Quality assessment of the human genome sequence

Jeremy Schmutz, Jeremy Wheeler, Jane Grimwood, Mark Dickson, Joan Yang, Chenier Caoile, Eva Bajorek, Stacey Black, Yee Man Chan, Mirian Denys, Julio Escobar, Dave Flowers, Dea Fotopulos, Carmen Garcia, Maria Gomez, Eidelyn Gonzales, Lauren Haydu, Frederick Lopez, Lucia Ramirez, James Retterer, Alex Rodriguez, Stephanie Rogers, Angelica Salazar, Ming Tsai & Richard M. Myers

Stanford Human Genome Center, Department of Genetics, Stanford University School of Medicine, 975 California Avenue, Palo Alto, California 94304, USA

As the final sequencing of the human genome has now been completed, we present the results of the largest examination of the quality of the finished DNA sequence. The completed study covers the major contributing sequencing centres and is based on a rigorous combination of laboratory experiments and computational analysis.

From the beginning, a primary objective of the Human Genome Project (HGP) was to generate a highly accurate reference sequence for the human genome. This sequence is now essentially complete and is available in its entirety as a reference for biomedical researchers. High-throughput genome sequencing has created a fundamental shift in the paradigm for biological research. Whereas gene discovery once drove DNA sequencing, now the sequencing of entire genomes drives gene discovery. As such, it is essential that the scientific community be informed about the accuracy of this reference sequence and of its fidelity to the biological templates from which it was derived.

Box 1

Large-scale sequencing terms for this study

Accuracy The measure of how likely the base pairs in a consensus are to be the correct base call. For a 99.99% accurate DNA sequence, it must contain only one incorrect base per 10,000 bp. Accuracy is also sometimes referred to as the 'quality' of a base pair, because estimated base-pair qualities are assigned by the assembly software when it creates the consensus.

Consensus The final reconstructed DNA sequence built by assembling the sequence reads and generating a consensus base call for each position in the assembly. In the case of a finished clone, there is only one consensus.

Contiguity The measure of how many pieces are contained within the assembly. A contiguous assembly would typically have multiple overlapping sequence reads the entire length of the consensus. The finishing rules allow a consensus in more than one piece to be called contiguous (no gaps) in certain difficult situations if the break point is annotated in the database entry.

Fidelity The fidelity of a consensus is how similar the consensus is to the underlying biological template from which the sequence reads were derived. Fidelity for a genome at the single base-pair level is difficult to measure without identifying and sequencing a different clone from the same position on the same chromosome and examining the difference between the sequences. In this study we evaluated the fidelity of a sequence in reference to the large-insert clone from which the sequence was derived, not the genomic template.

Finishing The process of collecting data, performing computational manipulation to a data set to convert a shotgun assembly into a single high-quality contiguous DNA sequence, and verifying the fidelity of the consensus.

World standards for sequence fidelity (known as the Bermuda Standards) were established at the meeting of HGP principal investigators in 1997 (http://www.gene.ucl.ac.uk/hugo/bermuda2.htm). These standards stated that finished sequence should contain less than one error per 10,000 DNA bases (99.99% accuracy), and that the sequence should be contiguous (without gaps). Compliance with the base-pair (bp) accuracy standard was measured by error probability assessments generated by DNA base-calling software[1–3] and by examining discrepancies between overlapping clone sequences. Compliance with the contiguity standard was an internal measurement based on each centre's complex sequence-finishing methodology. Over the course of the project, additional standards were created to ensure sequence fidelity (http://www.genome.wustl.edu/Overview/g16stand.php).

Although more than 2.8 billion base pairs of unique finished sequence has been generated by the sequencing centres comprising the International Human Genome Sequencing Consortium (IHGSC), until the present study was performed fewer than 5,000,000 bp of this sequence has been verified independently for compliance with the finishing standard[4]. Finished chromosome sequence papers have now been published for 9 of the 24 human chromosomes[5–13], with most of these papers estimating that the chromosomal sequence exceeds the 99.99% accuracy measure. To provide a more uniform picture of the finished sequence quality of the human genome, the National Human Genome Research Institute (NHGRI) solicited us to perform a detailed evaluation of the DNA sequence data that was generated for the HGP by seven of the IHGSC centres. We examined more than 34 megabases (Mb) of sequence data for accuracy, contiguity and fidelity (see Box 1), and participated in a computational data exchange with the Wellcome Trust Sanger Institute. This paper contains the results of our analysis of the quality of finished sequence data deposited by these centres in the public human genome databases from February 2001 through to July 2002.

Overview and procedure

Our quality assessment of finished human bacterial artificial chromosome (BAC) sequences was conducted in two rounds. For the first round of analysis, we evaluated the finished sequence produced by the three largest NHGRI-funded sequencing centres: the Baylor Human Genome Sequencing Center, the Washington University Genome Sequencing Center and the Whitehead Institute Center for Genome Research. We selected 120 BAC clones (about 6.7 Mb from each centre) from sequence submissions spanning the six-month period from 15 February 2001 through to 15 August 2001. The second round of analysis evaluated the sequence produced by the four smaller sequencing centres that individually

contributed more than 30 Mb to the human genome: the Genome Therapeutics Corporation, the French National Sequencing Center Genoscope, the University of Washington Genome Center and the RIKEN Genomic Sciences Center. We selected 80 BAC clones (about 3.4 Mb from each centre) from these sequencing centres, spanning the 17-month period from 15 February 2001 through to 30 June 2002 (Supplementary Fig. S1 and Table S1).

We sampled clones throughout the two time periods, and adjusted the number of clones that we selected to be a percentage of clones finished by each centre in each month. The sequencing centres provided us with sequencing read data and glycerol stocks of the large-insert clone. In contrast to previous quality assessments[4], we created a new subclone library for each clone and sequenced this library to 3–4 times coverage in high-quality base pairs. We generated these reads from both ends of sized plasmid subclones, which gave us the ability to evaluate independently the centre's submission regardless of how the original data were generated. We combined our new reads with the original data and then finished the resulting assembly to a high degree of accuracy, performing directed sequencing reactions from the large-insert clone when necessary. These directed reads included reactions performed with alternative chemistries such as dGTP and Invitrogen sequencing enhancers. All finishing quality analysis was performed using the Phred/Phrap/Consed[3] pipeline. We then compared our 'gold standard' consensus to the original submitted consensus and then verified and classified any discrepancies. For each discrepancy, we counted the number of error events and base-pair errors and as necessary classified the error as a significant error or misassembly (see Box 2). We counted an error only if original data generated by the submitting centre supported the correct consensus; in this way, we avoided classifying any large-insert clone growth variations as sequencing errors.

Accuracy results and base-pair errors

Our analysis indicates that all of the sequencing centres surveyed met the standards for 99.99% accuracy over the time period studied. Figure 1 shows the plot of the error events and the base-pair errors for each clone that we assessed. These are plotted as rates, normalized per 10 kilobases (kb) over the length of each clone, and include all incorrect base pairs. Most (184 out of 197) of the clones have less than 1 bp error per 10 kb, with 59 of the clones having no identified errors. Twelve of the thirteen remaining clones exceed the 1 bp per 10 kb standard owing to significant errors. Disregarding the significant errors, only 1 of the 197 clones exceeded the target error rate because of base-pair errors alone. Cumulative error results for each of the rounds are shown in Table 1. The individual centre base-pair error rates ranged from 1 in 25,420 bp to 1 in 154,479 bp, and significant error rates ranged from none found to 1 in 1.2 Mb (Supplementary Table S2).

The vast majority of error events found in the finished human BAC sequences affected a single base pair in the consensus sequence (411 out of 466, 88.2%). Roughly half (48%) of these errors were single base-pair substitutions, with the remainder (52%) being single base-pair insertions or deletions. The substitution errors were primarily miscalled bases in regions of low quality. However, there are many positions where a miscalled base was incorporated into the consensus sequence despite the presence of multiple high-quality reads with the proper base calls; these are obvious finishing errors. Most of these errors occurred where a single discrepant subclone at that position was given a high-quality score by the base-calling algorithm and the miscalled base was incorporated into the consensus sequence. Additionally, we identified 42 (9% of error events) multiple base-pair insertions, deletions and substitutions of less than 20 bp, most of which were clone mutations in a single subclone that were erroneously included in the consensus.

Box 2

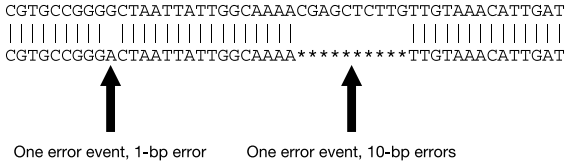
Analysis terms for this study

Base-pair errors The number of base-pair changes between our 'gold standard' consensus and the original submitted sequence.

Error events A count of the number of positions of change in the consensus discovered in the quality assessment process; a contiguous insertion, deletion or erroneous run of multiple base pairs is counted as a single error event because the multiple base-pair errors probably arose from a single process error.

Misassembly A rearrangement or deletion of the consensus caused by the incorrect joining of two similar pieces of sequence that are geographically separated in the true consensus.

Significant error A single error that causes at least 50 contiguous base pairs to be incorrect in the submitted consensus versus our gold standard consensus.



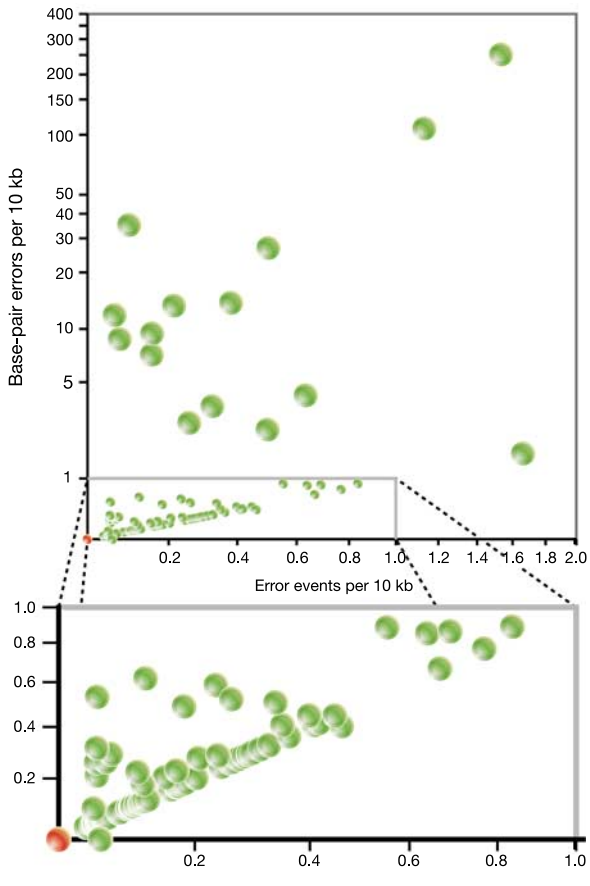


Figure 1 A plot of the error events per 10 kb versus the base-pair errors per 10 kb for the clones surveyed. Each green circle represents a different surveyed clone. A detailed view of the boxed area (less than one error event per 10 kb and less than 1-bp error per 10 kb) shows the diagonal distribution of all of the clones containing only single base-pair errors. The red circle indicates 59 clones with no errors.

Table 1 **Cumulative results of each of the quality assessment rounds**

Analysis results	Round 1	Round 2	Total
Sequence analysed (kb)	20,303	13,887	34,190
Clones analysed	117	80	197
Error events	183	283	466
Substitution events	73	135	208
Insertion/deletion events	110	148	258
Error event rate (bp)	1/110,948	1/49,069	1/73,369
Base-pair errors*	255	381	636
Base-pair error rate	1/79,621	1/36,448	1/53,758
Significant errors†	5	8	13
Significant error rate (bp)	1/4,060,688	1/1,735,828	1/2,630,005

*Does not include significant insertions, deletions or rearrangements of sequence.
†Insertions or deletions of greater than 50 bp or significant rearrangements of sequence.

Significant errors

We found a significant error in 12 out of the 197 (6.1%) BAC clones that we analysed. There were 13 total significant errors in these clones (2.8% of the total error events). Most of these were identifiable as potential problems from the initial assembly of only the contributing centre's data set. We found large consensus deletions that were derived from deleted subclone templates or polymerase chain reaction (PCR) amplified products. Long stretches of sequence were also deleted as a result of incorrect joins made in repetitive regions, through which sequencing was difficult, and joins were based on minimal sequence overlap. The distribution of these sequence areas that were more difficult to sequence varies across the human genome[12], and consequently, we did not survey difficult clones from every centre.

Potential error-prone finishing techniques

In the course of this quality assessment we identified finishing techniques that in some cases directly contributed to consensus errors that were not corrected before submission by the centre. A large number of the single base-pair-deletion errors were the result of G + C compressions from dye-primer chemistry (now phased out of use in most centres) or dGTP chemistry (a chemistry for difficult-to-sequence regions), or from A or T base drop-out errors on the Megabace platform. Some of the larger deletions in simple sequence regions were from PCR-generated templates or from single subclones that had deleted a portion of the repeat copies. Clones consisting of mostly single-direction M13 reads had more serious assembly issues in repetitive areas. Higher assembly stringencies would have reduced greatly the number of incorrect joins and improved the overall accuracy for the identified misassembled repeat structures.

Computational quality assessment of two contributors

In addition to the quality assessments detailed in this paper, the Wellcome Trust Sanger Institute and the Joint Genome Institute/Stanford Human Genome Center exchanged 38 finished clones over the same time period as round one in this study. These two centres examined only trace data and built new assemblies to compare to the submitted assemblies, and they did not add additional sequencing data. Suspected errors were verified by the original submitting centre. This study found that, for these two centres, there was on average 1 bp error per 651,000 bp and one potential significant error in 11.1 Mb. Together these centres contributed about 39% of the human genome sequence. Although this analysis is not directly comparable with our more detailed study—because computational analysis alone is unable to detect all of the errors found with additional sequencing (see Supplementary text)—this provides a reviewed estimate of error rates for these two centres.

Quality of the finished human genome

We believe that the quality evaluation methodology outlined in this paper provides a uniform framework to evaluate sequence produced by the disparate finishing systems used by the IHGSC sequencing centres in relation to the standards for finished sequence quality. Of the 197 clones analysed, we found that 182 (92.4%) significantly exceed the 99.99% accuracy standard, on the basis of a calculation of base-pair errors per 10 kb (Fig. 1). If the sampled data set is applicable to the entire genome, we can conclude that the base-pair accuracy standards have been exceeded tenfold, as there is less than 1 bp error per 100,000 bp of finished sequence. If we normalize for the relative amounts of sequence contributed by each centre, we should expect to find on average seven error events with nine incorrect bases per 1 Mb and one significant error per 6 Mb.

We believe that caution should be exercised in extrapolating our data beyond the specific regions of the genome that were surveyed in our study. This quality assessment is an evaluation of process, as it was based on a methodology of sampling sequence production over time, not sampling uniformly from the finished product. As such, our results are a reflection of the finishing methodologies used by the centres for the time period evaluated in our study, and these methodologies were subject to continuous improvement. For the centres investigated in round one, we sampled from a single production period, whereas clones submitted early and late in the HGP were not sampled; these clones are more likely to contain a higher error count. Along with improvements to knowledge and technology, the goals of the overall HGP changed over the course of the project, and the quality threshold used by the sequencing centres fluctuated in response to these production goals. In addition, our thorough quality evaluation methodology (which included additional shotgun sequencing) was applied only to sequencing centres contributing 55% of the total human sequence, with an additional 39% assessed by computational evaluation alone. No sequence was surveyed from the 6% of the genome finished by many smaller contributors.

As a result of differences in the application of finishing methodologies, the most significant factor correlated to sequence quality is centre-to-centre variation (Supplementary Table S2). The sequencing centres had different thresholds at which they determined a given region to be 'finished', and some centres intentionally exceeded the required quality levels (Table S2). The stringency of the contiguity threshold (the cause of most of the significant errors identified) applied by each group was the result of an admixture of production pressures, 'regional' complexity of their genomic territory, personnel experience in addressing the variations in finishing difficulty, and degree of communication and standardization of the group with the larger HGP community. The nature of the HGP as a pilot project for large-scale genomic sequencing makes it difficult to describe the quality of the human genome sequence as a singular entity, although this evaluation has provided valuable insights into the process of producing a complete, complex, finished genome sequence.

Applications to future projects

Well-defined finishing standards specifying targets for accuracy, singular contiguity and fidelity—coupled with descriptions of processes that enable greater compliance with these standards—will enable future genome sequencing projects to generate a more uniform quality product. Continuous sampling of finished sequence for quality evaluation throughout the production process of future genome sequencing projects would better enable global quality statements to be made, and subsequent incorporation of quality assessment feedback into the process could further enhance the quality of the product. Standardizing or minimizing a number of variables—genome source, cloning and library construction platforms, hierarchical sequencing strategies, definitions of finished product—will help further. Such procedures will enhance not only verification ability on a clone or regional basis, but will also be tremendously helpful in solving recalcitrant problems such as the resolution of large duplicated genomic structures. As new genome-sequencing techniques emerge over the course of a sequencing project (for example, cloning vectors, sequence chemistries, detection platforms, finishing techniques), a centralized quality-control centre could serve as a resource for evaluating the technique's relative ability to ensure fidelity with the genomic sequence, rather than each centre independently examining and evaluating all new technologies. In this capacity the quality-control centre would serve as a distributor of reviews and test performance reports for technological developments, which would allow all sequencing centres equal access to information about these techniques. A central trace data repository, such as the NCBI trace archive, is a positive step towards making all raw sequencing trace data available, but also storing the final assemblies would enable central coordination of gap-closing efforts and allow centres to concentrate on the finishing problems that they have developed pipelines to address, instead of expecting each centre to apply these complicated techniques to an equal standard. □

Received 24 October 2003; accepted 26 January 2004; doi:10.1038/nature02390.

1. Ewing, B. & Green, P. Base-calling of automated sequencing traces using *Phred*. II. Error probabilities. *Genome Res.* **8,** 186–194 (1998).
2. Ewing, B., Hiller, L., Wendl, M. & Green, P. Base-calling of automated sequence traces using *Phred*. I. Accuracy assessment. *Genome Res.* **8,** 175–185 (1998).
3. Gordon, D., Abajian, C. & Green, P. Consed: A graphical tool for sequence finishing. *Genome Res.* **8,** 195–202 (1998).
4. Felsenfeld, A., Peterson, J., Schloss, J. & Guyer, M. Assessing the quality of the DNA sequence from the Human Genome Project. *Genome Res.* **9,** 1–4 (1999).
5. Deloukas, P. *et al.* The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414,** 865–871 (2001).
6. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402,** 489–495 (1999).
7. Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405,** 311–319 (2000).
8. Helig, R. *et al.* The DNA sequence and analysis of human chromosome 14. *Nature* **421,** 601–607 (2003).
9. Hillier, L. *et al.* The DNA sequence of human chromosome 7. *Nature* **424,** 157–164 (2003).
10. Mungall, A. J. *et al.* The DNA sequence and analysis of human chromosome 6. *Nature* **425,** 805–811 (2003).
11. Skaletsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423,** 825–837 (2003).
12. International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
13. Dunham, A. *et al.* The DNA sequence and analysis of human chromosome 13. *Nature* **428,** 493–521 (2004).
14. Grimwood, J. *et al.* The DNA sequence and biology of human chromosome 19. *Nature* **428,** 529–535 (2004).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements We thank the participating centres for providing clone stocks and sequence data sets, and for their feedback about our quality assessment process. We also thank C. Lloyd and C. Bagguley for their detailed computational assessment of our finished sequence.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.S. (jeremy@shgc.stanford.edu) or R.M.M. (myers@shgc.stanford.edu).