

SPECIFIC AIMS

This competing renewal proposal focuses on the continued development of the Gene Wiki, a crowdsourced community resource for genetics and genomics. The goal during the first project period was to create a continuously-updated, community-reviewed, and collaboratively-written review article for every human gene. The Gene Wiki was created directly within Wikipedia as an informal collection of 10,646 gene-specific articles.

While the three specific aims from the first project period are detailed in the Progress Report section, the overarching objective was to cultivate the positive feedback loop between utility, usage, and contributions that underlie all successful crowdsourcing applications [2]. In short, we established the utility of the Gene Wiki by systematically harvesting and displaying data from structured databases. This utility led to a robust critical mass of readers who collectively viewed the Gene Wiki 68 million times in the last year. Drawn from this population of readers was a healthy critical mass of editors who collectively made over 15,000 edits in the last year. These edits improve the utility of the Gene Wiki, and the positive feedback loop repeats.

Having established this solid foundation during the previous three years of funding, we now propose the following specific aims:

Aim 1: Extend the Gene Wiki to systematically create and maintain Wikipedia pages for diseases and drugs. The expanded goal of this proposal is to create a continuously-updated, community-reviewed, and collaboratively-written review article for every human gene, disease, and drug.

- A. Define the set of “notable” genes, diseases and drugs, and curate their relevant biomedical properties. The work in this sub-aim will define the universe of entities for which we will create/maintain Wikipedia articles, and the annotations of those entities that will populate their corresponding infoboxes.
- B. Mine structured databases for relationships between genes, diseases and drugs. These links will provide a powerful mechanism for readers to traverse the network of biomedical knowledge, and also promote community editing of all of these important articles.

Aim 2: Implement an outreach plan to align Gene Wiki contributions with traditional scientific incentives and educational opportunities. Since this project is almost entirely based on community contributions, outreach is essential to ensure continued growth of the Gene Wiki.

- A. Partner with a peer-reviewed journal to solicit invited review articles with a parallel publication track on Wikipedia. This partnership ensures that the effort to create well-written Gene Wiki articles will also receive recognition through a traditional journal publication.
- B. Create an educational framework that will meaningfully harvest student effort to improve the Gene Wiki. This initiative will focus on learning in the context of real-world scientific problems.

Aim 3: Build a community-maintained Centralized Model Organism Database (CMOD) for microbial genomics. As the rate of whole genome sequencing is rapidly increasing, this aim will focus on the challenge of organizing gene and genome annotation across the tree of life.

- A. Develop import tools for all microbial gene models, gene annotations, and genome features in Wikidata based on common file formats. These tools will ensure that all existing genomics data will exist within Wikidata.
- B. Modify the data interface layer for the JBrowse genome browser and Web Apollo annotation editor to use the Wikidata API. Adapting these two established tools will provide a familiar interface to interact with Wikidata genome annotations.
- C. Develop centralized infrastructure that accepts GO annotations from the microbiology community. Organizing and centralizing annotations of gene function fills an important void in microbial genomics.

Aim 4: Leverage patient-aligned crowdsourcing for real-time annotation of the biomedical literature.

This aim engages a novel community of contributors while addressing a critical bottleneck in biomedical knowledge management.

- A. Create a web interface for concept recognition and normalization in biomedical text. Concept annotations provide a starting point for deeper mining and integration of biological knowledge.
- B. Create a web interface for relationship annotation. This sub-aim will focus on contributing to the community’s biological knowledge network.

APPROACH

This proposal includes four specific aims (outlined in **Figure 1**). **Aims 1** and **2** focus on expanding the core Wikipedia-based crowdsourcing infrastructure that was established in the initial grant period. **Aims 3** and **4** explore two novel applications of crowdsourcing in genetics and genomics. Because space constraints prevent a description of every detail, we outline here several overarching guiding principles that we have applied and will continue to apply in our work. First, we practice open science by freely discussing ideas via blog posts, social media, Wikipedia “talk pages”, and personal communication. Second, we practice collaborative science by aligning with complementary groups and efforts. Third, we broadly disseminate data and software, as detailed in the Resource Sharing plan. Fourth, our entire team has a strong commitment to structured data, interoperability, and reusability through the use of biomedical ontologies and semantic web standards.

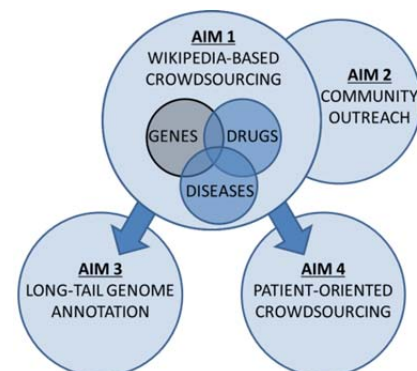


Figure 1. Overview of Specific Aims.

Aim 1: Extend the Gene Wiki to systematically create and maintain Wikipedia pages for diseases and drugs. The expanded goal of this proposal is to create a continuously-updated, community-reviewed, and collaboratively-written review article for every human gene, disease, and drug.

Significance. Wikipedia bills itself as a “free encyclopedia that anyone can edit” whose content is entirely the product of over 3 million monthly edits from the community. Wikipedia is a highly-used reference resource for virtually any topic, including scientific topics. In addition to the 10,646 gene articles currently maintained by the Gene Wiki project, there are 5,400 Wikipedia articles on drugs and 5,623 Wikipedia articles on diseases. Those gene, drug, and disease articles have collectively been viewed over 1.3 billion times in the last 12 months (68 million, 300 million, and 950 million views for each group, respectively). Clearly, a substantial amount of scientific content already exists on Wikipedia, and that the demand for scientific information is high. The significance of this aim can be presented on three levels:

Improving scientific literacy: Drug and disease articles already exist, and they are being read in massive numbers. Viewership is especially enhanced when these topics are relevant to current news events [2, 3]. Therefore, there is substantial value in ensuring that the basic data on those pages is uniformly displayed, derived from trustworthy sources, and continually updated.

Improving scientific discourse: It is overly simplistic to assume that all one billion pageviews come from non-scientists. We believe that this aim also improves the efficiency of scientific communication and dissemination among scientists. As the depth and detail of scientific knowledge continually increases, the overall scope of what any individual scientist can reasonably be considered an “expert” on is shrinking. As genome-scale profiling is becoming increasingly popular, it is now perfectly common for one scientist’s list of candidate genes to be populated by genes that are completely unfamiliar. Tools like the Gene Wiki are resources for working scientists to get an overview of a new topic and find pointers to the primary literature.

Recruiting new contributors: Systematically maintaining the structured data within Wikipedia pages is merely a means to an end. High-quality data attract readers, and from that community of readers contributors emerge. The Gene Wiki currently receives close to 1,500 edits per month [4], and expanding to diseases and drugs will increase contributions to all domain areas.

Innovation. The focus of this aim is on the design, creation, and maintenance of Wikipedia “infoboxes” that are found on every Gene Wiki page and contain structured data values. Though this goal is relatively straightforward, there is significant innovation in how these data will be stored. Currently, all infobox data are stored directly within Wikipedia using “templates”, semi-structured containers for standardizing display. Importantly however, data in templates are not accessible to querying or further data integration.

In the past two years, the Wikimedia Foundation (the parent organization to Wikipedia) has initiated a sister effort called Wikidata. In the words of the organizers, “Wikidata is a free knowledge base that can be read and edited by humans and machines alike”. In short, Wikipedia is to free-text what Wikidata is to structured data.

This community-maintained database can be edited and queried by anyone, either manually or using computer programs. The data model for Wikidata is described in more detail in **Aim 3** where it also figures prominently. Suffice it to say that in this aim, all structured data for genes, diseases, and drugs will be migrated/deposited in Wikidata, a process that is currently the subject of a Google Summer of Code project [5-7].

Wikipedia will be the first consumer of the data we upload (to populate the infoboxes), but it is certainly not the only consumer. Storing these data in Wikidata will enable other users and applications to perform rich, integrative queries. For example: *“What proteins with a kinase domain are encoded by genes on human chromosome 6p, are implicated in any type of epithelial cancer, and for which a small molecule agonist is available?”* While it is possible for a bioinformatician to answer this question by importing and joining data across many resources, it will be much simpler to execute this query on Wikidata because all the data exist in a single knowledgebase. Moreover, because Wikidata is compatible with the Semantic Web, even broader queries incorporating additional Linked Data sources are readily accessible.

To seed this knowledgebase, we will upload data from many key biomedical resources into Wikidata. However, we will undoubtedly miss important data sources. While these omissions are normally liabilities in a project proposal, this inevitable fact underscores the true value of a community-maintained database. *Anyone* is empowered to add new data to Wikidata, and contributions can range in size from one singular fact to a large data set. Importantly, once one person spends the effort to structure that new knowledge, it is then subsequently available for everyone. A community database asymptotically approaches completion through cumulative individual efforts, rather than duplication and fragmentation of effort across many local databases.

A. Define the set of “notable” genes, diseases and drugs, and curate their relevant biomedical properties. The work in this sub-aim will define the universe of entities for which we will create/maintain Wikipedia articles, and the annotations of those entities that will populate their corresponding infoboxes.

Although the set of human genes and proteins is reasonably well-defined in databases like NCBI Entrez Gene, Ensembl, and UniProt, the set of all human diseases and drugs has much fuzzier boundaries. Defining the scope for all three data types and defining the properties that are of broad interest to the community are the first steps in building the bioinformatics infrastructure proposed here. Note that all of these decisions are made in close collaboration with the relevant Wikipedia user communities, including the WikiProjects for Molecular and Cellular biology [7], for Medicine [8], and for Pharmacology [9].

Genes: The Gene Wiki is currently comprised of 10,646 articles on human genes. The set of properties in the gene infoboxes is largely complete based on decisions in the first funding period. The only enhancement proposed here is to add more complete coverage of SNPs and pathways relevant to each gene. Knowing the number, type, and physiological effect of SNPs within a gene’s boundaries is highly relevant for human genetics research. These data will be extracted from dbSNP and displayed in a simple infobox at the bottom of each gene page. Pathways are also a useful means to understand a gene’s role in human health and disease. These data will be extracted from the Reactome database (see **Letter of Support from Lincoln Stein**) and displayed within the existing gene infobox.

Diseases: There are currently 5,623 Wikipedia articles on human diseases. While this represents an impressive cumulative product of lots of manual effort, there are clearly many important diseases that do not yet have articles. Therefore, we will start by building off of the **Disease Ontology (DO)**, a standard ontology of human diseases developed by **Lynn Schriml** and **Warren Kibbe** (see **Letters of Support**) that currently has 8,683 classes. Schriml and Kibbe will lead the development of a “slim” version of Disease Ontology (“DOSlim”) that focuses on the set of diseases that are suitable for having a dedicated article in Wikipedia. This effort will include many additions through integration with disease resources shown in **Table 1** as well as subtraction of organizational “placeholder” nodes that appear in the full DO.

Initially, we will have a particular emphasis on rare diseases. Although their individual prevalence is infrequent, rare diseases collectively affect almost 25 million Americans. They are also the subject of intense biomedical research because discovery of the causative mutations is greatly aided by recent advances in sequencing technologies [10]. We have partnered with the Orphanet database (see **Letter of Support from Ségolène Aymé**) to take advantage of their effort to generate expert-written disease

definitions for all rare diseases. These authoritative disease definitions will be used as seed text for creation of the ~4,000 rare diseases that currently have no Wikipedia entry. This initial emphasis on rare diseases represents a timely and biomedically important focus. Since the infrastructure developed herein will be extensible to any disease with little or no modifications, the latter half of the grant period will be devoted to extending our effort to all diseases in DOslim.

We will also add several new properties to the gene infoboxes, including links to Disease Ontology [11] and Orphanet, and categorizations of disease subtypes based on the DOslim hierarchy. We will also improve how OMIM data and links are presented, particularly as they relate to OMIM Phenotypic Series (see **Letter of Support from Ada Hamosh**). Our effort will also have a special emphasis on descriptions of the associated clinical phenotypes in collaboration with the Human Phenotype Ontology (HPO) (see **Letter of Support from Peter Robinson**). The HPO currently has over 100,000 annotations to 7,000 distinct diseases, providing a detailed description of the signs and symptoms. Prototypes are being actively developed in collaboration with the Wikipedia community. A snapshot of the working prototype is shown in **Figure 2**, and the most current version can always be viewed at <http://tinyurl.com/GeneWikiPhase3>.

Drugs: There are currently 5,400 drug infoboxes in Wikipedia. While the data contained therein represent a valuable resource, many important drugs do not yet have Wikipedia pages or infoboxes. We will substantially expand the number of drug articles in Wikipedia. The list of drugs will be generated by aggregating the drug data sources shown in **Table 1**. After removing redundancies, we anticipate that there will be tens of thousands of compounds for which Wikipedia articles could be created.

We will also maintain an exhaustive list of chemical annotations in the drug infoboxes. This list will include the wide range of properties that already exist in the infoboxes (e.g. CAS number, molecular formula, pregnancy category), as well as new properties mined from the data sources in **Table 1** (including mechanism of action, biological target, and therapeutic uses). Again, a snapshot of the working prototype is shown in **Figure 2**, and the most current version can always be viewed at <http://tinyurl.com/GeneWikiPhase3>.

Table 1. Data sources for Specific Aim 1.

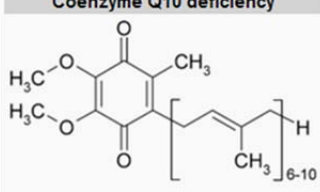
	Entities			Relationships			Notes
	Human Genes	Diseases	Drugs	Gene - Disease	Gene - Drug	Disease - Drug	
Current Wikipedia	10646	5623	5400	25074	37908	27839	
OMIM	14289	7445	---	3815	---	---	
OrphaNet	2899	5954	***	5153	---	---	
DrugBank	---	---	6805	---	9894	---	
PubChem	---	---	11690	---	---	---	(1)
Neurocarta	6198	1853	---	26379	---	---	
NDF-RT	---	4698	21071	---	---	55879	
DO	---	8683	---	---	---	---	
CTD	35319	12056	149474	27397	869902	186986	(2)
STITCH	---	---	262083	---	251320	---	(2)
DGI-DB	---	---	---	---	16391	---	
PharmGKB	26934	3421	3091	---	---	---	(3)

(1) only counting compounds with annotated pharmacological action

(2) will require significant filtering to identify highest-confidence interactions

(3) The license on PharmGKB relationships is incompatible with Wikipedia's CC-BY-SA license.

Coenzyme Q10 deficiency



Ubiquinone

DiseasesDB [32701](#)

MeSH [C564403](#)

Related genes [COQ2^{\[2\]}](#), [COQ6^{\[3\]}](#), [PDSS2^{\[4\]}](#), [PDSS1^{\[5\]}](#), [ADCK3^{\[6\]\[7\]}](#), [COQ9^{\[8\]}](#)

Compounds/Drugs [Ubidecarenone](#)

Orphanet [ORPHA35656](#)

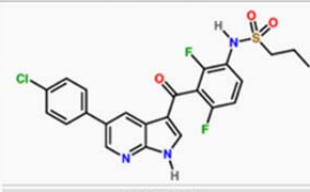
Disease [inherited metabolic disorder](#)

Classification [disorder](#)

OMIM Phenotypic Series [2]

Coenzyme Q10 deficiency, primary, 1	607426
Coenzyme Q10 deficiency, primary, 2	614651
Coenzyme Q10 deficiency, primary, 3	614652
Coenzyme Q10 deficiency, primary, 4	612016
Coenzyme Q10 deficiency, primary, 5	614654
Coenzyme Q10 deficiency, primary, 6	614650

vemurafenib



Clinical data

Trade names [Zelboraf](#)

AHFS/Drugs.com [FDA Professional Drug Information](#)

Licence data [US FDA link](#)

Pregnancy cat. [D \(US\)](#)

Legal status [R-only \(US\)](#)

Routes [Oral](#)

Mechanism & use

Mechanism [Protein kinase inhibitor](#)

Biological target [BRAF](#)

Therapeutic use [Melanoma](#)

Identifiers

CAS number [1029872-54-5](#)

ATC code [L01XE15](#)

PubChem [CID 42611257](#)

ChemSpider [24747352](#)

UNII [207SMY3FQT](#)

ChEMBL [ChEMBL1229517](#)

PDB ligand ID [032 \(PDBe, RCSB PDB\)](#)

Chemical data

Formula [C₂₃H₁₈ClF₂N₃O₃S](#)

Mol. mass [489.92 g/mol](#)

InChI [\[show\]](#)

[\(what is this?\) \(verify\)](#)

Figure 2. Prototype infoboxes for diseases (left) and drugs (right). Aim 1 will focus on updating and maintaining the 5000+ disease and drug infoboxes that already exist in Wikipedia, as well as creating thousands more pages for high priority diseases and drugs that do not yet exist. The most recent prototypes (and community discussion) are at <http://tinyurl.com/GeneWikiPhase3>.

Building robust and reliable infrastructure to maintain these infoboxes is equally important as the selection and design of the articles themselves. Making sure that data are current and complete with the state of biological consensus is essential to maintain credibility with both scientists and the general public. In this aim, we will

extend our code for automated maintenance of gene infoboxes to also handle disease and drug infoboxes. Updates from primary data sources will be run on at least a weekly basis.

B. Mine structured databases for relationships between genes, diseases and drugs. These links will provide a powerful mechanism for readers to traverse the network of biomedical knowledge, and also promote community editing of all of these important articles. There are many databases and meta-databases that describe biomedical relationships between genes, diseases and drugs (outlined in **Table 1**). A significant level of effort in this aim will be devoted to mapping between these resources and filtering for the highest quality data, a process that will involve both automated and manual review.

Gene-disease: Links between genes and diseases will be assembled from several large-scale resources like OMIM, Orphanet and CTD, and also more disease-specific resources like AlzGene (Alzheimer's disease), PDGene (Parkinson's disease), MSGene (multiple sclerosis), and SFARI (autism). In addition to the interactions themselves, OMIM, Orphanet and CTD also annotate interactions by the interaction type. These types (e.g., "Disease-causing germline mutation", "Major susceptibility factor", "protection against", and "resistance to") will be tracked and noted in the Wikipedia infoboxes whenever possible. This overall effort will be organized and led by the team developing Neurocarta, a framework for the aggregation and dissemination of gene-disease relationships (see **Letter of Support from Paul Pavlidis**).

Gene-drug: Links between genes and drugs will primarily be drawn from DGIdb [12], which in turn harvests interactions from a variety of other databases [13-18]. DGIdb also normalizes the nature of the interaction according to ~40 types, including inhibition, antagonism, agonism, potentiation, and binding (the top 5 most used types). Our team will also integrate gene-drug interaction data from CTD [19] and STITCH [20], though both of these resources will require significant filtering for the highest quality interactions.

Drug-disease: Links between diseases and drugs will primarily be drawn from the National Drug File - Reference Terminology (NDF-RT). This resource also categorizes interaction types ("may_treat", "may_prevent", "may_diagnose", and "induces"). CTD will also be a secondary source of drug-disease interactions (typed by "marker/mechanism" or "therapeutic"), though again filtering will need to be applied to identify the highest-confidence interactions. This effort will be organized and led by **Lynn Schriml** (see **Letter of Support**) who is already collaborating with **Evan Bolton** from PubChem (see **Letter of Support**) on mapping between the Disease Ontology, NDF-RT, UMLS and MeSH.

Prototypes showing how these relationships will be displayed in the context of Wikipedia infoboxes are included in **Figure 2**, and links will be added to articles on both entities participating in each interaction.

Aim 2: Implement an outreach plan to align Gene Wiki contributions with traditional scientific incentives and educational opportunities. Since this project is almost entirely based on community contributions, outreach is essential to ensure continued growth of the Gene Wiki.

Significance. The usage statistics presented in the Progress Report clearly demonstrate that the Gene Wiki has a critical mass of users and contributors. Nevertheless, it is essential to continually explore new ways to attract new editors and contributors. This aim briefly describes two such approaches.

A. Partner with a peer-reviewed journal to solicit invited review articles with a parallel publication track on Wikipedia. This partnership ensures that the effort to create well-written Gene Wiki articles will also receive recognition through a traditional journal publication. The most common reason for scientists not to contribute to Wikipedia is the perceived lack of reward (a perspective that is particularly true among more senior scientists whose domain expertise and perspective is often most valuable). In short, contributions to Wikipedia do not align with traditional metrics of scientific productivity.

Recently, we initiated a partnership with the journal GENE to solicit invited review articles on genes whose Wikipedia pages were underdeveloped. To date, three such review articles have been published on *MIR155* [21, 22], *EPHX2* [23, 24], and *SFTPA1* and *SFTPA2* [25-27]. In addition to the PubMed-indexed article, the authors also wrote significant text for the corresponding Wikipedia page. GENE has made a financial commitment to support the publishing side of this partnership, and the PI has joined GENE as a co-Executive

Editor. To ensure this initiative can maintain a continuous stream of articles, this sub-aim will formalize the effort to identify appropriate articles, to recruit qualified authors to write those articles, and to assist those authors when writing the Wikipedia versions of their review.

B. Create an educational framework that will meaningfully harvest student effort to improve the Gene Wiki.

This initiative will focus on learning in the context of real-world scientific problems. Students at all levels expend an extraordinary amount of human effort on class projects, but with only a few exceptions, the products of that effort are discarded at the end of term. Recently, however, there is growing enthusiasm for class projects that have lasting scientific benefit. These efforts have ranged from biocuration of the literature [28], to sequencing and annotation of bacteriophages [29], to annotation of metagenomics sequence [30]. There have even been a series of high school AP biology classes who have focused on making contributions to Wikipedia every year since 2008 (see [31] and links therein). Not only do students learn about important scientific topics, they also get a flavor for collaborative science and writing. In this aim, we will pursue similar initiatives to develop articles on genes, diseases, and drugs. This effort will involve identification of important and underdeveloped sets of Wikipedia articles, pairing those articles with aligned classes, and developing rubrics for evaluation. We will seek partners in all areas of higher education, including undergraduate, graduate, pharmacy and medical students.

Aim 3: Build a community-maintained Centralized Model Organism Database (CMOD) for microbial genomics. As the rate of whole genome sequencing is rapidly increasing, this aim will focus on the challenge of organizing gene and genome annotation across the tree of life.

Significance. The number of sequenced genomes is growing rapidly (**Figure 3**). There are currently ~3,000 species (bacteria, eukaryotes, and archaea) that have a completely sequenced genome. If the trend over the last 10 years holds, we will pass 10,000 sequenced genomes in 2015 and 100,000 completely sequenced species within the next 10 years. And given recent efforts toward “Preserving the Genomic Diversity of Life on Earth” [32, 33], it seems likely that genomic sequencing will continue even well beyond. The organization of genomic data has in large part been the domain of **Model Organism Databases (MODs)**. These community resources are fantastic resources for biomedical research, spanning a spectrum of model organisms from mouse to worm to *E. coli*. MODs fulfill key roles for their respective communities, from warehousing key genomic data, to providing query and visualization tools, to performing biocuration.

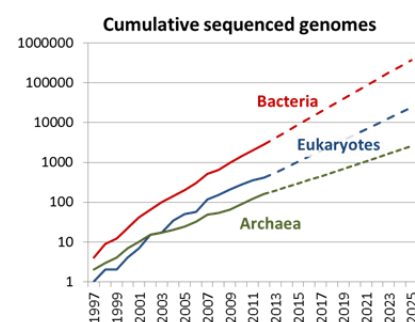


Figure 3. Growth in sequenced genomes. Dashed lines show future projections. Data are based on genome completion data in GOLD database.

The greatest growth in sequenced genomes is found among bacteria. This growth is driven in large part by several large-scale microbial genomics efforts from the Broad Institute, the J. Craig Venter Institute, and the Joint Genome Institute. Genome sequencing as a mechanism for classroom education is also increasingly popular [29, 30, 34-36]. The human microbiome in particular has been an area of intense recent interest (aided by substantial funding through the NIH’s Human Microbiome Project). For example, the gut microbiome has been shown to be involved in a diverse set of human disorders including irritable bowel disease, obesity, and colon cancer [37-39]. We have chosen to focus this aim on microbial genomics because of its important role in human health and because of the fragmentation and inaccessibility of knowledge.

Analyses of these newly sequenced genomes depend on effective tools for structured knowledge management. The current model is based on bioinformatics and biocuration staff dedicated to each organism, but it is impractical to scale this approach to current sequencing throughput. Efforts like the **Generic Model Organism Database (GMOD)** project create open source software to lower the activation barrier for setting up genomic infrastructure. However, creating and maintaining a GMOD instance still requires a significant and ongoing investment.

So while the number of species with a MOD is extensive, it is by no means exhaustive. In this aim, we focus on building a **Centralized Model Organism Database (CMOD)**, a single community-maintained database that easily scales with the rapid growth in genome sequences. CMOD will target the “Long Tail” of sequenced microbes that do not have and likely will never have a dedicated MOD. CMOD will implement two key features

that are essential for genomics -- gene and genome annotation. These two features only satisfy a subset of the functions that fully-staffed MODs fulfill, so CMOD is certainly not a replacement for traditional MODs. Nevertheless, CMOD offers a general solution to the knowledge management challenge in microbial genomics, and it will serve a valuable role in cases where no MOD, infrastructure, or biocurators currently exist.

Genome annotation refers to the description of genomic features and their coordinates. These features include genes, exons, promoters, and operons. NCBI, Ensembl Bacteria, and PATRIC are three of the largest and most systematic providers of microbial genome annotations. For example, NCBI annotates over 17 million genomic features on 2,515 bacterial genomes, Ensembl Bacteria annotates 79 million genomic features on 6,305 genomes, and PATRIC annotates 35 million features on 8,913 genomes.

Although these data repositories provide a valuable resource for microbial genome annotations, they are certainly not complete. Most notably, genome annotations are heavily biased in the feature types that are annotated. Over 97% of all features relate to the gene boundaries themselves ('gene', 'CDS', 'exon'). In contrast, only 57 of the 2515 bacterial species at NCBI had any annotated operon, promoter, attenuator, regulatory region, or terminator feature (0.1% of all features), and no such annotations were found in Ensembl Bacteria or PATRIC. While more specialized databases do exist (e.g., RegulonDB [40]), these resources often only operate on a subset of available genomes, they often have restrictive usage and dissemination licenses, and they often use custom file formats that make integration with other diverse resources difficult and time-consuming.

As a case in point, the transcriptional regulation of the *Listeria monocytogenes* genome was extensively characterized in a prominent Nature paper [41], revealing (among other findings) the existence of 517 operons and 103 small regulatory RNAs. However, those data are not visible through any genome browser, nor are they available for download at any of the data repositories examined. In fact, those data are currently only available in PDF format as supplementary info on the journal website. Even the canonical *lac* and *trp* operons in *Escherichia coli* are not annotated in any of the microbial resources listed above. In the absence of an official MOD, genome annotation data often end up buried in journal publications and inaccessible to computational analyses and structured queries.

Gene annotations refer to the description of genes and their protein products. Most often, gene annotations are based on the **Gene Ontology (GO)**, and managing the generation, storage, and dissemination of those annotations represents another core function of MODs. Among bacterial species, *Mycobacterium tuberculosis* and *Escherichia coli* are by far the most well-annotated species, each with over 10,000 experimentally-derived gene annotations. Overall, there are 189 bacterial species with at least five such annotations, and 310 bacterial species with at least one annotation. In total, there are ~55,000 experimentally-derived, bacterial GO annotations.

Although these existing annotations represent a tremendously valuable knowledgebase of microbial gene function, it represents only a small percentage of the biological knowledge that could (and should) be structured. As a very rough estimate of the number of species that could have GO annotations, we found 1,107 species that had at least 10 PubMed hits when searching the species name plus "gene". For example, consider *Borrelia burgdorferi*, the bacterial species that causes Lyme Disease whose genome was sequenced in 1997 [42]. Searching PubMed for "'Borrelia burgdorferi' AND gene" reveals over 1,400 articles. However, there are no GO annotations for any of the ~1,300 coding sequences, despite the discovery of many functional roles for genes in the literature (e.g., [43-48]).

For both genome and gene annotations, the analyses and examples above demonstrate that there is substantial knowledge of microbial genomics that is simply left unstructured and/or unlinked. When this knowledge is not easily aggregated and integrated, it is virtually inaccessible to the computational analyses and tools that are so important to genome-scale science. For example, a comprehensive microbial database could lead to better prediction models for operons [49], or better phylogenetic inference of functional annotations [50], or more accurate functional enrichment analyses for microbiome studies [51].

Although this aim will initially focus on the Long Tail of microbial species, the CMOD infrastructure will be equally applicable to any newly sequenced genome from any kingdom of life. For example, there are currently

over 400 sequenced eukaryotic genomes, but there are over 2,000 eukaryotic genomes with “annotatable content” (using the same PubMed search criteria described above). The projected growth rate in eukaryotic genome sequences is very close to what is observed for bacteria (**Figure 3**). A more comprehensive and integrated view of genome and gene annotation for non-microbial species will undoubtedly aid efforts for gene function prediction across all species based on phylogenetic inference [49, 52, 53].

Innovation. We will build CMOD on Wikidata, a sister project to Wikipedia. Just as Wikipedia crowdsources the creation of encyclopedic articles, Wikidata crowdsources the creation of structured database entries. Wikidata offers a flexible and universal data model by which knowledge can be represented. At its core, Wikidata is populated by items (e.g., genes, diseases, and drugs) and statements about those items. Statements are defined by a property, a value, one or more optional qualifiers, and one or more references (**Figure 4**). By incorporating the “reference” field directly into the data model, the provenance of every statement can be immediately seen (and used in queries).

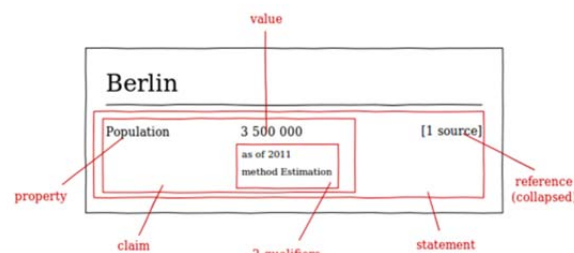


Figure 4. Wikidata data model.

Wikidata has huge potential for knowledge integration across many diverse domain areas. For example, consider a question like “*What gram negative bacteria have genes encoding both a phosphatidylcholine synthase and a phospholipid N-methyltransferase (both part of lecithin biosynthesis) within a single predicted operon?*” The data to perform integrative queries like these of course exist, but assembling those data sources would be a significant undertaking. Loading all those data sources into Wikidata would greatly simplify the query, and importantly, those data would be available for any other integrative query in the future.

We are excited to continue an existing partnership with the Wikidata community (see **Letter of Support from Denny Vrandečić**). However, we do recognize that Wikidata is a rather new initiative. As a backup, we have also done significant prototyping with Freebase, a “community-curated database” that is similar to Wikidata. Freebase has been in public release since 2007 and was acquired by Google in 2010. Although we prefer Wikidata based on some relatively minor design decisions, Freebase offers a perfectly viable backup. The development of Wikidata and Freebase are representative of a broader trend in online web development toward centralized, community resources. Just as many locally-maintained compute clusters are being retired in favor of centralized cloud computing, it is also just now becoming feasible for all biological knowledge (including gene and genome annotations) to be integrated in a CMOD.

A. Develop import tools for all microbial gene models, gene annotations, and genome features in Wikidata based on common file formats. These tools will ensure that all existing genomics data will exist within Wikidata.

As described previously, NCBI, Ensembl Bacteria, and PATRIC all provide a substantial amount of genome annotation data on a wide breadth of organisms. Before extending the state of genome annotation through crowdsourcing, it is important to ensure that we are capturing the majority of data that are already available in other databases. This process begins with defining the Wikidata data model. We partially prototyped the system in Freebase (see code [54]) prior to the creation of Wikidata, and we will follow this same general model. Next, we will create an importer that maps GFF3 files into the Wikidata gene model. There are over 17000 GFF files available through these three resources which extensively annotate basic genomic data like gene and exon boundaries, tRNAs and rRNAs. In addition, we will actively solicit from the community other reference data sets. For example, **Paul Gardner** has already agreed to test our system as a *data producer* (by uploading a dataset of annotated ncRNAs), and also to help identify other large-scale data sets that he would use as a *data consumer* (see **Letter of Support**).

Although the coding and community consensus building required in this sub-aim is time-consuming, we anticipate that the process overall will be uncontroversial and straightforward.

B. Modify the data interface layer for the JBrowse genome browser and Web Apollo annotation editor to use the Wikidata API. Adapting these two established tools will provide a familiar interface to interact with Wikidata genome annotations.

JBrowse and WebApollo are two of the most well-established and well-used genome browsers and genome annotation tools available [55, 56]. Both are part of the broader GMOD effort, so community adoption is well-established. Typically, both of these tools retrieve data from either local flat files or local databases using the `Bio::DB::*` Perl interfaces. In this sub-aim, we will create an alternate data access layer that instead utilizes the Wikidata API for genome annotation data. JBrowse is a read-only application, so it will only require use of the Query API [57]. Web Apollo has the added functionality of being a genome annotation tool, so it too will utilize the Wikidata Write API [58].

This work will be conducted in collaboration with both **Ian Holmes** (PI of JBrowse project) and **Suzanna Lewis** (PI of Web Apollo project); see **Letters of Support**. Given the degree of synergy between the Wikidata initiative described here and complementary community data-sharing efforts within the JBrowse and Web Apollo projects, our three groups have agreed to participate in a yearly hackathon.

C. Develop centralized infrastructure that accepts GO annotations from the microbiology community. Organizing and centralizing annotations of gene function fills an important void in microbial genomics.

A large collection of microbial GO annotations are generated and disseminated using the GAF 2.0 file format defined by the GO Consortium [59]. This standard captures a complete set of metadata for each annotation, including an evidence code, the primary reference, qualifiers and the annotation authority. The first step in this sub-aim will be to develop the Wikidata data model to capture this rich breadth of information on each GO annotation. It is also particularly important to distinguish authoritative annotations by professional biocurators from community contributions.

Next, we will develop a GAF import tool that translates annotation files into Wikidata statements. This tool will be used to import all annotation files currently distributed by the GO Consortium [60], and our infrastructure will keep Wikidata continuously in sync with official data releases. This GAF import tool will also be one of the primary input mechanisms for community contributions. Research groups with large-scale annotation efforts (either by computational or experimental means) can easily format their data into this simple tab-delimited text format. For contributors of individual GO annotations, we will also provide a simple web form with fields for all parameters in the GAF standard. Although not within the scope of this proposal, this general infrastructure also enables participation by other innovative initiatives to encourage community GO annotations (e.g., CACAO [28] and Gene Wiki text mining [61]).

We anticipate that other groups will use the Wikidata query API to build query, analysis, and visualization tools based on GO annotations in CMOD. For example as the phylogenetic breadth of GO annotations grows in CMOD, it is not unreasonable to update the PAINT algorithm to use Wikidata as a source of data. All the work described here will be done in close coordination with the GO Consortium through the PI's role on the Scientific Advisory Board (see also **Letter of Support from Judy Blake** on behalf of the GO consortium).

Aim 4: Leverage patient-aligned crowdsourcing for real-time annotation of the biomedical literature.

This aim engages a novel community of contributors while addressing a critical bottleneck in biomedical knowledge management.

Significance. The biomedical literature is massive and growing explosively. There are over one million new biomedical articles every year indexed in PubMed, which roughly equals 3000 every day, or one every 30 seconds. The growth of knowledge is of course incredibly valuable, but in the form of free-text publications, that knowledge is almost completely unstructured. Scientists cannot effectively compute on that new knowledge or incorporate it into informatics analyses. In this aim, annotation refers to the systematic and structured description of the biomedical literature. We focus particularly on annotation of relevant gene, disease and drug concepts, as well as the relationships between those concepts.

The current processes for annotating the literature are valuable, but incomplete. Most notably, the National Library of Medicine (NLM) performs indexing of all PubMed articles according to the MeSH vocabulary. However, this impressive service has limitations. The granularity is limited since ~26,000 MeSH terms are meant to cover all chemicals, genes (multiple species), diseases, and anatomic regions. The depth of annotations is limited because each article is only seen by a single indexer for an average of 15 minutes [62].

Indexers by necessity are generalists and not domain experts. In addition, there are no annotations of the relationships between concepts.

Other resources for biological networks also exist. Pathway databases like Reactome perform valuable biocuration services, but they primarily focus on creating interpretable abstractions of biological processes rather than systematically cataloging links [63]. Commercial entities like Ingenuity do systematic annotation of links, but they impose limitations on access and reuse. Interaction databases based on high-throughput data [64-70] are a valuable complement to (but not a substitute for) knowledge derived from hypothesis-driven studies. Text mining of relationships from free text is a nascent field. While there have been exciting recent advances (e.g., SemMedDB [71]), this is clearly an area of active research in the Bio-NLP community [72, 73].

To fill this notable gap in the landscape of biomedical resources, this aim will harness the collective efforts of patient-aligned individuals to perform real-time annotation of relationships in the biomedical literature. The result of this effort (tentatively named **Mark2Cure**) will be a continuously-updated network of biomedical concepts (genes, diseases, and drugs). We expect there to be at least three significant applications:

Enhanced literature searching: Scientists continuously exert significant effort to remain aware of the most relevant literature to them. This difficulty for individual scientists to precisely identify relevant research papers in a timely fashion is one major inefficiency in our system for biomedical publishing. With our emphasis on the most recent literature, Mark2Cure data can offer researchers literature alerts that are based on not only annotated concepts but also relationships between concepts.

Data mining in knowledge networks: Analyses of existing interaction databases have been mined in a wide range of productive ways, including the prediction of drug mechanisms and targets [74-78], identification of metastatic biomarkers [79] and pathway regulators [80-82], and even the refinement of biomedical ontologies [83]. We anticipate that the same network mining methods will be similarly informative when applied to our crowdsourced knowledge network.

Construction of a training set for text mining: The limited success of text-mining in extracting relationships is due in large part to the lack of high-quality training sets. In collaboration with **Karin Verspoor** (see **Letter of Support**), we will ensure our data collection will use the appropriate annotation standards so the data can be used for further method development in text mining.

It is difficult to estimate the number of annotators and annotations that will result from Mark2Cure. We acknowledge this is the highest-risk aim, and we clearly do not want to overstate its significance given the rich ecosystem of tools, resources and research groups in this area. As just one aim in an R01 grant, our goal is simply to create a tool based on community crowdsourcing that is complementary in approach and in scope to these existing efforts. Nevertheless, we recognize the potential that this system could “go viral”. Anecdotally, we know that many people touched by disease are immersing themselves in the science [84], and this tool absolutely taps into that emerging trend.

Innovation. In **Aims 1, 2 and 3**, our “crowds” are dominated by individuals who have a formal scientific background – students, educators, and professionals – because most contributions to these efforts require biomedical domain knowledge. However, the challenge of engaging this community is aligning the act of contributing with some explicit reward to the contributor. Unfortunately, traditional metrics of success in academic, educational and industrial circles typically do not recognize contributions to community resources. In short, this community has a very high level of expertise but a relatively low motivation to contribute.

We hypothesize that patient-aligned communities represent a complementary population from which we can assemble a crowd. Recently there has been an explosion in the number of online communities for patients, their families, and patient advocates. These range from disease-specific sites (e.g., Alzheimer’s disease [85], Celiac disease [86], Lyme Disease [87]) to more general sites that attract a broader spectrum of individuals (e.g., PatientsLikeMe [88-91], MyHealthTeams [92], SmartPatients [93]). The growth of these sites reflects a broader trend of patients and families becoming more actively involved in the management, understanding, treatment and research of their illness. This trend is particularly notable for those suffering from rare diseases (for example, [94-98]). In short, patient-aligned individuals have a very high level of motivation albeit with a

lower level of expertise. We expect that if we provide a clear, direct, and tangible mechanism for these individuals to advance scientific research, they would welcome that opportunity to contribute to science. Such an effort would align perfectly with their motivation to both advance research and to learn about their disease.

Despite their desire to contribute, conventional wisdom held that there is simply no mechanism by which non-scientists can advance research. However, recent work has shown that English-speaking non-scientists can use their reading and language skills to annotate a range of biomedical concepts in scientific texts [99-102]. These studies collectively demonstrate that non-scientists crowds compare very favorably to automated text-mining and expert curation in terms of both accuracy and cost. So while non-scientists may not have the background to fully comprehend the meaning of scientific text, they can (through either training or their innate ability to parse grammar) effectively identify and annotate biomedical concepts.

During the early development of Mark2Cure, **Josh Sommer** from the Chordoma Foundation (see **Letter of Support**), has volunteered his time and will encourage members of his organization to provide feedback and input. Once released, we think Mark2Cure will be broadly engaging to patient groups across all diseases. To best engage each community, we will create disease-specific portals that pre-filter research articles based on their relevance to each disease. This system of tailoring the articles shown to each individual volunteer will also encourage smaller foundations and groups to participate, since their efforts will be focused on the biomedical literature around their disease of interest.

A. Create a web interface for crowdsourcing concept recognition and normalization in biomedical text. Concept annotations provide a starting point for deeper mining and integration of biological knowledge.

The first task that Mark2Cure will address is disease recognition (**Figure 5**). During the online training phase, users will learn how to annotate mentions of diseases in abstracts using real-world examples. Use of online resources will be demonstrated and encouraged. After passing a test given at the end of training, users are then empowered to start performing real annotation tasks on actual text in need of annotation. Users will also have the option to get further training to “unlock” additional tools that allow them to do different types of concept annotation. For example, users can take training to identify drugs and/or genes. After passing those training modules, users gain more tools that they can use to more fully annotate text.

The basic task focuses on simply recognizing concepts in text by highlighting the author’s words. An advanced feature would be concept normalization – the process of relating a word or phrase to a specific entry in an ontology, controlled vocabulary or database entry. Other tools (e.g., [103, 104]) have reasonable interfaces for a two-stage process of concept recognition and then normalization, and we will generally follow this workflow. This work will heavily utilize the existing partnership and tools of the National Center for Biomedical Ontology (see **Letter of Support from Mark Musen**).

After a user submits their annotation for a given text, the system will show the user how their annotations agreed or disagreed with other users who annotated that same text (**Figure 6**). This feedback mechanism allows each user to continue their learning process as guided not by a gold standard, but by community consensus. There are several important points to note (briefly). First, community consensus is only shown after they have submitted their annotation, thereby ensuring that each article gets multiple independent votes. Second, if we determine that the community is adopting rules that lead to undesired consensus, then we also have the ability to manipulate that feedback mechanism and thereby reeducate the community. Third, each article will be shown to as many users as is necessary to reach a stable set of

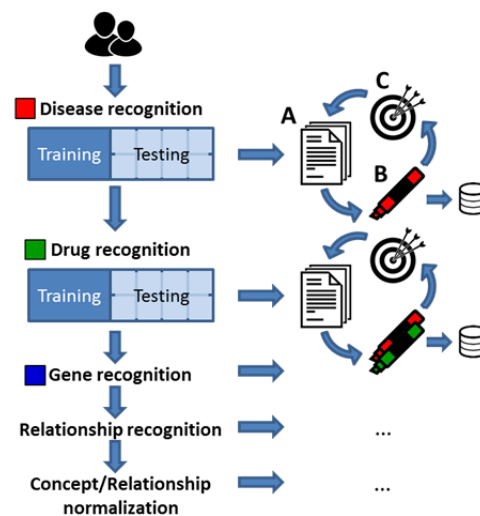


Figure 5. Overall Mark2Cure workflow. Users are first directed to the module on disease recognition. After a training session on how to recognize and annotate diseases in biomedical text, users are given a simple test to ensure they understood the instructions. After passing that test, (A) users are given a real-world journal abstract, (B) users annotate the document by highlighting disease terms (the results of which are stored in a database), and (C) the user is shown feedback comparing their annotations to the community consensus. The user is then shown a new article, and this cycle repeats as long as and at whatever rate the user desires. Users also have the option of receiving further training on additional annotation tasks, which unlock more tools with which the user can annotate. These additional training can be for drug/gene recognition, concept normalization, or relationship recognition/normalization.

consensus annotations. And fourth, we will keep detailed logs that will allow us to adjust the definition of consensus at any time, and there is no requirement that all contributors be weighted equally.

Concept recognition using text-mining tools for biomedical terms and text are reasonably good [105], but it is by no means perfect. Previous work using crowds for biomedical text annotation have shown measurable improvements over text mining tools, with F values in the range of 0.75 to 0.87 [99, 100, 102]. This aim also builds the community that will participate in the relationship annotation in **Aim 4B**, where the differential between human and computational annotation is even greater.

B. Create a web interface for relationship annotation.

This sub-aim will focus on contributing to the community's biological knowledge network.

While some individuals will choose only to participate in the concept annotation process in **Aim 4A**, we expect that some volunteers will want to pursue even deeper annotation of abstracts. For those users, we will offer training modules on relationship recognition and normalization.

When presented with a biomedical text, users will first perform concept identification and normalization. (Candidate concepts can also be pre-filled using text-mining or the output of concept recognition by other users.) Next, the user will be able to join those concepts by highlighting the word or phrase that describes the nature of the relationship (**Figure 7**). Again, relations can often be inferred based only on the sentence construction, making this task suitable for patient-aligned volunteers. As described for the concept annotation task, each abstract will be shown to multiple volunteers until a consensus set of relationships emerges. The consensus networks from each individual article will be combined into one massive community-maintained and continuously-updated knowledgebase. Each edge of this knowledgebase will have a quantitative score based on the degree of consensus, and the provenance will be explicitly tracked in the form of the PubMed IDs used to annotate the link.

In the process of annotating relationships, users will be building their own personal biological knowledge network (**Figure 8**), which we will manage and visualize for them. We think this knowledge browser will be a powerful motivator for many patient-aligned users as they explore the scientific literature related to their disease of interest. These volunteers not only want to help science, but they want to know and learn about the diseases that affect them and their loved ones. The personal knowledge management tools we provide will allow them to organize facts around topics they are interested in, and to compare their personal network to the aggregate community network to identify gaps in their knowledge.

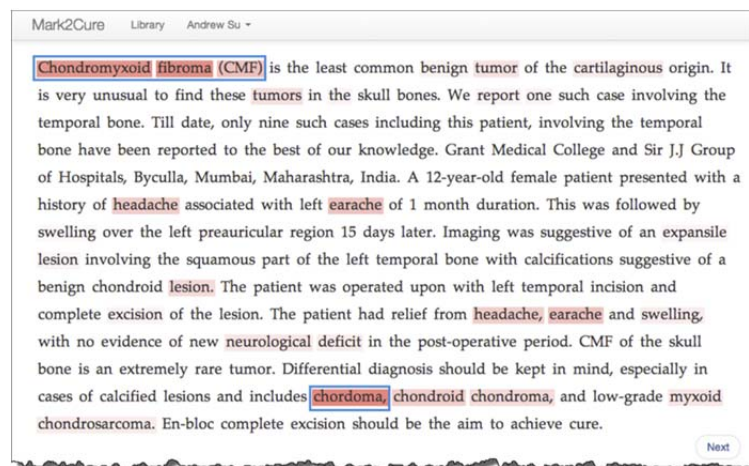


Figure 6. Prototype of the feedback mechanism for Mark2Cure. This screenshot shows the result after the user submits their annotations of diseases. On the feedback screen, the user's annotations are shown as blue boxes, and the community consensus annotations are shown as red highlights.

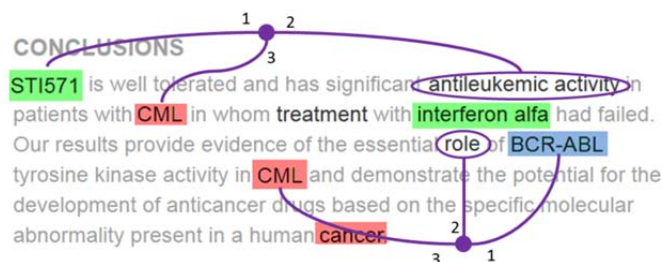


Figure 7. Annotation of biomedical relationships. Using the concluding paragraph of the abstract in [1] as an example, key biomedical concepts for drugs (green), diseases (red), and genes (blue) have been highlighted. Annotated relationships are shown in purple (1 – subject, 2 – predicate, 3 – object). An additional possible relationship (interferon alfa – treatment – CML) is omitted for clarity. Note that the annotations shown are not the most precise structuring of the knowledge since many nuances and qualifiers are not captured. This example shows the level of annotation we expect from our patient-aligned crowd, and we believe that systematically annotating articles at this level of structure would be very valuable.

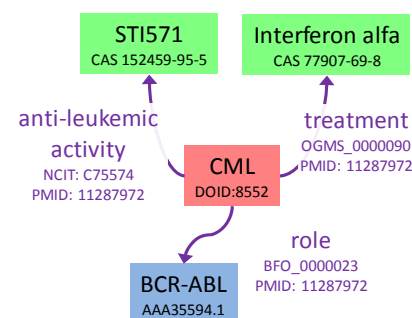


Figure 8. Biological knowledge network. This network diagram shows the structured view corresponding to the free text in **Figure 7**. In addition, all the concepts and relationships have been normalized using standard biomedical ontologies and database identifiers, and the provenance of all edges is recorded using the PMID.

PROGRESS REPORT: The proposal for the original funding period (8/2010 – 5/2014) included three aims. The funding level was 80% of the requested amount and 80% of the requested time period.

Original Aim 1: Cultivate a critical mass of users by ensuring that Gene Wiki content is both information-rich and timely. In this aim, we proposed the original idea of creating a continuously-updated, community-reviewed, and collaboratively-written review article for every human gene. Each article would be maintained within Wikipedia, and our group would ensure that the “infobox” was complete and current.

As proposed, we have expanded the scope of imported data displayed on the Gene Wiki to include a wide variety of requested content. (Note that Gene Wiki articles include information on both genes and their protein products. While semantically imprecise, this level of abstraction is appropriate for Wikipedia. Our planned work with Wikidata in **Aims 1** and **3** in this proposal will allow genes and proteins to be modeled separately.) We improved our automated update system so that content is now kept up to date automatically on a nightly basis, and we created a useful interface for creating new Gene Wiki articles on user demand [2]. While we believe these improvements are intrinsically valuable, the ultimate metric of success is community usage and participation. Our recent paper reported that the ~10,000 Gene Wiki articles are collectively viewed over 4 million times per month [4], and that number currently stands at over 5.5 million views per month. These articles are also edited almost 1,500 times per month by human editors plus an additional 1,500 times per month via automated “bots”. The high visibility of the Gene Wiki is actually the best means of assuring article quality, following the open source software mantra “Given enough eyeballs, all bugs are shallow” [106].

Original Aim 2: Integrate the Gene Wiki with WikiTrust, a system that assigns a confidence score to all contributed content in a Gene Wiki page. This aim proposed to build technical infrastructure to quantitatively assess the quality of individual Gene Wiki edits and of the Gene Wiki corpus as a whole.

WikiTrust is based on the principle that wiki text that remains unchanged over many subsequent views and edits tends to be trustworthy, and that users who historically have made trustworthy edits will tend to make trustworthy edits in the future. WikiTrust is the formalization of this principle into a quantitative score at word-level resolution, plus an easy method of visualization for that score [107-109]. WikiTrust has been deployed for all of Wikipedia, including the Gene Wiki, in the form of a Firefox plugin [110]. This system provides a unique, proven approach for automatically identifying untrustworthy content in wikis. The most important result is simply the hard evidence of the very low number and short duration of vandalism that exists in the articles of the Gene Wiki, amounting to just 0.032% of the time in a vandalized state [4].

Anecdotally, we find that the biggest challenge with Gene Wiki articles most often is that relevant information is missing, not that the presented material is inaccurate. Of course, the community editing model means that new material can be added at any time. Moreover, the presence of many excellent pages (e.g., reelin [111], *TBR1* [112] or protein C [113]) give us confidence that each Gene Wiki article will asymptotically approach our goal of a complete and accurate gene-specific review article.

Original Aim 3: Facilitate “computability” of the Gene Wiki by conversion of unstructured content to structured data. While the Gene Wiki is highly useful in its human-readable free-text form, this aim focused on additionally making community-contributed content computable. Much of our work in this area revolved around very simple text mining approaches to extract thousands of novel gene annotations relative to the Gene Ontology and the Disease Ontology [61]. We showed that these Gene Wiki-derived annotations systematically improved ontology-based enrichment analyses (also serving as a reflection of the high quality of Gene Wiki content) [61]. This approach was generalized in an automated framework to assess the quality of Gene Ontology annotations [114]. Finally, we created a semantically-explicit version of the Gene Wiki called Gene Wiki Plus that, when mashed-up with data from SNPedia [115], could be used to make complex queries on gene-disease relationships [116].

We also explored the idea of “semantic wiki links” as a way for the community to directly contribute structured data [117]. While we were strongly convinced of the scientific merits, it became clear this idea would have required a significant investment in community outreach and evangelism that was not possible given the funding reductions. However, with the recent emergence of the Wikidata project, the early design ideas and prototypes will be applied to that more robust technical framework (described more in **Aims 1** and **3**).

PROGRESS REPORT PUBLICATION LIST

1. Good BM, Su AI. Crowdsourcing for Bioinformatics. *Bioinformatics*. 2013 Jun 19. [Epub ahead of print]. PMID: In progress
2. Wu C, Macleod I, Su AI. BioGPS and MyGene.info: organizing online, gene-centric information. *Nucleic Acids Res*. 2013 Jan;41 (Database issue):D561-5. doi:10.1093/nar/gks1114. Epub 2012 Nov 21. PMID: PMC3531157
3. Grogan SP, Duffy SF, Pauli C, Koziol JA, Su AI, D'Lima DD, Lotz MK. Zone-specific gene expression patterns in articular cartilage. *Arthritis Rheum*. 2013 Feb;65(2):418-28. doi: 10.1002/art.37760. PMID: PMC3558601
4. Clarke EL, Loguercio S, Good BM, Su AI. A task-based approach for Gene Ontology evaluation. *J Biomed Semantics*. 2013 Apr 15; 4 Suppl 1:S4. doi:10.1186/2041-1480-4-S1-S4. Epub 2013 Apr 15. PMID: PMC3633003
5. Good BM, Clarke EL, Loguercio S, Su AI. Building a biomedical semantic network in Wikipedia with Semantic Wiki Links. *Database (Oxford)*. 2012 Mar 20;2012:bar060. doi: 10.1093/database/bar060. Print 2012. PMID: PMC3308151
6. Good BM, Clarke EL, Loguercio S, Su AI. Linking genes to diseases with a SNPedia-Gene Wiki mashup. *J Biomed Semantics*. 2012 Apr 24;3 Suppl 1:S6. doi:10.1186/2041-3-S1-S6. PMID: PMC3337266
7. Adler BT, de Alfaro L, Kulshreshtha A, Pye I. Reputation Systems for Open Collaboration. *Commun ACM*. 2011 Aug;54(8):81-87. No abstract available. PMID: PMC3615714
8. Good BM, Clarke EL, de Alfaro L, Su AI. The Gene Wiki in 2011: community intelligence applied to human gene annotation. *Nucleic Acids Res*. 2012 Jan;40(Database issue): D1255-61. doi: 10.1093/nar/gkr925. Epub 2011 Nov 10. PMID: PMC3245148
9. Good BM, Howe DG, Lin SM, Kibbe WA, Su AI. Mining the Gene Wiki for functional genomic knowledge. *BMC Genomics*. 2011 Dec 13;12:603. doi: 10.1186/1471-2164-12-603. PMID: PMC3271090
10. Good BM, Su AI. Games with a scientific purpose. *Genome Biol*. 2011 Dec 28;12(12):135. doi: 10.1186/gb-2011-12-12-135. PMID: PMC3334605
11. Adler T, de Alfaro L, Mola-Velasco SM, Rosso P, West AG. Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features. *Proceedings of CICLing, Conference on Intelligent Text Processing and Computational Linguistics*. 2011. PMID: Not applicable.

REFERENCES

1. Druker BJ, Talpaz M, Resta DJ, Peng B, Buchdunger E, Ford JM, Lydon NB, Kantarjian H, Capdeville R, Ohno-Jones S, Sawyers CL. **Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia.** *The New England journal of medicine.* 2001;**344**(14):1031-7.
2. Huss JW, 3rd, Lindenbaum P, Martone M, Roberts D, Pizarro A, Valafar F, Hogenesch JB, Su AI. **The Gene Wiki: community intelligence applied to human gene annotation.** *Nucleic Acids Res.* 2010;**38**(Database issue):D633-9. (PMC2808918)
3. Osborne M, Petrovic S, McCreadie R, Macdonald C, Ounis I. **Bieber no more: First Story Detection using Twitter and Wikipedia.** *SIGIR 2012 Workshop on Time-aware Information Access (#TAIA2012).* 2012.
4. Good BM, Clarke EL, de Alfaro L, Su AI. **The Gene Wiki in 2011: community intelligence applied to human gene annotation.** *Nucleic Acids Res.* 2012;**40**(Database issue):D1255-61. (PMC3245148)
5. **sulab / GeneBot — Bitbucket.** Available from: <https://bitbucket.org/sulab/genebot>.
6. **User:Chinmay26 - Wikidata.** Available from: <http://www.wikidata.org/wiki/User:Chinmay26>.
7. **Wikidata:Molecular biology task force - Wikidata.** Available from: <http://www.wikidata.org/wiki/WD:MBTF>.
8. **Wikipedia:WikiProject Medicine - Wikipedia, the free encyclopedia.**
9. **Wikipedia:WikiProject Pharmacology - Wikipedia, the free encyclopedia.** Available from: http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Pharmacology.
10. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. **Exome sequencing as a tool for Mendelian disease gene discovery.** *Nature reviews Genetics.* 2011;**12**(11):745-55.
11. Schriml LM, Arze C, Nadendla S, Chang YW, 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. (PMC3245088)
12. **DGIdb - Mining the Druggable Genome.** Available from: <http://dgidb.org>
13. Zhu F, Han B, Kumar P, Liu X, Ma X, Wei X, Huang L, Guo Y, Han L, Zheng C, Chen Y. **Update of TTD: Therapeutic Target Database.** *Nucleic Acids Res.* 2010;**38**(Database issue):D787-91. (PMC2808971)
14. Yeh P, Chen H, Andrews J, Naser R, Pao W, Horn L. **DNA-Mutation Inventory to Refine and Enhance Cancer Treatment (DIRECT): a catalog of clinically relevant cancer mutations to enable genome-directed anticancer therapy.** *Clinical cancer research : an official journal of the American Association for Cancer Research.* 2013;**19**(7):1894-901.
15. Somaiah N, Simon GR. **Molecular targeted agents and biologic therapies for lung cancer.** *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer.* 2011;**6**(11 Suppl 4):S1758-85.
16. Rask-Andersen M, Almen MS, Schioth HB. **Trends in the exploitation of novel drug targets.** *Nature reviews Drug discovery.* 2011;**10**(8):579-90.
17. McDonagh EM, Whirl-Carrillo M, Garten Y, Altman RB, Klein TE. **From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource.** *Biomarkers in medicine.* 2011;**5**(6):795-806. (PMC3339046)
18. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djombou Y, Eisner R, Guo AC, Wishart DS. **DrugBank 3.0: a comprehensive resource for 'omics' research on drugs.** *Nucleic Acids Res.* 2011;**39**(Database issue):D1035-41. (PMC3013709)
19. Davis AP, Murphy CG, Johnson R, Lay JM, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, King BL, Rosenstein MC, Wiegers TC, Mattingly CJ. **The Comparative Toxicogenomics Database: update 2013.** *Nucleic Acids Res.* 2013;**41**(Database issue):D1104-14. (PMC3531134)

20. Kuhn M, Szklarczyk D, Franceschini A, von Mering C, Jensen LJ, Bork P. **STITCH 3: zooming in on protein-chemical interactions**. *Nucleic Acids Res.* 2012;**40**(Database issue):D876-80. (PMC3245073)
21. **miR-155 - Wikipedia, the free encyclopedia**. Available from: <http://en.wikipedia.org/wiki/MiR-155>.
22. Elton TS, Selemo H, Elton SM, Parinandi NL. **Regulation of the MIR155 host gene in physiological and pathological processes**. *Gene*. 2012.
23. **Epoxide hydrolase 2 - Wikipedia, the free encyclopedia**. Available from: http://en.wikipedia.org/wiki/Epoxide_hydrolase_2.
24. Harris TR, Hammock BD. **Soluble epoxide hydrolase: Gene structure, expression and deletion**. *Gene*. 2013.
25. **SFTPA2 - Wikipedia, the free encyclopedia**. Available from: <http://en.wikipedia.org/wiki/SFTPA2>.
26. **Pulmonary surfactant-associated protein A1 - Wikipedia, the free encyclopedia**. Available from: <http://en.wikipedia.org/wiki/SFTPA1>.
27. Silveyra P, Floros J. **Genetic complexity of the human surfactant-associated proteins SP-A1 and SP-A2**. *Gene*. 2012. (PMC3570704)
28. **Category:CAAO - GONUTS**. Available from: <http://gowiki.tamu.edu/wiki/index.php/Category:CAAO>.
29. Hatfull GF, Pedulla ML, Jacobs-Sera D, Cichon PM, Foley A, Ford ME, Gonda RM, Houtz JM, Hryckowian AJ, Kelchner VA, Namburi S, Pajcini KV, Popovich MG, Schleicher DT, Simanek BZ, Smith AL, Zdanowicz GM, Kumar V, Peebles CL, et al. **Exploring the mycobacteriophage metaproteome: phage genomics as an educational platform**. *PLoS genetics*. 2006;**2**(6):e92. (PMC1475703)
30. Hingamp P, Brochier C, Talla E, Gautheret D, Thieffry D, Herrmann C. **Metagenome annotation using a distributed grid of undergraduate students**. *PLoS biology*. 2008;**6**(11):e296. (PMC2586363)
31. **Wikipedia:WikiProject AP Biology Bapst 2013 - Wikipedia, the free encyclopedia**. Available from: http://en.wikipedia.org/wiki/Wikipedia:WikiProject_AP_Biology_Bapst_2013.
32. **Global Genome Initiative**. Available from: <http://www.mnh.si.edu/ggi>.
33. **What Genomic Research Can Tell Us About the Earth's Biodiversity | Science & Nature | Smithsonian Magazine**. Available from: <http://www.smithsonianmag.com/science-nature/What-Genomic-Research-Can-Tell-Us-About-the-Earths-Biodiversity-207249761.html>.
34. **Undergraduate Research: Built Environment Genomes | microBEnet: The microbiology of the Built Environment network**. Available from: <http://www.microbe.net/undergraduate-research-built-environment-genomes/>.
35. **Science Education News: Twelve New Schools Will Offer Year-Long Phage Genomics Course | Howard Hughes Medical Institute (HHMI)**. Available from: <http://www.hhmi.org/news/seacohort420010127.html>.
36. Pope WH, Jacobs-Sera D, Russell DA, Peebles CL, Al-Atrache Z, Alcoser TA, Alexander LM, Alfano MB, Alford ST, Amy NE, Anderson MD, Anderson AG, Ang AA, Ares M, Jr., Barber AJ, Barker LP, Barrett JM, Barshop WD, Bauerle CM, et al. **Expanding the diversity of mycobacteriophages: insights into genome architecture and evolution**. *PLoS ONE*. 2011;**6**(1):e16329. (PMC3029335)
37. Manichanh C, Borrueal N, Casellas F, Guarner F. **The gut microbiota in IBD**. *Nature reviews Gastroenterology & hepatology*. 2012;**9**(10):599-608.
38. Tremaroli V, Backhed F. **Functional interactions between the gut microbiota and host metabolism**. *Nature*. 2012;**489**(7415):242-9.
39. Dejea C, Wick E, Sears CL. **Bacterial oncogenesis in the colon**. *Future Microbiol.* 2013;**8**(4):445-60.
40. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muniz-Rascado L, Garcia-Sotelo JS, Weiss V, Solano-Lira H, Martinez-Flores I, Medina-Rivera A, Salgado-Orsorio G, Alquicira-Hernandez S, Alquicira-Hernandez K, Lopez-Fuentes A, Porron-Sotelo L, Huerta AM, Bonavides-Martinez C, Balderas-Martinez YI, Pannier L, et al. **RegulonDB v8.0: omics data sets, evolutionary**

- conservation, regulatory phrases, cross-validated gold standards and more.** *Nucleic Acids Res.* 2013;**41**(Database issue):D203-13. (PMC3531196)
41. 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 MA, Soubigou G, Regnault B, Coppee JY, Lecuit M, Johansson J, Cossart P. **The *Listeria* transcriptional landscape from saprophytism to virulence.** *Nature.* 2009;**459**(7249):950-6.
 42. Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, White O, Ketchum KA, Dodson R, Hickey EK, Gwinn M, Dougherty B, Tomb JF, Fleischmann RD, Richardson D, Peterson J, Kerlavage AR, Quackenbush J, Salzberg S, et al. **Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*.** *Nature.* 1997;**390**(6660):580-6.
 43. Imai D, Holden K, Velazquez EM, Feng S, Hodzic E, Barthold SW. **Influence of arthritis-related protein (BBF01) on infectivity of *Borrelia burgdorferi* B31.** *BMC microbiology.* 2013;**13**(1):100.
 44. Karna SL, Prabhu RG, Lin YH, Miller CL, Seshu J. **Contributions of environmental signals and conserved residues to the functions of carbon storage regulator A of *Borrelia burgdorferi* (csrABb).** *Infection and immunity.* 2013.
 45. Patton TG, Brandt KS, Nolder C, Clifton DR, Carroll JA, Gilmore RD. ***Borrelia burgdorferi* bba66 Gene Inactivation Results in Attenuated Mouse Infection by Tick Transmission.** *Infection and immunity.* 2013;**81**(7):2488-98.
 46. Tilly K, Bestor A, Rosa PA. **Lipoprotein succession in *Borrelia burgdorferi*: similar but distinct roles for OspC and VlsE at different stages of mammalian infection.** *Molecular microbiology.* 2013.
 47. Troxell B, Ye M, Yang Y, Carrasco SE, Lou Y, Yang XF. **Manganese and Zinc Regulate Virulence Determinants in *Borrelia burgdorferi*.** *Infection and immunity.* 2013.
 48. Wood E, Tamborero S, Mingarro I, Esteve-Gassent MD. **BB0172, a *Borrelia burgdorferi* Outer Membrane Protein that Binds Integrin alpha3beta1.** *Journal of bacteriology.* 2013.
 49. Perteua M, Ayanbule K, Smedinghoff M, Salzberg SL. **OperonDB: a comprehensive database of predicted operons in microbial genomes.** *Nucleic Acids Res.* 2009;**37**(Database issue):D479-82. (PMC2686487)
 50. Gaudet P, Livstone MS, Lewis SE, Thomas PD. **Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium.** *Briefings in bioinformatics.* 2011;**12**(5):449-62. (PMC3178059)
 51. Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R, Angenent LT, Ley RE. **Succession of microbial consortia in the developing infant gut microbiome.** *Proc Natl Acad Sci U S A.* 2011;**108** Suppl 1:4578-85. (PMC3063592)
 52. Pimmer CR, Papakostas S, Leder EH, Davis MJ, Ragan MA. **Annotated genes and nonannotated genomes: cross-species use of Gene Ontology in ecology and evolution research.** *Molecular ecology.* 2013.
 53. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A, Pandey G, Yunes JM, Talwalkar AS, Repo S, Souza ML, Piovesan D, Casadio R, Wang Z, Cheng J, et al. **A large-scale evaluation of computational protein function prediction.** *Nature methods.* 2013;**10**(3):221-7. (PMC3584181)
 54. **sulab / MicrobeBase — Bitbucket.** Available from: <https://bitbucket.org/sulab/microbebase>.
 55. Lee E, Harris N, Gibson M, Chetty R, Lewis S. **Apollo: a community resource for genome annotation editing.** *Bioinformatics.* 2009;**25**(14):1836-7. (PMC2705230)
 56. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. **JBrowse: a next-generation genome browser.** *Genome research.* 2009;**19**(9):1630-8. (PMC2752129)
 57. **WikidataQuery API.** Available from: http://208.80.153.172/api_documentation.html.
 58. **Manual:Pywikipediabot/Wikidata - MediaWiki.** Available from: <http://www.mediawiki.org/wiki/Manual:Pywikipediabot/Wikidata>.

59. **GO Annotation File GAF 2.0 Format Guide**. Available from: http://www.geneontology.org/GO.format.gaf-2_0.shtml.
60. **Current Annotations**. Available from: <http://www.geneontology.org/GO.downloads.annotations.shtml>.
61. Good BM, Howe DG, Lin SM, Kibbe WA, Su AI. **Mining the Gene Wiki for functional genomic knowledge**. *BMC Genomics*. 2011;**12**:603. (PMC3271090)
62. Personal Communication, Diane Boehr, Head of Cataloging, National Library of Medicine
63. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, Jupe S, Kalatskaya I, Mahajan S, May B, Ndegwa N, Schmidt E, Shamovsky V, Yung C, Birney E, et al. **Reactome: a database of reactions, pathways and biological processes**. *Nucleic Acids Res*. 2011;**39**(Database issue):D691-7. (PMC3013646)
64. Danielsen JM, Sylvestersen KB, Bekker-Jensen S, Szklarczyk D, Poulsen JW, Horn H, Jensen LJ, Mailand N, Nielsen ML. **Mass spectrometric analysis of lysine ubiquitylation reveals promiscuity at site level**. *Mol Cell Proteomics*. 2011;**10**(3):M110 003590. (PMC3047152)
65. Havugimana PC, Hart GT, Nepusz T, Yang H, Turinsky AL, Li Z, Wang PI, Boutz DR, Fong V, Phanse S, Babu M, Craig SA, Hu P, Wan C, Vlasblom J, Dar VU, Bezginov A, Clark GW, Wu GC, et al. **A census of human soluble protein complexes**. *Cell*. 2012;**150**(5):1068-81. (PMC3477804)
66. Kristensen AR, Gsponer J, Foster LJ. **A high-throughput approach for measuring temporal changes in the interactome**. *Nature methods*. 2012;**9**(9):907-9.
67. Linding R, Jensen LJ, Ostheimer GJ, van Vugt MA, Jorgensen C, Miron IM, Diella F, Colwill K, Taylor L, Elder K, Metalnikov P, Nguyen V, Pasculescu A, Jin J, Park JG, Samson LD, Woodgett JR, Russell RB, Bork P, et al. **Systematic discovery of in vivo phosphorylation networks**. *Cell*. 2007;**129**(7):1415-26. (PMC2692296)
68. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, et al. **Towards a proteome-scale map of the human protein-protein interaction network**. *Nature*. 2005;**437**(7062):1173-8.
69. Vinayagam A, Stelzl U, Foulle R, Plassmann S, Zenkner M, Timm J, Assmus HE, Andrade-Navarro MA, Wanker EE. **A directed protein interaction network for investigating intracellular signal transduction**. *Science signaling*. 2011;**4**(189):rs8.
70. Wang J, Huo K, Ma L, Tang L, Li D, Huang X, Yuan Y, Li C, Wang W, Guan W, Chen H, Jin C, Wei J, Zhang W, Yang Y, Liu Q, Zhou Y, Zhang C, Wu Z, et al. **Toward an understanding of the protein interaction network of the human liver**. *Mol Syst Biol*. 2011;**7**:536. (PMC3261708)
71. Kilicoglu H, Shin D, Fiszman M, Roseblat G, Rindflesch TC. **SemMedDB: a PubMed-scale repository of biomedical semantic predications**. *Bioinformatics*. 2012;**28**(23):3158-60. (PMC3509487)
72. **BioNLP 2012 Program**. Available from: <http://compbio.ucdenver.edu/BioNLP2012/program.shtml>.
73. **Tasks - BioNLP-ST 2013**. Available from: <https://sites.google.com/site/bionlpst2013/tasks>.
74. Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, Murino L, Tagliaferri R, Brunetti-Pierri N, Isacchi A, di Bernardo D. **Discovery of drug mode of action and drug repositioning from transcriptional responses**. *Proc Natl Acad Sci U S A*. 2010;**107**(33):14621-6. (PMC2930479)
75. Laenen G, Thorrez L, Bornigen D, Moreau Y. **Finding the targets of a drug by integration of gene expression data with a protein interaction network**. *Molecular bioSystems*. 2013;**9**(7):1676-85.
76. Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tumer Z, Pociot F, Tommerup N, Moreau Y, Brunak S. **A human phenome-interactome network of protein complexes implicated in genetic disorders**. *Nat Biotechnol*. 2007;**25**(3):309-16.
77. Pritchard JR, Bruno PM, Hemann MT, Lauffenburger DA. **Predicting cancer drug mechanisms of action using molecular network signatures**. *Molecular bioSystems*. 2013;**9**(7):1604-19.

78. Suthram S, Dudley JT, Chiang AP, Chen R, Hastie TJ, Butte AJ. **Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets.** *PLoS Comput Biol.* 2010;**6**(2):e1000662. (PMC2816673)
79. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. **Network-based classification of breast cancer metastasis.** *Mol Syst Biol.* 2007;**3**:140. (PMC2063581)
80. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK, Leonardson A, Castellini LW, Wang S, Champy MF, Zhang B, Emilsson V, Doss S, Ghazalpour A, Horvath S, et al. **Variations in DNA elucidate molecular networks that cause disease.** *Nature.* 2008;**452**(7186):429-35. (PMC2841398)
81. Mummery-Widmer JL, Yamazaki M, Stoeger T, Novatchkova M, Bhalerao S, Chen D, Dietzl G, Dickson BJ, Knoblich JA. **Genome-wide analysis of Notch signalling in Drosophila by transgenic RNAi.** *Nature.* 2009;**458**(7241):987-92. (PMC2988197)
82. Zhang B, Gaiteri C, Bodea LG, Wang Z, McElwee J, Podtelezhnikov AA, Zhang C, Xie T, Tran L, Dobrin R, Fluder E, Clurman B, Melquist S, Narayanan M, Suver C, Shah H, Mahajan M, Gillis T, Mysore J, et al. **Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease.** *Cell.* 2013;**153**(3):707-20. (PMC3677161)
83. Dutkowsk J, Kramer M, Surma MA, Balakrishnan R, Cherry JM, Krogan NJ, Ideker T. **A gene ontology inferred from molecular networks.** *Nat Biotechnol.* 2013;**31**(1):38-45. (PMC3654867)
84. **Non-scientists researching diseases (with tweets) - andrewsu - Storify** Available from: <http://storify.com/andrewsu/non-scientists-and-researching-diseases>.
85. **Alzheimer Research Forum Home.** Available from: <http://www.alzforum.org/>.
86. **Celiac Disease Foundation.** Available from: <http://www.celiac.org/>.
87. **LymeMD :: Lyme Disease Research Foundation.** Available from: <http://www.lymemd.org/>
88. Frost J, Okun S, Vaughan T, Heywood J, Wicks P. **Patient-reported outcomes as a source of evidence in off-label prescribing: analysis of data from PatientsLikeMe.** *Journal of medical Internet research.* 2011;**13**(1):e6. (PMC3221356)
89. Wicks P, Massagli M, Frost J, Brownstein C, Okun S, Vaughan T, Bradley R, Heywood J. **Sharing health data for better outcomes on PatientsLikeMe.** *Journal of medical Internet research.* 2010;**12**(2):e19. (PMC2956230)
90. Wicks P, Vaughan TE, Massagli MP, Heywood J. **Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm.** *Nat Biotechnol.* 2011;**29**(5):411-4.
91. **Live better, together! | PatientsLikeMe.** Available from: <http://www.patientslikeme.com/>, .
92. **MyHealthTeams.** MyHealthTeams]. Available from: <http://www.myhealthteams.com>.
93. **Smart Patients.** Available from: <https://www.smartpatients.com/>.
94. **Hunting down my son's killer.** Available from: <http://matt.might.net/articles/my-sons-killer/>.
95. **Parents of twins with rare disease fight for FDA approval of treatment – USATODAY.com.** Available from: <http://usatoday30.usatoday.com/news/health/story/health/story/2012-03-06/Parents-of-twins-with-rare-disease-fight-for-FDA-approval-of-treatment/53383696/1>.
96. **Our Story | Prion Alliance.** Available from: <http://www.prionalliance.org/our-story/>.
97. **Opening a Can of Worms | The Scientist Magazine®.** Available from: <http://www.the-scientist.com/?articles.view/articleNo/30802/title/Opening-a-Can-of-Worms/>.
98. **Founder's Story | Adenoid Cystic Carcinoma Research Foundation.** Available from: <http://www.accrf.org/about-accrf/founders-story/>.
99. Haijun Z, Lingren T, Deleger L, Qi L, Kaiser M, Stoutenborough L, Solti I, editors. **Cheap, Fast, and Good Enough for the Non-biomedical Domain but is It Usable for Clinical Natural Language Processing? Evaluating Crowdsourcing for Clinical Trial Announcement Named Entity**

Annotations. *Healthcare Informatics, Imaging and Systems Biology (HISB), 2012 IEEE Second International Conference on; 2012 27-28 Sept. 2012.*

100. MacLean DL, Heer J. **Identifying medical terms in patient-authored text: a crowdsourcing-based approach.** *Journal of the American Medical Informatics Association.* 2013.
101. Burger JD, Doughty E, Bayer S, Tresner-Kirsch D, Wellner B, Aberdeen J, Lee K, Kann MG, Hirschman L. **Validating candidate gene-mutation relations in MEDLINE abstracts via crowdsourcing.** *Proceedings of the 8th international conference on Data Integration in the Life Sciences; College Park, MD.* 2368527: Springer-Verlag; 2012. p. 83-91.
102. Yetisgen-Yildiz M, Solti I, Xia F, Halgrim SR. **Preliminary experience with Amazon's Mechanical Turk for annotating medical named entities.** *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk; Los Angeles, California.* 1866724: Association for Computational Linguistics; 2010. p. 180-3.
103. Salgado D, Krallinger M, Depaule M, Drula E, Tendulkar AV, Leitner F, Valencia A, Marcelle C. **MyMiner: a web application for computer-assisted biocuration and text annotation.** *Bioinformatics.* 2012;**28**(17):2285-7.
104. Wei CH, Kao HY, Lu Z. **PubTator: a web-based text mining tool for assisting biocuration.** *Nucleic Acids Res.* 2013;**41**(Web Server issue):W518-22. (PMC3692066)
105. Shah NH, Bhatia N, Jonquet C, Rubin D, Chiang AP, Musen MA. **Comparison of concept recognizers for building the Open Biomedical Annotator.** *BMC bioinformatics.* 2009;**10** Suppl 9:S14. (PMC2745685)
106. **Linus's Law - Wikipedia, the free encyclopedia.** Available from: http://en.wikipedia.org/wiki/Linus's_Law.
107. Alfaro LD, Kulshreshtha A, Pye I, Adler BT. **Reputation systems for open collaboration.** *Commun ACM.* 2011;**54**(8):81-7.
108. Alfaro Ld, Shavlovsky M. **Attributing authorship of revisioned content.** *Proceedings of the 22nd international conference on World Wide Web; Rio de Janeiro, Brazil.* 2488419: International World Wide Web Conferences Steering Committee; 2013. p. 343-54.
109. Adler BT, Alfaro LD, Mola-Velasco SM, Rosso P, West AG. **Wikipedia vandalism detection: combining natural language, metadata, and reputation features.** *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part II; Tokyo, Japan.* 1964776: Springer-Verlag; 2011. p. 277-88.
110. Adler BT, Alfaro Ld. **A content-driven reputation system for the wikipedia.** *Proceedings of the 16th international conference on World Wide Web; Banff, Alberta, Canada.* 1242608: ACM; 2007. p. 261-70.
111. **Reelin - Wikipedia, the free encyclopedia.** Available from: <http://en.wikipedia.org/wiki/Reelin>.
112. **TBR1 - Wikipedia, the free encyclopedia.** Available from: <http://en.wikipedia.org/wiki/TBR1>.
113. **Protein C - Wikipedia, the free encyclopedia.** Available from: http://en.wikipedia.org/wiki/Protein_C.
114. Clarke EL, Loguercio S, Good BM, Su AI. **A task-based approach for Gene Ontology evaluation.** *Journal of biomedical semantics.* 2013;**4** Suppl 1:S4. (PMC3633003)
115. Cariaso M, Lennon G. **SNPedia: a wiki supporting personal genome annotation, interpretation and analysis.** *Nucleic Acids Res.* 2012;**40**(Database issue):D1308-12. (PMC3245045)
116. Good BM, Clarke EL, Loguercio S, Su AI. **Linking genes to diseases with a SNPedia-Gene Wiki mashup.** *Journal of biomedical semantics.* 2012;**3** Suppl 1:S6. (PMC3337266)
117. Good BM, Clarke EL, Loguercio S, Su AI. **Building a biomedical semantic network in Wikipedia with Semantic Wiki Links.** *Database : the journal of biological databases and curation.* 2012;**2012**:bar060. (PMC3308151)

RESOURCE SHARING PLAN

Data sharing plan: Because the vast majority of our data handling happens directly on Wikipedia, all the content generated is licensed under the Creative Commons Attribution-ShareAlike License (CC BY-SA 3.0) and can be accessed via the Wikipedia API. For the results of the text mining associated with Gene Wiki Plus, we additionally generate monthly RDF dumps that are available for download at <http://genewikiplus.org/wiki/GeneWiki:Data>. For the proposed work, the same general principles apply. Wikidata is also licensed under CC BY-SA 3.0, so the data associated with Aims 1 and 3 will be available under those terms. The data generated in Aim 4 will be disseminated as RDF dumps using OWL ontologies.

Software sharing plan: All code generated as part of this project is hosted in public code repositories. For the current project period, the relevant code repositories are here:

- pygenewiki (<https://bitbucket.org/sulab/pygenewiki>): perform the nightly sync between source database and Gene Wiki templates
- gwsync (<https://bitbucket.org/sulab/gwsync>): code to sync between Gene Wiki pages and Gene Wiki Plus
- genewiki-miner (<https://bitbucket.org/sulab/genewiki-miner>): Information extraction and parsing related to the Gene Wiki

A similar pattern will be followed for the proposed project period.