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

to determine the structure of genes — which parts (the exons) make their way into a functional mRNA molecule and which do not (the introns). The high degree of alternative splicing in vertebrates makes this comparative approach particularly important. Gene-finding computational algorithms cannot easily predict the existence of alternative forms of an mRNA without experimental information, but this information is difficult to come by in the case of rare mRNAs. For example, an exon that is used in only a few cells of the human brain might never be experimentally detected in an mRNA. But that exon's sequence would probably be conserved in the mouse genome.

Comparing the genomes of closely related species can also help in identifying gene-control regions. This approach has been used for over two decades[11], and has been validated by showing that the conserved sequences indeed correspond to functional control elements in individual genes[12]. But this computational problem is more difficult than identifying exons, and it will be challenging to scale up to a genome-wide level. The proteins that control gene expression by recognizing regulatory regions often detect sequence features that elude the best computer algorithms, and may use information from contacts with other proteins that is difficult to model. Proteins are simply cleverer than computers.

That said, our knowledge of the DNA-binding properties of individual proteins, as well as the structural features of the DNA sites to which they bind, continues to increase. Moreover, we can use experimental evidence; for example, genes that are expressed together might be expected to share control elements. And, as methods for comparing sequences continue to improve, we can expect to learn more about elusive features of the genome, such as genes encoding RNAs that do not encode proteins[13], start points of DNA replication, and genetic elements that control chromosome structure. ■

Gerald M. Rubin is in the Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, California 94708-3200, and the Howard Hughes Medical Institute, 4000 Jones Bridge Road, Chevy Chase, Maryland 20815-6789, USA.
e-mail: gerry@fruitfly.berkeley.edu

1. International Human Genome Sequencing Consortium *Nature* **409,** 860–921 (2001).
2. Venter, J. C. *et al. Science* **291,** 1304–1351 (2001).
3. Adams, M. D. *et al. Science* **287,** 2185–2195 (2000).
4. The *C. elegans* Sequencing Consortium *Science* **282,** 2012–2018 (1998).
5. The Arabidopsis Genome Initiative *Nature* **408,** 796–815 (2000).
6. Rubin, G. M. *et al. Science* **287,** 2204–2215 (2000).
7. Spring, J. *FEBS Lett.* **400,** 2–8 (1997).
8. Hentze, M. W. & Kulozik, A. E. *Cell* **96,** 307–310 (1999).
9. Ashburner, M. *et al. Genetics* **153,** 179–219 (1999).
10. Fraser, A. G. *et al. Nature* **408,** 325–330 (2000).
11. Ravetch, J. V., Kirsch, I. R. & Leder, P. *Proc. Natl Acad. Sci. USA* **77,** 6734–6738 (1980).
12. Fortini, M. E. & Rubin, G. M. *Genes Dev.* **4,** 444–463 (1990).
13. Lee, R. C., Feinbaum, R. L. & Ambros, V. *Cell* **75,** 843–854 (1993).

Single nucleotide polymorphisms

From the evolutionary past…

Mark Stoneking

Single nucleotide polymorphisms are the bread-and-butter of DNA sequence variation. They provide a rich source of information about the evolutionary history of human populations.

Studies of genetic variation in human populations began inauspiciously[1]. The first such study — of ABO blood-group frequencies — was carried out by two Polish immunologists, Ludwik and Hanka Hirszfeld, at the end of the First World War. This work was notable for its broad coverage of the world's populations, large sample sizes and scrupulous attention to anthropological details. Yet the Hirszfelds still ran into difficulties in publishing in *The Lancet*, the premier medical journal of the time. The editor could not see the relevance of their work, and so this seminal study of human genetic variation first appeared in an obscure anthropological journal[2]. The relevance became abundantly clear when Felix Bernstein subsequently used the Hirszfelds' data to demonstrate that the ABO blood-group frequencies were better explained by a single gene with three variants (alleles), and not — as prevailing wisdom then held — two genes each with two alleles[3].

Happily, times have changed, diversity is now all the rage[4,5], and editors have become more appreciative of the importance of human genetic variation. The latest evidence of that is the paper on page 928 of this issue[6], which reports the identification and mapping of 1.4 million single nucleotide polymorphisms (SNPs, pronounced 'snips') in the human genome. The paper is the result of the labours of a large collaboration, The International SNP Map Working Group.

So, what are SNPs? Quite simply, they are the bread-and-butter of DNA sequence variation — polymorphism, to those in the business. A DNA sequence is a linear combination of four nucleotides; compare two sequences, position by position, and wherever you come across different nucleotides at the same position, that's a SNP (see Fig. 1 on page 823). So SNPs reflect past mutations that were mostly (but not exclusively) unique events, and two individuals sharing a variant allele are thereby marked with a common evolutionary heritage. In other words, our genes have ancestors, and analysing shared patterns of SNP variation can identify them.

However, the real importance of SNPs is that there are so many of them. One estimate[7] is that comparing two human DNA sequences results in a SNP every 1,000–2,000 nucleotides. That may not sound like much until you realize that there are 3.2 billion nucleotides in the human genome, which translates into 1.6 million–3.2 million SNPs. And that's just from comparing two sequences — the total number of SNPs in humans is obviously much more. Most human variation that is influenced by genes can be traced to SNPs, especially in such medically (and commercially) important traits as how likely you are to become afflicted with a particular disease, or how you might respond to a particular pharmaceutical treatment, as discussed by Chakravarti[8] on the following page. And even when a SNP is not directly responsible, the sheer number of SNPs means they can also be used to locate genes that influence such traits.

The deluge of SNPs reported by the SNP working group[6] also promises great things for those of us who analyse patterns of molecular genetic variation to reconstruct the evolutionary history of human populations. Our genes contain the signature of an expansion from Africa within the past 150,000 years or so[9]. But there is still debate as to whether the modern humans from Africa completely replaced archaic non-African populations with no interbreeding, or whether we perhaps carry the vestiges of Neanderthal or other archaic non-African genes.

Demonstrating a recent African origin for every single one of our 3.2 billion nucleotides goes beyond the bounds of reason or necessity, but there is still much to be learned. For a start, most of our insights into molecular anthropology arise from DNA in mitochondria and (more recently) polymorphisms of the Y chromosome. This is because these DNA sequences are haploid — that is, represented just once in each cell, in contrast to the other chromosomes, which are represented twice — and they are inherited from just one parent, so they do not undergo the usual sequence shuffling (recombination) during egg and sperm production. This makes them easier to analyse and extremely informative. But both suffer from the drawback that, in the absence of recombination, they behave as single genes, and the history of any single gene can differ from that of a population or species because of natural selection or chance events involving that gene.

Accurate inferences concerning popula-