https://doi.org/10.1093/nar/gkac966

FAVOR: functional annotation of variants online resource and annotator for variation across the human genome

Hufeng Zhou[1,*,†], Theodore Arapoglou[1,†], Xihao Li[1], Zilin Li[1,2], Xiuwen Zheng[3], Jill Moore[4], Abhijith Asok[5], Sushant Kumar[6,7], Elizabeth E. Blue[8,9], Steven Buyske[10], Nancy Cox[11], Adam Felsenfeld[12], Mark Gerstein[13,14], Eimear Kenny[15,16,17], Bingshan Li[18], Tara Matise[19], Anthony Philippakis[20], Heidi L. Rehm[21,22], Heidi J. Sofia[12], Grace Snyder[12], NHGRI Genome Sequencing Program Variant Functional Annotation Working Group, Zhiping Weng[4], Benjamin Neale[21,23], Shamil R. Sunyaev[21,24] and Xihong Lin[1,21,25,*]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA, [2]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN, USA, [3]Department of Biostatistics, University of Washington, Seattle, WA 98195, USA, [4]Program in Bioinformatics and Integrative Biology, University of Massachusetts Chan Medical School, Worcester, MA, USA, [5]Microsoft Inc. Redmond, WA, USA, [6]Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada, [7]Princess Margaret Cancer Centre, Toronto, ON, Canada, [8]Division of Medical Genetics, University of Washington, Seattle, WA, USA., [9]Brotman Baty Institute for Precision Medicine, Seattle, WA, USA., [10]Department of Statistics, Rutgers, The State University of New Jersey, Piscataway, NJ, USA, [11]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA, [12]National Human Genome Research Institute, Bethesda, DC, USA, [13]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA, [14]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA, [15]Department of Genetics and Genomic Science, Icahn School of Medicine at Mount Sinai, New York, NY, USA, [16]Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA, [17]Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA, [18]Department of Molecular Physiology and Biophysics, Vanderbilt University Medical Center, Nashville, TN, USA, [19]Department of Genetics, Rutgers, The State University of New Jersey, Piscataway, NJ, USA, [20]Data Science Platform, Broad Institute of Harvard and MIT, Cambridge, MA, USA, [21]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA, [22]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA, [23]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA, [24]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA and [25]Department of Statistics, Harvard University, Cambridge, MA, USA

Received August 15, 2022; Revised September 25, 2022; Editorial Decision October 08, 2022; Accepted October 14, 2022

ABSTRACT

Large biobank-scale whole genome sequencing (WGS) studies are rapidly identifying a multitude of coding and non-coding variants. They provide an unprecedented resource for illuminating the genetic basis of human diseases. Variant functional annotations play a critical role in WGS analysis, result interpretation, and prioritization of disease- or trait-associated causal variants. Existing functional annotation databases have limited scope to perform online queries and functionally annotate the genotype data of large biobank-scale WGS studies. We develop the Functional Annotation of Variants Online Resources (FAVOR) to meet these pressing needs. FAVOR provides a comprehensive multi-faceted variant functional annotation online portal that summarizes and visualizes findings of all possible nine billion single nucleotide variants (SNVs) across the genome. It allows for rapid variant-, gene- and region-level queries of variant functional annotations.

[*]To whom correspondence should be addressed. Tel: +1 617 432 2914; Fax: +1 617 432 5619; Email: hzhou@hsph.harvard.edu
Correspondence may also be addressed to Xihong Lin. Email: xlin@hsph.harvard.edu

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

FAVOR integrates variant functional information from multiple sources to describe the functional characteristics of variants and facilitates prioritizing plausible causal variants influencing human phenotypes. Furthermore, we provide a scalable annotation tool, FAVORannotator, to functionally annotate large-scale WGS studies and efficiently store the genotype and their variant functional annotation data in a single file using the annotated Genomic Data Structure (aGDS) format, making downstream analysis more convenient. FAVOR and FAVORannotator are available at *https:// favor.genohub.org*.

INTRODUCTION

A rapidly increasing number of large biobank-scale Whole Genome/Exome Sequencing (WGS/WES) studies are being conducted. They provide rich opportunities for understanding the genetic bases of complex human diseases and traits. Examples of large WGS/WES studies include the Trans-Omics Precision Medicine Program (TOPMed) of the National Heart, Lung and Blood Institute (NHLBI) (1), the Genome Sequencing Program (GSP) of the National Human Genome Research Institute (NHGRI), UK biobank (2) and All of Us (3). These large WGS/WES studies have identified hundreds of millions of coding and non-coding genetic variants across the human genome from hundreds of thousands of individuals and provided opportunities to evaluate their associations to diseases and traits.

Variant functional annotation provides functional information from many different sources to elucidate the multi-faceted functions of genetic variants. It empowers a wide range of analyses of array-based Genome-Wide Association Studies (GWAS) and large-scale WGS/WES studies (4–14). A variety of functional annotations have been developed to measure multiple aspects of biological functionality of variants, including protein function (15–17), conservation (18,19), epigenetics (20,21), spatial genomics (22,23), network biology (24), mappability (25), local nucleotide diversity (26), gene location and sequence (27) and integrative composite annotations (5,14,28–30). These annotations have successfully prioritized plausible causal variants of underlying GWAS signals to facilitate studying their functional impact in experimental studies following GWAS findings (10). They have also been used for identifying causal variants in fine-mapping studies (4,8,11), estimating partitioned heritability (6), calculating polygenic risk scores (PRSs) (7), and empowering rare variant (RV) association analysis of WGS studies (9,12,13,31). For example, large-scale WGS/WES studies (1,3,32) assess the associations between complex diseases/traits and coding and non-coding rare variants across the genome. The recently developed STAAR method incorporates multi-faceted variant functional annotations to boost the power of rare variant association tests in WGS/WES studies (12,13).

There is a pressing need to develop a comprehensive whole genome variant functional annotation database and browser for online queries to facilitate analysis and interpretation of GWAS and WGS/WES studies, as well as software that functionally annotates any GWAS and WGS/WES study for downstream statistical genetic analysis. Although there are several well-established variant functional annotation databases, such as CADD (5,33), VEP (34), Annovar (35), WGSA (36), SnpEff (37), and recently developed functional databases VarSome (38) and VannoPortal (39), there are several limitations. First, these resources have limited online query capabilities, and do not provide user-friendly variant function annotation browsers that summarize and visualize multi-faceted functional annotations of a single variant and/or multiple variants in a gene or a region. For example, WGSA does not provide an online browser for querying variant functional annotations. VEP provides a browser with only a few annotations. CADD allows for querying a single variant or variants in a region but displays the annotation results in a large table that is difficult to navigate. The recently developed tool VannoPortal has several attractive features, including a responsive and interactive web interface with rich functional annotations, but it currently only supports single variant query. Most of these resources do not allow for gene- and region-level variant annotations and have limited capacities in summarizing and visualizing query results.

Second, these databases miss some annotations that are useful for WGS analysis and result interpretation. For example, the commonly used databases, e.g. CADD (5,33), VEP (34), Annovar (35), WGSA (36), SnpEff (37), do not provide overall and ancestry-specific allele frequencies (AFs) from large WGS studies such as gnomAD (40) and TOPMed (1), and are lack of ClinVar information (41). The recently developed databases VarSome and VannoPortal provide ClinVar information and use the older version of gnomAD (v2). These resources do not provide functional category-specific annotation Principal Components (aPCs) (12), cCREs (42), Nuclear Diversity (26) and Recombinant Rate (26), which are important for WGS analysis.

Third, there is a lack of scalable and easy-to-use tools that satisfy the need of functionally annotating large-scale WGS/WES studies. Existing functional annotation databases and tools are not scalable for functionally annotating a massive number of variants in large-scale WGS/WES studies. Moreover, few of the currently available functional annotation tools can provide organized output in a format that is both storage efficient and ready to be used in downstream statistical genetic analyses, such as fine-mapping (4,11), heritability (6), rare variant association tests (12,13). There is a pressing community need to develop a convenient and comprehensive functional annotation tool that annotates any WGS study dataset at scale and generates a functionally annotated genotype file in an organized and compressed format, that can be readily integrated into the downstream analysis.

We developed Functional Annotation of Variants Online Resources (FAVOR), a comprehensive whole genome variant annotation database and a variant browser that provides hundreds of functional annotation scores from a variety of biological functional dimensions for all possible 9 billion Single Nucleotide Variants (SNVs) and observed short insertions/deletions (indels). FAVOR provides a fast, convenient, and user-friendly web interface that features online single variant, gene- and region-level variant queries. Search results are well-organized and conveniently visualized, according to their major functional categories. FAVOR distinguishes itself from the limitations of existing tools by providing functional annotation information that can be easily viewed through multiple functional category-based blocks and tables directly on its web interface. On top of that, FAVOR automatically generates dynamic summaries of search results by identifying important functional scores of the queried variant. These FAVOR unique features grant users immediate and intuitive insight into the search results while still maintaining users' access to the comprehensive display of multi-faceted functional scores. We have provided a comparison between FAVOR with the existing annotation databases (Supplementary Table S1).

We have also developed FAVORannotator, a tool that functionally annotates the genotype data of any WGS/WES study at scale using the FAVOR database (GRCh38 build) and stores the genotype data and their aligned functional annotation data in an annotated Genomic Data Structure (aGDS) file. The proposed aGDS data format extends the Genomic Data Structure (GDS) format (43), by storing the genotype data and the corresponding functional annotation data in a single file, making downstream integrative analysis of variants with their functional annotations more efficient and convenient. The GDS format is highly storage-efficient, with a compression rate of a thousand times compared with the VCF format. FAVORannotator is scalable and computationally efficient for functionally annotating large biobank-scale WGS/WES studies, for example, it completes the functional annotation of 1 billion variants of 184 878 multi-ethnic WGS samples in 38 CPU hours and storing those data in an aGDS file of size 488 GB. FAVORannotator automatically exports annotation results into aGDS format and achieves high storage efficiency (use CCDG Freeze 2 and TOPMed Freeze 8 datasets as examples, see Supplementary Table S2).

FAVOR DATABASE

The FAVOR relational functional annotation database provides comprehensive multi-faceted variant functional annotations of all possible 9 billion SNVs in the whole genome by integrating data from multiple different sources, including CADD v1.5 (5,33), GENCODE v31 (44), Annovar (35), WGSA (36), ClinVar (41), ENCODE (42), SnpEff (37), 1000 Genome (45), TOPMed Bravo Freeze 8 (1), gnomAD v3 (40) and other individual studies (25,28,46–49). The preprocessing stage assigns the functional annotation values to each variant by using the variant as the primary key of the relational database.

The FAVOR database is built using the PostgreSQL relational DBMS (Database Management System) for storing and retrieving variant annotation data and using a multi-table design that supports efficient integration of different types of scores. Specifically, it stores 160 functional annotation values for all possible 8,892,915,237 SNVs, and 79,997,898 observed indels in 20 TB of space. These functional annotations are organized into 12 major types, including Variant Category, Allele Frequencies (AFs), ClinVar, Integrative Scores, Protein Functions, Conservation, Epigenetics, Chromatin States, Local Nucleotide Diversity, Mutation Density, Mappability and Proximity (Supplementary Table S3). The FAVOR database can be downloaded from the FAVOR website.

FAVOR ONLINE PORTAL

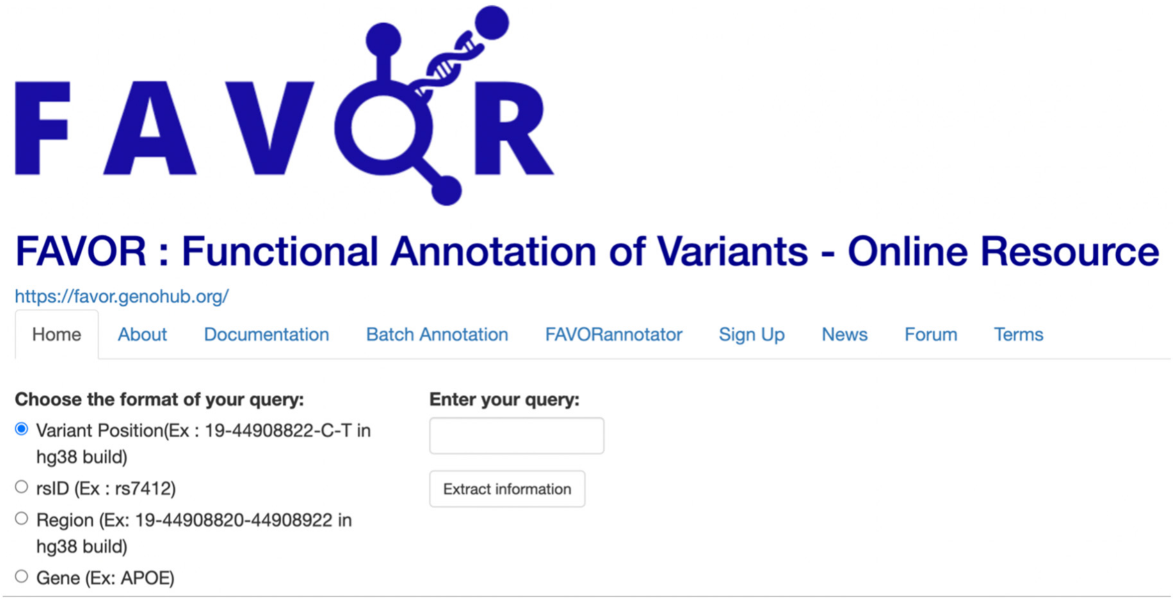
The online FAVOR portal facilitates fast and convenient online functional annotation query using an R shiny app (Figure 1). It allows users to search for a single variant (either in position format or rsID), multiple variants in a gene or genomic region (either in position format or gene name), or batches of tens of thousands of variants. The variant functional annotation results are displayed in tabular overviews in a summary tab (Figure 2), a full tables tab (Figure 3), and visualized using histograms (Figure 4).

The FAVOR web interface is exceptionally nimble. Single Variant Search (both variant position and rsID) renders results on the webpage immediately, while Gene-based and Region-based Variant Search takes just a few seconds to display results, and Batch annotation directly generates the annotation results for up to 10 000 variants allowing for a range of input file formats. This fast response speed is the product of its backend database indices and table design. The indices employ a diverse set of data structures, each tailored toward specific functionalities. The table design relies upon an original primary key (a combined string that consists of variant chromosome position and reference and alternative allele, e.g. 19-44908822-C-T) that efficiently relates the tables with regard to both computation and storage. This implementation enables the fast query of 160 annotations for all 9 billion SNVs at the variant, gene and region levels. Single Variant Search organizes functional annotation results in blocks defined by annotation types (Figure 3 and Supplementary Table S3), and Gene-based and Region-based Variant Search results display in large tables (Supplementary Table S4), all the query results display on web interface can be downloaded from the "Download query results" button at the bottom.

Compared with the other existing variant functional annotation online portals, FAVOR provides more comprehensive query options (Supplementary Table S5) including Single Variant, Gene-based and Region-based Searches, and Batch annotation. For the variant-level query, FAVOR has a similar query speed compared to CADD and is much faster than the other functional annotation online portals. FAVOR provides gene/region-level variant functional annotations, which are lacking in other portals. FAVOR is a little slower in batch annotation, as it provides much more functional annotations compared to the other portals that allow for batch annotation (Supplementary Table S5).

Single Variant Search

For Single Variant Search, users can input a variant position (in hg38 build) or an rsID. The retrieved functional annotation results are displayed in three tabs: Summary, Full Table, Figures. The Summary Tab gives an overview of the biological functionality of a variant by providing the filtered annotations that flag the variant as plausibly functional, for example, Polyphen scores equal to 1 (probably_damaging), SIFT score equals to 0 (deleterious), or ClinVar Significance is Pathogenic, and the integrative scores greater than 10 on



FAVOR : Functional Annotation of Variants - Online Resource

https://favor.genohub.org/

Choose the format of your query:

- ● Variant Position(Ex : 19-44908822-C-T in hg38 build)
- ○ rsID (Ex : rs7412)
- ○ Region (Ex: 19-44908820-44908922 in hg38 build)
- ○ Gene (Ex: APOE)

Enter your query:

Figure 1. FAVOR web interface. This online portal provides a convenient web interface allowing for variant-, gene-, and region-level annotation queries. The home page displays the supported query methods, and examples of the expected input.

the PHRED scale (in the top 10% of the genome). By selecting and presenting the most informative functional annotation of a queried variant in the summary tab avoids overwhelming users with a large amount of information.

The Full Tables tab displays all functional annotation scores—organized into 17 blocks of annotation groups (Figure 4). These blocks are Basic, ClinVar, Variant Category, Overall Allele Frequencies (AFs), Ancestry-Specific AF, Gender AF, Integrative Score, Protein Function, Conservation, Epigenetics, Transcription Factors, Chromatin States, Local Nucleotide Diversity, Mutation Density, Mapability and Proximity Table.

Different groups of functional annotation depict the variants from multiple functional perspectives. For example, ClinVar reports the relationships between genetic variants and phenotypes (41). FAVOR provides critical information from ClinVar, including Clinical Significance, Disease Name, Review Status, Disease Database ID, and Gene Reported related to the variants. Variant Category annotations provide the consequences of the genetic variants in the context of gene, categorical regulatory information, and the relative location of the variant with the closest gene (Supplementary Table S3).

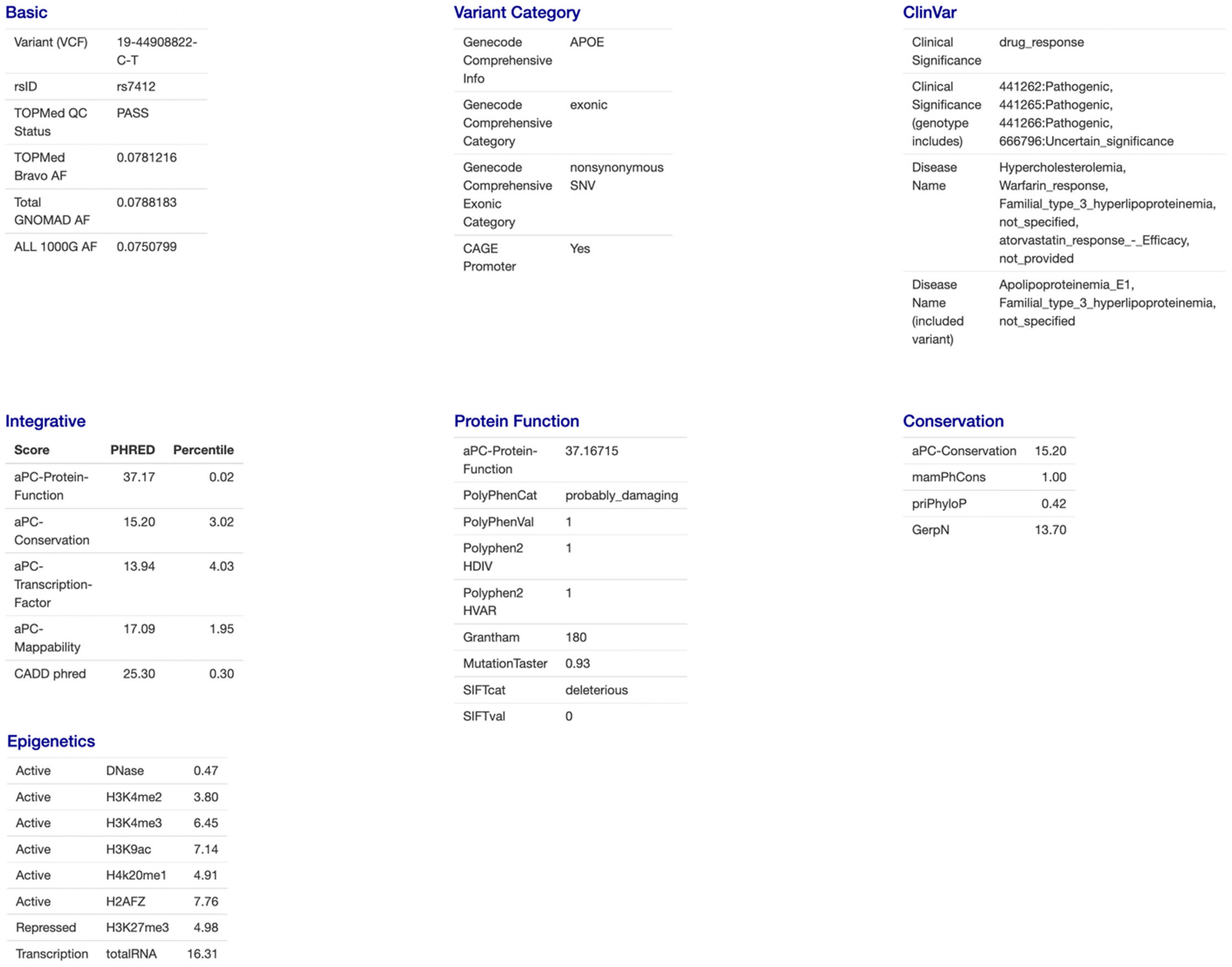
FAVOR integrates multiple AFs of observed variants from multiple variant databases, including the overall AFs, from 1000 Genome (45), TOPMed Bravo Freeze 8 (1) and gnomAD (40), and ancestry-specific AFs and gender-specific AFs from 1000 Genome (45) and gnomAD (40). FAVOR provides multiple integrative scores for both coding and non-coding variants, including CADD v1.5 (5,33), LINSIGHT (50), FATHMM-XF (29), FunSeq (47), Aloft (28) and annotation Principal Components (aPCs) (12). The aPCs summarize multiple aspects of variant function by calculating the first variant-specific PC from the individual functional annotation scores in a functional category (12). For example, aPC-conservation is the first PC of the eight individual standardized conservation scores.

Furthermore, FAVOR displays category-specific individual functional annotations that represent multiple biological functionalities of each variant in a given functional category (Supplementary Table S3). For example, protein function scores describe various impact scores of the variant's damages to protein function. Conservation scores summarize the conservation functional annotation of the variants (both within and between species). Epigenetics scores summarize the signals of the open chromatin markers, close chromatin markers, and transcription markers. FAVOR also provides individual annotation scores of local nucleotide diversity, mutation density and mappability (e.g. using the unconverted genome Umap and the bisulfite-converted genome Bismap) (Supplementary Table S3). Results can be visualized using histograms in the Figures tab (Figure. 4).

Region/Gene-based Search

For Region/Gene-based Search, users input either a gene name (official symbol), or region (starting and ending positions using the hg38 build). FAVOR will instantaneously output the functional annotation summary results of the variants in the gene or the region, as well as variant-specific annotations in a range of annotation categories. The fast display of the retrieved results of the Region/Gene-based Search is enabled through indexing and efficient multi-table database management.

The Region/Gene-based Search summary tab provides the summary statistics of the variants in a region or a gene using several key summary tables and histograms, including Allele Frequency Distribution, GENCODE Category, ClinVar Clinical Significance, Functional Consequences and High Integrative Functional Scores (Figure 5).



Basic

Variant (VCF)	19-44908822-C-T
rsID	rs7412
TOPMed QC Status	PASS
TOPMed Bravo AF	0.0781216
Total GNOMAD AF	0.0788183
ALL 1000G AF	0.0750799

Variant Category

Genecode Comprehensive Info	APOE
Genecode Comprehensive Category	exonic
Genecode Comprehensive Exonic Category	nonsynonymous SNV
CAGE Promoter	Yes

ClinVar

Clinical Significance	drug_response
Clinical Significance (genotype includes)	441262:Pathogenic, 441265:Pathogenic, 441266:Pathogenic, 666796:Uncertain_significance
Disease Name	Hypercholesterolemia, Warfarin_response, Familial_type_3_hyperlipoproteinemia, not_specified, atorvastatin_response_-_Efficacy, not_provided
Disease Name (included variant)	Apolipoproteinemia_E1, Familial_type_3_hyperlipoproteinemia, not_specified

Integrative

Score	PHRED	Percentile
aPC-Protein-Function	37.17	0.02
aPC-Conservation	15.20	3.02
aPC-Transcription-Factor	13.94	4.03
aPC-Mappability	17.09	1.95
CADD phred	25.30	0.30

Protein Function

aPC-Protein-Function	37.16715
PolyPhenCat	probably_damaging
PolyPhenVal	1
Polyphen2 HDIV	1
Polyphen2 HVAR	1
Grantham	180
MutationTaster	0.93
SIFTcat	deleterious
SIFTval	0

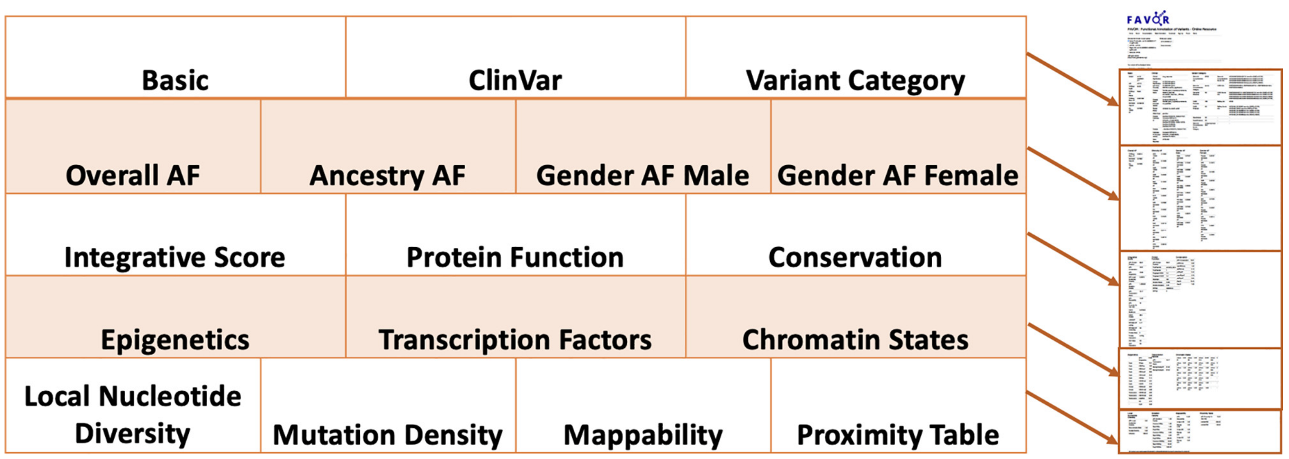
Conservation

aPC-Conservation	15.20
mamPhCons	1.00
priPhyloP	0.42
GerpN	13.70

Epigenetics

Active	DNase	0.47
Active	H3K4me2	3.80
Active	H3K4me3	6.45
Active	H3K9ac	7.14
Active	H4k20me1	4.91
Active	H2AFZ	7.76
Repressed	H3K27me3	4.98
Transcription	totalRNA	16.31

Figure 2. Single Variant Query Summary. The Single Variant Query Summary tab shows a dynamic overview of the filtered annotations with evidence for plausible functional consequences. For example, the annotations are displayed if Polyphen scores equal to 1 (probably_damaging), SIFT score equals to 0 (deleterious), ClinVar Significance is Pathogenic, and the integrative scores that are greater than 10 on the PHRED scale.



Basic	ClinVar	Variant Category	
Overall AF	Ancestry AF	Gender AF Male	Gender AF Female
Integrative Score	Protein Function	Conservation	
Epigenetics	Transcription Factors	Chromatin States	
Local Nucleotide Diversity	Mutation Density	Mappability	Proximity Table

Figure 3. Single Variant Query Functional Annotation Tabulation. The Full Tables tab in the FAVOR Single Variant Query organizes functional annotation results in blocks defined by annotation types.

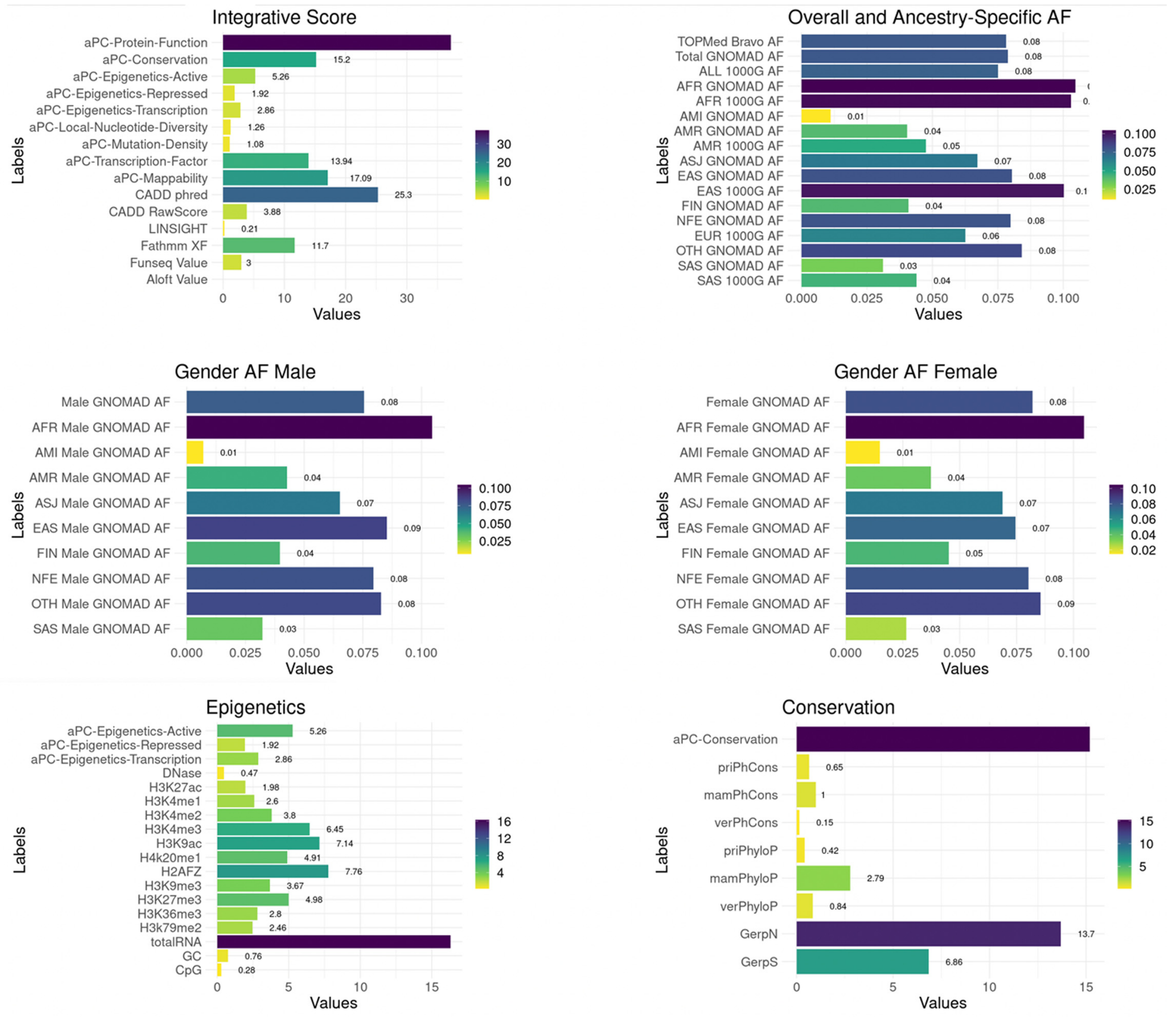


Figure 4. Single Variant Query Functional Annotation Visualization. The Figures tab in the FAVOR Single Variant Online Query displays a visualization of the functional annotation results of a queried variant in the histogram.

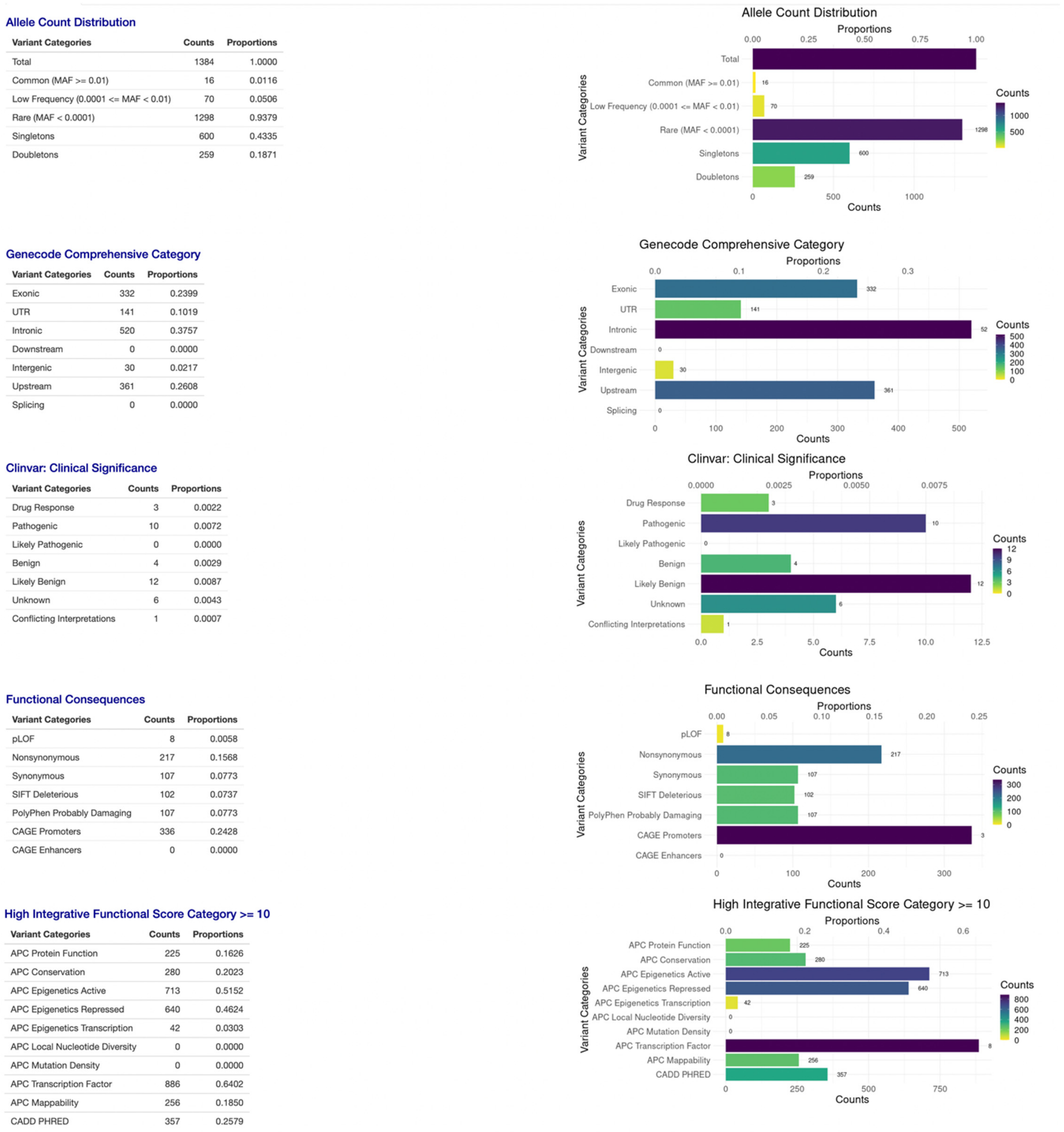
In the Region/Gene-based individual variant annotation table, 32 commonly used annotations (Supplementary Table S4) are displayed for each variant in the gene/region. The variants can be sorted by their values in any column. It also has a convenient search feature that allows users to filter the variants in the region/gene based on specified features and keywords. For example, typing 'pathogenic' in the search box above the displaying table provides only the pathogenic variants of the region/gene.

Batch annotation

Batch annotation provides functional annotations of a list of variants submitted by users in a file. It supports multiple file formats as input, including CSV, TSV, VCF, XLS and RDS. Multiple formats and IDs of variants are also supported. For example, each row of a text file can specify a variant's chromosome, position, reference, and alternative allele value (e.g. 1-10253-CTA-C), or a variant's chromosome and position values (e.g. 1-10253), or rsIDs (e.g. rs868413313). Users can upload the variants list using the above file formats on the FAVOR batch annotation page. Batch annotation files are currently limited to 10,000 variants in the interest of online wait time. It takes a few minutes to annotate 1000 variants. The annotation results containing 160 annotations of the variants in the submitted variant list are available for download. FAVORannotator, discussed below, can be used to handle functional annotations of a larger number of variants, e.g. hundreds of millions of variants in a WGS/WES study.

ANNOTATED GENOMIC DATA STRUCTURE (AGDS)

Variant Call Format (VCF) (51) has been frequently used for storing variant call data of sequencing studies. However, VCF is text-based and thus inefficient with regard to storage, particularly for large-scale WGS data of hundreds of thousands to millions of subjects that have hundreds



Allele Count Distribution

Variant Categories	Counts	Proportions
Total	1384	1.0000
Common (MAF >= 0.01)	16	0.0116
Low Frequency (0.0001 <= MAF < 0.01)	70	0.0506
Rare (MAF < 0.0001)	1298	0.9379
Singletons	600	0.4335
Doubletons	259	0.1871

Genecode Comprehensive Category

Variant Categories	Counts	Proportions
Exonic	332	0.2399
UTR	141	0.1019
Intronic	520	0.3757
Downstream	0	0.0000
Intergenic	30	0.0217
Upstream	361	0.2608
Splicing	0	0.0000

Clinvar: Clinical Significance

Variant Categories	Counts	Proportions
Drug Response	3	0.0022
Pathogenic	10	0.0072
Likely Pathogenic	0	0.0000
Benign	4	0.0029
Likely Benign	12	0.0087
Unknown	6	0.0043
Conflicting Interpretations	1	0.0007

Functional Consequences

Variant Categories	Counts	Proportions
pLOF	8	0.0058
Nonsynonymous	217	0.1568
Synonymous	107	0.0773
SIFT Deleterious	102	0.0737
PolyPhen Probably Damaging	107	0.0773
CAGE Promoters	336	0.2428
CAGE Enhancers	0	0.0000

High Integrative Functional Score Category >= 10

Variant Categories	Counts	Proportions
APC Protein Function	225	0.1626
APC Conservation	280	0.2023
APC Epigenetics Active	713	0.5152
APC Epigenetics Repressed	640	0.4624
APC Epigenetics Transcription	42	0.0303
APC Local Nucleotide Diversity	0	0.0000
APC Mutation Density	0	0.0000
APC Transcription Factor	886	0.6402
APC Mappability	256	0.1850
CADD PHRED	357	0.2579

Figure 5. Region/Gene-based Query Summary tab. The Summary tab of the Region/Gene-based Query shows the multi-faceted functional annotation summary statistics of the variants in a gene or a region.

of millions to billions of variants. The recently developed Genomic Data Structure (GDS) format (43) provides a storage-efficient format to store WGS data. The storage efficiency is even more prominent when the sample size is large, for example, the annotated TOPMed Freeze 8 ($n = 140\ 306$ samples) genotype data shows a better file compression rate compared with the annotated GSP CCDG Freeze 2 ($n = 60\ 545$ samples) genotype data. The GDS format has a compression rate of 1000 times compared to the VCF format in large WGS studies. However, it does not incorporate variant functional annotations.

We developed the annotated Genomic Data Structure (aGDS) format (Figure 6), that extends the GDS format by integrating both genotypes in a WGS study and variant functional annotations in a single file. There are three main advantages of the aGDS format. First, it provides fast query

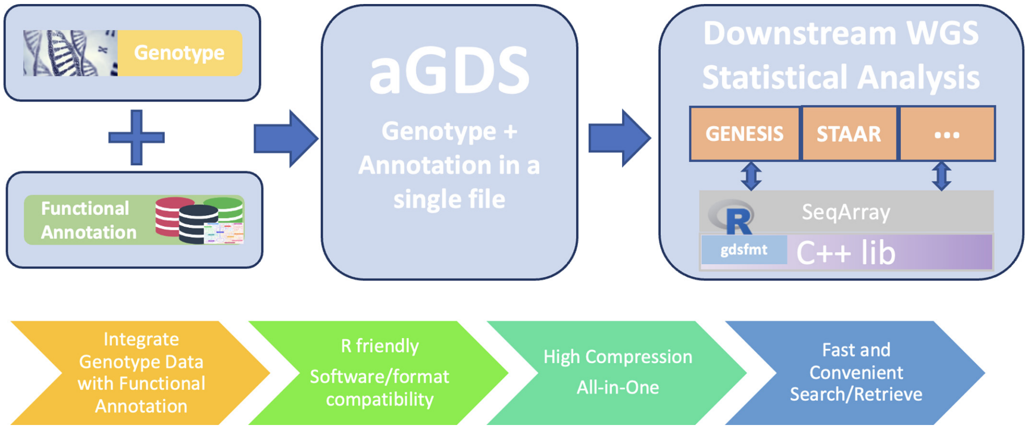

Genotype
+
Functional Annotation
aGDS
Genotype + Annotation in a single file
Downstream WGS Statistical Analysis
GENESIS
STAAR
...
R
SeqArray
gdsfmt
C++ lib
Integrate Genotype Data with Functional Annotation
R friendly Software/format compatibility
High Compression All-in-One
Fast and Convenient Search/Retrieve

Figure 6. Features of the annotated Genomic Data Structure (aGDS) format. This figure shows the features of the aGDS format and the process of creating aGDS files by combining functional annotations with genotype data.

and simultaneous retrieval of genotype and matched functional annotation data defined by flexible filtering criteria. Second, it is convenient to integrate an aGDS file into functionally informed downstream analysis pipelines, such as STAARpipeline for rare variant association analysis. Third, it is also highly storage-efficient for genotype and their functional annotation data. An aGDS file containing TOPMed Freeze 8 WGS data, including both genotype and their functional annotations of 140,306 samples, only takes 478 GB, that is three orders of magnitude smaller compared to VCF files (Supplementary Table S2).

The GDS format is designed to host large genotype data and can achieve extremely highly efficient random access of compressed data through independently compressed data blocks. It stores genotypes in a 2-bit array with ploidy, sample, and variant dimensions. An index vector associated with genotypes is used to indicate the number of bits (43). An aGDS file uses SeqArray to build functional annotation data in an GDS file. Variable-length annotation vectors are organized in an array. Functional annotation build-in and retrieval are available for efficient random access (43). Lempel-Ziv Markov chain (LZMA) or zlib are the lossless compression algorithms supported by aGDS. LZMA offers a higher compression ratio, but requires more memory allocation and run time (43). Functional annotation data are recorded alongside genotype data in a highly compressed format that significantly reduces storage consumption. Fast random access of the compressed functional annotation of selected variant sets can be efficiently performed, making aGDS attractive to host functionally annotated large-scale WGS/WES data for convenient downstream analysis.

Several existing WGS association analysis tools support the aGDS format, e.g. STAAR (12) and STAARpipeline. Several other tools support the GDS format, e.g. GENESIS (56), SeqArray (43), SeqVarTools and SNPRelate (57). As aGDS files are fully compatible with the tools supporting GDS files, the analytic tools that support the GDS format can be extended to support the aGDS format.

FAVORANNOTATOR

FAVORannotator is an open-source tool that uses the FAVOR database to functionally annotate and efficiently store genotype and variant functional annotation data of a WGS/WES study in an aGDS file, making downstream association analysis convenient (Figure 7). FAVORannotator only requires genotype data or a variant list as input and automatically annotates the genotype data or the variant list, generating an aGDS file as an output. An aGDS file with both genotypes and their functional annotations facilitates rare variant association analysis using individual-level data, e.g. using STAAR (12), while an aGDS file with only a variant list and their functional annotations facilitates rare variant meta-analysis using WGS summary statistics.

Time and memory resources for annotating a large number of variants using FAVORannotator are very attractive, especially for large-scale WGS/WES datasets, such as TOPMed, GSP and UK Biobank. For example, FAVORannotator produces an annotated genotype file in the aGDS format for n = 184 878 whole genome samples with 1 billion variants of the TOPMed Freeze 10a WGS data in 38 hours, and for n = 60 545 whole genome samples of 450 million variants of the GSP-CCDG Freeze 2 WGS data within 30 CPU hours. FAVORannotator has also been implemented as a workflow in the cloud-based platforms, including DNAnexus (UK Biobank), AnVIL (NHGRI) and BioData Catalyst (NHLBI) (Figure 8) (52). FAVORannotator's efficiency keeps cloud computing costs low. For example, it costs ∼$25 to annotate the TOPMed Freeze 10a WGS data by chromosome in parallel, e.g. in 3 CPU hours for chromosome 1.

Users can add customized functional annotations to an aGDS file by adding new columns to the FAVOR database using either the CSV or SQL format and then running FAVORannotator.

Both speed and storage efficiency of annotation results are crucial for downstream analysis. As existing functional annotation databases and tools, such as Annovar (35) and VEP (34), store variant annotation results in text tables (TSV, CSV), they are much less efficient in query speed and storage than FAVORannotator which uses the aGDS format (Supplementary Table S6). Several variant functional annotation tools, such as SnpEff (37), Vcfanno (55) and VarNote (54), use the VCF format. As VCF stores the same annotation variable names repeatedly for a large number of times in the INFO column, it is much less storage-efficient compared with aGDS (Supplementary Table S6). FAVORannotator, SnpEff (37), Vcfanno (55) and VarNote (54)

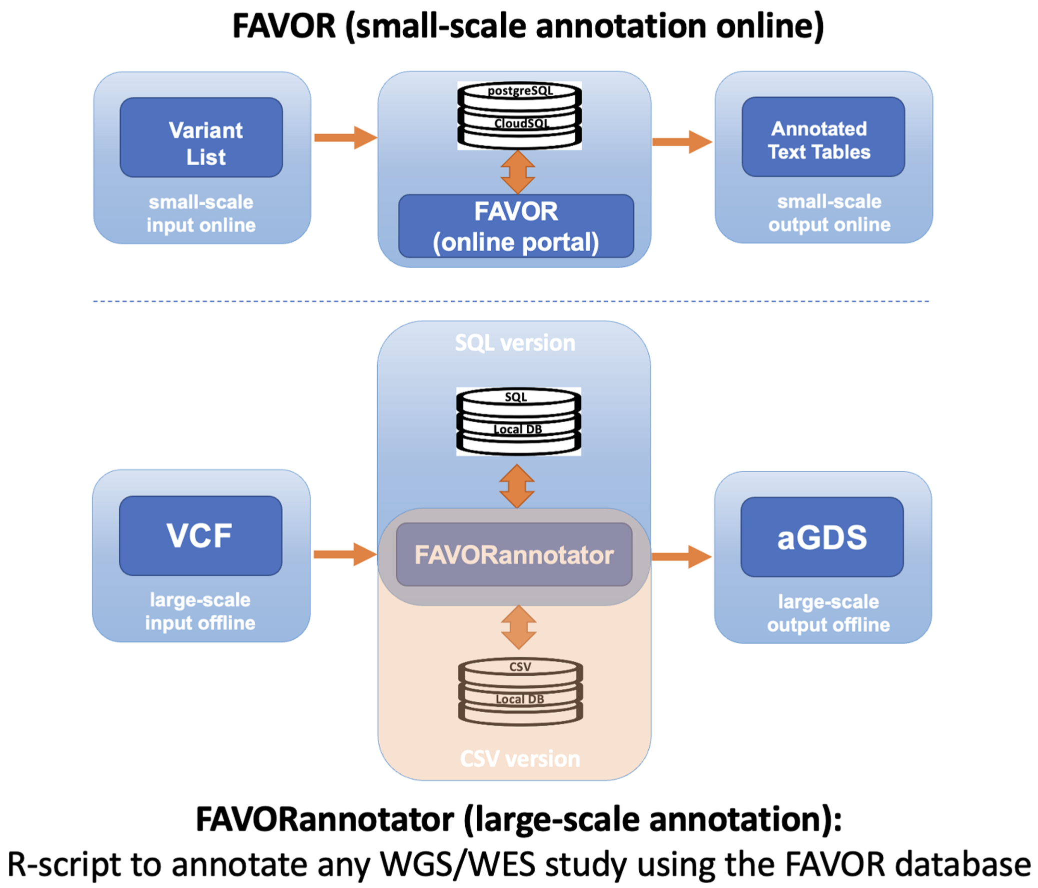
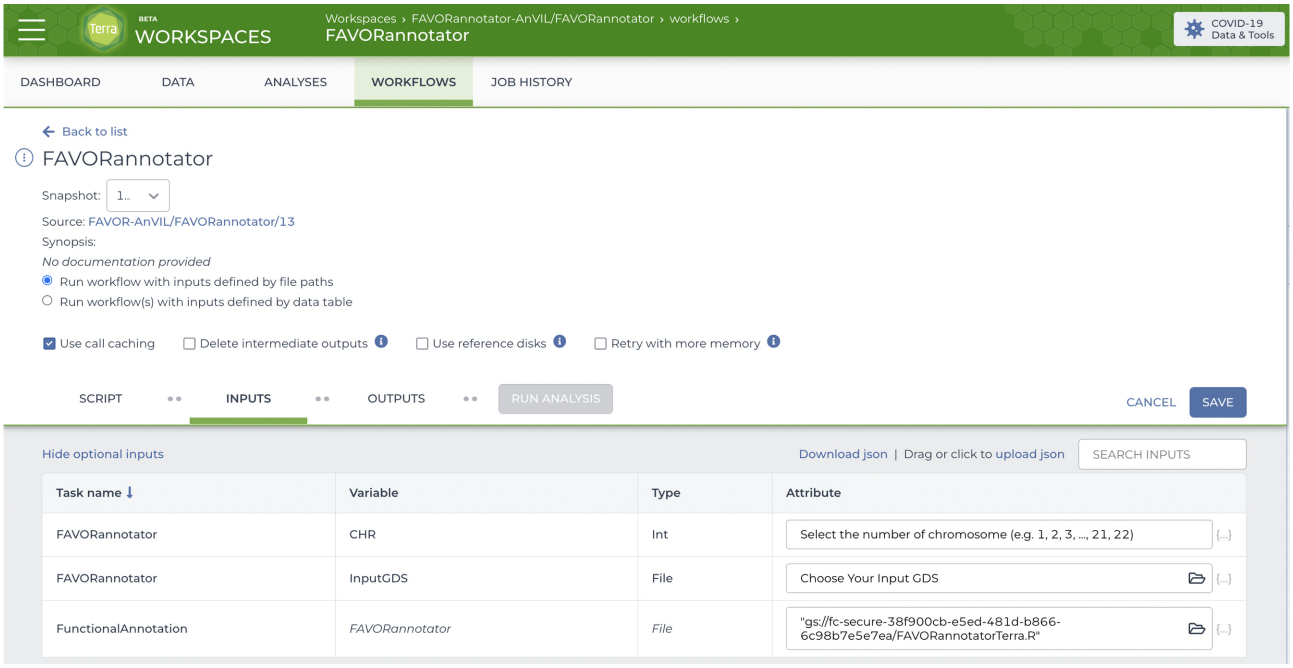

FAVOR (small-scale annotation online)
Variant List
small-scale input online
postgreSQL
CloudSQL
FAVOR (online portal)
Annotated Text Tables
small-scale output online
SQL version
SQL
Local DB
VCF
large-scale input offline
FAVORannotator
aGDS
large-scale output offline
CSV
Local DB
CSV version
FAVORannotator (large-scale annotation):
R-script to annotate any WGS/WES study using the FAVOR database

Figure 7. Graphical representation of the features of FAVOR batch annotation and FAVORannotator. For small-scale annotation (up to 10 000 variants), batch annotation can be used for online annotation at the FAVOR website. For large-scale annotation, e.g. hundreds of millions of variants in a Whole Genome/Exome Sequencing (WGS/WES) study, FAVORannotator can be used for annotation in a local cluster or a cloud platform, e.g. Amazon Web Services (AWS) and Google Cloud Platform (GCP). FAVORannotator uses the FAVOR backend database, which is available in the SQL or CSV formats, and outputs an aGDS file that integrates genotype and annotation data in a single file.



Workspaces › FAVORannotator-AnVIL/FAVORannotator › workflows ›
FAVORannotator

FAVORannotator

Snapshot: 1...

Source: FAVOR-AnVIL/FAVORannotator/13

Synopsis:

No documentation provided

● Run workflow with inputs defined by file paths
○ Run workflow(s) with inputs defined by data table

☑ Use call caching ☐ Delete intermediate outputs ☐ Use reference disks ☐ Retry with more memory

Task name ↓	Variable	Type	Attribute
FAVORannotator	CHR	Int	Select the number of chromosome (e.g. 1, 2, 3, ..., 21, 22)
FAVORannotator	InputGDS	File	Choose Your Input GDS
FunctionalAnnotation	*FAVORannotator*	*File*	"gs://fc-secure-38f900cb-e5ed-481d-b866-6c98b7e5e7ea/FAVORannotatorTerra.R"

Figure 8. Cloud-Native FAVORannotator Workflow. The interface of the FAVORannotator Workflow on Terra.bio.

store annotations alongside genotype data, and are convenient for downstream analysis. Supplementary Table S6 shows the aGDS format based FAVORannotator is much more storage-efficient than the existing tools export annotation results in text table or annotated VCF, such as Annovar, CADD, VarNote and Vcfanno, and is hence more efficient for downstream analysis.

DISCUSSION

FAVOR offers a comprehensive solution for the application of whole genome variant functional annotations, including open access and downloadable database, a user-friendly browser, and a tool FAVORannotator, to annotate large-scale WGS/WES data. The FAVOR database is a large relational data structure of multi-faceted functional annotations of all possible 9 billion SNVs and 80 million observed indels in the human genome. It is built using a storage-efficient postgreSQL database with indexed and relational tables, that provide fast query speeds. The FAVOR web interface provides fast variant-, gene-, region-level online multi-faceted functional annotations, as well as batch annotation. It emphasizes responsiveness while providing dynamic display and visualization features, and uses combined approaches, including visualizations, block organizations by categories, and convenient search and sorting functions, to provide a fast and convenient summary of the major functional impact of variants.

The FAVORannotator software enables researchers to use the FAVOR database to efficiently functionally annotate large WGS/WES studies at scale, and build a highly compressed and well-organized aGDS file. An aGDS file includes both genotype data and their annotations and can be easily integrated into downstream analysis pipelines. Together, FAVOR and FAVORannotator provide a valuable tool to facilitate downstream analysis and interpretation of WES/WGS studies and array based GWAS studies.

Although several compression methods are available for storing WGS data, such as gzip (vcf.gz), Bgzip or BCF (53), they are subject to two major limitations. First, they are not efficient for storing large-scale WGS data. Second, they are difficult to read while compressed. For instance, although the BCF format is more storage-efficient than the VCF format, the compression rate is 100 times. In contrast, the GDS format has a compression rate of 1000 times. Furthermore, both VCF and BCF formats do not store variant annotations efficiently nor support retrieval of annotations efficiently. The aGDS format resolves both limitations successfully. FAVORannotator is currently developed as a standalone annotation tool optimized for fast query performance using the FAVOR database. Users who would like to do functional annotation directly from commonly used public functional annotation databases can use general-purposed functional annotation tools and aligners, such as BCFTools (53), VarNote (54) and Vcfanno (55). These tools produce annotated VCF files, which are often quite large for biobank-size WGS studies. FAVORannotator can then be used to convert annotated VCF files generated by these annotations aligners to more storage-efficient aGDS files. It is of future research interest to extend FAVORannotator to be a general-purpose aligner that can perform efficient functional annotation directly using public functional annotation databases. It is also of future interest to port the FAVOR database to be used by general-purpose functional annotation tools, such as BCFTools, VarNote and Vcfanno.

In summary, FAVOR and FAVORannotator provide an intuitive and indispensable infrastructure for facilitating downstream analysis and result interpretation of large-scale WES/WGS studies. FAVOR currently provides non-tissue specific epigenetic functional annotations for non-coding variants. It is of future interest to integrate tissue and cell-type specific epigenetic functional annotations in FAVOR. As functional annotations continue to grow in depth and breadth, we will continue to improve and expand FAVOR by integrating more and state-of-art annotations and supporting more analytical scenarios.

DATA AVAILABILITY

FAVORannotator is an open-source annotation tool available in the GitHub repository (https://github.com/zhouhufeng/FAVORannotator).

The FAVOR essential database (containing 20 essential functional annotation scores) for all possible SNVs (8 812 917 339) and observed Indels (79 997 898) in Build GRCh38/hg38 is hosted on Harvard Dataverse (https://doi.org/10.7910/DVN/1VGTJI).

The FAVOR full database (containing 160 essential functional annotation scores) for all possible SNVs (8 812 917 339) and observed Indels (79 997 898) in Build GRCh38/hg38 is hosted on Harvard Dataverse (https://doi.org/10.7910/DVN/KFUBKG).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Human Genome Research Institute [U01-HG009088, U01-HG012064]; National Cancer Institute [R35-CA197449, U19-CA203654]; National Heart, Lung, and Blood Institute [R01-HL163560]. Funding for open access charge: NHGRI [U01-HG009088] and NHLBI [R01-HL163560].

Conflict of interest statement. A.P. is a Venture Partner at GV, a subsidiary of Alphabet corporation. He has received funding from Verily, MSFT, Intel, IBM, Bayer, Novartis, Pfizer, Biogen, Abbvie. B.M.N. is on the Scientific Advisory Board of Deep Genomics, a consultant for Camp4 Therapeutics, Takeda Pharmaceutical and Biogen. S.R.S. is consultant to NGM Biopharmaceuticals and Inari agriculture. He is also on Scientific Advisory Board of Veritas Genetics. X. Lin. is a consultant of AbbVie Pharmaceuticals and Verily Life Sciences. X.Z. is currently an employee of AbbVie Pharmaceuticals. Z.W. co-founded and serves as a scientific advisor for Rgenta Inc.

REFERENCES

1. Taliun,D., Harris,D.N., Kessler,M.D., Carlson,J., Szpiech,Z.A., Torres,R., Taliun,S.A.G., Corvelo,A., Gogarten,S.M., Kang,H.M. *et al.* (2021) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature*, **590**, 290–299.

2. Halldorsson,B.V., Eggertsson,H.P., Moore,K.H.S., Hauswedell,H., Eiriksson,O., Ulfarsson,M.O., Palsson,G., Hardarson,M.T., Oddsson,A., Jensson,B.O. *et al.* (2022) The sequences of 150,119 genomes in the UK biobank. *Nature*, **607**, 732–740.
3. All of Us Research Program Investigators, Denny,J.C., Rutter,J.L., Goldstein,D.B., Philippakis,A., Smoller,J.W., Jenkins,G. and Dishman,E. (2019) The “All of us” research program. *N. Engl. J. Med.*, **381**, 668–676.
4. Kichaev,G., Yang,W.Y., Lindstrom,S., Hormozdiari,F., Eskin,E., Price,A.L., Kraft,P. and Pasaniuc,B. (2014) Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLos Genet.*, **10**, e1004722.
5. Kircher,M., Witten,D.M., Jain,P., O’Roak,B.J., Cooper,G.M. and Shendure,J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
6. Finucane,H.K., Bulik-Sullivan,B., Gusev,A., Trynka,G., Reshef,Y., Loh,P.R., Anttila,V., Xu,H., Zang,C., Farh,K. *et al.* (2015) Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.*, **47**, 1228–1235.
7. Hu,Y., Lu,Q., Powles,R., Yao,X., Yang,C., Fang,F., Xu,X. and Zhao,H. (2017) Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLoS Comput. Biol.*, **13**, e1005589.
8. Kichaev,G., Roytman,M., Johnson,R., Eskin,E., Lindstrom,S., Kraft,P. and Pasaniuc,B. (2017) Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics*, **33**, 248–255.
9. Morrison,A.C., Huang,Z., Yu,B., Metcalf,G., Liu,X., Ballantyne,C., Coresh,J., Yu,F., Muzny,D., Feofanova,E. *et al.* (2017) Practical approaches for whole-genome sequence analysis of Heart- and Blood-Related traits. *Am. J. Hum. Genet.*, **100**, 205–215.
10. Lee,P.H., Lee,C., Li,X., Wee,B., Dwivedi,T. and Daly,M. (2018) Principles and methods of in-silico prioritization of non-coding regulatory variants. *Hum. Genet.*, **137**, 15–30.
11. Schaid,D.J., Chen,W. and Larson,N.B. (2018) From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.*, **19**, 491–504.
12. Li,X., Li,Z., Zhou,H., Gaynor,S.M., Liu,Y., Chen,H., Sun,R., Dey,R., Arnett,D.K., Aslibekyan,S. *et al.* (2020) Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat. Genet.*, **52**, 969–983.
13. Gaynor,S.M., Westerman,K.E, Zhou,H., Ackovic,L.L., Selvaraj,M.S., Li,X., Li,Z., Manning,A.K., Philippakis,A. and Lin,X. (2022) STAAR Workflow:A cloud-based workflow for scalable and reproducible rare variant analysis. *Bioinformatics*, **38**, 3116–3117.
14. Li,X., Yung,G., Zhou,H., Sun,R., Li,Z., Hou,K., Zhang,M.J., Liu,Y., Arapoglou,T., Wang,C. *et al.* (2022) A multi-dimensional integrative scoring framework for predicting functional variants in the human genome. *Am. J. Hum. Genet.*, **109**, 446–456.
15. Ng,P.C. and Henikoff,S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.
16. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
17. Adzhubei,I., Jordan,D.M. and Sunyaev,S.R. (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.*, **Chapter 7**, Unit7.20.
18. Cooper,G.M., Goode,D.L., Ng,S.B., Sidow,A., Bamshad,M.J., Shendure,J. and Nickerson,D.A. (2010) Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat. Methods*, **7**, 250–251.
19. Goode,D.L., Cooper,G.M., Schmutz,J., Dickson,M., Gonzales,E., Tsai,M., Karra,K., Davydov,E., Batzoglou,S., Myers,R.M. *et al.* (2010) Evolutionary constraint facilitates interpretation of genetic variation in resequenced human genomes. *Genome Res.*, **20**, 301–310.
20. Skipper,M., Eccleston,A., Gray,N., Heemels,T., Le Bot,N., Marte,B. and Weiss,U. (2015) Presenting the epigenome roadmap. *Nature*, **518**, 313.
21. Davis,C.A., Hitz,B.C., Sloan,C.A., Chan,E.T., Davidson,J.M., Gabdank,I., Hilton,J.A., Jain,K., Baymuradov,U.K., Narayanan,A.K. *et al.* (2018) The encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, **46**, D794–D801.
22. Rao,S.S., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D., Lander,E.S. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
23. Tang,Z., Luo,O.J., Li,X., Zheng,M., Zhu,J.J., Szalaj,P., Trzaskoma,P., Magalska,A., Wlodarczyk,J., Ruszczycki,B. *et al.* (2015) CTCF-Mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell*, **163**, 1611–1627.
24. Yu,H., Kim,P.M., Sprecher,E., Trifonov,V. and Gerstein,M. (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput. Biol.*, **3**, e59.
25. Karimzadeh,M., Ernst,C., Kundaje,A. and Hoffman,M.M. (2018) Umap and bismap: quantifying genome and methylome mappability. *Nucleic Acids Res.*, **46**, e120.
26. Gazal,S., Finucane,H.K., Furlotte,N.A., Loh,P.R., Palamara,P.F., Liu,X., Schoech,A., Bulik-Sullivan,B., Neale,B.M., Gusev,A. *et al.* (2017) Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.*, **49**, 1421–1427.
27. O’Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
28. Balasubramanian,S., Fu,Y., Pawashe,M., McGillivray,P., Jin,M., Liu,J., Karczewski,K.J., MacArthur,D.G. and Gerstein,M. (2017) Using ALoFT to determine the impact of putative loss-of-function variants in protein-coding genes. *Nat. Commun.*, **8**, 382.
29. Rogers,M.F., Shihab,H.A., Mort,M., Cooper,D.N., Gaunt,T.R. and Campbell,C. (2018) FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics*, **34**, 511–513.
30. Ionita-Laza,I., McCallum,K., Xu,B. and Buxbaum,J.D. (2016) A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.*, **48**, 214–220.
31. Quick,C., Wen,X., Abecasis,G., Boehnke,M. and Kang,H.M. (2020) Integrating comprehensive functional annotations to boost power and accuracy in gene-based association analysis. *PLoS Genet.*, **16**, e1009060.
32. Sudlow,C., Gallacher,J., Allen,N., Beral,V., Burton,P., Danesh,J., Downey,P., Elliott,P., Green,J., Landray,M. *et al.* (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.*, **12**, e1001779.
33. Rentzsch,P., Witten,D., Cooper,G.M., Shendure,J. and Kircher,M. (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.*, **47**, D886–D894.
34. McLaren,W., Gil,L., Hunt,S.E., Riat,H.S., Ritchie,G.R., Thormann,A., Flicek,P. and Cunningham,F. (2016) The ensembl variant effect predictor. *Genome Biol.*, **17**, 122.
35. Wang,K., Li,M. and Hakonarson,H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.
36. Liu,X., White,S., Peng,B., Johnson,A.D., Brody,J.A., Li,A.H., Huang,Z., Carroll,A., Wei,P., Gibbs,R. *et al.* (2016) WGSA: an annotation pipeline for human genome sequencing studies. *J. Med. Genet.*, **53**, 111–112.
37. Cingolani,P., Platts,A., Wang le,L., Coon,M., Nguyen,T., Wang,L., Land,S.J., Lu,X. and Ruden,D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: SNPs in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80–92.
38. Kopanos,C., Tsiolkas,V., Kouris,A., Chapple,C.E., Albarca Aguilera,M., Meyer,R. and Massouras,A. (2019) VarSome: the human genomic variant search engine. *Bioinformatics*, **35**, 1978–1980.
39. Huang,D., Zhou,Y., Yi,X., Fan,X., Wang,J., Yao,H., Sham,P.C., Hao,J., Chen,K. and Li,M.J. (2022) VannoPortal: multiscale functional annotation of human genetic variants for interrogating molecular mechanism of traits and diseases. *Nucleic Acids Res.*, **50**, D1408–D1416.
40. Karczewski,K.J., Francioli,L.C., Tiao,G., Cummings,B.B., Alfoldi,J., Wang,Q., Collins,R.L., Laricchia,K.M., Ganna,A., Birnbaum,D.P.

et al. (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, **581**, 434–443.

41. Landrum,M.J., Lee,J.M., Riley,G.R., Jang,W., Rubinstein,W.S., Church,D.M. and Maglott,D.R. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**, D980–D985.
42. Encode Project Consortium, Moore,J.E., Purcaro,M.J., Pratt,H.E., Epstein,C.B., Shoresh,N., Adrian,J., Kawli,T., Davis,C.A., Dobin,A. *et al.* (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, **583**, 699–710.
43. Zheng,X., Gogarten,S.M., Lawrence,M., Stilp,A., Conomos,M.P., Weir,B.S., Laurie,C. and Levine,D. (2017) SeqArray-a storage-efficient high-performance data format for WGS variant calls. *Bioinformatics*, **33**, 2251–2257.
44. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.*, **22**, 1760–1774.
45. The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
46. The FANTOM Consortium and the RIKEN PMI and CLST (DGT) (2014) A promoter-level mammalian expression atlas. *Nature*, **507**, 462–470.
47. Fu,Y., Liu,Z., Lou,S., Bedford,J., Mu,X.J., Yip,K.Y., Khurana,E. and Gerstein,M. (2014) FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.*, **15**, 480.
48. Abugessaisa,I., Noguchi,S., Hasegawa,A., Harshbarger,J., Kondo,A., Lizio,M., Severin,J., Carninci,P., Kawaji,H. and Kasukawa,T. (2017) FANTOM5 CAGE profiles of human and mouse reprocessed for GRCh38 and GRCm38 genome assemblies. *Sci Data*, **4**, 170107.
49. Fishilevich,S., Nudel,R., Rappaport,N., Hadar,R., Plaschkes,I., Iny Stein,T., Rosen,N., Kohn,A., Twik,M., Safran,M. *et al.* (2017) GeneHancer: genome-wide integration of enhancers and target genes in genecards. *Database (Oxford)*, **2017**, bax028.
50. Huang,Y.F., Gulko,B. and Siepel,A. (2017) Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.*, **49**, 618–624.
51. Danecek,P., Auton,A., Abecasis,G., Albers,C.A., Banks,E., DePristo,M.A., Handsaker,R.E., Lunter,G., Marth,G.T., Sherry,S.T. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
52. Schatz,M.C., Philippakis,A.A., Afgan,E., Banks,E., Carey,V.J., Carroll,R.J., Culotti,A., Ellrott,K., Goecks,J., Grossman,R.L. *et al.* (2022) Inverting the model of genomics data sharing with the NHGRI genomic data science analysis, visualization, and informatics Lab-space. *Cell Genom*, **2**, 100085.
53. Danecek,P., Bonfield,J.K., Liddle,J., Marshall,J., Ohan,V., Pollard,M.O., Whitwham,A., Keane,T., McCarthy,S.A., Davies,R.M. *et al.* (2021) Twelve years of SAMtools and BCFtools. *Gigascience*, **10**, giab008.
54. Huang,D., Yi,X., Zhou,Y., Yao,H., Xu,H., Wang,J., Zhang,S., Nong,W., Wang,P., Shi,L. *et al.* (2020) Ultrafast and scalable variant annotation and prioritization with big functional genomics data. *Genome Res.*, **30**, 1789–1801.
55. Pedersen,B.S., Layer,R.M. and Quinlan,A.R. (2016) Vcfanno: fast, flexible annotation of genetic variants. *Genome Biol.*, **17**, 118.
56. Gogarten,S.M., Sofer,T., Chen,H., Yu,C., Brody,J.A., Thornton,T.A., Rice,K.M. and Conomos,M.P. (2019) Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics*, **35**, 5346–5348.
57. Zheng,X., Levine,D., Shen,J., Gogarten,S.M., Laurie,C. and Weir,B.S. (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, **28**, 3326–3328.