British Journal of Nutrition (2004), **91**, 1–2
DOI: 10.1079/BJN20031012

Editorial

Of genes and genomes – and *dark matter*

The draft sequence of the human genome was published in February 2001 with simultaneous papers in *Nature* and in *Science* (Lander *et al.* 2001; Venter *et al.* 2001). For the scientific community this achievement was of enormous importance, and its broader cultural significance was emphasised by the involvement of political leaders on both sides of the Atlantic in the accompanying publicity. The draft sequence was replaced 2 years later, in April 2003, by the 'final' version. This was done not by publication in a journal but through an announcement that the sequence information had been deposited in publicly accessible databases. The sequence of the human genome undoubtedly represents the summit of the genome sequencing efforts that began more than a decade ago.

We now have the full sequence of the genome of a rapidly growing number of organisms. These range from bacteria such as *Escherichia coli* and *Salmonella enterica*, the nematode worm, *Caenorhabditis elegans*, the fruit fly, *Drosophila melangaster*, and plants such as rice (*Oryza sativa*). Priorities for the future include the sequencing of the honeybee (*Apis mellifera*), chicken (*Gallus gallus*), and chimpanzee (*Pan troglodytes*) genomes (see Couzin, 2003). Of particular importance is the deposition of the draft sequence of the genome of the mouse in May 2002 (see Marshall, 2002). The mouse is, of course, widely used as an experimental animal and as a model system for human disease, including in nutrition-related research. For example, genes critical for the regulation of energy balance, such as those encoding leptin and its receptor, have been identified through studies on mutant mice (Zhang *et al.* 1994; Chua *et al.* 1996; Lee *et al.* 1996).

More primitive organisms can also be valuable for investigating regulatory processes that impact on human function and disease, and this has been part of the underlying ethos of those studying *C. elegans*. The potential impact of such an approach is well illustrated by a recent report in which a genome-wide RNAi (RNA-mediated interference) analysis has identified a large number of genes implicated in normal fat storage (Ashrafi *et al.* 2003). The *C. elegans* genome is believed to encompass 16 757 genes, and 305 gene inactivations which reduce body fat and 112 that lead to increases in fat storage have been identified. A number of these *C. elegans* fat regulatory genes have mammalian homologues and some are already recognised as important in the control of body fat in man. But this approach has also led to the identification of genes encoding proteins not hitherto recognised as being involved in determining fat deposition, e.g. a lysosomal transporter and a peptide transporter (Ashrafi *et al.* 2003).

Although the number of genes in *C. elegans* is quoted as a very precise figure, the number of human genes continues to be a matter of uncertainty. A few years ago it was estimated that there were upwards of 100 000 protein-coding genes in the human genome. This number fell to as little as 32 000 when the draft sequence of the human genome was published (Lander *et al.* 2001), and the much lower than expected figure was one of the major surprises from the genome sequence. The difficulty in determining the precise number of human genes reflects the fact that exons – the protein coding regions – make up only around 2 % of the total DNA of the genome. This means that some 98 % of the genome does not directly code for proteins; some, of course, reflects regulatory sequences.

There are distinct regions of the genome which appear not to contain any genes – the so-called *dark matter* regions. There is continuing debate as to whether the dark matter does really contain genes; however, if present protein coding sequences are not recognised by the various software programmes that predict gene counts. These programmes themselves come up with very different estimates for the number of human genes outwith the regions of dark matter. At one end of the spectrum, GENSCAN predicts as many as 45 000 genes, while Ensembl/Genewise gives a prediction of just 24 500 – with other programmes providing figures closer to the lower end of this range (see Pennisi, 2003).

Three years ago the gene sequencing community set up a betting pool - GeneSweep - with a prize for the estimate which was nearest to the final agreed value. Last summer (2003) the decision was made to award the prize to the estimate closest to 24 500 – the Ensembl/Genewise prediction. In practise, even this currently accepted best value could be an over-estimate of the number of protein-coding genes, since it may include as many as 3 000 pseudogenes (see Pennisi, 2003). This suggests that the real number of human genes is little more than 20 000. Such a figure seems remarkable, given that *C. elegans* – a simple worm – is considered to contain nearly 17 000 genes.

Clearly, establishing the true number of protein-coding genes in the human genome is much more than of abstract interest; it is a pre-requisite for comprehending the full complexity of human function at the molecular, cellular and physiological levels – as well as for the

maturation of the rapidly burgeoning field of nutritional genomics (Trayhurn, 2003).

Paul Trayhurn
Editor-in-Chief
Liverpool Centre for Nutritional Genomics
Department of Medicine
University of Liverpool
Liverpool L69 3GA
UK
p.trayhurn@liverpool.ac.uk

References

Ashrafi K, Chang FY, Watts JL, *et al.* (2003) Genome-wide RNAi analysis of *Caenorhabditis elegans* fat regulatory genes. *Nature* **421**, 268–272.

Chua SC, Chung WK, Wupeng XS, *et al.* (1996) Phenotypes of mouse diabetes and rat fatty due to mutations in the *ob* (leptin) receptor. *Science* **271**, 994–996.

Couzin J (2003) GENOMICS: Sequencers examine priorities. *Science* **301**, 1176–1177.

Lander ES, Linton LM, Birren B, *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

Lee GH, Proenca R, Montez JM, *et al.* (1996) Abnormal splicing of the leptin receptor in diabetic mice. *Nature* **379**, 632–635.

Marshall E (2002) Public group completes draft of the mouse. *Science* **296**, 1005.

Pennisi E (2003) BIOINFORMATICS: Gene counters struggle to get the right answer. *Science* **301**, 1040–1041.

Trayhurn P (2003) Nutritional genomics – 'Nutrigenomics'. *Br J Nutr* **89**, 1–2.

Venter JC, Adams MD, Myers EW, *et al.* (2001) The sequence of the human genome. *Science* **291**, 1304–1351.

Zhang YY, Proenca R, Maffei M, Barone M, Leopold L & Friedman JM (1994) Positional cloning of the mouse obese gene and its human homolog. *Nature* **372**, 425–432.