*How does immediate sequence release affect commercial exploitation?* As observed by HUGO (6), it is important that the necessary incentives for commercial investment are preserved so that the development of products (particularly diagnostic tests and therapeutic agents) can continue without unduly interfering with scientific research. There has been much debate on the feasibility and advisability of protecting commercial interests by patenting gene sequences. It is now widely accepted that the patenting of raw human genomic DNA sequence or partial or complete gene sequences of unknown function is inappropriate (6–8). Such action might well discourage further research and development by others, for fear that future inventions downstream of the gene sequence itself could not be adequately protected. Given that raw human genomic sequence does not fulfill the requirement of patentability under existing patent law (that is, it must be novel, nonobvious, and have demonstrable utility), the best course of action is to release it freely. As a result, the value of the sequence will increase as it accrues additional information from other public domain sources, leading to the definition of novel gene structures, regulatory mechanisms, and functions. Free release of sequence data will also encourage exploitation by a maximum number of commercial and academic centers that are keen to compete in the development of new therapeutic agents. Encouraging such competition is healthy: The best possible advances, protected by the most appropriate well-defined patents, are more likely to emerge in a nonexclusive environment rather than in an environment in which a single company maintains an exclusive position to develop useful health care products at its own pace using its own preferred approaches. It is therefore vital that genomic sequence data are made immediately and freely available in the public domain to maximize their benefit to society.

## REFERENCES AND NOTES


1. The first International Strategy Meeting on Human Genome Sequencing, organized by the Wellcome Trust, was held in Bermuda, 25 to 28 February 1996. Participants included representatives of laboratories involved in human genome sequencing and of funding agencies, who met to discuss strategy, progress and plans, policies for data release, and the implications of such policies. The "Bermuda statement" was endorsed unanimously by all participants. See *Human Genome News* **7** (no. 6), 19 (1996).
2. The provision of genomic information to the public in three stages—as sequence-ready maps, as assembled shotgun sequence data, and as finished and annotated consensus sequence of each bacterial clone—was first practiced from the outset of the *Caenorhabditis elegans* genome project by the groups of R. Waterston (Washington University, St. Louis) and J. Sulston (the Sanger Centre). The same practice has been implemented for the release of human genomic sequence data at both centers. This data release policy is endorsed by both the Wellcome Trust and the National Institutes of Health. Other centers are also practicing or planning forms of data release, including the Whitehead Institute, The Institute for Genomic Research (TIGR), Baylor College of Medicine, and others. See also (*8*); E. Marshall and E. Pennisi, *Science* **272**, 188 (1996); National Science Council, *Report of the Committee on Mapping and Sequencing the Human Genome* (National Academy Press, Washington, DC, 1988).
3. In most cases, the preliminary assembled shotgun sequence data provide sequence representing around 90% of insert of the bacterial clone in a few large contiguous sequences that are virtually free of artifacts. Ongoing refinement of the shotgun strategy and sequencing biochemistry is resulting in further improvements to the quality and coverage of the initial assembled sequence. In addition to the prerelease of unfinished sequence at the FTP sites of the sequencing centers, the National Center for Biotechnology Information and the European Bioinformatics Institute are planning to provide centralized access to the unfinished sequence in the public domain.
4. R. Wooster *et al.*, *Nature* **378**, 789 (1995); S. Tavtigian *et al.*, *Nature Genet.* **12**, 333 (1996).
5. The proposal to determine the genomic sequence representing more than 90% of the human genome at an accuracy of 99.95% or greater was reported by E. Marshall, *Science* **267**, 783 (1995). The agreement to aim for completion of contiguous sequence at an accuracy of 99.99% (equivalent to the standard achieved for *Caenorhabditis elegans*) was promoted at the Bermuda meeting [see (*1*)].
6. Human Genome Organization Statement on Patenting of DNA Sequences, 1995.
7. BioIndustries Association, *The Patentability of Human Genes* (1995).
8. National Center for Human Genome Research, Policy on Availability and Patenting of Human Genomic DNA Sequence Produced by NCHGR Pilot Projects, 9 April 1996.
9. Written on behalf of the Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK, and Genome Sequencing Center, Washington University, St. Louis, MO 63108, USA.


# Should Non–Peer-Reviewed Raw DNA Sequence Data Release Be Forced on the Scientific Community?


Mark D. Adams and J. Craig Venter


The ability to sequence DNA accurately and efficiently has revolutionized biology and medicine and has ushered in the new era of genomic science, the study of genes and genomes. An argument has been made by some that the inherent value of DNA sequence data is so great, regardless of quality, that it should be downloaded nightly onto Internet sites (*1*). Coupled with this is the notion that it is somehow inappropriate for the scientific teams that generate sequence information to extract scientific value from their data before releasing these data to others. This proposal represents a radical departure from the way in which scientific research is traditionally conducted and should raise concerns in the scientific community.

The Human Genome Project has seen a wide range of conduct in the publication of research findings, particularly physical and genetic map resources as they relate to the highly competitive field of human genetics. These variations in data release prompted the National Institutes of Health (NIH) and the U.S. Department of Energy (DOE) to review data release policies and to set standards for genomic research. The current official NIH-DOE genome data release policy requires scientists to release their data within 6 months of generation (*2*). However, in conjunction with the awarding of pilot project grants for human genome sequencing, NIH asked each awardee to abide by a plan for the rapid release of data, essentially as proposed by the Sanger Centre (*3*).

At first glance, there might seem to be few, if any, compelling reasons for all genome research labs not to adopt the policy to immediately download sequence data directly from an Applied Biosystems sequencer to an Internet site (or to do so after a swift and cursory form of automated quality control, such as vector removal or partial assembly). After all, the modern molecular biologist is sophisticated enough to analyze unfinished DNA sequence data and incorporate it where appropriate into ongoing research projects, and much of the DNA sequence data available at genome laboratories' Internet sites comes with a user-beware warning and, in some cases, restrictions on use (*4*). However, this policy has not yet been subjected to a rigorous test of its true utility and benefit to the scientific community at large.

We believe there are substantial reasons why scientists should be cautious about using or releasing data and results that have been neither peer-reviewed nor extensively self-reviewed. Although we do not object to the policy of nightly data release adopted by some genome centers, we do object to having these terms applied across the board to all labs involved in genome research.

*Publication versus data release.* The peer-review process has been a fundamental part


The authors are at The Institute for Genomic Research (TIGR), Medical Center Drive, Rockville, MD 20850, USA.

of modern science and provides a mechanism for the critical review and evaluation of scientific work. This process distinguishes publication in a scientific journal from other forms of communication (such as news releases and public talks) and implies that certain criteria have been met. The development of alternative means of information distribution (on the Internet by means of FTP and the World Wide Web) has simplified the process of making data widely available and more easily interrogated. However, even though journals such as the *Journal of Biological Chemistry* are now available on the Internet, this change in distribution has not compromised their process of peer review.

We are concerned that raw or finished sequence data release without accompanying scientific publications will become the new end point of some groups' contributions to genomics. This could happen either intentionally or inadvertently. The policies of most scientific journals preclude publication of papers if the data already have been made available. For example, in the publication of our three genome analysis articles in *Science* (5), it was made clear to us that any release of the sequence information before the date of publication would violate *Science*'s embargo policy. Other journals either have rejected genome sequence articles from other groups or have made one-time major exceptions to existing policies. The scientific journals need to develop policies that either deal with prepublication release of data or make it clear to the community that current policies will remain in effect. Will peer reviewers in the future decide that the novelty of a manuscript has been compromised if the annotated sequence is fully available on the Internet?

Journals and the scientific community also need to develop standards with regard to data analysis papers where the authors are not associated with primary data generation. For example, the yeast sequencing consortium released the nearly completed genome sequence this year, and while their publications on the genome analysis were in preparation, an independent group submitted an article based on rapid analysis of the consortium's data. It might be argued that once the data are placed in public databases, it is every scientist for him- or herself. However, if those are the rules, then why would any scientists want to abdicate the fundamental analysis of their own data to others?

One unintended result of the release of unfinished, unannotated sequence to Internet sites is that many individuals rely on the public search services, such as the BLAST network service at the National Center for Biotechnology Information, to begin to understand what is encoded by each sequence. In self-defense, the public databases are considering accepting "prerelease" sequence and performing automated annotation themselves to reduce the load on the BLAST search services (6). This would not only place unfinished sequence alongside finished sequence in the public databases, it would clearly add to the burden and cost of management of public databases and would certainly add to the confusion of many people who are attempting to deal with genomic sequence data.

*Standards of quality and completeness.* Genome research centers use a wide variety of approaches to DNA sequencing and analysis. Some appear to believe that once the DNA sequence from a clone, region, or genome is obtained, the process is finished. At TIGR, DNA sequence is the raw data that begins our analysis of genome content, organization, comparative genomics, and evolution (5). DNA sequencing is not a mindless, robotic task that produces foolproof data. Extensive quality control measures are required at every step from initial library construction to gap closure, final assembly, and editing. We use database searches and other database comparisons, gene predictions, and frame-shift analysis as part of the final editing process. If, for example, we find a possible frame shift in a predicted gene, we reanalyze all the raw sequence data from that region of the genome and perform additional sequencing (with a different chemistry if necessary) to resolve any ambiguity. Despite the considerable effort put into sequence determination, we consider the sequence to be less than perfect.

Many have argued that accurate sequence data are not necessary for gene finding, and the success of the expressed sequence tag (EST) method (7) supports this argument. However, the worldwide effort to produce the first reference human genome sequence should have as its focus the goal of producing highly accurate, complete sequence information. To achieve this goal, it is necessary to use robust quality control and quality assurance procedures for data production, and the data must be in a user-friendly form for the scientific community. If genome sequence data differ from other forms of biological data, it is because they are fundamental data that will be used by scientists for centuries. Therefore, making diligent efforts to have the highest quality data possible is far more important than rushing the data out to save a few weeks.

Of what value is DNA sequence information alone, without a description of what it encodes? After all, the purpose of the genome project is not simply to print the book, but to elicit what meaning we can from the words and phrases. It is necessary that an attempt be made to define genes, regulatory sites, and repetitive elements and to provide the documentation for how the final sequence structure was determined (if questionable areas are present). Annotation is an integral and inseparable part of genomic sequencing, and it is the scientific justification for pursuing the genome project in academic laboratories rather than contracting it out to the lowest bidder.

*Danger of limited quality control.* Perhaps the most important issue raised by rapid data release is the danger to ongoing scientific research if quality controls, including peer review, are bypassed. The sequencing of the human, *Arabidopsis*, and other genomes will likely use bacterial or yeast artificial chromosomes (BACs or YACs) as sequencing substrates. Each clone to be sequenced must be first purified from a background of *Escherichia coli* or yeast genomic DNA as part of the subcloning process, which invariably results in some fraction of the primary sequence reads having a non–target species origin. Although these are relatively easy to screen out if done properly, and will become even easier once the *E. coli* genome is complete, it is not so easy to identify cases in which an incorrect clone is present as a contaminant—something that would be identified in the quality control process of finishing a project. For example, in an ongoing TIGR project to sequence the *Streptococcus pneumoniae* genome, we received a DNA sample purified from a clinical isolate of *S. pneumoniae* to construct a random small insert library. Preliminary database searches revealed sequences of *S. pneumoniae* origin, and the random phase of sequencing was begun in earnest. As sequencing progressed, the data began to assemble into two separate genomes, one with DNA sequence that clearly represented *S. pneumoniae* and another with sequence most closely related to a *S. viridans* species. Had we followed the proposal to nightly download raw DNA sequence from this project, the scientific community would have been misled into assuming that all of the sequences were from *S. pneumoniae*.

As another example, in 1993, one group submitted a set of "human" EST data to the European Molecular Biology Laboratory without adequate quality control; these ESTs were later found to include several thousand sequences of yeast and bacterial origin (8). These sequences remain in GenBank and are still annotated as being of human origin. The continued presence of these sequences in GenBank illustrates the difficulty of undoing an error when incorrect or misannotated sequence data are released. These will no longer be isolated incidents when a variety of large-scale DNA sequencing projects are under way. We question whether the benefits

of immediate data availability outweigh the potential serious errors that raw data may contain and the consequent waste of time, intellectual energy, and resources that the community will necessarily have to expend to bring scientific clarity to the data.

*Looking forward, not backward.* Fewer than 600 million base pairs of DNA sequence reside in the public databases, most of it redundant. If the human genome is to be completely sequenced over the next 7 to 10 years, then in each of these years the human genome sequencing community must produce accurate genomic sequence data and analysis equivalent to the sum of all DNA sequencing done to date. Such an effort would require the finishing and publication of, on average, 500 million base pairs of sequence each year—the equivalent of the *E. coli* genome being published every 2 days for the next 7 years. The nightly addition of the raw, unedited data to Web sites would double or triple the amount of information to be processed. This scenario considers only the Human Genome Project; however, projects are under way or planned for a large number of other genomes, including mouse, *Drosophila*, plants, parasites, and microbes. Given the enormous scope of the genome project, we feel that the sequencing labs need to focus on ensuring the highest quality data, analysis, and scientific interpretation, made available as soon as practicable upon completion and published in a timely fashion in peer-reviewed journals. In fact, early release of unedited, unfinished data may be detrimental to small molecular biology labs, which do not have the resources or computational tools to deal with the deluge of information.

Despite its tone of fairness, the argument for daily data release suggests indifference toward the intellectual effort that the scientific research community has set as a standard for itself in the publication and release of its work. Scientific custom has held that the scientist should be allowed to communicate to the research community what was achieved and how it was done, to analyze and comment, not only so that careful critical evaluation can be made, but also out of respect for the researcher and the achievement. Some have argued that genome sequencing is different from other scientific pursuits in that it is a public service—but what taxpayer-funded research is not a public service?

We propose that for genome sequencing projects that cannot be completed in 12 to 18 months, finished sequence data (of, for example, BAC clones) should be made available to the wider community immediately, without restriction or delay, as soon as these data have passed a series of rigorous quality control checks and have been annotated. Within a reasonable interval, these releases of data should be followed by complete scientific papers that not only describe the methods used for data generation and analysis, but also attempt to place the data in a broader biological context. We hope that the scientific journals, the scientific community, and the funding agencies will be tolerant or even encouraging of a variety of approaches to data generation, interpretation, and scientific publication to advance this exciting field of genomics.

## REFERENCES AND NOTES


1. E. Marshall, *Science* **272**, 477 (1996).
2. The NIH-DOE 6-month policy (NIH-DOE Guidelines for Access to Mapping and Sequencing Data and Material Resources) can be found at http://www.nchgr.nih.gov:80/Grant_info/Funding/Statements/data_release.html.
3. The National Center for Human Genome Research rapid release policy (Policy on Availability and Patenting of Human Genomic DNA Sequence Produced by NCHGR Pilot Projects, 9 April 1996) can be found at http://www.nchgr.nih.gov/Grant_info/Funding/Statements/patenting.html.
4. The Washington University Genome Sequencing Center FTP site (ftp://genome.wustl.edu/pub/gsc1/sequence/README) reads: "This archive site contains prerelease versions of the data, and we ask that you keep the following in mind:

   1) These data should be used for internal research purposes only and not for redistribution. If other researchers request a copy of the sequence, please have them contact the *Caenorhabditis elegans* sequencing project directly.

   2) This is not the final version of the sequence and may contain errors, this is especially true of the data in the shotgun and finishing directories. However, we feel it is of sufficient interest to be prereleased.

   3) Please consult with us before publication of any results related to this sequence until this sequence has been officially released. After it has been released, you are welcome to publish the results of your research as long as Washington University and the *C. elegans* sequencing project are acknowledged appropriately."
5. R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995); C. M. Fraser *et al.*, *ibid.* **270**, 397 (1995); C. J. Bult *et al.*, *ibid.* **273**, 1058 (1996).
6. D. Lipman, personal communication.
7. M. D. Adams *et al.*, *Science* **252**, 1651 (1991); M. D. Adams *et al.*, *Nature* **377** (suppl.), 3 (1995).
8. O. White *et al.*, *Nucleic Acids Res.* **21**, 3829 (1993); C. Anderson, *Science* **259**, 1685 (1993).


# The New Genomics: Global Views of Biology


Eric S. Lander


The Human Genome Project was designed as a three-step program to produce genetic maps, physical maps, and, finally, the complete nucleotide sequence map of the human chromosomes. In the past year, the first two milestones have essentially been reached (*1*) and pilot sequencing projects have begun with the aim of increasing speed and efficiency. Although only 1% of the human genome has been sequenced so far, there is growing confidence that the annual production rate can climb over the next 3 years to more than 500 megabases (Mb) worldwide—ensuring that the goal will be comfortably reached by the original, projection of 2005. The mouse, the leading biomedical model system, can likely be sequenced in parallel, although funding has not yet been committed.

With success in sight, thoughts are already turning to what should come next. The answer depends in part on how one understands the significance of the Human Genome Project. Commentators have sought to set the project in historical context by likening it to the Holy Grail, the Manhattan Project, and the moon shot. Each analogy is rich with implications about the appropriate follow-up. However, none of these precedents rings true.

Rather, the Human Genome Project is best understood as the 20th century's version of the discovery and consolidation of the periodic table. In the period from 1869 to 1889, chemists realized that it was possible to systematically enumerate all atoms and to arrange them in an array that captured their similarities and differences. The building blocks of chemistry were rendered finite, and the predictability of matter gave rise to the chemical industry on one hand and the theory of quantum mechanics on the other.

The Human Genome Project aims to produce biology's periodic table—not 100 elements, but 100,000 genes; not a rectangle reflecting electron valences, but a tree structure depicting ancestral and functional affinities among the human genes. The biological periodic table will make it possible to define unique "signatures" for each building block. Just as chemists can recognize atoms by mass and charge alone, biologists will be able to build detectors that allow each gene to be recognized from 20 well-chosen nucleotides or each protein from a distinctive fragment. Molecular biology has tended to


The author is at the Whitehead Institute for Biomedical Research and the Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142, USA. E-mail: lander@genome.wi.mit.edu