Rapid identification of new transcribed human sequences: a semi-ordered approach

Andrew Martin Dearlove

RAPID IDENTIFICATION OF NEW TRANSCRIBED HUMAN SEQUENCES: A SEMI-ORDERED APPROACH

Andrew Martin Dearlove

MRC UK Human Genome Mapping Project Resource Centre

Hinxton Hall, Hinxton, Cambridge, CB10 1RQ

A thesis submitted for an M.Sc. at the University of Durham

June 1995

RAPID IDENTIFICATION OF NEW TRANSCRIBED HUMAN SEQUENCES: A SEMI-ORDERED APPROACH

Andrew Martin Dearlove

MRC UK Human Genome Mapping Project Resource Centre

A thesis submitted for an M.Sc. at the University of Durham, June 1995

The sequencing of the DNA that is actually expressed and turned into protein has many applications to facilitate the discovery of new human genes, mapping of the human genome, and identification of coding regions in genomic sequences.

The redundancy encountered in systematic sequencing programmes necessitates the production of semi-ordered fragmentary cDNA libraries using various technologies, to subsequently generate high quality expressed sequence tags (ESTs). Modified standard protocols and new ideas and methodologies were utilised.

1. Intact and biologically active pure RNA was extracted from various fetal tissues using further optimised purification methods including ultracentrifugation. Standard commercially available kits were used for mRNA purification and cDNA synthesis.
2. A variation of the pBluescript KS II (+/-) vector was produced to subsequently clone the prepared restriction fragment sorted cDNA species.
3. Streptavidin bead capture of PCR products, and new methods involving the use of helper phage, filtration, and lacZ magnetic bead purification, were used to produce ssDNA templates for standard automated sequencing.

Over 75% of the cDNA fragment library clones contained inserts (average 500 bp). The EST data for 1167 clones (1552 sequences) showed the sorting technique subdivided the complex cDNA mixtures into distinct populations: sequences produced 636, 54% different sequence classes; and 523, 44% of the sequences were seen only once in this study. Database match searches identified 152 (13%) unknown ESTs, and 96 (8%) unique unknowns from the six different populations. No known intron or rRNA sequences, and only a small number of mitochondrial sequences were generated: suggesting the quality and integrity of the purified RNA was high.

Combining such semi-ordered populations from different tissues significantly increases the complexity of a cDNA library: this strategy rapidly identifies new transcribed sequences and genes, to forward the Human Genome Project; aiming to map and sequence all of the human genome.

ACKNOWLEDGEMENTS

I thank my supervisor Ross Sibson[1] for providing intellectual vision, some of the strategies used and copious advice throughout these experiments. Much guidance and support was also gratefully received from Mike Starkey[2] and Jacqueline Gross[3]. Thanks and appreciation are also due to other members of staff at the UK Human Genome Mapping Project Resource Centre, including Tony Vickers[4], Keith Gibson[2], Chris Mundy[2] and David Howells[5]. Assistance with data analysis was given by Computing Services including Claude Discala[6] and Yagnesh Umrania[2]. Various databases, sequence editing and analysis programs utilised were originated at the Resource Centre by Gary Williams[2] and Jeremy Parsons[7].

1 Formally UK HGMP Resource Centre, now Clatterbridge Cancer Research Trust, Clatterbridge Hospital, Bebington, Wirral, L63 4JY.

2 UK HGMP Resource Centre, Hinxton Hall, Hinxton, Cambridge, CB10 1RQ.

3 Formally UK HGMP Resource Centre, now Muscle Cell Biology Group, Clinical Sciences Centre, Royal Postgraduate Medical School, Du Cane Road, London, W12 0NN.

4 Formally UK HGMP Resource Centre.

5 Formally UK HGMP Resource Centre, now Applied Biosystems Ltd., a Division of The Perkin-Elmer Corporation, Kelvin Close, Birchwood Science Park North, Warrington, Cheshire, WA3 7PB.

6 Formally UK HGMP Resource Centre, now Généthon, 1, rue de l'internationale, 91002 Evry cédex, France.

7 Formally UK HGMP Resource Centre, now Department of Genetics, University of Washington School of Medicine, St. Louis, Missouri, USA.

CONTENTS

With 30 figures, 20 plates and 10 tables

Lists of figures, plates and tables

1. INTRODUCTION

1.1. The Human Genome Project

The Human Genome Project (HGP), mapping and sequencing the entire human genome, is perhaps the most exciting scientific event since the space race of the late 1960s, for man to walk on the moon. Around the world many hundreds of laboratories are analysing the genetic message encoded in DNA towards locating every gene, and deciphering all the instructions which specify a human being, within the next decade.

The race is primarily against time to get a complete understanding of the role and origins of all genomic components, and the prospect of being able to disentangle the complex causes of genetic disease. The HGP is a unique initiative in biology in terms of the scale: sequencing the 3000 Mb human genome is a huge task, and therefore much international co-ordination is required. From the concept of the HGP and its early development in the USA it was perceived that identification of all the genes involved in diseases including cancer would be necessary to be able to understand and combat them, although knowledge of the entire genome sequence would be even more useful.

The HGP was supported by the US Department of Energy (DOE) and the National Institutes of Health (NIH), and was promoted by the setting up of the Human Genome Organisation (HUGO) to influence and co-ordinate international contributions. Major progress depended on the co-operation and pooling of resources and the free exchange of data, with roles of co-ordination taken by the large scientific centres.

The goals of the HGP have evolved with much debate as to the priorities and strategies to be adopted, with the ultimate goal the determination of the nucleotide sequence of the human genome and the genome of model organisms. The first priorities are the development of improved technologies for mapping and sequencing, and informatics systems for handling the vast amount of data that will be generated over the course of the project.

Mapping the human genome provides a linear representation to describe the organisation of a set of landmarks, using a defined system of measurement based on co-ordinates. Physical maps can be cytogenetic, ordering loci by *in situ* hybridisation based on their relative position on the chromosome; or molecular, using sequence features as the landmarks. Genetic maps represent the genetic distance between markers: linkage maps base distance on centiMorgan (cM) units, a measure of the frequency of recombination

between the specific genes. Genetic distance does not have a constant relationship with physical distance, because recombination frequencies vary between different regions of the genome. The initial phase of the HGP is focusing on establishing such physical (Cohen *et al.*, 1993) and genetic maps (Weissenbach *et al.*, 1992, NIH/CEPH Collaborative mapping group, 1992, and Gyapay *et al.*, 1994), of landmarks such as polymorphisms, genes and specific DNA sequences.

One strategy used in the HGP starts with the intact genome to produce long range maps from the maps of the DNA markers to generate cloned regions to be sequenced. The opposite approach is to start with the sequence from random cosmid or bacteriophage clones, and proceed to assemble longer range sequences by identifying overlapping sets of cloned sequences (contigs).

Screening large libraries is tedious and time consuming, although multiplexing and simultaneous analysis techniques are amenable to automation. With the recent development of new vectors screening became easier, and the resolution obtained by genetic mapping and the maximum length of contig sequence that can be assembled became more compatible. Cloning technology had been limited by small cosmids that were able to carry exogenous DNA inserts of less than 50 kb, which was not big enough to contain genes from higher organisms (Little, 1992). YAC vectors which are capable of carrying much larger inserts of about 1 Mb, equivalent to about 1 cM on a linkage map was a significant development (Anand, 1992). They are not a great panacea though as the YAC libraries are often not ideal: showing high levels of chimaerism, there is also still a need to subclone into other vectors for sequencing.

Although great advances have been made, sequencing is still a tedious and slow process: further increased automation (robotics) and development of alternative technologies or chemistries are required to dramatically improve the present speed and capacity in order to finish the complete sequence of the human genome to schedule by 2005. Until such radically different approaches that would make todays methods look primitive are introduced to reduce the scale of the project to a more manageable task, the question of what to sequence is paramount.

The entire genome (100 Mb, 10 000-20 000 genes) of a much simpler organism, the microscopic nematode worm *Caenorhabditis elegans* is currently being sequenced (Sulston *et al.*, 1992). The development and lineage of all of the worms 945 cells have been tracked, so the biological purpose of the 90% "junk" intron and intergenic DNA may become clear. Other model genomes such as *Drosophila melanogaster* (fruit fly),

Arabidopsis thaliana (a simple plant) and *Saccharomyces cerevisiae* (yeast) are also being studied (Hood *et al.*, 1992). The human genome is much greater in size (3000 Mb) than that of these model organisms being sequenced, and so it makes sense to target those regions which will yield information of greatest immediate value. Indeed it is difficult to justify large scale blanket sequencing of the human genome at current rates and costs, when it could be done much quicker and cheaper in the near future.

1.2. The value of cDNA sequencing and ESTs

It was Sydney Brenner and Ed Southern, dominant figures in British molecular biology, who advised the UK Medical Research Council (MRC) that it would be expedient with their relatively modest resources relative to the Americans, to concentrate on analysing the DNA that is active: the DNA that is actually expressed and turned into protein. It is more cost-effective to turn to this complementary DNA or cDNA, instead of sorting through some 90 per cent of the "junk" DNA of the human genome. When a cell makes a protein, it first transcribes the relevant gene into messenger RNA, which carries the genetic information out of the nucleus to where the protein is made in the cytoplasm. The RNA is spliced, to remove the "junk" introns before protein synthesis. The cDNA is made from extracted mRNA using the enzyme reverse transcriptase to obtain only the "useful" sequence that encodes for protein.

The main problem is one of sorting through the many types of mRNA present, accepting the ones desired may be less abundant than others. The efficient sequencing of human (*Homo sapiens*) cDNA clones to generate new expressed sequence tags (ESTs) with the minimal amount of redundancy, has therefore to be concentrated upon. An EST is simply part of a sequence from a cDNA that corresponds to a mRNA, and is hence by definition a transcribed sequence as sought after in this study. They have many applications including the discovery of new human genes, mapping of the human genome, and identification of coding regions in genomic sequences.

1.3. Progress made using cDNAs and ESTs

In the early 1990s when less than 200 human brain mRNAs were listed in GenBank, a pilot project was initiated in the USA to test the use of ESTs in a comprehensive study of expressed genes (Adams *et al.*, 1991). Single-run sequence data was obtained for over 600 randomly chosen cDNA clones from various human brain libraries. Double-stranded cDNA clones in pBluescript vector were sequenced by primer (M13 universal or reverse) cycle sequencing. Double-stranded templates were produced by plasmid preparation (497), alkali lysis and Qiagen columns, or by the polymerase chain reaction (PCR), which exponentially copies and amplifies specific nucleotide sequences between two convergent primers that hybridise to opposite strands (117). About 60%, 90%, and 50% respectively, of templates produced by each method generated data. ESTs were examined for similarities in the GenBank nucleic acid database using BLAST (Altschul *et al.*, 1990). Results showed 197 (32%) of the sequences matched to human sequences, repetitive elements, mitochondrial genes, ribosomal RNA genes and other nuclear genes. Matches to non human entries were found in 48 (8%) of the sequences, whilst 230 (38%) had no significant matches. No inserts were found in the other 134 (22%) of sequences.

It was argued ESTs can serve the same purpose as random genomic DNA sequence-tagged-sites (STS) for the physical mapping of the human genome, whilst providing the additional feature of pointing directly to an expressed gene. PCR was used to screen a series of somatic cell hybrid cell lines containing defined human chromosomes or chromosome fragments for the presence of a given EST. Only hybrids containing the human gene corresponding to the EST yielded an amplified fragment, hence assigning the EST to a chromosome by analysis of the distribution of PCR products from hybrid DNA templates. It was found ESTs longer than 150 bp were most useful for mapping as well as the similarity searches: forty six ESTs were mapped to chromosomes.

Sequence data was shown to be more accurate nearer the primer, decreasing rapidly beyond 400 bases. Average accuracy obtained for the double stranded sequencing runs was about 98% for up to 400 bases. ESTs submitted to GenBank (348: M61953-M62300) were over 150 bases in length with <3% ambiguity. Overall sequence accuracy was over 97% based on matches to known human genes.

Single run DNA sequencing was shown to be an efficient method of obtaining preliminary data on cDNA clones: enough information was obtained in 150-400 bases of

a nucleotide sequence for early identification of the cDNA and localisation to a chromosome. It was shown that new sequences (230) representing previously uncharacterised genes could rapidly be identified. The random selection methods used revealed a high level of redundancy in the cDNA libraries, emphasising the importance of library pre-processing or normalisation techniques as the number of ESTs increases.
A basis for the use of ESTs had been provided, demonstrating their utilisation to fast cDNA characterisation to help economic tagging of human genes, with the provision of genetic markers and a genome resource.

The value of cDNA sequencing alongside systematic genomic sequencing was widely appreciated (Bergmann, 1992). The potential for rapid identification of unidentified genes particularly with regard to *C. elegans* was reviewed favourably with caution: suggesting that not all new cDNAs represent new proteins and genes, and that as databases grow, greater redundancy will be encountered.
A representative *C. elegans* library, sorted by the use of negative screening, of 1517 clones was investigated using single sequence reads (Waterston *et al.,* 1992). A resultant efficient identification of about 1200 of the estimated 15 000 genes was obtained. Over 30% of protein sequences obtained matched those already in databases, providing an *in vivo* analysis of known genes. YACs were used to map 670 of the cDNAs. The *C. elegans* research was focused on genomic sequencing to investigate the purpose of all DNA as mentioned previously.
Gene families and potential disease gene homologies between humans and *C. elegans* were identified using ESTs from a mixed stage *C. elegans* library (McCombie *et al.*, 1992). A large number (422) of new genes were again identified from only 720 ESTs, of which 26 were mapped.

Large scale sequencing of cDNA using PCR amplification was used to analyse the quantitative and qualitative aspects of gene expression, (Okubo *et al.,* 1991 & 1992). A cDNA library containing *Mbo* I cut (providing a constant reference point) 3' end fragments was constructed from the liver cancer cell line HepG2. The cDNA moieties of 1022 randomly chosen colonies were amplified by PCR and sequenced. The average size of inserts was 270 bp, but inserts of less than 20 bp were found in some clones: the sequence similarity of the remaining 982 clones was compared with each other and GenBank. Over half (514, 51%) of the sequences, representing 173 different species,

appeared more than once. The solitary sequence group consisted of 468 sequences. Many of the redundant group members (55 of the 514) consisted of only 3 species, which were identified by GenBank matches. Of the remainder of the redundant sequence group members, 107 (35 species) were identified and 352 (135 species) were novel. Of the solitary group members, 53 were identified, and 415 were novel. Hence in total from the 982 clones producing informative sequence 215 represented 91 known genes and 767 clones represented 550 novel genes. It was shown there was a higher probability of databank matches being found with redundant sequences than solitary sequences: the overall chance of database matches being found with the 3' end fragments used was 22%. The idea of "body mapping expressed genes" provided an efficient method of gene collection concentrating on sequencing small numbers (~1000) of cDNAs from each of the 200 or so different tissues of the human body, rather than normalisation or complexity of library.

Single pass sequencing of 1037 human brain clones of a library pre-screened with total brain cDNA yielded 1024 ESTs (Khan *et al.*, 1992): over 900 represented previously unidentified genes. A small subset (20) were physically mapped, assigning them to chromosomes by PCR amplification of human-rodent somatic cell hybrids using primers designed from the ESTs and analysis of the products. Hence ESTs were shown to potentially represent a rich source of informative gene associated polymorphic genetic markers.

It was reported in early 1993 (Sikela & Auffray, 1993) that the merits of large scale cDNA analysis and the role of ESTs in the systematic cataloguing of genes faster than ever was becoming generally appreciated. It was evident from the European Community Symposium on Strategies in cDNA Programs (Paris, France, October 1992), that rapid development and expansion had occurred. Over 25 000 ESTs had by now been generated, with over 80% derived from human sources representing over 10000 human genes. Large scale cDNA approaches were being applied to many areas of research, with strategies adapted to suit particular interests. Tissues from many sources were being used in different programs. Non-normalised libraries gave information on the profile of the relative expression levels of active genes, whilst other pre-screened or normalised libraries enriching for less common clones resulted in less redundancy. Clones were being sequenced from both the 5' end (more likely to yield protein

information) and the 3' end (providing a constant reference point for comparison). PCR or hybridisation based strategies were being used to map the ESTs and corresponding genes, adding further value to them. Obviously further co-ordination of future cDNA projects was deemed important to reduce redundancy and optimise output.

The partial sequence data from over 8591 human brain cDNAs was detailed by Adams *et al.*, (1993), with a further 3401 ESTs from 3013 clones being described. About 6000 distinct genes expressed in the brain were by now represented by ESTs, with over 700 different proteins identified. A further study showed advantages of a good quality directionally cloned human infant brain cDNA library (Adams *et al.*, 1993). The average length of these 1633 ESTs was 364 bp, and 37% of them were identified based on matches to over 320 different genes in the public databases. Construction of this directionally cloned library with size selection and improved purification methods of RNA led to a high coding content of the inserts of average length 1500 bp and less generation of non usable sequences.

The chromosomal distribution of 320 brain expressed genes was determined by studying the PCR products of human-rodent somatic cell hybrids (presence of human-specific product), and by genetically mapping polymorphic cDNAs using the CEPH pedigrees and database (Polymeropoulos *et al.*, 1993). Thence ESTs were being used further as gene markers on the physical map of the human gene providing a transcription map, suggesting a much greater future role for the polymorphic cDNAs in developing a high resolution genetic map. Correlation between areas of high GC content and gene density were noted, as was the non random distribution of genes.

Analysis of EST data, by matching to GenBank sequences and assuming 50% EST redundancy, was also used to get a more accurate estimate of the number of human genes (Fields *et al.*, 1994). The figure of 60 000-70 000 was consistent with others reached using genomic sequencing and CpG island data (Antequera & Bird, 1993): so was probably the most reliable to date. CpG islands are short stretches of DNA rich in unmethylated CpG, often marking the position of a gene, (Cross *et al.*, 1994).

1.4. Advances in technology

DNA sequencing was first made possible with the discovery of restriction enzymes and polymerases of the 1970s. Further developments included primed synthesis and gel electrophoresis, but a breakthrough in the rate of sequencing was made with the advent of the dideoxy chain termination reaction (Sanger *et al.*, 1977) and chemical degradation (Maxam & Gilbert, 1977) methods. Sequencing efficiencies were still relatively low though (1.5 kb per person year) due to difficulties in obtaining primers, single stranded templates cloned DNA and inadequate informatics. In the 1980s the introduction of oligonucleotide synthesisers, M13, plasmids, and computer assisted handling (Staden, 1982) helped in these areas. By the mid eighties sequencing reaction conditions had been adequately optimised, and other strategies initiated including PCR (Erlich, 1989) sequencing. Automated fluorescent sequencing (coupling automatic gel electrophoresis, raw data acquisition and base calling), and robotics were also in development.

Throughout the last decade there has been a move towards full automation with the objective of achieving 150 kb per person year, at a fraction of the cost. Dye primer or terminator chemistry is used with high quality templates, with a four label, single lane separation, 36 channel automated sequencer (Applied Biosystems 373A), often the preferred option (i.e. Adams *et al.*, 1991): although other hardware is available (Pharmacia A.L.F.).

The latest developments in autosequencers, data management and robotics enabled large throughput and analysis (i.e. Okubo *et al.*, 1991 & 1992): and the development of automation to facilitate high throughput DNA sequencing was now perceived a priority of many laboratories (Watson *et al.*, 1993). Both the production of good quality DNA templates and the sequencing process had been addressed with the objective of eventual integration. Developments had been made with regard to plaque or colony picking, solid phase recapture of purified DNA after manual centrifugation, and template sequencing using advanced liquid handling robots and thermal cyclers.

A model for high-throughput automated DNA sequencing and analysis core facilities, as set up at The Institute for Genomic Research (TIGR) (Adams *et al.*, 1994), was suggested. This system was based on automated sample preparation, template sequencing, and automated sequencers for data collection. Expansion had been achieved by evolution in DNA template production methods, sequencing chemistry,

instrumentation, data tracking and software for sequence analysis. More sequence data collection was possible by increasing samples (channels/lanes) per gel, with a shorter electrophoresis time and longer read length. Much manual intervention was still required for plate preparation, clone picking, actual running of robots, data editing and quality control. Integration of the data analysis and high capacity informatics was important for efficient feedback with regard to the earlier strategies. Balance in all areas of the project was required to optimise developments in each area and keep the efficiency as high as possible throughout.

Combining this automated sequencing technology with randomly selected cDNA clones representing mRNA distributions allows construction of a detailed picture of transcriptional activity of a cell or tissue, including the identification of transcribed genes, abundance levels and degree of overlap in gene expression amongst various tissues. Partial cDNA sequences (ESTs) provide the fastest way to obtain gene specific data for a large number of independent cDNA clones. Careful library design and sequencing strategies maximises the amount of information that can be derived from a particular mRNA population. Library source and method of construction are considered to produce a good library where every clone sequenced has potential to be an interesting new gene. Sequencing redundancy can be reduced by screening with total cDNA (eliminating abundant clones) or subtraction/normalisation techniques (Soares *et al.*, 1994). Extensive facilities for accumulation and analysis of the large amount of data produced are required to optimise the use of the partial gene sequences. Single pass sequencing is often used to maximise throughput to get the largest amount of information per raw base sequenced. This is accurate enough for sensitive similarity searches, analysing a large number of independent clones to identify different gene families quicker. It is obviously better to obtain a longer sequence to separate sequences from different gene families into their respective different genes if applicable: this can be achieved using the current technology by sequencing the cDNA from both ends to produce a longer good quality composite sequence.

1.5. The cDNA programme at the UK HGMP Resource Centre

The cDNA sequencing program was initiated at the MRC funded UK Human Genome Mapping Project Resource Centre located at Northwick Park Hospital, Harrow, in the autumn of 1990. cDNAs selected at random from commercial oligi-dT primed, nominally full length, libraries were systematically sequenced using A.L.F. (Pharmacia) and the now obsolete Genesis (Dupont) machines, generating over 1050 usable ESTs as seen below:

Library code	Tissue/cell type	No of usable sequences
H	Brain cerebral cortex (adult)	454
B	Liver (adult)	283
T	Rhabdomyosarcoma	88
V	Placenta	69
Y	Placenta	66
O	Bone marrow (adult)	61
X	Liver	28
W	Liver	20

The bias of these libraries towards abundant mRNAs, and 3' end sequences, as reflected in the redundancy encountered during sequencing, necessitated the construction of libraries more likely to be representative of the mRNA populations from which they were derived.

The "restriction fragment sorting" of complex cDNA populations was devised during 1991 as a means of increasing the number of mRNAs accessible via a given cDNA library. The procedure is based upon the partitioning of cohesive-ended cDNA restriction fragments according to the sequence of their ends, by successive base-specific adaptoring and base-specific PCR.

Fetal organ cDNA "libraries" totalling about 1600 clones (between 1-27 of a possible 256 "subsets" represented in each library) were generated by restriction fragment sorting, from brain, adrenal, kidney, liver, and two different organ mixture mRNA sources. Analysis of some fragmentary sequences produced at this time, selected at random from a small number of different "subsets", strongly suggested that the technique yielded distinct subpopulations of cDNA fragments.

Library code	Tissue/cell type	No of usable sequences
P	Whole brain (fetal)	410
S	Adrenal (fetal)	302
JG	Kidney (fetal)	26

A cDNA database, additionally containing sequences from other EC cDNA consortium laboratories, was established in 1992 using Sybase DBMS (originator Williams, G.). The facilities of the software included the detection and removal of spurious sequences like vector or adaptor; and the comparison of sequences with other nucleic acid sequences and protein sequences (in each possible reading frame) in the database, as well as those in EMBL and GenBank public databases. Further sophisticated software, namely ICATOOLS (originator Parsons, J.), was also produced to classify sequences into families.

About 4000 sequences had been produced by the EC cDNA consortium including the Resource Centre (Vickers, T.), Généthon in Paris (Auffray, C.), Genzentrum in Munich (Domdey, H. & Arnold, G.) and the Istuto di Biologia Cellulare in Rome (Tocchini-Valentini, G.) by mid 1992. Some 1435 ESTs generated at the Resource Centre were submitted to the EMBL database at the end of the year.

In order to make the fragmentary fetal organ cDNA clones accessible to a large number of laboratories, 15 024 *E. coli* clones (representing three fetal organ "libraries") were gridded at high density in 4x4 arrays on ten membrane filters using a Beckman Biomek automated workstation, facilitating hybridisation screening. Filter sets were dispatched to a total of 46 centres, and more than 750 individual clones have been issued to date.

This study details further work mainly between 1992 and 1994 incorporating some of the established "in house" protocols: endeavouring to add modifications to improve the quality and efficiency where possible, the development of the available technologies, and new ideas and methodologies. A second generation of fragmentary fetal organ cDNA "libraries" was made and some of the clones sequenced.

1. Intact and biologically active pure RNA free of chromosomal and mitochondrial DNA was fastidiously extracted from various fetal tissues using further optimised purification methods including ultracentrifugation. Standard methods and commercially available kits were used for mRNA purification and cDNA synthesis.

2. A variation of the pBluescript KS II (+/-) vector was produced to subsequently clone the prepared cDNA species which had been semi-ordered by restriction fragment sorting the cDNA population.
3. Streptavidin bead capture of PCR products, and new methods involving the use of helper phage, filtration, and lacZ magnetic bead purification, were used to produce ssDNA templates for standard automated fluorescent *Taq* cycle sequencing using ABI 373A DNA sequencers.

Stringent screening of the sequencing data of the "second generation" fragmentary fetal organ cDNAs means that only ESTs of defined quality (<4% ambiguities between bases 50-400) were processed for inclusion in the cDNA database and analysed further.

Simultaneously in what is likely to be part of a significant move towards the functional analysis of ESTs, the Resource Centre more recently independently and through collaborative projects, investigated the feasibility of analysing transcription patterns via the non-radioactive *in situ* tissue and cellular localisation of cDNA probes. Indeed the cDNA fragments produced were compatible with producing such probes to investigate the biological significance, purpose and location of each sequenced cDNA which adds great value to them; putting the Resource Centre in a strong position compared to other institutions in development in this area.

As a complement to the fragmentary cDNA libraries, 50 "full length" cDNA libraries were imported for further distribution of 100 ng ligation mixture aliquots for transformation. Functional cDNAs encoding cell-surface glycoproteins have been isolated from these libraries by expression screening, confirming the presence of full length cDNAs necessary for encoding properly folded and processed proteins. The identification of well characterised full length libraries, of established value, which would increase the repertoire of full length cDNA libraries available to the scientific community is in progress.

Over 15 000 PCR amplified mixed fetal organ cDNA fragments (representing 19 subsets) were produced in order to generate more gridded filters.

The Resource Centre relocated during 1994 to a new Genome campus also incorporating the Sanger Centre and European Bioinformatics Institute (EBI), in Hinxton, just outside Cambridge.

1.6. Overview of methods utilised

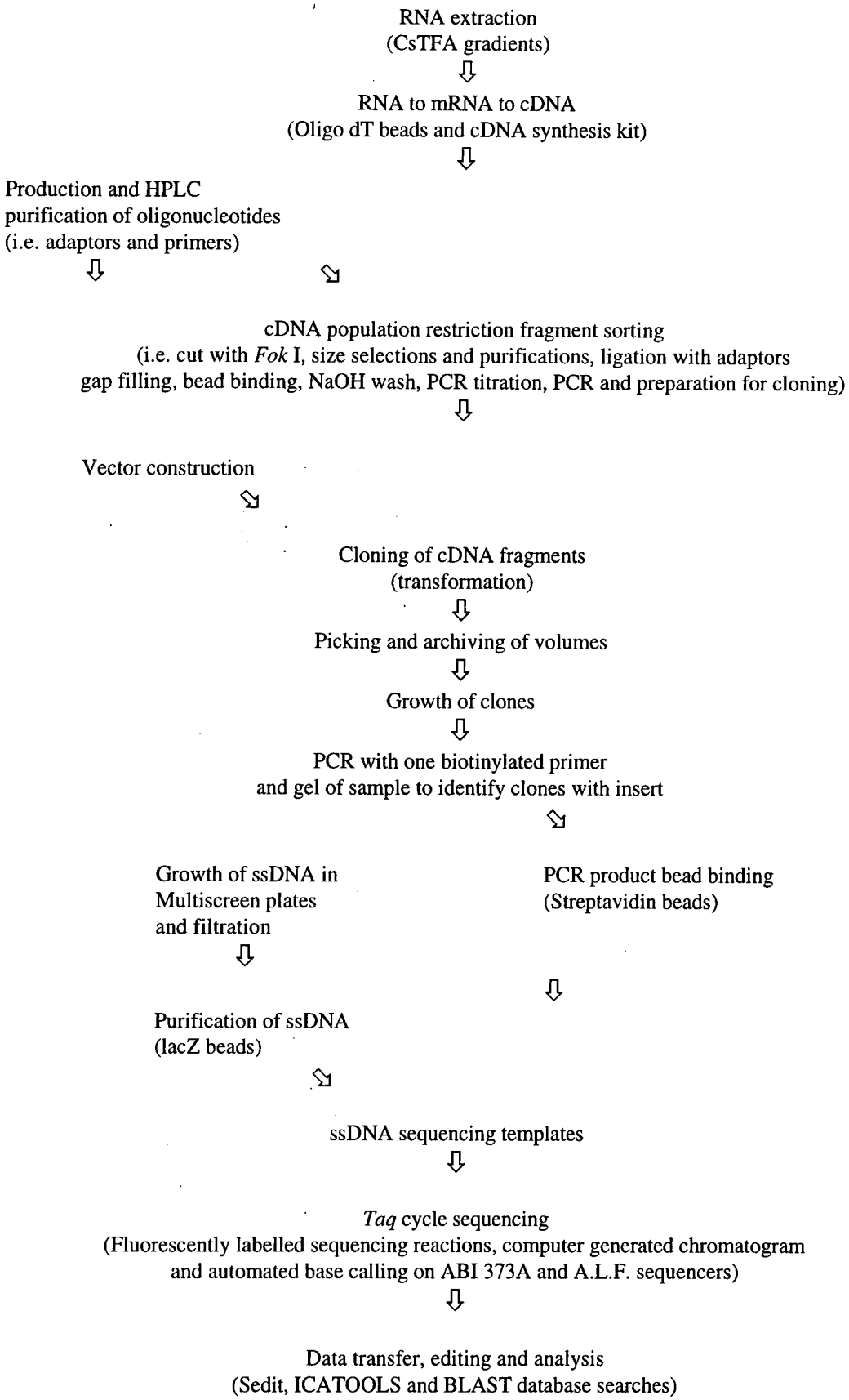


Figure 1. The rapid identification of new transcribed human sequences.

1.6.1. RNA extraction

Only genes that are being expressed are transcribed into messenger RNA (mRNA); so if this is used as starting material for a genomic library, the clones produced comprise only a selection of the number of genes in the cell, but all are actually being expressed in that cell type. Messenger RNA can not be ligated into a cloning vector, but it can be converted into DNA that can be, by complementary DNA (cDNA) synthesis utilising the properties of various enzymes.

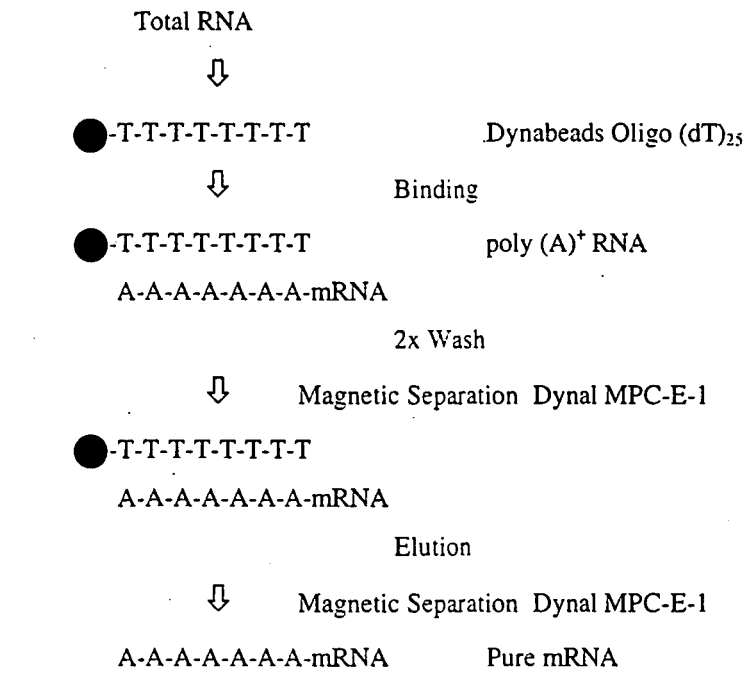
The isolation of pure good quality RNA from tissue is a prerequisite for cDNA synthesis. The main problem during the extraction procedure is protecting the RNA from degradation by ribonucleases, which are widely prevalent in the tissues and laboratory. The inclusion of the chaotropic salt guanidinium thiocyanate (GTC) in the primary extraction buffer provides a high degree of protection from RNase activity. RNA extracted in the GTC buffer can be purified further by equilibrium centrifugation in caesium chloride (CsCl), during which it becomes pelleted at the bottom of the centrifuge tube. In a recent improvement to this approach, CsCl was replaced with caesium trifluoroacetate (CsTFA) (Okayama *et al.*, 1987), this is more effective in inhibiting RNase activity and deproteinising nucleic acids (Carter *et al.*, 1983). This improved procedure with further modifications has been utilised.

The tissue of interest is homogenised in Extraction Buffer containing GTC. Cellular debris is removed by centrifugation. The extract is loaded in a bed of higher density CsTFA solution and centrifuged for over 36 hours. This leads to the formation of a density gradient, with protein floating on top ($\rho < 1.25$ g/ml), DNA banded within the gradient ($\rho \cong 1.52$ g/ml), and the RNA banded at a lower level within the gradient ($\rho \cong 1.62$ g/ml). Following fractionation the RNA is loaded onto an identical gradient and centrifuged again for over 16 hours. RNA is recovered by standard ethanol precipitation. Selective precipitation of the RNA with 3M lithium chloride (LiCl) is then utilised for further purification. Standard phenol/chloroform and chloroform extractions are carried out before final ethanol precipitation of the total RNA.

1.6.2. mRNA purification and cDNA synthesis

The use of the Dynabeads Oligo $(dT)_{25}$ mRNA Purification Kit is based on the unique Dynabeads magnetic separation technology, and relies on base pairing between the poly $(A)^+$ residues at the 3' end of most messenger RNA and the oligo (dT) residues covalently coupled to the surface of the Dynabeads Oligo $(dT)_{25}$. Other RNA species lacking a poly A segment (poly $(A)^-$ RNA), mainly rRNA and tRNA, do not form hydrogen bonds with the Dynabeads Oligo $(dT)_{25}$ and are readily washed off.

Purification of mRNA from total RNA



cDNA synthesis

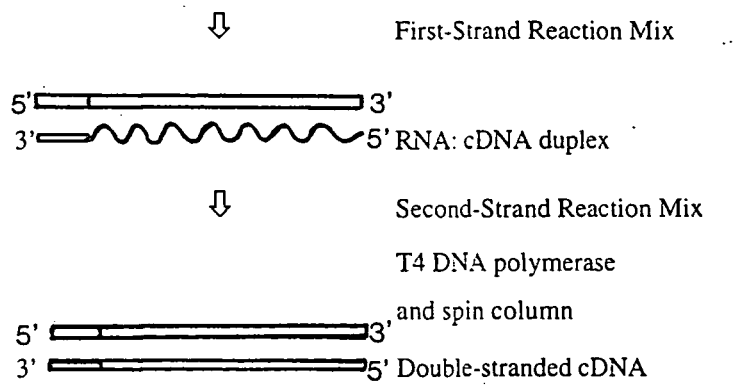


Figure 2. The use of the mRNA purification kit and the basis for the cDNA synthesis procedure.

First-strand cDNA synthesis is catalysed by Moloney Murine Leukemia Virus (MMLV) reverse transcriptase, using previously purified polyadenylated mRNA as template and oligo $(dT)_{12-18}$ as primer: the conditions permit full-length transcription of RNAs. A modified procedure is used for second strand synthesis, in which RNase H nicks the RNA strand of the RNA: cDNA duplex formed, and DNA polymerase I uses the nicks to replace the RNA with DNA by nick translation. T4 DNA polymerase was used in preference to Klenow fragment to produce blunt ended cDNA following second strand synthesis. The cDNA produced is extracted with phenol/chloroform before further purification and size selection using SizeSep 400 Spun Columns (Pharmacia). The columns remove small molecules such as oligonucleotides, nucleotides, phenol and other by-products that could interfere with subsequent reactions, as well as rapid selection of cDNAs larger than about 400 base pairs. Purification is achieved using bench top centrifugation with minimal dilution in a suitable buffer, often avoiding precipitation.

1.6.3. cDNA population restriction fragment sorting

As mentioned previously the sequencing of cDNAs, as opposed to genomic sequencing, represents a rapid route to gene identification. But as many cDNAs in man occur in more than one of the 50 tissues, or 200 possible cell types, the probability of repeatedly identifying the same cDNAs is quite high. This problem is exacerbated by the fact that the cDNA libraries generally available are poorly representative of the tissues from which they are derived, being biased towards abundant mRNAs.

Approaches for increasing the complexity of cDNA libraries (or normalisation) are necessary to minimise the redundancy encountered in systematic sequencing programmes. Restriction fragment sorting provides this means of subdividing complex cDNA mixtures into distinct semi-ordered subpopulations. In concept, individual subsets are of relatively low complexity, but combined provide access to a greater repertoire of genes than conventional cDNA libraries.

The restriction fragment sorting of cDNAs is based upon the partitioning of cohesive-ended cDNA restriction fragments according to the sequence of their ends. This necessitates the use of an enzyme which produces staggered cuts outside of its recognition sequence, so that any combination of bases is possible in the cohesive ends produced. The population of cDNA is digested with a type IIS restriction endonuclease,

generating fragments with non-identical cohesive ends. The number of different end sequences is 4^n, where n is the length of the overhang.

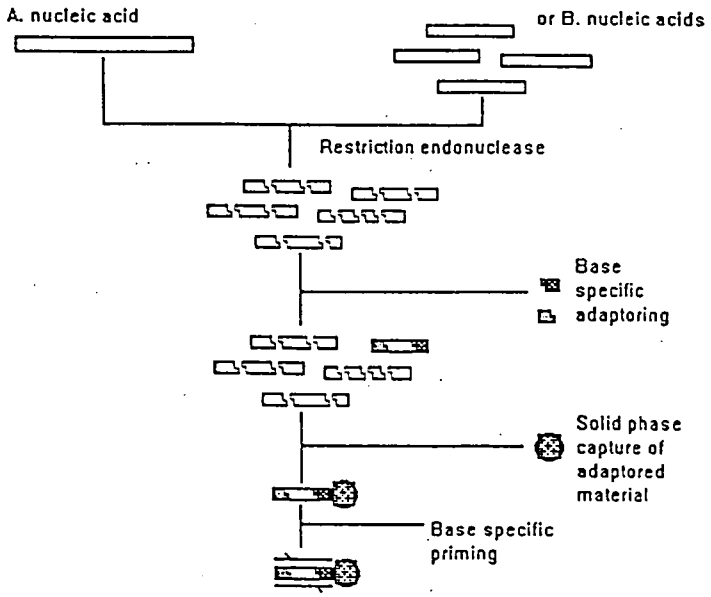


Figure 3. Basic scheme for selecting specific restriction fragments.

Fok I has an asymmetric pentanucleotide recognition sequence GGATG (N) 9/13, (cutting DNA every 512 bp on average), and generates fragments with 4 base 5' overhangs. If the two ends of a *Fok* I fragment are considered independently, there are potentially 4^8 (65,536) fragment classes, each with a different pair of cohesive ends. Each base of a given *Fok* I cohesive end represents an identifier for the cDNA fragment possessing that overhang. If 4 bases divided between the two ends of a *Fok* I fragment are considered, 4^4 (256) different classes of fragment can be identified.

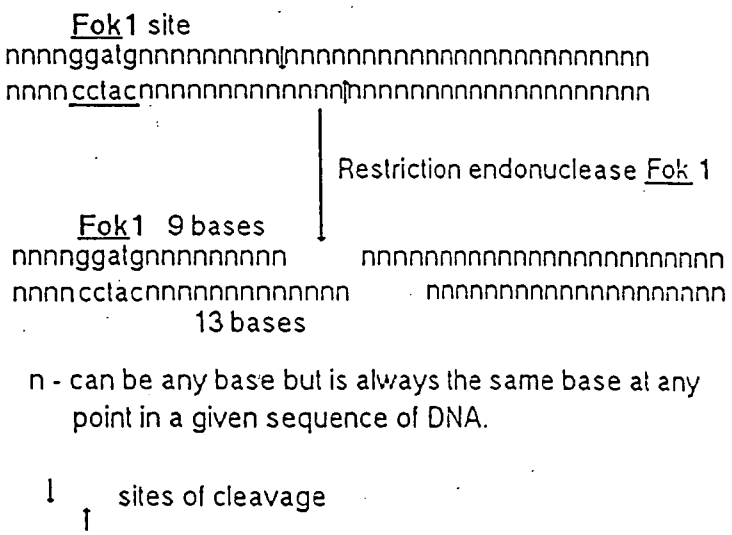


Figure 4. Nature of the cleavage produced by the restriction endonuclease *Fok* I.

cDNA fragments can be sorted into different subsets by successive base specific adaptoring, and base specific PCR. The selection of cDNA fragments by PCR serves to enrich the proportion of low abundance mRNAs, facilitating the subdivision of complex mRNA populations into subsets with a combined broad representation of cDNAs.

The procedure by which 256 *Fok* I cDNA restriction fragment classes are recognised is first by base specific adaptoring. The specificity of the T4 DNA ligase reaction is employed to select for cDNA fragments with particular cohesive termini. *Fok* I cDNA fragments are ligated to two types of adaptor. A given specific adaptor, with specified bases at two positions within the 4 base 5' cohesive end, is capable of specific ligation to $^{1}/_{16}$ of cDNA fragments which have complementary cohesive ends.

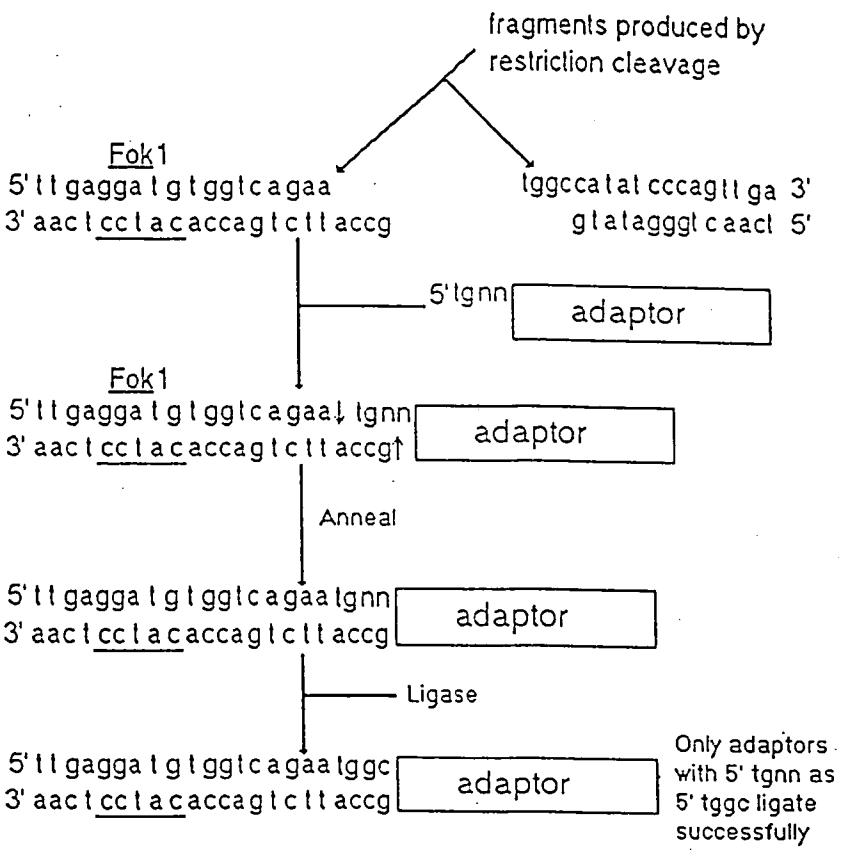


Figure 5. Base specific adaptoring at a *Fok* I cleavage site: specificity is introduced at two of the overhanging 5' bases of the adaptor.

Solid-phase capture of cDNA fragments containing at least a single specific adaptor provides a means of selecting $^{1}/_{16}$ of all *Fok* I cDNA fragments. Independent application of 16 different specific adaptors would enable a heterogeneous mixture of *Fok* I cDNA fragments to be fractionated into 16 "primary" subsets.

Base specific adaptoring and isolation of adaptored fragments

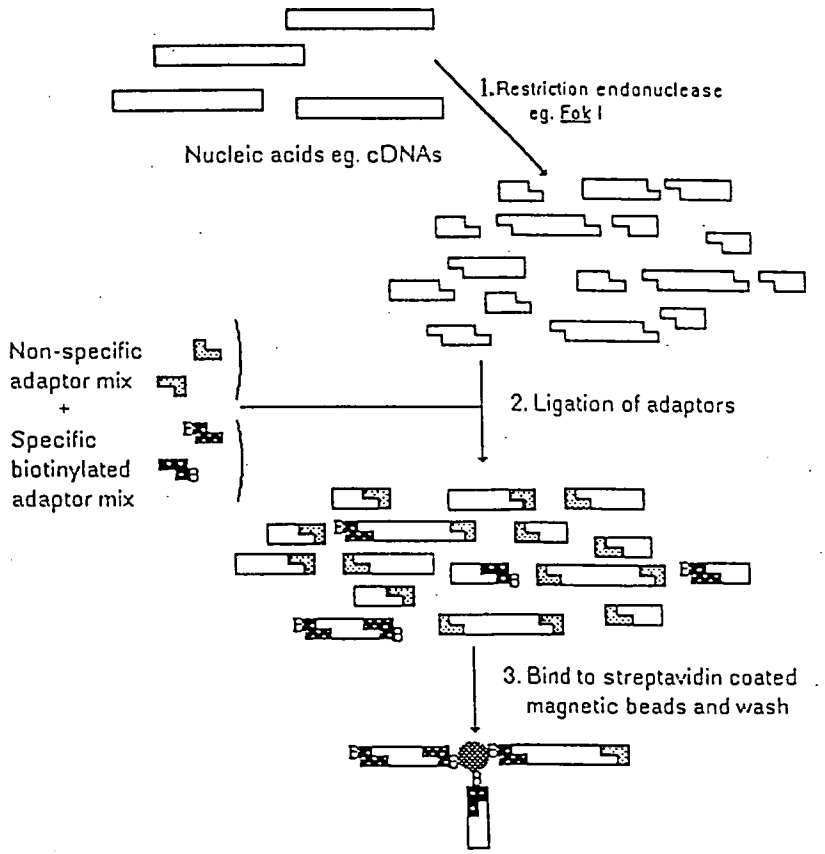


Figure 6. Selection for fragments produced by a type IIS restriction endonuclease: cuts are staggered and do not overlap the recognition site.

The specificity of primer annealing and extension, as part of the PCR reaction, can then be utilised to further subdivide adaptored *Fok* I cDNA fragment "primary" subsets. Asymmetric PCR, employing a primer with a specified base at its 3'-terminus, will theoretically discriminate in favour of solid-phase captured single-stranded cDNAs containing a complementary base at position four of the 5' cohesive terminus of a non-specific adaptor sequence. Four primers (3' "A", "C", "G" and "T"), used independently, would partition each "primary" subset into 4 "secondary" subsets.

A second selective round of amplification of specific solution phase single-stranded cDNAs can be achieved employing a primer 100% complementary to cDNAs with a given base at position four of the 5' cohesive end of a specific adaptor sequence. Amplification of each of the 64 "secondary" subsets with each of 4 specific primers (3' "A", "C", "G" and "T") would generate 256 "tertiary" subsets.

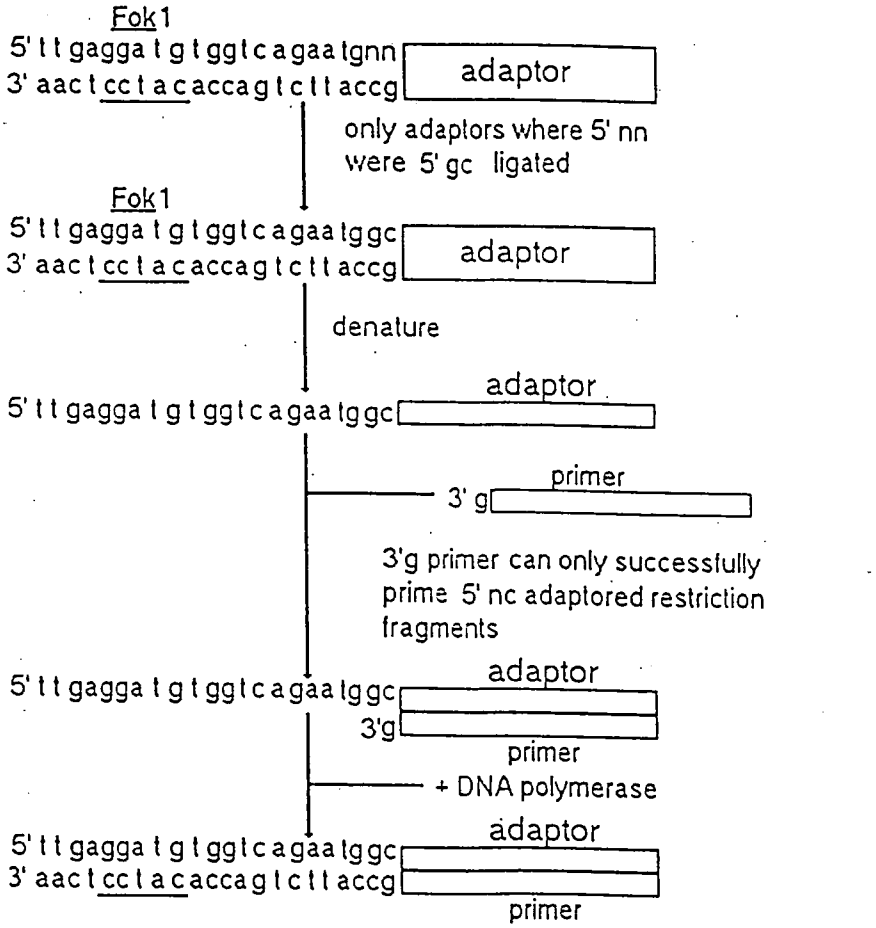


Figure 7. Base specific priming of base specifically adaptored *Fok* I fragment.

Base specific priming of DNA synthesis

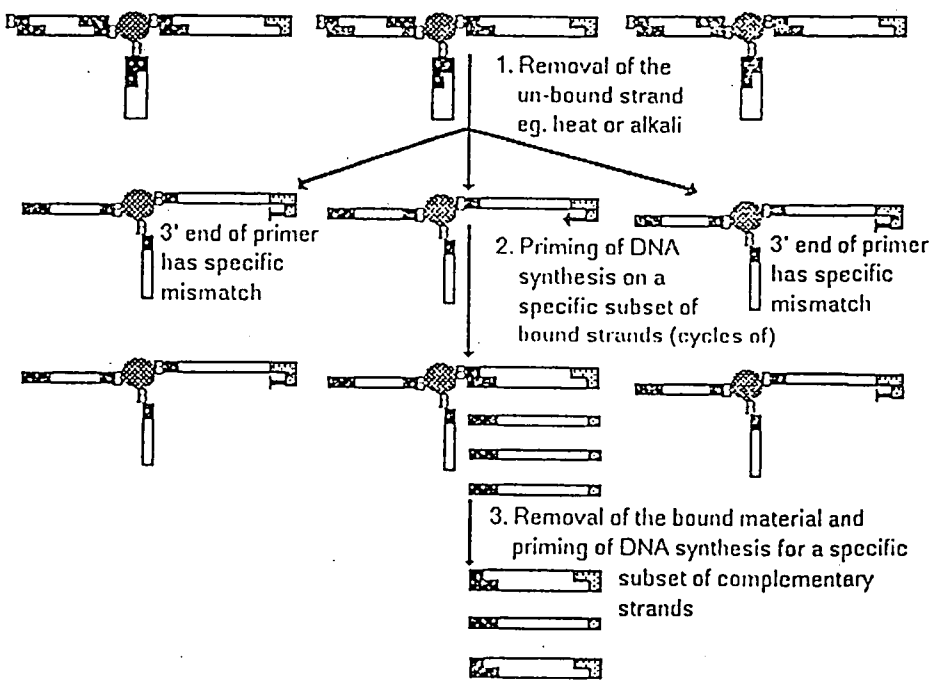


Figure 8. Selection for fragments produced by a type IIS restriction endonuclease: cuts are staggered and do not overlap the recognition site.

1.6.4. Vector construction

The oligonucleotide cassette was designed so it can be put into any vector containing unique *Hind* III and *Sac* I restriction sites within the multiple cloning site, and no *Bbs* I restriction site. Cutting with *Bbs* I (GAAGAC (N) 2/6) leaves non complementary overhangs for immediate directional cloning.

Bbs I

5' tcgccggtcttcggatccctgcaggaagactg 3'

3' gccagaagcctagggacgtccttctgacggct 5'

Bbs I

The PCR primers were designed so that when the amplified cDNAs were treated with T4 DNA polymerase and dTTP, the 4 base overhangs produced were complementary to those of the vector. Blue/white colour selection of the recombinants was retained in the appropriate vector.

Alkali lysis involves adding three buffers to the pelleted DNA, which weaken the cell walls, before lysis and neutralisation. Relative volumes can be adjusted as required for the size of preparation required, but times, temperatures and mixing methods are critical to good yields of plasmids of the correct formation: yields of over 1 µg per ml of culture should be obtained. Further removal of proteins and purification can be achieved by either phenol/chloroform methods, or using Qiagen tips which adsorb the DNA, washing off proteins and RNA from the salt column before eluting the plasmid DNA. Better quality DNA is obtained using columns, but is slower, more expensive, and not always necessary.

1.6.5. Cloning of cDNA fragments

The fetal liver cDNA library was constructed in the male (F') *E. coli* strain XL1-Blue by standard cloning (Hanahan, 1983), of the adaptored *Fok* I generated cDNA fragments (majority 440-570 bp) into the pBluescript II KS (+/-), appropriately prepared.

1.6.6. Template production and sequencing

Plasmid preparations could have been made by standard alkali lysis method, using Qiagen columns and protocols to manufacturers instructions, or any other commercially available products that produce good quality double stranded DNA for sequencing, but previous experience showed single stranded DNA (ssDNA) routinely reproducibly generates much better quality sequence data.

Solid phase preparation of sequencing single stranded templates from polymerase chain reaction (PCR) products (Uhlen, 1989) is an accepted protocol for producing many good quality ssDNA templates. Excess primers impede good bead binding and priming of sequencing templates, and so were removed using S-400 HR MicroSpin Columns (Pharmacia).

Single-stranded DNA represents the template of choice for dideoxy-chain termination DNA sequencing. The recombinant filamentous bacteriophage vector M13 is widely used to produce ssDNA for sequencing, since the DNA can be isolated with comparative ease and in high yields, for example using lacZ beads (Hawkins, 1992, and Watson *et al.*, 1993): but the unstability of the genomes of the bacteriophages in the presence of large inserts precludes the use of the conventional M13 vectors for the routine cloning and propagation of DNA.
Phagemid vectors (plasmids containing a filamentous bacteriophage origin of replication) provide a means of generating ssDNA, while obviating the necessity for the subcloning of fragments from plasmid to bacteriophage vectors. Superinfection of cells transformed with these plasmids using a suitable "helper" filamentous bacteriophage initiates packaging of single-stranded copies of these plasmids into progeny phage particles.
The major disadvantage of phagemids is generally the low and irreproducible yields of ssDNA isolated following superinfection. This problem can be largely alleviated by selection of appropriate vector and helper phage.

pBluescript II KS (+/-) (Short *et al.*, 1988) is a double-stranded phagemid, which is treated as a plasmid with regard to the cloning of foreign DNA and the *in vivo* propagation of foreign DNA, following conventional transformation of *E. coli*. The "KS" signifies the orientation of a synthetic polylinker within the N-terminal coding

region of the β-galactosidase gene. The vector carries a specific fragment from the intergenic region of the f1 filamentous phage, including the sequences required for the initiation and termination of viral DNA replication and for morphogenesis of bacteriophage particles. The (+/-) orientation of the f1 origin of replication enables rescue of the (+) phagemid strand, following superinfection.

Helper phage VCSM13 (Stratagene) is a derivative of M13K07 (Vierra & Messing, 1987), a M13 (Messing, 1983) recombinant containing mutations in the bacteriophage origin of replication (insertion of lacZ sequences) and gene II (nucleotide transition). The single-stranded genome also possesses a plasmid origin of replication and a kanamycin resistance gene (to facilitate selection for infected cells). VCSM13 is resistant to interference with the replication of the helper phage DNA, effected by certain f1 intergenic sequences present in some phagemids. The mutated VCSM13-encoded gene II product interacts less efficiently with the M13 origin of replication than with the pBluescript f1 origin of replication, and hence the majority of progeny single-stranded DNAs are derived from the phagemid.

Since DNA purity and quantity are important prerequisites for "good" sequencing templates, diligent purification of the single stranded DNA is essential. The standard template preparation procedure involves the removal of bacterial cells from the phage culture by centrifugation, precipitation of phage, and simultaneous phage lysis and deproteinisation. The laborious nature of this procedure is not readily compatible with the automation or simultaneous processing of a large number of templates.

Elimination of the need for centrifugation and precipitation permits the handling of samples in microtitre plate format and renders the ssDNA preparation adaptable to automation. The innovative growth of infected *E. coli* cultures in the 96-well filtration plate assemblies enables the efficient removal of bacterial cells by filtration, while selective purification is achieved by the specific hybridisation to the lacZ complementary paramagnetic bead-linked oligonucleotide.

Helper phage in Millipore Multiscreen 96-well filtration plates (3)

⇩

Inoculation (96 at once), with cDNA clones

⇩

Overnight growth at 37°C

⇩

Application of vacuum and collection of filtrates

⇩

Dynabeads lacZ Vector Purification Kit

Lysation and hybridisation in 1.5 ml tubes in water baths

⇩

Transference to a 96 format for washing of beads

⇩

Elution of ssDNA in 10 µl using Perkin-Elmer Cetus 9600 thermal cycler

⇩

Checking of amount and purity on an agarose gel, (0.5 µl sample)

⇩

Reconditioning and reuse of Dynabeads lacZ

⇩

Sequencing of ssDNA template using

PRISM Ready Reaction DyeDeoxy Terminator Cycle Sequencing

Figure 9. The purification of ssDNA using lacZ beads.

The PRISM™ Ready Reaction DyeDeoxy™ Terminator Cycle Sequencing Kit (Applied Biosystems) is used specifically for preparing samples for sequence analysis on the ABI 373A DNA Sequencer. Only the correct amounts of single or double stranded (ss or ds) template and primer (M13 (-20) GTAAAACGACGGCCAGT, an improved KS TCGAGGTCGACGGTATC, and T7 AATACGACTCACTATAG were used), needs to be added to the DyeDeoxy terminator reagent to perform fluorescence based dideoxy sequencing in a single tube. The four dye-labelled dideoxy nucleotides: G, A, T and C DyeDeoxy terminators, replace the standard dideoxy nucleotides in enzymatic sequencing, incorporating a dye label into the DNA with the terminating base. This chemistry eliminates labelled false stops, and all four termination reactions can be performed in one tube. Thermal cycling of sequencing reactions increases signal intensity and decreases sensitivity to reaction conditions. Excess dye-labelled dideoxy terminators have to be removed before the dye-labelled extension products are analysed. This is often done by the standard successive phenol/chloroform extractions although more efficient methods using various types of and spin-column chromatography including Biogel P6DG and S200, small and large columns are investigated.

The *Taq* Dye Primer Cycle Sequencing Kit is also used specifically for preparing samples for sequence analysis on the ABI 373A DNA Sequencer. Four labelled primers (each labelled with a different ABI fluorescent dye) and standard dideoxynucleotides are used in standard enzymatic sequencing.

Using standard overnight and quicker Basesprinter electrophoresis runs optimises the use of the available 373A sequencers for real time data collection.

1.6.7. Sequence manipulation and analysis

The ABI 373A sequence files, containing cDNA sequences in GCG format, are automatically categorised by "CDtotal", (originator Discala, C.), as "Good" (<4% ambiguities between bases 50-400), "Intermediate" (4-6.6% ambiguities between bases 50-400), or "Bad" (>6.6% ambiguities between bases 50-400). The artificial classification ascribed to each given sequence determines the nature of subsequent processing of that sequence. The "CDtotal" program additionally generates an experimental log displaying the continually updated progression in the accumulation of cDNA sequences (for each cDNA library, recording the number of cDNA sequences

falling into each of the three specified categories, the mean length of cDNA sequences, and the mean % of ambiguities between bases 50-400).

Removal of cloning vector and adaptor sequences via a two stage screening procedure, involving the program "Sedit", (originator Williams, G.), is necessary prior to entry into the cDNA database and a standard "BLAST" search (Altschul *et al.*, 1990), against the GenBank and EMBL databases respectively is completed.

cDNAs from the fragmentary fetal organ cDNA libraries are sequenced in the "forward" and/or the "reverse" direction (using mainly the M13 Universal and KS primers, respectively). The manner in which the "Good" (<4% ambiguities between bases 50-400) cDNA sequences are analysed is a reflection on the "*Fok* I cDNA fragment sorting" approach employed to construct the libraries. The restriction fragment sorting provides a means of subdividing complex cDNA mixtures into distinct subpopulations, and is based upon the partitioning of cohesive-ended cDNA restriction fragments according to the sequence of their ends. cDNA fragments are sorted into different subsets by successive base-specific adaptoring, and base-specific PCR. The analysis essentially involves evaluating the effectiveness of the restriction fragment sorting procedure at generating essentially distinct cDNA subpopulations.

2. MATERIALS AND METHODS

2.1. RNA extraction

2.1.1. Homogenisation and ultracentrifugation

RNA was isolated and purified from various human fetal tissues, stored at -80°C, by homogenisation in Extraction Buffer containing guanidinium thiocyanate (GTC) (Pharmacia Biotech Limited RNA Extraction Kit), and equilibrium centrifugation in caesium trifluoroacetate (CsTFA) solution (Pharmacia Biotech Limited RNA Extraction Kit), (Okayama *et al.*, 1987).

A total tissue mass of between 0.5 and 1.0 grams for each of the six organ types was obtained, using samples for one organ type as necessary.

Tissue	Specimen number	Menstrual age (weeks)	Mass (g)	Total mass (g)	Extraction fluid (mls)
Adrenal	11178	12.7	0.31		
	11289	13.4	0.51	0.82	14.8
Hand	11205	12.1	0.27		
	11281	15.1	0.68	0.95	17.1
Liver	11294	10.6	0.32		
	11296	14.6	0.31		
	11297	13.1	0.36	0.99	17.9
Lung	11285	14.6	1.08	1.08	19.5
Spleen	11285	14.6	0.25		
	11347	14.0	0.10		
	11384	15.1	0.15		
	11391	15.1	0.18		
	11393	15.4	0.13	0.81	14.6
Stomach	11165	15.8	0.36		
	11281	15.1	0.39	0.75	13.5

Table 1. Mass and age of fetal tissues used.

Tissues were added to 18 ml of RNA Extraction Buffer per gram of tissue (Pharmacia Biotech Limited buffered aqueous solution containing guanidinium thiocyanate, N-lauryl sarcosine, and EDTA; density =1.15 g/ml). The tissues were in turn passed through an Ultra-turrax blade, (Ultra-turrax T 25), for about 30 seconds each, until a homogenate was formed. These were centrifuged at 2500 revolutions per minute for 20 minutes at 15°C, in a Sorvall T6000B Centrifuge using a balanced swinging bucket rotor, to remove cell debris in preparation for ultracentrifugation.

Modifications were made to the standard Pharmacia Biotech Limited RNA Extraction Kit procedure involving the use of CsTFA aqueous solution with a greater density of 2.0 g/ml (Carter *et al.* 1983). The CsTFA was diluted to achieve an overall average density of approximately 1.55 g/ml with Extraction Buffer and tissue. CsTFA is hydroscopic and so was not left uncovered as a change in density would result. Using Beckman Polyallomer Quick Seal centrifuge tubes, 16x76 mm: 6.0 mls of CsTFA solution was used per tube (6x12=72 mls), as the total capacity of the each tube was 12.5 mls:[(6.0x2)+(6.5x1.15)=12.5x1.558].

Taking care not to disturb the pelleted material at the bottom of the tube: the required amount of the clarified homogenate in the extraction buffer, (6.5 mls), was removed and layered on to the CsTFA, filling each of the 12 tubes, 2 per tissue type. They were then sealed using a tube sealing device: in which the necks were melted with a metal cap on, moulded to seal the tubes; and left to cool to reset.

The nucleic acids were centrifuged in CsTFA solutions at 120,000 g_{av} (36K), for 42 hours, at 15°C in a Beckman L7-65 Ultracentrifuge using a balanced Ti50 fixed angle rotor with metal caps on the tubes.

2.1.2. Recovery of RNA

The tubes were carefully removed from the ultracentrifuge, making sure not to disturb the density gradients. Fractionation of the resultant bands were completed by collection of the contents of each tube from the bottom upwards through a 1.1 mm x 50 mm needle. The RNA band below the band of DNA was identified by a standard electrophoresed 0.85% (w/v) tris borate (1xTBE) agarose gel labelled with 0.5 µg/ml ethidium bromide, loading a 10 µl sample from each vial and 1 µl of DNA markers numbers 6 and 3, mixed with 2 µl of 0.15% ficoll with bromophenol blue loading dye. A potential difference of 80 Volts was then applied across the gel running from negative

(black terminal), to positive (red terminal), for about an hour. The gel was then studied under a Ultra Violet light trans-illuminator with care wearing a protective face mask at all times: in which the ethidium bromide which intercalates with the RNA and DNA fluoresce. An image was taken, to examine the results of the electrophoresis further. From the location of luminescence on the gel, it was calculated which vials contain the required total RNA for each organ: these were then used further.

The located RNA near the base of the tubes in the early fractions was then reloaded in an identical gradient in only 6 tubes, (the same tissue types were pooled), and re-centrifuged at 120,000 g_{av} (36K), for 16 hours, at 15°C in a Beckman L7-65 ultracentrifuge using a Sw 41 Ti swing out rotor.

Similar fractionation and electrophoresis procedures were again utilised to locate the bands of RNA. Standard ethanol precipitation purified and concentrated the RNA, whilst removing the CsTFA: 1/10 th volume of 2 M potassium acetate and 2.5 x volume of ethanol was added to the positively identified aliquots. Mixed samples were placed at -20°C to aid precipitation.

Samples were centrifuged for 30 minutes in a Sorvall T6000B bench top centrifuge at 2500 rpm. The supernatant was poured off, and the pellet resuspended in a total of about 1 ml 70% ethanol and transferred to a 1.5 ml tube. The RNA was then repelleted by microfuging at 13000 rpm for 30 minutes. The majority of the ethanol was poured off before pulse spinning and removing the remainder of supernatant by pipette. The RNA was briefly air dried, before dissolving in 360 µl of TE buffer, (10 mM Tris-HCl, 1 mM EDTA, pH 7.5), on ice. 10 µl could be examined by standard agarose gel electrophoresis on a 0.85% agarose gel. electrophoresis. 1050 µl (3 x volume) of 3 M lithium chloride was added to the remainder, and placed at 4°C overnight for selective precipitation of the RNA.

Samples were microfuged at 13000 rpm for 30 minutes, and the supernatant was poured off. 500 µl of 80% ethanol was added to each sample and mixed before another 10 minutes microfuging. Again the supernatant was poured off, and after a brief spin excess fluid is pipetted off. The pellet of RNA was then air dried before redissolving in the appropriate volume of water to allow for pooling of samples from the same tissues as necessary, to a total volume of 400 µl.

Standard phenol/chloroform and chloroform (Sigma), extractions were also carried out before final ethanol precipitation of the total RNA. 400 µl of phenol/chloroform was added and mixed to each sample, and then microfuged at 13000 rpm for 5 minutes. The

supernatant was pipetted off and kept. The procedure was then repeated in full with chloroform. The RNA was again precipitated by ethanol precipitation in 1.5 ml tubes and placed at -20°C.

The tubes were mixed and 40 µl samples taken, with the remainder of RNA stored at -20°C. The samples were microfuged at 13000 rpm for 20 minutes, and the supernatant pipetted off. Fifty microlitres of 80% ethanol was added to each sample and mixed before another 10 minutes microfuging. Again the supernatant was pipetted off. The pellet of RNA was then air dried before redissolving in 40 µl of water.

20 µl was then examined by electrophoresis on a 0.85% agarose gel, whilst the other half was used to take an optical density reading by adding water to a 0.4 ml total. Readings were taken at wavelengths of 260 and 280 nanometres to calculate the concentration, amount and purity of the RNA available for subsequent production of cDNA.

2.2. mRNA purification and complementary DNA (cDNA) synthesis

The Dynabeads Oligo (dT)25 mRNA Purification Kit (Dynal), was used as instructed using 200 µl of beads for each 150 µg of total RNA (a mix containing 25 µg adrenal, hand, liver, lung, spleen and stomach RNA, and 150 µg liver RNA).
RNA was diluted to final volume of 200 µl binding buffer (20 mM Tris-HCl, 1.0 M LiCl, 2 mM EDTA, pH 7.5), and heated to 65°C for 2 minutes to disrupt secondary structures. This was added to 200 µl of washed beads also in binding buffer, to hybridise for 5 minutes at room temperature. The supernatant was removed after the tube had been in a Dynal MPC-E-1 magnet for over 30 seconds. The beads were washed twice out of the magnet with washing buffer (10 mM Tris-HCl, 0.15 M LiCl, 1 mM EDTA, pH 7.5), before removal of the supernatant as before The mRNA was eluted in 12.5 µl elution buffer (2 mM EDTA), after heating to 65°C for 2 minutes, and immediately placing the tube in a magnet. After positive gel electrophoresis results of a sample from the first wash and elute the purified mRNA was used immediately.

The standard cDNA Synthesis Kit (Pharmacia Biotech Limited), protocol was followed except T4 DNA polymerase was used instead of Klenow fragment for the production of blunt ended cDNA following second strand synthesis.
First strand synthesis was completed by simply adding the 1-5 µg of polyadenylated RNA in a total volume of 20 µl (diluted with RNase free water) that had previously been incubated at 65°C for 10 minutes and chilled on ice, to the First-Strand Reaction Mix (FPLC*pure* Cloned Murine Reverse Transcriptase, RNAguard, RNase- and DNase-Free BSA, oligo$(dT)_{12\text{-}18}$ primer, dATP, dCTP, dGTP, and dTTP, in aqueous buffer) with 1 µl of aqueous DTT solution added, and incubation at 37°C for 1 hour.
The first-strand reaction (33 µl in total) was subsequently added to the Second-Strand Reaction Mix (*E. coli* RNase H and *E. coli* DNA polymerase I in aqueous buffer containing dNTPs) and the new volume of 100 µl mixed by pipette action. This was incubated at precisely 12°C for 1 hour and 22°C for 1 hour, before inactivation of the enzymes at 70°C for 10 minutes and immediate chilling on ice. Eight units of T4 DNA polymerase (Boehringer Mannheim) were added and mixed before 30 minutes incubation at 37°C. At room temperature 100 µl of phenol/chloroform (Sigma) was added, the mix was vortexed before microfugation for 1 minute and collection of the upper aqueous layer.

After the phenol/chloroform extraction the double-stranded cDNA was purified and size selected using a standard SizeSep 400 Spun column (Pharmacia Biotech Limited), equilibrated in 50 mM NaCl, 10 mM Tris-HCl, pH 7.5. Each centrifugation was completed using a swinging bucket rotor in a Sorvall T6000B centrifuge at 1400 rpm for 3 minutes precisely including acceleration and deceleration time.

2.3. cDNA population restriction fragment sorting

Half of each of the cDNA products (60 µl) were cut to completion with 8 µl 4 U/µl *Fok* I enzyme (Boehringer Mannheim) in a total volume of 200 µl restriction buffer (10 mM Tris-HCl, 10 mM $MgCl_2$, 50 mM NaCl, 1 mM DTE, pH 7.5), at 37° for 2 hours, to generate cDNA fragments with unpredictable sequence at their termini. The activity of the restriction enzyme and reaction was tested on lambda DNA simultaneously and fragments observed by ethidium bromide agarose gel electrophoresis before proceeding. Purification was completed by two phenol/chloroform extractions and passing through SizeSep 400 Spun columns.

Oligonucleotides were synthesised on a 1 µM scale using an Applied Biosystems 380B DNA Synthesiser, separated from ammonium hydroxide by rotary evaporation and purified by high performance liquid chromatography (HPLC) on a Beckman reverse phase column. They were quantified by U.V. absorbance spectroscopy using a conversion factor of 20 µg/ml/A.$U._{260}$.

Half of the cDNA fragments were then mixed with 200 pmoles of each of the following adapters:

Universal π biotinylated adaptor	5' Bio GTTCTCGGAGCACTGTCCGAGA 3'
Non specific θ adapter	5' $(N)_4(N)_4(N)_4(N)_4$ TCCTTCTCCTGCACAGACA 3'
Universal θ adapter	5' TGTCTGTGCAGGAGAAGGA 3'

and 200 pmoles of 1 from the possible 16 specific π adaptors:

5' $AA(N)_4(N)_4$TCTCGGACAGTGCTCCGAGAAC 3'
5' $CA(N)_4(N)_4$TCTCGGACAGTGCTCCGAGAAC 3'
5' $GA(N)_4(N)_4$TCTCGGACAGTGCTCCGAGAAC 3'
5' $TA(N)_4(N)_4$TCTCGGACAGTGCTCCGAGAAC 3'
5' $AC(N)_4(N)_4$TCTCGGACAGTGCTCCGAGAAC 3'
5' $CC(N)_4(N)_4$TCTCGGACAGTGCTCCGAGAAC 3'
5' $GC(N)_4(N)_4$TCTCGGACAGTGCTCCGAGAAC 3'
5' $TC(N)_4(N)_4$TCTCGGACAGTGCTCCGAGAAC 3'
5' $AG(N)_4(N)_4$TCTCGGACAGTGCTCCGAGAAC 3'
5' $CG(N)_4(N)_4$TCTCGGACAGTGCTCCGAGAAC 3'
5' $GG(N)_4(N)_4$TCTCGGACAGTGCTCCGAGAAC 3'
5' $TG(N)_4(N)_4$TCTCGGACAGTGCTCCGAGAAC 3'
5' $AT(N)_4(N)_4$TCTCGGACAGTGCTCCGAGAAC 3'
5' $CT(N)_4(N)_4$TCTCGGACAGTGCTCCGAGAAC 3'
5' $GT(N)_4(N)_4$TCTCGGACAGTGCTCCGAGAAC 3'

5' $TT(N)_4(N)_4TCTCGGACAGTGCTCCGAGAAC$ 3'

These were denatured at 65°C for 3 minutes before incubation at 14°C for 16 hours with 1 µl (2.4 U/µl) T4 DNA ligase (Boehringer Mannheim) in a total volume of 90 µls ligation buffer (50 mM NaCl, 0.5 M Tris-HCl, 0.1 M $MgCl_2$, 0.1 M DTT, 1 mM spermidine, 10 mM ATP, pH 7.4).

Excess adapters and reagents were removed by two phenol/chloroform extractions and passing through SizeSep 400 Spun columns. The adapted cDNA was heated to 78°C in a thermalcycler in a total volume of 200 µl PCR buffer (50 mM KCl, 10 mM Tris-HCl, 2.5 mM $MgCl_2$) and dNTPs (0.27 mM), and 2 µl (5 U/µl) *Taq* polymerase was added. The reaction was incubated first at 78°C for 5 minutes, and then 72°C for 10 minutes.

80 µl of Streptavidin beads (Dynal) were washed three times with 1 M NaCl, 100 mM Tris-HCl, pH 8.3, and resuspended in 200 µl 50 mM NaCl, 100 mM Tris-HCl, pH 8.3.

The beads were added to the cDNA reaction and incubated at 28°C for 30 minutes (mixing every 5 minutes). The cDNAs on the beads were washed twice with 50 mM NaCl, 100 mM Tris-HCl, pH 8.3, four times with 0.15 M NaOH (with incubation for 5 minutes at 28°C for each), twice in water and once in PCR buffer (2.5 mM Mg^{2+}, 10 mM Tris, 50 mM KCl, pH 8.3); before resuspension in PCR buffer and dNTPs (0.2 mM dATP, dCTP, dGTP and dTTP).

Two pmoles of 1 of the 4 possible specific θ primers from:

5' TGTCTGTCGCAGGAGAAGGAA 3'

5' TGTCTGTCGCAGGAGAAGGAC 3'

5' TGTCTGTCGCAGGAGAAGGAG 3'

5' TGTCTGTCGCAGGAGAAGGAT 3'

and 0.5 µl (5U/µl) *Taq* polymerase was added to the cDNA on beads in a total volume of 60 µl. Under oil 16 cycles of PCR were completed: 95°C 30s, 65°C 2 min, 72°C 3 min.

30µl of Streptavidin beads were washed three times with 1 M NaCl, 100 mM Tris-HCl, pH 8.3, and resuspended in 30 µl 50 mM NaCl, 100 mM Tris-HCl, pH 8.3.

The beads were added to the completed PCR and incubated at 28°C for 30 minutes (mixing every 5 minutes). A magnet was applied to the beads, the supernatant removed, and used further after two phenol/chloroform extractions and passing through SizeSep 400 Spun columns

A fraction of the cDNA was set up in a PCR total volume of 40 µl including PCR buffer, dNTPs, and 2 pmoles of 1 of the 4 possible specific θ primers from:

5' TGTCTGTCGCAGGAGAAGGAA 3'

5' TGTCTGTCGCAGGAGAAGGAC 3'

5' TGTCTGTCGCAGGAGAAGGAG 3'

5' TGTCTGTCGCAGGAGAAGGAT 3'

the same as before for that subset, and 2 pmoles of 1 of the 4 possible specific π primers from:

5' GTTCTCGGAGCACTGTCCGAGAA 3'

5' GTTCTCGGAGCACTGTCCGAGAC 3'

5' GTTCTCGGAGCACTGTCCGAGAG 3'

5' GTTCTCGGAGCACTGTCCGAGAT 3'

0.5 µl (5U/µl) *Taq* polymerase was also added. Under oil 5 cycles of PCR were then completed: 95°C 30s, 65°C 2 min, 72°C 3 min on a Techne thermal-cycler.

20 pmoles of each of the relevant primers was added and a further 36 cycles were initiated. Every 4 cycles a sample of the PCR was taken and observed by standard agarose gel electrophoresis. The chosen PCRs (24, 28, or 32 cycles) were repeated under the same conditions with a further 10 minute 72°C extension cycle, again observing results before two phenol/chloroform extractions and passing the products through SizeSep 400 Spun columns.

The cDNA products were incubated in a total volume of 100 µl buffer (10 mM Tris-HCl, 10 mM $MgCl_2$, 50 mM NaCl, 1 mM DTE, 0.05 mM dTTP, pH 7.5), with 16 units T4 DNA polymerase at 37° for 30 minutes. After two phenol/chloroform extractions the products were passed through SizeSep 300 Spun columns, ready for cloning.

2.4. Vector construction

5 µg of pBluescript II KS (+/-) phagemid was cut with 0.1 units *Hind* III and 0.3 units *Sac* I in a volume of 100 µl 33 mM Tris acetate, 10 mM Mg-acetate, 66 mM K-acetate, 0.5 mM DTT, pH 7.9 for 2 hours at 37°C, after a suitable titration to optimise conditions. After two phenol/chloroform extractions and passing through SizeSep 400 Spun columns; 1 µg of product was added to 100 pmoles of both:

5' AGCTTTCGGCGGTCTTCGGATCCCTGCAGGAAGACTGGCGAGAGCT 3'

5' CTCGCCAGTCTTCCTGCAGGGATCCGAAGACCGCCGAA 3'

or

5' AGCTTTCGCCGGTCTTCGGATCCCTGCAGGAAGACTGCCGAGAGCT 3'

5' CTCGGCAGTCTTCCTGCAGGGATCCGAAGACCGGCGAA 3'

This was denatured at 65°C for 3 minutes before adding 0.24 units T4 DNA ligase to a total volume of 40 µl ligation buffer (50 mM NaCl, 0.5 M Tris-HCl, 0.1 M $MgCl_2$, 0.1 M DTT, 1 mM spermidine, 10 mM ATP, pH 7.4), and incubation at 16°C for 16 hours. After purification 1 unit of T4 polynucleotide kinase was added to a total volume of 100 µl buffer (50 mM NaCl, 0.5 M Tris-HCl, 0.1 M $MgCl_2$, 0.1 M DTT, 1 mM spermidine, 10 mM ATP, pH 7.4) and incubated at 37°C for 30 minutes.

Standard transformation protocols (Hanahan, 1983), were then utilised to grow plasmid colonies containing the cassette into *E. coli* on colour selective LB agar plates. *E. coli* (strain XL1-Blue) was incubated at 37°C in 100 ml LB broth to log phase ($O.D._{550} = 0.5$ A.U.; approx.10^8 cells/ml) in preparation. The cells pelleted by centrifugation and the supernatent poured off. They were then resuspended in 50 ml 50 mM $CaCl_2$ transformation buffer, and left on melting ice for 2 hours or more.

Ten microlitres of each ligation mixture was diluted with 100 µl of T.C.M. (10 mM Tris-HCl, 10 mM $CaCl_2$ and 10 mM $MgCl_2$), and cooled on ice in 15 ml polypropylene tubes (Falcon). Competent *E. coli* cells were again pelleted by centrifugation and the supernatant poured off. They were then resuspended in 5 ml 50 mM $CaCl_2$. 200 µl of competent cells were added with stirring, the tubes rolled, and incubated on ice for 30 minutes. Following heat shock at 42°C for precisely 1 minute, 1 ml of LB broth was added, and the transformed cells incubated with shaking at 37°C for 1 hour. The cells were then spun down, the supernatant poured off, and the pellet resuspended in the residual liquid. This was spread onto pre-warmed LB agar plates containing 50 µg/ml

ampicillin, 12.5 μg/ml tetracycline, 0.02 g/ml X-gal in DMF, and 0.024 g/ml IPTG. The plates were incubated for 15 hours at 37°C and then transferred to 4°C for 1-2 hours. Cells transformed by Bluescript plasmid (Tet^R) containing inserts interrupting the lacZ gene would be expected to grow as white colonies unless the reading frame was preserved.

In this case blue colonies were picked from the transformation plates into 2 ml of LB broth with ampicillin in bijous. These were incubated with shaking at 37°C overnight, from which plasmid preparations were made by alkali lysis methods (Sambrook *et al.*, 1989). The cultures were microcentrifuged to form a bacterial pellet, which were resuspended in 100 μl ice cold solution 1 (50 mM glucose, 25 mM Tris-HCl, 10 mM EDTA, pH 8.0). Twice the volume fresh ice cold solution 2 (0.2 M NaOH, 1% SDS), were added, and the tubes inverted several times. To each sample 150 μl of ice cold solution 3 (3 M potassium acetate) was added before brief mixing, and incubation on ice for 3-5 minutes. After microfugation at 13000 rpm for 5 minutes at 4°C the supernatants were removed for further phenol/chloroform and chloroform extraction purifications. The pure DNA samples were then precipitated by standard ethanol precipitation.

After the plasmid containing the cassette had been positively identified by restriction analysis, large scale plasmid preparations were made using standard Qiagen methods. These involved a protocol similar to alkali lysis: equal suitable volumes of the supplied solutions P1, P2 and P3 were used and added in the same manner. After centrifugation the supernatants were loaded on to a Qiagen tip, and adsorbed, suitable volumes of QB were applied and eluted before three separate additions and elutions to each tip of the same volume of QC. The required elutions produced after the applications of QF were collected in fresh tubes. Isopropanol (0.7 x volume) was added to precipitate the DNA, which was washed with 70% ethanol after centrifugation before redissolving in TE.

About 32 μg of product was cut with 8 μl *Bbs* I (4 U/μl) in a total volume of 160 μl 10 mM Bis Tris Propane-HCl, 10 mM $MgCl_2$, 1 mM DTT, pH 7.0 at 37°C for 2 hours; to produce 4 base overhangs in the plasmid ready for insertion of cDNAs prepared for cloning. A subsequent incubation with ligase at 16°C for 16 hours was set up to show no ligation occurs between the non complementary endings: 5' GCGA 3' and 5' CCGA 3', or 5' CCGA 3' and 5'GCGA 3'; other controls were also set up.

2.5. Cloning of cDNA fragments

Ten microlitres of the prepared cDNAs were ligated into 2 µl of vector using 2 units of T4 DNA ligase in a total volume of 20 µl buffer (50 mM NaCl, 0.5 M Tris-HCl, 0.1 M $MgCl_2$, 0.1 M DTT, 1 mM spermidine, 10 mM ATP, pH 7.4), and incubating at 14°C for 16 hours. The vector containing insert were cloned into *E. coli* strain XL1-Blue host cells using standard competence and transformation protocols (Hanahan, 1983). These were slightly more refined than those used previously involving Hanahan transformation buffer instead of $CaCl_2$ solution. 87.5 µl of DMSO was added to competent cells resuspended in 2.5 mls buffer and kept on ice for 5 minutes, before addition of the same volume of DTT, and incubation on ice for 10 minutes. Finally 87.5 µl DMSO was added, mixed gently and kept on ice for a further 5 minutes, before adding 200 µl aliquots of cells to the various ligation mixes including controls. After incubation on ice for half an hour, exactly 90 seconds at precisely 42°C, and placing on ice for 2-3 minutes, 800 µl of SOC broth was added. Colonies and controls were plated out in the residual after incubation at 37°C with shaking, before centrifugation and removal of supernatant. They were grown at 37°C for 16 hours on LB agar plates containing antibiotics (ampicillin and tetracycline) and blue/white colour selection (X-gal and IPTG) as before. White colonies were picked into wells containing 100 µl of LB broth and 50 µg/ml ampicillin per well. The colonies were duplicated into another microtitre plate and also stamped onto a LB agar plate, before incubation for 16 hours at 37°C. Standard glycerol stocks were made and stored frozen (Sambrook *et al.*, 1982). The clones on the agar plates were ready for subsequent PCR, helper phage inoculation or standard plasmid preparations to produce suitable template for sequencing.

2.6. Template production and sequencing

2.6.1. Template preparation from PCR products

The protocol used was as documented in detail by Sibson, D.R., in Chapter Twenty-One: Solid Phase Preparation of Sequencing Templates from PCR Products, of *Automated DNA Sequencing and Analysis* (Adams, Fields & Venter (eds), 1994), with working improvements added in accordance with the author.

Standard PCRs were set up from the clones grown on selective media using 20 pmoles of forward (5' GTAAAACGACGGCCAGT 3') and reverse (5' AACAGCTATGACCA TG 3') primers, one of which was 5' biotinylated, in a total volume of 40 µl PCR buffer and dNTPs, and 1 unit of *Taq* polymerase. These were kept at 95°C for 1 minute and then cycled 35 times: 95°C 30s, 60°C 30s, 72°C 40s, before a 5 minutes extension cycle. Excess primers were removed by passing the products made up to 60 µl with water through Pharmacia S-400 HR MicroSpin Columns as instructed using Beckman inserts to hold 1.5 ml tubes for use in the Sorvall benchtop centrifuge for ease and efficiency of throughput (24x4=96 at once).

M-280 Streptavidin beads (Dynal) were prepared for use by washing them in 1 M NaCl, 10 mM Tris-HCl, pH 8.3, (30 µl of beads per reaction), and resuspending them in 50 mM NaCl, 10 mM Tris-HCl, pH 8.3. The biotinylated products were then bound to the streptavidin coated beads by incubating at 28°C for 30 minutes in a Techne 96 well plate on a Techne hotplate with regular mixing. After washing the beads-double stranded DNA twice with 50 mM NaCl, 10 mM Tris-HCl, pH 8.3 using a Dynal MPC-96 magnet the non biotin containing strand was removed by incubation twice at 28°C for 5 minutes in fresh 0.15 M NaOH. The bead bound products were finally washed twice with water before resuspension in a suitable volume of water (10 µl) ready for sequencing.

2.6.2. DNA purification using lacZ beads

Each well of a 96-well plate (Falcon 3072 Microtest III tissue culture plate), containing 100 µl of LB (supplemented with 50 µg/ml ampicillin and 12.5 µg/ml tetracycline) per well, was inoculated with an *E. coli* fetal liver cDNA clone from a 25% glycerol stock microtitre plate maintained at -20°C. The ninety six clones were grown overnight at 37°C with shaking.

The clones were stamped onto LB + 1.5% bacto-agar (supplemented with 50 µg/ml ampicillin and 12.5 µg/ml tetracycline) in a single well plate (Dynatech Laboratories Inc. MIC-2000 Inoculum tray) and cultured for a further 16 hours at 37°C.

The *E. coli* clones were stamped into three 96-well filtration plate assemblies (Millipore Multiscreen 0.45 µm, triton-free, mixed cellulose acetate/cellulose nitrate ester filter), standing on microtitre plate lids (Dynatech Laboratories Inc): each well contained 250 µl of VCSM13 (~1 x10^8 pfu/ml) in 2 x YT broth (1.0% bacto-yeast extract, 1.6% bacto-tryptone, 0.5% NaCl, pH 7.0), supplemented with 50 µg/ml ampicillin. The cultures were grown with agitation (170 rpm) for 2 hours at 37°C. Kanamycin was added to each culture (final concentration 50 µg/ml), and the 3 x 96 clones were grown for a further 22 hours at 37°C.

The 96-well filtration plates were placed, in turn, onto a vacuum manifold and a vacuum applied to effect removal of bacterial cells from the phage supernatants by filtration. The filtrates from the three plates (3x~200 µl) were collected in to a single 96 deep-well titre plate (Beckman), (i.e. each of the ninety six phage supernatants was the product of three 250 µl VCSM13-infected *E. coli* cultures). Aliquots (10 µl) of a representative number of phage filtrates were analysed by agarose gel electrophoresis, to confirm that adequate yields of single-stranded phagemid DNA had been obtained before proceeding to purify the phagemid DNA further. The use of filtrate as opposed to the phage supernatant separated from the bacteria by centrifugation, with regard to both quality and throughput was evaluated. Single-stranded phagemid DNA (+ strand) was purified from the cleared phage filtrates by hybridisation to Dynabeads lacZ vector purification kit protocol (Fry *et al.*, 1992), unless otherwise indicated*. Dynal lacZ beads are streptavidin-coated paramagnetic beads, linked to which are biotinylated 40-mers complementary to a (+) strand sequence within the 5' regulatory region of the lacZ gene, present in M13 and M13-derived cloning vectors. The ninety six 600 µl phage filtrates were transferred from the deep well microtitre plate into individual 1.5 ml microcentrifuge tubes. The phage particles were lysed by the addition of 100 µl of 1.5% SDS, 0.125 M EDTA, and incubation (with mixing) at 75°C for 15 minutes. To each lysate, 200 µl of 20% PEG was added and 0.9 mg* (90 µl) of Dynabeads lacZ (determined by suitable titration to be sufficient), resuspended in 180 µl* of 6.5M sodium perchlorate. The phagemid DNA was hybridised with the lacZ beads by incubation at 45°C* for 30 minutes, mixing periodically to resuspend the beads. The majority of the supernatant (~1000 µl) was

removed from each tube using a Dynal MPC-M Magnetic Particle Concentrator (holds up to ten separate 1.5ml microcentrifuge tubes), to concentrate the paramagnetic beads, and the lacZ beads in each tube were resuspended in the residual ~80 µl volume and transferred to individual microAmp tubes (Perkin-Elmer) held in a GeneAmp 9600 microtube rack (Perkin-Elmer). The remaining 80 µl of supernatant was removed from each tube, and the lacZ beads were washed twice in 300 µl of 10 mM Tris-HCl (pH 8.0), 1.0 mM EDTA. The purified single-stranded phagemid DNA was eluted from the lacZ beads by the addition of 10 µl of 10 mM Tris-HCl (pH 8.0) to each sample, and incubation (Perkin-Elmer 9600 thermal cycler) at 70°C for 5 minutes and removal of elute whilst using a Dynal MPC-9600 Magnetic Particle Concentrator (holds Perkin-Elmer GeneAmp 9600 microtube rack). An aliquot (0.5 µl) of each of the ninety six elutes was examined by agarose gel electrophoresis to assess the purity and quantity of single-stranded phagemid DNA prior to sequencing. Beads were reconditioned using 0.1 M NaOH before reuse at least once.

Customised home made beads (Hawkins, 1992) were also made and evaluated briefly. The biotinylated oligonucleotide complementary to a part of the lacZ-region in different vectors was linked to Streptavidin beads through the biotin-streptavidin system. The DNA probe was an oligonucleotide 40-mer (5' TT ATC CGC TCA CAA TTC CAC ACA ACA TAC GAG CCG GAA GC 3'), complementary to the (+) strand of pBluescript II KS (+/-). One milligram per 100 µl of stock streptavidin beads were transferred to a 1.5ml Eppendorf tube, and put on a MPC-E magnet, to remove the supernatant, before washing with 200 µl of TTL buffer (100 mM Tris-HCl, pH 8.0, and 0.1% Tween 20, 1 M LiCl), and resuspension in 20 µl of TTL buffer. 25 pmoles of the HPLC purified biotin oligonucleotide was coupled to 1 mg (100 µl) of streptavidin, and incubated for 15 minutes at room temperature in the TTL buffer. The lacZ beads were then placed in a magnet to remove the supernatant, before washing with 0.15 M NaOH to remove any non-specifically bound probe. They were washed twice in TT buffer (250 mM Tris-HCl, pH 8.0, and 0.1% Tween 20) before incubation at 80°C for 10 minutes in TT buffer. They were then placed in a magnet and the supernatant removed again to remove any unstable biotin streptavidin couplings. Finally the lacZ beads complex were taken up in 100 µl hybridisation mixture, (6.5 M $NaClO_4$) or 100 µl storage buffer, (250 mM Tris-HCl, pH 8.0, 20 mM EDTA, 0.1 % Tween-20, 0.02% Sodium azide) to be ready for use.

2.6.3. Automated fluorescent sequencing

Sequencing of the ssDNA template was performed using standard ABI 373A sequencers with some Taq Dye Primer Cycle Sequencing but mostly PRISM Ready Reaction Dye Deoxy Terminator Sequencing Kit chemistry.

Slight alterations were made to improve the thermocycling protocols (30 cycles: 96°C, 20s; 50°C, 10s; 60°C, 4 minutes), (Rosenthal, 1992). Excess terminator removal methods utilised varied: one method used phenol/chloroform extraction and spin columns. First the columns were prepared for use: fresh Biorad Biogel P6DG gel (10 g/100ml in 50 mM NaCl, 10 mM Tris-HCl, pH 7.5-HCl) or Pharmacia Sephacryl-200 (MicroSpin S-200 HR by Pharmacia were also used), was loaded in to new large empty spin columns to a slurry depth of 3 cm or small columns to a depth of 2 cm. After elution, 3 ml of H_20 was loaded to each column and eluted (or centrifuged through). Columns were centrifuged at 1400 rpm for 3 minutes. For ease of use the smaller columns were preferred as these could be centrifuged (not microfuged as instructed) in 6x4 arrays (Beckman) allowing 96 samples to be processed at once. 20 µl of water was added to the thermal cycled extension products to be purified which were removed from storage at 4°C or -20°C (as necessary). 40 µl of phenol:H_2O: chloroform (68:18:14), was added to each tube, before replacing lids on tightly and vortexing for 60 seconds. They were centrifuged in the Sorvall centrifuge at 3000 rpm for 5 minutes carefully balanced. The 40 µl supernatants from the samples were loaded on to the columns before centrifugation at 1400 rpm for 3 minutes. Samples were placed in the Aquavac to dry down. The pellets were carefully resuspended in 3.5 microlitres formamide/50 mM EDTA (5:1 v/v) by vortexing and microfuging. Prior to loading the samples are heated on a heating block at 90°C for 2 minutes, and then immediately placed on ice. They were stored for at 4°C at this stage as necessary. Columns were reusable after gel matrix was cleaned with 3 volumes 0.2% NaAzide, and washed through with water. Other protocols were evaluated throughout for yield and quality of cleaned products.

Standard 24 cm well to read, and 0.4mm thick sequencing plates were used throughout. The 373A DNA Sequencer (Applied Biosystems) was set up as instructed by the manufacturers. Other settings and parameters were chosen and altered as required:

Parameter	Full Scan	BaseSprinter
Time (hours)	12	6
Power (watts)	30	40
Wells	24 or 36	18
PMT voltage	780-820	880-920
Gel concentration	6%	4.75%

2.7. Sequence manipulation and analysis

After necessary retracking and reanalysis (see ABI373A Users Manual), the raw data was transferred from the sequencer assigned Apple Macintosh to an edit Apple Macintosh where a hard copy was made on to optical disk using a Panasonic WORM system. The data was then transferred in gcg format with trace files using an ethernet connection to a Unix directory accessed by a DECpc system.

The ABI 373A sequence files, containing cDNA sequences in GCG format, were used as the input files for a program, "CDtotal", (originator Discala, C.), which categorised cDNA sequences as "Good" (<4% ambiguities between bases 50-400), "Intermediate" (4-6.6% ambiguities between bases 50-400), or "Bad" (> 6.6% ambiguities between bases 50-400). The artificial classification ascribed to each given sequence determined the nature of subsequent processing of that sequence. The "CDtotal" program additionally generated an experimental log displaying the continually updated progression in the accumulation of cDNA sequences (for each cDNA library, recording the number of cDNA sequences falling into each of the three specified categories, the mean length of cDNA sequences, and the mean % of ambiguities between bases 50-400).

cDNA sequences of "Intermediate" quality were edited (where possible), so that upon re-characterisation (re-evaluation using "CDtotal") they could be ascribed the "Good" classification. Editing featured the use of the Staden program, "Trace EDitor", to view a given trace file for possible characteristic DNA sequencer-specific base-calling errors, combined with the application of the GCG program, "Seqed", to appropriately edit the corresponding sequence file.

cDNA sequences classified as "Good" (0-3.7% ambiguities between bases 50-400), were validated prior to entry into the cDNA database through the removal of cloning vector and adaptor sequences via a two stage screening procedure, involving the program "Sedit", (originator Williams, G.), and a "BLAST" (Altschul *et al.*, 1990). search against the GenBank and EMBL databases, respectively.

The analysis essentially involved evaluating the effectiveness of the restriction fragment sorting procedure at generating essentially distinct cDNA subpopulations.

cDNA sequences (prior to the removal of adaptor sequences) were screened for the expected subset-specific adaptor and PCR primer pair selecting bases, and for the presence of *Fok* I sites (both at the 5'-end and internal), using the GCG program

"findpatterns". The program "CDpattern" displayed the output of "findpatterns" in tabular format.

The reverse compliments (= forward sequences) of cDNAs sequences (following the removal of vector and adaptor sequences) generated using the KS sequencing primer (reverse direction sequences) were obtained employing the GCG program, "REVERSE". The GCG programs, "FASTA" and "GELMERGE", were utilised to identify overlaps between the forward sequence and the reverse compliment of the reverse sequence of cDNAs for which both strands were sequenced. Where significant overlaps were identified, the forward sequence and the reverse compliment of the reverse sequence of a given cDNA were merged to construct a composite sequence, either utilising the consensus file output of "GELMERGE", or by manual editing. Where genuine overlaps could not be detected, the longer of the two sequences (forward sequence, or the reverse compliment of the reverse sequence) was selected for analysis.

The cDNA fragment composition of each subset was assessed by analysis of sequences (composite, or forward/reverse compliment of the reverse) using a program that indexes sequences sharing similarities into classes ("*ICAass*", originator Parsons, J., groups sequences which are approximately repeated within the length of another). Each class (family of sequences, or unique sequence) was characterised by similarity search ("BLAST") of its members against the GenBank database.

The cDNA fragment composition of each subset was then assessed by clustering the sequences into classes (family of sequences, or unique sequence) according to the highest scoring (best alignment) GenBank database similarity of each sequence ("CDcheck" program).

3. RESULTS

3.1. Quality of RNA and cDNA libraries

The first ultracentrifugation fractionated the RNA from the chromosomal and mitochondrial DNA successfully (Plates 1 & 2), with the second ultracentrifugation removing further DNA still present.

Plate 1. Ultracentrifugation density gradient fractionation of liver fetal tissue.

0.85 %TBE agarose gel stained with ethidium bromide: gel affected by CsTFA salt in samples.

20 µl samples: 1-19 fractionations from the base up: 3-8 RNA, 11-19 DNA and proteins.

1 µl DNA molecular weight marker VI.

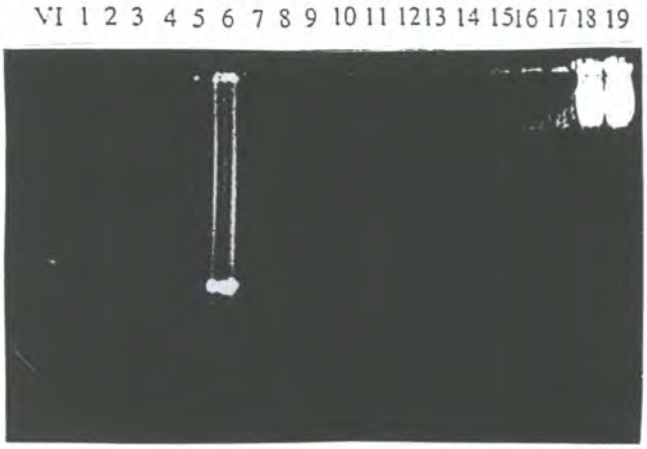


Plate 2. Ultracentrifugation density gradient fractionation of lung fetal tissue.

0.85 % TBE agarose gel stained with ethidium bromide: gel affected by CsTFA salt in samples.

20 µl samples: 1-19 fractionations from the base up: 6 RNA, 15-19 DNA and proteins.

1 µl DNA molecular weight marker VI.

The high quality of the purified total RNA was observed (Plate 3), and values between 1.8 and 2.0 for (OD_{260}/OD_{280}) indicated pure samples. Yields were variable between 47.0 µg and 682.1 µg (Table 2) depending on the nature of the fetal tissue.

Tissue	OD_{260}	OD_{280}	Purity (OD_{260}/OD_{280})	Concentration (µg/ml)	Total (µg)
Adrenal	0.074	0.034	2.18	59.2	82.9
Hand	0.042	0.022	1.91	33.6	47.0
Liver	0.609	0.315	1.93	487.2	682.1
Lung	0.328	0.162	2.02	262.4	367.4
Spleen	0.075	0.040	1.88	60.0	84.0
Stomach	0.125	0.066	1.89	100.0	140.0

Table 2. Purity and yield of RNA extracted.

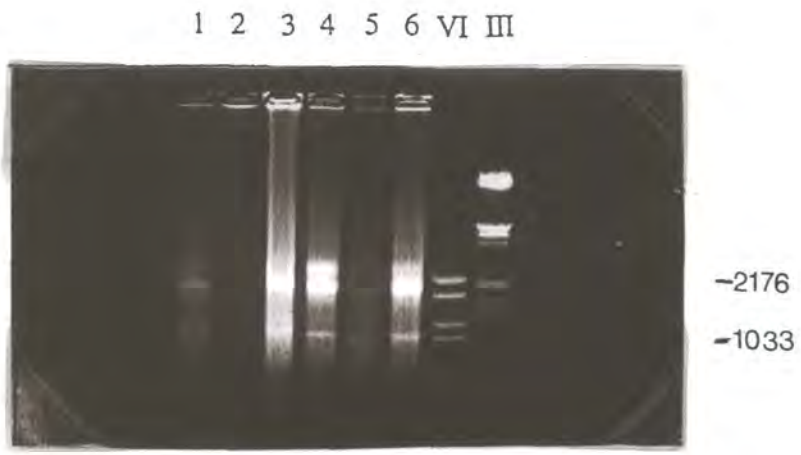


Plate 3. Total RNA.

5 µl samples: 1 Adrenal, 2 Hand, 3 Liver, 4 Lung, 5 Spleen, 6 Stomach.

1 µl DNA molecular weight marker VI (Boehringer Mannheim), pBR322 DNA·*Bgl* I + pBR322 DNA·*Hinf* I, solution in 10 mM Tris-HCl, 1 mM EDTA, pH 8.0, 0.25 µg/µl.

Fragment sizes: 2176, 1766, 1230, 1033, 653, 517, 453, 394, 298, 234, 220, 154 base pairs.

1 µl DNA molecular weight marker III (Boehringer Mannheim), λ DNA·*Eco*R I + *Hind* III, solution in 10 mM Tris-HCl, 1 mM EDTA, pH 8.0, 0.25 µg/µl.

Fragment sizes: 21226, 5148, 4973, 4268, 3530, 2027, 1904, 1584, 1375, 947, 831, 564, 125 base pairs.

The mRNA purification was shown to be successful by observing a sample of the first supernatant from the beads and a sample from the elute. The supernatant contains 95-97% of the total RNA, and the elute only about 3-5%.

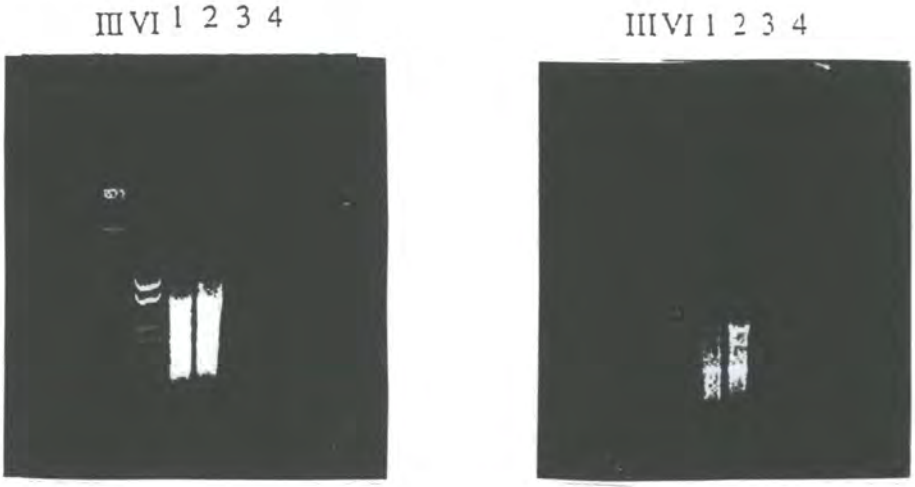


Plate 4. mRNA purification: two different exposures.

20 µl supernatant containing total RNA: 1 Mix: adrenal, hand, liver, lung, spleen and stomach, 2 Liver.

2.5 µl elute containing mRNA: 1 Mix: adrenal, hand, liver, lung, spleen and stomach, 2 Liver.

2 µl DNA molecular weight markers VI and III.

Much experimentation and quality control of products was required throughout the construction and production of the new vector ready for insertion of the prepared cDNAs. Firstly a titration of *Hind* III and *Sac* I restriction enzymes to be used had to be performed to optimise the cutting of the original pBluescript II KS (+/-) vector to produce a site suitable for insertion of the oligonucleotide cassette: and an actual check of the double digestion before ligation with adaptors

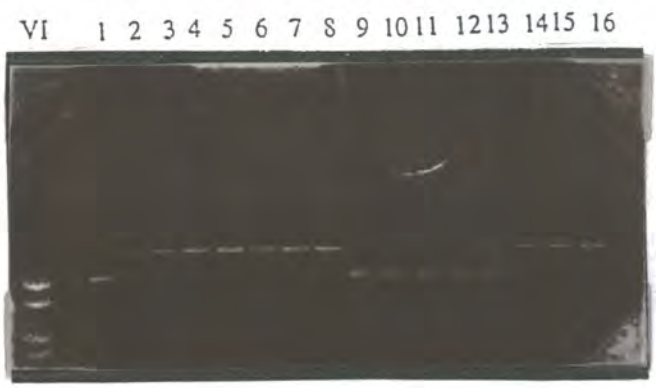


Plate 5. Restriction endonuclease titration.

2 µl DNA molecular weight marker VI.

15 µl samples: 1 & 9 no enzyme; 2-8 *Hind* III after 0, 1, 2, 4, 8, 16, & 32 minutes; 10-16 *Sac* I after 0, 2, 5, 10, 20, 30 & 60 minutes.

Plate 6. pBluescript II KS (+/-) cut with *Hind* III and *Sac* I.

2 µl DNA marker VI.

1 1 µl uncut Bluescript KS; 2 20 µl sample KS cut with *Hind* III and *Sac* I.

After transformation a small plasmid preparation and restrictions confirmed some of the picked clones contained the correct insert in the cloning site. Large scale Qiagen plasmid preparations then produced a greater yield of pure plasmids containing the insert. From fluorimeter readings taken concentrations of each preparation were calculated.

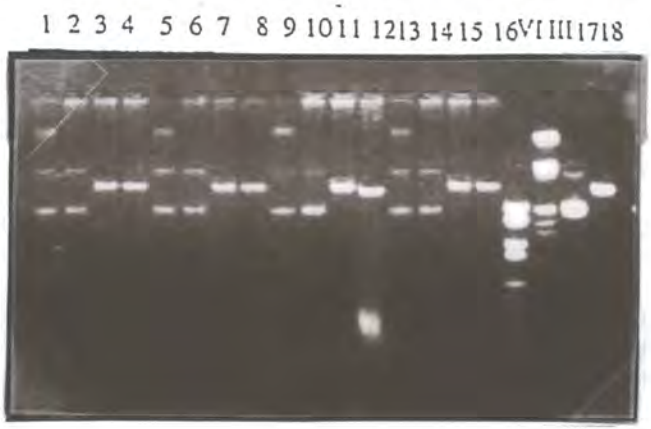


Plate 7. Positive identification of vector containing insert.

2 µl DNA molecular weight markers VI & III.
1-4 clone 1, 5-8 clone 2, 9-12 clone 3 & 13-16 clone 4.
1, 5, 9 & 13: no enzyme, not cut.
2, 6, 10 & 14: *Eco*R I cut, cuts original KS only.
3, 7, 11 & 15: *Pst* I cut, cuts original and vector with insert.
4, 8, 9 & 16: *Bbs* I cut, cuts only vector with insert.
17 uncut circular KS, 18 cut linear KS.

Plate 8. Plasmid preparations of vector containing insert.

2 µl DNA markers VI & III.
1-4 1 µl different clone preparations.
1 0.97 mg/ml, 2 2.38 mg/ml,
3 2.16 mg/ml, 4 3.67 mg/ml.
5 1 µl uncut KS.

The plasmid containing insert was cut with *Bbs* I ready for use. Other controls were included for quality control of the final product before ligation with the prepared cDNAs. These controls included cutting with other restriction enzymes and the addition of ligase to prove the *Bbs* I cut vector with its non complementary ends did not religate.

Plate 9. Restriction of inserted vector to prepare for use and controls.

2 µl DNA molecular weight markers VI & III.

1, 5 & 9: clone 1; 2, 6 & 10: clone 2;

3, 7 & 11: clone 3; 4, 8 & 9: clone 4.

1-4 *Pst* I cut; 5-8 *Bbs* I cut; 9-12 uncut.

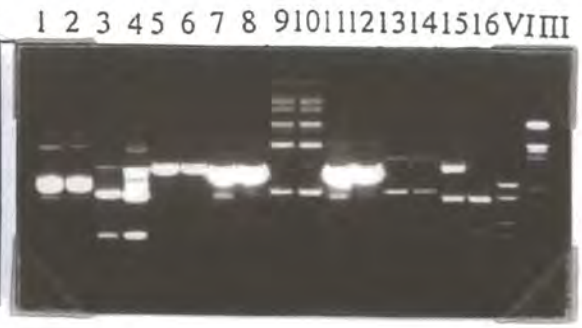


Plate 10. Controls of prepared vector.

2 µl DNA markers VI & III.

Odd nos. clone 1; even nos. clone 2.

1 & 2 uncut circular vector with insert,

3 & 4 *Pvu* I cut, cuts twice, 5 & 6 *Pst* I cut, cuts once,

7 & 8 *Bbs* I cut, cuts once, 9 &10 *Pst* I cut + ligase,

cuts and religates, concatamers,

11 & 12 *Bbs* I cut + ligase, cut once no religation,

13 & 14 *Pst* I cut + ligase, *Pvu* I cut, cuts twice

15 & 16 *Bbs* I cut + ligase, *Pvu* I cut, cuts twice

The generation of the specific PCR products was titrated and observed for optimisation before concatamerisation occurred before proceeding with purification and preparation for cloning and subsequent ligation with vector.

Plate 11. Selective PCR of cDNA fragments prior to cloning.

1 Mix, aa adaptor, c and a primers; 2 Liver, aa, cc, subset CB; 3 M aa ga; 4 L aa gc, subset CE; 5 L tt ca; 6 M tt cc, subset M; 7 L tt cc, subset CA; 8 L tt ga, subset CD; 9 L tt gc, subset CC.

2 µl DNA molecular weight marker VI.

After transformation, white colonies were picked and it was shown that over 75% of clones contained cDNAs between 340 and 750 bp (mostly 440-570 bp, average ~500 bp), and produced good clean PCR products suitable for further study.

3.2. Sequencing templates and data

PCR reactions were observed before column purification of clean products to remove excess primers, and before bead binding. The Streptavidin bead wash of the eluted strand was also observed before proceeding with sequencing of the bead bound template strand. Only the clones that produced good PCR products were used for lacZ methods.

Plate 12. PCR products after column purification.

1-24 5 µl samples of PCR products of different clones from plate C10, subset CA.

2 µl DNA molecular weight marker VI.

Before using the lacZ method for purification the multiplicity of infection (m.o.i. phage to clone) to use for optimal growth had to be titrated: a sample of filtrate was always observed to check for good growth. It was found that a m.o.i. of ~x20, growing for 16 hours using fresh colonies for inoculation gave the best results.

Plate 13. Titration of culture growth.

1-3 No helper phage; 4-6 m.o.i. x10; 7-9 m.o.i. x20,

10-12 m.o.i. x100.

1, 2, 4, 5, 7, 8, 10 & 11 grown in filtration plate;

3, 6, 9 & 12 grown in bijou.

DNA molecular weight markers VI & III.

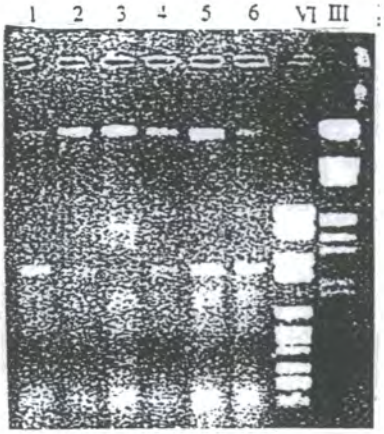


Plate 14. Growth of ssDNA.

1-6 10 µl samples of filtrates.

DNA molecular weight markers VI & III.

Filtering the cultures was compared with centrifugation, for removal of cells, and the subsequent purification using lacZ beads. Results obtained showed that the use of filtrates was better, although supernatants were often used to produce good templates.

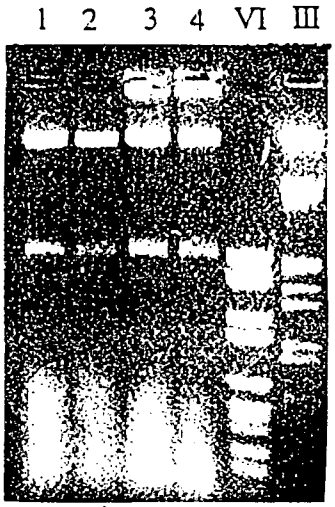


Plate 15. Filtration compared to supernatant: growth.

1 & 2 10 µl sample of filtrates;

3 & 4 10 µl sample of supernatants.

2 µl DNA markers VI & III.

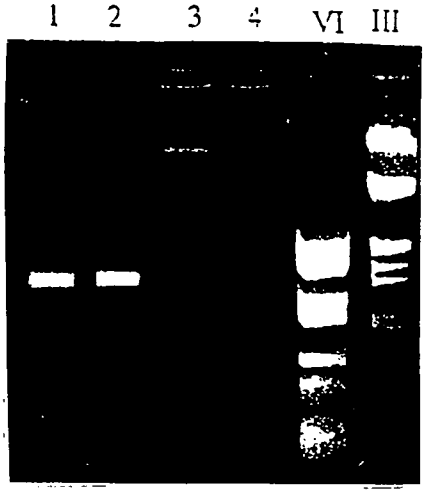


Plate 16. Filtration compared to supernatant: purified ssDNA.

1 & 2 2 µl sample of purified elute from filtrates;

3 & 4 2 µl sample of purified elute from supernatants.

2 µl DNA molecular weight markers VI & III.

It was shown that about 600 µl of filtrate and 90 µl of beads produced enough purified ssDNA template (~500 ng) for sequencing. Beads could be reused, and home made beads appeared to produce acceptable templates.

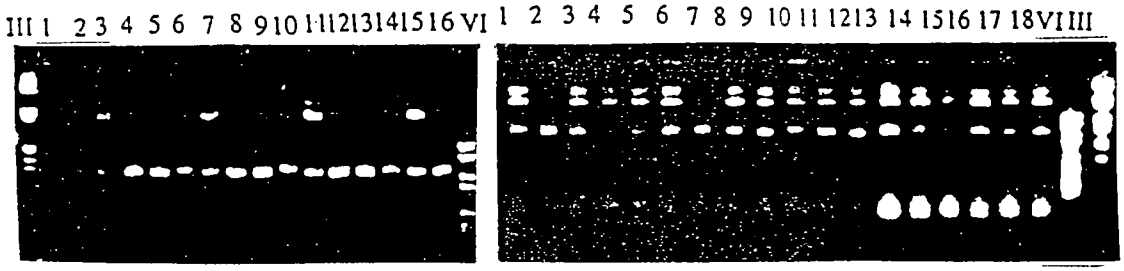


Plate 17. LacZ bead titration.

1-4 2 µl elute using 200 µl filtrate, 30 µl beads;

5-8 2 µl elute using 400 µl filtrate, 60 µl beads;

9-12 2 µl elute using 600 µl filtrate, 90 µl beads;

13-16 2 µl elute using 800 µl filtrate, 120 µl beads.

2 µl DNA markers VI & III.

Plate 18. Evaluation of lacZ beads.

1-6 2 µl elute using new beads;

7-12 2 µl elute reusing beads;

13-18 2 µl elute using customised beads;

2 µl DNA markers VI & III.

Plate 19. LacZ purified ssDNA.

1-8 2 µl elutes using 600µl filtrate, 90µl beads, clones from plate C7 (subset CA).

2 µl DNA molecular weight marker VI.

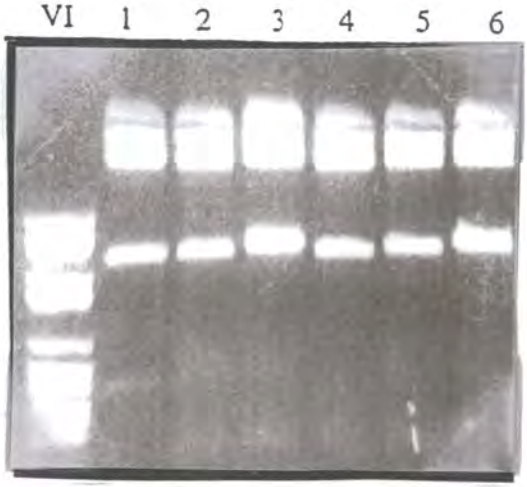


Plate 20. LacZ purified ssDNA.

1-6 2 µl elutes using 600 µl filtrate, 90 µl beads, clones from plate C7 (subset CA).

2 µl DNA molecular weight marker VI.

In total over 2000 ssDNA templates were produced by these two methods. The number of Ns (ambiguities called) between base 50 and 400 in the sequence was counted automatically by a small Unix program CDtotal, which only accepted the sequence into "good" if there was <14 Ns (4%) ambiguity.

There was some degree of artificial wastage due to experimentation with terminator clean up methods. It was found that a column alone was sufficient to clean reactions if the template was prepared by lacZ methods, but a phenol/chloroform purification and column (or two phenol chloroform purifications) was required for the Streptavidin templates, otherwise fluorescent artefacts were observed in the gel. Much variation was seen in the yield and purities eluted off the various columns tested: many parameters including buffer used and centrifugation speeds had a profound effect. Triethyl ammonium acetate (TEAA) was found to be one of the best buffers for use. Yields of sequencing products were sometimes low if purification procedures were not done to specification.

A total of 1552 "good" ESTs were generated, sequencing some clones in both directions. While using "Sedit" to cut pieces of vector and ambiguities from the sequences, Bluescript and adaptor sequences were often found at the beginning of the sequences, but ambiguities were normally the limiting factor at the end.

Preliminary investigations show that the ligation and PCR selected bases in each subset are seen in most cases. Where the start of the sequences are recognised, *Fok* I sites have been searched for: initial figures suggest cutting sites are seen in under the expected 50% of sequences at the expected position after the adaptor sequence at the start.

In general a high proportion of the ssDNA templates produced high quality fluorescent signals of EST data with low background impurities and ambiguity.

```
Final
CA
Number of good for CA.FOR:      303
Number of good for CA.REV:      345
Average of length:              677.41
Average of number of N:         6.78

CB
Number of good for CB.FOR:      234
Number of good for CB.REV:      88
Average of length:              645.18
Average of number of N:         5.97

CC
Number of good for CC.FOR:      42
Number of good for CC.REV:      62
Average of length:              600.30
Average of number of N:         6.49

CD
Number of good for CD.FOR:      66
Number of good for CD.REV:      64
Average of length:              784.35
Average of number of N:         5.55

CE
Number of good for CE.FOR:      67
Number of good for CE.REV:      37
Average of length:              753.02
Average of number of N:         8.92

M
Number of good for  M.FOR:      152
Number of good for  M.REV:      90
Average of length:              715.28
Average of number of N:         7.45

Number of good:                 1552
```

Figure 10. Final: The output produced in Unix by the program CDtotal.

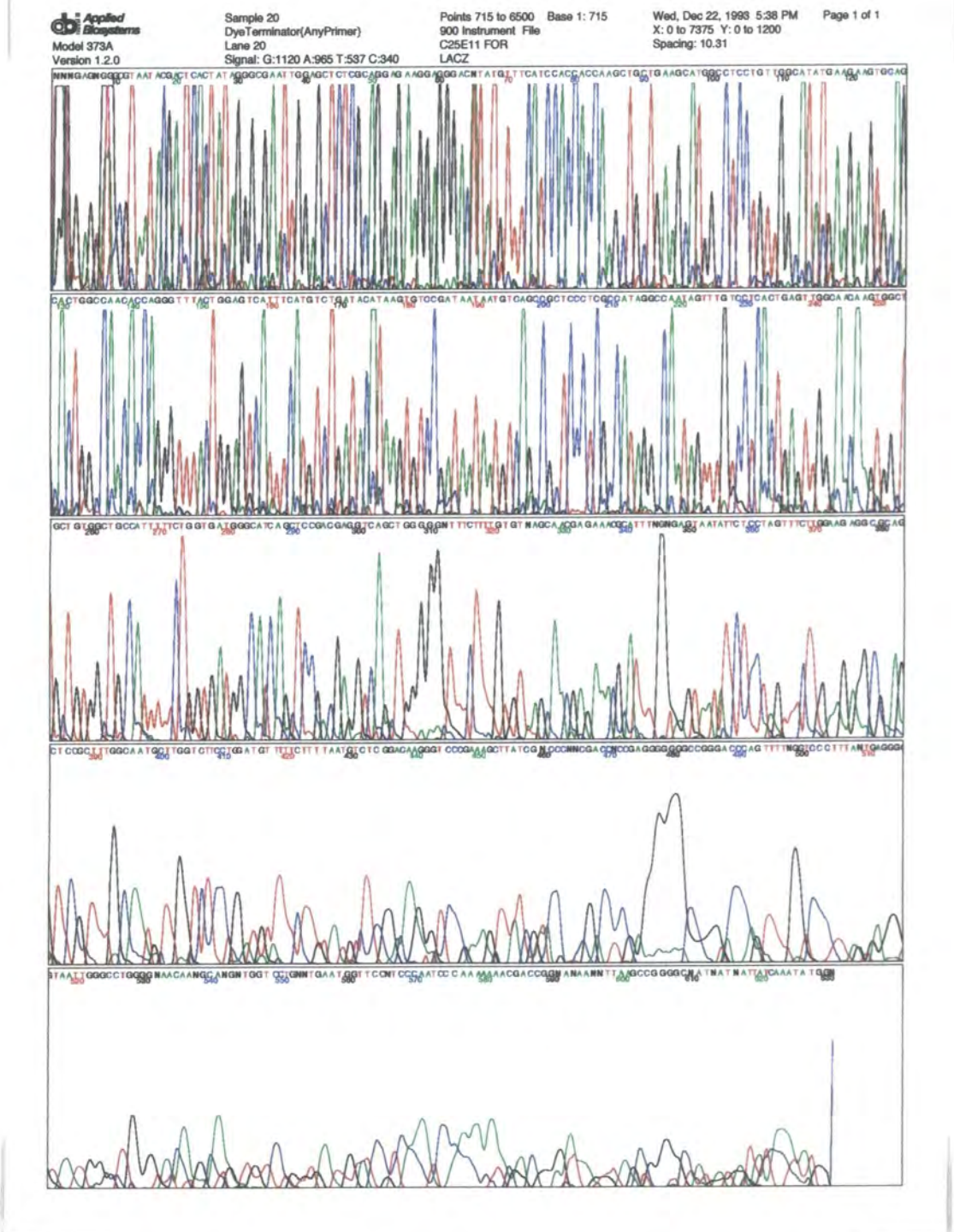


Figure 11. Automatic sequencer output for clone C25E11 (subset CE).

LacZ template, forward primer & terminator sequencing.

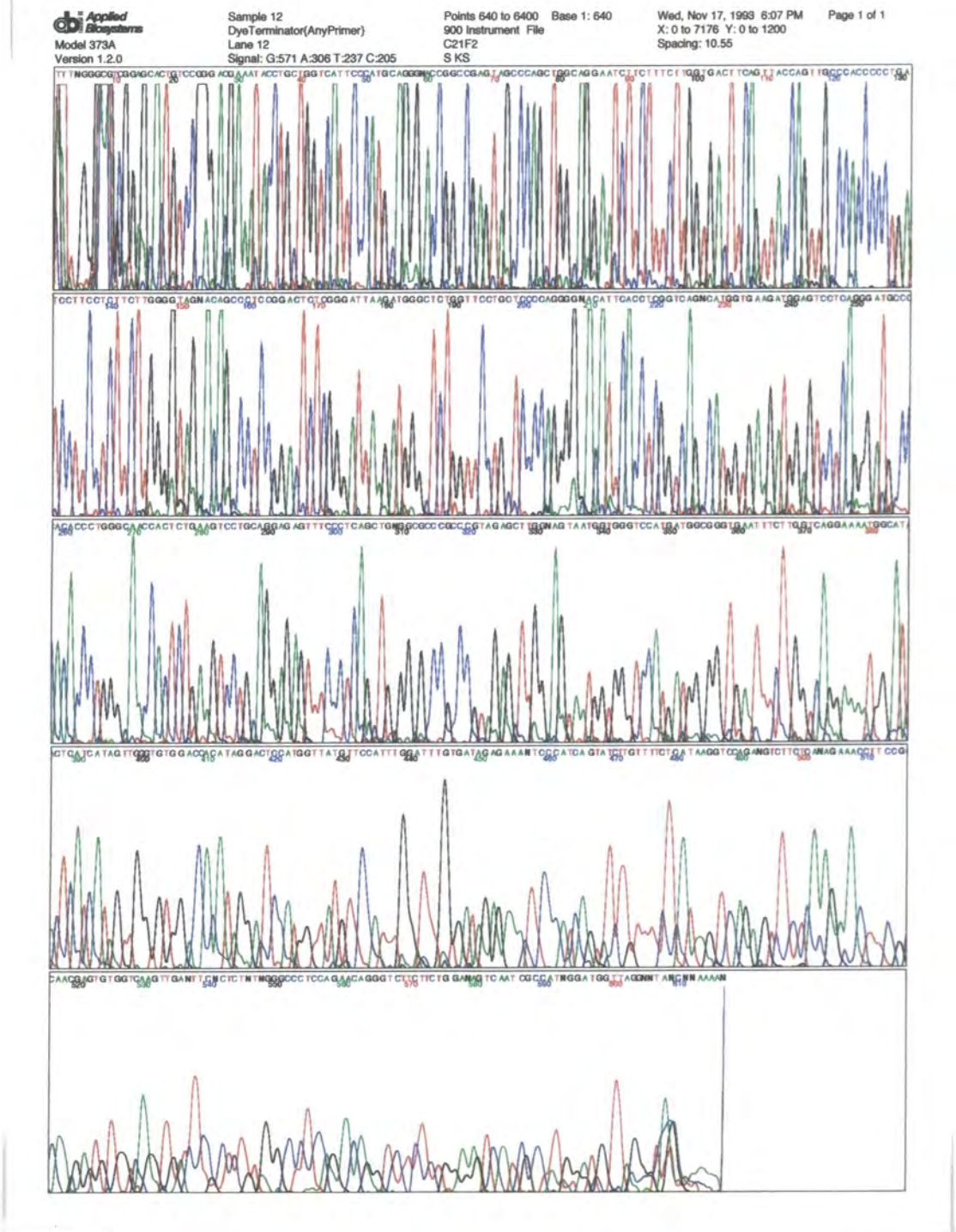


Figure 12. Automatic sequencer output for clone C21F02 (subset CC).

PCR streptavidin template, KS primer & terminator sequencing.

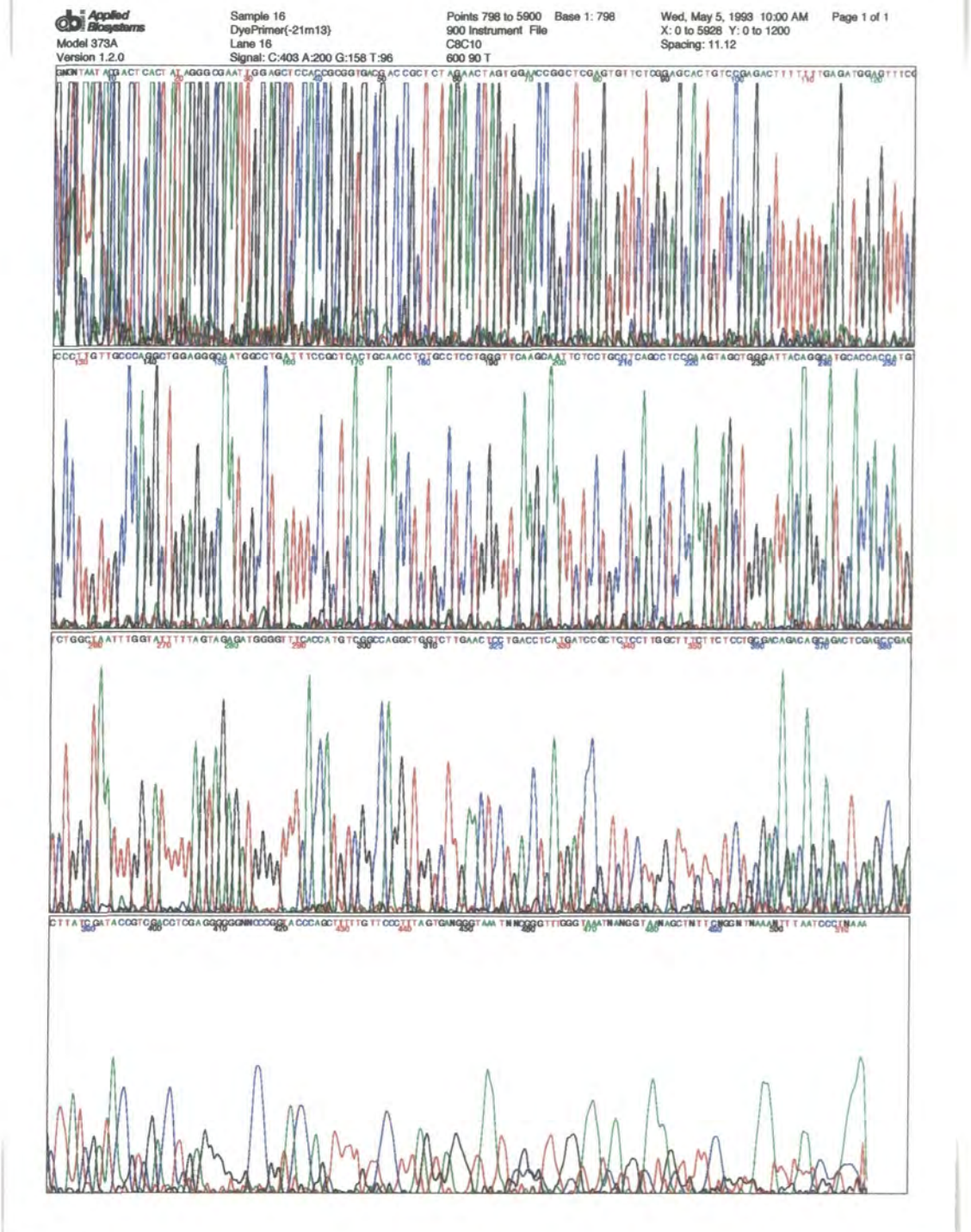


Figure 13. Automatic sequencer output for clone C08C10 (subset CA).

LacZ template, forward primer & primer sequencing.

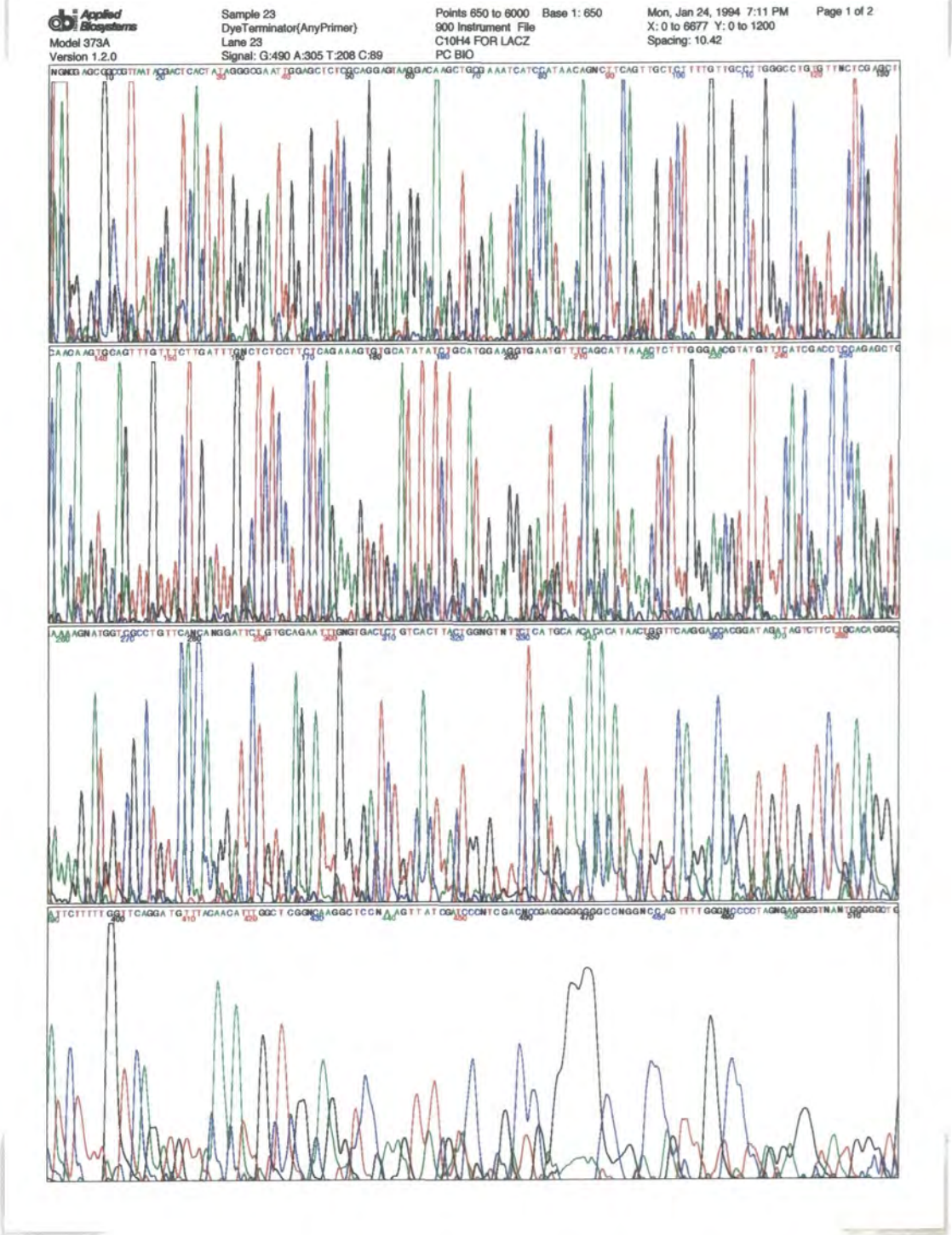


Figure 14. Automatic sequencer output for clone C10H04 (subset CA).

LacZ template, forward primer & terminator sequencing.

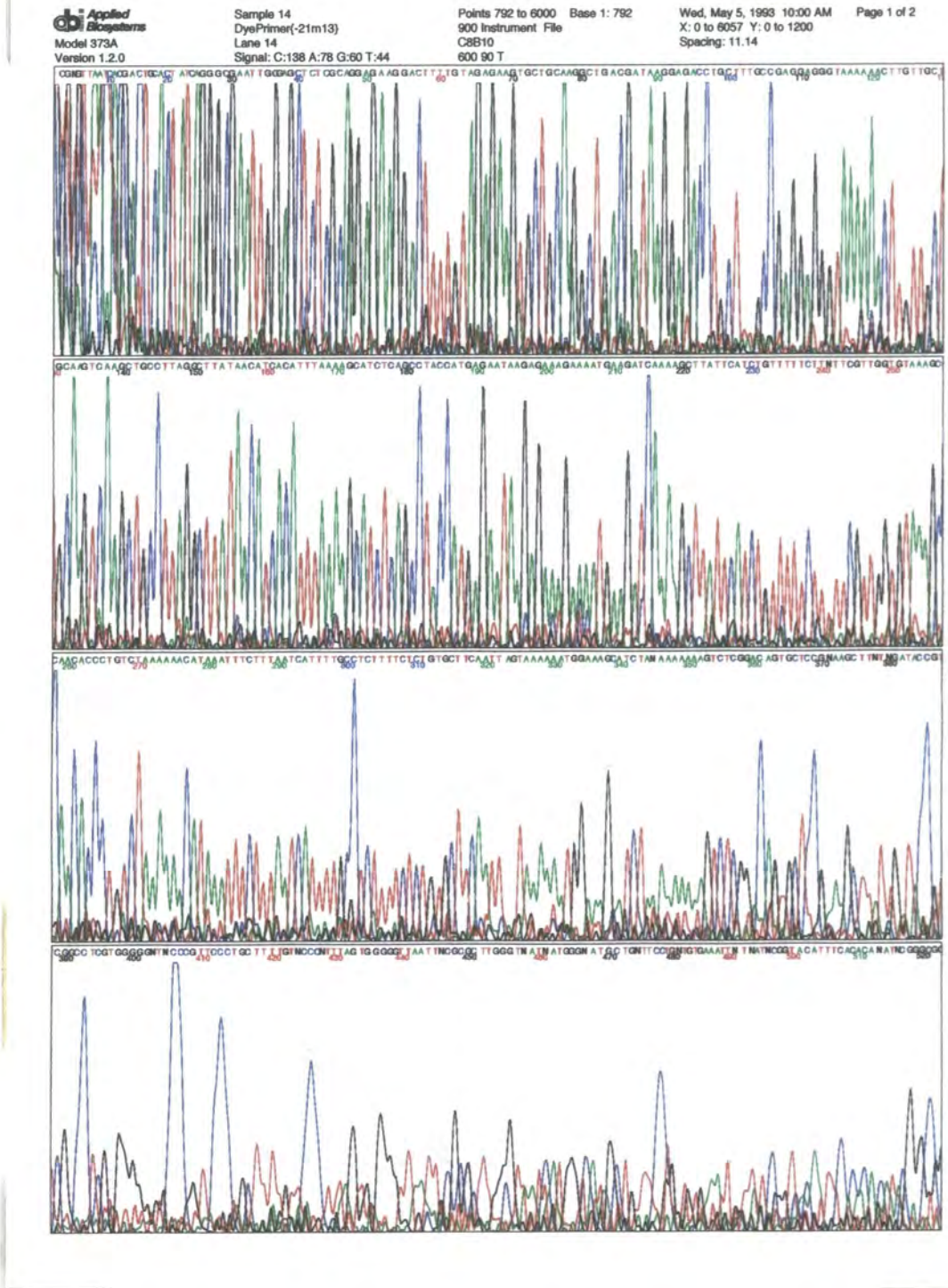


Figure 15. Automatic sequencer output for clone C08B10 (subset CA).

LacZ template, forward primer & primer sequencing.

3.3. Analysis of data

The 1552 "clean finished good" ESTs were put into the cDNA database: using BLAST they were compared with the sequences in the public databases for similarities, (GenBank November 1993 release contained 217 377 sequences and 219 858 629 total letters). Many significant similarities over 95% were obtained: the matches of 98-100% show the general quality of sequencing was very high.

```
Results of Database Search

Details of request AAAFRGR


Laboratory      : HGMP-RC
Library         : CE
Plate           : 25
Position        : E11
Species         : Human
Tissue          : Liver
Direction       : forward
Sequence        : GGGACATATGTTTCATCCACCACCAAGCTGCTGAAGCATGGCCTCCTGTT
                  GGCATATGAAGAAGTGCAGCACTGGCCAACACCAGGGTTTACTGGAGTCA
                  TTTCATGTCTGATACATAAGTGTCCGATAATAATGTCAGCCGCTCCCTCG
                  CCATAGGCCAATAGTTTGTCCTCACTGAGTTGGCAACAAGTGGCTGTTGT
                  GGCTGCCATTTTTCTGGTGATGGCCATCCAGCTCCGACGAGGTCAGCTGG
                  GGGGNTTTCTTTTGTGTAAGCAACGAGAAACGCATTTTGGGAGTAATATT
                  CTCCTAGTTTCTTGGAAGGAGGCCGCAGNTTCCGTTTTGGCAATGCTTGG
                  TCTCCTGGATGTTTTTCTTGTAATGTCTCGGANAAGGNTCCCGAAAGCTT
                  ATCGTCCCNNCGACCCCGAGGGGGGCCGGGNCCCAGTTTTGGTCCCCTTA
                  NNGAGGGTAAATGGGGNTGGGGNACAAGGCANGNTGGTCCNGNNTNAATG
                  GATCCTCANATCCCAAAAAACAACCGGNGAAANNNTAAGCCGGGGCTATN
                  ATNATATCAAATATGGGNGGGCATNCCTTCNGCGGAACCNNGGCANTNAT
                  ATATTGCACCGGNGGGGGGGTTGGNCNCCNCCCCCATCTCCC
```

Figure 16. Sequence of C25E11, AAAFRGR.

```
Results of Database Search

The following best matches are found for your requested sequence

Seqid    Database  Percent  Matching  Length  Description
AAAFRGR  genbank       99       227     228   human mrna encoding alpha-fetoprotein (afp)
AAAFRGR  genbank       99       227     228   human alpha-fetoprotein (afp) mrna, complet
AAAFRGR  genbank      100       141     141   human alpha-fetoprotein gene, complete cds.
AAAFRGR  genbank       99       140     141   gorilla alpha-fetoprotein (afp) gene, compl
AAAFRGR  genbank       71       164     228   mouse mrna encoding alpha-fetoprotein (a fe
AAAFRGR  genbank       70       160     228   rat alpha-fetoprotein mrna (partial).
AAAFRGR  genbank       70       160     228   rat messenger rna coding for alpha-fetoprot
AAAFRGR  genbank       70       160     228   rat mrna for alpha-fetoprotein (afp) (this
AAAFRGR  genbank       75       106     141   mouse alpha-fetoprotein (afp) gene, exon 12
AAAFRGR  genbank       73       109     149   rat alpha-fetoprotein 3-prime end mrna.
AAAFRGR  genbank       73       109     149   messenger rna for rat alpha-fetoprotein.
AAAFRGR  swissprot     98        74      75   alpha-fetoprotein precursor (alpha-fetoglob
AAAFRGR  swissprot     98        74      75   alpha-fetoprotein precursor (alpha-fetoglob
AAAFRGR  swissprot     54        48      88   alpha-fetoprotein precursor (alpha-fetoglob
AAAFRGR  swissprot     52        46      87   alpha-fetoprotein precursor (alpha-fetoglob
AAAFRGR  swissprot     43        36      83   serum albumin precursor.
AAAFRGR  swissprot     43        31      72   serum albumin precursor (fragment).
AAAFRGR  swissprot     43        32      73   serum albumin precursor.
AAAFRGR  swissprot     38        34      88   serum albumin precursor.
AAAFRGR  swissprot     37        33      88   serum albumin precursor.
AAAFRGR  swissprot     42        31      73   serum albumin precursor.
AAAFRGR  swissprot     45        32      71   serum albumin precursor.
AAAFRGR  cdna                                 matches to other cDNA's in the database.


23 entries in the database matched the search request.
```

Figure 17. The listed BLAST matches of AAAFRGR.

```
Results of Database Search

  AAAFRGR genbank
        > HSFETO V01514 Human mRNA encoding alpha-fetoprotein (AFP). AFP is a major
        serum protein (MG: 70000) synthesized during fetal life.
        Length = 2029

  Minus Strand HSPs:

 Score = 1131 (312.5 bits), Expect = 9.1e-87, P = 9.1e-87
 Identities = 227/228 (99%), Positives = 227/228 (99%), Strand = Minus

Query:   228 GATGGCCATCACCAGAAAAATGGCAGCCACAGCAGCCACTTGTTGCCAACTCAGTGAGGA 169
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  1385 GATGGCCATCACCAGAAAAATGGCAGCCACAGCAGCCACTTGTTGCCAACTCAGTGAGGA 1444

Query:   168 CAAACTATTGGCCTATGGCGAGGGAGCGGCTGACATTATTATCGGACACTTATGTATCAG 109
             ||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  1445 CAAACTATTGGCCTGTGGCGAGGGAGCGGCTGACATTATTATCGGACACTTATGTATCAG 1504

Query:   108 ACATGAAATGACTCCAGTAAACCCTGGTGTTGGCCAGTGCTGCACTTCTTCATATGCCAA 49
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  1505 ACATGAAATGACTCCAGTAAACCCTGGTGTTGGCCAGTGCTGCACTTCTTCATATGCCAA 1564

Query:    48 CAGGAGGCCATGCTTCAGCAGCTTGGTGGTGGATGAAACATATGTCCC 1
             ||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  1565 CAGGAGGCCATGCTTCAGCAGCTTGGTGGTGGATGAAACATATGTCCC 1612

 Score = 263 (72.7 bits), Expect = 9.0e-29, Poisson P(2) = 9.0e-29
 Identities = 59/67 (88%), Positives = 59/67 (88%), Strand = Minus

Query:   316 TCCAAGAAACTAGGAGAATATTACTCCCAAAATGCGTTTCTCGTTGCTTACACAAAAGAA 257
             | | ||||||||||||||||||||  ||||||||||||||||||||||||||||||  ||
Sbjct:  1299 TTCCAGAAACTAGGAGAATATTACTTACAAAATGCGTTTCTCGTTGCTTACACAAAGAAA 1358

Query:   256 ANCCCCC 250
               |||||
Sbjct:  1359 GCCCCCC 1365


> HUMALBAF4 J00077 Human alpha-fetoprotein (AFP) mRNA, complete cds.
          Length = 2032

  Minus Strand HSPs:

 Score = 1131 (312.5 bits), Expect = 9.1e-87, P = 9.1e-87
 Identities = 227/228 (99%), Positives = 227/228 (99%), Strand = Minus

Query:   228 GATGGCCATCACCAGAAAAATGGCAGCCACAGCAGCCACTTGTTGCCAACTCAGTGAGGA 169
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  1388 GATGGCCATCACCAGAAAAATGGCAGCCACAGCAGCCACTTGTTGCCAACTCAGTGAGGA 1447

Query:   168 CAAACTATTGGCCTATGGCGAGGGAGCGGCTGACATTATTATCGGACACTTATGTATCAG 109
             ||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  1448 CAAACTATTGGCCTGTGGCGAGGGAGCGGCTGACATTATTATCGGACACTTATGTATCAG 1507

Query:   108 ACATGAAATGACTCCAGTAAACCCTGGTGTTGGCCAGTGCTGCACTTCTTCATATGCCAA 49
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  1508 ACATGAAATGACTCCAGTAAACCCTGGTGTTGGCCAGTGCTGCACTTCTTCATATGCCAA 1567

Query:    48 CAGGAGGCCATGCTTCAGCAGCTTGGTGGTGGATGAAACATATGTCCC 1
             ||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  1568 CAGGAGGCCATGCTTCAGCAGCTTGGTGGTGGATGAAACATATGTCCC 1615
```

Figure 18. The best BLAST matches for AAAFRGR.

Results of Database Search

Details of request AAAFAYS

Laboratory	: HGMP-RC
Library	: CC
Plate	: 21
Position	: F02
Species	: Human
Tissue	: Liver
Direction	: reverse
Sequence	: ATTTAGNGCGTCGGAGCACTGTCCGGGACGAAATACCTGCTGGTCATTCC
	CATGCAGGGNACCGGCCGAGTAGCCCAGCTGGCAGGAATCTTCTTTCTTG
	GTGACTTCAGTTACCAGTTGCCCACCCCCTGATCCTTCCTCTTCTTGGGG
	TAGNACAGCCCTCCGGACTCTCGGGATTAAGATGGGCTCTGGTTCCTGCT
	CCCCAGGGGNACATTCACCTCGGTCAGCCATGGTGAAGATGGAGTCCTCA
	GGGATGCCCACACCCTGGGCAACCACTCTGAAGTCCTGCAGGAGAGTTTC
	CCTCAGCTGNGGCGCCCGCCCGTAGAGCTTGGCAGTAATGGTGGGTCCAT
	GATGGCGGGTGAATTTCTTGGTCAGGAAAATGGCATACTCATCATAGTTG
	GGTGTGGACCACATAGGACTCCATGGTTATGTTCCATTTGGATTTGTGAT
	AGAGAAANTCCCATCAGTATCTTGTTTCTCATAAGGTCCAGANGTCTTCT
	CANAGAAACCTTCCGGCAACGAGTGTGGTCAAGTTGATTTCCGCTCTNTG
	GGGCCCTCCAGAACAGGGTCTTCTCTGGGANAGTCAATCGCNNTGGGATN
	GTTAGGGGT

Figure 19. Sequence of C21F02, AAAFAYS.

Results of Database Search

The following best matches are found for your requested sequence

Seqid	Database	Percent	Matching	Length	Description
AAAFAYS	genbank	98	338	344	human mrna for alpha-1-microglobulin and
AAAFAYS	genbank	98	338	344	human mrna for protein hc (alpha-1-microg
AAAFAYS	genbank	83	280	334	porcine mrna for alpha-1 microglobulin/hi
AAAFAYS	genbank	83	280	334	porcine mrna for alpha1-microglobulin-bik
AAAFAYS	genbank	79	280	354	mesocricetus auratus mrna for precursor c
AAAFAYS	genbank	69	337	483	polyprotein 1-microglobulin/bikunin [rats
AAAFAYS	genbank	69	328	472	meriones unguiculatus mrna for precursor
AAAFAYS	genbank	67	319	472	m.musculus ambp mrna for alpha-1-microglc
AAAFAYS	genbank	97	106	109	(repeat region: alu) human gene for alpha
AAAFAYS	genbank	97	106	109	human inter-alpha-trypsin inhibitor light
AAAFAYS	genbank	77	128	165	rat alpha-1-microglobulin mrna.
AAAFAYS	swissprot	77	127	164	alpha-1-microglobulin / inter-alpha-tryps
AAAFAYS	swissprot	64	105	164	alpha-1-microglobulin / inter-alpha-tryps
AAAFAYS	swissprot	36	29	80	alpha-1-microglobulin / inter-alpha-tryps
AAAFAYS	swissprot	28	20	69	prostaglandin-h2 d-isomerase precursor (e
AAAFAYS	swissprot	36	16	44	von ebners gland protein precursor (veg p
AAAFAYS	swissprot	56	14	25	inter-alpha-trypsin inhibitor (iti) (ei-1
AAAFAYS	swissprot	33	9	27	progesterone receptor (pr) (forms a and b
AAAFAYS	swissprot	51	16	31	gooseberry distal protein (bsh9).
AAAFAYS	swissprot	55	10	18	major centromere autoantigen b (centromer
AAAFAYS	swissprot	31	18	57	immediate-early protein ie68 (orf4) (frag
AAAFAYS	swissprot	39	11	28	phosphoenolpyruvate carboxykinase, mitoch
AAAFAYS	cdna				matches to other cDNA's in the database.

23 entries in the database matched the search request.

Figure 20. The listed BLAST matches of AAAFAYS.

```
BLASTN 1.3.12 [29-Oct-93] [Build 15:48:47 Nov 25 1993]

Reference:  Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers,
and David J. Lipman (1990).  Basic local alignment search tool.  J. Mol. Biol.
215:403-410.

Notice:  this program is optimized to find nearly identical sequences rapidly.
To identify weak similarities encoded in nucleic acid, use BLASTX or TBLASTN.

Query=  CC-21-F02.REV.SEQ.comp, 609 bases, 4EC52932 checksum.
        (609 letters, both strands)

Database:  genbank
           217,377 sequences; 219,858,629 total letters.
Searching.....................................................done

                                                                    Smallest
                                                                    Poisson
                                                          High      Probability
Sequences producing High-scoring Segment Pairs:           Score     P(N)      N

HSA1MICR   X04494 Human mRNA for alpha-1-microglobulin an...  1666  1.1e-131  1
HSHCR      X04225 Human mRNA for protein HC (alpha-1-micr...  1666  1.1e-131  1
SSAMGBIK   X53685 Porcine mRNA for alpha1-microglobulin-b...  1184  2.6e-91   1
SSAMHI30   X52087 Porcine mRNA for alpha-1 microglobulin/...  1184  2.6e-91   1
HAMPCFTB   D31814 Mesocricetus auratus mRNA for precursor...  1104  3.8e-84   1
S87544     S87544 polyprotein 1-microglobulin/bikunin [ra...  1101  9.5e-84   1
MUIPCFTA   D31813 Meriones unguiculatus mRNA for precurso...  1064  6.1e-80   1
MMAMBP     X68680 M.musculus AMBP mRNA for alpha-1-microg...   983  9.9e-73   1
HSA1MBG1   X54816 Human gene for alpha-1-microglobulin-bi...   518  2.7e-51   2
HUMITILC05 M88246 Human inter-alpha-trypsin inhibitor lig...   518  2.4e-33   1
RATMGBA    J02600 Rat alpha-1-microglobulin mRNA.              492  3.6e-33   1
HSA1MBG3   X54818 Human gene for alpha-1-microglobulin-bi...   408  6.3e-24   1
HUMITILC08 M88249 Human inter-alpha-trypsin inhibitor lig...   408  6.4e-24   1
HUMITILC04 M88245 Human inter-alpha-trypsin inhibitor lig...   401  9.4e-24   1
HSA1MBG2   X54817 Human gene for alpha-1-microglobulin an...   248  1.3e-10   1
HUMITILC06 M88247 Human inter-alpha-trypsin inhibitor lig...   238  7.0e-10   1
SMOSPA1    L26598 Salmo salar alpha-1-microglobulin/bikun...   221  2.7e-08   1
HUMITILC03 M88244 Human inter-alpha-trypsin inhibitor lig...   210  2.1e-07   1

>HSA1MICR X04494 Human mRNA for alpha-1-microglobulin and HI-30.
          Length = 1221

  Plus Strand HSPs:

 Score = 1666 (460.3 bits), Expect = 1.1e-131, P = 1.1e-131
 Identities = 338/344 (98%), Positives = 338/344 (98%), Strand = Plus

Query:   209 CCAACTATGATGAGTATGCCATTTTCCTGACCAAGAAATTCACCCGCCATCATGGACCCA 268
            |||||||||||||||||||||||||||||||||||||||||||| ||||||||||||||||
Sbjct:   419 CCAACTATGATGAGTATGCCATTTTCCTGACCAAGAAATTCAGCCGCCATCATGGACCCA 478

Query:   269 CCATTACTGCCAAGCTCTACGGGCGGGCGCCNCAGCTGAGGGAAACTCTCCTGCAGGACT 328
            |||||||||||||||||||||||||||||||| ||||||||||||||||||||||||||
Sbjct:   479 CCATTACTGCCAAGCTCTACGGGCGGGCGCCGCAGCTGAGGGAAACTCTCCTGCAGGACT 538

Query:   329 TCAGAGTGGTTGCCCAGGGTGTGGGCATCCCTGAGGACTCCATCTTCACCATGGCTGACC 388
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   539 TCAGAGTGGTTGCCCAGGGTGTGGGCATCCCTGAGGACTCCATCTTCACCATGGCTGACC 598

Query:   389 GAGGTGAATGTNCCCCTGGGGAGCAGGAACCAGAGCCCATCTTAATCCCGAGAGTCCGGA 448
            ||||||||||| ||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   599 GAGGTGAATGTGTCCCTGGGGAGCAGGAACCAGAGCCCATCTTAATCCCGAGAGTCCGGA 658

Query:   449 GGGCTGTNCTACCCCAAGAAGAGGAAGGATCAGGGGGTGGGCAACTGGTAACTGAAGTCA 508
            ||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   659 GGGCTGTGCTACCCCAAGAAGAGGAAGGATCAGGGGGTGGGCAACTGGTAACTGAAGTCA 718

Query:   509 CCAAGAAAGAAGATTCCTGCCAGCTGGGCTACTCGGCCGGTNCC 552
            |||||||||||||||||||||||||||||||||||||||||| ||
Sbjct:   719 CCAAGAAAGAAGATTCCTGCCAGCTGGGCTACTCGGCCGGTCCC 762
```

Figure 21. The listed and best BLAST match for AAAFAYS.

Results of Database Search

Details of request AAAFJAU

Laboratory	: HGMP-RC
Library	: CA
Plate	: 8
Position	: C10
Species	: Human
Tissue	: Liver
Direction	: forward
Sequence	: CAAGCTGCGAAATCATCCATAACAGCTTTCAGTTGCTNTTTTGTTGCCTT GGGCTTGTGTTTCACAAGNTCAACAAGTGCAGTTTGTTTCTTGATTTGTC TCTCCTTCTCAGAAAGTGTGCATATATCTGCATGGAAGGTGAATGTCTCA GCATTAAACTCTTTGGGAACGTATGTTTCATCGACTTCCAGAGCTGAAAA GCATGGTCGCCTGTTCACCAAGGATTCTGTGCAGCATTTGGTGACTCTGT CNTNTACTGGCGTTTTCTCATGCAACACACATAACTGGCTCAGGACCACG GATAGATAGTCTTCTGCACAGGGCATTCTTTTTGCTTCAGGATGTTTACA ACATTTNGTC

Figure 22. Sequence of C08B10, AAAFJAU.

Results of Database Search

The following best matches are found for your requested sequence

Seqid	Database	Percent	Matching	Length	Description
AAAFJAU	genbank	98	348	355	sequence 1 from patent us 4914027.
AAAFJAU	genbank	98	348	355	h.sapiens (pxl53) gene for serum albumin.
AAAFJAU	genbank	98	348	355	h.sapiens (pxl53) gene for serum albumin
AAAFJAU	genbank	98	348	355	artificial sequence for the protein fusio
AAAFJAU	genbank	97	346	355	human serum albumin.
AAAFJAU	genbank	97	346	355	human messenger rna for serum albumin (hs
AAAFJAU	genbank	97	346	355	h.sapiens mrna for serum albumin.
AAAFJAU	genbank	97	346	355	human serum albumin (alb) mrna, complete
AAAFJAU	genbank	97	346	355	h.sapiens mrna for albumin.
AAAFJAU	genbank	97	335	342	artificial sequence for human serum album
AAAFJAU	genbank	92	330	357	macaca mulatta serum albumin mrna, 3 end.
AAAFJAU	swissprot	95	114	119	serum albumin precursor.
AAAFJAU	swissprot	77	91	118	serum albumin precursor.
AAAFJAU	swissprot	73	87	119	serum albumin precursor.
AAAFJAU	swissprot	73	87	119	serum albumin precursor.
AAAFJAU	swissprot	69	83	119	serum albumin precursor.
AAAFJAU	swissprot	72	86	119	serum albumin precursor (fragment).
AAAFJAU	swissprot	42	51	119	serum albumin precursor.
AAAFJAU	swissprot	47	55	116	alpha-fetoprotein precursor (alpha-fetogl
AAAFJAU	swissprot	46	54	116	alpha-fetoprotein precursor (alpha-fetogl
AAAFJAU	swissprot	40	47	116	alpha-fetoprotein precursor (alpha-fetogl
AAAFJAU	swissprot	41	48	115	74 kd serum albumin precursor.
AAAFJAU	cdna				matches to other cDNA's in the database.

23 entries in the database matched the search request.

Figure 23. The listed BLAST matches of AAAFJAU.

```
Results of Database Search

  AAAFJAU genbank
        > I01999 I01999 Sequence 1 from patent US 4914027.
       Length = 1880

  Minus Strand HSPs:

 Score = 1712 (473.1 bits), Expect = 3.1e-135, P = 3.1e-135
 Identities = 348/355 (98%), Positives = 348/355 (98%), Strand = Minus

Query:   356 AAATGTTGTAAACATCCTGAAGCAAAAAGAATGCCCTGTGCAGAAGACTATCTATCCGTG 297
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  1419 AAATGTTGTAAACATCCTGAAGCAAAAAGAATGCCCTGTGCAGAAGACTATCTATCCGTG 1478

Query:   296 GTCCTGAGCCAGTTATGTGTGTTGCATGAGAAAACGCCAGTANANGACAGAGTCACCAAA 237
            ||||||| ||||||||||||||||||||||||||||||||    ||||||||||||||
Sbjct:  1479 GTCCTGAACCAGTTATGTGTGTTGCATGAGAAAACGCCAGTAAGTGACAGAGTCACCAAA 1538

Query:   236 TGCTGCACAGAATCCTTGGTGAACAGGCGACCATGCTTTTCAGCTCTGGAAGTCGATGAA 177
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  1539 TGCTGCACAGAATCCTTGGTGAACAGGCGACCATGCTTTTCAGCTCTGGAAGTCGATGAA 1598

Query:   176 ACATACGTTCCCAAAGAGTTTAATGCTGAGACATTCACCTTCCATGCAGATATATGCACA 117
            |||||||||||||||||||||||||||| ||||||||||||||||||||||||||||||
Sbjct:  1599 ACATACGTTCCCAAAGAGTTTAATGCTGAAACATTCACCTTCCATGCAGATATATGCACA 1658

Query:   116 CTTTCTGAGAAGGAGAGACAAATCAAGAAACAAACTGCACTTGTTGANCTTGTGAAACAC 57
            |||||||||||||||||||||||||||||||||||||||||||||| ||||||||||||
Sbjct:  1659 CTTTCTGAGAAGGAGAGACAAATCAAGAAACAAACTGCACTTGTTGAGCTTGTGAAACAC 1718

Query:    56 AAGCCCAAGGCAACAAAANAGCAACTGAAAGCTGTTATGGATGATTTCGCAGCTT 2
            |||||||||||||||||| ||||||||||||||||||||||||||||||||||||
Sbjct:  1719 AAGCCCAAGGCAACAAAAGAGCAACTGAAAGCTGTTATGGATGATTTCGCAGCTT 1773


> A03758 A03758 H.sapiens (pXL53) gene for serum albumin.
       Length = 2136

  Minus Strand HSPs:

 Score = 1712 (473.1 bits), Expect = 3.5e-135, P = 3.5e-135
 Identities = 348/355 (98%), Positives = 348/355 (98%), Strand = Minus

Query:   356 AAATGTTGTAAACATCCTGAAGCAAAAAGAATGCCCTGTGCAGAAGACTATCTATCCGTG 297
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  1419 AAATGTTGTAAACATCCTGAAGCAAAAAGAATGCCCTGTGCAGAAGACTATCTATCCGTG 1478

Query:   296 GTCCTGAGCCAGTTATGTGTGTTGCATGAGAAAACGCCAGTANANGACAGAGTCACCAAA 237
            ||||||| ||||||||||||||||||||||||||||||||    ||||||||||||||
Sbjct:  1479 GTCCTGAACCAGTTATGTGTGTTGCATGAGAAAACGCCAGTAAGTGACAGAGTCACCAAA 1538

Query:   236 TGCTGCACAGAATCCTTGGTGAACAGGCGACCATGCTTTTCAGCTCTGGAAGTCGATGAA 177
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  1539 TGCTGCACAGAATCCTTGGTGAACAGGCGACCATGCTTTTCAGCTCTGGAAGTCGATGAA 1598

Query:   176 ACATACGTTCCCAAAGAGTTTAATGCTGAGACATTCACCTTCCATGCAGATATATGCACA 117
            |||||||||||||||||||||||||||| ||||||||||||||||||||||||||||||
Sbjct:  1599 ACATACGTTCCCAAAGAGTTTAATGCTGAAACATTCACCTTCCATGCAGATATATGCACA 1658

Query:   116 CTTTCTGAGAAGGAGAGACAAATCAAGAAACAAACTGCACTTGTTGANCTTGTGAAACAC 57
            |||||||||||||||||||||||||||||||||||||||||||||| ||||||||||||
Sbjct:  1659 CTTTCTGAGAAGGAGAGACAAATCAAGAAACAAACTGCACTTGTTGAGCTTGTGAAACAC 1718
```

Figure 24. The best BLAST matches for AAAFJAU.

Results of Database Search

Details of request AAAFOCG

Laboratory : HGMP-RC
Library : CA
Plate : 10
Position : H04
Species : Human
Tissue : Liver
Direction : forward
Sequence : CAAGCTGCGNAAATCATCCATAACAGCCTTCAGTTGCTCTTTTGTTGCCT
TGGGCCTGTGTTNCTCGAGCTCAACAAGTGCAGTTTGTTTCTNGATTTGN
CTCTCCTTCTCAGAAAGTGTGCATATATCTGCATGGAAGGTGAATGTTTC
AGCATTAAACTCTTTGGGAACGTATGTTTCATCGACCTCCAGAGCTGAAA
AGNATGGTCGCCTGTTCACCANGGATTCTGTGCAGAATTTGGGTGACTCT
GTCACTTACTGGCGTTTTCTCATGCAACACACATAACTGGTTCAAGGACC
ACGGATAGATAGTCTTCTTGCACAGGGCATTCTTTTTGGTTCAGGATGTT
TACAACATTTGGCTTCGGNCAAGGGCTCCNAAAGGTTATCGATACCNTCG
ACNCCGAGGGGGGGGCCCGNACCCAGTTTTGGGNCCCTAGNGAGGGGTNA
NTGGGCCTGGGGNATCAATGGCANANTTGTCCCNGGTNAATTGGNACCCC
NCANAATCCCANNAAANCGGCCGGGGGATAAAGNTNTAACNGGGGGGCCA
ANNNTNANTANCAATAATGGNTGGGNTCNCTCCCTCNCCGGGACCTNNGN
CAATTATANNTTCGCACCCGGNGGGGNGGGGTTTGNNCCCCTCCCCCA

Figure 25. Sequence of C10H04, AAAFOCG.

Results of Database Search

The following best matches are found for your requested sequence

Seqid	Database	Percent	Matching	Length	Description
AAAFOCG	genbank	95	222	232	human messenger rna for serum albumin (hs
AAAFOCG	genbank	95	221	232	sequence 1 from patent us 4914027.
AAAFOCG	genbank	95	221	232	artificial sequence for human serum album
AAAFOCG	genbank	95	221	232	h.sapiens (pxl53) gene for serum albumin.
AAAFOCG	genbank	95	221	232	h.sapiens (pxl53) gene for serum albumin
AAAFOCG	genbank	95	221	232	artificial sequence for the protein fusic
AAAFOCG	genbank	94	219	232	human serum albumin.
AAAFOCG	genbank	94	219	232	h.sapiens mrna for serum albumin.
AAAFOCG	genbank	94	219	232	human serum albumin (alb) mrna, complete
AAAFOCG	genbank	94	219	232	h.sapiens mrna for albumin.
AAAFOCG	genbank	90	210	232	macaca mulatta serum albumin mrna, 3 end.
AAAFOCG	swissprot	89	68	76	serum albumin precursor.
AAAFOCG	swissprot	72	55	76	serum albumin precursor.
AAAFOCG	swissprot	72	54	74	serum albumin precursor.
AAAFOCG	swissprot	71	54	76	serum albumin precursor.
AAAFOCG	swissprot	69	53	76	serum albumin precursor.
AAAFOCG	swissprot	70	52	74	serum albumin precursor (fragment).
AAAFOCG	swissprot	47	37	78	alpha-fetoprotein precursor (alpha-fetogl
AAAFOCG	swissprot	46	36	78	alpha-fetoprotein precursor (alpha-fetogl
AAAFOCG	swissprot	35	26	73	serum albumin precursor.
AAAFOCG	swissprot	39	31	78	alpha-fetoprotein precursor (alpha-fetogl
AAAFOCG	swissprot	38	30	78	alpha-fetoprotein precursor (alpha-fetogl
AAAFOCG	cdna				matches to other cDNA's in the database.

23 entries in the database matched the search request.

Figure 26. The listed BLAST matches of AAAFOCG.

```
Results of Database Search

  AAAFOCG genbank
        > HSALB1 V00494 Human messenger RNA for serum albumin (HSA).
       Length = 2055

  Minus Strand HSPs:

 Score = 1070 (295.7 bits), Expect = 2.8e-80, P = 2.8e-80
 Identities = 222/232 (95%), Positives = 222/232 (95%), Strand = Minus

Query:   242 CCAAATTCTGCACAGAATCCNTGGTGAACAGGCGACCATNCTTTTCAGCTCTGGAGGTCG 183
            |||||| |||||||||||||| |||||||||||||||||| |||||||||||||| ||||
Sbjct:  1532 CCAAATGCTGCACAGAATCCTTGGTGAACAGGCGACCATGCTTTTCAGCTCTGGAAGTCG 1591

Query:   182 ATGAAACATACGTTCCCAAAGAGTTTAATGCTGAAACATTCACCTTCCATGCAGATATAT 123
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  1592 ATGAAACATACGTTCCCAAAGAGTTTAATGCTGAAACATTCACCTTCCATGCAGATATAT 1651

Query:   122 GCACACTTTCTGAGAAGGAGAGNCAAATCNAGAAACAAACTGCACTTGTTGAGCTCGAGN 63
            |||||||||||||||||||||| |||||| |||||||||||||||||||||||||||| |
Sbjct:  1652 GCACACTTTCTGAGAAGGAGAGACAAATCAAGAAACAAACTGCACTTGTTGAGCTCGTGA 1711

Query:    62 AACACAGGCCCAAGGCAACAAAAGAGCAACTGAAGGCTGTTATGGATGATTT 11
            |||||| |||||||||||||||||||||||| ||||||||||||||||
Sbjct:  1712 AACACAAGCCCAAGGCAACAAAAGAGCAACTGAAAGCTGTTATGGATGATTT 1763

 Score = 272 (75.2 bits), Expect = 2.8e-30, Poisson P(2) = 2.8e-30
 Identities = 56/58 (96%), Positives = 56/58 (96%), Strand = Minus

Query:   296 CTTGAACCAGTTATGTGTGTTGCATGAGAAAACGCCAGTAAGTGACAGAGTCACCCAA 239
            | ||||||||||||||||||||||||||||||||||||||||||||||||||||||| ||
Sbjct:  1479 CCTGAACCAGTTATGTGTGTTGCATGAGAAAACGCCAGTAAGTGACAGAGTCACCAAA 1536

 Score = 206 (56.9 bits), Expect = 8.5e-32, Poisson P(3) = 8.5e-32
 Identities = 42/43 (97%), Positives = 42/43 (97%), Strand = Minus

Query:   361 CAAATGTTGTAAACATCCTGAACCAAAAAGAATGCCCTGTGCA 319
            ||||||||||||||||||||| ||||||||||||||||||||
Sbjct:  1416 CAAATGTTGTAAACATCCTGAAGCAAAAAGAATGCCCTGTGCA 1458


> I01999 I01999 Sequence 1 from patent US 4914027.
       Length = 1880

  Minus Strand HSPs:

 Score = 1061 (293.2 bits), Expect = 1.8e-79, P = 1.8e-79
 Identities = 221/232 (95%), Positives = 221/232 (95%), Strand = Minus

Query:   242 CCAAATTCTGCACAGAATCCNTGGTGAACAGGCGACCATNCTTTTCAGCTCTGGAGGTCG 183
            |||||| |||||||||||||| |||||||||||||||||| |||||||||||||| ||||
Sbjct:  1534 CCAAATGCTGCACAGAATCCTTGGTGAACAGGCGACCATGCTTTTCAGCTCTGGAAGTCG 1593

Query:   182 ATGAAACATACGTTCCCAAAGAGTTTAATGCTGAAACATTCACCTTCCATGCAGATATAT 123
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  1594 ATGAAACATACGTTCCCAAAGAGTTTAATGCTGAAACATTCACCTTCCATGCAGATATAT 1653

Query:   122 GCACACTTTCTGAGAAGGAGAGNCAAATCNAGAAACAAACTGCACTTGTTGAGCTCGAGN 63
            |||||||||||||||||||||| |||||| |||||||||||||||||||||||||||| | |
Sbjct:  1654 GCACACTTTCTGAGAAGGAGAGACAAATCAAGAAACAAACTGCACTTGTTGAGCTTGTGA 1713

Query:    62 AACACAGGCCCAAGGCAACAAAAGAGCAACTGAAGGCTGTTATGGATGATTT 11
            |||||| |||||||||||||||||||||||| ||||||||||||||||
Sbjct:  1714 AACACAAGCCCAAGGCAACAAAAGAGCAACTGAAAGCTGTTATGGATGATTT 1765
```

Figure 27. The best BLAST matches for AAAFOCG.

Results of Database Search

Details of request AAAFIYE

Laboratory	: HGMP-RC
Library	: CA
Plate	: 8
Position	: B10
Species	: Human
Tissue	: Liver
Direction	: forward
Sequence	: CGAGCGGCACCAACAGCGGCGCGAGCAAGAAGCGCTTTGAAGTGAAAAAG
	TGGAATGCAGTAGCCCTCTGGNCGTGGGATATTGTGGTTGATAACTGTGC
	CATCTGCAGGAACCACATTATGGATCTTTGCATAGAATGTCAAGCTAACC
	AGGCGTCCGCTACTTCAGAAGAGTGTACTGTCGCATGGGGAGTCTGTAAC
	CATGCTTTTCACTTCCACTGCATCTCTCGCTGGCTCAAAACACGACAGGT
	GTGTCCATTGGCAACAGAGAGTGGGAATTCCAAAAGTATGGGCACTAGGA
	AAAGACTTCTTNCATCAAGCTTTAATTGTTTTGTTATTCATTTAATGACT
	TTCCTNCTGTTACCTAATTACAAATTGGATGGGACTGTGTTTT

Figure 28. Sequence of C08B10, AAAFIYE.

Results of Database Search

The following best matches are found for your requested sequence

Seqid	Database	Percent	Matching	Length	Description
AAAFIYE	genbank	98	258	261	testican [human, testis, mrna, 3484 nt].
AAAFIYE	genbank	77	183	236	5c01h06-t7 zea mays cdna clone 5c01h06 5
AAAFIYE	swissprot	47	10	21	hypothetical 18.8 kd protein zk637.14 in
AAAFIYE	swissprot	34	16	46	goliath protein (g1 protein).
AAAFIYE	swissprot	59	13	22	trans-acting transcriptional protein icp0
AAAFIYE	swissprot	59	13	22	trans-acting transcriptional protein icp0
AAAFIYE	swissprot	50	11	22	trans-acting transcriptional protein icp0
AAAFIYE	swissprot	40	9	22	hypothetical 87.9 kd protein precursor in
AAAFIYE	swissprot	50	11	22	trans-acting transcriptional protein icp0
AAAFIYE	swissprot	27	6	22	beta-galactosidase (ec 3.2.1.23) (lactase
AAAFIYE	swissprot	58	7	12	60s ribosomal protein l9.
AAAFIYE	cdna				matches to other cDNA's in the database.

12 entries in the database matched the search request.

Figure 29. The listed BLAST matches of AAAFIYE.

```
Results of Database Search

  AAAFIYE genbank
        > S62147 S62147 testican [human, testis, mRNA, 3484 nt].
       Length = 3484

  Minus Strand HSPs:

 Score = 1278 (353.1 bits), Expect = 6.2e-99, P = 6.2e-99
 Identities = 258/261 (98%), Positives = 258/261 (98%), Strand = Minus

Query:   261 CCAATGGACACACCTGTCGTGTTTTGAGCCAGCGAGAGATGCAGTGGAAGTGAAAAGCAT 202
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:    14 CCAATGGACACACCTGTCGTGTTTTGAGCCAGCGAGAGATGCAGTGGAAGTGAAAAGCAT 73

Query:   201 GGTTACAGACTCCCCATGCGACAGTACACTCTTCTGAAGTAGCGGACGCCTGGTTAGCTT 142
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:    74 GGTTACAGACTCCCCATGCGACAGTACACTCTTCTGAAGTAGCGGACGCCTGGTTAGCTT 133

Query:   141 GACATTCTATGCAAAGATCCATAATGTGGTTCCTGCAGATGGCACAGTTATCAACCACAA 82
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   134 GACATTCTATGCAAAGATCCATAATGTGGTTCCTGCAGATGGCACAGTTATCAACCACAA 193

Query:    81 TATCCCACGNCCAGAGGGCTACTGCATTCCACTTTTTCACTTCAAAGCGCTTCTTGCTCG 22
             ||||||| | ||||||||||||||||||||||||||||||||||||||||||||||| ||
Sbjct:   194 TATCCCAGGCCCAGAGGGCTACTGCATTCCACTTTTTCACTTCAAAGCGCTTCTTGCCCG 253

Query:    21 CGCCGCTGTTGGTGCCGCTCG 1
             |||||||||||||||||||||
Sbjct:   254 CGCCGCTGTTGGTGCCGCTCG 274


> T18274 T18274 5c01h06-t7 Zea mays cDNA clone 5c01h06 5' end.
       Length = 450

  Plus Strand HSPs:

 Score = 703 (194.3 bits), Expect = 1.0e-50, P = 1.0e-50
 Identities = 183/236 (77%), Positives = 183/236 (77%), Strand = Plus

Query:    26 CAAGAAGCGCTTTGAAGTGAAAAAGTGGAATGCAGTAGCCCTCTGGNCGTGGGATATTGT 85
             ||| |||||||| ||  | || |||||||| || || || |||||| | |||||||| ||
Sbjct:   198 CAACAAGCGCTTCGAGATCAAGAAGTGGAACGCCGTCGCGCTCTGGGCATGGGATATCGT 257

Query:    86 GGTTGATAACTGTGCCATCTGCAGGAACCACATTATGGATCTTTGCATAGAATGTCAAGC 145
             || || ||||| ||||||||| | |||||||| |||||||| ||||| || || || ||
Sbjct:   258 CGTCGACAACTGCGCCATCTGCCGCAACCACATCATGGATCTATGCATCGAGTGCCAGGC 317

Query:   146 TAACCAGGCGTCCGCTACTTCAGAAGAGTGTACTGTCGCATGGGGAGTCTGTAACCATGC 205
             ||||| ||   ||| ||    || ||||| |||||||| | ||| |||||||| |||||
Sbjct:   318 GAACCAAGCTAGCGCGACCAGCGAGGAGTGCACTGTCGCTTNGGGTGTCTGTAATCATGC 377

Query:   206 TTTTCACTTCCACTGCATCTCTCGCTGGCTCAAAACACGACAGGTGTGTCCATTGG 261
             |||||||||||||||||    | ||||| || || || || ||||| ||||| |
Sbjct:   378 TTTTCACTTCCACTGCATCAGCAGGTGGCTTAAGACTCGCCAAGTGTGCCCATTAG 433
```

Figure 30. The best BLAST matches for AAAFIYE.

From the 1552 sequences consensus sequences were produced from ESTs sequenced from both ends. These 1167 consensus sequences (screened rigorously for unwanted sequences including vector and alu repeats) were then further analysed in their specific semi-ordered adaptor and primer selective "subsets" with ICATOOLS which looks for family type similarities. Sequences were grouped into classes (with a parent sequence) showing similarity; these sequences matched the same sequences in the database as would be expected.

Subset	**CA**	**CB**	**CC**	**CD**	**CE**	**M**	**TOTAL**
Adaptor	**TT**	**AA**	**TT**	**TT**	**AA**	**TT**	**TT or AA**
Primer 1	**C**	**C**	**G**	**G**	**G**	**C**	**C or G**
Primer 2	**C**	**C**	**C**	**A**	**C**	**C**	**A or C**
Plates	**C1-C16**	**C17-C20**	**C21-C22**	**C23-C24**	**C25-C26**	**M1-M4**	**C1-C26 M1-M4**
Number of sequences	**529**	**221**	**75**	**79**	**71**	**192**	**1167**
Number of different sequence classes	209	145	53	49	52	120	636
% different sequence classes	**40**	**66**	**71**	**62**	**73**	**67**	**54**
Number of unique sequences	159	124	46	38	43	103	513
% unique sequences	**30**	**56**	**61**	**48**	**61**	**54**	**44**
Number of unknown cDNAs	88	18	7	9	3	27	152
% unknown cDNAs	**17**	**8**	**9**	**11**	**4**	**14**	**13**
Number of unique unknown new sequences	54	14	5	6	2	15	96
% unique unknown new sequences	**10**	**6**	**7**	**8**	**3**	**8**	**8**
Number of mitochondrion/ mitochondria sequences	5	1	10	7	3	1	27
% mitochondrion/ mitochondria sequences	**0.9**	**0.5**	**13.3**	**8.9**	**4.2**	**0.5**	**2.3**

Table 3. Composition of fetal tissue cDNA fragment subset sequences.

Unique sequences were of no similarity to any others sequenced in this study. Sequence classes were defined by similarity between families by ICATOOLS. Unknowns showed no match to a cDNA of known purpose, whilst unique unknowns showed no match at all.

Subset CA

Adaptor: TT

Primer 1: C

Primer 2: C

Number of sequences: 529

Parent	Family	Number	%	Description
CA-10-H04	205	113	21.4	Human messenger RNA for serum albumin (1532-1763)
CA-10-F05	188	29	5.5	Human mRNA for alpha-1-microglobulin (245-575)
CA-10-B04	202	23	4.3	Human coagulation factor V mRNA, complete cds
CA-10-H02	204	12	2.3	Human mRNA for ankyrin (variant 2.1)
CA-11-H03	153	11	2.1	*A. officinalis* mRNA for asparagine
CA-10-D02	199	9	1.7	Human 90-kDa heat-shock protein gene (1977-2245)
CA-13-C11	169	9	1.7	Mouse mRNA for protein C, complete cds
CA-10-E03	120	9	1.7	Testican [human, testis, mRNA, 3484 nt]
CA-10-H09	158	8	1.5	Rat mRNA for ribosomal protein S5
CA-11-H04	146	8	1.5	*H. sapiens* partial cDNA sequence; clone 96A07
CA-11-C02	135	8	1.5	IB2318 *H. sapiens* cDNA 3' end similar to EST06268

Sequence	Database	%	Matching	Length	ID and Brief Description
CA-10-H04	genbank	95	222	232	V00494 serum albumin
CA-10-F05	genbank	90	301	331	X04494 alpha-1-microglobulin
CA-10-B04	genbank	98	191	193	M16967 coagulation factor V
CA-10-H02	genbank	95	62	65	X16609 ankyrin (variant 2.1)
CA-11-H03	genbank	67	42	62	X67958 asparagine
CA-10-D02	genbank	98	264	269	M16660 heat-shock protein)
CA-13-C11	genbank	77	38	49	D10445 protein C
CA-10-E03	genbank	99	238	240	S62147 testican
CA-10-H09	genbank	86	227	262	X58465 ribosomal protein S5
CA-11-H04	genbank	92	197	214	Z25122 partial cDNA 96A07
CA-11-C02	genbank	97	117	120	T15970 cDNA EST06268

Table 4. Composition of fetal tissue cDNA fragments in subset CA.

The eleven most frequent sequence classes, and percent best match results of a database search for these are listed. Base co-ordinates defining GenBank entries, which characterise cDNA sequence classes, are given in parenthesis to distinguish between different fragments of a given transcript.

Subset CB

Adaptor: AA

Primer 1: C

Primer 2: C

Number of sequences: 221

Parent	Family	Number	%	Description
CB-18-C03	97	25	11.3	Seq2269 *Homo sapiens* cDNA clone b4HFLSK
CB-17-F03	130	11	5.0	Human mRNA for aldolase B
CB-17-B11	101	10	4.5	Sequence 1 from patent US 4839283
CB-17-A06	142	5	2.3	Human mRNA for hepatic triglyceride lipase (HTGL)
CB-17-G08	116	5	2.3	mRNA for translationally controlled tumour protein
CB-18-D10	7	4	1.8	Human mRNA for 1-alpha-1-antitrypsin
CB-17-H10	121	4	1.8	Human mRNA for S-protein
CB-17-B09	111	4	1.8	*H. sapiens* NAP (nucleosome assembly protein)
CB-17-D08	91	3	1.4	EST06257 *Homo sapiens* cDNA clone HIBBD2 5' end
CB-18-A08	53	3	1.4	Human transaldolase mRNA

Sequence	Database	%	Matching	Length	ID and Brief Description
CB-18-C03	genbank	96	128	132	T24081 cDNA clone b4HFLSK
CB-17-F03	genbank	98	202	205	X02747 aldolase B
CB-17-B11	genbank	98	302	306	I01352 US patent 4839283
CB-17-A06	genbank	93	293	313	X07228 HTGL
CB-17-G08	genbank	99	347	348	X16064 tumour protein
CB-18-D10	genbank	99	262	263	A01846 1-alpha-1-antitrypsin
CB-17-H10	genbank	90	220	243	X03168 mRNA for S-protein
CB-17-B09	genbank	91	216	236	M86667 *H. sapiens* NAP
CB-17-D08	genbank	97	38	39	T08366 cDNA clone HIBBD2
CB-18-A08	genbank	95	112	117	L19437 human transaldolase

Table 5. Composition of fetal tissue cDNA fragments in subset CB.

The ten most frequent sequence classes, and percent best match results of a database search for these are listed. Base co-ordinates defining GenBank entries, which characterise cDNA sequence classes, are given in parenthesis to distinguish between different fragments of a given transcript.

Subset CC

Adaptor: TT

Primer 1: G

Primer 2: C

Number of sequences: 75

Parent	Family	Number	%	Description
CC-21-G08	17	10	13.3	Human messenger RNA for serum albumin (1424-1734)
CC-21-H05	53	5	6.7	Human mitochondrion, complete genome (6382-6827)
CC-21-F02	51	5	6.7	Human mRNA for alpha-1-microglobulin (419-762)
CC-21-D09	10	3	4.0	EST276 *Homo sapiens* cDNA clone 10G11
CC-21-B03	48	2	2.7	Human ribosomal protein L12 mRNA complete cds
CC-21-F03	43	2	2.7	Human HL60 3' directed *Mbo* I cDNA HUMGS01951
CC-21-C03	35	2	2.7	Human HL60 3' directed *Mbo* I cDNA HUMGS01373

Sequence	Database	%	Matching	Length	ID and Brief Description
CC-21-G08	genbank	92	288	311	V00494 serum albumin
CC-21-H05	genbank	94	217	230	J01415 mitochondrion
CC-21-F02	genbank	98	338	344	X04494 alpha-1-microglobulin
CC-21-D09	genbank	97	102	105	T24701 cDNA clone 10G11
CC-21-B03	genbank	96	212	219	L06505 ribosomal protein L12
CC-21-F03	genbank	83	51	61	D20969 HL60 3' cDNA clone
CC-21-C03	genbank	79	51	64	D20399 HL60 3' cDNA clone

Table 6. Composition of fetal tissue cDNA fragments in subset CC.

The seven most frequent sequence classes, and percent best match results of a database search for these are listed. Base co-ordinates defining GenBank entries, which characterise cDNA sequence classes, are given in parenthesis to distinguish between different fragments of a given transcript.

Subset CD

Adaptor: TT

Primer 1: G

Primer 2: A

Number of sequences: 79

Parent	Family	Number	%	Description
CD-23-F02	17	10	12.7	*H. sapiens* (pXL53) for serum albumin (1409-1773)
CD-23-F05	45	8	10.1	Human mRNA for alpha-1-microglobulin (245-647)
CD-23-D02	40	7	8.9	Human mitochondrion, complete genome (7028-7246)
CD-23-B08	49	2	2.5	
CD-23-A05	37	2	2.5	Human mRNA for alpha 1-antitrypsin (899-1316)
CD-23-F01	35	2	2.5	*H. sapiens* partial cDNA sequence clone HECD09
CD-23-D01	28	2	2.5	Human HepG2 3' directed *Mbo* I cDNA clone hm01d05
CD-23-C06	27	2	2.5	EST276 *Homo sapiens* cDNA clone 10G11
CD-23-B05	26	2	2.5	Human insulinoma rig-analogue mRNA encoding DBP
CD-23-E12	2	2	2.5	Human protein kinase mRNA (1472-1734)
CD-23-G05	13	2	2.5	Human H3.3 histone class C mRNA complete cds

Sequence	Database	%	Matching	Length	ID and Brief Description
CD-23-F02	genbank	97	356	365	A03758 serum albumin
CD-23-F05	genbank	96	387	403	X04494 alpha-1-microglobulin
CD-23-D02	genbank	99	218	219	J01415 mitochondrion
CD-23-B08	genbank				
CD-23-A05	genbank	96	403	418	X01683 alpha 1-antitrypsin
CD-23-F01	genbank	92	165	178	Z17841 partial cDNA sequence
CD-23-D01	genbank	97	131	135	D11765 HepG2 cDNA
CD-23-C06	genbank	97	102	105	T24701 cDNA clone 10G11
CD-23-B05	genbank	95	313	327	J02984 insulinoma rig-analogue
CD-23-E12	genbank	97	256	263	M59287 protein kinase mRNA
CD-23-G05	genbank	97	334	343	M11353 H3.3 histone class C

Table 7. Composition of fetal tissue cDNA fragments in subset CD.

The eleven most frequent sequence classes, and percent best match results of a database search for these are listed. Base co-ordinates defining GenBank entries, which characterise cDNA sequence classes, are given in parenthesis to distinguish between different fragments of a given transcript.

Subset CE

Adaptor: AA

Primer 1: G

Primer 2: C

Number of sequences: 71

Parent	Family	Number	%	Description
CE-25-E11	35	12	16.9	Human mRNA encoding alpha-fetoprotein
CE-25-A03	44	2	2.8	Human glycine-gamma-globin, 3' end (272-536)
CE-25-D03	43	2	2.8	EST346 Homo sapiens cDNA clone 12H7
CE-25-E03	41	2	2.8	Human mRNA for transferrin receptor
CE-25-H10	37	2	2.8	Human mRNA for asialoglycoprotein receptor (597-837)
CE-25-E09	30	2	2.8	Human mitochondrion, complete genome (15360-15632)
CE-25-C08	3	2	2.8	EST01423 *Homo sapiens* cDNA clone HRBBA08
CE-25-G02	13	2	2.8	Human humFib mRNA for fibrillarin
CE-25-E05	11	2	2.8	Human messenger RNA for alpha-globin (189-553)

Sequence	Database	%	Matching	Length	ID and Brief Description
CE-25-E11	genbank	99	227	228	V01514 alpha-fetoprotein
CE-25-A03	genbank	90	241	265	M15386 glycine-gamma-globin
CE-25-D03	genbank	94	121	128	T24771 cDNA clone 12H7
CE-25-E03	genbank	97	81	83	X01060 transferrin receptor
CE-25-H10	genbank	99	239	241	X55284 asialoglycoprotein
CE-25-E09	genbank	97	266	273	J01415 mitochondrion
CE-25-C08	genbank	98	276	279	M79268 cDNA HRBBA08
CE-25-G02	genbank	96	170	176	X56597 fibrillarin
CE-25-E05	genbank	96	353	365	V00493 alpha-globin

Table 8. Composition of fetal tissue cDNA fragments in subset CE.

The nine most frequent sequence classes, and percent best match results of a database search for these are listed. Base co-ordinates defining GenBank entries, which characterise cDNA sequence classes, are given in parenthesis to distinguish between different fragments of a given transcript.

Subset M

Adaptor: TT

Primer 1: C

Primer 2: C

Number of sequences: 192

Parent	Family	Number	%	Description
M-03-E06	126	24	12.5	*H. sapiens* (pXL53) gene for serum albumin (1548-1773)
M-02-E12	110	7	3.6	Bovine rotavirus outer capsid protein (VP4) mRNA
M-03-G12	95	4	2.1	Human 90-kDa heat-shock protein gene (1985-2245)
M-02-C11	121	4	2.1	IB3518 *H. sapiens* cDNA 3' end similar to EST04951
M-03-D11	109	4	2.1	*H. sapiens* putatively transcribed partial cDNA Z21370
M-03-A05	96	3	1.6	Human RNA polymerase subunit hRPB 33, mRNA
M-02-H02	82	3	1.6	Testican [human, testis, mRNA, 3484 nt]
M-02-B08	45	3	1.6	Seq2269 *Homo sapiens* cDNA clone b4HFLSK
M-02-D06	125	3	1.6	Human phospholipase C-gamma mRNA, complete cds
M-02-F05	123	3	1.6	Human mRNA for alpha-1-microglobulin (245-694)
M-02-G08	117	3	1.6	*H. sapiens* partial cDNA sequence; clone 96A07

Sequence	Database	%	Matching	Length	ID and Brief Description
M-03-E06	genbank	96	219	226	A03758 serum albumin
M-02-E12	genbank	72	35	48	M92986 outer capsid protein
M-03-G12	genbank	96	251	261	M16660 heat-shock protein
M-02-C11	genbank	94	287	304	T16120 IB3518 cDNA
M-03-D11	genbank	94	149	158	Z21370 partial cDNA
M-03-A05	genbank	87	249	283	J05448 RNA polymerase
M-02-H02	genbank	97	191	195	S62147 testican
M-02-B08	genbank	95	239	251	T24081 cDNA clone b4HFLSK
M-02-D06	genbank	66	47	71	M34667 phospholipase
M-02-F05	genbank	97	375	384	X04495 alpha-1-microglobulin
M-02-G08	genbank	91	62	68	Z25122 partial cDNA

Table 9. Composition of fetal tissue cDNA fragments in subset M.

The eleven most frequent sequence classes, and percent best match results of a database search for these are listed. Base co-ordinates defining GenBank entries, which characterise cDNA sequence classes, are given in parenthesis to distinguish between different fragments of a given transcript.

After the rigorous BLAST and ICATOOLS sequence analysis 636 (54%) different sequence classes were identified; 513 (44%) of these ESTs were only seen once. Various percentages of different sequence classes were generated for each of the subsets, averaging about 67% except for CA with only 40%. The same pattern was seen with the percentages of unique sequences, averaging 56% except for 30% of the ESTs being unique in the CA subset. The mixed tissue subset data was not significantly different to that generated by the liver tissue subsets.

The biological purpose of 152 (13%) of the ESTs generated in this study was unknown, whilst 96 (8%) of the sequences were totally new: unique to the public databases and still of unknown function representing yet to be determined genes. Significant differences in the proportions of unknowns were seen across the different subsets.

Only a very small amount of mitochondrion sequences (27, 2.3%) were produced. These were again unevenly distributed in the different subsets suggesting selection for these cDNAs being exhibited in some cases giving a figure inproportionately high for the amount of mitochondrial species present. No ESTs generated showed similarity matches with database intron sequences, showing successful removal of DNA

Within each subset no antisense sequences were identified, suggesting positive semi-ordering of the cDNA fragments was achieved. The profiles of the compositions of each of the subsets was also very significantly different as seen in the previous six tables. Indeed the populations sequenced from each subset were quite different; ICATOOLS was used to show this by analysis of some subsets combined together as well as separately.

Subsets	**CA &CB**	**CA & CC**	**CA & CD**
Adaptor	**TT & AA**	**TT & TT**	**TT & TT**
Primer 1	**C & C**	**C & G**	**C & G**
Primer 2	**C & C**	**C & C**	**C & A**
Number of different sequence classes	209 & 145	209 & 53	209 & 49
Total number of different sequence classes	354	262	258
Combined number of different sequence classes	328	249	253
% overlap between sequence classes	**7.3**	**5.0**	**1.9**

Table 10. Overlap between cDNA fragment subset sequences.

Unique sequences were of no similarity to any others sequenced in this study. Sequence classes were defined by similarity between families by ICATOOLS.

The specificity of the cDNA fragment sorting was effective as the amount of overlap of sequence classes between pairs of subsets was shown to be only 5% on average. Different parts of the sequences such as those for serum albumin, alpha-fetaprotein, and alpha-1-microglobulin were seen in a number of sequences in different subsets; other sequence families were perhaps also biased towards and identified more than once. In general many different ESTs were generated of much biological value.

4. DISCUSSION

4.1. Methods utilised and results obtained

Whilst the fastidious extraction of high quality RNA by two fractionations, and other purification stages is a lengthy process, it is imperative to the production of a good cDNA library. The oligo $(dT)_{25}$ beads method of mRNA purification from total RNA produces a very pure product, using many washes before the final elution to wash off and dilute any possible contaminants. Other quicker methods and similar bead methods exist to produce mRNA direct from tissue, and are successful but obviously prone to contamination with other types of RNA and DNA.

As PCR is used subsequently any small amounts of unwanted products would be magnified to observable levels, and so the inclusion of traditional based methods involving centrifugation is expedient. Considering the amount of time spent utilising the library, all expenditure in adding value to it by quality production procedures is energy well directed.

The method used was shown to be worthwhile and successful as only small numbers of mitochondrial sequences were generated, and these might have been selected for in some of the subsets. This is a significant achievement as any mitochondrial contamination present would have high levels of expression; the redundancy of sequencing is also reduced significantly by its absence. The fact that no intron matching ESTs were seen was good, but even more indisputable evidence that no genomic DNA derived cDNA was present could be shown. Firstly the control would be a PCR using the cDNA as template between two facing primers on a known regular house keeping gene within the gene exon, before moving both primer sites down a number of bases until one is within an identified intron for another PCR. No exponentially generated product would be expected to be seen in the second instance if no intron DNA was present. Other EST projects have sometimes generated known intron sequences.

The production of a ready to use, quality controlled vector on a large scale was of great importance. Previous methods involved ligating four different adaptors on to a relatively small amount of double digested vector to create the necessary overhangs for subsequent ligation with the prepared cDNAs fragments. This obviously introduced the inherent risks involved at a late stage, including no guarantee of 100% successful ligation or

removal of unsuccessful substandard vector. The new vector is easily produced in large quantities, with a simple cut and column purification. The use of this new vector reduced the occurrence of no cDNA insert in white clones; this reduced wastage and increased efficiency.

In this study ssDNA templates were generated from the cDNA clones because in general they produce better quality and longer ESTs. Plasmid preparation methods were still very labour intensive often involving centrifugation, were not 96 format compatible and therefore of low throughput. The template purification methods used were 96 format compatible, and automation could often be used to reduce manpower required. Perkin-Elmer 9600 thermal cyclers were used for the PCRs in a 96 format, and then the Beckman Biomek automated workstation could be used for the capture of the PCR products and washing steps.

In contrast with other large studies, helper phage was used with pBluescript to grow ssDNA for purification to generate ESTs. This meant the cDNAs could be kept in a good storage plasmid vector whilst being able to produce ssDNA with out recloning. The innovative growth of the DNA within the filtration plates and subsequent filtration reduced the transference steps and the need for inefficient awkward centrifugation. A Tecan Robotic Sample Processor with disposable tip option available at the Resource Centre was used for some transference steps. Robotic steps are often slow but do not involve user real time, so ultimately increase throughput and reproducibility. Much titration of relative amounts of helper and beads required was needed to optimise the protocols to a robust procedure. Customised lacZ beads could be made, but not to such a guaranteed quality, that gives a reliable sequence template.

As the templates were ssDNA and of good general quality, *Taq* terminator cycle sequencing with dye labelled primer in a one tube reaction could be used. Removal of excess terminators was not a trivial task, but was decided the preferred option rather than four separate reactions requiring pooling and ethanol precipitation required for primer cycle sequencing. KS dye labelled primer kits were not commercially available either, whereas this preferred customised primer could be used with the available dye labelled terminator kits. The length of read and even signal may have been compromised by the use of terminator cycle sequencing, but this chemistry completely eliminates labelled false stops. Sequenase T7 sequencing could have been utilised for increased accuracy of base calling and longer read length, but a much greater amount of ssDNA template would have been required and costs would have increased. The Applied

Biosystems four dye per lane technology was mainly used to achieve a greater throughput. Up to two ABI 373A DNA sequencers were available for real time data collection, and were used for full scans (24 or 36 samples) and the quicker basesprinter (18 samples) runs.

Various excess terminator removal methods were evaluated: both phenol/chloroform extractions and spin column purification were both found to be successful, the preferred option dependent on the user and daily throughput. Columns were found more suitable once the apparatus for large numbers was set up, specifications for column type was known and large numbers were routinely required. Using columns reduced the user variability found using phenol/chloroform and ethanol precipitations, provided that the exact successful specifications for the use of columns was strictly adhered to. Purification columns were used 96 (24x4 with the appropriate bench top centrifuge inserts) at a time for efficiency. The quality of the ssDNA could be checked prior to the sequencing reaction to reduce wastage, some cDNAs would not produce suitable templates. The PCR based method complemented the lacZ method which could only generate the one strand suitable for sequencing. The rate limiting step varied mainly from template production, the excess terminator removal or the running of samples on the automatic sequencers.

As the public sequence databases get bigger and contain more sequences (GenBank April 1995 release contained 352 414 sequences and 286 094 556 bases, about 20% human), it obviously becomes much harder to generate new sequences of unknown biological purpose.

The restriction fragment sorting yielded distinct semi-ordered subpopulations of the cDNA fragments, produced by cutting the cDNA with a type IIS restriction endonuclease, *Fok* I. The fidelity of the T4 DNA ligase catalysed reaction is the basis for the sorting of the cDNA fragments into distinct subsets by two base specific adaptoring. Specific primer annealing and extension are the prerequisites for the partitioning of the cDNA fragments by PCR employing two single base specific primers.

The definition of a different sequence was rigorous in this study as ICATOOLS was additionally used to give families of sequence classes, but on average 54% of all ESTs generated were of a different class. A total percentage of 44% were generated once only. The variation across the subsets might be due to the final numbers in each; it may

be that as more of each subset are sequenced, the chance of sequencing the same species is greater, explaining the lower figures for subset CA. Data needs be analysed in real time with regard to sequencing, so numbers against different classes could be plotted for each subset to be in direct comparison to confirm this. A more complex population would produce redundant ESTs at a later stage, this might be the case for subset M: the complexity of a given subset may be further enlarged by increasing the heterogeneity of the source material from which the RNA was isolated. From the data it is suggested that sequencing no more than 500 sequences from each subset would be optimal for the greatest efficiency in finding different sequence classes in over 50% of ESTs, with a high proportion of unique sequences. Indeed the rationale for the restriction fragment sorting technology is that combining distinct cDNA subpopulations, of relatively low complexity, ensures a greater representation of transcribed sequences than that achieved by conventional libraries. Subtraction methods or pre-screening methods would be required for conventional libraries to avoid redundancy increasing significantly as numbers of ESTs generated from the library increases.

If small numbers (hundreds) of different subsets are continually sequenced no great increase in the present low rate of redundancy of output should be encountered, as different populations are proven to have overlap of only about 5% (1.9, 5.0 and 7.3 for those studied). Many different subsets involving different selective adaptors, primers, and different tissues can easily be generated for sequencing purposes. The small amount of overlap between subsets is attributable to misligation, and the extension of PCR primers with a 3' end mismatch. The amplification of adaptored fragments may exaggerate the otherwise undetectable effects of misligation. The incorporation of PCR into the sorting procedure provides a means of enriching the proportion of low abundance transcripts: the proportions of unknown sequences and mitochondrial sequences may reflect this effect.

In total 13% of the ESTs generated represented potentially 152 unidentified genes, and 8%, 96 new gene sequences. This is a relatively high return for the amount of sequencing, which still shows the overall value of ESTs at this stage in the HGP compared to the slower genome sequencing. More interesting and biologically significant sequences of greater use to mapping strategies are efficiently realised, which is important until the required technologies are in place to sequence much faster and cheaper.

Variation of unknown cDNAs seen in each of the subpopulations is explained by the distribution of selecting bases created by the *Fok* I cuts. The differences in the numbers of mitochondrial sequences in each subset can be explained by this specific selection for them in certain fragment populations. This also explains the high numbers of some species in some subsets too, which is not a true reflection of expression levels. For example 21% of the sequences in CA shows similarities to serum albumin which is produced in fetal liver, but is not evident in subset CE where alpha-fetoprotein sequences were predominant (17%). Proof of the correct base selection and hypothesis for positive subdivision could be further proven by searching for the *Fok* I restriction sites in the exon sites, looking at the four base overhang produced and seeing whether this is as would be expected to have been selected for by the respective primers and adaptors. The sorting technology including base specific ligation may have many other applications in molecular biology including fragment profiling and sequencing.

Successful data management for collection, early analysis, editing, database searching and analysis is required for a productive EST program. Such large amounts of data are produced, that efficient and proficient use of it must be made to optimise its biological value. This was achieved quite successfully, with some parts evolving as required and deemed appropriate. Complete integration of the system so the informatics worked on a biological level to produce exactly what was required, whilst being able to adapt to future possible so far unidentified needs was necessary. Much forethought, planning and collaboration was obviously prudent in this area. The cDNA database is accessible to all Resource Centre users and will soon be available on the world wide web (www) via the HGMP RC information services. The provision of this data and the biological resources in the form of cDNA clones or PCR products on high density nylon membranes is important for future advances to be made in adding biological value such as mapping information or tissue localisation to the ESTs.

Good quality RNA was fastidiously made to make the semi-ordered restriction fragment sorted cDNA library. An improved vector was used to produce the cDNA clones, from which many quality ssDNA templates were routinely made for automatic sequencing using further developed methods. Much good quality EST data was successfully generated from semi-ordered populations using these protocols and the latest technologies. Detailed analysis revealed the great value of the innovative library

production and template technologies, and the overall strategy. Indeed combining data from such semi-ordered populations and different tissues increases the complexity of a cDNA library immensely and hence minimises the redundancy encountered in systematic sequencing programmes, which was the objective. Ultimately this provides a faster route to identification of new transcribed sequences, and genes to significantly forward the Human Genome Project.

4.2. Further experiments and future developments

A concerted effort towards a public-domain human gene map and amassing information on human genes is now being made by many scientific institutions and private companies working together (*Genome Digest*, 1995). Shared materials and co-ordination of activities minimise overlap; and the information gained will be put into public databases. Various consortia are exchanging information whilst pursuing different but complementary strategies. cDNA libraries and the short identifying partial sequences of cDNAs, ESTs, are the starting material for many of the strategies: mapping them back to their origin in the genome by hybridisation or PCR approaches. By amplifying the corresponding sequence from genomic DNA using suitably designed primers to the EST, it is converted to a "sequence tagged site" (STS) which can be mapped to a location on the genome. The EST database (dbEST) at the US National Center for Biotechnology Information (NCBI) contains over 8500 although the number of unique genes represented is much smaller than this as more than one EST can be derived from a single gene. The mapping of ESTs has only recently begun in earnest but over 1100 have been mapped so far. This is seen as the best way forward to rapid assembly of a partial transcription map of the genome; to complement the HGP sequencing of the whole human genome.

The Merck Expressed Gene Initiative is aiming to isolate one cDNA clone for each expressed human gene, firstly obtaining ESTs from over 200 000 clones. Highly expressed genes are likely to be over represented, so techniques such as oligonucleotide fingerprinting and subtractive hybridisation procedures will be used to identify clones representing more rare expressed genes. A set of unique genes (the Merck Gene Index), using the longest cDNAs will then be built.

The IMAGE consortium is using shared arrayed cDNA libraries for gene sequencing, mapping and expression studies, to gain the maximum amount of information from the same resource. One pass sequencing of the 3' ends generated nearly 28 000 usable sequences, with many not identical or similar to known genes. About 10 000 of the new human gene transcripts were registered with the European Bioinformatics Institute (EBI), and over 3000 of these had been assigned to chromosomes by somatic cell hybridisation. All information on these transcripts has been registered with the Genome Data Base (GDB): IMAGE is also providing the cDNA clones for the Merck Initiative.
The international EST consortium co-ordinated by the Sanger Centre, includes laboratories from around the world, and is working in concert with IMAGE. The goal is to complete a comprehensive framework map of the human genome including the majority of genes, integrated with genetic markers such as microsatellite repeats within the next couple of years. The map should be much denser than current genetic maps with a resolution of 0.5 Mb. The strategy is to map genes by using tag sequences from the 3' ends of cDNAs (ESTs) and from CpG islands (genomic regions that signal the presence of a gene). These gene tag sequences are used to develop STSs which are mapped by PCR with radiation hybrid panels or against YACs.

Integration of the information emerging from the different mapping approaches is vital in order to build up a complete gene map. The same STSs will be mapped onto the CEPH mega-YACs and the Caltech bacterial artificial chromosome (BAC) clones as well as the radiation hybrids. As progress is made on the gene map it will provide reagents and much information that can be applied towards the eventual goal of sequencing the genome. Indeed the sequence level map of the entire human genome with 99% coverage and 99.99% accuracy in the coding regions is still estimated to be complete within the next 5-10 years. The information generated by the public domain gene mapping projects is accessible from the dbEST at the NCBI, which is mirrored at the EBI. Further development of databases to keep up with all the data produced and possible other information like expression data is necessary. Support in funding for these projects is still forthcoming from many sources including the NIH National Center for Human Genome Research, the UK Medical Research Council (MRC), the US DOE and the Wellcome Trust. The European Commissions new Biomedicine and Health Research Programme (Biomed 2) is also injecting money towards transnational genome research

projects. Rapid developments are presently being made in the human gene map initiative, with much hard data being produced with these strategies.

The many ESTs generated are simultaneously being used for many other purposes, including the determination of full length cDNA and genomic sequences of the mapped genes. These can be used to obtain expression of the proteins, to determine their normal biological functions, accelerate disease gene associations, and their RNA and protein products. The gene or gene fragments may then be used for many other developments including mapping, tissue typing, individual or forensic identification, and locating gene regions associated with genetic disease. Sequence information culminates in the complete genetic sequence but other advances in biological discovery such as understanding of function and the use of the genes in diagnosis and treatment of disease requires further imagination.

There have been many recent developments in sequencing technologies and automation (robotics) that enable the faster generation of greater numbers of better ESTs more economically. This will lead to rapid completion of sequencing unknown cDNAs and the eventual sequencing of the whole human genome: indeed the emphasis is already turning from mapping to large scale sequencing (*Science*, 1995). Some specific advances have been made with regard to high throughput ds DNA and ss DNA template preparation methods, including 96 format plasmid preparations (i.e. Qiagen). Complete automation using robots such as the Amersham Vistra and Perkin-Elmer Turbo Catalyst are perceived to generate a large number of quality templates. Sequencing chemistries have continually been improving to give a better performance and higher productivity. For example formulations including inosine or deaza compounds helped to eliminate compression problems: and a new enzyme formulation, Perkin-Elmer AmpliTaq DNA Polymerase CS+ should significantly reduce false stops and drop out peaks encountered in Dye Primer Cycle Sequencing. It contains a single point mutation to eliminate the inherent 5'-3' nuclease activity, and a thermostable inorganic pyrophosphatase. Extended read lengths are being gained by the use of these further optimised chemistries, adapted autosequencers with longer well to read distances and improved computer base calling algorithms. The capacity of machines can be increased by using the four dye technology and more channels per gel. The new ABI PRISM 377 DNA Sequencer incorporates many significant improvements including a more flexible sensitive simultaneous multicolour detection system using a CCD camera, longer thinner gels run

faster under actively controlled temperature conditions. It is perceived that major new developments will be made in this area or perhaps radical new methods such as capillary electrophoresis (CE), sequencing by hybridisation (SBH) to oligonucleotides, multiplex sequencing or microscopic sequencing will be utilised (Jones, 1995).

Either way within a year or two most large cDNA or genome sequencing institutions may resemble a factory with technology and robotics at the forefront generating data efficiently with reduced labour and costs. An automated laboratory produces over 10 million DNA bases a year, with projected costs critically now at between 7 and 25 pence per base (*Science*, 1995). The biological strategies on which these energies are directed are critical to the continued success of the Human Genome Project: collaboration and co-ordination must continue. Informed decisions must be made by the relevant leading scientists regarding evolving strategies, and future issues such as ethics relating to personal genetic information and gene therapy as required.

In conclusion it can be seen that in mans race to walk on the moon in the Human Genome Project and sequence the complete human genome, the spacecraft has been built and has taken off. ESTs have been generated with the current technology and have been a biological breakthrough in gaining the correct co-ordinates of the destination; and will continue to do so, helping to map all genes. Further advances in technology should lead to a faster easier journey through less interesting space, the intron DNA, giving rapid genomic sequencing in the years to come so that the project is completed on time. On arrival the actual use of all this specific genetic information generated to gain a ubiquitous clinical advantage will be a large step for mankind.

5. REFERENCES

- Adams, M.D. *et al.* Complementary DNA sequencing: Expressed sequence tags and the human genome project. *Science* **252**, 1651-1656 (1991).
- Adams, M.D. *et al.* Sequence identification of 2,375 human brain genes. *Nature* **355**, 632-634 (1992).
- Adams, M.D., Kerlavage, A.R., Fields, C. & Venter, J.C. 3400 expressed sequence tags identify diversity of transcripts in human brain. *Nature Genet.* **4**, 256-267 (1993).
- Adams, M.D., Soares, M.B., Kerlavage, A.R., Fields, C. & Venter, J.C. Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain brain cDNA library. *Nature Genet.* **4**, 373-380 (1993).
- Adams, M.D. *et al.* A model for high-throughput automated DNA sequencing and analysis core facilities Sequence identification of 2,375 human brain genes. *Nature* **368**, 474-475 (1994).
- Adams, M.D., Fields, C,. & Venter, J. (eds) *Automated DNA Sequencing and Analysis*. Academic Press, London (1994).
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410 (1990).
- Anand, R. Yeast artificial chromosomes (YACS) and the analysis of complex genomes. *TIBTECH* **10**, 35-40 (1992).
- Antequera, P. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc. Natl. Acad. Sci. U.S.A.* **90**, 11995-11999 (1993).
- Bargmann, C.I. cDNA sequencing: a report from the worm front. *Nature Genet.* **1**, 79-80 (1992).
- Carter, C., Britton, V.J., and Haff, L. A centrifugation medium for nucleic acid isolation and purification. *Biotechniques* **1**, 142 (1983).
- Cohen, D., Chumakov, I., & Weissenbach, J. A first-generation physical map of the human genome. *Nature* **366**, 698-701 (1993).
- Cross, S.H., Charlton, J.A., Nan, X., & Bird, A.P. Purification of CpG islands using a methylated DNA binding column. *Nature Genet.* **6**, 236-244 (1994).
- Erlich, H.A. (ed) *PCR technology: principles and applications for DNA amplification*. Stockton Press, New York (1989).

- Fields, C., Adams, M.D., White, O., & Venter, J.C. How many genes in the human genome? *Nature Genet.* **7**, 345-346 (1994).
- Fry, G. *et al.* A new approach to template purification for sequencing applications using paramagnetic particles. *BioTechniques* **13**, 124-131 (1992).
- Gyapay, G. *et al.* 1993-1994 Généthon human genetic linkage map. *Nature Genet.* **7**, 246-248 (1994).
- Hanahan, D. Studies on transformation of *E. coli* with plasmids. *Journal of Molecular Biology* **166**, 557 (1983).
- Hawkins, T. M13 single strand purification. *J. DNA Sequence Mapping.* **3**, 65-69 (1992).
- Hood, L. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674-679 (1986).
- Hood, L. *et al.* Model genomes: the benefits of analysing homologous human and mouse sequences. *TIBTECH* **10**, 19-22 (1992).
- Jones, S.J.M. An update and lessons from whole-genome sequencing projects. *Current Opinion in Genetics and Development* **5**, 349-353 (1995).
- Khan, A.S. *et al.* Single pass sequencing and physical and genetic mapping of human brain cDNAs. *Nature Genet.* **2**, 173-179 (1992).
- Little, P. Generating "cloned DNA maps". *TIBTECH* **10**, 33-35 (1992).
- Máxam, A., & Gilbert, W. A new method for sequencing DNA. *Proc. Natl Acad. Sci. U.S.A.* **74**, 560-564 (1977).
- McCombie, W.R. *et al.* *Caenorhabditis elegans* expressed sequence tags identify gene families and potential disease gene homologues. *Nature Genet.* **1**, 124-131 (1992).
- Messing, J. New M13 vectors for cloning. *Methods Enzymol.* **101**, 20-78 (1983).
- NIH/CEPH Collaborative mapping group. A comprehensive genetic linkage map of the human genome. *Science* **258**, 67-86 (1992).
- Okubo, K. *et al.* A novel system for large-scale sequencing of cDNA by PCR amplification. *J. DNA Sequence Mapping.* **2**, 137-144 (1991).
- Okubo, K. *et al.* Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nature Genet.* **2**, 180-185 (1992).

- Okayama, H., *et al.* High-efficiency cloning of full-length cDNA-constuction and screening of cDNA expression libraries for mammalian cells *Methods Enzymol.* **154**, 3-28 (1987).
- Polymeropoulos, M.H., Xiao, H., Sikela, J.M., Adams, M.D., Venter, J.C., & Merril, C.R. Chromosomal distribution of 320 genes from a brain cDNA library. *Nature Genet.* **4**, 381-386 (1993).
- Rosenthal, A., and Charnock-Jones, D.S. New protocols for DNA sequencing with dye terminators. *J. DNA Sequence Mapping.* **3**, 61-64 (1992).
- Sambrook, J., Fritsch, E.F. & Maniatis, T. (eds) (1989) *Molecular Cloning: A Laboratory Manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Sanger, F., Nicken, S., & Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. U.S.A.* **74**, 5463-5467 (1977).
- Short, J.M., *et al.* Lambda-zap-a bacteriophage lambda-expression vector with *in vivo* excision properties *Nucl. Acids Res.* **16**, 7583-7600 (1988).
- Sikela, J.M., & Auffray, C. Finding new genes faster than ever. *Nature Genet.* **3**, 189-190 (1993).
- Soares, M.B., Bonaldo, M de F., Jelene, P., Su, L., & Efstratiadis, A. Construction and characterisation of a normalised cDNA library. *Proc. Natl Acad. Sci. U.S.A.* **91**, 9228-9232 (1994).
- Staden, R. Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. *Nucl. Acids Res.* **10**, 4731-4751 (1982).
- Sulston, J. *et al.* The *C. elegans* genome sequencing project-a beginning. *Nature* **356**, 37-41 (1992).
- Uhlen, M. Magnetic separation of DNA. *Nature* **340**, 733-744 (1989).
- Vierra, J. and Messing, J. Production of single-stranded plasmid DNA. *Methods Enzymol.* **151**, 3-11 (1987).
- Waterston, R. *et al.* A survey of expresses genes in *Caenorhabditis elegans. Nature Genet.* **1**, 114-123 (1992).
- Watson, A., Smaldon, N., Lucke, R., & Hawkins, T. The *Caenorhabditis elegans* genome sequencing project: first steps in automation. *Nature* **362**, 567-568 (1993).
- Weissenbach, J. *et al.* A second generation linkage map of the human genome. *Nature* **359**, 794-801 (1992).

- Wilkie, T. *Perilous Knowledge*. Faber and Faber Limited, London (1993).
- The human gene map initiative. *Genome Digest* **2**, 1-4 (1995).
- Emphasis turns from mapping to large-scale sequencing. *Science* **268**, 1270-1271 (1995).

6. APPENDIX

Centrifugation data

RCF = $(rw^2)/g$

rpm = $(1000)(\sqrt{357/r})$

Beckman 50 Ti rotor max radius = 80.8 mm

DNA conversion data

Average molecular weight of a deoxynucleotide base = 324.5 Daltons

1 A_{260} unit of double-stranded DNA = 50 µg

1 A_{260} unit of single-stranded DNA = 37 µg

1 A_{260} unit of single-stranded RNA = 40 µg

Purity of RNA can be measured by A_{260}/A_{280}: a ratio of 1.8-2.0 is good

(A_{260} = absorbance taken at 260 nanometres, A_{280} = absorbance taken at 280 nanometres)

DNA molecular weight marker VI (Boehringer Mannheim), pBR322 DNA *Bgl* I + pBR322 DNA *Hinf* I, solution in 10 mM Tris-HCl, 1 mM EDTA, pH 8.0, 0.25 µg/µl
Fragment sizes: 2176, 1766, 1230, 1033, 653, 517, 453, 394, 298, 234, 220, 154 base pairs

DNA molecular weight marker III (Boehringer Mannheim), λ DNA·*Eco*R I + *Hind* III, solution in 10 mM Tris-HCl, 1 mM EDTA, pH 8.0, 0.25 µg/µl
Fragment sizes: 21226, 5148, 4973, 4268, 3530, 2027, 1904, 1584, 1375, 947, 831, 564, 125 base pairs

PRISM™ Ready Reaction DyeDeoxy™ Terminator Cycle Sequencing Kit Premix:
(1.58 µM A-DyeDeoxy, 94.74 µM T-DyeDeoxy, 0.42 µM G-DyeDeoxy, 47.37 µM C-DyeDeoxy, 78.95 µM dITP, 15.79 µM dATP, 15.79 µM dCTP, 15.79 µM dTTP, 168.42 mM Tris-HCl (pH 9.0), 4.21 mM $(NH_4)_2SO_4$, 42.10 mM $MgCl_2$, 0.42 units/µl AmpliTaq DNA polymerase)

160 mls of 6% acrylamide-urea gel solution

80 g Urea

24 mls 40% Acrylamide stock solution

50 mls MilliQ water

2 g mixed-bed, ion-exchange resin

160 mls of 4.75% acrylamide-urea gel solution

80 g Urea

19 mls 40% Acrylamide stock solution

47 mls MilliQ water

2 g mixed-bed, ion-exchange resin

0.2 μm filtered and degassed for five minutes

16 mls of f.s. 10xTBE (108 g Tris, 55.0 g Boric acid, 8.3 g EDTA, H_2O to 1 litre)

Filter sterile MilliQ water to a final volume of 160 mls

Polymerisation of 50 mls gel: 250 microlitres 10% ammonium persulphate (1 g/10mls), and 25 microlitres TEMED