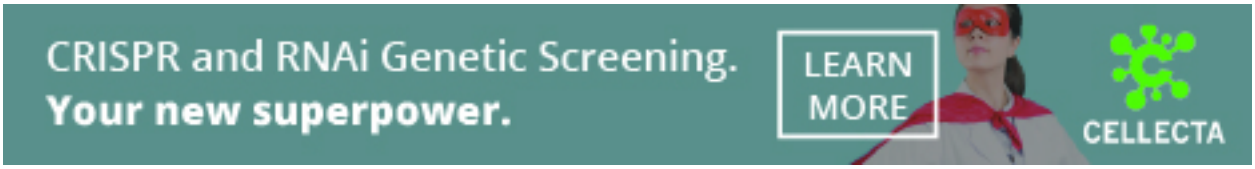
Assessing the Quality of the DNA Sequence from The Human Genome Project

Adam Felsenfeld, Jane Peterson, Jeffery Schloss, et al.

Genome Res. 1999 9: 1-4
Access the most recent version at doi:10.1101/gr.9.1.1



Assessing the Quality of the DNA Sequence from The Human Genome Project

Adam Felsenfeld,[1] Jane Peterson, Jeffery Schloss, and Mark Guyer

National Human Genome Research Institute (NHGRI), National Institutes of Health, Bethesda, Maryland 20892-6050 USA

It is sometimes hard to remember that the first DNA sequence of the entire genome of a free-living organism, *Hemophilus influenzae*, was reported <4 years ago (Fleischmann et al. 1995). Since then, the genomes of >17 other prokaryotes (http://linkage.rockefeller.edu/wli/seq/), a unicellular eukaryote, *Saccharomyces cerevisiae* (Nature 1996), and a multicellular organism, *Caenorhabditis elegans* (The C. elegans Sequencing Consortium 1998), have been completely sequenced. Progress toward determination of the human DNA sequence has also become more rapid; at the time of this writing, the public databases contain 227.2 Mb of nonredundant, finished sequence available in contigs of >30 kb (and another 152.7 Mb of unfinished sequence) (http://www.ncbi.nlm.nih.gov/genome/seq/weekly_report.html). In comparison, there was 84.4 Mb of finished data (http://www.ebi.ac.uk/~sterk/genome-MOT/) in February 1998. It is increasingly likely that the human sequence will be complete by 2003, and a working draft will be in hand even sooner (Collins et al. 1998; Venter et al. 1998). One consequence of our increased sequencing capacity is that within the next couple of years, we expect the rate of deposition of sequence data to increase from the current ~3 Mb per week, to an average of well over 10 Mb per week worldwide. Very few scientific fields can measure progress as easily as can be done for large-scale genomic sequencing, quantifiable as it is into base pairs per unit time.

However, mere numbers can be deceptive—the essential ''production'' nature of large-scale genomic sequencing leaves it susceptible to errors in ways other scientific endeavors are not. Because of the rapid accumulation of human genomic sequence data, there is little opportunity for, or even possibility of, direct peer review of data prior to publication. The major venue for primary publication of genomic data is not the peer-reviewed literature at all, but public databases. This is appropriate: Current peer-reviewed biological journals could not handle this much primary data, nor would they want to, nor would the community be likely to entrust this resource only to the printed medium. But more critically, the community has made the important decision that these data must be accessible very rapidly. For publicly funded laboratories throughout the world, genome sequence data are supposed to be released into a public database within 24 hr of being generated (Collins et al. 1998), a standard that is, as far as we are aware, unmatched by any other scientific discipline. This rapid release is in many ways at odds with what is normally understood to be peer review. Finally, the bulk of the work will probably not be directly replicated, especially for the human sequence and that of other large genomes. There is little doubt, however, that the data will be heavily relied on. For all of these reasons, it is important that the Human Genome Project (HGP) devise a way of measuring and reporting the quality of sequence data deposited in the public databases.

Not All Base Pairs Are Created Equal

Discussions about data quality properly focus on what is desirable for scientific purposes. For some purposes, such as identifying coding regions, reconstruction experiments suggest that low coverage (and therefore less accurate and less expensive) data would be sufficient (Bouck et al. 1998). For other uses, low quality data would be nearly useless. For example, the frequency of common single nucleotide polymorphisms in human DNA is ~1 in 1000; sequence that is less accurate than this will be of little help in identifying such polymorphisms, which are increasingly important for mapping and identifying genes contributing to complex disease and other phenotypes. Similarly, successful design of PCR primers for use in the amplification of genomic DNA requires high sequence quality. Studies employing sequence alignments, as well as those modeling protein structures based on sequence, benefit from accurate sequence. The degree of continuity, and the fidelity of the determined sequence assemblies to the target genome, also affect what the data can be used for. For example, the identification of distant regulatory regions or the ability to ask questions about genome evolution or correlates of disease (rearrangements, etc.) benefit from accurate determination of sequence over a relatively long range. Finally, there are likely to be features of the genome and uses that we do not currently appreciate to which a complete, accurate, and faithful sequence can be put.

These considerations suggest three convenient parameters that can be used to describe the quality of finished genomic sequence, which we will describe here as accuracy, contiguity, and fidelity. The sequencing community has discussed standards for each (see Box 1). Accuracy refers to the reliability of any single base pair. The current standard for

[1]Corresponding author.
E-MAIL adam_felsenfeld@nih.gov; FAX (301) 480-2770.

9:1–4 ©1999 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/99 $5.00; www.genome.org

Box 1. Origins of the QA Exercise

A workshop on DNA sequence validation (http://www.nhgri.nih.gov/HGP/Reports/dna_sequence_workshop.html.), held in the Spring of 1996, offered several recommendations regarding sequence quality, including

- Genomic DNA sequencing projects should attempt to reach an error rate of no more than 1:10,000 bases.
- The quality of individual base calls and of the sequence assembly should be assessed in a validation study and, to the extent possible, methods for validation should be independent of those used in the initial sequence determination.
- To demonstrate the fidelity of the clones sequenced to the native genomic structure, clear evidence should be presented that the genomic region is represented by the same restriction digestion pattern in at least two clones derived from independent transformation events. Greater depth of coverage is highly desirable and is expected in most cases.
- The public databases should store quality measures on each base pair.

These conclusions were confirmed at the Second International Strategy Meeting on Large-scale Sequencing in February 1997. At that time, it was also proposed that the HGP adopt a goal of producing human DNA sequence with no gaps (though it should be recognized that the ability to do this over truly large distances is unproven).

accuracy is that finished genomic sequence should have no more than one error per 10,000 bases. Contiguity refers to the length of an assembled sequence contig, and whether there are gaps or unresolved ambiguities. At present, the NHGRI operational standards for contiguity are that contigs should have no gaps or ambiguities and must be at least 30 kb to ''count'' as finished sequence. The standard for the minimum size of a ''finished'' contig is expected to increase as the HGP progresses. Fidelity refers to the precision of the assembly within a contig, relative to the genome from which the sequenced clones were derived. Fidelity errors can arise by misassembly of shotgun data, or rearrangement of the insert of the clone that was sequenced. Errors in assembling large, multiclone contigs can also be thought of as fidelity errors; ideally such long-range assemblies should be faithful to the target genome. However, there are at present relatively few large contigs (>1 Mb), and a quantitative description of the standard for fidelity has not been articulated.

Errors can be expected to occur with some frequency in even the most ideal circumstances. To add a complication, the human genomic sequence is being determined by several laboratories, each with its own approach. Although some internal quality standards and measures for assuring adherence to them have been adopted and are extremely useful [e.g., quality score cutoffs for data as determined by base-calling and assembly programs such as Phred and Phrap (Ewing and Green 1998; Ewing et al. 1998; http://bozeman.genome.washington.edu/phrap.docs/phrap.html)], these may not be able to overcome variability between centers because of differences in specific strategy (degree of shotgun coverage, finishing, and even mapping strategy, large-insert clone library used, etc.). Given all of these factors, it is very important to have an independent way of measuring and ensuring quality of the overall sequence product of the HGP.

Quality Assessment

In the Spring of 1997, a group of nine NHGRI grantees participated in an experiment designed to test whether the quality standards of 10,000 bp accuracy and zero gaps could be measured practically. The design of the experiment, based on discussions at the Second International Strategy Meeting on Large-scale Sequencing and further discussion with NHGRI sequencing centers (see Box 1), involved the electronic exchange of sequence data files between sequencing laboratories. Two completed large-insert clones from each of the NHGRI-funded large-scale human sequencing pilot projects were identified at random. For each, the data files generated by the automated sequencing instruments were sent, along with other pertinent information, to each of two other laboratories (the ''checkers''). Using that electronic information, the checkers attempted to determine how reproducibly the sequence could be reassembled. The number of discrepancies between the reassembled sequence and the sequence that had been submitted to the public databases by the producer laboratories was taken to indicate the level of quality of the submitted sequence. In summary, although the results from the two checkers were not identical (either to each other or to the data as represented in the database), in most cases, it was clear that poor data (many discrepancies) could be distinguished from high quality data (few discrepancies). In those cases in which the data were found to be of high quality, the rate of discrepancy was in the range of 1 in 10,000 bases. The participants agreed that this initial experiment suggested that an effective approach to evaluating the quality of DNA sequence data could be developed.

On the basis of this experience, a second quality assessment (QA) exercise was begun in September 1997. In this exercise, the electronic analysis was repeated and augmented with actual resequencing to resolve the discrepancies between the sequence submitted by the producing laboratory and the analyses of the checkers. The participants included all of the NHGRI-funded human sequencing centers, plus the *Drosophila* Genome Sequencing Center. At least four different clones that had been submitted to the public databases as finished, comprising a total of at least 200 kb, were checked from each laboratory. If the electronic reassembly of the trace data suggested an error rate of >1 in 2000,

this was documented and the checker was not obligated to analyze such poor sequence any further. If the error rate was better, the checker proceeded to actually resolve all of the discrepancies. The checkers were also instructed to attempt to resolve any ambiguities (Ns) and to close any gaps left by the data producer; this was done to determine whether the reported gaps or ambiguities actually resulted from truly difficult-to-sequence regions. Each data producer was given a chance to respond to the checkers' reports. The results of this exercise are summarized in Box 2.

The results of this second QA exercise, using data finished prior to September 1997, show that most of the sequencing centers were attaining the standard for single-base discrepancy rates. Excluding the three clones with error rates >1 in 2000 (which contained almost half the total number of single-base errors), the total number of single-base discrepancies was 120 in 1.59 Mb (note that this number uses the worse of two checkers' reports in those cases in which the checkers' reports differed and does not include consideration of the producer's response). The results also show what kind of single-base errors occur. A majority of the errors were single-base substitutions, which some consider to be less serious than insertions or deletions. The results of the exercise also tentatively suggest that many of the single-base errors occurred in regions where the quality of the original sequence trace data is high—more than half of the single-base errors appeared to be manual editing errors and were unambiguously resolved by re-editing the original data without the need for resequencing.

Besides single-base errors, several other kinds of error were revealed. Four of the inserts were misassembled. Subsequent analysis suggested that some of these errors were due to small deletions that occurred during clone growth, rather than to computational misassemblies. Three of the four were in cosmid vectors. Since the time of this exercise, essentially all of the public large-scale sequencing efforts have switched to the use of BACs or PACs as the cloning vector of choice because these vectors are thought to be less susceptible to rearrangement. In any case, the exercise points out that small deletions or other small rearrangements may be difficult to detect routinely (e.g., by single-enzyme fingerprinting of clones).

There are a number of caveats to the interpretation of these results. Although the concordance between the checkers' reports was good with regard to the general quality of each project (agreeing 28 of 37 times, according to the bins in the table in Box 2), there were often differences in the individual errors identified in a clone, especially when the original trace data supplied by the producers were poor. There were also differences in the methods used by each checker and the depth of the check. In addition to the variability introduced by having so many different checkers, there is a degree of variability that could be introduced by sampling. For example, at the time of this exercise, one producer had only two clones listed as finished in the public database while another had >100. Also, all clones, no matter how long since they had been completed, were eligible for testing. Considering the rapid progress in the field, such ''old projects'' probably did not even reflect the performance of a producer at the time of the exercise.

A third QA exercise was begun in November 1998; it is scheduled for completion in February 1999. The methodology will be essentially the same as for the second exercise but will focus on data produced within the previous 6 months. This time, 17 producers are participating, including the Sanger Centre and the Department of Energy (DOE)-sponsored Joint Genome Institute.

Although they have been very instructive, the quality assessment exercises undertaken so far have not adequately addressed some important issues of sequence quality. For example, they have not allowed determination of the existence of gaps between sequenced clones. Detecting such gaps will require a different approach, for example, restriction map comparison between large clone contigs and the corresponding genomic DNA. Related quality issues, such as long-range fidelity of largeinsert clones to the genome, will also have to be

Box 2. The Second NHGRI QA Exercise

Data eligible for being checked were selected from that deposited as finished as of September 1997.

Total number of clones available for checking as of September 1997: 420.
Total number of clones selected for the exercise: 37 (a total of 1.7 Mb tested).

Table 1. Single-Base Discrepancies—Number of Clones at Error Rate

<1/10000	1/10000–1/5000	1/5000–1/2000	>1/2000	Total
22	10[a]	1	3	36[b]

These numbers are based on results that indicated the higher error rate among the two checkers' reports, for each individual clone; the numbers do not take into account the producer's responses.
[a]For 7 of the 10 clones in this category, one of the two checkers actually evaluated those clones as having <1 in 10000 errors.
[b]One of the clones sent did not correspond to the clone originally sequenced; this was a clone-tracking error.

About two-thirds (133) of the single-base discrepancies were substitutions; one-third (73) were insertions or deletions, based on 206 cases of single-base errors where precise information was provided.

Other errors (not exclusive of single-base errors)
4 misassemblies, some likely to be due to small deletions during regrowth (~250–1900 bp) in the large-insert clone; 3 of these were in cosmids.
1 annotated gap closed (75 bp)
1 wrong clone sent (clone tracking error

addressed and may be incorporated into future QA exercises.

A single, specialized QA center would offer a number of advantages over the approach of round-robin exercises. First, a single QA center would ameliorate some of the uncertainties associated with variability due to the assessment being carried out in multiple locations. Second, the round-robin exercises have been an enormous amount of work for the centers, which are already operating at peak capacity to meet production goals. A dedicated QA center would also be a logical site to pursue research into improved ways to measure sequence quality. (Of course, a dedicated QA center in no way eliminates the need for each production group to have internal quality control measures in place.) It should be emphasized that an absolute prerequisite to having an independent measure of sequence quality is the availability of the raw data files underlying the consensus sequence in the public database.

In summary, the HGP has begun to develop effective means of assessing the quality of the genomic sequence being produced to determine the actual quality of the product, to inform review decisions, and to learn whether current technology is adequate for meeting sequence quality standards. This information will also be important to understand the complicated relationship between quality and cost, which must be taken into account in adoption of new quality standards. Finally, we believe that participation in the past QA exercises has encouraged individual centers to improve their processes, and more indirectly, we believe that it has fostered useful communication among the sequencing centers.

ACKNOWLEDGMENTS

The QA exercise would not have been possible without the initiative and enthusiastic participation of the large-scale sequencing centers in the design of the exercise or the time-consuming effort of their staff in doing the assessments. The individuals are, unfortunately, too numerous to thank individually here. The participating sequencing centers are the Washington University Genome Sequencing Center; The Whitehead Institute/MIT Genome Sequencing Project, The Genome Science and Technology Center at University of Texas Southwestern Medical Center, The Advanced Center for Genome Technology at the University of Oklahoma, The University of Washington Genome Center, The Stanford Genome Sequencing Center, The Advanced Center for Genetic Technology, Applied Biosystems Division, Perkin-Elmer, the DOE Joint Genome Institute, The Berkeley *Drosophila* Genome Project, The Baylor College of Medicine Human Genome Sequencing Center, and The Institute for Genomic Research. We thank Dr. Francis Ouellette and the National Center for Biotechnology Information for help in selecting the clones to be tested. We thank Drs. Susan Celniker, Lee Rowen, Bruce Birren, and David Bentley for useful conversations about sequence quality assessment. We thank Francis Collins, Elke Jordan, and Bettie Graham for their indispensable comments on this manuscript.

REFERENCES

Bouck, J., W. Miller, J.H. Gorrell, D. Muzny, and R.A. Gibbs. 1998. *Genome Res.* **8:** 1074–1084.

Collins, F.S., A. Patrinos, E. Jordan, A. Chakravarti, R. Gesteland, L. Walters, and members of the DOE and NIH planning groups. 1998. *Science* **282:** 682–689.

Ewing, B. and P. Green. 1998. *Genome Res.* **8:** 186–194.

Ewing, B., L. Hillier, M.C. Wendl, and P. Green. 1998. *Genome Res.* **8:** 175–185.

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Butt, J.-F. Tomb, B.A. Dougherty, J.M. Merrick et al. 1995. *Science* **269:** 496–512.

Guyer, M. 1998. *Genome Res.* **8:** 413.

Nature (Suppl.) 1996. **387:** 5–105.

The C. elegans Sequencing Consortium. 1998. *Science* **282:** 2012–2018.

Venter, J.C., M.D. Adams, G.G. Sutton, A.R. Kerlavage, H.O. Smith, and M. Hunkapiller. 1998. *Science* **280:** 1540–1542.