

# Mining the Gene Wiki to prioritize literature curation efforts

Andrew I. Su, Ph.D.

Genomics Institute of the Novartis Research Foundation  
10675 John Jay Hopkins Drive  
San Diego, CA 92121  
1-858-812-1500  
asu@gnf.org

Douglas G. Howe, Ph.D.

The Zebrafish Information Network  
5291 University of Oregon  
Eugene, OR, 97403  
1-541-346-0120  
dhowe@zfin.org

## ABSTRACT

The Gene Wiki is an effort to harness the Long Tail of biomedical scientists toward the goal of collaboratively and comprehensively annotating the function of human genes. The Gene Wiki was initiated within the online encyclopedia Wikipedia to take advantage of a critical mass of readers, contributors, and content. Previous analyses have shown that viewing and editing of the Gene Wiki already have a robust and growing user base. Here, we describe the systematic mining of Gene Wiki content to identify inline citations to the biomedical literature. We demonstrate that this analysis provides a useful tool for the Model Organism Database (MOD) community and their gene annotation curators. These efforts are part of a broader goal to convert community-contributed unstructured content in Gene Wiki into structured gene annotations. More information on the Gene Wiki can be found at [http://en.wikipedia.org/wiki/Portal:Gene\\_Wiki](http://en.wikipedia.org/wiki/Portal:Gene_Wiki).

## Categories and Subject Descriptors

H.1.2 [Models and Principles]: User/Machine Systems – *Human information processing*. H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing – *Dictionaries*. H.3.5 [Information Interfaces and Presentation]: Group and Organization Interfaces – *Computer-supported cooperative work, Evaluation/methodology*. J.3 [Computer Applications]: Life and Medical Sciences – *Biology and genetics*.

## General Terms

Management, Documentation, Human Factors.

## Keywords

Gene Wiki, community intelligence, gene annotation, biocuration, model organism database

## 1. INTRODUCTION

### 1.1 Gene Annotation

Comprehensively describing the function of all ~25 000 human genes is a critical but formidable task for the biomedical community. Related efforts in model organisms to also describe the function of all genes magnifies this challenge. Currently, these gene annotation efforts rely in large part on centralized efforts to manually curate the biomedical literature at Model Organism Databases (MOD) [1-4].

These manual curation efforts have been shown to produce very high quality gene annotation data, typically utilizing common

biomedical ontologies like the Gene Ontology (GO) [5]. Structured annotations like these enable many downstream bioinformatics and statistical analyses of biological data (e.g., [6]).

However, there is growing recognition that these centralized curation efforts do not scale with the rapid growth of the biomedical literature [7]. Further, as whole genome sequencing is becoming rapid and economical, new animal models are being utilized for genetics research at an increasing rate. Research utilizing most of these newer model organisms lacks the resources to support a dedicated professional curation staff. Consequently, the curation effort for these organisms falls squarely on the research community.

PubMed, the primary database of biomedical articles, currently has over 19 million citations and grows at a rate of over 800 000 articles per year. Comprehensively curating of all these articles for knowledge relevant to gene function is a daunting task for even the most dedicated and skilled organizations. One alternative curation model to help address this issue is community curation in wiki systems. Direct community curation has the advantage of utilizing the large number of biomedical scientists to help manage the large volume of curation.

### 1.2 The Gene Wiki

Recently, the popularity of a principle called the Long Tail has been increasing as a way to address issues of massive scale. In the context of information processing and knowledge generation, the Long Tail harnesses a large population of individuals, each of whom contributes relatively little. Although the contributions are individually small, they can collectively be quite substantial. The online encyclopedia Wikipedia is perhaps the most notable and successful example of a Long Tail system.

In contrast to the Long Tail, the current system of manual curation relies on a relatively small number of professional curators each contributing a large amount of content. To apply the Long Tail principle to the gene annotation challenge, we initiated an effort called the Gene Wiki [8]. We envisioned a wiki page for each human gene, where the scientific community at large would collaboratively summarize the current knowledge about each gene. To seed this process, we created gene "stubs" that contained a systematic baseline of gene annotation mined from existing annotation databases.

We also specifically chose to create the Gene Wiki directly within Wikipedia itself. The primary motivation was to benefit from the critical mass of readers, contributors, and content already existing within Wikipedia. A recent analysis of the Gene Wiki's usage statistics suggests that this goal was at least partially achieved, with approximately 3.9 million total page views and 1100 total edits per month [9]. In addition, greater than 85% of Gene Wiki

pages were found on the first page of Google search results when searching by gene symbol. In total, the Gene Wiki contains over 57 megabytes of text content, and between January and June of 2009 grew by 2.28 megabytes (approximately equal to the text in 19 research articles in *PLoS Biology*) [9].

Other community gene annotation efforts have also been developed on stand-alone wiki instances [10, 11]. Although these efforts have more difficulties creating and maintaining critical mass, they have the advantage of having greater control over the data model. Specifically, these efforts can natively accept and store structured gene annotation data.

The Gene Wiki (and Wikipedia as a whole) is fundamentally an unstructured data repository. The majority of contributed material is in the form of free text, supplemented by figures, diagrams, photos, and tables. The greatest challenge for the Gene Wiki is to convert the wealth of unstructured content into structured data suitable for downstream computations.

Here, we describe one approach for locating publications cited in Gene Wiki that should be prioritized for structured annotation through existing literature curation pipelines.

## 2. APPROACH AND RATIONALE

Like most scientific papers, the majority of contributions to Wikipedia are unstructured. However, editors are highly encouraged to cite reliable resources to support their contributed statements [12]. In standard biomedical literature, statements often reference PubMed-indexed articles, and this practice has been continued in biomedical articles in Wikipedia. In addition, online tools have been created to facilitate the simple and consistent formatting of these references based on a PubMed Identifier [13].

Here, we mined all available Gene Wiki articles for inline citations to PubMed. We envisioned that these data could be used in at least two ways.

First, inline citations in the Gene Wiki can be used to identify and prioritize articles in the older literature. Systematic curation efforts by MODs have only been established in the last 5-10 years, and curation efforts have generally focused on newly-published papers. However, many seminal papers on gene function occur in the older literature, and the absence of curated annotations derived from these articles often leaves notable information gaps in the manually curated gene annotation databases. We hypothesized that inline citations in the Gene Wiki would be useful to curators as a means to identify and prioritize specific articles in the older literature that are likely to contain gene annotation information.

Second, when combined with the sentence immediately preceding the PubMed citation, we hypothesized that the Gene Wiki report can lead to more efficient determination of specific gene annotations. Specifically targeting inline citations means that the contributor has already provided scientific context as to why the particular article is relevant to the gene whose page is being edited. We expect that a citation report that includes, among other information, the entire Gene Wiki sentence citing the reference will improve the throughput of centralized curation efforts.

The overall envisioned workflow is summarized in Figure 1.

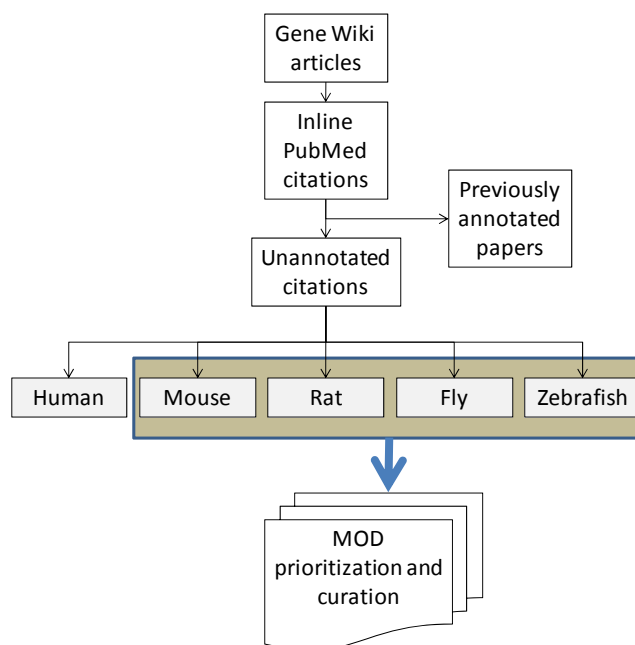


Figure 1. Overview of the Gene Wiki inline citation mining.

## 3. RESULTS

### 3.1 Citation Mining

We parsed all 9797 Gene Wiki pages and isolated all occurrences of inline citations to PubMed. All citations that were previously added in automated edits by our team were removed, leaving 6344 citations that were contributed by Wikipedia editors. As expected, the distribution of references among articles is highly skewed, with many articles having a few citations and a few articles having many citations (Figure 2). This power law-like distribution mirrors the broader pattern of gene annotation found in PubMed and NIH-funded grant abstracts [14].

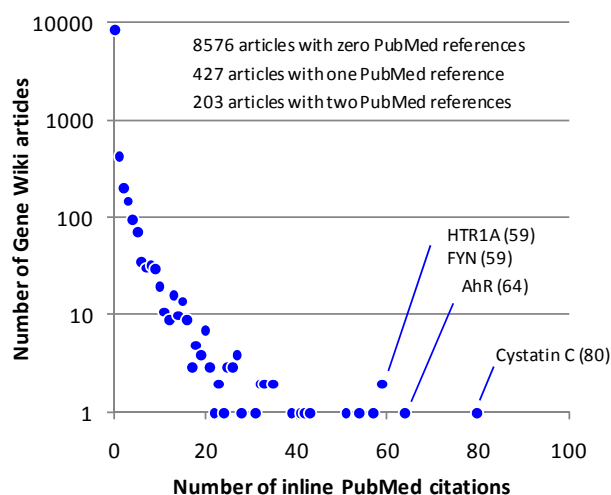
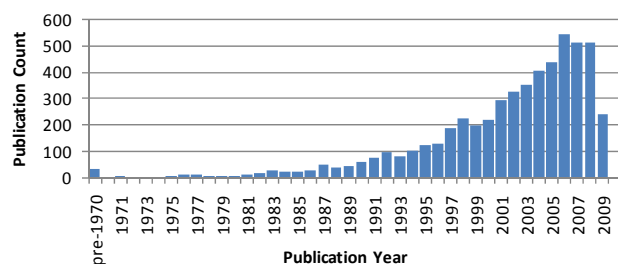


Figure 2. Analysis of the distribution of inline PubMed citations in all Gene Wiki articles.

We then analyzed the distribution of publication years from the inline Gene Wiki citations to PubMed (Figure 3). Of the 5454 unique PubMed articles cited, there were 1624 (30%) that were published before 2000, and 346 (6%) that were published before 1990.



**Figure 3. Distribution of publication years for all inline PubMed citations in the Gene Wiki.**

To determine which of these cited PubMed articles had been previously curated for gene functional data, we examined the GO annotation files (obtained from the GO consortium web site [15]) for five organisms: human (downloaded Nov 30 2009), mouse (Dec 11 2009), rat (Nov 21 2009), fly (Nov 28 2009), and zebrafish (Dec 14 2009). Although the Gene Wiki has a focus on human gene function, contributors often add content based on experiments in model organisms when those data are presumed to be relevant to the human gene. We felt these five organisms were the most likely to be relevant in this context. Of the 6344 inline PubMed references in the Gene Wiki, 746 (11.8%) were already cited in support of a GO annotation in at least one of the above organisms. The remaining 5598 were considered to potentially contain novel annotations. (Since a manuscript that has already been curated by one MOD could still have contain relevant data for curation at another MOD, this count represents a lower bound of the total potential novel annotations.)

We next sought to categorize these PubMed citations according to the organism(s) to which they were relevant. Based on MeSH terms, we found that 4068 were relevant to human, 1380 to mouse, 818 to rat, 66 to fly, and 21 to zebrafish.

Because manual inspection of the MeSH categorization of articles suggested that application of MeSH Terms had a high rate of both false positive and false negative species associations for our needs, we also investigated two alternative methods to determine the species relevance of each PubMed article. First, all MODs mentioned above maintain an index of PubMed articles relevant to their specific species. Using these indices, we found 653 mouse articles, 183 rat articles, 33 fly articles, and 11 zebrafish articles, as indexed by the Mouse Genome Informatics (MGI), the Rat Genome Database (RGD), FlyBase, and the Zebrafish Information Network (ZFIN) respectively. (Note that there is no human count because there is no official human curation organization.) Second, we consulted NCBI's "gene2pubmed" file as another way to associate PubMed articles with specific species. This method resulted in 1988 human, 827 mouse, 179 rat, 23 fly, and 5 zebrafish articles. There was generally good overlap between these two methods.

Interestingly, there were 500, 4752, and 2999 citations that matched none of our selected organisms according to MeSH terms, MOD indices, and NCBI, respectively. This finding suggests that there may be a significant false negative rate in assigning articles to their relevant organisms.

### 3.2 Curation Proof-of-Concept

In a real-world situation, specific inline citations could be sent to curators from MODs for official review in existing curation pipelines. To aid a curator in assessing whether an article is likely to have findings appropriate for structured gene annotation, we also extracted the full sentence from the Gene Wiki that included the inline citation. This context would enable a curator to efficiently prioritize candidate articles for curation. In the envisioned work flow, the full text of selected articles would then be curated according to existing standards. The output of that process may include a formalized representation of the unstructured Gene Wiki text, or annotations different than what the Gene Wiki author originally intended (and potentially even on different genes), or both.

To assess whether these inline citations would be useful for curators, we performed a scan through this list for specific candidate GO annotations. (The preliminary nature of this analysis did not yet warrant full curation.) A partial sampling of our findings is shown in Table 1. These candidate annotations were clearly drawn from a diverse selection of articles, species, and GO terms. Based on a the zebrafish-related articles, we estimate that ~50-60% of the articles in the Gene Wiki report contained information on gene function that could be encoded by GO annotations. In contrast, only 33% (1912 out of 5771) of the journal articles added to ZFIN between December 2003 (the inception of manual GO curation at ZFIN) and January 2010 have been cited for a GO annotation. This finding supports the contention that the Gene Wiki report is an enriched source of publications to target for structured functional gene annotation.

## 4. DISCUSSION

Although the Gene Wiki emphasizes human gene function, results from model organisms are often cited in the Gene Wiki. This property reflects a common perspective by researchers in which they often consider orthologous genes to have conserved functions. MODs, on the other hand, specifically base gene annotations on data derived from their model organism of interest. In the long term, Gene Wiki contributors need to be encouraged to be more explicit in specifying the organism from which the stated results are derived. Nevertheless, the representation of results across many organisms provides an opportunity to interface with MODs and aid their curation pipelines.

Mouse and rat are the most commonly cited organisms after human, which is not surprising given the wide range of human physiology that is modeled in these mammalian species. Nevertheless, even organisms as distant as *Drosophila* and zebrafish are represented in the Gene Wiki references. These more distant organisms provide excellent model systems in which to study, for example, vision and embryonic development. In the absence of direct human data, experimental results in these species are often the best available resource for understanding human gene function. In turn, the inclusion of these results in the Gene Wiki provides an opportunity to aid those MOD curation efforts.

**Table 1: Candidate gene annotations mined from the Gene Wiki.**

Gene Wiki article	PMID	Gene Wiki citing sentence	Species	Entrez Gene ID	Candidate annotation
HES5	17093926	Human HES5 gene binds to Notch receptor and expression of HES5 decreases during cartilage differentiation.	Human	388585	GO:0002062 -- chondrocyte differentiation
Tetherin	19091864	Initially discovered as an inhibitor to HIV-1 infection in the absence of Vpu, tetherin has also been shown to inhibit the release of other viruses such as the Lassa and Marburg virions.	Human	684	GO:0019076 -- release of virus from host
Syntaxin 3	10080545	The protein encoded by this gene is a member of the syntaxin family of cellular receptors for transport vesicles which participate in exocytosis in neutrophils.	Human	6809	GO:0006887 -- exocytosis
MEX3D	14769789	Upon binding, MEX3D has a negative regulatory action on Bcl-2 expression at the posttranscriptional level.	Human	399664	GO:0010608 -- posttranscriptional regulation of gene expression
ITM2A	10432285	ITM2A is also involved in activation of T-cells in the immune system [1] and in myocyte differentiation [2].	Mouse	16431	GO:0042110 -- T cell activation
	14984746				GO:0042692 -- muscle cell differentiation
SULF1	17920055	In adult mice, Sulf1 and Sulf2 have overlapping functions in regulating muscle regeneration.	Mouse	240725	GO:0043403 -- skeletal muscle tissue regeneration
TOX	18195075	Knockout mice that lack TOX have a severe defect in development of certain subsets of T lymphocytes.	Mouse	252838	GO:0030217 -- T cell differentiation
TRPM3	18978782	The activation causes calcium influx and subsequent insulin release, therefore it is suggested that TRPM3 modulates glucose homeostasis.	Rat	309407	GO:0050796 -- regulation of insulin secretion
					GO:0042593 -- glucose homeostasis
GPR119	16517404	Activation of the receptor has been shown to cause a reduction in food intake and body weight gain in rats.	Rat	302813	GO:0007631 -- feeding behavior
Synapsin I	323254	Synapsin I was shown to be phosphorylated by this calcium influx.	Rat	24949	GO:0071277 -- cellular response to calcium ion
Smoothed	14636583	SMO has also been shown to bind the kinesin motor protein Costal-2 and play a role in the localization of the Ci (Cubitus interruptus transcription factor) complex.	Drosophila	33196	GO:0005667 -- transcription factor complex
Frataxin	18258192	An overexpression of frataxin in Drosophila has shown an increase in antioxidant capability, resistance to oxidative stress insults and longevity.	Drosophila	31845	GO:0006979 -- response to oxidative stress
Sonic hedgehog	8684485	Of the hh homologues, shh has been found to have the most critical roles in development, acting as a morphogen involved in patterning many systems, including the limb [1] and midline structures in the brain [2], spinal cord, the thalamus by the zona limitans intrathalamica and the teeth .	Zebrafish	42737	GO:0035108 -- limb morphogenesis
	12606280				GO:0048854 -- brain morphogenesis

It is critical to emphasize that the Gene Wiki is not a replacement for manual curation efforts. Even when the Gene Wiki provides a detailed and specific context for citing an article, curators will still need to fully review the article to ensure an accurate and precise annotation, as well as to harvest any additional annotations that may be indicated. Moreover, there are many examples where a curator's experience will be needed to resolve more difficult situations. For example in Table 1, the gene "Sonic hedgehog" illustrates a classic example of a nomenclature problem. While most other organisms have only a single *Shh* gene, zebrafish have two *shh* genes, *shha* and *shhb*. The Gene Wiki discusses the paper from the context of the single human *Shh* gene. In this case, a curator would need to examine the source publication to determine which zebrafish gene should be annotated with the indicated function.

Converting unstructured Gene Wiki content into structured gene annotations is essential to ensure maximum utility and usability of the contributed content. Structured data enable efficient and precise queries and bioinformatic analyses (for example, [6]). The PubMed mining effort described here is the first step toward structuring unstructured Gene Wiki content. In parallel, we are pursuing several strategies to directly annotate biological objects (genes, chemicals, diseases), as well as the semantic relationships

between those objects, using established vocabularies and ontologies.

While the Gene Wiki is not a panacea for the scientific community's goal of comprehensive gene annotation, we believe that it will play an important complementary role to existing manual curation efforts.

## 5. ACKNOWLEDGMENTS

We acknowledge members of the Molecular and Cellular Biology WikiProject for coordinating the enthusiastic editing of Gene Wiki pages among both new and experienced Wikipedia users. This work was supported by grant number GM083924 (to AIS) from the National Institute of General Medical Sciences (NIGMS) at the National Institutes of Health, and by the Novartis Research Foundation. DGH is supported by grant number P41 HG002659 from the National Human Genome Research Institute of the National Institutes of Health

## 6. REFERENCES

- [1] C. J. Bult, J. T. Eppig, J. A. Kadin, J. E. Richardson, and J. A. Blake, "The Mouse Genome Database (MGD): mouse biology and model systems," *Nucleic Acids Res*, vol. 36, pp. D724-8, 2008.

- [2] J. Sprague, L. Bayraktaroglu, D. Clements, T. Conlin, D. Fashena, K. Frazer, M. Haendel, D. G. Howe, P. Mani, S. Ramachandran, K. Schaper, E. Segerdell, P. Song, B. Sprunger, S. Taylor, C. E. Van Slyke, and M. Westerfield, "The Zebrafish Information Network: the zebrafish model organism database," *Nucleic Acids Res*, vol. 34, pp. D581-5, 2006.
- [3] S. Tweedie, M. Ashburner, K. Falls, P. Leyland, P. McQuilton, S. Marygold, G. Millburn, D. Osumi-Sutherland, A. Schroeder, R. Seal, and H. Zhang, "FlyBase: enhancing Drosophila Gene Ontology annotations," *Nucleic Acids Res*, vol. 37, pp. D555-9, 2009.
- [4] S. N. Twigger, M. Shimoyama, S. Bromberg, A. E. Kwitek, and H. J. Jacob, "The Rat Genome Database, update 2007--easing the path from disease to data and back again," *Nucleic Acids Res*, vol. 35, pp. D658-62, 2007.
- [5] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet*, vol. 25, pp. 25-9, 2000.
- [6] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proc Natl Acad Sci U S A*, vol. 102, pp. 15545-50, 2005.
- [7] D. Howe, M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, D. P. Hill, R. Kania, M. Schaeffer, S. St Pierre, S. Twigger, O. White, and S. Yon Rhee, "Big data: The future of biocuration," *Nature*, vol. 455, pp. 47-50, 2008.
- [8] J. W. Huss, 3rd, C. Orozco, J. Goodale, C. Wu, S. Batalov, T. J. Vickers, F. Valafar, and A. I. Su, "A gene wiki for community annotation of gene function," *PLoS Biol*, vol. 6, pp. e175, 2008.
- [9] J. W. Huss, 3rd, P. Lindenbaum, M. Martone, D. Roberts, A. Pizarro, F. Valafar, J. B. Hogenesch, and A. I. Su, "The Gene Wiki: community intelligence applied to human gene annotation," *Nucleic Acids Res*, 2009.
- [10] R. Hoffmann, "A wiki for the life sciences where authorship matters," *Nat Genet*, vol. 40, pp. 1047-51, 2008.
- [11] B. Mons, M. Ashburner, C. Chichester, E. van Mulligen, M. Weeber, J. den Dunnen, G. J. van Ommen, M. Musen, M. Cockerill, H. Hermjakob, A. Mons, A. Packer, R. Pacheco, S. Lewis, A. Berkeley, W. Melton, N. Barris, J. Wales, G. Meijssen, E. Moeller, P. J. Roes, K. Borner, and A. Bairoch, "Calling on a million minds for community annotation in WikiProteins," *Genome Biol*, vol. 9, pp. R89, 2008.
- [12] Wikipedia:Verifiability. <http://en.wikipedia.org/wiki/Wikipedia:Verifiability>.
- [13] Wikipedia template filling. <http://diberri.dyndns.org/cgi-bin/templatefiller/>.
- [14] A. I. Su and J. B. Hogenesch, "Power-law-like distributions in biomedical publications and research funding," *Genome Biol*, vol. 8, pp. 404, 2007.
- [15] Current Annotations. <http://www.geneontology.org/GO.current.annotations.shtml>.