Genomic Sequence Information Should Be Released Immediately and Freely in the Public Domain

David R. Bentley

Progress in understanding is best achieved by the free exchange of knowledge and ideas. To understand the biology of humans and other organisms through genome interpretation, all genomic DNA sequence information should be "freely available and in the public domain in order to encourage research and development and to maximise its benefit to society" (*1*). This statement (applied to human sequence generated by large-scale sequencing centers) was unanimously endorsed by participants at the International Strategy Meeting on Human Genome Sequencing, held in February of this year. The key question in the current debate is whether to immediately release sequence information and, if so, in what form. The answer presented in this article is yes. The finished sequence should be released directly upon completion. Furthermore, there should be an earlier prerelease of unfinished sequence and additional mapping information. This is required to optimize coordination, independent checking, and exploitation in both academic and commercial laboratories (*2*).

Immediate sequence (and map) release permits coordination. Genomic sequence is typically produced in 40- to 200-kilobase segments, each of which is represented by a single bacterial clone [for example, a cosmid, fosmid, bacterial artificial chromosome (BAC), or P1 artificial chromosome (PAC)]. DNA from the clone is prepared and subcloned, 800 to 2000 templates are sequenced and assembled in a random shotgun phase, and ambiguities are resolved in a final directed phase ("finishing"). The completed consensus sequence is then annotated and submitted to the public database. The entire process can be done in 4 to 6 weeks, but can take longer, depending on the problems encountered during finishing of each clone. At large centers, 500 or more clones may be at intermediate stages of the process at any given time. To optimize coordination, it is therefore important to make the status of each clone as visible as possible to the rest of the world. It is not adequate to rely purely on release of the finished sequence of each clone as an indicator of progress; the risk of accidental duplication is high and any duplication is costly. This risk is minimized by providing regularly updated maps of all clones as soon as they enter the process, if not earlier still. Actual progress is then monitored conveniently by prerelease of the unfinished sequence (that is, the assembled shotgun sequence data) of each clone. This informal prerelease also provides most of the sequence information to the public promptly, before the stage that is subject to possible delays (the "finishing" stage).

Unfinished sequence is of immediate value to others and is not misleading. The biochemical process of sequence determination is extremely robust, and each template provides raw data of high quality. As a routine precaution, raw shotgun data are assembled and orphan reads (along with vector sequences and any reads of poor quality) are removed, thus eliminating virtually all artifacts before the prerelease. The resulting sequence data are thus of defined quality and contain information of sufficient accuracy for many biological and genetic studies (*3*). For example, the availability of the unfinished genomic sequence allowed the determination of the complete structure of the *BRCA2* gene as well as the detection of mutations that provided conclusive proof of the association of this gene with familial breast cancer (*4*).

Unfinished sequence does not clutter public databases. It is an internationally agreed aim that human genomic sequence will be finished to high accuracy (99.99%) (*5*). Given this commitment, the unfinished sequence is not a substitute for the finished product but constitutes a transient, dynamic buffer of finite size. The sequencing center has the responsibility of ensuring that sequence does not languish in the unfinished category (which would inflate the buffer unnecessarily). As the buffer of unfinished sequence is finite, it is possible to set up mechanisms to handle the data adequately. The situation would be quite different if the decision had been taken to skim the entire genome first. This would have resulted in an unmanageable amount of information of much lower intrinsic value requiring long-term storage.

Does immediate sequence release promote or hamper its use? The major aim is to promote maximum accessibility of the human genome sequence for interpretation and exploitation. These activities should flourish in both the academic and commercial sectors. Immediate release of sequence provides valuable data as quickly as possible to laboratories focusing on specific biological or clinical problems (usually associated with one or more limited regions of the genome). Historically, the absence of a reference map or sequence of the genome has meant that mapping and sequencing forms a major part of the effort of the researchers undertaking such a targeted study. In some cases, patents have been filed on gene sequences in an attempt to establish exclusive ownership and thus protect biotechnological inventions arising from knowledge of the gene sequence. Alternatively, the interests of the funding agency and participating laboratories have been protected by maintaining confidentiality.

We have now entered a transition phase in which the mapping and sequencing is increasingly being carried out in large-scale centers. However, because only a fraction of the human genome sequence (currently around 1%) has been determined as yet, collaborations are being established between the sequencing centers and groups pursuing specific targets in regions not yet sequenced. In this situation, the data release policy of the sequencing center cannot be compromised to protect the interests of the targeted study: If small parts of the emerging genome sequence were held back for this reason, the sequencing centers would be exercising control in using the resources of the large-scale public domain program to confer a selective advantage on certain laboratories over other groups. This is not acceptable. There must be no opportunity for selective retention. It can only be avoided effectively by adopting a strict policy of immediate data release along the lines defined above, so that a transparent view of the process of data generation is provided to the rest of the world and the data are made available freely and simultaneously to all. Furthermore, imposing any form of selective access of the data to certain groups will tend to impede progress. Identification of a target is not achieved from the genomic sequence alone. It requires expert interpretation of sequence by specialist knowledge derived, for example, from patient collections, biochemical knowledge, or further experiments using complementary DNA analysis, exon trapping, or mutation screening techniques. Withholding the genomic sequence ingredient from any academic or commercial laboratory with such knowledge impedes scientific progress and is not in the international public interest.

The author is at the Sanger Centre, Wellcome Trust Genome Campus, Hinxton Hall, Hinxton, Cambridge CB10 1SA, UK.

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*).

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

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