

# On the sequencing of the human genome

Robert H. Waterston\*<sup>†</sup>, Eric S. Lander<sup>‡</sup>, and John E. Sulston<sup>§</sup>

\*Genome Sequencing Center, Washington University, Saint Louis, MO 63108; <sup>†</sup>Whitehead Institute/Massachusetts Institute of Technology Center for Genome Research, Cambridge, MA 02142; and <sup>§</sup>Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom

Communicated by Aaron Klug, Medical Research Council, Cambridge, United Kingdom, December 21, 2001 (received for review November 7, 2001)

**Two recent papers using different approaches reported draft sequences of the human genome. The international Human Genome Project (HGP) used the hierarchical shotgun approach, whereas Celera Genomics adopted the whole-genome shotgun (WGS) approach. Here, we analyze whether the latter paper provides a meaningful test of the WGS approach on a mammalian genome. In the Celera paper, the authors did not analyze their own WGS data. Instead, they decomposed the HGP's assembled sequence into a "perfect tiling path", combined it with their WGS data, and assembled the merged data set. To study the implications of this approach, we perform computational analysis and find that a perfect tiling path with 2-fold coverage is sufficient to recover virtually the entirety of a genome assembly. We also examine the manner in which the assembly was anchored to the human genome and conclude that the process primarily depended on the HGP's sequence-tagged site maps, BAC maps, and clone-based sequences. Our analysis indicates that the Celera paper provides neither a meaningful test of the WGS approach nor an independent sequence of the human genome. Our analysis does not imply that a WGS approach could not be successfully applied to assemble a draft sequence of a large mammalian genome, but merely that the Celera paper does not provide such evidence.**

Two scientific papers (1, 2) recently appeared reporting "draft" sequences of the human genome. One was the product of the international Human Genome Project (HGP), and the other was the product of the biotechnology firm Celera Genomics. The two groups set out by using different methodologies, and each collected independent data sets.

In principle, the availability of two papers on the human genome has much potential scientific benefit. In addition to the comparison of two independently derived genome sequences, it should also allow methodological analysis of the sequencing strategies used for insights concerning the design of future genome-sequencing efforts.

Here, we focus on the methodological issues of genome sequence assembly. In general, genomic sequencing projects employ the same basic technique of shotgun sequencing developed by Sanger and others shortly after the invention of DNA sequencing around 1980 (e.g., see ref. 3). To determine the sequence of a large DNA molecule, the method begins by breaking up the DNA into smaller random overlapping fragments, obtaining sequence "reads" from these fragments, then using computer analysis to reassemble the random reads into "contigs". Because of cloning biases and systematic failures in the sequencing chemistry, the random data alone are usually insufficient to yield a complete, accurate sequence. Instead, it is usually more cost-effective to supplement the random data with the collection of sequence data directed to close the gaps and solve remaining problems. The technique has been refined over the ensuing two decades. The initial version, for example, involved sequencing from one end of each fragment. Ansorge and others (4) extended this approach in 1990 to include the sequencing of both ends (paired-end shotgun sequencing), thereby obtaining linking information that could be used to connect contigs separated by gaps into "scaffolds".

The shotgun sequencing technique can be directly applied to genomes with relatively few repeat sequences. The assembly

problem is straightforward, because reads with overlapping sequence can typically be merged together without risk of misassembly. The relatively few gaps and problems can be solved to produce complete sequences. The approach has been applied successfully to produce complete sequences of simple genomes such as plasmids, viruses, organelles, and bacteria. Whole-genome shotgun data alone also has been applied with an almost 15-fold redundancy (5) to produce a draft sequence of the euchromatic portion of the *Drosophila* genome (3% repeat content), although a clone-based strategy is being applied to convert this to a finished sequence.

A greater challenge arises in tackling complex genomes with a large proportion of repeat sequences that can give rise to misassembly. Two alternative approaches (Fig. 1) can be taken.

**Hierarchical shotgun (HS) assembly.** In this approach, the genome is first broken up into an overlapping collection of intermediate clones such as bacterial artificial chromosomes (BACs). The sequence of each BAC is determined by shotgun sequencing, and the sequence of the genome is obtained by merging the sequences of the BACs. The HS approach provides a guaranteed route for producing an accurate finished genome sequence, because the sequence assembly is local and anchored to the genome. But it requires some additional preliminary work, including selecting overlapping BACs and preparing shotgun libraries from each BAC.

**Whole-genome shotgun (WGS) assembly.** In this approach, the genome is decomposed directly into individual random reads. One then attempts to assemble the genome as a whole. The WGS approach avoids the preliminary work but has potential disadvantages: there is a greater risk of long-range misassembly. The resulting sequence components must be individually anchored to the genome, and the resulting assembly may be difficult to convert to a finished sequence.

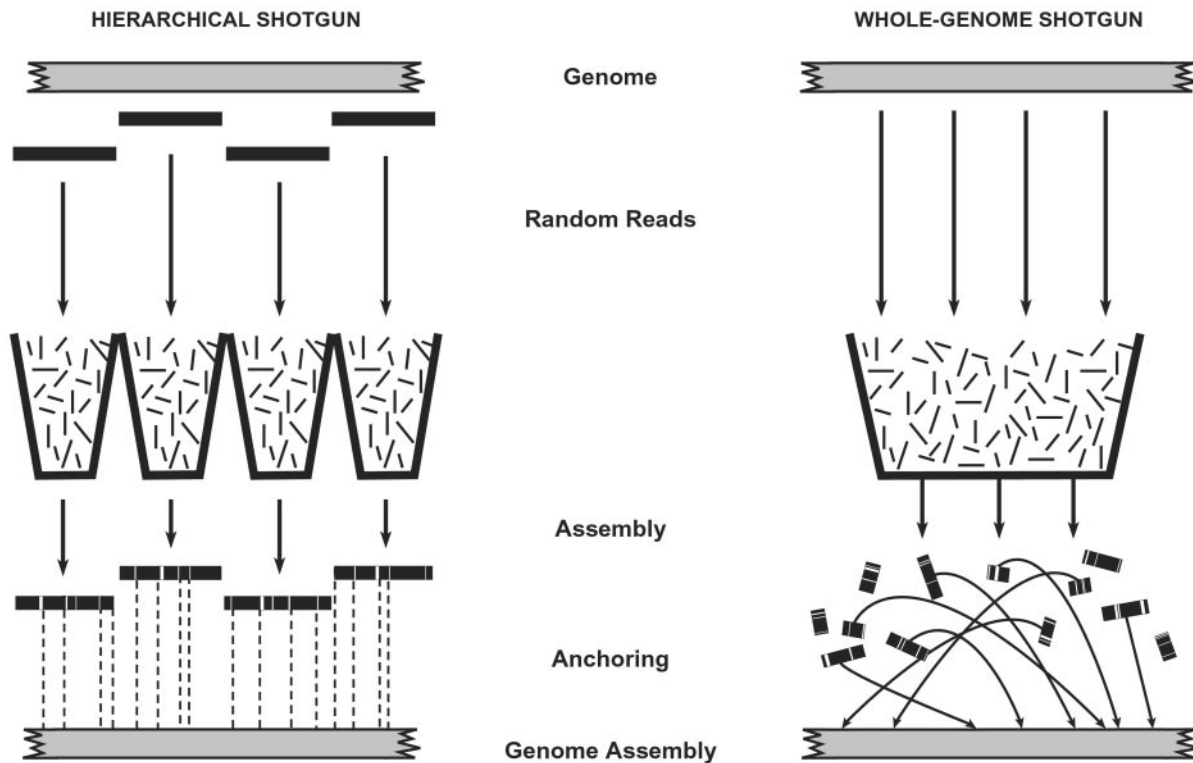
Whether to tackle the sequencing of the human genome with the HS or WGS approach was extensively debated in the scientific literature in 1996 and 1997 (6, 7). There was no doubt that the WGS approach could yield a large amount of the human sequence, but there was serious concern that the ultimate cost of producing a finished human reference sequence would be much greater. In fact, the potential cost savings in producing a draft sequence was unclear, because map construction and library production account for a minor fraction (<10%) of the total sequencing costs. For these reasons, the HGP elected to use the HS approach.

In 1998, Celera Genomics was formed with the goal of applying the alternative WGS approach to the human genome. Differing opinions were expressed concerning the likely product. Venter (8) projected that the WGS approach would suffice to assemble the entire genome in a small number of pieces. He

Abbreviations: HGP, Human Genome Project; HS, hierarchical shotgun; BACs, bacterial artificial chromosomes; WGS, whole-genome shotgun; STS, sequence-tagged sites; CSA, compartmentalized sequence assembly.

<sup>†</sup>To whom reprint requests should be addressed at: Genome Sequencing Center, Box 8501, 4444 Forest Park Boulevard, Saint Louis, MO 63108. E-mail: bwaterst@watson.wustl.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.



**Fig. 1.** Sequencing strategies. (*Left*) The hierarchical shotgun (HS) strategy involves decomposing the genome into a tiling path of overlapping BAC clones, performing shotgun sequencing on and reassembling each BAC, and then merging the sequences of adjacent clones. The method has the advantage that all sequence contigs and scaffolds derived from a BAC belong to a single compartment with respect to anchoring to the genome. (*Right*) Whole-genome shotgun (WGS) strategy involves performing shotgun sequencing on the entire genome and attempting to reassemble the entire collection. With the WGS method, each contig and scaffold is an independent component that must be anchored to the genome. In general, many scaffolds may not be anchored without directed efforts. (Contigs are contiguous blocks of sequence; scaffolds are sets of contigs joined by paired reads from both ends of a plasmid insert.)

estimated that an assembly based on 10-fold coverage would have fewer than 5,000 contigs separated by gaps of  $\approx 60$  bp. Such an assembly would include the nucleotide sequence of 99.99% of the human euchromatic genome and would consist of a small number of components that could then be positioned, or anchored, in the genome. Relatively little work would be required to produce a finished sequence. Olson (9) foresaw a different outcome, projecting there would be more than 100,000 components. That number of components would be far too many to anchor in the genome by using sequence-tagged sites (STS) as landmarks, resulting in a significant portion of the genome being unanchored. Moreover, he argued that it would be difficult or impossible to use such results as a foundation for producing a finished sequence.

With the recent publications on the human genome sequence (1, 2), it is possible to examine the reported results for insight concerning sequencing methodology. In particular, the Celera authors reported that their paper constituted a successful application of the WGS approach to a mammalian genome and that they had provided a genome sequence based primarily on their own data. Here, we examine the validity of these conclusions.

### Analysis and Results

**HGP Data and HS Assembly.** The HGP strategy is based on the sequencing of overlapping BACs ( $\approx 170$  kb) with known locations in the human genome. BACs are subjected to increasing levels of sequence coverage and completion: draft at  $\approx 5$ -fold coverage, deep shotgun at  $\approx 10$ -fold coverage, and finished resulting from directed gap closure.

At the time of publication (1), the BACs sequenced to at least draft status covered  $\approx 94\%$  of the euchromatic genome. The

merged sequence itself had fairly large contigs (half the sequence resided in contigs of  $>80$  kb) and represented  $\approx 90\%$  of the euchromatic portion of the human genome, roughly equally divided among draft, full shotgun, and finished status. The total sequence coverage was 7.5-fold.

**Celera Data and WGS Assembly.** The Celera strategy was based on assembling the genome from random sequences generated from both ends of whole-genome plasmid libraries with 2-, 10-, and 50-kb inserts. The authors generated a total sequence coverage of 5.1-fold.

The Celera authors presented no genome assembly based on their own WGS data (2). Thus, their paper provides neither a direct experimental test of the WGS method nor a direct assessment of the Celera data.

**Joint Assemblies.** The Celera paper presented only joint analyses based on a combined data set including both the HGP data (which had been made available on the world wide web, in keeping with the HGP's policy of free and immediate data release) and Celera's own data. The paper reported two joint analyses: a "faux" WGS assembly and a compartmentalized sequence assembly (CSA).

The methods are discussed below, and their output is summarized in Table 1. Notably, the joint assemblies do not contain dramatically more total sequence than the HGP assemblies that were used as input. Both the HGP assembly and the joint assembly based on the HGP and Celera data contain  $\approx 90\%$  of the human euchromatic genome. To be sure, the joint assembly contains some additional sequence (estimated to be a few percent) and adds additional ordering information. But the

**Table 1. Reported statistics for genome assemblies in the HGP and Celera papers**

Category	Celera			HGP
	WGS	Faux WGS	Faux CSA	
Sequence coverage	5.1 × Celera	5.1 × Celera + 7.5 × HGP	5.1 × Celera + 7.5 × HGP	7.5 × HGP
Length (in Gb) of draft genome assembly, counting only bases with known sequence*	NR	12.6 × total 2.587	12.6 × total 2.654	2.693
Length (in Gb) of draft genome assembly, including unknown nucleotides in gaps <sup>†</sup>	NR	2.848	2.906	2.916
Proportion of sequence in euchromatic genome present in draft genome assembly, % <sup>‡</sup>	NR	89	91	92
Number of contigs <sup>§</sup>	NR	221,036	170,033	149,821
Number of scaffolds <sup>§</sup>	NR	118,968	53,591	87,757
Number of components, to be anchored in genome <sup>  </sup>	NR	118,968	3,845	942

NR, not reported. The HGP and Celera papers differ in how assembly statistics are generally described. The HGP paper typically cites the number of known nucleotides in the assembly, excluding the unknown bases in the gaps. The Celera paper generally cites the total length “covered” or “spanned” by the assembly; this figure includes the roughly 0.25 billion unknown bases in gaps. However, comparable numbers can be extracted from the two papers.

\*For Celera, see table 3 in ref. 2. For HGP, see table 8 in ref. 1.

<sup>†</sup>For Celera, see table 3 in ref. 2. For HGP, see second footnote to table 8 in ref. 1.

<sup>‡</sup>Number of known bases of draft genome sequence divided by total length of euchromatic portion of human genome (2.92 Gb).

<sup>§</sup>For Celera, see table 3 in ref. 2. For HGP, see table 7 in ref. 1.

<sup>||</sup>For Celera’s faux WGS, see table 3 in ref. 2. For Celera’s CSA, see ref. 2, p. 1313, column 2, paragraph 3. For HGP, see tables 7 and 8 in ref. 1.

<sup>||</sup>Components refers to the units that must be independently anchored in the genome. These are scaffolds in the case of WGS and clone contigs in the case of CSA and HGP.

differences in the results using the combined data sets and the HGP data alone are relatively slight.

**Faux WGS Reads.** The manner in which the joint assemblies used the HGP data are noteworthy. The Celera authors stated that they combined 2.9-fold coverage from the HGP with 5.1-fold coverage from Celera. However, a close reading of the paper shows that the 2.9-fold coverage was derived in an unusual manner that is very unlike shotgun data and implicitly preserved much of the HGP assembly information.

The authors “shredded” the HGP’s assembled sequence data into simulated reads of 550 bp, which they termed “faux reads”. Each BAC was shredded to yield 2-fold coverage; given the overlaps between BAC clones, this yielded a total of 2.9-fold coverage. The authors then fed these faux reads into their assembly program, together with their own WGS reads. The stated purpose of shredding the HGP data was to break any misassemblies; this goal is a reasonable one. However, the shredding was done in such a manner that the resulting assembly bore no relation to a WGS assembly. Specifically, **the faux reads were not a random 2-fold sampling, but instead comprised perfect 2× coverage across the assembled HGP contigs. In other words, the faux reads were perfectly spaced, with each overlapping the next by half its length, thereby completely avoiding the problems of gaps, small overlaps, and errors that arise in realistic data (Fig. 2).**<sup>¶</sup>

**Faux WGS Assembly.** The first joint assembly (faux WGS) was modeled on the WGS approach in the sense that the sequence reads (both actual and faux) were fed into a genome assembly program without additional information. In particular, the genome assembly program did not have explicit information about the overlaps or location of the faux reads in the HGP sequence contigs.

We were curious, however, whether the HGP assembly infor-

mation was *implicitly* preserved through use of a perfect 2-fold tiling path. We tested this hypothesis by creating simulated data sets from the finished sequence of human chromosome 22 and performing genome assemblies by using a WGS assembly program, ARACHNE (10, 11).

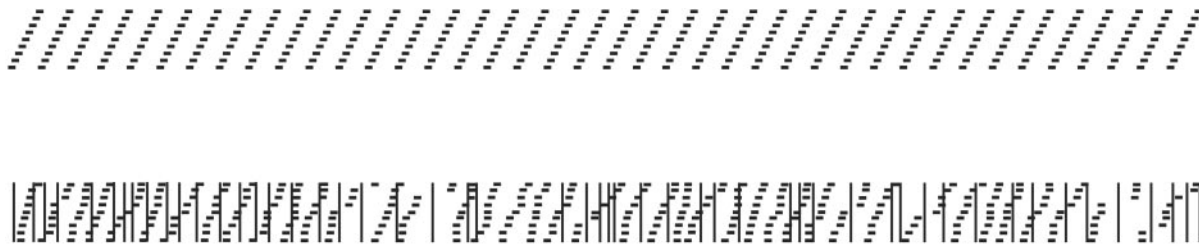
- Set A (2× perfect tiling path) consisted of a perfect tiling path of 550-bp reads, each overlapping the next by 275 bp, across the finished sequence.
- Set B (2× random coverage) consisted of randomly chosen reads of 550 bp, providing a total of 2-fold coverage.
- Set C (5× random coverage) consisted of randomly chosen reads of 550 bp, providing a total of 5-fold coverage.

We first compared the results of assembling sets A and B (Table 2). The perfect 2× tiling path (set A) yields an assembly covering essentially the entire chromosome in huge sequence contigs. More than 99% of the sequence lies in contigs >10 kb, and more than 43% in contigs >500 kb. The N50 contig length (the length  $L$  such that 50% of the sequence lies in contigs of at least  $L$ ) is 421 kb. The huge size of the contigs is not surprising, because the underlying data have no true gaps, and the reads have large overlaps that are readily detected by a shotgun assembly algorithm. By contrast, the random 2× coverage (set B) yields tiny contigs. The N50 contig length is <2 kb for the random coverage compared with 421 kb for the perfect tiling path. This result is also not surprising, because the random 2-fold coverage necessarily leaves many true gaps and small overlaps (12). Thus, shotgun assembly programs cannot assemble long contigs from such data.

We next compared the results for sets A and C (Table 2). The 2× perfect tiling path yields dramatically larger contigs than 5× random shotgun data. The N50 length is almost 40-fold larger for the perfect tiling path than for the 5× random data (421 vs. 11 kb).

These results show that a fundamental limitation in genome sequence assembly is the random nature of the data. A perfect tiling path avoids this issue and implicitly preserves the underlying assembly information.

<sup>¶</sup>See ref. 2. The nature of “shredding” is mentioned in passing on p. 1309, column 3, paragraph 4; the implications for assembly are not discussed in the paper.



**Fig. 2.** Random vs. perfect spacing in 100 kb. (*Upper*) Two-fold coverage in perfectly spaced reads. (*Lower*) Two-fold coverage in randomly selected reads. There are frequent regions in which the adjacent reads either fail to overlap or the overlap is too small to allow reliable detection (< 40 bp). These breaks in continuity are indicated by vertical lines.

The results have several clear implications for the faux WGS assembly. First, dissection of the HGP contigs into a  $2\times$  perfect tiling path suffices to allow their nearly complete reconstruction (even up to hundreds of kilobases) without the explicit need for positional information. Second, a  $2\times$  perfect tiling path contains much more inherent assembly information than  $5\times$  random coverage. It is not surprising, then, that the combined assembly quite faithfully (although not completely) reproduced the finished sequence of chromosome 22.

Thus, for the portions of the genome where sequence was available from the HGP, it is impossible to learn from the paper the respective contributions of the two data sets. On the one hand, the WGS reads can fill gaps, link contigs, and correct misassemblies. On the other hand, sequence from the BAC contigs was used directly in a final assembly step to fill gaps (in a process the authors referred to as “external gap walking”). In any case, the fact is that there is very little difference in the coverage and continuity of the faux WGS assembly using the two data sets and the HGP genome sequence.

**Anchoring the Faux WGS Assembly.** Although the faux WGS assembly is not a meaningful test of a true WGS assembly, it is nonetheless worth examining for insight about other issues pertinent to genome assembly. In particular, a key measure of the faux WGS assembly is the proportion of the genome sequence contained in large, anchored components. Although all sequence produced by the HS strategy can be localized to a small region of the genome (namely, a mapped BAC clone), the sequence components produced by the WGS assembly are initially free-floating islands that do not contribute to a genome assembly until they have been anchored. In the current case, the Celera authors aimed to anchor their sequence components by using the HGP’s STS maps.

The faux WGS assembly yielded 221,036 contigs linked to-

gether into 118,986 scaffolds (sets of contigs joined by paired reads from both ends of a plasmid insert) and contained a total of 88.6% of the euchromatic genome (Table 1). If one focuses only on scaffolds larger than 30 kb, their number is small enough to allow them to be anchored. But these scaffolds contain only  $\approx 2.334 \times 10^9$  bases (see table 3 in ref. 2) or  $\approx 79.9\%$  of the euchromatic genome, leaving  $\approx 20\%$  of the euchromatic genome unanchored.

To anchor more of the genome, the full set of nearly 119,000 scaffolds must be used. But this quantity vastly exceeds the number that can be anchored by using existing STS maps. Localizing so many scaffolds would require a large directed mapping effort, which could only commence after the assembly was completed.

Moreover, these 119,000 scaffolds still contain only 88.6% of the euchromatic genome. The remaining 11.4% lies in the 25% of the reads that remain unassembled or is missing entirely from the WGS coverage. Clearly, it is not feasible to anchor this portion of the genome.

**Compartmentalized Sequence Assembly.** The second joint assembly (CSA) used a local clone-based rather than a genome-wide WGS approach. In fact, the CSA was conceptually identical to the HGP’s HS approach, and it explicitly used all of the HGP’s clone-based sequence and map data.

Briefly, the CSA analysis began by assigning Celera’s WGS reads to individual HGP BAC clones (by matching their sequences to the assembled sequence contigs) then to overlapping sets of HGP BACs, dubbed “compartments” (there was a total of 3,845 HGP compartments). The CSA then performed separate local sequence assembly on each of the compartments (using both the Celera WGS reads and faux HGP reads corresponding to the compartment). The number of compartments was small enough that nearly all could be readily anchored to the chromosomes by using the available HGP map resources.

**Table 2. Implications on assembly using perfect tiling paths of faux data based on simulated data from finished sequence from human chromosome 22\***

	$2\times$ perfect tiling path	$2\times$ random coverage	$5\times$ random coverage
N50 contig length <sup>†</sup> (kb)	421 kb	<2 kb	11 kb
Sequence in contigs > 5 kb, %	>99	6.4	80
Sequence in contigs > 10 kb, %	99	0.2	55
Sequence in contigs > 20 kb, %	98	†	20
Sequence in contigs > 50 kb, %	94	†	0.5
Sequence in contigs > 100 kb, %	87	†	†
Sequence in contigs > 500 kb, %	43	†	†
Sequence in contigs > 1000 kb, %	29	†	†

\*The finished sequence of human chromosome 22 was decomposed into random reads of length 550 bp, with the indicated total coverage. The reads were either randomly selected or chosen to comprise a perfect tiling path with overlaps between consecutive reads having constant size. The reads were assembled into contigs using the WGS assembly program ARACHNE (10,11).

<sup>†</sup>Refers to the length  $L$  such that 50% of all nucleotides are contained in contigs of length  $\geq L$ .

† = less than 0.1%.

The construction of compartments closely mirrored the HGP's construction of clone contigs and used the HGP clone sequences, STS content, and fingerprint clone maps.<sup>||</sup> The local sequence assembly was similarly straightforward. The compartments had an average size of  $\approx 760$  kb (only a few times larger than a BAC clone) and were readily assembled by using standard computer programs such as PHRAP (by P. Green, available at <http://www.phrap.org/>). The resulting scaffolds were anchored and then further ordered and oriented by using the HGP's STS content and fingerprint clone maps.

The CSA thus provides a revised version of the HGP assembly based on the addition of WGS reads to the individual clone contigs. All of the biological analyses reported in the Celera paper were based on the CSA sequence.

## Discussion

Here, our primary purpose is to examine whether the recently published paper from Celera Genomics (2) provides insight into the performance of the WGS approach on mammalian genomes. Our analysis indicates that it is not possible to draw meaningful conclusions about the WGS approach because the authors did not perform an analysis of their own data by itself. Instead, they used an unorthodox approach to incorporate simulated data from the HGP. In particular, the paper presented only joint assemblies of Celera's 5.1-fold WGS data together with a perfect tiling path of faux reads that implicitly retained the full information inherent in the HGP's 7.5-fold coverage. Furthermore, the joint assemblies were anchored to the genome by using the HGP's clone and marker maps.

We should emphasize that our analysis does not imply that a WGS approach cannot, in principle, produce a valuable draft sequence of a mammalian genome. To the contrary, we are optimistic that this approach will be possible with improved computational algorithms. Indeed, we expect that the WGS approach can play a useful role in obtaining a draft sequence from various organisms, including the mouse (13). For the mouse, with 6-fold WGS coverage (13) and the use of improved algorithms (11), it has yielded initial assemblies with N50 scaffold lengths exceeding 1 Mb and N50 contig lengths of 16 kb. Conceivably, such improved algorithms could yield similar results from human WGS data.

Whether the WGS approach alone can provide an efficient technique for producing a finished genome sequence, however, remains an open question. To obtain the complete sequence of the mouse genome (13), the WGS data are being used in conjunction with clone-based data. The WGS assembly has been anchored to an extensive BAC-based physical map by matching the sequence contigs to BAC-end sequences, thereby localizing numerous small contigs and providing longer range continuity. Moreover, the BACs serve as substrates for producing a finished sequence, as is being done in the human and fly. Current experience suggests that clone-based sequencing will remain an essential aspect of producing finished sequence from large, complex genomes.

Although the Celera paper leaves open many methodological issues, it does demonstrate one of the HGP's core tenets, the value of making data freely accessible before publication. As the analysis above shows, the availability of the HGP data contributed to Celera's ability to assemble and publish a human genome sequence in a timely fashion. When speed truly matters, openness is the answer.

<sup>||</sup>See ref. 2, p. 1313, column 2, paragraph 3 and p. 1314, column 2, paragraph 2.

1. International Human Genome Sequencing Consortium (2001) *Nature (London)* **409**, 860–921.
2. 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.
3. Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. & Petersen, G. B. (1982) *J. Mol. Biol.* **162**, 729–773.
4. Edwards, A., Voss, H., Rice, P., Civitello, A., Stegemann, J., Schwager, C., Zimmermann, J., Erfle, H., Caskey, C. T. & Ansorge, W. (1990) *Genomics* **6**, 593–608.
5. Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., *et al.* (2000) *Science* **287**, 2185–2195.
6. Weber, J. L. & Myers, E. W. (1997) *Genome Res.* **7**, 401–409.
7. Green, P. (1997) *Genome Res.* **7**, 410–417.
8. Marshall, E. & Pennisi, E. (1998) *Science* **280**, 994–995.
9. Marshall, E. (1999) *Science* **284**, 1906–1909.
10. Batzoglou, S. (2000) Ph.D. thesis (Massachusetts Institute of Technology, Cambridge).
11. Batzoglou, S., Jaffe, D. B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J. & Lander, E. S. (2001) *Genome Res.* **12**, 177–189.
12. Lander, E. S. & Waterman, M. S. (1988) *Genomics* **2**, 231–239.
13. Mouse Genome Sequencing Consortium (2001) *Genesis* **31**, 137–141.

# 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.

**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

**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  $\geq L$ .

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).<sup>†</sup>

**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.

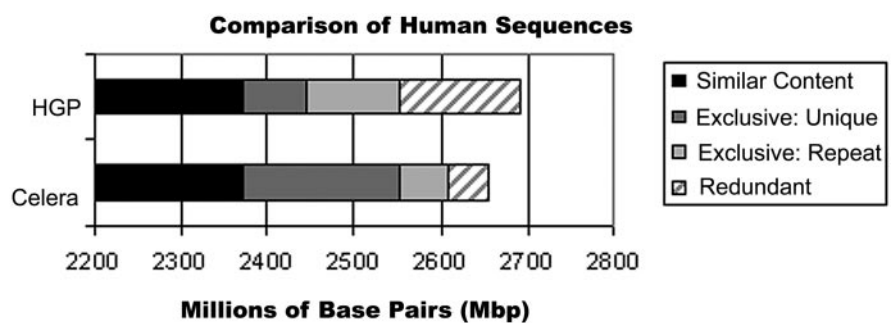
\*To whom reprint requests should be addressed. E-mail: MyersGW@celera.com.

<sup>†</sup>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 bac-tig ends and bactic 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



**Fig. 1.** Celera and HGP 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.

showed the HGSC sequence to have an ordering mistake every 70 kbp (figure 7 of ref. 1).

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 non-repetitive 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 out-

puts 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](http://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.

- 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.
- International Human Genome Sequencing Consortium (2001) *Nature (London)* **409**, 860–921.
- Waterston, R. H., Lander, E. S. & Sulston, J. E.

- (2002) *Proc. Natl. Acad. Sci. USA* **99**, 3712–3716.
- 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.
- 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.
- Young, J. M., Friedman, C., Williams, E. M., Ross, J. A., Tonnes-Priddy, L. & Trask, B. J. (2002) *Hum. Mol. Genet.* **11**, 535–546.
- Zhang, X. & Firestein, S. (2002) *Nat. Rev. Neurosci.* **5**, 124–133.

# Whole-genome disassembly

Phil Green\*

Howard Hughes Medical Institute and University of Washington, Seattle, WA 98195

The race to sequence the human genome has garnered a level of popular attention unprecedented for a scientific endeavor. This fascination has partly been caused of course by the importance of the goal; but it also reflects the Olympian nature of the contest, which opposed two capable teams with sharply contrasting cultures (public and private), personalities, and strategies. Titanic struggles being the stuff of mythology, it should perhaps not surprise us that a number of myths regarding this race have already emerged. In a recent issue of PNAS, Waterston *et al.* (1), leaders of the public effort, help to dispel one of these myths, involving the controversial “whole-genome shotgun” strategy used by Celera.

Issues surrounding sequencing strategies will no doubt seem arcane to most readers but are worth considering if only because they may significantly influence the pace and cost of DNA sequencing during the remainder of the Genome Era. That a strategy is needed at all arises from the fact that a sequencing “read,” the tract of data obtainable in a single experimental run, is only a few hundred bases in length and contains errors. Getting reliable sequence of a larger DNA segment therefore requires a method for generating and piecing together a number of reads covering the segment. Since its introduction by Sanger and colleagues over 20 years ago, the favored method for this purpose has comprised the following steps: an initial “shotgun” phase in which reads are derived from subclones essentially randomly located within the targeted region; an assembly phase, in which read overlaps are determined (the main challenge here being to identify and discard false overlaps arising from repeated sequences) and used to approximately reconstruct the underlying sequence; and a finishing phase in which additional reads are obtained in directed fashion to close gaps and shore up data quality where needed. The shotgun phase usually involves obtaining a substantial redundancy of read coverage of the target, typically at least 6–8-fold, to minimize the amount of work required during the labor-intensive finishing phase.

For the human genome, which comprises some 3 billion base pairs, the public effort adopted a well-tested modular ap-

proach in which large fragments of the genome (roughly 150,000 bp in size) were first cloned into a bacterial host (as bacterial artificial chromosomes or BACs) and then sequenced individually by the shotgun method. Among other advantages, this “clone by clone” strategy simplifies the assembly problem (by reducing its scale and the likelihood of errors caused by repeats), generates substantial sequence tracts of known contiguity that can be mapped relatively efficiently back to the genome, and yields resources that are useful in the finishing stage and for independent tests of assembly accuracy. A “draft” version of the genome sequence (based on a somewhat lower shotgun depth coverage for most of the clones) obtained in this way was published last year (2).

In contrast, Celera adopted a whole-genome shotgun approach, which purports to accelerate the above process by bypassing the intermediate step of cloning large fragments and instead derives reads directly from the whole genome. The process is clearly riskier because of the significantly greater possibility of assembly error, but had been successfully used by Celera to produce a near-complete sequence of the *Drosophila* genome (3, 4) with about 2,500 gaps. Its ability to cope with the human genome, which is 30-fold larger and much richer in repetitive sequences than *Drosophila*, remained unclear. Against all odds, Celera demonstrated that it worked (5), producing an independent human genome sequence of comparable or higher quality than that obtained by the public effort.

Or did they?

This is the myth that Waterston *et al.* (1) overturn. Far from constructing an independent sequence, Celera incorporated the public data in three important ways into their “whole genome assembly.” (i) The assembled BAC sequences from the public project were “shredded” in a manner that (as Waterston *et al.* show) retained nearly all of the information from the original sequence, and used as input. (ii) In a process called “external gap walking,” unshredded, assembled, public BAC sequences were used to close gaps. (iii) Public mapping data were used to anchor sequence islands to the genome. As a

result, the assembly reported by Celera cannot be viewed as a true whole-genome shotgun assembly. Moreover, accuracy tests in ref. 5, which involved comparison of Celera’s assembly to finished portions of the public sequence, are virtually meaningless because the finished sequence was itself used in constructing the Celera assembly.

We are left with no idea how a true whole-genome assembly would have performed. It is striking, however, that even with this use of the public data, what Celera calls a whole-genome assembly was a failure by any reasonable standard: 20% of the genome is either missing altogether or is in the form of 116,000 small islands of sequence (averaging 2.3 kb in size) that are unplaced, and for practical purposes unplaceable, on the genome.

Several other myths beyond the one discussed by Waterston *et al.* have become widely accepted. One is that the whole genome shotgun approach was in large measure responsible for Celera’s rapid pace at sequencing the *Drosophila* and human genomes. In fact, their great speed was mainly because of the acquisition of a huge, unprecedented sequencing capacity (some 200+ capillary machines, each able to produce 500–1000 reads per day) as a result of their corporate ties with a manufacturer of these machines. That this was really the key factor is evident from the fact that when the public effort acquired similar capacity, they were able to attain a comparable or higher throughput by using the clone by clone approach.

A third myth is that the whole-genome approach saves money. Although definitive judgement here should await a rigorous cost accounting, the basic economics of sequencing by the clone by clone approach have apparently not changed greatly over the past 5 or 6 years. Less than 10% of the overall cost goes to BAC mapping and subclone library construction, 50–60% to the shotgun itself (assuming a coverage of 6–8 $\times$ ), and the remaining 30–40% goes to finishing. Even if it works as intended, the whole-genome approach can save at best the 10% involved

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

\*E-mail: phg@u.washington.edu.

in BAC mapping and subclone libraries; but, as was argued in ref. 6, even this minimal savings is likely to be negated, or worse, by inefficiencies created at the shotgun or finishing stage. The *Drosophila* project (3, 4) is a case in point (in fact, the only case we have). Celera generated shotgun coverage of nearly 15×, approximately double what is used in a clone by clone approach, which was necessitated in part by the effective loss of about 1/4 of the reads (“chaff”) that could not be incorporated into the whole-genome assembly. Moreover, the finishing process (being carried out by G. Rubin and colleagues) has involved generating reads on a clone by clone basis from a minimally overlapping set of mapped BACs spanning the genome. Thus, none of the costs that were supposed to be saved by the whole-genome shotgun in fact were, and the effective doubling of the cost of the shotgun itself significantly increased the cost of the whole project beyond that of a clone by clone approach.

A widespread view among many observers has been that, issues like the above aside, the genome race has in any case at least been good for science. In my view this also is a myth. Competition does have the beneficial effects of motivating the competitors to work harder and to critically challenge their opponents’ work (as

with the current paper by Waterston *et al.*), but it also has the downside of encouraging shortcuts that may compromise the ultimate result. In the case of the genome race, the downside seems to have outweighed the benefits. For example, Celera reduced the amount of shotgun data it generated from the originally intended 10× (7), which is probably the minimal amount necessary to afford any hope of success with a true whole-genome approach, to a mere 3.8× (their article reports 5.1×, but this must be reduced by the 26% lost as “chaff”), which incidentally is only about 1/2 the amount reported for the public project. As a result, it became impossible to objectively compare the two approaches. The competition probably did induce Celera to eventually provide greater access to their data than they otherwise would have, although this access is under terms that fall substantially short of the original promise (7) to deposit the data in the public databases, and fails to uphold the essential principle that scientific discoveries should form a basis on which other scientists are free to build.

In my view, the effect of the competition on the public side was also undesirable. Although their rate of sequence production accelerated greatly after the competition was engaged, this was mainly attributable to the availability of the

higher throughput new technology (capillary sequencers). Partially offsetting the throughput gains were apparently gross inefficiencies in the process of sequence acquisition that resulted from the pressure to rapidly process BACs before a minimally overlapping set had been identified. Furthermore, it remains quite unclear whether the decision to produce an intermediate quality product (the draft) will prove wise in the long run; although the major centers have stated a commitment to finish the genome, motivation of many participants has surely been reduced now that the project is regarded by the public as complete. It remains to be seen whether a truly finished genome will appear by next year as promised.

Is there a lesson in this? I am not sure there is one. Competition is of course a basic fact of nature that we cannot and should not eliminate. The undesirable results it may have produced in this case—widespread misinformation, exaggerated claims, and a compromised product—are mostly due to the high-profile nature of the contest, and perhaps also to the fact that a significant amount of corporate money was riding on the perceived success of one team. The best that those of us on the sidelines can do is to continue to scrutinize the results.

1. Waterston, R. H., Lander, E. S. & Sulston, J. E. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 3712–3716.
2. International Human Genome Sequencing Consortium (2001) *Nature (London)* **409**, 860–921.
3. Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer,

- S. E., Li, P. W., Hoskins, R. A., Galle, R. F., *et al.* (2000) *Science* **287**, 2185–2195.
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. 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**, 1304–1391.
6. Green, P. (1997) *Genome Res.* **7**, 410–417.
7. Venter, J. C., Adams, M. D., Sutton, G. G., Kerlavage, A. R., Smith, H. O. & Hunkapiller, M. (1998) *Science* **280**, 1540–1542.

# Revisiting the independence of the publicly and privately funded drafts of the human genome

The field of human genetics had a landmark year in 2001. Two groups published a draft sequence of the human genome at the same time (1, 2). One paper was from the International Human Genome Project (HGP) (1), a public group that provided open access to their sequence as it was obtained. The other paper was authored by a private group, Celera Genomics (2), who do not provide immediate or free access to their sequence data. Accordingly, Celera used both the publicly available data and their independently generated data in compiling their draft of the human genome. The public group used a conservative, divide-and-conquer strategy in which segments of the genome were first cloned into bacteria to yield bacteria artificial chromosomes (BACs). These BACs were mapped and then sequenced. The private group used a bolder, whole-genome shotgun (WGS) method in which genome-wide sequence fragments are assembled computationally.

In March of 2002, Waterston, Lander, and Sulston, leaders in the HGP, claimed that the public data played a critical enabling role in the Celera assembly (3). They argued that Celera's highly ordered shredding of the HGP data preserved so much of its long-range assembly that the Celera assembly was not independent and was not achieved by the WGS method. PNAS invited two commentaries on this paper. One, by Green (4), concluded that be-

cause Celera's work was anchored by sequence islands obtained from the public database, the assembly reported by Celera could not be viewed as a true WGS assembly. The second commentary, by some of the leaders of the Celera initiative (5), strongly disagreed with the allegations. They argued that Celera started with shredded bactigs and not with reassembled HGP data. They also criticized the simulation Waterston *et al.* used to support their argument and said it was based on incorrect and oversimplified assumptions.

From the feedback PNAS received after the publication of these papers, it was clear that the views presented were so divergent that it was very difficult for the scientific community to evaluate the claims. The exact areas of disagreement were not even clear. The major issues in the disagreement were submerged in the incompatible claims. The controversy is not about the validity of the WGS method for sequencing very large genomes, but about whether Celera used this method to assemble the human genome in 2001. If their results depended on the HGP, which is derived from a hierarchical method, then they did not assemble an independent draft sequence. There is no problem in Celera using the HGP data; it was there for everyone. The disagreement is in how it was used. Was it shredded so well that it only provided additional sequence reads to supplement the extensive ones by Celera? Or was much of the sequence integrity retained? There is also no

question that Celera produced a high-quality draft. The question is only whether the long-range assembly was independent.

Accordingly, we thought that one more round of comments would clarify the situation. Representatives from both the public HGP group and the private Celera group agreed to author a second and final round of perspectives. These papers can be found on pages 3022 and 3025 (6, 7). We used the following ground rules for these invited papers. In the first round, because the HGP group initiated the debate (3), we provided Celera an advance copy of the HGP paper so they could prepare their rebuttal commentary (5). We did not show the rebuttal to the HGP group. This second round of the debate has followed similar protocol. The HGP contribution (6), which addresses the first response by Celera and provides new arguments, was sent to Celera for their reply (7). We felt that Celera should have the last word as it was their assembly that was being questioned. Accordingly, the HGP authors did not see, and thus could not comment on, the Celera response.

We asked both groups to stick to the scientific issues, to avoid recrimination, and to explain technical terms clearly. Both contributions were reviewed and revised. We believe that this second round of the debate has clarified the issues surrounding this epochal contribution to science.

Nicholas R. Cozzarelli

1. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001) *Nature* **409**, 860–921.
2. 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.

3. Waterston, R. H., Lander, E. S. & Sulston, J. E. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 3712–3716.
4. Green, P. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 4143–4144.
5. Myers, E. W., Sutton, G. G., Smith, H. O., Adams,

M. D. & Venter, J. C. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 4145–4146.

6. Waterston, R. H., Lander, E. S. & Sulston, J. E. (2003) *Proc. Natl. Acad. Sci. USA* **100**, 3022–3024.
7. Adams, M. D., Sutton, G. G., Smith, H. O., Myers, E. W. & Venter, J. C. (2003) *Proc. Natl. Acad. Sci. USA* **100**, 3025–3026.

# More on the sequencing of the human genome

Robert H. Waterston<sup>\*†</sup>, Eric S. Lander<sup>‡</sup>, and John E. Sulston<sup>§</sup>

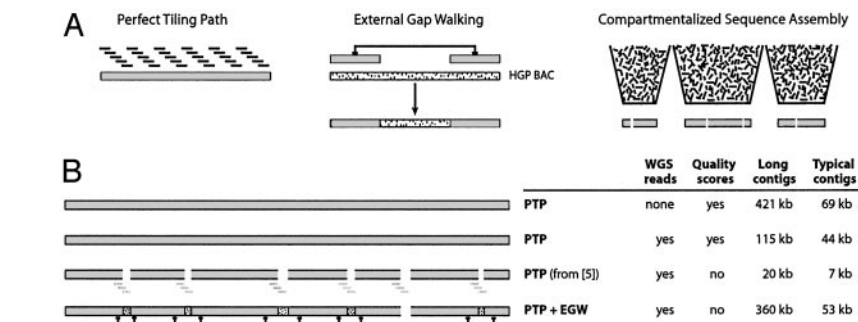
<sup>\*</sup>Department of Genome Sciences, University of Washington, Seattle, WA 98195; <sup>‡</sup>Whitehead Institute/Massachusetts Institute of Technology Center for Genome Research, Cambridge, MA 02142; and <sup>§</sup>Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom

The international Human Genome Project (HGP) and Celera Genomics published articles last year on the sequence of the human genome (1, 2). In a recent article (3), we analyzed aspects of the Celera article.

We noted that the article did not report an assembly of Celera's own data but rather reported only joint assemblies based on a data set that included the assembled genome sequence of the HGP. Approximately 60% of the underlying sequence data and 100% of the mapping data used in Celera's analysis came from the HGP, and the HGP genome assembly itself contained 90% of the euchromatic sequence of the human genome. We also noted that Celera used various approaches for using the HGP data (referred to as perfect tiling, gap filling,<sup>†</sup> and compartmentalized assembly; see Fig. 1) that implicitly preserved much of the HGP assembly information. We concluded that Celera's assemblies made extensive and inextricable use of the HGP genome information and thus were not an independent assembly of the human genome.

Our critique was not concerned with whether the Celera authors could have produced an independent assembly with their own data; it simply noted that the article (2) did not do so. It did not address the potential utility of the whole-genome shotgun (WGS) approach; it simply noted that the article was not a meaningful application of the approach. (The utility of WGS for producing draft genome sequence has not been the question. The issue has been whether it provides the best route for producing a finished genome sequence of a complex mammalian genome such as the human or whether clone-based sequencing is more effective; experience to date strongly suggests the latter.) Also, it was not concerned with which strategy or assembly was better; because Celera's assembly made use of both data sets, meaningful comparisons are impossible.

Our report elicited two commentaries. One, by Green (4), concurred with our analysis. The other, by Myers *et al.* (five of the Celera authors), raised certain issues about our analysis (5). Specifically, they acknowledge that their approaches preserved the HGP assembly to some extent, but they contend that



**Fig. 1.** Uses of the HGP genome assemblies in Celera genome assemblies and impact on assembly. (A) HGP data were used in three ways. (i) Perfect tiling. HGP's contigs were decomposed ("shredded") into perfect tiling paths of uniformly overlapping "faux" reads with no gaps (3). (ii) Gap filling. Gaps between linked contigs (gray) in assemblies were filled by directly taking sequence from unshredded assembled HGP BACs (stippled). (iii) Compartmentalized assembly. The main assembly in ref. 2, used for all biological analyses, was obtained by first matching the WGS reads of Celera to small compartments corresponding to overlapping HGP BACs and then assembling each compartment locally. (B) Average N50 contig lengths for reconstructing a perfect tiling from either a long (>1 megabase) or typical (length distribution as in the HGP assembly) contig. The four lines correspond to assembly of the perfect tiling reads alone or with WGS reads (>100 $\times$ ) and either with or without quality scores. The third line (from ref. 5) has short contigs because lack of quality scores limits overlap detection. The fourth line shows that these limitations are substantially overcome by gap filling (see *Gap Filling and Compartmentalized Assembly Extend the Reconstruction*).

the role of the HGP data in the Celera joint assemblies was minor.

Here we address the technical issues raised by Myers *et al.* We show that the analysis of Myers *et al.* underestimates the role of the HGP genome assembly in their work because they focus on only one of the ways in which the HGP data were used. Moreover, we note that the major role of the HGP sequence can be directly seen from the properties of the Celera assembly.

## Assembly Reconstruction

Genome assemblies are characterized by the degree of sequence continuity, measured by N50 contig length.<sup>||</sup> The HGP draft genome sequence had an N50 contig length of  $\approx 82$  kb (see table 7 of ref. 1).

Our main point in ref. 3 was that the approaches of Celera implicitly preserve most HGP assembly information at such length scales. That is, typical-size contigs can be reconstructed largely from the HGP data.

By contrast, Myers *et al.* assert that Celera's analysis only preserved the HGP assembly up to scales of  $\approx 7$  kb in their WGS assembly (see table 1 of ref. 5).

Even this degree of implicit reconstruction is quite substantial. It is  $\approx 13$ -fold larger than typical read lengths ( $\approx 0.5$  kb) and only  $\approx 12$ -fold smaller than assembled HGP contigs ( $\approx 82$  kb). It dramatically simplifies further assembly.

Thus, Myers *et al.* substantially underestimate the use of the HGP assembly, because the analysis focuses only on the impact of perfect tiling but does not consider the other uses of the HGP data.

## Perfect Tiling Begins the Reconstruction

In ref. 3 we illustrated that a genome assembly could be largely reconstructed from a perfect tiling. Specifically, we showed that a WGS assembly program (6) can reconstruct long sequence contigs from a perfect tiling of the completed chromosome 22 sequence. Most

<sup>†</sup>To whom correspondence should be addressed at: Department of Genome Sciences, University of Washington, Box 357730, Seattle, WA 98195-7730. E-mail: waterston@gs.washington.edu.

<sup>||</sup>Gap filling is referred to as "external gap walking." The device is mentioned in passing on pg. 1312 of ref. 2; the implications for assembly are not discussed in the article.

<sup>||</sup>The N50 length is the length  $x$  such that 50% of the sequence is contained in contigs of length  $x$  or greater.

remaining gaps could then be filled by gap filling.

Myers *et al.* suggest that this example was not fully realistic, because perfect tilings would need to be reassembled among a vast sea of WGS reads. We tested this point using our previous criteria (3) but found it had no significant impact on the ability to reconstruct perfect tilings. Specifically, we compared reconstructing perfect tilings alone vs. amid a huge excess ( $>100\times$ ) of WGS reads. Even under such an extreme scenario, the N50 contig length remains extremely large (Fig. 1*Aii*, lines 1 and 2).

Myers *et al.* also remark that (i) a finished chromosome is not representative of typical HGP contigs, and (ii) Celera's perfect tilings were constructed from individual HGP bacterial artificial chromosomes (BACs) rather than merged BAC sequences. Although true, these points do not affect the analysis. Computational analysis shows that small contigs are more readily reconstructed than large contigs: The ratio of reconstructed N50 length to input contig length increases as contig size decreases, approaching 1 for contig sizes  $\leq 40$  kb. Similarly, computational analysis shows that the N50 length of reconstructed perfect tilings is comparable whether one begins with individual BACs or merged BAC sequences.

In short, we confirm our previous conclusion (3) that perfect tilings implicitly contain most assembly information and that this information is readily extracted by typical assembly programs (6).

Why then do Myers *et al.* report that they can only partially reconstruct perfect tilings? The answer seems to lie in the details of the Celera assembler. Most current assembly programs distinguish between true overlaps and close but spurious matches with related sequences by exploiting sequence-quality scores (reflecting error rates for each base); even a modest sequence difference at high-quality bases is sufficient to reject spurious matches. Because the Celera assembler does not use quality scores, it employs a much more conservative, and thus less powerful, approach for detecting true overlaps; an apparent overlap is accepted only if there is no other sequence in the genome with  $\geq 94\%$  sequence identity.\*\* Accordingly,

fewer true overlaps are accepted in the initial assembly stage.

As a result, the Celera assembler cannot fully exploit the assembly information inherent in the perfect tilings and instead yields only a partial reconstruction of the HGP assemblies (N50  $\approx 7$  kb).

### Gap Filling and Compartmentalized Assembly Extend the Reconstruction

In focusing only on the initial step of overlap detection, Myers *et al.* do not consider the subsequent uses of the assembled HGP data that further extend the N50 length. In later stages of the Celera assembly, contigs are linked together by using mate-pair information, and the resulting gaps are then filled by various methods that may use sequence not included in the initial stages. One of these methods is gap filling, which fills gaps between linked sequence contigs by directly using assembled HGP BACs. This device and others largely eliminate any gaps arising from incomplete overlap detection in the HGP-derived perfect tilings in the first stage of assembly.

To exploit gap filling, only a small amount of mate-pair information is needed to establish linkage between contigs. We found that the equivalent of 10% of the Celera WGS data ( $<0.5\times$  coverage) readily yielded enough linking information to allow gap filling to fill most gaps in the assembled HGP sequence (Fig. 1*Aii*, line 4). The approach ensures that the continuity of Celera's joint assemblies should not be worse than that of the HGP assemblies.

Similarly, the compartmentalized assembly, the main assembly method used in the article, dramatically decreases the problem of overlap detection due to spurious matches by restricting the analysis to tiny local regions (rather than the whole genome). For example, Myers *et al.* report that Celera's assembler can readily reconstruct perfect tilings to N50 lengths of 256 kb when focusing even on a large region such as chromosome 22 (see table 1 of ref. 3).

In short, Celera's approaches implicitly reconstruct the HGP sequence to a much greater extent than suggested by Myers *et al.*

### Direct Evidence from N50 Length

The most direct evidence of the importance of the HGP data for the Celera work comes from considering the properties of the Celera assembly itself (Table 1).

**Table 1. N50 contig lengths**

Celera human assembly	$\approx 86$ kb
HGP human assembly	$\approx 82$ kb
Expectation for $5\times$ WGS	$\approx 11$ kb
Celera mouse assembly ( $5.3\times$ WGS)	14.6 kb

Celera's joint assembly has an N50 contig length<sup>††</sup> of  $\approx 86$  kb, which is nearly identical to that of the HGP assembly at  $\approx 82$  kb. The similarity between the assemblies is no coincidence.

In fact, it is mathematically impossible to obtain such a large N50 from Celera's own  $5.1\times$  WGS coverage. Indeed, our computer simulations (3) show that  $5.0\times$  WGS should yield N50 of only  $\approx 11$  kb. Consistent with this estimate, Celera recently reported (7) that its draft sequence of the mouse genome based on  $5.3\times$  WGS coverage had N50 of  $\approx 14$  kb.

The large N50 could not even occur from  $\approx 7.1\times$  WGS, as Celera implies it used in the joint assemblies by combining  $5.1\times$  of its WGS data with  $2\times$  sampling of the HGP assemblies. Simulations indicate an expected N50 in the range of 30 kb. Consistent with this, the recent  $\approx 7\times$  WGS assembly by the public Mouse Genome Sequencing Consortium (8) obtained an N50 of  $\approx 24$  kb.

The large N50 ( $\approx 82$  kb) of the HGP assemblies was due partly to the HGP's deeper underlying coverage and the use of localized assembly with programs exploiting quality scores and crucially to the directed reads used in finishing for the clone-based approach. This continuity was carried over to Celera's assembly through various devices such as perfect tiling and gap filling.

Thus, contrary to the suggestion of Myers *et al.*, the sequence continuity of Celera's assemblies was vastly extended by use of the HGP genome assemblies.

### Assembly Comparison

Myers *et al.* cite various statistics to suggest that the Celera assembly is radically different from the HGP assembly. In fact, the difference is modest even by their own analysis. Total coverage differs by 2% of the human genome, with Celera adding  $\approx 7\%$  to the HGP and omitting  $\approx 5\%$  it was unable to assemble. Sequence continuity was nearly identical (86 vs. 82 kb), and the long-range connectivity was similar (3.6 vs. 2.3 megabases).<sup>‡‡</sup>

<sup>††</sup>The N50 contig length can be calculated from figure 2 of ref. 9 or directly from the assembly.

<sup>‡‡</sup>Myers *et al.* suggest that the HGP has lower long-range connectivity by focusing on connectivity by paired-end links rather than BAC contigs (see table 6 of ref. 1).

\*\*Myers *et al.* suggest that Celera's assembler used weak criteria for overlap detection to cope with low-quality HGP data. In fact, the same criteria were used for Celera's own work on fly and mouse. The error rate for HGP raw reads was  $<0.1\%$  at most bases and for assembled contigs was  $<0.1\%$  at  $>95\%$  of bases.

## Conclusion

Our goal is to respond to the suggestion of Myers *et al.* (5) that Celera's assembly was largely independent of the HGP genome assembly that was used as input. The data (with a majority of the underlying sequence information and all mapping information coming from the HGP), the methodology (with approaches that preserve assembly information), and the properties of the as-

sembly (extensive sequence continuity, as shown by N50) all indicate that the assembly reported in the Celera article is instead appropriately viewed as a refinement built on the HGP assemblies. This is not meant to suggest that the Celera article did not add some sequence, as well as additional order and orientation, beyond the HGP sequence that was used as input to their assembly process.

Of course, the ultimate goal is a finished sequence of the human genome to

serve as a lasting foundation for medicine. Perhaps the most important difference between the public and private genome efforts is that only the HGP has chosen to take on the task of converting the draft genome sequence to a finished genome sequence. This goal seems well within reach. As of this writing, the HGP has produced finished sequence covering  $\approx 98\%$  of the euchromatic human genome.

1. International Human Genome Sequencing Consortium (2001) *Nature* **409**, 860–921.
2. 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.
3. Waterston, R. H., Lander, E. S. & Sulston, J. E. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 3712–3716.
4. Green, P. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 4143–4144.
5. Myers, E. W., Sutton, G. G., Smith, H. O., Adams, M. D. & Venter, J. C. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 4145–4146.
6. Batzoglu, S., Jaffe, D. B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J. & Lander, E. S. (2001) *Genome Res.* **12**, 177–189.
7. Mural, R. J., Adams, M. D., Myers, E. W., Smith, H. O., Gabor Miklos, G. L., Wides, R., Halpern, A., Li, P. W., Sutton, G. G., Nadeau, J., *et al.* (2002) *Science* **296**, 1661–1671.
8. Mouse Genome Sequencing Consortium (2002) *Nature* **420**, 520–562.
9. Aach, J., Bulyk, M. L., Church, G. M., Comander, J., Derti, A. & Shendure, J. (2001) *Nature* **409**, 856–859.

# The independence of our genome assemblies

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

<sup>\*</sup>Celera Genomics, 45 West Gude Drive, Rockville, MD 20850; <sup>‡</sup>Institute for Biological Energy Alternatives, 1901 Research Boulevard, Suite 600, Rockville, MD 20850; <sup>§</sup>Department of Electrical Engineering and Computer Sciences, 231 Cory Hall, University of California, Berkeley, CA 94720; and <sup>¶</sup>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  $\approx 0.0003\%$  of the size of the human genome.<sup>¶</sup> 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  $\approx 35,000$  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  $\approx 5$ -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 $\times$  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 $\times$  sequence coverage in mate pairs from a combination of large and small insert libraries. Second, the 5 $\times$  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 assem-

bly (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

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

<sup>¶</sup>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 success-

ful 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](http://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.