SPECIFIC AIMS

Scientific research is most efficient when new hypotheses are based on the sum total of all human knowledge to date. Unfortunately, we as a research community do a relatively poor job of organizing existing biomedical knowledge. There are at least two key inefficiencies in our current system of knowledge management. First, our primary mechanism for communicating new findings continues to be unstructured, free-text publications. This corpus of documents is very difficult to utilize computationally (for example, to build tools to analyze and visualize relationships between biomedical entities), and therefore individual scientists are generally limited to the tiny proportion of knowledge contained in the papers they are personally able to read. Second, the landscape of biomedical knowledge is highly fragmented. This fragmentation of knowledge is observed in the > 1 million new research articles published every year, as well as among the hundreds of databases that aim to structure information in specific areas of biology.

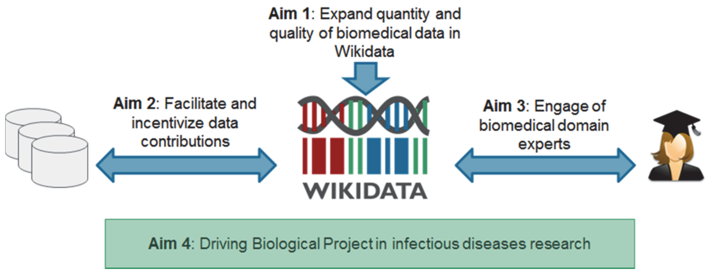

Aim 1: Expand quantity and quality of biomedical data in Wikidata
Aim 2: Facilitate and incentivize data contributions
Aim 3: Engage of biomedical domain experts
WIKIDATA
Aim 4: Driving Biological Project in infectious diseases research

Figure 1. Schematic overview of the proposal's aims.

This proposal addresses both of these challenges, focusing on making all biomedical knowledge **Findable, Reusable, Accessible, and Interoperable (FAIR)**. We will continue our work leveraging Wikidata, a centralized and structured knowledge base that is openly edited and maintained by the community. Our vision to make Wikidata the largest repository of open biomedical knowledge can be achieved through the following three aims (**Figure 1**):

Aim #1: Build an ecosystem of tools that advances both the quantity and quality of biomedical data in Wikidata. Having accurate and complete information is essential to Wikidata's continued growth as a trusted knowledge source.

A) Enhance our data ingestion ecosystem for adding valuable sources of biomedical information to Wikidata. This work will address the entire data pipeline from an informatics, software, and legal perspective.
B) Create computable data models to ensure data quality and consistency. Based on Shape Expressions (ShEx), these data models will support robust data validation and surveillance tools.

Aim #2: Create Wikidata-based tools to facilitate and incentivize third-party data contributions. These tools are specifically aimed at convincing biomedical resource owners to contribute to Wikidata.

A) Extend our Wikidata Integrator (WDI) python library into a fully-featured database connector. WDI will support data modeling and validation checks, as well as advanced logging and reporting functionality.
B) Generate automated reports of community contributions to relevant Wikidata items. These reports will notify curators when additions or edits are made, effectively serving as a universal community curation interface.

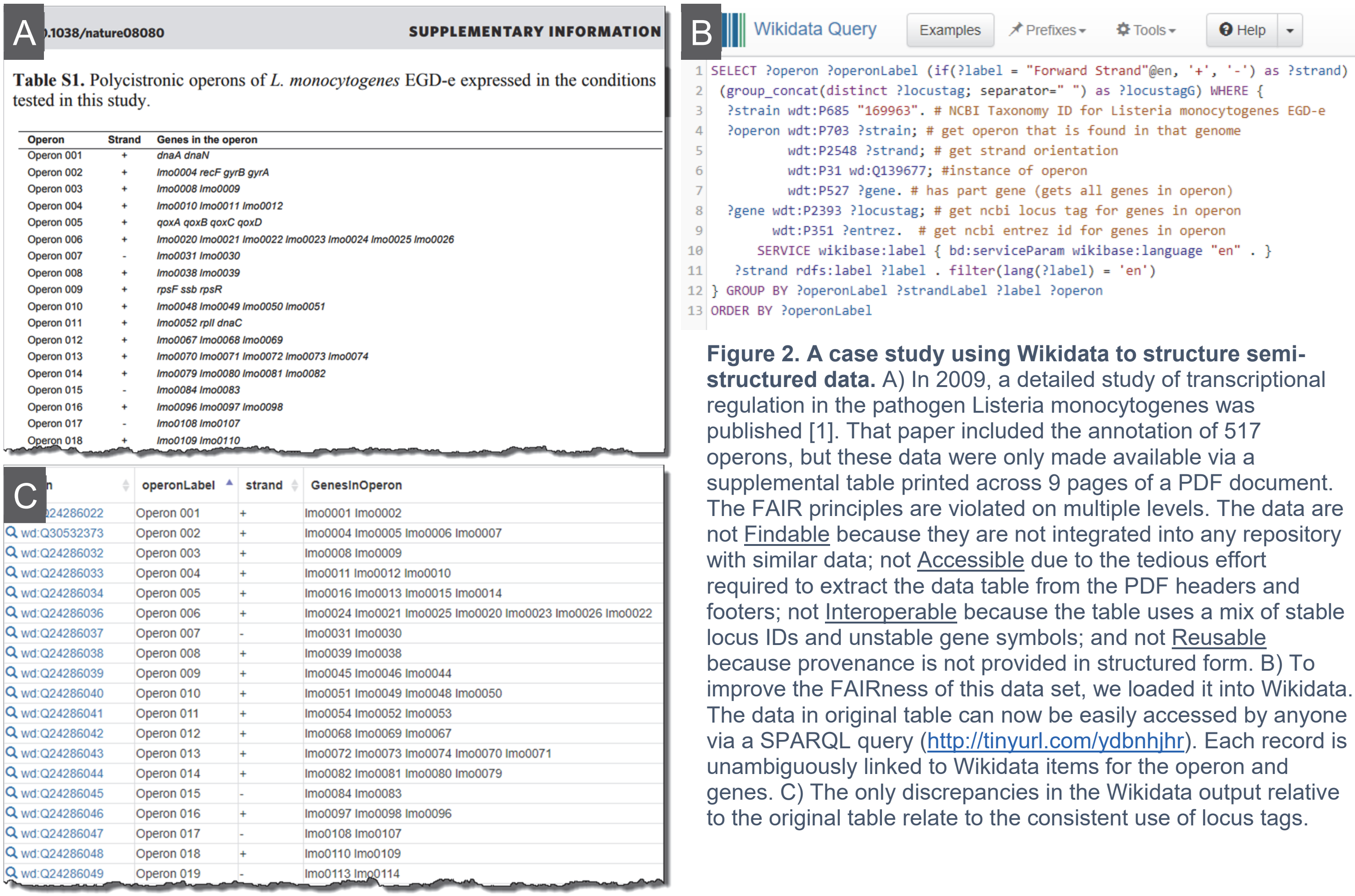
Aim #3: Solicit "deep and narrow" contributions from targeted communities of domain experts. In contrast to the broad and shallow focus of most database imports, this aim will target information from experts that is not currently captured in any structured format.

A) Integrate our Wikidata-backed WikiGenomes application with the Apollo genome annotation tool. This synergy will combine our complementary strengths of a powerful user interface and robust community infrastructure.
B) Generalize the Chlambase application into a deployable Software Development Kit. This SDK will enable organism-specific genome databases to be created with minimal resources and administration effort.
C) Capture structured data from review article authors through our partnership with the journal GENE. This existing partnership utilizes a dual publication model to incentivize expert contributions.

Crossing all of these Specific Aims will be a **Driving Biological Project focusing on infectious disease research**. We will apply the tools and approaches developed in this proposal to build the knowledge management foundation for a project to detect pathogens from unbiased metagenomic sequence data.

Overall Significance. Dealing with unstructured data in biology is one of the greatest information management challenge researchers currently face. The problem can be best demonstrated by example. In 2009, Toledo-Arana et al. published a now widely cited paper on transcriptional regulation in *Listeria monocytogenes* [1]. Among their findings was a comprehensive characterization of the operon structure in this organism's genome. One can imagine many secondary analyses based on these data including, for example, inference of metabolic pathways, or comparative analyses of transcriptional regulation.

Unfortunately, the critical data table with these operon data was only published in the supplemental info, and then only embedded in a PDF file (**Figure 2**). When evaluating this paper through the lens of FAIR data principles [3], this paper does far more to inhibit reuse than facilitate it. So despite its scientific importance, this data set is likely being used far less widely and less efficiently than it could be. At best, many scientists are redundantly performing the same data discovery and data cleaning tasks over and over again. At worst, these data are not being reused simply because of the many obstacles to FAIRness.



.1038/nature08080 **SUPPLEMENTARY INFORMATION**

Table S1. Polycistronic operons of *L. monocytogenes* EGD-e expressed in the conditions tested in this study.

Operon	Strand	Genes in the operon
Operon 001	+	*dnaA dnaN*
Operon 002	+	*lmo0004 recF gyrB gyrA*
Operon 003	+	*lmo0008 lmo0009*
Operon 004	+	*lmo0010 lmo0011 lmo0012*
Operon 005	+	*qoxA qoxB qoxC qoxD*
Operon 006	+	*lmo0020 lmo0021 lmo0022 lmo0023 lmo0024 lmo0025 lmo0026*
Operon 007	-	*lmo0031 lmo0030*
Operon 008	+	*lmo0038 lmo0039*
Operon 009	+	*rpsF ssb rpsR*
Operon 010	+	*lmo0048 lmo0049 lmo0050 lmo0051*
Operon 011	+	*lmo0052 rplI dnaC*
Operon 012	+	*lmo0067 lmo0068 lmo0069*
Operon 013	+	*lmo0070 lmo0071 lmo0072 lmo0073 lmo0074*
Operon 014	+	*lmo0079 lmo0080 lmo0081 lmo0082*
Operon 015	-	*lmo0084 lmo0083*
Operon 016	+	*lmo0096 lmo0097 lmo0098*
Operon 017	-	*lmo0108 lmo0107*
Operon 018	+	*lmo0109 lmo0110*

```
SELECT ?operon ?operonLabel (if(?label = "Forward Strand"@en, '+', '-') as ?strand)
 (group_concat(distinct ?locustag; separator=" ") as ?locustagG) WHERE {
  ?strain wdt:P685 "169963". # NCBI Taxonomy ID for Listeria monocytogenes EGD-e
  ?operon wdt:P703 ?strain; # get operon that is found in that genome
          wdt:P2548 ?strand; # get strand orientation
          wdt:P31 wd:Q139677; #instance of operon
          wdt:P527 ?gene. # has part gene (gets all genes in operon)
  ?gene wdt:P2393 ?locustag; # get ncbi locus tag for genes in operon
        wdt:P351 ?entrez.  # get ncbi entrez id for genes in operon
      SERVICE wikibase:label { bd:serviceParam wikibase:language "en" . }
   ?strand rdfs:label ?label . filter(lang(?label) = 'en')
} GROUP BY ?operonLabel ?strandLabel ?label ?operon
ORDER BY ?operonLabel
```

n	operonLabel	strand	GenesInOperon
Q24286022	Operon 001	+	lmo0001 lmo0002
wd:Q30532373	Operon 002	+	lmo0004 lmo0005 lmo0006 lmo0007
wd:Q24286032	Operon 003	+	lmo0008 lmo0009
wd:Q24286033	Operon 004	+	lmo0011 lmo0012 lmo0010
wd:Q24286034	Operon 005	+	lmo0016 lmo0013 lmo0015 lmo0014
wd:Q24286036	Operon 006	+	lmo0024 lmo0021 lmo0025 lmo0020 lmo0023 lmo0026 lmo0022
wd:Q24286037	Operon 007	-	lmo0031 lmo0030
wd:Q24286038	Operon 008	+	lmo0039 lmo0038
wd:Q24286039	Operon 009	+	lmo0045 lmo0046 lmo0044
wd:Q24286040	Operon 010	+	lmo0051 lmo0049 lmo0048 lmo0050
wd:Q24286041	Operon 011	+	lmo0054 lmo0052 lmo0053
wd:Q24286042	Operon 012	+	lmo0068 lmo0069 lmo0067
wd:Q24286043	Operon 013	+	lmo0072 lmo0073 lmo0074 lmo0070 lmo0071
wd:Q24286044	Operon 014	+	lmo0082 lmo0081 lmo0080 lmo0079
wd:Q24286045	Operon 015	-	lmo0084 lmo0083
wd:Q24286046	Operon 016	+	lmo0097 lmo0098 lmo0096
wd:Q24286047	Operon 017	-	lmo0108 lmo0107
wd:Q24286048	Operon 018	+	lmo0110 lmo0109
wd:Q24286049	Operon 019	-	lmo0113 lmo0114

Figure 2. A case study using Wikidata to structure semi-structured data. A) In 2009, a detailed study of transcriptional regulation in the pathogen Listeria monocytogenes was published [1]. That paper included the annotation of 517 operons, but these data were only made available via a supplemental table printed across 9 pages of a PDF document. The FAIR principles are violated on multiple levels. The data are not Findable because they are not integrated into any repository with similar data; not Accessible due to the tedious effort required to extract the data table from the PDF headers and footers; not Interoperable because the table uses a mix of stable locus IDs and unstable gene symbols; and not Reusable because provenance is not provided in structured form. B) To improve the FAIRness of this data set, we loaded it into Wikidata. The data in original table can now be easily accessed by anyone via a SPARQL query (http://tinyurl.com/ydbnhjhr). Each record is unambiguously linked to Wikidata items for the operon and genes. C) The only discrepancies in the Wikidata output relative to the original table relate to the consistent use of locus tags.

To bring this Listeria operon data set into the Open Data ecosystem, we loaded it into Wikidata, a sister project to the popular online encyclopedia, Wikipedia. Like Wikipedia, Wikidata empowers the community at large to collaboratively create and maintain an information resource. Whereas Wikipedia focuses on unstructured encyclopedia articles, Wikidata focuses on a knowledge base of structured content. By loading this data set into Wikidata, it can now be easily retrieved via the Wikidata API in a fully structured format and by anyone in the community. No one should have to extract this table from the PDF ever again. In short, this data set is now completely FAIR-compliant (**Figure 2**).

This example is of course generalizable far beyond this single data set (we are currently processing another data set where operon membership is specified by alternating row colors in Excel), and even beyond just structured data tables in publications. Wikidata is in principle able to represent biomedical knowledge at massive breadth as well as fine-grained granularity. Wikidata can fundamentally change how we communicate, disseminate, and integrate scientific knowledge. Of course, in an ideal world, the publishing industry would innovate to empower/require authors to provide structured data as a core part of each publication. But this prospect seems exceedingly unlikely in the foreseeable future, and in the meantime, Wikidata offers a compelling solution in which anyone (authors, informaticians, curators, domain experts) can properly structure a data set once and then easily share the output of their effort with the entire scientific community.

The discussion above describes the significance of Wikidata as a FAIR repository for biomedical knowledge *in general*. The significance of this proposal's Aims *in particular* lies in addressing the primary concerns of three target communities of scientists. Specifically, **Aim 1** addresses the needs of data consumers, who need a solid foundation of content in terms of both quantity and quality. **Aim 2** appeals to data contributors by lowering technical barriers and increasing positive incentives. **Aim 3** facilitates contributions from domain experts through focused web interfaces and initiatives.

In summary, science is most efficient when new hypotheses are based on the entirety of knowledge known to date. However, the current state of fragmentation and inaccessibility of biomedical information is greatly impeding the efficiency of scientific research. The scientific premise of this proposal is that building a FAIR repository for biomedical knowledge will improve the efficiency of research, and that Wikidata will be a valuable tool to engage the community in achieving this goal.

Innovation. The common theme of this grant has always revolved around the use of crowdsourcing for biomedical knowledge management. While knowledge management continues to be dominated by centralized efforts in biocuration and database creation, there is growing appreciation that crowdsourcing can be a complementary approach to these traditional methods. We and others have investigated and demonstrated the use of community platforms that enable researchers to collaboratively make biomedical knowledge more FAIR.

In addition to this emphasis on crowdsourcing, our effort was relatively unique in its focus on using existing community platforms, rather than creating a new crowdsourcing site *de novo*. This grant initially focused on the use of Wikipedia as a crowdsourcing platform, but that emphasis has shifted completely in this proposal to Wikidata. Wikidata currently has items for over 27 million items, and has received over 500 million edits since its inception in 2012.

Relative to the traditional approach of creating new resources from scratch, our use of Wikidata has several important advantages. Specifically:

- The Wikimedia Foundation has a proven track record of serving massive resources with high reliability and performance,
- The platform requires no additional resources or effort from us to maintain the infrastructure,
- The scope is essentially boundless, which enables greater opportunities for data integration that cross traditional boundaries,
- Both Wikipedia and Wikidata have an existing critical mass of users and contributors.

Finally, the Wikidata platform itself is also unique and innovative in several important ways. First, it is one of the first graphical user interfaces for read/write access to the Semantic Web, which is potentially as transformative as user-friendly publication tools (e.g., blogging platforms, YouTube) were for the World Wide Web. Second, Wikidata improves data science reproducibility by making structured qualifiers and references part of the core data model, meaning provenance is easily computable. Third, Wikidata adopted an explicit CC0 "Public Domain" license, which means that downstream uses are completely unencumbered by licensing issues (which have arisen as significant roadblocks to science in the past [4]). Finally, Wikidata shifts the burden of aligning data models in data integration efforts from the consumer (repetitively done by each end user) to the contributor, thereby greatly simplifying downstream data reuse. For all these reasons, we are excited to build upon the Wikidata platform, and Wikidata is similarly excited to develop biomedical use cases within their framework (see **Letter of Support from Lydia Pintscher**).

In summary, we believe that innovation is a core strength of this proposal based on the high-level emphasis on crowdsourcing, on the use of existing community platforms, and specifically on our use of Wikidata.

Progress report. The prior project period of this grant extends from July 2014 to April 2018. That proposal contained four aims. Three of those prior aims are directly relevant to the current Aims in this proposal, and <u>the relevant progress from the previous project period is interleaved in the Aims below</u>. The prior aim that does *not* directly relate to an Aim in this proposal described our effort to apply citizen science to challenges in biocuration. Specifically, we sought to test whether volunteers with no specialized training could (and would) perform biomedical entity recognition at an accuracy comparable to professional biocurators. We first tested this hypothesis using Amazon Mechanical Turk, a paid microtask marketplace. We showed that while a single volunteer did not perform as well as a single professional, the consensus vote of *six* volunteers was roughly equivalent to an expert biocurator on a disease recognition task [5]. We then created a citizen science application called Mark2Cure and repeated this experiment with unpaid volunteers with qualitatively equivalent results [6]. Together, these results demonstrated that citizen scientists are both capable of and willing to perform biocuration tasks with a high degree of accuracy. While we consider Mark2Cure to be a successful, ongoing project with over 1000 volunteers, we have decided that it is out of scope of the current proposal.

Aim #1: Build an ecosystem of tools that advances both the quantity and quality of biomedical data in Wikidata. Having accurate and complete information is essential to Wikidata's continued growth as a trusted knowledge source.

Progress report / preliminary data. We featured Wikidata prominently in our last funded proposal, and its importance (and our enthusiasm) grew substantially over the course of the project period. The basic data model for Wikidata can be distilled into a statement describing a relationship (subject, property, value) with qualifiers and references (**Figure 3**). We previously proposed to create Wikidata items for key biomedical entities – genes, proteins, drugs, and diseases – and load key annotations for and relationships between those items. (Separately, we also proposed to load a substantial amount of microbial genomics data, which is described more in **Aim 3**.) This work involved data modeling discussions with the Wikidata community (including lengthy discussion and detailed patterns for recording provenance and evidence [7]), creating "bots" to load data sets, and automating synchronization with the source databases [8].

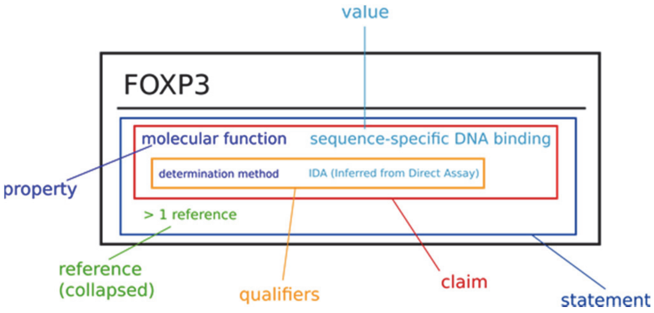


Figure 3. The basic structure of a Wikidata statement. At its core, a Wikidata statement is composed of a triple comprised of a subject, a predicate (or "property"), and an object (or "value). In the above example, the elements of the triple are FOXP3, "molecular function", and "sequence-specific DNA binding", respectively. Statements can also optionally have references (which are designed to capture provenance of the assertion) and/or qualifiers (that capture, for example, the determination method or biological context in which the assertion is valid).. Importantly, all the entities shown above are assigned unique identifiers (e.g., FOXP3 = Q426188, molecular function = P680, sequence-specific DNA binding = Q14818107).

These proposed data loading aims from the last project period are now substantially complete. We have systematically imported data from a wide range or resources, including NCBI Gene [9], Ensembl [10], UniProt [11], InterPro [12], Disease Ontology [13], GWAS Catalog [14], ChEMBL [15], NDF-RT [16] and PubChem [17]. Collectively, these bots have created or edited well over one million Wikidata items on biomedical entities. Although these bots are being regularly and automatically run in our production environment, they are also undergoing continual refinement to improve the fidelity of the data models, to provide more detail in evidence and provenance, and to account for corner cases.

What we did not anticipate, and what we were pleasantly surprised by, was the emergence of other related data providers who also wanted to be included within Wikidata. Some of those data resources chose to work in varying degrees of coordination with our team (e.g., WikiPathways [18], Reactome [19], CIVIC [20], ECO [21]), while other initiatives were executed completely independently of us (e.g., chemical toxicity data from the CDC, bibliographic data for scientific literature). The result of our collective efforts is a rich network of biomedical knowledge (a portion of which is summarized in **Figure 4**).

Wikidata provides a powerful query interface based on the SPARQL query language, which enables rich, integrative queries over this knowledge base. For example, a relatively <u>simple query over the Wikidata network</u> can retrieve "drugs for cancers that target genes related to cell proliferation and that physically interact with

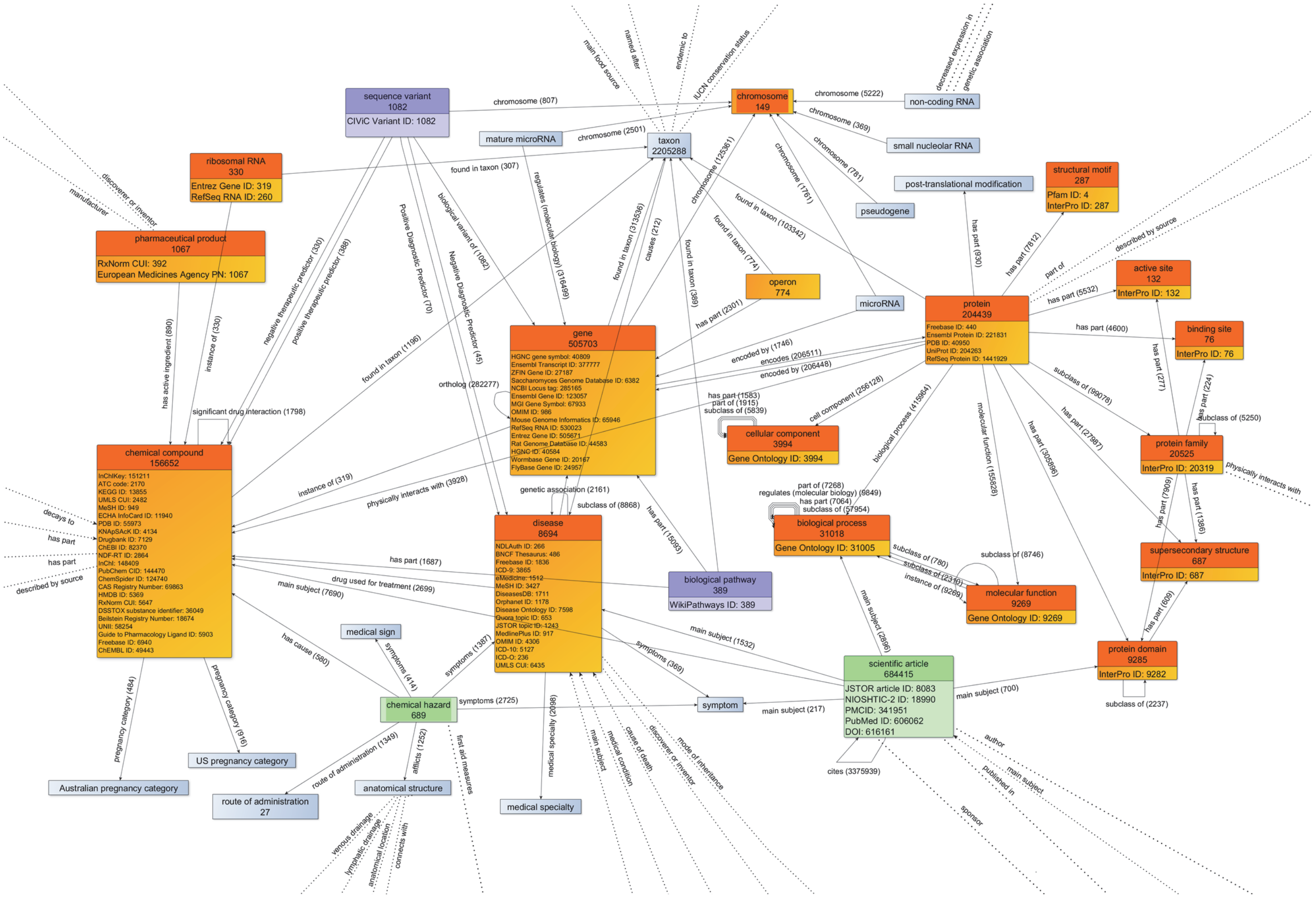


Figure 4. A class-level diagram depicting the core biomedical knowledge network in Wikidata. Every box represents one entity type, which is noted in the box header with the number of Wikidata items of that class. The attributes listed in the box are Wikidata properties (and property counts) for each entity type. Properties that link to other Wikidata items are depicted as edges in the graph. The core entity types that were loaded by our team loaded during the previous project period are shown in orange. Data for human genetic variants and biological pathways are being loaded by other research groups, but in coordination with our team and using our Python library (purple boxes). A team from NIOSH has loaded data on chemical hazards and environmental exposures, and a separate team of researchers is loading bibliographic data for the scientific literature (green boxes). A larger version of this figure can be viewed at https://www.wikidata.org/wiki/User:ProteinBoxBot.

gene products known to be genetically associated to a disease", or "all organisms that are located in the female urogenital tract and that have a gene involved in indole metabolism". More complex queries based on references and evidence can be filtered based on specific determination methods, evidence codes, disputed claims, or bibliographic data. Finally, federated queries can dynamically combine data from other SPARQL endpoints (e.g., UniProt, WikiPathways) to incorporate data that may not be appropriate for loading to Wikidata. (The specific SPARQL queries for these examples and others are available in [22].)

Significance. The significance of this proposal is exemplified by the SPARQL queries described (and linked) above. Based on the data that we and others have seeded, Wikidata is already a unique and powerful resource for answering biomedical questions. As Wikidata continues to grow, the results of these queries will continue to improve in terms of both precision and recall.

However, the significance *of this aim* specifically lies in ensuring that the foundation of biomedical data in Wikidata is robust, both in terms of quantity and quality. If data consumers do not have a high degree of confidence in this foundation, all the technical and architectural advantages of the Wikidata platform will be of no practical value. Therefore, this aim proposes two mechanisms to ensure that Wikidata has a firm foundation of biomedical knowledge – by directly adding commonly-used data resources of strategic importance, and by creating rigorous data models that ensure data quality and consistency.

A) Enhance our data ingestion ecosystem for adding valuable sources of biomedical information to Wikidata. This work will address the entire data pipeline from an informatics, software, and legal perspective.

This subaim will involve at least three specific components. First, we will load the **Medical Subject Headings (MeSH)** resource that was created by the National Library of Medicine (NLM) [23] into Wikidata. MeSH was created over 50 years ago as a “hierarchically-organized terminology for indexing and cataloging of biomedical information,” and it is one of the most widely used vocabularies across the landscape of biomedical resources [24]. Although MeSH itself is not an ontology, it nevertheless is a key vocabulary to which many ontologies are mapped, and therefore serves a foundational role in biomedical information management. As one metric of its centrality, among the 566 resources indexed in BioPortal [25], MeSH has the third highest number of mappings to other biomedical vocabularies (666,954), and easily the highest number among resources with an open license.

Here, we will create a Wikidata bot that will both perform an initial loading of MeSH and keep Wikidata in sync with the source data files. However, even more challenging than writing the bot itself will be ensuring that any existing entries are reliably detected (so that duplicate Wikidata items are not created). Given the nature of the MeSH records (many concepts that are likely to already have Wikidata items) and the number of records (27,921 Class 1 Descriptors and 241,851 Supplementary Chemical Records), it is likely that significant overlap will already exist. Nevertheless, string matching alone is not reliable. This challenge is similar to what we faced in loading the Disease Ontology [13] into Wikidata (albeit on a much smaller scale), so we will employ similar strategies. Specifically, in addition to string matching, we will develop concept matching heuristics based on secondary identifiers, as well as Wikidata’s subclass relationships. Any uncertain or ambiguous cases will be flagged for manual inspection. Initially, manual review will be handled by our team, but if the numbers are too large then we will employ other crowd-based curation methods like the Wikidata Game [26, 27].

The second component of this subaim will be to build a generic importer for resources in the Open Biomedical Ontologies (OBO) Foundry [28, 29]. The goal of the OBO Foundry is to develop a suite of biomedical ontologies that adhere to a common set of principles, and collectively they cover a broad cross section of biological domain areas. Like for MeSH, systematically loading these ontologies into Wikidata would provide a solid foundation for biomedical knowledge management. Our generalized importer for OBO Foundry ontologies will be based on our current bot that loaded and maintains the **Disease Ontology (DO)** in Wikidata. Since DO is a part of the OBO Foundry, much of the technical infrastructure for data modeling, duplication checking, and manual review will be reused. In addition, we will provide a written SOP to guide ontology developers through the entire Wikidata loading process.

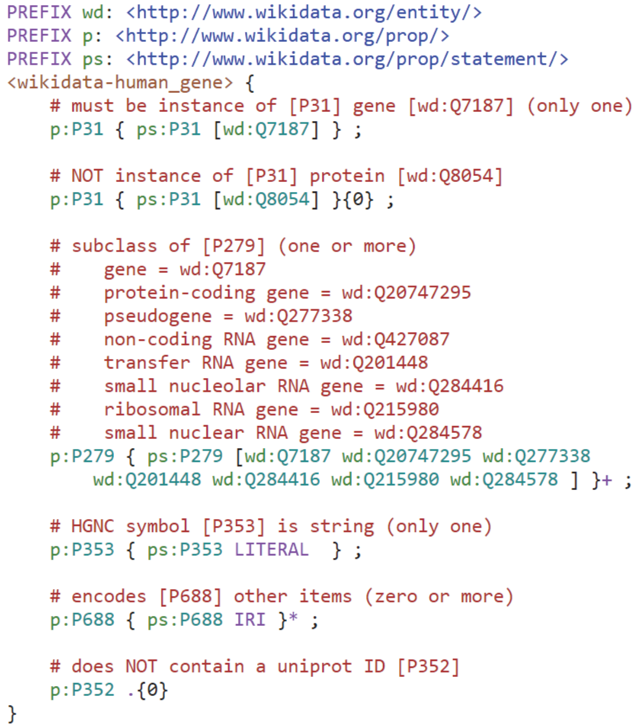
However, there is one formidable challenge with importing ontologies from the OBO Foundry – data licensing. There are currently 154 non-obsolete ontologies listed in the OBO Foundry [28]. Of these, only 81 provide an explicit data license, and only five are licensed under a CC0 Public Domain license (and hence would be eligible for inclusion in Wikidata). Nevertheless there are two reasons why we are pursuing an OBO Foundry importer as part of this subaim. First, a robust discussion is ongoing (led in part by our team) within the bioinformatics community on the subject of data licensing [30-32]. So we believe that the 73 ontologies that currently lack licenses eventually be compelled to explicitly choose a license (in compliance with the Foundry’s stated principles [33]). Second, we believe that momentum toward CC0 licensing for data resources is growing relative to CC-BY (by far the most popular license chosen in the OBO Foundry). This shift is largely based on the relatively recent realization that the scientific community already has a strong tradition of citation and attribution, and legally requiring attribution merely puts up legal roadblocks to reuse [34, 35]. Just within the past year, several groups have converted from CC-BY to CC0 based at least in part on these discussions, including Disease Ontology [13], Evidence and Conclusion Ontology [21], Symptom Ontology [36], Pathogen Transmission Ontology [37], CIViC [20], WikiPathways [38], and the Cancer Genome Interpreter [39], and several others are currently contemplating the same migration [40, 41].

In light of this data licensing issue and recent trends, the third component of this subaim is to continue our CC0 advocacy among data resource providers. The examples of conversions to CC0 mentioned in the previous paragraph demonstrate that personal engagement works, and that the argument for CC0 is often chosen when the pros and cons are actually laid out for consideration. Since the principle of open data has always been at the core of this Gene Wiki project, we believe that advocacy efforts within the community to promote CC0 for data licensing are well within the scope of this aim and of this proposal.

We believe that the three components in this subaim are complementary and synergistic. In the event that we encountered any unexpected obstacles in any one of these components, we believe that the Aim will still be successfully accomplished by focusing on the remaining two components.

<u>B) Create computable data models to ensure data quality and consistency.</u> Based on Shape Expressions (ShEx), these data models will support robust data validation and surveillance tools.

Relatively early in our Wikidata effort, we discovered that while community editing has many advantages, we also needed mechanisms to detect well-intentioned but incorrect edits. For example, we created separate Wikidata items for each gene and the protein(s) it encodes. This semantic precision between genes and proteins was an important part of our early data modeling efforts. However, we found that other editors would sometimes merge the gene and protein items based on their similar names, not realizing the importance of that distinction. Those actions resulted in semantically incorrect statements of genes having protein domains, and proteins having genomic coordinates.



```
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX p: <http://www.wikidata.org/prop/>
PREFIX ps: <http://www.wikidata.org/prop/statement/>
<wikidata-human_gene> {
    # must be instance of [P31] gene [wd:Q7187] (only one)
    p:P31 { ps:P31 [wd:Q7187] } ;

    # NOT instance of [P31] protein [wd:Q8054]
    p:P31 { ps:P31 [wd:Q8054] }{0} ;

    # subclass of [P279] (one or more)
    #    gene = wd:Q7187
    #    protein-coding gene = wd:Q20747295
    #    pseudogene = wd:Q277338
    #    non-coding RNA gene = wd:Q427087
    #    transfer RNA gene = wd:Q201448
    #    small nucleolar RNA gene = wd:Q284416
    #    ribosomal RNA gene = wd:Q215980
    #    small nuclear RNA gene = wd:Q284578
    p:P279 { ps:P279 [wd:Q7187 wd:Q20747295 wd:Q277338
        wd:Q201448 wd:Q284416 wd:Q215980 wd:Q284578 ] }+ ;

    # HGNC symbol [P353] is string (only one)
    p:P353 { ps:P353 LITERAL  } ;

    # encodes [P688] other items (zero or more)
    p:P688 { ps:P688 IRI }* ;

    # does NOT contain a uniprot ID [P352]
    p:P352 .{0}
}
```

Figure 5. A simple ShEx model for validating human gene items in Wikidata. For example, this model declares that valid items have only one HGNC symbol, have one or more subclass relationships to an enumerated list of gene types, have any number of linked proteins that it encodes, and NOT be a subclass of a protein or have a UniProt ID (either of which would indicate an erroneous merging of Wikidata items). A more complete ShEX model for human genes can be found at [2].

We discovered that the Semantic Web community was also focused on this challenge of expressing and validating data models. In particular, **Shape Expressions (ShEx)** had emerged as a formal means of expressing data models as graph structures. The ShEx standard is being developed within the W3C community [42].

Here, we propose to create an entire library of ShEx models for the biomedical entity types within Wikidata. These would include (but not be limited to), models for genes, proteins, diseases, drugs, genetic variants, and pathways. These models would assert, for example, that one item cannot be a subclass of both a gene and a protein, and that a gene can have multiple Gene Ontology annotations but only one linked NCBI Gene ID. ShEx models will also allow us to formally define the evidence and reference model that is currently only expressed in free text [7]. A simple version of a ShEx model for a human gene is shown in **Figure 5**, and a more complete model can be found at [2].

We will maintain this library of ShEx data models in a public Github repository so that anyone in the community can use them and propose changes. (We also believe that these models will help new users and tool developers become familiar with the biomedical data in Wikidata.) Based on these ShEx models, we will maintain a continually running script that validates Wikidata items against the relevant ShEx models. Any Wikidata items that fail to validate would be flagged for manual inspection and/or notification of a relevant data maintainer (see also **Aim 2B**).

We are confident that the ShEx specification is well-suited to our aim, and that we will have the support of the ShEx community in defining our models and adapting to our requirements (see **Letter of Support from Eric Prud'hommeaux**). However, if unexpected and unsurmountable obstacles arise, we will investigate alternative systems for authoring and validating data models, including SHACL [43] and the built-in Wikidata constraint checking system. Once the models are defined in any language, translating the expressions to another syntax will be straightforward.

Aim #2: Create Wikidata-based tools to facilitate and incentivize third-party data contributions. These tools are specifically aimed at enabling biomedical resource owners to easily contribute to Wikidata.

Progress report / preliminary data. During the previous funding period, we made significant progress on the technical infrastructure to read from and write to Wikidata. These developments were initially developed in the context of our Wikidata "bot", which was implemented in Python and served as our interface to the Wikidata API. Although the Wikidata API offers basic functions for creating and editing items and statements, our bot needed to perform many higher level functions. First, and most notably, we implemented detailed logic for item normalization to determine whether a record to be added could be mapped to an existing Wikidata item, whether a new item should be created, or whether a record needed manual review. (Incorrect resolution of an existing item or unnecessary creation of new items are both extremely problematic for data integration.) Second, we also added to our bot an efficient "fast run" mode that intelligently computed incremental edits, rather than completely writing the entire record on every update (increasing efficiency by over 10-fold). We also implemented many additional features related to authentication, logging, error handling, and scheduling.

Because we realized that this technical infrastructure would be generally useful for any bot performing massive database loading or editing (whether those bots were written by our team or other bot developers), we abstracted these features into a python library that we called **WikiDataIntegrator (WDI)** [44]. Now, WDI greatly simplifies the creation of bots to load new data sources. WDI has enabled new members of our team to quickly become productive bot developers. In addition, at least six additional external bot developers use WDI as the foundation of their code. (For example, see **Letters of Support from Dragan Espenschied**, **James Hare**, **Alex Pico,** and **Obi and Malachi Griffith**).

Significance. Our team was the first to perform systematic loading of biomedical data in Wikidata. Since then, the biomedical Wikidata community has grown, largely based on our team's direct collaborations and relationships. However, for Wikidata to succeed long-term, it cannot be based on our team's direct contributions and partnerships. Therefore, our effort to create a comprehensive knowledge base within Wikidata must also include activities that will stimulate adoption by the broader community. This aim is divided into two components: to decrease the technical barriers to data contribution, and to increase the positive incentives for data providers. If successful, we believe this effort will be of great interest to data contributors (see **Letters of Support from Alex Pico,** and **Obi and Malachi Griffith**) and significantly increase the rate, scale, and scope of biomedical data contributions to Wikidata.

A) Extend our Wikidata Integrator (WDI) python library into a fully-featured database connector. WDI will support data modeling and validation checks, as well as advanced logging and reporting functionality.

This subaim will focus on continued development of the WDI library to make it technically easy for data owners to quickly contribute data. In its current state, WDI simplifies the process of data loading relative to directly accessing the Wikidata API, and this subaim will further streamline that process. In addition to various software engineering improvements (like implementing multi-threading and batch processing capabilities), we will focus on implementing two key functionalities.

First, WDI will be extended to perform data validation based on any ShEx model, including the ones created in **Aim 1C**. When complete, WDI will be able to validate each record to be added or updated to confirm that the result will still conform to the model. Any edits that produce invalid records relative to the ShEx data model will throw an exception and not be written to Wikidata. This data validation step is analogous to validation relative to a table schema in a traditional relational database. This functionality will provide developers an additional level of confidence that loading of a data set is proceeding as intended, and avoiding the messy process of cleaning up erroneous edits after the fact.

Second, we will improve the logging and reporting features within WDI. Because we are utilizing an open community resource, we believe that it is important that our bots maintain full transparency in our operation. That means having easy traceability to the last time each bot task was run, seeing the success/failure status of each run, and the ability to examine any exceptions identified in the loading process. We have implemented a rudimentary logging system that outputs the raw bot logs to the bot's user page [45]. This system will be improved to include more user friendly views of the logs, including summary statistics. We will also create better capabilities for flagging records that need human review (for example, for records that fail ShEx validation). In many cases, volunteers from the Wikidata community are willing and able to resolve these edge cases, thereby reducing or eliminating the manual effort required by data providers.

We believe that these two steps will significantly decrease the technical friction for becoming a Wikidata data contributor, and that they will be welcomed by both bot developers and the broader Wikidata community. However, in the event that other roadblocks or issues are identified that prevent wider adoption of WDI (for example, through our Github issue tracker), then we would reprioritize our activities as appropriate to successfully achieve the broader Aim.

B) Generate automated reports of community contributions to relevant Wikidata items. These reports will notify curators when additions or edits are made, effectively serving as a universal community curation interface.

As both the complexity and the volume of curation tasks is rapidly increasing, many biocuration organizations have explored the use of community curation as a complement to professional curation [46-48], and still other efforts are based entirely on community curation [38, 49, 50]. However, building a community curation platform and stimulating community momentum are significant challenges. Here, we explore the potential of Wikidata to serve as a universal and extensible platform for community curation. The core of this platform will be an automated reporting system that will actively notify data owners of edits and additions to relevant Wikidata items. For example, WikiPathways and Reactome curators may want to be notified of edits to Wikidata Pathway items, and CIVIC curators would receive notification of edits on human genetic variants. This report will also include any data modeling issues that were identified using the relevant ShEx models created in **Aim 1C**. Database owners might then choose one of several options: reverting the edit if it is deemed to be incorrect, approving the edit for official inclusion in the database if it is correct, or approving a modified version of the edit based on refinement or improvement of the community contribution.

In addition to the technical benefit of using our reporting system, basing a community curation platform on Wikidata also has substantial benefits in terms of community building. Wikidata is already a large a growing community of volunteers, all of whom are interested in efficient and structured data management. The Wikidata community includes a diverse membership, and different groups already have efficient mechanisms for contributing to Wikidata. For example, individual domain experts can contribute individual edits using the web interface. Information scientists can perform bulk edits and imports of other structured data sources. And tool developers can create custom interfaces that utilize the Wikidata API and target specific user communities (e.g., WikiGenomes/Chlambase, described in **Aim 3**).

Finally, we are also in early talks with the Wikimedia Foundation to compute usage metrics for each item and statement within Wikidata based on the SPARQL query logs. When/if these plans materialize, these usage metrics will also be included in the data provider report, which in turn can be used by the data providers as evidence of community adoption in their own funding proposals.

Aim #3: Solicit “deep and narrow” contributions from targeted communities of domain experts. In contrast to the broad and shallow focus of most database imports, this aim will target information from experts that is not currently captured in any structured format.

Progress report / preliminary data. In the previous project period, one aim specifically addressed the information management challenges faced by the “Long Tail” of sequenced genomes. As the rate of genome sequencing increases, the number of organisms for which no model organism database exists will continue to increase. For these communities, we proposed to create a centralized and editable genome portal based on Wikidata. The result of that effort is WikiGenomes.org [51]. We loaded into Wikidata genome annotations for the 120 NCBI reference genomes [52] and the 394k genes contained therein. (New organisms can be easily added through our existing pipeline based on an NCBI Taxonomy ID.) WikiGenomes provides a web interface to query and browse those data, and also to contribute new annotations through a simple graphical user interface, which are then also written to Wikidata and available to other users. WikiGenomes is a fully hosted system – accessing and editing annotation data requires no computational expertise or infrastructure.

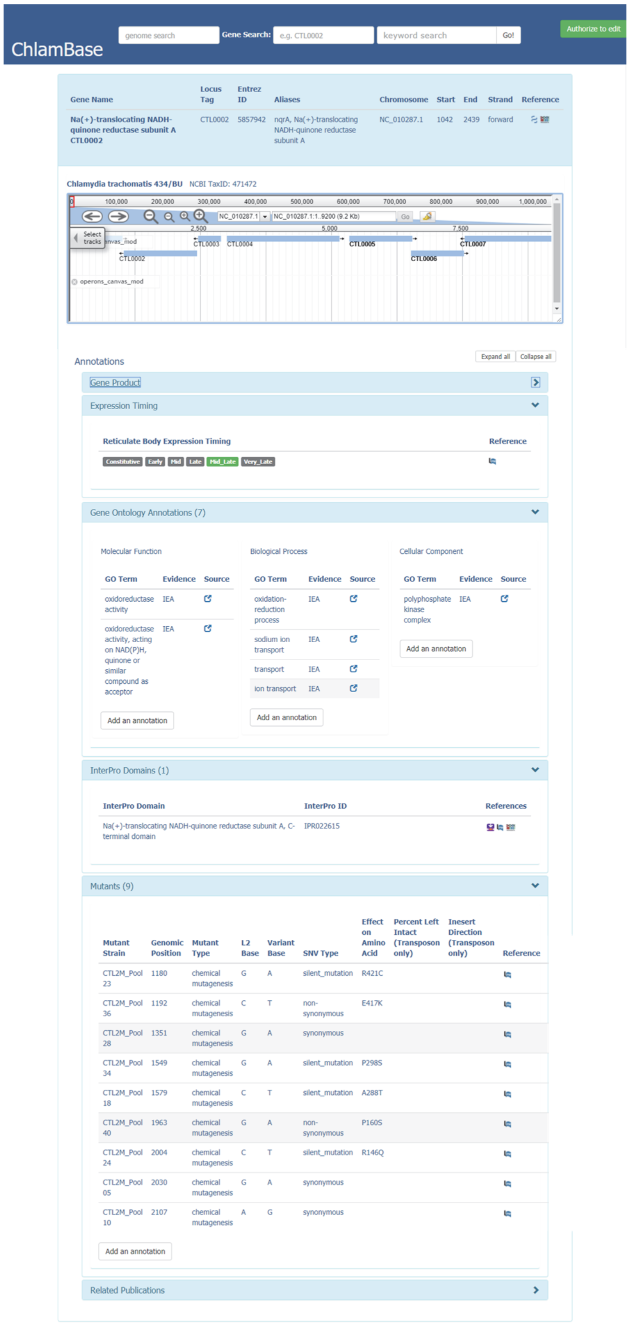
In a related initiative, we also created Chlambase.org, a fork of WikiGenomes that is tailored to the Chlamydia research community. Chlambase was created to serve a more focused research community than the general-purpose scope of WikiGenomes. Chlambase tests a complementary model in which communities can assume some of the computational burden in exchange for the flexibility to customize a portal for their researchers' specific needs. For example, Chlambase allows users to view (and add) annotations of available mutant lines [53], strain-specific orthologs, and developmental gene expression [54] (**Figure 6**).

Significance. Structured data and knowledge that have been captured in databases are unquestionably valuable. Loading these data sets directly (**Aim 1**) or facilitating their loading by data providers (**Aim 2**) will play a critical role in growing Wikidata into a comprehensive biomedical knowledge base. However, few would argue that this is a *complete* solution. Except for a few pockets of focused, deep curation, there is still a huge amount of biomedical knowledge that is not yet captured in any structured database and hence cannot be loaded into Wikidata.

This Aim tests the hypothesis that the best source of deep and focused knowledge is individual domain experts. Intuitively we know this, since we seek out those experts to give seminars and write review articles as means to educate the community. However, we currently do not have efficient interfaces for these experts to contribute to *structured* biocuration efforts, nor do we have effective incentives to motivate participation. This Aim addresses both of these issues.

The vision behind **Aims 3A** and **3B** is to attract expert contributors by providing them more efficient access to information than through any other tool that is currently available. WikiGenomes and Chlambase fill an information access need that is currently not being served, and are designed to appeal to a user's selfish interests in wanting to do research more efficiently. WikiGenomes and Chlambase then follow Wikipedia's model by pairing a useful tool with a very low barrier to contribution. Even if only a small percentage of visitors make a contribution, this structure sets up a positive feedback loop between increasing usage, increasing contributions, and increasing utility.

(NOTE: For clarity, preliminary data and significance related to **Aim 3C** is presented in that subaim below.)



Gene Name	Locus Tag	Entrez ID	Aliases	Chromosome	Start	End	Strand	Reference
Na(+)-translocating NADH-quinone reductase subunit A CTL0002	CTL0002	5857942	nqrA, Na(+)-translocating NADH-quinone reductase subunit A	NC_010287.1	1042	2439	forward	

Chlamydia trachomatis 434/BU NCBI TaxID: 471472

Reticulate Body Expression Timing	Reference
Constitutive Early Mid Late Mid_Late Very_Late	

Molecular Function

GO Term	Evidence	Source
oxidoreductase activity	IEA	
oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor	IEA	

Biological Process

GO Term	Evidence	Source
oxidation-reduction process	IEA	
sodium ion transport	IEA	
transport	IEA	
ion transport	IEA	

Cellular Component

GO Term	Evidence	Source
polyphosphate kinase complex	IEA	

InterPro Domain	InterPro ID	References
Na(+)-translocating NADH-quinone reductase subunit A, C-terminal domain	IPR022615	

Mutant Strain	Genomic Position	Mutant Type	L2 Base	Variant Base	SNV Type	Effect on Amino Acid	Percent Left Intact (Transposon only)	Inesert Direction (Transposon only)	Reference
CTL2M_Pool 23	1180	chemical mutagenesis	G	A	silent_mutation	R421C			
CTL2M_Pool 36	1192	chemical mutagenesis	C	T	non-synonymous	E417K			
CTL2M_Pool 28	1351	chemical mutagenesis	G	A	synonymous				
CTL2M_Pool 34	1549	chemical mutagenesis	G	A	silent_mutation	P298S			
CTL2M_Pool 18	1579	chemical mutagenesis	C	T	silent_mutation	A288T			
CTL2M_Pool 40	1963	chemical mutagenesis	G	A	non-synonymous	P160S			
CTL2M_Pool 24	2004	chemical mutagenesis	C	T	silent_mutation	R146Q			
CTL2M_Pool 05	2030	chemical mutagenesis	G	A	synonymous				
CTL2M_Pool 10	2107	chemical mutagenesis	A	G	synonymous				

Figure 6. The Chlambase.org portal for the Chlamydia research community. This interface was designed in close collaboration with Chlamydia researchers. The goal is to provide domain experts with an integrated gene report that is not available at any other online resource, including Chlamydia-specific data on genetic mutants and gene expression timing. Importantly, Chlambase facilitates both read and write access. All data, including the genome annotation data for JBrowse, are retrieved directly from Wikidata via the SPARQL interface.

A) Integrate our Wikidata-backed WikiGenomes application with the Apollo genome annotation tool. This synergy will combine our complementary strengths of a powerful user interface and robust community infrastructure.

JBrowse is a popular, open-source, web-based genome browser [55], and Apollo is a JBrowse plugin that empowers instantaneous and collaborative annotation of genomic features [56]. Both projects are mature initiatives with enthusiastic user communities. Importantly, they both already are widely used in the expert communities that this Aim seeks to harness.

Typical JBrowse and Apollo installations rely on creating and maintaining a local database. During the previous project period, we together modified the JBrowse data access layer to be able to also read from the Wikidata API [51]. This work was a core component of our WikiGenomes initiative, in which we embedded JBrowse in our web portal as a generic interface to browse genome annotations in Wikidata (currently the 120 NCBI reference genomes). We paired the read-only JBrowse application for genome annotations (e.g., gene boundaries) with a simple read-write interface for gene annotations (e.g., Gene Ontology annotations, loss of function mutants, operons) [51]. (In the context of our Driving Biological Project, we will also enable annotation of the organisms themselves.)

In the next project period, we will work on extending this work to a full read-write application for both genome and gene annotations in a more complete integration of JBrowse, Apollo, and WikiGenomes (**Figure 7**). We will do this by also modifying the data access layer of Apollo to optionally write to Wikidata instead of a local database. WikiGenomes will then continue to serve as a wrapper application, providing a unified interface to access annotations for any organism that has been loaded to Wikidata. We believe that the focus of each project is highly complementary, and that the combined strengths will result in a powerful tool for genomics research and community curation (see **Letters of Support from Suzanna Lewis and Ian Holmes**).

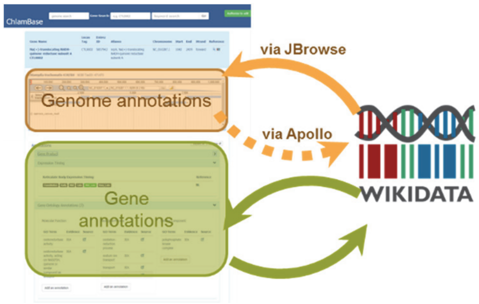
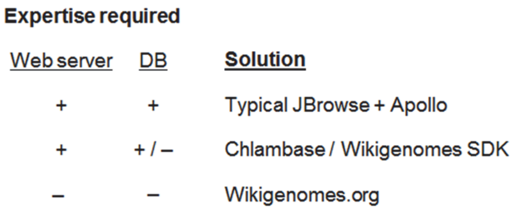


Figure 7. Schematic overview of the WikiGenomes / JBrowse / Apollo mashup. In the prior project period, we created a read-write interface between WikiGenomes (and Chlambase) and Wikidata for gene annotations. We also created a read-only interface between JBrowse and Wikidata. In Aim 3A, we will complete the final connector by refactoring Apollo to write genome annotation edits to Wikidata (dotted arrow).

B) Generalize the Chlambase application into a deployable Software Development Kit. This SDK will enable organism-specific genome databases to be created with minimal resources and administration effort.

Relative to WikiGenomes, our Chlambase application is targeted at deeper engagement and richer annotations, but tailored to a smaller and more focused user community. In this subaim, we will generalize the Chlambase website into a deployable software development kit (SDK). Success will be minimally defined by two criteria. First, a new focused genome portal must be easy to deploy. For example, we currently plan to base our SDK on a Docker container, which will greatly simplify the deployment process. We will also abstract all of the configurable elements of the application into a single configuration file. Second, the base application must be extensible via a simple plugin system. Initial examples of these plugins will be based on the Chlamydia-specific extensions (for mutant lines, strain-specific orthologs, and developmental gene expression) that we already added to Chlambase, though the system will be generic enough to allow many other types of annotations as well (e.g., properties of the organism itself, relevant literature notifications). Importantly, this plugin system will also allow the creation of hybrid instances that combine data in Wikidata with local databases (e.g., less-public data, or data not suitable for Wikidata). (See **Letter of Support from Kevin Hybiske**.)

From a data science and crowdsourcing perspective, we are interested in empirically testing which model produces more community annotation – a single generic interface (WikiGenomes) or many focused portals (like Chlambase). However, from a bioinformatics infrastructure perspective, it is likely that both models (and variations of them) will be useful depending on the specific community (**Figure 8**).



Expertise required

Web server	DB	Solution
+	+	Typical JBrowse + Apollo
+	+ / –	Chlambase / Wikigenomes SDK
–	–	Wikigenomes.org

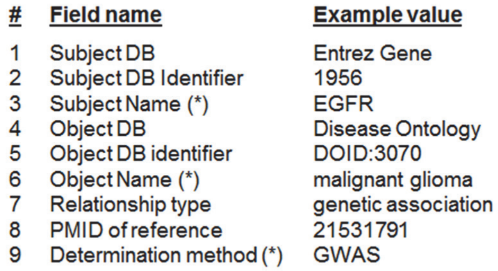
Figure 8. Options for model organism database configuration. Upon completion of Aim 3B, organism-specific communities will have three options based on the JBrowse / Apollo / Wikidata stack. If users would like the responsibility for maintaining a web server and a database (and the customizability that it enables), then the typical JBrowse / Apollo installation with local computer resources is the best option. If users want to assume no responsibility for maintaining any infrastructure, then the best option is to use our centrally hosted wikigenomes.org instance for both viewing and editing annotations. And users who want to be able to customize their interface (and possibly also their data sources) can do so with the WikiGenomes SDK as a starting point for their application.

We do not anticipate any technical issues with implementing the next phases of WikiGenomes and the Chlambase SDK. However, while we believe our interface designs are good and historically have been successful, it is certainly possible that these interfaces still do not substantially stimulate expert community annotations. In this event, we will divert some of our developer effort to perform a more detailed user-centered design process, starting with scientists in the Chlamydia research community.

<u>C) Capture structured data from review article authors through our partnership with the journal GENE.</u> This existing partnership utilizes a dual publication model to incentivize expert contributions.

Progress report / preliminary data. During the previous review period, we initiated the <u>Gene Wiki Reviews</u> series of invited review articles. In partnership with the journal GENE, we invited experts on individual genes or gene families to write a review article under a dual publication model. One article was peer-reviewed and published via the normal GENE editorial process, and the second was written in the form of a gene-specific Wikipedia article. This approach allowed us to align incentives by pairing contribution to Wikipedia with a traditional metric of academic credit [57]. To date, GENE has published 68 Gene Wiki Review articles and made significant edits to the corresponding 92 gene-specific Wikipedia pages [58].

Significance. While **Aims 3A** and **3B** focus on motivating community contribution by creating useful tools and lowering the barrier to editing, this subaim appeals to a different motivation. Here, we <u>pair community curation with traditional metrics of academic credit</u>. During the previous project period, we have seen how the dual publication model of the Gene Wiki Reviews series resulted in substantial community participation. Although the overall number of articles is relatively modest, each article is of high quality and covers the subject matter with considerable depth. This subaim will leverage the same invited reviews model and focus that effort toward aggregating structured biomedical knowledge.



#	Field name	Example value
1	Subject DB	Entrez Gene
2	Subject DB Identifier	1956
3	Subject Name (*)	EGFR
4	Object DB	Disease Ontology
5	Object DB identifier	DOID:3070
6	Object Name (*)	malignant glioma
7	Relationship type	genetic association
8	PMID of reference	21531791
9	Determination method (*)	GWAS

Figure 9. Data fields to be collected in the structured data template for the Gene Wiki Reviews series. Determination method will be an optional field, and we will provide a list of possible values for each relationship type. Subject Name and Object Name will be populated automatically using the corresponding identifiers, and their inclusion in the data table will be for sanity checking by the authors.

Approach. Here, we will extend the Gene Wiki Reviews series of invited review articles (see **Letter of Support from Andre van Wijnen**). Currently, there are two requirements for authors to participate in this initiative: an article for peer-review and publication via GENE, and an expanded article for Wikipedia. We will add a third requirement for authors to contribute <u>a structured data table that describes key biomedical relationships</u>. We will standardize a template that specifically targets the most common relationship types that Gene Wiki Reviews authors write about – associations between genes and diseases, variants and diseases, gene products and drugs, and genes and pathways.

For each relationship, we will require the author to provide enough information to create a corresponding Wikidata statement. Minimally, each entity will be described by a database type (e.g., UniProt, NCBI Gene) and database identifier for both the subject and the object, a relationship type that will be drawn from the list of Wikidata properties, and the PMID of a reference supporting the relationship (**Figure 9**). Initially, performing quality control on the data table may be a semi-manual process to ensure accuracy and fidelity with respect to the authors' intent. Once the data table is properly structured, our bot team will load it into Wikidata, which will trigger a second round of quality control checks based on the relevant ShEx models (**Aim 1C**).

Based on our experience with the existing Gene Wiki Reviews series, we believe that invited authors will be willing and able to provide the structured data table requested. However, if user feedback says otherwise, or if we see a substantial decline in participation in the Gene Wiki Reviews series, then we will divert some resources to create user-friendly tools and interfaces for data entry. In particular, we believe the metadata authoring tools from the CEDAR center may also be useful for this task [59].

We fully acknowledge that the data tables we receive will not be exhaustive. Nevertheless, we believe that they will represent positive progress toward our goal of getting high quality contributions directly from domain experts, and if successful, this subaim will serve as a generalizable model for an expanded initiative.

Driving Biological Project: Wikidata as a knowledge base for unbiased pathogen detection. Focusing on this use case in infectious disease research will ensure that our Aims remain rooted in efforts to solve real-world biomedical problems.

Unexplained acute febrile illnesses (UAFIs) account for millions of visits to hospitals and clinics every month, and the majority of these diseases (30-40% of overall cases) are believed to be caused by infections [60]. The typical diagnostic strategy involves targeted testing of only suspected pathogens, and sequentially testing for other pathogens if the initial results are negative. Unfortunately, the lack of definitive and timely diagnosis can lead to costly hospital stays and even result in deadly outcomes (for example, [61]). Moreover, it can also result in delays in the public health response in the critical early stages of infectious disease outbreaks, such as the severe delays in detecting Ebola virus during the 2013-2016 epidemic [62, 63], and Zika virus during the ongoing epidemic [64-66].

As the cost and assay times of genomic sequencing continues to fall, the potential of unbiased metagenomic sequencing for infectious disease diagnosis becomes increasingly feasible [67]. We are partnering with the Andersen Lab at TSRI, who is developing exactly such a method (see **Letter of Support from Kristian Andersen**). While metagenomic sequencing data from blood will be the primary mechanism for detecting the presence of potential pathogens, additional factors will also play an important role in differentiating causative from non-causative infections. Those other factors may include associated diagnostic clues (e.g., pattern of organ involvement, lab test results) and phenotypes (e.g., chills, weight loss, vasculitis) [68], or the prevalence of the disease and/or vector in the geographic region in which the case originated.

In this DBP, we will explore the use of Wikidata to organize and integrate all the knowledge relevant to infectious disease pathogens. The tools and approaches of each aim will be adapted to the goal of developing a knowledge base that supports a method for unbiased pathogen detection. Such a knowledge base will also be useful studies of genetic epidemiology [62, 66] and phylogeographic analyses [69].

In the context of **Aim 1**, we will load into Wikidata several key data sets and vocabularies that are essential to unbiased pathogen detection and minimally covering the NIAID Priority Pathogen list [70]. Specifically, we will load historical outbreak data for viral and bacterial pathogens that was previously curated in the GeMInA database [71] and the gazetteer ontology of geographic locations [72] (both led or co-led by Co-PI Schriml), as well as relevant data from the NIAID Bioinformatics Resource Centers [73]. The result of this effort will be the most comprehensive knowledge base of infectious disease relationships, including disease-pathogen, disease-phenotype, disease-drug, pathogen-vector, disease-prevalence, and vector-prevalence links.

In the context of **Aim 2**, the automated reports of Wikidata edits relevant to infectious disease items and relationships will be the model on which our generalized reporting framework will be based. This feedback is especially important to this DBP because while the GeMInA database has defined a useful schema for modeling outbreaks, the database itself has not been updated since 2010. There will be a significant “catch up” period in which many additions will be made by community members, and this reporting mechanism will be the primary mechanism by which new edits get reviewed by other domain experts and curators.

In the context of **Aim 3**, we will populate WikiGenomes with all NIAID pathogens to facilitate community curation by domain experts. We will also extend the WikiGenomes interface to enable annotation of not just genes and genomic elements, but also of the organism itself and its relevance to disease (using the data models and ontologies described for **Aim 1**). Based on our collective expertise running hackathons and annotation jamborees, we will also run conference-based events modeled after past successful examples [74].

Summary. This proposal focuses on the challenge of making data more Findable, Accessible, Interoperable, and Reusable (FAIR). Biomedical research is continually accelerating in terms of its ability to produce new data and new knowledge, and FAIRness ensures that they have maximal impact on the speed and efficiency of scientific research. We believe that Wikidata is an ideal platform to advance FAIR principles, and that this proposal offers a robust mechanism to harness Wikidata to advance biomedical discovery.

Progress report publication list

1: Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, Ainscough BJ, Ramirez CA, Rieke DT, Kujan L, Barnell EK, Wagner AH, Skidmore ZL, Wollam A, Liu CJ, Jones MR, Bilski RL, Lesurf R, Feng YY, Shah NM, Bonakdar M, Trani L, Matlock M, Ramu A, Campbell KM, Spies GC, Graubert AP, Gangavarapu K, Eldred JM, Larson DE, Walker JR, Good BM, Wu C, Su AI, Dienstmann R, Margolin AA, Tamborero D, Lopez-Bigas N, Jones SJ, Bose R, Spencer DH, Wartman LD, Wilson RK, Mardis ER, Griffith OL. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet. 2017 Jan 31;49(2):170-174. doi: 10.1038/ng.3774. PubMed PMID: 28138153; PubMed Central PMCID: PMC5367263.

2: Tsueng G, Good BM, Ping P, Golemis E, Hanukoglu I, van Wijnen AJ, Su AI. Gene Wiki Reviews-Raising the quality and accessibility of information about the human genome. Gene. 2016 Nov 5;592(2):235-8. doi: 10.1016/j.gene.2016.04.053. Epub 2016 May 2. PubMed PMID: 27150585.

3: Putman TE, Burgstaller-Muehlbacher S, Waagmeester A, Wu C, Su AI, Good BM. Centralizing content and distributing labor: a community model for curating the very long tail of microbial genomes. Database (Oxford). 2016 Mar 28;2016. pii: baw028. doi: 10.1093/database/baw028. Print 2016. PubMed PMID: 27022157; PubMed Central PMCID: PMC4822648.

4: 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 Mar 17;2016. pii: baw015. doi: 10.1093/database/baw015. Print 2016. PubMed PMID: 26989148; PubMed Central PMCID: PMC4795929.

5: Hettne KM, Thompson M, van Haagen HH, van der Horst E, Kaliyaperumal R, Mina E, Tatum Z, Laros JF, van Mulligen EM, Schuemie M, Aten E, Li TS, Bruskiewich R, Good BM, Su AI, Kors JA, den Dunnen J, van Ommen GJ, Roos M, 't Hoen PA, Mons B, Schultes EA. The Implicitome: A Resource for Rationalizing Gene-Disease Associations. PLoS One. 2016 Feb 26;11(2):e0149621. doi: 10.1371/journal.pone.0149621. eCollection 2016. PubMed PMID: 26919047; PubMed Central PMCID: PMC4769089.

6: Khare R, Good BM, Leaman R, Su AI, Lu Z. Crowdsourcing in biomedicine: challenges and opportunities. Brief Bioinform. 2016 Jan;17(1):23-32. doi: 10.1093/bib/bbv021. Epub 2015 Apr 17. Review. PubMed PMID: 25888696; PubMed Central PMCID: PMC4719068.

7: Good BM, Loguercio S, Griffith OL, Nanis M, Wu C, Su AI. The cure: design and evaluation of a crowdsourcing game for gene selection for breast cancer survival prediction. JMIR Serious Games. 2014 Jul 29;2(2):e7. doi: 10.2196/games.3350. PubMed PMID: 25654473; PubMed Central PMCID: PMC4307816.

8: Good BM, Nanis M, Wu C, Su AI. Microtask crowdsourcing for disease mention annotation in PubMed abstracts. Pac Symp Biocomput. 2015:282-93. PubMed PMID: 25592589; PubMed Central PMCID: PMC4299946.

9: Good BM, Ainscough BJ, McMichael JF, Su AI, Griffith OL. Organizing knowledge to enable personalization of medicine in cancer. Genome Biol. 2014 Aug 27;15(8):438. doi: 10.1186/s13059-014-0438-7. PubMed PMID: 25222080; PubMed Central PMCID: PMC4281950.

10: Su AI, Good BM, van Wijnen AJ. Gene Wiki Reviews: marrying crowdsourcing with traditional peer review. Gene. 2013 Dec 1;531(2):125. doi: 10.1016/j.gene.2013.08.093. Epub 2013 Sep 5. PubMed PMID: 24012870.

11: Good BM, Su AI. Crowdsourcing for bioinformatics. Bioinformatics. 2013 Aug 15;29(16):1925-33. doi: 10.1093/bioinformatics/btt333. Epub 2013 Jun 19. Review. PubMed PMID: 23782614; PubMed Central PMCID: PMC3722523.

REFERENCES

1. Toledo-Arana A, Dussurget O, Nikitas G, Sesto N, Guet-Revillet H, Balestrino D, Loh E, Gripenland J, Tiensuu T, Vaitkevicius K, Barthelemy M, Vergassola M, Nahori M-A, Soubigou G, Régnault B, Coppée J-Y, Lecuit M, Johansson J, Cossart P. The Listeria transcriptional landscape from saprophytism to virulence. Nature. 2009;459(7249):950-6. doi: 10.1038/nature08080. PMID: 19448609.
2. Genewiki-ShEx/wikidata_human-genes.shex at master · SuLab/Genewiki-ShEx 2017. Available from: https://github.com/SuLab/Genewiki-ShEx/blob/master/genes/wikidata_human-genes.shex.
3. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, t Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. doi: 10.1038/sdata.2016.18. PMID: 26978244; PMCID: PMC4792175.
4. Oxenham S. Legal confusion threatens to slow data science. Nature. 2016;536(7614):16-7. doi: 10.1038/536016a. PMID: 27488781.
5. Good BM, Nanis M, Wu C, Su AI. Microtask crowdsourcing for disease mention annotation in PubMed abstracts. Pac Symp Biocomput. 2015:282-93. PMID: 25592589; PMCID: PMC4299946.
6. Tsueng G, Nanis SM, Fouquier J, Good BM, Su AI. Citizen Science for Mining the Biomedical Literature. Citizen Science: Theory and Practice. 2016;1(2):14. doi: 10.5334/cstp.56.
7. User:ProteinBoxBot/evidence - Wikidata [2017/6/22]. Available from: https://www.wikidata.org/wiki/User:ProteinBoxBot/evidence.
8. 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. doi: 10.1093/database/baw015. PMID: 26989148; PMCID: PMC4795929.
9. Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, Ermolaeva O, Tolstoy I, Tatusova T, Pruitt KD, Maglott DR, Murphy TD. Gene: a gene-centered information resource at NCBI. Nucleic Acids Res. 2015;43(Database issue):D36-42. doi: 10.1093/nar/gku1055. PMID: 25355515; PMCID: PMC4383897.
10. Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Gil L, Giron CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Juettemann T, Keenan S, Laird MR, Lavidas I, Maurel T, McLaren W, Moore B, Murphy DN, Nag R, Newman V, Nuhn M, Ong CK, Parker A, Patricio M, Riat HS, Sheppard D, Sparrow H, Taylor K, Thormann A, Vullo A, Walts B, Wilder SP, Zadissa A, Kostadima M, Martin FJ, Muffato M, Perry E, Ruffier M, Staines DM, Trevanion SJ, Cunningham F, Yates A, Zerbino DR, Flicek P. Ensembl 2017. Nucleic Acids Res. 2017;45(D1):D635-D42. doi: 10.1093/nar/gkw1104. PMID: 27899575; PMCID: PMC5210575.
11. The UniProt C. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45(D1):D158-D69. doi: 10.1093/nar/gkw1099. PMID: 27899622; PMCID: PMC5210571.
12. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY, Dosztanyi Z, El-Gebali S, Fraser M, Gough J, Haft D, Holliday GL, Huang H, Huang X, Letunic I, Lopez R, Lu S, Marchler-Bauer A, Mi H, Mistry J, Natale DA, Necci M, Nuka G, Orengo CA, Park Y, Pesseat S, Piovesan D, Potter SC, Rawlings ND, Redaschi N, Richardson L, Rivoire C, Sangrador-Vegas A, Sigrist C, Sillitoe I, Smithers B, Squizzato S, Sutton G, Thanki N, Thomas PD, Tosatto SC, Wu CH, Xenarios I, Yeh LS, Young SY, Mitchell AL. InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res. 2017;45(D1):D190-D9. doi: 10.1093/nar/gkw1107. PMID: 27899635; PMCID: PMC5210578.
13. Schriml LM, Arze C, Nadendla S, Chang Y-WW, Mazaitis M, Felix V, Feng G, Kibbe WA. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res. 2012;40(Database issue):D940-6. doi: 10.1093/nar/gkr972. PMID: 22080554; PMCID: PMC3245088.
14. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, Pendlington ZM, Welter D, Burdett T, Hindorff L, Flicek P, Cunningham F, Parkinson H. The new NHGRI-

EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017;45(D1):D896-D901. doi: 10.1093/nar/gkw1133. PMID: 27899670; PMCID: PMC5210590.

15. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrian-Uhalte E, Davies M, Dedman N, Karlsson A, Magarinos MP, Overington JP, Papadatos G, Smit I, Leach AR. The ChEMBL database in 2017. Nucleic Acids Res. 2017;45(D1):D945-D54. doi: 10.1093/nar/gkw1074. PMID: 27899562; PMCID: PMC5210557.
16. National Drug File - Reference Terminology Source Information: U.S. National Library of Medicine. Available from: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT/.
17. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, Bryant SH. PubChem Substance and Compound databases. Nucleic Acids Res. 2016;44(D1):D1202-13. doi: 10.1093/nar/gkv951. PMID: 26400175; PMCID: PMC4702940.
18. Kutmon M, Riutta A, Nunes N, Hanspers K, Willighagen EL, Bohler A, Mélius J, Waagmeester A, Sinha SR, Miller R, Coort SL, Cirillo E, Smeets B, Evelo CT, Pico AR. WikiPathways: capturing the full diversity of pathway knowledge. Nucleic Acids Res. 2016;44(D1):D488-94. doi: 10.1093/nar/gkv1024. PMID: 26481357; PMCID: PMC4702772.
19. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S, Matthews L, May B, Milacic M, Rothfels K, Shamovsky V, Webber M, Weiser J, Williams M, Wu G, Stein L, Hermjakob H, D'Eustachio P. The Reactome pathway Knowledgebase. Nucleic Acids Res. 2016;44(D1):D481-7. doi: 10.1093/nar/gkv1351. PMID: 26656494; PMCID: PMC4702931.
20. Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, Ainscough BJ, Ramirez CA, Rieke DT, Kujan L, Barnell EK, Wagner AH, Skidmore ZL, Wollam A, Liu CJ, Jones MR, Bilski RL, Lesurf R, Feng Y-Y, Shah NM, Bonakdar M, Trani L, Matlock M, Ramu A, Campbell KM, Spies GC, Graubert AP, Gangavarapu K, Eldred JM, Larson DE, Walker JR, Good BM, Wu C, Su AI, Dienstmann R, Margolin AA, Tamborero D, Lopez-Bigas N, Jones SJM, Bose R, Spencer DH, Wartman LD, Wilson RK, Mardis ER, Griffith OL. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet. 2017;49(2):170-4. doi: 10.1038/ng.3774. PMID: 28138153; PMCID: PMC5367263.
21. Chibucos MC, Mungall CJ, Balakrishnan R, Christie KR, Huntley RP, White O, Blake JA, Lewis SE, Giglio M. Standardized description of scientific evidence using the Evidence Ontology (ECO). Database. 2014;2014. doi: 10.1093/database/bau075. PMID: 25052702; PMCID: PMC4105709.
22. User:ProteinBoxBot/SPARQL Examples - Wikidata. Available from: https://www.wikidata.org/wiki/User:ProteinBoxBot/SPARQL_Examples.
23. Medical Subject Headings - Home Page: U.S. National Library of Medicine; 1999 [updated 1999/9/1]. Available from: https://www.nlm.nih.gov/mesh/meshhome.html.
24. Sewell W. MEDICAL SUBJECT HEADINGS IN MEDLARS. Bull Med Libr Assoc. 1964;52:164-70. PMID: 14119288; PMCID: PMC198088.
25. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, Jonquet C, Rubin DL, Storey M-A, Chute CG, Musen MA. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. 2009;37(Web Server issue):W170-3. doi: 10.1093/nar/gkp440. PMID: 19483092; PMCID: PMC2703982.
26. The Whelming › The Game Is On. Available from: http://magnusmanske.de/wordpress/?p=203.
27. The Whelming › Enter the Distributed Game. Available from: http://magnusmanske.de/wordpress/?p=362.
28. The OBO Foundry. Available from: http://www.obofoundry.org/.
29. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007;25(11):1251-5. doi: 10.1038/nbt1346. PMID: 17989687; PMCID: PMC2814061.
30. Obofoundry. Create guidelines for OBO maintainers who want to be included in Wikidata · Issue #285 · OBOFoundry/OBOFoundry.github.io. Available from: https://github.com/OBOFoundry/OBOFoundry.github.io/issues/285.

31. Su A. Open Data should mean CC0, not CC-BY | The Su Lab 2016 [updated 2016/8/2]. Available from: http://sulab.org/2016/08/open-data-should-mean-cc0/.
32. Haendel M, Mungall C, Su A, Robinson P, Chute C, B Altman R, Payne PRO, Lawler M, Oprea TI, Willbanks J, Srinivasan S, Hunter L, Sim I, McDonald S, Mooney S, Smedley D, Ganley E, Kenall A, Clark T, Goble C, Dumontier M, Holmes K, Diekans M, Zell A, Overby C, Glusman G, Carmody L, Jiang G, Munos-Torres M, Hoatlin M, Goecks J, Jongeneel V, Bittker J, Gourdine J-P, Brush MH, Zhu RL, Mangravite L, Tyler B, D Wilkinson M, Crusoe MR, Mazumder R, P Tatonetti N, D'Eustachio P, Vasilevsky N, McMurry J, Champieux R. Request for Community partnership in data resource licensing planning. figshare. 2017. doi: 10.6084/m9.figshare.4972709.v1.
33. Open: OBO Technical WG. Available from: http://www.obofoundry.org/principles/fp-001-open.html.
34. Cohen D. CC0 (+BY) | Dan Cohen 2013. Available from: http://www.dancohen.org/2013/11/26/cc0-by/.
35. Villa L. Copyleft and data: databases as poor subject 2016. Available from: http://lu.is/blog/2016/09/14/copyleft-and-data-databases-as-poor-subject/.
36. Symptom Ontology: OBO Technical WG. Available from: http://www.obofoundry.org/ontology/symp.html.
37. Pathogen Transmission Ontology: OBO Technical WG. Available from: http://www.obofoundry.org/ontology/trans.html.
38. Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo CT, Pico AR. WikiPathways: building research communities on biological pathways. Nucleic Acids Res. 2012;40(Database issue):D1301-7. doi: 10.1093/nar/gkr1074. PMID: 22096230; PMCID: PMC3245032.
39. Cancer Genome Interpreter - Identification of therapeutically actionable genomic alterations in tumors. Available from: https://www.cancergenomeinterpreter.org/biomarkers.
40. Consider switch to CC-0 or export of a distinct CC-0 subset for wikidata · Issue #147 · oborel/obo-relations. Available from: https://github.com/oborel/obo-relations/issues/147.
41. proposed licenses · EnvironmentOntology/environmental-exposure-ontology@24806ab 2017. Available from: https://github.com/EnvironmentOntology/environmental-exposure-ontology/commit/24806aba2f3b3514873c83f547fd766bf7dd5551.
42. Shape Expressions Language 2.0. Available from: http://shex.io/shex-semantics/index.html.
43. Shapes Constraint Language (SHACL). Available from: https://www.w3.org/TR/shacl/.
44. SuLab. SuLab/WikidataIntegrator. Available from: https://github.com/SuLab/WikidataIntegrator.
45. User:ProteinBoxBot/Bot Status - Wikidata. Available from: https://www.wikidata.org/wiki/User:ProteinBoxBot/Bot_Status.
46. Howe KL, Bolt BJ, Cain S, Chan J, Chen WJ, Davis P, Done J, Down T, Gao S, Grove C, Harris TW, Kishore R, Lee R, Lomax J, Li Y, Muller H-M, Nakamura C, Nuin P, Paulini M, Raciti D, Schindelman G, Stanley E, Tuli MA, Van Auken K, Wang D, Wang X, Williams G, Wright A, Yook K, Berriman M, Kersey P, Schedl T, Stein L, Sternberg PW. WormBase 2016: expanding to enable helminth genomic research. Nucleic Acids Res. 2016;44(D1):D774-80. doi: 10.1093/nar/gkv1217. PMID: 26578572; PMCID: PMC4702863.
47. Bunt SM, Grumbling GB, Field HI, Marygold SJ, Brown NH, Millburn GH, FlyBase C. Directly e-mailing authors of newly published papers encourages community curation. Database. 2012;2012:bas024. doi: 10.1093/database/bas024. PMID: 22554788; PMCID: PMC3342516.
48. Bhartiya D, Laddha SV, Mukhopadhyay A, Scaria V. miRvar: A comprehensive database for genomic variations in microRNAs. Hum Mutat. 2011;32(6):E2226-45. doi: 10.1002/humu.21482. PMID: 21618345.
49. Zhang Z, Sang J, Ma L, Wu G, Wu H, Huang D, Zou D, Liu S, Li A, Hao L, Tian M, Xu C, Wang X, Wu J, Xiao J, Dai L, Chen L-L, Hu S, Yu J. RiceWiki: a wiki-based database for community curation of rice genes. Nucleic Acids Res. 2014;42(Database issue):D1222-8. doi: 10.1093/nar/gkt926. PMID: 24136999; PMCID: PMC3964990.
50. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, Porto C, Bouslimani A, Melnik AV, Meehan MJ, Liu W-T, Crüsemann M, Boudreau PD, Esquenazi E, Sandoval-Calderón M, Kersten RD, Pace LA, Quinn RA, Duncan KR, Hsu C-C, Floros DJ, Gavilan RG, Kleigrewe K, Northen T, Dutton RJ, Parrot D, Carlson EE, Aigle B, Michelsen CF,

Jelsbak L, Sohlenkamp C, Pevzner P, Edlund A, McLean J, Piel J, Murphy BT, Gerwick L, Liaw C-C, Yang Y-L, Humpf H-U, Maansson M, Keyzers RA, Sims AC, Johnson AR, Sidebottom AM, Sedio BE, Klitgaard A, Larson CB, Boya P CA, Torres-Mendoza D, Gonzalez DJ, Silva DB, Marques LM, Demarque DP, Pociute E, O'Neill EC, Briand E, Helfrich EJN, Granatosky EA, Glukhov E, Ryffel F, Houson H, Mohimani H, Kharbush JJ, Zeng Y, Vorholt JA, Kurita KL, Charusanti P, McPhail KL, Nielsen KF, Vuong L, Elfeki M, Traxler MF, Engene N, Koyama N, Vining OB, Baric R, Silva RR, Mascuch SJ, Tomasi S, Jenkins S, Macherla V, Hoffman T, Agarwal V, Williams PG, Dai J, Neupane R, Gurr J, Rodríguez AMC, Lamsa A, Zhang C, Dorrestein K, Duggan BM, Almaliti J, Allard P-M, Phapale P, Nothias L-F, Alexandrov T, Litaudon M, Wolfender J-L, Kyle JE, Metz TO, Peryea T, Nguyen D-T, VanLeer D, Shinn P, Jadhav A, Müller R, Waters KM, Shi W, Liu X, Zhang L, Knight R, Jensen PR, Palsson BØ, Pogliano K, Linington RG, Gutiérrez M, Lopes NP, Gerwick WH, Moore BS, Dorrestein PC, Bandeira N. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol. 2016;34(8):828-37. doi: 10.1038/nbt.3597. PMID: 27504778; PMCID: PMC5321674.

51. Putman TE, Lelong S, Burgstaller-Muehlbacher S, Waagmeester A, Diesh C, Dunn N, Munoz-Torres M, Stupp GS, Wu C, Su AI, Good BM. WikiGenomes: an open web application for community consumption and curation of gene annotation data in Wikidata. Database (Oxford). 2017;2017(1). doi: 10.1093/database/bax025. PMID: 28365742; PMCID: PMC5467579.
52. Reference and Representative Genomes. Available from: https://www.ncbi.nlm.nih.gov/genome/browse/reference/.
53. Kokes M, Dunn JD, Granek JA, Nguyen BD, Barker JR, Valdivia RH, Bastidas RJ. Integrating chemical mutagenesis and whole-genome sequencing as a platform for forward and reverse genetic analysis of Chlamydia. Cell Host Microbe. 2015;17(5):716-25. doi: 10.1016/j.chom.2015.03.014. PMID: 25920978; PMCID: PMC4418230.
54. Nicholson TL, Olinger L, Chong K, Schoolnik G, Stephens RS. Global stage-specific gene regulation during the developmental cycle of Chlamydia trachomatis. J Bacteriol. 2003;185(10):3179-89. PMID: 12730178; PMCID: PMC154084.
55. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. JBrowse: a next-generation genome browser. Genome Res. 2009;19(9):1630-8. doi: 10.1101/gr.094607.109. PMID: 19570905; PMCID: PMC2752129.
56. Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, Stein L, Holmes IH, Elsik CG, Lewis SE. Web Apollo: a web-based genomic annotation editing platform. Genome Biol. 2013;14(8):R93. doi: 10.1186/gb-2013-14-8-r93. PMID: 24000942; PMCID: PMC4053811.
57. Tsueng G, Good BM, Ping P, Golemis E, Hanukoglu I, van Wijnen AJ, Su AI. Gene Wiki Reviews-Raising the quality and accessibility of information about the human genome. Gene. 2016;592(2):235-8. doi: 10.1016/j.gene.2016.04.053. PMID: 27150585.
58. Gene Wiki Review - Virtual Special Issue. Available from: http://www.sciencedirect.com/science/journal/03781119/vsi.
59. Musen MA, Bean CA, Cheung KH, Dumontier M, Durante KA, Gevaert O, Gonzalez-Beltran A, Khatri P, Kleinstein SH, O'Connor MJ, Pouliot Y, Rocca-Serra P, Sansone SA, Wiser JA, team C. The center for expanded data annotation and retrieval. J Am Med Inform Assoc. 2015;22(6):1148-52. doi: 10.1093/jamia/ocv048. PMID: 26112029; PMCID: PMC5009916.
60. Cunha BA. Fever of unknown origin. Infect Dis Clin North Am. 1996;10(1):111-27. PMID: 8698986.
61. Victoria Knight C. Tick-borne disease suspected in 2-year-old's death 2017 [updated 2017/6/8]. Available from: http://www.cnn.com/2017/06/08/health/indianapolis-girl-dies-tick-bite-bn/index.html.
62. Holmes EC, Dudas G, Rambaut A, Andersen KG. The evolution of Ebola virus: Insights from the 2013-2016 epidemic. Nature. 2016;538(7624):193–200. doi: 10.1038/nature19790. PMID: 27734858.
63. Gire SK, Goba A, Andersen KG, Sealfon RS, Park DJ, Kanneh L, Jalloh S, Momoh M, Fullah M, Dudas G, Wohl S, Moses LM, Yozwiak NL, Winnicki S, Matranga CB, Malboeuf CM, Qu J, Gladden AD, Schaffner SF, Yang X, Jiang PP, Nekoui M, Colubri A, Coomber MR, Fonnie M, Moigboi A, Gbakie M, Kamara FK, Tucker V, Konuwa E, Saffa S, Sellu J, Jalloh AA, Kovoma A, Koninga J, Mustapha I, Kargbo K, Foday M, Yillah M, Kanneh F, Robert W, Massally JL, Chapman SB, Bochicchio J, Murphy C, Nusbaum C, Young S, Birren BW, Grant DS, Scheiffelin JS, Lander ES, Happi C, Gevao SM, Gnirke A,

Rambaut A, Garry RF, Khan SH, Sabeti PC. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 2014;345(6202):1369–72. doi: 10.1126/science.1259657. PMID: 25214632; PMCID: PMC4431643.

64. Metsky HC, Matranga CB, Wohl S, Schaffner SF, Freije CA, Winnicki SM, West K, Qu J, Baniecki ML, Gladden-Young A, Lin AE, Tomkins-Tinch CH, Ye SH, Park DJ, Luo CY, Barnes KG, Shah RR, Chak B, Barbosa-Lima G, Delatorre E, Vieira YR, Paul LM, Tan AL, Barcellona CM, Porcelli MC, Vasquez C, Cannons AC, Cone MR, Hogan KN, Kopp EW, Anzinger JJ, Garcia KF, Parham LA, Ramírez RMG, Montoya MCM, Rojas DP, Brown CM, Hennigan S, Sabina B, Scotland S, Gangavarapu K, Grubaugh ND, Oliveira G, Robles-Sikisaka R, Rambaut A, Gehrke L, Smole S, Halloran ME, Villar L, Mattar S, Lorenzana I, Cerbino-Neto J, Valim C, Degrave W, Bozza PT, Gnirke A, Andersen KG, Isern S, Michael SF, Bozza FA, Souza TML, Bosch I, Yozwiak NL, MacInnis BL, Sabeti PC. Zika virus evolution and spread in the Americas. Nature. 2017;546(7658):411-5. doi: 10.1038/nature22402. PMID: 28538734.

65. Faria NR, Quick J, Claro IM, Thézé J, de Jesus JG, Giovanetti M, Kraemer MUG, Hill SC, Black A, da Costa AC, Franco LC, Silva SP, Wu CH, Raghwani J, Cauchemez S, du Plessis L, Verotti MP, de Oliveira WK, Carmo EH, Coelho GE, Santelli ACFS, Vinhal LC, Henriques CM, Simpson JT, Loose M, Andersen KG, Grubaugh ND, Somasekar S, Chiu CY, Muñoz-Medina JE, Gonzalez-Bonilla CR, Arias CF, Lewis-Ximenez LL, Baylis SA, Chieppe AO, Aguiar SF, Fernandes CA, Lemos PS, Nascimento BLS, Monteiro HAO, Siqueira IC, de Queiroz MG, de Souza TR, Bezerra JF, Lemos MR, Pereira GF, Loudal D, Moura LC, Dhalia R, França RF, Magalhães T, Marques ET, Jr., Jaenisch T, Wallau GL, de Lima MC, Nascimento V, de Cerqueira EM, de Lima MM, Mascarenhas DL, Neto JPM, Levin AS, Tozetto-Mendoza TR, Fonseca SN, Mendes-Correa MC, Milagres FP, Segurado A, Holmes EC, Rambaut A, Bedford T, Nunes MRT, Sabino EC, Alcantara LCJ, Loman NJ, Pybus OG. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Nature. 2017;546(7658):406-10. doi: 10.1038/nature22401. PMID: 28538727.

66. Grubaugh ND, Ladner JT, Kraemer MUG, Dudas G, Tan AL, Gangavarapu K, Wiley MR, White S, Thézé J, Magnani DM, Prieto K, Reyes D, Bingham AM, Paul LM, Robles-Sikisaka R, Oliveira G, Pronty D, Barcellona CM, Metsky HC, Baniecki ML, Barnes KG, Chak B, Freije CA, Gladden-Young A, Gnirke A, Luo C, MacInnis B, Matranga CB, Park DJ, Qu J, Schaffner SF, Tomkins-Tinch C, West KL, Winnicki SM, Wohl S, Yozwiak NL, Quick J, Fauver JR, Khan K, Brent SE, Reiner RC, Jr., Lichtenberger PN, Ricciardi MJ, Bailey VK, Watkins DI, Cone MR, Kopp EWt, Hogan KN, Cannons AC, Jean R, Monaghan AJ, Garry RF, Loman NJ, Faria NR, Porcelli MC, Vasquez C, Nagle ER, Cummings DAT, Stanek D, Rambaut A, Sanchez-Lockhart M, Sabeti PC, Gillis LD, Michael SF, Bedford T, Pybus OG, Isern S, Palacios G, Andersen KG. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. Nature. 2017;546(7658):401-5. doi: 10.1038/nature22400. PMID: 28538723.

67. Mulcahy-O'grady H, Workentine ML. The Challenge and Potential of Metagenomics in the Clinic. Front Immunol. 2016;7:29. doi: 10.3389/fimmu.2016.00029. PMID: 26870044; PMCID: PMC4737888.

68. Cunha BA, Lortholary O, Cunha CB. Fever of unknown origin: a clinical approach. Am J Med. 2015;128(10):1138.e1-.e15. doi: 10.1016/j.amjmed.2015.06.001. PMID: 26093175.

69. Dudas G, Carvalho LM, Bedford T, Tatem AJ, Baele G, Faria NR, Park DJ, Ladner JT, Arias A, Asogun D, Bielejec F, Caddy SL, Cotten M, D'Ambrozio J, Dellicour S, Di Caro A, Diclaro JW, Duraffour S, Elmore MJ, Fakoli LS, Faye O, Gilbert ML, Gevao SM, Gire S, Gladden-Young A, Gnirke A, Goba A, Grant DS, Haagmans BL, Hiscox JA, Jah U, Kugelman JR, Liu D, Lu J, Malboeuf CM, Mate S, Matthews DA, Matranga CB, Meredith LW, Qu J, Quick J, Pas SD, Phan MVT, Pollakis G, Reusken CB, Sanchez-Lockhart M, Schaffner SF, Schieffelin JS, Sealfon RS, Simon-Loriere E, Smits SL, Stoecker K, Thorne L, Tobin EA, Vandi MA, Watson SJ, West K, Whitmer S, Wiley MR, Winnicki SM, Wohl S, Wölfel R, Yozwiak NL, Andersen KG, Blyden SO, Bolay F, Carroll MW, Dahn B, Diallo B, Formenty P, Fraser C, Gao GF, Garry RF, Goodfellow I, Günther S, Happi CT, Holmes EC, Kargbo B, Keïta S, Kellam P, Koopmans MPG, Kuhn JH, Loman NJ, Magassouba Nf, Naidoo D, Nichol ST, Nyenswah T, Palacios G, Pybus OG, Sabeti PC, Sall A, Ströher U, Wurie I, Suchard MA, Lemey P, Rambaut A. Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature. 2017;544(7650):309-15. doi: 10.1038/nature22040. PMID: 28405027.

70. NIAID Emerging Infectious Diseases/Pathogens | NIH: National Institute of Allergy and Infectious Diseases. Available from: https://www.niaid.nih.gov/research/emerging-infectious-diseases-pathogens.

71. Schriml LM, Arze C, Nadendla S, Ganapathy A, Felix V, Mahurkar A, Phillippy K, Gussman A, Angiuoli S, Ghedin E, White O, Hall N. GeMInA, Genomic Metadata for Infectious Agents, a geospatial surveillance pathogen database. Nucleic Acids Res. 2010;38(Database issue):D754-64. doi: 10.1093/nar/gkp832. PMID: 19850722; PMCID: PMC2808878.
72. Gazetteer | NCBO BioPortal. Available from: https://bioportal.bioontology.org/ontologies/GAZ.
73. Bioinformatics Resource Centers (BRC) Awards | NIH: National Institute of Allergy and Infectious Diseases. Available from: https://www.niaid.nih.gov/research/bioinformatics-resource-centers-awards.
74. Abarenkov K, Adams RI, Laszlo I, Agan A, Ambrosio E, Antonelli A, Bahram M, Bengtsson-Palme J, Bok G, Cangren P, Coimbra V, Coleine C, Gustafsson C, He J, Hofmann T, Kristiansson E, Larsson E, Larsson T, Liu Y, Martinsson S, Meyer W, Panova M, Pombubpa N, Ritter C, Ryberg M, Svantesson S, Scharn R, Svensson O, Töpel M, Unterseher M, Visagie C, Wurzbacher C, Taylor AFS, Kõljalg U, Schriml L, Henrik Nilsson R. Annotating public fungal ITS sequences from the built environment according to the MIxS-Built Environment standard – a report from a May 23-24, 2016 workshop (Gothenburg, Sweden). Annotating public fungal ITS sequences from the built environment according to the MIxS-Built Environment standard – a report from a May 23-24, 2016 workshop (Gothenburg, Sweden). 2016;16:1. doi: 10.3897/mycokeys.16.10000.