SOFTWARE TOOL ARTICLE

Providing gene-to-variant and variant-to-gene database identifier mappings to use with BridgeDb mapping services.

[version 1; peer review: 1 approved, 1 not approved]

Friederike Ehrhart ^{1,2}, Jonathan Melius¹, Elisa Cirillo¹, Martina Kutmon ^{1,3}, Egon L. Willighagen ¹, Susan L. Coort ¹, Leopold M.G. Curfs², Chris T. Evelo ¹⁻³

¹Department of Bioinformatics - BiGCaT, NUTRIM, Maastricht University, Maastricht, 6200 MD, The Netherlands

²GKC-Rett Expertise Centre, Maastricht University Medical Center, Maastricht, 6200 MD, The Netherlands

³Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, Maastricht, 6200 MD, The Netherlands

V1 First published: 03 Sep 2018, 7(ELIXIR):1390
<https://doi.org/10.12688/f1000research.15708.1>
 Latest published: 03 Sep 2018, 7(ELIXIR):1390
<https://doi.org/10.12688/f1000research.15708.1>

Abstract

Database identifier mapping services are important to make database information interoperable. BridgeDb offers such a service. Available mapping for BridgeDb link 1. genes and gene products identifiers, 2. metabolite identifiers and InChI structure description, and 3. identifiers for biochemical reactions and interactions between multiple resources that use such IDs while the mappings are obtained from multiple sources. In this study we created BridgeDb mapping databases for selections of genes-to-variants (and variants-to-genes) based on the variants described in Ensembl. Moreover, we demonstrated the use of these mappings in different software tools like R, PathVisio, Cytoscape and a local installation using Docker. The variant mapping databases are now described on the [BridgeDb website](#) and are available from the [BridgeDb mapping database repository](#) and updated according to the regular BridgeDb mapping update schedule.

Keywords

database identifier mapping, gene variant, BridgeDb, interoperability

Open Peer Review

Approval Status

	1	2
version 1		
03 Sep 2018	view	view
1. Patrice Godard , UCB (Union Chimique Belge), Brussels, Belgium		
2. Osman Ugur Sezer , Acibadem University, Istanbul, Turkey		
Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the [ELIXIR](#) gateway.

Corresponding author: Friederike Ehrhart (friederike.ehrhart@maastrichtuniversity.nl)

Author roles: **Ehrhart F:** Funding Acquisition, Investigation, Methodology, Project Administration, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Melius J:** Formal Analysis, Software, Writing – Review & Editing; **Cirillo E:** Conceptualization, Data Curation, Visualization, Writing – Original Draft Preparation; **Kutmon M:** Data Curation, Formal Analysis, Investigation, Supervision, Visualization, Writing – Review & Editing; **Willighagen EL:** Investigation, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Coort SL:** Conceptualization, Supervision, Writing – Review & Editing; **Curfs LMG:** Supervision, Writing – Review & Editing; **Evelo CT:** Conceptualization, Funding Acquisition, Project Administration, Resources, Supervision, Validation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was funded by ELIXIR, the European research infrastructure for life-science data
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2018 Ehrhart F *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Ehrhart F, Melius J, Cirillo E *et al.* **Providing gene-to-variant and variant-to-gene database identifier mappings to use with BridgeDb mapping services. [version 1; peer review: 1 approved, 1 not approved]** F1000Research 2018, 7 (ELIXIR):1390 <https://doi.org/10.12688/f1000research.15708.1>

First published: 03 Sep 2018, 7(ELIXIR):1390 <https://doi.org/10.12688/f1000research.15708.1>

Introduction

Many bioinformatics software tools rely on database identifier mapping, for instance for 1) recognition and mapping of identifiers used in experimental data to the corresponding identifiers present in secondary sources like pathways or ontology classes or 2) simply to combine data from different sources that use different identifiers. BridgeDb is a database identifier mapping tool that is available as a Java framework and as an installable web service ([van Iersel *et al.*, 2010](#)). Tools that integrate BridgeDb are for instance: the community curated pathway resource [WikiPathways](#) ([Slenter *et al.*, 2018](#)), the modular pathway editor and pathway analysis tool [PathVisio](#) ([Kutmon *et al.*, 2015](#)), and the network tool [Cytoscape](#) used to visualize, extend and evaluate biological networks. Depending on the available mappings BridgeDb can provide the mapping between identifiers from various data sources, also when these link to different molecular levels, *e.g.* gene to protein. BridgeDb can also be deployed as a web service. Moreover, it is available in a semantic web version, the Identifier Mapping Service (IMS), which can be used inside the Open PHACTS platform but can also be deployed from a software container ([Gray *et al.*, 2014](#)) ([link to tutorial](#) and [link to GitHub](#)). Mappings for BridgeDb are already available for gene products for many species (produced from the respective Ensembl genome annotations ([Aken *et al.*, 2017](#))), for metabolite identifiers (produced from HMDB ([Wishart *et al.*, 2013](#))) and ChEBI ([Hastings *et al.*, 2013](#))), and for reaction identifiers (produced from Rhea ([Morgat *et al.*, 2017](#))).

The BridgeDb mapping databases are linking pins between tools that support genetic variants, genes, and pathways analysis helping to visualize a complex biological context such those typical of the multifaceted (genetic) diseases. Gene-to-variant mapping was not yet available for BridgeDb. Such mappings can be especially useful to work with genetic variations, for instance when evaluating traits with a complicated genetic background like blood pressure, susceptibility to heart failure, or diabetes type 2 development. Single nucleotide polymorphism (SNP) can be responsible for phenotypic variations. In extreme cases this can be the cause of rare genetic disorders. For example, several SNPs in the human *DMD* gene can cause Duchenne muscular dystrophy (DMD), a severe congenital disorder which leads to severe physical impairment ([Magri *et al.*, 2011](#)). Since BridgeDb can stack mappings, the combination of the new gene-to-variant mapping database with the collection that was already available offers versatile mappings for variants to a large set of different human gene and gene product identifiers.

The main objective of this work was to provide mappings between gene identifiers and variant identifiers in both directions. The steps needed to achieve this were: 1) select the best source for the mappings, 2) collect data from the selected source, 3) annotate the result with provenance data about the process, the source, and the source version, and 4) finally to release the new BridgeDb mapping database and integrate that in the regular BridgeDb mapping database update schedule.

Target users for the resulting mappings are 1) bioinformaticians and developers, working on new approaches for data integration, if these use human genetic (variant) information; 2) members and users of ELIXIR data interoperability services, including the implementations in the tools mentioned that perform analyses based on human genetic variant data, for instance for the analysis of common multifaceted genetic diseases or in the rare disease field; and 3) researchers who access and query molecular data resulting from the analysis above.

Methods

The gene-to-variant database uses mappings between Ensembl and dbSNP ([Kitts *et al.*, 2013](#)). The Ensembl gene-to-dbSNP variant mappings present in Ensembl were used as the source. The released database is based on Ensembl r91, dbSNP b150, and the human genome assembly GRCh38. Although Ensembl provides more genetic variation from different sources, we focused on dbSNP as this variation database is regularly updated and adjusted to the actual Ensembl genome build. We compared both sources (Ensembl and dbSNP) and made sure that Ensembl provides all dbSNP available variants. So, we are able to rely only on the Ensembl API as a source for the extraction of the data necessary for creation of this mapping database. To prevent problems introduced by the user interfaces we used database dumps for this comparison.

The data dump was obtained from the Ensembl ftp server ([link for download](#)). For the first version, we used Ensembl 91, gene annotation with Gencode 27. The vcf (variant call format) file is the one relevant for our mapping. It contains the dbSNP identifier with its additional attributes and the associated Ensembl transcript identifier. By querying the Ensembl platform web service, we can access the gene identifier of the transcript. Combined, this leads to mappings between variants and genes. The size of the complete mapping database exceeded 150 Gb (for Ensembl 91), so we decided to create several different subsets: exonic variants, missense variants, protein truncating variants (PTV), PTV and missense variants, and variants with a PolyPhen score >0.908 indicating “Probably Damaging”. Other selections can be created easily on individual demand.

The created database contains the link between the Ensembl gene identifiers and the dbSNP variant identifiers including a selection of attributes (MAF (minor allele frequency), chromosome, variant alleles, and chromosome position start/end).

For the rare cases where a variation is associated to more than one gene, the variant is also associated to these genes in the BridgeDb database. For example, [rs199773918](#) overlaps in the exons of two genes ([ENSG00000173366](#) and [ENSG00000239732](#)), and in the exonic variant BridgeDb mapping both genes show up. Nevertheless, in our selection of variants it may happen that not all of them show up due to different variant effect classifications in the different genes. As an example, [rs199773918](#) is a variant that overlaps in the following genes: *TPR* ([ENSG00000047410](#)) and *PRG4* ([ENSG00000116690](#)). This variant is a “3’ prime UTR variant” of *TPR* and a “missense variant” of *PRG4*. It can be

found in both genes variant tables but due to our selection it will show up only once in the missense variant dataset.

Implementation

Database creation. An open-source Java program to create the gene-to-variant database is available on [GitHub](#). After downloading the vcf file from Ensembl, users create a configuration file with several parameters. Then the database creation program will parse the vcf file, retrieve additional information through the Ensembl web service and create the BridgeDb mapping database. Due to the large amount of mappings, the tool commits the mappings to the database in batches to keep the required memory low.

Operation. The database creation workflow is depicted in [Figure 1](#). The vcf file can be downloaded from the [Ensembl FTP](#). The “Homo_sapiens_incl_consequences.vcf.gz” file is used.

System requirements. The database creation tool runs with Java and requires more memory than usually given to a Java process. We advise users to allocate 3–4GB of memory at least when running the database creation tool (-Xmx4G).

Results

The resulting BridgeDb mapping databases are available as a Derby database from [here](#). The new mappings are available for all the BridgeDb implementations mentioned above (PathVisio,

Cytoscape, R package, web service, and the IMS). The mapping databases are freely available for download under CC-BY license. Application examples of the use of the variant BridgeDb database are given in the following section. We created gene-to-variant mapping databases for the variant classes given in [Table 1](#). Any other subset of variant classes can be created on demand using the tool described in the Methods section.

Use cases

To test and demonstrate the application of the variant BridgeDb database, we downloaded the database from BridgeDb. The gene-to-variant (and variant-to-gene) queries are shown in four different tools: R command line ([Team, 2014](#)), PathVisio ([Kutmon et al., 2015](#)), Cytoscape ([Shannon et al., 2003](#)) and the local IMS installation using Docker, in order to provide an overview of the flexibility of the mapping database in different environments. A genetic variant of the rare disease Duchenne muscular dystrophy (DMD) was selected from the gene-disease association database DisGeNET ([Piñero et al., 2017](#)). The rs104894790 ([Lenk et al., 1993](#)) SNP was chosen because it presented a high number of citations and a stop gain damaging effect on the gene’s protein product.

R

The SNP, rs104894790, as described above was used to query the Ensembl identifier for the gene(s) in which it is located (variant-to-gene query). The query was performed in R command

Figure 1. Database creation workflow. The gene-variant mapping database is built on the variant call format (vcf) file provided by Ensembl. After running the database creation tool, the database can be used in all the different use cases.

Table 1. Gene-to-variant mapping databases (status Ensembl 91, to be updated regularly).

SNP selection	File	Date	Size (zipped)
Exonic variants	SNP_r91_Exon.bridge.zip	2018-06-04	1.1Gb
Missense variants	SNP_r91_Missence.bridge.zip	2018-06-07	620Mb
Protein truncating variants	SNP_r91_PTV.bridge.zip	2018-06-07	75Mb
Protein truncating variants and missense:	SNP_r91_PTV_Missence.bridge.zip	2018-06-07	620Mb
All variants with a PolyPhen score > 0.908 indicating “Probably Damaging”	SNP_r91_PolyPhen.bridge.zip	2018-07-18	260Mb

line, after the installation of the BridgeDb R package ([link to BridgeDb R package](#)) (example R script in [Supplementary File 1](#)) (R version 3.5.1). The result shows that the variant is positioned only in one gene: dystrophin (*DMD*, [ENSG00000198947](#)). *DMD* is one of the largest genes in the human DNA (about 2.2 Mb), and is composed of 79 exons and has 32 known transcripts of which 20 are protein coding. Because the output is identifiers, it can be easily linked to other R packages such as mygene ([Mark et al., 2014](#)) which normally wraps around the mygene.info web service ([Xin et al., 2016](#)).

PathVisio

We used PathVisio (version 3.3.0) ([Figure 2](#)), a biological pathway analysis tool that allows drawing, editing and analyzing biological pathways, to demonstrate how the new gene-variant database can be used to evaluate variants in a pathway context. PathVisio, like Cytoscape, has the BridgeDb functionality integrated in the core. For the purpose of the demonstration, we first selected pathways that contain the *DMD* gene from the R example. Five pathways were found: two striated muscle contraction pathways (WP3795 and [WP383](#)), Ectoderm differentiation ([WP2858](#)), Extracellular matrix organization ([WP2703](#))

and Arrhythmogenic right ventricular cardiomyopathy ([WP2118](#)). In principle, a new PathVisio plugin could now be developed that searches pathways that contain genes with selected variants automatically, or the plugin could show all variants from an analysis sets on a given pathway. For the example, one of the striated muscle contraction pathways ([WP383](#)) was selected and visualized. Next, the BridgeDb variant database was loaded, using the [BridgeDbConfig](#) plugin. After selecting a gene in the pathway, the backpage tab of the right hand side panel now shows the list of hyperlinks obtained from the BridgeDb database that point to different information sources linked to the gene selected. [Figure 2](#) shows the backpage with the list of the 720 SNPs (from the BridgeDb with a PolyPhen score > 0.908, file `SNP_r91_PolyPhen.bridge`) for the selected *DMD* gene. All the SNPs in the backpage have a hyperlink to the corresponding dbSNP page.

Cytoscape

An alternative gene-to-variant visualization is provided using Cytoscape (version 3.6.1), a popular tool for (biological) network analysis and visualization ([Figure 3](#)). The BridgeDb app for Cytoscape is available here. A node with the Ensembl gene

Figure 2. PathVisio shows the diagram of the pathway WP383 from WikiPathways collection in the left panel of the tool. When the *DMD* gene is selected a list of hyperlinks from different sources are displayed in the back page of the left panel. In this case, a list of SNPs located in the gene is visualized.

Figure 3. Gene-variant network in Cytoscape. Using the BridgeDb app for Cytoscape, a gene-variant network for the *DMD* gene (blue rounded rectangle) and its 720 probably damaging variants (PolyPhen score > 0.908) was created. The node color of the variants represents the PolyPhen score as a gradient (white-red), the darker the red, the higher the PolyPhen score.

identifier of *DMD* was created and the 720 SNPs were mapped to the gene using the BridgeDb app interface. A gene-variant network was created using the list of variants mapped. Moreover, the app can be used to configure the selection of several attribute columns related to the variant nodes such as: chromosome location, minor allele frequency, and variant allele. In this example figure, we visualize the PolyPhen score as the node fill color of the variants. For simplicity, the rs-numbers are not displayed.

BridgeDb Identifier Mapping Service (IMS)

Finally, we here show that identifier mapping linking variants to genes and *vice versa* can also be done at a semantic web level, we here demonstrate how an online BridgeDb Identifier Mapping

Service (IMS) can be set up. The IMS technology was developed in the Open PHACTS project to link drug discovery related data sets, including a Docker image (Batchelor *et al.*, 2014; van Iersel *et al.*, 2010; Williams *et al.*, 2012). Here, identifier mappings are defined by link sets, which specify which identifiers are mapped. However, unlike traditional BridgeDb mapping files, these link sets also specify why the two identifiers are mapped, allowing them to be used as scientific lenses (Batchelor *et al.*, 2014).

Because the IMS works at a semantic web level, identifiers are represented by uniform resource identifiers (URIs). Moreover, the IMS is aware of URI equivalence defined by the MIRIAM registry (Juty *et al.*, 2012). This means that even when a mapping

file does not provide mappings for a certain URI, one would still get a number of equivalent URIs, following knowledge from MIRIAM database. And, when a single mapping is found in the link sets, equivalent URIs for the mapped URIs it returned. The IMS provide a *targetUriPattern* parameter allowing you to restrict the number of mapped URIs.

We developed a tutorial explaining how to set up an IMS instance with the variant-gene mappings (available from GitHub). The instance is run locally using a Docker container developed by Open PHACTS, which is available from [DockerHub](#). After the Docker image is started, it provides a web interface and an API. The web interface has a “Check Mapping for an URI” page where the URI can be given to be mapped, the return format (XML, JSON, or HTML), and optionally a *lensUri* (see (Batchelor *et al.*, 2014), and the aforementioned *targetUriPattern*).

However, it is more convenient to use this API from other tools, as demonstrated with a second R script (Supplementary File 1). This R script uses the curl (Ooms, 2017 link) and jsonlite (Ooms, 2014) packages to interact with the IMS. The first package is used to call the IMS webservice and the second to convert the returned JSON into a data model more easily handled in R. The example consists of two API calls: the first part finds 603 variants for the *DMD* gene (Ensembl ID ENSG00000198947); the second example takes a single variant (dbSNP ID rs769658853) and looks up the matching gene.

Discussion

The BridgeDb toolset provides several apps and tools designed for different purposes, while mapping databases are available to link different database IDs for genes and gene products, metabolites, and reactions and interactions. A mapping database in the BridgeDb software environment, capable of linking genes to their variants and *vice versa*, was not yet available. The new database is expected to be useful to enhance the biological interpretation of genetic variant data (as shown with the example of the *DMD* gene) for instance when using apps that evaluate biological pathways, use the classification of genes according to ontology terms, or in the R environment when performing gene and variant related statistical evaluation.

With this newly created mapping database and the transitivity function of BridgeDb, the user can map between three different

layers: e.g. variant-gene-protein. This approach can support multi-omics analysis for various biomedical applications, and tools like Cytoscape and PathVisio can be used immediately to benefit from this.

We intend to keep the content up-to-date by regular updates. The human variant mapping database is already incorporated into the quarterly BridgeDb mapping database update. Also other variant sets including more than only the currently included protein truncating and missense variants can be created on user community (or individual) demand.

Data availability

The new gene-to-variant mapping databases are available here: http://bridgedb.org/data/gene_database/

Available under a Apache 2.0 licence (<http://www.apache.org/licenses/LICENSE-2.0.html>)

Software availability

Source code for making of the mapping databases is available from GitHub: <https://github.com/BiGCAT-UM/BridgeDbVariantDatabase>

Archived source code at time of publication: <http://doi.org/10.5281/zenodo.1326514> (Willighagen & Melius, 2018)

License: Apache 2.0

Competing interests

No competing interests were disclosed.

Grant information

This work was funded by ELIXIR, the European research infrastructure for life-science data.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

The authors would like to thank the BridgeDb development team. This work heavily leaned on previous work done by the dbNP and Ensembl teams who curated the actual mappings and on the original BridgeDb development team, especially Martijn van Iersel.

Supplementary material

Supplementary File 1 – R code and instructions for setting up the BridgeDb IMS docker.

[Click here to access the data.](#)

References

- Aken BL, Achuthan P, Akanni W, *et al.*: **Ensembl 2017**. *Nucleic Acids Res.* 2017; **45**(D1): D635–D642.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Batchelor C, Breninkmeijer CYA, Chichester C, *et al.*: **Scientific Lenses to Support Multiple Views over Linked Chemistry Data**. *International Semantic Web Conference.* 2014; **2014**: 98–113.
[Publisher Full Text](#)
- Gray AJ, Groth P, Loizou A, *et al.*: **Applying linked data approaches to pharmacology: Architectural decisions and implementation**. *Semant Web.* 2014; **5**(2): 101–113.
[Publisher Full Text](#)
- Hastings J, de Matos P, Dekker A, *et al.*: **The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013**. *Nucleic Acids Res.* 2013; **41**(Database issue): D456–63.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Juty N, Le Novère N, Laibe C: **Identifiers.org and MIRIAM Registry: community resources to provide persistent identification**. *Nucleic Acids Res.* 2012; **40**(Database issue): D580–6.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kitts A, Phan L, Ward M, *et al.*: **The NCBI Handbook - The Database of Short Genetic Variation (dbSNP)**. 2 ed. Bethesda (MD): National Center for Biotechnology Information (US). 2013.
[Reference Source](#)
- Kutmon M, van Iersel MP, Bohler A, *et al.*: **PathVisio 3: an extendable pathway analysis toolbox**. *PLoS Comput Biol.* 2015; **11**(2): e1004085.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Lenk U, Hanke R, Kraft U, *et al.*: **Non-isotopic analysis of single strand conformation polymorphism (SSCP) in the exon 13 region of the human dystrophin gene**. *J Med Genet.* 1993; **30**(11): 951–4.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Magri F, Govoni A, D'Angelo MG, *et al.*: **Genotype and phenotype characterization in a large dystrophinopathic cohort with extended follow-up**. *J Neurol.* 2011; **258**(9): 1610–23.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Mark A, Thompson R, Afrasiabi C, *et al.*: **Access MyGene.Info_ services**. *Bioconductor.* 2014.
[Publisher Full Text](#)
- Morgat A, Lombardot T, Axelsen KB, *et al.*: **Updates in Rhea - an expert curated resource of biochemical reactions**. *Nucleic Acids Res.* 2017; **45**(D1): D415–D418.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ooms J: **The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects**. arXiv. 2014.
[Reference Source](#)
- Piñero J, Bravo À, Queralt-Rosinach N, *et al.*: **DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants**. *Nucleic Acids Res.* 2017; **45**(D1): D833–D839.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Shannon P, Markiel A, Ozier O, *et al.*: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Res.* 2003; **13**(11): 2498–504.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Slenter DN, Kutmon M, Hanspers K, *et al.*: **WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research**. *Nucleic Acids Res.* 2018; **46**(D1): D661–D667.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Team RC: **R: A language and environment for statistical computing**. R Foundation for Statistical Computing. 2014.
- van Iersel MP, Pico AR, Kelder T, *et al.*: **The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services**. *BMC Bioinformatics.* 2010; **11**: 5.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Williams AJ, Harland L, Groth P, *et al.*: **Open PHACTS: semantic interoperability for drug discovery**. *Drug Discov Today.* 2012; **17**(21–22): 1188–98.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Willighagen E, Melius J: **BIGCAT-UM/BridgeDbVariantDatabase: Gene-Variant database builder**. *Zenodo.* 2018.
<http://www.doi.org/10.5281/zenodo.1326514>
- Wishart DS, Jewison T, Guo AC, *et al.*: **HMDB 3.0--The Human Metabolome Database in 2013**. *Nucleic Acids Res.* 2013; **41**(Database issue): D801–7.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Xin J, Mark A, Afrasiabi C, *et al.*: **High-performance web services for querying gene and variant annotation**. *Genome Biol.* 2016; **17**(1): 91.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 01 November 2018

<https://doi.org/10.5256/f1000research.17143.r39322>

© 2018 Ugur Sezerman O. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Osman Ugur Sezerman

Acibadem University, Istanbul, Turkey

Database identifier mapping services are necessary to make the information interoperable to be able to link to other resources. In the present work tbridgeDB added a new feature enabling mapping databases for genes-to variants and vice versa for the variants described in Ensemble.

Implementation stages are explained in detail making the work reproducible. The use case scenario clearly demonstrates the value added from the service. The work is certainly scientifically sound.

I have two points to be addressed:

1. How do they handle the variants in case of splice variants of the same gene are present
2. It is difficult to query and search the available gene to variant mapping databases

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 22 October 2018

<https://doi.org/10.5256/f1000research.17143.r39323>

© 2018 Godard P. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Patrice Godard

UCB (Union Chimique Belge), Brussels, Belgium

This article describes the use of the BridgeDb framework and associated tools to map gene and gene variant identifiers. The authors produced and made available 5 ready to use mapping databases focused on different categories of human gene variants extracted from the Ensembl database (version 91). They also document in the article the use of these databases in 4 different environments. As stated by the authors BridgeDb is integrated or used by different tools or resources such as wikipathway and pathvision. In this context, it makes sense to enrich the BridgeDB ecosystem with additional mapping databases such as those produced by the authors and focused on gene/gene variant associations, and it makes sense to publish an article describing the new available features.

However there are fundamental flaws in the article that seriously undermine the findings and conclusions:

- The different databases which are provided are not described sufficiently. The point here is about informing the user about the level of exhaustivity (relative to the original data source) he can expect when performing a mapping between genes and gene variants.
 - What is the number of SNPs and genes covered by each of them?
 - What were the formal criteria used to include a SNP in one or the other database? (They can be found in the config file but they are not described and the list of other possible options is not given)
 - The authors do not report if and how they check the content of the databases they produced according to the original data source (Ensembl 91).
 - What is the overlap between the different bridge databases?
 - The selection of the database for mapping gene identifiers to gene variants can be done according to variant criteria under focus But how to select the relevant database to achieve the opposite task: finding a gene associated to a gene variant?
 - This concern is exemplified by the IMS use case provided by the authors: the number

of variants found associated to the ENSG00000198947 gene using the IMS method (603) is smaller than the number of variants found with the methods relying on the PolyPhen bridge database (720). An greater, or at least an equal, number of SNPs was expected since no specific database is specified in the IMS query. Does it rely on the same information than available in bridge databases?

- The authors claim that the Bridge databases they provide contain a selection of attributes that can be retrieved. I could only test this feature using the cytoscape BridgeDb app, trying to reproduce the Figure 3. I could install the SNP_r91_PolyPhen.bridge database and use it to retrieve 720 variants associated to the ENSG00000198947 gene. Then, I tried to get all the attributes for the 720 SNPs.
 - The query took more than 3 hours to run (Processor: Intel i5-6300U 2.40Ghz; RAM: 8GB; OS: Windows 10 Enterprise 64-bit; Cytoscape 3.6.1; BridgeDb app 1.1.0.2). Such long runtime should be mentioned in the article.
 - The Polyphen score was not available among the listed attributes which prevented me to reproduce the figure 3. Also, the Polyphen scores are reported by transcript in Ensembl and the authors do not document how it is recorded by gene in the Bridge database: do they take the average, the maximum or the minimum score?
 - Moreover I got only empty arrays for the MAF attribute (no value).
 - Only one allele was returned as Variant Alleles attribute for each SNP.

In addition to these flaws, other major issues need to be addressed:

- Introduction
 - The authors do not cite services/tools already available for finding SNP/gene cross-references. Among possible candidates: Ensembl BioMart (<https://www.ensembl.org/biomart>) and MyVariant (<https://myvariant.info/>). The authors should explain in the introduction why they develop a new resource and describe in the discussion the advantage or the differentiating features of their solution.
 - The authors list 3 categories of users but they do not describe their needs and how those needs would be fulfilled by their solution (this could be part of the discussion).
- Method
 - In the introduction the authors mention the selection of the best source for mappings as the first step to build the bridge database. However, they selected the dbSNP information provided by Ensembl without explaining why they made this choice. What were the criteria used to define this resource as the best one? Why not use the files provided by dbSNP directly?
 - As mentioned above, the author claim the following attributes are available for the SNPs: MAF, chromosome and chromosome position and variant alleles. Beside the flaws identified above, these attributes should be described more precisely: In which population the MAF has been measured? What are the ancestral and minor alleles of each SNP? What is the genome version used for chromosome position.
- Use cases
 - I could download and use the databases to reproduce the code provided by the authors. However, the SNP provided as an example in the script is not the one described in the article (rs5927022 in the script and rs104894790). It is annoying since the SNP provided in the script could not be found in any of the 5 bridge databases (it's an intronic SNP actually: http://www.ensembl.org/Homo_sapiens/Variation/Explore?db=core;r=3:52224867-

52225867;source=dbSNP;v=rs5927022;vdb=variation;vf=15937754). Also it was not easy to find in which database the variant described in the article was available. I've tried all of them and could find this variant in the "SNP_r91_Exon.bridge" and "SNP_r91_PTV.bridge" databases. The script should be in accordance with the article. The authors should also provide a strategy to identify the relevant database to map a SNP or several SNPs to a gene (as mentioned above).

- Is it possible to get SNP attributes from the R interface as it is from cytoscape?

Finally minor issues could also be addressed to improve the quality of the article

- Method
 - The authors write that they are able to rely on Ensembl API. But they've used files downloaded from Ensembl site and not the API. This sentence should be modified accordingly.
 - The author mention problems introduced by Ensembl user interfaces. What are these problems?
- Implementation
 - How long does it take to create each bridge database? Why not creating a complete database with more attribute for variant annotation?
 - The vcf file mentioned in the article is not available anymore. It has been split by chromosome since the Ensembl release 93. Is the database creation workflow compatible with this new organization of the original files?
 - The figure 1 is not very informative. It does not describe the database creation workflow which is only a box in this figure. I think it would be more informative to focus on this box and to explain what are the different steps in this box. Indeed, according to the information found on github, it seems that there are 2 java programs "VariantReader" and "VariantCreator" which are called sequentially in order to produce the database.
- Results
 - The dates in table 1 are misleading. They probably refer to the date of the database creation in June and July 2018. However the Ensembl version used as data source is from December 2017/April 2018. The authors should clarify this point in the table legend.
- Use cases
 - It would be very useful to add the attributes of the SNPs in the PathVisio backpage in addition to the hyperlinks.
 - Being able to access SNP information from cytoscape is a nice feature. However I don't think that the use case provided by the author is very relevant. Indeed I don't know how a network with 1 gene linked to 720 variants can be used or interpreted as such (In this case a table with all the variants related to the gene and their attributes should be sufficient). Maybe, an example of a network with more gene would be more interesting.
 - The link to the cytoscape app is missing.
 - The 3 first paragraphs of the "BridgeDb identifier Mapping Service (IMS)" should go in Methods.

Is the rationale for developing the new software tool clearly explained?

Partly

Is the description of the software tool technically sound?

Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

No

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bioinformatics, genomics, genetics, R

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research