CURRENT SCIENCE

Volume 80 Number 4 25 February 2001

EDITORIAL

Publishing the genome

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution.

– International Human Genome Sequencing Consortium

Decoding of the DNA that constitutes the human genome has been widely anticipated for the contribution it will make towards understanding human evolution, the causation of disease, and the interplay between the environment and heredity in defining the human condition.

– J. C. Venter *et al.*

The completion of the 'rough draft' of the human genome sequence has been announced again. The first time this happened was when the commanding generals of the sequencing armies, of the publicly-funded International Human Genome Sequencing Consortium and the private company Celera Genomics, Francis Collins and Craig Venter, appeared at a joint news conference last summer; an event that appeared to have been brokered by the then American President, Bill Clinton. The popular press went wild in the days that followed; some reports extravagantly concluding that as a consequence of genome sequencing all of us would lead disease-free lives extending to 150 years. Now the two genome groups have done it again, this time by publishing their results in scientific journals. But, getting the articles into print simultaneously, in a manner that the credit is equally shared between the two groups, appears to have involved considerable backroom effort. The final outcome is the appearance of two mammoth papers in the 15 February issue of *Nature* (International Human Genome Consortium, *Nature*, 2001, **409**, 860–921) and the 16 February issue of *Science* (J. C. Venter *et al.*, *Science*, 2001, **291**, 1304–1351). While the raw sequence data of the publicly-funded consortium have been flowing into the GenBank database, providing unrestricted access, the sequence data generated by Celera Genomics are available to academic scientists only in a restricted manner. The data-sharing debate has been contentious in the past and is likely to be even more rancourous in the future; ironically the remarkably rapid progress of the Celera group was in considerable measure aided by the large volume of data deposited in public databases. The editors of *Science* have presumably agreed to the access limitations insisted upon by the Celera group, so that the journal does publish the first major results of biology's most visible project. The honours seem to be equally divided between the two competing groups and journals; future historians may indeed reveal some fascinating insights into the processes by which the authors and editors reached a conclusion that would have done King Solomon proud.

But, what of the genome? The two groups have indeed assembled a draft sequence that covers over 90% of the genome; the International Consortium reports about 2.69 billion bases (to most of us, an incomprehensible string of the letters A, T, G and C), while the Celera group registered 2.91 billion bases. There is, quite simply, too much information to be digested easily. In the two major papers and several ancillary analyses that appear in the issues of *Science* and *Nature* there are many observations highlighted that will provide the impetus for new lines of biological research in the coming years. The most dramatic outcome of the first pass at 'annotating' (the term preferred by bioinformatics researchers for deciphering the encoded information) the genome, has been the significantly smaller number of genes that have now been found. In the early days of the sequencing program, gene number estimates fluctuated widely, most estimates varying from 60,000 to 75,000 genes, with a one-time peak of 120,000 genes. But, now that the dust is settling, the numbers are dramatically smaller. The Celera group has found '26,588 protein-encoding transcripts for which there was strong corraborating evidence and an additional ~12,000 computationally-derived genes with mouse matches or other weak supporting evidence'. The International Consortium concludes that 'there appear to be about 30,000–40,000 protein-coding genes in the human genome – only about twice as many as in worm or fly'. In the first phase of reactions that have appeared in the press, there appears to be some measure of discomfort in the fact that humans do not appear to possess a dramatically larger complement of genes than the fruitfly, *Drosophila melanogaster* (13,600), the worm, *Caenorhabditis elegans* (18,500) and the tiny plant *Arabidopsis thaliana* (25,000). Curiously, classical

biology has always lived quite comfortably with the 'C-value paradox', where there is no simple correlation between genome size and apparent organismal complexity. We now appear to be moving towards a refinement of the problem, as the number of protein coding genes does not appear to reconcile easily with our notions of complexity. Indeed, Venter and his colleagues draw attention to J. B. S. Haldane's 1937 speculation 'that a population of organisms might have to pay a price for the number of genes it can possibly carry'. In Haldane's view, large gene complements can reach a stage where 'each zygote carries so many new deleterious mutations that the population simply cannot maintain itself'. Remarkably, early estimates of gene loci in humans by Hermann Muller (1967) and J. F. Crow and M. Kimura (1970), hovered around a figure of 30,000 genes.

As long anticipated, the human genome is an enormously barren stretch of DNA bases that do not code for anything; a veritable 'desert', rarely punctuated by a gene oasis. The Y-chromosome, which is a male attribute, is particularly barren, containing relatively few, albeit important, genes. However, the size of the 'human proteome', the total complement of proteins, quite literally the workhorses of cellular biochemistry, maybe significantly higher; current estimates climb into the neighbourhood of 60,000. This would mean, of course, that the final translation of the 'message', encoded in human genes and genomes, would require considerably greater use, than hitherto anticipated, of 'alternative splicing mechanisms'; a process in which the final messages are constructed by piecing together coding regions in different ways. This molecular sleight of hand, would allow considerably more gene products (proteins) than predicted from the number of identified genes. A considerable part of the difficulty in decoding the information implicit in the human genome is the presence of tracts of non-coding DNA of widely varying length (introns), which separate segments of coding sequences (exons). The human genome sequence reveals large stretches of 'parasitic DNA', which appear to have come about by 'reverse transcription from RNA. In places the genome looks like a sea of reverse-transcribed DNA with a small admixture of genes'. (Baltimore, D., *Nature*, 2001, **409**, 814). These stretches of DNA might provide a link to our evolutionary past, containing an imprint of mankind's continuing interaction with viruses. Surprisingly, the genome sequence also suggests that as many as a couple of hundred human genes may be a result of direct transfer from bacteria, as distinct from evolution from bacterial genes. Clearly, we are all genetically modified ('engineered') organisms, natural transgenics created over eons of evolution. A Nature News Service summary alludes to the ongoing debate on genetically modified (GM) organisms: '... it is perhaps ironic that all humans, including those in the anti-GM lobby are GM organisms'.

In thinking about the impact of the unveiling of the genome, it is useful to consider David Baltimore's assessment: 'I've seen a lot of exciting biology emerge over the past 40 years. But chills still ran down my spine when I first read the paper that describes the outlines of our genome.... Not that many questions are definitively answered. For conceptual impact, it does not hold a candle to Watson and Crick's 1953 paper describing the structure of DNA. Nonetheless, it is a seminal paper launching the era of post-genomic science'. How do the authors of these, massive and at times forbiddingly complex, papers view the future? For Venter *et al.* the issues are clear. Low gene numbers limit the extent of complexity that can be explained purely on the basis of simple interactions between protein components, a cornerstone of reductionist biochemistry. The newer descriptions of 'robust networks', with which the rapidly growing literature of complexity theory abounds, may be called into play in biology, to rationalize the sometimes seemingly incomprehensible relationship between genotype and phenotype. For Venter *et al.* 'the real challenge of human biology, beyond the task of finding out how genes orchestrate the construction and maintenance of the miraculous mechanism of our bodies, will lie ahead as we seek to explain how our minds have come to organize thoughts sufficiently well to investigate our own existence'. The paper by the International Consortium ends (borrowing a memorable phrase from the classic paper of Watson and Crick): 'Finally, it has not escaped our notice that the more we learn about the human genome, the more there is to explore'. For these authors the journey has been long and T. S. Eliot provides inspiration: 'We shall not cease from exploration. And the end of all our exploring will be to arrive where we started, and know the place for the first time'.

The publication of the first papers to emerge from the human genome sequencing project marks a major milestone in biology. The papers represent a technological and organizational *tour de force*, which sets new standards for biological research. The Celera achievement of reaching a preset goal in less than a year is truly remarkable; competition undoubtedly catalysing the final sprint to the finish line. The first year of the 20th century saw the rediscovery of Mendel's Laws of Heredity and Max Planck's formulation of the quantum hypothesis. The opening weeks of the new millennium clearly belong to biology. What lies ahead is hardly predictable, but one certainty has emerged; the practice of biology will never be the same again.

P. Balaram