Nucleic Acids Research, 2026, **54**, D1437–D1445
https://doi.org/10.1093/nar/gkaf1341
Advance access publication date: 8 December 2025
Database issue

The IGVF catalog—from genetic variation to function

Daofeng Li[1,*,†], Shane Liu[1,†], Pedro R. Assis[2,†], Mingjie Li[2,†], Shengcheng Dong[2], Ian Whaling[2], Otto Jolanki[2], Meenakshi Kagda[2], Wenjin Zhang[1], Juan F. Macias-Velasco[1], Tianjie Liu[1], Sarah Cody[3], Lucinda Antonacci-Fulton[3], Yuanhao Huang[4], Jie Liu[4], Michael T. Montgomery[2], Daniel Zeiberg[5], Shantanu Jain[5], Vikas Pejaver[6], Timothy Bergquist[6], Yile Chen[7], Predrag Radivojac[8], Charles A. Gersbach[9], Rintsen N. Sherpa[4], Christopher P. Castro[4], Alan P. Boyle[4], Lea M. Starita[10], Douglas M. Fowler[10], Nadav Ahituv[11], Kushal K. Dey[12], William H. Majoros[13], Timothy E. Reddy[13], Mark Craven[14], Riya Sinha[2], Yuriy Sverchkov[14], Xiangmeng Cai[2], Mpathi Z. Nzima[15], Michael A. Calderwood[16], Joel Rozowsky[17], Mark Gerstein[17,18], Jian Ma[19], Feng Yue[20], J. Michael Cherry[2], Michael I. Love[21,22], Jesse M. Engreitz[2,23,24,25], Benjamin C. Hitz[2,*], Ting Wang[1,3,*]

[1]Department of Genetics, The Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO 63110, United States
[2]Department of Genetics, Stanford University, Stanford, CA 94305, United States
[3]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO 63108, United States
[4]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, United States
[5]Institute for Experiential AI, Northeastern University, Boston, MA 02115, United States
[6]Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY 10029, United States
[7]Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA 98195, United States
[8]Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115, United States
[9]Department of Biomedical Engineering, and Center for Advanced Genomic Technologies, Duke University, Durham, NC 27708, United States
[10]Department of Genome Sciences, University of Washington, Seattle, WA 98195, United States
[11]Department of Bioengineering and Therapeutic Sciences, Institute for Human Genetics, University of California San Francisco, San Francisco, CA 94158, United States
[12]Computational and Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY 10065, United States
[13]Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC 27710, United States
[14]Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI 53706, United States
[15]Cecil H. and Ida Green Center for Reproductive Biology Sciences, University of Texas Southwestern Medical Center, Dallas, TX 75390, United States
[16]Center for Cancer Systems Biology (CCSB) & Department of Cancer Biology, Dana–Farber Cancer Institute and Department of Genetics, Blavatnik Institute, Harvard Medical School, Boston, MA 02215, United States
[17]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, United States
[18]Department of Statistics and Data Science, Yale University, New Haven, CT 06520, United States
[19]Ray and Stephanie Lane Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, United States
[20]Department of Biochemistry and Molecular Genetics, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, United States
[21]Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, United States
[22]Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, United States
[23]Basic Science and Engineering Initiative, Stanford, Children's Health, Betty Irene Moore Children's Heart Center, Stanford, CA 94304, United States
[24]The Novo Nordisk Foundation Center for Genomic Mechanisms of Disease, Broad Institute, Cambridge, MA 02141, United States
[25]Stanford Cardiovascular Institute, Stanford University, Stanford, CA 94305, United States

[*]To whom correspondence should be addressed. Email: dli23@wustl.edu
Correspondence may also be addressed to Benjamin C. Hitz. Email: hitz@stanford.edu
Correspondence may also be addressed to Ting Wang. Email: twang@wustl.edu
[†]The first four authors should be regarded as Joint First Authors.

Abstract

Genomic variation between individuals is essential for understanding how differences in the genome sequence affect molecular and cellular processes. The Impact of Genomic Variation on Function (IGVF) Consortium aims to uncover the relationships among genomic variation, genome function, and phenotypes by combining experimental techniques, such as single-cell mapping and genomic perturbation assays, with computational approaches such as machine learning-based predictive modeling. The IGVF Data and Administrative Coordinating Centers collect, analyze, and disseminate data and results from across the consortium through an open-source platform called the IGVF Catalog. This resource includes, but is not limited to, data on the effects of coding variants on protein abundance and function, noncoding variants on enhancer activity (measured by MPRA or predicted computationally), and associations between variants and quantitative traits. All data are organized within a graph database comprising over 50 types of data collections with nearly 3 billion nodes and over 7.5 billion edges. The Catalog offers public API endpoints (https://api.catalog.igvf.org/) and a user-friendly interface for exploring, querying, and visualizing the data at https://catalog.igvf.org. We expect that this open-access platform will support the broader scientific community to advance our understanding of how genomic variation influences biology and disease.

Received: August 31, 2025. **Revised:** October 26, 2025. **Accepted:** October 28, 2025

Introduction

Genomic variation between individuals provides the foundation for understanding how differences in the genome sequence influence molecular and cellular processes, ultimately shaping human phenotypes and disease susceptibility. While large-scale genomic sequencing projects have cataloged millions of variants, the functional interpretation of these variants, particularly in noncoding regions, remains a major challenge in biological and biomedical sciences [1, 2].

Several widely used resources provide invaluable observational data. For example, gnomAD characterizes population allele frequencies [1], GWAS Catalog aggregates genotype–phenotype associations [3], GTEx links genetic variation to gene expression across tissues [2], ClinVar curates clinically observed variant interpretations [4], and Open Targets integrates multi-omic evidence for drug discovery [5]. However, these databases are primarily descriptive or observational in nature: they document associations but rarely establish mechanistic causality.

The Impact of Genomic Variation on Function (IGVF) Consortium, established by the NHGRI in 2021, was created to fill this gap by systematically uncovering causal relationships between genomic variation, genome function, and phenotypes [6]. IGVF employs a dual strategy: (i) experimental approaches, including single-cell mapping technologies and genomic perturbation assays such as Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) gene editing, massively parallel reporter assays (MPRAs), and saturation mutagenesis; and (ii) computational modeling, leveraging machine learning and artificial intelligence to predict the functional consequences of genomic variants.

To coordinate and disseminate these efforts, the Data and Administrative Coordinating Centers (DACCs) collect, standardize, and integrate data generated across the consortium. These outputs are made available through the IGVF Catalog, an open-source, graph-based resource designed to serve both researchers and computational biologists. The Catalog contains diverse datasets, including the effects of coding variants on protein abundance and function, noncoding variants on enhancer activity (measured experimentally with MPRAs or inferred computationally), and variant-to-trait associations derived from integrative analyses.

All data are organized within a large-scale graph database currently encompassing over 50 types of data collections with nearly 3 billion nodes and over 7.5 billion edges, enabling users to explore interconnected layers of genomic information. To maximize accessibility, the Catalog provides both a public API (https://api.catalogkg.igvf.org) and an interactive user-friendly interface (https://catalog.igvf.org) for searching, querying, and visualizing data in real time. By moving beyond association-based resources toward a framework built on causal experimental evidence and predictive modeling, IGVF aims to accelerate the discovery of mechanisms linking genomic variation to function and to empower the broader scientific community in advancing our understanding of human biology and disease.

Implementation

Backend database design

The knowledge graph

The IGVF Catalog database is designed as a graph in which the nodes represent biological entities or concepts, whereas the edges represent relationships (usually measurements or predictions) between these nodes (Table 1). Many nodes are represented by genomic coordinates (in either GRCh38 for human or GRCm39 for mouse), e.g. variants, genes, transcripts, and genomic elements. Other biochemical entities, such as proteins, protein (or other coding) variants, protein complexes, and drugs, are also represented as nodes. Finally, biological concepts such as tissues, cells, cell lines, phenotypes, diseases, biochemical functions, etc., are represented as nodes derived from common public ontologies (Supplementary Table 1). Informational entities such as GWAS studies, pathways, or protein–DNA binding motifs are also represented as nodes. Edges relate various nodes together; for example, genes are linked to transcripts, which are linked to proteins. Edges between ontology terms are loaded directly from the ontologies themselves, representing parent–child relationships between concepts. Experimental and predicted data (from both the public domain and IGVF) are also represented by edges between nodes; protein–protein interactions from BioGrid [7] and IntAct [8] are edges between two proteins; expression quantitative trait loci (eQTLs) are edges between variants and genes; enhancers predicted to affect gene expression are edges between genomic elements and genes, while chromatin accessibility QTLs (caQTLs) are edges between variants and genomic elements. GWAS results and functional effect predictions are edges between variants and phenotypes (ontology terms). Supplementary Table 2 provides a list of IGVF-generated datasets currently integrated into the knowledgebase.

Graph database implementation

Using a knowledge graph representation has a few advantages over an object or relational database. While the simple API functions that return nodes and edges over 1–2 "hops" could be served from nearly any database, the graph database formulation allows many more deep hop queries and can also be used as input tokens with a large language model (LLM) for retrieval-augmented generation (RAG). We have chosen to implement this graph database in a package called ArangoDB (https://arangodb.com/), which has several advantages over other systems. ArangoDB is not a pure graph database; it can act as an object store or relational database as well, which gives the system significant flexibility. Foremost is the inclusion of an "MDI" multidimensional index, which allows us to efficiently query based on chromosomal coordinate ranges. This is used to find, for example, which genomic elements overlap the position of a given nucleotide variant, or which genomic elements are within 10 kb of a given gene. Without this feature, we would have to precalculate linkages in genomic coordinate space for all nodes with a genomic coordinate. The system runs on a cluster of four memory-optimized cloud computing nodes. We employ node types that are suitable for high-performance databases using the NVMe protocol for highly parallel data transfers with local storage. Data ingestion is done by parsing flat text files (usually TSV, or tab-separated values) using Python scripts into JSONL (https://jsonlines.org/) from all the various data sources, which are then imported into the database.

Table 1. Summary statistics of the data currently loaded into the backend graph database (version 1.0)

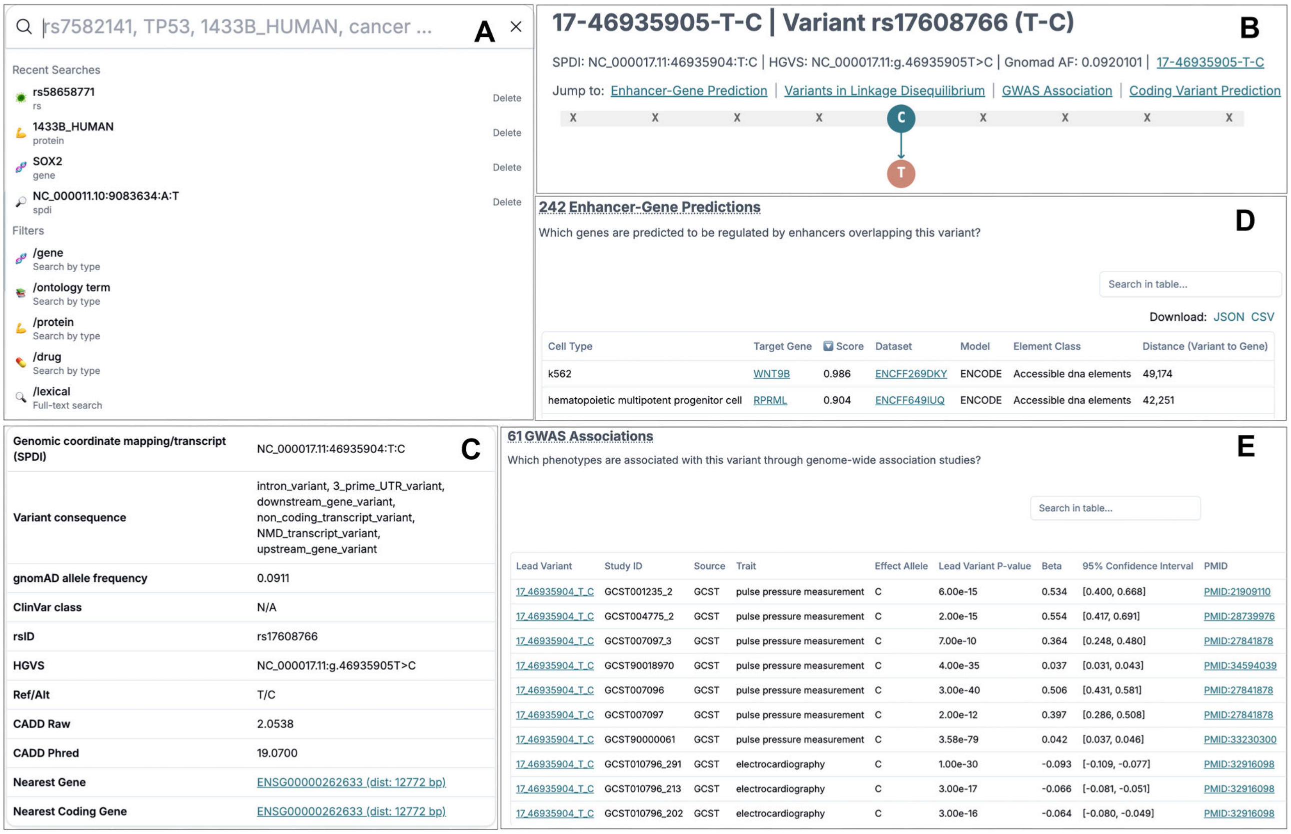
Entity	Type	Arango collection	Source	v1.0
Variants	node	variants	FAVOR (dbSNP 155) [9, 10]+ dbNSFP [11]	1 870 783 419
Non-synonymous variants	node	coding_variants	dbNSFP	942 417 704
Genes	node	genes	GENCODE (v43) [12]	68 881
Transcripts	node	transcripts	GENCODE (v43)	272 993
Proteins	node	proteins	UniProtKB/TrEMBL [13, 14]	188 013
Protein–protein interactions	edge	proteins_proteins	IntAct, BioGRID	11 486 365
Protein complexes	node	complexes	EBI Complex Portal [15]	1687
cCREs	node	genomic_elements	ENCODE [16]	2 348 854
Accessible elements	node	genomic_elements	E2G elements	81 532 493
Accessible elements	node	genomic_elements	caQTLs	24 545
Tested elements	node	genomic_elements	**ENCODE/IGVF MPRA CRISPR**	165 978
Ontology terms	node	ontology_terms	*Various*	620 729
Gene annotations	edge	gene_products_terms	GO	2 518 911
eQTL	edge	variants_genes	eQTL catalogue (30 studies) [17]	13 681 389
Splice–QTL	edge	variants_genes	eQTL catalogue (30 studies)	5 244 210
CRISPR variant effects	edge	variants_genes	**IGVF Variant-EFFECTS**	737
caQTL	edge	variants_genomic_elements	regulomeDB [18]	5597
Variant effect predictions	edge	variants_genomic_elements	**BlueSTARR** [19]	7 553 977
New caQTLs	edge	variants_genomic_elements	AFGR [20], EBI catalogue	460 196
Variant effects	edge	variants_genomic_elements	**IGVF MPRA (5 datasets)**	24 287
Element–gene	edge	variants_biosamples	STARR-seq [21]	36 481 085
LD (4 ancestries)	edge	variants_variants	topLD [22]	5 939 629 733
Gene–transcript	edge	genes_transcripts	GENCODE (v43)	551 368
Protein–transcript	edge	transcripts_proteins	GENCODE + UniProtKB	188 086
Ontology connections	edge	ontology_terms_ontology_terms	all ontologies	2 300 978
Element effect on Gex	edge	genomic_elements_biosamples	element level MPRA	10 814
Element–gene	edge	genomic_elements_genes	ENCODE-E2G [23]	117 039 661
Element–gene	edge	genomic_elements_genes	**CRISPR derived for training**	10 412
Element–gene	edge	genomic_elements_genes	**Perturb-Seq**	1 748 304
Coding variant prediction	edge	coding_variants_phenotypes	**MutPred2** [24]	641 276 378
Coding variant prediction	edge	coding_variants_phenotypes	**ESM1v**	454 527 037
Coding variant prediction	edge	coding_variants_phenotypes	**VAMP-seq (inc Multi)** [25]	183 696
Coding variant effect	edge	variants_phenotypes_coding_variants	**SGE** [26]	15 143
Functional effect prediction	edge	variants_phenotypes	**cV2F** [27]	2 046 460
GWAS	edge	variants_phenotypes_studies	OpenTargets	315 159
GWAS studies	node	studies	OpenTargets	22 690
eQTL studies	node	studies	EBI QTL catalogue	33
Pathways	node	pathways	Reactome [28]	2711
Pathway hierarchy	edge	pathways_pathways	Reactome	2730
Gene–pathway	edge	genes_pathways	Reactome	146 697
Motifs	node	motifs	HOCOMOCO v11/SEMpl [29]	624
Motif–proteins	edge	motifs_proteins	HOCOMOCO v11/SEMpl	2 282
TF allele SB	edge	variants_proteins	AdAstra [30]	1 147 815
TF allele SB	edge	variants_proteins	GVATdb [31]	4 288 875
UKBB plasma pQTLs	edge	variants_proteins	PPP [32]	42 154
Predicted TF allele SB	edge	variants_proteins	**SEMpl** [33]	357 795 432
Co expression	edge	genes_genes	CoXPresdb [34]	3 538 462
Genetic interactions	edge	genes_genes	BioGrid	15 085
Drugs	node	drugs	PharmGKB [35]	4613
Drug–variant	edge	variants_drugs_genes	PharmGKB	24 695
Disease–gene	edge	diseases_genes	Orphanet [36]	8218
Disease–gene–variant	edge	variants_diseases_genes	ClinGen [37]	3820

An entity refers to a specific type of data entry, where an “element” denotes a genomic region. Type indicates whether the entry is represented as a node or an edge in the graph database, with nodes corresponding to individual data entries and edges representing connections between nodes. Arango collection specifies the identifier of the collection in the Arango database, analogous to a table in a relational database. Source denotes the origin of the data, with IGVF-generated datasets highlighted in bold. The final column reports the total counts of each data entity.

Frontend user interface

Unified search bar

The catalog features a unified search bar designed to let users “search by anything” (Fig. 1A). It supports both free-text names and structured identifiers across diverse entities such as genes, proteins, diseases, and studies. The search system provides real-time autocomplete with type-aware suggestions, and a deterministic parser recognizes standard identifiers—including rsIDs, genomic coordinates, SPDI strings [38], PMIDs, and more—directly routing users to the appropriate page. Users can further refine queries with slash commands (e.g. /genes) to confine results to a specific entity type or enter a lexical mode for full-text searching. A persistent search history enables quick re-querying, and the interface showcases scrollable example queries to guide exploration.



rs7582141, TP53, 1433B_HUMAN, cancer ...

Recent Searches

- rs58658771 rs
- 1433B_HUMAN protein
- SOX2 gene
- NC_000011.10:9083634:A:T spdi

Filters

- /gene Search by type
- /ontology term Search by type
- /protein Search by type
- /drug Search by type
- /lexical Full-text search

17-46935905-T-C | Variant rs17608766 (T-C)

SPDI: NC_000017.11:46935904:T:C | HGVS: NC_000017.11:g.46935905T>C | Gnomad AF: 0.0920101 | 17-46935905-T-C

Jump to: Enhancer-Gene Prediction | Variants in Linkage Disequilibrium | GWAS Association | Coding Variant Prediction

X X X X C X X X X

T

242 Enhancer-Gene Predictions

Which genes are predicted to be regulated by enhancers overlapping this variant?

Search in table...

Download: JSON CSV

Cell Type	Target Gene	Score	Dataset	Model	Element Class	Distance (Variant to Gene)
k562	WNT9B	0.986	ENCFF269DKY	ENCODE	Accessible dna elements	49,174
hematopoietic multipotent progenitor cell	RPRML	0.904	ENCFF649IUQ	ENCODE	Accessible dna elements	42,251

Genomic coordinate mapping/transcript (SPDI)	NC_000017.11:46935904:T:C
Variant consequence	intron_variant, 3_prime_UTR_variant, downstream_gene_variant, non_coding_transcript_variant, NMD_transcript_variant, upstream_gene_variant
gnomAD allele frequency	0.0911
ClinVar class	N/A
rsID	rs17608766
HGVS	NC_000017.11:g.46935905T>C
Ref/Alt	T/C
CADD Raw	2.0538
CADD Phred	19.0700
Nearest Gene	ENSG00000262633 (dist: 12772 bp)
Nearest Coding Gene	ENSG00000262633 (dist: 12772 bp)

61 GWAS Associations

Which phenotypes are associated with this variant through genome-wide association studies?

Search in table...

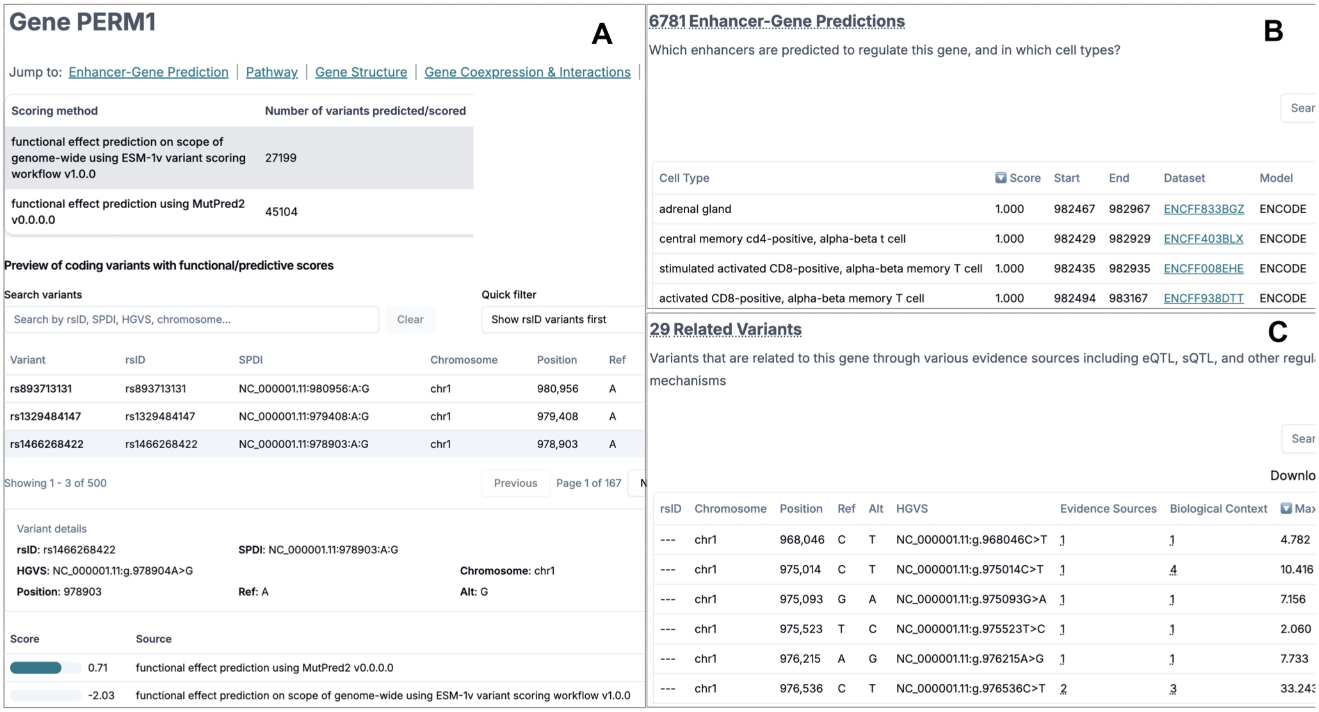
Lead Variant	Study ID	Source	Trait	Effect Allele	Lead Variant P-value	Beta	95% Confidence Interval	PMID
17_46935904_T_C	GCST001235_2	GCST	pulse pressure measurement	C	6.00e-15	0.534	[0.400, 0.668]	PMID:21909110
17_46935904_T_C	GCST004775_2	GCST	pulse pressure measurement	C	2.00e-15	0.554	[0.417, 0.691]	PMID:28739976
17_46935904_T_C	GCST007097_3	GCST	pulse pressure measurement	C	7.00e-10	0.364	[0.248, 0.480]	PMID:27841878
17_46935904_T_C	GCST90018970	GCST	pulse pressure measurement	C	4.00e-35	0.037	[0.031, 0.043]	PMID:34594039
17_46935904_T_C	GCST007096	GCST	pulse pressure measurement	C	3.00e-40	0.506	[0.431, 0.581]	PMID:27841878
17_46935904_T_C	GCST007097	GCST	pulse pressure measurement	C	2.00e-12	0.397	[0.286, 0.508]	PMID:27841878
17_46935904_T_C	GCST90000061	GCST	pulse pressure measurement	C	3.58e-79	0.042	[0.037, 0.046]	PMID:33230300
17_46935904_T_C	GCST010796_291	GCST	electrocardiography	C	1.00e-30	-0.093	[-0.109, -0.077]	PMID:32916098
17_46935904_T_C	GCST010796_213	GCST	electrocardiography	C	3.00e-17	-0.066	[-0.081, -0.051]	PMID:32916098
17_46935904_T_C	GCST010796_202	GCST	electrocardiography	C	3.00e-16	-0.064	[-0.080, -0.049]	PMID:32916098

Figure 1. Unified search interface and table-based variant data visualization in the IGVF Catalog. (**A**) The unified search bar on the Catalog homepage provides a centralized entry point for querying any data stored in the Catalog backend database, supporting both variant- and gene-centric searches. (**B**) Start of a variant page, which begins with basic descriptive information about the variant, followed by a navigation panel linking to all associated data tables on the page. (**C**) Example of a variant detail table. (**D**) The table displayed is the enhancer–gene predictions table, which summarizes predicted regulatory interactions. (**E**) The GWAS associations table, which displays which phenotypes are associated with this variant from genome-wide association studies.

Table-based data display

To effectively organize and present the highly interconnected information within the IGVF Catalog graph database, we employ table-based views as the primary mechanism for displaying data associated with individual database items, such as variants (Fig. 1B–E) or genes (Fig. 2). Tables were selected as the default representation because each item in the underlying graph database is typically linked to many other items, and a tabular format provides a structured, scalable, and user-friendly way to summarize these relationships. For example, on a variant detail page, the Catalog displays enhancer–gene predictions associated with that variant in a table format (Fig. 1D). Each row corresponds to a prediction within a specific cell type, while the columns capture key attributes, including the predicted target gene(s), prediction score, and links to the relevant supporting datasets. The table is further enhanced with interactive features such as grouping predictions by cell type, sorting rows by prediction score, and filtering entries by attributes, enabling users to explore the data dynamically. Multiple other types of information related to a variant are also presented in tabular form. For instance, more information about this variant (Fig. 1B and C), GWAS associations (Fig. 1E), coding variant functional predictions, and allelic effects on transcription factor (TF) binding, among others. Each of these tables provides a concise yet comprehensive view of the associated evidence, allowing users to seamlessly navigate between data categories. To facilitate access, in-page navigation controls are provided, enabling users to jump directly to the table of interest without excessive scrolling. By adopting this consistent, table-centric display strategy, the Catalog ensures that diverse and heterogeneous datasets can be explored within a unified framework, improving both accessibility for casual exploration and efficiency for hypothesis-driven research.

Figure 2 illustrates examples of table-based displays within the Gene page of the Catalog. Like the Variant page, in-page navigation is provided to facilitate quick access to different tables containing gene-associated information. At the top of the page, summary tables present coding variants linked to the selected gene, including predicted functional impacts and computational scores (Fig. 2A). Downstream tables provide additional layers of annotation: predicted enhancer–gene links derived from the ENCODE-rE2G model [23], with each row corresponding to a predicted enhancer and its relevant cell type (Fig. 2B); and variants associated with the gene through multiple lines of evidence, such as eQTL, splicing QTL (sQTL), and other regulatory relationships (Fig. 2C). Additional resources, including associated pathways, gene–gene interactions, and co-expression patterns, are also available, ensuring a comprehensive view of the gene's functional and regulatory landscape.



Gene PERM1

Jump to: Enhancer-Gene Prediction | Pathway | Gene Structure | Gene Coexpression & Interactions |

Scoring method	Number of variants predicted/scored
functional effect prediction on scope of genome-wide using ESM-1v variant scoring workflow v1.0.0	27199
functional effect prediction using MutPred2 v0.0.0.0	45104

Preview of coding variants with functional/predictive scores

Search variants: Search by rsID, SPDI, HGVS, chromosome... Clear

Quick filter: Show rsID variants first

Variant	rsID	SPDI	Chromosome	Position	Ref
rs893713131	rs893713131	NC_000001.11:980956:A:G	chr1	980,956	A
rs1329484147	rs1329484147	NC_000001.11:979408:A:G	chr1	979,408	A
rs1466268422	rs1466268422	NC_000001.11:978903:A:G	chr1	978,903	A

Showing 1 - 3 of 500 Previous Page 1 of 167

Variant details

rsID: rs1466268422 **SPDI**: NC_000001.11:978903:A:G

HGVS: NC_000001.11:g.978904A>G **Chromosome**: chr1

Position: 978903 **Ref**: A **Alt**: G

Score	Source
0.71	functional effect prediction using MutPred2 v0.0.0.0
-2.03	functional effect prediction on scope of genome-wide using ESM-1v variant scoring workflow v1.0.0

6781 Enhancer-Gene Predictions

Which enhancers are predicted to regulate this gene, and in which cell types?

Cell Type	Score	Start	End	Dataset	Model
adrenal gland	1.000	982467	982967	ENCFF833BGZ	ENCODE
central memory cd4-positive, alpha-beta t cell	1.000	982429	982929	ENCFF403BLX	ENCODE
stimulated activated CD8-positive, alpha-beta memory T cell	1.000	982435	982935	ENCFF008EHE	ENCODE
activated CD8-positive, alpha-beta memory T cell	1.000	982494	983167	ENCFF938DTT	ENCODE

29 Related Variants

Variants that are related to this gene through various evidence sources including eQTL, sQTL, and other regul mechanisms

rsID	Chromosome	Position	Ref	Alt	HGVS	Evidence Sources	Biological Context	Max
---	chr1	968,046	C	T	NC_000001.11:g.968046C>T	1	1	4.782
---	chr1	975,014	C	T	NC_000001.11:g.975014C>T	1	4	10.416
---	chr1	975,093	G	A	NC_000001.11:g.975093G>A	1	1	7.156
---	chr1	975,523	T	C	NC_000001.11:g.975523T>C	1	1	2.060
---	chr1	976,215	A	G	NC_000001.11:g.976215A>G	1	1	7.733
---	chr1	976,536	C	T	NC_000001.11:g.976536C>T	2	3	33.243

Figure 2. Table-based displays in the Gene page of the IGVF Catalog. (**A**) Summary of coding variants associated with the selected gene, including functional predictions and annotations, with most data generated directly by the IGVF Consortium. Two methods, esm-1v-workflow [19] and MutPred2 [24], are used to show the number of variants predicted/scored. (**B**) Predicted enhancer–gene links from the ENCODE-rE2G model, where each row represents a predicted enhancer and its corresponding cell type. (**C**) Variants linked to the gene through diverse evidence sources, including eQTL, sQTL, and other regulatory relationships.

Other visualization components

The IGVF Catalog provides a diverse suite of interactive, HTML canvas-based visualizations designed to facilitate intuitive exploration of genomic data (Fig. 3A–D). All visualizations retrieve data dynamically from the IGVF API and render responsively to the user's screen, allowing real-time interaction with large datasets. Complementing the tabular displays of gene- and variant-related information (Figs 1 and 2), these visual modules provide an integrated, graphical perspective that highlights functional and mechanistic relationships. On variant pages, users can view an animated edit diagram that depicts nucleotide substitutions, along with a population allele frequency bar chart enhanced with hover-enabled metadata for each population (Fig. 3A). A linkage disequilibrium (LD) [39] heatmap is also available (Fig. 3B), enabling users to assess correlation structures between variants within the same genomic region. On gene pages, multiple complementary visualization modules are provided. These include distribution plots of functional and predictive scores across diverse annotation sources, a fully zoomable and interactive gene-to-gene interaction network that captures co-expression and regulatory relationships (Fig. 3C), and a tree-like plot highlighting pathway enrichments (Fig. 3D).

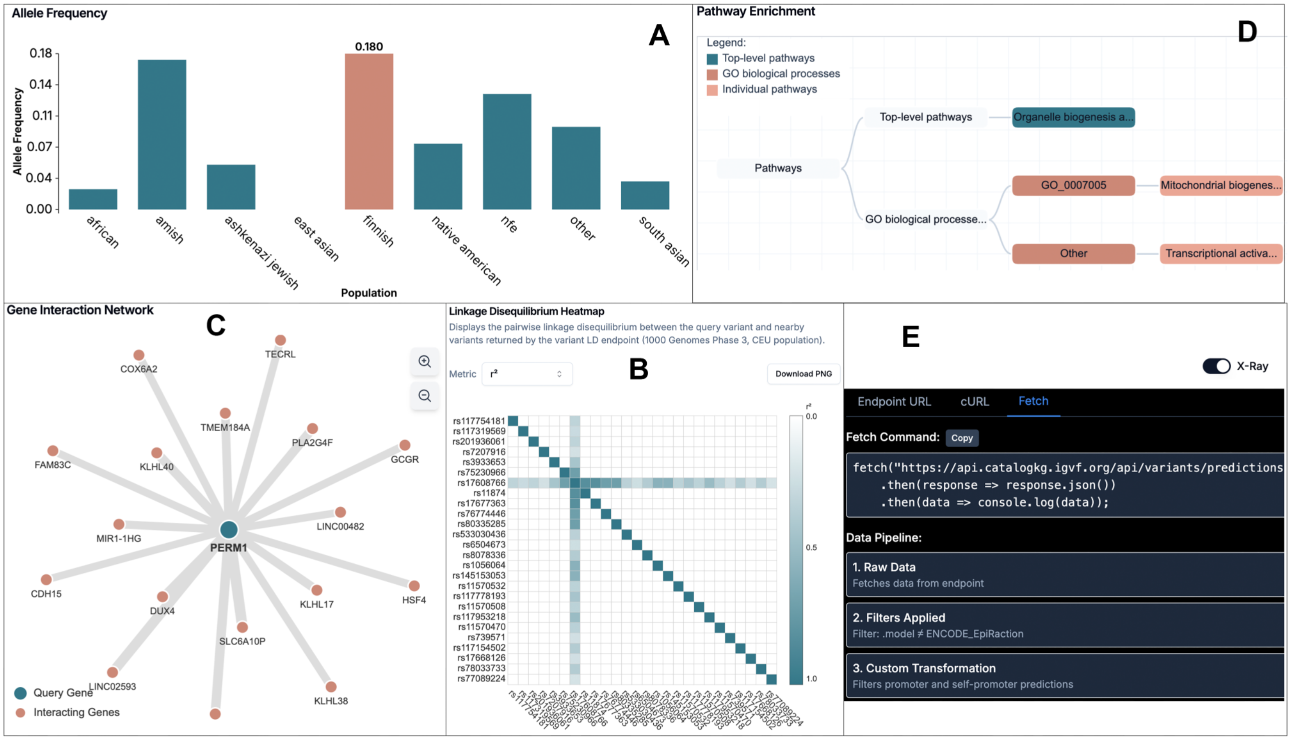
Feedback and glossary system

The Catalog includes a glossary that provides clear contextual definitions for scientific terms used throughout the application. Users can access additional details via hover-enabled tooltips, and each glossary entry includes a "Suggest an Edit" option that opens a prepopulated form, allowing users to propose updated definitions, add notes, and share contact information. Submissions are stored in a database for review, and their status can be tracked as feedback is processed (Supplementary Fig. S1). A password-protected Glossary Editor enables approved maintainers to incorporate updates directly, ensuring that definitions remain current and community-informed. This system not only enhances user comprehension but also serves as a collaborative platform, engaging sister consortia and researchers worldwide in refining and standardizing scientific terminology within the Catalog.

API service and information

The preferred way to access the data in the backend Graph Database is via the API. The IGVF Catalog API exposes access to the data from the Graph Database using internal optimized queries. We provide an HTTP access and a TRPC-based access. The HTTP access is made public using the OpenAPI specification, allowing REST calls from several clients. The TRPC-based access is made available using the tRPC (typescript remote protocol call) protocol written in Typescript. This allows direct RPC calls using the Typescript language as a client, a widely used language to build frontend applications, speeding up development and ingestion of a fully type-safe API. Direct access to the backend Graph Database is supported through the Arango query language (AQL), with a demonstration provided in a Jupyter notebook (Supplementary File 1).

In addition to backend APIs, the IGVF Catalog's frontend includes an "X-ray" feature that exposes the underlying API calls used to generate each table, providing users with full transparency into how data are retrieved and displayed (Fig. 3E). For every supported table, users can toggle an X-



Fetch Command:

```
fetch("https://api.catalogkg.igvf.org/api/variants/predictions
    .then(response => response.json())
    .then(data => console.log(data));
```

Data Pipeline:

1. Raw Data
Fetches data from endpoint

2. Filters Applied
Filter: .model ≠ ENCODE_EpiRaction

3. Custom Transformation
Filters promoter and self-promoter predictions

Figure 3. Interactive visualization components and the X-Ray feature in the IGVF Catalog. (**A**) Dynamic population allele frequency bar chart, where hovering over each bar displays detailed metadata for the corresponding population. **(B)** LD heatmap, with a metrics dropdown menu that enables users to select the measure (r^2 or D^2) used to generate the plot. (**C**) Zoomable gene interaction network, allowing users to explore co-expression and regulatory relationships among genes. (**D**) A tree-like plot illustrating pathway enrichment for a selected gene, providing a hierarchical view of functional annotation. (**E**) X-Ray feature, which exposes the underlying code and query pipeline used to fetch data from the Catalog API, including API endpoints, parameters, and optional client-side transformations, enabling reproducibility and transparency.

ray panel to view the fully resolved query URL (including parameters, pagination, and filters), alongside ready-to-use examples such as a cURL command and a JavaScript “fetch()” snippet for reproducing the query externally. When client-side filtering or transformations are applied, the panel further documents the processing pipeline, showing how raw data are mapped, filtered, enriched, or otherwise transformed before display. Because the table data fetcher and X-ray panel share the same data source, consistency is ensured between what is shown and what is retrieved. This design makes IGVF not only a data exploration platform but also a developer- and researcher-friendly resource, lowering barriers to reusing and integrating its data into external workflows, something rarely offered by other genomic databases.

For users working in R, the *rigvf* package (https://igvf.github.io/rigvf/) offers a convenient interface to explore and analyze Catalog data directly within R/Bioconductor. Users can query the Catalog by variants, genomic elements, and genes using several accepted identifiers, and retrieve the other nodes that are connected in the graph database, with variants, elements, and genes represented using standard Bioconductor classes such as GenomicRanges. Genomic elements, which tend not to have stable identifiers as do variants or genes, may be queried by specifying an extent of the genome.

We have also implemented an experimental “free text” query endpoint within the API. This system employs an LLM–RAG (LLM with RAG) framework via LangChain (https://www.langchain.com/). Specifically, input text is first classified to infer a reduced graph database schema, which is then passed as additional tokens to the ArangoGraphQAChain. This process generates AQL queries that are executed on the database, with the resulting output returned to the LLM to produce a plain-text response (in addition to a data table). The current implementation uses OpenAI’s GPT-4.1 (https://openai.com/) to generate both the graph queries and the final output. While results can be further enhanced by employing larger models or incorporating a broader representation of the graph schema, these approaches require substantially greater computational resources. To facilitate easier navigation of the IGVF Catalog resources, we implemented a chatbot (https://catalog.igvf.org/chat). The chatbot attempts to translate users’ questions into AQL queries whenever possible and returns results directly from our knowledge graph database. This provides a more intuitive interface for users who may not be familiar with query languages. An example of such a query and its corresponding output is shown in Supplementary Fig. S2.

Relationship to the IGVF data portal

The IGVF Data Portal (https://data.igvf.org/) hosts the raw and processed datasets generated by the IGVF consortium, while the IGVF Catalog provides a user-facing interface for querying, visualizing, and integrating the analyzed results produced by consortium working groups alongside public annotations from external databases. IGVF data on the portal is organized primarily around Files and FileSets (representing raw experimental data, processed results, analytical outputs, predictions, and models) covered by extensive metadata.

Release files from AnalysisSets, along with data from public sources, are the raw material that is loaded into the Catalog Knowledge Graph. Metadata is explored via a faceted search interface where results, resources, and metadata can be viewed and files downloaded. Results are presented in a configurable, paginated table with adjustable columns, ontology-linked references, and export options in JSON or CSV. In addition to serving as the basis for the IGVF Catalog, the IGVF data portal is a web application that enables the exploration of raw data, integrated annotations, and streamlines access to the diverse experimental outputs generated by IGVF.

Comparison with other similar databases

Most widely used genomic resources, such as gnomAD [1] (population-scale variant frequencies), GWAS Catalog [3] (variant–trait associations), GTEx [2] (expression QTLs across tissues), ClinVar [4] (clinical variant annotations), and Open Targets Platform [5, 40] (integration of genetic associations with functional genomics to prioritize causal variants and genes), focus on different sets of data or user groups. These databases provide critical insights about naturally occurring human genetic variants and their associations with molecular phenotypes or disease.

The IGVF Catalog builds on some of these resources and incorporates unique experimental datasets and predictive models generated by the IGVF Consortium. These datasets include genomic perturbation experiments that build on observational resources to directly probe causal mechanisms of genomic variation.

While MaveDB [41] focuses on results from multiplexed assays of variant effects, which experimentally assess the functional consequences of thousands of variants in parallel, IGVF expands on this concept with a broader suite of experimental strategies—including CRISPR-based perturbations, single-cell transcriptomics, and chromatin accessibility profiling. By integrating these complementary methods, IGVF is uniquely positioned to move beyond single-variant effects and uncover the regulatory architecture and gene networks that shape phenotypic outcomes.

All this information is organized into a unified, graph-based framework that connects functional, genetic, and genomic resources, enabling cross-linked exploration across data types. In this way, the IGVF Catalog enables linking genomic variation to molecular and cellular phenotypes through both experimental and predictive insights.

Limitations and outlook

The IGVF Catalog is an actively evolving resource, and several limitations remain. Data coverage is still incomplete, with ongoing integration of multiple datasets (e.g. Biobank OR results and MaveDB-derived assays). Some interactive features and autocomplete functionality, are being refined to improve performance and search accuracy. Clearer labeling of IGVF-generated versus public data is also needed to enhance interpretability. Finally, documentation and user guidance are under active development and revision, and feedback mechanisms have only recently been expanded to accommodate broader community participation.

The DACCs are continually enhancing both the backend database and frontend interface of the IGVF Catalog; thus, future updates will address these limitations by incorporating additional datasets (including both new datasets generated by IGVF consortium and other resources from the community), optimizing query and visualization performance, implementing explicit data source labels, embedding contextual documentation, and refining feedback and glossary systems. Regular updates and subsequent publications will ensure the Catalog remains an up-to-date, comprehensive, and user-friendly resource.

Acknowledgements

We are deeply grateful to the anonymous reviewers for their constructive feedback, which has greatly enhanced both the quality of the manuscript and the usability of the Catalog application. We also sincerely thank the IGVF members who contributed to the development of the data processing pipelines, file formats, and data analyses, as well as those who submitted data to the Catalog knowledge graph database. Finally, we are thankful to the NHGRI staff for their invaluable guidance and support.

Author contributions: Software and visualization: D.L., S.L., P.R.A., M.L., S.D.

Data curation: I.W., M.K., M.T.M., Y.H., J.L., D.Z., S.J., V.P., T.B., Y.C., P.R., C.A.G, R.N.S., C.P.C., A.P.B., L.M.S., D.M.F., N.A., K.K.D., W.H.M., T.E.R., M.C., R.S., X.C., M.Z.N., M.A.C., Y.S.

Resources: W.Z., J.F.M.V., T.L., S.C., L.A.F., J.R., M.G., J.M., F.Y., J.M.C.

Writing—original draft and Writing—review & editing: D.L., M.I.L., J.M.E., B.C.H, T.W.

Supervision: M.I.L., J.M.E., B.C.H, T.W.

Funding acquisition: B.C.H., T.W.

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This work was supported by the NIH NHGRI IGVF Program (UM1HG011966, UM1HG011969, UM1HG011972, UM1HG011989, UM1HG011996, UM1HG012003, UM1HG012010, UM1HG012053, UM1HG011986, UM1HG012076, UM1HG012077, U01HG011952, U01HG011967, U01HG012009, U01HG012022, U01HG012039, U01HG012064, U01HG012069, U01HG012041, U01HG012047, U01HG012051, U01HG012059, U01HG012079, U01HG012103, U24HG012012, U24HG012070) Funding to the Open Access publication charges for this article is provided by NHGRI HG012070.

Data availability

The IGVF Catalog is freely available at https://catalog.igvf.org/, with detailed documentation and user guidance provided at https://docs.catalog.igvf.org. The Catalog's underlying data can also be accessed programmatically through its API service at https://api.catalogkg.igvf.org/, which includes comprehensive reference materials. All of the code for the IGVF catalog backend, including database and API, is available at Zenodo (10.5281/zenodo.17468372).

References

1. Karczewski KJ, Francioli LC, Tiao G *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;581:434–43. https://doi.org/10.1038/s41586-020-2308-7
2. Consortium GT. The GTEx Consortium Atlas of genetic regulatory effects across human tissues. *Science* 2020;369:1318–30. https://doi.org/10.1126/science.aaz1776
3. Cerezo M, Sollis E, Ji Y *et al.* The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity. *Nucleic Acids Res* 2025;53:D998–D1005. https://doi.org/10.1093/nar/gkae1070
4. Landrum MJ, Lee JM, Benson M *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 2018;46:D1062–7. https://doi.org/10.1093/nar/gkx1153
5. Buniello A, Suveges D, Cruz-Castillo C *et al.* Open Targets Platform: facilitating therapeutic hypotheses building in drug discovery. *Nucleic Acids Res* 2025;53:D1467–75. https://doi.org/10.1093/nar/gkae1128
6. Consortium IGVF. Deciphering the impact of genomic variation on function. *Nature* 2024;633:47–57. https://doi.org/10.1038/s41586-024-07510-0
7. Oughtred R, Stark C, Breitkreutz BJ *et al.* The BioGRID interaction database: 2019 update. *Nucleic Acids Res* 2019;47:D529–41. https://doi.org/10.1093/nar/gky1079
8. Orchard S, Ammari M, Aranda B *et al.* The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucl. Acids Res.* 2014;42:D358–63. https://doi.org/10.1093/nar/gkt1115
9. Zhou H, Arapoglou T, Li X *et al.* FAVOR: functional annotation of variants online resource and annotator for variation across the human genome. *Nucleic Acids Res* 2023;51:D1300–11. https://doi.org/10.1093/nar/gkac966
10. Sherry ST, Ward MH, Kholodov M *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001;29:308–11. https://doi.org/10.1093/nar/29.1.308
11. Liu X, Li C, Mou C *et al.* dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med* 2020;12:103. https://doi.org/10.1186/s13073-020-00803-9
12. Mudge JM, Carbonell-Sala S, Diekhans M *et al.* GENCODE 2025: reference gene annotation for human and mouse. *Nucleic Acids Res* 2025;53:D966–75. https://doi.org/10.1093/nar/gkae1078
13. Boeckmann B, Bairoch A, Apweiler R *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 2003;31:365–70. https://doi.org/10.1093/nar/gkg095
14. Boutet E, Lieberherr D, Tognolli M *et al.* UniProtKB/Swiss-Prot. *Methods Mol Biol* 2007;406:89–112.
15. Meldal BHM, Perfetto L, Combe C *et al.* Complex Portal 2022: new curation frontiers. *Nucleic Acids Res* 2022;50:D578–86. https://doi.org/10.1093/nar/gkab991
16. Project Consortium ENCODE. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489:57–74. https://doi.org/10.1038/nature11247
17. Kerimov N, Tambets R, Hayhurst JD *et al.* eQTL Catalogue 2023: new datasets, X chromosome QTLs, and improved detection and visualisation of transcript-level QTLs. *PLoS Genet* 2023;19:e1010932. https://doi.org/10.1371/journal.pgen.1010932
18. Dong S, Zhao N, Spragins E *et al.* Annotating and prioritizing human non-coding variants with RegulomeDB v. *Nat Genet* 2023;55,724–6. https://doi.org/10.1038/s41588-023-01365-3
19. Sverchkov Y. ESM-1v predictions for all AA substitutions in all MANE proteins. February 2025, *Zenodo*, https://zenodo.org/doi/10.5281/zenodo.14828609
20. DeGorter MK, Goddard PC, Karakoc E *et al.* Transcriptomics and chromatin accessibility in multiple African population samples. bioRxiv, https://doi.org/10.1101/2023.11.04.564839, 6 November 2023, preprint: not peer reviewed.
21. Arnold CD, Gerlach D, Stelzer C *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 2013;339:1074–7. https://doi.org/10.1126/science.1232542
22. Taliun D, Harris DN, Kessler MD *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 2021;590:290–9. https://doi.org/10.1038/s41586-021-03205-y
23. Gschwind AR, Mualim KS, Karbalayghareh A *et al.* An encyclopedia of enhancer-gene regulatory interactions in the human genome. bioRxiv, https://doi.org/10.1101/2023.11.09.563812, 23 February 2025, preprint: not peer reviewed.
24. Pejaver V, Urresti J, Lugo-Martinez J *et al.* Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nat Commun* 2020;11:5918. https://doi.org/10.1038/s41467-020-19669-x
25. Matreyek KA, Starita LM, Stephany JJ *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat Genet* 2018;50:874–82. https://doi.org/10.1038/s41588-018-0122-z
26. Findlay GM, Daza RM, Martin B *et al.* Accurate classification of BRCA1 variants with saturation genome editing. *Nature* 2018;562:217–22. https://doi.org/10.1038/s41586-018-0461-z
27. Fabiha T, Raine I, Kundu S *et al.* A consensus variant-to-function score to functionally prioritize variants for disease. bioRxiv, https://doi.org/10.1101/2024.11.07.622307, 10 Novermber 2024, preprint: not peer reviewed.
28. Milacic M, Beavers D, Conley P *et al.* The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Res* 2024;52:D672–8. https://doi.org/10.1093/nar/gkad1025
29. Vorontsov IE, Eliseeva IA, Zinkevich A *et al.* HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors. *Nucleic Acids Res* 2024;52:D154–63. https://doi.org/10.1093/nar/gkad1077
30. Abramov S, Boytsov A, Bykova D *et al.* Landscape of allele-specific transcription factor binding in the human genome. *Nat Commun* 2021;12:2751. https://doi.org/10.1038/s41467-021-23007-0
31. Yan J, Qiu Y, Ribeiro Dos Santos AM *et al.* Systematic analysis of binding of transcription factors to noncoding variants. *Nature* 2021;591:147–51. https://doi.org/10.1038/s41586-021-03211-0
32. Sun BB, Chiou J, Traylor M *et al.* Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* 2023;622:329–38. https://doi.org/10.1038/s41586-023-06592-6
33. Nishizaki SS, Ng N, Dong S *et al.* Predicting the effects of SNPs on transcription factor binding affinity. *Bioinformatics* 2020;36:364–72. https://doi.org/10.1093/bioinformatics/btz612
34. Obayashi T, Kodate S, Hibara H *et al.* COXPRESdb v8: an animal gene coexpression database navigating from a global view to detailed investigations. *Nucleic Acids Res* 2023;51:D80–7. https://doi.org/10.1093/nar/gkac983
35. Barbarino JM, Whirl-Carrillo M, Altman RB *et al.* PharmGKB: a worldwide resource for pharmacogenomic information. *WIREs Mechanisms of Disease* 2018;10:e1417. https://doi.org/10.1002/wsbm.1417
36. Weinreich SS, Mangon R, Sikkens JJ *et al.* [Orphanet: a European database for rare diseases]. *Ned Tijdschr Geneeskd* 2008;152:518–9.
37. Rehm HL, Berg JS, Brooks LD *et al.* ClinGen—the Clinical Genome Resource. *N Engl J Med* 2015;372:2235–42. https://doi.org/10.1056/NEJMsr1406261

38. Holmes JB, Moyer E, Phan L *et al.* SPDI: data model for variants and applications at NCBI. *Bioinformatics* 2020;36:1902–7. https://doi.org/10.1093/bioinformatics/btz856
39. Slatkin M. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* 2008;9:477–85. https://doi.org/10.1038/nrg2361
40. Ochoa D, Hercules A, Carmona M *et al.* Open Targets Platform: supporting systematic drug-target identification and prioritisation. *Nucleic Acids Res* 2021;49:D1302–10. https://doi.org/10.1093/nar/gkaa1027
41. Esposito D, Weile J, Shendure J *et al.* MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol* 2019;20:223. https://doi.org/10.1186/s13059-019-1845-6

Received: August 31, 2025. **Revised:** October 26, 2025. **Accepted:** October 28, 2025