sequencing project carried to completion by the methods described in this issue. Genome sequencing will get easier from here.

Looking ahead, there are two threats to producing a quality finished product. One is simple exhaustion on the part of the consortium's members: each new round of press conferences announcing that the human genome has been sequenced saps the morale of those who must come to work each day actually to do what they read in the newspapers has already been done.

We may also expect to hear the argument that the current sequence is good enough for most purposes, and that remaining problems should be resolved by users as the need for accurate sequence in specific regions arises. What we have now is certainly a lot better than what we had yesterday. But biologists in the future will be comparing vast data sets to the reference sequence of the human genome. They must be able to do so with confidence that the discrepancies they encounter are due to the limitations of their own data or, more interestingly, to biology. They should not need to expend time, energy and imagination compensating for a failure now to pursue the Human Genome Project to a grand conclusion. We must move on and finish the job, even as the bright lights of media attention shift elsewhere. ■

Maynard V. Olson is in the Departments of Medicine and Genetics, Fluke Hall, Mason Road, University of Washington, Seattle, Washington 98195-2145, USA.
e-mail: mvo@u.washington.edu

1. National Research Council *Mapping and Sequencing the Human Genome* (National Academy Press, Washington DC, 1988).
2. The International Human Genome Mapping Consortium *Nature* **409,** 934–941 (2001).
3. Coulson, A., Sulston, J., Brenner, S. & Karn, J. *Proc. Natl Acad. Sci. USA* **83,** 7821–7825 (1986).
4. Olson, M. V. *et al. Proc. Natl Acad. Sci. USA* **83,** 7826–7830 (1986).
5. Yu, A. *et al. Nature* **409,** 951–953 (2001).
6. The BAC Resource Consortium *Nature* **409,** 953–958 (2001).
7. Montgomery, K. T. *et al. Nature* **409,** 945–946 (2001).
8. Brüls, T. *et al. Nature* **409,** 947–948 (2001).
9. Bentley, D. R. *et al. Nature* **409,** 942–943 (2001).
10. Venter, J. C. *et al. Science* **291,** 1304–1351 (2001).
11. Tilford, C. A. *et al. Nature* **409,** 943–945 (2001).

The draft sequences

Filling in the gaps

Peer Bork and Richard Copley

Two rough drafts of the human genome sequence are now published. Completion of the sequences lies ahead, but the implications for studying human diseases and for biotechnology are already profound.

With the publication of the human genome sequence — described and analysed on page 860 of this issue[1] and in this week's *Science*[2] — we cross a border on the route to a better understanding of our biological selves. But unlike the previously published sequences of human chromosomes 21 and 22 (refs 3,4), the present sequences of the whole human genome are not considered complete. The bulk of the data make up what is called a 'rough draft'. So what is all the fuss about? What exactly does 'rough draft' mean, and what can we learn from sequences such as this?

In the draft from the publicly funded International Human Genome Sequencing Consortium[1], around 90% of the gene-rich — euchromatic — portion of the genome has been sequenced and 'assembled', the term used to describe the process of using a computer to join up bits of sequence into a larger whole. Each base pair of this 90% was sequenced four times on average, ensuring reasonable precision. Only about a quarter of the whole genome is considered 'finished' — another bit of genomics jargon, which basically means that each base pair has been sequenced eight to ten times on average, with gaps in the sequence existing only because of the limitations of present technology. Nonetheless, the sequence of base pairs in the draft is very accurate, and is unlikely to change much; 91% of the euchromatin sequenced has an error rate of less than one base in 10,000 (ref. 1).

For the other draft, that produced by Celera Genomics[2], a variety of methods suggest that between 88% and 93% of the euchromatin has been sequenced and assembled. But direct comparison of these numbers with the public consortium's draft is almost impossible — different procedures and measures were used to process the data and to estimate accuracy. Both projects also have sequence data that were not used in the assembly process, raising the real level of coverage by a few percentage points.

These numbers might seem rather arbitrary, but even when the first genome of an animal species was published[5], it was clear that simple, practical finish lines do not exist (Box 1, Fig. 1). The present level of coverage of the human genome reflects the point where a shift of focus occurs, from sequencing the genome many times over to producing a high-quality, continuous sequence[6]. There is some way to go yet.

Essentially, 'rough draft' refers to the fact that the sequences are not continuous — there are gaps (Box 1). If there are too many gaps, it can be impossible to order and orientate the many small strings of bases that are the raw products of genome sequencing. This might, for example, hamper projects that seek to identify genes involved in inherited diseases. A first step to finding such genes is to work out which region of which chromosome they are on. The complete genome sequence should be immensely useful for the next step — identifying the relevant gene at that region. But gaps and errors in ordering and placing the strings of sequence will make this difficult.

Another problem of incompleteness is that it is difficult to make definitive

Box 1 What makes a completely sequenced genome?

When is sequencing work on a genome complete? No genome for a eukaryotic organism — roughly, those organisms whose cells contain a nucleus — has been sequenced to 100%. There are regions, often highly repetitive, that are difficult or impossible to clone (one of the initial steps in a sequencing project) or sequence with current technology. Fortunately, such regions are expected to contain relatively few protein-coding genes[4,10].

The extent of these regions varies widely in different species. So, rather than applying a universal gold standard, each sequencing project has made pragmatic decisions as to what constitutes a sufficient level of coverage for a particular genome. For example, as much as one-third of the sequence of the fruitfly *Drosophila melanogaster* was not stable in the cloning systems used, and so was not sequenced. But 97% of the so-called euchromatic portion — where most genes are thought to reside — was sequenced[11] (Fig. 1).

For the human genome, one definition of 'finished' is that fewer than one base in 10,000 is incorrectly assigned[6]; more than 95% of the euchromatic regions are sequenced; and each gap is smaller than 150 kilobases[12]. Such standards represent realistic goals given current technology. By this standard, over a quarter of the public consortium's sequence[1] is considered finished at present, including the previously published long arms of chromosomes 21 and 22 (refs 3,4; Fig. 1). The Celera sequences of chromosomes 21 and 22 are slightly more gappy than those from the public consortium, but the converse seems to be true for the other chromosomes[2]. But again, as different protocols were used, it is not easy to compare the overall status of the two assemblies. In the longer term, as much of the heterochromatin — which is harder to sequence, and contains few genes — as possible must be sequenced, because we might otherwise miss important features.

P.B. & R.C.

Organism	Year	Millions of bases sequenced	Total coverage (%)	Coverage of euchromatin (%)	Predicted number of genes	Number of genes per million bases sequenced
Saccharomyces cerevisiae	1996	12	93	100	5,800	483
Caenorhabditis elegans	1998	97	99	100	19,099	197
Drosophila melanogaster	2000	116	64	97	13,601	117
Arabidopsis thaliana	2000	115	92	100	25,498	221
Human chromosome 21	2000	34	75	100	225	7
Human chromosome 22	1999	34	70	97	545	16
Human genome rough draft (public sequence)	2001	2,693	84	90	31,780	12
Human genome rough draft (Celera sequence)	2001	2,654	83	88–93	39,114	15

Figure 1 **Sequenced eukaryotic genomes. Total coverage uses an estimate of the total genome size and includes heterochromatin (condensed genomic areas that were originally characterized by staining techniques, and are thought to be highly repetitive and gene-poor). The gene-rich areas make up euchromatin. Gene numbers are taken from the original sequence publications[1–6,14,15]; most numbers have since changed slightly and different sources give different estimates depending on protocols. The data for the public consortium's rough draft of the human genome are taken from ref. 1, Table 8, page 872. The estimate of total coverage for the Celera data is based on the public consortium's estimate of the full genome size (3,200 million base pairs); the percentage of euchromatin covered is taken from ref. 2. The predicted numbers of human genes are discussed further in the text.**

statements about which genes are unique to other species and do not have relatives in the human genome. So it might be prudent not to place too much emphasis on such 'missing' genes at this stage. Even so, they are running out of places to hide, particularly because the level of coverage of the human genome is probably higher than reported here[1,2] — there are other chunks of unassembled genome sequence in public databases, such as in independent collections of so-called expressed sequence tags.

But ensuring high quality and high coverage are only two aspects of producing a finished genome. For most biologists, the real interest is in the genes themselves. Here, the picture is less rosy, although the problems are caused not so much by the draft nature of the sequence as by the difficulty in finding genes among the other genomic DNA (Box 2).

Even coming up with a rough count of the number of genes is not straightforward. The public consortium's initial set contains about 32,000 genes, made up of around 15,000 known genes and 17,000 predictions. But these 32,000 genes are estimated to come from around 24,500 actual genes — some predicted genes could be 'pseudogenes', or just fragments of real genes. On the other hand, the sensitivity of prediction tends to be only about 60%, so it is reasonable to assume that another 6,800 or so genes (40% of 17,000) have been overlooked. This is how the present estimate of about 31,000 genes (6,800 plus 24,500) was reached[1]. Celera predicts that there are around 39,000 genes, but warns that the evidence for some 12,000 of these is weak[2]. The two groups use different gene-identification techniques, so these numbers are not directly comparable. Minor changes in procedures or data could alter either figure considerably. For example, such changes led to a recent estimate being lowered[7,8] from 120,000 to fewer than 81,000 — and both now seem untenable. Much is a matter of interpretation.

Fortunately, there is every reason to believe that the quality of gene prediction will rapidly improve, and an experimental technique for doing so is discussed on page 922 (ref. 9). With the sequencing of the genomes of other vertebrates, our ability to detect genes by their similarity to known sequences will get better. This is because, thanks to natural selection, gene sequences tend to be altered less during evolution than the DNA surrounding them. In a couple of years we should have at least a more complete list of testable gene candidates.

Despite all this, the information now available has profound implications. For example, there are already many heavily hunted disease-associated genes that have been identified using the public draft (ref. 1, Table 26, page 912). Together with studies of single nucleotide polymorphisms — the base differences from human to human — the draft also provides a framework for understanding the genetic basis and evolution of many human characteristics.

With the draft in hand, researchers have a new tool for studying the regulatory regions and networks of genes. Comparisons with other genomes should reveal common regulatory elements, and the environments of

Box 2 When is a predicted gene a gene?

How many genes are encoded in the human genome? This is a simple question without — as yet — a straightforward answer[13]. The density of genes in the human genome is much lower than for any other genome sequenced so far (Fig. 1), making it particularly difficult to predict where genes are.

Both Celera and the public sequencing consortium used computational algorithms to model genes and make predictions, but such methods are far from perfect. Not only can the start and end positions of a predicted gene be wrong, but exons (the coding parts of a gene) can be missed entirely or wrongly predicted to exist. To reduce this latter effect, the public sequencing consortium required the exons of predicted genes to be 'confirmed', by showing significant similarity to a known sequence (DNA or protein) in a database. But this requirement might be too conservative, making it difficult to predict the presence of new gene families. Celera has required similar confirmation of predictions, but its mouse-genome sequencing project may have provided evidence for further vertebrate-specific genes.

Spurious prediction is also a problem. All genes are expressed by being copied (transcribed) into messenger RNA; most messenger RNAs are then translated into proteins. But even evidence that a stretch of DNA is transcribed does not definitively show that stretch to be a gene. We do not know how efficiently cells control transcription; indeed, it seems likely that non-gene DNA sequences are transcribed relatively frequently[12]. Nor do we know how well the cell identifies transcripts that cannot be translated into a functioning protein. Moreover, proteins that cannot serve any useful function (for example, because they cannot fold correctly) could be made, but rapidly removed. To arrive at a true set of protein-encoding genes, we cannot rely on computational techniques alone, but must continue to characterize proteins and their functions.

These problems provide scope for estimates of human gene number to vary widely. Although recent estimates are converging in the 30,000–40,000 range (as opposed to earlier estimates of 100,000 or so), it could be many years before we have the final answer. P.B & R.C.

genes shared with other species may offer insight into function and regulation beyond the level of individual genes. The draft is also a starting point for studies of the three-dimensional packing of the genome into a cell's nucleus. Such packing is likely to influence gene regulation.

On a more applied note, the information can be used to exploit technologies such as chips made using DNA or proteins, complementing more traditional approaches. Such chips could now, for instance, contain all the members of a protein family, making it possible to find out which are active in particular diseased tissues. A new world of biotechnology will provide tools and information by exploiting genome data.

Sequencing the tough leftovers of the human genome will be essential. Without a finished sequence, we will not know what we are missing. Each missed gene is potentially a missed drug target, and even gene-poor areas might be critical for gene regulation. Nevertheless, we must now confront the fact that the era of rapid growth in human genomic information is over. The challenge we face is nothing less than understanding how this comparatively small set of genes creates the diversity of phenomena and characteristics that we see in human life. The human genome lies before us, ready for interpretation.

Peer Bork and Richard Copley are at EMBL, Meyerhofstrasse 1, 69012 Heidelberg, Germany. Peer Bork is at the Max-Delbrück Center for Molecular Medicine, Robert-Rössle-Strasse 10, 13125 Berlin-Buch, Germany.
e-mails: Peer.Bork@EMBL-Heidelberg.de
Richard.Copley@EMBL-Heidelberg.de

1. International Human Genome Sequencing Consortium *Nature* **409,** 860–921 (2001).
2. Venter, J. C. *et al. Science* **291,** 1304–1351 (2001).
3. Dunham, I. *et al. Nature* **402,** 489–495 (1999).
4. The Chromosome 21 Mapping and Sequencing Consortium *Nature* **405,** 311–319 (2000).
5. The *C. elegans* Sequencing Consortium *Science* **282,** 2012–2018 (1998).
6. Collins, F. S. *et al. Science* **282,** 682–689 (1998).
7. Liang, F. *et al. Nature Genet.* **25,** 239–240 (2000).
8. Liang, F. *et al. Nature Genet.* **26,** 501 (2000).
9. Shoemaker, D. D. *et al. Nature* **409,** 922–927 (2001).
10. The Arabidopsis Sequencing Consortium *Cell* **100,** 377–386 (2000).
11. Adams, M. D. *et al. Science* **287,** 2185–2195 (2000).
12. Normile, D. & Pennisi, E. *Science* **285,** 2038–2039 (1999).
13. Aparicio, S. *Nature Genet.* **25,** 129–130 (2000).
14. Goffeau, A. *et al. Nature* **387** (suppl.), 1–105 (1997).
15. The Arabidopsis Genome Initiative *Nature* **408,** 796–815 (2000).

The draft sequences

Comparing species

Gerald M. Rubin

Comparing the human genome sequences with those of other species will not only reveal what makes us genetically different. It may also help us understand what our genes do.

How are the differences between humans and other organisms reflected in our genomes? How similar are the numbers and types of proteins in humans, fruitflies, worms, plants and yeast? And what does all of this tell us about what makes a species unique? With the publication of the draft human genome sequences, on page 860 of this issue[1] and in this week's *Science*[2], we can start to compare the sequences of vertebrate, invertebrate and plant genomes in an attempt to answer these questions.

An obvious place to start our comparison is the total number of genes in each species. Here is a real surprise: the human genome probably contains between 25,000 and 40,000 genes, only about twice the number needed to make a fruitfly[3], worm[4] or plant[5]. We know that there is a higher degree of 'alternative splicing' in humans than in other species. In other words, there are often many more ways in which a gene's protein-coding sections (exons) can be joined together to create a functional messenger RNA molecule, ready to be translated into protein. So more proteins are encoded per gene in humans than in other species.

Even so, we cannot escape the conclusion — drawn previously from comparisons of simpler genomes[6] — that physical and behavioural differences between species are not related in any simple way to gene number. Many researchers, struck by the fact that there are four times as many genes in some gene families in the human genome compared with fruitflies[7], extrapolated from these cases and suggested that the human genome might be the product of two doublings of the whole of a simpler genome found in the common ancestor of fruitflies and humans. But, as the analyses of the human genome show[1,2], if such doublings did occur, the evidence for them has since been obscured by massive gene loss and amplification of particular gene families in the human genome.

Individual proteins often feature discrete structural units, called domains, that are conserved in evolution. More than 90% of the domains that can be identified in human proteins are also present in fruitfly and worm proteins, although they have been shuffled to create nearly twice as many different arrangements in humans[1,2]. Thus, vertebrate evolution has required the invention of few new domains. Of the human proteins that are predicted to exist, 60% have some sequence similarity to proteins from other species whose genomes have been sequenced. Just over 40% of the predicted human proteins share similarity with fruitfly or worm proteins. And 61% of fruitfly proteins, 43% of worm proteins and 46% of yeast proteins have sequence similarities to predicted human proteins.

But what about the proteins whose sequences show no strong similarity to known proteins from other species? Over a third of the yeast, fruitfly, worm and human proteins fall into this class. These proteins might retain similar functions, even though their sequences have diverged. Or they might have acquired species-specific functions.

Alternatively, we may need to entertain the possibility that the open reading frames that encode these proteins are maintained in a new way, one that is independent of the precise amino-acid sequence and thus is free to evolve rapidly. (An open reading frame is the part of a gene encoding the amino-acid sequence of its protein product.) After all, we know that cells have at least one mechanism, called nonsense-mediated decay of mRNA, for detecting imperfect open reading frames irrespective of the amino-acid sequence that they encode[8].

It will be interesting to see the extent to which the number of human proteins in this rapidly evolving class decreases as the genomes of other vertebrates, such as mice, are sequenced. This will give us an indication of just how fast these proteins are changing. Indeed, there is already evidence from studies of flies[9] and worms[10] that these rapidly evolving proteins are less likely to have essential functions, consistent with their being less likely to be conserved during evolution.

Such comparisons of distantly related genomes are fascinating from an evolutionary point of view. But comparison of closely related genomes will be much more important in addressing the key problem now facing genomics — determining the function of individual DNA segments. The concept is simple: segments that have a function are more likely to retain their sequence during evolution than non-functional segments. So DNA segments that are conserved between species are likely to have important functions. The ideal species for comparison are those whose form, physiology and behaviour are as similar as possible, but whose genomes have evolved sufficiently that non-functional sequences have had time to diverge. In practice, there may be no one ideal species, because different genes and regulatory sites evolve at different rates. Nevertheless, this approach has a long history of success, and becomes progressively more efficient as the cost of DNA sequencing declines.

One use of such sequence comparisons is