Title: NIAGADS Alzheimer's GenomicsDB: A resource for exploring Alzheimer's Disease genetic and genomic knowledge

Authors

Emily Greenfest-Allen[24], Conor Klamann[123], Prabhakaran Gangadharan[123], Amanda Kuzma[123], Yuk Yee Leung[123], Otto Valladares[123], Christian J. Stoeckert Jr. [124], Li-San Wang

Affiliations

Corresponding Author

Emily Greenfest-Allen

Research in Context (<=150 words; separate to own document)

Systematic Review

Interpretation

Future Directions

Abstract

INTRODUCTION

The NIAGADS Alzheimer's Genomics Database is an interactive knowledgebase for AD genetics that provides access to GWAS summary statistics datasets deposited at NIAGADS, a national genetics data repository for AD and related dementia (ADRD).

METHODS

The website makes available >70 genome-wide summary statistics datasets from GWAS and genome sequencing analysis for AD/ADRD. Variants identified from these datasets are mapped to up-to-date variant and gene annotations from a variety of resources and linked to functional genomics data.

The database is powered by a big-data optimized relational database and ontologies to consistently annotate study designs and phenotypes, facilitating data harmonization and efficient real-time data analysis and variant or gene report generation.

RESULTS

Detailed variant reports provide tabular and interactive graphical summaries of known ADRD associations, as well as highlight variants flagged by the Alzheimer's Disease Sequencing Project (ADSP). Gene reports provide summaries of co-located ADRD risk-associated variants and have been expanded to include meta-analysis results from aggregate association tests performed by the ADSP allowing us to flag genes with genetic-evidence for AD.

DISCUSSION

The GenomicsDB makes available >100 million variant annotations, including ~30 million (5 million novel) variants identified as AD-relevant by ADSP, for browsing and real-time mining via the website or programmatically through a REST API. With a newly redesigned, efficient, search interface and comprehensive record pages linking summary statistics to variant and gene annotations, this resource makes these data both accessible and interpretable, establishing itself as valuable tool for AD research.

1 Background

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that affects 5.8 million people in US in 2018, is effectively untreatable, and invariably progresses to complete incapacitation and death 10 or more years after onset. Early work in the 1990s identified mutations in the amyloid precursor protein (APP) gene, presenilins 1 and 2 that cause AD, and alleles of the apolipoprotein E gene (APOE) that increase (ε4) or decrease (ε2) susceptibility to late-onset Alzheimer's disease (LOAD). Heritability of AD is high, ranging from near 60% to 80% in the best fitting model [1,2]. However, apart from APOE, there is no simple pattern of inheritance for LOAD. Instead, it is likely caused by a complex combination of common, polygenic variants [3] acting together with a small number of rare variants with a large effect [4,5].

Our current understanding of genetic risk for AD has resulted mainly from massive genotyping and sequencing efforts such as the Alzheimer's Disease Genetics Consortium (ADGC), the International Genomics of Alzheimer's Project (IGAP), and the Alzheimer's Disease Sequencing Project (ADSP). Large-scale genome wide association studies (GWAS) and GWAS-derived meta-analyses have been performed by each of these groups [4–7], the results of which are deposited at the National Institute of Aging (NIA) Genetics of Alzheimer's Disease Data Storage Site (NIAGADS) at the University of Pennsylvania [8]. NIAGADS is an NIA-designated essential national infrastructure, providing a one-stop access portal for Alzheimer's disease 'omics datasets. Qualified investigators can submit data use requests to access protect personal genetic information. NIAGADS also disseminates meta-analysis results and GWAS summary statistics for researchers to explore knowledge on AD genetic risks. However, substantive bioinformatics expertise and compute power are required to annotate and mine these datasets, which is a significant hurdle for many researchers planning to explore this large and ever-increasing volume of data. Assembly of unrestricted genomic knowledge into an integrated, interactive web resource would help overcome this barrier.

Here, we introduce the NIAGADS Alzheimer's Genomics Database (GenomicsDB), which was developed in collaboration with the ADGC and ADSP with this goal in mind. The GenomicsDB is a user-friendly workspace for data sharing, discovery, and analysis designed to facilitate the quest for better understanding of the complex genetic underpinnings of AD neurodegeneration and accelerate the progress of research on AD and AD related dementias (ADRD). It accomplishes this by making summary genetic evidence for Alzheimer's disease both accessible to and interpretable by molecular biologists, clinicians and bioinformaticians alike regardless of computational skills.

2 Methods

2.1 Genomics Datasets

2.1.1 NIAGADS GWAS summary statistics

As of August 2020, the GenomicsDB provides unrestricted access to genome-wide summary statistics p-values from >70 GWAS and ADSP meta-analysis. Summary statistic results are linked to >100 million ADSP annotated single-nucleotide variants (SNVs) and indels. GWAS summary statistics datasets deposited at NIAGADS are integrated into the GenomicsDB as they become publicly available via publication or permission of the submitting researchers. These include studies that focus specifically on AD and late-onset AD (LOAD), as well as those on ADRD-related neuropathologies and biomarkers. A full listing of the summary statistics datasets currently available through the NIAGADS GenomicsDB is provided in **Table 1**.

Prior to loading in the database, the datasets are annotated (e.g. provenance, phenotypes, study design) and variant representation normalized (chr:pos:ref_allele:alt_allele on the forward strand in a 1-based coordinate system) to ensure consistency with ADSP analysis pipelines and facilitate harmonization with third-party annotations. To ensure data privacy, the NIAGADS GenomicsDB website only makes *p*-values from the summary statistics available for browsing (on dataset, gene, and variant reports and as genome browser tracks) and analysis. Access to the full summary statistics (including genome-wide allele frequencies and effect sizes) and corresponding GWAS or sequencing results is managed via formal data-access requests made to NIAGADS. All datasets included in the GenomicsDB are properly credited to the submitting researchers or sequencing project.

2.1.2 NHGRI-EBI GWAS Catalog

Variants and summary statistics curated in the NHGRI-EBI GWAS catalog [9] are listed in NIAGADS GenomicsDB variant reports and a track is available on the genome browser. Results linking variants to AD/ADRD are highlighted.

2.1.3 ADSP meta-analysis results

The NIAGADS GenomicsDB has recently expanded its scope to include meta-analysis results offering genetic evidence for gene-level and single-variant risk associations for AD. Currently available are case/control association results recently published by the ADSP [7] and deposited at NIAGADS (Accession number NG00065).

2.2 Variant annotation

2.2.1 Variant identification

Single nucleotide polymorphisms (SNPs) and short-indels are uniquely identified by position and allelic variants. All variants are reported on the forward strand in a 1-based coordinate system. This allows accurate mapping of risk-association statistics to specific mutations and to external variant annotations from resources such as gnomAD (https://gnomad.broadinstitute.org/) [10] and GTex (https://www.gtexportal.org/home/) [11]. All variants are mapped to dbSNP (https://www.ncbi.nlm.nih.gov/snp/) [12] and linked to refSNP identifiers when possible.

2.2.2 ADSP variant annotations

Annotated variants in the NIAGADS GenomicsDB include the >29 million SNPs and ~50,000 short-indels identified during the ADSP Discovery Phase whole-genome (WGS) and whole-exome sequencing (WES) efforts [13]. These variants are highlighted in variant and dataset reports and their quality control status is provided. As part of this sequencing effort, the ADSP developed an annotation pipeline that builds on Ensembl's VEP software [14] to efficiently integrate standard annotations and rank potential variant impacts according to predicted effect (such as codon changes, loss of function, and potential deleteriousness) [13,15]. Variant tracks annotated by these results are available for both the WES and WGS variants on the GenomicsDB genome browser.

The pipeline has been applied to all variants in the database; annotations can be browsed on variant reports or used to filter search results. User uploaded lists of variants are automatically annotated in real-time.

2.2.3 Allele frequencies

The NIAGADS GenomicsDB includes allele frequency data from 1000 Genomes (phase 3, version 1) (https://www.internationalgenome.org/home) [16], ExAC (http://exac.broadinstitute.org/) [17], and gnomAD [10].

2.2.4 Linkage disequilibrium

Linkage-disequilibrium (LD) structure around annotated variants is estimated using phase 3 version 1 (11 May 2011) of the 1000 Genomes Project [16]. LD estimates were made using PLINK v1.90b2i 64-bit [18]. Only LD-scores meeting a correlation threshold of $r^2 \geq 0.2$ are stored in the database. *Locuszoom.js* [19,20] is used to render LD-scores in the context of the GWAS summary statistics datasets.

2.3 Gene and transcript annotation

2.3.1 Gene identification

Gene and transcript models are obtained from the GENCODE Release 19 (GRCh37.p13) reference gene annotation; a GRCh38 version of the GenomicsDB will be released in early 2021 [21]. Standard gene nomenclature is imported from the HUGO Gene Nomenclature Committee at the European Bioinformatics Institute [22] and used to link annotated genes to external resources such as UniProt (https://www.uniprot.org/) [23], the UCSC Genome Browser (http://genome.ucsc.edu)[24], and Online Mendelian Inheritance in Man (OMIM) database (https://omim.org/) [25,26].

2.3.2 Functional annotation

Annotations of the functions of genes and gene products are taken from packaged releases of the Gene Ontology (GO; http://geneontology.org) and GO-gene associations [27] and are updated regularly. GO-gene associations are reported in summary tables on gene reports and include details on annotation sources, as well as new information from the GO causal modeling (GO-CAM) framework that allows better understanding of how different gene products work together to effect biological processes [28].

Users can run functional enrichment analysis on gene search results or uploaded gene lists. Geneset enrichment and semantic similarity scores are calculated using the *goatools* Python library for GO analysis [29].

2.4.3 Pathways

Gene membership in molecular and metabolic pathways is provided from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (https://www.genome.jp/kegg/) [30] and Reactome (https://reactome.org/) [31]. Users can run pathway enrichment analysis on gene search results or uploaded gene lists. Pathway enrichment statistics are calculated using a multiple hypothesis corrected Fisher's exact test implemented using the *SciPy*, *pandas*, and *statsmodels* Python packages.

2.4 Functional genomics

More than 50,000 functional genomics tracks have been integrated into the NIAGADS Alzheimer's Genomics Database and mapped against AD/ADRD-associated variants. Functional genomics data are queried from the NIAGADS Functional genomics repository (FILER; https://lisanwanglab.org/FILER), which provides harmonized functional genomics datasets that have been GIGGLE indexed [32] for quick lookups. FILER tracks made available through the GenomicsDB have been pulled from multiple functional genomics resources, including the Encyclopedia of DNA Elements (ENCODE) [33,34], the Functional Annotation of the Mouse/Mammalian Genome (FANTOM5) enhancer atlas [35], and the NIH Roadmap Epigenomics Mapping Consortium [36]. Genome browser tracks are available for all functional genomics datasets and are organized by data source, biotype (e.g., cell, tissue, or cell line), type of functional annotation (e.g., expressed enhancers, transcription factor binding sites, histone modifications) and platform or assay type to facilitate track selection.

2.5 Overview of database design

The GenomicsDB is powered by a PostgreSQL relational database system that has been optimized for parallel big-data querying, allowing for efficient real-time data mining. Data are organized using the modular Genomics Unified Schema version 4 (GUS4), which was designed for scalable integration and dissemination of large-scale 'omics datasets. Loading of all data, including summary statistics datasets and sequence feature annotations, is managed by the

GUS4 application layer (https://github.com/VEuPathDB/GusAppFramework), which ensures the accuracy and consistency of the data integration.

2.6 Overview of website design and organization

The NIAGADS GenomicsDB is powered by an open-source database system and web-development kit (WDK; https://github.com/VEuPathDB/WDK) developed and successfully deployed by the Eukaryotic Pathogen, Vector and Host Informatics (VEuPathDB) Bioinformatics Resource Center [37,38]. The VEuPathDB WDK provides a query engine that ties the database system to the website via an easily extensible data model. The data model automatically generates and organize searches, search results, and detailed gene and variant reports, with concepts and data organized by topics from the EMBRACE Data And Methods (EDAM) ontology, which defines a comprehensive set of well-established, familiar concepts that are prevalent within bioinformatics and computational biology [39]. This allows for seamless updates of third-party data and rapid integration of new datasets as they become publicly available.

The WDK also provides a framework for lightweight Java/Jersey representational state transfer (REST) services for data querying. This allows search results and dataset or feature reports to be returned in multiple file formats (e.g., delimited-text, XML, and JSON) in addition to browsable, interactive web pages. This new feature of GenomicsDB facilitates programmatic access and has enabled the inclusion of sophisticated JavaScript visualizations for summarizing search results and interactive graphical filters and data exploration tools in our gene and variant reports. The GenomicsDB uses a combination of an in-house JavaScript genomics visualization toolkit and established third-party visualization tools, including the *HighCharts.js* (https://www.highcharts.com/) charting library for rendering scatter, pie, and bar charts, *ideogram.js* (https://github.com/eweitz/ideogram) for chromosome visualization, *LocusZoom.js* for rendering LD annotations in the context of NIAGADS GWAS summary statistics datasets, and an *IGV.js* powered genome browser [40].

All code used to generate the WDK website, including the JavaScript genomics visualization toolkit are available on GitHub (https://github.com/NIAGADS).

2.7 Overview of the NIAGADS genome browser

The NIAGADS genome browser provides an interactive graphical display that enables researchers to visually inspect and browse entire GWAS summary statistics datasets in a genomic context. An embedded browser snapshot is included on every GenomicsDB variant and gene report, providing a quick overview of the sequence feature relative to the genome sequence and any proximal variants or genes. The GenomicsDB website also includes a fully functional genome browser that allows user to compare NIAGADS GWAS summary statistics tracks to each other, against annotated gene or variant tracks, or to the more than >50,000 functional genomics tracks from the NIAGADS FILER functional genomics repository. This tool is powered by *IGV.js*, with track data and interactive track metadata underlying our customized, user-friendly track selection tool queried in real-time via NIAGADS GenomicsDB REST services.

Documentation for the REST endpoints serving data tracks and code for the NIAGADS-customized IGV browser tracks and the JavaScript track selection tool are available on the GitHub (https://github.com/NIAGADS), allowing users to deploy their own *IGV.js* or *IGV-Jupyter* browsers with the facility to render GenomicsDB data tracks.

3. Results

The NIAGADS GenomicsDB creates a public forum for sharing, discovery, and analysis of genetic evidence for Alzheimer's disease that is a made them accessible via an interface designed for easy mastery by biological researchers, regardless of background.

3.1 Finding variants, genes, and datasets

The GenomicsDB homepage and navigation menu contain a quick site search allowing users to quickly find variants, genes, and datasets of interest by identifier or keyword (**Fig XXXA**). The GenomicsDB also provides a dataset browser that allows users to search for GWAS summary statistics datasets by AD/ADRD phenotype, population, genotype, attribution, and sequencing center (**Fig. XXXB**).

3.1 Browsing and mining NIAGADS GWAS summary statistics

A detailed report is provided for each of the GWAS summary statistics datasets in the NIAGADS GenomicsDB (**Fig. X**). These reports allow users to interactively browse the genetic variants with genome-wide significance in the dataset (p-value $\leq 5 \times 10^{-8}$ to account for false positives due to testing associations of millions of variants simultaneously) via tables and interactive plots that provide an overview of the distribution and potential functional or regulatory impacts of the top variants (and proximal gene-loci) across the genome. All genes and variants listed in a dataset report are linked to reports in the GenomicsDB that provide detailed information about genetic evidence for AD for the sequence feature (see next sections). Dataset reports also provide quick links back to their parent accession in the NIAGADS repository, where users can download the complete p-values or make formal data access requests for the full summary statistics, or related GWAS, expression, or sequencing data.

Users have several options for exploring a dataset in more depth. The NIAGADS genome browser can be used to visually inspect the dataset in a broader genomic context and compare against other 'omics tracks in the GenomicsDB or FILER (see section 2.7). Dataset reports also provide an inline search (**Fig. X, inset A**) allowing users to mine the summary statistics in real-time via the website, setting their own p-value cut-off. Variants meeting the cut-off are reported in an interactive table that can be sorted or filtered by annotations (e.g., variant type, predicted effect, deleteriousness) (**Fig. XXA**). A per-chromosome genome view is also available (**Fig. XXB**) allowing users to explore an interactive ideogram depicting the distribution of variants meeting the search and filter criteria across the genome and allowing inspection of LD structure among proximal variants.

Tables of results can be downloaded or requested via the API for programmatic processing. Registered users also have the option to save and share search results both privately and publicly; publicly shared search results are assigned a stable URL that can be referenced in publications.

3.2 Detailed variant reports

Variant reports include a basic summary about the variant (alleles, variant type, flanking sequence, genomic location) and a genome browser snapshot placing the variant in a broader genomic context (**Fig. XXXX**). Links to the NIAGADS Alzheimer's Disease Variant Portal (ADVP) and external resources for additional information (e.g., dbSNP, ClinVar) are also provided. All other information in the report is subdivided into multiple sections that can be expanded or hidden at the user's discretion. These sections include sub-reports on genetic variation (e.g., allele population frequencies and LD), function prediction determined via the ADSP annotation pipeline (incl. transcript and regulatory consequences), and comprehensive listings of GWAS inferred disease or trait associations from both NIAGADS summary statistics and the NHGRI-EBI GWAS Catalog. Tables listing summary statistics results can be dynamically filtered by *p*-value, dataset, phenotypes, or covariates, and the filtered results are downloadable. Links to the source datasets for each reported statistic are also provided, leading to detailed dataset reports (e.g., NIAGADS GWAS summary statistics) or to the source publication (e.g., curated variant catalogs). These tables are paired with browsable *LocusZoom.js* views of the LD structure surrounding the variant in the context of selected GWAS summary statistics datasets.

3.3 Detailed gene reports

Like the variant reports, gene reports provide basic summary information about the gene (nomenclature, gene type, genomic span) and a genome browse snapshot highlighting ADSP annotated and NIAGADS GWAS summary statistics-linked variants proximal to or within the footprint of the gene (**Fig. XXXXX**). Links to the NIAGADS ADVP and to external resources (e.g., UniprotKB, OMIM, and ExAC) are also included in the summary. Two types of gene-linked genetic evidence for AD are provided in the GenomicsDB gene reports. First, we have surveyed the top risk-associated variants from the NIAGADS GWAS summary statistics datasets and provide a comprehensive listing of and links to those contained within ±200kb of each gene. Second, we report meta-analysis results from gene-based rare variant aggregation tests performed as part of the ADSP discovery phase case/control analysis [41]. Genes found to have a significant *p*-value in these results are flagged as being associated with genetic-evidence for AD. Also provided on the gene report are sections reporting function prediction (Gene Ontology associations and evidence) and pathway membership (KEGG and Reactome). Tables reporting these results or annotations can be dynamically filtered or downloaded.

3.4 Analysis Tools

TBD

4 Discussion

The NIAGADS Alzheimer's Genomics Database is a user-friendly platform for interactive browsing and real-time in-depth mining of published genetic evidence and genetic risk-factors for AD. It provides open, real-time access to >50 summary statistics datasets from genome-wide association analysis (GWAS) of Alzheimer's disease and related neuropathologies. Flexible search options allow users to easily retrieve AD risk-associated variants, conditioned on phenotypes such as ethnicity and age of onset. Users can compare the NIAGADS datasets against personal gene or variant lists.

Every entry in the GenomicsDB has been linked with relevant external resources and functional genomics annotations to supply further information and assist researchers in interpreting the potential functional or regulatory role of risk-associated variants and susceptibility loci. The GenomicsDB is updated periodically with enhanced features and new datasets and annotations when they are reported. The AD research community is actively encouraged through outreach and collaboration to submit data to NIAGADS to keep this public platform updated and timely.

The GenomicsDB is integrated with other resources available at NIAGADS. Users can follow links back to the NIAGADS repository to view comprehensive details about all GWAS summary statistics datasets from NIAGADS accession or request access to the primary data. The REST services used to query the database and generate data or feature reports provide the foundation of an API that allows programmatic access to the database, which we plan to integrate with cloud based NIAGADS analysis pipelines.

The GenomicsDB is regularly updated to keep up with advances in Alzheimer's disease genomics research. New AD-related GWAS summary statistics datasets and meta-analysis results from the ADSP are added as they become available. Reference databases are updated yearly. All genomics data in the current version of the GenomicsDB are aligned and mapped to the GRCh37.p13 genome build. A GRCh38 version of the database is planned for release in early 2021.

GenomicsDB can be a potent platform for the AD genetics community to host comprehensive AD genetic and genomic findings. It uses the latest web and database technologies to allow integration with new tools, and NIAGADS is constantly improving. As more data and tools are available the NIAGADS Alzheimer's Genomics Database will become a central hub for AD/ADRD research and data analysis.

6 References

[1] Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, et al. Role of genes and environments for explaining Alzheimer disease. Arch Gen Psychiatry 2006;63:168–74. https://doi.org/10.1001/archpsyc.63.2.168.

[2] Jansen IE, Savage JE, Watanabe K, Bryois J, Williams DM, Steinberg S, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nature Genetics 2019;51:404–13. https://doi.org/10.1038/s41588-018-0311-9.
[3] Hollingworth P, Harold D, Sims R, Gerrish A, Lambert J-C, Carrasquillo MM, et al. Common variants in ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. Nat Genet 2011;43:429–35. https://doi.org/10.1038/ng.803.
[4] Lambert J-C, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nature Genetics 2013;45:1452–8. https://doi.org/10.1038/ng.2802.
[5] Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. Nat Genet 2019;51:414–30. https://doi.org/10.1038/s41588-019-0358-2.
[6] Naj AC, Jun G, Beecham GW, Wang L-S, Vardarajan BN, Buros J, et al. Common variants at MS4A4/MS4A6E , CD2AP , CD33 and EPHA1 are associated with late-onset Alzheimer's disease. Nature Genetics 2011;43:436–41. https://doi.org/10.1038/ng.801.
[7] Bis JC, Jian X, Kunkle BW, Chen Y, Hamilton-Nelson KL, Bush WS, et al. Whole exome sequencing study identifies novel rare and common Alzheimer's-Associated variants involved in immune response and transcriptional regulation. Molecular Psychiatry 2018:1–17. https://doi.org/10.1038/s41380-018-0112-7.
[8] Kuzma A, Valladares O, Cweibel R, Greenfest-Allen E, Childress DM, Malamon J, et al. NIAGADS: The NIA Genetics of Alzheimer's Disease Data Storage Site. Alzheimer's & Dementia 2016;12:1200–3. https://doi.org/10.1016/j.jalz.2016.08.018.
[9] Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 2019;47:D1005–12. https://doi.org/10.1093/nar/gky1120.
[10] Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. BioRxiv 2019:531210. https://doi.org/10.1101/531210.
[11] Gamazon ER, Segrè AV, van de Bunt M, Wen X, Xi HS, Hormozdiari F, et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. Nature Genetics 2018;50:956–67. https://doi.org/10.1038/s41588-018-0154-4.
[12] Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001;29:308–11.
[13] Butkiewicz M, Blue EE, Leung YY, Jian X, Marcora E, Renton AE, et al. Functional annotation of genomic variants in studies of late-onset Alzheimer's disease. Bioinformatics 2018;34:2724–31. https://doi.org/10.1093/bioinformatics/bty177.
[14] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol 2016;17. https://doi.org/10.1186/s13059-016-0974-4.
[15] Wheeler NR, Benchek P, Kunkle BW, Hamilton-Nelson KL, Warfe M, Fondran JR, et al. Hadoop and PySpark for reproducibility and scalability of genomic sequencing studies. Pac Symp Biocomput 2020;25:523–34.

[16] Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature 2015;526:68–74. https://doi.org/10.1038/nature15393.
[17] Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016;536:285–91. https://doi.org/10.1038/nature19057.
[18] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–75. https://doi.org/10.1086/519795.
[19] Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 2010;26:2336–7. https://doi.org/10.1093/bioinformatics/btq419.
[20] Clark CP, Flickinger M, Welch R, VandeHaar P, Taliun D, Boehnke M, et al. LocusZoom.js: Web-based plugin for interactive analysis of genome and phenome wide association studies. Presented at the 66th Annual Meeting of The American Society of Human Genetics, Vancouver: 2016, p. 189T.
[21] Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res 2019;47:D766–73. https://doi.org/10.1093/nar/gky955.
[22] Braschi B, Denny P, Gray K, Jones T, Seal R, Tweedie S, et al. Genenames.org: the HGNC and VGNC resources in 2019. Nucleic Acids Res 2019;47:D786–92. https://doi.org/10.1093/nar/gky930.
[23] UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47:D506–15. https://doi.org/10.1093/nar/gky1049.
[24] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Res 2002;12:996–1006. https://doi.org/10.1101/gr.229102.
[25] Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res 2015;43:D789-798. https://doi.org/10.1093/nar/gku1205.
[26] Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype-gene relationships. Nucleic Acids Res 2019;47:D1038–43. https://doi.org/10.1093/nar/gky1151.
[27] The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res 2019;47:D330–8. https://doi.org/10.1093/nar/gky1055.
[28] Thomas PD, Hill DP, Mi H, Osumi-Sutherland D, Auken KV, Carbon S, et al. Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. Nature Genetics 2019;51:1429–33. https://doi.org/10.1038/s41588-019-0500-1.
[29] Klopfenstein DV, Zhang L, Pedersen BS, Ramírez F, Warwick Vesztrocy A, Naldi A, et al. GOATOOLS: A Python library for Gene Ontology analyses. Scientific Reports 2018;8:1–17. https://doi.org/10.1038/s41598-018-28948-z.
[30] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27–30. https://doi.org/10.1093/nar/28.1.27.

[31] Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Res 2020;48:D498–503. https://doi.org/10.1093/nar/gkz1031.
[32] Layer RM, Pedersen BS, DiSera T, Marth GT, Gertz J, Quinlan AR. GIGGLE: a search engine for large-scale integrated genome analysis. Nat Methods 2018;15:123–6. https://doi.org/10.1038/nmeth.4556.
[33] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57–74. https://doi.org/10.1038/nature11247.
[34] Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res 2018;46:D794–801. https://doi.org/10.1093/nar/gkx1081.
[35] Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature 2014;507:455–61. https://doi.org/10.1038/nature12787.
[36] Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature 2015;518:317–30. https://doi.org/10.1038/nature14248.
[37] Fischer S, Aurrecoechea C, Brunk BP, Gao X, Harb OS, Kraemer ET, et al. The strategies WDK: a graphical search interface and web development kit for functional genomics databases. Database (Oxford) 2011;2011. https://doi.org/10.1093/database/bar027.
[38] Aurrecoechea C, Barreto A, Basenko EY, Brestelli J, Brunk BP, Cade S, et al. EuPathDB: the eukaryotic pathogen genomics database resource. Nucleic Acids Res 2017;45:D581–91. https://doi.org/10.1093/nar/gkw1105.
[39] Ison J, Kalas M, Jonassen I, Bolser D, Uludag M, McWilliam H, et al. EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. Bioinformatics 2013;29:1325–32. https://doi.org/10.1093/bioinformatics/btt113.
[40] Robinson JT, Thorvaldsdóttir H, Turner D, Mesirov JP. igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). BioRxiv 2020:2020.05.03.075499. https://doi.org/10.1101/2020.05.03.075499.
[41] Bis JC, Jian X, Kunkle BW, Chen Y, Hamilton-Nelson KL, Bush WS, et al. Whole exome sequencing study identifies novel rare and common Alzheimer's-Associated variants involved in immune response and transcriptional regulation. Mol Psychiatry 2018. https://doi.org/10.1038/s41380-018-0112-7.