Title

Toward unrestricted use of public genomic data

Journal

Science, 363(6425)

ISSN

0036-8075

Authors

Amann, Rudolf I

Baichoo, Shakuntala

Blencowe, Benjamin J

et al.

Publication Date

2019-01-25

DOI

10.1126/science.aaw1280

Peer reviewed

POLICY FORUM

DATA ACCESS

Toward unrestricted use of public genomic data

Publication interests should not limit access to public data

By **Rudolf I. Amann, Shakuntala Baichoo, Benjamin J. Blencowe, Peer Bork, Mark Borodovsky, Cath Brooksbank, Patrick S. G. Chain, Rita R. Colwell, Daniele G. Daffonchio, Antoine Danchin, Victor de Lorenzo, Pieter C. Dorrestein, Robert D. Finn, Claire M. Fraser, Jack A. Gilbert, Steven J. Hallam, Philip Hugenholtz, John P. A. Ioannidis, Janet K. Jansson, Jihyun F. Kim, Hans-Peter Klenk, Martin G. Klotz, Rob Knight, Konstantinos T. Konstantinidis, Nikos C. Kyrpides, Christopher E. Mason, Alice C. McHardy, Folker Meyer, Christos A. Ouzounis, Aristides A. N. Patrinos, Mircea Podar, Katherine S. Pollard, Jacques Ravel, Alejandro Reyes Muñoz, Richard J. Roberts, Ramon Rosselló-Móra, Susanna-Assunta Sansone, Patrick D. Schloss, Lynn M. Schriml, João C. Setubal, Rotem Sorek, Rick L. Stevens, James M. Tiedje, Adrian Turjanski, Gene W. Tyson, David W. Ussery, George M. Weinstock, Owen White, William B. Whitman, Ioannis Xenarios**

Despite some notable progress in data sharing policies and practices, restrictions are still often placed on the open and unconditional use of various genomic data after they have received official approval for release to the public domain or to public databases. These restrictions, which often conflict with the terms and conditions of the funding bodies who supported the release of those data for the benefit of the scientific community and society, are perpetuated by the lack of clear guiding rules for data usage. Existing guidelines for data released to the public domain recognize but fail to resolve tensions between the importance of free and unconditional use of these data and the "right" of the data producers to the first publication. This self-contradiction has resulted in a loophole that allows different interpretations and a continuous debate between data producers and data users on the use of public data. We argue that the publicly available data should be treated as open data, a shared resource with unrestricted use for analysis, interpretation, and publication.

SHARING, PUBLISHING, PARADOX

The landmark 2003 Fort Lauderdale Agreement (*1*) supports the free and unrestricted use of genome sequencing data by the scientific community after the data and related phenotype information have had ethical approval for release but before those data are used for publication. In the years since the agreement, the issue of 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 have fine-tuned policies focused on specific platforms (such as genome-wide association studies) or even with wider spectra of data being covered (for example, the 2014 National Institutes of Health Genome Data Sharing Policy). A number of widely adopted developments [such as open-access, FAIR (findable, accessible, interoperable, and reusable) principles (*5*)] 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 being organized. For example, the European Bioinformatics Institute has launched a community survey to determine what most investigators want for open data in microbiome research.

However, despite improvements in aspects of data sharing policies in the past 15 years, with much focus on determining when data should be made publicly available (for example, the ENCODE project has recently eliminated the 9-month moratorium on data usage, applied in earlier phases of the project), policies have not adequately resolved a critical dilemma, regarding how data are to be used once made publicly available. The Fort Lauderdale Agreement contains a self-contradictory proposition, proposing that the scientist leading the generation of new data should "make the data generated by the resource immediately and freely available without restriction," yet at the same time recommending 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." 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 agreement's acknowledgment of academic fair play, encouraging users of data publicly released in this fashion to "appropriately cite the source of the data analysed and acknowledge the resource producers."

> ***"Unrestricted use of public data should be aligned with a reward system in research and academia."***

In light of this, supporters of restricted use of public genomic data point to the agreement to argue that the first use 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. This has been frequently implemented as official data release policy from various institutes or research consortia who make the data publicly available but restrict the analysis by the larger community (*6*). Even when the data have become public following the data release policy of the funding agency, proponents of this view argue that 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.

Justification of these restrictions is commonly that (i) prepublication data are not validated and may contain errors and (ii) generating new data typically involves years of preparation, including

The list of author affiliations is available in the supplementary materials. Email: jioannid@stanford.edu

project design, setting up collaborations, and sampling—most of which requires the extended and strenuous efforts of several people, including students and postdoctorals, not to mention the operational cost of processing samples for sequencing. Several other fields of biomedical research beyond genomics have also continued to have very limited data availability; for example, many large epidemiological cohorts [such as EPIC (European Prospective Investigation into Cancer and Nutrition), the Rotterdam Study, and Nurses' Health Study] retain the data for use only by the PI and his or her 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 coexist in the same project, such as when a nutritional epidemiology cohort (a field with a tradition of limited sharing) undertakes microbiome analyses (a field with a strong tradition of sharing).

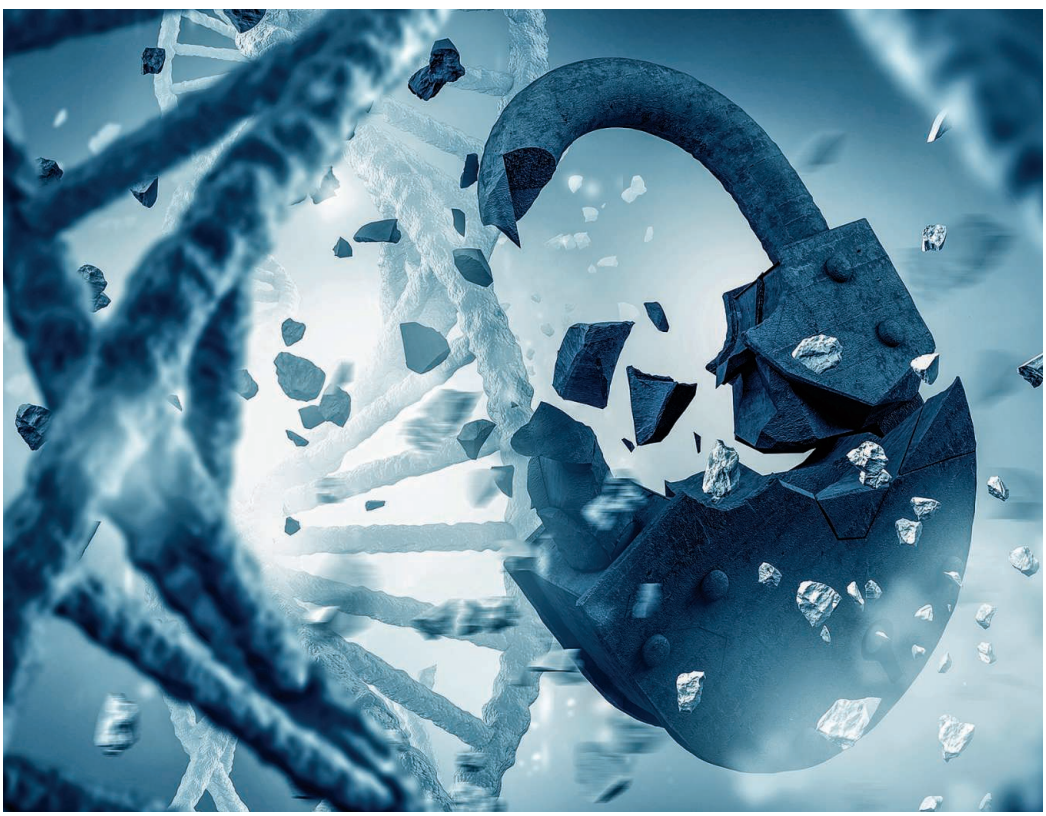

ILLUSTRATION: SERGEY NIVENS/SHUTTERSTOCK

OUTSIDERS AND VALUE ADDED

Many developments have rendered the Fort Lauderdale Agreement rather outdated and in need of a revision to reflect the current state of science and technology. Past recommendations have been typically restricted to well-defined community resource projects, and none of them covers all sequencing projects. As new, larger, and more complex datasets become available, the need to optimize and to bring up to date the existing sharing practices to meet current challenges becomes even greater.

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 Framework. The large number of datasets integrated into these and 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 reach a "genomical" scale of yottabytes (*7*), it has become challenging to appropriately acknowledge or cite every dataset that has been included in such an analysis.

Whereas in the past, immediate release of data may not have enabled an "outsider" to publish before the scientists who produced the original data, analytical capability advances and the availability of so many datasets may now place scientists other than the originator(s) 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 by means of numerous value-added platforms and tools from multiple groups.

Although publication of other researchers' data can lead to claims of "data parasitism," many acknowledge that such use of data can add value (*8*)—for example, owing 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.

ATTRIBUTION AND INCENTIVES

We argue that once 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. Our recommendation rests on the following guiding principles: (i) Public genomics data that have ethics approval for release should be open data—available for unrestricted use, together with associated metadata—with the exception of sensitive human data to which additional ethics restrictions may apply (*9*); (ii) science advances through open competition with clear-cut, transparent rules, not through posing restrictions and limitations; and (iii) credit should be given appropriately to resource producers and should be transparent.

These recommendations should not impede protection of sensitive human data. 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 meantime, resistance to sharing sensitive data, such as those from clinical trials, is gradually being curbed. Some major clinical journals such as *PLOS Medicine*

and *The BMJ* have policies to not publish clinical trials unless the authors pledge to share the raw data (*10*). An empirical evaluation showed that trialists shared raw data from 46% of these trials upon request, and reanalyses of the raw data did not change any of the major conclusions (*10*). Across the entire biomedical literature, in 2015 to 2017, about one-fifth of published articles shared raw data, a major increase over previous years (*11*).

Unrestricted use of public data should be aligned with a reward system in research and academia (*12*). Institutions and funders need to recognize the coinage of open data sharing and confer the proper credit on scientists who generate the data. Universities and research institutions may offer promotion and tenure on the basis of 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 with outsiders who can leverage the impact of those data—a dichotomy that occurs frequently in international science.

Attribution is particularly important when there are power imbalances in science. Digital object identifiers (DOIs) have been used to monitor and track changes for any data type (such as figshare) and are now used quite broadly. These can be readily incorporated into digital content prospectively or retrospectively because they provide a solid framework to monitor and track the use of all sequencing projects, including unpublished datasets (*13*).

It is also important to identify efficient ways to give credit for the generation of protocols that describe the process of data production and the physical effort and thinking that was invested toward producing specific datasets. Such 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, enabling the community to advance 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. Efforts to facilitate data deposition by data generators by improving the process of data submission to public repositories and recognizing the effort to generate such public data sets are critical. Although we acknowledge the need for an improved credit system for data generators, asking for data generators to be, by default, the data analysts as well is like requiring a screenplay writer also to be the director of the movie. Although these functions may sometimes coexist, they don't have to.

The scale of data generation makes it imperative to revisit data release policies that funding agencies and journal publishers have implemented for sequence data and associated metadata (*14*). Although a lot of DNA sequence data are released in publicly accessible databases within 24 hours after generation, this principle was never extended to encompass other sequence data—for example, microbial genomes or metagenomes and associated metadata. The U.S. Department of Energy has an immediate data release policy (the data become publicly available immediately upon generation); however, other funding agencies allow 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 with that of the past, when these policies were instituted. Thus, revisions to data release policies are necessary to ensure that public data can be used by the entire community in a timely fashion, without losing value and impact.

> ***"...asking for data generators to be, by default, the data analysts as well is like requiring a screenplay writer also to be the director of the movie."***

Last, journal publishers need to revisit their publication policies, with respect to the availability of the data when a manuscript is submitted for publication. Publishers should equip the peer-review process for their journals with editorial tools that enable authors to acknowledge generators of data deposited in the public domain. Likewise, data repositories should develop mechanisms that enable authors to reference such multiple-study– or multiple-author–generated datasets without excessively extending reference lists. 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 after publication (*9*).

The need for clear policy that protects public data from restrictions has become even more important with recently proposed changes to the Nagoya Protocol to the Convention on Biological Diversity and the ongoing efforts to include "digital sequence information" in an international agreement against biopiracy (*15*). Wider data sharing is likely to allow more participation in the research enterprise of the many scientists who work in resource-poor settings and may be less able to compete in generating expensive new data.

Advancing the genomics field requires strong affirmative policies toward open and unrestricted data sharing that promote inclusive community-driven research and training. The intention of the funding agencies who require prepublication 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. ■

REFERENCES AND NOTES

1. W. Trust, meeting report, 14–15 January 2003, Fort Lauderdale, USA (2003); www.genome.gov/pages/research/wellcomereport0303.pdf.
2. E. Birney *et al.*; Toronto International Data Release Workshop Authors, *Nature* **461**, 168 (2009).
3. D. Field *et al.*, *Science* **326**, 234 (2009).
4. P. N. Schofield *et al.*; CASIMIR Rome Meeting participants, *Nature* **461**, 171 (2009).
5. M. D. Wilkinson *et al.*, *Sci. Data* **3**, 160018 (2016).
6. E. Pennisi, *Science* (2018); 10.1126/science.aav4025.
7. Z. D. Stephens *et al.*, *PLOS Biol.* **13**, e1002195 (2015).
8. C. S. Greene, L. X. Garmire, J. A. Gilbert, M. D. Ritchie, L. E. Hunter, *Nat. Genet.* **49**, 483 (2017).
9. M. G. I. Langille, J. Ravel, W. F. Fricke, *Microbiome* **6**, 8 (2018).
10. F. Naudet *et al.*, *BMJ* **360**, k400 (2018).
11. J. D. Wallach, K. W. Boyack, J. P. A. Ioannidis, *PLOS Biol.* **16**, e2006930 (2018).
12. D. Moher *et al.*, *PLOS Biol.* **16**, e2004089 (2018).
13. G. M. Garrity *et al.*, *Stand. Genomic Sci.* **1**, 78 (2009).
14. R. Cook-Deegan, R. A. Ankeny, K. Maxson Jones, *Annu. Rev. Genomics Hum. Genet.* **18**, 389 (2017).
15. K. Kupferschmidt, *Science* **361**, 14 (2018).

ACKNOWLEDGMENTS

The views expressed in this paper are those of the authors and do not reflect their affiliated centers or any sponsoring federal agencies. M.B. is supported by NIH R01GM128145. P.H. is co-founder, board member, and equity holder of Microba. K.T.K. is supported by NSF 1759831. F.M. is supported by NIH 1R01AI123037-01 and NSF 1645609. C.A.O. is supported by Elixir-GR (MIS#5002780) which is funded by the Operational Programme NSRF 2014-2020 and cofinanced by Greece and the European Union. A.A.N.P is an advisor to LunaDNA LLC, Synthetic Genomics and a consultant at Oak Ridge National Lab. K.S.P. is a consultant for Tenaya Therapeutics, Phylagen, and uBiome.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/363/6425/350/suppl/DC1

10.1126/science aaw1280

Toward unrestricted use of public genomic data

Rudolf I. Amann, Shakuntala Baichoo, Benjamin J. Blencowe, Peer Bork, Mark Borodovsky, Cath Brooksbank, Patrick S. G. Chain, Rita R. Colwell, Daniele G. Daffonchio, Antoine Danchin, Victor de Lorenzo, Pieter C. Dorrestein, Robert D. Finn, Claire M. Fraser, Jack A. Gilbert, Steven J. Hallam, Philip Hugenholtz, John P. A. Ioannidis, Janet K. Jansson, Jihyun F. Kim, Hans-Peter Klenk, Martin G. Klotz, Rob Knight, Konstantinos T. Konstantinidis, Nikos C. Kyrpides, Christopher E. Mason, Alice C. McHardy, Folker Meyer, Christos A. Ouzounis, Aristides A. N. Patrinos, Mircea Podar, Katherine S. Pollard, Jacques Ravel, Alejandro Reyes Muñoz, Richard J. Roberts, Ramon Rosselló-Móra, Susanna-Assunta Sansone, Patrick D. Schloss, Lynn M. Schriml, João C. Setubal, Rotem Sorek, Rick L. Stevens, James M. Tiedje, Adrian Turjanski, Gene W. Tyson, David W. Ussery, George M. Weinstock, Owen White, William B. Whitman and Ioannis Xenarios

Science **363** (6425), 350-352.
DOI: 10.1126/science.aaw1280

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.