nature genetics August

The EST express gathers speed

Thousands of human genes are being characterized by analysing complementary DNA sequences. But the job of making full use of this wealth of information is just beginning.

THE task of the human genome project — to identify and map all 50,000–100,000 human genes — seems herculean. Substantial headway is being made towards the first of those goals, however, and at an accelerating pace. A good ten per cent of the genes have already been cloned, sequenced (if only partially)[1] and published, due in large measure to the automated complementary DNA sequencing endeavours of several groups, among them Craig Venter's at The Institute for Genomic Research (TIGR). In the July and August issues of *Nature Genetics*, Venter and colleagues have followed up their two earlier reports describing almost 4,000 human brain cDNA sequences[2,3], or 'expressed sequence tags' (ESTs), with another 4,500 clones[4,5], reflecting an increase in the speed of data gathering and the benefits of subtle changes in sequencing strategy. The August issue also reveals progress in the vital tasks of mapping these sequences[6] and communicating this wealth of information[7].

Certain aspects of genome analysis seem suited to private industry, and automated DNA sequencing is one of them. TIGR is barely one year old, but is already generating sequence data at an unprecedented rate. Thirty Applied Biosystems sequencers (which come at $120,000 apiece) work around the clock to yield up to 1,000,000 nucleotides of sequence per week — about the total submitted to GenBank in the same time. The present output resembles a sort of glorified laundry list — nearly 200 different human cDNA libraries are being analysed and that number may eventually double — but such a compilation is a necessary prerequisite to the vastly more interesting biological analyses that such data will make possible.

Last month, Adams *et al.*[4] described 3,401 ESTs derived from 3,013 brain cDNA clones, including homologues of the very low density lipoprotein and epidermal growth factor receptors, and two ESTs bearing significant similarities to the human *KUP* and *Drosophila tramtrack* and *Broad-Complex* genes, which encode zinc-finger proteins. Another newly discovered zinc-finger gene, *LAZ-3*, bearing homology to these *Drosophila* transcription factors, is disrupted in lymphoma translocations involving chromosome 3 (ref. 8). Adams *et al.* also offer a glimpse of things to come with a 'computerized northern blot' — a comparison of the sequence profile from various tissues, such as brain versus liver. Structural and regulatory proteins are more abundant in brain, but secreted proteins and polypeptides involved in protein synthesis predominate in liver.

In this month's issue, Adams *et al.*[5] describe how they found almost the same number of homologies while analysing only half as many clones. The trick is to use a directionally cloned brain cDNA library, which optimizes the amount of coding information obtained from an EST (and hence the likelihood of detecting a significant homology) and also minimizes the chances of contamination or of generating chimaeras. Adams *et al.* identified clones of a second amyloid precursor-like protein (APLP), a cousin of the amyloid precursor protein so central to Alzheimer's disease pathology. The complete characterization of this APLP2 cDNA isolated by more traditional means will soon be reported[9].

The impact of this strategy is considerable: including a wealth of unpublished data, Venter estimates that some 25,000 human gene sequences have been identified, the vast majority ESTs. As recent findings from large-scale genomic sequencing projects point to there being a likely total of about 75,000 human genes[10], perhaps as many as one-third of human genes have already been isolated. But lest researchers be discouraged that the more laborious pursuit of single genes is now redundant, they may take heart from the fact that Venter's group has yet to spot the recently discovered Huntington's gene among their brain cDNAs.

But as these EST compilations grow, the geneticist may ask what use they serve if there is no cohesive attempt to map them onto the genome. Some groups are rising to the challenge. Polymeropolous *et al.*[6] have mapped more than 300 brain ESTs to whole chromosomes using the polymerase chain reaction on somatic cell hybrids. Their distribution, however, does not fully correspond with the cytogenetic length of the chromosomes — for example, chromosomes 2 and 19 contain similar numbers of genes despite their large difference in size. And chromosomes 13 and 18 were relatively under-represented in mapped genes, suggesting that the only viable human trisomies are those of chromosomes 13, 18 and 21, due to their paucity of genes. Some researchers are beginning to localize ESTs to sub-chromosomal regions, especially on the X chromosome. But even though this huge resource begs to be exploited, there is still no systematic undertaking to map ESTs. Even if it takes several years to compile the precise localization of thousands of ESTs, such knowledge would change the face of medical genetics: mapping rare disease traits, rather than isolating the corresponding genes, would become the rate determining factor.

Aside from accelerating the pace of gene discovery, the EST bandwagon holds great promise for work with other organisms[1] (including *Caenorhabditis elegans* and *Saccharomyces cerevisiae*) and in other disciplines (such as the study of changes in gene expression associated with cell division and tumorigenesis). The need for direct access to EST data will hasten the day when biologists embrace computer databases as a regular part of their research. Such a database, called dbEST, is now available[7]; as well as being a repository for EST data, dbEST periodically conducts homology searches against new entries in nucleotide and protein databases. That still leaves the unenviable task of deciding how long to devote sizeable portions of journals (not the CD-ROM variety) to expanding lists of ESTs. But then the life of an editor is never easy.

Kevin Davies

Kevin Davies is Editor of Nature Genetics.

Also in this month's *Nature Genetics*: three reports on the correlation between repeat size and age of onset in Huntington's disease; a single gene for Hirschsprung disease located on chromosome 10; the first missense mutation identified in dystrophin; verifying the accuracy of chromosome physical maps; and mapping the ovine fecundity gene.

1. Sikela, J.M. & Auffray, C. *Nature Genet.* **3**, 189–191 (1993).
2. Adams, M.D. *et al. Science* **252**, 1651–1656 (1991).
3. Adams, M.D. *et al. Nature* **355**, 632–634 (1992).
4. Adams, M.D., Kerlavage, A.R., Fields, C. & Venter, J.C. *Nature Genet.* **4**, 256–267 (1993).
5. Adams, M.D. *et al. Nature Genet.* **4**, 373–380 (1993).
6. Polymeropolous, M.H. *et al. Nature Genet.* **4**, 381–386 (1993).
7. Boguski, M.S., Lowe, T.M.J. & Tolstoshev, C.M. *Nature Genet.* **4**, 331–332 (1993).
8. Kerckaert, J.P. *et al. Nature Genet.* (in the press).
9. Wasco, W. *et al. Nature Genet.* (in the press).
10. Martin-Gellardo, A. *et al. Nature Genet.* **1**, 34–39 (1992).