most genes are located outside the heterochromatic regions; interestingly, however, in regions of the genome rich in GC bases, the gene density is greater and the average intron size is lower. These introns — made up of largely meaningless sequence that breaks up the protein-coding sequences (exons) of genes — are much longer in human DNA than in the genomes previously sequenced. Their dilution of the coding sequence is one element that makes finding genes by computer so difficult in human DNA.

A major interest of the genome sequence to many biologists will be the opportunity it provides to discover new genes in their favourite systems — for instance, cell biologists will search for new genes for signalling proteins, and neurobiologists will look for new ion channels. This data-mining exercise was carried out by various groups which report their initial findings in papers that appear on pages 824–859 of this issue. They found some new and interesting genes, but surprisingly few, and occasionally could not find the full extent of genes that they knew were there. The paucity of discoveries reflects their concentration on systems that were previously heavily studied.

Gene-regulatory sequences are now there for all to see, but initial attempts to find them were also disappointing. This is where the genomic sequences of other species — in which the regulatory sequences, but not the functionally insignificant DNA, are likely to be much the same — will open up a cornucopia. Basically, the human sequence at its present level of analysis allows us to answer many global questions fairly well, but the detailed questions remain open for the future.

What interested me most about the genome? The number of genes is high on the list. The public project estimates that there are 31,000 protein-encoding genes in the human genome, of which they can now provide a list of 22,000. Celera finds about 26,000. There are also about 740 identified genes that make the non-protein-coding RNAs involved in various cell housekeeping duties, with many more to be found. The number of coding genes in the human sequence compares with 6,000 for a yeast cell, 13,000 for a fly, 18,000 for a worm and 26,000 for a plant. None of the numbers for the multicellular organisms is highly

## Genome speak

**Allele** Humans carry two sets of chromosomes, one from each parent. Equivalent genes in the two sets might be different, for example because of *single nucleotide polymorphisms*. An allele is one of the two (or more) forms of a particular gene.

**Bacterial artificial chromosome (BAC)** A chromosome-like structure, constructed by genetic engineering, that carries genomic DNA to be *cloned*.

**Centromere** Chromosomes contain a compact region known as a centromere, where sister chromatids (the two exact copies of each chromosome that are formed after replication) are joined.

**Cloning** The process of generating sufficient copies of a particular piece of DNA to allow it to be sequenced or studied in some other way.

**Complementary DNA (cDNA)** A DNA sequence made from a *messenger RNA* molecule, using an enzyme called reverse transcriptase. cDNAs can be used experimentally to determine the sequence of messenger RNAs after their introns (non-protein-coding sections) have been *spliced* out.

**Conservation** Genes that are present in two distinct organisms are said to be conserved. Conservation can be detected by measuring the similarity of the two sequences at the base (RNA or DNA) or amino-acid (protein) level. The more similarities there are, the more highly conserved the two sequences.

**Euchromatin** The gene-rich regions of a genome (see also *heterochromatin*).

**Eukaryote** An organism whose cells have a complex internal structure, including a nucleus. Animals, plants and fungi are all eukaryotes.

**Expressed sequence tag (EST)** A short piece of DNA sequence corresponding to a fragment of a *complementary DNA* (made from a cell's *messenger RNA*). ESTs have been used to hunt for genes, so hundreds of thousands are present in sequence databases.

**Genome** The complete DNA sequence of an organism.

**Genotype** The set of genes that an individual carries; usually refers to the particular pair of alleles (alternative forms of a gene) that a person has at a given region of the genome.

**Haplotype** A particular combination of alleles (alternative forms of genes) or sequence variations that are closely linked — that is, are likely to be inherited together — on the same chromosome.

**Heterochromatin** Compact, gene-poor regions of a genome, which are enriched in simple sequence repeats. As it can be impossible to *clone*, heterochromatin is often ignored when calculating the percentage of a genome that has been sequenced. Heterochromatin was originally identified as regions of the genome that stained differently to euchromatin (gene-rich regions).

**Introns and exons** Genes are *transcribed* as continuous sequences, but only some segments of the resulting *messenger RNA* molecules contain information that codes for the gene's protein product. These segments are called exons. The regions between exons are known as introns, and are *spliced* from the RNA before the product is made.

**Long and short arms** The regions either side of the centromere, a compact part of a chromosome, are known as arms. As the centromere is not in the centre of the chromosome, one arm is longer than the other.

**Messenger RNA (mRNA)** Proteins are not synthesized directly from genomic DNA. Instead, an RNA template (a precursor mRNA) is constructed from the sequence of the gene. This RNA is then processed in various ways, including *splicing*. Spliced RNAs destined to become templates for protein synthesis are known as mRNAs.

**Mutation** An alteration in a genome compared to some reference state. Mutations do not always have harmful effects.

**Phenotype** The observable properties and physical characteristics of an organism.

**Polymorphism** A region of the genome that varies between individual members of a population. To be called a polymorphism, a variant should be present in a significant number of people in the population.

**Prokaryote** A single-celled organism with a simple internal structure and no nucleus. Bacteria and archaebacteria are prokaryotes.

**Proteome** The complete set of proteins encoded by the *genome*.

**Pseudogene** A region of DNA that shows extensive similarity to a known gene, but which cannot itself function, either because it has lost the signal required for *transcription* (the promoter sequence) or because it carries mutations that prevent it from being *translated* into protein.

**Recombination** The process by which DNA is exchanged between pairs of equivalent chromosomes during egg and sperm formation. Recombination has the effect of making the chromosomes of the offspring distinct from those of the parents.

**Restriction endonuclease** An enzyme that cleaves DNA at every location at which a particular short sequence occurs. Different types of restriction endonuclease cleave at different target sequences.

**Single nucleotide polymorphism (SNP)** A *polymorphism* caused by the change of a single nucleotide. Most genetic variation between individual humans is believed to be due to SNPs.

**Splicing** The process that removes introns (non-protein-coding portions) from *transcribed* RNAs. Exons (protein-coding portions) can also be removed. Depending on which exons are removed, different proteins can be made from the same initial RNA or gene. Different proteins created in this way are 'splice variants' or 'alternatively spliced'.

**Transcription** The process of copying a gene into RNA. This is the first step in turning a gene into a protein, although not all transcripts lead to proteins.

**Transcriptome** The complete set of RNAs *transcribed* from a genome.

**Translation** The process of using a *messenger RNA* sequence to build a protein. The messenger RNA serves as a template on which transfer RNA molecules, carrying amino acids, are lined up. The amino acids are then linked together to form a protein chain.

Peer Bork and Richard Copley