PLOS ONE

RESEARCH ARTICLE

HuVarBase: A human variant database with comprehensive information at gene and protein levels

Kaliappan Ganesan[1]*, **A. Kulandaisamy**[1], **S. Binny Priya**[1], **M. Michael Gromiha**[1,2]*

1 Department of Biotechnology, Bhupat and Jyoti Mehta School of BioSciences, Indian Institute of Technology Madras, Chennai, Tamilnadu, India, **2** Advanced Computational Drug Discovery Unit (ACDD), Institute of Innovative Research, Tokyo Institute of Technology, Midori-ku, Yokohama, Kanagawa, Japan

* gromiha@iitm.ac.in (MMG); kganeshnew@gmail.com (KG)

Citation: Ganesan K, Kulandaisamy A, Binny Priya S, Gromiha MM (2019) HuVarBase: A human variant database with comprehensive information at gene and protein levels. PLoS ONE 14(1): e0210475. https://doi.org/10.1371/journal.pone.0210475

Editor: Ozlem Keskin, Koç University, TURKEY

Received: September 17, 2018

Accepted: December 25, 2018

Published: January 31, 2019

Data Availability Statement: https://www.iitm.ac.in/bioinfo/huvarbase

Funding: This project work was partially supported by the Department of Biotechnology, Government of India to MMG (No. BT/PR16710/BID/7/680/2016). There was no additional external funding received for this study.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Human variant databases could be better exploited if the variant data available in multiple resources is integrated in a single comprehensive resource along with sequence and structural features. Such integration would improve the analyses of variants for disease prediction, prevention or treatment. The HuVarBase (HUmanVARiantdataBASE) assimilates publicly available human variant data at protein level and gene level into a comprehensive resource. Protein level data such as amino acid sequence, secondary structure of the mutant residue, domain, function, subcellular location and post-translational modification are integrated with gene level data such as gene name, chromosome number & genome position, DNA mutation, mutation type origin and rs ID number. Disease class has been added for the disease causing variants. The database is publicly available at https://www.iitm.ac.in/bioinfo/huvarbase. A total of 774,863 variant records, integrated in the HuVarBase, can be searched with options to display, visualize and download the results.

Introduction

Human variant databases are being created frequently with specific scopes and contents. Their significance ranges from accurately predicting the disease [1] to achieving personalized medicine [2]. However, as the scopes of these databases differ, variant related data is inevitably spread across a number of databases such as 1000 Genomes [3], COSMIC [4], ClinVar [5], SwissVar [6], Humsavar (https://www.uniprot.org/docs/humsavar) etc. These databases have various limitations, including the fact that the data structures of these databases are not compatible with each other. As a result, obtaining comprehensive information on a variant of interest is still challenging for geneticists, biologists and clinicians [7]. Although integrating resources for the analysis of variant data obtained through Next Generation Sequencing (NGS) has been reported earlier [8], the integration is for tools and pipelines and not for the variant data.

Further, currently available variant databases do not include protein level data namely sequence, structural or functional information about the protein which has the variant.

Moreover, for disease causing variants, the disease class information is not available in the current databases. Recently, Kulandaisamy et al. [9] reported the MutHTP, wherein, gene level and protein level information related to disease causing and neutral variants have been compiled in a comprehensive manner. However, MutHTP is limited to variants reported in human membrane proteins only.

We present here the HuVarBase (HUmanVARiant dataBASE), which is a comprehensive database for collating human variant data along with protein level data such as secondary structure of the residue in which the mutation has occurred, protein domain, subcellular localization, post-translational modification and function of the protein in which the variant is reported. In addition, if a variant leads to disease, then the disease class information also is included. The database has been implemented in a searchable server and made available online

Materials and methods

Datasets and curation

An outline of the datasets and curation steps are given in Fig 1. The current human variant datasets available in 1000 Genomes [3], ClinVar [4], COSMIC [5], SwissVar [6] and Humsavar (https://www.uniprot.org/docs/humsavar) were downloaded and merged based on their UniProt [10] identifiers. If the UniProt identifiers were not available in the dataset, then the same were obtained from UniProt database using the Gene Name or sequence identifiers. For the COSMIC [5] dataset, variants reported in two or more tumor samples were designated as cancer causing (driver mutations) [11] and included in our databases. Remaining COSMIC variant data were not included in our database. If a variant is reported in more than one database, then the variant data is merged and the respective databases are mentioned in the source column. Protein sequences corresponding to the UniProt identifiers were obtained from the UniProt server. For a given variant, the neighboring residue information(three residues each, before and after the mutated residue) was taken from either the UniProt canonical protein sequence, or the UniProt isoform protein sequence, whichever had the amino acid residue in which the variant was reported. Protein Data Bank—PDB [12] identifiers of proteins and the secondary structure information of the variant residue were obtained from the SIFTS [13] database. Chromosome number and & genome position corresponding to the variants were collected from the neXtProt [14] database.

Subcellular localization, function and post-translational modifications of the proteins were obtained from the UniProt database. Protein domains in which the variant residues were present were obtained from the Pfam (https://pfam.xfam.org) database [15]. The disease classes were obtained from the KEGG [16] database based on the disease description. If the disease description does not match with that of the KEGG database, then other sources like Genetic and Rare Diseases Information Center (https://rarediseases.info.nih.gov/) [17] and Genetic Testing Registry (https://www.ncbi.nlm.nih.gov/gtr/) [18] were referred to obtain the name of the disease or synonyms of the disease. Then using that particular information, disease class was obtained from KEGG database. The present scope of the HuVarBase is to include small-scale variants of 'missense', 'insertion' and 'deletion' types along with the 'non-sense' type. Large-scale variations like large copy number variations will be added in future versions of HuVarBase.

Web server and site

HuVarBase is available online at https://www.iitm.ac.in/bioinfo/huvarbase/ and it is meant to be used for non-clinical academic purposes only. The server works on a Linux-Apache-

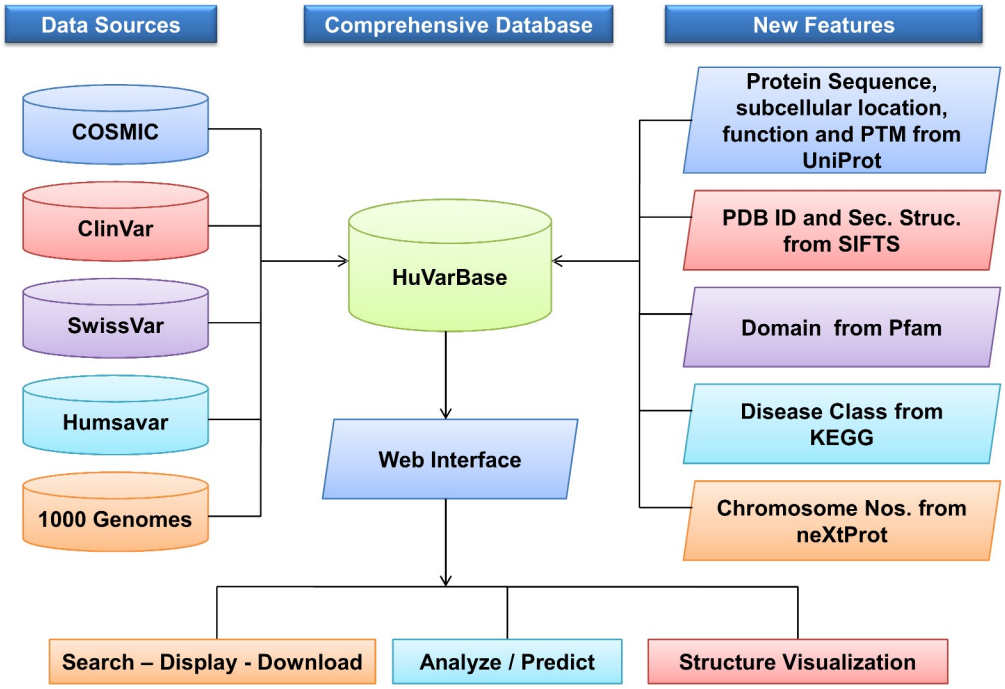

Fig 1. Schematic diagram describing the data collection, features and applications of HuVarBase.

https://doi.org/10.1371/journal.pone.0210475.g001

MySQL-PHP (LAMP) architecture. Most fields in this database are searchable by either entering keywords or by choosing a keyword in the drop down menu. Users can also choose the required fields to be shown in the search results. Hyperlinks are given in the search results to, GeneCards [19] for information on the gene in which the given variant has occurred, dbSNP server [20] for rs ID numbers, UniProt server for information on the protein of the variant, Jmol (http://www.jmol.org) to view the 3D structure of the protein with the variant residue, the protein sequence with color-coded residues to differentiate neutral and disease causing variants and to the source databases. The search results can also be downloaded as a single file. The Frequently Asked Questions section contains answers to common questions raised by the users. A brief tutorial is also available on how to search and obtain variant data from HuVarBase.

Results

HuVarBase database includes 702,048 disease causing variants of which, 652,399 are 'missense', 10,174 are 'nonsense', 8,850 are 'insertion', and 30,625 are 'deletion' variants. There are 72,815 neutral variants (i.e. limited or no evidence on the pathogenic role of the mutation) of which 66,191 are 'missense', 2,885 are 'nonsense', 259 are 'insertion', and 3,480 are 'deletion' variants. In total, there are 774,863 variants reported from 18,318 proteins. The number of variants contributed from each of the data sources to the HuVarBase is depicted in Fig 2.

For a given variant record, the fields included are; i) Gene level data: name of the gene which has the variant, chromosomal coordinates, DNA mutation, type of the mutation, dbSNP [20] rs ID number if available and origin, ii) Protein level data: protein mutation, UniProt and PDB IDs of the corresponding protein, UniProt ID of the canonical or isoform of the protein in which the mutation has been reported, neighboring amino acids of the mutated amino acid, secondary structure of the particular amino acid residue, conservation score of the

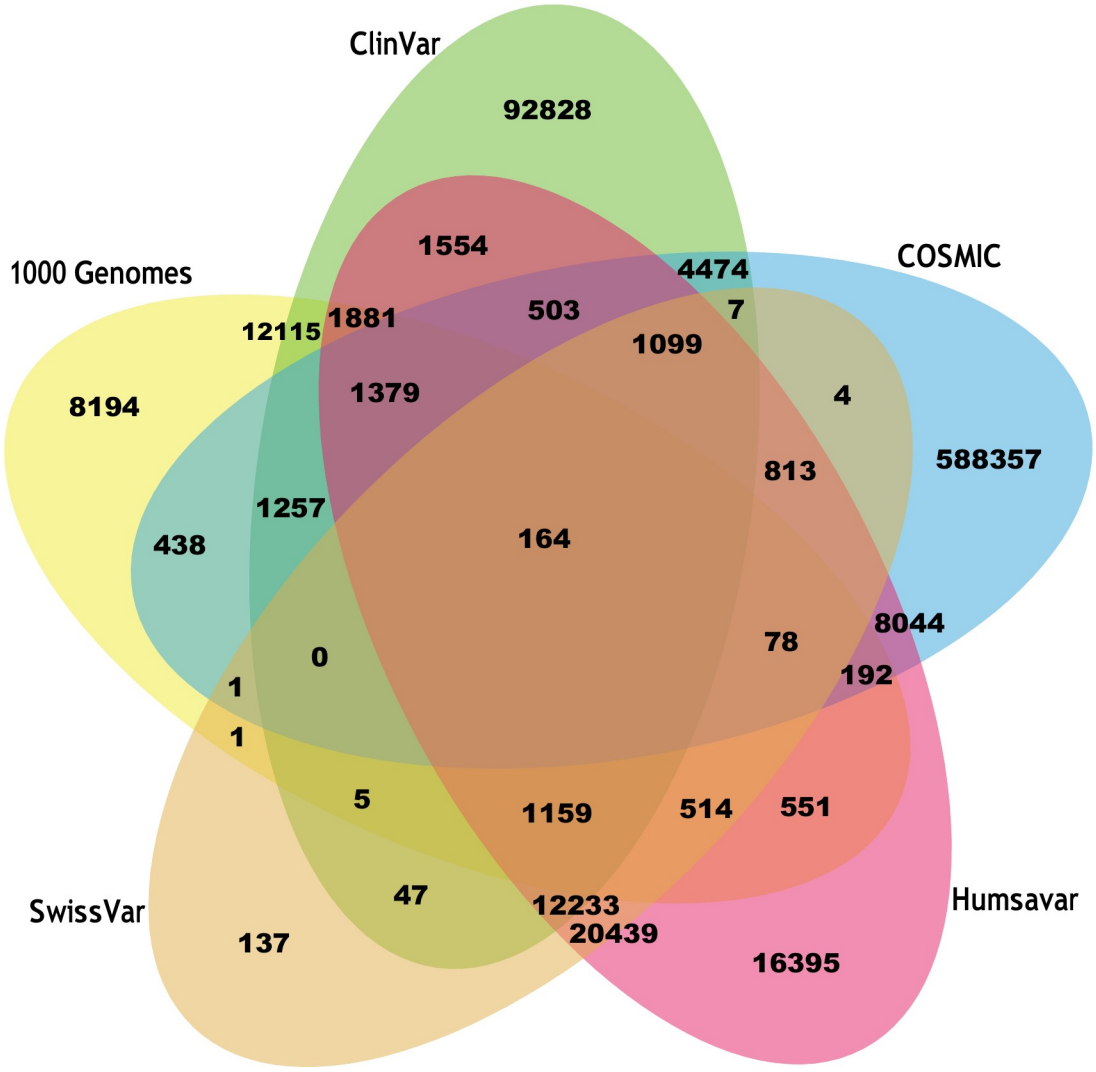

Fig 2. Venn diagram depicting the number of variants contributed from each of the data sources.

https://doi.org/10.1371/journal.pone.0210475.g002

residue change, function of the protein, protein domain in which the mutation has occurred, subcellular localization of the protein and post-translational modifications of the protein, and iii) others: Disease or tissue type in which the variant has been reported, disease class and the source database. Statistics regarding the database such as disease class frequency matrices for the disease causing and neutral mutations, etc. are given in the web server. Table 1 gives a comparison of features included in currently available human variant databases with HuVarBase.

Applications

The comprehensive HuVarBase facilitates searching for a variant and obtaining variant-related sequence and structural information for viewing or downloading. The HuVarBase includes small scale variant types such as 'missense', 'insertion'and 'deletion' along with the type 'nonsense' in the human genome. The applications of HuVarBase range from analysis, disease prediction to personalized medicine as exemplified by earlier efforts [21–23]. Analysis of the variants can be performed at sequence and structural level, in order to understand the effects of

Table 1. Comparison of features in HuVarBase with existing databases.

Features	Humsavar	SwissVar	1000 Genomes	COSMIC	ClinVar	MutHTP	HuVarBase
Gene name	Yes	Yes	Yes	Partial	Yes	Yes	Yes
Chromosome number	No	No	Yes	Yes	Yes	Yes	Yes
Origin of mutation	No	Yes	No	Yes	Yes	Yes	Yes
DNA Mutation	No	No	No	Yes	Yes	Yes	Yes
Type of mutation	Missense	Missense	Missense	All	All	Missense, Insertion, Deletion	Missense, Nonsense, Insertion, Deletion
rs ID number	Yes	Yes	Yes	No	Yes	No	Yes
UniProt ID	Yes	Yes	Yes	Partial	Partial	Yes	Yes
3D structure (PDB)	No	Yes	No	Yes	No	Yes	Yes
Disease class	No	No	No	Yes	No	Yes	Yes
Conservation score	No	No	No	No	No	Yes	Yes
Neighbouring residues	No	No	No	No	No	Yes	Yes
UniProt ID of Isoforms	No	No	Yes	No	No	Yes	Yes
Protein Domain	No	No	No	No	No	Yes	Yes
Protein Function	No	No	No	No	No	No	Yes
Subcellular Location	No	No	No	No	No	No	Yes
PTM*	No	No	No	No	No	No	Yes
Secondary structure	No	No	No	No	No	No	Yes
Organism	Human	Human	Human	Human	Human	Human (membrane proteins)	Human

* PTM—Post Translational Modifications

https://doi.org/10.1371/journal.pone.0210475.t001

mutations leading to disease. Prediction algorithms requiring a comprehensive variant dataset can make use of the vast dataset available with the HuVarBase. The database will be updated periodically. The updates will be on a quarterly basisand the update information will be reflected in the 'What's New' section of the web server.

Acknowledgments

We are grateful to the Bioinformatics facility, Department of Biotechnology and Indian Institute of Technology Madras for computational facilities.

Author Contributions

Conceptualization: M. Michael Gromiha.

Data curation: Kaliappan Ganesan, A. Kulandaisamy, S. Binny Priya.

Formal analysis: M. Michael Gromiha.

Funding acquisition: M. Michael Gromiha.

Investigation: M. Michael Gromiha.

Methodology: Kaliappan Ganesan, A. Kulandaisamy, S. Binny Priya, M. Michael Gromiha.

Project administration: M. Michael Gromiha.

Resources: M. Michael Gromiha.

Software: Kaliappan Ganesan.

Supervision: M. Michael Gromiha.

Validation: Kaliappan Ganesan, M. Michael Gromiha.

Writing – original draft: Kaliappan Ganesan.

Writing – review & editing: Kaliappan Ganesan, M. Michael Gromiha.

References

1. International Alport Mutation Consortium, Savige J, Ars E, Cotton RG, Crockett D, Dagher H, et al. DNA variant databases improve test accuracy and phenotype prediction in Alport syndrome. Pediatr Nephrol. 2014 Jun; 29(6):971–7. https://doi.org/10.1007/s00467-013-2486-8 Epub 2013 May 30. Review. PMID: 23720012.
2. Ritter DI, Roychowdhury S, Roy A, Rao S, Landrum MJ, Sonkin D, et al. ClinGen Somatic Cancer Working Group. Somatic cancer variant curation and harmonization through consensus minimum variant level data. Genome Med. 2016 Nov 4; 8(1):117. https://doi.org/10.1186/s13073-016-0367-z PMID: 27814769; PubMed Central PMCID: PMC5095986.
3. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015 Oct 1; 526(7571):68–74. https://doi.org/10.1038/nature15393 PMID: 26432245; PubMed Central PMCID: PMC4750478.
4. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017 Jan 4; 45(D1):D777–D783. https://doi.org/10.1093/nar/gkw1121 Epub 2016 Nov 28. PMID: 27899578; PubMed Central PMCID: PMC5210583.
5. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018 Jan 4; 46(D1):D1062–D1067. https://doi.org/10.1093/nar/gkx1153 PMID: 29165669; PubMed Central PMCID: PMC5753237.
6. Mottaz A, David FP, Veuthey AL, Yip YL. Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. Bioinformatics. 2010 Mar 15; 26(6):851–2. https://doi.org/10.1093/bioinformatics/btq028 Epub 2010 Jan 26. PMID: 20106818; PubMed Central PMCID: PMC2832822.
7. Li J, Shi L, Zhang K, Zhang Y, Hu S, Zhao T, et al. VarCards: an integrated genetic and clinical database for coding variants in the human genome. Nucleic Acids Res. 2018 Jan 4; 46(D1):D1039–D1048. https://doi.org/10.1093/nar/gkx1039 PMID: 29112736; PubMed Central PMCID: PMC5753295.
8. Thangam M, Gopal RK. CRCDA—Comprehensive resources for cancer NGS data analysis. Database (Oxford). 2015 Oct 8;2015. pii: bav092. https://doi.org/10.1093/database/bav092 Print 2015. PMID: 26450948; PubMed Central PMCID:PMC4597977.
9. Kulandaisamy A, Binny Priya S, Sakthivel R, Tarnovskaya S, Bizin I, Hönigschmid P, et al. MutHTP: mutations in human transmembrane proteins. Bioinformatics. 2018 Jul 1; 34(13):2325–2326. https://doi.org/10.1093/bioinformatics/bty054 PMID: 29401218.
10. Pundir S, Martin MJ, O'Donovan C. UniProt Protein Knowledgebase. Methods Mol Biol. 2017; 1558:41–55. https://doi.org/10.1007/978-1-4939-6783-4_2 PMID: 28150232; PubMed Central PMCID: PMC5565770.
11. Gnad F, Baucom A, Mukhyala K, Manning G, Zhang Z. Assessment of computational methods for predicting the effects of missense mutations in human cancers. BMC Genomics. 2013; 14 Suppl 3:S7. https://doi.org/10.1186/1471-2164-14-S3-S7 Epub 2013 May 28. PMID: 23819521; PubMed Central PMCID: PMC3665581.
12. Burley SK, Berman HM, Christie C, Duarte JM, Feng Z, Westbrook J, et al. RCSB Protein Data Bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. Protein Sci. 2018 Jan; 27(1):316–330. https://doi.org/10.1002/pro.3331 Epub 2017 Nov 11. Review. PMID: 29067736; PubMed Central PMCID: PMC5734314.
13. Velankar S, Dana JM, Jacobsen J, van Ginkel G, Gane PJ, Luo J, et al. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Res. 2013 Jan; 41(Database issue): D483–9. https://doi.org/10.1093/nar/gks1258 Epub 2012 Nov 29. PMID: 23203869;PubMed Central PMCID: PMC3531078.
14. Gaudet P, Michel PA, Zahn-Zabal M, Britan A, Cusin I, Domagalski M, et al. The neXtProt knowledgebase on human proteins: 2017 update. Nucleic Acids Res. 2017 Jan 4; 45(D1):D177–D182. https://doi.org/10.1093/nar/gkw1062 Epub 2016 Nov 29. PMID: 27899619; PubMed Central PMCID: PMC5210547.
15. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016 Jan 4; 44(D1):D279–85. https://doi.org/10.1093/nar/gkv1344 Epub 2015 Dec 15. PMID: 26673716; PubMed Central PMCID: PMC4702930.

16. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017 Jan 4; 45(D1):D353–D361. https://doi.org/10.1093/nar/gkw1092 Epub 2016 Nov 28. PMID: 27899662; PubMed Central PMCID: PMC5210567.

17. Lewis J, Snyder M, Hyatt-Knorr H. Marking 15 years of the Genetic and Rare Diseases Information Center. Transl Sci Rare Dis. 2017 May 25; 2(1–2):77–88. https://doi.org/10.3233/TRD-170011 PMID: 29152459; PubMed Central PMCID: C5685198.

18. Rubinstein WS, Maglott DR, Lee JM, Kattman BL, Malheiro AJ, Ovetsky M, et al. The NIH genetic testing registry: a new, centralized database of genetic tests to enable access to comprehensive information and improve transparency. Nucleic Acids Res. 2013 Jan; 41(Database issue):D925–35. https://doi.org/10.1093/nar/gks1173 Epub 2012 Nov 27. PMID: 23193275; PubMed CentralPMCID: PMC3531155.

19. Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Curr Protoc Bioinformatics. 2016 Jun 20; 54:1.30.1–1.30.33. https://doi.org/10.1002/cpbi.5 PMID: 27322403.

20. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan 1; 29(1):308–11. PMID: 11125122; PubMed Central PMCID: PMC29783.

21. Shen J, Song K, Slater AJ, Ferrero E, Nelson MR. STOPGAP: a database for systematic target opportunity assessment by genetic association predictions. Bioinformatics. 2017 Sep 1; 33(17):2784–2786. https://doi.org/10.1093/bioinformatics/btx274 PMID: 28472345.

22. Gosalia N, Economides AN, Dewey FE, Balasubramanian S. MAPPIN: a method for annotating, predicting pathogenicity and mode of inheritance for nonsynonymous variants. Nucleic Acids Res. 2017 Oct 13; 45(18):10393–10402. https://doi.org/10.1093/nar/gkx730 PMID: 28977528; PubMed Central PMCID: PMC5737764.

23. Singhal A, Simmons M, Lu Z. Text Mining Genotype-Phenotype Relationships from Biomedical Literature for Database Curation and Precision Medicine. PLoS Comput Biol. 2016 Nov 30; 12(11):e1005017. https://doi.org/10.1371/journal.pcbi.1005017 eCollection 2016 Nov. PMID: 27902695; PubMed Central PMCID: PMC5130168.