EJSO 2003; **29**: 693–696
doi:10.1016/S0748-7983(03)00082-9

EJSO
European Journal of Surgical Oncology

EDUCATIONAL SECTION

Quaternary basic man

D. A. Rew

Royal South Hants Cancer Centre, Southampton University Hospitals, Southampton SO14 0YG, UK

This article is one of a series of educational features on contemporary developments in cell and molecular sciences of relevance to cancer specialists. It describes significant aspects of the human genome and related technologies, as revealed in the publication of draft sequences by two consortia in 2001.

Key words: human genome; cancer; therapeutics; experimentation.

INTRODUCTION

The sequence of the human genome was published in early 2001.[1–4] This marked the culmination of a remarkable series of intellectual and technical achievements. Less than 50 years after Watson, Crick and colleagues proposed the structure of DNA,[5] the formula for Man was captured in a computer as digital list of some three billion C, A, G and T nucleotide bases. Our perspectives on disease would be transformed.

The dust has now settled on the hype and publicity of this scientific tour de force, and we can better understand the implications for cancer biology and therapeutics, and the practical limitations of the data as currently presented. The scientific momentum has if anything increased since 2001, with progress in the sequencing of other mammalian genomes on an industrial scale; in the study of the transcriptome, the global mRNA product of the genome; and of the proteome, the global protein product of the genome.

The public drafts of the human genome were published in February 2001 in concurrent issues of Nature and Science. The core sequence information produced by the academic multinational, publicly funded human genome project (HGP) consortium was reported in Nature, and that of the privately funded Celera Corporation was published in Science. The HGP data can be inspected on the Internet at a number of sites, using a variety of publicly available software tools, which allow access to various levels of detail. The Celera data is also available, but under greater restrictions consistent with its commercially funded origins.

Sequencing is the science of determining the sequence of the genetic code, C/A/T/G nucleotide by nucleotide, accurately and reproducibly. Each consortium has published a draft sequence based on sequencing human DNA a number of times and from a small number of individuals to minimise error. The draft sequences are still being completed and cross-checked in final detail, for example where problems arise in linking together long sections of highly repetitive DNA sequence. Raw DNA sequence data is a long string of nucleotide code, which requires massive computational power and corroborative data for its deconvolution. Functional human genes and gene fragments are interspersed with long lengths of repetitive and non-coding DNA, and with much evolutionarily 'redundant' genetic code.

THE TECHNIQUES OF SEQUENCING

HGP and Celera adopted different and competitive approaches to sequencing. HGP used a methodical approach to sequencing more than 20 000 predetermined fragments of human DNA known as large insert clones. These were separately amplified in bacterial DNA for sequence analysis, with different chromosomes and

Correspondence to: David A. Rew, University Surgical Unit, Royal South Hants Hospital, Brintons Terrace, Southampton SO14 0YG, UK. Tel.: +44-2380-767823; Fax: +44-2380-825148; E-mail: dr1@soton.ac.uk

0748–7983/03/$30.00

sections of the genome allocated to different academic units around the world. The grid of information was then reconstructed with increasing layers of detail.

Celera used a faster, computer intensive technique known as the 'whole genome shotgun (WGS)' approach. Human chromosomes were broken up into random fragments of DNA, which were individually sequenced and then reconstructed using computer algorithms. Each consortium has benefited from the other's approach and the stimulus of competition to expedite the end result. The existence of two data sets from separate sources gives additional confidence in the quality of the data, which are broadly concordant from each source.

THE DESCRIPTION OF THE HUMAN GENOME

The total genome comprises between 2.9 and 3.2 gigaBases (thousand million nucleotides), of which less than 1.5% appears to encode for proteins. Twenty-eight percent of the genome appears to be transcribable into RNA. The sequences are approximately 90% finished in the gene rich regions. There are some 30 000 demonstrable gene sequences. This seems to some to be surprisingly few for such a complex organism as man, and is considerably less than early estimates of 100 000 genes. In comparison, there are some 6000 genes in yeast, 13 000 in the fruit fly, 18 000 in the nematode worm and 26 000 in the mustard seed plant. Genomic size is not proportional to organismal complexity, because of the varying proportions of redundant and junk DNA in different species. Species such as the puffer fish have a much smaller proportion of non-functional DNA, and yet the organism develops normally.

THE CLASSIFICATION OF INDIVIDUAL GENES

Unique gene coding sequences in Man appear to code less than 5% of the genome. Genes code for proteins, or for the ribosomal and transfer RNAs, which transcribe mRNA into protein. Genes may be identified by specific Start and Stop sequences, which aid transcription. Many genes are broken up on the chromosome by sequences of non-functional DNA known as introns, and some gene fragments may recombine in different ways, thus extending the number of possible combinations. Identification of genes is thus not a precise science. Nevertheless, sufficient data is now in place to make a reasonable estimate of the number of genes, and the proportional of 'non-functional' DNA in the code. Genes vary considerably in size. The gene for the muscle protein dystrophin is 2.4 million bases long, including non-functional DNA. The muscle protein titin is coded by more than 80 000 bases, assembled from almost 180 separate exons.

Computer aided comparative studies of the known sequence data of similar genes from other species help in the identification and functional classification of the putative human genes. Large, public, internet accessible databases of genes, mRNA sequences and proteins have been collated. In many cases, the identity of the gene will be immediately apparent or readily deduced from archival data. In other cases, the function may be deduced from functional components or motifs recognised in other genes. The position on the genome and proximity to other genes may also help establish identity. Confirmation of the identity and function of many putative genes may have to await further genomic studies on other species. Sequences will rarely be identical in interspecies and inter-individual comparisons, and they are often discontinuous. Software algorithms must thus be sophisticated to effect matches.

NON-GENE DNA

Repeat sequences of DNA appear to code for at least 50% of the genome. They include a mixture of inactivated genes and partially copied genes (pseudogenes), short segment repeats $(\text{single base})_n$, sectional duplications of up to 300 kilobases (kB), and blocks of repeats of functional genes, such as for ribosomal proteins. Transposons are particularly common. These are evolutionarily mobile elements, so called 'parasitic' segments of genetic information, of DNA or RNA, which can be transferred between species such as bacteria and incorporated into chromosomes of the host species either directly as DNA or by reverse transcription of RNA. The precise functions and the reasons why all this seemingly non-functional DNA is tolerated in the genome are as yet unclear. Internal genomic evolution has clearly been taking place in parallel with the evolution of species.

GENOMIC ARCHAEOLOGY

The DNA in all organisms on Earth must represent an absolute continuity from the earliest self-replicating molecules and simple organisms, added to, modified and deleted over countless generations, and producing evolutionary diversification, speciation and mutation. Nature appears to be highly conservative in the use of genes and proteins, modifying them gradually through mutation and Darwinian competition to new functions. The fundamental genes for cell structure and function, DNA preservation and accurate replication are highly conserved across all living species, as evidenced by the DNA sequencing studies from a variety of organisms. Only 94 of 1278 protein families in the human genome

are unique to vertebrates. The sequence in functioning genes compared with other species, and the nature of the inactive sequences, repeats, transposons and pseudogenes, is in effect a fossil record going back over thousands of generations. The estimated rate of accumulation of sequence mutations allows us to calculate the time of evolutionary divergence of various species and phyla, and of the acquisition and mutation rates in individual chromosomes, such as, for example, the human male Y chromosome.

Large fractions of human DNA seem to have been built up over millennia from RNA viruses incorporated into the genome by reverse transcription. The HIV-1 virus is such a parasite. Several hundred other genes appear to have been incorporated from bacteria. There also appear to be evolutionarily mobile genetic elements which resemble retroviruses (those which transcribe RNA to DNA) and which move around the genome and copy themselves as they go. Some become extinct, and lose the ability to self-replicate. Others continue to modify the genome. Thus, for example, the Alu sequence is replicated one million times in the genome, while the self-replicating LINE1 sequence accounts for 17% of human DNA. They are concentrated in the X chromosome and are rare in the 'homeobox' regions, those groups of genes which regulate body development and whose disruption would not be compatible with life. A picture thus emerges of a relatively condensed ancestral DNA which has become populated and interspersed with huge quantities of exogenous DNA.

THERAPEUTIC IMPLICATIONS OF THE HUMAN GENOME PROJECT

It is too early to discern practical therapeutic applications of the HGP. Advances will come in a number of ways. Single dominant and recessive genes for specific and rare hereditary (monogenic) diseases are likely to be identified more quickly, and at least 30 such genes have been so, for example the CNGA3 gene for cone photoreceptor function in colour blindness. Multifactorial genetic diseases will take longer to sort out.

Of particular practical interest is the identification of drug targets among metabolic and receptor genes and their products. For example, a gene coding for a serotonin neuroreceptor 5-HT-3B, has been found within the draft genome and may be a therapeutic target for the newly recognised 5HT pathway.

HUMAN INDIVIDUALITY AND SINGLE NUCLEOTIDE POLYMORPHISMS (SNPS)

Individual humans differ from each other by approximately one base pair per thousand. These mutations and substitutions are known as single nucleotide polymorphisms (SNPs). When expressed in genes, they can subtly or dramatically alter the function of the protein product, and give rise to dysfunction and disease, including neoplasia. Examples include the BRCA and p53 genes, in each of which many SNPs have been identified. Other SNPs may help predict individual responses to therapeutic interventions. The scale of the challenge in understanding SNPs is massive. More than 2.5 million SNPs have already been archived. It has been suggested that personalised SNP profiles might define the health risks and disease susceptibilities of individuals.

THE FUTURE

The analysis of draft DNA sequence data will be a continuing challenge for computational biology. Cross-checking and filling in of gaps will take several years. The speed with which the human genome was sequenced by the application of technology on the industrial scale portends a great increase in the rate at which other organisms will be sequenced and compared, starting with the most economically, scientifically and commercially important. For example, the mouse genome is being finalised, and we may expect all the primate genomes to be worked through in due course. Eventually, we will be able to subtract the common elements in each genome and compare the differences, which give each species its specific characteristics, form and function. We will be able to identify those features of gene and chromosome structure which link and distinguish one species from another.

CONCLUSIONS

There is a massive amount of work now to be done to finalise the draft sequences and then to understand how the raw linear data sets are converted into the complex three dimensional form and function of proteins and hence of living and diseased cells and organisms. The limitations on human experimentation will dictate the completion of whole genome analysis of the mouse to allow detailed and exhaustive studies of development and disease on a gene by gene and mutation by mutation basis.

For all its technical virtuosity, whole genome DNA sequence data is only one step to the detailed understanding of the molecular construction and workings of

the human body. It is a dictionary rather than a language, or a census rather than a description of the working of a human community. The controls to the expression of individual genes and patterns of genes at each stage of the life of cells, tissues and organs are not understood. The proteome poses a much greater analytical challenge than does the genome, because proteins are much more transient species than DNA, and because multi-component proteins can be constructed from the templates of numerous genes, a phenomenon known as alternative splicing. Some 35% of genes seem to be readable in more than one way, and their expression is likely to be contextual, much as words such with common spelling such as 'bear' and 'lie' can only be understood in specific context.

There will be no instant insights into cancer and its treatment, but we are at last able to place finite boundaries on the size of the genome and its content, and to work through the problems posed in a logical fashion. Nevertheless, the language of human genomics and proteinomics will become more common in the applied cancer literature with time, and it is thus important that we can understand the terminology and evaluate the literature as it encroaches upon everyday professional practice.

KEY EDUCATIONAL POINTS

The draft sequence of the human genome has been published in duplicate by two competing consortia.

The core human genome comprises 2.9 Gb nucleotides and 30 000 genes.

Less than 5% of the genome codes for true genes. Much is composed of 'junk' DNA acquired through evolution and through self-replication of otherwise non-functional segments.

Patterns in the base sequence reveals the 'archeology' of molecular evolution of the genome.

Genome sequence is insufficient to understand form and function. Knowledge of the proteome, the full range of protein product of the genome, and its control is needed.

Many genes are believed to code for more than one protein through a building block system.

Advances in the technology of sequencing herald the early sequencing of the genomes of many other organisms.

Comparative genomics will allow new forms of investigation of hereditary and genetic diseases, including neoplasia.

REFERENCES

1. International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature* 2001; **409**: 860–921.
2. Venter JC *et al.* The sequence of the human genome. *Science* 2001; **291**: 1304–51.
3. Baltimore D. Our genome unveiled. *Nature* 2001; **409**: 814–6.
4. Wolfsberg TG, Mcentyre J, Schuler GD. Guide to the draft human genome. *Nature* 2001; **409**: 824–6.
5. Watson JD, Crick FHC. Molecular structure of nucleic acids. A structure for deoxyribosenucleic acid. *Nature* 1953; **171**: 737–8.

FURTHER READING

A number of other articles and commentary can be found in Nature 2001, vol. 409, pp. 814 et seq, and in Science 2001, vol. 291, issue 5507.
www.nhgri.nih.gov/genome_hub a list of sites of access to HGP data.
www.nature.com/genomics.
www.nature.com/nsu (articles on Nature Science Update).
www.sciencemag.org/content/vol291/issue5507.