On the sequencing and assembly of the human genome

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

Celera Genomics, 45 W. Gude Drive, Rockville, MD 20850

On June 26, 2000, Celera Genomics and the International Human Genome Sequencing Consortium (HGSC) announced at the White House the completion of the first assembly of the human genome and the completion of a rough draft, respectively. In February of 2001, the two teams simultaneously published their analyses of the genome sequences generated (1, 2). The joint announcement and subsequent publications were a result of long discussions among Celera and HGSC scientists on reducing the negative rhetoric and demonstrating to the public that both teams were working for the public good. Now three laboratory leaders from the public consortium, Waterston, Lander, and Sulston (WLS), argue that Celera did not produce an independent sequence of the human genome or meaningfully demonstrate the whole-genome shotgun (WGS) technique (3). This conclusion is based on incorrect assumptions and flawed reasoning.

Table 1. Shredded data does not inherently reassemble

Data set	Overlap criterion	No. of contigs	Mean size, kbp	N50 size,* kbp
2× shred of chromosome 22	100	781	43.2	2,488.5
2× shred of chromosome 22	94	2,433	13.8	256.0
Reconstruction of chromosome 22 in a 2× shred of all HGSC data	94	10,142	3.6	20.4
2× shred of all HGSC data	94	2,081,677	1.7	6.8

In isolation, a perfect 2× shred of chromosome 22 reassembles well. In the context of the entire genome and when a provision is made for imperfect overlaps, the degree of reassembly is much lower.
*Refers to the minimum length L such that 50% of all nucleotides are contained in contigs of length $\geq L$.

Our Starting Point Was a Shredding of Several Hundred Thousand Bactigs, Not of the HGSC Genome Assembly. The key assertion of WLS is that by using information from the HGSC, Celera's method implicitly retained the full assembly structure produced by the HGSC. This is incorrect. As described in table 2 of ref. 1, we combined our data with a uniformly spaced 2× shredding of 677,708 individual bactigs, contigs of bacterial artificial chromosomes (BAC) clones shotgun sequenced by the HGSC, *not* the genome assembly reported in ref. 2. The goal of including this sequence was to take advantage (with attribution) of the work of the HGSC to the extent that it would contribute additional sequence coverage. The global order and the overall sequence of the genome were determined by using the set of 27 million mate-paired reads generated at Celera. Mate-pairs are sets of reads that are adjacent to one another in the genome and serve to link together nearby segments to promote assembly. The 38.7-fold genome coverage spanned by these mate-pairs provided the long-range order (over millions of basepairs) of both assembly methods reported in ref. 1. Without the Celera data, the best assembly that we could have produced would have been the 677,708 completely unordered bactigs, assuming that every shredded bactig would reconstitute itself during assembly as is claimed by WLS.

Simulation Using Chromosome 22 Alone Leads to a Distorted View of Assembly. WLS use a simulation to argue that a uniformly spaced 2× shredding would naturally result in such a reassembly of the HGSC bactig data. However, this exercise was not applied to the genome. Rather, it was applied to a single finished high-quality chromosome, constituting only 1% of the genome. It is thus misleading for the following reasons. First, the assembly problem is 100 times more complex for the genome than for a single chromosome, as the complete genome contains approximately 100 times more copies of each repetitive element than chromosome 22. Second, the majority of the HGSC data was in 6–8 kbp bactigs that were sometimes overlapping and occasionally misassembled, and whose sequence accuracy was as poor as 4% error near the tips. So assembling a shredding of such sequence must permit differences in read overlaps, whereas assembling a shredding of a finished sequence need not. Celera's assembler considers all overlaps at 94% or greater similarity as *equivalent* (4) and uses the pairing of end-sequence reads as the principal factors for achieving accurate order. Traditional assemblers that make local decisions based on the degree of overlap similarity are intrinsically too error prone to be reliable at the scale of mammalian WGS. Third, unlike the contiguous sequence of chromosome 22 used in the simulation, the HGSC data available in September of 2000 consisted of 5% predraft sequence consisting of 1×–3× light-shotgun reads of BACs, 75% rough-draft unordered bactigs of BACs derived from 3×–5× shotgun data of each BAC, and only 20% finished sequences of individual BACs (table 2 of ref. 1).†

Assembly Simulation with a Real-World Scenario Shows No Implicit Reassembly. We repeated the simulation experiment of WLS, but under a progression of conditions to demonstrate the impact of these real-world factors. With 100% identity (Table 1, first row) required for overlap, chromosome 22 is reconstituted from shredded reads by Celera's whole-genome assembler to the same degree as in the simulation reported by WLS. But when imperfect overlaps are permitted (94% identity, second row), as is required to truly accommodate sequencing errors in the HGSC data, the impact of near-identity repeats just within chromosome 22 becomes apparent: a much larger number of contigs are generated. When assembled in the context of the remaining 99% of the

See companion article on page 3712 in issue 6 of volume 99.

*To whom reprint requests should be addressed. E-mail: MyersGW@celera.com.

†The HGSC data are described by WLS as a 7.5× data set, but it is not a 7.5× *random* shotgun data set. Different regions of the genome were represented by BACs that had been sequenced to different fold coverage. Having finished 12× sequence in one part of the genome does not improve the result in regions where there is only 2× or no data at all.

genome (third row), the reassembled sequence for chromosome 22 is even more fractured. Finally, if one looks at contig sizes over a shredding of all of the HGSC data, 80% of which is rough draft (fourth row), the picture is even worse. When one (*i*) permits error in the overlaps, (*ii*) expands the problem to 100% of the genome, (*iii*) considers that most of HGSC data is rough draft, and (*iv*) includes another 5.1× of Celera data, data that further involves polymorphic variation across 5 individuals, WLS's claim of "implicit reassembly" is seen to be completely unfounded.

We shredded the HGSC dataset to overcome errors inherent in HGSC unfinished sequence including low-quality bactig ends and bactig misassemblies, and we were not under any illusions that this was akin to random coverage of the genome. A 2× shredding was the minimal way to incorporate all the HGSC data while giving it the least weight in an assembly involving 5× of Celera data. The mate-pair data from Celera's whole-genome libraries was the driving force for assembly by both methods presented in ref. 1. Thus while neither the Compartmentalized Shotgun Assembly (CSA) nor Whole-Genome Assembly (WGA) represents a completely "pure" application of whole-genome sequencing, the whole-genome sequence dataset produced at Celera determined the structure and content of the genome assemblies.

There Are Substantial Quantitative and Qualitative Differences Between Celera's Published Sequence and That of the HGSC. Because Celera and the HGSC both published sequences giving about 2.6 Gbp of the genome and we used the HGSC data in GenBank, one might mistakenly conclude that the two results published in February of 2001 are identical. But parity in the amount of sequence does not imply equality in terms of order or content. Celera's assembly had substantially better long-range contiguity (half of Celera's sequence was in scaffolds over 3.56 Mbp long, whereas half of the HGSC sequence was in scaffolds under 0.27 Mbp long). Moreover, Celera's end-sequence reads empirically validated the high accuracy of the contigs and their order in scaffolds and showed the HGSC sequence to have an ordering mistake every 70 kbp (figure 7 of ref. 1).

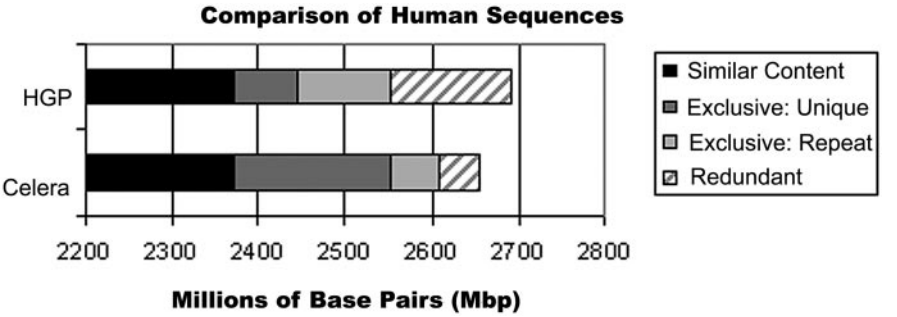

Fig. 1. Celera and HGSC reconstructions are not the same. The black segments, about 2.34 Gbp, agree between the two reported sequences. The dark gray segments represent sequence unique to an assembly that is essentially nonrepetitive, whereas the light gray segments represent repetitive DNA unique to each assembly. The gray hatched segments represent redundant data that should have been assembled with other sequences, and should therefore not be counted.

A sequence-level, whole-genome comparison further shows that there is substantial difference in the content of the two assemblies as summarized in Fig. 1. The HGSC assembly contains about 140 Mbp of redundant data that should have been assembled into the remainder of the genome. Celera's CSA assembly by contrast had only 50 Mbp of redundant data. If one removes these artifacts, Celera had 2.61 Gbp and the HGSC had 2.55 Gbp with about 420 Mbp represented in only one assembly or the other, a difference of 15.0% of the combined sequence. A majority of sequence unique to the HGSC assembly is in short segments of 1–3 kbp and is predominantly interspersed repetitive sequence. By contrast, most of the sequence unique to the Celera assembly is in large (>30 kbp) segments and is nonrepetitive in nature. The genome sequences reported (1, 2) are thus quite different, demonstrating that having 2.6 Gbp of data is not the same as having it properly assembled.

Celera's assembly was missing the interiors of highly similar repetitive elements and the extremely dense repeat regions near the centromeres, whereas the HGSC reconstruction was missing as much as 10% of the unique, information-rich parts of the genome. The basic explanation is that although we input the sum of the two data sets, the Celera assembler only outputs that portion of the genome that it can assemble with confidence.

Whole-Genome Shotgun Sequencing as the Paradigm for the Future. We have built the first of a new breed of assemblers for putting together ultra-large shotgun data sets. In 1995, when the *Haemophilus influenzae* genome was sequenced with a WGS approach (5), the assembler available at the time was not perfect, but it produced a result sufficient to finish the genome with a real economy of effort. Now prokaryotic genomes are routinely sequenced this way (www.tigr.org). The scenario today is the same as that of 1995 with respect to the WGS sequencing of large vertebrate genomes. We agree with the optimism of WLS that WGS will "play a useful role in obtaining a draft sequence from various organisms, including the mouse" (3). We produced a draft sequence of the mouse genome in June 2001 that has subsequently been of great use in permitting whole-genome analyses (e.g., refs. 6 and 7).

We remain resolute in our goal of providing the most accurate and complete version of the human genome for scientists to use in making scientific and medical breakthroughs. A careful, independent reevaluation of the approaches taken by the publicly funded labs could lead to many more genomes being accurately and rapidly sequenced to the benefit of the entire community.

1. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al*. (2001) *Science* **291,** 1304–1351.
2. International Human Genome Sequencing Consortium (2001) *Nature (London)* **409,** 860–921.
3. Waterston, R. H., Lander, E. S. & Sulston, J. E. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 3712–3716.
4. Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H., Remington, K. A., *et al*. (2000) *Science* **287,** 2196–2204.
5. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., *et al*. (1995) *Science* **269,** 496–512.
6. Young, J. M., Friedman, C., Williams, E. M., Ross, J. A., Tonnes-Priddy, L. & Trask, B. J. (2002) *Hum. Mol. Genet.* **11,** 535–546.
7. Zhang, X. & Firestein, S. (2002) *Nat. Rev. Neurosci.* **5,** 124–133.