Comment

After the (draft) sequence

Sydney Brenner

Address: The Salk Institute, North Torrey Pines Road, La Jolla, CA 92037-1099, USA. E-mail: sbrenner@molsci.org

Published: 17 April 2001

Genome ***Biology*** 2001, **2(5)**:comment1006.1–1006.2

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/5/comment/1006

We now have the two reports on the human genome sequence, one, in *Nature* (2001, **409:**860-921), by the International Human Genome Sequencing Consortium (GC), the other, in *Science* (2001, **291:**1304-1351), by Celera Genomics (CG). Each is accompanied by a flurry of secondary papers analyzing different aspects of the sequence, and these will no doubt be followed in the future by more analyses as, at least, the public sequence is available with no restrictions whatsoever to those who wish to examine it. The amount of information is enormous and all we have here is the surface. I have spent some time on both papers but the material will require much deeper reading in months to come.

I will begin by saying something about the two approaches used. GC based their approach on sequencing ordered, large-insert bacterial artificial chromosome (BAC) libraries which had previously been shown to produce data with 99.99% accuracy and no gaps. It makes sense to talk about such levels of accuracy when dealing with single cloned stretches of DNA. A collection of such BACs would not provide a total human genome sequence with that accuracy, however, but instead a singular mosaic, not representing any existing human genome. With a level of polymorphism of 0.1%, it is clearly not possible to talk about this level of accuracy for the genome sequence. The insistence on high accuracy may have carried with it certain costs, and it is clear that the switch to a draft sequence because of CG's entry into the field speeded up the work considerably. In retrospect it might have been more reasonable to have aimed at that from the start. There were some heretics who thought that this would be an important first step rather than wasting resources on the precise sequencing of all those *Alu* repeats. By October 2000, when the GC data were assembled, 900 megabases had been sequenced with 20-25 X coverage (each base sequenced an average of 20-25 times) and were considered finished, 3000 megabases were in draft form (12 X coverage) and a minority, 270 megabases, were in pre-draft form (6 X coverage). All of these data were available to CG.

CG's assembly was based on a 'mate-pair' strategy. This is not completely random, as many believe, but provides results in which one half of the data is spatially correlated with the other half, since two sequences are collected from the two ends of clones with inserts of several sizes (2, 10 and 50 kilobases). CG used two approaches in their assembly. In the first, the publicly available sequence from more than 30,000 BACs was shredded, pooled with CG's own data and assembled. In the second, the known BAC clustering was preserved and these clusters were added to those generated from CG's data together with an additional set of more than 104,078 BAC end sequences. The two methods, it is stated, gave essentially the same results. CG tested their computational assembler by shredding the sequences of chromosomes 21 and 22, both of which had been sequenced to high accuracy, and found that they assembled to give much the same structure but there were gaps. About half of the gaps contained sequences with a large fraction of repetitive elements, which could account for the failure of the assembly. The question of whether the CG assembly would have worked without the public data is already being hotly debated. It is a great pity that CG did not attempt an assembly of their own data first, to see how far they could do it without recourse to the public data. In the interests of scientific objectivity this would have been the wisest procedure even if it failed to provide results up to expectation. It is also clear that mapping data were also used to order the segments on the larger scale.

According to the public press, the most significant result of both groups is the small number of genes (better called gene loci). About 26,000 were found by CG while the estimate of GC ranges from 30,000 to 40,000. This is very far from the 100,000 many people expected and far from the 120,000 predicted from cDNA sequencing. The latter number can be explained by alternatively spliced products, but I think that the number of gene loci will turn out to be at the higher end of the estimates, possibly as high as 50,000. The low numbers could be accounted for by the inadequacy of the

gene-finding programs in regions which are gene-poor, and where a few kilobases of exons might be scattered through hundreds of kilobases of intron sequences. Time will tell. In any event, when we have the loci defined we will still need to characterize the gene products for each one, and to analyze the repertoire in the many different cell types. This will be a task even greater than that of sequencing the genome.

I found Table 19 in the CG paper very interesting, as it compared the numbers of genes of different types in the various genomes sequenced. In particular, there are a large number of genes encoding regulatory proteins, including 607 zinc-finger-containing proteins, as opposed to 232 in flies. Of course, we do not know how complexity scales with this increase, and clearly understanding this will be central to formulating theories of the complexity of organisms. The GC report has used the global information arising from having most of the sequence to analyze the evolution of transposable elements, mutation rates, CpG islands and many other sequence features. Both reports discuss single-nucleotide polymorphism (SNP) data, and the availability of large numbers of these should enable a large number of association studies in man, as well as the measurement of linkage disequilibrium over the entire genome.

It is clear that a massive amount of work must still be done on the sequence itself. It has to be finished, to make sure that all the genes have been found and that all the pieces are correctly ordered. I hope that there will be no relaxation from this task. I also hope that these hastily assembled papers will not be taken as anything but a first look. When all of that is done, we will have really finished sequencing the human genome, and we can look forward to the even more daunting task of understanding the regulation of genes, and the host of novel problems that will be uncovered by the sequencing of other mammals and other vertebrates.