A "Quality-First" Credo for the Human Genome Project

Maynard Olson and Phil Green

Genome Res. 1998 8: 414-415
Access the most recent version at doi:10.1101/gr.8.5.414

A ‘‘Quality-First’’ Credo for the Human Genome Project

Maynard Olson[1] and Phil Green[2,3]

[1]Departments of Medicine and Genetics and [2]Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195 USA

The Human Genome Project is lurching toward large-scale genomic sequencing. Although we find the arguments in favor of this path compelling, there remain sobering uncertainties about the cost of producing genomic sequence on a gigabase-pair scale, the rate at which the needed sequencing capacity can be developed, and the extent to which compromises will be required in the quality of the final product. Of course, it is these very uncertainties that make the sequencing of the human genome a scientific and managerial challenge worthy of the committed attention of many hard-working and talented people. Policy decisions are now being made that will greatly affect how this talent is deployed during the years ahead. We argue here, on both scientific and managerial grounds, that it is essential that the Human Genome Project adopt a ‘‘quality-first’’ credo.

Scientific arguments for quality are rooted in a view of how the sequence will be used. The most common uses of reference sequences involve comparisons with other data. A vast number of sequences, derived from many sources—human and nonhuman—will be compared with the human reference sequence by future scientists. We should not prejudge either the comparison methods or the questions that the comparisons will address. Our goal should be to produce a reference sequence sufficiently good that comparisons made with it will rarely fail or produce misleading results because the reference sequence is inaccurate or incomplete. The sequence differences that are detected should, in nearly all cases, reflect real biological effects or limitations on future experimental or theoretical methods—not errors in the reference data. Viewed from this perspective, the commonly accepted base-pair accuracy goal of 0.9999 is a minimum quality target. Current data suggest that two randomly chosen human genomes are likely to show sequence similarity at the 0.999 level. Hence, even at a 10^{-4} error rate, on the order of 10% of the discrepancies with the reference sequence that will arise during intraspecies comparisons will be due to errors in the reference data. Each such artifactual discrepancy will lower the efficiency or diminish the discrimination power of future scientific inquiries.

Base-pair accuracy is only one dimension of sequence quality. Indeed, with current technology it is the least problematic one. The issues of sequence contiguity, clone validation, and assembly checking are both more challenging and more crucial from the perspective of sequence users. Contiguity has proven particularly elusive. Highly fragmented sequence, such as virtually all of the human genomic sequence currently in GenBank, is difficult either to use or to validate. At a contiguity standard of 100 kb, which a substantial portion of current GenBank entries fail to meet, many human genes are not even in a single piece. At a 1-Mb standard, candidate regions associated with positional cloning projects typically still contain gaps. Gaps not only leave users of the data uncertain as to the precise genomic origins and biological content of particular sequences, but they also make the sequences difficult to validate. Problems of all types accumulate in data adjacent to gaps because the self-consistency tests on which sequence validation depends fail in these regions. Gap sizes are difficult to measure reliably and often prove larger than expected. For sequence segments appreciably below 1 Mb in size, the relative orientations of adjacent segments—and even the ordering of the segments—can be difficult to determine.

Clone validation involves demonstrating that a particular recombinant DNA molecule is a faithful replica of the genome from which it was derived. At present, the only practical method for validating large-insert clones is to check their consistency with other overlapping clones from the same region. This process can be carried out at various resolutions, ranging from restriction-fragment mapping to complete sequencing. Clone aberrations that occur reproducibly in most or all clones cannot be detected by these methods. However, experience indicates that such aberrations are rare. A more serious problem is that the self-consistency tests are difficult to carry out rigorously. Because nearly every clone in current libraries has a unique pair of end points and a given clone can come from either of two human haplotypes, self-consistency checking is often confounded by mapping errors and haplotype differences.

Assembly checking involves experimental tests to ensure that sequence assembly programs have correctly melded individual sequence tracts to form the composite sequence. The most widely used method of assembly checking is to compare restriction digests of a recombinant DNA molecule with the pattern of fragment sizes predicted from the assembled sequence. Although we suspect that undetected clone validation errors are more common than undetected assembly errors, some trends in this area bear watching. For example, the increasing reliance on shotgun sampling of relatively large clones (e.g., 150- to 200-kb BAC clones) both increases the likelihood of assembly errors and makes them more difficult to detect.

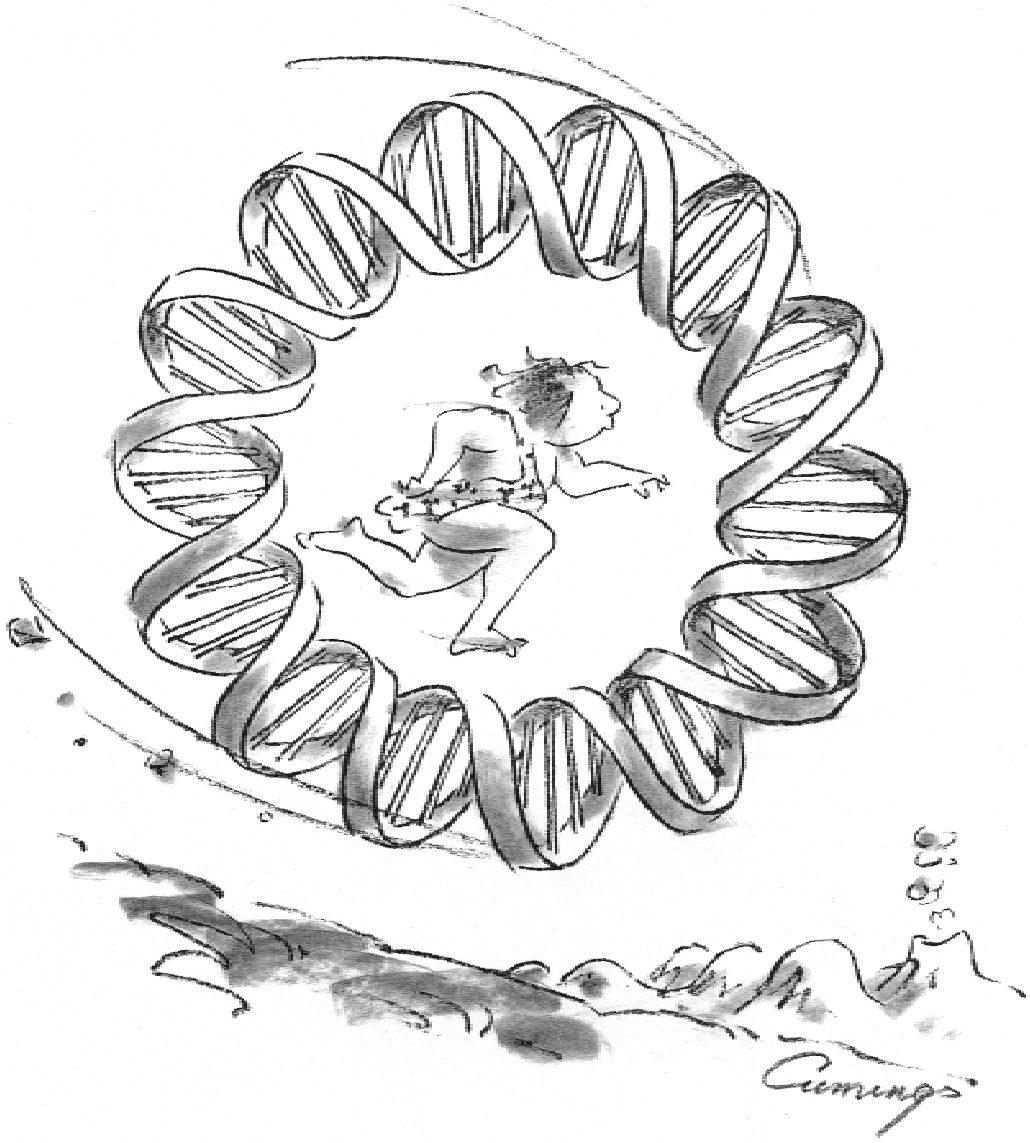
As this brief survey indicates, there are significant technical challenges associated with assessing the quality of genomic sequence data. However, these challenges can and must be met. In addition to the scientific rationale for

[3]Corresponding author.
E-MAIL phg@u.washington.edu; FAX (206) 685-7344.

8:414–415 ©1998 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/98 $5.00; www.genome.org

meeting them, which was discussed above, there is a strong managerial rationale for adopting a quality-first credo for the Human Genome Project. The central managerial challenges now facing the Project are to choose strong genome centers to do the sequencing, to encourage them to implement well chosen production methods on a large scale, and to control costs. None of these challenges can be rationally approached without adopting stringent quality standards now. Decisions about the relative merits of different data production processes can only be made in the context of a well-defined goal. The alternative is to build raw data collection capacity with no assurance that this capacity can ever produce high-quality sequence at a practical cost.

Cost will almost certainly emerge as the main driver of technical, managerial, and policy decisions about large-scale sequencing. Costs cannot be controlled until they are reliably measured, and they cannot be measured in an environment that encourages indefinite deferral of large and unknown costs. It will only become practical to define costs meaningfully if we shorten the time lag between raw-data collection and the production of finished sequence meeting high contiguity, accuracy, and genome-fidelity standards. In such an environment, constructive competition between data production centers will lead to meaningful cost reductions. It will also provide solid experience about cost–quality tradeoffs. We may ultimately find that aspects of our current quality standards have unfavorable cost–benefit ratios. However, we should start by aiming high. In this way, we will push the technology in the right directions, put the Human Genome Project on solid managerial footing, and maximize the chance that the Project's legacy will be a sequence that will meet the test of time.



Rendezvous with Destiny