Check for updates

SOFTWARE TOOL ARTICLE

# REVISED GeneDB and Wikidata [version 2; peer review: 2 approved]

Magnus Manske, Ulrike Böhme, Christoph Püthe, Matt Berriman

Parasites and Microbes, Wellcome Trust Sanger Institute, Cambridge, CB10 1SA, UK

**v2** **First published:** 01 Aug 2019, **4**:114 (https://doi.org/10.12688/wellcomeopenres.15355.1)

**Latest published:** 14 Oct 2019, **4**:114 (https://doi.org/10.12688/wellcomeopenres.15355.2)

## Abstract

Publishing authoritative genomic annotation data, keeping it up to date, linking it to related information, and allowing community annotation is difficult and hard to support with limited resources.
Here, we show how importing GeneDB annotation data into Wikidata allows for leveraging existing resources, integrating volunteer and scientific communities, and enriching the original information.

## Keywords

GeneDB, Wikidata, MediaWiki, Wikibase, genome, reference, annotation, curation

**Open Peer Review**

**Reviewer Status** ✓ ✓

| | Invited Reviewers | |
|---|---|---|
| | **1** | **2** |
| REVISED<br>**version 2**<br>published<br>14 Oct 2019 | ✓<br>report | ✓<br>report |
| **version 1**<br>published<br>01 Aug 2019 | ✓<br>report | ?<br>report |

1 **Sebastian Burgstaller-Muehlbacher**, Medical University of Vienna, Vienna, Austria

2 **Andra Waagmeester**, Micelio, Antwerp, Belgium

Any reports and responses or comments on the article can be found at the end of the article.

**Corresponding author:** Magnus Manske (mm6@sanger.ac.uk)

**Author roles: Manske M**: Conceptualization, Methodology, Software, Writing – Original Draft Preparation; **Böhme U**: Data Curation, Validation, Writing – Review & Editing; **Püthe C**: Project Administration, Writing – Review & Editing; **Berriman M**: Resources, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** This work was supported by the Wellcome Trust through a Biomedical Resources Grant to MB [108443].

**Copyright:** © 2019 Manske M *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:** Manske M, Böhme U, Püthe C and Berriman M. **GeneDB and Wikidata [version 2; peer review: 2 approved]** Wellcome Open Research 2019, **4**:114 (https://doi.org/10.12688/wellcomeopenres.15355.2)

**First published:** 01 Aug 2019, **4**:114 (https://doi.org/10.12688/wellcomeopenres.15355.1)

**REVISED** **Amendments from Version 1**

We added/clarified text about Apollo, long-term sustainability, data license, and Rfam, as requested by reviewers. We also added the requested reference for Putman *et al.* (2017) and updated the Burgstaller-Muehlbacher citation.

**Any further responses from the reviewers can be found at the end of the article**

## Introduction

The GeneDB website has presented genome annotation data from eukaryotic and prokaryotic pathogens[1] sequenced by the Wellcome Sanger Institute for more than 15 years. The underlying data are stored in a database, designed using the Chado[2] schema. The project was established to display genomes sequenced and annotated by the former Pathogen Sequencing Unit at the Sanger Institute, but over time the usage has changed. Now, genomes are stored and displayed if they are undergoing some level of curation or ongoing improvement. The site provides a way for curators and researchers to see changes to annotation long before those changes are integrated with other data types in a number of collaborating databases. To reflect the change of usage, where the website is often not the primary access point for many users, GeneDB has recently undergone a redesign and simplification[3]. In particular, the web-based genome annotation tool Apollo[4] has been adopted as a major entry point for viewing genome data. While this delivers a structured, multi-track view of the genome and annotated genomic features (genes, ncRNAs, etc), the current version of Apollo has a limited capability for displaying the rich functional descriptions of individual genes that were a major feature of the previous GeneDB website.

Wikidata is a collaboratively edited, machine-readable and -writable knowledge base hosted by the Wikimedia Foundation, which also runs the collaboratively edited encyclopedia Wikipedia. Wikipedia has become the most accessed online encyclopedia and is unique in both its open, community-based editing, and a first port-of-call for public access to curated knowledge. Several bioinformatics projects make use of Wikipedia. The most successful of these is the Rfam project, where Wikipedia has been used to successfully manage free-text descriptions of RNA families[5] for over a decade. The Rfam-associated journal requires authors of new RNA families to create the matching Wikipedia page, tightly integrating Wikipedia into an entire field of research.

Wikidata currently contains 55 million items, which represent a superset of all Wikipedia article topics in over 300 languages, including biographical items, locations, species, artworks, scientific publications, etc. Amongst these items, Wikidata already stores human and mouse genes and proteins, as part of the Gene Wiki project[6], which originally started on Wikipedia[7], and many prokaryotic genes, as part of the WikiGenome project[8].

Wikidata offers various application programming interfaces (APIs) to read or write information in an automated way, including a query service using SPARQL, a query language for data on the Semantic Web[9]. All these services are freely accessible by third-party users.

In the present study, we describe how we have exported the contents of GeneDB into Wikidata to ensure the long term sustainability of high value curated information and to make the annotated gene and protein information available to a wider audience. Within Wikidata, potentially anyone can contribute to the annotation, for instance by adding further external cross-references to third-party databases, linking gene and proteins to the scientific literature, or even short free-text descriptions. These community changes can be detected, checked, and, in appropriate cases, imported back into GeneDB.

We also describe utilising the Wikidata APIs to create a new version of the GeneDB website with content created solely based on Wikidata items. The design of the new GeneDB website closely mirrors the old one but now provides continuity and stability for incoming links from other websites. Furthermore, by building the site from Wikidata components, the new GeneDB website benefits from additional information and queries harvested from Wikidata.

## Methods

### Importing GeneDB into Wikidata

GeneDB exports its Chado database monthly into publicly accessible files (ftp://ftp.sanger.ac.uk/pub/genedb/releases/) in general feature format version 3 (GFF3) and gene association file (GAF) format.

These files are regularly parsed by bespoke code to create or update Wikidata items, for both genes and their protein products.

This includes the addition of GO terms, as well as the creation and usage of Wikidata items about the scientific publications containing the respective findings.

An item about a gene (example: https://www.wikidata.org/wiki/Q19044775) or a protein on Wikidata consists of labels and aliases, descriptions, and a list of statements. Each statement is comprised of the following: a property from a community-controlled vocabulary (e.g., “chromosome”, “found in species”, “GeneDB ID”); a value, usually a plain string or a link to another Wikidata item, but also a date, a location, or a number, depending on the property; an optional list of qualifiers; and an optional list of references.

When updating an item, elements are added, altered, or removed on Wikidata if the current GeneDB information is different, and GeneDB is the authoritative source. All other elements of the Wikidata item remain unchanged during updates. Updates are performed automatically, utilizing the publicly exported GFF and GAF files based on Chado.

### Importing community changes from Wikidata into GeneDB

Community contributions on Wikidata can be divided into two parts. One part is the mass edit of items, by either bots (software-based robots that perform automated editing) or mass-editing tools. The other part are individual, usually manual edits, of low volume.

Only some edits are directly relevant to GeneDB; a new description of a protein in Dutch will not be imported into the Chado database, and can therefore be ignored. Likewise, the addition of external identifiers to sources not tracked by GeneDB can be ignored.

Individual edits that are both relevant to GeneDB, and not done by a Wikidata user on a "whitelist" of known, trusted users are summarized daily by an automated script, and sent to the GeneDB ticketing system for manual inspection. These changes are either ignored, reverted on Wikidata (e.g., vandalism), or imported into the GeneDB Chado database. The volume of such edits is quite low (~1/week) at this time, though we expect this to pick up with more members of the scientific community becoming aware of this venue into Wikidata.

## Implementation

The code to import and update Wikidata items was written in Rust[10] (rustc 1.36.0), using (amongst others) the rust-bio crate[11] for GFF and GAF reading. Rust was chosen as a language for its speed, security, low resource consumption, and available crates (libraries) to build on. Some of these crates, such as MediaWiki and Wikibase (Wikidata) API handling, were (co-)developed by the corresponding author independently. See software availability for source code[12,13]

## Operation

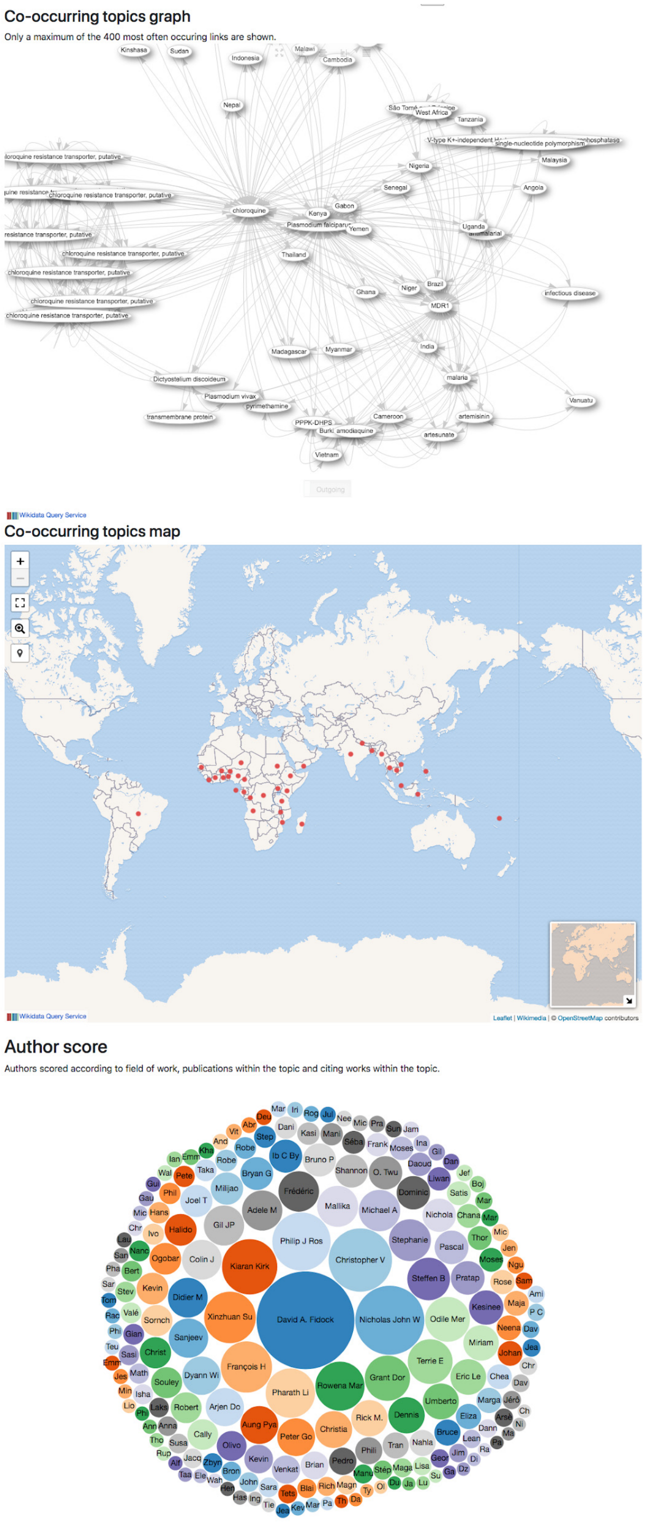
The Wikidata import code will run on any platform that Rust can be compiled for. Additional requirements are an internet connection, and login information for a Wikidata bot account.

The website operates client-side using JavaScript, and can be deployed on any standard web server. Besides Wikidata, no additional server-side support is required.

## Results/Use cases

At this time (26/07/2019), 409,219 Wikidata items for genes (including 9,031 pseudogenes), and an additional 397,979 items for proteins, have been created (http://tinyurl.com/y6wpfyrn), covering 45 taxa.

By exporting from GeneDB (Chado) into Wikidata, the data are intricately integrated into its ecosystem creating new functions with minimal project-specific development; for instance, links to and between publications, which in turn link to authors, institutions, etc. These connections between items allow for complex queries that were not possible before (Figure 1).



Co-occurring topics graph

Only a maximum of the 400 most often occuring links are shown.

Co-occurring topics map

Author score

Authors scored according to field of work, publications within the topic and citing works within the topic.

**Figure 1. Part of the Scholia tool website about the PfCRT gene (https://tools.wmflabs.org/scholia/topic/Q19044775).** These figures show connected topics, connected locations on a map, and the most prolific authors for this topic, respectively. All figures are generated live from Wikidata via its various APIs. All this information, and more, results from linking scientific publications to the genes in Wikidata.

## Gene pages

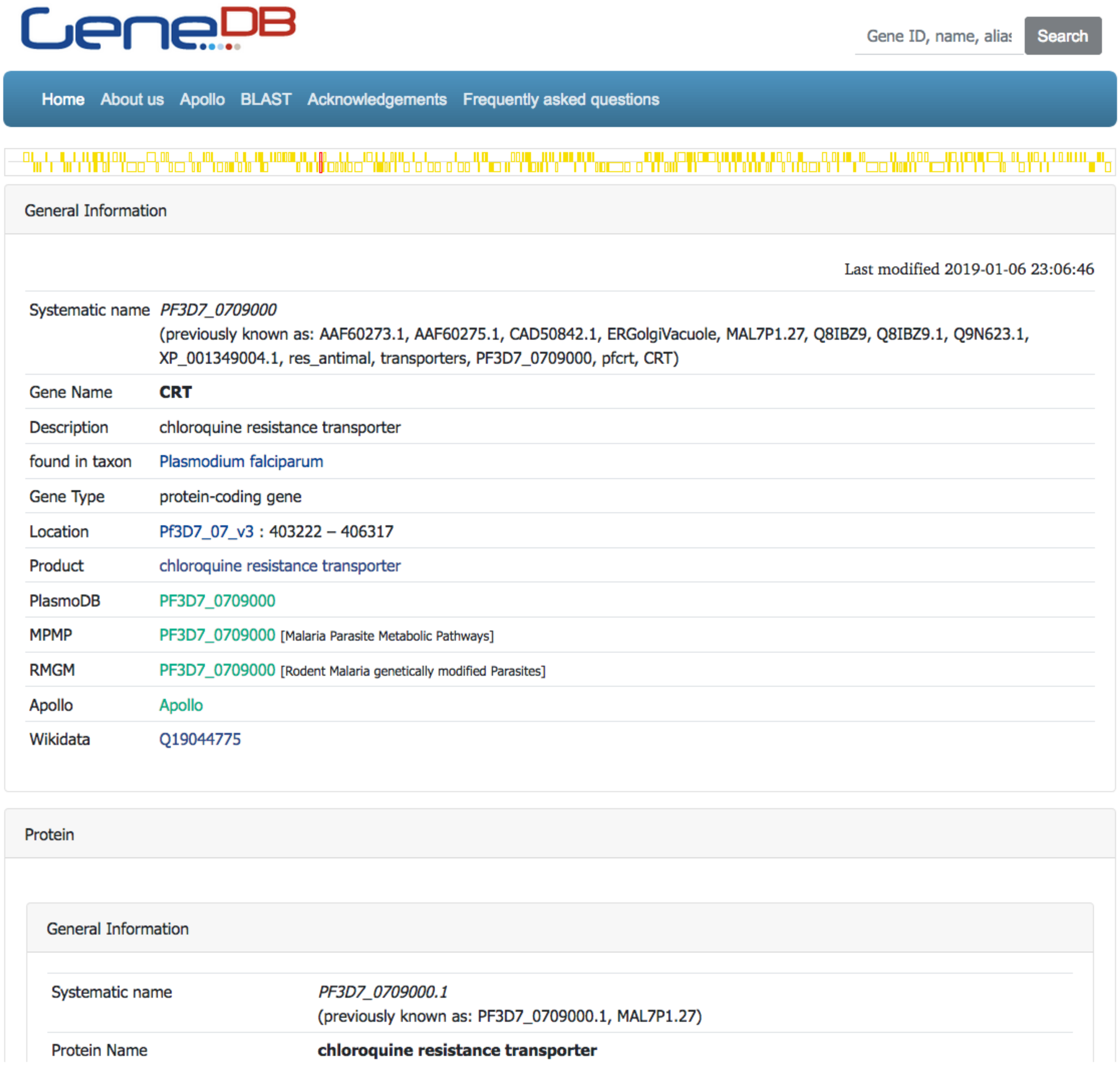
To replicate the way genes were represented on the previous version of the GeneDB website, a pure HTML/JavaScript site using vue.js was created. JavaScript components are written as modules. All pages and components are designed to work on both desktop and mobile. Web pages for genes, proteins, species, chromosomes, GO term[14] queries, and searches are generated on-the-fly utilising the Wikidata API and SPARQL interface, and Wikidata serves as the only back-end for these pages.

Each gene item, and its associated protein item(s), can be viewed on a separate page (Figure 2), that is rendered on-the-fly. This rendering includes a map of the gene on the chromosome, names, IDs, descriptions, links to other web resources (both from Wikidata statements, and auto-generated based on species and gene ID), a link to the Apollo browser view of the gene, a list of known orthologs.

Below that, each protein encoded by the gene is listed, as well as the known GO term ontology, complete with evidence and publication links, where available (Figure 3).

If Wikidata contains items about publications that have the gene or protein as a “main subject”, these publications are listed at the bottom of the page. This is an example of additional, on-topic information that Wikidata provides on top of the GeneDB dataset.

Gene/protein pages link out to other web-based databases via Wikidata-stored (e.g., UniProt) or computed (e.g., PlasmoDB)



General Information

Last modified 2019-01-06 23:06:46

| | |
|---|---|
| Systematic name | *PF3D7_0709000*<br>(previously known as: AAF60273.1, AAF60275.1, CAD50842.1, ERGolgiVacuole, MAL7P1.27, Q8IBZ9, Q8IBZ9.1, Q9N623.1, XP_001349004.1, res_antimal, transporters, PF3D7_0709000, pfcrt, CRT) |
| Gene Name | **CRT** |
| Description | chloroquine resistance transporter |
| found in taxon | Plasmodium falciparum |
| Gene Type | protein-coding gene |
| Location | Pf3D7_07_v3 : 403222 – 406317 |
| Product | chloroquine resistance transporter |
| PlasmoDB | PF3D7_0709000 |
| MPMP | PF3D7_0709000 [Malaria Parasite Metabolic Pathways] |
| RMGM | PF3D7_0709000 [Rodent Malaria genetically modified Parasites] |
| Apollo | Apollo |
| Wikidata | Q19044775 |

Protein

General Information

| | |
|---|---|
| Systematic name | *PF3D7_0709000.1*<br>(previously known as: PF3D7_0709000.1, MAL7P1.27) |
| Protein Name | **chloroquine resistance transporter** |

**Figure 2. Gene information on GeneDB (https://www.genedb.org/gene/PF3D7_0709000).** All information comes exclusively from Wikidata.

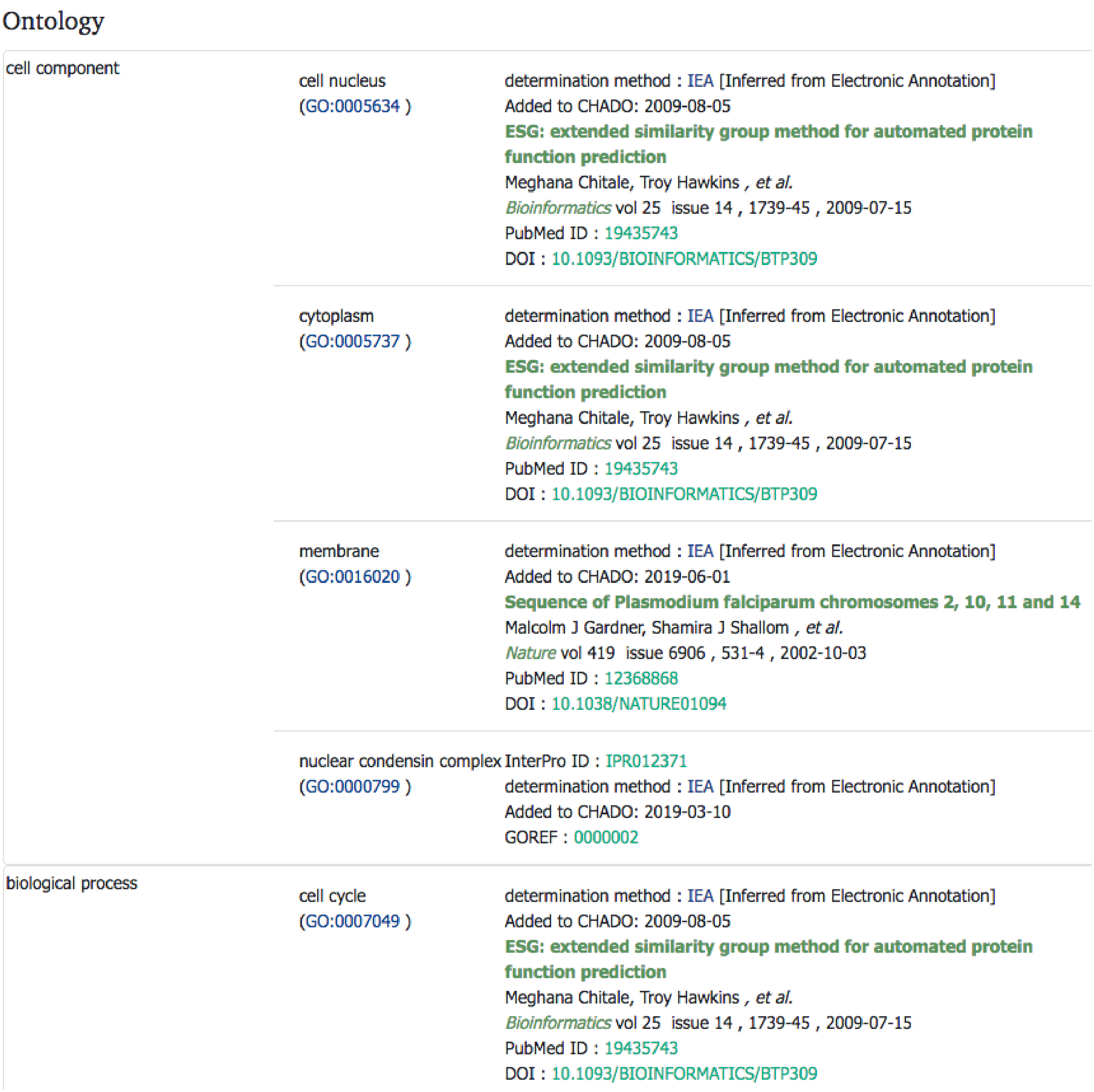
URLs. GO terms show supporting information, including citations of, and links to, the original publications. Also, a list of all publications on Wikidata with the respective gene as a subject is available on both GeneDB (example: https://www.genedb.org/gene/PF3D7_1200600) and the Scholia tool (example: https://tools.wmflabs.org/scholia/topic/Q18971176).

### Other functionality

The GeneDB search function utilizes Wikidata search, letting users find genes by name, alias, ID, and related information, across all covered species. The search will only return genes on Wikidata with a GeneDB identifier.

Each species in GeneDB has its own page (Figure 4), showing basic information about the species, links to other web resources, to the Apollo browser, and a list of chromosomes linking to the genes located there.

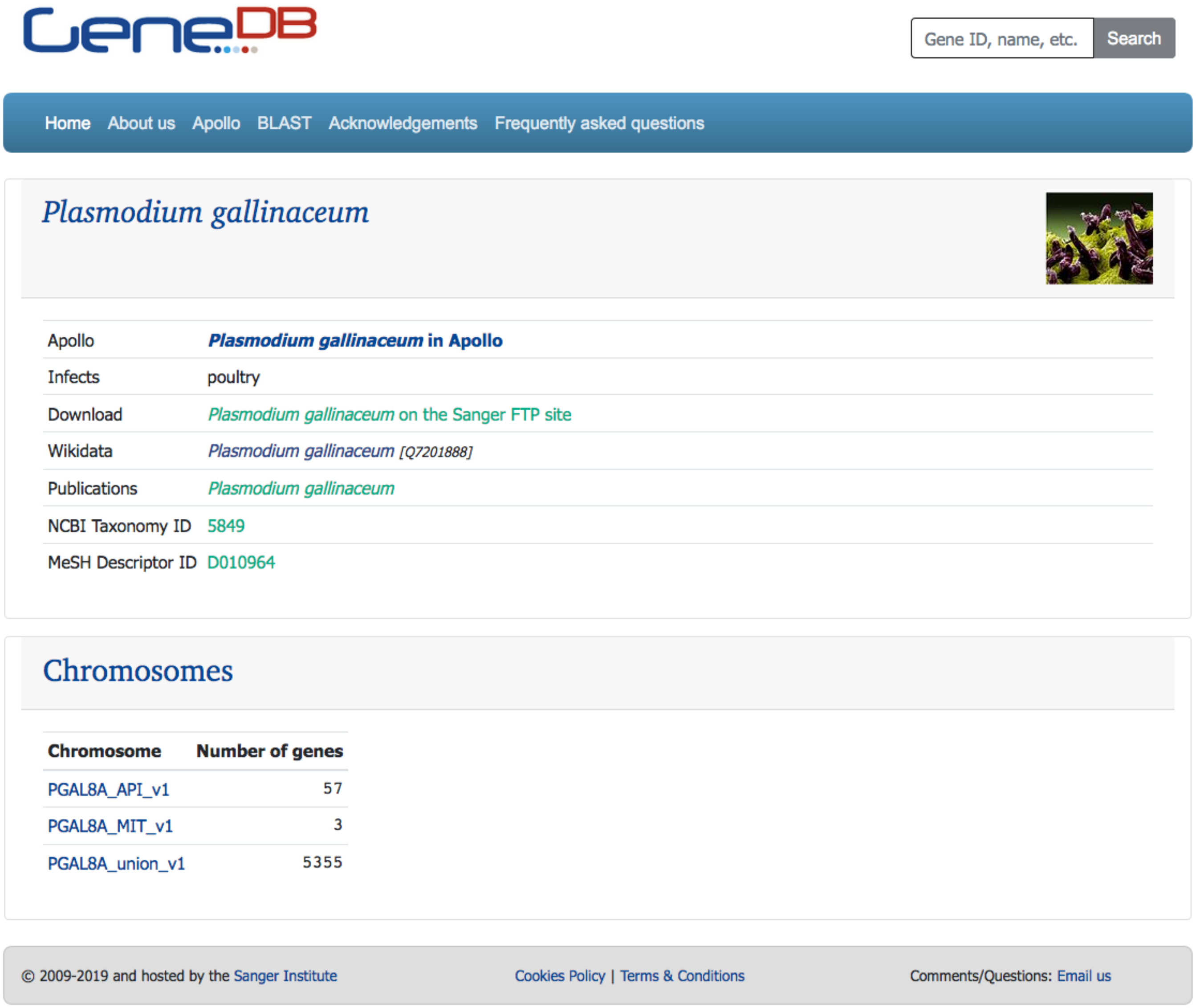
Genes/proteins annotated with specific GO terms can be listed, grouped by species (Figure 5). These lists are linked from every GO term on a gene/protein page.



Ontology

| | | |
|---|---|---|
| cell component | cell nucleus<br>(GO:0005634 ) | determination method : IEA [Inferred from Electronic Annotation]<br>Added to CHADO: 2009-08-05<br>**ESG: extended similarity group method for automated protein function prediction**<br>Meghana Chitale, Troy Hawkins , *et al.*<br>*Bioinformatics* vol 25 issue 14 , 1739-45 , 2009-07-15<br>PubMed ID : 19435743<br>DOI : 10.1093/BIOINFORMATICS/BTP309 |
| | cytoplasm<br>(GO:0005737 ) | determination method : IEA [Inferred from Electronic Annotation]<br>Added to CHADO: 2009-08-05<br>**ESG: extended similarity group method for automated protein function prediction**<br>Meghana Chitale, Troy Hawkins , *et al.*<br>*Bioinformatics* vol 25 issue 14 , 1739-45 , 2009-07-15<br>PubMed ID : 19435743<br>DOI : 10.1093/BIOINFORMATICS/BTP309 |
| | membrane<br>(GO:0016020 ) | determination method : IEA [Inferred from Electronic Annotation]<br>Added to CHADO: 2019-06-01<br>**Sequence of Plasmodium falciparum chromosomes 2, 10, 11 and 14**<br>Malcolm J Gardner, Shamira J Shallom , *et al.*<br>*Nature* vol 419 issue 6906 , 531-4 , 2002-10-03<br>PubMed ID : 12368868<br>DOI : 10.1038/NATURE01094 |
| | nuclear condensin complex<br>(GO:0000799 ) | InterPro ID : IPR012371<br>determination method : IEA [Inferred from Electronic Annotation]<br>Added to CHADO: 2019-03-10<br>GOREF : 0000002 |
| biological process | cell cycle<br>(GO:0007049 ) | determination method : IEA [Inferred from Electronic Annotation]<br>Added to CHADO: 2009-08-05<br>**ESG: extended similarity group method for automated protein function prediction**<br>Meghana Chitale, Troy Hawkins , *et al.*<br>*Bioinformatics* vol 25 issue 14 , 1739-45 , 2009-07-15<br>PubMed ID : 19435743<br>DOI : 10.1093/BIOINFORMATICS/BTP309 |

**Figure 3. GO term ontology (https://www.genedb.org/gene/PF3D7_1135600), with sources.**



*Plasmodium gallinaceum*

| | |
|---|---|
| Apollo | ***Plasmodium gallinaceum* in Apollo** |
| Infects | poultry |
| Download | *Plasmodium gallinaceum* on the Sanger FTP site |
| Wikidata | *Plasmodium gallinaceum* *[Q7201888]* |
| Publications | *Plasmodium gallinaceum* |
| NCBI Taxonomy ID | 5849 |
| MeSH Descriptor ID | D010964 |

Chromosomes

| Chromosome | Number of genes |
|---|---|
| PGAL8A_API_v1 | 57 |
| PGAL8A_MIT_v1 | 3 |
| PGAL8A_union_v1 | 5355 |

© 2009-2019 and hosted by the Sanger Institute

Cookies Policy | Terms & Conditions

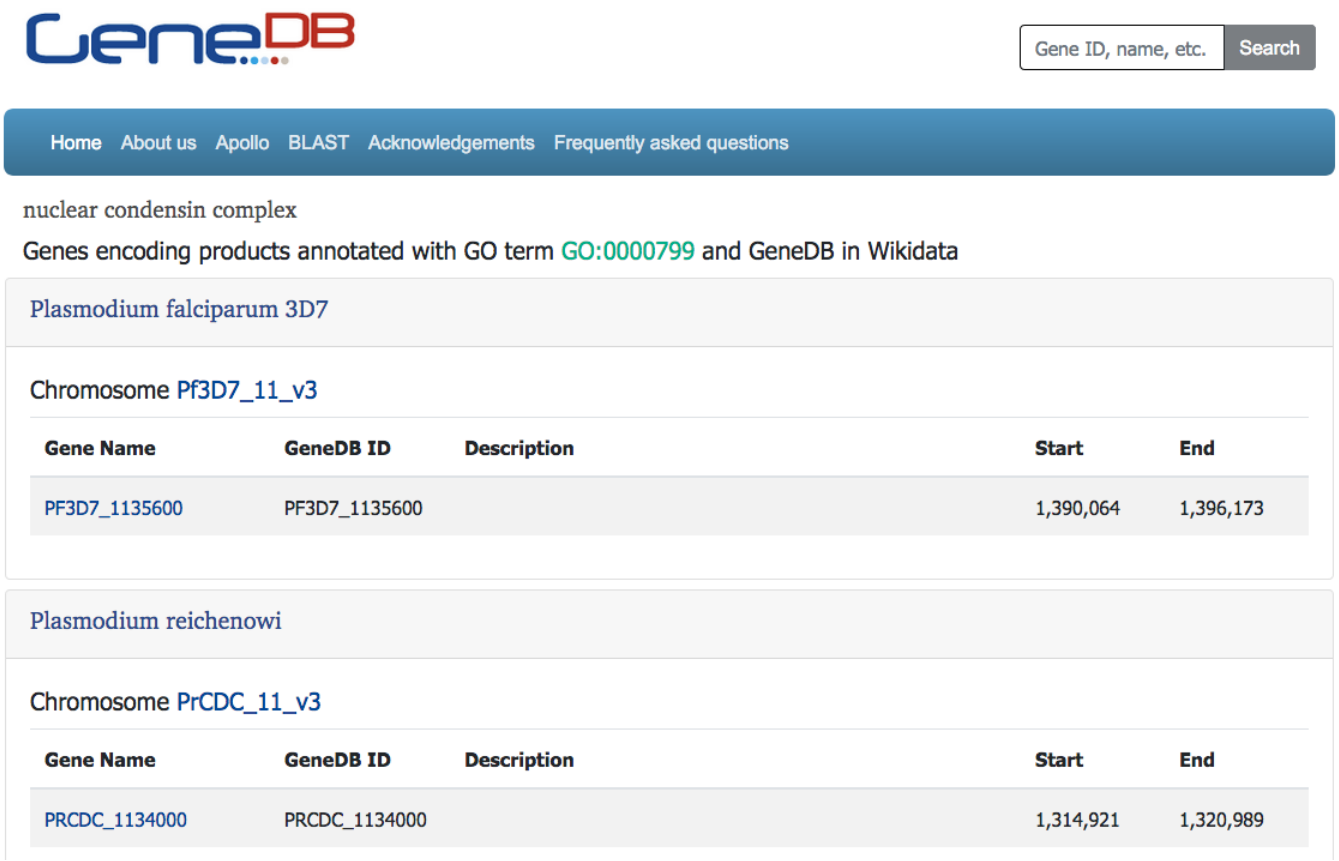
Comments/Questions: Email us

**Figure 4. Species page for *Plasmodium gallinaceum*, using information and image from Wikidata.**

## Discussion

Wikidata has provided GeneDB with a venue to host and publish its data, and to invite community edits, without giving up on GeneDB's curation authority. The simplified maintenance of the HTML/JS-only GeneDB website, compared to a previous one that combined a frontend and backend solution, frees technical and personnel resources. Linking genes and proteins to, and from, other Wikidata items allows for novel methods of querying the data, and for new questions to be asked. Publishing on Wikidata also exposes the data in new ways and potentially engages with a broader public audience.

For sustained operation, we are working on a unified, curated update mechanism that takes user-generated input from Wikidata, Apollo, and Artemis[15], and lets professional curators validate the changes before feeding them back to the Chado database. Changes on Wikidata that are not curated by the GeneDB project may be displayed on the GeneDB website without curation.



nuclear condensin complex

Genes encoding products annotated with GO term GO:0000799 and GeneDB in Wikidata

Plasmodium falciparum 3D7

Chromosome Pf3D7_11_v3

| Gene Name | GeneDB ID | Description | Start | End |
|---|---|---|---|---|
| PF3D7_1135600 | PF3D7_1135600 | | 1,390,064 | 1,396,173 |

Plasmodium reichenowi

Chromosome PrCDC_11_v3

| Gene Name | GeneDB ID | Description | Start | End |
|---|---|---|---|---|
| PRCDC_1134000 | PRCDC_1134000 | | 1,314,921 | 1,320,989 |

**Figure 5. Genes for proteins with a specific GO term (https://www.genedb.org/#/go/GO:0000799), in all species.** Results from Wikidata, limited to genes with a GeneDB ID.

## Data availability

### Source data

**Chado-based GFF and GAF files**

ftp://ftp.sanger.ac.uk/pub/genedb/releases/

**Accession numbers in use cases**

GeneDB IDs

Pathogen genomic data from GeneDB, Accession number PF3D7_0709000. https://identifiers.org/genedb:PF3D7_0709000

Pathogen genomic data from GeneDB, Accession number: PF3D7_1200600. https://identifiers.org/genedb:PF3D7_1200600

Pathogen genomic data from GeneDB, Accession number: PF3D7_1135600. https://identifiers.org/genedb:PF3D7_1135600

Wikidata items

Genomic data for VAR2CSA from Wikidata, Accession number: Q18971176. https://identifiers.org/wikidata:Q18971176

Genomic data for CRT from Wikidata, Accession number: Q19044775. https://identifiers.org/wikidata:Q19044775

*Plasmodium gallinaceum* data from Wikidata, Accession number: Q7201888. https://identifiers.org/wikidata:Q7201888

GO terms

Gene Ontology entry for nuclear condensin complex, Accession number: GO:0000799. https://identifiers.org/GO:0000799

### Underlying data

All data underlying the results are available as part of the article and no additional source data are required.

All data is in the public domain (GeneDB)[16] or CC-0 (Wikidata), which are effectively equivalent.

## Software availability

GeneDBot, the code that updates Wikidata from Chado:

Source code: https://github.com/sanger-pathogens/genedbot_rs

Archived source code at time of publication: http://doi.org/10.5281/zenodo.3352001[12]

License: GPL v3.0

The HTML/JS code for the GeneDB website (Wikidata display code is mostly in the htdocs/wd directory):

Source code: https://github.com/sanger-pathogens/GeneDB-static

Archived source code at time of publication: http://doi.org/10.5281/zenodo.3352003[13]

License: GPL v3.0

## References

1. Logan-Klumpler FJ, De Silva N, Boehme U, *et al.*: **GeneDB--an annotation database for pathogens.** *Nucleic Acids Res.* 2012; **40**(Database issue): D98–D108.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Mungall CJ, Emmert DB, FlyBase Consortium: **A Chado case study: an ontology-based modular schema for representing genome-associated biological information.** *Bioinformatics.* 2007; **23**(13): i337–i346.
PubMed Abstract | Publisher Full Text

3. Böhme U, Otto TD, Sanders M, *et al.*: **Progression of the canonical reference malaria parasite genome from 2002–2019 [version 1; peer review: 2 approved, 1 approved with reservations].** *Wellcome Open Res.* 2019; **4**: 58.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Dunn NA, Unni DR, Diesh C, *et al.*: **Apollo: Democratizing genome annotation.** *PLoS Comput Biol.* Darling AE, editor. 2019; **15**(2): e1006790.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Daub J, Gardner PP, Tate J, *et al.*: **The RNA WikiProject: community annotation of RNA families.** *RNA.* 2008; **14**(12): 2462–2464.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Waagmeester A, Schriml L, Su A: **Wikidata as a linked-data hub for Biodiversity data.** Pensoft Publishers; *Biodiversity Information Science and Standards.* 2019; **3**: e35206.
Publisher Full Text

7. Burgstaller-Muehlbacher S, Waagmeester A, Mitraka E, *et al.*: **Wikidata as a semantic framework for the Gene Wiki initiative.** *Database (Oxford).* 2016; 2016: pii: baw015.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Putman TE, Lelong S, Burgstaller-Muehlbacher S, *et al.*: **WikiGenomes: an open web application for community consumption and curation of gene annotation data in Wikidata.** *Database (Oxford).* 2017; **2017**(1).
PubMed Abstract | Publisher Full Text | Free Full Text

9. Segaran T, Evans C, Taylor J: **Programming the Semantic Web.** O'Reilly Media; 2009.
Reference Source

10. Contributors to Wikimedia projects: **Rust (programming language) - Wikipedia**. In: Wikimedia Foundation, Inc. 2010; [cited 15 May 2019].
Reference Source

11. Köster J: **Rust-Bio: a fast and safe bioinformatics library.** *Bioinformatics.* 2016; **32**(3): 444–446.
PubMed Abstract | Publisher Full Text

12. Manske M: **sanger-pathogens/genedbot v1.0.0 (Version v1.0.0).** *Zenodo.* 2019.
http://www.doi.org/10.5281/zenodo.3352001

13. pathdbpi, Offord V, Manske M, *et al.*: **sanger-pathogens/GeneDB-static v1.0.0 (Version v1.0.0).** *Zenodo.* 2019.
http://www.doi.org/10.5281/zenodo.3352003

14. Ashburner M, Ball CA, Blake JA, *et al.*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet.* 2000; **25**(1): 25–29.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Carver T, Harris SR, Berriman M, *et al.*: **Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data.** *Bioinformatics.* 2012; **28**(4); 464–9.
PubMed Abstract | Publisher Full Text | Free Full Text

16. **The data release policy on GeneDB, placing all data in the public domain.**
Reference Source

# Open Peer Review

## Current Peer Review Status: ✓ ✓

### Version 2

Reviewer Report 30 October 2019

https://doi.org/10.21956/wellcomeopenres.16983.r36761

**© 2019 Waagmeester A.** 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.

✓ **Andra Waagmeester**
Micelio, Antwerp, Belgium

The authors addressed the comments.

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Biomedical informatics, Wikidata, Semantic Web,

**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 15 October 2019

https://doi.org/10.21956/wellcomeopenres.16983.r36762

**© 2019 Burgstaller-Muehlbacher S.** 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.

✓ **Sebastian Burgstaller-Muehlbacher**
Max Perutz Labs, Vienna Biocenter, Medical University of Vienna, Vienna, Austria

The authors modified the original manuscript as suggested, no further comments from my side.

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Bioinformatics, sematic web, knowledge networks, deep learning

**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.**

Version 1

Reviewer Report 27 August 2019

https://doi.org/10.21956/wellcomeopenres.16772.r36100

**© 2019 Waagmeester A.** 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.

?

**Andra Waagmeester**
Micelio, Antwerp, Belgium

This paper describes work done in linking data from the GeneDB website in Wikidata. GeneDB captures genome annotation data. The article also describes how a feedback loop is created enriching GeneDB with content from Wikidata. This is an exciting development where the UI directly leverages content from Wikidata, instead of using in-house caching databases.
I have some minor comments.

1. Wikidata uses a CC0 license on captured data. It is not clear if this is compatible with the applicable Guidelines on the use of data in publications on GeneDB ( https://www.sanger.ac.uk/legal/). The terms for usage require explicitly attribution, which CC0 does not. I am curious how the authors solved this. A section on legalities of sharing GeneDB on Wikidata would be a welcome addition.

2. "The most successful of these is the Rfam project, where Wikipedia has been used to successfully manage free-text descriptions of RNA families[5] for over a decade." Why is this the most successful project, who are the second and third runner up?

**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?**

Partly

**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:*** Biomedical informatics, Wikidata, Semantic Web,

**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, however I have significant reservations, as outlined above.**

Author Response 10 Oct 2019

**Magnus Manske**, Wellcome Trust Sanger Institute, Cambridge, UK

Thank you for your review, and your constructive suggestions. In reply to your individual points:

- Data from GeneDB is in the public domain, which is compatible with CC-0. The paper and the "data release" page of the GeneDB website have been updated accordingly.
- Regarding Rfm, we use "most successful" to describe the uptake of Wikipedia as a central part of the RNA families journal, and thereby, scientific community. This has been added to the manuscript.

***Competing Interests:*** No competing interests were disclosed.

Reviewer Report 20 August 2019

https://doi.org/10.21956/wellcomeopenres.16772.r36099

**© 2019 Burgstaller-Muehlbacher S.** 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.

✓ **Sebastian Burgstaller-Muehlbacher**

Max Perutz Labs, Vienna Biocenter, Medical University of Vienna, Vienna, Austria

The authors describe a new implementation and technical foundation of GeneDB, a database for eukaryotic and prokaryotic parasites. As a new technical foundation, the authors use Wikidata, an open, semantic web enabled graph database operated by the Wikimedia Foundation. To integrate GeneDB data into Wikidata, the authors wrote bots (robots/database clients) using the language Rust and imported GeneDB genes, proteins and Gene Ontology annotations into Wikidata. Furthermore, they created a website which renders the data imported into Wikidata in a comprehensive and more user-friendly way than a raw Wikidata data item would. Also, they integrate genome track views using Apollo.

Overall, the study opens up a new dataset to the public domain entirely. Data are imported and made available via the Wikidata API and the Wikidata SPARQL endpoint, allowing users of the data powerful queries. Having these data available and easily accessible to the parasite research community worldwide seems crucial. Easy access is also enabled by the web interface the authors implemented.
Moreover, due to the nature of Wikidata, contributions by the (research) community are relatively easy to accomplish. The authors also provide a reasonable approach for feeding back (research) community contributions into GeneDB, while avoiding vandalism.

Integration with the Apollo genome browser also provides a genome view, typical for a gene/genome centered database. However, the authors should describe in some more detail how they setup Apollo and where Apollo gets its genome tracks from (e.g. DNA sequences). This is relevant as such data typically cannot be stored in Wikidata and thus need to be drawn from a different source.

The manuscript is well-written, however the introduction should also mention the work of Putman *et al.* (2017[1]), who follow a similar technical approach but for a different set of genomes. The discussion is rather brief and could profit from addressing the authors long-term maintenance plans, especially regarding curation of the inflow of community contributions to GeneDB.

In summary, this work by Manske *et al.* is very well performed, except for the minor points mentioned above. It's an important data release for the parasite research community and enables, for the first time, collaborative work on genomic data integration and data annotation within this community. This should lower greatly the hurdles for worldwide collaboration on alleviating the severe diseases associated with many of the parasites in GeneDB.

Ad citation 7 of the manuscript: This preprint has now been published as Burgstaller-Muehlbacher *et al.* (2016[2]).

**References**

1. Putman T, Lelong S, Burgstaller-Muehlbacher S, Waagmeester A, Diesh C, Dunn N, Munoz-Torres M, Stupp G, Wu C, Su A, Good B: WikiGenomes: an open web application for community consumption and curation of gene annotation data in Wikidata. *Database*. 2017; **2017**. Publisher Full Text
2. Burgstaller-Muehlbacher S, Waagmeester A, Mitraka E, Turner J, Putman T, Leong J, Naik C, Pavlidis P, Schriml L, Good BM, Su AI: Wikidata as a semantic framework for the Gene Wiki initiative.*Database (Oxford)*. 2016; **2016**. PubMed Abstract | Publisher Full Text

**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, sematic web, knowledge networks, deep learning

**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.**

Author Response 10 Oct 2019

**Magnus Manske**, Wellcome Trust Sanger Institute, Cambridge, UK

Thank you for your review, and your constructive suggestions. In reply to your individual points:

- The Apollo part of the GeneDB website is not part of this publication. Please see Böhme, Otto et al. (2019) for details.
- The reference to Putman et al. (2017) has been added.
- Some discussion of long-term sustainability has been added.
- The Burgstaller-Muehlbacher citation has been updated.

***Competing Interests:*** No competing interests were disclosed.