

Towards Unrestricted Usage of Public Genomic Data

Rudolf I. Amann¹, Shakuntala Baichoo², Benjamin J. Blencowe³, Peer Bork⁴, Mark Borodovsky⁵, Cath Brooksbank⁶, Patrick S.G. Chain⁷, Rita R. Colwell⁸, Daniele G. Daffonchio^{9,10}, Antoine Danchin^{11,12}, Victor de Lorenzo¹³, Pieter C. Dorrestein^{14,15}, Robert D. Finn⁶, Claire M. Fraser¹⁶, Jack A. Gilbert¹⁷, Steven J. Hallam^{18,19}, Philip Hugenholtz²⁰, John P.A. Ioannidis²¹, Janet K. Jansson²², Jihyun F. Kim²³, Hans-Peter Klenk²⁴, Martin G. Klotz^{25,26}, Rob Knight²⁷, Konstantinos T. Konstantinidis²⁸, Nikos C. Kyrpides²⁹, Christopher E. Mason^{30,31,32}, Alice C. McHardy³³, Folker Meyer³⁴, Christos A. Ouzounis³⁵, Aristides A. N. Patrinos³⁶, Mircea Podar³⁷, Katherine S. Pollard³⁸, Jacques Ravel¹⁶, Alejandro Reyes Muñoz³⁹, Richard J. Roberts⁴⁰, Ramon Rosselló-Móra⁴¹, Susanna-Assunta Sansone⁴², Patrick D. Schloss⁴³, Lynn M. Schriml¹⁶, João C. Setubal⁴⁴, Rotem Sorek⁴⁵, Rick L. Stevens^{46,47}, James M. Tiedje^{48,49}, Adrian Turjanski⁵⁰, Gene W. Tyson²⁰, David W. Ussery⁵¹, George M. Weinstock⁵², Owen White¹⁶, William B. Whitman⁵³, Ioannis Xenarios⁵⁴

¹Max Planck Institute for Marine Microbiology, Bremen, D-28359, Germany

²Department of Computer Science & Engineering, University of Mauritius, Réduit 80837, Mauritius

³Donnelly Centre and Department of Molecular Genetics, University of Toronto, 160 College Street, Toronto, ON M5S 3E1, Canada

⁴The European Molecular Biology Laboratory (EMBL), Structural and Computational Biology, Heidelberg, Germany

⁵Wallace H. Coulter Department of Biomedical Engineering, Georgia Tech, Atlanta, Georgia 30332, USA

⁶European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI) Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

⁷Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM, USA

⁸CosmosID, Rockville, MD, USA

⁹King Abdullah University of Science and Technology (KAUST), Biological and Environmental Sciences and Engineering Division (BESE), Thuwal 23955-6900, Saudi Arabia

¹⁰University of Milan, Department of Food Environmental and Nutritional Sciences, Milan 20133, Italy

¹¹Integromics, Institute of Cardiometabolism and Nutrition, Hospital de la Pitie-Salpetriere, 47 Boulevard de l'Hospital, 75013 Paris, France.

¹²School of Biomedical Sciences, Li KaShing Faculty of Medicine, Hong Kong University, 21 Sassoon Road, Pokfulam, Hong Kong.

¹³Centro Nacional de Biotecnología-CSIC, Campus de Cantoblanco, 28049, Madrid, Spain

¹⁴Collaborative Mass Spectrometry Innovation Center, University of California San Diego, La Jolla, CA, USA

¹⁵Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA

¹⁶Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland, USA

¹⁷The Microbiome Center, Surgery, University of Chicago, Chicago, IL, 60637, USA

¹⁸Department of Microbiology & Immunology, University of British Columbia, Vancouver, BC, V6T1Z1, Canada

¹⁹ECOSCOPE Training Program, University of British Columbia, Vancouver, British Columbia, V6T1Z3, Canada Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, Australia

²¹Departments of Medicine, Health Research and Policy, Biomedical Data Science, and Statistics, Stanford University and Meta-Research Innovation Center at Stanford (METRICS) , Stanford, CA 94305, USA

²²Biological Sciences Division, Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, WA, USA

²³Department of Systems Biology and Division of Life Sciences, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea

²⁴School of Natural and Environmental Sciences, Faculty of Science, Agriculture and Engineering, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK

²⁵School of Molecular Biosciences, College of Veterinary Medicine, Washington State University, Richland, WA, USA

²⁶State Key Laboratory of Marine Environmental Science, Institute of Marine Microbes & Ecospheres, College of Ocean & Earth Sciences, Xiamen University, Xiamen, China

²⁷Center for Microbiome Innovation, and Departments of Pediatrics, Bioengineering, and Computer Science & Engineering, UC San Diego, 9500 Gilman Drive, San Diego, CA 92023, USA

²⁸School of Civil and Environmental Engineering & School of Biological Sciences, Georgia Institute of Technology, 311 Ferst Dr. NW, Atlanta, GA, 30332, USA

²⁹Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

³⁰Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA

³¹The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA

³²The WorldQuant Initiative for Quantitative Prediction, Weill Cornell Medicine, New York, NY, USA

³³Department of Computational Biology of Infection Research, Helmholtz Centre for Infection Research, 38124 Braunschweig, Germany

³⁴Mathematics and Computer Science Division, Argonne National Laboratory, 9700 S Cass Avenue, Lemont, IL 60439, USA

³⁵Biological Computation & Process Laboratory, Chemical Process & Energy Resources Institute, Centre for Research & Technology Hellas, Thessalonica GR-57001, Greece

³⁶Novim, Kohn Hall, UC Santa Barbara, Santa Barbara, CA 93106

³⁷Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA

³⁸Gladstone Institutes, University of California, and Chan-Zuckerberg Biohub, San Francisco, CA 94158, USA

³⁹Departamento de Ciencias Biológicas, Universidad de los Andes, Bogotá, Colombia

⁴⁰New England Biolabs, Ipswich, MA 01938, USA

⁴¹Marine Microbiology Group, IMEDEA (CSIC-UIB), 07190 Esporles, Spain

⁴²Oxford e-Research Centre, Department of Engineering Science, University of Oxford, 7 Keble Road, OX1 3QG, Oxford, UK

⁴³Department of Microbiology & Immunology, University of Michigan, Ann Arbor, MI, USA

⁴⁴Department of Biochemistry, Institute of Chemistry, University of São Paulo, Av. Prof. Lineu Prestes, 78 room 909, 05508-000, São Paulo, SP, Brazil

⁴⁵Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel.

⁴⁶Computer Science Department and Computation Institute, University of Chicago, Chicago, Illinois, USA

⁴⁷Computing, Environment, and Life Sciences Directorate, Argonne National Laboratory, Argonne, Illinois, USA

⁴⁸Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, Michigan 48824, USA

⁴⁹Center for Microbial Ecology, East Lansing, Michigan 48824, USA

⁵⁰Departamento de Química Biológica, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Pabellón II, Buenos Aires, Argentina.

⁵¹Department of Biomedical Informatics University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA

⁵²Jackson Laboratory for Genomic Medicine, Farmington, CT.

⁵³Department of Microbiology, University of Georgia, Athens 30602, USA

⁵⁴Center for Integrative Genomics, University of Lausanne Lausanne, Switzerland

Despite long fermentation of ideas and some notable progress in data sharing, existing practices still place restrictions on the open and unconditional use of various genomic data after their official approval for release to the public domain or to public databases. These restrictions often come into conflict with the terms and conditions of the relevant funding bodies who supported the release of those data for the benefit of the scientific community and society. We argue that the publicly available data should be treated as open data, *i.e.*, a shared resource with unrestricted usage for analysis, interpretation, and publication.

Background

To address the issues of sharing genomic data, particularly the sharing of pre-publication data that already has ethical approval for public release, biomedical scientists agreed upon and created a public declaration in 2003, which came to be known as the Fort Lauderdale Agreement (1). This declaration supports the free and unrestricted use of genome sequencing data by the scientific community after the data and related phenotype information have ethical approval for release, but before those data are used for publication. It proposed that the scientist leading the generation of new data should “*make the data generated by the resource immediately and freely available without restriction*”. The agreement encouraged users of data publicly released in this fashion to “*appropriately cite the source of the data analysed and*

acknowledge the resource producers". Moreover, it recommended that the users of the data should *"recognize that the resource producers have a legitimate interest in publishing prominent peer-reviewed reports describing and analyzing the resource that they have produced"*.

In the last 15 years since the Fort Lauderdale Agreement, the issue of expanding and materializing wider, faster, more efficient data sharing has been a recurring theme (2-4). Data sharing policies have not been static and many funding agencies, including NIH, have fine-tuned policies focused on specific platforms, e.g. genome-wide association studies, or even with wider spectra of data being covered (e.g. the 2014 NIH Genome Data Sharing Policy). These efforts often attract major attention and commentaries in the literature. Issues of individual privacy and consent are also favorite topics in such discussions. However, despite clear progress in data sharing, as new and larger and more complex types of datasets become available, the need to optimize and to bring up-to-date the existing sharing practices to meet current challenges becomes even greater.

Motivation for policy changing

Past recommendations have been typically restricted to well-defined community resource projects and none of them covers all sequencing projects. Moreover, the Fort Lauderdale Agreement contained a self-contradictory proposition. With immediate release, resource producers are not always guaranteed that they can publish prominent peer-reviewed reports if others use their data first. This paradox is evident, despite the acknowledgement of academic fair-play. The data can either be available without restrictions (and therefore open), or they will be available with some restrictions (and therefore not open). Subsequent improvements in data sharing policies in the last 15 years have not tackled this dilemma in a satisfactory way.

Many advances have occurred in genomics and data science during the past 15 years. These include massive increases in both the production and scale of genomic and metagenomic datasets, as well as the development of open data initiatives¹ according to which *"some data should be freely available to everyone to use and re-publish as they wish, without restrictions from copyright, patents or other mechanisms of control"*. The sequencing revolution has resulted in the generation of myriad datasets, many of which are publicly released without an accompanying publication. These datasets are often processed and integrated into public databases and public repositories, or under scholarly commons such as the Open Science Framework². These resources provide users the opportunity to mine and analyze data that have been made public but are often still unpublished in the peer-review literature. The large number of datasets integrated into these and other similar resources (totaling hundreds of thousands of datasets) and the lack of comprehensive automated mechanisms to track the publication status of each dataset make it virtually impossible for these resources to provide this information to their users. Moreover, with the advent of large global data analysis studies, which include the mining of thousands of publicly available datasets and reaching a "genomical" scale of yottabytes that even surpasses measurements in astronomical data generation (5), it has

¹ https://en.wikipedia.org/wiki/Open_data

² <https://osf.io/>

become challenging to appropriately acknowledge or cite every dataset that has been included in such an analysis. While in the past immediate release of data may not have enabled an “outsider” to publish before the scientists producing the original data, analytical capability advances and the availability of so many datasets currently may place teams of scientists other than the originator of a given dataset in a better position to publish first. For example, the “outsider” team may have better analytical capabilities, and/or overarching protocols for analyzing more comprehensive sets of data, pre- or post-publication. Also, sequence datasets can be interrogated via numerous value-added platforms and tools from multiple groups. Taken together, these developments have rendered the Fort Lauderdale Agreement rather outdated and in dire need of a revision to reflect the current state of science and technology.

Public Data Controversy: open or not?

Ever since the Fort Lauderdale Agreement and the declaration that scientific data should be publicly available, we have seen a number of widely adopted developments, e.g. open-access, FAIR principles (6), etc., that have created a more refined data-sharing ecosystem that is not captured by the earlier agreements. In order to address the current complexities of data sharing, new community efforts are currently being organized. The European Bioinformatics Institute has recently launched a community survey to determine what most investigators want for open data in microbiome research (7). The twitter hashtag and social media campaign `#free_the_public_data_movement` was also launched to open the dialogue on these issues and increase awareness for the need to revisit the community recommendations about the usage of published or unpublished public data. Controversy has arisen, as different points of view have been expressed regarding data ownership and best data utilization practices. For example, studies such as the American Gut (7) GEBA (8), and the Earth Microbiome Project (9) have been open and available for years prior to formal publication, resulting in many publications by the community prior to the central documents describing the core projects of data generation in an asynchronous, yet widely accepted manner. While publication of other researchers’ data can lead to claims of ‘data parasitism’, many acknowledge that such use of data, adds value (10), for example, due to the increased knowledge obtained from meta analyses of multiple datasets with goals substantially different to, or unanticipated by, the original data generators. Indeed, many researchers have built global Consortia from these data sharing models and approaches, such as Tara Oceans, MetaSUB, and Earth BioGenome, respectively.

In contrast, the supporters of restricted utilization of public genomic data argue that: (a) pre-publication data are not validated and may contain errors, and (b) generating the new data typically involves years of preparation including project design, setting up collaborations, and sampling, most of which requires the extended and strenuous efforts of several people, including students and postdocs not to mention the operational cost of processing samples for sequencing. Accordingly, the first usage of the data after they become public should still be restricted so that the Principal Investigator (PI) under whom the data were generated should retain the rights to first publication. Even when the data have become public following the data release policy of the funding agency, outside investigators should still contact the scientist(s) that have generated the initial data and request permission for using them. Some supporters go

as far as extending this proposal to data that are not just public but also already published in research articles. In these cases, the PIs of those projects would like to maintain the prime (or even exclusive) rights for further analysis and publication of the data that they produced, even after their initial publication. Once we move beyond the genomic landscape, several other fields of biomedical research have sadly continued to have very limited data availability, e.g. many of the large epidemiological cohorts retain the data for use only by the PI and his/her inner circle of associates, even after hundreds of articles have been published in the literature. With increasing interdisciplinarity in science, it is becoming more common for very different traditions toward data sharing to co-exist in the same project, e.g. when a nutritional epidemiology cohort (a field with a tradition of limited sharing) undertakes microbiome analyses (a field with a strong tradition of sharing).

We fully understand the concerns of the scientists who produce the data and the need to acquire appropriate credit. However, we also argue that once the data are publicly released following the data release rules of the agency that funded the project, they should be freely available for use without any restrictions or conditions. Ultimately, the intention of the funding agencies who require pre-publication data sharing has always been to encourage the use of such data by the entire community and to encourage open competition to accelerate discovery and maximize the benefit for members of society who are paying for data generation. And most importantly, as was recently argued, freely available public data without restrictions is considered *“critical for progress in any scientific discipline and has been the cornerstone of sound and reproducible genomics research”* (11).

In addition, when new sequence data are submitted to any of the International Nucleotide Sequence Database Collaboration (INDSC) sites, they have already gone through rigorous Quality Controls (QC) and are further scrutinized by additional QC steps at the INSDC repositories.

Public Data: in need of a new usage policy and a new credit system

Our recommendation on the usage policy of publicly available data primarily rests on the following guiding principles:

- (i) public genomics data that have ethics approval for release should be open data, *i.e.* data that are available for unrestricted usage, together with their associated metadata – with the exception of sensitive human data where additional ethics restrictions may apply (11);
- (ii) science advances through open competition, not through posing restrictions and limitations;
- (iii) credit should be given appropriately to resource producers and should be transparent.

These recommendations for the revision of the public data usage policy should not in any way impede with the protection of sensitive human data. With increasing ability to identify people, it is essential to alert research participants about this possibility and try to obtain by default consent for widespread use of raw data and when possible to encourage collaboration. We acknowledge that for existing sensitive human data, some restrictions may be appropriate. However, moving forward, explicit informed consent can remove much of this obstacle, if

participants are told in advance that their data will be maximally, openly used and what the potential (if often minimal) risks inherent in this use are. In the meanwhile, resistance to share sensitive data, such as those from clinical trials is gradually being curbed. Some major clinical journals such as PLoS Medicine and BMJ have explicit policies not to publish clinical trials unless the authors pledge to share the raw data. An empirical evaluation showed that trialists shared the raw data from 46% of these trials upon request and re-analyses of the raw data did not change any of the major conclusions (12). Across the entire biomedical literature, in 2015-2017, about one-fifth of published articles shared raw data, a major increase over previous years (13).

The unrestricted usage of public data should be aligned with a reward system in research and academia (14). Institutions and funders need to recognize the coinage of open data sharing and confer the proper credit on scientists who have generated the data. Universities and research institutions may offer promotion and tenure based on different tracks that may suit data generators, data analysts, data translators, or scientists who combine two or more of these functions. This is particularly important when the data producer has less influence or resources compared to outsiders who can leverage the impact of those data, a dichotomy that occurs frequently in international science³. Attribution is particularly important when there are power imbalances in science. Digital Object Identifiers (DOIs) have been used to monitor and track the changes for any data types (e.g. figshare)⁴ and are now used quite broadly. As an example, in prokaryotic nomenclature and taxonomy, DOIs have been used for more than a decade and already provide persistent records. These can be readily incorporated into digital content prospectively or retrospectively as they provide a solid framework to monitor and track the use of all sequencing projects, including unpublished datasets (15). It is also important to identify efficient ways to give credit for the generation of background protocols that describe the process of production and the physical effort and thinking that was invested towards producing specific datasets. Some data production entails automated, routine processes, while in other cases intensive sample collection, processing and innovation to design a protocol and execute a study are required. Such logistic and protocols can also be efficiently linked to the datasets they generate. Furthermore, many datasets and projects are currently created by merging multiple smaller sets of data. The challenge is to adopt smart strategies so that these iterative agglomerations can still carry the DOIs or other reference of the smaller sets that they have incorporated. This becomes increasingly important for the numerous databases that collect, integrate and improve the quality and value of the public raw data or are linked to time series studies. To a large extent, these resources are also generating new data and metadata, thus enabling the community to advance their research and make new discoveries. Eventually, this approach should aim at providing an independent means of evaluating the impact of the research products that are created by individual scientists, much like current bibliometric methods. In this scenario data becomes a public good or shared resource that enables the research community to conduct synergistic research and training activities. More efforts to facilitate data deposition by data generators by improving the process of data submission to

³ <http://icis.ucdavis.edu/?p=1169>

⁴ <https://figshare.com/>

public repositories and recognizing the effort to generate such public data sets is critical. Yet, while acknowledging the need for improved credit system for the data generators, it is important to note that asking for data generators to be also by default the data analysts is like asking for a screenplay writer to also be necessarily the director of a movie. While these functions may sometimes co-exist, they don't have to.

The scale of the data generation makes it imperative to revisit data release policies that funding agencies and journal publishers have implemented both for sequence data and their associated metadata (16). While a lot of DNA sequence data are released in publicly accessible databases within twenty-four hours after generation, this principle was never extended to encompass other sequence data as, for instance, microbial genomes or metagenomes and associated metadata. Here, the US Department of Energy (DOE) has led the way with an immediate data release policy (i.e. the data become publicly available immediately upon generation), however, other funding agencies allow for the data to remain private until the time of publication. Yet, data analysis may take years before publication; this delay coupled with the increasing speed of generating new data creates a very different landscape compared to the past, when these policies were instituted. Thus, revisions to data release policies, are also necessary to ensure that public data can be used by the entire community in a timely fashion, without losing its value and therefore impact. Finally, journal publishers also need to revisit their publication policies, with respect to the availability of the data when a manuscript is submitted for publication. Following the recommendations of *Microbiome's* editors, the sequence data and their associated metadata need to be freely available together with detailed protocols at the time a manuscript is submitted for peer review, rather than post-publication (11).

The need for a clear policy protecting public data from restrictions, has become even more important with the recently proposed changes to the Nagoya Protocol⁵ and the ongoing efforts to include “digital sequence information” in an international agreement against biopiracy (17). Wider data sharing is also likely to allow more participation in the research enterprise of the large number of good scientists who work in resource-poor settings and otherwise cannot compete in generating expensive new data.”

In summary, when genomic data become publicly available, they need to be more widely shared without strings attached to the sharing process. Advancing the genomics field forward requires strong affirmative policies towards open and unrestricted data sharing that promote inclusive community-driven research and training activities.

Important notice: The views expressed in this paper are those of the authors and do not reflect their affiliated centers or any sponsoring Federal Agencies

References

⁵ <https://www.cbd.int/abs/about/>

1. Report of a meeting organized by the Wellcome Trust and held on 14–15 January 2003 at Fort Lauderdale, USA. (Publication 2003 <https://www.genome.gov/pages/research/wellcomereport0303.pdf>) [the easiest access to this source is via the URL]
2. E. Birney, *et al.*, Prepublication data sharing. *Nature* **461**(7261):168-70 (2009).
3. D. Field, *et al.* Megascience. 'Omics data sharing. *Science* **326**(5950):234-6 (2009).
4. P. N. Schofield, *et al.*, Post-publication sharing of data and tools. *Nature* **461**(7261):171-3 (2009)
5. Z. D. Stephens, *et al.*, Big Data: Astronomical or Genomical? *PLoS Biol.* **13**(7):e1002195. (2015).
6. M. D. Wilkinson, *et al.*, The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018 (2016).
7. D. McDonald, *et al.*, American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems*. **3**(3), pii: e00031-18 (2018).
8. N. C. Kyrpides, *et al.*, Genomic encyclopedia of bacteria and archaea: sequencing a myriad of type strains. *PLoS Biol.* **12**(8), e1001920 (2014).
9. L.R. Thompson, *et al.*, A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**(7681):457-463 (2017).
10. C. S. Greene, *et al.*, Celebrating parasites. *Nature Genetics* **49**, 483–484 (2017).
11. M. G. I. Langille, *et al.*, "Available upon request": not good enough for microbiome data! *Microbiome* **6**(1), 8 (2018).
12. Naudet F, *et al.*, Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in The BMJ and PLOS Medicine. *BMJ*. **360**:k400 (2018)
13. Wallach J *et al.*, *PLoS Biology*, **in press** (2018)
14. D. Moher, *et al.*, Assessing scientists for hiring, promotion, and tenure. *PLoS Biol.* **16**(3), e2004089 (2018).
15. G. M. Garrity, *et al.*, Studies on Monitoring and Tracking Genetic Resources: An Executive Summary. *Standards in Genomic Sciences* **1**(1), 78–86 (2009).
16. Cook-Deegan R, *et al.*, Sharing Data to Build a Medical Information Commons: From Bermuda to the Global Alliance. *Annu Rev Genomics Hum Genet.* **18**:389-415.(2017)
17. K. Kupferschmidt, Biologists raise alarm over changes to biopiracy rules. *Science* **361**, 14 (2018).