Commentary

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 approach 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×), 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.

COMMENTARY

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.