CURRENT SCIENCE

Volume 78 Number 12 25 June 2000

EDITORIAL

Genes and genomes

Large multiinstitutional and multinational projects have a tendency to run behind schedule and are usually characterized by substantial cost overruns. A major exception has been the Human Genome sequencing program, which was originally slated to end in about 2005, but appears to be reaching the finish line five years earlier. When whole genome sequencing was contemplated in the 1980s, most serious molecular biologists were loath to support a project, which appeared to be a monotonous, technology-driven exercise that might cause a large diversion of funds in the United States from the traditional 'small science' of individual investigator-driven projects. There was a general impression that the search for disease genes may be accomplished by approaches that did not involve the enormous labour of complete sequencing. But the skeptics were to be proved wrong, on almost all fronts, as sequencing technologies and data handling capabilities were revolutionized in the 1990s. The last few years have seen a dramatic acceleration of the rate of sequencing by consortia in Europe, Japan and the United States. Genome after genome has succumbed to this onslaught. The most notable recent victories on the battlefield of sequencing include the genome of the fruitfly, *Drosophila* (*Science*, 24 March, 2000), and a host of small genomes of pathogens, including *Mycobacterium leprae* (*Science*, 2000, **288**, 800). The publication of the sequences of chromosome 21 (M. Hattori *et al.*, *Nature*, 2000, **405**, 311) and chromosome 22 (I. Dunham *et al.*, *Nature*, 1999, **402**, 489), the smallest of the 24 human chromosomes, marks the beginning of the end of the sequencing phase of the Human Genome project. It is clear now that we will soon have available the sequence of the 3 billion plus DNA bases, that constitute the core of human inheritance. The last battles for the human genome have been marked by intense competition between publicly funded groups spearheaded by the US National Institutes of Health and a private company, Celera Genomics founded by Craig Venter, which appears to be repeatedly demonstrating that the best catalyst for scientific progress in the modern age may be the smell of commercial success. The race for an almost mythical finish line and the growing controversy over the release and sharing of data, have obscured the issue of what happens after all the sequences are in and safely stored in digital databases. One of the more difficult aspects of war is managing the peace that follows. The genome wars of today are likely to throw up a host of new problems that might still be unanticipated.

In pondering on genomes it might be useful to ask: 'what are genes?' Some years ago on the verge of the biodiversity convention at Rio de Janeiro, a newspaper featured a cartoon of the inevitable minister about to head for the jamboree, plaintively asking an assistant: 'what is a gene'. At first glance this appears a ridiculously simple question. The acceptable answers in a quiz or an examination are many. Some will hedge their bets and say generally, that 'a gene is an unit of inheritance'; classical genetics, of course, associates genes with specific phenotypic characteristics of organisms. A stiff and formal answer found in the book *Molecular Biology of the Cell* by Alberts *et al.*, states that 'a gene is defined as a nucleotide sequence in a DNA molecule that acts as a functional unit for the production of an RNA molecule'. And, elementary molecular biology tells us that if RNA is around the corner, proteins cannot be far behind. But, in today's context the genome gold diggers are primarily looking for sequences that will translate into protein products. After all, as Arthur Kornberg said some years ago, 'in my theater, the nucleic acids write the script but the enzymes (proteins) do the acting'. If all the protein sequences in an organism ('the proteome') are the real target of the genome programs, it is clear that the post-sequencing era will be dominated by efforts to decipher the segments of DNA that translate into protein products. The human genome, of course, contains an enormous amount of non-coding DNA (once dismissively labelled as 'junk DNA' by Francis Crick); the task ahead will be to sift the wheat from the chaff. If the human genome is viewed as an incredible book written with over three billion letters, based on a four letter alphabet, it is clear that the task of annotation will

involve clever procedures to recognize genes. While this appears a simple task once the grammar of nucleic acid sequences is clear and the punctuation marks defined, in reality the problem is formidable. The field of genome analysis today is the happy hunting ground of computer scientists eager to develop ever more efficient algorithms for decoding sequence data; even IBM has now directed its most powerful computer towards deciphering genome data, presumably a more useful activity than matching wits with Gary Kasparov or Vishwanathan Anand.

Since living cells rely very substantially on proteins to carry out all their housekeeping functions, it is natural to ask: 'how many distinct proteins do human beings have?' Since we do not have a definite answer, as yet, educated guesses abound in the literature. To set the problem in context it is useful to recall that the bacterium *Escherichia coli* has ~ 4000 genes, yeast ~ 6000 and the fruitfly, *Drosophila*, ~ 13,600. Gene numbers do not necessarily readily correlate with functional complexity; the worm *Caenorhabditis elegans* eclipses the fruitfly with a complement of ~18,400 genes. While the number of genes in humans may appear to be of interest primarily to biologists, it is worth noting that the shares of some biotechnology companies, on occasion, zoom upwards when gene estimates are raised and plummet when the number is lowered. Clearly, there is money to be made in genes; the larger the number, the greater are the chances for commercial exploitation in the post-genome era.

The present status of human gene counts is highlighted by a set of three reports that appeared in the same issue of *Nature Genetics* (Vol. 25, June 2000). Using an analysis of expressed sequence tags (ESTs), B. Ewing and P. Green (p. 232) obtain an estimate of 35,000 genes. Roest Crollius *et al.* (p. 235), using an approach based on comparative genomics employing the pufferfish genome, suggest that humans must possess ~28,000–34,000 genes. These estimates are considerably lower than the oft-quoted figures of 50,000–90,000, which have appeared in the literature over the years. For those whose hearts sink as gene estimates are lowered, there is still hope. F. Liang *et al.* (p. 239) from the Institute of Genomic Research use ESTs and come to the conclusion that the number of human genes is ~1,20,000. Undoubtedly, identifying genes definitively will provide the real answer, but for now we must live with delicious uncertainty. The tasks of eventually assigning functions to all the genes in genomes and monitoring their expression in different biological contexts will be major preoccupations in the future. The gene annotation area is already attracting many theoreticians and computer scientists, but may also be advanced by those whose computational skills are also backed by sound biochemical intuition. The sequences of chromosomes 21 and 22 are only the tip of the human genome iceberg. Already, there are signs that there are enormous non-coding regions which may have some bearing on our evolutionary past. On chromosome 21, there is a 7 million base pair segment that contains a single gene; this region is larger than the entire genome of *E. coli*, the bacterium with which we share much of our basic biochemistry.

The Indian contribution to the genome analysis program has been negligible. Even as the first step towards joining a Japanese-led consortium to sequence the rice genome has been taken, Monsanto has announced the first draft of this genome. It may, in fact, be worthwhile to worry if local efforts can contribute to understanding sequenced genomes, rather than compete head to head on important targets. The example of the Brazilian effort, which deciphered the 2.7 million base-pair genome of *Xylella fastidiosa*, a bacterial pathogen that causes citrus variegated chlorosis and threatens the citrus industry in São Paulo (*Science*, 2000, **288**, 800), may provide instructive lessons for Third World efforts in genome sequencing.

P. Balaram