Opinion
Twin peaks: the draft human genome sequence
Colin AM Semple, Kathryn L Evans and David J Porteous

Address: Medical Genetics Section, Department of Medical Sciences, The University of Edinburgh, Molecular Medicine Centre, Western General Hospital, Edinburgh EH4 2XU, UK.

Correspondence: Colin AM Semple. E-mail: Colin.Semple@ed.ac.uk

Abstract

Once thought to be impossible or a waste of resources, the initial high-volume stages of sequencing the human genome have been completed.

Extra, extra – read all about it

The simultaneous publication in *Nature* [1] and *Science* [2] magazines of the draft human sequence, generated, respectively, by the long-established and publicly funded International Human Genome Sequencing Consortium (IHGSC) [1] and the commercially backed new kids on the block, Celera Genomics [2,3], is a landmark worth celebrating, savoring and slowly digesting. The decoding of more than 90% of the human genome is the culmination of a massive scientific effort. The human genome is approximately 30 times the size of the recently sequenced genomes of the nematode worm and fruitfly, and 250 times that of yeast, and has reached this advanced stage in a fraction of the time taken to sequence these smaller genomes. As a publishing event alone, it is noteworthy on several counts. The usual editorial rationing, which in the past informed us in less than 1,000 words of, for example, the discovery of the structure of DNA [4], has been waived to give the authors full rein to describe how the sequence of human genomic DNA was obtained [1,2]. The sheer volume of background articles, instant commentaries, preliminary analyses, projections and follow-on studies leaves hardly a stone unturned, but this is just the beginning.

Ignoring for the moment the brouhaha over who exactly did what, how, when and where in this 'race' to sequence the human genome, it is perhaps worth reflecting on two key points. First, the human genome sequence is of singular scientific, medical, industrial and societal importance; its broad impact is already tangible and will surely pervade as-yet unimagined areas of human ingenuity. Second, the sequence of the human genome is at present patently incomplete. Key questions will remain unanswerable until it is freely available in full. The target date for a 'finished' genome sequence is still 2003, and the labour-intensive task of filling the remaining gaps in the draft have begun. Only the public effort is undertaking this 'finishing' process, and there is some concern that the large sequencing centers will lose focus on this relatively unrewarding work and will instead migrate to the greener pastures of functional genomics and sequencing other organisms [5].

The political skirmishes between the publicly funded and Celera projects have been presented in the popular press as an explosive feud. For example, in the UK we read about the fight between "brilliant US scientist and rapacious capitalist" Craig Venter and "second-hand car driving, bearded" John Sulston (*The Guardian*, 28 June 2000). What is beyond dispute is the importance of the intervention of the Wellcome Trust [6], leading to the 'Bermuda agreement' [7], which established the principle of immediate and complete release of the publicly funded human genome sequence. At the same time, the catalytic effect of the commercially driven and complementary approach taken by Celera led to a rationalization and ramping-up of the publicly funded effort. Politicians have courted representatives from both the public and private projects. As Director of the Wellcome-funded Sanger Centre, Sulston was architect of the Bermuda agreement and led Wellcome's contribution to the IHGSC, for which he received a well-deserved knighthood in the

Queen's New Year's Honours list. This honour is normally reserved for captains of industry, loyal civil servants and sporting heroes - not categories into which Sir John easily fits. Meanwhile, Venter and Francis Collins, Director of the US National Human Genome Research Institute, were invited to the White House for congratulatory handshakes.

Public versus private: a tale of two methods

The two draft sequences have been produced using different methods. The IHGSC started from a clone-based physical map of the genome [8], while Celera used the whole-genome shotgun method to produce the sequence. The IHGSC shotgun-sequenced each bacterial artificial chromosome (BAC) clone to an average fourfold depth (that is, each clone was sequenced an average of four times), assembled the sequence fragments on the basis of overlap, and then merged these assemblies into larger regions using information on how individual BACs overlapped. This method was chosen over the whole-genome shotgun approach on the basis it would minimize the problems associated with assembling a repeat-dense genome (now known to be more than 50% repetitive sequence) and that it would speed the process of finishing the remaining gaps. The use of haploid BACs, rather than a diploid genome (or genomes), would make assembly of individual clone sequences less prone to the uncertainties caused by polymorphism. In addition, this strategy allowed division of the work amongst the many collaborating Genome Centres and easy distribution of the emerging data in a form that would be useful long before the work was complete.

Celera carried out whole-genome shotgun sequencing on DNA from several individuals, and assembled the data with that produced by the IHGSC using two different assembly protocols. The first combined the Celera data with the entire IHGSC data set, while the second involved initially clustering the data to a chromosome, or chromosomal region, on the basis of mapping information, before combining the two data sets; the second method was reported to provide slightly greater sequence coverage. Prior to the assembly by Celera, the IHGSC data were computationally shredded into 500 base-pair (bp) fragments, to avoid problems caused by misassembly of BAC sequences, chimeric BACs and contamination of the sequence database with DNA from other organisms. It is important to note that the public data used by Celera were downloaded from GenBank [9] on 1 September 2000, so their assembly may well be improved by incorporating the sequence produced over the intervening months. Investigators working on a particular genomic region would be advised to download the relevant Celera sequence and the most recent IHGSC data sets and assemble them to provide the most complete coverage of their genomic region of interest. The incorporation of public sequence and mapping data by Celera means that it is difficult to determine which of the strategies has worked 'best',

but the Sanger Centre has published a comparison of the public and Celera's strategies on the web [10].

The average biologist is probably interested in practical aspects of the two draft genomes, such as how the assembled data produced by the two groups compare. Some effort has been made to answer this question by Aach *et al.* [11], who compared the Celera draft genome assembly [2] with an assembly ("HGP-nr") produced from the clone sequences generated by the IHGSC. Unfortunately no citation or construction protocol is given for the HGP-nr assembly and it appears not to be publicly available, although it was produced at the US National Center for Biotechnology Information (NCBI) [12]. In Table 1, we have combined the comparison by Aach *et al.* [11] of the Celera and HGP-nr assemblies with the publicly available draft genome assembly (7 October 2000 version) produced by David Haussler and Jim Kent of the University of California, Santa Cruz (UCSC) [13]. From Table 1 it would appear that HGP-nr is similar to the UCSC assembly. Both HGP-nr and UCSC consist of a few thousand contigs containing more than 100,000 gaps, whereas the Celera assembly consists of 21,000 relatively small contigs that contain only a fraction of the number of gaps in the other two assemblies. Aach *et al.* [11] report that the longest gap in the Celera assembly is 168 kilobases (kb). The UCSC assembly [13] is estimated to contain gaps of around 35 kb within contigs, and gaps of around 170 kb between contigs [1]. What differences there are between the assemblies are as likely to be explained by the different releases of HGP sequence data they incorporate as anything else. For those of us who hoped to use the Celera data to plug gaps in the public sequence, there is a glimmer of hope: Aach *et al.* [11] found that around 0.14% of Celera sequence was not present in the HGP-nr assembly. But on a note of caution, Aach *et al.* [11] also found evidence for possible Celera misassemblies (sequences assembled in the

Table 1

A comparison of three draft genome sequence assemblies

Assembly	Celera	HGP-nr	UCSC
Reference	[2]	[11]	[13]
Length (gigabases)	2.9	2.9	3.3
Sequenced bases (gigabases)	2.66	2.84	2.69
Number of gaps	21,684	181,079	145,514
Number of contigs	54,061	6,094	4,884
N50 length* (megabases)	0.8	2.8	2.3
HGP data version used	1/09/00	?/12/00	07/10/00

*N50 length is a measure of the contig length containing the 'typical' nucleotide. Specifically, it is the maximal length L such that 50% of all nucleotides lie in contigs of size at least L. See text for further details of the three assemblies.

wrong order or orientation), of the kind rigorously quanti-fied by the UCSC effort [13].

A reasonable expectation, given Celera's use of the public mapping and sequence data, might have been that Celera's assembly would be superior to the public ones. Taking the data in Table 1 with those in Aach *et al.* [11], it would appear that there is no clear 'winner'. It is somewhat clearer who the losers are, however. As an open letter from leading bioinfor-maticists points out [14], Celera's access restrictions have the effect of stopping any large-scale analysis of the Celera sequence data by public endeavours such as Ensembl [15]. Such restrictions on the public distribution and use of scien-tific data may set dangerous precedents [16].

The incredible shrinking genome
The headline-grabbing discovery of the draft sequence publi-cations has been the predicted number of genes: Celera found 26,000-38,000 [2] and IHGSC found 30,000-40,000 [1]. Many biologists had become accustomed to a ball-park figure of 100,000 human genes; estimates have varied but, as the IHGSC paper points out [1], studies using expressed sequence tags (ESTs), comparative genomics and the com-pleted human chromosomes have all suggested figures of 30,000-35,000. Much has been said about the relationship between the predicted number of human genes and the numbers in other species' genomes. Apparently "the low figure" of predicted human genes is to cause "serious prob-lems for scientists trying to explain the complexity of the human species" (*The Observer* [UK], 11 February 2001). Reportedly, Venter was even moved to comment that "we simply do not have enough genes for this idea of biological determinism to be right" (*The Observer*, 11 February 2001). This line of argument rather ignores the diversity and inter-actions of downstream products of the genes - the lesser-studied worlds of RNA and proteins. Even in the simplest case imaginable, where each gene is simply turned on or off, a genome with 30,000 genes can encode $2^{30,000}$ states [17], but we know that things are not so simple. The IHGSC detected alternative transcript splicing in 59% of genes on chromosome 22 [1] and a previous study based on EST data estimated that 38% of all human genes are alternatively spliced [18]. This leaves a lot of room for the generation of complexity, and as large-scale assays for protein interactions are developed [19] our view of the path from genotype to phenotype is unlikely to become simpler. It would already appear that human proteins have more 'complex' architec-tures (more often containing domains involved in different functions) than those in other sequenced eukaryotic genomes [1].

Early estimates of 100,000 human genes were based on an assumption of uniform gene density. In both draft sequences gene density is in fact found to be highly variable, both within and between chromosomes. Around 20% of the

Celera assembly is estimated to be 'desert' (regions larger than 500 kb devoid of genes) and the total length of desert per chromosome varies twofold. The density of single nucleotide polymorphisms (SNPs) was also found to vary significantly across both draft genomes, so the even coverage desirable for whole-genome studies of variation remains on the wish list. The public effort has discovered 1.42 million SNPs [20], which are publicly available in the NCBI's dbSNP database [21], and it is hoped that these SNPs will lead to the identification of many disease-associated genes. As a mapping resource they are already valuable, but the real promise of these variations lies in the detection of function-ally important SNPs or haplotypes of SNPs. With only around two exonic SNPs per gene publicly available [20], however, much of the functional SNP discovery will probably be left to those interested in specific genes.

Large SNP collections should also enable more detailed genetic studies of ancient human history. Patterns of sequence conservation across the genome have already shed light on the more distant evolutionary past: the public effort estimates that the genome contains 183 conserved segments with an average length of 15.4 Mb [1]. Using less stringent thresholds, Venter *et al.* [2] have found evidence for 1,077 duplicated blocks of sequence, including ancient duplica-tions approximating chromosome-length that are likely to date from the point of the emergence of vertebrates. Further analyses of these data may reveal whether an ancient whole-genome duplication is the explanation for these duplicated blocks [22].

And now for something completely different
Now that we have most of the genome sequence, the imme-diate task is to improve the description of its contents. Most genes identified in either draft genome are computational predictions supported by homology to known expressed sequences. This kind of automatically generated draft genome annotation is available in the Ensembl database [15], which contains 25,790 genes (release 0.8.0) including 94% of known genes. A similar interface for viewing anno-tated features on genomic sequence is the Human Genome Browser at UCSC [13]. Celera use similar combinations of software for sequence annotation [2] but have not made any of their annotation data publicly available. The wide variety of gene-prediction software available has one defining char-acteristic, however: fallibility. Typical claims for sensitivity (the percentage of genes detected) are 77-98%, but esti-mates of accuracy (the percentage of genes/exons predicted correctly) are much lower. In addition, only protein-coding exons are predicted by most gene prediction programs; untranslated regions (UTRs) are ignored, and there is little success at all in predicting promoters and regulatory sites [23]. Most non-coding RNA genes, such as the one encod-ing the large *Xist* transcript involved in X-chromosome dosage compensation [24], will be missed altogether. This

means that inadequacies in the annotation of the genome will be a recurring problem for some time to come. Laborious work screening cDNA libraries to verify or refute gene predictions made for the first completed human chromosome, chromosome 22 [25], continues at the Sanger Centre [26]. Large-scale automation using microarray technology offers the hope of accelerating this process [27]. Valuable short-cuts to gene-structure determination for most genes will undoubtedly come from the sequence emerging from the Mouse Genome Sequencing Initiative [28,29]. Comparative analysis using mouse and other vertebrate sequences is also expected to uncover many non-coding features, such as promoters.

Aside from the definition of gene structures, the annotation of gene function is the other major obstacle on the path to an informative genome sequence. In both draft genome sequences it is estimated that around 40-60% of genes cannot (yet) be assigned molecular function. In many cases, the functions assigned on the basis of homology are only broad descriptions based on the presence of a known domain. Co-expression of uncharacterized genes with well-studied genes may provide further clues to function [30], but the aim must be a description of the processes and/or complexes in which gene products participate. Much information may come from building protein interaction maps for the genome, using the yeast two-hybrid protein-protein interaction assay, as has been done for many yeast proteins [31]. Ultimately, the intention is to provide an accurate picture of the three-dimensional structure of each human protein, and moves are underway to develop technologies for large-scale structure determination [32]. On top of any large data sets produced automatically, it is already evident that a mountain of hard-won data from the bench will be required before we fully describe the physiology of our genes. This work should also usher in a new set of approaches in biology, aiming to combine large data sets and their computational analysis with results from the wet lab.

References

1. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al.*: **The sequence of the human genome.** *Science* 2001, **291:**1304-1351.
3. **Celera Genomics** [http://www.celera.com/]
4. Watson JD, Crick FHC: **A structure for deoxyribose nucleic acid.** *Nature* 1953, **171:**737-738.
5. Pennisi E: **What's next for the genome centers?** *Science* 2001, **291:**1204-1207.
6. **The Wellcome Trust** [http://www.wellcome.ac.uk/]
7. **Summary of principles agreed at the international strategy meeting on human genome sequencing** [http://www.hugo-international.org/hugo/bermuda.htm]
8. International Human Genome Mapping Consortium: **A physical map of the human genome.** *Nature* 2001, **409:**934-941.
9. **GenBank** [http://www.ncbi.nlm.nih.gov/Genbank/]
10. **Comparison of draft human sequence versions from the public and private domain** [http://www.sanger.ac.uk/HGP/publication2001/comparison.shtml]
11. Aach J, Bulyk ML, Church GM, Comander J, Derti A, Shendure J: **Computational comparison of two draft sequences of the human genome.** *Nature* 2001, **409:**856-859.
12. Marshall E: **Comparison shopping.** *Science* 2001, **291:**1180-1181.
13. **Human Genome Project Working Draft** [http://genome.ucsc.edu/]
14. **Open letter to the bioinformatics community** [http://www.genetics.wustl.edu/eddy/people/eddy/openletter.html]
15. **Ensembl** [http://www.ensembl.org/]
16. Roos DS: **Bioinformatics – trying to swim in a sea of data.** *Science* 2001, **291:**1260-1261.
17. Claverie JM: **What if there are only 30,000 human genes?** *Science* 2001, **291:**1255-1257.
18. Brett D, Hanke J, Lehmann G, Haase S, Delbruck S, Krueger S, Reich J, Borka P: **EST comparison indicates 38% of human mRNAs contain possible alternative splice forms.** *FEBS Lett* 2000, **474:**83-86.
19. Legrain P, Selig L: **Genome-wide protein interaction maps using two-hybrid systems.** *FEBS Lett* 2000, **480:**32-36.
20. International SNP Map Working Group: **A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.** *Nature* 2001, **409:**928-933.
21. **A Database of Single Nucleotide Polymorphisms** [http://www.ncbi.nlm.nih.gov/SNP/]
22. Skrabanek L, Wolfe KH: **Eukaryote genome duplication – where's the evidence?** *Curr Opin Genet Dev* 1998, **8:**694-700.
23. Semple C: **Gene prediction: the end of the beginning.** *Genome Biology* 2000, **1:** reports4012.1-4012.3.
24. Brown CJ, Hendrich BD, Rupert JL, Lafreniere RG, Xing Y, Lawrence J, Willard HF: **The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus.** *Cell* 1992, **71:**527-542.
25. Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, Collins JE, Bruskiewich R, Beare DM, Clamp M, Smink LJ, *et al.*: **The DNA sequence of human chromosome 22.** *Nature* 1999, **402:**489-495.
26. **The Sanger Centre Chromosome 22 Site** [http://www.sanger.ac.uk/HGP/Chr22/]
27. Shoemaker DD, Schadt EE, Armour CD, He YD, Garrett-Engele P, McDonagh PD, Loerch PM, Leonardson A, Lum PY, Cavet G, *et al.*: **Experimental annotation of the human genome using microarray technology.** *Nature* 2001, **409:**922-927.
28. Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, Arakawa T, Hara A, Fukunishi Y, Konno H, *et al.*: **Functional annotation of a full-length mouse cDNA collection.** *Nature* 2001, **409:**685-690.
29. **Mouse Genome Sequencing Initiative** [http://www.informatics.jax.org/mgihome/MGS/mgp.shtml]
30. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95:**14863-14868.
31. Newman JR, Wolf E, Kim PS: **A computationally directed screen identifying interacting coiled coils from *Saccharomyces cerevisiae*.** *Proc Natl Acad Sci USA* 2000, **97:**13203-13208.
32. Service RF: **Structural genomics offers high-speed look at proteins.** *Science* 2000, **287:**1954-1956.