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

Jolene Ramsey[a,b,1], Brenley McIntosh[a,1], Daniel Renfro[a,1], Suzanne A. Aleksander[a,1,2], Sandra LaBonte[a,1], Curtis Ross[a], Adrienne E. Zweifel[a], Nathan Liles[a,3], Shabnam Farrar[a], Jason J. Gill[b,c], Ivan Erill[d,e], Sarah Ades[f], Tanya Z. Berardini[g], Jennifer A. Bennett[h], Siobhan Brady[i], Robert Britton[j,4], Seth Carbon[k], Steven M. Caruso[d], Dave Clements[l], Ritu Dalia[m,5], Meredith Defelice[f], Erin L. Doyle[n], Iddo Friedberg[o,6], Susan M.R. Gurney[m], Lee Hughes[p], Allison Johnson[q], Jason M. Kowalski[r,7], Donghui Li[g], Ruth C. Lovering[s], Tamara L. Mans[t,8], Fiona McCarthy[u,9], Sean D. Moore[v], Rebecca Murphy[w], Timothy D. Paustian[x], Sarah Perdue[r,10], Celeste N. Peterson[y], Birgit M. Prüß[z], Margaret S. Saha[aa], Robert R. Sheehy[bb], John T. Tansey[cc], Louise Temple[dd], Alexander William Thorman[ee], Saul Trevino[ff], Amy Cheng Vollmer[gg], Virginia Walbot[hh], Joanne Willey[ii], Deborah A. Siegele[jj], James C. Hu†[a,b]

[a]Department of Biochemistry & Biophysics, Texas A&M University, College Station, TX 77843
[b]Center for Phage Technology, Texas A&M University, College Station, TX 77843
[c]Department of Animal Science, Texas A&M University, College Station, TX 77843
[d]Department of Biological Sciences, University of Maryland Baltimore County, Baltimore, MD 21250
[e]Department of Computer Science and Electrical Engineering, University of Maryland Baltimore County, Baltimore, MD 21250
[f]Department of Biochemistry & Molecular Biology, The Pennsylvania State University, University Park, PA 16801
[g]The Arabidopsis Information Resource, Phoenix Bioinformatics, Newark, CA 94560
[h]Department of Biology and Earth Science, Otterbein University, Westerville, OH 43081
[i]Department of Plant Biology and Genome Center, University of California Davis, Davis, CA 95616
[j]Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48823
[k]Lawrence Berkeley National Laboratory, Berkeley, CA 94720
[l]Department of Biology, John Hopkins University, Baltimore, MD 21218
[m]Department of Biology, Drexel University, Philadelphia, PA 19104
[n]Biology Department, Doane University, Crete, NE 68333
[o]Department of Microbiology, Miami University, Oxford, OH 45056
[p]Department of Biological Sciences, University of North Texas, Denton, TX 76203
[q]Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA 23284
[r]Biological Sciences Department, University of Wisconsin-Parkside, Kenosha, WI 53144
[s]Institute of Cardiovascular Science, University College London, London WC1E 6BT, United Kingdom

43   tDepartment of Biochemistry and Biotechnology, Minnesota State University Moorhead,
44   Brooklyn Park, MN
45   uDepartment of Basic Science, College of Veterinary Medicine, Mississippi State
46   University, Starkville, MS 39762
47   vBurnett School of Biomedical Sciences, University of Central Florida, Orlando, FL
48   32819
49   wDepartment of Biology, Centenary College of Louisiana, Shreveport, LA 71104
50   xDepartment of Bacteriology, University of Wisconsin, Madison, WI 53715
51   yBiology Department, Suffolk University, Boston, MA 02108
52   zMicrobiological Sciences Department, North Dakota State University, Fargo, ND 58105
53   aaDepartment of Biology, College of William & Mary, Williamsburg, VA 23185
54   bbBiology Department, Radford University, Radford, VA 24142
55   ccDepartment of Biochemistry and Molecular Biology, Otterbein University, Westerville,
56   OH 43081
57   ddSchool of Integrated Sciences, James Madison University, Harrisonburg, VA 22807
58   eeDepartment of Environmental and Public Health Sciences, University of Cincinnati,
59   Cincinnati, OH 45221
60   ffDepartment of Chemistry, Math, and Physics, Houston Baptist University, Houston, TX
61   77074
62   ggBiology Department, Swarthmore College, Swarthmore, PA 19081
63   hhDepartment of Biology, Stanford University, Stanford, CA 94305
64   iiDepartment of Science Education, Hofstra University, Hempstead, NY 11549
65   jjDepartment of Biology, Texas A&M University, College Station, TX 77843
66
67   1. J.R., B.M., D.R., S.A.A., S.L. contributed equally to this work.
68   2. Present address: Department of Genetics, Stanford University, Palo Alto, CA 94305
69   3. Present address: Advanced Research Computing, Virginia Polytechnic Institute and
70   State University, Blacksburg, VA 24061
71   4. Present address: Department of Molecular Virology and Microbiology, Baylor
72   College of Medicine, Houston, TX 77030
73   5. Present address: MaST Community Charter School, Philadelphia, PA 19116
74   6. Present address: Department of Veterinary Microbiology and Preventive Medicine,
75   Iowa State University, Ames, IA 50011
76   7. Present address: Department of Math and Natural Sciences, Marian University, Fond
77   du Lac, WI 54935
78   8. Present address: North Hennepin Community College, Brooklyn Park, MN 55445
79   9. Present address: Animal and Comparative Biomedical Sciences, The University of
80   Arizona, Tucson, AZ 85721
81   10. Present address: Department of Physics, University of Wisconsin, Madison, WI
82   53715
83   † Deceased
84

85   * To whom correspondence should be addressed: Deborah A. Siegele or Jolene Ramsey

86   **Email:** siegele@tamu.edu, jolenerr@tamu.edu

87 **Author Contributions:** Assigned according to the CReDIT taxonomy roles: B.M., D.R.,
88 and J.C.H. **Conceptualization**; J.R., B.M., D.R., S.A.A., S.L., C.R., A.E.Z., S.F., D.A.S.,
89 and J.C.H. **Data curation**; B.M., D.R., S.L., C.R. **Formal Analysis**; , D.A.S. and J.C.H.
90 **Funding Acquisition**; J.R., B.M., S.A.A., S.L., A.E.Z., I.E., S.A., T.B., J.A.B., S.B.,
91 R.B., S.C., S.M.C., D.C, R.D., M.D., E.L.D., I.F., S.M.R.G., L.H., A.J., A.M.K., D.L.,
92 R.C.L., T.L.M., F.M., S.D.M., R.M., T.D.P., S.P., C.N.P., B.M.P., M.S.S., R.R.S., J.T.T.,
93 L.T., A.W.T., S.T., A.C.V., V.W., J.W., D.A.S., and J.C.H. **Investigation**; B.M., D.R.,
94 S.A.A., S.L., S.F., and J.C.H. **Methodology**; J.R., B.M., D.R., S.A.A., and J.C.H. **Project
95 administration**; B.M., D.R., S.L., D.A.S., and J.C.H. **Resources**; J.R., B.M., A.E.Z.,
96 S.F., J.J.G., D.A.S., and J.C.H. **Supervision**; D.R., N.L., C.R., and J.C.H. **Software**; J.R.,
97 B.M., D.R., C.R., A.E.Z., S.F., D.A.S., and J.C.H. **Validation**; J.R., B.M., D.R., and C.R.
98 **Visualization**; J.R., B.M., and J.C.H. **Writing – original draft**; J.R., B.M., D.R., S.A.A.,
99 C.R., A.E.Z., I.E., T.B., S.C., S.M.C., A.C.V., D.C., R.D., S.M.R.G., L.H., D.L, R.C.L,
100 T.L.M., S.D.M., C.P., B.M.P., J.T.T., A.W.T., V.W., J.W., D.A.S. **Writing – review &
101 editing**.

102 **Keywords:** data curation, gene ontology, community participation, publication,
103 biomedical research

Abstract

Experimental data about known 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 5000 literature-based annotations. Many of those annotations are to organisms not currently well-represented within GO. Over a ten-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.

Significance Statement

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 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, represents the best return on investment by improving data accessibility across the biological sciences. To supplement professional curation, our CACAO project enabled annotation of the scientific literature by community annotators, in this case undergraduates, which resulted in contribution of thousands of validated entries to public resources. These annotations are now being used by scientists worldwide.

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 databases, 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 three 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's 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 that has kept up with manual and automated annotations, many fields are limited by low numbers of trained personnel and minimal participation, even by 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 successes include crowdsourcing the analysis of the 2011 Shiga-toxin producing *E. 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). Previously, Dr. James Hu was quoted in *Nature* describing the fundamental challenge for community participation in biocuration in terms of the 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 5000 high-quality community annotations added to the GO database. This effort was motivated by the clear parallels

5

185 between the foundational skills used in the professional biocuration field and the well-
186 defined goals for undergraduate training(30). A professional GO biocurator creates gene
187 annotations by finding relevant primary literature, extracting information about normal
188 gene function from it, and entering that information using the controlled GO vocabulary
189 into online databases(31). We hypothesized that university students, guided by their
190 instructors, could accomplish similar tasks and perform community GO annotation while
191 developing strong critical reading skills in a templated annotation task requiring rigorous
192 reading of primary scientific literature. The GONUTS wiki platform
193 (https://gowiki.tamu.edu/) was originally built as a framework for experts not familiar
194 with GO to annotate from literature in their field. We leverage GONUTS to allow student
195 structured GO annotation entry (Fig. 1)(32).
196
197
198 **Results**
199

200 **Sustainable community member contribution via an online intercollegiate**
201 **competition**
202   To address our need for broader participation and expansion beyond model
203 organism databases, we initiated an intercollegiate competition based at Texas A&M
204 University mainly for undergraduate students, called Community Assessment of
205 Community Annotation with Ontologies (CACAO). The specifics of teaching practice are
206 beyond the scope of this report (33). Here, we limit the discussion to details of the
207 competition that are relevant to annotation. Teams of students (competitors) participate in
208 the CACAO competition. Instructors (also called judges) assess all annotations entered by
209 competitors for accuracy and completeness, then give feedback. Peer review by the
210 competitors is incentivized by awarding points for challenges that correct an entry. Teams
211 earn points only for correct annotations and challenges. The team with the highest points
212 accumulated over the competition period wins. Vetted, high-quality annotations are then
213 submitted to the GO Consortium database. CACAO quickly expanded, hosting 39
214 competitions over eight years including 23 colleges and universities, with 792 community
215 annotators and 50 judges. After reading 2879 peer-reviewed journal articles, community
216 members submitted 11,123 annotations to GONUTS (Fig. 1). Many were rejected, usually
217 as unsuitable for GO annotation. Following careful review of each facet for every
218 annotation submitted through online CACAO competitions, 4913 diverse annotations were
219 added to the GO Consortium database after 2018 (Fig. 1). Those annotations are maintained
220 as mandated by updates or changes in the ontology.
221

222 **Annotations generated through CACAO are diverse and novel**
223   The 4913 annotations contributed through GONUTS have spanned all domains of
224 life plus viruses, with the majority being skewed towards eukaryotes, in particular model
225 organisms among the chordates (human, mouse, rat, *etc*.), *Streptophyta* (plants including
226 *Arabidopsis*), and *Ascomycota* (such as budding yeast) (Fig. 2A). As only unique
227 annotations are accepted, this demonstrates that community members can help fill the gaps
228 left by professional biocurators working for model organism databases. Archaea and
229 archaeal viruses are sparsely annotated in GO and represent the smallest fractions within

6

230 our set, with only 24 and six annotations each, respectively. In contrast, 285 eukaryotic
231 viruses are represented, and 45% of the viral annotations cover 384 viruses that infect
232 bacteria (phages include *Siphoviridae*, *Myoviridae*, *Podoviridae*, and *Tectiviridae*). Nearly
233 half of the 1000 annotations listed for bacterial viruses (phages) in QuickGO list CACAO
234 as the source. Annotations for bacterial proteins make up only 5% of total GO annotations,
235 but 30% of CACAO annotations. At the Order level, the top five bacterial categories
236 (*Enterobacterales*, *Bacillales*, *Lactobacillales*, *Pseudomonales*, *Vibrionales*) are heavily
237 studied Gram-negative and Gram-positive organisms of importance to microbiology
238 research and the medical community. The microbial (virus and bacteria) entities herein
239 described represent high genetic diversity and often serve as the basis for significant
240 automated propagation to eukaryotic gene products. Thus, we conclude that not only do
241 CACAO annotators fill gaps for model organisms, but also expand coverage to a wide array
242 of otherwise poorly curated species.
243   Interestingly, GO process terms are used in more than half of the CACAO
244 annotations (Fig. 2B). The top three terms used within each aspect (Process, Component,
245 and Function) are only a small proportion of the total for that branch, an indicator that
246 community members annotate to a wide variety of terms. The cellular component terms
247 are relatively general (nucleus), likely reflecting the ambiguity of experimental methods
248 typically reported in papers. In contrast, the top process and function terms are near leaf-
249 level, having few to no child terms, indicating specific annotations. To better understand
250 the level of detail captured in annotations made by CACAO users, we used GOATOOLS,
251 a python package developed by Klopfenstein *et al.* for representing where terms fall within
252 the ontology hierarchical graph(34). Given the variety of annotation types in our set (*e.g.,*
253 aspects, species), we used a measure that counts the number of descendants (*dcnt*), or child
254 terms, for each entry. Higher level terms will have a larger score and are considered general
255 or global. More descriptive terms with no descendants, or leaf-level terms, are more precise
256 or detailed and receive the lowest *dcnt* value. The *dcnt* analysis quantitatively demonstrates
257 that CACAO annotations are made to specific terms (Fig. 2C). That pattern is consistent
258 with the way annotations were reviewed, where only the most specific term that could be
259 chosen based on the details reported in the paper was counted correct. For comparison, we
260 performed the same *dcnt* analysis on all manual GO annotations available through 2019.
261 The distributions of *dcnt* values for GO annotations are broader, and statistically different
262 from CACAO within each aspect (Fig. 2C). These data demonstrate that community users
263 can contribute high-quality, precise, and scientifically relevant annotations to GO.
264

CACAO community curators enrich ontology development

266   Over time, GO terms and relationships adapt to reflect research progress(35). Small
267 and large-scale rearrangements result from changes in relationships between GO terms to
268 improve the representation of biological knowledge. Regular updates to the ontology are
269 critical for the database to remain relevant and current. The GO Consortium tracks requests
270 to change the ontology as issues via their GitHub repository accessible on the Helpdesk
271 (http://help.geneontology.org/). CACAO users have submitted >50 tickets via this system,
272 resulting in the creation of 49 new GO terms, many of which now have child terms added
273 by others. Given the diverse literature areas read by community curators, many of these
274 new terms are breaking ground in the ontology. At time of writing, the new terms added

275 based on CACAO feedback had been used >650 times by curators. In addition, at least 14
276 non-term changes, such as clarified definitions and relationships for current terms, have
277 also occurred. A beneficial, unintended consequence of CACAO is that curators are
278 compelled to resolve issues within the ontology and incorporate new knowledge from areas
279 that are not traditionally covered by model organism databases.
280
281
282 **Discussion**
283
284 **Community member annotations through CACAO add long-term value to**
285 **GO.** The resources available through the GO ecosystem are among the computational
286 tools most cited by biologists(6). Automatically inferred annotations, those made without
287 curator intervention, are temporary but make up a significant dynamic proportion of the
288 total GO annotations at any given time. However, there are an astonishing >6 million
289 manual GO annotations in the Aug. 2020 release of GO files. The quality of
290 computationally assigned annotations relies on a solid undergirding of manual
291 annotations performed by a dedicated biocuration community(36). The efforts described
292 here are not meant to rival the volume produced by dedicated biocurators, nor are they
293 suggested to replace that organized effort. Instead, we demonstrate how small
294 contributions from many individual community members over time accumulate into a
295 valuable and unique resource. By virtue of its decoupling from the traditional funding
296 model, community curation supplements professional biocuration, especially in under-
297 funded areas(17).
298
299 **Targeted crowd-sourcing with attribution makes CACAO annotation**
300 **sustainable.** Recognizing the need to pull expertise from diverse bench scientists, various
301 other initiatives have been implemented to encourage community participation, including
302 asking non-expert 'crowds' to help correct the ontology with lower cost and similar
303 accuracy to experts(37). Another natural by-product of this crowdsourcing influx is the
304 diversification of the biocuration workforce. Such introduction of new expertise and
305 perspectives, as is so often the case with trainees, is analogous to the workplace
306 observation that diverse teams innovate and produce more than homogenous ones(38).
307 While the majority in the 'crowd' may be unlikely to participate(39), the CACAO
308 implementation of GONUTS is a sustainable model for community contribution of vetted
309 GO annotations in areas of current interest because it caters to a nonrandom crowd,
310 primarily students in an academic course setting.
311 In a resource-limited environment, the need to incentivize data curation has been
312 creatively approached with different methods such as the micropublication format(40–
313 42). The PomBase community curation project took form as an online annotation tool
314 called Canto where researchers can curate their own publications. Canto has garnered up
315 to an impressive 50% response rates for co-annotation from authors within their
316 community(43). Yet, motivating researchers to weigh in on ontology structure is a long-
317 standing challenge(20). Recognizing the need to credit individuals for their annotation
318 efforts, UniProt now offers a portal for submitting literature-based curation linked to an
319 ORCiD (https://community.uniprot.org/bbsub/bbsub.html)(44), as does the new Generic

320 Online Annotation Tool built for the plant community
321 (http://goat.phoenixbioinformatics.org/). Importantly, the GONUTS wiki provides a web-
322 based public record of CACAO contributions on the website, allowing individuals to cite
323 their efforts.
324
325   **CACAO contributions are valuable because they are unique.** As NIH-funded
326 resources for microbes of public health importance are being consolidated into broad
327 bioinformatic resource centers, community investment into annotation through a standard
328 pipeline is warranted(45–48). On the one hand, community curators can spend the time to
329 read and extract information from redundant papers (those with information highly
330 similar to already curated literature and conclusions) thus enhancing eukaryotic model
331 organism annotation depth and increasing confidence in existing annotations. On the
332 other hand, community curators sample from a vast literature space outside the typical
333 biocurator's expertise, expanding overall organism coverage, including in microbial
334 organisms as demonstrated here(49). Because microbial genomes are typically smaller,
335 groups of students can make a major contribution. A significant instance is adding ~50%
336 of all phage GO annotations available in the GO annotation files. CACAO has also
337 spurred updates to ontology relationships. For example, a large rearrangement of biofilm
338 GO terms occurred after CACAO users initiated discussion about their parentage and
339 definitions.
340
341   **Community curation through CACAO meets modern open-source research**
342 **and education goals.** With online education thrust to the forefront in this era of the
343 global COVID-19 pandemic, sustainable and authentic education-driven engagement
344 solutions are critically needed(30, 50, 51). Direct individual contributions, community-
345 driven research, and classroom-focused efforts in any number of formats (*e.g.* CACAO,
346 Adopt-a-genome(52)) have been useful in developing student skills and in serving the
347 scientific community. From an educational perspective, the competition format is an
348 engaging format that models real-world scientific skill development with regards to
349 critical reading, iterative editing of a product, and peer review. We hypothesize that this
350 mini biocurator experience may have similar benefits with regards to recruitment,
351 retention, and graduation observed with undergraduate research (53, 54). The biocuration
352 model is highly applicable to scientists and trainees worldwide and complies with FAIR
353 (Findable, Accessible, Interoperable, Reusable(55)) data principles, making its results
354 accessible to all. GO annotation for SARS-CoV-2 and its infection of human cells was
355 immediately pursued to aid strategic planning of the pandemic response
356 (http://geneontology.org/covid-19.html). We appeal to scientists to participate in
357 biocuration efforts through GONUTS, UniProt, or a model organism database/the
358 Alliance of Genome Resources where users can contribute from the comfort of any
359 computer(56).
360

Materials and Methods

361
362

363   CACAO competitions for intercollegiate teams are hosted on GONUTS
364 (https://gonuts.tamu.edu). Raw data for all users and every annotation history are
365 maintained by custom extensions to the MediaWiki software used by GONUTS(32).
366 Additional information about competition rules can be found at
367 https://gowiki.tamu.edu/wiki/index.php/Category:CACAO. The data presented here
368 encompass annotations generated from 2010-2018, with expanded taxon information
369 retrieved using the UniProt application programming interface (API) as well as the ETE
370 (v3.1.1) module and various tools from BioPython (v1.74) (57, 58). Summary statistics
371 for CACAO annotations given in Fig. 1 were mined from our local database storage.
372   Fully correct annotation data are transferred from GONUTS regularly via the
373 current Gene Association File (GAF) or Gene Product Association Data (GPAD) file
374 format, as outlined in GO requirements, directly to the European Bioinformatics
375 Institute's Protein2GO for incorporation into the complete GO annotation files. All
376 currently included annotations are accessible on GONUTS or via the search engine
377 QuickGO (https://www.ebi.ac.uk/QuickGO/annotations) by filtering for parameter
378 "assigned by" CACAO, and are also provided as a supplementary dataset in GPAD
379 format (Supp Dataset 1) (59).
380   The 01-01-2020 non IEA GAF (goa_uniprot_all_noiea.gaf.gz) and ontology file
381 (go.obo) were downloaded from http://release.geneontology.org/ for the *dcnt* analysis.
382 Values for *dcnt* were calculated according to GOATOOLS on all manual annotations not
383 assigned by CACAO(34). The Mann-Whitney test with a two-sided p-value was used to
384 compare GO and CACAO *dcnt* distributions within each aspect using SciPy(60, 61).
385   For the phage analyses, the GAF was filtered into a subset using the following
386 TaxIDs from the NCBI Taxonomy browser: 12333 (unclassified bacterial viruses),
387 1714267 (*Gammasphaerolipovirus*), 10656 (*Tectiviridae*), 10472 (*Plasmaviridae*), 10659
388 (*Corticoviridae*), 10841 (*Microviridae*), 10860 (*Inoviridae*), 28883 (Caudovirales),
389 11989 (*Leviviridae*), and 10877 (*Cystoviridae*).
390   Changes to the ontology initiated by CACAO users were tallied by searching
391 through the GO issue tracker at GitHub (https://github.com/geneontology/go-
392 ontology/issues) for user handles: @jimhu-tamu, @suzialeksander, @sandyl27, @jrr-cpt,
393 @ivanerill, and/or the query text "CACAO" for open and closed issues, then manually
394 reviewed for accuracy. The final list of GO terms used to calculate the annotations is
395 included as [supplemental file 2]. Matplotlib (v3.1.1) and Seaborn (v0.9.0) were used to
396 generate pie charts, box plots, and bar graphs (62, 63). Figures were compiled and
397 rendered with the open-source program Inkscape 0.92.2.

Acknowledgments

References

1. N. Skunca, A. Altenhoff, C. Dessimoz, Quality of computationally inferred gene ontology annotations. *PloS Comput Biol* **8**, e1002533 (2012).

2. L.-L. Sun, *et al.*, Global analysis of fission yeast mating genes reveals new autophagy factors. *PLoS Genetics* **9**, e1003715 (2013).

3. P. Denny, *et al.*, Exploring autophagy with Gene Ontology. *Autophagy* **14**, 419–436 (2018).

4. M. E. Lee, S. F. Rusin, N. Jenkins, A. N. Kettenbach, J. B. Moseley, Mechanisms connecting the conserved protein kinases Ssp1, Kin1, and Pom1 in fission yeast cell polarity and division. *Curr Biol* **28**, 84-92.e4 (2018).

5. M. Ashburner, *et al.*, Gene Ontology: tool for the unification of biology. *Nat Genet* **25**, 25–29 (2000).

6. T. G. O. Consortium, *et al.*, The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res* **49**, D325–D334 (2020).

7. V. K. Khodiyar, D. Howe, P. J. Talmud, R. Breckenridge, R. C. Lovering, From zebrafish heart jogging genes to mouse and human orthologs: using Gene Ontology to investigate mammalian heart development. *F1000Res* **2**, 242 (2013).

8. C. Zhang, W. Zheng, P. L. Freddolino, Y. Zhang, MetaGO: predicting Gene Ontology of non-homologous proteins through low-resolution protein structure prediction and protein-protein network mapping. *J Mol Biol* **430**, 2256–2265 (2018).

9. C. Zhang, X. Wei, G. S. Omenn, Y. Zhang, Structure and protein interaction-based Gene Ontology annotations reveal likely functions of uncharacterized proteins on human chromosome 17. *J Proteome Res* **17**, 4186–4196 (2018).

10. C. Zhang, P. L. Freddolino, Y. Zhang, COFACTOR: improved protein function prediction by combining structure, sequence and protein-protein interaction information. *Nucleic Acids Res* **45**, W291–W299 (2017).

11. R. C. Lovering, *et al.*, Improving interpretation of cardiac phenotypes and enhancing discovery with expanded knowledge in the Gene Ontology. *Circ Genom Precis Med* **11**, e001813 (2018).

12. R. J. Andrew, K. Fisher, K. J. Heesom, K. A. B. Kellett, N. M. Hooper, Quantitative interaction proteomics reveals differences in the interactomes of amyloid precursor protein isoforms. *J Neurochem* **161**, 41 (2019).

454   13. D. Li, T. Z. Berardini, R. J. Muller, E. Huala, Building an efficient curation workflow
455   for the Arabidopsis literature corpus. *Database (Oxford)* **2012**, bas047 (2012).

456   14. H. J. Drabkin, J. A. Blake, M. G. I. Database, Manual Gene Ontology annotation
457   workflow at the Mouse Genome Informatics Database. *Database (Oxford)* **2012**, bas045–
458   bas045 (2012).

459   15. R. Bastow, S. Leonelli, Sustainable digital infrastructure. *EMBO Rep* **11**, 730–734
460   (2010).

461   16. C. Dessimoz, N. Škunca, Eds., *The Gene Ontology Handbook* (Springer New York,
462   2017).

463   17. S. Poux, *et al.*, On expert curation and scalability: UniProtKB/Swiss-Prot as a case
464   study. *Bioinformatics* **33**, 3454–3460 (2017).

465   18. H. Tang, R. D. Finn, P. D. Thomas, TreeGrafter: phylogenetic tree-based annotation
466   of proteins with Gene Ontology terms and other annotations. *Bioinformatics* **35**, 518–520
467   (2018).

468   19. P. Gaudet, M. S. Livstone, S. E. Lewis, P. D. Thomas, Phylogenetic-based
469   propagation of functional annotations within the Gene Ontology consortium. *Brief
470   Bioinform* **12**, 449–462 (2011).

471   20. E. Ong, Y. He, Community-based ontology development, annotation and discussion
472   with MediaWiki extension Ontokiwi and Ontokiwi-based Ontobedia. *AMIA Jt Summits
473   Transl Sci Proc* **2016**, 65–74 (2016).

474   21. I. S. for Biocuration, Biocuration: Distilling data into knowledge. *PLoS Biol* **16**,
475   e2002846 (2018).

476   22. H. Rohde, *et al.*, Open-source genomic analysis of shiga-toxin–producing E. coli
477   O104:H4. *New Engl J Med* **365**, 718–724 (2011).

478   23. F. C. Group, *et al.*, Crystal structure of a monomeric retroviral protease solved by
479   protein folding game players. *Nat Struct Mol Biol* **18**, 1175–1177 (2011).

480   24. J. Giles, Internet encyclopaedias go head to head. *Nature* **438**, 900–901 (2005).

481   25. B. M. Good, E. L. Clarke, L. de Alfaro, A. I. Su, The Gene Wiki in 2011: community
482   intelligence applied to human gene annotation. *Nucleic Acids Res* **40**, D1255-61 (2011).

483   26. N. J. Reavley, *et al.*, Quality of information sources about mental disorders: a
484   comparison of Wikipedia with centrally controlled web and printed sources. *Psychol Med*
485   **42**, 1753–1762 (2011).

13

27. W. Arroyo-Machado, D. Torres-Salinas, E. Herrera-Viedma, E. Romero-Frías, Science through Wikipedia: A novel representation of open knowledge through co-citation networks. *PloS One* **15**, e0228713 (2020).

28. B. Mons, *et al.*, Calling on a million minds for community annotation in WikiProteins. *Genome Biol* **9**, R89 (2008).

29. E. Callaway, No rest for the bio-wikis. *Nature* **468**, 359–60 (2010).

30. C. Bauerle, *et al.*, *Vision and Change in Undergraduate Biology Education: A Call to Action*, C. Brewer, D. Smith, Eds. (AAAS, 2011).

31. R. Balakrishnan, M. A. Harris, R. Huntley, K. V. Auken, J. M. Cherry, A guide to best practices for Gene Ontology (GO) manual annotation. *Database* **2013**, bat054 (2013).

32. D. P. Renfro, B. K. McIntosh, A. Venkatraman, D. A. Siegele, J. C. Hu, GONUTS: the Gene Ontology Normal Usage Tracking System. *Nucleic Acids Res* **40**, D1262-9 (2012).

33. I. Erill, S. Caruso, J. C. Hu, "Gamifying Critical Reading through a Genome Annotation Intercollegiate Competition". *Tested Studies in Laboratory Teaching* **39**, 1-18 (2018).

34. D. V. Klopfenstein, *et al.*, GOATOOLS: A Python library for Gene Ontology analyses. *Scientific reports* **8**, 10872 (2018).

35. S. Leonelli, A. D. Diehl, K. R. Christie, M. A. Harris, J. Lomax, How the gene ontology evolves. *Bmc Bioinformatics* **12**, 325 (2011).

36. The Gene Ontology Consortium, The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* **47**, D330–D338 (2019).

37. J. M. Mortensen, *et al.*, Using the wisdom of the crowds to find critical errors in biomedical ontologies: a study of SNOMED CT. *J Am Medical Informatics Assoc* **22**, 640–8 (2014).

38. T. H. Swartz, A.-G. S. Palermo, S. K. Masur, J. A. Aberg, The science and value of diversity: closing the gaps in our understanding of inclusion and diversity. *J Infect Dis* **220**, S33–S41 (2019).

39. P. D. Karp, Crowd-sourcing and author submission as alternatives to professional curation. *Database* **2016**, baw149 (2016).

14

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.30.440339; this version posted May 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a

CC-BY 4.0 International license.

40. D. Raciti, K. Yook, T. W. Harris, T. Schedl, P. W. Sternberg, Micropublication: incentivizing community curation and placing unpublished data into the public domain. *Database* **2018**, bay013- (2018).

41. P. E. Bourne, J. R. Lorsch, E. D. Green, Perspective: Sustaining the big-data ecosystem. *Nature* **527**, S16–S17 (2015).

42. P. D. Karp, How much does curation cost? *Database* **2016**, baw110 (2016).

43. A. Lock, M. A. Harris, K. Rutherford, J. Hayles, V. Wood, Community curation in PomBase: enabling fission yeast experts to provide detailed, standardized, sharable annotation from research publications. *Database* **2020** (2020).

44. U. Consortium, UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**, D506–D515 (2019).

45. C. Aurrecoechea, *et al.*, EuPathDB: the eukaryotic pathogen genomics database resource. *Nucleic Acids Res* **45**, D581–D591 (2016).

46. A. R. Wattam, *et al.*, Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Res* **45**, D535–D542 (2016).

47. B. E. Pickett, *et al.*, ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res* **40**, D593-8 (2011).

48. Y. Zhang, *et al.*, Influenza Research Database: An integrated bioinformatics resource for influenza virus research. *Nucleic Acids Res* **45**, D466–D474 (2016).

49. M. G. Giglio, C. W. Collmer, J. Lomax, A. Ireland, Applying the Gene Ontology in microbial annotation. *Trends Microbiol* **17**, 262–8 (2009).

50. S. G. Hoskins, D. Lopatto, L. M. Stevens, 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* **10**, 368–378 (2011).

51. J. E. Round, A. M. Campbell, Figure facts: encouraging undergraduates to take a data-centered approach to reading primary literature. *CBE Life Sci Educ* **12**, 39–46 (2013).

52. C. A. Kerfeld, The Joint Genome Institute's microbial genome annotation program for undergraduates. *FASEB J* **23**, 84.2-84.2 (2009).

53. T. C. Jordan, *et al.*, A broadly implementable research course in phage discovery and genomics for first-year undergraduate students. *mBio* **5**, e01051-13 (2014).

549   54. D. I. Hanauer, *et al.*, An inclusive Research Education Community (iREC): Impact of
550   the SEA-PHAGES program on research outcomes and student learning. *Proc National*
551   *Acad Sci* **114**, 13531–13536 (2017).

552   55. M. D. Wilkinson, *et al.*, The FAIR Guiding Principles for scientific data management
553   and stewardship. *Scientific data* **3**, 160018 (2016).

554   56. T. A. of G. R. Consortium, *et al.*, Alliance of Genome Resources Portal: unified
555   model organism research platform. *Nucleic Acids Res* **48**, D650–D658 (2019).

556   57. J. Huerta-Cepas, F. Serra, P. Bork, ETE 3: Reconstruction, Analysis, and
557   Visualization of Phylogenomic Data. *Mol Biol Evol* **33**, 1635–8 (2016).

558   58. P. J. A. Cock, *et al.*, Biopython: freely available Python tools for computational
559   molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

560   59. R. P. Huntley, *et al.*, The GOA database: Gene Ontology annotation updates for 2015.
561   *Nucleic Acids Res* **43**, D1057–D1063 (2015).

562   60. P. Virtanen, *et al.*, SciPy 1.0: fundamental algorithms for scientific computing in
563   Python. *Nat Methods* **17**, 261–272 (2020).

564   61. H. B. Mann, D. R. Whitney, On a Test of Whether one of Two Random Variables is
565   Stochastically Larger than the Other. *Ann Math Statistics* **18**, 50–60 (1947).
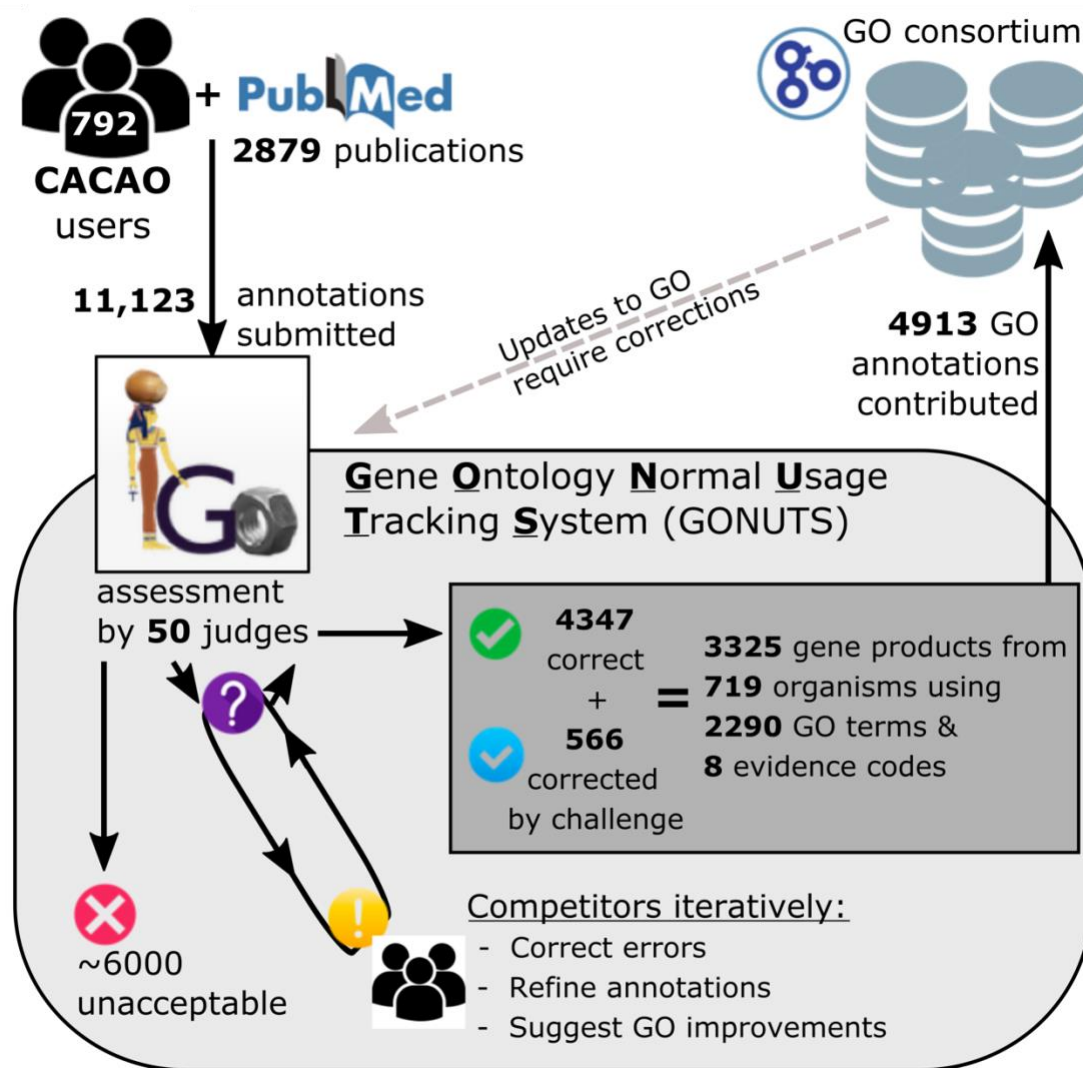
566   62. J. D. Hunter, Matplotlib: A 2D Graphics Environment. *Comput Sci Eng* **9**, 90–95
567   (2007).

568   63. M. Waskom, *et al.*, mwaskom/seaborn: v0.9.0 (July 2018) (Version v0.9.0) (2018).

569

570

571

572
Figure 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 Gene Ontology. 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.
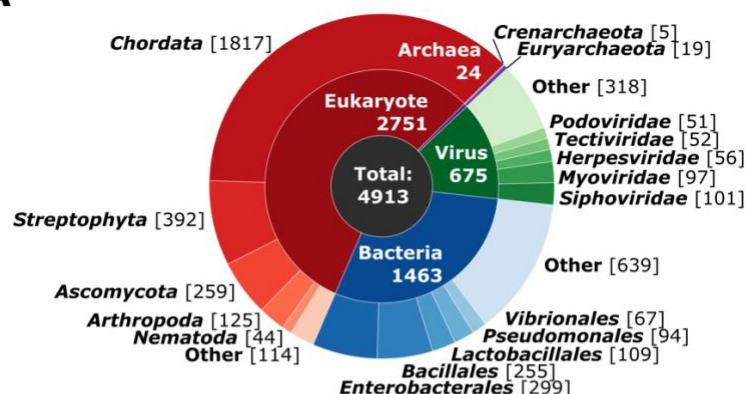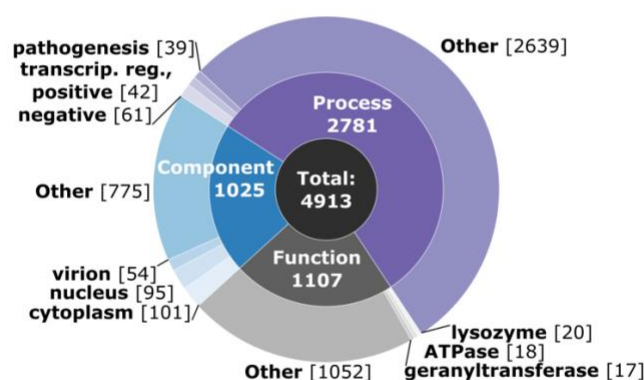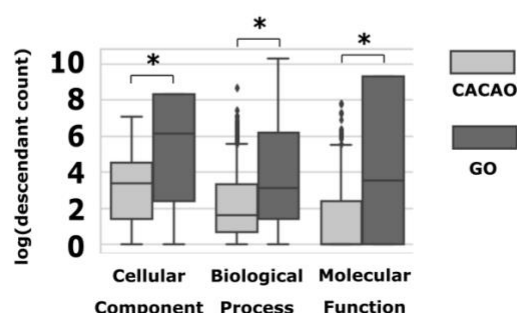
584
585
586 **Figure 2: The GO annotations contributed by CACAO users are diverse and specific.** A)
587 Proteins annotated by CACAO users are depicted by species domain. The organisms most highly
588 represented in each domain are displayed on the outer ring of the chart divided by the following
589 rank: Phylum for eukaryotes and archaea, Order for bacteria, and Family for viruses. The number
590 of GO annotations in each category is indicated in brackets. B) The distribution of GO terms used
591 for CACAO annotations are graphed by aspect within the ontology. The top three terms within
592 each aspect are labeled on the outer ring. For clarity, "activity" was dropped from each function
593 term, and the process terms were abbreviated from "positive/negative regulation of transcription,
594 DNA-templated" to "transcript. reg., positive or negative". The number of GO annotations for
595 each term is indicated in brackets. C) The descendant counts, corresponding to depth within the
596 ontology, for CACAO annotations (n = 4913) and all other manual GO annotations through 2019
597 (n = 255,958) are graphed. Significant differences measured by the Mann-Whitney test with
598 p<0.001 are marked with an *.