A Gene Wiki for Community Annotation of Gene Function

Jon W. Huss, III[©], Camilo Orozco[©], James Goodale[©], Chunlei Wu, Serge Batalov, Tim J. Vickers, Faramarz Valafar, Andrew I. Su[*]

Gene portals (e.g., Entrez Gene [1] and Ensembl [2]) and model organism databases (e.g., Mouse Genome Database [3], Rat Genome Database [4], FlyBase [5]) are popular and useful tools for researching gene annotation and enforcing data standards. These databases provide a large volume and diversity of information on each gene, including protein and transcript sequences, genome location, genomic structure, aliases, links to literature, and gene function. These sites are considered to be the definitive sources for these types of gene annotation. However, by their very nature as authoritative annotation sources, the data displayed on these sites must be subjected to a high degree of oversight by expert curators. In short, the data model used by gene portals and model organism databases focuses on large contributions from a relatively small number of contributors.

In contrast, the online encyclopedia Wikipedia uses a different model for collaboratively synthesizing knowledge, commonly referred to as the "Long Tail" [6]. Originally coined in reference to the power law relationship observed in Internet commerce, the Long Tail is typified by Wikipedia's relatively open data model that targets small contributions from a large population of contributors. Articles in Wikipedia can be freely edited by all users, including anonymous editors, and any registered user can create new articles. Established in 2001, the English Wikipedia currently contains over two million articles edited by over six million user accounts. A recent study found that the number of contributions from new editors (less than 100 total edits) in total equals the number of contributions from the most established editors (greater than 10,000 edits) [7], illustrating the collective importance of the Long Tail. Equally importantly, previous studies have shown that Wikipedia content on scientific topics rivals the online Encyclopedia Britannica in accuracy [8].

Despite the widespread use of Wikipedia for general interest topics, its use for scholarly subjects has been uneven. The potential power of applying the Long Tail model to gene annotation has been previously noted [9–11]. A loose organization of Wikipedia editors has spearheaded the creation and expansion of several thousand articles related to molecular and cellular biology (the "MCB Wikiproject"), including many gene-specific pages. These articles vary widely in quality, format, and completeness, ranging from relatively complete encyclopedic entries (e.g., "enzyme," "oxidative phosphorylation," and "RNA interference") to very short collections of information called "stubs" (e.g., "amphinase" and "glomus cell"). As an example of the collaborative writing process, the article on RNAi has been edited 708 times by 232 unique editors since its initial creation in October 2002. On the subject of human genes, generally only the most well-characterized of genes and proteins have highly developed entries (e.g., "HSP90" and "NF-κB").

In principle, a comprehensive gene wiki could have naturally evolved out of the existing Wikipedia framework, and as described above, the beginnings of this process were already underway. However, we hypothesized that growth could be greatly accelerated by systematic creation of gene page stubs, each of which would contain a basal level of gene annotation harvested from authoritative sources. Here we describe an effort to automatically create such a foundation for a comprehensive gene wiki. Moreover, we demonstrate that this effort has begun the positive-feedback loop between readers, contributors, and page utility, which will promote its long-term success.

Laying the Foundation for a Comprehensive Gene Wiki

We first designed a gene stub with information based primarily on data from Entrez Gene (Figure 1). Each gene stub consists of a sidebar detailing the symbols and aliases, external identifiers, gene function (as represented in Gene Ontology), and genomic location. Although gene stubs are primarily focused on human genes, links to their mouse orthologs are also provided. When available, links to the Protein Data Bank are displayed under a thumbnail ribbon diagram, and gene expression patterns across diverse human tissues are shown as thumbnail bar charts [12]. Links to the primary databases are included when available. In addition, the central area of the gene stub shows a gene summary and a list of relevant references in the literature, both of which were provided by Entrez Gene.

After having finalized the content and format of the gene stubs with input from many Wikipedia users,

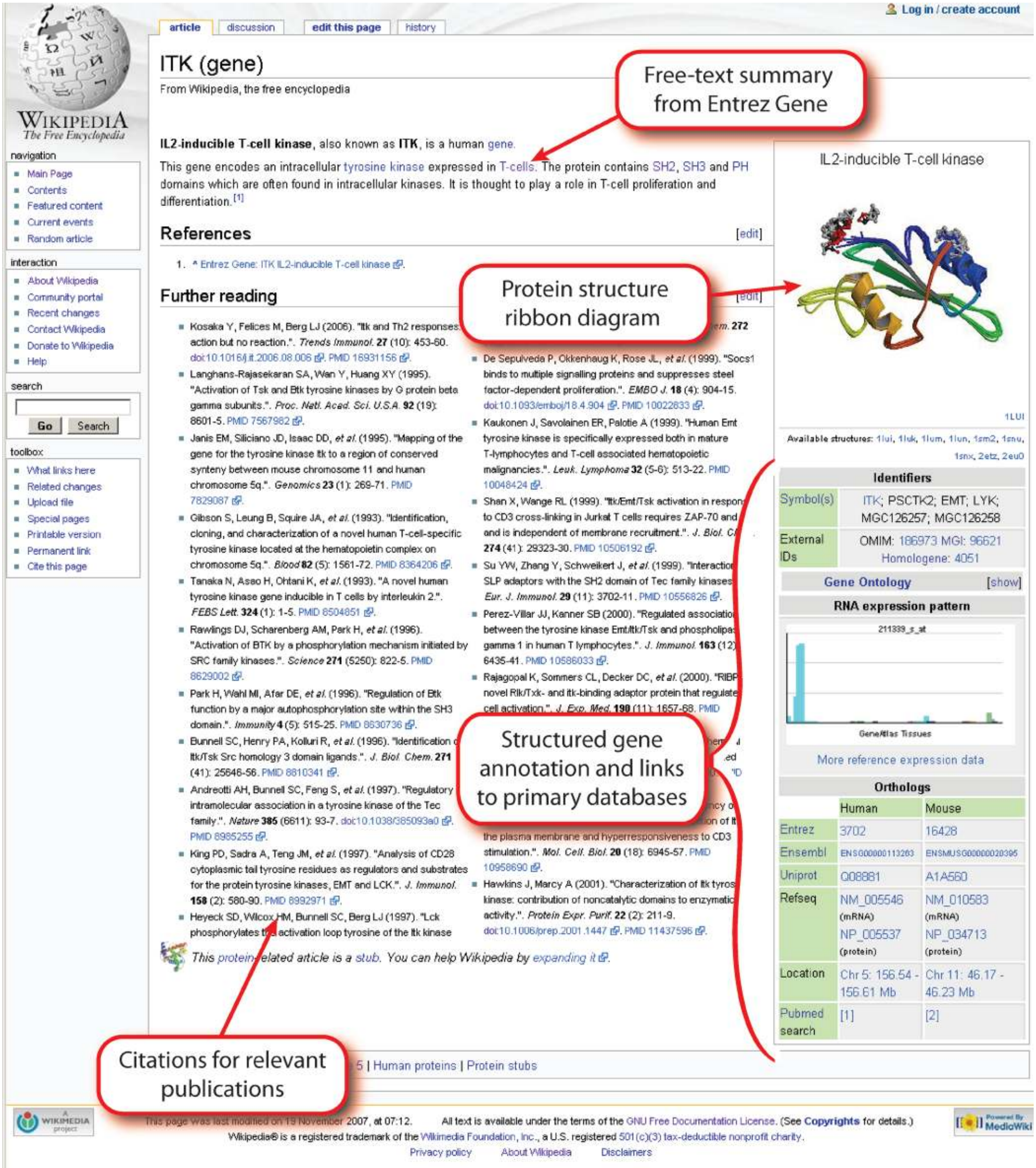
The Community Page is a forum for organizations and societies to highlight their efforts to enhance the dissemination and value of scientific knowledge.

Citation: Huss JW III, Orozco C, Goodale J, Wu C, Batalov S, et al. (2008) A gene wiki for community annotation of gene function. PLoS Biol 6(7): e175. doi:10.1371/journal.pbio.0060175

Jon W. Huss III and Faramarz Valafar are with San Diego State University, Bioinformatics and Medical Informatics Graduate Program, San Diego, California, United States of America. Camilo Orozco, James Goodale, Chunlei Wu, Serge Batalov, and Andrew I. Su are with the Genomics Institute of the Novartis Research Foundation, San Diego, California, United States of America. Tim J. Vickers is with Washington University School of Medicine, Department of Molecular Microbiology, St. Louis, Missouri, United States of America.

[*] To whom correspondence should be addressed. E-mail: asu@gnf.org

[©] These authors contributed equally to this work.



navigation

- Main Page
- Contents
- Featured content
- Current events
- Random article

interaction

- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia
- Donate to Wikipedia
- Help

search

Go | Search

toolbox

- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link
- Cite this page

ITK (gene)

From Wikipedia, the free encyclopedia

IL2-inducible T-cell kinase, also known as **ITK**, is a human gene.

This gene encodes an intracellular tyrosine kinase expressed in T-cells. The protein contains SH2, SH3 and PH domains which are often found in intracellular kinases. It is thought to play a role in T-cell proliferation and differentiation.[1]

References [edit]

1. ^ Entrez Gene: ITK IL2-inducible T-cell kinase.

Further reading [edit]

- Kosaka Y, Felices M, Berg LJ (2006). "Itk and Th2 responses: action but no reaction.". *Trends Immunol.* **27** (10): 453-60. doi:10.1016/j.it.2006.08.006. PMID 16931156.
- Langhans-Rajasekaran SA, Wan Y, Huang XY (1995). "Activation of Tsk and Btk tyrosine kinases by G protein beta gamma subunits.". *Proc. Natl. Acad. Sci. U.S.A.* **92** (19): 8601-5. PMID 7567982.
- Janis EM, Siliciano JD, Isaac DD, *et al.* (1995). "Mapping of the gene for the tyrosine kinase Itk to a region of conserved synteny between mouse chromosome 11 and human chromosome 5q.". *Genomics* **23** (1): 269-71. PMID 7829087.
- Gibson S, Leung B, Squire JA, *et al.* (1993). "Identification, cloning, and characterization of a novel human T-cell-specific tyrosine kinase located at the hematopoietin complex on chromosome 5q.". *Blood* **82** (5): 1561-72. PMID 8364206.
- Tanaka N, Asao H, Ohtani K, *et al.* (1993). "A novel human tyrosine kinase gene inducible in T cells by interleukin 2.". *FEBS Lett.* **324** (1): 1-5. PMID 8504851.
- Rawlings DJ, Scharenberg AM, Park H, *et al.* (1996). "Activation of BTK by a phosphorylation mechanism initiated by SRC family kinases.". *Science* **271** (5250): 822-5. PMID 8629002.
- Park H, Wahl MI, Afar DE, *et al.* (1996). "Regulation of Btk function by a major autophosphorylation site within the SH3 domain.". *Immunity* **4** (5): 515-25. PMID 8630736.
- Bunnell SC, Henry PA, Kolluri R, *et al.* (1996). "Identification of Itk/Tsk Src homology 3 domain ligands.". *J. Biol. Chem.* **271** (41): 25646-56. PMID 8810341.
- Andreotti AH, Bunnell SC, Feng S, *et al.* (1997). "Regulatory intramolecular association in a tyrosine kinase of the Tec family.". *Nature* **385** (6611): 93-7. doi:10.1038/385093a0. PMID 8985255.
- King PD, Sadra A, Teng JM, *et al.* (1997). "Analysis of CD28 cytoplasmic tail tyrosine residues as regulators and substrates for the protein tyrosine kinases, EMT and LCK.". *J. Immunol.* **158** (2): 580-90. PMID 8992971.
- Heyeck SD, Wilcox HM, Bunnell SC, Berg LJ (1997). "Lck phosphorylates the activation loop tyrosine of the Itk kinase
- *…em.* **272**
- De Sepulveda P, Okkenhaug K, Rose JL, *et al.* (1999). "Socs1 binds to multiple signalling proteins and suppresses steel factor-dependent proliferation.". *EMBO J.* **18** (4): 904-15. doi:10.1093/emboj/18.4.904. PMID 10022833.
- Kaukonen J, Savolainen ER, Palotie A (1999). "Human Emt tyrosine kinase is specifically expressed both in mature T-lymphocytes and T-cell associated hematopoietic malignancies.". *Leuk. Lymphoma* **32** (5-6): 513-22. PMID 10048424.
- Shan X, Wange RL (1999). "Itk/Emt/Tsk activation in respon… to CD3 cross-linking in Jurkat T cells requires ZAP-70 and… and is independent of membrane recruitment.". *J. Biol. C…* **274** (41): 29323-30. PMID 10506192.
- Su YW, Zhang Y, Schweikert J, *et al.* (1999). "Interactio… SLP adaptors with the SH2 domain of Tec family kinases… *Eur. J. Immunol.* **29** (11): 3702-11. PMID 10556826.
- Perez-Villar JJ, Kanner SB (2000). "Regulated associatio… between the tyrosine kinase Emt/Itk/Tsk and phospholipa… gamma 1 in human T lymphocytes.". *J. Immunol.* **163** (12)… 6435-41. PMID 10586033.
- Rajagopal K, Sommers CL, Decker DC, *et al.* (2000). "RIBP… novel Rlk/Txk- and itk-binding adaptor protein that regulate… cell activation.". *J. Exp. Med.* **190** (11): 1657-68. PMID
- …the plasma membrane and hyperresponsiveness to CD3 stimulation.". *Mol. Cell. Biol.* **20** (18): 6945-57. PMID 10958690.
- Hawkins J, Marcy A (2001). "Characterization of Itk tyros… kinase: contribution of noncatalytic domains to enzymatic activity.". *Protein Expr. Purif.* **22** (2): 211-9. doi:10.1006/prep.2001.1447. PMID 11437596.

This protein-related article is a stub. You can help Wikipedia by expanding it.

5 | Human proteins | Protein stubs

IL2-inducible T-cell kinase

1LUI

Available structures: 1lui, 1luk, 1lum, 1lun, 1sm2, 1snu, 1snx, 2etz, 2eu0

Identifiers	
Symbol(s)	ITK; PSCTK2; EMT; LYK; MGC126257; MGC126258
External IDs	OMIM: 186973 MGI: 96621 Homologene: 4051

Gene Ontology [show]

RNA expression pattern

More reference expression data

Orthologs

	Human	Mouse
Entrez	3702	16428
Ensembl	ENSG00000113263	ENSMUSG00000020395
Uniprot	Q08881	A1A560
Refseq	NM_005546 (mRNA) NP_005537 (protein)	NM_010583 (mRNA) NP_034713 (protein)
Location	Chr 5: 156.54 - 156.61 Mb	Chr 11: 46.17 - 46.23 Mb
Pubmed search	[1]	[2]

This page was last modified on 19 November 2007, at 07:12. All text is available under the terms of the GNU Free Documentation License. (See Copyrights for details.) Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a U.S. registered 501(c)(3) tax-deductible nonprofit charity.

Privacy policy About Wikipedia Disclaimers

Free-text summary from Entrez Gene

Protein structure ribbon diagram

Structured gene annotation and links to primary databases

Citations for relevant publications

doi:10.1371/journal.pbio.0060175.g001

Figure 1. Example Gene Stub Created at Wikipedia for the ITK Gene

All gene stubs include annotation harvested from public databases, including genome location, protein structures, anatomic gene expression patterns, and gene functions.

we developed a computer program to generate gene stubs in automated fashion. (The program was developed using the Java programming language, and source code was released via the Apache License 2.0 at http://protein-box-bot.googlecode.com.) Gene annotation from the Entrez Gene database was downloaded for each gene, and properly formatted "wikitext" corresponding to the layout described above was created. The program then attempted to check if a gene page already existed. If not, the wikitext was automatically uploaded to Wikipedia, creating a new gene page titled according to the official gene symbol. If an existing gene page was detected, the wikitext was written to a log file for manual inspection and integration. All automatically generated content was added in templates that were made specifically in the context of this project so that subsequent updates of these data can be performed without disrupting manual edits made by other Wikipedia users. Gene pages were created in approximate order of decreasing number of references as indexed by Entrez Gene. As of February 2008, there were 7,500 new gene stubs created automatically and approximately 650 existing pages that were amended with content from our program.

Evaluating Community Participation in Gene Annotation

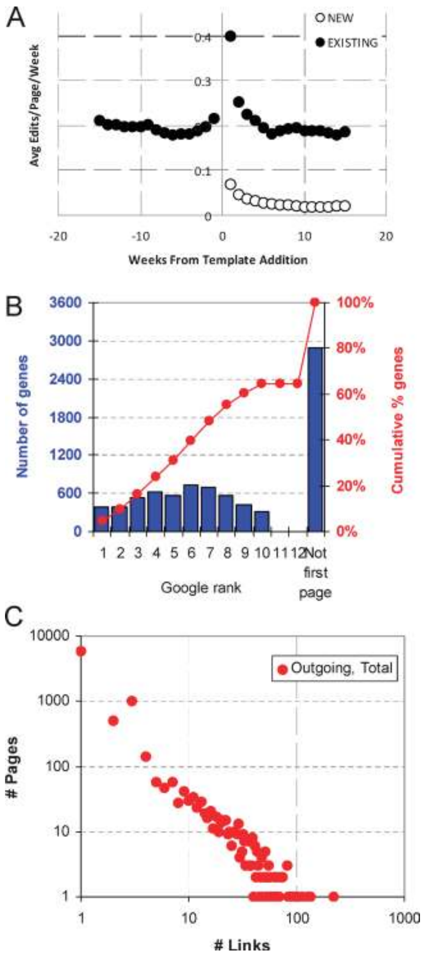
To assess the effect of these gene stubs on Wikipedia, we first surveyed the edit logs for each of these pages that track contributions from the community (Figure 2A). For the 650 gene pages that were previously existing and amended in this effort, we found that the edit rates were roughly equal when comparing activity both before and after our automated efforts. Among the 7,500 new gene stubs created, edit rates were on average 10-fold less than for the pre-existing pages. However, due to the substantial difference in size of these groups, approximately 50% of all edits to gene pages were made on the newly created pages. These results demonstrate that in terms of the absolute number of edits, this effort roughly doubled the amount of mammalian gene annotation activity in Wikipedia. Given the relatively short period of time for which these entries have been available and the fact that this effort has not been previously announced in publication, we expect the rate of editing activity to continue to grow.

As another indication of the current and potential impact of these gene pages, we examined the ranking of these gene pages by the search engine Google. When searching by gene symbol, over 60% of gene pages at Wikipedia are listed in the first page of hits (Figure 2B). (Google ranks separated by pre-existing versus new pages are shown in Figure S1.) Previous research has shown that ranking in search results strongly influences users' navigation choices, and web users rarely navigate to lower-ranked search results [13]. Since any effort utilizing the Long Tail is dependent on a critical mass of users, the high visibility of these gene pages strongly suggests that the Wikipedia gene wiki will continue to grow, in the number of both readers and editors.

Finally, because Wikipedia entries allow for simple hyperlinking from gene pages to related concepts (as well as between two related gene pages), we examined the characteristics of the network topology surrounding these gene pages. Edges in the network denote hyperlinks, which can be further characterized by their directionality. When examining links that were directly referenced in the text (due to technical limitations we focused solely on outgoing links), we found that the network degree roughly followed a power law (with a slight excess of pages with low degree), indicative of a scale-free network (Figure 2C). (Network degree separated by pre-existing versus new pages is shown in Figure S2.) This scale-free property has been observed in many "real-world" networks, including the organization of the Internet [14] and of social networks [15]. In biology, scale-free networks have been observed in networks derived from protein interactions [16] and metabolic pathways [17,18].

The existence of scale-free properties in the Wikipedia gene network has at least two implications. First, scale-free networks are known to have a small network diameter (represented by the shortest path between any two concepts), suggesting that the Wikipedia gene network facilitates the efficient "browsing" of related topics in the local gene neighborhood. These related articles can link to other genes (linked by protein family or biological pathway, for example), as well as basic biological processes, disease states, specific researchers, and experimental methodologies. Second, the current scale-free property in gene wiki connectivity has potential implications for the gene wiki's ability to grow. We expect that gene wiki users will add links that reflect relationships between genes, including physical, functional, and regulatory interactions. As noted above, many of these types of biological networks are also thought to be scale-free. Because wiki pages have been shown to easily handle a large number of links (usually as inline hyperlinks) between related concepts, we expect the gene wiki to naturally accommodate the scale of relevant biological networks. In support of this hypothesis, recent analysis of the power law relationship in Wikipedia links [19] reveals a network having roughly the same degree distribution as seen in protein interaction networks [16].

As alluded to previously, the success of this gene wiki effort relies on a



doi:10.1371/journal.pbio.0060175.g002

Figure 2. Analysis of Gene Wiki Impact

(A) For the ~650 existing gene pages at Wikipedia that were amended with structured annotation, the cumulative rate of edits before and after addition was approximately the same. The rate of edits to the ~7,500 new gene pages was approximately 10-fold less than for the existing pages, though in total these edits accounted for approximately half of all edits to gene pages at Wikipedia. (B) When searching by gene symbol, Wikipedia gene pages are highly ranked by the Google search engine. Approximately 66% of all gene pages are found on the first page of search results. (C) The gene wiki can be analyzed in a network representation, where hyperlinks to other Wikipedia pages constitute network edges. Analysis of the network topology shows that connectivity follows a power law, indicative of a scale-free network.

positive-feedback loop between page utility, the number of readers, and the number of editors. It is commonly recognized that new wiki editors are more likely to edit an existing page rather than to create a new one. Therefore, this stub-creation effort represents the critical first step in that positive feedback loop—the creation of useful initial wiki pages. As suggested in Figure 2B, these stubs will attract some reasonable population of readers, a small fraction of whom will additionally edit and improve the page. It is then hoped and expected that this positive-feedback loop will become self-sustaining.

Leveraging the "Long Tail" for a Successful Gene Wiki

By leveraging the Long Tail of scientists in the annotation process, we believe that the gene wiki described here will harness a powerful and untapped source of gene annotation that is complementary to existing resources. Moreover, the type of data displayed by this gene wiki is also complementary in nature to other databases. Whereas existing gene portals are primarily focused on structured annotation data (ontologies, genome coordinates, etc.), the gene wiki is primarily focused on unstructured content in the form of free text, images, and diagrams that are more typical of review articles. Finally, the boundless nature of Wikipedia also allows for a single source to be appropriate to multiple levels of readers, from a lay audience to students to working professionals and academics.

However, the Long Tail and Wikipedia also have potential liabilities. Most notably, the completely open and anonymous nature of Wikipedia raises potential concerns about the completeness and accuracy of the articles, and in turn, the potential to recruit the broader scientific community to participate. We believe that the success of a gene wiki relies on the same foundation underlying the success of the rest of Wikipedia. Specifically, a simple editing syntax and a detailed version history for all pages enable effective collaboration and quick correction of incorrect or misleading content. A sizable population of readers then serves simultaneously as consumers, reviewers, and editors of content. Wikipedia has also managed to maintain an effective culture of collaboration by adherence to and promotion of its five core pillars (included in which is a code of conduct and maintenance of a neutral point of view) [20]. Finally, editors tend to add pages of interest to their "watchlist," which highlights further changes to the article by other editors. Regardless of the specific reasoning, it is difficult to contest Wikipedia's success thus far in the breadth and accuracy of articles [8].

In addition, viable alternatives to Wikipedia exist. For example, a parallel gene wiki effort at Citizendium (http://www.citizendium.org) is also being considered, which, among other differences, requires authors to use real names and provides an explicit role for expert editors [21]. Despite potential concerns, this gene wiki project was initiated at Wikipedia to take advantage of the existing critical mass of articles that can be linked to and from gene pages, and for the critical mass of editors who are skilled in other aspects of online collaboration (copy editing, dispute resolution, governance, etc.) Although several other gene wikis already exist, none currently has access to a large user base and favorable search engine rankings of Wikipedia.

Importantly, this gene wiki effort is not meant to be a substitute for existing resources. Gene portals and model organism databases will continue to serve as authoritative references with a specific role for data curation and enforcement of data standards. Moreover, the structured and typed data in gene portals is amenable to incorporation into pipelines and systematic analyses in a way the information in a gene wiki cannot [22]. Most importantly, because articles are dynamic and not subject to rigorous peer review, the gene wiki is not intended to be a reference that is cited in a traditional peer-reviewed article or used exclusively as a source of gene annotation. Nevertheless, we believe that this gene wiki will be a valuable launch pad for collaboratively summarizing knowledge, and we expect that scientists will synergistically use the gene wiki with traditional gene portals.

Despite its infancy, the gene wiki effort has already had a substantial and growing impact on the Wikipedia community. We believe that this effort will encourage further contributions from scientists around the world and become a robust, cross-referenced tool for students, educators, and researchers everywhere. With the entire community's input, we envision this gene wiki evolving into a collection of collaboratively created, continually updated, community-reviewed review articles for every gene in the human genome. ■

Supporting Information

Figure S1. Analysis of Google Rank as Stratified by New Versus Pre-Existing Wiki Pages

These data are presented as described in Figure 2B.

Found at doi:10.1371/journal.pbio.0060175.sg001 (40 KB PPT).

Figure S2. Analysis of the Outgoing Network Degree as Stratified by New Versus Pre-Existing Wiki Pages

These data are presented as described in Figure 2C.

Found at doi:10.1371/journal.pbio.0060175.sg002 (38 KB PPT).

Acknowledgments

We thank John Hogenesch and Michael Cooke for helpful comments, and the MCB community at Wikipedia (particularly Christopher Day, Diane Schroeder, Emanuele Plebani, and Konrad Koehler) for feedback on gene stub design and enthusiastic editing of gene pages.

Funding. This work was supported by the Novartis Research Foundation and the Alfred P. Sloan Foundation (FV).

References

1. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2008) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 36: D13-21.
2. Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, et al. (2008) Ensembl 2008. Nucleic Acids Res 36: D707-714.
3. Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA (2008) The Mouse Genome Database (MGD): mouse biology and model systems. Nucleic Acids Res 36: D724-728.
4. Twigger SN, Shimoyama M, Bromberg S, Kwitek AE, Jacob HJ (2007) The Rat Genome Database, update 2007–easing the path from disease to data and back again. Nucleic Acids Res 35: D658-662.
5. Wilson RJ, Goodman JL, Strelets VB (2008) FlyBase: integration and improvements to query tools. Nucleic Acids Res 36: D588-593.
6. Anderson C (2004) The Long Tail. Wired 12(10). Available: http://www.wired.com/wired/archive/12.10/tail.html. Accessed 12 June 2008.
7. Kittur A, Chi EH, Pendleton BA, Suh B, Mytkowicz T (2007) Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. 25th Annual ACM Confernce on Human Factors in Computing Systems (CHI 2007). 28 April–3 May 2007; San Jose, California, United States.
8. Giles J (2005) Internet encyclopaedias go head to head. Nature 438: 900-901.
9. Giles J (2007) Key biology databases go wiki. Nature 445: 691.

10. Wang K (2006) Gene-function wiki would let biologists pool worldwide resources. Nature 439: 534.
11. Salzberg SL (2007) Genome re-annotation: a wiki solution? Genome Biol 8: 102.
12. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 101: 6062-6067.
13. Pan B, Hembrooke H, Joachims T, Lorigo L, Gay G, et al. (2007) In Google we trust: Users' decisions on rank, position, and relevance. J Comp-Mediated Comm 12(3): article 3.

Note Added in Proof

Recently, Mons et al. [23] described another effort that uses community intelligence, called WikiProteins.

14. Barabasi A, Albert R, Jeong H (2000) Scale-free characteristics of random networks: the topology of the World-Wide Web. Phys A 281: 69-77.
15. Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286: 509-512.
16. Yook SH, Oltvai ZN, Barabasi AL (2004) Functional and topological characterization of protein interaction networks. Proteomics 4: 928-942.
17. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL (2000) The large-scale organization of metabolic networks. Nature 407: 651-654.
18. Wagner A, Fell DA (2001) The small world inside large metabolic networks. Proc Biol Sci 268: 1803-1810.
19. Kamps J, Koolen M (2008) The Importance of link evidence in Wikipedia. ECIR. Berlin: Springer Berlin / Heidelberg. pp. 270-282.
20. Wikipedia: Five pillars. Available: http://en.wikipedia.org/wiki/Wikipedia:Five_pillars. Accessed 12 June 2008.
21. Giles J (2006) Wikipedia rival calls in the experts. Nature 443: 493.
22. Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L (2001) The distributed annotation system. BMC Bioinformatics 2: 7.
23. Mons B, Ashburner M, Chichester C, van Mulligan E, Weeber M, et al. (2008) Calling on a million minds for community annotation in WikiProteins. Genome Biol 9: R89. doi:10.1186/gb-2008-9-5-r89