

# Public bioinformatics resources for genomic analysis interpretation

Clara Tang



# “Public” resources in HKU : HPCF

- CPOS HPCF

The screenshot shows the homepage of the LKS Faculty of Medicine Centre for PanorOmic Sciences. At the top right, there is an email link "enquiry.cpos@hku.hk | 2831-5500" and social media icons for Instagram, Facebook, and Twitter. The main header features the HKU Med logo and the text "LKS Faculty of Medicine" and "Centre for PanorOmic Sciences" in English, with the Chinese name "香港大學泛組學科研中心" below it. Below the header is a navigation menu with links to "Home", "Core Services", "Other Equipment", "News", "Events", "Publications", "FAQ", "About CPOS", "Careers", "Contact", and "Brochures". A search icon is also present. The main content area has a large banner image of a server room with the text "High Performance Computing Facility" overlaid. Below the banner is a sub-navigation menu with links to "Overview", "Data Analysis Services", "Data Analysis Resources", "Commercial Data Analysis Tools", "Online Data Delivery System (ODDS)", "High Performance Computing Facility" (which is highlighted in blue), and "Contact".

## Overview

CPOS IT operates the High Performance Computing Facility (HPCF), comprising of 54 nodes of 3,266 CPU threads, 20,121GB memory and 4.2PB storage capacity, on a high speed network.

# “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/>)

The screenshot shows the main interface of the ITS HPC2021 web portal. At the top is a navigation bar with links for Apps, Clusters, Desktop, Files, IDE, Interactive Apps, Jobs, and a download icon. Below the navigation bar is a yellow banner with the text: "Please close all browser tab(s) to perform a proper log-out rather than clicking the \"Log out\" button." To the right of the banner is a logo placeholder labeled "logo". Below the banner, a section titled "OnDemand provides an integrated, single access point for all of your HPC resources." is visible. On the left side, there is a "Pinned Apps" section showing a grid of eight application icons with their names and status: Remote Desktop (System Installed App), Home Directory (System Installed App), HPC2021-io2 Shell Access (System Installed App), Active Jobs (System Installed App), Visual Studio Code (System Installed App), RStudio (System Installed App), MATLAB (System Installed App), and ParaView (System Installed App). To the right of the pinned apps is a "Disk Quota and Usage" table with the following data:

Folder	Disk Quota	Used Quota	Quota Usage
/home/clalatsm	100G	28G	28%
/lustre1/u/clalatsm	500G	388G	77.59%
/lustre1/g/surg_claratsm	20480G	10915G	53.30%

At the bottom, there is a section titled "Resource usage efficiency of recent jobs" with a table header:

Job ID	End	Job Name	State	Requested CPU	CPU Efficiency(%)	Requested RAM(GB)	RAM Efficiency(%)	Walltime
--------	-----	----------	-------	---------------	-------------------	-------------------	-------------------	----------

# “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
<b>Path</b>	/home/\$USER	/lustre1/u/\$USER	/lustre1/g/\$PI_GROUP	/tmp
<b>Usage</b>	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
<b>Availability</b>	Available across any nodes in HPC2021	Available across any nodes in HPC2021	Available across any nodes in HPC2021	Available on the attached node only
<b>Capacity</b>	100GB per user	500GB per user	<ul style="list-style-type: none"><li>• 5TB (Default) per PI group</li><li>• Up to 15TB at no charge</li><li>• Above 15TB may be charged at standard storage rates</li></ul>	<ul style="list-style-type: none"><li>• ~400GB per general purpose compute node</li><li>• 8.6TB per GPU node</li><li>• 11TB per large memory node</li></ul>
<b>Performance</b>	Moderate – Not appropriate for workload requiring high throughput or small file operations	High throughput for large files and IO: <ul style="list-style-type: none"><li>• Good for sequential access to files of moderately large size (&gt;1MB)</li><li>• Not so good for small or random file access (e.g source code, scripts, software, temporary files)</li></ul>	High throughput for large files and IO: <ul style="list-style-type: none"><li>• Good for sequential access to files of moderately large size (&gt;1MB)</li><li>• Not so good for small or random file access (e.g source code, scripts, software, temporary files)</li></ul>	High performance, especially in terms of operation of small files
<b>Clean-up Policy</b>	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 <b>job submission/management</b>
hpc2021-io1.hku.hk	IO Node	Reserved for <b>file transfer</b> , file management and data analysis/visualization
hpc2021-io2.hku.hk		

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

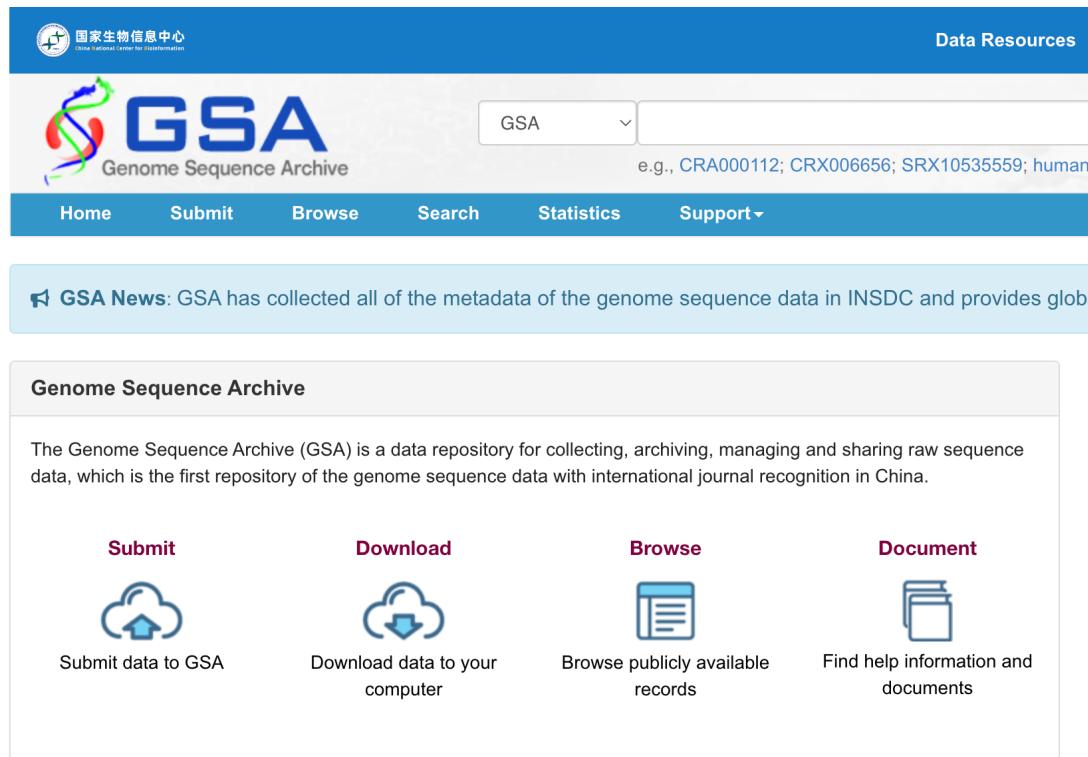
- Genome Browsers
- Genetic Variations Databases
- Human Traits & Diseases Databases
- Nucleotide Sequence Databases
- Gene Expression Databases
- Protein Sequence & domains: Databases and Search Tools
- Phylogeny & Taxonomy
- Databases of other organisms
- Gene Prediction
- Metabolic, Gene Regulatory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



The screenshot shows the homepage of the GSA (Genome Sequence Archive). At the top, there's a blue header bar with the GSA logo and a search bar. Below the header, a banner displays a news item about GSA collecting metadata from INSDC. The main content area is titled "Genome Sequence Archive" and contains a paragraph about the repository's purpose. It features four main buttons: "Submit", "Download", "Browse", and "Document". Each button has an icon and a brief description below it.

国家生物信息中心  
GSA  
Genome Sequence Archive

Data Resources

GSA e.g., CRA000112; CRX006656; SRX10535559; human

Home Submit Browse Search Statistics Support

**GSA News:** GSA has collected all of the metadata of the genome sequence data in INSDC and provides global

**Genome Sequence Archive**

The Genome Sequence Archive (GSA) is a data repository for collecting, archiving, managing and sharing raw sequence data, which is the first repository of the genome sequence data with international journal recognition in China.

**Submit** **Download** **Browse** **Document**

Submit data to GSA Download data to your computer Browse publicly available records Find help information and documents



The image contains three screenshots of biological databases. The top two are side-by-side: the left one shows the SRA (Sequence Read Archive) interface with a search bar and a blue background image of a DNA helix; the right one shows the NIH/National Library of Medicine's SRA page with a similar layout. The bottom part shows the ENA (European Nucleotide Archive) homepage, featuring its logo and a teal header with navigation links like "Home", "Submit", "Search", "Rulespace", "About", and "Support". A message at the bottom encourages users to subscribe to the ENA-announce mailing list.

An official website of the United States government [Here's how you know](#)

NIH National Library of Medicine  
National Center for Biotechnology Information

SRA SRA Advanced

SRA - N

Sequence Read

EMBL-EBI home Services

**ENA** European Nucleotide Archive

Home Submit ▾ Search ▾ Rulespace About ▾ Support ▾

We recommend that you subscribe to the ENA-announce mailing list for updates on ENA services.

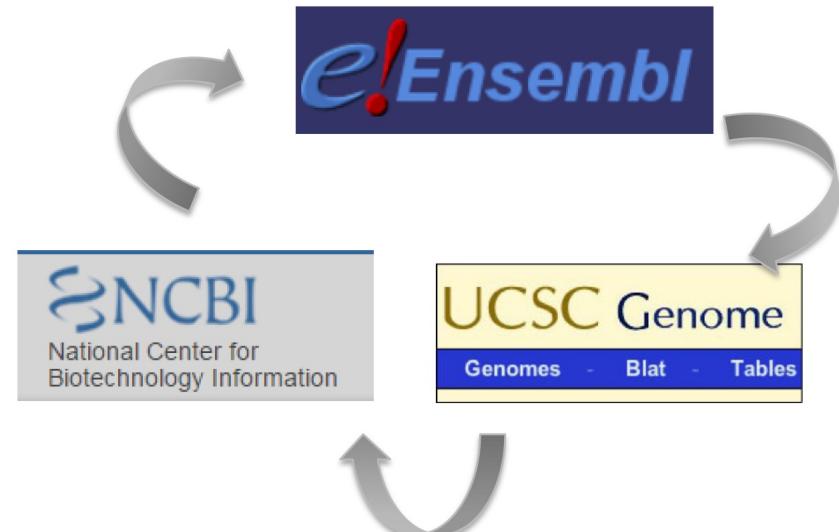
# Publicly available bioinformatics resources

- 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

**ClinVar** Genomic variation as it relates to human health  [Search ClinVar](#)

[Advanced search](#)

[About](#) [Access](#) [Submit](#) [Stats](#) [FTP](#) [Help](#) [Was this helpful?](#)  

[Follow](#)    [Print](#) [Download](#)

[Cite this record](#) 

**NM\_174936.3(PCS<sub>K</sub>9):c.2037C>A (p.Cys679Ter)**

**Interpretation:** Benign

**Review status:**  criteria provided, multiple submitters, no conflicts

**Submissions:** 4 (Most recent: Mar 14, 2019)

**Last evaluated:** Jan 19, 2019

**Accession:** VCV000002877.4

**Variation ID:** 2877

**Description:** single nucleotide variant

# Genetic variation databases - ClinVar

**ClinVar** Genomic variation as it relates to human health  [Search ClinVar](#)

[Advanced search](#)

[About](#) [Access](#) [Submit](#) [Stats](#) [FTP](#) [Help](#) [Was this helpful?](#)  

[Follow](#)    [Print](#) [Download](#)

**NM\_174936.3(PCS<sub>K</sub>9):c.2037C>A (p.Cys679Ter)**  [Cite this record](#)

<b>Interpretation:</b>	Benign
<b>Review status:</b>	 criteria provided, multiple submitters, no conflicts
<b>Submissions:</b>	4 (Most recent: Mar 14, 2019)
<b>Last evaluated:</b>	Jan 19, 2019
<b>Accession:</b>	VCV000002877.4
<b>Variation ID:</b>	2877
<b>Description:</b>	single nucleotide variant

# 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 <a href="#">(Nykamp K et al. (Genet Med 2017))</a> 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 <a href="#">(ACMG Guidelines, 2015)</a> 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](#) [?](#)

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

**Variant details** [?](#)

Conditions

Gene(s)

[FEEDBACK](#)

<b>Allele ID:</b>	46375												
<b>Variant type:</b>	single nucleotide variant												
<b>Variant length:</b>	1 bp												
<b>Cytogenetic location:</b>	13q13.1												
<b>Genomic location:</b>	13: 32337464 (GRCh38) <a href="#">GRCh38 UCSC</a> 13: 32911601 (GRCh37) <a href="#">GRCh37 UCSC</a>												
<b>HGVS:</b>	<table><thead><tr><th>Nucleotide</th><th>Protein</th><th>Molecular consequence</th></tr></thead><tbody><tr><td>NC_000013.11:g.32337464C&gt;T</td><td></td><td></td></tr><tr><td>NC_000013.10:g.32911601C&gt;T</td><td></td><td></td></tr><tr><td>LRG_293t1:c.3109C&gt;T</td><td>LRG_293p1:p.Gln1037Ter</td><td></td></tr></tbody></table> <a href="#">... more HGVS</a>	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	
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												
<b>Protein change:</b>	Q1037*												
<b>Other names:</b>	p.Q1037*:CAA>TAA 3337C>T												

# Genetic variation databases - ClinVar

**NM\_000059.3(BRCA2):c.3109C>T (p.Gln1037Ter)**

[Cite this record](#) [?](#)

<b>Interpretation:</b>	Pathogenic
<b>Review status:</b>	★★★☆ reviewed by expert panel
<b>Submissions:</b>	12 (Most recent: Apr 24, 2019)
<b>Last evaluated:</b>	Sep 8, 2016
<b>Accession:</b>	VCV000037819.5
<b>Variation ID:</b>	
<b>Description:</b>	

**Variant details**

**Conditions**

**Gene(s)**

**FEEDBACK**

	Interpretation (Last evaluated)	Review status (Assertion criteria)	Condition (Inheritance)	Submitter	Supporting information (See all)
<b>NM_000059.3(BRCA2):c.3109C&gt;T (p.Gln1037Ter)</b>	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)	<b>Evidence details</b> Comment: Variant allele predicted to encode a truncated non-functional protein.
<b>Allele</b>					
<b>Varia</b>					
<b>Varia</b>					
<b>Cyto</b>					
<b>Geno</b>					
<b>HGVS</b>					
<b>Prote</b>					
<b>Othe</b>					

**FEEDBACK**

# Genetic variation databases - GnomAD

gnomAD browser    gnomAD v4.1.0    Search    About Team Stats Policies Publications Blog Changelog Downloads Forum Contact Help/FAQ

Help us continue to improve gnomAD by taking 5 minutes to fill out our [user survey](#).

**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

gnomAD v4.1.0    Search by gene, region, or variant

Or

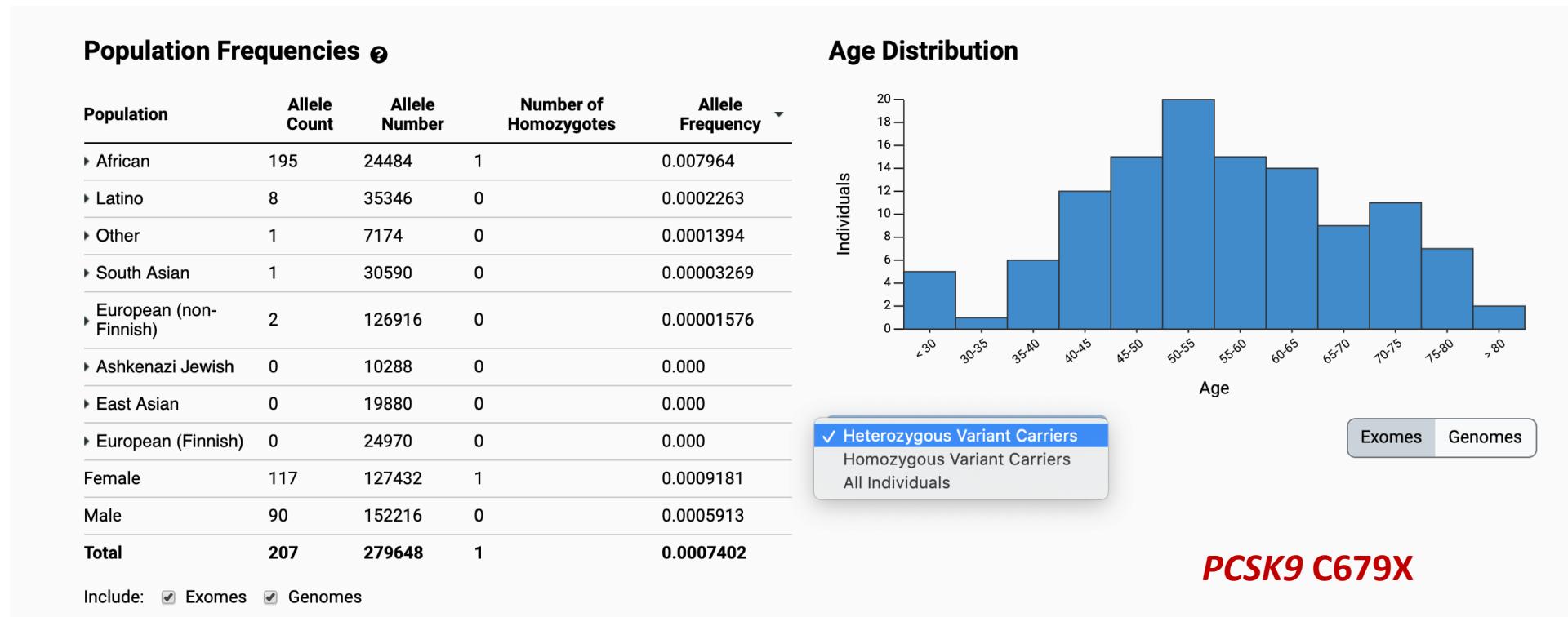
- Download gnomAD data
- Read gnomAD publications
- Find co-occurrence of two variants
- Browse tandem repeats in gnomAD
- Locate features not yet in gnomAD v4

Sample size across major ExAC/gnomAD releases [↗](#)

Release	Exomes	Genomes	Total
ExAC	60,706	0	60,706
gnomAD v2	125,748	0	125,748
gnomAD v3	0	76,156	76,156
gnomAD v4	654,732	76,215	730,947

# Genetic variation databases - GnomAD

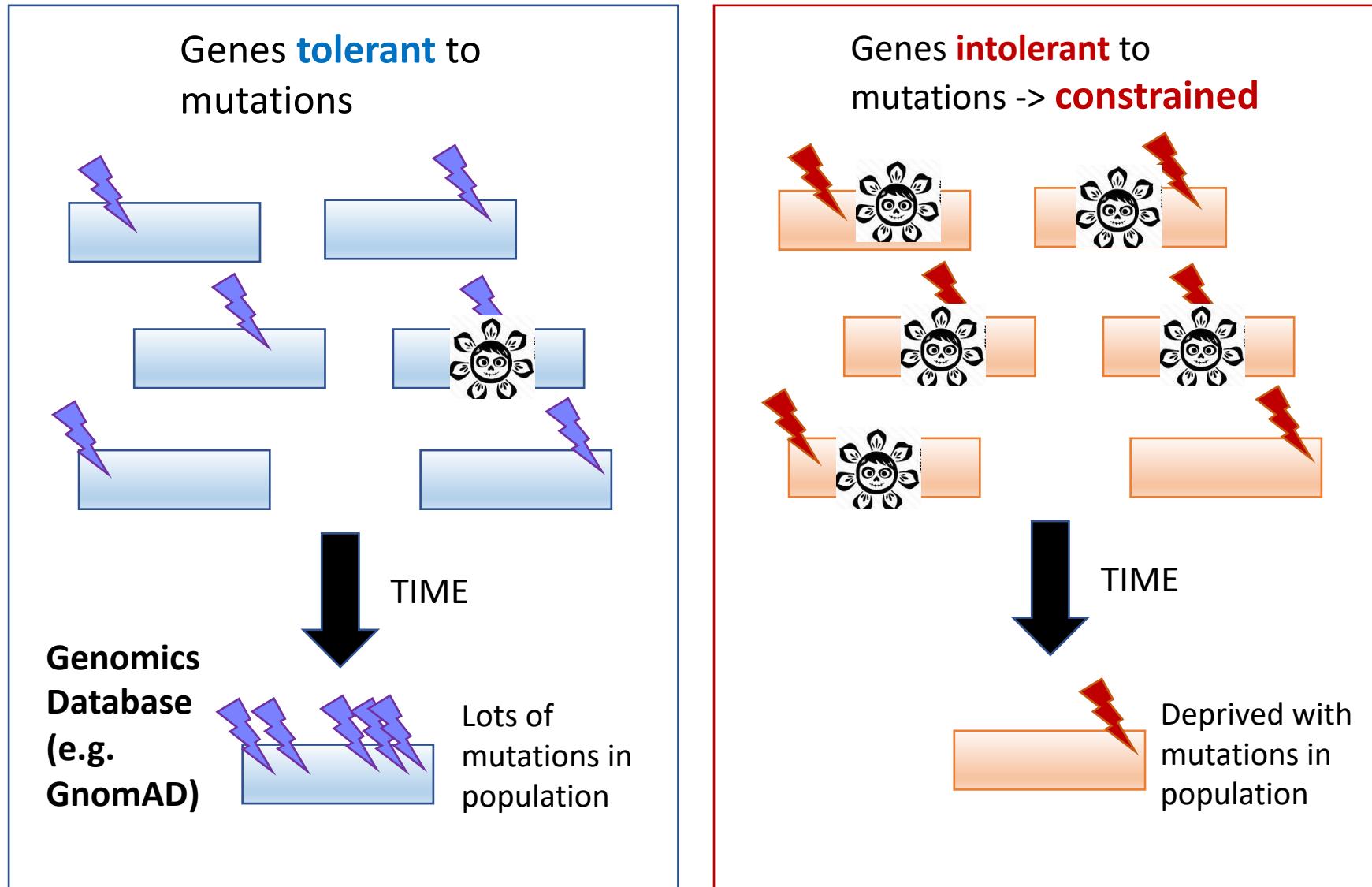
- Known / novel variants (using **dbSNP/gnomAD/others**)
- Allele and genotype frequencies (using **gnomAD/others**)



**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

Category	Expected SNVs	Observed SNVs	Constraint metrics
Synonymous	1104	1012	Z = 1.51 o/e = 0.92 (0.87 - 0.97) 0 ⚡ 1
Missense	3052.9	1988	Z = 7.04 o/e = 0.65 (0.63 - 0.68) 0 ⚡ 1
pLoF	234.8	26	pLI = 1 o/e = 0.11 (0.08 - 0.15) 0 ⚡ 1

## 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

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

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

# Human Traits & Diseases Databases

- OMIM



The screenshot shows the official OMIM website. At the top, there's a blue header with the NIH National Library of Medicine logo and a "Log in" button. Below the header is a search bar with dropdown menus for "OMIM" and "Search", and links for "Limits" and "Advanced". A large banner features a green circular illustration of a classical figure and the word "OMIM". To the right of the banner, the text describes OMIM as a comprehensive compendium of human genes and genetic phenotypes. The main navigation menu below the banner includes links for "About", "Statistics", "Downloads", "Contact Us", "MIMmatch", "Donate", "Help", and a question mark icon. On the right side, there's a "Resources" section with several empty lines for links. The central content area has a large "OMIM®" logo and the text "An Online Catalog of Human Genes and Genetic Disorders". Below this, it says "Updated June 3rd, 2024". A search bar at the bottom allows users to search for clinical features, phenotypes, genes, and more, with a magnifying glass icon. Additional links for "Advanced Search", "Need help?", and "Mirror site" are provided.

NIH National Library of Medicine  
National Center for Biotechnology Information

OMIM OMIM Search Limits Advanced Help

**OMIM**

OMIM is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily. OMIM is authored and edited at the McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, under the direction of Dr. Ada Hamosh. Its official home is [omim.org](http://omim.org).

About Statistics Downloads Contact Us MIMmatch Donate Help ? Resources

**50 YEARS**  
**OMIM**  
Human Genetics Knowledge for the World

**OMIM®**

An Online Catalog of Human Genes and Genetic Disorders

Updated June 3rd, 2024

Search OMIM for clinical features, phenotypes, genes, and more... 

Advanced Search : OMIM, Clinical Synopses, Gene Map

Need help? : Example Searches, OMIM Search Help,  OMIM Video Tutorials

Mirror site : <https://mirror.omim.org>

# Human Traits & Diseases Databases

- <https://www.clinicalgenome.org/>

The screenshot shows the ClinGen Clinical Genome Resource website. At the top, there is a navigation bar with links to Get Started, About Us, Curation Activities, Working Groups, Expert Panels, Documents & Announcements, Tools, and a search icon. Below the navigation bar, a large header reads "Explore the clinical relevance of genes & variants". A subtext explains that ClinGen is a National Institutes of Health (NIH)-funded resource dedicated to building a central resource that defines the clinical relevance of genes and variants for use in precision medicine and research. A search bar at the top allows users to enter a gene symbol or HGNC ID. Below the search bar, a navigation menu includes tabs for All Curated Genes, Gene-Disease Validity (which is highlighted with a red border), Dosage Sensitivity, Clinical Actionability (also highlighted with a red border), Curated Variants (highlighted with a red border), Statistics, More, and a help icon. A prominent teal banner below the menu states "ClinGen is defining the clinical relevance of genes and variants". Below the banner, two gene entries are displayed: Aceruloplasminemia and Acute Intermittent Porphyria. Each entry includes a green circular icon with a plus sign, a unique identifier (AC157 and AC095 respectively), the date it was last updated (Mon, 10 Dec 2018 and Wed, 14 Sep 2016), the source (CP and HMBS), the disease name, its status (Released), associated symptoms (Morbidity due to iron accumulation for Aceruloplasminemia and Neurovisceral attacks for Acute Intermittent Porphyria), and treatment (Iron chelation and avoidance of iron supplementation for Aceruloplasminemia and Optimal clinical management to reduce risk of for Acute Intermittent Porphyria).

ClinGen Clinical Genome Resource

Get Started About Us Curation Activities Working Groups Expert Panels Documents & Announcements Tools

Explore the clinical relevance of genes & variants

ClinGen is a National Institutes of Health (NIH)-funded resource dedicated to building a central resource that defines the clinical relevance of genes and variants for use in precision medicine and research.

Gene Enter a gene symbol or HGNC ID (Examples: ADNP, HGNC:15766)

All Curated Genes **Gene-Disease Validity** Dosage Sensitivity Clinical Actionability Curated Variants Statistics More

**ClinGen is defining the clinical relevance of genes and variants**

Gene	Date	Source	Disease	Status	Symptom	Treatment
AC157	Mon, 10 Dec 2018	CP	Aceruloplasminemia	Released	Morbidity due to iron accumulation	Iron chelation and avoidance of iron supplementation
AC095	Wed, 14 Sep 2016	HMBS	Acute Intermittent Porphyria	Released	Neurovisceral attacks	Optimal clinical management to reduce risk of

# Human Traits & Diseases Databases

## - Gene-disease validity

Gene  Search

All Curated Genes Gene-Disease Validity ▾ Dosage Sensitivity ▾ Clinical Actionability ▾ Curated Variants ▾ Statistics Downloads More ▾ ? ▾

 Hypertrophic Cardiomyopathy Expert Panel 42 Total Curations [Return to Listing](#)

Search in table  Showing 1 to 25 of 42 rows 25 rows per page 1 2

Gene	Disease	MOI	SOP	Contribution	Classification	Last Eval.
ACTA1	hypertrophic cardiomyopathy	AD	SOP7	Primary	No Known Disease Relationship	05/01/2020
ACTC1	hypertrophic cardiomyopathy	AD	SOP8	Primary	Definitive	06/23/2021
ACTN2	intrinsic cardiomyopathy	AD	SOP5	Primary	Moderate	08/06/2018
ALPK3	hypertrophic cardiomyopathy	AR	SOP4	Primary	Strong	02/07/2017
ANKRD1	hypertrophic cardiomyopathy	AD	SOP7	Primary	Limited	05/01/2020

# Human Traits & Diseases Databases

- Clinical actionability: the findings should be medically actionable

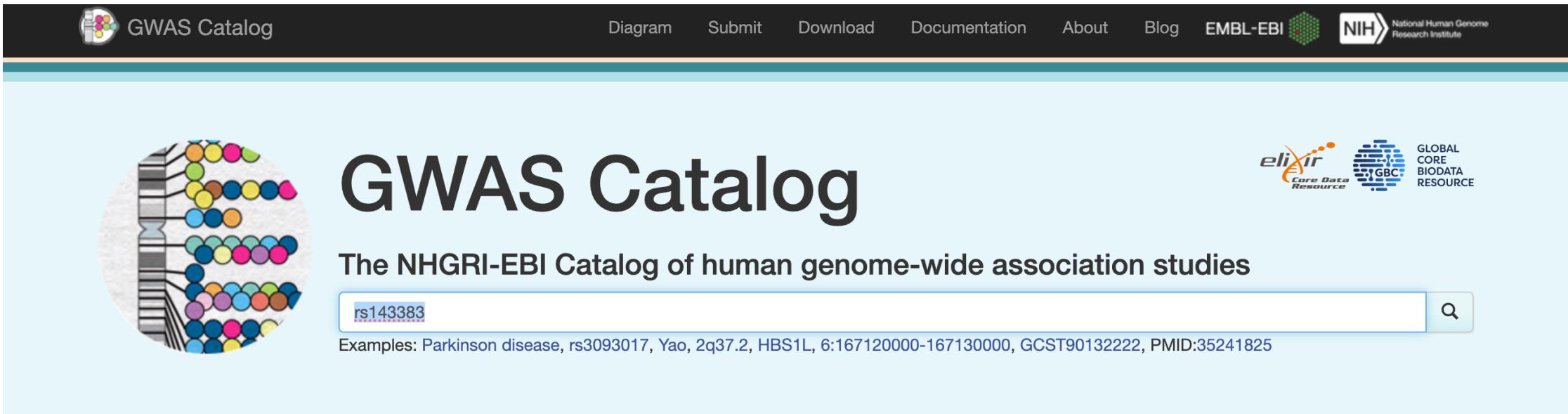
The screenshot shows a web browser displaying the ClinGen Actionability Knowledge Repository. The page title is "ACTIONABILITY KNOWLEDGE REPOSITORY". The main content area is titled "Adult Actionability Reports - Outcome-Intervention Pairs" and includes tabs for "Adult Summaries" and "Pediatric Summaries". A "Home" button is located in the top right corner of the main header.

The main content area features a table with the following columns:

Doc	Latest Search	Genes-scoring Group	Disease	Status-overall	Outcome	Intervention	Final-severity
Id	Date						
Aceruloplasminemia							
AC157	Mon, 10 Dec 2018	CP	Aceruloplasminemia	Released	Morbidity due to iron accumulation	Iron chelation and avoidance of iron supplementation	2
Acute Intermittent Porphyria							
AC095	Wed, 14 Sep 2016	HMBS	Acute Intermittent Porphyria	Released	Neurovisceral attacks	Optimal clinical management to reduce risk of	2

# Human Traits & Diseases Databases

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



The screenshot shows the homepage of the GWAS Catalog. At the top, there is a dark navigation bar with the "GWAS Catalog" logo, links for "Diagram", "Submit", "Download", "Documentation", "About", "Blog", and logos for "EMBL-EBI" and "NIH". Below the navigation bar, there is a large graphic of a chromosome with colored dots representing genetic variants. To the right of the graphic, the title "GWAS Catalog" is displayed in large, bold, black font. Below the title, the subtitle "The NHGRI-EBI Catalog of human genome-wide association studies" is shown. A search bar contains the text "rs143383" and a magnifying glass icon. Below the search bar, there is a link to an example entry: "Examples: Parkinson disease, rs3093017, Yao, 2q37.2, HBS1L, 6:167120000-167130000, GCST90132222, PMID:35241825". On the right side of the page, there are logos for "elixir Core Data Resource" and "GLOBAL CORE BIODATA RESOURCE".

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

Refine search results ^

V Variants 1  
G Genes 1

Catalog stats

- Last data release on 2024-05-20
- 6868 publications
- 330006 SNPs
- 619964 top associations
- 86880 full summary statistics
- Genome assembly GRCh38.p14
- dbSNP Build 156
- Ensembl Build 112
- EFO Version v3.66.0

## Search results for rs143383

V rs143383

Location: 20:35438203 Cytogenetic region:20q11.22 Most severe consequence: 5 prime utr variant Mapped gene(s): GDF5

Associations 11 Studies 11

G GDF5

Description: growth differentiation factor 5

Location: 20:35433347-35454746 Cytogenetic region: 20q11.22 Biotype: protein coding

Associations 135 Studies 102

# Human Traits & Diseases Databases

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

The screenshot shows the GWAS Catalog website interface. At the top, there is a navigation bar with links for Search, Diagram, Submit, Download, Documentation, About, Blog, EMBL-EBI, and NIH. Below the navigation bar, there are tabs for Available data: Associations (11), Studies (11), Traits (5), and Linkage disequilibrium (LD). On the right, there is a "Download Associations" button. The main content area is titled "Associations (11)". It includes a search bar with "Show 5 entries" and buttons for Column visibility, Export, and Clear search. A feedback link is located on the right side of this section. The data is presented in a table with the following columns: Variant and risk allele, P-value, P-value annotation, RAF, OR, Beta, CI, Mapped gene, Reported trait, Trait(s), Background trait(s), Study accession, and Location. The table contains five rows of data for variants rs143383-T, rs143383-G, rs143383-?, rs143383-?, and rs143383-A. The last row is partially visible. At the bottom, it says "Showing 1 to 5 of 11 entries" and has a page navigation bar with links for 1, 2, 3, and ».

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	$8 \times 10^{-18}$	-	0.64	1.09287	-	[NR]	GDF5	Knee osteoarthritis	osteoarthritis, knee	-	GCST006925	20:35438203
rs143383-G	$1 \times 10^{-17}$	-	0.3608	-	-	-	GDF5	Cortical surface area	cortical surface area measurement	-	GCST90091060	20:35438203
rs143383-?	$5 \times 10^{-7}$	-	NR	1.0495905	-	[1.03-1.07]	GDF5	Knee osteoarthritis	osteoarthritis, knee	-	GCST007090	20:35438203
rs143383-?	$1 \times 10^{-9}$	-	NR	-	0.48 unit increase	[0.32-0.64]	GDF5	Height	body height	-	GCST90090967	20:35438203
rs143383-A	$1 \times 10^{-12}$	-	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

The screenshot shows the GWAS Catalog website interface. At the top, there is a navigation bar with links for Search, Diagram, Submit, Download, Documentation, About, Blog, EMBL-EBI, and NIH. Below the navigation bar, there are tabs for Available data: Associations (11), Studies (11), Traits (5), and Linkage disequilibrium (LD). On the right, there is a "Download Associations" button. The main content area is titled "Associations (11)". It includes a search bar with "Show 5 entries" and buttons for Column visibility, Export, and Clear search. A feedback link is located on the right side of this section. The data is presented in a table with the following columns: Variant and risk allele, P-value, P-value annotation, RAF, OR, Beta, CI, Mapped gene, Reported trait, Trait(s), Background trait(s), Study accession, and Location. The table contains five rows of data for variants rs143383-T, rs143383-G, rs143383-?, rs143383-?, and rs143383-A. The last row is partially visible. At the bottom, it says "Showing 1 to 5 of 11 entries" and has a page navigation bar with links for 1, 2, 3, and ».

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	$8 \times 10^{-18}$	-	0.64	1.09287	-	[NR]	GDF5	Knee osteoarthritis	osteoarthritis, knee	-	GCST006925	20:35438203
rs143383-G	$1 \times 10^{-17}$	-	0.3608	-	-	-	GDF5	Cortical surface area	cortical surface area measurement	-	GCST90091060	20:35438203
rs143383-?	$5 \times 10^{-7}$	-	NR	1.0495905	-	[1.03-1.07]	GDF5	Knee osteoarthritis	osteoarthritis, knee	-	GCST007090	20:35438203
rs143383-?	$1 \times 10^{-9}$	-	NR	-	0.48 unit increase	[0.32-0.64]	GDF5	Height	body height	-	GCST90090967	20:35438203
rs143383-A	$1 \times 10^{-12}$	-	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/>

The image shows the UK Biobank website homepage. At the top, there is a navigation bar with three buttons: "Researcher log in" (purple), "Participant log in" (orange), and "Contact us" (grey). Below the navigation bar, there is a search bar with the placeholder "Search" and a magnifying glass icon.

The main banner features the UK Biobank logo and tagline "Enabling scientific discoveries that improve human health". It also includes the text "The world's most important health research database" and a description of the database's purpose: "Data drives discovery. We have curated a uniquely powerful biomedical database that can be accessed globally by approved researchers. Explore de-identified data from half a million UK Biobank participants to enable new discoveries to improve public health." Below the banner are two buttons: "About our data" and "About us".

On the right side of the banner, there is a photograph of a large-scale storage facility with multiple rows of shelving units filled with blue containers.

Below the banner, there are three featured sections:

- Community and Forum**: A dark-themed card with the UK Biobank logo and the text "Community and Forum Online now".
- 500,000 genomes**: A light-colored card featuring a DNA double helix icon and the text "500,000 genomes".
- Global Researcher Access Fund**: A teal-themed card showing a world map with white lines indicating global connectivity.

# Human Traits & Diseases Databases - UKBiobank

- <https://pheweb.org/>

A screenshot of a web browser showing the URL "pheweb.org" in the address bar. Below the address bar is a navigation bar with several links: "Reload via HKUL", "RmBooking", "HPC2021 Dashboa...", "Online Simplified...", "PLINK 1.9", "Capture Reference", "gnomAD browser", "LDlink | An Interac...", and "GWAS Catalog".

## 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/>

A screenshot of a web browser showing the URL "pheweb.org" in the address bar. Below the address bar is a navigation bar with several links: "Reload via HKUL", "RmBooking", "HPC2021 Dashboa...", "Online Simplified...", "PLINK 1.9", "Capture Reference", "gnomAD browser", "LDlink | An Interac...", and "GWAS Catalog".

## 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>

The screenshot shows the 'About' page of the All of Us Research Program website. At the top, there is a dark blue header with the NIH logo, a search bar, and 'LOG IN' and 'Join Now' buttons. Below the header, the All of Us logo is on the left, followed by navigation links: About, Get Involved, Funding and Program Partners, Protecting Data and Privacy, and News and Events. The main title 'About' is centered above the content. The content area begins with a paragraph about the program's goal of better health for all of us. It then describes the mission to accelerate health research and medical breakthroughs through three focus areas: nurturing partnerships, delivering datasets, and catalyzing an ecosystem. A diagram at the bottom illustrates these three interconnected areas.

The *All of Us* Research Program is a historic effort to collect and study data from one million or more people living in the United States. The goal of the program is better health for all of us.

Our mission is to accelerate health research and medical breakthroughs, enabling individualized prevention, treatment, and care for all of us. This mission is carried out through three connected focus areas that are supported and made possible by a team that maintains a culture built around the program's core values.

**Nurture partnerships** for decades with at least a **million participants** who reflect the diversity of the U.S.

**Deliver one of the largest, richest biomedical datasets** that is broadly available and secure

**Catalyze an ecosystem** of communities, researchers, and funders who make *All of Us* an **indispensable** part of health research

# Human Traits & Diseases Databases - All of Us

- <https://allofus.nih.gov>

The screenshot shows the homepage of the All of Us Research Hub Data Browser. At the top, there is a navigation bar with the "All of Us" logo, "Research Hub", "NIH National Institutes of Health All of Us Research Program", and links for "ABOUT", "DATA & TOOLS", "DISCOVER", "SUPPORT", a search icon, and a prominent orange "REGISTER" button. To the right of the register button is a "RESEARCHER LOGIN" button. Below the navigation bar, the page title "Data Browser" is centered in a large, bold, dark blue font. A descriptive paragraph follows, explaining that the data is aggregate-level, derived from multiple sources, and protects participant privacy by removing personal identifiers and rounding counts to 20. It also mentions that summary demographic information is included and individual-level data is available in the Researcher Workbench. Below this text is a section titled "Search Across Data Types" with a "Keyword Search" input field containing the placeholder "Keyword Search". To the right of the search field is a small "x" icon. To the right of the search field is a "FAQ" button enclosed in a light gray box. At the bottom left, there is a link to "EHD Domains".

Home > Data Browser

## Data Browser

Browse aggregate-level data contributed by *All of Us* research participants. Data are derived from multiple [data sources](#). To protect participant privacy, we have removed personal identifiers, rounded aggregate data to counts of 20, and only included summary demographic information. Individual-level data are available for analysis in the [Researcher Workbench](#).

Search Across Data Types i

Keyword Search x

FAQ

Data includes 409,420 participants as of 2/15/2023.

EHD Domains

# Human Traits & Diseases Databases - PheWAS

- <https://www.phewascatalog.org/>

## Phenome Wide Association Studies

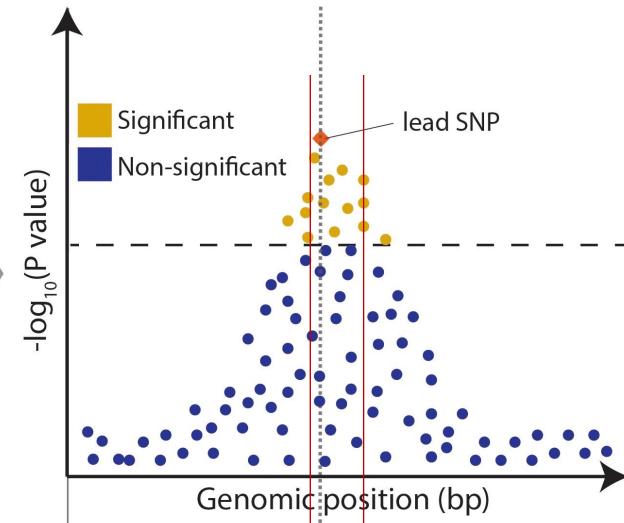
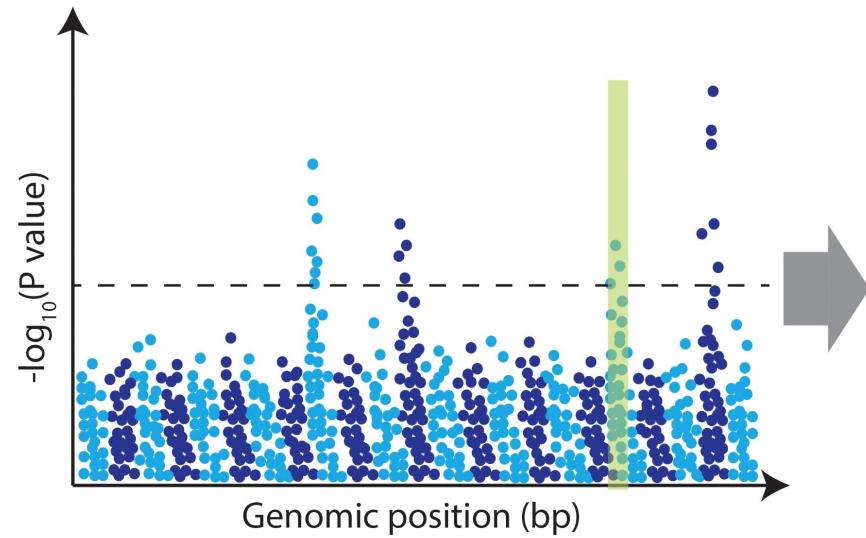
**Phenome-wide association studies (PheWAS)** analyze many phenotypes compared to a single genetic variant (or other attribute). This method was originally described using electronic medical record (EMR) data from EMR-linked in the Vanderbilt DNA biobank, BioVU, but can also be applied to other richly phenotyped sets.

### PheWAS Catalogs

PheWAS of GWAS Catalog of SNPs

Neanderthal PheWAS:  
Discovery & Replication  
Results

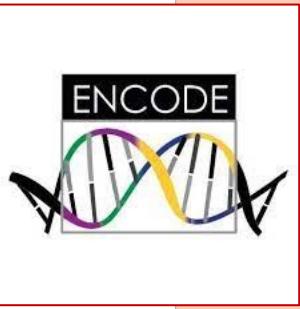
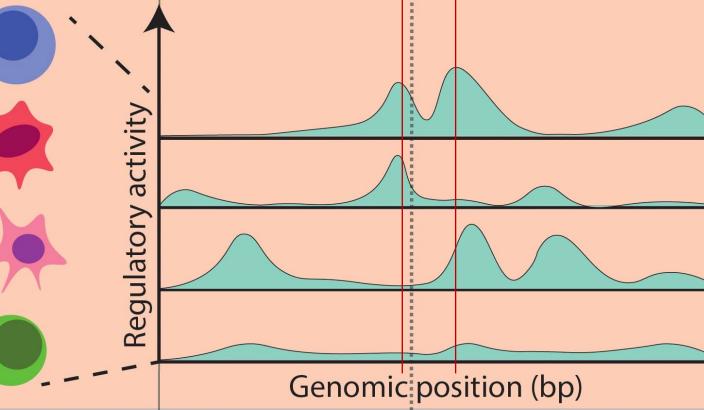
Combined PheWAS - GWAS,  
Neanderthal, HLA



Which are the causal variants?  
**Fine-mapping**

LD ( $r^2$ )

0 1

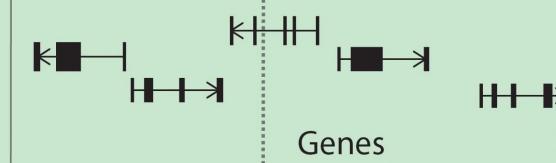


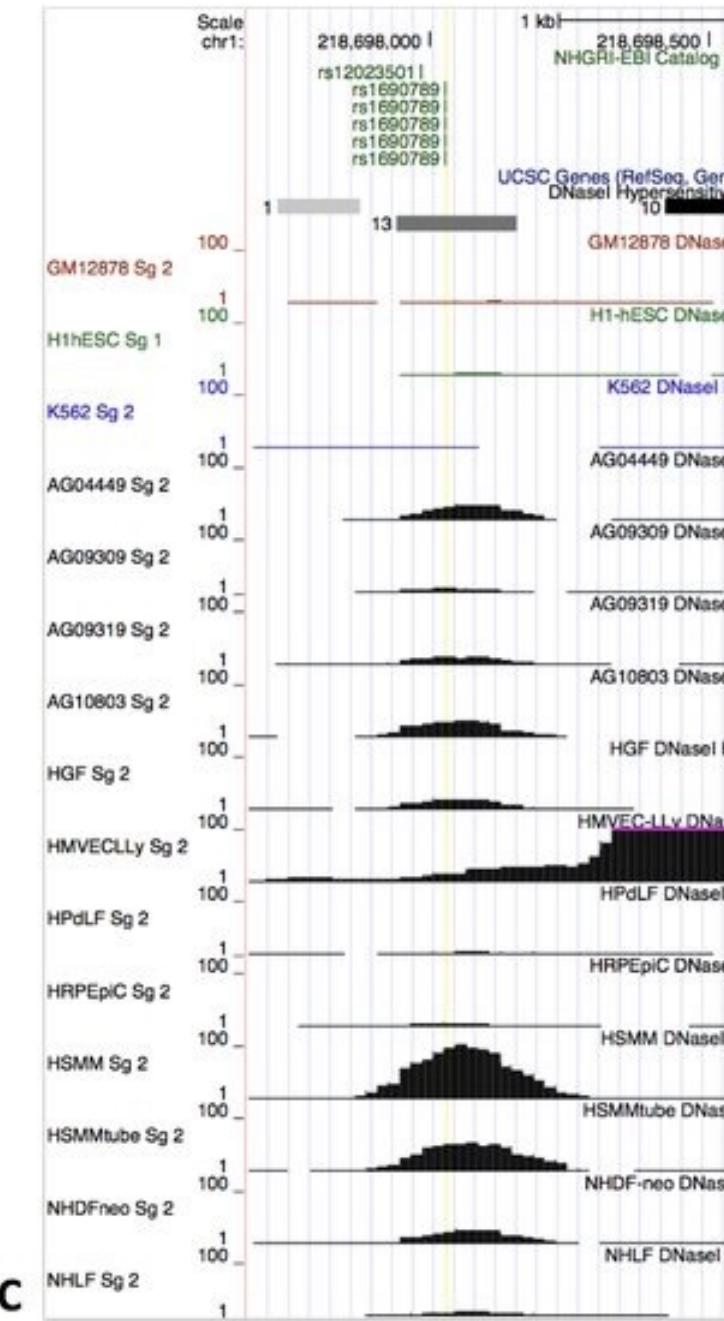
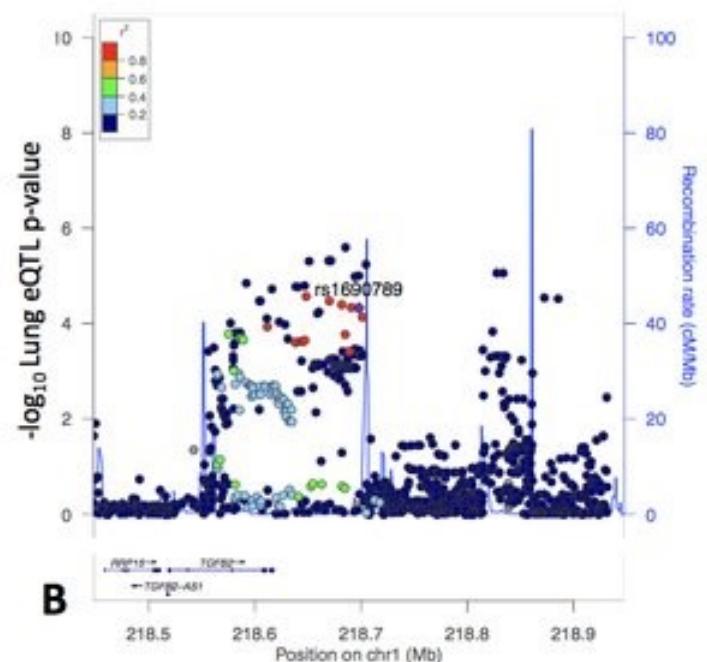
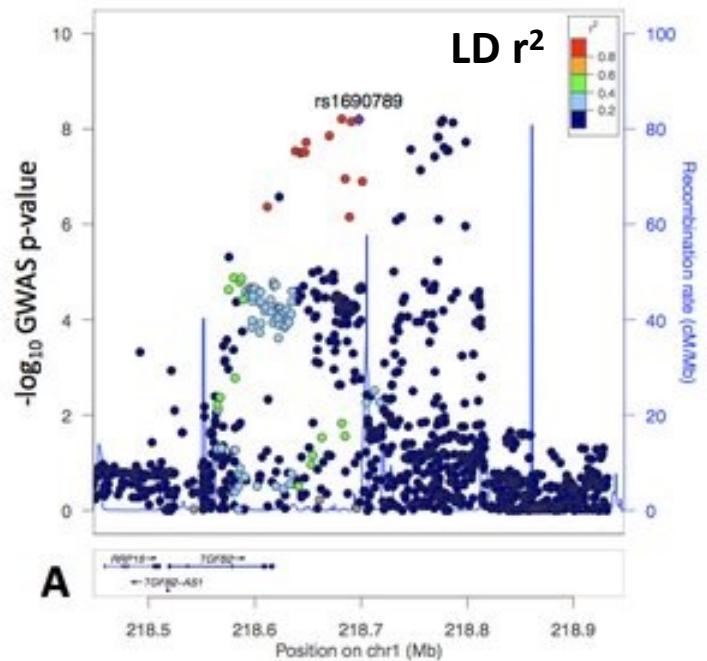
In which cell types do the variants act?  
**SNP enrichment**

<https://www.encodeproject.org/>



Which genes are regulated by the variants?  
**Colocalization**





# Gene Expression Databases - GTEx



Home Downloads ▾ Expression ▾ Single Cell ▾ QTL ▾ IGV Browser Tissues & Histology ▾ Documentation ▾

About GTEx

GTEx

GTEx Consortium

Data Release and  
Publication Policy

## The GTEx Project

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

The Genotype-Tissue Expression (GTEx) project aims to provide to the scientific community which to study human gene expression and regulation and its relationship to genetic variation. GTEx will collect and analyze multiple human tissues from donors who are also densely genotyped, allowing for the analysis of genetic variation within their genomes. By analyzing global RNA expression within individual tissues, correlations between the expression levels of genes as quantitative traits, variations in gene expression correlated with genetic variation can be identified as expression quantitative trait loci, or eQTLs.

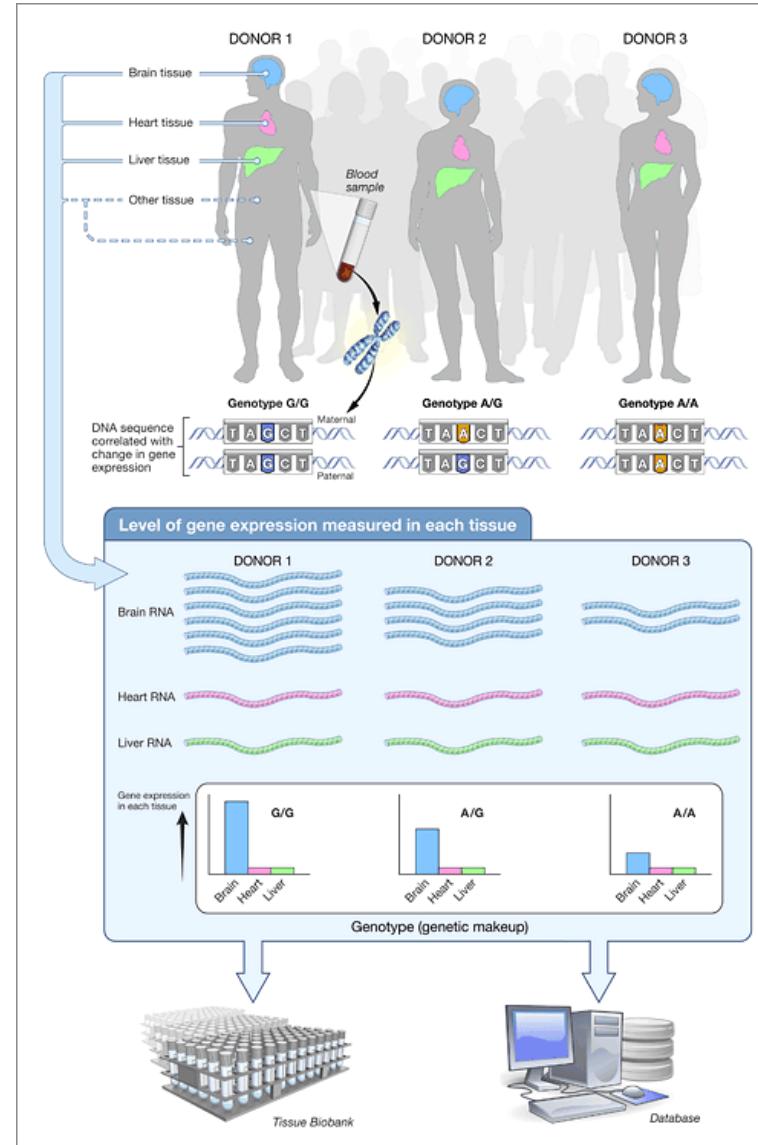
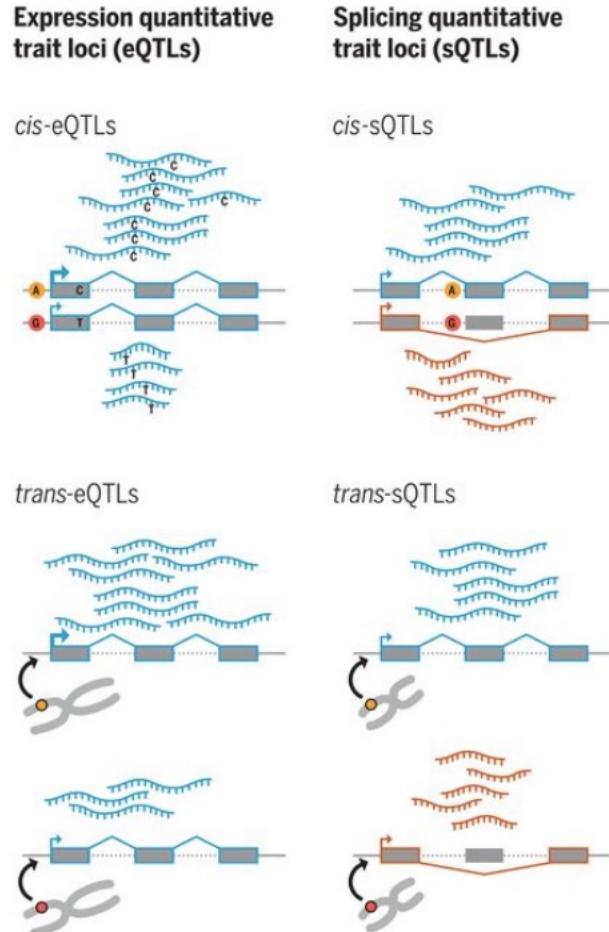
Despite the rapid progress achieved using genome-wide association studies (GWAS; See: <https://www.ebi.ac.uk/gwas/>) to identify genetic changes associated with common diseases such as heart disease, cancer, diabetes, asthma, and stroke, a large majority of these genetic variants lie outside of the protein-coding regions of genes and often even outside of the genes themselves. It is difficult to discern which genes are affected and by what mechanism. The comprehensive catalog of human eQTLs will greatly help to identify genes whose expression is affected by genetic variation and 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 implications of the research, support for statistical methods development, and creation of a database to house the results.

<https://gtexportal.org/home/>

# 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



Home Downloads ▾ Expression ▾ Single Cell ▾ QTL ▾ IGV Browser Tissues & Histology ▾ Documentation ▾

About GTEx

GTEx

GTEx Consortium

Data Release and  
Publication Policy

## The GTEx Project

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

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

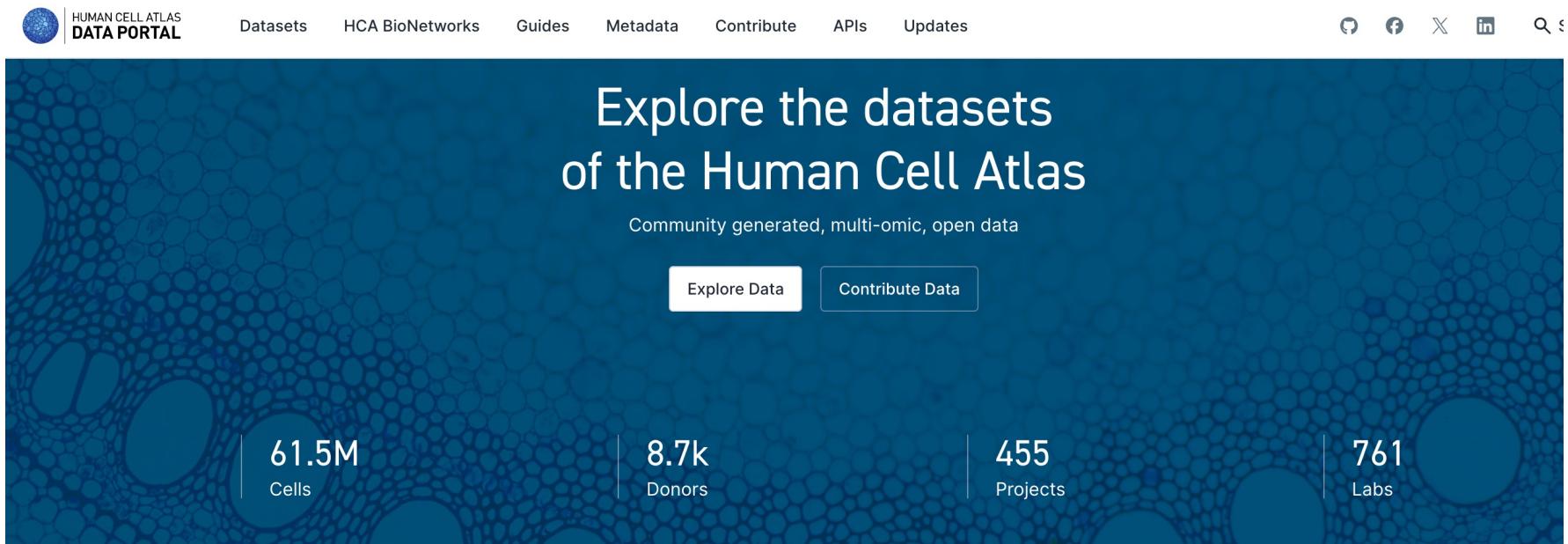
Despite the rapid progress achieved using genome-wide association studies (GWAS; See: <https://www.ebi.ac.uk/gwas/>) to identify genetic changes associated with common diseases such as heart disease, cancer, diabetes, asthma, and stroke, a large majority of these genetic variants lie outside of the protein-coding regions of genes and often even outside of the genes themselves. It is difficult to discern which genes are affected and by what mechanism. The comprehensive catalog of human eQTLs will greatly help to identify genes whose expression is affected by genetic variation and 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 implications of the work, research, support for statistical methods development, and creation of a database to house the data collected.

<https://gtexportal.org/home/>

# Gene Expression Databases - Single cell Atlas

- <https://data.humancellatlas.org/>



## HCA Biological Network Atlases



# Gene Expression Databases - Single cell Atlas

The image shows two screenshots of gene expression databases. The top screenshot is the homepage of the Gut Cell Survey ([gutcellatlas.org](http://gutcellatlas.org)). It features a large logo for "Gut Cell Survey" with a stylized white intestine icon. Below the logo, it says "by Teichlab" and "Wellcome Sanger Institute Part of Human Cell Atlas". The bottom screenshot is the Liver Single Cell Atlas ([liveratlas.vilarinhola.med.yale.edu](http://liveratlas.vilarinhola.med.yale.edu)). It displays a Dimensionality Reduction Plot Options panel with dropdown menus for "Which Plot?" (UMAP) and "Color by" (Cell Type). To the right is a UMAP plot showing clusters of data points color-coded by cell type. A legend on the right lists various cell types: B Cells, Cholangiocytes, Cycling, Endothelial Cells-LSECs, Endothelial Cells-Lymphatic, Endothelial Cells-Macrovacular Arterial, Endothelial Cells-Macrovacular Venous, Hepatic Stellate Cells and Fibroblasts, Hepatocytes, Mononuclear Phagocytes, NK,NKT,T Cells, pDC, Plasma Cells, Undetermined, and Vascular Smooth Muscle Cells. At the bottom is a Bar Plot Options panel with dropdown menus for "Group by" (Cluster) and "Split by".