Published in final edited form as:

Hum Mutat. 2020 February ; 41(2): 387–396. doi:10.1002/humu.23942.

The Clinical Genome and Ancestry Report (CGAR): An Interactive Web Application for Prioritizing Clinically-implicated Variants from Genome Sequencing Data with Ancestry Composition.

In-Hee Lee¹, Jose A. Negron¹, Carles Hernandez-Ferrer¹, William Jefferson Alvarez¹, Kenneth D. Mandl^{1,2,3}, Sek Won Kong^{1,2,*}

¹Computational Health Informatics Program, Boston Children's Hospital, Boston, MA 02115

²Department of Pediatrics, Harvard Medical School, Boston, MA 02115

³Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115

Abstract

Genome sequencing is positioned as a routine clinical work-up for diverse clinical conditions. A commonly used approach to highlight candidate variants with potential clinical implication is to search over locus- and gene-centric knowledge databases. Most web-based applications allow a federated query across diverse databases for a single variant; however, sifting through a large number of genomic variants with combination of filtering criteria is a substantial challenge. Here we describe the Clinical Genome and Ancestry Report (CGAR), an interactive web application developed to follow clinical interpretation workflows by organizing variants into seven categories: (1) reported disease-associated variants, (2) rare and high-impact variants in putative disease-associated genes, (3) secondary findings which the American College of Medical Genetics and Genomics recommends reporting back to patients, (4) actionable pharmacogenomic variants, (5) focused reports for candidate genes, (6) de novo variant candidates for trio analysis, and (7) germline and somatic variants implicated in cancer risk, diagnosis, treatment and prognosis. For each variant, a comprehensive list of external links to variant-centric and phenotype databases are provided. Furthermore, genotype-derived ancestral composition is used to highlight allele frequencies from a matched population since some disease-associated variants show a wide variation between populations. CGAR is an open-source software and is available at <https://tom.tch.harvard.edu/apps/cgar/>.

*To whom correspondence should be addressed at: Sek Won Kong, MD, 300 Longwood Avenue, Boston Children's Hospital, Boston, MA 02115, T: 617-919-2689, F: 617-730-0817, sekwon.kong@childrens.harvard.edu.

Conflict of Interest

The authors declare that they have no competing interests.

Data availability Statement

The web application developed in this study is available in <https://tom.tch.harvard.edu/apps/cgar/>. And the source code is available from https://bitbucket.org/gnome_pipeline/cgar_pub.

Keywords

Whole genome sequencing; variant annotation; clinical interpretation; ancestry; pharmacogenomics; cancer

INTRODUCTION

The clinical utility of whole-genome and -exome sequencing (WGS and WES) has been supported by multiple independent studies across diverse human diseases (Farnaes et al., 2018; Splinter et al., 2018; Tan et al., 2017). Clinical genome sequencing is now routinely used to discover disease-associated variants in rare diseases (Bick, Jones, Taylor, Taft, & Belmont, 2019; Boycott et al., 2019; Gahl, Wise, & Ashley, 2015; Posey, 2019; Posey et al., 2017; Prokop et al., 2018; Wright et al., 2015). A hundred variants need be reviewed for each individual as an individual typically carries about one hundred loss-of-function variants that are not necessarily associated with disease phenotypes (MacArthur et al., 2012). Moreover, other features such as allele frequencies across populations and published evidences need to be considered for identifying potentially pathogenic variants worth clinical reporting (Richards et al., 2015). Thus, identifying candidate variants associated with particular phenotypes remains a challenge.

To this end, diverse locus-, gene-, and phenotype-centric databases provide information on phenotype-associated variants and genes (Amberger, Bocchini, Scott, & Hamosh, 2019; Landrum et al., 2016; Stenson et al., 2017). The Model organism Aggregated Resources for Rare Variant ExpLoration (MARRVEL) and VarSome provide comprehensive information for each variant by integrating multiple resources (Kopanios et al., 2018; Wang et al., 2017). The work required to use these databases is substantial, including narrowing down the list of variants by searching multiple databases, verifying detailed information in variant-centric databases, and incorporating other criteria such as minor allele frequency in an ancestry-matched population. It thus remains difficult to identify putative disease-associated variants among the variants found in an individual (Hiatt et al., 2018; Taylor et al., 2015).

A single page summary report of WES/WGS has been considered as a practical choice for conveying genomic information to ordering clinicians (Vassy et al., 2015). Clinical genome reports are usually delivered to the point-of-care in a format resembling other clinical laboratory test reports that are supported by electronic health records (EHRs) (Shirts et al., 2015). However, clinical laboratories and clinicians could have different opinions regarding candidate variants listed in such static summary (Bland et al., 2018). Moreover, this type of genome report does not take full advantage of genome sequencing over targeted gene panel testings. Initial positive hit rates for medical conditions still vary widely (3 – 79%) (Schwarze, Buchanan, Taylor, & Wordsworth, 2018). Re-analysis of sequence and re-interpretation is often required to identify phenotype-associated variants (Cowley et al., 2019; Eldomery et al., 2017). There is a clear need for a method to report genome sequencing results in an interactive, intuitive, and user-friendly manner that allows continuous exploration of all variants with comprehensive and up-to-date annotation.

Here we describe the Clinical Genome and Ancestry Report (CGAR) that filters and highlights clinically implicated variants using an intuitive graphical user interface. This approach contrasts with the traditional static text format for clinical genome reports. CGAR provides a unified view of information gathered from diverse databases as well as links to external variant-centric databases, and highlights allele frequencies in matching populations according to genotype-derived ancestry. Users can browse variants and iteratively refine the initial selection of variants. Figure 1A summarizes the workflow in CGAR.

METHODS

Variant annotation using Variant Effect Predictor

CGAR accepts variant data in variant call format (VCF) (Danecek et al., 2011) and uses Variant Effect Predictor (VEP; release 94) (McLaren et al., 2016) and its plugins for variant annotation. We used the option ‘--everything’ to collect: (1) the predicted functional consequences according to gene models of RefSeq (Pruitt et al., 2014), Ensembl (Zerbino et al., 2018), and Consensus Coding Sequence (CCDS) (Pruitt et al., 2009); (2) the accession identifiers for protein products from the UniProt databases (UniProt, 2019); (3) overlapping protein domain information if available; (4) pathogenicity predictions from Sorting Intolerant From Tolerant (SIFT) (Sim et al., 2012) or Polymorphism Phenotyping-2 (PolyPhen-2) (Adzhubei et al., 2010); (5) variant allele frequencies from various population sequencing projects such as the 1000 Genomes Project (1000 Genomes Project et al., 2015) and Genome Aggregation Database (gnomAD) (Lek et al., 2016); (6) PubMed unique identifiers (PMIDs) for published papers describing variants. Also, we used the following VEP plugins (https://github.com/Ensembl/VEP_plugins) to enrich annotations:

‘Downstream (predicts detailed downstream effects of a frameshift variant on protein sequence),’ ‘TSSDistance (calculates the distance from the transcription start site for upstream variants),’ ‘SpliceRegion (provides more granular predictions of splicing effects),’ ‘dbNSFP (presents various pathogenicity prediction scores for missense variants from data files by (Liu, Wu, Li, & Boerwinkle, 2016)),’ ‘dbSNV (presents scores for effects by splicing variants from (Jian, Boerwinkle, & Liu, 2014)),’ ‘CADD (for predicted variant pathogenicity scores by (Rentzsch, Witten, Cooper, Shendure, & Kircher, 2019)),’ ‘Condel (for predicted variant pathogenicity scores by (Gonzalez-Perez & Lopez-Bigas, 2011)),’ ‘LoFTool (provides genic score for intolerance to functional variation as in (Fadista, Oskolkov, Hansson, & Groop, 2017)),’ and ‘ExACpLI (provides the probability of a gene intolerant to loss-of-function variants as shown in (Lek et al., 2016)).’

In addition to VEP and its plugins, we used custom annotations for (1) sequence conservation scores from phastCons (Siepel et al., 2005), phyloP (Pollard, Hubisz, Rosenbloom, & Siepel, 2010), and GERP++ (Davydov et al., 2010); (2) low sequence complexity regions marked by RepeatMasker (www.repeatmasker.org); (3) locus-specific coverage metrics from gnomAD exomes. For (1) and (2), we downloaded tables ‘phastCons100way’, ‘phyloP100wayAll’, ‘GERP++’, and ‘RepeatMasker’ from UCSC Genome Browser (latest available data in the site: Dec. 12, 2014) and converted them to custom annotation tracks for VEP. For (3), we used the coverage metric files available from gnomAD browser, which provided per-locus percentage of exome samples that reached

target depth of coverage – e.g., 10x, 20x or 30x. Low coverage metric values (i.e., < 80%) suggest sub-optimal depth of coverage with WES for the locus, which could imply either an inaccurate allele frequency estimation or an unreliable variant call. Since > 20x depth of coverage is thought to be sufficient for attaining 99% sensitivity for calling heterozygous variants (Meynert, Ansari, FitzPatrick, & Taylor, 2014), CGAR displays coverage metric values for 20x, i.e., the proportion of exomes reached 20x or more at the loci.

Genotype-based ancestral proportion analysis

Variant allele frequencies need to be correctly compared with those from population of matching ancestry, since allele frequencies vary across populations. We implemented an ancestral proportion analysis method for WGS, EIGMIX (Zheng & Weir, 2016), an accurate and efficient algorithm that can handle millions of variants and large number of query individuals. EIGMIX derives principal components from surrogate populations with reported ancestry and projects an individual of interest to the principal components to determine its ancestry. Assuming center coordinates of each surrogate population as unit vectors in ancestral proportion space (e.g., (1,0,0), (0,1,0), (0,0,1) for three ancestral populations), EIGMIX builds linear transformation from the principal component space to ancestral proportions, which is used to calculate the proportions of ancestral populations for an individual.

As surrogate populations for EIGMIX, we used the five continental level populations in the 1000 Genomes Project Phase 3 dataset – i.e., African, American, East Asian, European, and South Asian. For each surrogate ancestral group, we selected populations showing relatively lower admixed ancestral structure than other population within the same continental group (1000 Genomes Project et al., 2015). The Peruvian individuals from Lima, Peru (PEL; N=85) were used as surrogates for American, Yoruba in Ibadan, Nigeria (YRI; N=108) for African, Han Chinese in Beijing, China (CHB; N=103) for East Asian, Utah Residents with Northern and Western European Ancestry (CEU; N=99) for European, and Indian Telugu from the UK (ITU; N=102) for South Asian. We selected 85 individuals from each surrogate group to balance between groups. Next, we selected common bi-allelic SNPs (minor allele frequency 5% or higher in 1000 Genomes Project) and pruned to select 1,058,271 SNPs that were at least 2,000 bases apart from each other. In each analysis, only the variants found in both the input WGS VCF file and the pre-selected 1,058,271 SNPs are used for EIGMIX analysis. The derived ancestral proportions are stored in database along with variants for each individual.

Identification of clinically implicated variants for interactive filtering

After VCF files are processed by CGAR, users can begin their analysis by selecting an individual sample name. CGAR lists variants that satisfy one of the following criteria: (1) the variants that were previously reported as disease-associated from the Human Gene Mutation Database (HGMD) (Stenson et al., 2017) or pathogenic variants from ClinVar (Landrum et al., 2016); (2) variants located in rare disease-associated genes as identified in the OrphaNet database (Pavan et al., 2017); (3) variants located in the 59 medically actionable genes recommended for reporting incidental findings in clinical sequencing according to the American College of Medical Genetics and Genomics (ACMG)

recommendations (ACMG SF v2.0) (Kalia et al., 2017); (4) pharmacogenomic (PGx) variants; (5) *de novo* germline variant candidates found in the proband but not in parents; (6) variants that match with previously reported somatic variants in the Catalogue of Somatic Mutations In Cancer (COSMIC) (Forbes et al., 2017) or which are located in cancer-related genes listed in the Cancer Gene Census (Futreal et al., 2004); and (7) variants related to user-defined candidate genes or phenotypes.

PGx variants and genes were compiled from the Pharmacogenomics Knowledge Base (PharmGKB) (Sanguhl, Berlin, Altman, & Klein, 2008), Clinical Pharmacogenetics Implementation Consortium (CPIC) (Relling et al., 2013) and Pharmacogenomics Research Network (PGRN) project (Gordon et al., 2016). For tumor-normal paired samples, tumor-specific somatic variants are reported in criterion 6 (above). Overall, commonly-used criteria in clinical sequencing laboratories were applied to list clinically implicated variants satisfying any of criteria 1 through 6. Criterion 7 is met when users incorporate domain-specific knowledge to find variants not captured before. To identify genes associated with user-defined phenotypes, we used FindZebra – a database of genotype and phenotype associations based on machine-learning of literature (Dragusin et al., 2013). FindZebra uses a text search engine to retrieve and rank genes according to gene-specific scores for phenotype based on various sources of curated documents such as Online Mendelian Inheritance in Men (OMIM) (Amberger et al., 2019) or Genetic and Rare Diseases Information Center (GARD, <https://rarediseases.info.nih.gov/>). From the ranked list of genes, CGAR selects the genes with top scores within a user-specified threshold. FindZebra allows any terms describing disease, phenotype or symptoms without requiring conversion to standardized ontology terms.

The matching variants are further filtered by criteria such as allele frequencies, predicted functional consequences, and additional context-specific criteria. Results are presented in a graphical interface where users can interact to refine their search by adjusting filtering parameters. Figure 1B details the process of the variant filtering in CGAR.

Implementation of web application

CGAR was written in Python using the Flask 1.0.2 micro-framework (<http://flask.pocoo.org>) and Bootstrap 3.3.7 (<http://getbootstrap.com>) for web interface development and could be hosted using an Apache 2.4 (<http://www.apache.org>) web server. A public server (<https://tom.tch.harvard.edu/apps/cgar/>) is pre-loaded with publicly available WGS/WES VCF files and all features are fully functional without registration. An individual account can be set up upon request for users from non-profit academic institutions who does not need local set up. CGAR protects its variant data with automatic Advanced Encryption Standard (AES) encryption at-rest feature in MySQL as well as by securing its communication with client browser with Hypertext Transfer Protocol Secure (HTTPS) protocol. Nonetheless, for additional security, setting up a local CGAR server behind a firewall is recommended. The source code is available at https://bitbucket.org/gnome_pipeline/cgar_pub and Docker container for local deployment is also available upon request.

RESULTS

Interactive user interface in web application

Figure 2 shows an example screenshot from CGAR. In the top left, a pie chart summarizes predicted ancestral proportions for the selected sample. For trio analysis, results from parental genomes are also shown. On the right side, the ancestral proportions for 278 individuals from the Simons Genome Diversity Project (Mallick et al., 2016) are shown. These individuals were recruited from 127 populations from disparate locations around the world, and pie charts were plotted on the matching geographies. This map of ancestral proportions for 278 individuals is intended to provide a guideline for the ancestral origin of the sample. The ancestral group of largest proportion could be directly used to select allele frequencies of matching population from gnomAD.

In each tab, the main table displays essential variant information such as the Human Genome Organization (HUGO) gene symbol, gene-level score of being loss-of-function intolerant (pLI score) for the gene, variant representation according to Human Genome Variation Society (HGVS) nomenclature, zygosity, calculated variant consequences, the highest allele frequency, and a coverage metric (Figure 2A). Additionally, each tab contains several columns to provide information specific to it. For instance, the columns such as 'ClinVar phenotype', 'Extra phenotype links', and 'Review status' are displayed for variants in the 'ClinVar' tab (Figure 2A), while the columns such as 'Pharmacogenomic information' are displayed for variants in the 'Pharmacogenomics' tab (Figure 2B). Variants selected using HGMD and OrphaNet databases are available only to users with valid license agreements for them.

Further details on each variant can be found by clicking the plus button on the left of each gene symbol. Expanded child row (Figure 2B, bottom) contains the links to a pop-up window for detailed information on the variant and links to external server including gnomAD, MARRVEL (Wang et al., 2017), VarSome (Kopanos et al., 2018), Beacon Network (Global Alliance for Genomics and Health, 2016), WEScover, a web application showing gnomAD WES coverage data organized by genes and populations (Alvarez et al., 2018), and VarSite (Stephenson, Laskowski, Nightingale, Hurles, & Thornton, 2019). The link to detailed view opens a new window with further details for each variant including: (1) calculated variant consequence for each transcript with predicted pathogenicity scores from multiple algorithms; (2) allele frequencies from the 1000 Genomes Project, gnomAD, and the Exome Sequencing Project summarized by ancestral populations; (3) scores for predicted effect on splice sites; (4) locus-specific conservation scores; (5) protein families or domains overlapping with the locus; and (6) a list of PubMed entries on the variants extracted from VEP annotation.

The side menu in the left lists controls to customize and configure variant filtering criteria for the current tab (Figure 2A). The set of available controls changes contextually according to the contents for each tab while filtering criteria for zygosity, allele frequency and calculated consequences are available for all tabs. A default set of filtering criteria includes rare heterozygous and homozygous variants (allele frequency < 0.5% in any population) with high impacts – i.e., transcript_ablation (SO: 0001893), splice_acceptor_variant (SO:

0001574), splice_donor_variant (SO: 0001575), stop_gained (SO:0001587), frameshift_variant (SO:0001589), stop_lost (SO: 0001578), start_lost (SO: 0002012), and transcript_amplification (SO: 0001889) – following the classification of VEP's prediction of functional consequences. Users can change these settings to include more or less variants accordingly. The current applied filtering criteria can be found in the top of main table (grey boxes under the heading 'Your query' in Figure 2).

Comparison with other applications

We compared the features available in CGAR with three recently published tools for variant interpretation – MARRVEL (Wang et al., 2017), VarSome (Kopanov et al., 2018), and VarCards (J. Li et al., 2018) (Table 1). MARRVEL was developed for interpreting filtered candidate variants from patients with rare diseases and can take gene symbol, DNA or protein variant as input. Unique features of MARRVEL include: (1) ortholog search across model organisms including alignment of protein domains in ortholog proteins (2) integration with the Geno2MP server (<https://geno2mp.gs.washington.edu/>) that links variants to phenotypic information from multiple Mendelian gene discovery projects and (3) listing reported alleles when a gene symbol is used as input. With rich phenotype information, MARRVEL is an essential tool for interpreting the potential impacts of variants from patients with rare diseases once a candidate variant is selected. VarSome enables community users to add comments on variant interpretation similar to ClinVar. Some features of VarSome are freely available to public; however, a paid subscription is required to use VCF as input. VarCards provides a variety of genetic and clinically relevant information for a single and multiple coding variant; however, a very wide tabular output makes it difficult to navigate without detailed knowledge on annotation tracks.

CGAR provides a comprehensive overview of multiple variants from WES and WGS, organized based on their clinical implications, while the others aim to provide comprehensive annotation on one-variant-at-a-time mode. Among the compared tools, CGAR provides the most comprehensive use of resources providing disease/phenotype-associated variants and genes – i.e., HGMD, ClinVar, OMIM, GWAS catalog (Buniello et al., 2019), and OrphaNet. Also, for cancer-related information, CGAR provides both gene-level (CENSUS) and variant-level (COSMIC) information, while the others use variant-level information from COSMIC and/or the International Cancer Genome Consortium (ICGC, <https://icgc.org>). Finally, CGAR provides more unique features (listed under 'other features' in Table 1). For instance, the secondary findings by ACMG recommendations and the pharmacogenomic information such as CPIC guidelines are additionally provided by CGAR.

Variant analysis using CGAR

Next, we describe two use cases to highlight various aspects in CGAR: (1) finding phenotype-associated variants in rare genetic disorders and (2) finding de novo variants in trio. The samples we used in these use cases are also available in the public server. The user guide can provide a more detailed step-by-step explanation for use cases.

As the first example, we used a semisynthetic WGS data (Sifrim et al., 2013) created, from a publicly available VCF file for a healthy individual, by injecting published disease-causing

variants for Miller syndrome (available for download at <http://extasy.esat.kuleuven.be>). Miller syndrome [MIM# 263750] is a very rare autosomal recessive disorder characterized by abnormal face and limb development. Previous WES and WGS studies found compound heterozygous mutations in the *DHODH* gene [MIM# 126064] in affected individuals (Ng et al., 2010; Roach et al., 2010). First, we checked the ClinVar tab with default settings (i.e., all pathogenic or likely pathogenic variants regardless of zygosity, allele frequency, consequences or review status in ClinVar) and found six variants. Given that the incidence of Miller syndrome is one in a million newborns (<https://ghr.nlm.nih.gov/condition/miller-syndrome>), we restricted allele frequency filtering criteria to less than 0.5% to find very rare variants, resulting in two missense variants in the *DHODH* gene (ENST00000219240.4:c.454G>A and ENST00000219240.4:c.605G>C) as well as one pathogenic variant associated with proline dehydrogenase deficiency [MIM# 239500] (ENST00000357068.6:c.1292G>A). Independently, we also used the ‘Phenotype associated’ tab with query keyword “Miller”. To limit our search for the most relevant genes according to literature, we set CGAR to return only the top 1% of the genes. Again, we found the same two missense variants in the *DHODH* gene among high or moderate impact variants. This example illustrates how CGAR could quickly identify known disease-associated variants or rare variants in the genes associated with phenotype.

The second use case illustrates how to filter for *de novo* variant candidates and to reduce false positive finding using coverage metrics in CGAR. Identification of *de novo* variant candidates in proband requires use of a specialized analysis pipeline such as PolyMutt (B. Li et al., 2012), VarScan (Koboldt et al., 2012), or joint variant calling using pedigree information to reduce false discoveries and false negatives. However, when only VCF files are available for a trio, rare or novel heterozygous variants reported only in proband could be selected for further evaluation. Additionally, coverage metrics calculated using a large number of WES datasets in the gnomAD server can be used to reduce false discoveries in loci with suboptimal coverage in large sequencing efforts. We used a publicly available WGS dataset from a family (CEPH/Utah Pedigree 1463, Coriell Institute, Camden, NJ) to demonstrate. Three family members – daughter (NA12878, as an index case), father (NA12891) and mother (NA12892) – were chosen for trio analysis because the list of validated *de novo* variants were available from literature.(Conrad et al., 2011) The VCF files for this family members were downloaded from the Genome in a Bottle Consortium’s FTP site (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/technical/platinum_genomes/). Since the validated variants were reported in the human reference genome assembly NCBI36/hg18, we used *liftOver* tool to transpose them to the GRCh37/hg19 coordinates (Kuhn, Haussler, & Kent, 2013), resulting in 49 germline and 952 somatic variants including 7 *de novo* variants located in the protein coding sequence genes (1 germline and 6 non-germline variants).

In trio-based analysis, the variants in the index case were compared to the ones from both parental samples and only the variants unique to the index case were selected. By default, only rare heterozygous variants on autosomes with high impact are reported by CGAR. However, to compare with the set of all validated variants, we removed the filtering on variant consequences. CGAR found 142 *de novo* variant candidates, including all validated variants in protein coding genes. A single validated variant (non-germline) not reported by

CGAR was not reported in the three VCF files. The minimum coverage metrics for the validated variants that were discovered by CGAR was 77.47%. As such, filtering the other *de novo* candidate variants in CGAR with lower coverage metrics than the validated variants resulted in 34 out of 142 candidates, suggesting false positive candidates due to insufficient coverage in parents. This illustrated the importance of input variant call quality as well as the utility of the coverage metric from a population scale WES and WGS database, when prioritizing *de novo* variants. None of the candidates were found in *denovo-db* (Turner et al., 2017). Nonetheless pedigree-aware joint variant calling is imperative to reduce the amount of false discovery for family-based analysis.

CONCLUSION

CGAR was designed to meet the needs of the academic community as a simple and interactive tool for the analysis of WGS and WES results in clinical research settings. We aimed to provide a streamlined and intuitive solution to variant annotation and filtering for healthcare professionals without extensive experience in genomic data analysis tools. To make the results self-explanatory, we simplified the main screen in each tab to provide only essential information and prepared collapsible child-rows and pop-up windows for the links to the other resources and detailed variant information, respectively. To facilitate a WGS/WES interpretation work flow, we put a set of default filtering criteria for each category that can provide good starting candidates, which could be interactively changed by users. A public server is available for healthcare professionals to try out with or without creating account; however, we strongly recommend deploying CGAR within local secure network environments using a Docker container. CGAR will support the human reference genome GRCh38 as reprocessed 1000 Genome Project datasets and additional annotation tracks become available, instead of lifted-over information from GRCh37. Additionally, we will adopt Integrative Genomics Viewer (Robinson, Thorvaldsdottir, Wenger, Zehir, & Mesirov, 2017) to enable visual inspection of short-reads alignment at variant loci and flanking regions if analysis ready BAM files are provided with URLs.

Acknowledgements

S.W.K. was supported in part by grants from the National Institutes of Health (R01MH107205, U01HG007530, R24OD024622, and U01TR002623) and by the Boston Children's Hospital Precision Link initiative.

References

- 1000 Genomes Project, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, ... Abecasis GR (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68–74. doi:10.1038/nature15393 [PubMed: 26432245]
- Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, ... Sunyaev SR (2010). A method and server for predicting damaging missense mutations. *Nat Methods*, 7(4), 248–249. doi:10.1038/nmeth0410-248 [PubMed: 20354512]
- Alvarez WJ, Lee I-H, Hernandez-Ferrer C, La Rosa JN, Mandl KD, & Kong S (2018). WEScover: whole exome sequencing vs. gene panel testing. *bioRxiv*. doi:10.1101/367607
- Amberger JS, Bocchini CA, Scott AF, & Hamosh A (2019). [OMIM.org](https://omim.org): leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res*, 47(D1), D1038–D1043. doi:10.1093/nar/gky1151 [PubMed: 30445645]

- Bick D, Jones M, Taylor SL, Taft RJ, & Belmont J (2019). Case for genome sequencing in infants and children with rare, undiagnosed or genetic diseases. *J Med Genet*. doi:10.1136/jmedgenet-2019-106111
- Bland A, Harrington EA, Dunn K, Pariani M, Platt JCK, Grove ME, & Caleshu C (2018). Clinically impactful differences in variant interpretation between clinicians and testing laboratories: a single-center experience. *Genet Med*, 20(3), 369–373. doi:10.1038/gim.2017.212 [PubMed: 29240077]
- Boycott KM, Hartley T, Biesecker LG, Gibbs RA, Innes AM, Riess O, ... Baynam G (2019). A Diagnosis for All Rare Genetic Diseases: The Horizon and the Next Frontiers. *Cell*, 177(1), 32–37. doi:10.1016/j.cell.2019.02.040 [PubMed: 30901545]
- Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, ... Parkinson H (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*, 47(D1), D1005–D1012. doi:10.1093/nar/gky1120 [PubMed: 30445434]
- Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, Casals F, ... Genomes P (2011). Variation in genome-wide mutation rates within and between human families. *Nat Genet*, 43(7), 712–714. doi:10.1038/ng.862 [PubMed: 21666693]
- Cowley MJ, Liu YC, Oliver KL, Carvill G, Myers CT, Gayevskiy V, ... Roscioli T (2019). Reanalysis and optimisation of bioinformatic pipelines is critical for mutation detection. *Hum Mutat*, 40(4), 374–379. doi:10.1002/humu.23699 [PubMed: 30556619]
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, ... Genomes Project Analysis, G. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156–2158. doi:10.1093/bioinformatics/btr330 [PubMed: 21653522]
- Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, & Batzoglou S (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol*, 6(12), e1001025. doi:10.1371/journal.pcbi.1001025 [PubMed: 21152010]
- Dragusin R, Petcu P, Lioma C, Larsen B, Jorgensen HL, Cox IJ, ... Winther O (2013). FindZebra: a search engine for rare diseases. *Int J Med Inform*, 82(6), 528–538. doi:10.1016/j.ijmedinf.2013.01.005 [PubMed: 23462700]
- Eldomery MK, Coban-Akdemir Z, Harel T, Rosenfeld JA, Gambin T, Stray-Pedersen A, ... Lupski JR (2017). Lessons learned from additional research analyses of unsolved clinical exome cases. *Genome Med*, 9(1), 26. doi:10.1186/s13073-017-0412-6 [PubMed: 28327206]
- Fadista J, Oskolkov N, Hansson O, & Groop L (2017). LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals. *Bioinformatics*, 33(4), 471–474. doi:10.1093/bioinformatics/btv602 [PubMed: 27563026]
- Farnaes L, Hildreth A, Sweeney NM, Clark MM, Chowdhury S, Nahas S, ... Kingsmore SF (2018). Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization. *NPJ Genom Med*, 3, 10. doi:10.1038/s41525-018-0049-4 [PubMed: 29644095]
- Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, ... Campbell PJ (2017). COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res*, 45(D1), D777–D783. doi:10.1093/nar/gkw1121 [PubMed: 27899578]
- Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, ... Stratton MR (2004). A census of human cancer genes. *Nat Rev Cancer*, 4(3), 177–183. doi:10.1038/nrc1299 [PubMed: 14993899]
- Gahl WA, Wise AL, & Ashley EA (2015). The Undiagnosed Diseases Network of the National Institutes of Health: A National Extension. *JAMA*, 314(17), 1797–1798. doi:10.1001/jama.2015.12249 [PubMed: 26375289]
- Global Alliance for Genomics and Health. (2016). A federated ecosystem for sharing genomic, clinical data. *Science*, 352(6291), 1278–1280. doi:10.1126/science.aaf6162 [PubMed: 27284183]
- Gonzalez-Perez A, & Lopez-Bigas N (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet*, 88(4), 440–449. doi:10.1016/j.ajhg.2011.03.004 [PubMed: 21457909]
- Gordon AS, Fulton RS, Qin X, Mardis ER, Nickerson DA, & Scherer S (2016). PGRNseq: a targeted capture sequencing panel for pharmacogenetic research and implementation. *Pharmacogenet Genomics*. doi:10.1097/FPC.0000000000000202

- Hiatt SM, Amaral MD, Bowling KM, Finnila CR, Thompson ML, Gray DE, ... Cooper GM (2018). Systematic reanalysis of genomic data improves quality of variant interpretation. *Clin Genet*, 94(1), 174–178. doi:10.1111/cge.13259 [PubMed: 29652076]
- Jian X, Boerwinkle E, & Liu X (2014). In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res*, 42(22), 13534–13544. doi:10.1093/nar/gku1206 [PubMed: 25416802]
- Kalia SS, Adelman K, Bale SJ, Chung WK, Eng C, Evans JP, ... Miller DT (2017). Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet Med*, 19(2), 249–255. doi:10.1038/gim.2016.190 [PubMed: 27854360]
- Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, ... Wilson RK (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*, 22(3), 568–576. doi:10.1101/gr.129684.111 [PubMed: 22300766]
- Kopanos C, Tsiolkas V, Kouris A, Chapple CE, Albarca Aguilera M, Meyer R, & Massouras A (2018). VarSome: The Human Genomic Variant Search Engine. *Bioinformatics*. doi:10.1093/bioinformatics/bty897
- Kuhn RM, Haussler D, & Kent WJ (2013). The UCSC genome browser and associated tools. *Brief Bioinform*, 14(2), 144–161. doi:10.1093/bib/bbs038 [PubMed: 22908213]
- Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, ... Maglott DR (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res*, 44(D1), D862–868. doi:10.1093/nar/gkv1222 [PubMed: 26582918]
- Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, ... Exome Aggregation, C. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616), 285–291. doi:10.1038/nature19057 [PubMed: 27535533]
- Li B, Chen W, Zhan X, Busonero F, Sanna S, Sidore C, ... Abecasis GR (2012). A likelihood-based framework for variant calling and de novo mutation detection in families. *PLoS Genet*, 8(10), e1002944. doi:10.1371/journal.pgen.1002944 [PubMed: 23055937]
- Li J, Shi L, Zhang K, Zhang Y, Hu S, Zhao T, ... Sun Z (2018). VarCards: an integrated genetic and clinical database for coding variants in the human genome. *Nucleic Acids Res*, 46(D1), D1039–D1048. doi:10.1093/nar/gkx1039 [PubMed: 29112736]
- Liu X, Wu C, Li C, & Boerwinkle E (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum Mutat*, 37(3), 235–241. doi:10.1002/humu.22932 [PubMed: 26555599]
- MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, ... Tyler-Smith C (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, 335(6070), 823–828. doi:10.1126/science.1215040 [PubMed: 22344438]
- Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, ... Reich D (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, 538(7624), 201–206. doi:10.1038/nature18964 [PubMed: 27654912]
- McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, ... Cunningham F (2016). The Ensembl Variant Effect Predictor. *Genome Biol*, 17(1), 122. doi:10.1186/s13059-016-0974-4 [PubMed: 27268795]
- Meynert AM, Ansari M, FitzPatrick DR, & Taylor MS (2014). Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics*, 15, 247. doi:10.1186/1471-2105-15-247 [PubMed: 25038816]
- Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, ... Bamshad MJ (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet*, 42(1), 30–35. doi:10.1038/ng.499 [PubMed: 19915526]
- Pavan S, Rommel K, Mateo Marquina ME, Hohn S, Lanneau V, & Rath A (2017). Clinical Practice Guidelines for Rare Diseases: The Orphanet Database. *PLoS One*, 12(1), e0170365. doi:10.1371/journal.pone.0170365 [PubMed: 28099516]
- Pollard KS, Hubisz MJ, Rosenbloom KR, & Siepel A (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*, 20(1), 110–121. doi:10.1101/gr.097857.109 [PubMed: 19858363]

- Posey JE (2019). Genome sequencing and implications for rare disorders. *Orphanet J Rare Dis*, 14(1), 153. doi:10.1186/s13023-019-1127-0 [PubMed: 31234920]
- Posey JE, Harel T, Liu P, Rosenfeld JA, James RA, Coban Akdemir ZH, ... Lupski JR (2017). Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. *N Engl J Med*, 376(1), 21–31. doi:10.1056/NEJMoa1516767 [PubMed: 27959697]
- Prokop JW, May T, Strong K, Bilinovich SM, Bupp C, Rajasekaran S, ... Lazar J (2018). Genome sequencing in the clinic: the past, present, and future of genomic medicine. *Physiol Genomics*, 50(8), 563–579. doi:10.1152/physiolgenomics.00046.2018 [PubMed: 29727589]
- Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, ... Ostell JM (2014). RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res*, 42(Database issue), D756–763. doi:10.1093/nar/gkt1114 [PubMed: 24259432]
- Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, ... Lipman D (2009). The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res*, 19(7), 1316–1323. doi:10.1101/gr.080531.108 [PubMed: 19498102]
- Relling MV, Gardner EE, Sandborn WJ, Schmiegelow K, Pui CH, Yee SW, ... Clinical Pharmacogenetics Implementation, C. (2013). Clinical pharmacogenetics implementation consortium guidelines for thiopurine methyltransferase genotype and thiopurine dosing: 2013 update. *Clin Pharmacol Ther*, 93(4), 324–325. doi:10.1038/clpt.2013.4 [PubMed: 23422873]
- Rentzsch P, Witten D, Cooper GM, Shendure J, & Kircher M (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*, 47(D1), D886–D894. doi:10.1093/nar/gky1016 [PubMed: 30371827]
- Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, ... Committee, A. L. Q. A. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*, 17(5), 405–424. doi:10.1038/gim.2015.30 [PubMed: 25741868]
- Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, ... Galas DJ (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, 328(5978), 636–639. doi:10.1126/science.1186802 [PubMed: 20220176]
- Robinson JT, Thorvaldsdottir H, Wenger AM, Zehir A, & Mesirov JP (2017). Variant Review with the Integrative Genomics Viewer. *Cancer Res*, 77(21), e31–e34. doi:10.1158/0008-5472.CAN-17-0337 [PubMed: 29092934]
- Sangkuhl K, Berlin DS, Altman RB, & Klein TE (2008). PharmGKB: understanding the effects of individual genetic variants. *Drug Metab Rev*, 40(4), 539–551. doi:10.1080/03602530802413338 [PubMed: 18949600]
- Schwarze K, Buchanan J, Taylor JC, & Wordsworth S (2018). Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genet Med*. doi:10.1038/gim.2017.247
- Shirts BH, Salama JS, Aronson SJ, Chung WK, Gray SW, Hindorff LA, ... Overby CL (2015). CSER and eMERGE: current and potential state of the display of genetic information in the electronic health record. *J Am Med Inform Assoc*, 22(6), 1231–1242. doi:10.1093/jamia/ocv065 [PubMed: 26142422]
- Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, ... Haussler D (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8), 1034–1050. doi:10.1101/gr.3715005 [PubMed: 16024819]
- Sifrim A, Popovic D, Tranchevent LC, Ardeschirdavani A, Sakai R, Konings P, ... Moreau Y (2013). eXtasy: variant prioritization by genomic data fusion. *Nat Methods*, 10(11), 1083–1084. doi:10.1038/nmeth.2656 [PubMed: 24076761]
- Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, & Ng PC (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res*, 40(Web Server issue), W452–457. doi:10.1093/nar/gks539 [PubMed: 22689647]

- Splinter K, Adams DR, Bacino CA, Bellen HJ, Bernstein JA, Cheattle-Jarvela AM, ... Undiagnosed Diseases, N. (2018). Effect of Genetic Diagnosis on Patients with Previously Undiagnosed Disease. *N Engl J Med*, 379(22), 2131–2139. doi:10.1056/NEJMoa1714458 [PubMed: 30304647]
- Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, ... Cooper DN (2017). The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet*, 136(6), 665–677. doi:10.1007/s00439-017-1779-6 [PubMed: 28349240]
- Stephenson JD, Laskowski RA, Nightingale A, Hurles ME, & Thornton JM (2019). VarMap: a web tool for mapping genomic coordinates to protein sequence and structure and retrieving protein structural annotations. *Bioinformatics*. doi:10.1093/bioinformatics/btz482
- Tan TY, Dillon OJ, Stark Z, Schofield D, Alam K, Shrestha R, ... White SM (2017). Diagnostic Impact and Cost-effectiveness of Whole-Exome Sequencing for Ambulant Children With Suspected Monogenic Conditions. *JAMA Pediatr*, 171(9), 855–862. doi:10.1001/jamapediatrics.2017.1755 [PubMed: 28759686]
- Taylor JC, Martin HC, Lise S, Broxholme J, Cazier JB, Rimmer A, ... McVean G (2015). Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nat Genet*, 47(7), 717–726. doi:10.1038/ng.3304 [PubMed: 25985138]
- Turner TN, Yi Q, Krumm N, Huddleston J, Hoekzema K, HA FS, ... Eichler EE (2017). denovo-db: a compendium of human de novo variants. *Nucleic Acids Res*, 45(D1), D804–D811. doi:10.1093/nar/gkw865 [PubMed: 27907889]
- UniProt C (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*, 47(D1), D506–D515. doi:10.1093/nar/gky1049 [PubMed: 30395287]
- Vassy JL, McLaughlin HM, MacRae CA, Seidman CE, Lautenbach D, Krier JB, ... Green RC (2015). A one-page summary report of genome sequencing for the healthy adult. *Public Health Genomics*, 18(2), 123–129. doi:10.1159/000370102 [PubMed: 25612602]
- Wang J, Al-Ouran R, Hu Y, Kim SY, Wan YW, Wangler MF, ... Bellen HJ (2017). MARRVEL: Integration of Human and Model Organism Genetic Resources to Facilitate Functional Annotation of the Human Genome. *Am J Hum Genet*, 100(6), 843–853. doi:10.1016/j.ajhg.2017.04.010 [PubMed: 28502612]
- Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, van Kogelenberg M, ... (2015). Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet*, 385(9975), 1305–1314. doi:10.1016/S0140-6736(14)61705-0 [PubMed: 25529582]
- Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, ... Flicek P (2018). Ensembl 2018. *Nucleic Acids Res*, 46(D1), D754–D761. doi:10.1093/nar/gkx1098 [PubMed: 29155950]
- Zheng X, & Weir BS (2016). Eigenanalysis of SNP data with an identity by descent interpretation. *Theor Popul Biol*, 107, 65–76. doi:10.1016/j.tpb.2015.09.004 [PubMed: 26482676]

Figure 1. Workflow in CGAR.

(A) A simplified workflow in CGAR. From selected samples ('Variants') and/or phenotypes/genes of interest, CGAR initially filters for clinically implicated variants. Then, users browse identified variants and adjust filters (the shaded box). (B) CGAR workflow in detailed list of components. CGAR takes variant call files (VCFs) from whole genome and exome sequencing (WGS and WES) with optional inputs for phenotype and a list of genes. A single index case, trio, or tumor-normal pair can be used to create variant and gene centric reports with the interactive graphical user interface (GUI), through which a user can change filtering criteria such as zygosity, allele frequency threshold and predicted impact on protein function. To retrieve genes associated with given signs and symptoms, we use FindZebra (Dragusin et al., 2013) a machine learning based disease-gene search engine. Phenotype-based and/or user provided genes are used to create a report for variants in such genes. Rare or novel variants (i.e., allele frequency less than 0.5% in any population) with high impact – i.e., splice site disrupt, frameshift, nonsense, misstart, and nonstop – are displayed by default; however, allele frequency threshold according to population and predicted impacts can be easily set in the GUI by the user to select variants. Databases providing variant level annotation information are shown in red boxes and those with gene-level annotation information are shown in blue boxes. The disease-causing mutations reported in HGMD and rare disease genes collected in Orphanet (the dotted boxes) are accessible only to licensed users.

Figure 2. User interfaces in CGAR. (A) The ClinVar category in an example report for the selected sample ('Miller'). Only seven out of nine report categories are available for the current user. The dashed box on the left shows controls available for the current tab. The configurations of filters used to generate current tab can be checked from the grey boxes on top of the main table. Currently, it is displaying three variants under the settings: any zygosity, either pathogenic or likely pathogenic, allele frequency < 0.5% in any population, and non-synonymous variants (high or moderate impact). In the main table, each row represents different variants with columns as annotations for the variant. Some of the columns – 'ClinVar phenotype', 'Extra phenotype

links', and 'Review status' – are specific to the current category. (B) In a different category ('Pharmacogenomics') for the same sample, the controls on the left and the columns in the main table are changed for the category. CGAR also provides menus for further details on the variant as well as various external links to search for more information on the variant.

Table 1.
Comparison of features available within programs for integrative variant analysis.

In each column, features are grouped in three colors: green for features exclusive for the program, yellow for features shared with other programs, and orange for features not implemented in the program.

Available features		CGAR	MARRVEL	VarSome	VarCards
Disease/ Phenotypes	HGMD (Stenson et al., 2017)	✓			
	ClinVar (Landrum et al., 2016)	✓	✓	✓	✓
	OMIM (Amberger et al., 2019)	✓	✓		
	GWAS-catalog (Buniello et al., 2019)	✓			✓
	OrphaNet (Pavan et al., 2017)	✓			
Cancer- related	Variants in COSMIC (Forbes et al., 2017)	✓		✓	✓
	Genes from CENSUS (Futreal et al., 2004)	✓			
	Variants in ICGC (https://dcc.icgc.org)	✓		✓	✓
Variant pathogenicity predictions	CADD (Rentzsch et al., 2019)	✓	✓		✓
	ClinPred (Alirezaie, Kernohan, Hartley, Majewski, & Hocking, 2018)	✓			✓
	Condel (Gonzalez-Perez & Lopez-Bigas, 2011)	✓			
	DANN (Quang, Chen, & Xie, 2015)	✓			✓
	Eigen (Ionita-Laza, McCallum, Xu, & Buxbaum, 2016)	✓			✓
	FATHMM (Shihab et al., 2014)	✓		✓	✓
	FATHMM-MKL (Shihab et al., 2015)	✓		✓	✓
	GenoCanyon (Lu et al., 2015)	✓		✓	✓
	LRT (Chun & Fay, 2009)	✓		✓	✓
	M-CAP (Jagadeesh et al., 2016)	✓	✓	✓	✓
	MetaLR and MetaSVM	✓		✓	✓
	MutationAssessor (Reva, Antipin, & Sander, 2011)	✓		✓	✓
	MutationTaster (Schwarz, Cooper, Schuelke, & Seelow, 2014)	✓		✓	✓
	PolyPhen2 (Adzhubei et al., 2010)	✓	✓	✓	✓
	PROVEAN (Choi & Chan, 2015)	✓		✓	✓
	ReVe (J. Li, Zhao, et al., 2018)	✓		✓	✓
	REVEL (Ioannidis et al., 2016)	✓	✓	✓	✓
	SIFT (Sim et al., 2012)	✓		✓	✓
	VEST3 (Carter, Douville, Stenson, Cooper, & Karchin, 2013)	✓		✓	✓
Sequence conservation scores	GERP (Davydov et al., 2010)	✓	✓	✓	✓
	fitCons (Gulko, Hubisz, Gronau, & Siepel, 2015)	✓		✓	✓
	phastCons (Siepel et al., 2005) 100-way or 20-way	✓		✓	✓
	phyloP (Pollard et al., 2010) 100-way or 20-way	✓	✓	✓	✓
	SiPhy (Garber et al., 2009) 29-way	✓		✓	✓
Other features	Beacon Network query	✓		✓	
	Coverage in gnomAD	✓		✓	
	De novo variants in denovo-db (Turner et al., 2017)	✓			✓
	Pharmacogenomic data from CPIC (Relling & Klein, 2011), PharmGKB (Sangukhl et al., 2008)	✓			
	Protein domains	✓	✓	✓	
Input type	Mapping to protein structure (Stephenson et al., 2019)	✓			
	Secondary findings	✓			
	Reported structural variants	✓	✓	✓	
	Single variant	✓	✓	✓	✓
	VCF file	✓	✓	✓	✓
Available features		CGAR	MARRVEL	VarSome	VarCards