The independence of our genome assemblies

Mark D. Adams*†, Granger G. Sutton*, Hamilton O. Smith‡, Eugene W. Myers§, and J. Craig Venter¶

**Celera Genomics, 45 West Gude Drive, Rockville, MD 20850; ‡Institute for Biological Energy Alternatives, 1901 Research Boulevard, Suite 600, Rockville, MD 20850; §Department of Electrical Engineering and Computer Sciences, 231 Cory Hall, University of California, Berkeley, CA 94720; and ¶Center for the Advancement of Genomics, 1901 Research Boulevard, Suite 600, Rockville, MD 20850*

The characterization, analysis, and conclusions of Waterston *et al.* (1) with regard to our published work (2, 3) are incorrect. Celera was founded with the goal of applying the whole genome shotgun (WGS) strategy to assemble the sequence of the human genome as rapidly as possible to advance the field of genomics (4). Despite our previous arguments to the contrary (3), Waterston *et al.* (1) persist in their claims that Celera's assembly of the human genome, reported in 2001 (2), was simply a "refinement built upon the [Human Genome Sequencing Consortium (HGSC)] assemblies." In fact, the Celera assemblies were constructed on the basis of mate-pair information from Celera sequence data, and the HGSC contribution to the structure and content was minimal.

Waterston *et al.* (1) use only one characteristic of genome assemblies as the basis for their arguments. The "contig N50 length" refers to the short-range character of an assembly, or the assembly quality in segments that are ≈0.0003% of the size of the human genome.‖ There are two other critically important parameters by which an assembly must be judged: scaffold N50 length (which measures how well contigs are put together in linear sets) and the correctness of the order of the base pairs in the assembled sequence of each chromosome. The HGSC assemblies are considerably worse by each of these measures than the Celera assemblies. The independence of the Celera assemblies is illustrated in these differences. The scaffold N50 length for the compartmentalized shotgun assembly was 2.96 Mbp (2), compared with 0.27 Mbp for the HGSC assembly (table 7 in ref. 5). Celera's assembly had ≈35,000 fewer ordering errors (segments of the genome either misplaced or rearranged) than the HGSC (figures 6 and 7 in ref. 2). Both the scaffold N50 length and the dramatic difference in assembly order are a direct result of the high density of mate-pair coverage (38-fold) present in Celera's whole-genome shotgun data set that provides the power to build accurate assemblies over long genomic distances.

The primary input to Celera's assembly was ≈5-fold sequence coverage of the human genome in sequences derived from both ends of randomly cloned shotgun fragments of the genome (mate pairs). This whole-genome component was augmented by "faux reads" produced by shredding the sequence of partially assembled contigs from bacterial artificial chromosome (BAC) clones sequenced by the HGSC (5). In the fall of 2000, two-thirds of the BAC clones sequenced by the HGSC were present in GenBank at Phase 1 coverage or less (meaning 3–5× coverage and partially assembled, with generally 10–50 individual contigs in random order). Far from "relying" on the "HGSC assemblies," Celera used a collection of >677,000 (table 2 in ref. 2) unlinked, unordered, and in many cases erroneously assembled contigs; these were shredded before input into assembly. The Celera assembler was designed to rely on Celera's mate-paired data as the primary determinant of assembly over the BAC data of the HGSC. Our reliance on mate pairs as the overriding determinant of assembly structure was based on three factors. First, our experience with several test assemblies of the *Drosophila* genome that demonstrated that excellent long-range assembly (multimegabase-sized scaffolds) could be obtained with 5× sequence coverage in mate pairs from a combination of large and small insert libraries. Second, the 5× whole-genome shotgun sequence from Celera was expected to contain 97% coverage of the genome; by shredding contigs from BAC clones, we expected to fill small gaps that were present in the shotgun coverage, essentially the remaining up to 3% of the genome not covered by Celera data. Third, the inconsistent quality of contigs from BAC assemblies had a strong negative effect on assembly that we wanted to minimize. Contigs from BACs were often of poor quality at their edges, and were frequently misassembled (see ref. 2, notes 40, 41, and 47), both factors that result in breaks or inconsistencies when merged with the whole-genome shotgun data from Celera.

Waterston *et al.* (1) raised three technical points: (*i*) perfect tiling, (*ii*) gap filling, and (*iii*) N50 length. The latter two are intimately related and will be treated together. In an attempt to model the reassembly of shredded contigs ("perfect tilings"), Waterston *et al.* have pursued a simulation of genome assembly (figure 1 in ref. 1) that does not predict the behavior of our assembly algorithms. Most of their discussion focuses on the performance of the simulation using reads with quality values. This is irrelevant to a discussion of our method, because the Celera assembler did not use quality values to evaluate overlaps. By using the same software that we used for genome assembly, we showed that the ability to reassemble shredded reads is dramatically affected by the total amount of whole-genome sequence present in the assembly (table 1 in ref. 3). The inability of Waterston *et al.* to replicate this finding means that there are fundamental differences between their model and the true way in which shredded reads contributed to the Celera assembly.

Celera's WGS assembly algorithms depend on three things to be successful: construction of unitigs (contigs that assemble uniquely, with no conflicting information), identification of unique unitigs (those that are single copy in the genome), and sufficient mate pairs connecting the unique unitigs. The shredded reads from the HGSC data were treated as individual reads, and thus add no mate pair information. In practice the shredded reads also did not increase the number of base pairs in unique unitigs versus using only Celera fragments (1.67 Gbp versus 1.66 Gbp), and thus contributed only a minimal amount unique genome sequence that was not also represented in the Celera data set. Although the number of base pairs in unique unitigs was essentially the same with or without the shredded reads, the N50 size was slightly larger with the shredded reads than without (3 kbp versus 2.5 kbp). The size range of the unitigs is thus dominated by the presence of intervening repeats, but there is a slight increase in unitig size because of an increase in effective coverage due to the shredded reads. This same increase in unitig size is analogous to what would have been seen with more random

†To whom correspondence should be addressed. E-mail: mark.adams@celera.com.

‖The N50 length is the length *x* such that 50% of the sequence is contained in segments of length *x* or greater. Contig refers to contiguously assembled regions. Scaffolds are sets of contigs linked together by mate pairs, which are pairs of reads from the ends of subclones, where one mate is in one contig and the other is in the adjacent contig.

whole genome sequencing coverage, which was what the shredded reads were intended to approximate. There was no large-scale reconstruction of the shredded reads as stated by Waterston *et al.*, and the character of the unitigs was consistent with a pure WGS data set. The ordering of unitigs within contigs and scaffolds and the size of scaffolds in the whole genome assembly were entirely dependent on the mate-pair data from the WGS data set.

The remainder of their argument refers to gap filling and the contig N50 length. Gap filling is only effective once a scaffold exists with gaps to be filled between adjacent linked contigs. Scaffold construction was driven entirely by Celera mate-pair data, which resulted in very accurate contig order. After gap filling, we observed a $<1\%$ increase in total sequence length because the gaps were small. The critically important factor was knowing which gaps could be filled by using BAC-derived data without introducing assembly errors. Filling a reasonable *number* of gaps with a small *amount* of sequence serves to increase the contig N50 size while having a negligible impact on the overall assembly; there is no change in the base pair order or contig order, and a $<1\%$ increase in sequence length. This does not represent a substantive hidden contribution of HGSC data to the Celera assemblies.

One of the typical ways in which disputes about analytical techniques can be resolved is by exploring whether a method has been proven to work in other areas. The clear and definitive answer in reference to whole-genome shotgun sequencing is yes. The successful assembly of the mouse genome using a WGS strategy with an $\approx 5\times$ data set sequenced at Celera (6), and subsequently by a consortium that includes Waterston and coauthors (7), should remove any doubt about the applicability of the whole-genome strategy for a mammalian genome. We are pleased to note that Francis Collins (Director of the National Human Genome Research Institute at National Institutes of Health) commented that the consortium's mouse assembly was significantly better than the initial HGSC human assembly (www.sanger.ac.uk/Info/Press/020506.shtml). Our mouse genome assembly (6) was also better in many respects than the human assembly reported in ref. 2. It exhibited longer scaffolds, a higher fraction of the genome in scaffolds >1 Mbp, and only $\approx 3\%$ less sequence coverage than the human assembly (6). Comparison of the two mouse genome assemblies (8) and analysis of the finished *Drosophila* sequence (9) have provided additional documentation of the effectiveness of the whole-genome strategy.

In summary, Celera did produce an independent assembly of the human genome. In fact, it could readily be argued that the HGSC contribution to the Celera assembly was $\approx 1\%$ of sequence length and $\approx 35{,}000$ errors of order (figures 6 and 7 in ref. 2) that were overcome by reliance on Celera's mate-pair data. Further validation of the WGS method has now been demonstrated through sequencing of the mouse genome by a whole-genome shotgun strategy. The core principles that we applied to the *Drosophila,* human, and mouse assemblies include identification of unique and repeated sequences and use of mate pairs to link together adjacent regions. These features also formed the basis of an assembly program from the Lander group (10), for which a patent application has been filed (11).

Finally, we commend the HGSC on its continued efforts to finish the euchromatic portion of the human genome sequence. The improvements in quality and, more importantly, contiguity will serve all users of genome information, both public and private. Rather than "compete" in closure activities with the HGSC to obtain the last few percent of the genome, we decided 2 years ago that our scientific efforts were better spent in developing resources for interpreting the genome. After preparing an initial assembly and annotation of the mouse genome to facilitate comparative genomics (6), Celera and Applied Biosystems have gone on to develop and validate genome-wide reagents that are available to all to facilitate gene expression and genetics studies (http://store.appliedbiosystems.com) and to identify a large number of new polymorphisms that affect protein-coding regions.

We all share the same genome. Through the considerable creativity, dedication, and technical efforts of hundreds of scientists, the human genome continues to become a more effective tool in the study of human physiology and disease. We believe that it is time to get on with that important work (12, 13).

We are grateful for assistance with genome assembly comparisons from Art Delcher, Aaron Halpern, Daniel Huson, Clark Mobarry, Jason Miller, and Ross Lippert.

1. Waterston, R. H., Lander, E. S. & Sulston, J. E. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 3022–3024.
2. Venter, J. C., Adams, M. D., Myers, E. W., Li, P., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al.* (2001) *Science* **291,** 1304–1351.
3. Myers, E. W., Sutton, G. G., Smith, H. O., Adams, M. D. & Venter, J. C. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 3713–3714.
4. Venter, J. C., Adams, M. D., Sutton, G. G., Kerlavage, A. R., Smith, H. O. & Hunkapiller, M. (1998) *Science* **280,** 1540–1542.
5. International Human Genome Sequencing Consortium (2001) *Nature* **409,** 860–921.
6. Mural, R. J., Adams, M. D., Myers, E. W., Smith, H., Gabor Miklos, G. L., Wides, R., Halpern, A., Li, P. W., Sutton, G. G., Nadeau, J., *et al.* (2002) *Science* **296,** 1661–1671.
7. Mouse Genome Sequencing Consortium (2002) *Nature* **420,** 520–562.
8. Celniker, S. E., Wheeler, D. A., Kronmiller, B., Carlson, J. W., Halpern, A., Patel, S., Adams, M. D., Champe, M., Dugan, S. P., Frise, E., *et al.* (2002) *Genome Biol.* **3,** RESEARCH0079.1–0079.14.
9. Xuan, Z., Wang, J. & Zhang, M. Q. (2002) *Genome Biol.* **4,** R1.1–R1.10.
10. Batzoglou, S., Jaffe, D. B., Stanley, K., Butler, J, Gnerre, S., Mauceli, E., Berger, B., Mesirov, J. P. & Lander, E. S. (2002) *Genome Res.* **12,** 177–189.
11. Batzoglou, S., Berger, B., Mesirov, J. P. & Lander, E. (2002) U.S. Patent Appl. 20020049547.
12. Subramanian, G., Adams, M. D., Venter, J. C. & Broder, S. (2001) *J. Am. Med. Assoc.* **286,** 2296–2307.
13. Collins, F. S. & Guttmacher, A. E. (2001) *J. Am. Med. Assoc.* **286,** 2322–2324.