

# More on the sequencing of the human genome

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

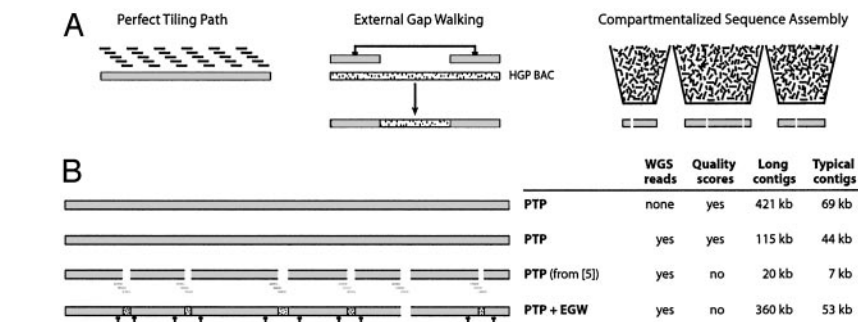
\*Department of Genome Sciences, University of Washington, Seattle, WA 98195; <sup>‡</sup>Whitehead Institute/Massachusetts Institute of Technology Center for Genome Research, Cambridge, MA 02142; and <sup>§</sup>Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA,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