PLOS COMPUTATIONAL BIOLOGY
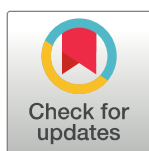
EDUCATION

Crowdsourcing biocuration: The Community Assessment of Community Annotation with Ontologies (CACAO)

Jolene Ramsey[1,2‡*], Brenley McIntosh[1‡], Daniel Renfro[1‡], Suzanne A. Aleksander[1‡¤a], Sandra LaBonte[1‡], Curtis Ross[1,2], Adrienne E. Zweifel[1], Nathan Liles[1¤b], Shabnam Farrar[1], Jason J. Gill[2,3], Ivan Erill[4,5], Sarah Ades[6], Tanya Z. Berardini[7], Jennifer A. Bennett[8], Siobhan Brady[9], Robert Britton[10¤c], Seth Carbon[11], Steven M. Caruso[4], Dave Clements[12], Ritu Dalia[13¤d], Meredith Defelice[6], Erin L. Doyle[14], Iddo Friedberg[15¤e], Susan M. R. Gurney[13¤f], Lee Hughes[16], Allison Johnson[17], Jason M. Kowalski[18¤g], Donghui Li[7], Ruth C. Lovering[19], Tamara L. Mans[20¤h], Fiona McCarthy[21¤i], Sean D. Moore[22], Rebecca Murphy[23], Timothy D. Paustian[24], Sarah Perdue[18¤j], Celeste N. Peterson[25], Birgit M. Prüß[26], Margaret S. Saha[27], Robert R. Sheehy[28], John T. Tansey[29], Louise Temple[30], Alexander William Thorman[31], Saul Trevino[32], Amy Cheng Vollmer[33], Virginia Walbot[34], Joanne Willey[35], Deborah A. Siegele[36*], James C. Hu[1,2†]

1 Department of Biochemistry & Biophysics, Texas A&M University, College Station, Texas, United States of America, 2 Center for Phage Technology, Texas A&M University, College Station, Texas, United States of America, 3 Department of Animal Science, Texas A&M University, College Station, Texas, United States of America, 4 Department of Biological Sciences, University of Maryland Baltimore County, Baltimore, Maryland, United States of America, 5 Department of Computer Science and Electrical Engineering, University of Maryland Baltimore County, Baltimore, Maryland, United States of America, 6 Department of Biochemistry & Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania, United States of America, 7 The Arabidopsis Information Resource, Phoenix Bioinformatics, Newark, California, United States of America, 8 Department of Biology and Earth Science, Otterbein University, Westerville, Ohio, United States of America, 9 Department of Plant Biology and Genome Center, University of California Davis, Davis, California, United States of America, 10 Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, Michigan, United States of America, 11 Division of Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, California, United States of America, 12 Department of Biology, John Hopkins University, Baltimore, Maryland, United States of America, 13 Department of Biology, Drexel University, Philadelphia, Pennsylvania, United States of America, 14 Biology Department, Doane University, Crete, Nebraska, United States of America, 15 Department of Microbiology, Miami University, Oxford, Ohio, United States of America, 16 Department of Biological Sciences, University of North Texas, Denton, Texas, United States of America, 17 Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, Virginia, United States of America, 18 Biological Sciences Department, University of Wisconsin-Parkside, Kenosha, Wisconsin, United States of America, 19 Institute of Cardiovascular Science, University College London, London, United Kingdom, 20 Department of Biochemistry and Biotechnology, Minnesota State University Moorhead, Brooklyn Park, Minnesota, United States of America, 21 Department of Basic Science, College of Veterinary Medicine, Mississippi State University, Starkville, Mississippi, United States of America, 22 Burnett School of Biomedical Sciences, University of Central Florida, Orlando, Florida, United States of America, 23 Department of Biology, Centenary College of Louisiana, Shreveport, Louisiana, United States of America, 24 Department of Bacteriology, University of Wisconsin, Madison, Wisconsin, United States of America, 25 Biology Department, Suffolk University, Boston, Massachusetts, United States of America, 26 Microbiological Sciences Department, North Dakota State University, Fargo, North Dakota, United States of America, 27 Department of Biology, College of William & Mary, Williamsburg, Virginia, United States of America, 28 Biology Department, Radford University, Radford, Virginia, United States of America, 29 Department of Biochemistry and Molecular Biology, Otterbein University, Westerville, Ohio, United States of America, 30 School of Integrated Sciences, James Madison University, Harrisonburg, Virginia, United States of America, 31 Department of Environmental and Public Health Sciences, University of Cincinnati, Cincinnati, Ohio, United States of America, 32 Department of Chemistry, Math, and Physics, Houston Baptist University, Houston, Texas, United States of America, 33 Department of Biology, Swarthmore College, Swarthmore, Pennsylvania, United States of America, 34 Department of Biology, Stanford University, Stanford, California, United States of America, 35 Department of Science Education, Donald and Barbara

Zucker School of Medicine at Hofstra/Northwell, Hempstead, New York, United States of America,
36 Department of Biology, Texas A&M University, College Station, Texas, United States of America

† Deceased.
¤a Current address: Department of Genetics, Stanford University, Palo Alto, California, United States of America
¤b Current address: Advanced Research Computing, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, United States of America
¤c Current address: Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, Texas, United States of America
¤d Current address: MaST Community Charter School, Philadelphia, Pennsylvania, United States of America
¤e Current address: Department of Veterinary Microbiology and Preventive Medicine, Iowa State University, Ames, Iowa, United States of America
¤f Current address: School of Biology, University of St Andrews, St Andrews, United Kingdom
¤g Current address: Department of Math and Natural Sciences, Marian University, Fond du Lac, Wisconsin, United States of America
¤h Current address: North Hennepin Community College, Brooklyn Park, Minnesota, United States of America
¤i Current address: Animal and Comparative Biomedical Sciences, The University of Arizona, Tucson, Arizona, United States of America
¤j Current address: Department of Physics, University of Wisconsin, Madison, Wisconsin, United States of America
‡ Co-first authors.
* ecoliwiki@gmail.com (JR); d-siegele@tamu.edu (DAS)

Abstract

Experimental data about gene functions curated from the primary literature have enormous value for research scientists in understanding biology. Using the Gene Ontology (GO), manual curation by experts has provided an important resource for studying gene function, especially within model organisms. Unprecedented expansion of the scientific literature and validation of the predicted proteins have increased both data value and the challenges of keeping pace. Capturing literature-based functional annotations is limited by the ability of biocurators to handle the massive and rapidly growing scientific literature. Within the community-oriented wiki framework for GO annotation called the Gene Ontology Normal Usage Tracking System (GONUTS), we describe an approach to expand biocuration through crowdsourcing with undergraduates. This multiplies the number of high-quality annotations in international databases, enriches our coverage of the literature on normal gene function, and pushes the field in new directions. From an intercollegiate competition judged by experienced biocurators, Community Assessment of Community Annotation with Ontologies (CACAO), we have contributed nearly 5,000 literature-based annotations. Many of those annotations are to organisms not currently well-represented within GO. Over a 10-year history, our community contributors have spurred changes to the ontology not traditionally covered by professional biocurators. The CACAO principle of relying on community members to participate in and shape the future of biocuration in GO is a powerful and scalable model used to promote the scientific enterprise. It also provides undergraduate students with a unique and enriching introduction to critical reading of primary literature and acquisition of marketable skills.

Author summary

The primary scientific literature catalogs the results from publicly funded scientific research about gene function in human-readable format. Information captured from those studies in a widely adopted, machine-readable standard format comes in the form of Gene Ontology (GO) annotations about gene functions from all domains of life. Manual annotations based on inferences directly from the scientific literature, including the evidence used to make such inferences, represent the best return on investment by improving data accessibility across the biological sciences and allowing novel insights between evolutionarily related organisms. To supplement professional curation, our Community Assessment of Community Annotation with Ontologies (CACAO) project enabled annotation of the scientific literature by community annotators, in this case undergraduates, which resulted in the contribution of thousands of unique, validated entries to public resources. Importantly, the annotations described here initiated by non-experts often deal with topics not typically covered by the experts. These annotations are now being used by scientists worldwide in their research efforts.

Introduction

Biocuration captures information from the primary literature in a computationally accessible fashion. The biocuration process generates annotations connecting experimental data with unique identifiers representing precisely defined ontology terms and logical relationships. While the majority of existing annotations are computational predictions built on knowledge from human biocuration, manually curated annotations from published experimental data are still the gold standard for functional annotations [1]. Universal access to well-curated data-bases, such as UniProt and those maintained by model organism consortia, allows scientists worldwide to leverage computational approaches to solve pressing biological problems. New insights on complex cellular processes such as autophagy, cell polarity, and division can be clarified after assessing relationships in curated data [2–4]. The Gene Ontology (GO; http://geneontology.org/) is an evolving biocuration resource that provides the framework for capturing attributes of gene products within 3 aspects or main branches: biological process, molecular function, and cellular component [5,6]. Importantly, connections can be made between model organism genes and human genes with comprehensive GO coverage [7]. Additionally, using GO data generates testable hypotheses in areas with little direct experimentation [8–10]. Application to high-throughput and systems biology, for instance, has led to insights and better methods for identification and analysis of the genes involved in cardiac and Alzheimer disease [11,12].

Without question, GO is a critical scientific resource, but manual annotation is an extremely labor-intensive process [13,14]. The pace at which the information is generated in the literature exceeds the capacity of professional biocurators to perform manual curation and the willingness of funding agencies to pay for a larger biocurator labor force [15]. Although the general Swiss-Prot protein database (https://www.uniprot.org/) model is one example of a scalable process for targeted manual and automated annotations from the (annotatable) literature, most fields are limited by low numbers of trained personnel and minimal participation from trained scientists [16,17]. The problem is most severe for communities studying organisms without a funded model organism database. Nevertheless, curation of the experimental literature from as many species as possible strengthens inference of function when there is substantial evolutionary conservation [18,19]. Several groups are developing tools to facilitate

community engagement, such as the Gene Ontology Normal Usage Tracking System (GONUTS) site described here. These efforts stem from the realization that, while most scientists acknowledge the importance of data curation, it is hard to motivate individuals to volunteer their knowledge [20,21]. Spectacular crowdsourcing successes include the analysis of Shiga toxin–producing *Escherichia coli* [22], the solution of the structure of an HIV protease by the FoldIt player community [23], and science content within Wikipedia [24–27]. In other cases, high-profile community annotation efforts have been less successful [28], which we attribute to the disconnect between traditional incentives for funding and promotion in academia [29].

Here, we describe the successful implementation over nearly a decade of a university instruction-based model resulting in nearly 5,000 high-quality community annotations added to the GO database. This effort was motivated by the clear parallels between the foundational skills used by professional biocurators and the well-defined goals for undergraduate training [30]. A professional GO biocurator creates gene annotations by finding relevant primary literature, extracting information about normal gene function from it, and entering that information using the controlled GO vocabulary into online databases [31]. We demonstrate that university students, guided by their instructors, could accomplish similar tasks and perform community GO annotation while developing strong critical reading skills in a templated annotation task requiring rigorous reading of primary scientific literature.

Results

Sustainable community member contribution via an online intercollegiate competition

To address the need for broader participation and expansion beyond model organism databases, we initiated an intercollegiate competition based at Texas A&M University mainly for undergraduate students, called Community Assessment of Community Annotation with Ontologies (CACAO). Here, we limit the discussion to details of the competition that are relevant to annotation as the specifics of teaching practice were previously reported [32]. Leveraging the GONUTS wiki platform (https://gowiki.tamu.edu/), a framework for experts not familiar with GO to annotate from literature in their field, teams of students (competitors) participate in the CACAO competition (Fig 1) [33]. Instructors (also called judges) assess all annotations entered by competitors for accuracy and completeness, then give feedback. Peer review by the competitors is incentivized by awarding points for challenges that correct an entry. Teams earn points only for correct annotations and challenges. The team with the highest points accumulated over the competition period wins. Vetted, high-quality annotations are then submitted to the GO Consortium database. CACAO quickly expanded, hosting 39 competitions over 8 years including 23 colleges and universities, with 792 community annotators and 50 judges. After reading 2,879 peer-reviewed journal articles, community members submitted 11,123 annotations to GONUTS (Fig 1). Following careful review through 2018, 4,913 diverse annotations were added to the GO Consortium database (Fig 1). Those annotations are maintained as mandated by updates or changes in the ontology.

Annotations generated through CACAO are diverse, novel, and specific

The 4,913 annotations contributed through GONUTS have spanned all domains of life plus viruses, with the majority being skewed toward eukaryotes, in particular model organisms among the chordates (human, mouse, rat, etc.), *Streptophyta* (plants including *Arabidopsis*), and *Ascomycota* (such as budding yeast) (Fig 2A). As only unique annotations are accepted,
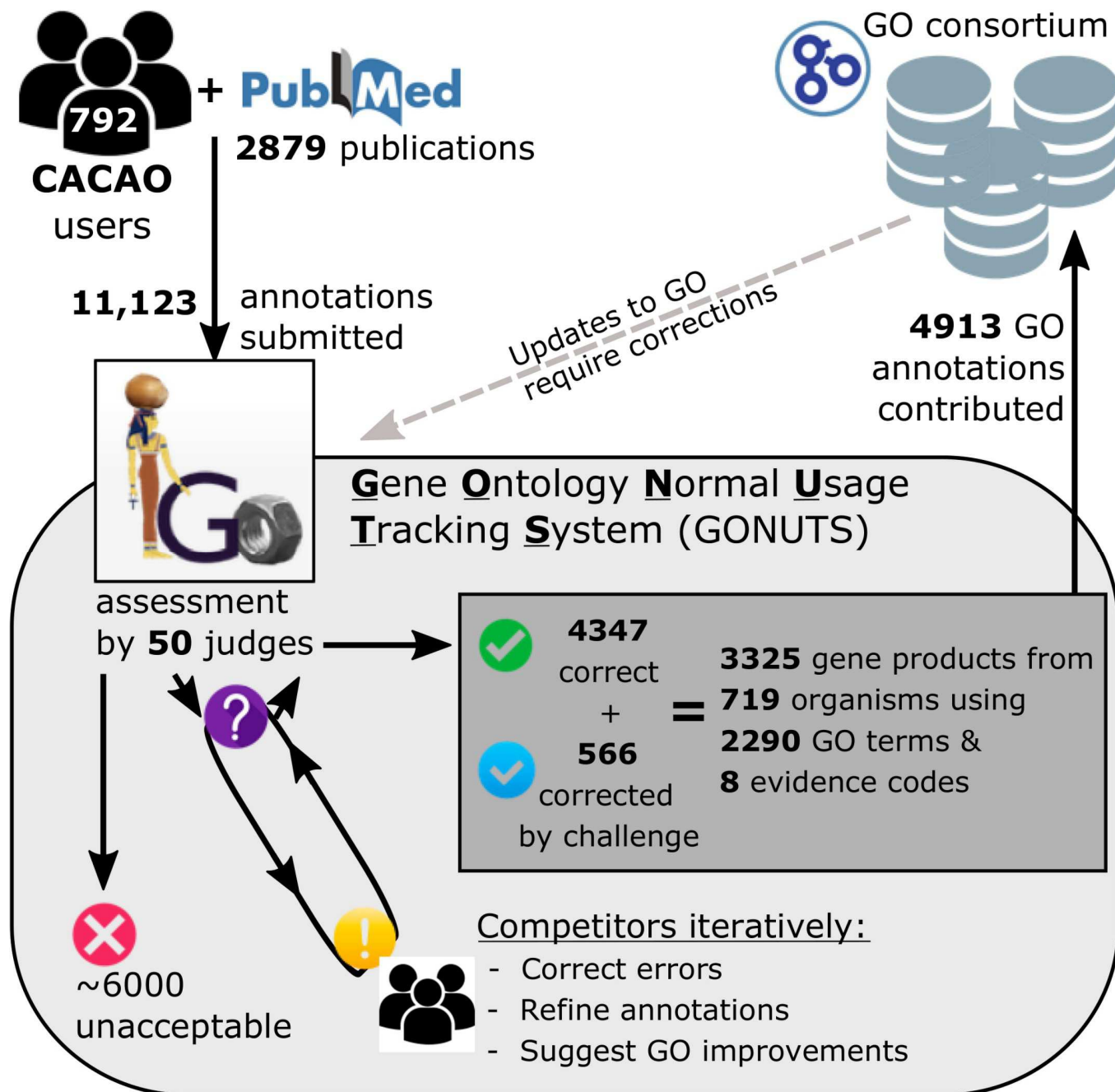
Fig 1. CACAO competitors contribute a large number of GO annotations. Overall CACAO contributions are summarized in the context of the workflow for quality control and submission to the GO Consortium. CACAO users consume the primary literature, collect information about normal gene functions from the paper study subjects, and capture the evidence and conclusions using the GO. Those annotations are reviewed by trained judges and marked as unacceptable (red X), requiring changes (yellow!, or purple? flagged for further review), or acceptable (green check, or blue check after correction) within the GONUTS framework. Competitors challenge entries and engage in peer review until an annotation is corrected or marked unacceptable. Fully vetted annotations are deposited into the public GO database maintained by professional biocurators and used by scientists worldwide. As required, CACAO-submitted annotations will be updated to reflect rearrangements and changes in GO. CACAO, Community Assessment of Community Annotation with Ontologies; GO, Gene Ontology; GONUTS, Gene Ontology Normal Usage Tracking System.

https://doi.org/10.1371/journal.pcbi.1009463.g001

A
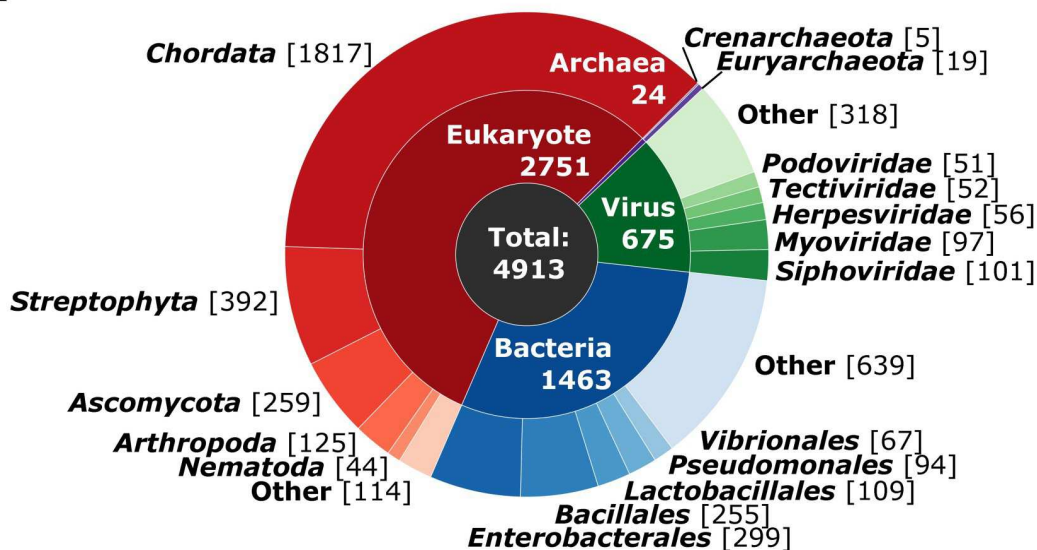
B
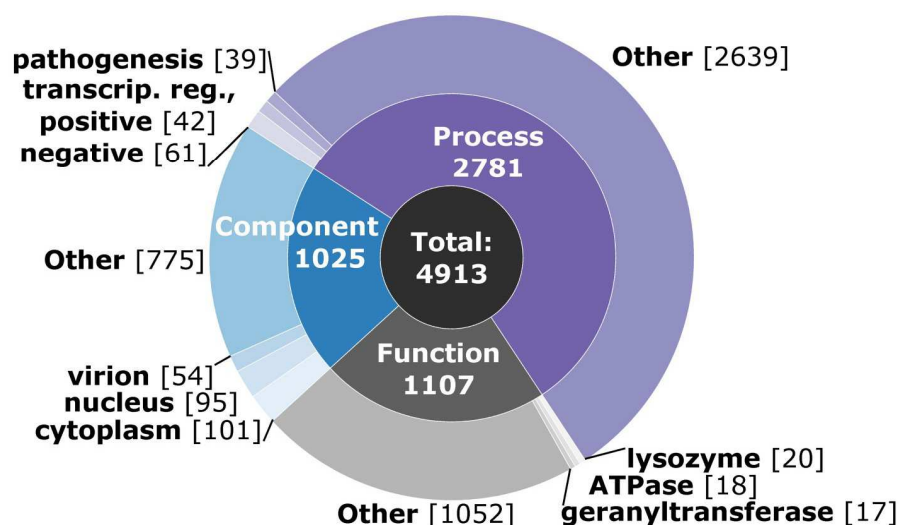
C
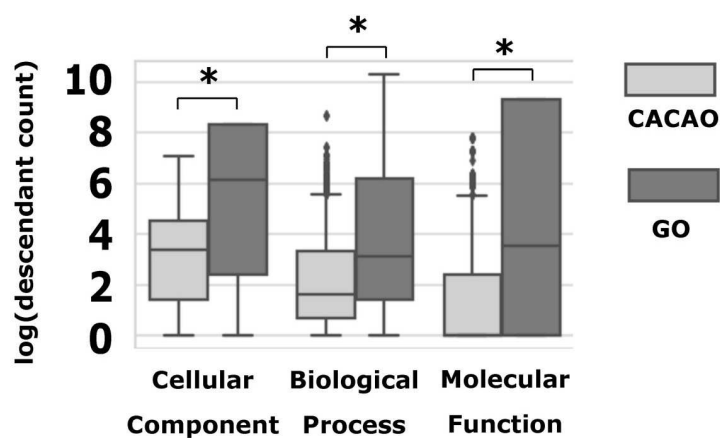
Fig 2. The GO annotations contributed by CACAO users are diverse and specific. (A) Proteins annotated by CACAO users are depicted by species domain. The organisms most highly represented in each domain are displayed on the outer ring of the chart divided by the following rank: phylum for eukaryotes and archaea, order for bacteria, and family for viruses. The number of GO annotations in each category is indicated in brackets. **(B)** The distribution of GO terms used for CACAO annotations are graphed by aspect within the ontology. The top 3 terms within each aspect are labeled on the outer ring. For clarity, "activity" was dropped from each function term, and the process terms were abbreviated from "positive/negative regulation of transcription, DNA-templated" to "transcript. reg., positive or negative." The number of GO annotations for each term is indicated in brackets. **(C)** The descendant counts, corresponding to depth within the ontology, for CACAO annotations ($n = 4,913$) and all other manual GO annotations in UniProt through 2019 ($n = 255,958$) are graphed. Significant differences measured by the Mann–Whitney test with $p < 0.001$ are marked with an *. CACAO, Community Assessment of Community Annotation with Ontologies; GO, Gene Ontology.

this demonstrates that community members can help fill gaps left by professional biocurators working for model organism databases. CACAO annotations also go beyond model organisms. The 614 annotations to viral genes made by CACAO participants represented 285 eukaryotic viruses and 384 viruses that infect bacteria (bacteriophages). Nearly half of the approximately 1,000 annotations listed for bacteriophages in QuickGO list CACAO as the source. Annotations for bacterial proteins make up only 5% of total GO annotations, but 30% of CACAO annotations. At the order level, the top 5 bacterial categories (Enterobacterales, Bacillales, Lactobacillales, Pseudomonales, and Vibrionales) are heavily studied Gram-negative and Gram-positive organisms of importance to microbiology research and the medical community. The microbial (virus and bacteria) entities herein described represent high genetic diversity and often serve as the basis for significant automated propagation to eukaryotic gene products. Thus, we conclude that not only do CACAO annotators fill gaps for model organisms, but also expand coverage to a wide array of otherwise poorly curated species.

CACAO participants annotate to a wide variety of specific terms (Fig 2B). The top 3 most used terms within each aspect, approximately 5% for biological process and molecular function or approximately 24% for cellular component, are only a small proportion of the total for that branch. While the cellular component terms used are relatively general (i.e., nucleus), the top process and function terms are near leaf-level and more specific, having few to no child terms. To better understand the level of detail captured in annotations made by CACAO users, we used GOATOOLS, a Python package for representing where terms fall within the ontology hierarchical graph [34]. Based on the variety of annotation types in our set (e.g., aspects and species), we selected a measure that counts the number of descendants ($dcnt$), or child terms, for each entry. Higher level terms will have a larger score and are considered general or global. More descriptive terms with no descendants, or leaf-level terms, are more precise or detailed and receive the lowest $dcnt$ value. The $dcnt$ analysis quantitatively demonstrates that CACAO annotations are made to specific terms (Fig 2C). That pattern is consistent with the way annotations were reviewed, where only the most specific term that could be chosen based on the details reported in the paper was counted correct. For comparison, we performed the same $dcnt$ analysis on all manual GO annotations available through 2019. The distributions of $dcnt$ values for GO annotations are broader and statistically different from CACAO within each aspect (Fig 2C). These data demonstrate that community users can contribute high-quality, precise, and scientifically relevant annotations to GO.

CACAO community curators enrich ontology development

The GO changes over time to reflect research progress and improve the representation of biological knowledge [35]. The GO Consortium tracks requests to change the ontology via their GitHub repository accessible on the Helpdesk (http://help.geneontology.org/). CACAO users have submitted >50 tickets via this system, resulting in the creation of 49 new GO terms,

many of which now have child terms added by others. Given the diverse literature areas read by community curators, many of these terms are breaking new ground in the ontology. At time of writing, the new terms added based on CACAO feedback had been used >650 times by curators. In addition, at least 14 nonterm changes, such as clarified definitions and relationships for current terms, have also occurred. A beneficial, unintended consequence of CACAO is that curators are compelled to resolve issues within the ontology and incorporate new knowledge from areas that are not traditionally covered by model organism databases.

Discussion

Community member annotations through CACAO add long-term value to GO

The GO resource is among the computational tools most cited by biologists [6]. Automatically inferred annotations, those made without curator intervention, are temporary but make up a significant dynamic proportion of the total GO annotations at any given time. However, the quality of computationally assigned annotations relies on a solid undergirding of manual annotations where the data are reviewed and then annotations are created by a dedicated biocuration community [36]. Of the >6 million manual annotations in the July 2021 release of GO files, approximately one-sixth of the human-curated manual annotations come from traditional experimental evidence, and most of the rest of them come from sequence similarity and phylogenetic similarity evidence [19]. The efforts described here are not meant to rival the volume produced by dedicated biocurators, nor to replace that organized effort. Instead, we demonstrate how small contributions from many individual community members over time accumulate into a unique and valuable resource. By virtue of its decoupling from the traditional funding model, community curation supplements professional biocuration, especially in underfunded areas [17].

Targeted crowdsourcing with attribution makes CACAO annotation sustainable

Recognizing the need to pull expertise from diverse bench scientists, various other initiatives have been implemented to encourage community participation with lower cost [37,38]. For example, the PomBase community curation project called Canto has garnered up to an impressive 50% response rate for co-annotation from authors within their community [38]. Another natural by-product of crowdsourcing is the diversification of the biocuration workforce. Such introduction of new expertise and perspectives is analogous to the workplace observation that diverse teams innovate and produce more than homogenous ones [39]. While the majority in the "crowd" may be unlikely to participate [40], the CACAO implementation of GONUTS is a sustainable model for community contribution of vetted GO annotations in areas of current interest because it caters to a nonrandom crowd, primarily students in an academic course setting.

In a resource-limited environment, the need to incentivize data curation has been creatively approached with different methods such as the micropublication format [41–43]. Yet, motivating researchers to weigh in on ontology structure is a long-standing challenge [20]. Recognizing the need to credit individuals for their annotation efforts, UniProt now offers a portal for submitting literature-based curation linked to an ORCiD (https://community.uniprot.org/bbsub/bbsub.html) [44], as does the new Generic Online Annotation Tool built for the plant community (http://goat.phoenixbioinformatics.org/). Importantly, the GONUTS wiki provides a web-based public record of CACAO contributions, allowing individuals to cite their efforts.

CACAO contributions are valuable because they are unique

On the one hand, community curators can spend the time to read and extract information from redundant papers (those with information highly similar to already curated literature and conclusions) to enhance model organism annotation depth and increase confidence in existing annotations. On the other hand, community curators sample from a vast literature space outside the typical biocurator's expertise, expanding overall organism coverage, such as shown for microbial organisms here [45–49]. Because microbial genomes are typically smaller, groups of students can make a major contribution. A significant instance is adding approximately 50% of all phage GO annotations available in the GO annotation files. CACAO has also spurred updates to ontology relationships. For example, a large rearrangement of biofilm GO terms occurred after CACAO users initiated discussion about their parentage and definitions.

Community curation through CACAO meets modern open-source research and education goals

With online education thrust to the forefront during the global Coronavirus Disease 2019 (COVID-19) pandemic, sustainable and authentic education-driven engagement solutions are critically needed [30,50,51]. Community-driven skills-based classroom research in any number of formats (e.g., CACAO, genome annotation [52–54]) serves the scientific community. From an educational perspective, the competition aspect is an engaging format that models real-world scientific skill development with regard to critical reading, iterative editing of a product, and peer review. We hypothesize that this mini biocurator experience may have similar benefits for recruitment, retention, and graduation observed with undergraduate research [55,56]. The biocuration model is highly applicable to scientists and trainees worldwide and complies with Findable, Accessible, Interoperable, and Reusable (FAIR) [57] data principles, making its results accessible to all. GO annotation for Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) and its infection of human cells was immediately pursued to aid strategic planning of the pandemic response (http://geneontology.org/covid-19.html). We appeal to scientists to participate in biocuration efforts through GONUTS, UniProt, or a model organism database/the Alliance of Genome Resources where users can contribute from the comfort of any computer [58].

Materials and methods

CACAO competitions for intercollegiate teams are hosted on GONUTS (https://gowiki.tamu.edu/). Raw data for all users and every annotation history are maintained by custom extensions to the MediaWiki software used by GONUTS [33]. Additional information about competition rules can be found at https://gowiki.tamu.edu/wiki/index.php/Category:CACAO. The data presented here encompass annotations generated from 2010 to 2018, with expanded taxon information retrieved using the UniProt application programming interface (API) as well as the ETE (v3.1.1) module and various tools from BioPython (v1.74) [59,60]. Summary statistics for CACAO annotations given in Fig 1 were mined from our local database storage.

Fully correct annotation data are transferred from GONUTS regularly via the current Gene Association File (GAF) or Gene Product Association Data (GPAD) file format, as outlined in GO requirements, directly to the European Bioinformatics Institute's Protein2GO for incorporation into the complete GO annotation files. All currently included annotations are accessible on GONUTS or via the search engine QuickGO (https://www.ebi.ac.uk/QuickGO/annotations) by filtering for parameter "assigned by" CACAO and are also provided as a Supporting information dataset in GPAD format (S1 File) [61].

The 01-01-2020 non IEA GAF (goa_uniprot_all_noiea.gaf.gz) and ontology file (go.obo) were downloaded from http://release.geneontology.org/ for the *dcnt* analysis. Values for *dcnt* were calculated according to GOATOOLS on all manual annotations not assigned by CACAO [34]. The Mann–Whitney test with a 2-sided *p*-value was used to compare GO and CACAO *dcnt* distributions within each aspect using SciPy [62,63].

For the phage analyses, the GAF was filtered into a subset using the following TaxIDs from the NCBI Taxonomy browser: 12333 (unclassified bacterial viruses), 1714267 (Gammasphaerolipovirus), 10656 (Tectiviridae), 10472 (Plasmaviridae), 10659 (Corticoviridae), 10841 (Microviridae), 10860 (Inoviridae), 28883 (Caudovirales), 11989 (Leviviridae), and 10877 (Cystoviridae).

Changes to the ontology initiated by CACAO users were tallied by searching through the GO issue tracker at GitHub (https://github.com/geneontology/go-ontology/issues) for user handles: @jimhu-tamu, @suzialeksander, @sandyl27, @jrr-cpt, @ivanerill, and/or the query text "CACAO" for open and closed issues, then manually reviewed for accuracy. Matplotlib (v3.1.1) and Seaborn (v0.9.0) were used to generate pie charts, box plots, and bar graphs [64,65]. Figures were compiled and rendered with the open-source program Inkscape 0.92.2.

Supporting information

S1 File. CACAO GPAD data. The full GPAD format file (in an Excel workbook) with all annotations used for the analysis presented. CACAO, Community Assessment of Community Annotation with Ontologies; GPAD, Gene Product Association Data.
(XLSX)

Acknowledgments

References

1. Škunca N, Altenhoff A, Dessimoz C. Quality of computationally inferred gene ontology annotations. PLoS Comput Biol. 2012; 8(5):e1002533. https://doi.org/10.1371/journal.pcbi.1002533 PMID: 22693439

2. Sun L-L, Li M, Suo F, Liu X-M, Shen E-Z, Yang B, et al. Global analysis of fission yeast mating genes reveals new autophagy factors. PLoS Genet. 2013; 9(8):e1003715. https://doi.org/10.1371/journal.pgen.1003715 PMID: 23950735

3. Denny P, Feuermann M, Hill DP, Lovering RC, Plun-Favreau H, Roncaglia P. Exploring autophagy with Gene Ontology. Autophagy. 2018; 14(3):419–36. https://doi.org/10.1080/15548627.2017.1415189 PMID: 29455577

4. Lee ME, Rusin SF, Jenkins N, Kettenbach AN, Moseley JB. Mechanisms connecting the conserved protein kinases Ssp1, Kin1, and Pom1 in fission yeast cell polarity and division. Curr Biol. 2018; 28 (1):84–92.e4. https://doi.org/10.1016/j.cub.2017.11.034 PMID: 29249658

5. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000; 25(1):25–9. https://doi.org/10.1038/75556 PMID: 10802651

6. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. 2021; 49(D1):D325–34. https://doi.org/10.1093/nar/gkaa1113 PMID: 33290552

7. Khodiyar VK, Howe D, Talmud PJ, Breckenridge R, Lovering RC. From zebrafish heart jogging genes to mouse and human orthologs: using Gene Ontology to investigate mammalian heart development. F1000Res. 2013; 2:242. https://doi.org/10.12688/f1000research.2-242.v2 PMID: 24627794

8. Zhang C, Zheng W, Freddolino PL, Zhang Y. MetaGO: predicting Gene Ontology of non-homologous proteins through low-resolution protein structure prediction and protein-protein network mapping. J Mol Biol. 2018; 430(15):2256–65. https://doi.org/10.1016/j.jmb.2018.03.004 PMID: 29534977

9. Zhang C, Wei X, Omenn GS, Zhang Y. Structure and protein interaction-based Gene Ontology annotations reveal likely functions of uncharacterized proteins on human chromosome 17. J Proteome Res. 2018; 17(12):4186–96. https://doi.org/10.1021/acs.jproteome.8b00453 PMID: 30265558

10. Zhang C, Freddolino PL, Zhang Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein-protein interaction information. Nucleic Acids Res. 2017; 45(W1): W291–9. https://doi.org/10.1093/nar/gkx366 PMID: 28472402

11. Lovering RC, Roncaglia P, Howe DG, Laulederkind SJF, Khodiyar VK, Berardini TZ, et al. Improving interpretation of cardiac phenotypes and enhancing discovery with expanded knowledge in the Gene Ontology. Circ Genom Precis Med. 2018; 11(2):e001813. https://doi.org/10.1161/CIRCGEN.117.001813 PMID: 29440116

12. Andrew RJ, Fisher K, Heesom KJ, Kellett KAB, Hooper NM. Quantitative interaction proteomics reveals differences in the interactomes of amyloid precursor protein isoforms. J Neurochem. 2019; 149(3):399–412. https://doi.org/10.1111/jnc.14666 PMID: 30664241

13. Li D, Berardini TZ, Muller RJ, Huala E. Building an efficient curation workflow for the *Arabidopsis* literature corpus. Database. 2012; 2012:bas047. https://doi.org/10.1093/database/bas047 PMID: 23221298

14. Drabkin HJ, Blake JA for the Mouse Genome Informatics Database. Manual Gene Ontology annotation workflow at the Mouse Genome Informatics Database. Database. 2012; 2012:bas045. https://doi.org/10.1093/database/bas045 PMID: 23110975

15. Bastow R, Leonelli S. Sustainable digital infrastructure. EMBO Rep. 2010; 11(10):730–4. https://doi.org/10.1038/embor.2010.145 PMID: 20847740

16. Lewis SE. The vision and challenges of the Gene Ontology. In: Dessimoz C, Škunca N, editors. The Gene Ontology Handbook. Methods in Molecular Biology, vol 1446. Humana Press, New York, NY; 2017. p. 291–302.

17. Poux S, Arighi CN, Magrane M, Bateman A, Wei C-H, Lu Z, et al. On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. Bioinformatics. 2017; 33(21):3454–60. https://doi.org/10.1093/bioinformatics/btx439 PMID: 29036270

18. Tang H, Finn RD, Thomas PD. TreeGrafter: phylogenetic tree-based annotation of proteins with Gene Ontology terms and other annotations. Bioinformatics. 2019; 35(3):518–20. https://doi.org/10.1093/bioinformatics/bty625 PMID: 30032202

19. Gaudet P, Livstone MS, Lewis SE, Thomas PD. Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. Brief Bioinform. 2011; 12(5):449–62. https://doi.org/10.1093/bib/bbr042 PMID: 21873635

20. Ong E, He Y. Community-based ontology development, annotation and discussion with MediaWiki extension Ontokiwi and Ontokiwi-based Ontobedia. AMIA Jt Summits Transl Sci Proc. 2016; 2016:65–74. PMID: 27570653

21. International Society for Biocuration. Biocuration: Distilling data into knowledge. PLoS Biol. 2018; 16(4): e2002846. https://doi.org/10.1371/journal.pbio.2002846 PMID: 29659566

22. Rohde H, Qin J, Cui Y, Li D, Loman NJ, Hentschke M, et al. Open-source genomic analysis of Shiga-toxin–producing E. coli O104:H4. N Engl J Med. 2011; 365(8):718–24. https://doi.org/10.1056/NEJMoa1107643 PMID: 21793736

23. Khatib F, DiMaio F, Foldit Contenders Group, Foldit Void Crushers Group, Cooper S, Kazmierczyk M, et al. Crystal structure of a monomeric retroviral protease solved by protein folding game players. Nat Struct Mol Biol. 2011; 18(10):1175–7. https://doi.org/10.1038/nsmb.2119 PMID: 21926992

24. Giles J. Internet encyclopaedias go head to head. Nature. 2005; 438(7070):900–1. https://doi.org/10.1038/438900a PMID: 16355180

25. Good BM, Clarke EL, de Alfaro L, Su AI. The Gene Wiki in 2011: community intelligence applied to human gene annotation. Nucleic Acids Res. 2011; 40(Database issue):D1255–61. https://doi.org/10.1093/nar/gkr925 PMID: 22075991

26. Reavley NJ, Mackinnon AJ, Morgan AJ, Alvarez-Jimenez M, Hetrick SE, Killackey E, et al. Quality of information sources about mental disorders: a comparison of Wikipedia with centrally controlled web and printed sources. Psychol Med. 2011; 42(8):1753–62. https://doi.org/10.1017/S003329171100287X PMID: 22166182

27. Arroyo-Machado W, Torres-Salinas D, Herrera-Viedma E, Romero-Frías E. Science through Wikipedia: A novel representation of open knowledge through co-citation networks. PLoS ONE. 2020; 15(2): e0228713. https://doi.org/10.1371/journal.pone.0228713 PMID: 32040488

28. Mons B, Ashburner M, Chichester C, Mulligen E, Weeber M, den J, et al. Calling on a million minds for community annotation in WikiProteins. Genome Biol. 2008; 9(5):R89. https://doi.org/10.1186/gb-2008-9-5-r89 PMID: 18507872

29. Callaway E. No rest for the bio-wikis. Nature. 2010; 468(7322):359–60. https://doi.org/10.1038/468359a PMID: 21085149

30. Bauerle C, DePass A, Lynn D, O'Connor C, Singer S, Withers M, et al. Vision and change in undergraduate biology education: a call to action [Internet]. Washington, D.C.: AAAS;2011. [cited 2021 June 24] Available from: http://visionandchange.org/finalreport.

31. Balakrishnan R, Harris MA, Huntley R, Van Auken K, Cherry JM. A guide to best practices for Gene Ontology (GO) manual annotation. Database. 2013; 2013(0):bat054. https://doi.org/10.1093/database/bat054 PMID: 23842463

32. Erill I, Caruso S, Hu JC. Gamifying critical reading through a genome annotation intercollegiate competition. Tested Studies in Laboratory Teaching. 2018; 39:Article 6.

33. Renfro DP, McIntosh BK, Venkatraman A, Siegele DA, Hu JC. GONUTS: the Gene Ontology Normal Usage Tracking System. Nucleic Acids Res. 2012 Jan; 40(Database issue):D1262–9. https://doi.org/10.1093/nar/gkr907 PMID: 22110029

34. Klopfenstein DV, Zhang L, Pedersen BS, Ramírez F, Vesztrocy AW, Naldi A, et al. GOATOOLS: A Python library for Gene Ontology analyses. Sci Rep. 2018 Jul 18; 8(1):10872. https://doi.org/10.1038/s41598-018-28948-z PMID: 30022098

35. Leonelli S, Diehl AD, Christie KR, Harris MA, Lomax J. How the gene ontology evolves. BMC Bioinformatics. 2011 Aug 5; 12:325. https://doi.org/10.1186/1471-2105-12-325 PMID: 21819553

36. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019 Jan 8; 47(D1):D330–8. https://doi.org/10.1093/nar/gky1055 PMID: 30395331

37. Mortensen JM, Minty EP, Januszyk M, Sweeney TE, Rector AL, Noy NF, et al. Using the wisdom of the crowds to find critical errors in biomedical ontologies: a study of SNOMED CT. J Am Med Inform Assoc. 2015 May; 22(3):640–8.

38. Lock A, Harris MA, Rutherford K, Hayles J, Wood V. Community curation in PomBase: enabling fission yeast experts to provide detailed, standardized, sharable annotation from research publications. Database. 2020 Jan 1;2020:baaa028. https://doi.org/10.1093/database/baaa028 PMID: 32353878

39. Swartz TH, Palermo AS, Masur SK, Aberg JA. The science and value of diversity: closing the gaps in our understanding of inclusion and diversity. J Infect Dis. 2019 Aug 20; 220(Issue Supplement_2):S33–41. https://doi.org/10.1093/infdis/jiz174 PMID: 31430380

40. Karp PD. Crowd-sourcing and author submission as alternatives to professional curation. Database. 2016 Dec 26;2016:baw149. https://doi.org/10.1093/database/baw149 PMID: 28025340

41. Raciti D, Yook K, Harris TW, Schedl T, Sternberg PW. Micropublication: incentivizing community curation and placing unpublished data into the public domain. Database. 2018 Jan 1;2018:bay013. https://doi.org/10.1093/database/bay013 PMID: 29688367

42. Bourne PE, Lorsch JR, Green ED. Perspective: Sustaining the big-data ecosystem. Nature. 2015 Nov 5; 527(7576):S16–7. https://doi.org/10.1038/527S16a PMID: 26536219

43. Karp PD. How much does curation cost? Database. 2016 Aug 7;2016:baw110. https://doi.org/10.1093/database/baw110 PMID: 27504008

44. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019 Jan 8; 47(D1):D506–15. https://doi.org/10.1093/nar/gky1049 PMID: 30395287

45. Aurrecoechea C, Barreto A, Basenko EY, Brestelli J, Brunk BP, Cade S, et al. EuPathDB: the eukaryotic pathogen genomics database resource. Nucleic Acids Res. 2017 Jan 4; 45(D1):D581–91. https://doi.org/10.1093/nar/gkw1105 PMID: 27903906

46. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, et al. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. Nucleic Acids Res. 2017 Jan 4; 45 (D1):D535–42. https://doi.org/10.1093/nar/gkw1017 PMID: 27899627

47. Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, Hunt V, et al. ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Res. 2012; 40(Database issue): D593–8. https://doi.org/10.1093/nar/gkr859 PMID: 22006842

48. Zhang Y, Aevermann BD, Anderson TK, Burke DF, Dauphin G, Gu Z, et al. Influenza Research Database: An integrated bioinformatics resource for influenza virus research. Nucleic Acids Res. 2017 Jan 4; 45(D1):D466–74. https://doi.org/10.1093/nar/gkw857 PMID: 27679478

49. Giglio MG, Collmer CW, Lomax J, Ireland A. Applying the Gene Ontology in microbial annotation. Trends Microbiol. 2009 Jul; 17(7):262–8. https://doi.org/10.1016/j.tim.2009.04.003 PMID: 19577473

50. Hoskins SG, Lopatto D, Stevens LM. The C.R.E.A.T.E. approach to primary literature shifts undergraduates' self-assessed ability to read and analyze journal articles, attitudes about science, and epistemological beliefs. CBE Life Sci Educ. 2011; 10(4):368–78. https://doi.org/10.1187/cbe.11-03-0027 PMID: 22135371

51. Round JE, Campbell AM. Figure facts: encouraging undergraduates to take a data-centered approach to reading primary literature. CBE Life Sci Educ. 2013; 12(1):39–46. https://doi.org/10.1187/cbe.11-07-0057 PMID: 23463227

52. Ditty JL, Kvaal CA, Goodner B, Freyermuth SK, Bailey C, Britton RA, et al. Incorporating genomics and bioinformatics across the life sciences curriculum. PLoS Biol. 2010 Aug 10; 8(8):e1000448. https://doi.org/10.1371/journal.pbio.1000448 PMID: 20711478

53. Hosmani PS, Shippy T, Miller S, Benoit JB, Munoz-Torres M, Flores-Gonzalez M, et al. A quick guide for student-driven community genome annotation. PLoS Comput Biol. 2019 Apr 3; 15(4):e1006682. https://doi.org/10.1371/journal.pcbi.1006682 PMID: 30943207

54. Ramsey J, Rasche H, Maughmer C, Criscione A, Mijalis E, Liu M, et al. Galaxy and Apollo as a biologist-friendly interface for high-quality cooperative phage genome annotation. PLoS Comput Biol. 2020 Nov 2; 16(11):e1008214. https://doi.org/10.1371/journal.pcbi.1008214 PMID: 33137082

55. Jordan TC, Burnett SH, Carson S, Caruso SM, Clase K, DeJong RJ, et al. A broadly implementable research course in phage discovery and genomics for first-year undergraduate students. mBio. 2014 Feb 4; 5(1):e01051–13. https://doi.org/10.1128/mBio.01051-13 PMID: 24496795

56. Hanauer DI, Graham MJ, SEA-PHAGES, Betancur L, Bobrownicki A, Cresawn SG, et al. An inclusive Research Education Community (iREC): Impact of the SEA-PHAGES program on research outcomes and student learning. Proc National Acad Sci U S A. 2017 Dec 19; 114(51):13531–6. https://doi.org/10.1073/pnas.1718188115 PMID: 29208718

57. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15; 3:160018.

58. Alliance of Genome Resources Consortium. Alliance of Genome Resources Portal: unified model organism research platform. Nucleic Acids Res. 2020 Jan 8; 48(D1):D650–8. https://doi.org/10.1093/nar/gkz813 PMID: 31552413

59. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. Mol Biol Evol. 2016 Jun; 33(6):1635–8. https://doi.org/10.1093/molbev/msw046 PMID: 26921390

60. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009 Jun 1; 25(11):1422–3. https://doi.org/10.1093/bioinformatics/btp163 PMID: 19304878

61. Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, et al. The GOA database: Gene Ontology annotation updates for 2015. Nucleic Acids Res. 2015 Jan; 43(Database issue): D1057–63. https://doi.org/10.1093/nar/gku1113 PMID: 25378336

62. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020 Mar; 17(3):261–72. https://doi.org/10.1038/s41592-019-0686-2 PMID: 32015543

63. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat. 1947 Mar; 18(1):50–60.

64. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng. 2007 May-Jun; 9(3):90–5.

65. Waskom M, Botvinnik O, O'Kane D, Hobson P, Ostblom J, Lukauskas S, et al. mwaskom/seaborn: v0.9.0. Version v0.9.0 [software]. 2018 Jul 16 [cited 2021 Jun 24]. Available from: https://zenodo.org/record/1313201.