Recounting a genetic story

Roger H. Reeves

The DNA sequence of human chromosome 21, now published, provides indications that the total number of human genes has been overestimated, and is a valuable resource for research into Down syndrome.

The sequencing of whole genomes means that genetic information is increasingly considered in functional rather than structural terms. But for the moment, the principal structural element of genomes — the chromosome — remains the predominant unit by which to measure progress in the Human Genome Project.

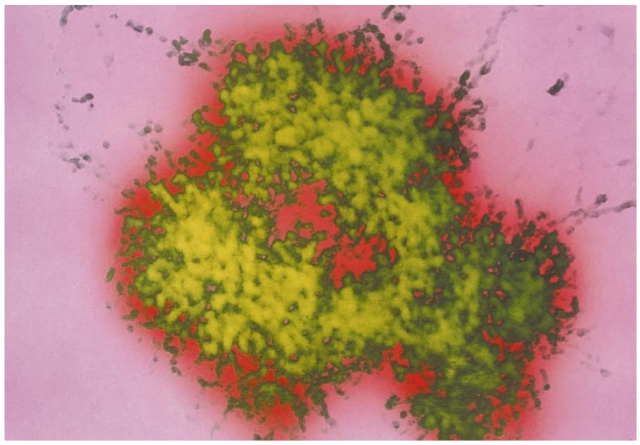
On page 311 of this issue[1], a multinational consortium reports the complete sequence of chromosome 21, the smallest human chromosome. The results indicate that previous estimates of the total number of human genes may need to be revised downwards. Meanwhile, the small number of genes, and a catalogue that identifies them, provide a boost for those endeavouring to define all of the primary molecular players in Down syndrome. Affecting one in 700 live births, Down syndrome occurs when three copies of chromosome 21 are inherited instead of two (Fig 1, top). The condition is the most common known genetic cause of mental retardation and the leading cause of congenital heart disease, and results in a wide variety of other developmental and health problems.

The consortium's report[1] describes several new technical achievements. The total length of the sequence reported is 33.55 million base pairs (or megabases, Mb). This covers 99.7% of the long arm of chromosome 21 (Fig 1, bottom), and just exceeds the 33.46 Mb reported for the slightly larger long arm of chromosome 22 (ref. 2). The paper includes the longest continuous DNA sequence reported to date, extending 28.5 Mb. The entire chromosome sequence has only three gaps (totalling 100 kilobases), compared with the ten gaps (totalling about 1 Mb) for the long arm of chromosome 22.

The new sequence also includes 281 kilobases from chromosome 21's short arm — mapping and cloning of which posed a challenge because it contains several classes of highly repetitive sequences[3]. The length of this short arm can vary greatly among individuals. So this sequence is the first example of a large genome region that can expand or contract on a scale of many megabases.

The sequencing of the long arm of chromosome 21 provides a somewhat arbitrary, but nonetheless worthwhile, basis for deriving conclusions about the general organization of the human genome. The most striking difference is the reduced gene content of chromosome 21 — 225 genes identified, compared with 545 on chromosome 22. The two consortia responsible for these sequences used somewhat different criteria to identify the genes within their respective chromosomes (Box 1, overleaf). But the differences may well balance each other out, meaning that a comparison of gene numbers is valid.

Chromosome 21 was expected to be relatively gene-poor, but it seems that it is even more impoverished than anticipated. The



CNRI/SCIENCE PHOTO LIBRARY

RICHARD J. GREEN/SCIENCE PHOTO LIBRARY

Figure 1 **Chromosome 21 in context. Top, triplication of chromosome 21 is the genetic defect underlying Down syndrome. Bottom, transmission electron micrograph of chromosome 21, showing the long and short arms.**

long arm of chromosome 21 represents about 1% of the human genome, but was predicted to contain less than 1% of the total number of human genes. The Unigene project[4] suggested that chromosome 21 would contain only 80% of the number of genes that would be expected on the basis of its size. If the total number of human genes were 100,000, as predicted, chromosome 21 would still be expected to contain 800–1,000 genes. The 225 genes now identified[1] stand in stark contrast to this prediction.

Combining data from the long arms of the two completely sequenced chromosomes, the chromosome 21 consortium estimates that the human genome may contain as few as 40,000 genes. However, this is based on complete sequences for just 2% of the human genome, and could be low for a variety of reasons. For example, other human chromosomes may be more gene-rich. The major histocompatibility complex (MHC) region on chromosome 6 — a region essential to the immune system — spans only 3.6 Mb, but contains 128 genes and 96 pseudogenes[5].

Another measure of gene richness is provided by the number of 'CpG islands' on the long arms of chromosomes 21 and 22. These islands are DNA sequences of a few hundred base pairs that have a high amount (more than 55%) of cytosine and guanine nucleotides. They are associated with about 60% of known human genes, and might be useful in gene identification. The two sequencing consortia[1,2] again applied different criteria to count CpG islands (Box 1), and these differences probably produce a total that is higher — by an unknown amount — for chromosome 22's long arm. Even so, chromosome 21 appears to be even poorer in CpG islands than in genes when compared with chromosome 22.

The chromosome 22 sequencing consortium suggested that its identification of 545 genes on the long arm was low — a conclusion based in part on the fact that 271 of the 553 identified CpG islands have not yet been associated with genes. In fact, nearly all of the 115 conservatively predicted CpG islands on chromosome 21's long arm are associated with genes. Analysis of both chromosomes using the same methods will help to determine the accuracy of identifying genes by counting CpG islands.

The chromosome 21 sequencing consortium also compared the chromosome 21 sequence with data in the available mouse genome database. No new conserved syntenies — regions where the same genes are 'linked' on chromosomes in different species — were identified. The previously known conserved syntenies are with mouse chromosomes 10, 16 and 17. The chromosome 21 consortium suggests that discrepancies in the gene order predicted by comparing the sequence to mouse gene-linkage maps may result from the differing resolution of these maps. In fact, the higher-resolution physical map[6] of mouse chromosome 10 shows that all 24 genes known to be shared between mouse chromosome 10 and human chromosome 21 occur in the same order. The high degree of conservation between human and mouse is important, because comparing the two sequences — as more of the mouse sequence becomes available — is likely to increase our ability to pick out genes and other significant features from the welter of sequence information.

The availability of the chromosome 21 sequence will have an immediate impact on the study of human single-gene disorders. For example, the genes responsible for five of those monogenic disorders that map to chromosome 21 — including two forms of deafness, Usher and Knobloch's syndromes — have not yet been identified. But having the complete sequence will obviate the labour-intensive step of identifying candidate genes. The genes responsible for these disorders are likely to be found rapidly.

But the greatest impact of the chromosome 21 gene catalogue will be in assessing the contributions of specific genes to traits seen in Down syndrome. The small number of genes on chromosome 21 is likely to be part of the reason why the presence of three copies of this chromosome — unlike so many chromosome defects — is not fatal at a very young age, or even before birth. Yet there are varying ideas about which genes are associated with particular features of Down syndrome, and the mechanisms by which an imbalance in the number of genes might produce the more than 80 physical and mental disorders that can be seen in this trisomy. Obtaining a comprehensive catalogue of the genes on chromosome 21 has been a goal of Down syndrome researchers for many years, and is realized in this landmark contribution. ■

Box 1 Finding genes in a DNA haystack

The two groups[1,2] involved in sequencing chromosomes 21 and 22 used a similar combination of methods to search the sequences for genes. But they set different criteria to arrive at the numbers of genes and CpG islands — regions of DNA with more than 55% of cytosine and guanine nucleotides, which often mark the 5′ ends of genes.

Sixty per cent of mammalian genes are reported to have a CpG island near their 5′ end. But the percentage of CpG islands associated with genes is unknown, and can only be determined by knowing the total number of genes. Chromosome 22's long arm[2] is reported to have 553 CpG islands. The corresponding number is not given for chromosome 21. Rather, the total of 115 CpG islands reported for chromosome 21 includes only the subset that are not associated with repetitive DNA elements. This lowers the CpG island total on chromosome 21 by an unknown amount.

The chromosome 21 consortium[1] used a conservative criterion for identifying genes amid DNA sequences. This criterion demanded that regions matching the short sequences representing the transcripts of genes should show evidence of having been spliced from multiple protein-coding gene portions (exons).

This would lower the predicted number of genes relative to the calculation for chromosome 22 (ref. 2), which included matches of these expressed sequence tags (ESTs) to single putative exons. Genes with large 3′ untranslated regions or large introns (non-coding parts of genes), or those represented in EST databases only by untranslated sequences, are likely to be excluded by the chromosome 21 criterion.

However, the estimate of 225 genes on the long arm of chromosome 21 includes those that are identified only by computer algorithms that can predict exons from a variety of sequence features (*in silico* prediction). Those with a gene-like structure identified by two or more algorithmic methods were added into the chromosome 21 gene total. Gene-prediction programs can give high false-positive rates of exon prediction, even in combination, and could inflate the gene number.

For chromosome 22, genes predicted only by algorithmic methods do not contribute to the total of 545 genes. But the analysis of chromosome 22 included a projection that was corrected for false positives resulting from *in silico* predictions. If genes predicted only by algorithm were included, the chromosome 22 total might rise by about 100. So the differences in the gene-identification strategies used by the two consortia will tend to cancel each other out. The degree to which this is true could be determined by re-analysis of each sequence using the other method. **R.H.R.**

Roger H. Reeves is in the Department of Physiology, The Johns Hopkins University School of Medicine, 725 North Wolfe Street, Baltimore, Maryland 21205-2105, USA.
e-mail: rreeves@welch.jhu.edu

1. The Chromosome 21 Mapping and Sequencing Consortium *Nature* **405,** 311–319 (2000).
2. Dunham, I. *et al. Nature* **402,** 489–495 (1999).
3. Wang, S. Y. *et al. Genome Res.* **9,** 1059–1073 (1999).
4. Deloukas, P. *et al. Science* **282,** 744–746 (1998).
5. The MHC Sequencing Consortium *Nature* **401,** 921–923 (1999).
6. Wiltshire, T. *et al. Genome Res.* **9,** 1214–1222 (1999).