Published online 15 September 2009 *Nucleic Acids Research, 2010, Vol. 38, Database issue* ***D633–D639***
doi:10.1093/nar/gkp760

The Gene Wiki: community intelligence applied to human gene annotation

Jon W. Huss III[1], Pierre Lindenbaum[2], Michael Martone[3], Donabel Roberts[4], Angel Pizarro[5], Faramarz Valafar[4], John B. Hogenesch[5] and Andrew I. Su[1,*]

[1]Genomics Institute of the Novartis Research Foundation, San Diego, CA 92121, USA, [2]Department of Bioinformatics, CEPH/Fondation Jean-Dausset, Paris, France, [3]Rush University Medical College, Chicago, IL 60612, [4]San Diego State University, Bioinformatics and Medical Informatics Graduate Program, San Diego, CA 92182 and [5]Department of Pharmacology, Institute for Translational Medicine and Therapeutics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA

Received August 14, 2009; Revised August 26, 2009; Accepted August 29, 2009

ABSTRACT

Annotating the function of all human genes is a critical, yet formidable, challenge. Current gene annotation efforts focus on centralized curation resources, but it is increasingly clear that this approach does not scale with the rapid growth of the biomedical literature. The Gene Wiki utilizes an alternative and complementary model based on the principle of community intelligence. Directly integrated within the online encyclopedia, Wikipedia, the goal of this effort is to build a gene-specific review article for every gene in the human genome, where each article is collaboratively written, continuously updated and community reviewed. Previously, we described the creation of Gene Wiki 'stubs' for approximately 9000 human genes. Here, we describe ongoing systematic improvements to these articles to increase their utility. Moreover, we retrospectively examine the community usage and improvement of the Gene Wiki, providing evidence of a critical mass of users and editors. Gene Wiki articles are freely accessible within the Wikipedia web site, and additional links and information are available at http://en.wikipedia.org/wiki/Portal:Gene_Wiki.

INTRODUCTION

The sequencing and analysis of the human genome have largely elucidated the 'parts list' of the cellular machinery in the form of approximately 25 000 genes. However, the comprehensive annotation of gene function remains a formidable challenge. The scale of the task ahead is illustrated by two simple analyses of links between PubMed and the Entrez Gene database.

The first analysis showed that while there are several well-studied genes that have thousands of indexed citations in the literature, that degree of functional annotation falls off steeply (1). Almost 80% of Entrez Gene entries had five or fewer linked references in PubMed; almost 50% had zero linked references (Figure 1A). This pattern was even more pronounced when examining links to the Gene Ontology (also shown in Figure 1A). Clearly, there remains much work to be done to functionally annotate the human genome and to comprehensively catalog these findings in gene annotation databases.

A second analysis examined the rate at which PubMed entries were being linked to Entrez Gene (Figure 1B). Between 1970 and 2008, the number of publications added to PubMed grew at an annualized rate of ~3.4%. On examining that same time period, it was found that the number of articles that are currently linked to at least one Entrez Gene entry 'grew' by ~18% per year, but still <4% of all PubMed entries in recent years (and <1.5% overall) are linked to Entrez Gene. Assuming that >4% of PubMed-indexed articles have relevance to human gene function, this finding suggests that the rate limiting step is not generating the data, but capturing the derived knowledge in gene annotation databases.

Currently, the process of annotating gene function typically entails large-scale efforts by the model organism community (2–4) and genome annotation centers (5). Formal annotation of gene function often utilizes controlled vocabularies like Gene Ontology (6). While the annotation process can be aided by the use of computational tools, ultimately the assignment of gene function is a manual process requiring the attention of one or more domain experts (7). This centralized model

*To whom correspondence should be addressed. Tel: 858-812-1500; Fax: 858-812-1570; Email: asu@gnf.org

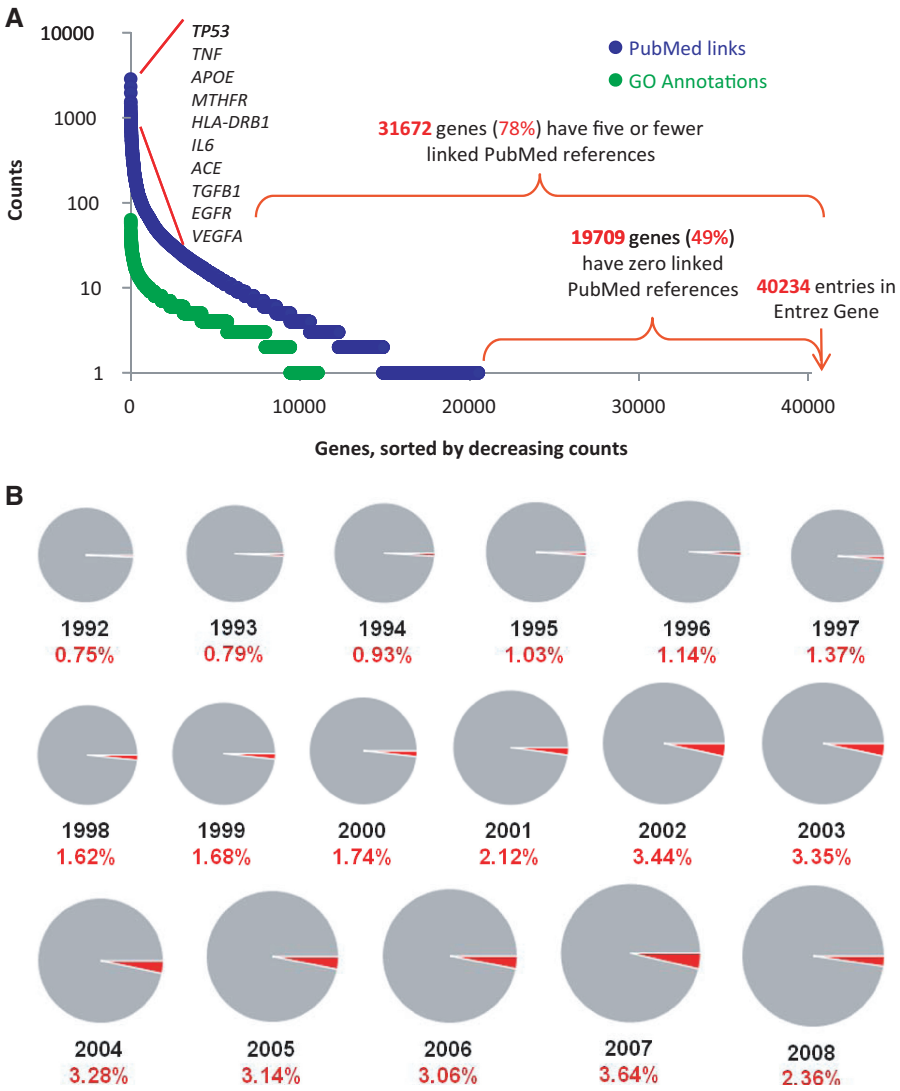

Figure 1. Analysis of links between the Entrez Gene and PubMed databases. (**A**) Examining the degree of gene annotation from the perspective of Entrez Gene, we found that while a few genes are very well annotated with links to PubMed references, the vast majority of genes have few or no linked references. (**B**) Examining links from the perspective of PubMed, we found that only a small fraction of published articles are linked to human genes. Taken together, these findings suggest that the traditional model of centralized curation is not scaling well with the rate of scientific research, and that complementary approaches based on community intelligence may be worth exploring.

has been very successful in its goal to systematically advance gene annotation, creating essential tools and ontologies in the process.

However, this model alone may not be sufficient to efficiently and systematically annotate gene function. Many leading voices in the gene annotation and model organism communities recently wrote a feature article in *Nature* describing the current state and future of biocuration (8). They noted the immense challenge to the curator community (typically numbering in tens to hundreds of people) to keep pace with the biomedical literature (currently 18 million articles in PubMed, roughly 750 000 new articles per year). Specifically, these curation experts suggest that merely preserving the existing models of gene annotation will lead to an increasing lag between curated data and biological knowledge, and that 'sooner or later, the research community will need to be involved in the annotation effort to scale up to the rate of data generation' (8).

Thus, although leaders in the curation community have successfully set up a robust pipeline and infrastructure, and although the individuals in the curation community are clearly skilled in the annotation process, the amount of resources devoted to this important task may be simply insufficient relative to the volume of biomedical data being generated.

Recently, several efforts have been published, which attempt to harness the principle of 'community intelligence' (9–15). In particular, we introduced the Gene Wiki (11), an effort to systematically annotate articles in the online encyclopedia, Wikipedia, for approximately 9000 human genes. Articles were created or amended with content mined from structured gene annotation databases, including Entrez Gene, Ensembl, UniProt and

the Protein Data Bank (PDB). Although the emphasis of the Gene Wiki is on describing human gene function, data from model organisms is often contributed as appropriate.

Here, we present an update describing the recent systematic improvements to the Gene Wiki. Moreover, we report on a retrospective analysis of Gene Wiki usage and editing. Finally, we offer some concluding remarks on general progress and challenges facing efforts to collaboratively engage the entire community of scientists.

DATABASE CONTENT

As described earlier (11), the initial Gene Wiki effort focused on creating or amending gene pages to include a free-text summary, an 'infobox' with links to public databases, Gene Ontology annotations and when available, protein structure identifiers. These data were primarily harvested from the Entrez Gene database (16). Since that publication, we have introduced several other systematic improvements to the Gene Wiki.

Protein interactions

Our previously described Gene Wiki effort resulted in approximately 9000 Wikipedia pages on human genes. However, with the exception of a few well-developed gene pages that existed prior to our involvement, these pages were accessible only through search engines and not through links from other articles. In the parlance of Wikipedia, these stubs were 'orphans' that were disconnected from the Wikipedia network defined by links between articles.

To better link the pages in the Gene Wiki to other biomedically relevant pages, we systematically created links between gene pages based on known protein–protein interactions in the literature. Interactions were downloaded from the BioGRID database (http://thebiogrid.org). We conservatively filtered for interactions that were supported by two independent techniques or two separate publications. A new section for 'Interactions' was created on each Gene Wiki page with at least one entry, which contained both links to the partner's gene page and inline references to the relevant publications. In total, we added 12 628 links on 3389 gene pages.

Better integration of the Gene Wiki into the larger network of Wikipedia articles greatly improves navigation between related topics. For example, readers can now easily browse from the breast cancer article, to the Gene Wiki page for the commonly mutated *BRCA2* gene, to the page for *EMSY* (*C11ORF30*), the protein product of which has been shown to interact with *BRCA2* and silence its transcriptional activity (17).

PDB galleries

Recognizing the importance of structural biology data, we undertook a focused effort to increase links between the Gene Wiki and the PDB (18). We first uploaded thumbnail images of all PDB structures to the Wikimedia Commons, a repository for freely usable media. Images were downloaded from the PDBe (http://www.ebi.ac.uk/pdbe/), and in total, 66 693 images were uploaded to the Wikimedia Commons. To aid in browsing and searching, PDB images were also categorized according to their assignments in the Structural Classification of Proteins (SCOP) database (19). The set of PDB structures can be browsed at http://commons.wikimedia.org/wiki/SCOP.

The easy availability of thumbnail images for almost all PDB structures will encourage their incorporation into relevant Gene Wiki and Wikipedia articles. To begin this process, we added an image gallery of PDB thumbnails to every Gene Wiki page with solved structures. To maintain balance with the rest of the pages' content, the image galleries were shown in an expandable window at the bottom of each Gene Wiki page. In total, PDB image galleries were added to 2852 Gene Wiki pages with a total of 16 018 links to PDB structures. For example, the PDB gallery for *MDM2* (http://en.wikipedia.org/wiki/Mdm2) shows PDB structures corresponding to the unbound protein, as well as structures in complex with its p53 peptide ligand and two small molecule inhibitors.

On-demand web service

There are currently approximately 9000 pages in the Gene Wiki collection. To satisfy Wikipedia's 'notability' criterion, we initially limited our effort to genes with the most linked references in PubMed (as indexed in Entrez Gene). However, to enable other Wikipedia editors to easily create Gene Wiki pages for other genes of interest, we created a simple web tool to generate the properly formatted 'wikitext' for any arbitrary gene of interest. This tool can either be used to update existing content to the most recent data, or to create a new page where none previously existed.

This Gene Wiki formatting tool has been implemented as a BioGPS plugin, accessible at http://biogps.gnf.org/GeneWikiGenerator. By utilizing BioGPS as the search interface, users can search for their gene or genes of interest using most public identifiers and keywords. Upon clicking on a gene, the web tool returns the wikitext in three distinct text boxes, together with links and instructions on how to create a new Gene Wiki page. To allow programmatic usage of this web service, the output is returned as XML and formatted to HTML using an XML style sheet.

RETROSPECTIVE USAGE ANALYSIS

Previously, we suggested that the long-term success of community intelligence resources is dependent on a positive feedback among page utility, readers and editors (11). In the ideal case, each Gene Wiki page provides some baseline level of useful content, which then attracts a certain number of readers. Some (likely small) percentage of those readers will then become contributors, where their contribution could be something as trivial as fixing a typo or as substantial as summarizing a recent paper. Contributions improve the Gene Wiki page, which then draws more readers, and then a larger core of contributors. In other words, usage is directly

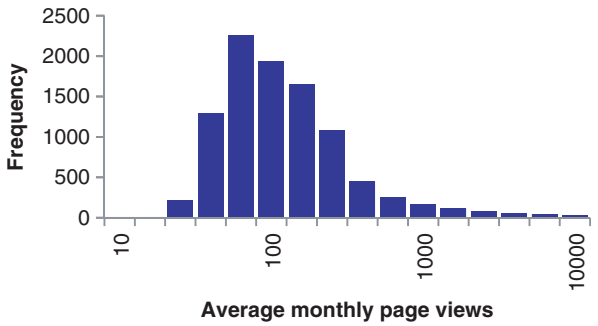

Figure 2. Average monthly page views of pages in the Gene Wiki. Calculated over the first 6 months of 2009, this histogram shows the average number of page views per month over all Gene Wiki articles. Each page receives an average of 304 page views per month (median = 80).

proportional to utility, contribution rate is directly proportional to usage rate and utility is directly proportional to contribution rate.

The first step in this process, creating article 'stubs' that had general utility, was the focus of both our first Gene Wiki effort and the systematic improvements described earlier. In addition, we now have the necessary data on usage and editing patterns to retrospectively assess the other two edges of the positive feedback loop.

Usage was analyzed for the 6-month period between 1 January and 30 June 2009, over the 9678 current Gene Wiki pages. In total, these pages were viewed over 17 million times (3.9 million total page views per month). On a per article basis, the Gene Wiki averaged over 300 pages views per page per month (Figure 2; see also Supplementary Table S1). Closer analysis of these statistics revealed a broad range of usage levels (Table 1). The top-viewed articles are primarily related to genes of general societal interest (e.g. insulin, erythropoietin) and are viewed tens of thousands of times per month, likely dominated by non-scientists. Near the 100th most-popular pages are gene pages that cross many areas of biology (e.g. interleukin 10, c-Met), and these pages are viewed thousands of times per month. Finally, near the 1000th most-viewed pages are genes that are likely of interest to a relatively small population of scientists (e.g. IGSF8, TRPC6), and these pages receive approximately 300 page views per month. We believe that these statistics are indicative of Gene Wiki usage by both scientists and non-scientists.

Supporting the future growth in usage of the Gene Wiki, we also found that >85% of all Gene Wiki pages are found within the top eight Google hits when searching by gene symbol (Figure 3). This figure represents a substantial increase over the ~60% observed shortly after the gene stubs were created (11).

We next examined the third leg of the positive feedback loop by analyzing the editing logs of Gene Wiki pages. During the same period between 1 January and 30 June 2009, there were a total of 6848 edits to 1893 Gene Wiki pages by 1923 unique users or Internet Protocol (IP) addresses. (In addition, automated edits by 'bots' accounted for 11 912 edits.) Edits over this period were

Table 1. Representative Gene Wiki articles ranked and grouped by number of monthly page views between January and June 2009

Rank	Average monthly page views	Gene Wiki page
1	109 116	Insulin
2	64 582	Titin
3	63 673	Human chorionic gonadotropin
4	50 198	Vasopressin
5	49 908	ANKH
6	36 998	CLOCK
7	34 367	Catalase
8	32 382	Erythropoietin
9	30 933	Glucagon
10	29 617	Parathyroid hormone
101	4173	Tau protein
102	4167	Interleukin 10
103	4162	APC (gene)
104	4151	C-Met
105	4106	Factor V
106	4082	Interleukin 8
107	4034	CD44
108	4019	Histamine H1 receptor
109	3980	Kappa Opioid receptor
110	3845	Dihydrofolate reductase
1001	308	CSDA
1002	308	CNTNAP2
1003	308	IGSF8
1004	307	Adenosine A3 receptor
1005	307	RYR1
1006	306	ETV6
1007	306	Small heterodimer partner
1008	306	5-HT1D receptor
1009	306	TRPC6
1010	306	Interleukin-6 receptor

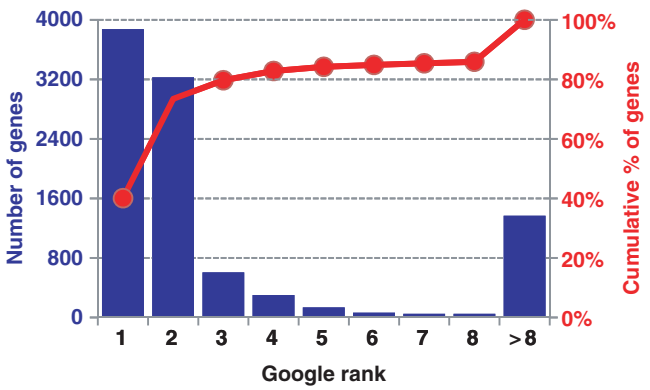

Figure 3. Google rank of Gene Wiki pages. When searching by gene symbol, >85% of Gene Wiki pages are found within the first eight hits by the search engine Google.

quite constant at an average of about 1100 edits per month (SD = 171), 263 edits per week (SD = 69) and 38 edits per day (SD = 21). The cumulative effect of these edits was to increase the size of the text in the Gene Wiki by 2.28 MB (4.1%), approximately the equivalent to the text of 19 research articles in *PLoS Biology*. For individual articles, changes in page size are plotted as a function of current page size and Google rank in Figure 4.

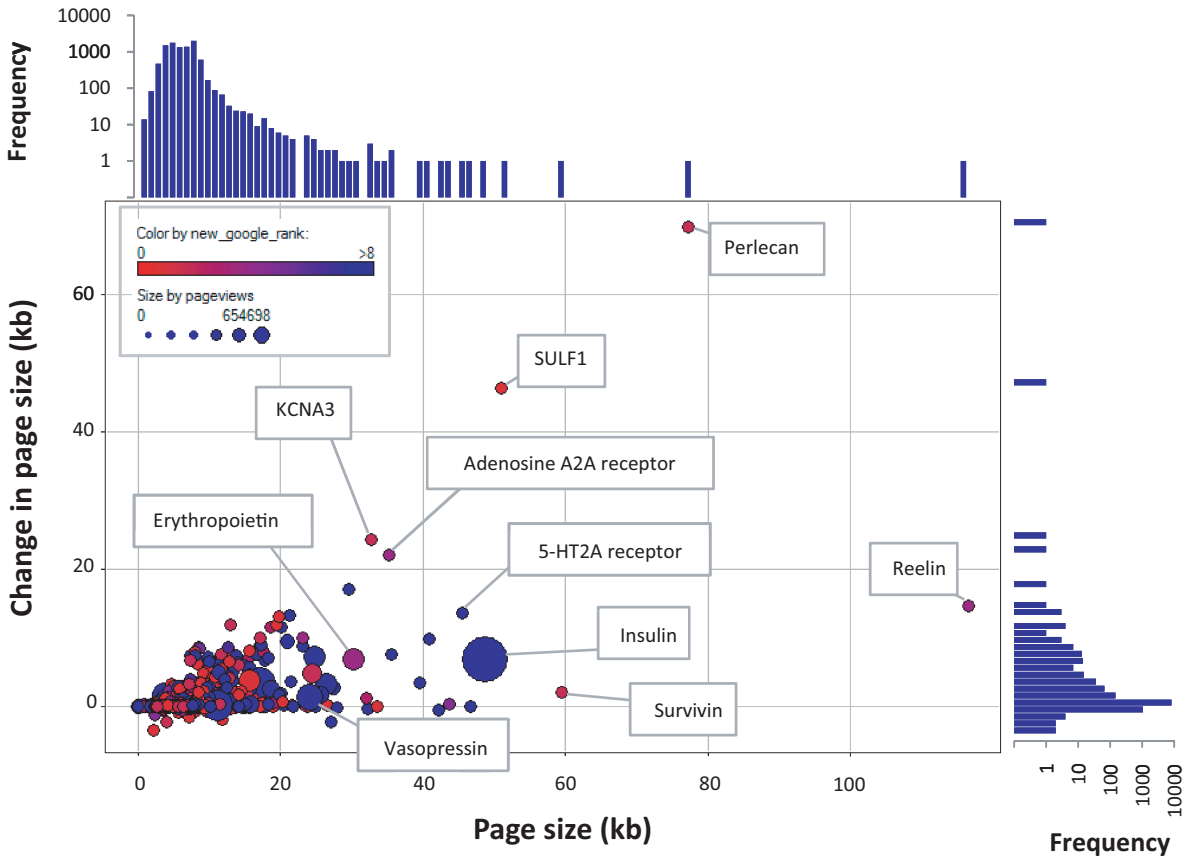

Figure 4. Analysis of page size versus change in page size. The change in size of each Gene Wiki article during the first 6 months of 2009 is plotted as a function of the page size at the end of June 2009. Larger markers indicate more page views, and marker color indicates the Google rank of the page when searching by gene symbol. Labels are shown for a few representative pages.

When examining the usage of statistics, we noticed spikes in the viewing of certain genes, especially those mentioned in the popular press. To explore this observation, we identified the 771 Gene Wiki pages with the most recent variability in monthly page views. Of these, 69 had been searched often enough to have data in Google Trends (http://www.google.com/trends), a service that quantifies how many Google searches have been done for a particular term over time relative to the total number of Google searches. The correlation between Gene Wiki page views and Google Trends over time is readily apparent, with 43% of examined pages having significant correlation ($R > 0.3$; $P < 0.01$).

In many cases, the strong relationship between page views and Google Trends was driven by articles in the popular press (Figure 5). For example, the Wikipedia article for human chorionic gonadotropin (HCG) is one of the most frequently viewed articles in the Gene Wiki, presumably for its common usage in pregnancy tests. In May 2009, the Wikipedia article for this gene experienced a sharp spike in views (and edits) when Manny Ramirez was suspended for using HCG as a performance-enhancing drug. Similarly, catalase is frequently viewed article for its relevance to many areas of biology including aging and cancer. However, following a scientific report linking catalase function to premature gray hair in February 2009 (20), a prominent spike occurred in the viewing and editing of its Gene Wiki entry. Taken in sum, these data show a dynamic relationship between scientific publications, reports on this science in the popular press and usage of the Gene Wiki. These observations also underscore the potential opportunity and effectiveness of using the Gene Wiki for public outreach and scientific education.

DISCUSSION AND FUTURE DIRECTIONS

With the explosion in biological wikis, it is clear that the community intelligence model resonates with the biology and scientific community (9–15). Despite the enthusiasm in the potential of this model, it is also clear that realizing this potential is not trivial. Many of these biological wikis appear to suffer from a lack of participation. Establishing a critical mass of users and useful content appears to be the most common obstacle in these efforts.

By integrating directly with Wikipedia, establishing critical mass has not been an issue for the Gene Wiki. Clearly, Wikipedia already had a critical mass of users and articles, and the Gene Wiki has been able to effectively leverage those resources as demonstrated by the usage and editing metrics presented above. Moreover, within the last year, the American Society for Cell Biology, the Society

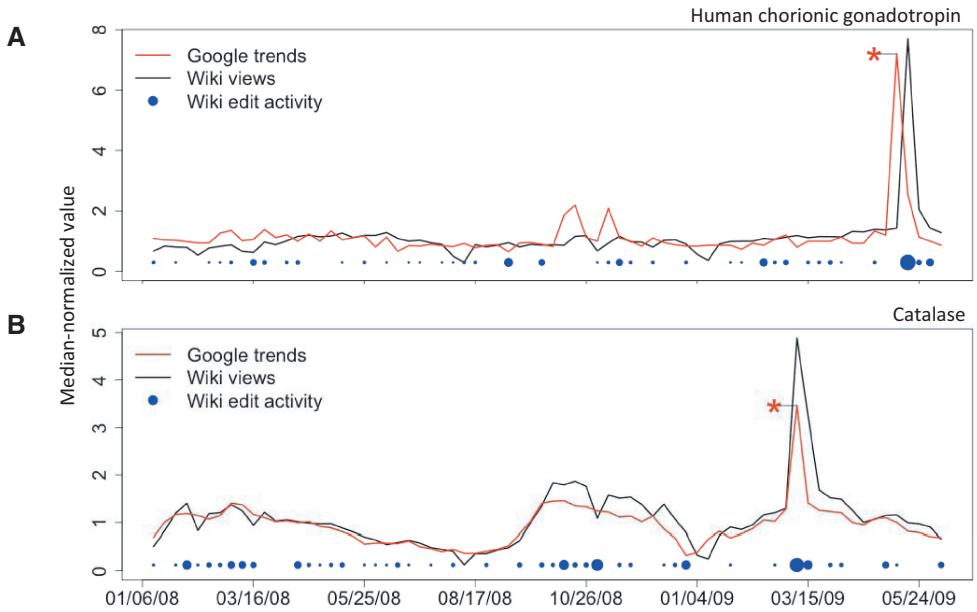

Figure 5. Timelines of Wikipedia views and Google Trends information for HCG (**A**) and catalase (**B**). Blue dots represent editing events, with bigger size corresponding to more editing events. The asterisks mark major events in the popular press for HCG and catalase, where Manny Ramirez was linked to performance-enhancing drugs and catalase was linked to premature gray hair, respectively.

for Neuroscience and the National Institutes of Health have all held workshops or initiated efforts focused on science articles in Wikipedia. However, the Gene Wiki inherited a completely different set of challenges. First and most notably, Wikipedia allows users to remain completely anonymous, which often leads to fears of inaccuracy and bias. And second, Wikipedia is primarily focused on building unstructured articles (free text, images, diagrams, etc.) with relatively little attention to how contributed knowledge can be structured for downstream analyses in the way that Gene Ontology annotations, for example, can be utilized (21).

We intend to focus on these issues in future developments of the Gene Wiki. Although previous studies have suggested that Wikipedia is of comparable accuracy to traditionally curated works (22), other efforts have been developed to explicitly account for trustworthiness of content based on historical editing patterns of each user (23). Moreover, while we still believe that a completely unstructured Gene Wiki article is useful to the community (similarly to a gene-specific review article), we are also investigating methods to integrate community intelligence with data structure using novel technical solutions [e.g. Semantic MediaWiki (24)] and biomedical ontologies (25).

It is essential to emphasize that community intelligence efforts are not a replacement for traditionally curated gene annotation authorities (16,26–28). In contrast, we believe that community intelligence resources are complementary to existing databases and offer a different set of strengths and weaknesses. Certainly, the data generation model is very different, and users of the Gene Wiki need to recognize that the Gene Wiki, like Wikipedia itself, should be treated differently than the primary literature and expert-curated databases.

Ultimately, we believe that a variety of solutions in the area of community intelligence are worth exploring. Future Gene Wiki development will focus on addressing the challenges described above, and we are also very enthusiastic about complementary efforts as they work to build critical mass and encourage participation. Regardless, the usage metrics presented above demonstrate that the Gene Wiki is relevant right now, certainly to the general public and also to a growing number of scientists. We hope that the scientific community embraces this opportunity both to collaboratively annotate gene function and to directly communicate with the public in science education and outreach.

ACCESSIBILITY

Wikipedia is freely available for viewing at http://wikipedia.org, and the Gene Wiki Portal page can be accessed at http://en.wikipedia.org/wiki/Portal:Gene_Wiki. All text is licensed under the Creative Commons Attribution/Share-Alike License 3.0 (Unported).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors acknowledge Konrad F. Koehler for helpful suggestions and enthusiastic editing, Jeff Janes and Julia Turner for technical assistance, as well as the entire community of Wikipedia editors and the Molecular and Cellular Biology WikiProject (http://en.wikipedia.org/wiki/WP:MCB) for contributions and feedback.

FUNDING

Funding for this work and for the open access charge was provided by the Novartis Research Foundation and the National Institutes of Health [Grant Number 1R01GM083924 to A.S.].

Conflict of interest statement. None declared.

REFERENCES

1. Su,A.I. and Hogenesch,J.B. (2007) Power-law-like distributions in biomedical publications and research funding. *Genome Biol.*, **8**, 404.
2. Blake,J.A., Eppig,J.T., Bult,C.J., Kadin,J.A. and Richardson,J.E. (2006) The Mouse Genome Database (MGD): updates and enhancements. *Nucleic Acids Res.*, **34**, D562–D567.
3. Cherry,J.M., Ball,C., Weng,S., Juvik,G., Schmidt,R., Adler,C., Dunn,B., Dwight,S., Riles,L., Mortimer,R.K. *et al.* (1997) Genetic and physical maps of Saccharomyces cerevisiae. *Nature*, **387**, 67–73.
4. Grumbling,G. and Strelets,V. (2006) FlyBase: anatomical data, images and queries. *Nucleic Acids Res.*, **34**, D484–D488.
5. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, **32**, D262–D266.
6. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
7. Hill,D.P., Smith,B., McAndrews-Hill,M.S. and Blake,J.A. (2008) Gene Ontology annotations: what they mean and where they come from. *BMC Bioinformatics*, **9(Suppl. 5)**, S2.
8. Howe,D., Costanzo,M., Fey,P., Gojobori,T., Hannick,L., Hide,W., Hill,D.P., Kania,R., Schaeffer,M., St Pierre,S. *et al.* (2008) Big data: the future of biocuration. *Nature*, **455**, 47–50.
9. Daub,J., Gardner,P.P., Tate,J., Ramsköld,D., Manske,M., Scott,W.G., Weinberg,Z., Griffiths-Jones,S. and Bateman,A. (2008) The RNA WikiProject: community annotation of RNA families. *RNA*, **14**, 2462–2464.
10. Hoffmann,R. (2008) A wiki for the life sciences where authorship matters. *Nat. Genet.*, **40**, 1047–1051.
11. Huss,J.W., Orozco,C., Goodale,J., Wu,C., Batalov,S., Vickers,T.J., Valafar,F. and Su,A.I. (2008) A gene wiki for community annotation of gene function. *PLoS Biol.*, **6**, e175.
12. Mons,B., Ashburner,M., Chichester,C., van Mulligen,E., Weeber,M., den Dunnen,J., van Ommen,G., Musen,M., Cockerill,M., Hermjakob,H. *et al.* (2008) Calling on a million minds for community annotation in WikiProteins. *Genome Biol.*, **9**, R89.
13. Pico,A.R., Kelder,T., van Iersel,M.P., Hanspers,K., Conklin,B.R. and Evelo,C. (2008) WikiPathways: pathway editing for the people. *PLoS Biol.*, **6**, e184.
14. Stokes,T.H., Torrance,J.T., Li,H. and Wang,M.D. (2008) ArrayWiki: an enabling technology for sharing public microarray data repositories and meta-analyses. *BMC Bioinformatics*, **9**, S18 Suppl. 6.
15. Hodis,E., Prilusky,J., Martz,E., Silman,I., Moult,J. and Sussman,J.L. (2008) Proteopedia - a scientific 'wiki' bridging the rift between three-dimensional structure and function of biomacromolecules. *Genome Biol.*, **9**, R121.
16. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D31.
17. Hughes-Davies,L., Huntsman,D., Ruas,M., Fuks,F., Bye,J., Chin,S., Milner,J., Brown,L.A., Hsu,F., Gilks,B. *et al.* (2003) EMSY links the BRCA2 pathway to sporadic breast and ovarian cancer. *Cell*, **115**, 523–535.
18. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
19. Andreeva,A., Howorth,D., Chandonia,J., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res*, **36**, D419–D425.
20. Wood,J.M., Decker,H., Hartmann,H., Chavan,B., Rokos,H., Spencer,J.D., Hasse,S., Thornton,M.J., Shalbaf,M., Paus,R. *et al.* (2009) Senile hair graying: H2O2-mediated oxidative stress affects human hair color by blunting methionine sulfoxide repair. *FASEB J.*, **23**, 2065–2075.
21. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
22. Giles,J. (2005) Internet encyclopaedias go head to head. *Nature*, **438**, 900–901.
23. Adler,B.T., Chatterjee,K., de Alfaro,L., Faella,M. and Pye,I. (2008) Assigning trust to Wikipedia content. In WikiSym 08: Proceedings of the International Symposium on Wikis, Porto, Portugal; 150.
24. Krötzsch,M., Vrandečić,D. and Völkel,M. (2006) *The Semantic Web – ISWC 2006*. Springer, Heidelberg, Berlin.
25. Noy,N.F., Shah,N.H., Whetzel,P.L., Dai,B., Dorf,M., Griffith,N., Jonquet,C., Rubin,D.L., Storey,M., Chute,C.G. *et al.* (2009) BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.*, **37**, W170–W173.
26. Flicek,P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2008) Ensembl 2008. *Nucleic Acids Res.*, **36**, D707–D714.
27. Amberger,J., Bocchini,C.A., Scott,A.F. and Hamosh,A. (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.*, **37**, D793–D796.
28. Bairoch,A., Bougueleret,L., Altairac,S., Amendolia,V., Auchincloss,A., Argoud-Puy,G., Axelsen,K., Baratin,D., Blatter,M., Boeckmann,B. *et al.* (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, **37**, D169–D174.