Publication rights in the era of open data release policies

Lee Rowen, Gane K. S. Wong, Robert P. Lane, Leroy Hood

Abstract

The open data release policy adopted by the large-scale DNA sequencing centers has made accessible valuable information that facilitates research. Herein, we argue that the data producers' rights to receive credit for at least some portion of the analyses of the data must be protected. We suggest that this protection take the form of a specification of the probable content of the primary paper the data producers intend to publish when the data gathering is complete. Rights to publish that paper ought then be restricted to the producers unless they give permission otherwise.

In 1996, an international group of principal investigators from the large sequencing centers working on the Human Genome Project adopted a set of data release principles known as the Bermuda Conventions (1). The key principle states that "all human genomic sequence information generated by centers funded for large-scale sequencing should be freely available and in the public domain in order to encourage research and development and to maximize its benefit to society." Adherence to an immediate data release policy was subsequently formalized as a condition for receiving sequencing funds from the National Human Genome Research Institute, the Department of Energy, the Wellcome Trust, and other granting agencies. The Bermuda Conventions have thus set a precedent, and a possible norm, for other future large-scale data gathering projects.

Immediate data release generates a conflict between the interests of the data producers and third parties using that data. At the heart of this conflict are two interrelated issues. First, data producers often draw a distinction between preliminary and final data. Second, the official notification that the data are final is often captured in the publication of comprehensive analyses by the data producers in peer-reviewed journals. Thus, publication of analyses by third parties, before the data producers have officially signed off, preempts what the data producers consider to be their prerogative. This conflict underlies the recent controversy, discussed in Nature (2), regarding who has the right to publish analyses of the malaria and sleeping sickness parasite genomes that are currently being sequenced by a consortium of genome centers.

The underlying problem is that the data are "out there" with no formal restrictions on their use. Nothing except a journal's editorial policy or the peer-

review process prevents a third party from publishing sequence analyses accompanied merely by a reference to the accession numbers assigned to the sequences by the database. An alternative practice is to include an acknowledgement to the appropriate genome center(s). Neither form of attribution does much to benefit the careers of the individual data producers, particularly as accession numbers are not considered prior publications (3).

In the past, considerations of etiquette have guided the appropriateness of publishing analyses of other people's sequences or annotations. Often, but not always, contact between the third party and the data producers results in permission to publish, co-authorship, back-to-back papers, or other agreeable options. Acrimony arises when significant analyses are published against the wishes of, or without consulting, the data producers . For example, Immunology Today recently published two review articles (4) based on the gene content of ~1.5 megabases of annotated sequence from the mouse major histocompatibility complex (MHC). As directors of this sequencing effort, we (L.R and L. H.) consider these reviews to be a violation of both the spirit of the open data release policy and the normal workings of scientific communication, wherein reviews come after primary publications in peer-reviewed journals, not before.

The MHC example indicates that etiquette considerations may no longer hold much power in governing proper behavior regarding the publication of analyses of data posted on the internet. The web fosters a climate of anonymity in which the content of data is divorced from its context of acquisition. Yet, a situation in which it becomes the norm that no credit is given to data producers is untenable. To the extent that any community benefits from freely shared information, that same community must also accept the responsibility to ensure that the producers of this information are appropriately rewarded. In science, all rewards (to both career and ego) flow from publications in peer-reviewed journals.

What to do?

Consider a precedent from another scientific discipline. In space sciences, there is a mechanism to reward the people who oversee the construction of the large observatories, like the Hubble Space Telescope, that everyone shares. Scientists involved in developing the instruments are given a guaranteed amount of observing time, with an opportunity to specify their research objectives and the objects that they plan to observe. Their program is made publicly available at the time of a general Call for Proposals, and other proposals are not allowed to duplicate their stated plans. They then have a proprietary period of time after the observations are completed to analyze the data and publish their results (5).

What is different about large-scale sequencing is that the data are already freely available on the web. Since there is no way to control what people do with these

data, a new method must be created to determine what they can publish based on this data. We propose a policy that formalizes what is now frequently done informally – a requirement of *permission* from the data producers before third parties are allowed to publish certain types of analyses (6). In a sense, citing an accession number is analogous to referencing a "personal communication." Personal communications are accompanied by a letter of support granting the manuscript authors permission to use the unpublished data, and that permission is not generally granted if the proposed paper precludes from publication a similar paper that the producers themselves intend to write.

One might immediately object that the phrase "certain types of analyses" is unclear and, therefore, that the conditions under which permission to publish is required are not specifiable. To some extent, this is true. Sequence analysis is inherently open-ended. Disputes are inevitable, given the vagueness with which boundary conditions on legitimate ownership can be formulated. Nonetheless, we believe that the research community must develop a policy and set specific guidelines on the kinds of analyses that the data producers can justifiably claim priority. Two obvious criteria come to mind.

(1) The analysis must be based directly on sequence from a limited number of producers (e.g. the malaria consortium). This eliminates from consideration cross-species or global analyses, which typically require unpublished data from many unrelated producers, and where it is almost impossible not to use data produced outside of one's own laboratory. If permission has to be obtained for every such sequence, many worthwhile projects will never be done, to the detriment of biology.

(2) The analysis must address a question that the sequence producers could reasonably have planned. This eliminates from consideration unexpected new discoveries, like the realization that horizontal gene transfer is common in bacterial genomes. This result could not have been anticipated by the sequence producers, and cannot plausibly be claimed. However, a description of all the genes in a particular organism or a major locus should be restricted because it is an "obvious" paper.

Let us also not forget the fundamental reason why these open data release policies exist. It is widely accepted that hundreds, indeed thousands, of papers will arise from the genomic data produced by these large-scale sequencing projects. There is more to be learned from the sequence than any one laboratory can accomplish. Therefore, it is extremely shortsighted for the scientific community not to reserve at least a few papers for the people who worked so hard to produce the data. Is it so unreasonable to insist that third parties should focus on the *other* thousands of papers?

Implementation

First, we suggest that the databases add a tag or qualifier to an accessioned sequence entry that indicates whether the submitters require permission for publication of analyses based directly on that sequence. If the data producers do not require a request for permission, then no restrictions would be placed on what is done with the sequence, and the accession number would be considered an acceptable reference.

Second, for accessioned sequences tagged as requiring permission, an indication of the sort of publication(s) that is (are) planned must be included as part of the database entry. These plans can, of course, change over the course of the project. Titles like "Analysis of the 750-kilobase mouse major histocompatibility complex (MHC) class III region and comparison to its human counterpart" would inform third parties of the intention of the data producers. A brief abstract would also be required, in order to allow potential journal editors and reviewers to decide when or if a publication claim right has been violated.

Third, if data producers require permission for third party publication of analyses of their sequences, then a negotiation must occur over the contents and timing of the proposed publication, and a permission release form must be submitted to the journal editor, as is done now for many journals regarding "personal communications." Once the producers have published the sequence analysis in accordance with their specified plans, then no restrictions on third party publications would remain.

Fourth, there must be a mechanism for resolving disputes and enforcing the policy. There will be instances where third parties insist that sequence producers have claimed too much territory or where sequence producers have failed to publish their analyses after some reasonable period of time. Conflict resolution could plausibly be left to the journal editors and reviewers, who are already entrusted with enforcement of a related issue – ensuring that appropriate papers are cited in a manuscript.

Given where the final responsibility for enforcement lies, it is imperative that journal editors agree on a set of policies so that the "if we don't publish it someone else will" argument will lose its force. Moreover, given the volume of data being released for the Human Genome Project, and other organisms, it is hoped that an agreed upon policy could be developed soon, in order to avoid the conflicts and bad feelings that are arising out of the current ambiguities regarding claim rights to sequence analyses.

The benefits of a successful policy would have implications well beyond sequencing. Other examples of unpublished data that have been freely shared

include expressed sequence tags (ESTs), single nucleotide polymorphisms (SNPs), and many well-known bioinformatics tools. Undoubtedly, these same issues will soon arise in functional genomics and proteomics. If the issue of rewarding data producers is not adequately addressed, then the laudable precedent set by the Human Genome Project is less likely to be adopted by other data gathering efforts. This would be a loss to both science and society.

References and Notes

1. The Bermuda Conventions are described in http:/ / www.wellcome.ac.uk/ en/ 1/ biopoldat.html and http:/ / www.nhgri.nih.gov/ Grant_info/ Funding/ Statements/ RFA/ data_release.html.

2. See Opinion, Nature 405, 719 (2000); C. Macilwain, Nature 405, 601 (2000); Gottlieb et al., Nature 406, 121 (2000) for a discussion of the dispute between the parasite sequence producers and third parties who want to publish analyses based on the preliminary data.

3. Nature 405, 719 (2000). "We believe that genomics databases, like preprint servers and conferences, represent a form of intra-community networking from which all researchers benefit. Nature does not count them as prior publications. This policy applies not only to raw sequence data but also to 'annotations': proposals for the functions of the genes in the database. These often represent substantial pieces of research in themselves."

4. R.J.N. Allcock et al, Immunology Today 21, 328 (2000); C.Y. Yu et al, Immunology Today 21, 320 (2000). The following acknowledgement appears at the bottom of a table in which all of the genes from the available mouse MHC sequence were delineated: "Most of the genomic sequence was generated and annotated by Dr. Lee Rowen and her colleagues at the University of Washington."

5. The data release policies for the Hubble Space Telescope are detailed at http:/ / www.stsci.edu/ ftp/ proposer/ cycle10/ cp/ cp10-5.html#pgfId-1818439. We thank Mike Hauser, Deputy Director of the Space Telescope Science Institute, for explaining these policies to us.

6. For example, the malaria genome sequencing project's web site http:/ / e2kroos.cis.upenn.edu/ release_oq.html states that: "Given that this information is considered preliminary and may contain inaccuracies, it is expected that no one will publish analyses, based on the preliminary data contained in this web site, of genes on a whole chromosome or genome scale

without permission of this sequencing center. This sequencing center plans on publishing the completed and annotated sequence in a peer-reviewed journal."

L. Rowen and L. Hood are at the Institute for Systems Biology, Seattle, WA 98105, USA. G.K.S. Wong and R.P. Lane are at the Department of Genetics and Department of Molecular Biotechnology, respectively, at the University of Washington, Seattle, WA 98195 USA.