

Mining the draft human genome

Ewan Birney*, Alex Bateman†, Michele E. Clamp† & Tim J. Hubbard†

* The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

† The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

Now that the draft human genome sequence is available, everyone wants to be able to use it. However, we have perhaps become complacent about our ability to turn new genomes into lists of genes. The higher volume of data associated with a larger genome is accompanied by a much greater increase in complexity. We need to appreciate both the scale of the challenge of vertebrate genome analysis and the limitations of current gene prediction methods and understanding.

In this issue, accompanying the description of the sequence¹, there are nine data-mining papers that interrogate the genome from distinct biological perspectives. These range from broad topics—cancer², addiction³, gene expression⁴, immunology⁵ and evolutionary genomics⁶—to the more focused: membrane trafficking⁷, cytoskeleton⁸, cell cycle⁹ and circadian clock¹⁰. The findings reported by these authors are likely to be indicative of many people's experiences with the draft human genome: frustrating and rewarding in equal measures.

The current data set

The human genome—the first vertebrate genome sequence to be determined—seems likely to be quite representative of what we will find in other vertebrate genomes. It is around 30 times larger than the recently sequenced worm and fly genomes, and 250 times larger than that of yeast, the first eukaryotic genome to be sequenced¹¹. Despite its size, it seems likely to have only two or three times as many genes as the fly and worm genomes, with the coding regions of genes accounting for only 3% of the DNA. Repeat sequences form a large proportion of the remaining DNA, around 46%. These repeats may or may not have a function, but they are certainly characteristic of large vertebrate genomes. The rest of the sequence contains promoters, transcriptional regulatory sequences and other features, as yet unknown.

The International Human Genome Sequencing Consortium has been sequencing the genome in fragments of about 100–200 kilobases (kb). These fragments exist as bacterial artificial chromosome (BAC) clones, which are derived from sequences whose chromosomal location is known. Each newly generated sequence is deposited in the high-throughput genome sequence (HTGS) division of the International Nucleotide Database (GenBank/EMBL/DDBJ) within 24 hours of being assembled and is assigned a unique identifier (its accession number). For the working draft, about 75% of clones are 'unfinished': each still consists of about 10–20 unassembled sequence fragments. Sequencing centres are continuously reading new sequence data from these clones until all the gaps are eliminated, at which point the sequence is declared 'finished'. As the HTGS entries are updated they retain the same accession numbers, but their version numbers increase.

There is a great deal of overlap between BAC clones, so it is typically more convenient to view a cleaned up version of the raw data, in which the sequences of the clones are correctly ordered and overlapped to remove redundancy and create a contiguous DNA sequence for each chromosome. These virtual chromosome sequences change continuously as gaps are closed and fragment ordering is refined.

Finding genes

With over 30 genomes sequenced, the casual observer could be forgiven for thinking that gene prediction, or annotation, was a problem filed neatly under 'solved'. Unfortunately this is far from

true. The large size of the genome makes finding the genes much more difficult. The protein-coding parts of human genes, called exons, are split into pieces in the genome and these pieces are separated by non-coding sequence called introns. Nearly all of the increase in gene size in human compared with fly or worm is due to the introns becoming much longer (about 50 kb versus 5 kb). The protein-encoding exons, on the other hand, are roughly the same size. This decrease in signal (exon) to noise (intron) ratio in the human genome leads to misprediction by computational gene-finding strategies.

Many methods for predicting genes are based on compositional signals that are found in the DNA sequence. These methods detect characteristics that are expected to be associated with genes, such as splice sites and coding regions, and then piece this information together to determine the complete or partial sequence of a gene. Unfortunately, these *ab initio* methods tend to produce false positives, leading to overestimates of gene numbers, which means that we cannot confidently use them for annotation. They also do not work well with unfinished sequence that has gaps and errors, which may give rise to frameshifts, when the reading frame of the gene is disrupted by the addition or removal of bases.

Thankfully, there is a wealth of data that we can use to produce more reliable gene predictions. Information on expressed sequences (expressed sequence tags (ESTs) and complementary DNAs) and proteins from humans and other organisms provide a more accurate resource for resolving gene structures against the vast genomic background. The most effective algorithms integrate gene-prediction methods with similarity comparisons. Such algorithms are integral to software programs such as GeneWise¹², Genomescan¹³ and Genie¹⁴, which provide accurate, automatic predictions, whereas BLAST or FASTA programs typically require considerable manual effort to determine the complete structure of a single gene.

The most powerful tool for finding genes may be other vertebrate genomes. Comparing conserved sequence regions between two closely related organisms will enable us to find genes and other important regions in both genomes with no previous knowledge of the gene content of either. The next couple of years should see the sequencing of the mouse, zebrafish and *Tetraodon* genomes. The preliminary sequence of *Tetraodon* has already proved useful in estimating gene numbers¹⁵, and shows much promise for the use of comparative genomics in gene prediction.

Resources available to the user

There are a number of resources currently available for perusing the human genome. 'Human Genome Central'¹⁶ attempts to gather together the most useful web sites (see <http://www.ensembl.org/genome/central/> or <http://www.ncbi.nlm.nih.gov/genome/central/>). The best starting point for the uninitiated will be a site such as those of NCBI, Ensembl or the University of Santa Cruz (UCSC). These sites offer a mixture of genomic viewers and web-searchable

datasets, and allow analysis of the human genome sequence without the need to run complex software locally.

For more involved analysis, it might be necessary to download some of the data locally. Useful downloadable sequence-oriented datasets include protein datasets (available from Ensembl and NCBI) and the assembled DNA sequence for regions of the genome, available at UCSC. Other genomic datasets are also available, such as the global physical map from The Genome Sequencing Center in St Louis and the single nucleotide polymorphism (SNP) database from NCBI. Raw sequence data is available from the International Nucleotide Database (GenBank/EMBL/DDBJ), but this data is generally more difficult to handle because it is very fragmentary, can contain contaminating non-human DNA and may include misleading information such as incorrect map assignment.

This loose network of sites will probably coalesce into a more coordinated network of sites offering informative web pages and resources. NCBI, Ensembl and UCSC are developing new, more accessible resources that will become available within the next year.

How to use the resources

There are two main ways to use the human genome sequence. First, we can look for a homologue of a protein that is known from another organism. For example, Clayton *et al.*² looked for relatives of the *Drosophila* period clock protein and found the three known relatives and a possible fourth cousin on chromosome 7. Or we can try and find all of the proteins belonging to a particular family—in ref. 4, Tupler *et al.* catalogue all homeobox domains⁴. The easiest way to approach these problems is to use a protein set. This sidesteps the frustration of predicting genes, but makes the researcher reliant on the quality of the predictions being provided. For most of the accompanying reports, a single protein set was the most useful resource provided. For example, Nestler *et al.* searched for G-protein receptor kinases³ using PSI-BLAST, which searches only protein datasets.

What are the potential pitfalls of the data? Human genes are hard to predict and are often fragmented. If each end of a query protein matches to a different predicted protein, we should suspect that the query sequence may in fact be two parts of a fragmented gene. The two matched human genes should be in the same or adjacent genomic locations. Pollard⁸ discovered that fragmentation complicated the analysis of myosin genes. In addition, the unfinished human genomic DNA may contain contamination, particularly from bacteria but also from other sources. Contaminating DNA is routinely removed from finished sequence, but some is still present in unfinished sequence. If the predicted gene matches a bacterial gene more closely than any vertebrate gene then it will almost always be a contaminant. Futreal *et al.*² were led up a blind alley for a week before they discovered that cDNA contamination in draft genomic

sequences was giving the false impression of multiple p53 proteins in the genome.

During the assembly of unfinished human genomic data it is possible to create artificial duplications, which can result in artefacts in the subsequent analysis. Very similar gene sequences found within the same clone may represent duplicate genes, but could also be the result of an assembly error. This also means that predicted protein sets may contain artificial duplications, leading to overestimation of the number of members in a family.

What does this analysis tell us? For Bock *et al.*⁷, the draft genome revealed a list of the molecular players involved in membrane trafficking, providing a platform for experiments that may complete our understanding of this area of biology. In contrast, Murray and Marks⁹ found no new cyclin-dependent kinases, indicating that they were all found by traditional experimental techniques. Futreal *et al.* had a similar experience for known cancer genes, but suggest that with new techniques the genome will provide new avenues of cancer research².

The interpretation of unfinished draft genomic data may seem like hard work. But it is something to become accustomed to, because we expect future vertebrate genomes to be released initially in draft form. The database providers must develop better ways of viewing the data; and researchers need to be educated in how to use them. That said, there are many undiscovered treasures in the current data set waiting to be found by intuition, hard work and experimental verification. Good luck, and happy hunting! □

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Futreal, A., Wooster, R., Kasprzyk, A., Birney, E. & Stratton, S. Cancer and genomics. *Nature* **409**, 850–852 (2001).
3. Nestler, E. J. & Landsman, E. Learning about addiction from the genome. *Nature* **409**, 834–835 (2001).
4. Tupler, R., Perini, G. & Green, M. R. Expressing the human genome. *Nature* **409**, 832–833 (2001).
5. Fahrner, A. M., Bazan, J. F., Papathevasiou, P., Nelms, K. A. & Goodnow, C. C. A genomic view of immunology. *Nature* **409**, 836–838 (2001).
6. Li, W.-H., Gu, Z., Wang, H. & Nekrutenko, A. Evolutionary analyses of the human genome. *Nature* **409**, 847–849 (2001).
7. Bock, J. B., Matern, H. T., Peden, A. A. & Scheller, R. H. A genomic perspective on membrane compartment organization. *Nature* **409**, 839–841 (2001).
8. Pollard, T. D. Genomics, the cytoskeleton and motility. *Nature* **409**, 842–843 (2001).
9. Murray, A. W. & Marks, D. Can sequencing shed light on cell cycling? *Nature* **409**, 844–846 (2001).
10. Clayton, J. D., Kyriacou, C. P. & Reppert, S. M. Keeping time with the human genome. *Nature* **409**, 829–831 (2001).
11. Goffeau, A. *et al.* The Yeast Genome Directory. *Nature* **387** (suppl.), 1–105 (1997).
12. Birney, E. & Durbin, R. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* **10**, 547–548 (2000).
13. Burge *et al.* *Nature Genet.* (submitted).
14. Reese, M. G., Kulp, D., Tammana, H., Haussler, D. Genie—gene finding in *Drosophila melanogaster*. *Genome Res.* **10**, 529–538 (2000).
15. Crollius, H. R. *et al.* Characterization and repeat analysis of the compact genome of the freshwater pufferfish *Tetraodon nigroviridis*. *Genome Res.* **10**, 939–949 (2000).
16. Genome website set up to help with sequence analysis. *Nature* **406**, 929 (2000).

Correspondence should be addressed to E.B. (e-mail: birney@ebi.ac.uk).