ATAV - a comprehensive platform for population-scale genomic analyses

Zhong Ren[1], Gundula Povysil[1], David B. Goldstein[1]

[1] Institute for Genomic Medicine, Columbia University Irving Medical Center, New York, New York

Abstract

Summary:

We present ATAV, an analysis platform for large-scale whole-exome and whole-genome sequencing projects. ATAV stores variant and coverage data for all samples in a centralized database, which is then efficiently queried by ATAV to support diagnostic analyses for trios and singletons, as well as rare-variant collapsing analyses for finding disease associations in complex diseases. Runtime logs ensure full reproducibility and the modularized ATAV framework makes it extensible to continuous development of new functions. In recent years ATAV has not only been helpful for identifying disease-causing variants for a range of diseases, but has also enabled the discovery of novel genes by rare-variant collapsing on datasets containing more than 20,000 samples. Analyses to date have been performed on more than 110,000 sequenced individuals demonstrating that the framework is robust to large-scale studies.

Availability and implementation:

ATAV is open source, cross-platform compatible, and is available under the MIT license at https://github.com/igm-team/atav

Introduction

Diagnostic and cohort sequencing studies benefit from the analysis of a large number of samples combined with similarly processed controls. A common approach to reach the necessary scale for analysis is to use a joint-calling procedure and store all samples in a single VCF file. While effective in allowing a single analysis of all samples included in the single VCF file, this approach has significant limitations. Perhaps most importantly, this approach is not amenable to ongoing analyses as new samples become available. Moreover, when projects combine multiple cohorts that were not sequenced together and in which controls might be re-used for several studies, the cost and time required to perform joint-calling for each analysis can become prohibitive. In addition to these considerations, typical sequencing file formats (VCF, BAM) place a sizeable overhead in moving these data from physical storage to the compute nodes for dynamic and multi user analysis needs. Furthermore, standard diagnostic and case-control studies leverage a range of filtering parameters, including variant calling (genotype quality, read coverage), variant annotation (gene, effect), internal population frequencies (minor allele frequency, genotype frequency) and external dataset filters (gnomAD[1], RVIS[2]) to identify "qualifying variants" that meet a specific set of user-defined criteria. These sophisticated needs place an additional burden on creating an audit trail for re-analyses and reproducibility. As the size and number of simultaneous users increase, ad-hoc analyses become prohibitively inefficient in the conventional single joint-genotyped VCF framework.

To address these constraints and dynamic analyses needs, we have developed ATAV to streamline genomic analysis needs ranging from the standard diagnostic case interpretation to large-scale cohort analyses for disease-associated gene discovery. The ATAV platform is built on an open source relational database. The database (ATAVDB) is configured with a feature allowing data replication across a cluster of nodes. ATAVDB contains sample variant data, read coverage data, variant annotation data, external annotation data, and metadata. A data pipeline toolkit extracts variants, annotations and associated quality data from VCF files and the coverage and genotype quality from bam files. Currently the Institute for Genomic Medicine (IGM) at Columbia University has data of over 100K whole exomes, and the coding-regions of over 10K whole genomes stored in ATAVDB. It contains over 23 billion variant calls from over 210 million distinct genomic co-ordinates and read coverage information for all samples.

Database

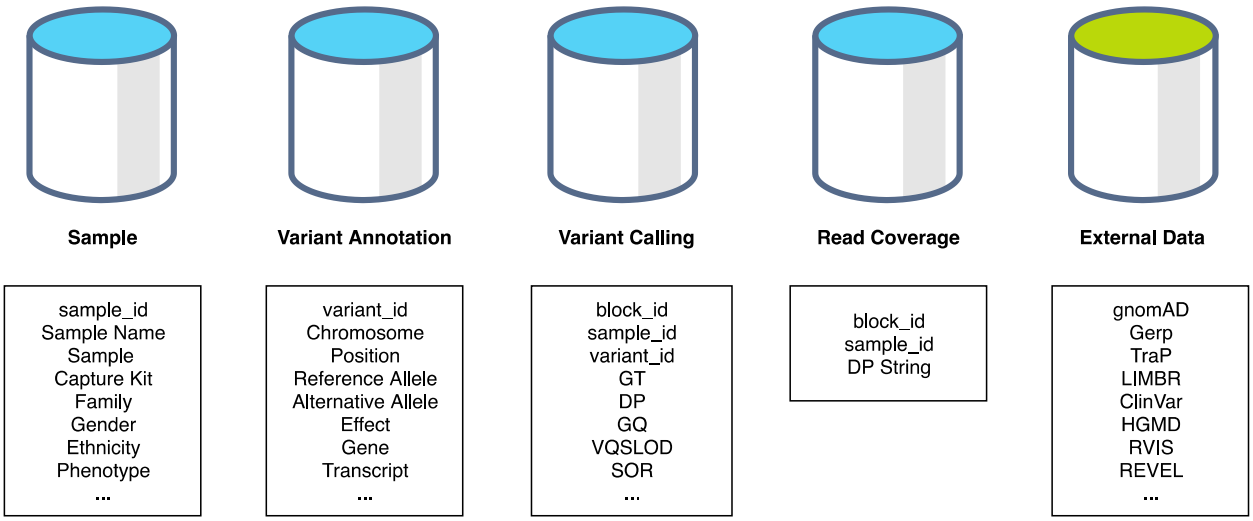


Figure 1: ATAV core database schema and external databases

At the IGM, we use Percona Server for MySQL and its high-performance storage engine Percona TokuDB to improve scalability and operational efficiency. In the database, we store a universal variant list across all samples, annotation data that is annotated through ClinEff[3], sample variants calls and associated quality metrics, as well as all site's coverage data for inferring reference alleles at non-call sites. In addition to that, ATAVDB stores external databases such as allele frequencies from gnomAD[1], ExAC[4], or, DiscovEHR[5], scores such as GERP++[6], TraP[7], LIMBR[8], MTR[9], RVIS[2], subRVIS[10], REVEL[11], PrimateAI[12], CCR[13], as well as clinical annotations from ClinVar[14,15], ClinGen[16], HGMD[17], and OMIM (see Figure 1). ATAV has an external data plugin code structure, which allows quick code integration of gene based, site based and variant based data.

For efficiently storing coverage information for every site and every sample, the ATAV data pipeline parses through the bam files to generate read coverage data and converts site coverage values into bin values: a [0-9]; b [10-19]; c [20-29]; d [30-49]; e [50-199]; f >=200. A Run-length encoding procedure is used to further compress data within fixed 1000 bp block regions (see Figure 2). This way the data size is reduced by about 1000 times making it possible to store the coverage information for more than 100K samples. The information that has been loaded into ATAVDB has been determined over many years of applied use to be that information that is most often required for thestandard genetic analyses performed as part of both diagnostic genetic studies and gene discovery. For example, in diagnostic analyses for identifying *de novo* mutations in affected children, it is necessary to know that the parental samples have sufficient coverage at the relevant site, but not necessary to know the precise number of reads, leading to the binning strategy on coverage described above. For the vast

majority of applications, we have found that the necessary information can be economically stored and retrieved as described.

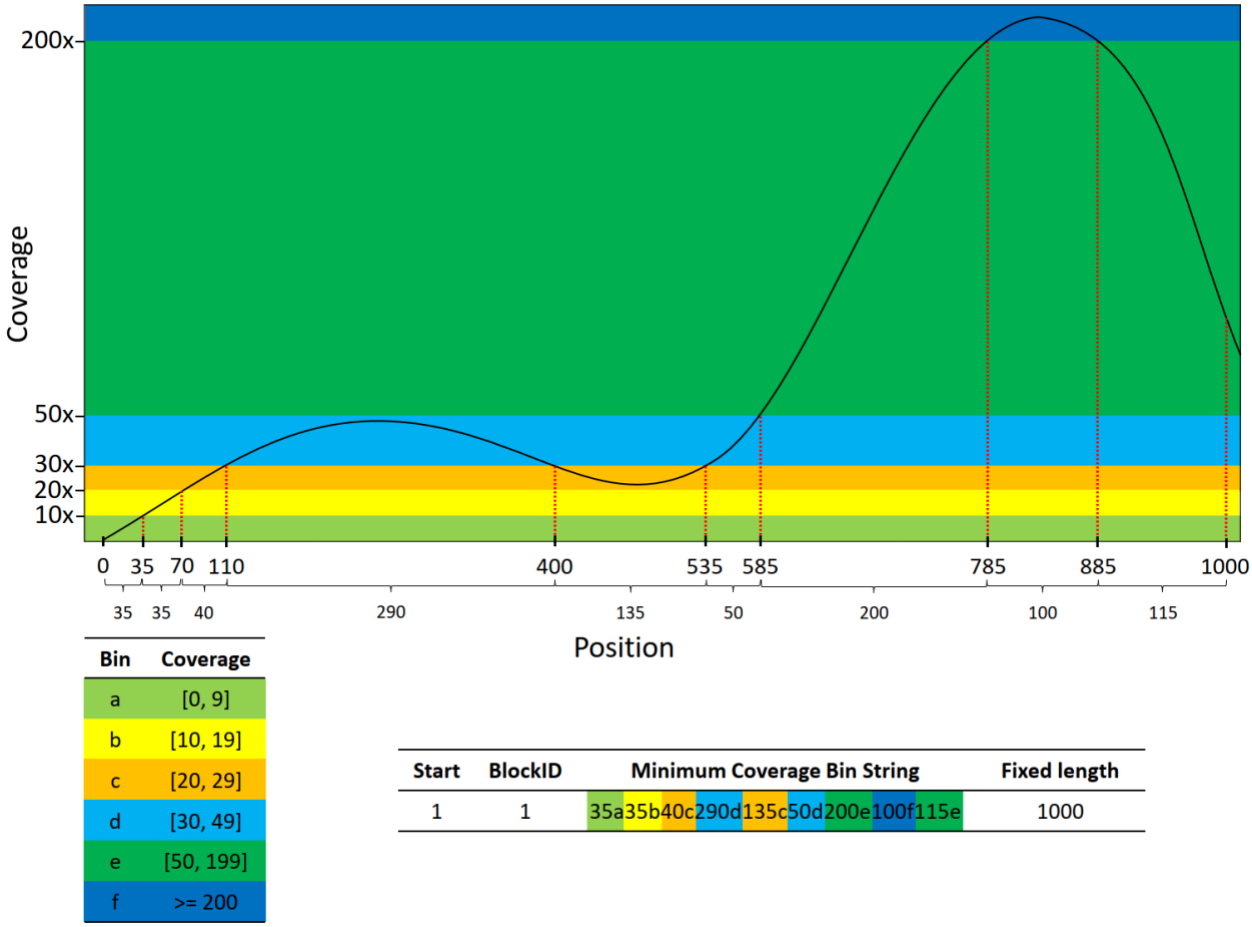


Figure 2: Per site coverage value converted into a fixed 1000 base pair length Bin string

Platform architecture

Users login to the head node to run ATAV jobs which will automatically allocate resources and submit jobs to the cluster (see Figure 3). A standard setup with a 6 node Sun Grid Engine (SGE) cluster (2x10 Cores, 128GB RAM) allows the running of at least 100 jobs at once. Each job will query a slave database with minimum database connections. Using a local customized bioinformatics pipeline, it is possible to continue loading new samples into the master database which will automatically replicate to all slave databases.

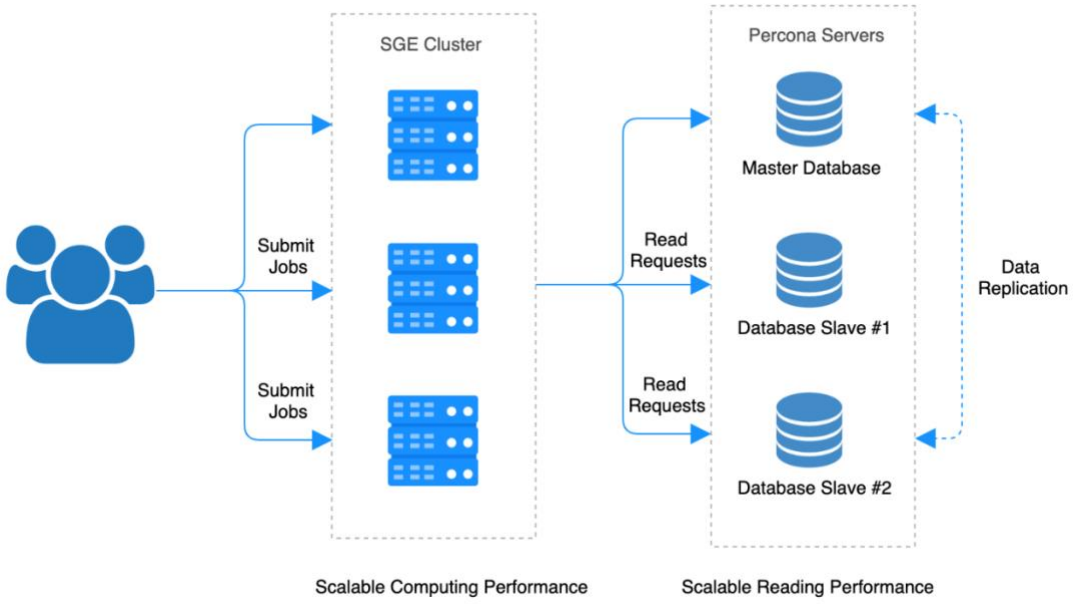


Figure 3: Platform architecture – The user submits jobs to an SGE cluster, the ATAV job will then query to get data from replicated slave database

Application

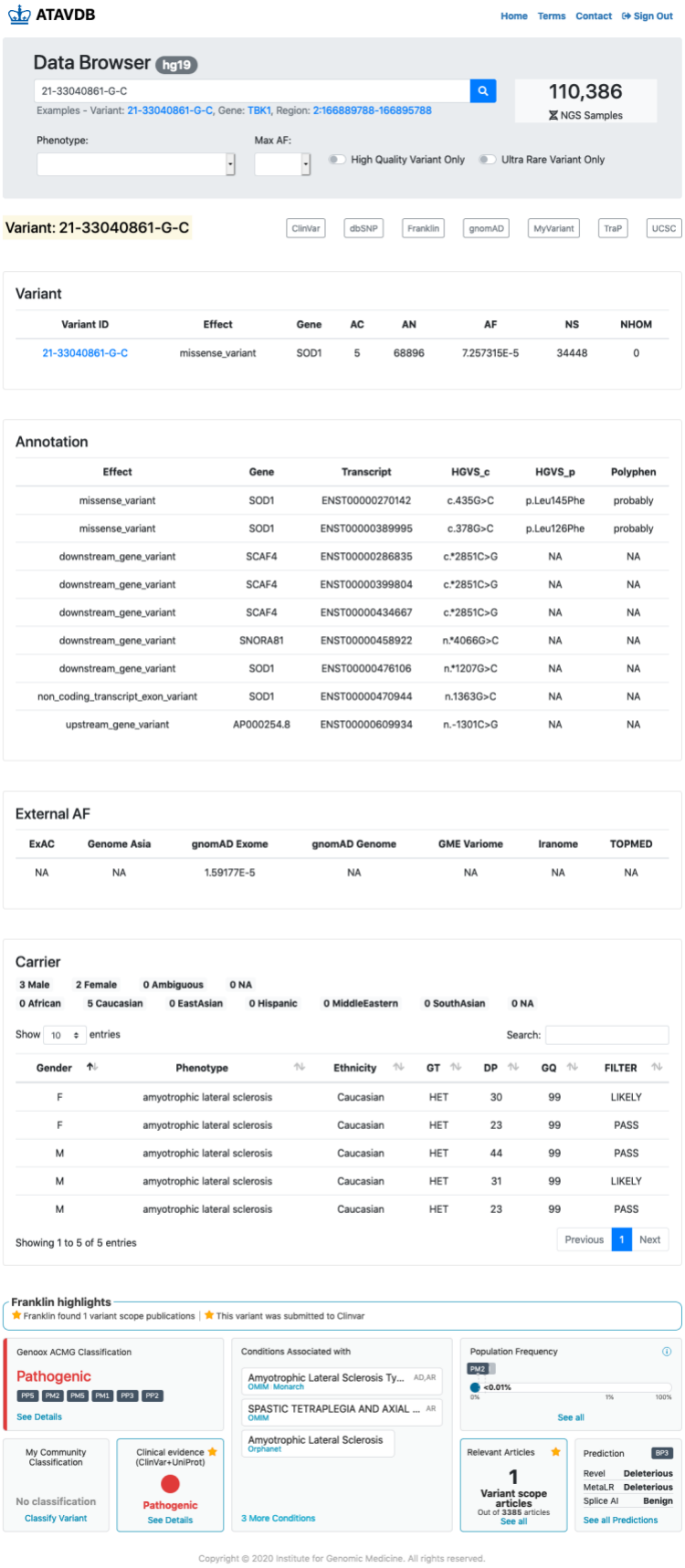
The ATAV command line tool is the interface to ATAVDB. Written in java, ATAV consists of three modules. (i) The command line parser and query engine translate user defined parameters and the input sample list (in PLINK's ped format[18]) into an efficient SQL query for interrogating the relational database, (ii) A runtime variant object creator parses SQL output into a collection of variant objects. Each variant object includes variant information (genomic coordinates, annotation), variant calls in sample list, sample genotype calls at co-ordinates without a called variant and external annotation data. (iii) A statistical analyses module iterates over the variant objection collection to perform downstream analyses. ATAV currently supports tests for diagnostic analyses such as identifying putative *de novo* and inherited genotypes of interest in trios, and a framework for performing region-based rare-variant collapsing analyses that identify genes or other genomic units that carry an excess of qualifying variants among cases in comparison to the background variation observed in internal controls of convenience in ATAVDB.

The modularized ATAV framework makes it extensible to continuously develop new functions that operate on sequencing/variant data sets. Critical to data integrity, all ATAV analyses allow an auditable log of software and database version, filter parameters adopted, the input sample lists used in the specific run and the runtime logs that ensure full reproducibility.

The analysts and researchers of the IGM, have run about 22,000 ATAV job within the last year. 10,000 jobs completed in minutes, 9,000 jobs completed in hours, and the remaining 3,000 jobs completed within two days.

The ATAV data browser is a web user interface that allows everyone within the network to access variant level data directly from the full data set in the ATAVDB. It supports the search of variants by gene, region and variant ID. The gene or region view displays a list of variants with allele count, allele frequency, number of samples, effect, gene etc. The variant view displays a set of annotations (effect, gene, transcript, polyphen) and details about variant carriers (gender, phenotype and quality metrics). It includes links to other public data resources such as Ensembl, gnomAD, ClinVar etc. and directly integrates additional annotations via APIs (e.g. Genoox Franklin API for clinical variant interpretation). The data browser has advanced filters such as a maximum allele frequency threshold to only search rare or ultra-rare variants, restriction to high quality variants or restriction to a certain phenotype. In contrast to many other platforms, the data browser is able to show newly added sample data in real time and is therefore evolving rapidly as more and more samples are sequenced.

Example variant view



Analysis

Rare-Variant collapsing

ATAV provides functions for all recommended steps of the rare-variant collapsing workflow recently summarized in Povysil et al. 2019[19].

For the sample pruning steps ATAV creates the necessary input files by pulling data out of ATAVDB and automatically calls existing standard tools such as KING[20], Eigenstrat[21], or FlashPCA[22]. Since the coverage information for every sample and site is already efficiently stored in ATAVDB, ATAV can efficiently compare coverage between cases and controls and provides two different tests to perform coverage harmonization: sites can be removed if cases and controls show differing proportions of individuals with enough coverage[23]; or if a binomial test shows that the case/control status and coverage are not independent[24]. The outputs of the sample pruning and coverage harmonization steps can be used as inputs for dominant or recessive collapsing models. Within the collapsing model call, ATAV selects qualifying variants (QVs) that pass filters based on variant quality (Phred quality (QUAL), genotype Phred quality (GQ), quality by depth (QD), mapping quality (MQ) and variant quality score log-odds (VQSLOD)), variant annotation (effect, pathogenicity prediction scores, intolerance scores), as well as internal and external minor allele frequencies (MAFs). All QVs are used for building the collapsing matrix, a gene-by-individual indicator matrix with a 0 if there is no qualifying variant found in that gene in that individual, and a 1 if there is at least one. This collapsing matrix is used for looking for associations between genes with QVs and the phenotype of interest by using a Fisher's exact test or Firth-based logistic regression. Finally, quantile-quantile (QQ) plots are created and the genomic inflation factor lambda is estimated using a permutation-based expected distribution of p-values.[23]

A standard collapsing analysis usually consists of several different models that all capture specific types of QVs. While quality control (QC) filters are used for all models, other filters, such as the predicted variant effects or population allele frequencies, depend on the specific model in use. In order to speed up computation, ATAV provides the option of running a general collapsing model first using the QC filters all models have in common and relaxed allele frequency thresholds. The output of this initial model can be used as input for a collapsing-lite function that makes it possible to run the individual collapsing models within minutes since additional filters can just be applied to the previous output and the variant database does not have to be queried again.

Diagnostic analysis

All annotations and filters mentioned in the previous subsection such as QC filters or internal or external MAFs are also important for diagnostic analyses especially for singletons where we

cannot use additional family information. In addition to that, ATAV provides special functions for trios and families to reduce the number of potential disease-causing variants in the final output.

ATAV leverages information about family structure and affectedness status that is provided by the sample file (PLINK-style ped file). Multiple families can be analyzed at once and related controls are for example automatically removed when calculating control frequencies. Furthermore, the affectedness status is used to decide whether to look for inherited or *de novo* variants.

In the standard trio case of one affected offspring and unaffected parents, ATAV uses a series of functions to extract *de novo* variants, newly compound-heterozygous or newly homozygous variants. For distinguishing compound-heterozygosity from variants that are in-phase, ATAV checks that both parents carry one of the qualifying variants. ATAV not only considers the genotype of the individuals, but also their coverage. If the coverage at a variant site is below a minimum threshold of 10 for any of the individuals the variant is still included in the output, but flagged as possibly *de novo*, possibly newly compound-heterozygous or possibly newly homozygous. Furthermore, ATAV identifies putative parental-mosaic variant transmissions. For each parent-child pair, it extracts all variants that were transmitted from parent to child where the variant in the parent has a low proportion of alternate alleles indicating mosaicism.

ATAV also leverages an external annotation dataset called KnownVar, which combines information from multiple variant and disease databases (e.g. ClinVar[14,15], HGMD[17], OMIM, ClinGen[16]). The data is stored in ATAVDB and regularly updated. KnownVar annotations are not only included if the "exact" variant has been reported before, but also if a different variant at the same site has been linked to disease. Typical annotations include the associated disease, ClinVar clinical significance, HGMD Class and Pubmed IDs of relevant papers. In addition to that, disease associated variants in close proximity are extracted from HGMD and ClinVar. On a gene level, annotations include the total number of likely pathogenic or pathogenic variants of each category (copy number variation, small insertion/deletion, splice, nonsense, missense) in ClinVar, disease associations and inheritance from OMIM and dosage sensitivity from ClinGen. All the information provided by KnownVar can be used as additional information in the diagnostic setting to evaluate whether a variant can be considered as diagnostic for a specific patient.

Result

The collapsing framework of ATAV has enabled the confirmation of known and the discovery of novel genes in a wide range of diseases such as epilepsies[25,26], sudden unexplained death in epilepsy[27], congenital kidney malformations[28], chronic kidney disease[29], amyotrophic lateral

sclerosis[30,31], Alzheimer's disease[24], retinal dystrophy[32], and idiopathic pulmonary fibrosis[23]. Furthermore the diagnostic framework has helped to identify both diagnostic genotypes in known genes and candidate genotypes in novel genes in a wide range of diseases including rare undiagnosed genetic disorders[33,34], epilepsies[35–37], alternating hemiplegia of childhood[38], and chronic kidney disease[39]

Conclusions and future directions

We present ATAV as an analysis platform for large-scale whole-exome and whole-genome sequencing projects. The most challenging aspect of the initial use of ATAV is that it needs to be used with ATAVDB and requires establishing a similarly structured database and loading into it the necessary data for retrieval. The advantages of the ATAV framework, however, are that 1) it allows continuous real time analyses of all samples loaded into the database without the need for computationally demanding joint calling preceding each analysis and 2) it allows convenient tracking of precise analyses performed.

Our experience with this platform on a database carrying more than 100,000 samples indicates that a relational database can be optimized in a way that makes it possible to analyze current large-scale genomic datasets. Our current data processing and storage framework is robust and flexible when combining data from multiple projects and mixing exomes and genomes. ATAV supports diagnostic analyses for trios and singletons, as well as rare-variant collapsing analyses for finding disease associations in complex diseases. Further optimizations are possible such as database sharding which is a horizontal partition of data in a database or search engine. Other potential solutions include storing the data in HDFS (Hadoop Distributed File System) and utilizing Apache Spark to do distributed cluster computing. This would allow the processing of large amounts of variant data in parallel at once speeding up computations and enabling an even further increase in sample sizes. The goal of ATAV is to work towards standardizing and optimizing storage and data processing for large scale sequencing data across multiple studies and to provide an easy to use interface for users with little computational experience while ensuring full reproducibility.

Reference:

1. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. BioRxiv 531210.

2. Petrovski, S., Gussow, A.B., Wang, Q., Halvorsen, M., Han, Y., Weir, W.H., Allen, A.S., and Goldstein, D.B. (2015). The Intolerance of Regulatory Sequence to Genetic Variation Predicts Gene Dosage Sensitivity. PLoS Genet. *11*, e1005492.

3. Cingolani, P., Platts, A., Wang, L.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. Fly (Austin). *6*, 80–92.

4. Lek, M., Karczewski, K.J., Minikel, E. V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

5. Dewey, F.E., Murray, M.F., Overton, J.D., Habegger, L., Leader, J.B., Fetterolf, S.N., O'Dushlaine, C., Van Hout, C. V, Staples, J., Gonzaga-Jauregui, C., et al. (2016). Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. Science *354*, aaf6814.

6. Davydov, E. V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. PLoS Comput. Biol. *6*, e1001025.

7. Gelfman, S., Wang, Q., McSweeney, K.M., Ren, Z., La Carpia, F., Halvorsen, M., Schoch, K., Ratzon, F., Heinzen, E.L., Boland, M.J., et al. (2017). Annotating pathogenic non-coding variants in genic regions. Nat. Commun. *8*, 236.

8. Hayeck, T.J., Stong, N., Wolock, C.J., Copeland, B., Kamalakaran, S., Goldstein, D.B., and Allen, A.S. (2019). Improved Pathogenic Variant Localization via a Hierarchical Model of Sub-regional Intolerance. Am. J. Hum. Genet. *104*, 299–309.

9. Traynelis, J., Silk, M., Wang, Q., Berkovic, S.F., Liu, L., Ascher, D.B., Balding, D.J., and Petrovski, S. (2017). Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. Genome Res. *27*, 1715–1729.

10. Gussow, A.B., Petrovski, S., Wang, Q., Allen, A.S., and Goldstein, D.B. (2016). The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. Genome Biol. *17*, 9.

11. Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., et al. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am. J. Hum. Genet. *99*, 877–885.

12. Sundaram, L., Gao, H., Padigepati, S.R., McRae, J.F., Li, Y., Kosmicki, J.A., Fritzilas, N.,

Hakenberg, J., Dutta, A., Shon, J., et al. (2018). Predicting the clinical impact of human mutation with deep neural networks. Nat. Genet. *50*, 1161–1170.

13. Havrilla, J.M., Pedersen, B.S., Layer, R.M., and Quinlan, A.R. (2019). A map of constrained coding regions in the human genome. Nat. Genet. *51*, 88–95.

14. 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*, D980-5.

15. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. *46*, D1062–D1067.

16. Rehm, H.L., Berg, J.S., Brooks, L.D., Bustamante, C.D., Evans, J.P., Landrum, M.J., Ledbetter, D.H., Maglott, D.R., Martin, C.L., Nussbaum, R.L., et al. (2015). ClinGen — The Clinical Genome Resource. N. Engl. J. Med. *372*, 2235–2242.

17. Stenson, P.D., Ball, E. V., Mort, M., Phillips, A.D., Shiel, J.A., Thomas, N.S.T., Abeysinghe, S., Krawczak, M., and Cooper, D.N. (2003). Human Gene Mutation Database (HGMD ®): 2003 update. Hum. Mutat. *21*, 577–581.

18. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., De Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

19. Povysil, G., Petrovski, S., Hostyk, J., Aggarwal, V., Allen, A.S., and Goldstein, D.B. (2019). Rare-variant collapsing analyses for complex traits: guidelines and applications. Nat. Rev. Genet. 1–13.

20. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.

21. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. Principal components analysis corrects for stratification in genome-wide association studies.

22. Abraham, G., Qiu, Y., and Inouye, M. (2017). FlashPCA2: principal component analysis of Biobank-scale genotype datasets. Bioinformatics *33*, 2776–2778.

23. Petrovski, S., Todd, J.L., Durheim, M.T., Wang, Q., Chien, J.W., Kelly, F.L., Frankel, C., Mebane, C.M., Ren, Z., Bridgers, J., et al. (2017). An Exome Sequencing Study to Assess the Role of Rare Genetic Variation in Pulmonary Fibrosis. Am. J. Respir. Crit. Care Med. *196*, 82–93.

24. Raghavan, N.S., Brickman, A.M., Andrews, H., Manly, J.J., Schupf, N., Lantigua, R., Wolock, C.J., Kamalakaran, S., Petrovski, S., Tosto, G., et al. (2018). Whole-exome sequencing in 20,197 persons for rare variants in Alzheimer's disease. Ann. Clin. Transl. Neurol. *5*, 832–842.

25. Allen, A.S., Bellows, S.T., Berkovic, S.F., Bridgers, J., Burgess, R., Cavalleri, G., Chung, S.-K., Cossette, P., Delanty, N., Dlugos, D., et al. (2017). Ultra-rare genetic variation in common epilepsies: a case-control sequencing study. Lancet Neurol. *16*, 135–143.

26. Zhu, X., Padmanabhan, R., Copeland, B., Bridgers, J., Ren, Z., Kamalakaran, S., O'Driscoll-Collins, A., Berkovic, S.F., Scheffer, I.E., Poduri, A., et al. (2017). A case-control collapsing analysis identifies epilepsy genes implicated in trio sequencing studies focused on de novo mutations. PLoS Genet. *13*, e1007104.

27. Bagnall, R.D., Crompton, D.E., Petrovski, S., Lam, L., Cutmore, C., Garry, S.I., Sadleir, L.G., Dibbens, L.M., Cairns, A., Kivity, S., et al. (2016). Exome-based analysis of cardiac arrhythmia, respiratory control, and epilepsy genes in sudden unexpected death in epilepsy. Ann. Neurol. *79*, 522–534.

28. Sanna-Cherchi, S., Khan, K., Westland, R., Krithivasan, P., Fievet, L., Rasouly, H.M., Ionita-Laza, I., Capone, V.P., Fasel, D.A., Kiryluk, K., et al. (2017). Exome-wide Association Study Identifies GREB1L Mutations in Congenital Kidney Malformations. Am. J. Hum. Genet. *101*, 789–802.

29. Cameron-Christie, S., Wolock, C.J., Groopman, E., Petrovski, S., Kamalakaran, S., Povysil, G., Vitsios, D., Zhang, M., Fleckner, J., March, R.E., et al. (2019). Exome-Based Rare-Variant Analyses in CKD. J. Am. Soc. Nephrol. *30*, 1109–1122.

30. Cirulli, E.T., Lasseigne, B.N., Petrovski, S., Sapp, P.C., Dion, P.A., Leblond, C.S., Couthouis, J., Lu, Y.-F., Wang, Q., Krueger, B.J., et al. (2015). Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. Science *347*, 1436–1441.

31. Gelfman, S., Dugger, S., de Araujo Martins Moreno, C., Ren, Z., Wolock, C.J., Shneider, N.A., Phatnani, H., Cirulli, E.T., Lasseigne, B.N., Harris, T., et al. (2019). A new approach for rare variation collapsing on functional protein domains implicates specific genic regions in ALS. Genome Res. *29*, 809–818.

32. Wolock, C.J., Stong, N., Ma, C.J., Nagasaki, T., Lee, W., Tsang, S.H., Kamalakaran, S., Goldstein, D.B., and Allikmets, R. (2019). A case–control collapsing analysis identifies retinal dystrophy genes associated with ophthalmic disease in patients with no pathogenic ABCA4 variants. Genet. Med. *21*, 2336–2344.

33. Zhu, X., Petrovski, S., Xie, P., Ruzzo, E.K., Lu, Y.-F., McSweeney, K.M., Ben-Zeev, B., Nissenkorn, A., Anikster, Y., Oz-Levi, D., et al. (2015). Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. Genet. Med. 2015 1710 *17*, 774.

34. Petrovski, S., Shashi, V., Petrou, S., Schoch, K., McSweeney, K.M., Dhindsa, R.S., Krueger, B., Crimian, R., Case, L.E., Khalid, R., et al. (2015). Exome sequencing results in successful riboflavin treatment of a rapidly progressive neurological condition. Mol. Case Stud. *1*, a000257.

35. Consortium, E., Allen, A.S., Berkovic, S.F., Cossette, P., Delanty, N., Dlugos, D., Eichler, E.E., Epstein, M.P., Glauser, T., Goldstein, D.B., et al. (2013). De novo mutations in epileptic encephalopathies. Nature *501*, 217–221.

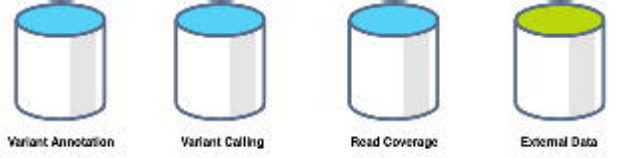
36. Myers, C.T., Stong, N., Mountier, E.I., Helbig, K.L., Freytag, S., Sullivan, J.E., Ben Zeev, B., Nissenkorn, A., Tzadok, M., Heimer, G., et al. (2017). De Novo Mutations in PPP3CA Cause Severe Neurodevelopmental Disease with Seizures. Am. J. Hum. Genet. *101*, 516–524.

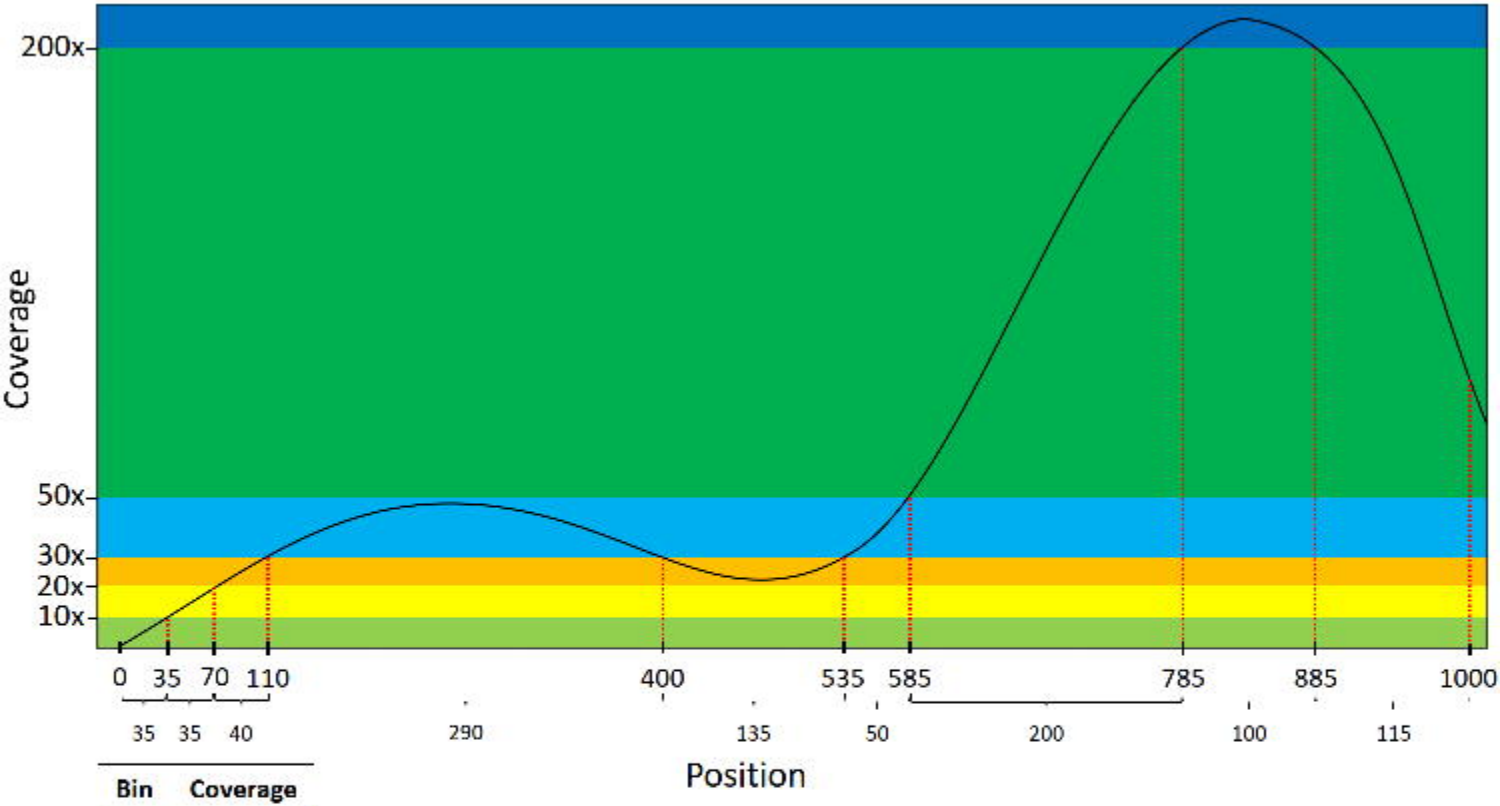
37. Petrovski, S., Küry, S., Myers, C.T., Anyane-Yeboa, K., Cogné, B., Bialer, M., Xia, F., Hemati,

P., Riviello, J., Mehaffey, M., et al. (2016). Germline de Novo Mutations in GNB1 Cause Severe Neurodevelopmental Disability, Hypotonia, and Seizures. Am. J. Hum. Genet. *98*, 1001–1010.

38. Heinzen, E.L., Swoboda, K.J., Hitomi, Y., Gurrieri, F., De Vries, B., Tiziano, F.D., Fontaine, B., Walley, N.M., Heavin, S., Panagiotakaki, E., et al. (2012). De novo mutations in ATP1A3 cause alternating hemiplegia of childhood. Nat. Genet. *44*, 1030–1034.

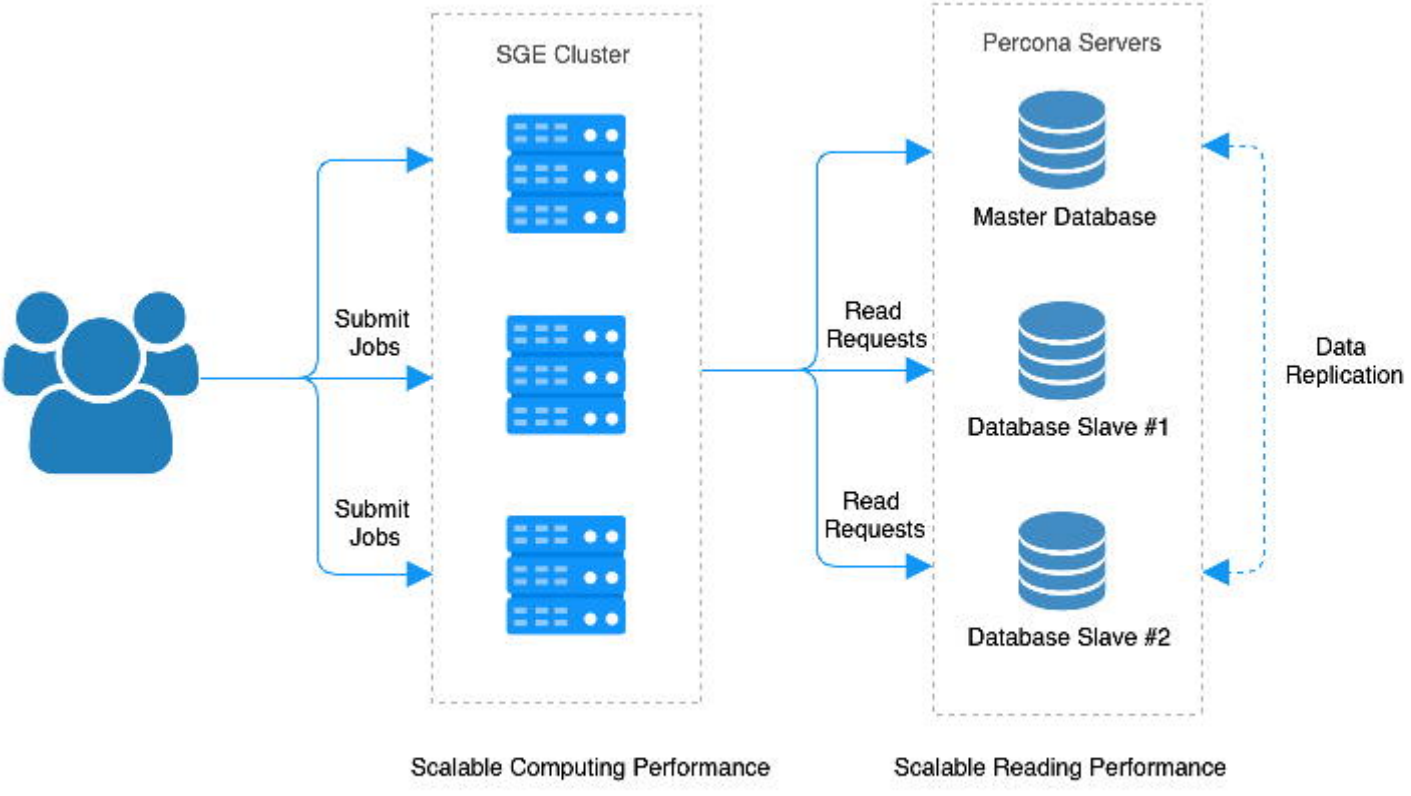
39. Groopman, E.E., Marasa, M., Cameron-Christie, S., Petrovski, S., Aggarwal, V.S., Milo-Rasouly, H., Li, Y., Zhang, J., Nestor, J., Krithivasan, P., et al. (2019). Diagnostic Utility of Exome Sequencing for Kidney Disease. N. Engl. J. Med. *380*, 142–151.


Variant Annotation
variant_id
Chromosome
Position
Reference Allele
Alternative Allele
Effect
Gene
Transcript
...
Variant Calling
block_id
sample_id
variant_id
GT
DP
GQ
VQSLOD
SOR
...
Read Coverage
block_id
sample_id
External Data
gnomAD
Genp
TraP
LINER
ClinVar
HGMD
RMS
REVEL
...



Bin	Coverage
a	[0, 9]
b	[10, 19]
c	[20, 29]
d	[30, 49]
e	[50, 199]
f	>= 200

Start	BlockID	Minimum Coverage Bin String	Fixed length
1	1	35a35b40c290d135c50d200e100f115e	1000


SGE Cluster
Percona Servers
Master Database
Submit Jobs
Submit Jobs
Read Requests
Read Requests
Data Replication
Database Slave #1
Database Slave #2
Scalable Computing Performance
Scalable Reading Performance