The End of the Beginning: The Race to Begin Human Genome Sequencing

Mark Boguski, Aravinda Chakravarti, Richard Gibbs, et al.

Genome Res. 1996 6: 771-772
Access the most recent version at doi:10.1101/gr.6.9.771

EDITORIAL

The End of the Beginning: The Race to Begin Human Genome Sequencing

The Human Genome Project has different meanings to different scientists, but the essence of it has always been the drive to obtain the total reference genome sequence for humans. In the first five years of its official existence, the Human Genome Project has concentrated on the development of genetic and physical maps that would ultimately support genome sequencing in the human and a few model organisms. The comprehensive genetic and physical maps, and the associated reagents, have already found use amongst a large, diverse, and enthusiastic group of geneticists who are using them to find genes for both simple and complex hereditary traits. One revolution is already under way: Geneticists are increasingly turning their attention to the dissection of complex traits per se using maps, rather than using the classical approach of analyzing only Mendelian traits associated with complex phenotypes. It is now time to set attention to the main task, that of large-scale human genome sequencing, and that is precisely what has been rumbling in the background.

In the past year there has been a clear transition to the thought that current technologies are adequate to begin the task of human genome sequencing, since radically new and revolutionary sequencing technologies are not evident and there is greater danger in not having the sequence sooner. It is now increasingly clear that a further revolution in biology will arise from the sequence itself, not necessarily from the technologies used to obtain it. Consistent with this view, some countries have already initiated large-scale human genome sequencing programs. Without much fanfare, the race to finish the human sequence has begun.

The transition to fulfilling the sequencing goals has been largely prompted by a proposal made by Robert Waterston at Washington University in the United States and John Sulston at the Sanger Centre in the United Kingdom, based on their experience in genome sequencing of *Caenorhabditis elegans*. Agencies funding the Human Genome Project seem to agree as much. In the United States, the National Institutes of Health (NIH) has funded six groups to establish pilot programs and develop production groups capable of sequencing large tracts of human DNA at high accuracy and low cost. In the United Kingdom, the Sanger Centre has been funded by the Wellcome Trust to support the production of first-generation sequence for up to one-third of the human genome. And there are evolving sequencing efforts in Germany and France.

Re-establishing human genome sequencing as *the* current priority has led to several discussions in the sequencing community on specific issues that require immediate attention: these relate to sequencing strategy, data quality, and data release. Over the last six months, three meetings have examined these issues. In February, the Wellcome Trust sponsored a meeting in Bermuda, attended by the major practitioners of large-scale sequencing, with the objective of establishing policies for data release and implementation of a system for coordinating sequencing targets. A general policy for coordinating efforts to prevent unnecessary duplication of sequencing was endorsed, but, importantly, there was widespread consensus on the desirability of "immediate and free" data release. A second meeting, at the NIH in March, emphasized that a premium should be placed on generating high-quality data. The ways in which such a standard would be maintained were less clear, nor were practical mechanisms for monitoring data quality established. A third gathering, at the annual Cold Spring Harbor Genome Mapping and Sequencing meeting in May, clarified the increasing convergence on strategies for performing large-scale genome sequencing. In fact, the differences in strategies are all subtle since virtually all major sequencing groups use fluorescent-based four-color chemistries and some variation of methods that employ both random shotgun and directed phases.

These meetings have led to a consensus view that the general strategies for sequencing are largely established, even though numerous technical details need to be resolved. With this in mind, a few groups have made strong commitments to sequence specific regions of the genome and a number of consortia have been formed. For some groups the path forward is clear, with little left to decide; the major impediment is obtaining funds to operate large-scale sequencing programs. These scientists argue that from this activity itself will emerge not only the required experience but also all of the required information

6:771–772 ©1996 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/96 $5.00

systems for handling the sequence data. Given this optimistic mood, we can expect exciting accomplishments over the next several years and the birth of a new sequence-based experimental era. Nevertheless, at least three issues require further detailed discussion as the sequencing phase proceeds—the generation of sequence-ready maps, the definition of sequence data quality and accuracy, and the mechanism of data dissemination.

Sequence-ready Maps

Progress in constructing physical maps of human chromosomes represents a highly successful component of the first phase of the Human Genome Project. However, at present available physical maps are not optimal for supporting production-level sequencing. Current physical maps of most human chromosomes have a resolution of 100–200 kb and are based largely on YAC clones. This resolution is insufficient for obtaining the desired cloned coverage in any of the preferable bacterial-based BACs, PACs, or P1 systems. Additionally, there is a shortage of well-characterized, highly redundant, large-insert bacterial-based libraries that are needed to construct the maps in the first place. These two shortcomings are not trivial since they can greatly affect the attainment of the sequencing goals, both in the short and long term. Thus, the considerable attention currently being paid to improving sequencing throughput should be complemented by an effort to build the high-resolution maps necessary to support genome sequencing.

Data Quality Standards

Near-perfect sequence data have been defined as less than one error per 10,000 nucleotides sequenced. There are arguments for encouraging sequence production at less than this near-perfect standard because the cost could decrease drastically for only a marginal decrease in quality. Concomitantly, this would lead to an increase in the rate of data production. However, without objective criteria for gauging sequence quality, a set of systematic rules by which quality might be reduced cannot be defined. The absence of such objective criteria could, in the extreme, result in the production of vast amounts of low-quality data, perhaps without necessarily achieving cost savings. This is a particularly grim possibility to the community at large since too much resequencing of the genome may be necessary if the initially generated sequence were of low quality. We emphasize the need for a concerted effort to define objective sequence quality standards and to relate the efficiency of sequence production with cost and quality.

Data Dissemination and Annotation

A central philosophy guiding genome researchers has been the rapid dissemination of all data; thus, the oft-cited goal of immediate and free release of sequence data is highly laudable. One recent proposal is for large-scale sequencing centers to distribute prefinished sequence data, as they emerge, on the World Wide Web, followed by public data base submissions at a later date after corrections, revisions, and annotations have been made. What appears exemplary at first sight in fact bodes trouble for the majority of end-users. The great majority of biologists depend upon public data bases for comprehensive and up-to-date sequence information, and simply have neither the time nor the resources to visit dozens of Web sites frequently to locate and analyze new, unannotated data. Therefore, as a practical matter, quick unannotated releases of data on the Web will likely benefit only large groups with considerable personnel and technical resources, particularly commercial entities. There are yet other interesting and important issues that remain unsolved: What types of annotation are necessary to make the data most useful for the greatest majority of biologists? Who will be responsible for the long-term annotation and maintenance of the reference sequence? What will constitute a publication when most of the traditional kinds of sequence discoveries will be documented by automated sequence analysis? We all believe that the availability of a reference human genome sequence will impact greatly on how we will do genetics and biology. To ensure that all scientists can participate on an equal footing, solutions to the management, maintenance, and dissemination of data are a very high priority.

These are challenging issues, but they all require solution, and this can come about only from critical scientific thought and attention. Investigation of these issues, in conjunction with the data generated during the initial escalation of human genomic sequencing, will represent key areas in genome research over the coming months. We look forward to seeing the fruits of this important effort, and their application by geneticists worldwide, in the pages of *Genome Research.*

Mark Boguski
Aravinda Chakravarti
Richard Gibbs
Eric Green
Richard M. Myers