VarFish - Collaborative and Comprehensive Variant Analysis for Diagnosis and Research

Manuel Holtgrewe[1,2,*], Oliver Stolpe[1,2], Mikko Nieminen[1,3], Stefan Mundlos[4,5], Alexej Knaus[6], Uwe Kornak[4,5], Dominik Seelow[7,8], Lara Segebrecht[4], Malte Spielmann[4,5], Björn Fischer-Zirnsak[4,5], Felix Boschann[4], Ute Scholl[9,7], Nadja Ehmke[4], Dieter Beule[1,3]

[1] CUBI – Core Unit Bioinformatics, Berlin Institute of Health, Berlin, 10117, Germany
[2] Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, 10117, Germany
[3] Max Delbrück Center for Molecular Medicine, Berlin, 13125, Germany
[4] Institute of Medical Genetics and Human Genetics, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, 13353, Germany
[5] Development and Disease Group, Max Planck Institute for Medical Genetics, Berlin, 14195, Germany
[6] Institute for Genomic Statistics and Bioinformatics, University Hospital Bonn, Bonn, 53127, Germany
[7] Centrum für Therapieforschung, Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health (BIH), Berlin, Germany
[8] Berlin Institute of Health (BIH), Anna-Louisa-Karsch-Str. 2, 10178 Berlin, Germany
[9] Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Department of Nephrology and Medical Intensive Care, BCRT – Berlin Institute of Health Center for Regenerative Therapies, 13353 Berlin, Germany

[*] To whom correspondence should be addressed. Tel: +49 30 450 543 607; Fax: +49 30 450 543 901; Email: manuel.holtgrewe@bihealth.de

ABSTRACT

VarFish is a user-friendly web application for the quality control, filtering, prioritization, analysis, and user-based annotation of panel and exome variant data for rare disease genetics. It is capable of processing variant call files with single or multiple samples. The variants are automatically annotated with population frequencies, molecular impact, and presence in databases such as ClinVar. Further, it provides support for pathogenicity scores including CADD, MutationTaster, and phenotypic similarity scores. Users can filter variants based on these annotations and presumed inheritance pattern and sort the results by these scores. Filtered variants are listed with their annotations and many useful link-outs to genome browsers, other gene/variant data portals, and external tools for variant assessment. VarFish allows user to create their own annotations including support for variant assessment following ACMG-AMP guidelines. In close collaboration with medical practitioners, VarFish was designed for variant analysis and prioritization in diagnostic and research settings as described in the software's extensive manual. The user interface has been optimized for supporting these protocols. Users can install VarFish on their own in-house servers where it provides additional lab notebook features for collaborative analysis and allows re-analysis of cases, e.g., after update of genotype or phenotype databases.

INTRODUCTION

Targeted sequencing (1) such as gene panel or whole exome sequencing (WES) has become common in clinical genetics research and diagnostic applications. Whole genome sequencing (WGS) is an emerging approach for such applications, yet interpretation of small non-coding variants remains challenging, and WES is still considered the most cost-efficient (2). Of course, the exome-like variants from WGS data can be analyzed in the same fashion as WES data. The interest in this area is shown by the large number of tools available for the scoring, filtering, and prioritization of exome-wide variants including Phen-Gen (3), OVA (4), BiERapp (5), QueryOR (6), wANNOVAR (7), MutationDistiller (8), eXtasy (9), or the Exomiser (10). These prior works cover different feature sets (shown in Table 1) and differ in stability and availability of their source code. The latter is particularly important when authors discontinue their service. Common features include variant pathogenicity and gene-phenotype similarity scores, annotations and filtering by population frequencies. In the light of database updates, we view certain topics to be important emerging themes in the field of variant filtering and prioritizing. These include joint filtering of multiple cases, collaboration in analysis, building databases of analysed cases with variant assessments, and the re-evaluation of cases. Approaching these topics commonly requires duplicate work or advanced bioinformatics skills. Here, we report on our web-based application VarFish developed to tackle these challenges. VarFish is freely available without login at https://varfish-kiosk.bihealth.org and a demo version showcasing features available on custom install is available at https://varfish-demo.bihealth.org. Its source code is available under the permissive MIT license at https://github.com/bihealth/varfish-server.

RESULTS

Feature comparison with State-of-the-Art Tools

Table 1 shows the features implemented in VarFish in comparison to state-of-the art web tools (based on Figure 1 from (8)). The year of latest update is important as databases on variants and genes are growing quickly. The support of the Human Phenotype Ontology (HPO) and Online Mendelian Inheritance in Man (OMIM) is relevant for prioritizing single exomes. Filtering of genes and regions is of interest when characterizing patient cohorts for variants in known disease-causing genes. Furthermore, support for multi-sample files or even multiple VCF files at once allows more advanced analyses. Implementing quality control features is important to gauge the quality of data in an integrated platform. Dynamic variant reports greatly improve usability over repeated submission of queries to batch systems. Tabular downloads (e.g., as spreadsheet files) make it possible to archive cases on the user’s computer and store them in clinical or laboratory information management system. Supporting users in annotating variants with flags, colour codes extends such systems in the manner of a laboratory notebook. Supporting users in the ACMG-AMP (11) classification of variants is useful for creating diagnostic reports. The possibility to organize cases in projects with project-based access control further fosters collaboration in data analysis and allows for building in-house databases. Custom installations on the user's own server can address data privacy issues. Finally, the availability of extensive documentation, tutorials and providing the tools without requiring user registration lowers the entry barrier.

The VarFish Workflow

There are two major parts in the VarFish data processing workflow: the data *preprocessing and import* and the *query construction and execution* step.

The input to the *preprocessing and import* step is a file in VCF (Variant Call Format) format (12) and optionally a PLINK (13) pedigree file. Each variant record is read from the file and annotated with molecular impact using the Jannovar (14) library for both RefSeq (15) and ENSEMBL (16) transcripts, with the distance to the closest exon in either database, with its population frequencies in the ExAC (17), gnomAD (18), and Thousand Genomes Project (19) databases, and its presence in ClinVar (20). The variants are assigned a case identifier and are then imported together with properties such as quality scores from the VCF file into a PostgreSQL database table following the star schema pattern common in data warehouse applications.

The input to the *query construction and execution* step consists of the identifier of the case and the query settings from the user (see below). It constructs a SQL database query for selecting the variants based on the input criteria and joins the central variant table to further metadata tables (e.g., providing information about genes, variants, or conservation information). This query is then submitted to the database system for execution. VarFish was developed for use in genetics of rare diseases where users desire to create short lists of variants (say less than 200) for further analysis based on population frequency, genotype/segregation in families, molecular impact, and other criteria (21). In particular, the first three criteria can be used to greatly reduce the number of resulting variants. By employing the star schema pattern, database indices can be created for the most common queries and the (small) number of rows returned from the query on the central variant table can be obtained fast. The extensive metadata acquisition can then be limited to this small number of rows. As a periodic background job, VarFish tabulates the number of samples that each variant occurs in heterozygous, hemizygous, and homozygous state. This allows for removing variants seen in many cases as is the case for local polymorphisms or artifacts not seen in the population databases because of differences in variant calling.

Quality Control Functionality

Another feature in VarFish is a global quality control function. Figure 1 shows an example of the quality control (QC) plots available. The three plots follow the Peddy methodology (22) and allow samples be examined for (un)expected relatedness, the sex derived from X-chromosomal variants, and depth-of-coverage vs. fraction of heterozygous variants. A detailed description and interpretation guide is available in the VarFish user manual. Further, VarFish allows users to provide quality control information for each sample that cannot be derived from VCF files, such as coverage information in JSON format (https://json.org). This allows an integrated display of QC information suitable for clinicians as suggested by Shyr, c.f. (23). Finally, the user can consider this information for all samples in a project to evaluate a sample in comparison to similarly processed ones or for a whole cohort.

Database- and User-Based Annotation

VarFish integrates a growing list of databases. Databases such as gnomAD provide information on the variant frequencies within (sub)populations, dbSNP provides identifiers for registered variants, ClinVar provides variant pathogenicity and PubMed identifiers. Protein-level conservation information is derived from UCSC genome browser (24) data, the NCBI gene database provides gene summaries and gene reference into function (RIF) information, and the HPO (25) provides phenotype information. The background databases need to be imported when installing VarFish locally. An archive file with this data is provided for download. We provide the full Snakemake (26) workflow for downloading the data from open and free sources for reproducibility.

The VarFish result display lists extensive links to databases and data portals providing additional information. Furthermore, functionality for remote control of the integrative genome viewer (IGV) (27) and assessment of variants by tools such as MutationTaster (28) are provided. Resulting variant lists can be directly uploaded into MutationDistiller (8) for a complementary analysis. Further, users can annotate variants with flags, color codes, and free-form text as shown in Figure 2. This figure also shows the support for computing and storing using the ACMG-AMP guidelines (11).

Filtering Interface

Figure 3 illustrates the filtering interface and workflow from the user's perspective. The aim is to provide an easy access for inexperienced users yet offer high flexibility for experts. This is realized by providing two levels of presets. With no preset, VarFish provides a high degree of configurability for genotype, population frequency, variant quality measures (including variant call quality and variant allelic balance), and ClinVar annotation. On the first preset level, it provides default settings for several categories. For example, there are separate "super strict", "strict", and "relaxed" population frequency settings under the assumption of dominant mode of inheritance, separate ones for assumed recessive mode of inheritance, and "strict" and "relaxed" settings for variant quality measures. On the second preset level, VarFish allows the user to select between settings such as "*de novo*", "dominant inheritances", and "recessive inheritance." For example, the second-level preset "recessive inheritance" will set the genotype filter to return homozygous variants and compound heterozygous variants in the index (enforcing appropriate genotypes for the parents if present) with appropriate (yet strict) population frequencies and strict quality thresholds. The rationale is that users prefer to start reviewing a few promising variants first and then relax the filters while the total number of variants remains manageable. The user browses each variant in the raw data using IGV as well as the gene summary information and gene phenotype links. The results of this research can then be documented for each variant and flags are generated for the different categories (see above). In contrast, the second-level preset “*de novo*” sets the genotype filter to “heterozygous” for the index, and “reference” for all other members in the pedigree, the population frequencies are set to very restrictive values, while the quality thresholds are relaxed and deeply intronic variants with a distance of up to 100 bp are included. The rationale is that *de novo* variants are very rare in trios and that even non-coding and low-covered variants are worth getting inspected, the latter also under the aspect of possible mosaic disease causing variants (29).

Joint Filtering of Multiple Cases

VarFish is capable of filtering the variants of multiple cases at once. The resulting variants are annotated with the number of cases that have at least one variant identified in a given gene. The result can then be sorted by that number to identify genes that carry rare variants of interesting impact and mode of inheritance. To demonstrate this, we performed a re-analysis of the original cohort used for identifying *TGDS* as a disease gene for Catel-Manzke syndrome published earlier (30). For this analysis, the second preset “recessive inheritance” was applied first, followed by relaxing the quality thresholds as the data was generated in the early days of WES sequencing. Figure 4 shows the top results. Of the resulting 388 variant records the only gene carrying variants in all six sequenced cases was *TGDS*, the next gene listed matched only two cases.

DISCUSSION

We here present VarFish, a flexible platform for the automated annotation of small variants, their filtering and prioritization. We demonstrated its use and effectiveness in a practical use case. The system aims at empowering biomedical and clinician researchers to perform complex, customizable variant prioritization and filtering in a maximally flexible way. Instead of implementing new custom scores, VarFish builds on and combines best-in-class scoring algorithms such as CADD and MutationTaster for variant pathogenicity prediction and gene-phenotype similarity computation. Annotation, filtering, and prioritization are shown to the user, allowing swift interactive processing by sorting variants according to different criteria. Comprehensive link-outs to external gene and variant databases are provided, and the integration of the IGV genome browser enables raw data inspection. Allowing the user to leave annotation flags, colour code, and free-text comments has proven highly useful for our in-house users. Data can be analyzed in an integrated platform up to ACMG-AMP variant assessment, followed by downloading a spreadsheet file for documentation in external systems.

Further advanced features such as collecting cases in projects and performing joint queries on multiple cases at once allow answering research questions that previously required bioinformatics expertise. For example, this allows an integrated characterization of disease cohorts by screening for pathogenic variants in known disease-associated genes followed by a joint analysis of the remaining cases, e.g., to identify jointly mutated genes.

Finally, the pipeline to generate background the database files from publicly available sources, and the pipeline for annotating VCF files and the filtering user interface is available under the permissive MIT open source license. This ensures full transparency, allows for setting up a fully reproducible variant analysis pipeline, and provides users with the advantage of open source systems in that there is no vendor-locking (in concordance with the FAIR (31) data management principles) and users are independent of the original software author. VarFish is used in the authors' daily work and actively maintained. We welcome questions, comments, and suggestions via email or the Github project's issue tracker.

CONCLUSION

VarFish is a flexible and powerful platform for variant filtering, prioritization, and user-based annotation. With its collaboration and laboratory notebook features, it promotes collaborative analysis of cases and the re-analysis of variants at multiple points in time. Future development will focus on the extension of re-analysis features and supporting the analysis of structural variants, and support for whole genome data.

WEB SERVER IMPLEMENTATION

VarFish is implemented in Python 3 with the Django web framework based on SODAR-core (https://github.com/bihealth/sodar_core) using PostgreSQL 11 for data storage and querying and VCFPy (32) for file parsing. In our installation, it runs on Linux container server with 128 GB of RAM, 16 cores, and 1 TB of disk.

AVAILABILITY

VarFish is available for public usage at https://varfish-kiosk.bihealth.org. A demonstration instance with full collaboration features has been setup at https://varfish-demo.bihealth.org. The source code is available under the permissive MIT license from https://github.com/bihealth/varfish-server.

ACCESSION NUMBERS

No new data was generated for this study. For demonstration purposes we used data from a previous publication (30).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENT

The authors thank Hannah Kemmer, Jirko Kühnisch, and Elisa Schäfer for their feedback on VarFish.

FUNDING

U.S. is funded by Stiftung Charité. N.E is funded by a Rahel-Hirsch-Stipendium.

CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

REFERENCES

1. Shendure,J. and Ji,H. (2008) Next-generation DNA sequencing. *Nat Biotechnol*, **26**, 1135–1145.

https://doi.org/10.1038/nbt1486

2. Sun,Y., Ruivenkamp,C.A.L., Hoffer,M.J.V., Vrijenhoek,T., Kriek,M., van Asperen,C.J., den Dunnen,J.T. and Santen,G.W.E. (2015) Next-Generation Diagnostics: Gene Panel, Exome, or Whole Genome? *Human Mutation*, **36**, 648–655.

https://doi.org/10.1002/humu.22783

3. Javed,A., Agrawal,S. and Ng,P.C. (2014) Phen-Gen: combining phenotype and genotype to analyze rare disorders. *Nat Methods*, **11**, 935–937.

https://doi.org/10.1038/nmeth.3046

4. Antanaviciute,A., Watson,C.M., Harrison,S.M., Lascelles,C., Crinnion,L., Markham,A.F., Bonthron,D.T. and Carr,I.M. (2015) OVA: Integrating molecular and physical phenotype data from multiple biomedical domain ontologies with variant filtering for enhanced variant prioritization. *Bioinformatics*, 10.1093/bioinformatics/btv473.

https://doi.org/10.1093/bioinformatics/btv473

5. Alemán,A., Garcia-Garcia,F., Salavert,F., Medina,I. and Dopazo,J. (2014) A web-based interactive framework to assist in the prioritization of disease candidate genes in whole-exome sequencing studies. *Nucleic Acids Research*, **42**, W88–W93.

https://doi.org/10.1093/nar/gku407

6. Bertoldi,L., Forcato,C., Vitulo,N., Birolo,G., De Pascale,F., Feltrin,E., Schiavon,R., Anglani,F., Negrisolo,S., Zanetti,A., *et al.* (2017) QueryOR: a comprehensive web platform for

genetic variant analysis and prioritization. *BMC Bioinformatics*, **18**, 225.

https://doi.org/10.1186/s12859-017-1654-4

7. Chang,X. and Wang,K. (2012) wANNOVAR: annotating genetic variants for personal genomes via the web. *J. Med. Genet.*, **49**, 433–436.

https://doi.org/10.1136/jmedgenet-2012-100918

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3556337

8. Hombach,D., Schuelke,M., Knierim,E., Ehmke,N., Schwarz,J.M., Fischer-Zirnsak,B. and Seelow,D. (2019) MutationDistiller: user-driven identification of pathogenic DNA variants. *Nucleic Acids Res*, **47**, W114–W120.

https://doi.org/10.1093/nar/gkz330

9. Sifrim,A., Popovic,D., Tranchevent,L.-C., Ardeshirdavani,A., Sakai,R., Konings,P., Vermeesch,J.R., Aerts,J., De Moor,B. and Moreau,Y. (2013) eXtasy: variant prioritization by genomic data fusion. *Nat Methods*, **10**, 1083–1084.

https://doi.org/10.1038/nmeth.2656

10. Smedley,D., Jacobsen,J.O.B., Jäger,M., Köhler,S., Holtgrewe,M., Schubach,M., Siragusa,E., Zemojtel,T., Buske,O.J., Washington,N.L., *et al.* (2015) Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat Protoc*, **10**, 2004–2015.

https://doi.org/10.1038/nprot.2015.124

11. Richards,S., Aziz,N., Bale,S., Bick,D., Das,S., Gastier-Foster,J., Grody,W.W., Hegde,M., Lyon,E., Spector,E., *et al.* (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**, 405–423.

https://doi.org/10.1038/gim.2015.30

12. Danecek,P., Auton,A., Abecasis,G., Albers,C.A., Banks,E., DePristo,M.A., Handsaker,R.E., Lunter,G., Marth,G.T., Sherry,S.T., *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

https://doi.org/10.1093/bioinformatics/btr330

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3137218

13. 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. *The American Journal of Human Genetics*, **81**, 559–575.

https://doi.org/10.1086/519795

14. Jäger,M., Wang,K., Bauer,S., Smedley,D., Krawitz,P. and Robinson,P.N. (2014) Jannovar: a java library for exome annotation. *Hum. Mutat.*, **35**, 548–555.

https://doi.org/10.1002/humu.22531

http://www.ncbi.nlm.nih.gov/pubmed/24677618

15. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D., *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733-745.

https://doi.org/10.1093/nar/gkv1189

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702849

16. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Girón,C.G., *et al.* (2018) Ensembl 2018. *Nucleic Acids Res*, **46**, D754–D761.

https://doi.org/10.1093/nar/gkx1098

17. 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.

https://doi.org/10.1038/nature19057

18. 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, 10.1101/531210, 13-August-2019, pre-print: not peer-reviewed Genomics.

19. A global reference for human genetic variation (2015) *Nature*, **526**, 68–74.

https://doi.org/10.1038/nature15393

20. 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.

https://doi.org/10.1093/nar/gkx1153

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5753237

21. Robinson,P.N., Piro,R.M. and Jäger,M. (2018) Computational exome and genome analysis CRC Press, Boca Raton.

22. Pedersen,B.S. and Quinlan,A.R. (2017) Who's Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy. *The American Journal of Human Genetics*, **100**, 406–413.

https://doi.org/10.1016/j.ajhg.2017.01.017

23. Shyr,C., Kushniruk,A., van Karnebeek,C.D.M. and Wasserman,W.W. (2016) Dynamic software design for clinical exome and genome analyses: insights from bioinformaticians, clinical geneticists, and genetic counselors. *J Am Med Inform Assoc*, **23**, 257–268.

https://doi.org/10.1093/jamia/ocv053

24. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler, and D. (2002) The Human Genome Browser at UCSC. *Genome Res.*, **12**, 996–1006.

https://doi.org/10.1101/gr.229102

http://www.ncbi.nlm.nih.gov/pubmed/12045153

25. Robinson,P.N., Köhler,S., Bauer,S., Seelow,D., Horn,D. and Mundlos,S. (2008) The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *Am J Hum Genet*, **83**, 610–615.

https://doi.org/10.1016/j.ajhg.2008.09.017

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2668030

26. Köster,J. and Rahmann,S. (2012) Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.

https://doi.org/10.1093/bioinformatics/bts480

27. Robinson,J.T., Thorvaldsdóttir,H., Winckler,W., Guttman,M., Lander,E.S., Getz,G. and Mesirov,J.P. (2011) Integrative Genomics Viewer. *Nat Biotechnol*, **29**, 24–26.

https://doi.org/10.1038/nbt.1754

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3346182

28. Schwarz,J.M., Cooper,D.N., Schuelke,M. and Seelow,D. (2014) MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods*, **11**, 361–362.

https://doi.org/10.1038/nmeth.2890

29. Cao,Y., Tokita,M.J., Chen,E.S., Ghosh,R., Chen,T., Feng,Y., Gorman,E., Gibellini,F., Ward,P.A., Braxton,A., *et al.* (2019) A clinical survey of mosaic single nucleotide variants in disease-causing genes detected by exome sequencing. *Genome Med*, **11**, 48.

https://doi.org/10.1186/s13073-019-0658-2

30. Ehmke,N., Caliebe,A., Koenig,R., Kant,S.G., Stark,Z., Cormier-Daire,V., Wieczorek,D., Gillessen-Kaesbach,G., Hoff,K., Kawalia,A., *et al.* (2014) Homozygous and Compound-Heterozygous Mutations in TGDS Cause Catel-Manzke Syndrome. *The American Journal of Human Genetics*, **95**, 763–770.

https://doi.org/10.1016/j.ajhg.2014.11.004

31. Wilkinson,M.D., Dumontier,M., Aalbersberg,Ij.J., Appleton,G., Axton,M., Baak,A., Blomberg,N., Boiten,J.-W., da Silva Santos,L.B., Bourne,P.E., *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, **3**, 160018.

https://doi.org/10.1038/sdata.2016.18

32. Holtgrewe,M. and Beule,D. (2016) VCFPy: a Python 3 library with good support for both reading and writing VCF. *JOSS*, **1**, 85.

https://doi.org/10.21105/joss.00085

TABLE AND FIGURES LEGENDS

Table 1. Feature comparison of state-of-the-art tools for variant filtering and prioritization.

	Exomiser	Phen-Gen	eXtasy	OVA	QueryOR	w-ANNOVAR	BiERapp	Mutation Distiller	VarFish
latest update	2019	2014	2013	2015	2017	2017	2017	2020	2020
HPO	✔	✔	✔	✔	✔	✔	✔	✔	✔
OMIM				✔	✔	✔		✔	✔
gene lists/regions	✔			✔	✔		✔	✔	✔
no registration	✔	✔	✔	✔		✔		✔	✔
gene info				✔	✔		✔	✔	✔
demo / tutorial		✔		✔	✔	✔	✔	✔	✔
dynamic report		✔			✔		✔	✔	✔
tabular download		✔	✔	✔	✔	✔	✔	✔	✔
multi-sample VCF	(✔)	✔			✔	(✔)	✔		✔
bring your server	✔	✔	✔				✔		✔
user annotations									✔
ACMG-AMP class									✔
quality control									✔
multiple VCFs									✔*
in-house DB									✔*
collaboration									✔*

A tick mark (✔) indicates that the feature is present. “Dynamic reports” allows users to interactively change filter and sorting options. “Tabular downloads” allows users to download a spreadsheet (e.g. Excel) file with their results. “User annotations” allow users to leave flag, color coded, or free-text comments, “ACMG-AMPs support” assists users in creating variant assessments following the ACMG guidelines. “Quality control” allows users to perform quality control, e.g., for depth of coverage or relatedness of individuals. “Multi-sample VCFs” supports cases with more than one sample while “multiple VCFs” supports querying multiple cases at the same time. “Bring your server” allows users to create their own installations on their own server. “In-house DB” allows to build a database of variants identified at the user’s institution. An asterisk (*) indicates that the feature is only available on installations on the user’s own server. Parentheses around the tick mark in “multi-sample VCF” row indicates that filtering is restricted to predefined models of inheritance.

Figure 1. Quality control plots following the Peddy (22) approach. The plots are described in the main text.

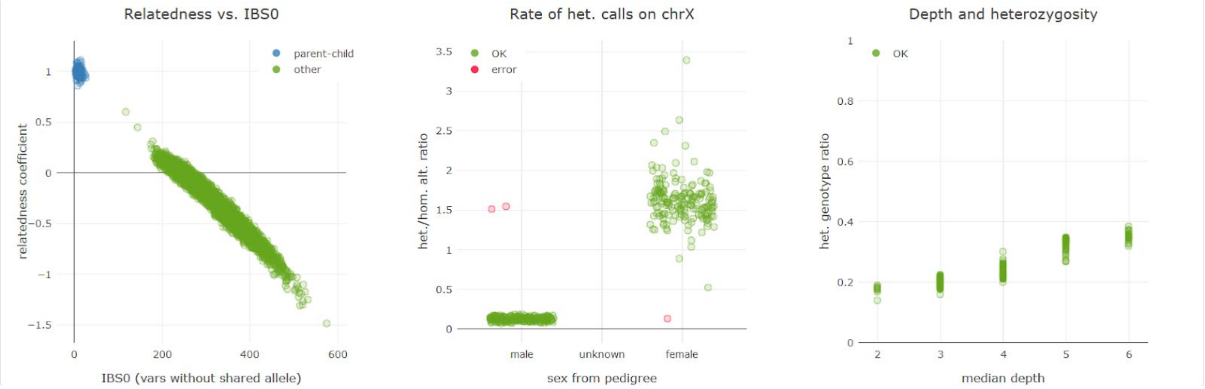
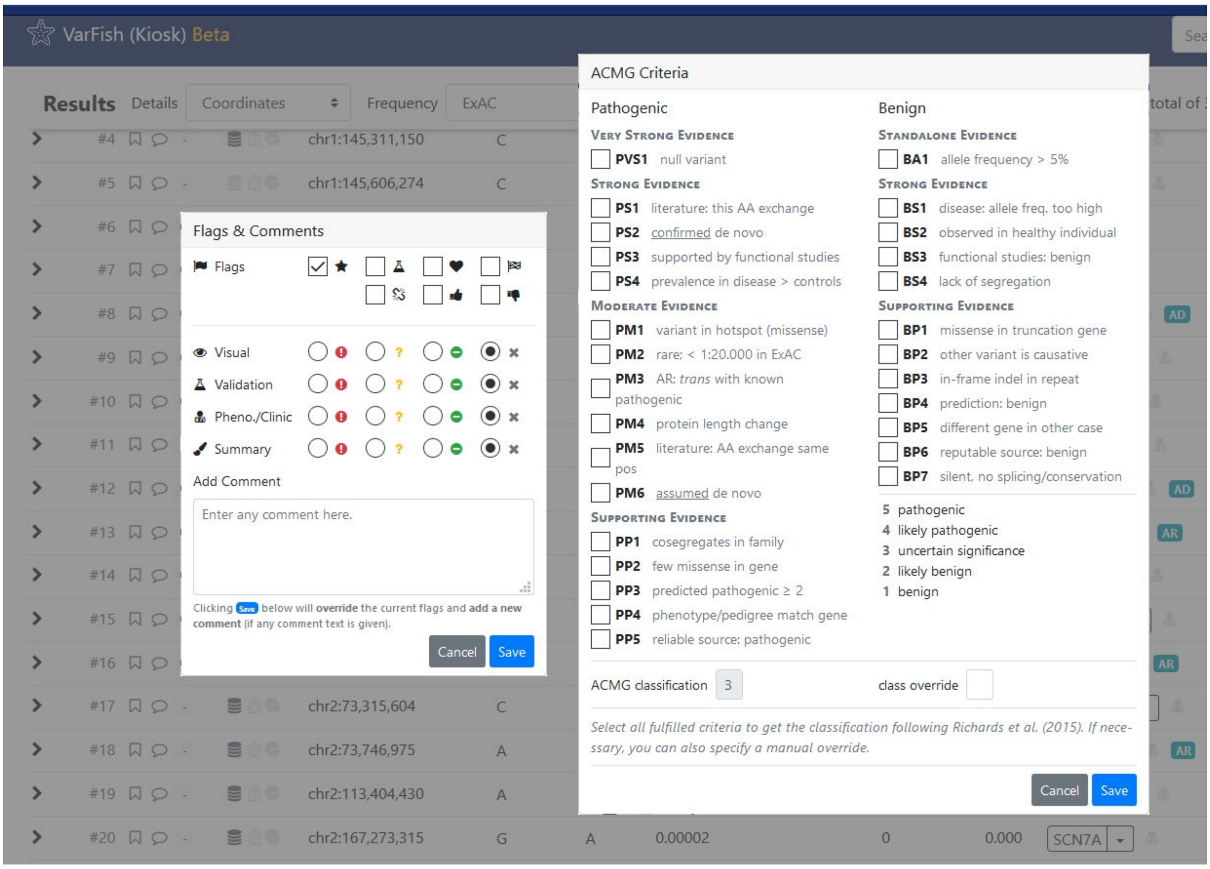


Figure 2. User annotation of variants. Users can apply flags and color codes to variants and leave free-text annotations. Flags include “bookmark”, “reported as candidate” and “final causative variant” as well as “no phenotype linked to gene”. Color codes can be assigned in categories “raw data visual inspection”, “gene clinical/phenotype match” and “validation results” as well as an overall summary color.



Flags & Comments

Flags ☑ ★ □ □ □ □ □ □ □

Visual ○ ○ ○ ●

Validation ○ ○ ○ ●

Pheno./Clinic ○ ○ ○ ●

Summary ○ ○ ○ ●

Add Comment

Enter any comment here.

Clicking Save below will **override** the current flags and **add a new comment** (if any comment text is given).

Cancel Save

ACMG Criteria

Pathogenic

VERY STRONG EVIDENCE

- □ **PVS1** null variant

STRONG EVIDENCE

- □ **PS1** literature: this AA exchange
- □ **PS2** confirmed de novo
- □ **PS3** supported by functional studies
- □ **PS4** prevalence in disease > controls

MODERATE EVIDENCE

- □ **PM1** variant in hotspot (missense)
- □ **PM2** rare: < 1:20.000 in ExAC
- □ **PM3** AR: *trans* with known pathogenic
- □ **PM4** protein length change
- □ **PM5** literature: AA exchange same pos
- □ **PM6** assumed de novo

SUPPORTING EVIDENCE

- □ **PP1** cosegregates in family
- □ **PP2** few missense in gene
- □ **PP3** predicted pathogenic ≥ 2
- □ **PP4** phenotype/pedigree match gene
- □ **PP5** reliable source: pathogenic

Benign

STANDALONE EVIDENCE

- □ **BA1** allele frequency > 5%

STRONG EVIDENCE

- □ **BS1** disease: allele freq. too high
- □ **BS2** observed in healthy individual
- □ **BS3** functional studies: benign
- □ **BS4** lack of segregation

SUPPORTING EVIDENCE

- □ **BP1** missense in truncation gene
- □ **BP2** other variant is causative
- □ **BP3** in-frame indel in repeat
- □ **BP4** prediction: benign
- □ **BP5** different gene in other case
- □ **BP6** reputable source: benign
- □ **BP7** silent, no splicing/conservation

5	pathogenic
4	likely pathogenic
3	uncertain significance
2	likely benign
1	benign

ACMG classification 3

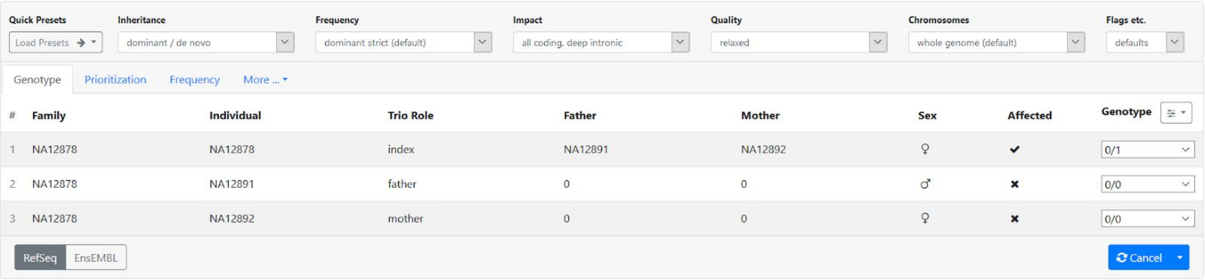
class override

Select all fulfilled criteria to get the classification following Richards et al. (2015). If necessary, you can also specify a manual override.

Cancel Save

.

Figure 3. The filtering interface. XXX



Quick Presets: Load Presets

Inheritance: dominant / de novo

Frequency: dominant strict (default)

Impact: all coding, deep intronic

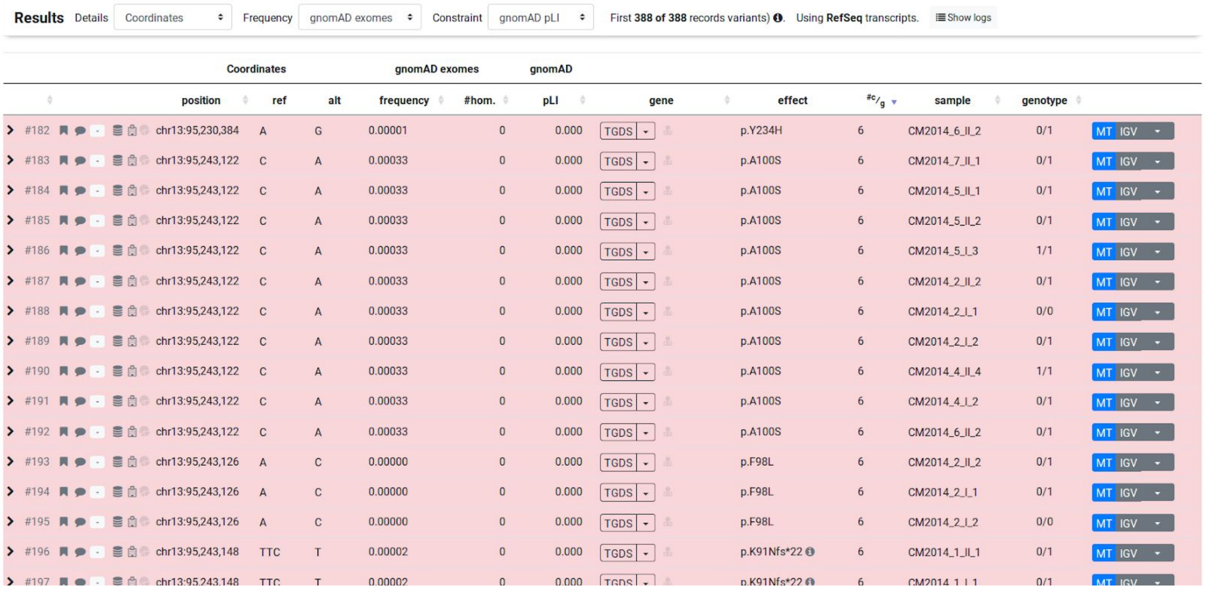
Quality: relaxed

Chromosomes: whole genome (default)

Flags etc.: defaults

#	Family	Individual	Trio Role	Father	Mother	Sex	Affected	Genotype
1	NA12878	NA12878	index	NA12891	NA12892	♀	✓	0/1
2	NA12878	NA12891	father	0	0	♂	✗	0/0
3	NA12878	NA12892	mother	0	0	♀	✗	0/0

Figure 4. Catel-Manzke cohort filtering results (first 15 variants shown) to reproduce the finding that TGDS is the most likely candidate for being the disease gene.



Results Details: Coordinates · Frequency: gnomAD exomes · Constraint: gnomAD pLI · First **388 of 388** records variants). Using **RefSeq** transcripts. · Show logs

	Coordinates			gnomAD exomes		gnomAD						
	position	ref	alt	frequency	#hom.	pLI	gene	effect	$^{\#c}/_{g}$	sample	genotype	
#182	chr13:95,230,384	A	G	0.00001	0	0.000	TGDS	p.Y234H	6	CM2014_6_II_2	0/1	MT IGV
#183	chr13:95,243,122	C	A	0.00033	0	0.000	TGDS	p.A100S	6	CM2014_7_II_1	0/1	MT IGV
#184	chr13:95,243,122	C	A	0.00033	0	0.000	TGDS	p.A100S	6	CM2014_5_II_1	0/1	MT IGV
#185	chr13:95,243,122	C	A	0.00033	0	0.000	TGDS	p.A100S	6	CM2014_5_II_2	0/1	MT IGV
#186	chr13:95,243,122	C	A	0.00033	0	0.000	TGDS	p.A100S	6	CM2014_5_I_3	1/1	MT IGV
#187	chr13:95,243,122	C	A	0.00033	0	0.000	TGDS	p.A100S	6	CM2014_2_II_2	0/1	MT IGV
#188	chr13:95,243,122	C	A	0.00033	0	0.000	TGDS	p.A100S	6	CM2014_2_I_1	0/0	MT IGV
#189	chr13:95,243,122	C	A	0.00033	0	0.000	TGDS	p.A100S	6	CM2014_2_I_2	0/1	MT IGV
#190	chr13:95,243,122	C	A	0.00033	0	0.000	TGDS	p.A100S	6	CM2014_4_II_4	1/1	MT IGV
#191	chr13:95,243,122	C	A	0.00033	0	0.000	TGDS	p.A100S	6	CM2014_4_I_2	0/1	MT IGV
#192	chr13:95,243,122	C	A	0.00033	0	0.000	TGDS	p.A100S	6	CM2014_6_II_2	0/1	MT IGV
#193	chr13:95,243,126	A	C	0.00000	0	0.000	TGDS	p.F98L	6	CM2014_2_II_2	0/1	MT IGV
#194	chr13:95,243,126	A	C	0.00000	0	0.000	TGDS	p.F98L	6	CM2014_2_I_1	0/1	MT IGV
#195	chr13:95,243,126	A	C	0.00000	0	0.000	TGDS	p.F98L	6	CM2014_2_I_2	0/0	MT IGV
#196	chr13:95,243,148	TTC	T	0.00002	0	0.000	TGDS	p.K91Nfs*22	6	CM2014_1_II_1	0/1	MT IGV
#197	chr13:95,243,148	TTC	T	0.00002	0	0.000	TGDS	p.K91Nfs*22	6	CM2014_1_I_1	0/1	MT IGV