# Bibliome Variant Database: Automated Identification and Annotation of Genetic Variants in Primary Literature

Samuel W. Baker and Arupa Ganguly

## AFFILIATIONS

Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.

## CONTACT

samuel.baker@pennmedicine.upenn.edu

## ABSTRACT

The Bibliome Variant Database (BVdb) is a freely available reference database containing over 1 million human genetic variants mapped to the human genome that have been mined from primary literature. The BVdb is designed to facilitate variant interpretation in clinical and research contexts by reducing or eliminating the time required to search for literature describing a given variant. Users can search the database using gene symbols, HGVS variant nomenclature, genomic positions, or rsIDs. Each variant page lists references in the database that describe the variant, as well as the exact gene symbol and variant text description identified in each reference.

## AVAILABILITY AND IMPLEMENTATION

The BVdb is freely available at http://bibliome.ai

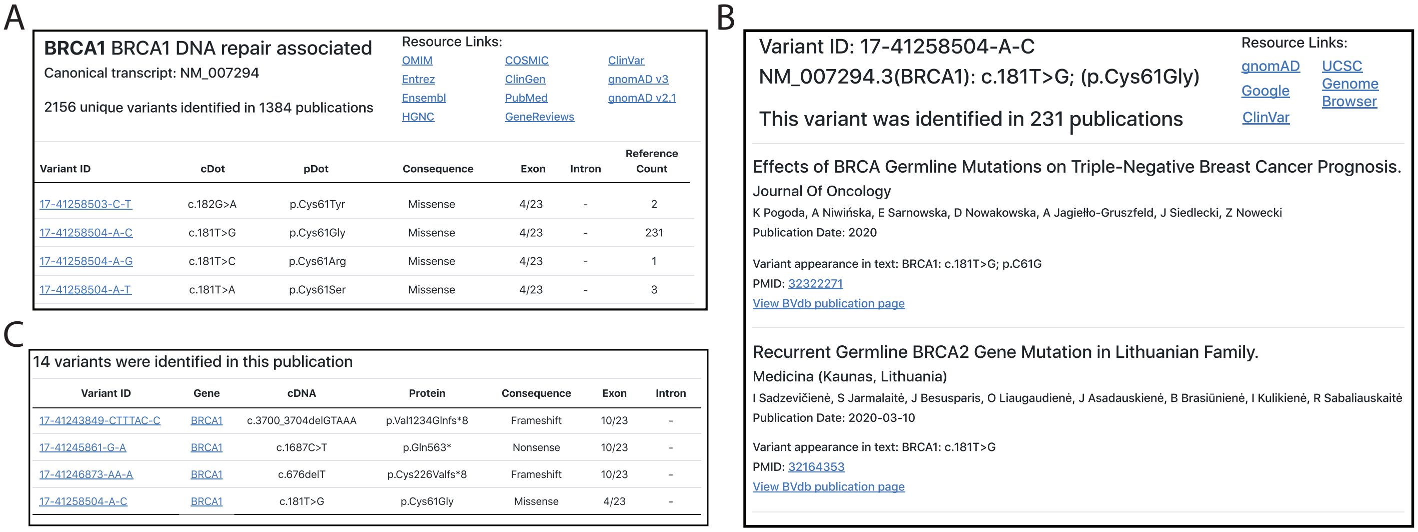
## INTRODUCTION

Clinical exome and genome sequencing are increasingly being used to diagnose rare genetic disease and in the treatment of cancer (Williams 2019). Such methods identify hundreds of genomic variants that must be evaluated to determine which are most relevant to a patient's disease (Retterer et al. 2016). In the majority of clinical laboratories, variant interpretation is a labor-intensive manual process and interpretation of each variant can require up to one hour or more of expert time (Dewey et al. 2014). Literature review plays a central role in variant interpretation and information gathered from primary sources describing a given variant is routinely used during application of germline and somatic variant classification guidelines (Richards et al. 2015; Li et al. 2017). Curated variant interpretation databases, such as ClinVar, Human Gene Mutation Database (HGMD), and Clinical Interpretation of Variants in Cancer (CIVIC), represent valuable resources linking genomic variants to primary literature, and their use enables more efficient variant interpretation by reducing or eliminating the time required to search for primary literature describing a given variant (Landrum et al. 2016; Stenson et al. 2017; Griffith et al. 2017). However, owing to the significant labor requirements of manual curation, such databases include only a proportion of all published variants (Stenson et al. 2017; Griffith et al. 2017).

Here, we present the Bibliome Variant Database (BVdb), a structured variant reference database generated through automated identification and annotation of genomic variants in open access literature. The BVdb is designed to facilitate variant interpretation in clinical, diagnostic, and research contexts by allowing rapid lookup of primary literature describing genetic variation across genes and genomic loci.

## MATERIALS AND METHODS

The BVdb consists of a text mining pipeline and website, both of which were written in Python. The PubMed Central Open Access Subset and Author Manuscript Collection, together containing over 3.6 million full-text publications, was downloaded from the NCBI FTP server on June 10th, 2020 (ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/). Additional open access publications referenced in the Unpaywall Database Snapshot were also downloaded (https://unpaywall.org/products/snapshot, last accessed June 10th, 2020). For each publication, regular expressions were used to identify rsIDs as well as text patterns corresponding to HGVS-compliant and commonly used HGVS-non-compliant variant nomenclature describing cDNA and protein sequence changes (den Dunnen et al. 2016). Each variant identified within a given publication was assigned to a gene symbol present in the same publication based on spatial, numeric, and contextual features within the text. Gene-assigned variants were added to the BVdb only if the cDNA and/or protein annotations could be mapped to both the canonical transcript, or transcript present in the text, of the assigned gene and the human genome version hg19. For incomplete variant annotations, i.e. cDNA annotations identified in the absence of a corresponding protein annotation and vise-versa, the absent cDNA or protein annotation was imputed and the combination of the identified and imputed variant annotations were added to the BVdb. Genomic coordinates for each gene-assigned variant were determined using the Transvar reverse annotation function, and variant annotations displayed in the BVdb browser were generated using VEP (Zhou et al. 2015; McLaren et al. 2016).



**BRCA1** BRCA1 DNA repair associated
Canonical transcript: NM_007294

2156 unique variants identified in 1384 publications

Resource Links:
OMIM COSMIC ClinVar
Entrez ClinGen gnomAD v3
Ensembl PubMed gnomAD v2.1
HGNC GeneReviews

| Variant ID | cDot | pDot | Consequence | Exon | Intron | Reference Count |
|---|---|---|---|---|---|---|
| 17-41258503-C-T | c.182G>A | p.Cys61Tyr | Missense | 4/23 | - | 2 |
| 17-41258504-A-C | c.181T>G | p.Cys61Gly | Missense | 4/23 | - | 231 |
| 17-41258504-A-G | c.181T>C | p.Cys61Arg | Missense | 4/23 | - | 1 |
| 17-41258504-A-T | c.181T>A | p.Cys61Ser | Missense | 4/23 | - | 3 |

14 variants were identified in this publication

| Variant ID | Gene | cDNA | Protein | Consequence | Exon | Intron |
|---|---|---|---|---|---|---|
| 17-41243849-CTTTAC-C | BRCA1 | c.3700_3704delGTAAA | p.Val1234Glnfs*8 | Frameshift | 10/23 | - |
| 17-41245861-G-A | BRCA1 | c.1687C>T | p.Gln563* | Nonsense | 10/23 | - |
| 17-41246873-AA-A | BRCA1 | c.676delT | p.Cys226Valfs*8 | Frameshift | 10/23 | - |
| 17-41258504-A-C | BRCA1 | c.181T>G | p.Cys61Gly | Missense | 4/23 | - |

Variant ID: 17-41258504-A-C
NM_007294.3(BRCA1): c.181T>G; (p.Cys61Gly)

This variant was identified in 231 publications

Resource Links:
gnomAD UCSC Genome Browser
Google
ClinVar

Effects of BRCA Germline Mutations on Triple-Negative Breast Cancer Prognosis.
Journal Of Oncology
K Pogoda, A Niwińska, E Sarnowska, D Nowakowska, A Jagiełło-Gruszfeld, J Siedlecki, Z Nowecki
Publication Date: 2020

Variant appearance in text: BRCA1: c.181T>G; p.C61G
PMID: 32322271
View BVdb publication page

Recurrent Germline BRCA2 Gene Mutation in Lithuanian Family.
Medicina (Kaunas, Lithuania)
I Sadzevičienė, S Jarmalaitė, J Besusparis, O Liaugaudienė, J Asadauskienė, B Brasiūnienė, I Kulikienė, R Sabaliauskaitė
Publication Date: 2020-03-10

Variant appearance in text: BRCA1: c.181T>G
PMID: 32164353
View BVdb publication page

**Figure 1.** (A) The BVdb gene view, listing results from for the *BRCA1* gene symbol search term. (B) The variant view, displaying information and references describing the *BRCA1* c.181T>G; (p.Cys61Gly) variant. (C). The publication view, listing a subset of all variants identified in (Pogoda et al. 2020).

## RESULTS AND DISCUSSION

The BVdb can be accessed at https://bibliome.ai using a desktop computer or mobile device. Contents of the BVdb are presented in three basic interconnected views, each of which is designed to provide information content in a format which facilitates variant interpretation and classification.

1) The gene view (Figure 1A) displays all variants in the BVdb assigned to a given gene and can be accessed by entering an HGNC-approved gene symbol in the search bar. Gene-specific information, including transcript used for annotation and links to external gene-specific resources, is located at the top of the page. Variants listed in the gene view are sorted by genomic position and information including cDNA annotation, protein annotation, predicted coding effect, exon/intron number, and number of references in which the variant was identified is displayed for each variant.

2) The variant view (Figure 1B) can be accessed either by clicking on a variant ID on the gene page or by entering a variant-specific search term in the search bar. Each variant page lists publications in which the variant was identified and provides links to variant-specific entries in external databases. Each publication entry contains information including, article title, author list, journal, publication date, and the specific variant representation identified in the article text (i.e. *KRAS* p.Gly12Asp versus *KRAS* p.G12D).

3) The publication view (Figure 1C) can be accessed by clicking on PMIDs in the variant view or by entering a PMID in the search bar. This view contains all publication-specific information displayed in the variant view, and a list of all variants identified within the publication. Clicking on a variant ID or gene symbol allows navigation to the variant and gene views, respectively.

Ideal use-cases of the BVdb include variant interpretation, curation, and systematic literature review in clinical and research contexts. Use of the BVdb to gather evidence can reduce the time required to identify publications describing a given variant, a necessary and often rate-limiting process in clinical laboratory variant interpretation workflows (Ravichandran et al. 2019). Additionally, use of the BVdb can promote variant classification harmonization by ensuring all users have access to the same open access literature during variant interpretation (Garber et al. 2016). Furthermore, the BVdb can also facilitate variant reanalysis, by allowing easy identification of newly published literature describing previously evaluated variants (Baker et al. 2019).

## CONCLUSION

The BVdb is purpose-built to facilitate variant interpretation and curation in clinical and research contexts. The BVdb is a powerful tool that enables rapid lookup of primary literature describing genetic variation. Finally, automation ensures the BVdb can keep pace with the increasing volume of published literature and can add references describing genetic variants shortly after they become available.

## REFERENCES

Baker, Samuel W., Jill R. Murrell, Addie I. Nesbitt, Kieran B. Pechter, Jorune Balciuniene, Xiaonan Zhao, Zhenming Yu, et al. 2019. “Automated Clinical Exome Reanalysis Reveals Novel Diagnoses.” *The Journal of Molecular Diagnostics: JMD* 21 (1): 38–48.

Dewey, Frederick E., Megan E. Grove, Cuiping Pan, Benjamin A. Goldstein, Jonathan A. Bernstein, Hassan Chaib, Jason D. Merker, et al. 2014. “Clinical Interpretation and Implications of Whole-Genome Sequencing.” *JAMA: The Journal of the American Medical Association* 311 (10): 1035–45.

Dunnen, Johan T. den, Raymond Dalgleish, Donna R. Maglott, Reece K. Hart, Marc S. Greenblatt, Jean McGowan-Jordan, Anne-Francoise Roux, Timothy Smith, Stylianos E. Antonarakis, and Peter E. M. Taschner. 2016. “HGVS Recommendations for the Description of Sequence Variants: 2016 Update.” *Human Mutation* 37 (6): 564–69.

Garber, Kathryn B., Lisa M. Vincent, John J. Alexander, Lora J. H. Bean, Sherri Bale, and Madhuri Hegde. 2016. “Reassessment of Genomic Sequence Variation to Harmonize Interpretation for Personalized Medicine.” *American Journal of Human Genetics* 99 (5): 1140–49.

Griffith, Malachi, Nicholas C. Spies, Kilannin Krysiak, Joshua F. McMichael, Adam C. Coffman, Arpad M. Danos, Benjamin J. Ainscough, et al. 2017. “CIViC Is a Community Knowledgebase for Expert Crowdsourcing the Clinical Interpretation of Variants in Cancer.” *Nature Genetics* 49 (2): 170–74.

Landrum, Melissa J., Jennifer M. Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, et al. 2016. “ClinVar: Public Archive of Interpretations of Clinically Relevant Variants.” *Nucleic Acids Research* 44 (D1): D862–68.

Li, Marilyn M., Michael Datto, Eric J. Duncavage, Shashikant Kulkarni, Neal I. Lindeman, Somak Roy, Apostolia M. Tsimberidou, et al. 2017. “Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists.” *The Journal of Molecular Diagnostics: JMD* 19 (1): 4–23.

McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. 2016. “The Ensembl Variant Effect Predictor.” *Genome Biology* 17 (1): 122.

Pogoda, Katarzyna, Anna Niwińska, Elżbieta Sarnowska, Dorota Nowakowska, Agnieszka Jagiełło-Gruszfeld, Janusz Siedlecki, and Zbigniew Nowecki. 2020. “Effects of BRCA Germline Mutations on Triple-Negative Breast Cancer Prognosis.” *Journal of Oncology* 2020 (January): 8545643.

Ravichandran, Vignesh, Zarina Shameer, Yelena Kemel, Michael Walsh, Karen Cadoo, Steven Lipkin, Diana Mandelker, et al. 2019. “Toward Automation of Germline Variant Curation in Clinicalcancer Genetics.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 21 (9): 2116–25.

Retterer, Kyle, Jane Juusola, Megan T. Cho, Patrik Vitazka, Francisca Millan, Federica Gibellini, Annette Vertino-Bell, et al. 2016. “Clinical Application of Whole-Exome Sequencing across Clinical Indications.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 18 (7): 696–704.

Richards, Sue, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W. Grody, 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.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 17 (5): 405–24.

Stenson, Peter D., Matthew Mort, Edward V. Ball, Katy Evans, Matthew Hayden, Sally Heywood, Michelle Hussain, Andrew D. Phillips, and David N. Cooper. 2017. “The Human Gene Mutation Database: Towards a Comprehensive Repository of Inherited Mutation Data for Medical Research, Genetic Diagnosis and next-Generation Sequencing Studies.” *Human Genetics* 136 (6): 665–77.

Williams, Marc S. 2019. “Early Lessons from the Implementation of Genomic Medicine Programs.” *Annual Review of Genomics and Human Genetics* 20 (August): 389–411.
Zhou, Wanding, Tenghui Chen, Zechen Chong, Mary A. Rohrdanz, James M. Melott, Chris Wakefield, Jia Zeng, et al. 2015. “TransVar: A Multilevel Variant Annotator for Precision Genomics.” *Nature Methods* 12 (11): 1002–3.