

FAVOR: Functional Annotation of Variants Online Resource and Annotator for Variation across the Human Genome

Hufeng Zhou^{1,#,*}, Theodore Arapoglou^{1,#}, Xihao Li¹, Zilin Li¹, Xiuwen Zheng², Jill Moore³, Abhijith Asok⁴, Sushant Kumar^{5,6}, Elizabeth E. Blue^{7,8}, Steven Buyske⁹, Nancy Cox¹⁰, Adam Felsenfeld¹¹, Mark Gerstein^{12,13}, Eimear Kenny^{14,15,16}, Bingshan Li¹⁷, Tara Matisse¹⁸, Anthony Philippakis¹⁹, Heidi Rehm^{20,21}, Heidi J. Sofia¹¹, Grace Snyder¹¹, NHGRI Genome Sequencing Program Variant Functional Annotation Working Group, Zhiping Weng³, Benjamin Neale^{20,22}, Shamil R. Sunyaev^{20,23}, and Xihong Lin^{1,20,24*}

¹ Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

² Department of Biostatistics, University of Washington, Seattle, WA 98195, USA

³ Program in Bioinformatics and Integrative Biology, University of Massachusetts Chan Medical School, Worcester, MA, USA

⁴ Microsoft Inc. Redmond, WA

⁵ Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada

⁶ Princess Margaret Cancer Centre, University of Toronto, Toronto, ON, Canada

⁷ Division of Medical Genetics, University of Washington, Seattle, Washington, USA.

⁸ Brotman Baty Institute for Precision Medicine, Seattle, Washington, USA.

⁹ Department of Statistics, Rutgers, The State University of New Jersey, Piscataway, NJ, USA

¹⁰ Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

¹¹ National Human Genome Research Institute, Bethesda, DC, USA

¹² Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA

¹³ Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA

¹⁴ Department of Genetics and Genomic Science, Icahn School of Medicine at Mount Sinai, New York, NY, USA

¹⁵ Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

¹⁶ Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA

¹⁷ Department of Molecular Physiology and Biophysics, Vanderbilt University Medical Center, Nashville, TN, USA

¹⁸ Department of Genetics, Rutgers, The State University of New Jersey, Piscataway, NJ, USA

¹⁹ Data Science Platform, Broad Institute of Harvard and MIT, Cambridge, MA, USA

²⁰ Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA

²¹ Department of Medicine, Massachusetts General Hospital, Boston, MA, USA

²² Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA

²³ Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

²⁴ Department of Statistics, Harvard University, Cambridge, MA, USA

The two authors contributed equally.

* To whom correspondence should be addressed.

Tel: +1-617-432-2914; Fax: +1-617-432-5619

Email: hzhou@hsph.harvard.edu, mlin@hsph.harvard.edu

ABSTRACT

Large-scale whole genome sequencing (WGS) studies and biobanks are rapidly generating 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 or are unable to functionally annotate the genotype data of large WGS studies and biobanks for downstream analysis. We develop the Functional Annotation of Variants Online Resources (FAVOR) to meet these pressing needs. FAVOR provides a comprehensive online multi-faceted portal with summarization and visualization of all possible 9 billion single nucleotide variants (SNVs) across the genome, and allows for rapid variant-, gene-, and region-level online queries. It integrates variant functional information from multiple sources to describe the functional characteristics of variants and facilitates prioritizing plausible causal variants influencing human phenotypes. Furthermore, a scalable annotation tool, FAVORannotator, is provided for functionally annotating and efficiently storing the genotype and variant functional annotation data of a large-scale sequencing study in an annotated GDS file format to facilitate downstream analysis. FAVOR and FAVORannotator are available at <https://favor.genohub.org>.

INTRODUCTION

A rapidly increasing number of large-scale Whole Genome/Exome Sequencing (WGS/WES) studies and biobanks are being conducted. They provide rich opportunities for understanding the genetic bases of complex human diseases and traits. Examples of large WGS/WES studies and biobanks 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 (<https://www.genome.gov/Funded-Programs-Projects/NHGRI-Genome-Sequencing-Program>), UK biobank (2) and All of Us (3). These large WGS/WES studies and biobanks have sequenced hundreds of millions of coding and non-coding genetic variants across the human genome from hundreds of thousands of individuals, evaluating their relationship to diseases and traits.

Variant functional annotation represents a powerful and effective approach to leveraging functional information from many different bioinformatics sources to elucidate the multi-faceted functions of genetic variants for a wide range of analyses of WGS/WES studies and Genome-Wide Association Studies (GWAS) (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) and integrative composite annotations (4,27-29). These annotations have successfully prioritized plausible causal variants of underlying GWAS signals according to their functional impact in experimental studies following GWAS findings (5,30), localizing causal variants in fine-mapping studies (4,8), partitioning heritability in GWAS (6), predicting genetic risk (6,7,9), and improving rare variant (RV) analysis of WGS association studies (12-14,31). For example, large-scale WGS/WES studies (1,3,32,33) 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 (12-14).

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 existing variant functional annotation databases, such as CADD (5,34), VEP (35), Annovar (36), WGsA (37), they have several limitations. First, these existing variant annotation databases have narrow scope. None of the databases provides overall and ancestry-specific allele frequencies from gnomAD (38) and TOPMed (1), and ClinVar information (39). Second, these resources have limited online query capabilities, and do not provide a user-friendly variant function annotation browser that summarizes and visualizes multi-faceted functional annotations of a

single variant and multiple variants in a gene or a region. For example, WGS does not provide a browser for querying variant functional annotations online. VEP provides a browser but 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. Furthermore, most of these resources do not allow for gene- and region- level variant annotations. None of these databases provides a summary or visualization of query results. 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-variants association tests (13,14). 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 for all possible 9 billion Single Nucleotide Variants (SNVs) and observed short insertions/deletions (indels) from a variety of biological functional aspects. FAVOR provides a fast, convenient, and user-friendly web interface that features online single variant and gene-/region-level variant queries. Search results are well-organized and 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 1).

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 Genomics Data Structure (GDS) format (40), 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-scale WGS/WES studies, for example, it completes the functional annotation of 1 billion variants in 38 CPU hours and storing those data in an aGDS file of size 488 GB.

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 (5,34), GENCODE v31(41), Annovar (36), WGS (37), ClinVar (39), ENCODE (42), SnpEff (43), 1000 Genome (44), TOPMed Bravo Freeze 8 (1), gnomAD v3(38), and other individual studies (25,28,45-48). 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 from TOPMed Freeze 8 in 20 TB of space (Supplementary Table 2). 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 3). 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

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 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), Ethnicity-Specific AF, Gender AF, Integrative Score, Protein Function, Conservation, Epigenetics, Transcription Factors, Chromatin States, Local Nucleotide Diversity, Mutation Density, Mapability, 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 (39). 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 3).

FAVOR integrates multiple AFs of observed variants from multiple variant databases, including the overall AF, ethnicity-specific AF and gender-specific AF from 1000 Genome (44), TOPMed Bravo Freeze 8 (1) and gnomAD v3 (38). FAVOR provides multiple integrative scores for both coding and non-coding variants, including CADD v1.5 (5,34), LINSIGHT (49), FATHMM-XF (29), FunSeq (46), Aloft (28) and annotation Principal Components (aPCs). 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 3). 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, mappability (e.g., using the unconverted genome Umap and the bisulfite-converted genome Bimap) (Supplementary Table 3). Data visualization is available using histograms in the Figures tab for a more intuitive look into functional importance (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). In the Region/Gene-based individual variant annotation table, 32 commonly used annotations (Supplementary Table 4) 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 key words. 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 an rsID (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 less than less than 1 minute to annotate 1,000

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) (50) 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 of millions to billions of variants. The recently developed Genomic Data Structure (GDS) format (40) provides a storage-efficient format to store WGS data. The GDS format has a compression rate of 1000 times compared to the VCF format. 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 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 (13). Third, it is also highly storage-efficient for genotype and functional annotation data. An aGDS file containing TOPMed Freeze 8 WGS data, including both genotype and functional annotations of 140,306 samples, only takes 487 GB, that is three orders of magnitude smaller compared to the VCF files (Supplementary Table 2).

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, and to facilitate downstream association analysis (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. The former facilitates rare variant association analysis using individual level data, while the latter facilitates rare variant meta-analysis using summary statistics, by incorporating functional annotations, e.g., using STAAR (13).

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 $n=180,000$ whole genome samples with 900 million variants of the TOPMed Freeze 10a WGS data in 38 hours, and for $n=60545$ whole genome samples of 450 million variants of 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) (51). FAVORannotator's efficiency keeps cloud computing costs low, for example, costing ~\$25 to annotate the TOPMed Freeze 10a WGS data by chromosome in parallel, e.g., 3 CPU hours for chromosome 1.

DISCUSSION

FAVOR offers a comprehensive solution for the application of whole genome variant functional annotations, including an open access and downloadable database, a user-friendly browser, and a tool FAVORannotator, to annotate large-scale genetic data. The FAVOR database is a large relational data structure of multi-faceted functional annotations of all possible 8,812,917,339 SNVs and 79,997,898 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 a genetic study at scale, such as GWAS, WGS and WES studies, and build a highly compressed and well organized aGDS file, that includes both genotype data and their annotations and can be easily integrated into downstream analysis pipelines. Together, FAVOR and FAVORannotator provide a critical infrastructure for facilitating downstream analysis and interpretation of GWAS/WGS/WES studies.

Although several compression methods are available for storing WGS data, such as gzip (vcf.gz), Bgzip or BCF(52), 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 only store genotype data and do not store variant annotations. The aGDS format resolves both limitations successfully.

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.

FUNDING

This work was supported by grant nos. R35-CA197449, P01-CA134294, U19-CA203654 and R01-HL113338 (to X. Lin), U01-HG012064 (to Z. Weng and X. Lin), U01-HG009088 (to X. Lin, S.R.S. and B.M.N.).

AVAILABILITY

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

CONFLICT OF INTEREST

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. G.R.A. is an employee of Regeneron Pharmaceuticals and owns stock and stock options for Regeneron Pharmaceuticals. X. L. is a consultant of AbbVie Pharmaceuticals. Z. W. co-founded and serves as a scientific advisor for Rgenta Inc. 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.

Reference

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*.
3. Investigators, A.o.U.R.P. (2019) The “All of Us” research program. *New England Journal of Medicine*, **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. Li, Z., Li, X., Zhou, H., Gaynor, S.M., Selvaraj, M.S., Arapoglou, T., Quick, C., Liu, Y., Chen, H., Sun, R. *et al.* (2021) A framework for detecting noncoding rare variant associations of large-scale whole-genome sequencing studies. *bioRxiv*, 2021.2011.2005.467531.
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.

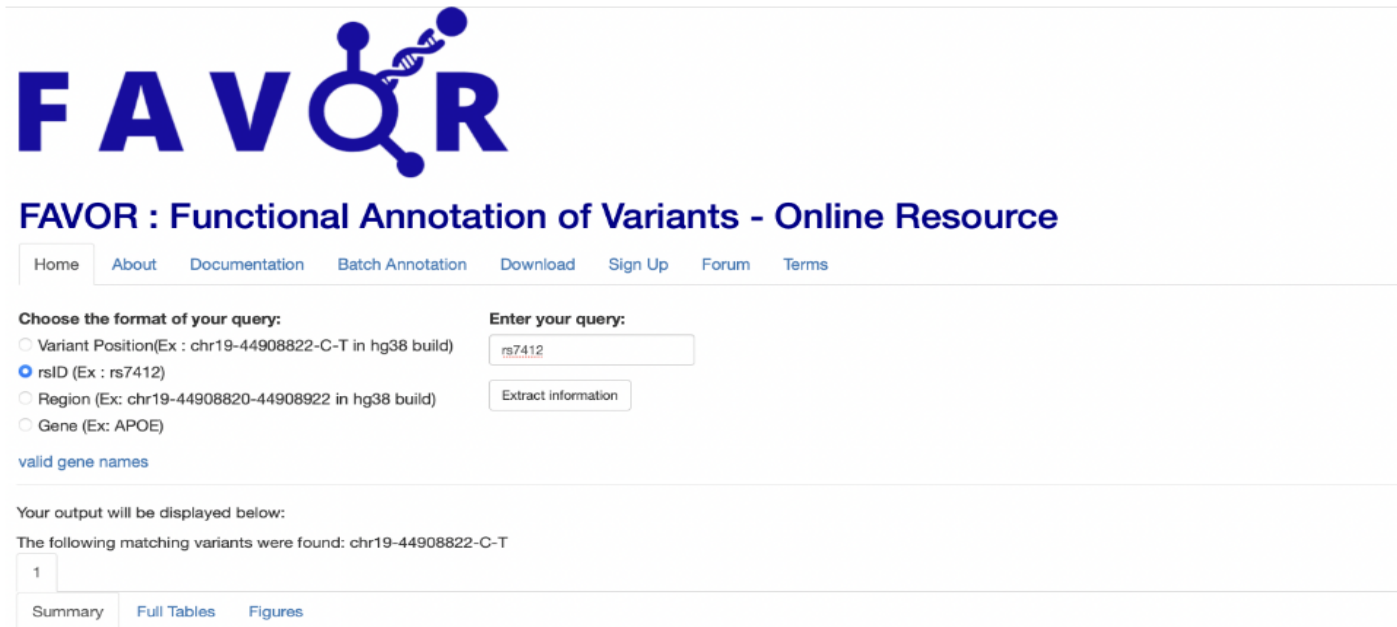
21. Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K. *et al.* (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res*, **46**, D794-D801.
22. Rao, S.S., Huntley, M.H., Durand, N.C., 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., Rusczycki, 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 Bimap: 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. Coordinators, N.R. (2014) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, **42**, D7-17.
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. Ewan Birney, R.C., Andy Clark, Daniele Fallin, Jonathan Haines, Monica Justice, Rod McInnes, Len Pennacchio. NHGRI Genome Sequencing Program, pp. <https://www.genome.gov/Funded-Programs-Projects/NHGRI-Genome-Sequencing-Program>.
33. 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.
34. 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.
35. 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.
36. Wang, K., Li, M. and Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, **38**, e164.

37. 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) WGS: an annotation pipeline for human genome sequencing studies. *J Med Genet*, **53**, 111-112.
38. 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.
39. 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**, D980-985.
40. Zheng, X., Gogarten, S.M., Lawrence, M., Stimp, 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.
41. 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.
42. Consortium, E.P., Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shores, 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. 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.
44. Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68-74.
45. Consortium, F., the, R.P., Clst, Forrest, A.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J., Haberle, V., Lassmann, T. *et al.* (2014) A promoter-level mammalian expression atlas. *Nature*, **507**, 462-470.
46. 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.
47. 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.
48. 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**.
49. 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.
50. 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.
51. 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**.

52. 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**.

Figures

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



The screenshot displays the FAVOR web interface. At the top, the logo 'FAVOR' is shown in blue, with a stylized DNA double helix integrated into the letter 'O'. Below the logo is the title 'FAVOR : Functional Annotation of Variants - Online Resource'. A navigation menu includes links for Home, About, Documentation, Batch Annotation, Download, Sign Up, Forum, and Terms. The main content area features a search form with the heading 'Choose the format of your query:' and four radio button options: 'Variant Position (Ex : chr19-44908822-C-T in hg38 build)', 'rsID (Ex : rs7412)', 'Region (Ex: chr19-44908820-44908922 in hg38 build)', and 'Gene (Ex: APOE)'. The 'rsID' option is selected. To the right, there is a text input field containing 'rs7412' and a button labeled 'Extract information'. Below the search form, there is a link for 'valid gene names'. The results section states 'Your output will be displayed below:' and 'The following matching variants were found: chr19-44908822-C-T'. A table with one row and one column containing the number '1' is shown. At the bottom, there are three tabs: 'Summary', 'Full Tables', and 'Figures', with 'Summary' being the active tab.

Figure 2. Single Variant Query Summary page: The Single Variant Query Summary page 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 | | | Variant Category | | ClinVar | |
|--------------------------|--------------------|------------|---------------------------------------|-------------------|---|---|
| Variant | chr19-44908822-C-T | | Gencode Comprehensive Info | APOE | Clinical Significance | drug_response |
| rsID | rs7412 | | Gencode Comprehensive Category | exonic | Clinical Significance (genotype includes) | 441262:Pathogenic, 441265:Pathogenic, 441266:Pathogenic, 666796:Uncertain_significance |
| TOPMed QC Status | PASS | | Gencode Comprehensive Exonic Category | nonsynonymous SNV | Disease Name | Familial_type_3_hyperlipoproteinemia, Warfarin_response, atorvastatin_response_-_Efficacy, not_provided |
| TOPMed Bravo AF | 0.0821388 | | CAGE Promoter | YES | Disease Name (included variant) | Apollipoproteinemia_E1, Familial_type_3_hyperlipoproteinemia, not_specified |
| GNOMAD Total AF | 0.0788183 | | | | | |
| ALL 1000G AF | 0.07508 | | | | | |
| Integrative | | | Protein | | Conservation | |
| Score | PHRED | Percentile | aPC-Protein-Function | 39.81 | aPC-Conservation | 15.81 |
| aPC-Protein-Function | 39.81 | 0.01 | PolyPhenCat | probably_damaging | mamPhCons | 1.00 |
| aPC-Conservation | 15.81 | 2.62 | PolyPhenVal | 1 | priPhyloP | 0.42 |
| aPC-Epigenetics | 10.85 | 8.22 | Polyphen2 HDIV | 1.0 | GerpN | 13.40 |
| aPC-Transcription-Factor | 13.17 | 4.82 | Polyphen2 HVAR | 1.0 | | |
| aPC-Mappability | 14.58 | 3.48 | Grantham | 180 | | |
| aPC-Proximity-To-TSS-TES | 19.00 | 1.26 | MutationTaster | 0.930 | | |
| CADD PHRED | 26.30 | 0.23 | SIFTcat | deleterious | | |
| | | | SIFTval | 0 | | |
| Epigenetics | | | | | | |
| Open | aPC-Epigenetics | 10.85 | | | | |
| Open | DNase | 0.47 | | | | |
| Open | H3K4me2 | 3.80 | | | | |
| Open | H3K4me3 | 6.45 | | | | |
| Open | H3K9ac | 7.14 | | | | |
| Open | H4K20me1 | 4.91 | | | | |
| Open | H2AFZ | 7.76 | | | | |
| Closed | H3K27me3 | 4.98 | | | | |
| Transcription | totalRNA | 16.31 | | | | |

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

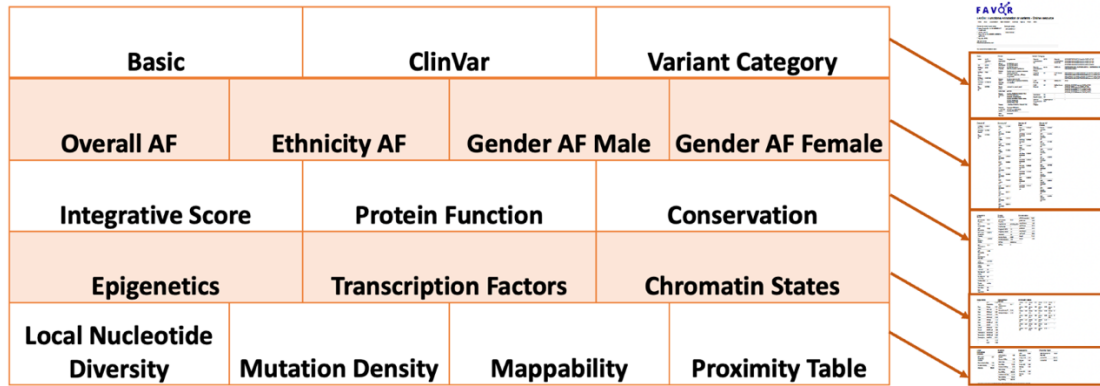


Figure 4. Single Variant Query Functional Annotation Visualization: The Figures tab in the FAVOR Single Variant Online Query displays a visualization of the functional annotation results of a queried variant in histogram.

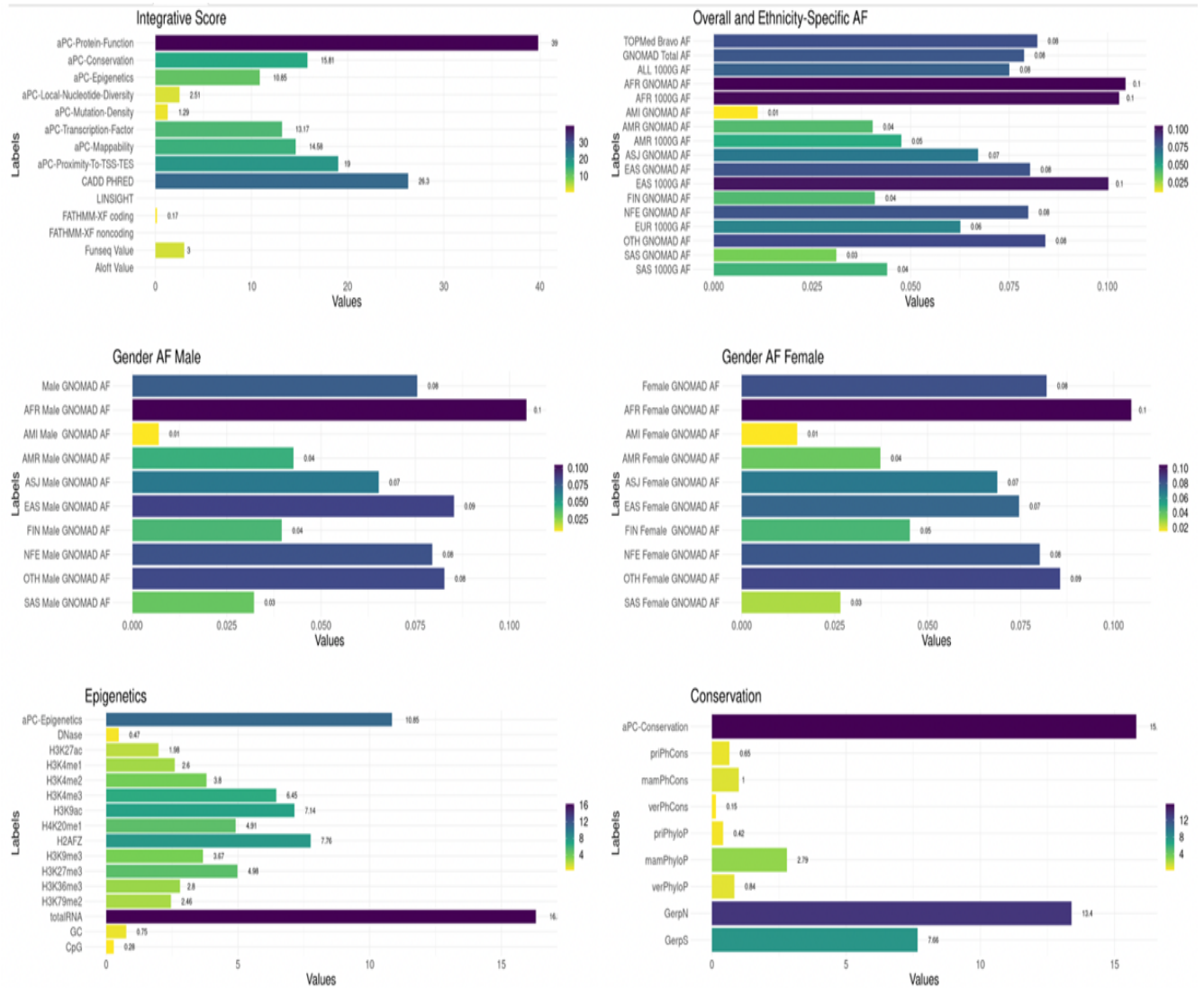


Figure 5. Region-/Gene-based Query Summary tab: The Summary tab of the Region-/Gene-based Query shows the multi-faceted functional annotation summary statistics of the variants in a gene or a region.

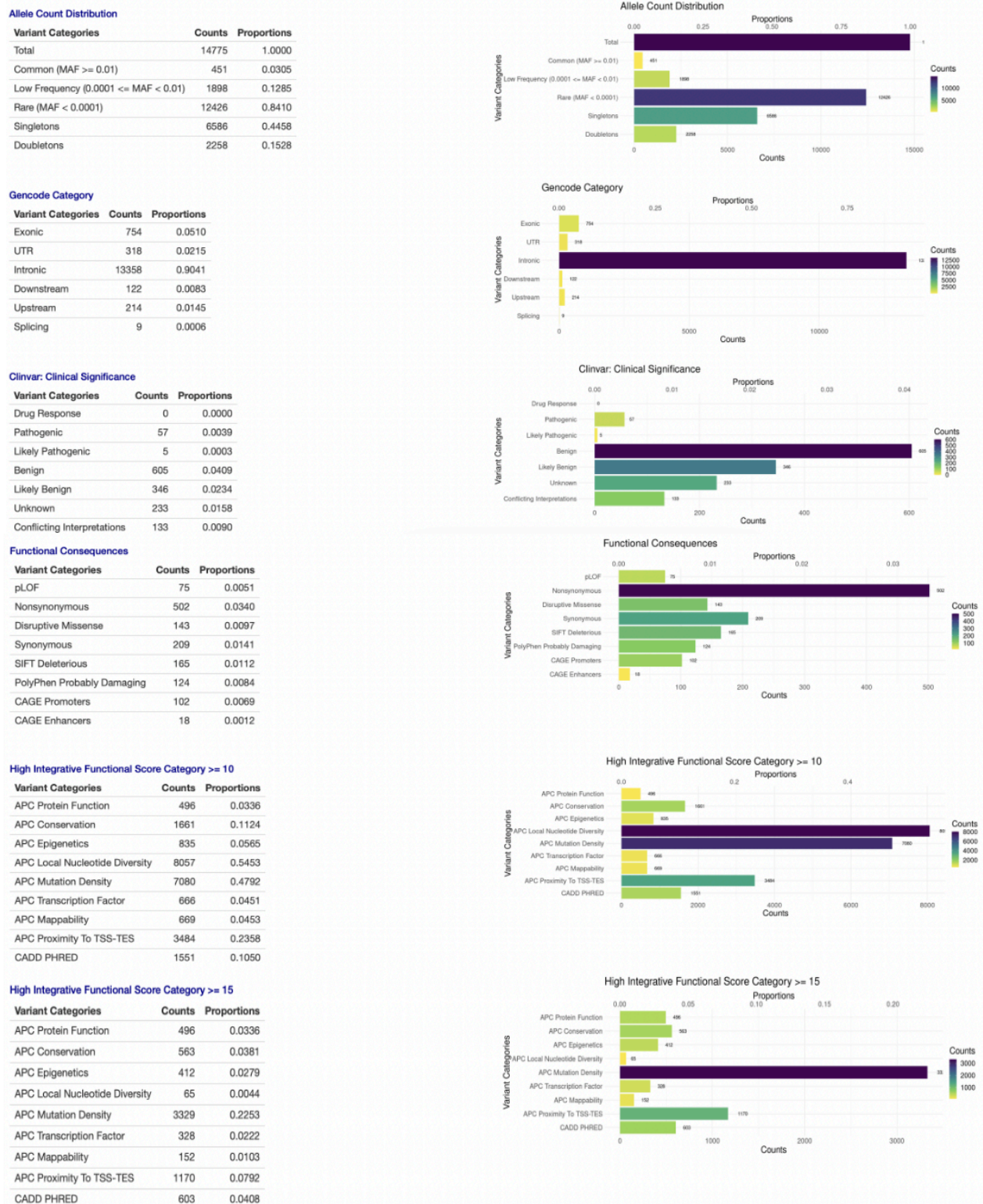


Figure 6. Features of the annotation Genomics Data Structure (aGDS) format: This figure shows the features of the aGDS format and the process of creating aGDS files by combining functional annotations with genotype data.

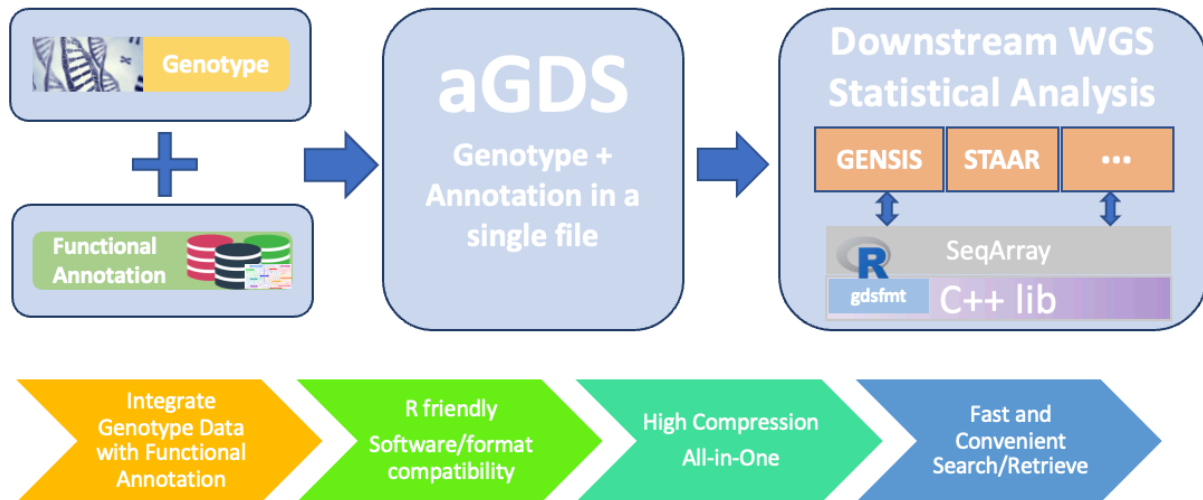
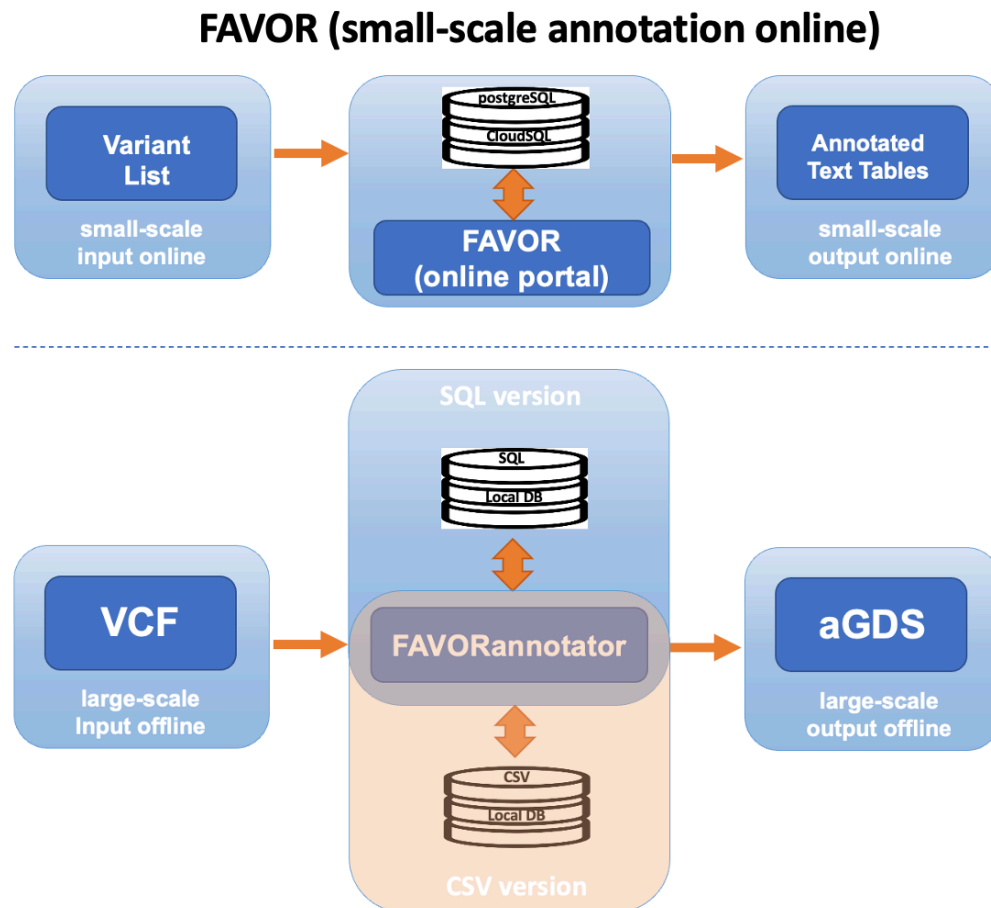


Figure 7. Graphical Representation of the features of FAVOR batch annotation and FAVORannotator: For small-scale annotation (up to 10,000 variants), batch annotation can be used for online annotation at the FAVOR website. For large-scale annotation, e.g., hundreds of millions of variants in a Whole Genome/Exome 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 backend database FAVOR, which is available in the SQL or CVS formats, and outputs an aGDS file that integrates genotype and annotation data in a single file.



FAVORannotator (Large-scale annotation):

R-script to annotate any WGS/WES study using the FAVOR database

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

WORKSPACES FAVORannotator

COVID-19 Data & Tools Cloud Environment None

DASHBOARD DATA NOTEBOOKS WORKFLOWS JOB HISTORY

← Back to list

FAVORannotator

Snapshot: 1

Source: FAVOR-AnVIL/FAVORannotator/12

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

SCRIPT INPUTS OUTPUTS RUN ANALYSIS CANCEL SAVE

Hide optional inputs Download json | Drag or click to upload json SEARCH INPUTS

| Task name ↓ | Variable | Type | Attribute |
|----------------------|----------------|------|----------------------------|
| FAVORannotator | CHRN | Int | 1 |
| FAVORannotator | InputaGDS | File | gs://UKBB200kWES.chr1.gds |
| FunctionalAnnotation | FAVORannotator | File | gs://FAVORannotatorTerra.R |