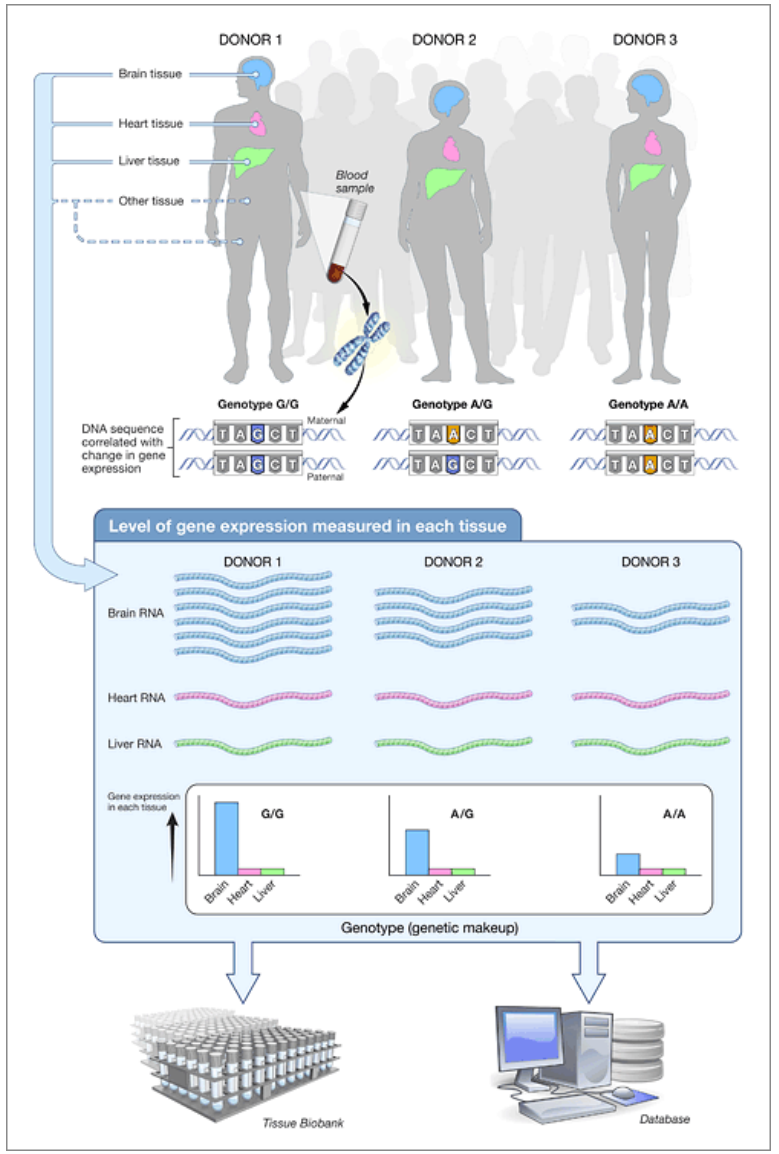
Despite the rapid progress achieved using genome-wide association studies (GWAS; See: https://www.ebi.ac.uk/gwas/) to identify genetic changes associated with comr such as heart disease, cancer, diabetes, asthma, and stroke, a large majority of these g outside of the protein-coding regions of genes and often even outside of the genes ther difficult to discern which genes are affected and by what mechanism. The comprehensi human eQTLs will greatly help to identify genes whose expression is affected by genetic provide a valuable basis on which to study the mechanism of that gene regulation.

The project will also involve consultation and research into the ethical, legal and social i research, support for statistical methods development, and creation of a database to ho

Gene Expression Databases - GTEx

https://gtexportal.org/home/





Browse

By Tissue

Histology Viewer

Single Cell

Data Overview

Multi-Gene Single Cell Query

Expression

Multi-Gene Query

Transcript Browser

QTL

Locus Browser (Gene-centric)

Locus Browser (Variant-centric)

Gene Expression Databases - GTEx

https://gtexportal.org/home/

About GTEx

- GTEx
- GTEx Consortium
- Data Release and Publication Policy

The GTEx Project

Correlations between genotype and tissue-specific gene expression levels will help ider genome that influence whether and how much a gene is expressed. GTEx will help rese understand inherited susceptibility to disease and will be a resource database and tissu studies in the future.

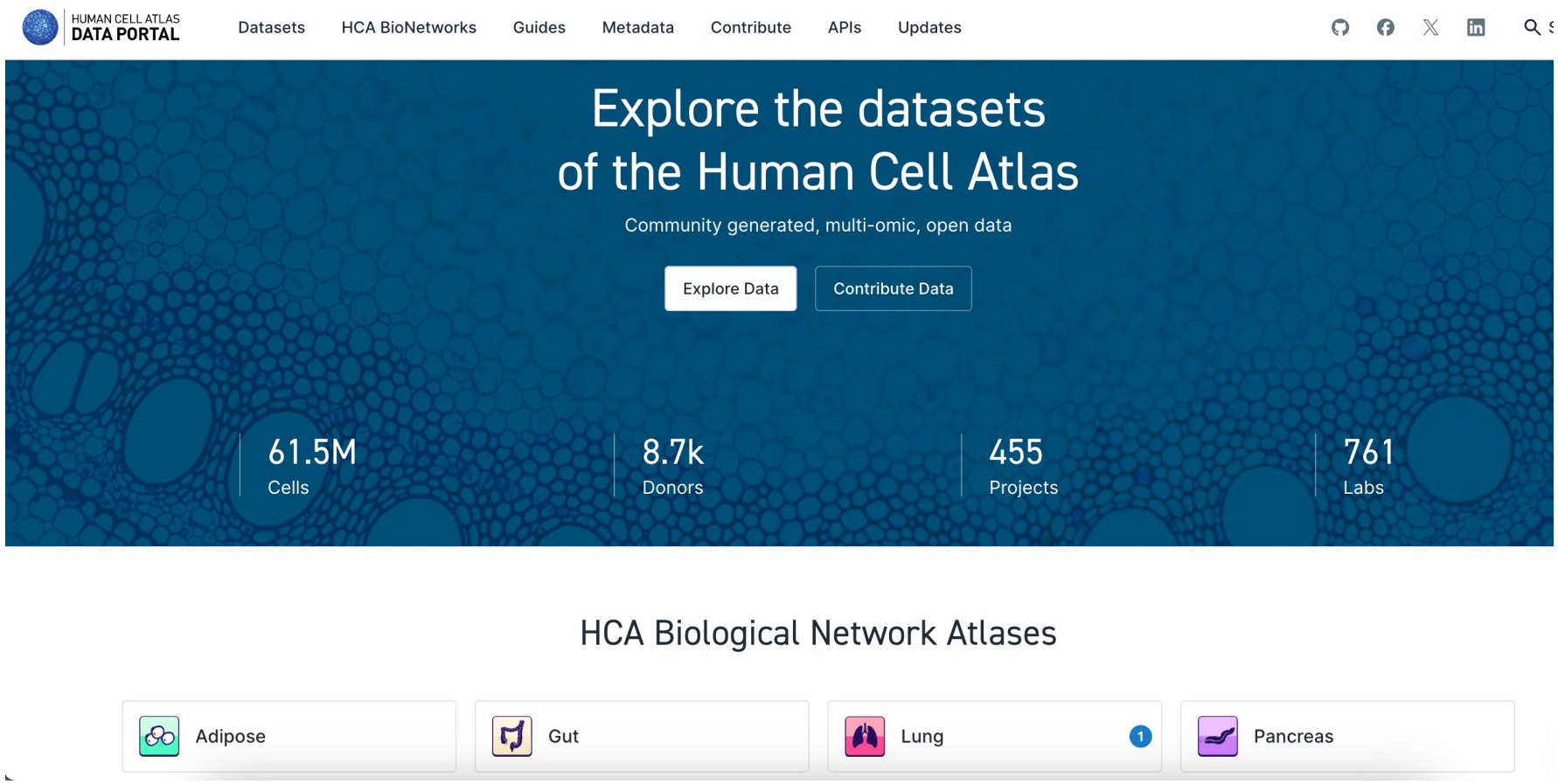
The Genotype-Tissue Expression (GTEx) project aims to provide to the scientific commu which to study human gene expression and regulation and its relationship to genetic va will collect and analyze multiple human tissues from donors who are also densely genot genetic variation within their genomes. By analyzing global RNA expression within indivi treating the expression levels of genes as quantitative traits, variations in gene expressi correlated with genetic variation can be identified as expression quantitative trait loci, o

Despite the rapid progress achieved using genome-wide association studies (GWAS; See: https://www.ebi.ac.uk/gwas/) to identify genetic changes associated with comr such as heart disease, cancer, diabetes, asthma, and stroke, a large majority of these g outside of the protein-coding regions of genes and often even outside of the genes ther difficult to discern which genes are affected and by what mechanism. The comprehensi human eQTLs will greatly help to identify genes whose expression is affected by genetic provide a valuable basis on which to study the mechanism of that gene regulation.

The project will also involve consultation and research into the ethical, legal and social i research, support for statistical methods development, and creation of a database to ho

Gene Expression Databases - Single cell Atlas

- https://data.humancellatlas.org/



Gene Expression Databases - Single cell Atlas