Title: Liberating host-virus knowledge from COVID-19 lockdown

Authors: Nathan S. Upham[1,*], Jorrit H. Poelen[2], Deborah Paul[3], Quentin J. Groom[4], Nancy B. Simmons[5], Maarten P. M. Vanhove[6], Sandro Bertolino[7], DeeAnn M. Reeder[8], Cristiane Bastos-Silveira[9], Atriya Sen[1], Beckett Sterner[1], Nico Franz[1], Marcus Guidoti[10], Lyubomir Penev[11], and Donat Agosti[12]

Affiliations:

[1]School of Life Sciences, Arizona State University, Tempe, AZ 85782, USA.

[2]Ronin Institute for Independent Scholarship, Montclair, NJ 07043, USA.

[3]Illinois Natural History Survey, University of Illinois, Urbana-Champaign, IL 61820, USA.

[4]Meise Botanic Garden, 1860 Meise, Belgium.

[5]Department of Mammalogy, Division of Vertebrate Zoology, American Museum of Natural History, New York, NY 10024, USA.

[6]Research Group Zoology: Biodiversity and Toxicology, Centre for Environmental Sciences, Hasselt University, 3590 Diepenbeek, Belgium.

[7]Department of Life Sciences and Systems Biology, University of Turin, 10123 Torino, Italy.

[8]Department of Biology, Bucknell University, Lewisburg, PA 17837, USA.

[9]Museu Nacional de História Natural e da Ciência & Centre for Ecology, Evolution and Environmental Changes (CE3C), Universidade de Lisboa, Lisbon, Portugal.

[10]Plazi, Porto Alegre, Brazil.

[11]Pensoft Publishers, Sofia, Bulgaria.

[12]Plazi, Bern, Switzerland.

*Correspondence to: nathan.upham@asu.edu

Abstract: 186 words; Main text (with figure legends): 2035 words; References: 23 total.

Abstract:

Connecting basic data about bats and other potential mammal hosts of SARS-CoV-2 with their ecological context is now critical for understanding the emergence and spread of COVID-19. However, when global lockdown started in March 2020, the world's bat experts were locked out of their research laboratories, which, in turn, locked up large volumes of offline ecological and taxonomic data. Global lockdown has put a magnifying glass on the long-standing problem of biological 'dark data' that is published but disconnected from digital knowledge resources, and thus unavailable for high-throughput analysis. Host-to-virus knowledge will be biased until this challenge is addressed. Here we outline two viable solutions: (i) how to interconnect published data about mammal hosts and viruses in the short term; and (ii) how to shift the publishing paradigm beyond unstructured text (e.g., PDFs) to labeled networks of digital knowledge. Biological taxonomy is foundational to both solutions as the indexing system for biodiversity data. Building digitally connected 'knowledge graphs' of host-virus interactions will establish the needed agility for quickly identifying reservoir hosts of novel zoonoses, allow for AI-based predictions of emergence, and thereby improve planetary health security.

Main text:

An irony of COVID-19 likely originating from a bat-borne coronavirus [1] is that the global lockdown to quell the pandemic also locked up physical access to much-needed knowledge about bats. Basic data about bat diversity, ecology, and geography, as well as that of other potential mammal hosts [1,2], was suddenly critical for understanding SARS-CoV-2's emergence and spread. However, with the world's bat experts locked out of their research laboratories, any undigitized or offline data was also locked up. For the first time, many scientists found themselves isolated from knowledge. Why, in this digitally connected age, was basic knowledge about species and their ecological interactions not already digitized, online, and freely accessible to all? What must be done to improve global access to public health-related biodiversity knowledge?

Understanding why biodiversity science was unprepared—and how to fix it before the next crisis—has been a hot topic, spawning multiple taskforces in the biodiversity research community since the pandemic began (e.g., [3–5]). Of key interest has been mending the chasm in knowledge transfer from the physical biocollections, which contain the preserved specimens, tissues, and associated material used to describe biodiversity, to biomedical scientists in health-related fields like infectious disease, epidemiology, and virology. Most biodiversity knowledge ever published remains effectively locked in textual, unstructured articles, and is thus isolated from efforts to synthesize global ecological interactions. These data are 'known' in publications but are digitally disconnected—revealing a striking 'knowledge frontier' that is preventing scientists from around the world from digitally discovering their existence. With human activities like land conversion spurring the emergence of zoonoses [6], it is increasingly urgent to build interconnected networks of digital knowledge.

ILLUMINATING BIODIVERSITY DARK DATA

Similar to how physicists believe that dark matter exists but have difficulty measuring it, biodiversity scientists cannot easily make use of 'dark data' that is published but either old and rare (e.g., inside archival or unpublished literature) or new and locked (e.g., behind paywalls, in digitally unreadable formats, or unlinked to other data). Traditionally, a particular research project might manually synthesize information from hundreds or thousands of articles in disparate formats over the course of years, yielding a comprehensive 'snapshot' of written knowledge. Still today, gathering the widely scattered biodiversity data relevant to mammal host-virus interactions would take years instead of the needed weeks for responding to a crisis like SARS-CoV-2. Remarkably, new publications continue increasing the volume of dark data, since the ubiquitous 'portable data format' (PDF) requires substantial efforts to make ecological phenomena like host-virus

interactions extractable for re-use. To address deeply interconnected global problems like COVID-19, we need new solutions rooted in building expansive digital knowledge (Fig. 1).

For data to form digital knowledge, it must first be published in datasets that are open access and FAIR — Findable on the web, digitally Accessible, Interoperable among different computing systems, and thus Reusable for later analyses. Satisfying all of those criteria opens the door for creating highly useful 'knowledge graphs' [7], in which digital open data are meaningfully linked together on massive scales, forming knowledge that is collectively greater than its sum. As Tim Berners-Lee presciently wrote in 2006, "it is the unexpected re-use of information which is the value added by the web" [8]. Illuminating the zoonotic origins of COVID-19 is exactly the kind of unexpected re-use of data that biodiversity science was ill-prepared to address at the start of the pandemic. Building a comprehensive host-virus knowledge graph will furthermore enable rapidly improving AI algorithms (e.g., in the field of natural language processing, NLP [9]) to flexibly learn from the structure of this newly digital knowledge.

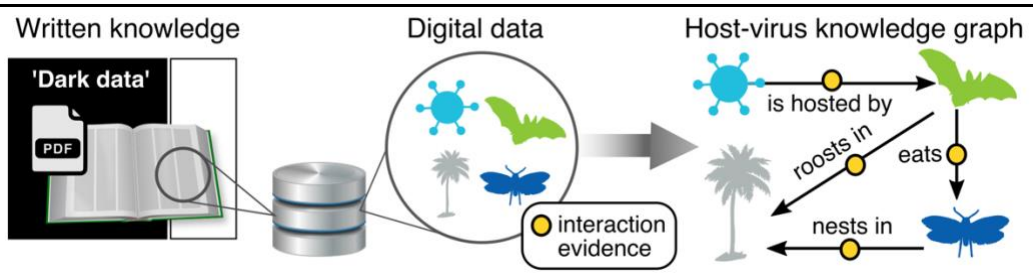

Fig. 1. The evolution of biodiversity knowledge from analog to digital. Extracting written knowledge from publications into databases is only the first step toward creating forms of digital, structured knowledge in which ecological interactions and evidence thereof are additionally annotated. Such 'knowledge graphs' include levels of confidence in each annotation as derived from evidence sources, which enable high-throughput integrative modeling of complex ecological dynamics like viral spillover. Much written knowledge is undigitized and digitally disconnected, and so is 'dark data' from the perspective of synthetic knowledge graphs.

TAXONOMY AS THE KEY TO HOST-VIRUS KNOWLEDGE

Linking viruses to animal hosts, hosts to environments, and hosts to other hosts is the raw material needed to build a host-virus knowledge graph (Fig. 2A). However, meaningfully connecting host species, viral species, and their ecological traits requires mastery of a fundamental but often overlooked discipline: biological taxonomy. For at least three centuries, mainstream science has used the names of species—most often the 'genus & species' pair of Linnaean taxonomy—to index research findings. Virtually all observations about organismal behaviors and functions, habitats, genomics, and pathogens are linked to species names via publications called 'taxonomic treatments,' in which authors describe the boundaries of species (and other taxa) based on physical evidence. Because that evidence—especially from preserved specimens but also their tissues, parasites, and derived data like DNA sequences—has improved along with the science of taxonomy through time, multiple names may have been used to refer to similar sets of organisms. Thus, making sense of biodiversity data requires keeping track of how the validity of taxonomic names has changed historically (e.g., synonyms, varying name usages).

How species names have been used by different authors over time is the 'taxonomic key' to finding otherwise hidden virus-to-host interactions in publications. By linking species names, evidence, and taxonomic treatments through time, it is possible to create 'taxonomic intelligence' services [10] that allow for flexible conversion of named species data across taxonomies. For example, SARS-like coronaviruses observed in horseshoe bats identified as *Rhinolophus sinicus* in 2013 [11] need to be resolved relative to the 2019 re-classification of portions of this species as *R. thomasi* and *R. rouxii* [12]. However, updating the taxonomy of named data is not yet possible aside from manually on small scales. Existing taxonomic infrastructures like the Catalogue of Life have not prioritized building scalable solutions to this problem, in large part because species name changes are often very rapid. Even in a relatively well known group like mammals, the global number of species recognized has changed by >40% in the last 25 years

[13] over which time the number of described viruses has increased by a staggering 400% [14]. Keeping track of mammal-to-virus interactions relative to that taxonomic flux has not been incentivized in proportion to its importance to understanding zoonotic emergence. Therefore, we must make efforts to prioritize the building taxonomic intelligence services, which will then enable the extraction and meaningful linking of named host-virus interaction data on planetary scales.

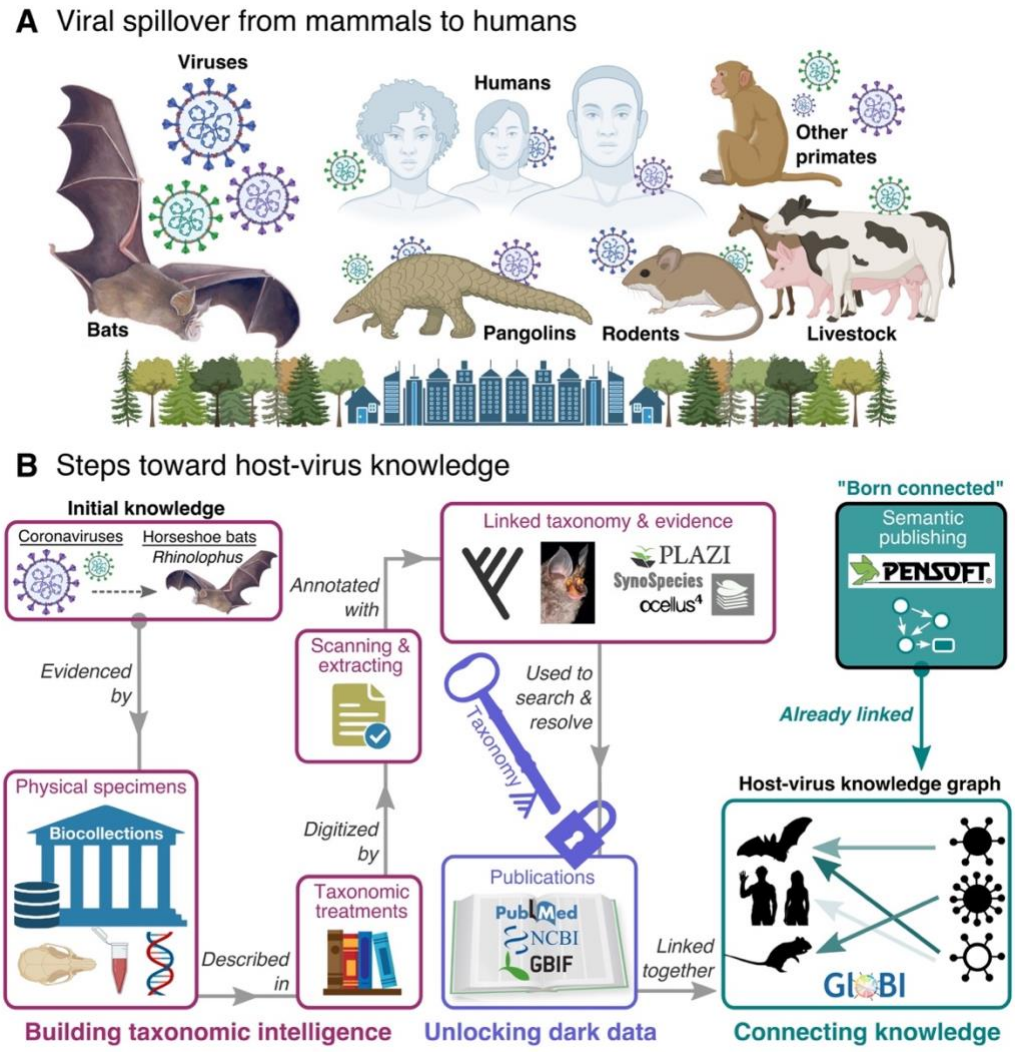

Fig. 2. Connecting digital knowledge of host-virus interactions. (A) The sharing of viruses among humans and other mammals is remarkably common, yet the ecological circumstances under which spillover occurs are poorly understood. (B) Digitally liberating ecological knowledge from locked publications requires building taxonomic intelligence—i.e., how and why species names have been used through time—and then using that taxonomic "key" to liberate and connect

"dark" interaction data hidden in publications. Alternatively, data can be "born connected" if new articles are published using computer-readable (semantic) tags for ecological interactions like "has host" or "pathogen of." Both pathways result enable newly comprehensive "knowledge graphs" connecting host-virus interactions and evidence.

TOWARD A HOST-VIRUS KNOWLEDGE GRAPH

Thankfully, nearly two decades of work in the digital knowledge arena (e.g., [7,8,10,15,16]) has established foundations for a two-pronged approach to building host-virus knowledge (Fig. 2B). First, dark data need to be liberated from existing publications. These efforts are being led by Plazi [15]—a pioneering platform for literature digitization, extraction, and linking—to create new flows of digital data from printed books, archives, and otherwise locked publications [17]. For example, the Plazi services Synospecies and Ocellus have recently indexed taxonomic names and images, respectively, from taxonomic treatments spanning from Linnaeus' initial 1758 publication *Systema Naturae* to the recent *Handbook of the Mammals of the World* series [18]. Once digitally indexed, taxonomic data can be annotated and connected to biocollection-based evidence to formally align taxonomic names with their biological meanings. This liberated taxonomic knowledge allows for more robust literature searches and subsequent name translation of host-virus interaction data. Such efforts have already discovered reliable data on 1,146 host-virus interactions from selected publications (Coronavirus-Host Community on Zenodo). Second, new articles need to be published without creating more dark data. Exemplary in this areas are efforts being led by Pensoft—publisher of biodiversity journals such as *ZooKeys*—to index computer readable terms called 'semantic metadata' [19] during the normal publishing process. For example, Pensoft responded to COVID-19 by beginning to index parts of speech such as "has host" and "pathogen of" to assist with mining biotic interactions from article texts and tables, which has netted over 2,000 biotic interactions now annotated as article metadata [20]. Such digital enhancements greatly streamline the process of data extraction, because new articles already

contain digital text, linking terms, and thus a native form of digital knowledge. These data are 'born connected' relative to the post-processing steps needed with traditional PDF publishing.

To build a singular host-virus knowledge graph requires a central hub for discovering relevant data, resolving disparate taxonomies, and connecting the resulting insights. Promising progress by the Global Biotic Interactions database GloBI—an open-access ecological network across all of life [16]—has led to new pipelines for ecological data to flow from sources of both 'old' [17,18] and 'new' literature [20]. From April to October 2020 alone, these pipelines resulted in adding >53,000 host-virus data points to GloBI (see dataset on Zenodo [21]). These associations involve 19% more valid species of mammals than were identified in a recent host-virus synthesis (897 vs. 754 species in [22]). Such a dramatic initial effort illustrates the potential for broad-scale data linking to yield new insights. Yet these are small steps relative to what could eventually comprise a comprehensive and taxonomically nimble graph of host-virus knowledge. What interconnected phenomena might be illuminated when such knowledge is freely available to the world's scientists and public health specialists?

BEYOND THE PDF: KNOWLEDGE THAT IS BORN CONNECTED

We have outlined ways to interconnect, and thus liberate, previously 'dark' host-virus interactions from publications. However, doing so is expensive and so is infeasible at scale if publishers continue to publish under the same paradigm. Therefore, we recommend three immediate policy changes: (i) major journals should switch to publishing formats that are not only FAIR, but also semantically tagged with terms relevant to broad-scale ecological interactions (especially host-symbiont and host-host relationships); (ii) academic institutions should incentivize (e.g., via tenure evaluations) publishing in such open access and semantically aware journals; and (iii) investments in data generation should be balanced with infrastructure enhancements for data reuse toward making increasingly complete biodiversity knowledge graphs. Taxonomists,

ecologists, data scientists, and policymakers have essential roles to play in this paradigm shift toward digital knowledge.

The value-added by digitally connected knowledge is tremendous, both for its potential to build nonlinear insights and to expand the capacity of biodiversity researchers around the world, especially in the Global South [23]. Limitations to accessing biodiversity information in developing countries can be diverse and sometimes seemingly trivial, including gaps in geographical knowledge; lack of data sharing among and between scientists and policy-makers; inaccessible presentations of information; and limited financial resources. Efforts are hence needed not only to increase, as is often called for, biodiversity monitoring, but also to develop the capacity of local scientific and citizen communities to mobilize the resulting data into digital knowledge infrastructures.

We cannot continue to waste resources to rediscover biodiversity a second time. Unprecedented reliability and completeness of knowledge about biological interactions are now required to address multiple socio-ecological challenges, from COVID-19 to runaway climate change, each of which exists on scales too massive and too detailed for any one individual to observe alone. COVID-19 lockdown should teach us that siloed science does not serve society as well as its alternative. Multiple solutions, including vaccines and treatments, will eventually arrive to free us from this pandemic. However, the solution for our limited ecological knowledge is already here. We already have much of the technology needed to liberate and connect biodiversity data across the entire tree of life—what is most lacking is the collective will to do so.

REFERENCES

1. Boni MF, Lemey P, Jiang X, Lam TT-Y, Perry BW, Castoe TA, et al. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat Microbiol. 2020; 1–10. doi:10.1038/s41564-020-0771-4

2. Xia X. Extreme Genomic CpG Deficiency in SARS-CoV-2 and Evasion of Host Antiviral Defense. Mol Biol Evol. 2020;37: 2699–2705. doi:10.1093/molbev/msaa094

3. CETAF-DiSSCo COVID-19 Taskforce. Communities Taking Action | CETAF – Consortium of European Taxonomic Facilities | Distributed Systems of Scientific Collections – DiSSCo. 2020 [cited 31 Aug 2020]. Available: https://cetaf.org/covid19-taf-communities-taking-action

4. ViralMuse. iDigBio Wiki- ViralMuse Task Force. In: iDigBio Wiki [Internet]. 2020 [cited 30 Oct 2020]. Available: https://www.idigbio.org/wiki/index.php/ViralMuse_Task_Force

5. Research Data Alliance. RDA-COVID19. In: RDA [Internet]. 2020 [cited 30 Oct 2020]. Available: https://www.rd-alliance.org/groups/rda-covid19

6. Faust CL, McCallum HI, Bloomfield LSP, Gottdenker NL, Gillespie TR, Torney CJ, et al. Pathogen spillover during land conversion. Ecol Lett. 2018;21: 471–483. doi:10.1111/ele.12904

7. Penev L, Dimitrova M, Senderov V, Zhelezov G, Georgiev T, Stoev P, et al. OpenBiodiv: A Knowledge Graph for Literature-Extracted Linked Open Data in Biodiversity Science. Publications. 2019;7: 38. doi:10.3390/publications7020038

8. Berners-Lee T. Linked Data - Design Issues. 2006 [cited 14 Sep 2020]. Available: https://www.w3.org/DesignIssues/LinkedData.html

9. Burgdorf A, Pomp A, Meisen T. Towards NLP-supported Semantic Data Management. ArXiv200506916 Cs. 2020 [cited 4 Nov 2020]. Available: http://arxiv.org/abs/2005.06916

10. Page RDM. Taxonomic names, metadata, and the Semantic Web. Biodivers Inform. 2006;3. doi:10.17161/bi.v3i0.25

11. Ge X-Y, Li J-L, Yang X-L, Chmura AA, Zhu G, Epstein JH, et al. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. Nature. 2013;503: 535–538. doi:10.1038/nature12711

12. Burgin CJ. Rhinolophus sinicus K. Andersen 1905. 2019 [cited 12 Nov 2020]. doi:10.5281/ZENODO.3808964

13. Burgin CJ, Colella JP, Kahn PL, Upham NS. How many species of mammals are there? J Mammal. 2018;99: 1–14. doi:10.1093/jmammal/gyx147

14. International Committee on Taxonomy of Viruses. ICTV Historical taxonomy releases. 2020 [cited 3 May 2020]. Available: https://talk.ictvonline.org/taxonomy/p/taxonomy_releases

15. Agosti D, Egloff W. Taxonomic information exchange and copyright: the Plazi approach. BMC Res Notes. 2009;2: 53. doi:10.1186/1756-0500-2-53

16. Poelen JH, Simons JD, Mungall CJ. Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. Ecol Inform. 2014;24: 148–159. doi:10.1016/j.ecoinf.2014.08.005

17. Agosti D, Catapano T, Sautter G, Kishor P, Nielsen L, Ioannidis-Pantopikos A, et al. Biodiversity Literature Repository (BLR), a repository for FAIR data and publications. Biodivers Inf Sci Stand. 2019;3: e37197. doi:10.3897/biss.3.37197

18. Agosti D. Time for an interim review of Plazi's Covid-19 related activities. 2020. Available: http://plazi.org/news/beitrag/time-for-an-interim-review-of-plazis-covid-19-related-activities/3e26b3bc95a4b39f0a2a9d7fccee8b19/

19. Senderov V, Simov K, Franz N, Stoev P, Catapano T, Agosti D, et al. OpenBiodiv-O: ontology of the OpenBiodiv knowledge management system. J Biomed Semant. 2018;9: 5. doi:10.1186/s13326-017-0174-5

20. Dimitrova M, Poelen J, Zhelezov G, Georgiev T, Agosti D, Penev L. Semantic Publishing Enables Text Mining of Biotic Interactions. Biodivers Inf Sci Stand. 2020;4: e59036. doi:10.3897/biss.4.59036

21. Poelen J, Upham N, Agosti D, Ruschel T, Guidoti M, Reeder D, et al. CETAF-DiSCCo/COVID19-TAF biodiversity-related knowledge hub working group: indexed biotic interactions and review summary. Zenodo; 2020. doi:10.5281/zenodo.3839098

22. Olival KJ, Hosseini PR, Zambrana-Torrelio C, Ross N, Bogich TL, Daszak P. Host and viral traits predict zoonotic spillover from mammals. Nature. 2017;546: 646–650. doi:10.1038/nature22975

23. Nagaraj A, Shears E, Vaan M de. Improving data access democratizes and diversifies science. Proc Natl Acad Sci. 2020;117: 23490–23498. doi:10.1073/pnas.2001682117

Acknowledgments: We thank A. Casino, D. Koureas, and W. Addink for organizing the CETAF-DiSSCo COVID-19 Taskforce that resulted in this research. Illustrations were created in Inkscape with images from BioRender.com and CC-BY licenses; Fig 2 reuses bat images from http://zenodo.org/record/3756730 and Michigan Science Art with permission. E. Florsheim provided valuable conversations. **Funding:** Our efforts were supported by CETAF (Consortium of European Taxonomic Facilities), DiSSCo (Distributed System of Scientific Collections), the Biodiversity Knowledge Integration Center at Arizona State University (N.S.U.), the

SYNTHESYS+ Research and Innovation action (Q.J.G., grant nos. H2020-EU.1.4.1.2 and 823827), the Arcadia charitable fund of Lisbet Rausing and Peter Baldwin (D.A.), the National Science Foundation Advancing Digitization of Biodiversity Collections Program DBI-1547229 (D.P.), and National Science Foundation award "Collaborative Research: Digitization TCN: Digitizing collections to trace parasite-host associations and predict the spread of vector-borne disease," Award numbers DBI:1901932 and DBI:1901926 (J.H.P); **Author contributions:** N.S.U, D.P., Q.P.G., N.B.S., J.H.P., and D.A. designed the conceptual arguments of this research. J.H.P., M.G., D.A., L.P., and N.S.U. worked on methods and created software for data liberation. Curation of the resulting data was performed by J.H.P., D.A., and L.P., while N.S.U. and J.H.P. performed validations. N.S.U. and D.A. wrote the initial draft, N.S.U. created the figures with help from C.B.S., and all authors reviewed and edited the manuscript. **Competing interests:** Authors declare no competing interests; **Data and materials availability:** All data liberated as a result of these efforts are available at https://doi.org/10.5281/zenodo.4068958.