Global expression mapping of mammalian genomes

by

Wolfgang Sebastian Meier-Ewert

a thesis submitted for the degree of

Doctor of Philosophy

in the University of London

Department of Biochemistry and Molecular Biology

University College London

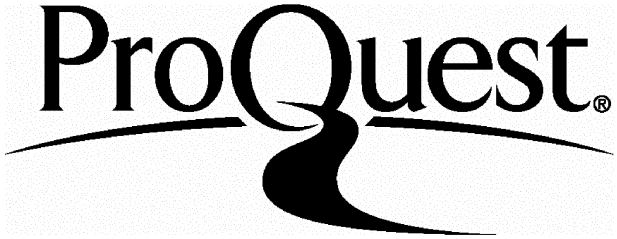
Gower Street, London WC1

Genome Analysis Laboratory

Imperial Cancer Research Fund

Lincoln's Inn Fields, London WC2A 3PX

September 1994



This thesis is dedicated to my parents and my wife.

Acknowledgments

First and foremost I would like to thank Hans Lehrach for the opportunity of carrying this work out under his guidance and for allowing me to benefit from his experience and seemingly endless enthusiasm.

I would also like to thank Elmar Maier and Leonard Schalkwyk for many long and stimulating discussions during the course of this work.

Richard Mott I would like to thank especially for making most of the data analysis possible and for teaching me much basic statistics.

For excellent technical assistance I would like to thank Michelle Charlesworth, Bernard Freeman and Jane Sandall. My gratitudes also to Jane Sandall for proof reading this document.

There are three kinds of lies: lies, damned lies and statistics.

Benjamin Disraeli (1804-1881)

Abstract

The aim of genome projects is to decipher all the information contained within the DNA of an organism and to study the way this information is processed in physiological processes. It is believed that more than 95% of the information content of the mammalian genome is represented in the protein coding sequnces that make up only approximately 2% of the DNA sequence. Consequently much effort is being invested in the study of coding sequences in the form of cDNA analysis. This thesis is concerned with the development of a new strategy for a highly parallel approach to analyse entire cDNA libraries. The strategy is based upon generating sufficient sequence information to identify uniquely more than 100,000 cDNA clones by hybridisation with short oligonucleotides, typically 7 - 10 mers. Each oligonucleotide is hybridised to all cDNA clones in parallel and under stringent conditions positively identifies a subset (3 - 10%) of clones. Oligonucleotides are designed in such a way that each will positively identify a different subset of clones and statistical simulations estimate that approximately 200 such hybridisation events are required to identify uniquely upto 100,000 cDNA sequences. Such a fingerprint can be generated from many cDNA libraries constructed from different tissue mRNAs and will not only lead to the identification of most sequecnes expressed from the genome but also indicate the level of expression by determining the number of times any given sequence is represented across different cDNA libraries. A human foetal brain cDNA library has been constructed and 100,000 clones arrayed into microtitre plates and on nylon membranes. All the required technological developments have been carried out successfully and are presented. In excess of 200 oligonucleotide hybridisations have been performed on a subset of 32,000 cDNA clones and 1,000 sequenced control clones. A detailed analysis of the data on the control clones is presented and the implications for cDNA fingerprinting discussed.

Abbreviations

ATP	Adenosine-5'-triphosphate
bp	base pairs
cM	centi morgan
CPU	central processing unit
DNA	deoxyribonucleicacid
cDNA	copy deoxyribonucleicacid
E. coli	*Escherichia Coli*
EDTA	ethylenediaminetetraaceticacid
cpm	counts per minute
g	gravitational acceleration
kb	kilo base pairs
Mb	mega base pairs
m-RNA	messenger ribonucleicacid
OD	optical density
PCR	polymerase chain reaction
PEG	polyethylene glycol
RNA	ribonucleicacid
r-RNA	ribosomal ribonucleicacid
SDS	sodium dodecyl sulphate
STS	sequence tagged site
t-RNA	transfer ribonucleicacid
Tris	tri-hydroxymethyl-aminomethane
U	unit
UV	ultraviolet
YAC	yeast artificial chromosome

List of Figures

List of Tables

1. Introduction

1.1 Molecular Genetics

Mammalian molecular genetics has experienced a technological revolution over the past fifteen years that is unparalleled in biological research. A myriad of techniques have now been developed that constitute some of the most powerful tools available in this discipline. In fact, over the last ten years technical innovation itself has been elevated to a scientific discipline and is culminating in the human genome project whose ultimate aim is to decipher all the information contained in the genetic material of the human organism.

A small selection of key technological innovations are: the discovery of restriction endonucleases, the development of DNA sequencing techniques, the use of restriction fragment length polymorphism (RFLP) to study DNA segregation in families, the use of pulsed field gel electrophoresis for the separation of large DNA fragments (> 50 kb), the development of a range of cloning systems in both *E.coli* and yeast, the development of site directed mutagenesis, the invention of the polymerase chain reaction (PCR) and the introduction of exogenous DNA into organisms to create transgenics. More detailed accounts have been published in numerous reviews. Many more technological developments will be necessary to achieve the aims of the genome projects. After all, even if the human genome can be mapped and sequenced with present technology, which is by no means entirely clear, there are several other genomes that are the subject of large scale analyses (e.g. mouse, drosophila, xenopus, c.elegans, zebra fish, puffer fish and arabidopsis to name but a few). A further question that will require attention is how many different genomes of each species should be analysed in order to gather sufficient information to allow us

to understand the complexities of genome function? It is realistic to assume that in future multiple genome equivalents will be analysed each year for most species under investigation and this scale of data generation will require far more efficient analysis systems. Given the enormous cost of the genome projects under way at the moment and the scale of the tasks mentioned above, continued technological developments will form an essential part of future molecular genetics.

While large scale genome analysis represents enough work for several generations of scientists the immediate aim of the human genome project of generating a complete map, sequencing of the entire genome and identifying all the genes present is really a service industry feeding the rest of the biological/medical sciences community. The data generated by the human genome project will serve as a library of information, stored in computers, that will be accessed and utilised by many scientists not in any way involved in its generation. As a result the emphasis of scientific research is likely to change drastically. Where scientists now devote anywhere from months to years of benchwork to isolating their genes of interest, in future this work will be replaced by database searches, followed by requests for the relevant clones from a central distribution service. This system will of course be not unlike that already in operation for obtaining DNA from families under the auspices of CEPH (Centre d'Etude du Polymorphisme Humain).

1.2 Positional Cloning (from Phenotype to Genotype)

Positional cloning has become the most powerful technique for the isolation of genes involved in genetically inherited traits and those diseases caused by major cytogenetic alterations. such as many cancers. Since the cloning of the genes involved in retinoblastoma (Friend et al. 1985), chronic granulomatous disease (Royer Pokora et al., 1986), and duchenne muscular dystrophy (Monaco et al., 1986) positional cloning

has emerged as the method of choice for the identification of disease genes. The ever more rapidly expanding list of cloned genes involved in genetic diseases, disease susceptibility and developmental control processes is testament to the advances that have been achieved over the last few years in positional cloning techniques (for review see (Ballabio, 1993)). Positional cloning strategies have evolved over the past ten years incorporating most of the advances in molecular technology. Early genes cloned by positional cloning were identified by studying various chromosomal deletions that narrowed the candidate region down to a manageable size and then cDNA clones were identified that spanned the deletion breakpoint, thus defining candidate genes (e.g. duchenne muscular dystrophy). Genes cloned later, such as cystic fibrosis (Kerem et al., 1989; Riordan et al., 1989; Rommens et al., 1989) made use of extensive chromosome walking and linkage disequilibrium to narrow a candidate region lacking any gross rearrangements. Most recent cloning efforts have incorporated even more technology. The eventual success in cloning the Huntington's disease gene (The Huntington's Disease Collaborative Research Group, 1993), which came more than ten years after the isolation of the first informative polymorphic DNA marker (Gusella et al., 1983), required contig building by walking experiments with yeast artificial chromosomes, P1 clones and cosmids over a region in excess of 2 Mb on chromosome 4. Due to lack of clear genetic data the candidate region could not be narrowed any further and an extensive program of systematic searching for candidate genes began, which employed techniques such as exon trapping, cDNA fishing and single stranded conformational polymorphism studies. Eventually a cDNA clone was found that contained a triplet repeat, the expansion of which is associated with the disease. Triplet repeats have emerged over the past three years as an important molecular mechanism in the causation of genetic disease. Seven genetic diseases have now been identified that contain expansions of trinucleotide repeats (for review see Mandel, 1994). In some cases the triplet is part of the coding sequence which is

transcribed and translated, in others the triplet is only transcribed and in another the repeat lies in non-coding DNA upstream of the promoter sequences.

1.3 From Genotype back to Phenotype

While the hunt for genes involved in human disease has continued apace so the search for new and more representative animal models continues. The generation of new phenotypes by induced mutation has long been established in the studies of unicellular organisms, but has only recently been applied to higher organisms, both invertebrate (e.g. drosophila) and vertebrate (e.g. xenopus, mouse). The use of gene targeting as a method of generating specific mutations in known genes has been applied with great success and has given exciting insights into the function of many genes involved in processes such as early development and tumorgenesis. So, while molecular genetics of higher organisms has relied upon naturally occurring phenotypes to study gene function for many years, there is now the possibility of more detailed analysis of gene function by the introduction of deliberate specific mutations in genes and the study of resulting phenotypes. Targeted disruption of murine *hox*-genes has detailed the previously known involvement of these genes in embryogenesis (e.g. (Carpenter et al., 1993) and (Condie and Capecchi, 1993)) and is leading to the elucidation of the complex interaction of genes in the *hox*-clusters. Several of the mutations associated with human genetic diseases have been recreated in mouse models with the aim of generating a resource which can be used to develop and test possible treatments and cures. This approach, though helpful, has proved complex in interpretation, and factors such as genetics background influence phenotypic expression. Generating genotypes in mice that are analogous to those in humans often mirror the human phenotype only partially or not at all. A mouse model for cystic fibrosis has shown many of the physiological characteristics of the

human disease, including intestinal obstruction, but fails to recreate the severe abnormalities in respiratory epithelium. The fact that there are more than 170 mutations in the cystic fibrosis transmembrane conductance regulator (CFTR) gene (Collins, 1992) (although 70% of patients carry a 3 base deletion in exon 10 (ΔF508)) that give rise to a widely varying range of symptoms of course, complicates analysis. Nevertheless animal models for cystic fibrosis have led to the development of promising strategies for gene therapy (Zabner et al., 1993 and Hyde et al., 1993).

Other examples of gene targeting do not yield immediately applicable animal models. The p53 gene is associated with many human cancers and is a classic example of a tumour suppressor gene (for review see (Weinberg, 1991)). Mutations resulting in gain of function, loss of heterozygosity and mutations in both p53 alleles (first proposed as disease mechanism in retinoblastoma (Hollstein et al., 1991)) have been confirmed in cases of p53 associated tumours (Knudson, 1971). However, transgenic mice that are homozygous for a p53 null-allele are developmentally normal (Donehower et al., 1992) and while being susceptible to spontaneous tumours show surprisingly high viability. So far therefore, the mouse p53 model does not hold as much promise for the development of effective treatments as the cystic fibrosis models.

Targeted gene disruption has also become a powerful method to confirm the function of a positionally cloned gene. For example, both the human and mouse gene for sex determination (*SRY* and *sry* respectively) were cloned (Sinclair et al., 1990; Gubbay et al., 1990) and later transgenic mice were generated that were chromosomally female but developed as males (Koopman et al., 1991), thus confirming that the *sry* gene is sufficient for testis differentiation.

1.4 Genome Mapping

1.4.1 Genetic Mapping

Human genetic analysis has always lagged behind that of other organisms such as fruit fly and maize mainly because crosses cannot be set up at will by geneticists and because only few genetic markers were known that were heterozygous within families. Since the early 1980s human genetics has come of age due to the discovery of highly polymorphic DNA markers.

Botstein and colleagues (Botstein et al., 1980) proposed a genetic mapping strategy based upon the segregation of restriction fragment length polymorphism (RFLP). These markers can be sufficiently polymorphic to allow the construction of useful linkage maps of the human genome. Using such markers the first real linkage maps of whole chromosomes were created (Dryana and White, 1985). Other sources of polymorphic markers have been discovered that have a higher rate of heterozygosity. RFLP markers usually detect the presence or absence of a restriction endonuclease site and therefore have a maximum heterozygosity of 50%.

Tandem repeat sequences have been found distributed throughout the genome and Jeffreys and colleagues cloned a myoglobin 'minisatellite' sequence, which when used as a hybridisation probe under low stringency conditions crosshybridises to other repeat sequences (Jeffreys et al., 1985). Variations in the number of tandem repeat units are detected on digested DNA that does not cleave the minisatellite internally. Minisatellites, although found throughout the genome, tend to cluster near telomers and the markers of choice now used are 'microsatellites'.

Weber and May (1989) proposed the use of polymorphic $(dC\text{-}dA)_n$ repeat lengths for genetic fingerprinting and mapping. By sequencing

genomic DNA flanking the $(CA)_n$ repeats, the polymerase chain reaction can be used to amplify DNA across the repeat and the products run on a sequencing gel.

In 1992 a 'comprehensive genetic linkage map of the human genome' was published that consisted of 1416 loci including 339 microsatellites (NIH/CEPH Collaborative Mapping Group 1992). Later in the same year a further 814 microsatellites were mapped with an average heterozygosity > 0.7 (Weissenbach et al., 1992). Microsatellites have the advantage that through PCR on the one hand only small amounts of genomic DNA are required for analysis and that they can be directly incorporated into physical maps as sequence tagged sites (STS) (Olson et al., 1989). In the most recent report of the Genethon group, a total of 2,066 microsatellites have been mapped, of which 60% show a heterozygosity greater than 0.7 (Gyapay et al., 1994).

1.4.2 Physical Mapping

Physical maps combined with genetic maps are required for the identification of candidate genes and the molecular characterisation of the genome. A physical map consists of ordered cloned DNA fragments and any physical map is only as good as the clones used in building it and the larger the clones are the fewer are needed, of course, to span any given distance. Yeast artificial chromosomes (YACs) (Burke et al., 1987) that contain recombinant DNA inserts up to > 1 megabase are the clones of choice for physical mapping.

There are several strategies that are used to build physical maps and they fall into three main categories: mapping by restriction digest analysis (Coulson et al., 1986), mapping based upon STS content (Olson et al., 1989) and the determination of overlap by direct hybridisation of clones as both probes and targets (Lehrach et al., 1990). The use of STS mapping yielded the first real successes of the human genome project with maps covering most of human chromosome

21 (Chumakov et al., 1992) and Y (Foote et al., 1992; Vollrath et al., 1992). Extensive mapping has been carried out using restriction digest fingerprinting of YACs (Bellanne-Chantelot et al., 1992). Mapping based purely on hybridisation techniques has so far been applied successfully to the mapping of the yeast *Schizosaccharomyces pombe* which is the first eukaryotic genome to be mapped completely in both YACs (Maier et al., 1992) and cosmid/P1 clones (Hoheisel et al., 1993).

A combination of all three techniques has been used in the construction of 'First generation physical map of the human genome' (Cohen et al., 1993). This map is a very significant achievement and also combines the genetic and physical map by incorporating over 2000 microsatellite markers. Importantly the YAC library that Cohen and colleagues use has been distributed widely throughout the genome community and is used in many large scale projects. This is an important consideration since in this way the information generated in many different labs can be collated into building one map and separately conducted and funded initiatives complement each other and become additive. Given the scale of the task of the human genome project, the cost of funding and the international co-operation that is involved, this is an appropriate way of proceeding. Although nationalistic priorities and competitive issues need to be overcome in such a system the potential benefits of a more significant useful scientific resource should be sufficient incentive.

To establish a set of common resources that can be used by the scientific community for genome analysis a reference library database (RLDB) system has be established at the Imperial Cancer Research Fund (Zehetner and Lehrach, 1994). Libraries can be screened by hybridisation in form of filter grids that contain up to 36,864 clones. Identified clones can then be obtained as cultures in exchange for information about the probe used for screening. The experimental information is stored in a relational database which can be generally accessed via a variety of computer network facilities (e.g. anonymous

ftp, gopher, www server, e-mail and CD-ROM). More than 25 libraries are currently available through the RLDB ranging from YACs, cosmids, P1s and cDNAs from human, mouse, drosophila and *S.pombe*. The list is constantly being expanded and includes libraries from other centres, such as the human YACs from CEPH and mouse YACs from both St.Marys hospital, London and the Whitehead institute. A similar scheme is also operated by the HGMP Resource Centre in Harrow, Middlesex.

The utility of any map is determined not only by the extent of its coverage but also by its resolution. The aim of a physical map is to facilitate the identification of genes and their control sequences and in the case of the human genome to provide a substrate for sequencing the genome. While the YAC maps go a long way towards providing a resource for further analysis, much work remains to locate genes and to generate a 'sequence ready' map. The largest clones being successfully sequenced on a routine basis at present are cosmids (~ 40 kilobases) and the physical maps of the human genome will have to reach this resolution before a true sequence resource is created, unless dramatic breakthroughs in sequencing technology are achieved.

1.4.3 Expression Mapping

With few exceptions, all cells of multicellular organisms share the same pool of genomic DNA. Tissue differences arise through the use of information of a subset of the DNA pool to drive protein expression. Proteins are responsible for cellular physiology, anatomy and function. Transcription regulation is a major control process for protein synthesis and experimental data suggests that much of tissue specific protein expression is reflected in the level of expression of cytoplasmic mRNA. The mammalian brain displays the highest complexity of function and many studies have found this reflected in the complexity of its mRNA population (Milner et al., 1987).

1.4.3.1 mRNA Studies

Studies of cytoplasmic poly(A)+ RNA populations have revealed a large variation in complexity between different cell types. Complexity of RNA isolates are measured by two methods: 1) Saturation hybridisation to radioactively labelled single stranded non-repetitive genomic DNA (Chikaraishi et al., 1978). (Non-repetitive DNA is prepared by three successive rounds of self annealing to C_0t 200 and selection of single stranded DNA by hydroxylapatite chromatography. After hybridisation to RNA and hydroxylapatite chromatography the proportion of non-repetitive DNA bound to RNA can be assessed.) 2) Reassociation kinetics of RNA to first strand cDNA prepared by reverse transcription.

Analyses of mammalian brain poly(A)+ RNA populations have shown that up to 65% of poly(A)+ RNA is not shared with other tissues used as non-neuronal controls, such as liver and kidney. These estimates come from both hybridisation studies (Chaudhari and Hahn, 1983) and cDNA clone analysis (Milner and Sutcliffe, 1983). Estimates of the number of genes expressed in tissues varies between studies, but all have found the complexity of brain poly(A)+ RNA to be 2-3 fold higher than non-neuronal tissues (Milner et al., 1987). To estimate the number of genes expressed in a tissue from its mRNA complexity the average length of the mRNA molecules must be taken into consideration. Between 14,000 and 128,000 genes are estimated to be expressed in rodent brain, using 1800 nucleotides as the average mRNA length. In a study of almost 200 brain cDNA clones, brain specific mRNAs were found to be significantly longer at 5,000 nucleotides (Milner and Sutcliffe, 1983) and the authors estimate approximately 30,000 genes to be expressed in the mammalian brain. Of these 20,000 are believed to be brain specific (Sutcliffe, 1988). Specifically expressed genes make up the majority of complexity but only a small fraction of the mass of poly(A)+ RNA. Since the mammalian brain is such a highly complex organ with a diverse physiology and histology, brain tissue is not truly an appropriate concept as applied to gene expression studies. Indeed

many mRNA species are estimated to be expressed at less than one copy per cell. The likely explanation for this observation is that only a sub population of cells in brain tissue express those very rare transcripts. Studies concerning gene expression in certain brain regions, such as cerebral cortex and cerebellum, have made the unsurprising observation that there is substantial regional variation in complexities and sequence overlap in mRNA populations.

No discussion of RNA complexity studies would be complete without reference to nonpolyadenylated RNA poly(A)- RNA. The significance of poly(A)- RNA is still rather unclear. Several studies have shown that poly(A)- RNA and poly(A)+ RNA in the cytoplasm of rodent brain tissue are approximately equal in complexity and substantially non-overlapping. Studies on HeLa cell mRNA indicate that poly(A)- RNA is a subset of the poly(A)+ RNA population when assessed by in vitro translation of both mRNA fractions and resolution of translation products on 2D-polyacrylamide gels (Kaufmann et al., 1977). However, although poly(A)- RNA is found in non-neuronal tissues (Milcarek, 1979) high complexity poly(A)- RNA is found mainly in brain tissue and is expressed predominantly postnatally (Chaudhari and Hahn, 1983). The relevance of the study, using HeLa cells, on poly(A)- RNA function is therefore in doubt. Postnatal changes in mRNA complexities have not been found in other tissues and may reflect the considerable postnatal development of brain tissues as opposed to other organs such as liver and kidney. Interestingly, most of the increased complexity of mRNA observed in postnatal mouse brain tissue is accounted for in the poly(A)- RNA fraction (Chaudhari and Hahn, 1983).

1.4.3.2 cDNAs as a means of studying gene expression

In the late 1970's detailed studies were performed on the activities of viral enzymes. Initial studies indicated that RNA driven DNA transcription produced products with a bimodal size distribution. Since

reverse transcriptase has an RNase activity as well as a DNA polymerisation activity, it was postulated that it was the degradation of RNA molecules and subsequent transcription that gives rise to the smaller sized DNA fraction which was shown to be produced after the larger size fraction has been synthesised. Several nuclease inhibitors were tested to prevent the degradation of the RNA template during reverse transcription reactions and it was shown that more or less complete cDNA copies of RNA molecules as long as 6 kb could be synthesised (Kacian and Myers, 1976). This study defined the exact conditions under which cDNA synthesis is optimal for generating long cDNAs with highest possible specific activity. A later study by van Ness and Hahn (1980) addressed the representation of cDNA synthesised from complex RNA templates both by oligo(dT) and random priming. Assessed both by saturation hybridisation to single copy genomic DNA and hybridisation kinetics to template mRNA, van Ness and Hahn found that approximately 98% of the complexity of the template mRNA was also represented in the cDNA. The study of cDNAs was therefore established as a valid system for the study of gene expression. Cloning the cDNAs was easily achieved by ligating linkers to the ends of the cDNAs and subsequently ligating into plasmid or phage hosts. In the early years of high complexity cDNA library construction phage vectors were used rather than plasmid vectors because of the higher efficiency of packaging and transfection, compared to transformation. Since mRNA sequences are represented at vastly different frequencies in total mRNA extractions, and cDNAs are a representative copy, it is desirable to have as many cDNA sequences in any given library as possible in order to maximise the probability of even rarely transcribed sequences being cloned. In the last five years there has been a transition to the use of arrayed clone libraries, where individual clones are stored in microtitre plate wells and therefore obtain a unique address (Nizetic et al., 1991b). These arrayed clones can be stored indefinitely in any number of copies so that they can be used in many experiments over a long period, accumulating information on a

constant resource. Due to handling complications when dealing with arrayed phage clones, the emphasis has shifted recently to the use of cDNA libraries cloned into plasmid vectors. Arrayed cDNAs can be grown as colonies on nylon membranes and DNA prepared in situ in the same way as described by Nizetic (Nizetic et al., 1991b).

1.4.3.3 cDNA library construction

The quality of a cDNA library is critically dependent on the intactness of the mRNA used as template in the reverse transcription reaction. Classically, mRNA is selected from total RNA extractions from tissue or cell lines. Extraction of good quality total RNA, that is RNA with as little as possible degradation either from nucleases or by physical breakage, is the first critical step in cDNA library construction. The main source of RNA degradation comes from RNases. Although RNase activity can be suppressed during the extraction procedure the enzyme is highly resistant to exposure to high temperatures and chaotropic agents, regaining much of its activity when returned to renaturing conditions. During RNA extraction great care therefore has to be exercised to denature and then remove all RNase activity. During cell lysis highly reducing conditions are used (e.g. 120 mM 2-mercaptoethanol), which inactivate RNases by breaking disulphide bridges (Chirgwin et al., 1979).

Poly(A)+ RNA selection is carried out by chromatography with cellulose immobilised oligo-d(T). Since poly(A)+ RNA only represents approximately 1 - 2% of total RNA extracted from cells, this enrichment is important especially when the first strand cDNA synthesis is random primed.

The synthesis of cDNA is performed by reverse transcriptase using one of two priming systems, oligo-d(T) or random hexadeoxynucleotides (Sambrook et al., 1989). Both systems have advantages and drawbacks and the decision on which to use must be based on the application for

which the cDNA library is to be used. Oligo-d(T) primed cDNA libraries have the advantage that all clones contain the 3' end of the gene's transcript and that inserts can easily be directionally cloned into expression vectors, increasing the proportion of clones whose inserts are correctly translated. Directional cloning has additional advantages such as reduced non-recombinant background because adapter ligations can be avoided (for review see (Kaiser, 1990)). The main drawback of oligo-d(T) primed cDNA libraries lies in the fact that long transcripts are unlikely to be cloned in their entirety not only due to broken RNA templates but the inefficiency of cloning long DNAs into many vectors, particularly plasmids. As a result it is often not possible to isolate the 5' end of long transcripts from oligo-d(T) primed cDNA libraries. To overcome this problem cDNA libraries can be synthesised by random priming of the mRNA. These libraries are generally cloned in a non directional way since after second strand synthesis the cDNA is blunt ended and linker molecules are ligated to both ends (however, it should be possible to prime with random oligomers that all have a restriction enzyme recognition site at their 5' end, such as *Not*I.). There are two main technical complications in making random primed cDNA libraries: Firstly, poly(A)- RNA, particularly ribosomal RNA which constitutes the majority of molecules in total RNA extractions, causes a higher proportion of ribosomal cDNA clones, thus reducing the complexity of a library; secondly, random primed libraries generally contain a higher proportion of chimeric clones probably arising during linker ligation.

One of the most difficult challenges is the construction of high complexity cDNA libraries form very small tissue samples. For many applications of expression studies it is desirable to synthesise cDNA from a highly purified selection of cells that are often in very short supply. Examples include developmental expression studies using mouse embryos (Rothstein et al., 1992) and neuronal expression studies of various brain sections (Travis and Sutcliffe, 1988). One of the ways to address this problem has been to use the PCR reaction to amplify cDNA synthesised from very small amounts of template mRNA

(Froussard, 1992; Dumas Milne Edwards et al., 1991) and even single cells (Saiki et al., 1988). Protocols that make use of PCR need to be designed very carefully to avoid unequal amplification of sequences. Although, as with a reverse transcription reaction, there is no inherent sequence bias in the polymerisation reaction, the nature of the PCR means that only fully transcribed molecules will be amplified exponentially. Since the chance of complete transcription is proportional to the length of the molecule, most PCR based protocols use a size selection for short cDNAs. As a result PCR amplified cDNA libraries mostly have inserts of average length < 1,000 bp and therefore contain few full length clones.

Many protocols have been developed to counter the problem of over representation in libraries of abundant cDNAs which often hamper the identification of differentiation (Mather et al., 1981), developmental (Sargent and Dawid, 1983), tissue (Hedrick et al., 1984) or tumour specific (Lee et al., 1991) transcripts. Subtraction procedures make use of cDNA from one source and 'driver' mRNA from another in a hybridisation system that leaves mainly those sequences not shared by both samples single stranded. Hydroxylapatite chromatography then allows the separation of double and single stranded nucleic acid fractions. Early applications of this technique used subtraction to generate probes (Mather et al., 1981) and later also to generate subtracted cDNA clones (Hedrick et al., 1984) and whole libraries (Sargent and Dawid, 1983). Typically however, these protocols require the use of ~100 µg driver RNA which is prohibitive for many applications. The use of phenol emulsion enhanced hybridisation allows far less driver RNA to be used (~6 µg) and has been applied for subtractive hybridisation of monkey cortex-specific cDNA clones (Travis and Sutcliffe, 1988).

Two reports have also been published on the generation of normalised cDNA libraries, in which abundant and rare transcripts are represented in approximately equal numbers (Ko, 1991; Patanjali et al., 1991). In

these cases the near second order kinetics of double stranded cDNA reassociation is utilised to deplete abundant sequences from a complex mixture of cDNAs. Patanjali et al. use random primed cDNAs whereas Ko used an oligo-d(T) primed library and each approach has advantages and disadvantages as discussed above. Ko's choice of oligo-d(T) primed cDNAs that have a short average length (several hundred bases) limits mis-representation introduced by one of the major difficulties associated with all annealing based strategies for both normalisation and subtraction. Rare sequences that differ only in a small proportion of their length such as members of gene families and alternative splice variants, will tend to be eliminated by hybridisation to a more abundant homologous species. The 3' untranslated region of mRNA is known to be specific for gene family members (e.g. actin genes) (Alonso et al., 1986) and often also for splice variants as well as transcripts with alternative polyadenylation sites. By using a cDNA population biased towards the 3' untranslated region of mRNAs Ko's system avoids a great deal of this kind of erroneous depletion in sequence complexity while sacrificing information due to very little coding sequence in each clone.

For some applications of expression studies the use of subtraction or normalised cDNA libraries is fraught with complications in the sense that the libraries have lost one of their useful qualities, that is that the level of expression of a sequence is correlated to the number of times it is represented within the library. After all, since neither subtraction nor normalisation is a quantitative procedure but merely an enrichment, what is gained is a reduction in the number of clones that need be analysed to find the sequence of interest. If one subtracts or normalises too much so that only very few sequences remain to be analysed then the risk is great that one has lost what one was looking for. On the other hand if one subtracts or normalises only a little, one has gained not much in terms of a reduction in screening effort with less risk of having discarded the desired transcript. Since the extent of subtraction or normalisation required is difficult to assess at the outset of an experiment a certain

amount of guessing is involved. If one could obtain a totally normalised gene catalogue in which each transcript was represented only once, screening would of course be greatly simplified. A cDNA library however, in which most sequences are present say half a dozen times, only reduces the total number of clones to screen since identified cDNAs will have to be put through a secondary screening cycle anyhow, to identify multiple copies of the same transcript.

It is intuitively reasonable that given the fact that most tissue specific transcripts are expressed at a low level in most cells and that much of the sequence complexity of mRNA is in the low abundance class (Sutcliffe, 1988) a normalised library would aid the identification of novel gene sequences in a random sampling approach. This has been verified experimentally by Höög (1991), who showed that eliminating abundant cDNAs from a random sequencing strategy dramatically increased the identification of previously unidentified transcripts. For this kind of application a high quality normalised cDNA library is an invaluable resource.

1.4.3.4 cDNA sequencing

The past few years have seen the inception of several large scale sequencing projects based exclusively on cDNA clones (Adams et al., 1991; Okubo et al., 1992; Kahn et al., 1992). The aims of these projects are really twofold: on the one hand they aim to identify all the transcribed sequences in the human genome and on the other identified genes can, by one additional step, be placed on the emerging physical maps of chromosomes. Both aims are not easily accommodated within one experimental strategy. In order to characterise a gene the coding sequence is most informative in terms of indicating homologies to other genes and therefore pointing to functional properties of the gene product. Using the single pass sequencing approach however to generate expressed sequence tags (EST) it is important to be able to design PCR

primers that not only amplify a DNA fragment of reasonable size so that it can be readily detected by ethidium bromide staining of agarose gels, but that do not fall into separate exons, since the introns that are present in the genomic DNA are unlikely to be amplifiable by PCR. The largest exon on average in human DNA is the 3' exon that also contains the 3' untranslated region (the 3' untranslated region only rarely contains exons) (Hawkins, 1988). It is generally possible to design PCR primers several hundred nucleotides apart that still lie in the 3' untranslated region (Wilcox et al., 1991). There is an added advantage in that the 3' untranslated region of genes is most informative for discriminating gene family members since the conservation is lower than in coding sequence. Since most mRNAs are longer than the sequence that can be determined by a single experiment, this kind of approach does mean however that much of the sequence obtained from the cDNAs will be non-coding, therefore yielding little information on the gene product.

To date more than 14,000 ESTs have been published (Adams et al., 1991; Adams et al., 1992; Okubo et al., 1992; Kahn et al., 1992; Adams et al., 1993a; Adams et al., 1993b) and most are available in form of a database (dbEST) (Boguski et al., 1993). These large scale projects are generating new gene sequences at an unprecedented rate, reports (given at the Sequencing and physical mapping meeting, Hilton Head, S.C. in 1993) suggest that approximately 200,000 bases of raw sequence are produced every day and that approximately 80,000 different sequences have been identified. Large scale projects are also underway to assign many ESTs to chromosomes (Polymeropoulos et al., 1992; Polymeropoulos et al., 1993).

1.4.3.5 Combining genetic, physical and expression maps

An important aspect of genome analysis is the integration of mapping information from all kinds of experimental approaches. A comprehensive map of a genome will incorporate a complete physical

map (that is a continuum of overlapping clones) with genetically polymorphic markers on most clones and information of the genes located on each clone, where and when they are expressed and their entire coding sequence. This kind of information cannot be generated in a single project and the way the human genome project has developed over the past few years the tendency has been for large centres to concentrate on a particular experimental strategy to address one aspect of genome analysis. It is clear that almost as much effort will need to be expended in combining the different kinds of data as was required to generate each individually. There are really two kinds of approaches being used that while both analysing similar or even identical clones are not directly complementary and have divided the genome community into two camps. On the one hand there is the use of sequence information to generate PCR primers that can be used as sequence tagged sites (STS) for physical mapping (Olson et al., 1989) or for genetic mapping from polymorphic sequences (Weber and May, 1989) and as expressed sequence tags (ESTs) when derived from cDNAs. And on the other hand there is the use of clones and short oligonucleotides in hybridisation procedures that can generate maps without any prior sequence information (Lehrach et al., 1990). For genetic mapping using microsatellite markers it seems clear that PCR based methods are the most efficient for generating this kind of information. Ideally however, techniques used for large scale analysis should be simple and easily automated. So far this does not apply to procedures involving sequencing reactions, oligonucleotide synthesis, subsequent PCR of thousands of templates and electrophoretic analysis. Although some of the individual steps have been automated, stringing them together into an automated production line has not been achieved so far.

The use of cDNAs as mapping probes is one efficient way of combining expression data with physical and genetics maps and is reviewed in detail by Southern (1992). The use of arrayed cDNA libraries (Lennon and Lehrach, 1991) makes it possible to use the clones both in expression and mapping studies correlating the two types of data

through the common medium used. Due to the unequal representation of sequences in cDNA libraries steps need to be taken to 'normalise' the library before clones can be used on a large scale for genome mapping. Several papers have been published that propose methods for achieving this (Ko, 1991; Patanjali et al., 1991; Meier Ewert et al., 1993).

1.5 Hybridisation Analysis

Hybridisation based clone analysis has been extensively developed in recent years (Lehrach et al., 1990). The use of arrayed clone libraries in form of high density filter arrays, allows a highly parallel approach to clone analysis. A hybridisation probe can be used to screen many thousands of clones in a single experiment. One of the powerful advantages of hybridisation based analysis is that clones can be used both as probes and targets. In a test study for hybridisation based physical mapping the genome of the yeast *Schizosaccharomyces pombe* was mapped with YACs, P1s and cosmids (Maier et al., 1992; Hoheisel et al., 1993). All clone types were used as both probes and targets through different phases of the project. Initially a YAC map was established using genetic markers to anchor clones along the genetic map and cosmids as probes on YACs in order to establish YAC overlap. End rescue techniques from end of contig YAC clones were used to connect overlapping contigs that had not been identified by cosmids. Once a YAC map had been established a subset of YAC clones were hybridised to both P1 and cosmid clones. A sampling without replacement strategy was then used in which single cosmids were hybridised back to all cosmids and P1s and previously non identified clones used as probes in the next hybridisation round. This is still the only complete physical map of any eukaryotic genome in which yeast and bacterial cloning systems combine to sequence ready map. Given that one of the functions of physical maps is as a resource for

sequencing projects and these still only reliably succeed on the cosmid level, large maps like those generated for the human chromosomes will need to contain cosmids and/or P1s. The STS approach being used by many groups to establish YAC maps will require a prohibitively large number of STSs to establish cosmid and P1 sequence ready maps and therefore a hybridisation approach will still be required.

1.5.1 Clone Mapping with Oligonucleotides

The use of synthetic oligonucleotides to map clones was proposed by Poustka et al.(1986). Like single copy probes oligonucleotides give information about the clones they hybridise to. Unlike single copy probes however they give far less information about each clone but there can be far more clones hit by the probe. This means that a hybridisation experiment with an oligonucleotide probe can yield far more information than a single copy probe, if the frequency of hybridisation of the oligonucleotide is high enough. The main advantageous feature of oligonucleotide fingerprinting is its favourable scaling characteristics compared to single copy hybridisation probes. An increase (n) in the number of clones to analyse, increases the number of required hybridisation experiments by n for single copy probes and by log(n) for oligonucleotide probes. The great disadvantage of oligo fingerprinting and the reason that it has not yet fulfilled its theoretical promise, is the fact that the experimental error rate is far greater than for single copy hybridisations. This need not be an insurmountable problem as long as one can determine the error rate for each experiment which in turn is only possible if information already exists of the target clones' expected behaviour with each oligonucleotide probe. Taking advantage of the highly parallel approach to which hybridisation experiments lend themselves one can use a mixture of known and unknown clones as a hybridisation target and assess the value of the data generated on both sets of clones from the correlation between expected and observed behaviour of the known clones. The assumption that has to be made in

this kind of scheme is that the error rate is very similar for both types of clones, which is probably valid since most of the experimental error is oligonucleotide dependent.

The cosmid and P1 maps of the three *S. pombe* chromosomes that have already been generated could provide exactly the kind of control discussed above for large scale oligonucleotide mapping of other genomes or chromosomes.

1.5.2 Sequencing by Hybridisation

The information generated on target clones by synthetic oligonucleotide hybridisations taken to its ultimate limit, that is hybridisation of all possible oligonucleotides of a given length, yields the entire sequence of the target clones in oligonucleotide sized blocks. It should therefore be possible to reconstruct the sequence of the target clones. Sequencing by hybridisation (SBH) has been proposed independently in two papers (Drmanac et al., 1989; Bains and Smith, 1988). Two formats for this approach are envisioned: format 1 in which experimental clones are immobilised on a solid support and oligonucleotides labelled with some kind of reporter molecule are hybridised serially to all target clones; format 2 whereby all oligonucleotides are immobilised on a solid support and experimental clones are hybridised serially. The two formats are suitable for different kinds of experimental strategies, really depending on whether there are more oligonucleotides or more target sequences involved.

One of the first experimental conditions that have to be fulfilled is the reliable sequence dependent hybridisation of short oligonucleotides in such a way that perfectly matched duplexes can be discriminated from mismatched ones. This is essential for the reconstruction of sequence from a series of hybridisation events. Drmanac et al. (1990b) showed that this was possible under easily achievable experimental conditions for oligonucleotides as short as hexamers. Two papers have been

published in which short clones of known sequence were hybridised with over 100 oligonucleotides in a format 1 scheme (Strezoska et al., 1991; Drmanac et al., 1993) illustrating that the necessary experimental and analytical procedures have been assembled into a working system. Since of course one hundred experiments to determine 100 bp of sequence is grossly inefficient when compared to gel sequencing methods, considerable scale up has to be achieved before a useful experimental strategy is developed. This poses a series of not inconsiderable problems both experimentally and analytically. Short oligonucleotide hybridisations require highly purified target DNA since even small amounts of host or vector DNA contamination can cause significant background signal. Therefore, a very efficient and labour unintensive method has to be used to generate many thousands of target DNAs. This is presently best achieved with the PCR which put a constraint on the size of target DNA that can be generated routinely. The size of the target DNA is an important consideration in terms of the length of oligonucleotide used and with regard to sequence reconstruction given a complete dataset. As pointed out by Bains (1991) there is an inherent limit to the length of sequence that can be reconstructed unambiguously from oligonucleotide hybridisation data. Sequence reconstruction algorithms operate on a likelihood system, that is they choose that sequence which is statistically most likely to give rise to the observed dataset. After a certain length a branch point is reached at which there are two equally likely possibilities. Bains showed by simulations that the length of sequence which can be reconstructed unambiguously is dependent on the length of the oligonucleotide used and that for example octamers allow on average 200 bp to be assembled before an ambiguity is reached. Secondary structure of the target DNA is a very important factor governing hybridisation behaviour and the longer the DNAs the greater is the probability of its effect being significant. One way to get around this problem is to fragment the target DNA into smaller pieces and as long as

this is done in a more or less random fashion, such as sonication, no one hybridisation site should be destroyed in all molecules.

There are several groups currently working on format 2 SBH, that is where a large array of oligonucleotides is immobilised and single experimental DNAs, or RNAs, hybridised to the array. An extensive set of evaluation experiments has been carried out by Southern and colleagues (Southern et al., 1992) in which they have developed a novel DNA synthesis system on glass whereby the resultant oligonucleotides can be used directly for hybridisation in form of a glass 'DNA chip' (Maskos and Southern, 1992a). In later studies they have characterised extensively the hybridisation behaviour of the large arrays of oligonucleotides (Maskos and Southern, 1992b; Maskos and Southern, 1993a; Case-Green and Southern, 1994). These have shown clearly that hybridisation characteristics vary greatly and are sequence dependent. The variable hybridisation characteristics of short oligonucleotides are a more acute problem for a format 2 approach since it involves hybridising many thousands of different sequences under the same conditions in parallel. Format 1 in contrast uses single oligonucleotides in each hybridisation and therefore makes it easier for the experimental conditions to be adapted to the particular probe in use. However, the same studies also show that the hybridisation characteristics while correlating with G+C content and some nearest neighbour predictions (Breslauer et al., 1986; Freier et al., 1986) are still not readily predictable.

Khrapko et al.(1991) have developed a system in which oligonucleotides are immobilised within a 30 μm polyacrylamide gel layer in turn covalently bound to glass. They propose a method of adjusting the concentration of each oligonucleotide in the array compensating for the differential hybridisation behaviour so that a single hybridisation protocol can be used to discriminate all sequences. So far they have not succeeded in generating such a 'normalised' chip partly at least due to the immense amount of work involved. At Affymetrix (Palo Alto, CA,

USA) oligo chips are also being generated by a photolythographic method (Jacobs and Fodor, 1994; Fodor et al., 1993) and preliminary evaluations have been carried out.

Many of the format 2 systems are being applied more and more not to the reconstruction of complete sequence, but for the detection of sequence variation within a known sequence (Maskos and Southern, 1993b; Mirzabekov, 1994; Fodor et al., 1993). This approach does not require the hybridisation of nearly all oligonucleotides of a given length but only a relatively small set of well characterised and experimentally matched sequences. It also has the huge potential as a diagnostic tool. As ever more genetic sequence variations, be they mutations or polymorphisms, are linked directly to diseases or other health factors a very large increase in the demand for the generation of sequence data is foreseeable. Most of this demand will be focused on obtaining sequence information for many individuals on regions of the genome previously characterised and will thus be eminently amenable to this kind of system.

1.5.3 Sequence fingerprinting with oligonucleotides

An application for oligonucleotide hybridisation that lies somewhere between physical mapping and sequencing is that of sequence fingerprinting using short oligonucleotides (Drmanac et al., 1990a; Meier Ewert et al., 1993). To illustrate this approach to clone analysis it is useful to think of each experiment as increasing the information gained on all clones under investigation. One can then ask how much information of this type is required to generate sufficient information. The amount of required information depends on the goal of the analysis. If, for example, a cDNA library is to be analysed, then the minimum information one would like to generate for each clone is whether it is the same as any other clone in the library and whether it is already present in a databank.

A very simple unsophisticated estimate of the required data can be made as follows:

Consider that there are a maximum of 100,000 gene sequences to discriminate between and that the clones under analysis are cDNAs of average length 1,500 bp. If all sequences were random then an oligonucleotide of length l has a probability

$$p = 0.5 \times 1{,}500 / 4^{l}$$

of occurring in any given clone. Provided that oligonucleotide hybridisation events are non correlated, that is the result of one hybridisation does not greatly affect the probability of another such as for highly overlapping probes, the probability of n given oligonucleotides occurring is p^n. One can now estimate a likelihood ratio of two clones sharing a given fingerprint p^n, that is a series of oligonucleotide hybridisation events, because they also share the same sequences or by random chance. The chance of random identity I is at least 10^{-5} since there are at most 100,000 different sequences in the analysis. Therefore, if one requires $I / p^n > 10^6$, then $p^n < 10^{-11}$. For p = 0.1 this means that 11 positive hybridisation events are required to each clone to have a likelihood ratio of 10^6 that all fingerprints are unique to any given sequence (this may be likened to the *lod* score, in this case 6, used in genetic linkage analysis). 110 hybridisations at p = 0.1 would be required to yield 11 positive hybridisation events per clone. The true number of hybridisations required in practice will of course be affected by the error rate of the data.

To illustrate the oligonucleotide fingerprinting approach a simple test was carried out using the Genbank database of DNA sequences. Six octanucleotide sequences were extracted from exon 25 of the cystic fibrosis transmembrane regulator gene (CFTR) (1: TGAGGTGG, 2: CATGGCCA, 3: ATGACAAA, 4: AGATTTTA 5: TAAAATGG, 6: AAAAATAT). All sequences in the combined databanks of Genbank and EMBL were searched for those entries that contained all six octamers,

using the 'findpattern' command. The following four entries were found:

Human growth hormone:	66,495 bp
Human HPRT gene:	56,737 bp
Human thymidilate synthase gene:	18,596 bp
EBV genome:	172,282 bp
Human CFTR exon 25:	983 bp

The result shows the exon 25 of the CFTR gene is the only sequence in either database that contains the six octamers in less than 18,596 bp. While this in no way proves the oligonucleotide fingerprinting approach working for entire cDNA libraries, since the sequence complexity of a cDNA library will be far greater than that of all the database entries, it illustrates well the power of fingerprint information. This test also illustrates one of the important considerations of any fingerprinting approach, which is the size of the target DNA fragments. A fingerprint can only be useful for discriminating sequences that of similar length. In the test described here, four octanucleotide sequences from exon 25 of the CFTR gene were in fact sufficient to achieve the same discrimination.

The expected hybridisation results for any database entry can be easily determined and compared to those produced by the experimental clones. The fingerprint therefore serves two functions, it determines those clones that share largely the same sequence and it can also match the experimental data against the predicted fingerprints generated by known sequences thus identifying previously sequenced cDNAs. Providing the hybridisation data are reproducible the sequence specificity of the interactions need not affect the ability of the data to identify all the different sequences in the analysis. However, in order to be able to compare the experimental oligonucleotide fingerprints with theoretical

ones determined for database sequences the error rate for the sequence specific hybridisation must be included in the analysis assuming that the data are not error free.

The amount of sequence information generated by a fingerprinting approach is not unlike that from 'tag' sequencing of cDNAs (Adams et al., 1991). In the tag sequencing approach, typically 200 - 300 bp are sequenced from the 3' ends of randomly selected cDNA clones. Several groups are involved in this kind of project (Okubo et al., 1992; Kahn et al., 1992; Adams et al., 1993a) generating hitherto unparalleled amounts of sequence data. What does tag sequencing really do? It generates enough information on unknown clones to be able to determine whether or not it has been identified previously, whether it belongs to a known gene family and/or contains certain sequence motifs that allude to gene product function. In addition PCR primers can be designed that can be used in physical mapping experiments. Clones of potential interest then need to be sequenced entirely, possibly new full length clones need to be isolated and their expression patterns determined. In other words, the initial sequencing is a prelude to further analysis that helps in identifying clones of particular interest, so that limited resources can be focused and efforts prioritised. All these criteria can also be fulfilled by a sequence fingerprint based on oligonucleotide hybridisations, which has the advantage of all hybridisation approaches in that it lends itself to massively parallel data generation.

It seems likely that most of the human genes expressed in most major tissues will be sequenced in the near future by conventional gel sequencing methods. A parallel approach, of a slightly different format, is currently being implemented at The Institute for Genomic Research (TIGR), where a factory approach has been taken to molecular biology, similar to that successfully employed a Genethon, in which 30 automated sequencing machines are run continuously (Adams et al., 1994). While this is presently undoubtedly the fastest way of generating sequence information it is not clear that the brute force approach will

continue to be viable. Once most of the expressed sequences in human tissues have been identified the next step will be to map their expression patterns across as many tissue types as possible. Given that there are over 50 tissues for which expression data will be of immediate interest, and that there are several model organisms for which homologous sequences and their respective expression patterns will be of interest the cost incurred so far is merely the thin edge of a wedge. There appears still to be a need for more large scale cost effective analysis systems. The large scale gel sequencing centres are likely to find their greatest utility in detailed analyses for which this technique is still the most powerful.

1.6 Aims of this thesis

Over the past years there has been substantial development in the technology of highly parallel hybridisation approaches, some of which have been used in the analysis of cDNA libraries (Lennon and Lehrach, 1991). cDNAs have been identified as one of the most suitable systems for the analysis of transcribed sequences and many varied approaches are being applied to their characterisation. Transcribed sequences contain the great majority of the information stored in the vertebrate genome. The information gleaned from the analysis of cDNA clone libraries is twofold, in that on the one hand the sequence of the clones yields the coding information and on the other hand the source tissue, from which the library was constructed, contributes both spatial and temporal information about the expression of genes. In order to achieve a comprehensive analysis of gene expression in vertebrate organisms, or non-vertebrate multicellular organisms for that matter, it will be necessary to analyse not only sufficient numbers of cDNA clones to identify most of the transcribed sequences, but also to cover a large number of tissues to obtain a significant amount of temporal and spatial expression data. Ultimately, cDNA analyses will have to be carried out

on more than a million clones to achieve both of these objectives. Present gel-sequencing technology is not yet well suited to such an application, not least of all because of the considerable cost. As an alternative, or at least a complementary approach, a hybridisation based analysis should provide a useful addition to the analytical tools available in molecular genetics. As introduced in the previous section of this *Introduction*, an oligonucleotide fingerprinting approach to the partial characterisation cDNA clones would theoretically required only of the order of two hundred hybridisations to generate sufficient data to discriminate all possible cDNA sequences. For the application of such a fingerprinting technique to many thousands of clones several experimental techniques need to be developed further to allow the required scale up that is involved.

The aim of this thesis is to develop the necessary tools, both experimental and analytical, required for a successful oligonucleotide fingerprinting approach to cDNA clone analysis. This involves mainly the automation and scale up of existing technologies on the experimental side, starting from the picking of randomly plated cDNA clones and the preparation of DNA to the arraying at high density and hybridisation with hundreds of oligonucleotides. For the analysis of hybridisation signals, both an automated image analysis and a robust data analysis system must be developed. Finally, all the steps involved in the entire fingerprinting approach must be tested in a controlled way to assess the performance of the system. The rest of this thesis describes the implementation and evaluation of a first generation oligonucleotide fingerprinting approach.

2. MATERIALS AND METHODS

2.1 Reagents

2.1.1 General Reagents and Materials

2.1.1.1 Sigma Chemicals Co.

Trizma hydrochloride

Trizma base

bovine serum albumin (BSA)

dithiothreitol (DTT)

phenylmethylsulphonylfluoride (PMSF)

agar

agarose

ampicillin (sodium salt)

salmon sperm DNA

yeast t-RNA

2.1.1.2 BDH Laboratories

general salts and chemicals

glucose

ethanol

iso-propylalcohol

chloroform

phenol

proteinase K

N-lauroyl sarcosine

2.1.1.3 Difco

bacto tryptone

bacto peptone

yeast extract

brain-heart infusion (BHI)

2.1.1.4 Pharmacia

deoxyadenosine triphosphate (di-sodium salt)

deoxythymidine triphosphate (sodium salt)

deoxycytidine triphosphate (sodium salt)

deoxyguanosine triphosphate (sodium salt)

2.1.1.5 Amersham International plc.

Hybond N+ (nylon)

[^{33}P-γ] ATP (3000 mCi mmol^{-1})

[^{32}P-γ] ATP (5000 mCi mmol^{-1})

[^{32}P-α] dATP (3000 mCi mmol^{-1})

2.1.1.6 Kodak

XAR-5 X-ray film

2.1.1.7 Genetix

384-well microtitre plates and lids (Q-plates)

384-pin replicators

Q-plate heat sealing film

2.1.2 Enzymes and enzyme buffer reagents

2.1.2.1 New England Biolabs

general restriction enzymes

T4 DNA ligase

T4 polynucleotide kinase

2.1.2.2 Boehringer Mannheim

calf intestinal phosphatase

E.coli DNA polymerase large fragment (Klenow)

digoxygenin antibody / alkaline phosphatase conjugate

2.1.2.3 Gibco BRL

Superscript cDNA synthesis kit

2.1.2.4 InVitrogen

Fast track poly A+ RNA isolation kit

2.1.3 Solutions and Media

Many of these recipes were derived from Hoheisel (1987).

2xYT-media:	1.6% (w/w) Bacto tryptone, 1% (w/w) yeast extract, 0.5% (w/w) NaCl, 96.9% H_2O (for 2xYT-agar + 9 gl^{-1} agar)
BHI-media:	36 g l^{-1} brain heart infusion in H_2O
50x TAE:	23.6% (w/w) Trizma base, 6.2% (w/w) acetic acid, 1.6% (w/w) EDTA, 84.3% H_2O.
10x TBE:	15.45% (w/w) Trizma base, 2.62% (w/w) boric acid, 0.9% (w/w) EDTA, 81.03% (w/w) H_2O.
1 M Tris pH 8.0:	5.16% (w/w) Trizma base, 8.64% (w/w) Trizma hydrochloride, 86.20% (w/w) H_2O.
Denaturing solution:	1.87% (w/w) NaOH, 8.15% (w/w) NaCl, 89.98% (w/w) H_2O.
0.5 M EDTA pH 8.0	16.9% (w/w) EDTA, 1.9% (w/w) NaOH, 81.2% (w/w) H_2O.

100x TE:	1 M Tris pH 8.0, 100 mM EDTA.
3 M Na-phosphate pH 7.2:	6.68% (w/w) Na_2HPO_4, 0.52% H_3PO_4, 92.80% (w/w) H_2O (1 M with respect to Na^+).
20x SSC:	15.36% (w/w) NaCl, 7.73% (w/w) tri-sodium citrate, 76.91% (w/w) H_2O, pH 7.5.
SSarc:	20% (v/v) 20x SSC, 25% (v/v) N-lauroyl sarcosine (30%), 55% (v/v) H_2O.

2.1.4 Bacterial strains

DH5 α	F'/*endA1 hsdR17* ($r_K^- m_K^+$) *supE44 thi-1 recA1 gyrA* (Nal^r) *relA1* Δ(*lacZYA-argF*)*U169 deoR* (Ø80 *dlac*Δ(*lacZ*) *M15*)
XL1 blue	F'::Tn*10 proA*$^+$*B*$^+$ *lacI*^{q} Δ(*lacZ*)*M15* / *recA1 endA1 gyrA96* (Nal^r) *thi hsdR17* ($r_K^- m_K^+$) *supE44 relA1 lac*

2.2 Experimental procedures

2.2.1 RNA isolation

RNA was isolated from frozen tissue using a protocol derived from Chirgwin et al. (1979). Up to 3 g of frozen tissue was ground in a pestle and mortar precooled in liquid nitrogen. The finely powdered tissue was poured into 600 ml lysis solution (4 M guanidinium isothiocyanate, 25 mM sodium acetate, 120 mM 2-mercaptoethanol, sterile filtered), thoroughly dispersed and homogenised in a sterile glass

dounce. The tissue lysate was passed three times through a sterile gauge 25 needle and six 8.3 ml aliquots were each layered onto 2.4 ml caesium chloride solution (5.7 M caesium chloride, 25 mM sodium acetate) in polypropylene ultracentrifuge tubes. Samples were centrifuged in a SW40 rotor (Beckman) at 35,000 rpm, for 20 hours, at 20°C, under vacuum. The supernatant was removed and each RNA pellet resuspended in 500 µl guanidinium hydrochloride solution (7.5 M guanidinium hydrochloride pH 7.0, 5 mM dithiothreitol). RNA was precipitated with 12.5 µl acetic acid (1 M) and 250 µl ethanol. After centrifugation the pellet was washed in 70% ethanol and resuspended in 100 µl sterile water.

2.2.2 poly A+ RNA isolation

The isolation of poly A+ RNA from total RNA extracts was carried out by oligo(dT)-cellulose chromatography as described by Sambrook et al. (1989). Up to 5 mg total RNA in water was heated to 68°C for 10 min mixed 1:1 with column loading buffer (20 mM Tris-HCl pH 7.6, 0.5 M sodium chloride, 1 mM EDTA, 0.1% sodium lauroyl sarcosinate) and loaded onto a 1 ml oligo(dT)-cellulose column. The column was prepared from 1 g oligo (dT)-cellulose (Sigma) which was dissolved in 0.1 M sodium hydroxide, poured into a Disposocolumn (Biorad) and washed with 6 column volumes of column loading buffer. The flow through was reheated to 68°C for 10 min, reloaded onto the column which was then washed with loading buffer until the eluate $A_{260} < 0.02$. Poly A+ RNA was eluted with 10 mM Tris-HCl pH 7.6, 1 mM EDTA and 0.05% SDS. 1 ml fractions were collected and their A_{260} measured in cuvettes that had been treated for 1 hour in concentrated hydrochloric acid / ethanol (1:1). RNA was precipitated by addition of sodium acetate to 0.3 M (pH 6.0) followed by 2.5 volumes ethanol.

2.2.3 cDNA cloning

cDNA was prepared from up to 5 µg poly A+ RNA using a commercial cDNA synthesis kit (Superscript cDNA kit, Gibco BRL). First strand cDNA was synthesised by oligo-dT priming using a *Not*I primer (5' -pGACTAGTTCTAGATCGCGAGCGGCCGCCC$(T)_{15}$-3'), double stranded cDNA size fractionated over a gel filtration column (Sephacryl S-500) and directionally cloned into a *NotI*, *Sal*I digested pSPORT plasmid vector.

2.2.3.1 First strand synthesis

The reverse transcription reaction is carried out with BRL's Superscript enzyme, which is isolated from a recombinant Moloney murine leukaemia virus reverse transcriptase gene lacking Rnase H activity (Kotewicz et al., 1988).

The RNA and *NotI*-primer were incubated together at 70°C for 10 min prior to addition of other reagents.

final concentration	reagent	volume	stock concentration
50 mM	Tris-HCl pH 8.3	4 µl	5x buffer
75 mM	KCl		
3 mM	$MgCl_2$		
10 mM	DTT	2 µl	0.1 M
500 µM	dNTPs(each)	1 µl	10 mM
	a^{32}P-dATP	1 µl	1 µCi μl^{-1}
50 µg ml^{-1}	*NotI* primer adapter	2 µl	500 ng μl^{-1}
≤250 µg ml^{-1}	poly-A+ RNA	5 µl	1 µg μl^{-1}
50,000 U ml^{-1}	superscript RT	5 µl	10,000 U μl^{-1}

The first strand reaction is carried out in a total volume of 20 µl. Prior to the addition of the superscript enzyme the reaction mixture was incubated at 37°C for 2 min. The reverse transcription reaction was performed at 37°C for 1 hour. Before the addition of second strand synthesis reagents, 2 µl of the first strand reactions were removed and

48 µl EDTA (0.5 M) added. This aliquot was stored at 4°C for later analysis.

2.2.3.2 Second strand synthesis

The first strand reaction was cooled on ice before the addition of second strand reagents.

final concentration	reagent	volume	stock concentration
25 mM	Tris-HCl pH 7.5	30 µl	5x-buffer
100 mM	KCl		
5 mM	$MgCl_2$		
10 mM	$(NH_4)_2SO_4$		
0.15 mM	β-NAD^+		
1.2 mM	DTT		
250 µM	dNTPs(each)	3 µl	10 mM
65 U ml^{-1}	DNA ligase	1 µl	10 U μl^{-1}
250 U ml^{-1}	DNA polymerase1	4 µl	10 U μl^{-1}
13 U ml^{-1}	RNaseH	1 µl	2 U μl^{-1}
	H_2O	93 µl	

The second strand synthesis was carried out in a total volume of 150 µl and the reaction mixture incubated at 16°C for 2 hours.

The termini were blunt ended by the addition of 2 µl T4 DNA polymerase (5 U ml^{-1}) and further incubation at 16°C for 5 min. The reaction was terminated by the addition of 10 µl EDTA (0.5 M), placed on ice and extracted with 150 µl phenol/chloroform/iso-amyl alcohol (25:24:1) and centrifuged at 11,000x g for 5 min. 140 µl of the aqueous phase was transferred to a fresh microfuge tube and the cDNA precipitated with 70 µl ammonium acetate (NH_4Ac) (7.5 M) and 0.5 ml 100% ethanol (-20°C). The sample was centrifuged immediately at 11,000 x g for 20 min, washed with 0.5 ml 80% ethanol (-20°C) and air dried at 37°C for 10 min.

2.2.3.3 *SalI adapter ligation*

To the dried cDNA pellet the following reagents were added on ice:

final concentration	reagent	volume	stock concentration
50 mM	Tris-HCl pH7.6	10 µl	5x buffer
10 mM	$MgCl_2$		
1 mM	ATP		
5% (w/v)	PEG 8000		
1 mM	DTT		
200 µg $µl^{-1}$	*SalI* adapters	10 µl	1 µg $µl^{-1}$
100 U ml^{-1}	T4 DNA ligase	5 µl	20 U $µl^{-1}$
	H_2O	25 µl	

The *SalI* adapter was made up of the following two oligonucleotides 1: 5'-TCGACCCACGCGTCCG-3'; 2: 5'-pCGGACGCGTGGG-3'. The ligation was carried in final volume of 50 µl and incubated at 16°C overnight. The ligation reaction was extracted with 50 µl phenol/chloroform/iso-amyl alcohol (25:24:1), centrifuged at 11,000x g for 5 min and 45 µl of the aqueous phase precipitated with 25 µl NH_4Ac (7.5M) and 150 µl 100% ethanol (-20°C). After centrifugation at 11,000 x g for 20 min the pellet was washed with 0.5 ml 80% ethanol and air dried at 37°C for 10 min.

2.2.3.4 *NotI digestion*

To the dried pellet of cDNA from adapter ligation the following reagents were added for *NotI* digestion in a total volume of 50 µl:

final concentration	reagent	volume	stock concentration
50 mM	Tris-HCl pH8.0	5 µl	React7
10 mM	$MgCl_2$		
50 mM	NaCl		
50 mM	KCl		
1200 U ml^{-1}	*Not*I	4 µl	15 U μl^{-1}
	H_2O	41 µl	

The reaction was incubated at 37°C for 2 hours and then extracted with 50 µl phenol/chloroform/iso-amyl alcohol (25:24:1), centrifuged at 11,000x g for 5 min and 45 µl of the aqueous phase precipitated with 25 µl NH_4Ac (7.5M) and 150 µl 100% ethanol (-20°C). After centrifugation at 11,000x g for 20 min the pellet was washed with 0.5 ml 80% ethanol and air dried at 37°C for 10 min.

2.2.3.5 Column Chromatography

The prepacked sephacryl S-500 HR columns were washed through four times with 800 µl TEN (10 mM Tris-HCl pH 7.6, 0.1 mM EDTA, 25 mM sodium chloride), letting each 800 µl wash drain completely before applying the next. The dried cDNA from the *Not*I digest was resuspended in 100 µl TEN and loaded onto the column. The eluate was collected in tube 1, and 100 µl TEN added to the column. The next eluate was collected in tube 2. There after a further 18 drops were collected in separate tubes (3 - 20) by sequentially adding 100 µl aliquots of TEN to the column.

2.2.3.6 cDNA yield estimation

The Cerenkov counts of the 20 fractions collected during size selection were determined separately in a scintillation counter. The entire counts in the tritium channel were read and the amount of cDNA calculated according to the following formula:

$$\text{cDNA } (\mu g) = \frac{(cpm)(50\mu l/10\mu l)(20\mu l/2\mu l)(4pmol\ dNTP/pmol\ dATP)}{(cpm/pmol\ dATP)(3030pmol\ dNTP/\mu g\ cDNA)}$$

The specific activity:(SA: cpm / pmol dATP) is given by the following formula:

$$SA = \frac{cpm/10\mu l}{200pmol\ dATP/10\mu l}$$

To determine the SA the Cerenkov counts in 10 µl of the stored first strand synthesis aliquot were measured and that value used in the above formula.

2.2.3.7 Vector ligation of cDNA

This ligation reaction is designed for the ligation of 10 ng cDNA in volume of 20 µl.

final concentration	reagent	volume	stock concentration
50 mM	Tris-HCl pH7.6	4 µl	5x-buffer
10 mM	$MgCl_2$		
1 mM	ATP		
5% (w/v)	PEG 8000		
1 mM	DTT		
2.5 µg ml^{-1}	pSPORT1	1 µl	50 ng μl^{-1}
0.5 µg ml^{-1}	cDNA	<10 µl	<1 ng μl^{-1}
50 U ml^{-1}	T4 DNA ligase	1 µl	1 U μl^{-1}
	H_2O	4-13 µl	

The ligation was allowed to proceed at room temperature for 3 hours. After ligation DNA was precipitated with 5 µl yeast t-RNA (1 µg μl^{-1}), 12.5 µl NH_4Ac (7.5M) and 70 µl 100% ethanol (-20°C). The pellet was washed in 0.5 µl 80% ethanol (-20°C), air dried at 37°C and resuspended in 5 µl water.

2.2.4 Preparation of electrocompetent *E.coli* cells

The protocol is adapted from Dower et al. (1988). The appropriate *E.coli* strain was streaked on an L-agar + relevant antibiotic and incubated at 37°C overnight. A single colony was inoculated into 50 ml L-broth in a 250 ml conical flask and incubated at 37°C in an orbital shaking incubator (200 rpm) overnight. 1 litre of prewarmed L-broth without antibiotic was inoculated with 1/100 volume of overnight culture and grown at 37°C with shaking at 200 rpm to an A_{600} = 0.5. Cells were chilled to 4°C in ice water and then pelleted by centrifugation at 2000x g for 10 min at 4°C. The cell pellet was resuspended in 1 litre cold water (4°C) and then centrifuged as in the previous step. The cells were washed one more time with 500 ml cold water then in 20 ml cold 10% glycerol and finally resuspended in 2 ml 10% glycerol. Aliquots of 40 µl were either used immediately for transformation or frozen in liquid nitrogen and stored at -70°C.

2.2.5 Transformation by electroporation

Electroporation was carried out according to Dower et al. (1988). Up to 5 µl of DNA was mixed into 40 µl of electrocompetent cells on ice. After 5 min on ice the mixture was transferred into a cooled 2 mm electroporation cuvette (Biorad) and the cells transformed using a Biorad Genepulser with the following settings: 2.5 kV, 25 µF and 200 Ω. Immediately after pulsing the cells were taken up in 1 ml prewarmed SOC medium (2% bacto tryptone (w/v) 0.5% bacto yeast extract (w/v), 10 mM sodium chloride, 2.5 mM potassium chloride, 10 mM magnesium chloride, 10 mM magnesium sulphate and 20 mM glucose), and incubated in a 15 ml Falcon tube at 37°C shaking at 200 rpm for 1 hour. Cells were then plated immediately.

2.2.6 Arraying and storage of clones into microtitre plates

E.coli transformed with cDNA were plated out on 22 cm x 22 cm (Bioassay trays, NUNC) 2xYT + 50 μg ml^{-1} ampicillin agar plates and incubated at 37°C for 16 hours. Individual colonies were inoculated automatically into wells of quadruple density microtitre plates filled with 40 μl 2xYT (+ 50 μg ml^{-1} ampicillin, Hogness modified freezing medium), using an automated picking robot, purpose built in house. After inoculation plates were incubated at 37°C for 16 - 20 hours. Plates were wrapped airtight and stored at -70°C.

2.2.7 TAQ DNA polymerase preparation

Modified from Engelke et al. (1990) by L.Schalkwyk and myself.

An *E.coli* clone carrying the Thermus aquaticus DNA polymerase I gene as described by Engelke et al. (1990) was streaked onto an agar plate containing 50 μg ml^{-1} ampicillin and incubated at 37°C overnight. A single colony was inoculated into 50 ml 2xYT + 50 μg ml^{-1} ampicillin in a 250 ml conical flask and incubated at 37°C in an orbital shaking incubator (200 rpm) overnight. 10 ml of the overnight culture was inoculated into 1 l 2xYT + 50 μg ml^{-1} ampicillin in a 2 l conical flask and incubated in a shaking orbital incubator (200 rpm) at 37°C. When A_{600} reached 0.2, IPTG was added to 670 μM and the culture incubated for a further 16 hours. The cells were pelleted by centrifugation at 2000 x g for 10 min in a Beckman J6/B centrifuge and then resuspended in 100 ml buffer A (50 mM Tris-HCl pH 7.9, 50 mM dextrose, 1 mM EDTA) containing 80 mg lysozyme (Sigma). After 10 min at room temperature phenylmethylsulphenylfluoride (PMSF) was added to 1 mM and then 20 ml of buffer B (10 mM Tris-HCl pH 7.9, 50 mM KCl, 1 mM EDTA, 0.5% Tween-20 (v/v), 0.5% Nonidet P-40 (v/v)) were added. Sample was mixed by vortexing and then incubated at 75°C for 45 min. The sample was then cooled rapidly on ice for 2 min and then

centrifuged in open top tubes at 11,200 x g for 10 min in a Sorval HB4 rotor. The supernatant was loaded onto a 50 ml Biorex 70 (Biorad) column, previously equilibrated with 6 bed-volumes (i.e. 300 ml) of Buffer C (20 mM Hepes pH 7.9, 1 mM EDTA, 1 mM PMSF, 0.5% Tween-20 (v/v), 0.5% Nonident P-40 (v/v)) + 50 mM KCl. The column was washed with 300 ml of buffer C + 50 mM KCl and the TAQ-protein eluted with buffer C + 200 mM KCl. Ten fractions of 10 ml were collected and separately assayed for polymerase activity by nicktranslation. Three fractions of the highest activity were pooled and dialysed twice against 1 l buffer D (20 mM Hepes pH 7.9, 0.1 M KCl, 0.1 mM EDTA, 1 mM DTT, 0.5% mM PMSF, 0.005% gelatine, 50% glycerol) at 4°C for one hour and overnight respectively. Polymerase isolates were stored at -20°C.

2.2.8 TAQ DNA polymerase assay

A DNA polymerase assay was adapted from Chien et al. (1976) and Aposhian & Kornberg (1962).

Activated salmon sperm DNA was prepared by digestion with pancreatic DNase. Salmon sperm DNA was incubated at 0.25 mg ml^{-1} in 50 mM Tris-HCl pH 7.6, 5 mM magnesium chloride, 0.5 mg ml^{-1} bovine serum albumin, 0.5 µg ml^{-1} pancreatic DNase at 37°C for 15 min. The reaction mixture was then transferred to a 78°C waterbath for 5 min and the DNA precipitated with 110 µl sodium acetate (2.5M) and 1.1 ml isopropanol. After centrifugation at 11,000x g for 20 min. the DNA pellet was resuspended to 2 µg $µl^{-1}$ in water.

TAQ DNA polymerase isolates were assayed in a total volume of 50 µl containing 25 mM Tris-HCl pH 7.6, 25 mM potassium chloride, 2 mM magnesium chloride, 1 mM 2-mercaptoethanol, 0.2 mM dATP, 0.2 mM dTTP, 0.2 mM dCTP, 0.2 mM dGTP, 0.1 mg ml^{-1} activated salmon sperm DNA, 30 µCi [^{32}P-α] dATP. The reaction was incubated at 74°C for 30 min and then cooled on ice. 5 µl aliquots of the reaction were

spotted onto 1 cm^2 Whatman 1 filters, dried and assayed in a scintillation counter by Cerenkov counting. The filter squares were then washed four times for 2 min in 15 ml 5% trichloroacetic acid, 20 mM sodium phosphate in a 50 ml tube. The filters were then assayed again by Cerenkov counting and the radioactive incorporation calculated by subtracting the counts after washing from those measured before.

2.2.9 Waterbath PCR amplification

cDNA was cloned into the *SalI / NotI* digested vector pSPORT (as described in the cDNA cloning section) and amplified by the polymerase chain reaction (PCR) directly from liquid cultures stored in quadruple density microtitre plates using primers flanking the cloning site (Sport/3 20mer: 5'CCGGTCCGGAATTCCCGGGT3', Sport/5 29mer: 5'GCACGCGTACGTAAGCTTGGATCCTCTAG3'). For the M13 control clones, which were cloned into mp18, primers flanking the cloning site were again used for PCR (MP18RSP 37mer: 5'GAGCGGATAACAATTTCACACAGGAAACAGCTATGAC3' and MP18FSP 30mer: 5'TTTCCCAGTCACGACGTTGTAAAACGACGG3'). Amplification was carried out in 384-well microtitre plates in a volume of 30 µl containing: 10 pmol of each primer, 50 mM KCl, 4 mM potassium phosphate pH 7.4, 1.5 mM $MgCl_2$, 0.01% gelatine, 200 µM dGTP 200 µM dCTP, 200 µM dATP, 200 µM dTTP, 0.5 units TAQ-polymerase. Reactions were inoculated with approximately 0.2 µl *E.coli* culture using a 384-pin transfer device (Genetix, Christchurch Dorset) and then heat sealed with a 45 µm bilaminar nylon / polypropylene film using a commercial plate-sealing device (Genetix, Christchurch Dorset).The sealed 384-well microtitre plates were cycled automatically 30 times between waterbaths at 96°C for 3 minutes and 73°C for 5 minutes. After cycling the plates were briefly centrifuged (Beckman J6/B) and the sealing film removed by reheating and thus melting the surface of the plates in a plate-sealer. Samples were covered with conventional microtitre plate lids and stored at -70°C.

2.2.10 PCR reactions in commercial PCR machines

PCR amplifications of cDNA clones were carried out mainly in a Cetus 9600 PCR machine. Oligonucleotide primer immediately flanking the cloning sites were used (Sport3/86 20mer: 5'-CCGGTCCGGAATTCCCGGGT-3', Sport5/86 29mer: 5'-GCACGCGTACGTAAGCTTGGATCCTCTAG-3'). Each reaction was inoculated with a small amount of liquid cDNA clone culture transferred by a yellow Gilson pipette tip without dispensing. A total volume of 30 µl contained: 100 pmol of each primer, 50 mM KCl, 10 mM Tris-HCl pH 8.55, 1.5 mM $MgCl_2$, 0.01% gelatine, 2 mM dGTP, 2 mM dCTP, 2 mM dATP, 2 mM dTTP, 2.5 units TAQ-polymerase. 30 cycles of 72°C for 3 min and 94°C for 1 min were preformed after an initial denaturation of 4 min at 94°C .

2.2.11 PCR product precipitation

PCR samples were completely evaporated in a 65°C fan oven. To each well 30 µl precipitation solution (6.25 mM Tris HCl pH 7.6, 0.625 mM EDTA, 250 mM sodium acetate, 37.5% (v/v) isopropanol) were added using an eight-channel pipette (Socorex). After 1 hour at room temperature the plates were centrifuged at 2000x g (Beckman J6/B) for 30 min. The supernatant was drained off and the plates allowed to air dry at room temperature for 15 min. To each well 3 µl 1x TE was added using an eight-channel pipette and the plates were again centrifuged briefly to ensure even coverage of the bottoms of all wells and the DNA allowed to dissolve at 4°C for 24 hours. Plates were wrapped airtight and stored at -70°C.

2.2.12 cDNA arraying onto nylon membranes

cDNA clones were arrayed onto nylon membranes (Hybond N+) either as in situ DNA filters or as PCR product filters.

2.2.12.1 In situ DNA filters

In situ cDNA filters were produced using a modification of a previously described protocol for the preparation of high density cosmid arrays (Nizetic et al., 1991a). Using a further developed spotting robot from the one described in Nitzetic et. al. (1991b), *E.coli* cells stored in 384-well storage plates were inoculated onto 15 nylon membranes (Hybond N+, Amersham) using a 384-pin transfer device. The 22 cm x 22 cm nylon membranes were wetted in 2xYT medium + 50 μg ml^{-1} ampicillin and placed onto 2 Whatman 3MM filters also soaked in 2xYT medium + 50 μg ml^{-1} ampicillin. Typically 54 384-well plates (20,736 cDNA clones) were arrayed onto each membrane arranged in six fields of 384 9-clone boxes. After completion of the spotting cycle each membrane was transferred onto 2xYT agar + 50 μg ml^{-1} ampicillin and incubated at 37℃ for 16 hours. Filters were then stored at 4℃ for up to three days before in situ DNA preparation.

2.2.12.1.1 In situ DNA preparation

In situ DNA preparation was carried out according to Nitzetic et al. 1990. Membranes were placed colony side up on a Whatman 3MM filter soaked in denaturant (0.5 M NaOH, 1.5 M NaCl) at room temperature for 4 min. The membranes were transferred onto a fresh Whatman 3M filter soaked in denaturant and placed into a steam bath at 84℃ for 4 min. To neutralise, the membranes were transferred onto fresh Whatman filters soaked in neutralisation buffer (1 M Tris-HCl pH 7.6, 1.5 M sodium chloride) for 4 min. Membranes were then air dried for 10 min on Whatman 3M filters and subsequently transferred into 600 ml proteinase K solution (1 M Tris-HCl pH 8.5, 50 mM EDTA, 100 mM NaCl, 1% (w/v) sodium lauroyl sarkosinate, 0.28 mg ml^{-1} proteinase K (BDH)) prewarmed and incubated at 37℃ for 30 min. Each 600 ml lot of proteinase K solution was used for up to five membranes. After proteinase K treatment membranes were completely air dried on

Whatman 3M filters. Before use, DNA was crosslinked under UV light for 2 min.

2.2.12.2 PCR product filters

The cDNA PCR products were arrayed directly onto nylon membranes (Hybond N+, Amersham) using the same spotting robot as for the preparation of in situ filters. Again, liquid was transferred onto up to 15 nylon membranes using a 384-pin transfer device. The pins are spring loaded and have a 0.2 mm diameter tip, which transfers approximately 0.2 µl liquid. When un-precipitated PCR product was spotted, each plate was transferred five times onto the same spot, immobilising 5 - 50 ng PCR product. When precipitated PCR product was spotted only a single transfer was performed. The nylon membranes were soaked in denaturant (0.5 M NaOH, 1.5 M NaCl) and placed onto a Whatman 3M filter also soaked in denaturant. DNA was arrayed in the same format as for in situ filters with genomic salmon sperm DNA in the centre of each 3x3 block. The salmon sperm DNA was spotted at 20 ng μl^{-1}.

After completion the filters were placed onto fresh Whatman 3M filters soaked in denaturant for 2 min and then transferred onto Whatman 3M filters soaked in neutralisation solution (1 M Tris-HCl pH 7.6, 1.5 M NaCl). for 4 min. Membranes were then submerged in neutralisation solution for 2 min and subsequently air dried.

2.2.13 DNA radiolabelling

2.2.13.1 Random primed labelling

Double stranded DNA was labelled by random priming according to Feinberg and Vogelstein (1983). Purified DNA was heated in a boiling waterbath in a volume of 35 µl for three minutes and then cooled rapidly

on ice. and the following reagents added: 5 µl 5 x labelling buffer (1 M Hepes pH 6.6, 250 mM Tris-HCl, 25 mM magnesium chloride, 50 mM 2-mercaptoethanol, 100 µM dTTP, 100 µM dCTP, 100 µM dGTP, 25 OD_{260} units ml^{-1} hexadeoxyribonucleotides), 2 µl bovine serum albumin (10 $mgml^{-1}$), 1µl *E.coli* DNA-polymerase I (Klenow fragment) (10 U $µl^{-1}$) and 2 µl [^{32}P-α] dATP (10 µCi $µl^{-1}$). The reaction was incubated either at 37°C for 1 hour or at room temperature overnight. 3 µl EDTA (0.5M) were added to the reaction and the polymerisation products precipitated with 2 µl yeast t-RNA (10 mg ml^{-1}), 45 µl ammonium acetate (5 M), 200 µl 100% ethanol, placed on dry ice for 5 min. and centrifuged at 11,000x g for 20 min. The pellet was resuspended in 100 µl 1x TE (10 mM Tris-HCl pH 7.6, 1 mM EDTA) and a 1µl aliquot assayed for radioactivity by Cerenkov counting. Unless used immediately, labelled DNA was stored at -20°C.

2.2.13.2 Terminal labelling

Oligonucleotides were labelled at their 5' termini by phosphate transfer using T4 polynucleotide kinase. 30 pmol oligonucleotide was labelled in a 30 µl reaction containing: 3 µl 10x buffer (700 mM Tris-HCl pH 7.6, 100 mM magnesium chloride, 50 mM dithiothreitol), 2 µl T4 polynucleotide kinase (10 $Uµl^{-1}$, NEB) and 5 µl [^{33}P-γ] ATP (10 µCi $µl^{-1}$. The reaction mixture was incubated at 37°C for 45 min and then terminated by the addition of 3.5 µl EDTA (0.5 M). To assess incorporation of radioisotope, 1 µl aliquots were spotted onto polyethyleneimine (PEI) chromatography sheets (Polygram, Macherey-Nagel, Düren, Germany), air dried and vertically chromatographed in 0.75 M potassium di-hydrogenphosphate pH 3.5. The solvent front was allowed to migrate at least 10 cm. The chromatography sheet was wrapped in saran wrap and exposed to X-ray film for 30 min. Unless used immediately, labelled oligonucleotides were stored at -20°C.

2.2.14 DNA hybridisation

2.2.14.1 Random primed DNA hybridisation

2.2.14.1.1 Hybridisation

DNA bearing nylon membranes (in situ YAC, cosmid, cDNA or southern blots) were hybridised according to Church and Gilbert (1984). Membranes were prehybridised in 'Church' buffer (0.5 M sodium phosphate pH 7.2 (note: 0.5 M with respect to Na^+), 7% (w/v) sodium dodecyl sulphate (SDS), 1 mM EDTA, 0.1 mg ml^{-1} yeast t-RNA) in a sealed plastic bag at 65°C for at least 30 min. Random prime labelled DNA was heated to 100°C for 3 min. and then added to Church buffer to a final concentration of 1 x 10^6 cpm ml^{-1}. Prehybridised membranes were drained, hybridisation solution added and hybridised at 65°C for 3 hours to overnight.

2.2.14.1.2 Washing

Hybridisation bags were unsealed the hybridisation solution drained off and the membranes transferred into 1 litre wash solution (40 mM sodium phosphate pH 7.2, 1% (w/v) SDS) at room temperature for 5 min. Membranes were then transferred into fresh wash solution at room temperature and incubated at 65°C for 20 min. If membranes still contained > 100 cps as measured by a hand held Geiger-monitor then the last wash was repeated a second time. Membranes were briefly air dried on Whatman 3M filter paper to remove excess liquid and then wrapped air tight in saran wrap and used for autoradiography.

2.2.14.1.3 Stripping

To remove bound radioactive material up to ten membranes were incubated in 1 litre 5 mM Sodium phosphate pH 7.2, 0.1% SDS at 90°C

for 30 min. Filters were dried between Whatman 3M filters and stored dry.

2.2.14.2 Oligonucleotide hybridisation

2.2.14.2.1 Hybridisation

For oligonucleotide hybridisation only filters with cDNA PCR products were used. The nylon membranes were briefly placed in SSarc buffer (600 mM sodium chloride, 60 mM sodium citrate, 7.2%(w/v) sodium lauroly sarcosinate (Sarkosyl N30, BDH)) to pre-equilibrate. Oligos were labelled at 5' ends with [^{33}P-γ] ATP using T4 polynucleotide kinase and hybridised at 4 nM in SSarc buffer, at 4°C for 3 hours - overnight. Typically two 22 cm x 22 cm membranes were hybridised in one 300 mm glass bottle with 30 mm diameter (Hybaid) in a volume of 10 ml. Filters were separated by two 23 cm x 23 cm nylon meshes (Hybaid) to allow free access to the hybridisation solution to all areas of the membranes. Bottles were sealed with a rubber bung and rotated horizontally on a purpose built roller at 5 rpm.

2.2.14.2.2 Washing

After hybridisation membranes were rinsed briefly in cold (4°C) SSarc buffer and washed together with the nylon meshes in 1 litre SSarc buffer in a polypropylene lunch box at 10°C for 15 - 30 min. Up to 8 membranes were washed together in 1 litre SSarc even if they had been hybridised with different oligonucleotides. After washing membranes were separated from the nylon meshes and drained of excess liquid on Whatman 3M paper at room temperature. Membranes were never allowed to dry completely while radioactive. Membranes were placed onto the polythene side of Benchkote (Whatman) and wrapped air tight in saran wrap and used for autoradiography.

2.2.14.2.3 Stripping

To remove all bound radioactive oligonucleotide, up to 20 membranes were incubated twice in 1 litre 0.1 x SSarc at 65°C for 10 min. The 0.1 x SSarc was boiled immediately prior to pouring onto the membranes. After stripping membranes were stored either in SSarc buffer or used immediately in another hybridisation.

2.2.15 DNA sequencing

Sequencing of cDNAs was performed by a dideoxy-termination method on double strand DNA. All sequencing reactions were carried out using a Sequenase v2.0 DNA sequencing kit (UBS, Cleveland, USA). DNA was prepared by alkaline lysis from 5 ml of an overnight culture and resuspended in 20 µl water. To denature the template DNA, 5 µl of a prep were diluted with 13 µl water and 2 µl of 2 M NaOH, 2 mM EDTA were added. The DNA was precipitated with 2 µl 3 M NaAc pH6 and 50 µl ethanol (100%). After centrifugation the pellet was washed once in 70% ethanol and resuspended in 7 µl water.

To anneal the sequencing primer, 1 µl primer and 2 µl sequencing buffer were added. The sample was then heated in a 65°C waterbath and allowed to cool to 35°C over a period of 30 min. Sequencing reactions were then carried out, using [^{35}S-α]dATP, as shown in the protocol supplied with the kit.

The sequencing reactions were run out on 30cm x 40 cm 6% polyacrylamide denaturing gels (6 M urea) at 60 mA (constant Watts) for 4 hours.

3. cDNA library construction

3.1 Mouse adult brain cDNA library construction

3.1.1 cDNA synthesis

In order to familiarise myself with the technique of cDNA cloning without wasting precious materials, I decided to generate a cDNA library from mouse adult brain tissue. Brain tissue was chosen in the expectation that the resulting cDNA would contain the highest possible complexity and therefore offer a valuable resource for the greatest number of applications.

Mouse adult brain poly-A+ RNA was kindly provided by Greg Lennon and used for the construction of cDNA using a commercial Superscript cDNA kit from Gibco-BRL. The detailed protocols are described in the *Materials and Methods* section.

5 µg poly-A+ RNA were used in the first strand synthesis and 1 µCi [^{32}P-α] dATP included. The radiolablled cDNA was easier to trace throughout the cloning procedure and allowed quantitation of the cDNA yield at the end of the process. After the size selection using a sephacryl S400 column, Cerenkov counts were determined for the collected fractions. The data are shown in table 3-1.

Table 3-1

Fraction Number	Fraction Volume (µl)	Total Volume (µl)	Cerenkov Counts (cpm)	cpm per µl	Amount of cDNA (ng)	Conc. of cDNA (ng μl^{-1})
1	93	93	20	0.22	2	
2	95	188	20	0.21	2	
3	35	223	20	0.57	2	
4	36	259	20	0.56	2	
5	26	285	0	0.00	0	
6	35	320	30	0.86	3	
7	35	355	230	6.57	24	0.69
8	25	380	920	36.80	96	3.84
9	37	417	2350	63.51	245	6.63
10	36	453	3360	93.33	351	9.74
11	36	489	3550	98.61	371	10.29
12	33	522	3890	117.88	406	12.30
13	35	557	3410	97.43	356	10.17
14			3640		380	
15			4290		448	
16			2030		212	
17			1420		148	
18			1280		134	
19			370		39	
20			370		39	

A total of 3.2 µg cDNA were synthesised which corresponds to an efficiency of approximately 30%. Fraction 8 was the largest size fraction with a high yield of cDNA and fraction 12 the last fraction for which the eluate volume was below 550 µl. The cDNA kit has been optimised in such a way that up to a volume of 550 µl of eluate contains cDNA of an average size in excess of 500 bp.

3.1.1.1 Transformation controls

In an initial test ligation and transformation experiment the number of expected non-recombinant background colonies was determined. In an identical ligation reaction to the one specified in the cDNA cloning protocol (see *Materials and Methods)*, 50 ng pSPORT1 vector DNA

(precut with *Not*I / *Sal*I and dephosphorylated) were ligated in the absence of cDNA and then transformed by electroporation into *E.coli* XL1-blue cells, using a Biorad Genepulser with the following settings: 2.5 kV, 200Ω, 25 μFad. The results are shown in table3-2:

Table 3-2

DNA transformed	# colonies	cfu μg-1
1 ng unligated pSPORT1	0	-
3 ng ligated pSPORT1	26	2.6×10^3
6 pg pBluescript	89	8.9×10^7

These results indicate that the vector DNA supplied with the cDNA cloning kit is completely linearised, but that there is a small fraction which yields closed circular DNA upon ligation. The most likely explanation for this is that there are some molecules which have been digested only with one of the two cloning restriction endonucleases (*Not*I, *Sal*I) and that some of these are incompletely dephosphorylated. The expected background of 2.6×10^3 cfu μg^{-1} of vector DNA is tolerable for the planned cloning experiments.

In a test ligation 10 ng of fraction 8 and fraction 12 cDNA were ligated separately to the precut pSPORT vector DNA as described in the *Materials and Methods* section. Using 2 mm cuvettes, 1/5 (1 μl) of each of the precipitated ligations were transformed into *E.coli* XL1-blue cells by electroporation with the same settings as for the control transformation described in the preceding paragraph. Two serial dilutions of 1:10 were made of the transformed cells and 1/4 of each dilution plated on agar petri dishes (+50 $\mu g\ \mu l^{-1}$ ampicillin) and then incubated at 37℃ overnight. Table 3-3 shows the results of these transformations. The results indicate that the total amount of cDNA will be sufficient to plate enough cDNA clones (100,000) for arraying a library of useful size into microtitre plates.

Analysis of a sample of cDNA inserts form both fractions #8 and #12, showed that the average sizes were approximately 1,500 bp and 300 bp respectively. This indicates that fraction #12 contains cDNA that are undesirable small for the purposes of a cDNA library.

Table 3-3

DNA transformed	# colonies in 1/4 transformation	total cfu in ligation
fraction #8		
dilution: 1/100	14	
dilution: 1/10	131	
dilution: 1/1	~1,400	~28,000
fraction #12		
dilution: 1/100	13	
dilution: 1/10	200	
dilution: 1/1	~2,000	~40,000
1 pg pBluescript	100	1×10^8 cfu μg^{-1}

3.1.2 cDNA library arraying into microtitre plates

In collaboration with Peter Jones at the LMB Cambridge 20,000 mouse brain cDNAs were picked using prototype colony picker which the Cambridge group had been developing (Jones et al., 1992). A ligation reaction of fraction #8 of the mouse adult brain cDNAs was transformed in XL 1 blue cells and plated on rectangular (11 cm x 8 cm) petri dishes, at a density of 600 colonies per dish. Single colonies were inoculated into separate wells of 96-well microtitre plates filled with 2xYT medium + Hogness modified freezing mix + 50 $\mu g\ ml^{-1}$ ampicillin. Plates were incubated at 37℃ overnight and then stored at -70℃ .

In order to assess the quality of the cDNA library and the efficiency of the colony picking robot, PCR amplification was performed on 96

clones and the products analysed on agarose gels. PCR reactions were carried out in a Cetus-9600 machine. Figure 3-1 shows an ethidium bromide stained agarose gel of 10 µl of each reaction. 87 out of 96 wells (90%) contained cDNA inserts amplifiable by PCR and the average size was 1,400 bp. The samples in figure 3-1 are considerably overloaded and show nonspecific artifacts as well as multiple bands in some cases. It is difficult from PCR reactions of cDNA clones to determine reliably whether a single clone has been amplified, or whether multiple clones are present in some samples. Most reactions show one major product. The extent of extra products obtained on a PCR reaction is greatly affected by the amount of template DNA used (observation over many PCR amplifications and personal communication). Since all the samples shown in figure 3-1 were inoculated directly from bacterial liquid cultures, it is likely that the inoculum varied substantially between reactions. The gels show at least 8 samples (8.3%) in which there are multiple bands of similar intensity, which might suggest that these wells contain more that one cDNA clone.

This cDNA library has been made publicly available through the ICRF Reference Library system (Zehetner and Lehrach, 1994) in the form of high density filter arrays. To date 18 filter sets have been distributed and 34 clones identified from the library.

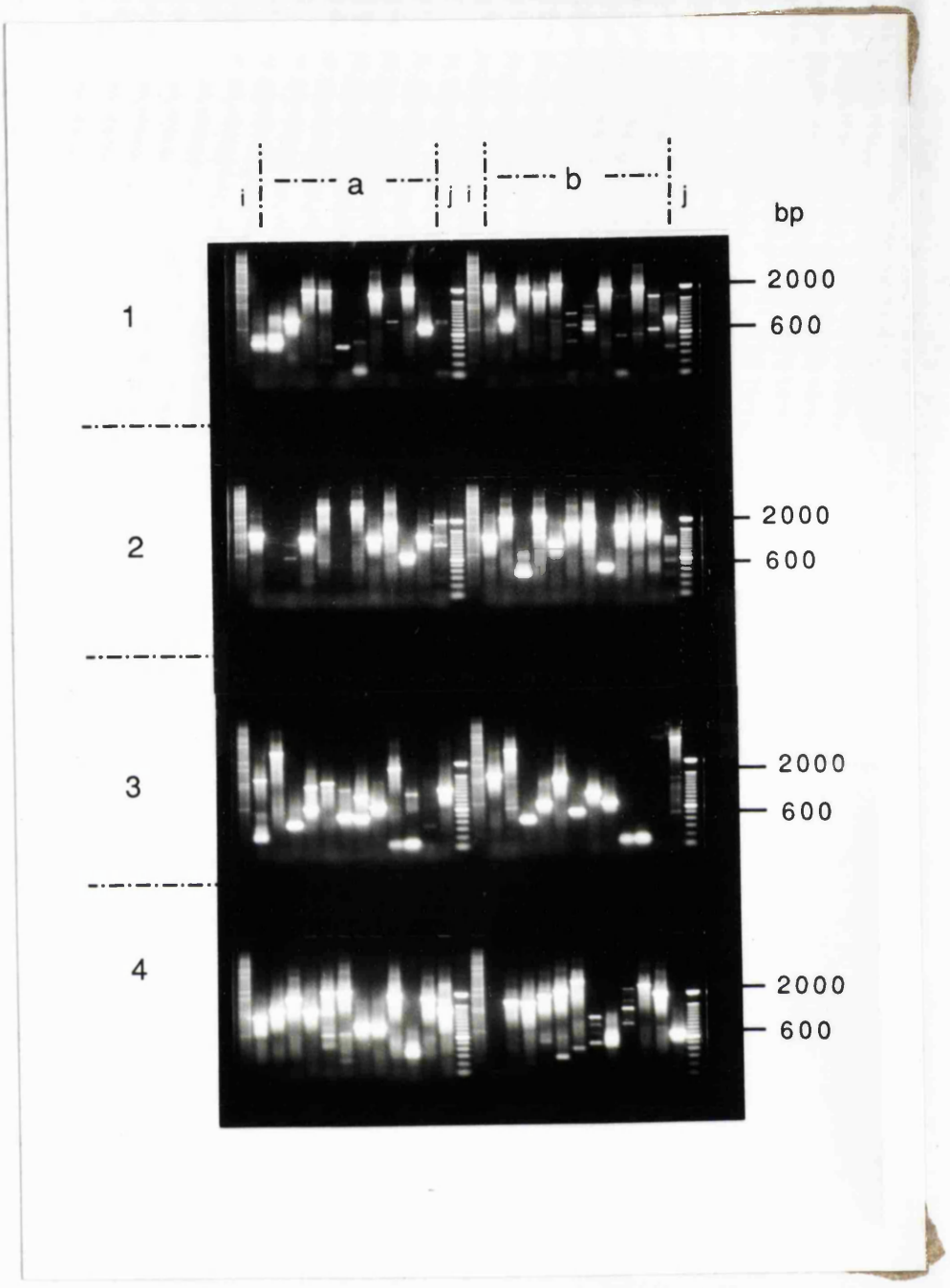


Fig. 3-1 Two 1% agarose gels of 96 human foetal brain cDNA PCR products. 30 cycles of 1 min 94°C and 2 min 73°C were performed on a Cetus 9600 machine. 10 µl of a 30 µl reaction were loaded in each lane. The size markers used were a 100 bp ladder (j) and *BstE*II digested lambda DNA.

3.2 Human foetal brain cDNA library construction

For the purpose of fingerprinting by oligonucleotide hybridisation, with the aim of studying the expression of many thousands of genes, brain tissue was chosen in order to generate a cDNA library with the maximum possible complexity. Due to the rapid degradation of RNA in dead tissue it is important to use tissue that has been frozen as soon as possible after extraction. The most readily available source of human tissue is from aborted foetuses. Frozen human foetal brain tissue (stored at -70°C) was therefore obtained from a 14 week foetus.

3.2.1 RNA isolation from frozen tissue

Total RNA was extracted form 1.19 g frozen human foetal brain tissue by guanidinium isothiocyanate lysis followed by centrifugation through CsCl as described in *Materials and Methods* section . A total of 350 µg RNA were extracted an resuspended in 500 µl RNase free water.

3.2.1.1 poly(A)+ RNA isolation

A commercial 'Fast Track' kit from In Vitrogen was used to isolate poly-A+ RNA from the total RNA extracted from the frozen foetal brain tissue. This protocol is based on poly-dT cellulose chromatography. All 350 µg total RNA extracted from the human foetal brain tissue were used in one poly-A+ RNA selection. Due to the small amount of poly-A+ RNA expected from 350 µg total RNA (approximately 2% (Sambrook et al., 1989)) it was not possible to assess the quality of the extracted poly-A+ RNA without losing the majority of the sample. After the final elution and subsequent precipitation the extracted RNA was resuspended in a volume of 6 µl RNase free water.

3.2.2 cDNA synthesis

cDNA was again synthesised using the same commercial cDNA kit as for the mouse adult brain library. 5 µl of the human foetal brain poly A+ RNA isolated as described above were used for first strand cDNA synthesis. The protocol followed was exactly as for the mouse adult brain cDNA library. Problems occurred however during the size fractionation with the prepacked Sephacryl S-500 column. The column became blocked and required repacking after the cDNA had been loaded. As a consequence, the size fractionation did not proceed as intended. Although the cDNA was radioactively labelled, multiple attempts to obtain readings by Cerenkov counting were unreproducible. Later investigation indicated that considerable interference with Cerenkov counting can be caused by static electricity that accumulates on polypropylene tubes when handled with certain types of latex gloves.

As an alternative method for determining the amount of cDNA in each of the size fractions, 1 µl of each fraction along with a series of DNA samples of known concentration were spotted onto 1% agarose containing 0.5 mg ml^{-1} ethidium bromide. Figure 3-2 shows a polaroid photograph of the samples of agarose taken under UV light. This method gives a reasonably accurate measure of the amount of cDNA in each fraction, which corresponded to approximately 1-2 ng μl^{-1} for fractions #11, 12, 13 and 14.

In a test ligation of 10 ng of fraction #13, 12,000 clones were obtained for 1/5 of the ligation. The average size of small sample of inserts for this ligation was determined to be approximately 1,110 bp by PCR amplification.

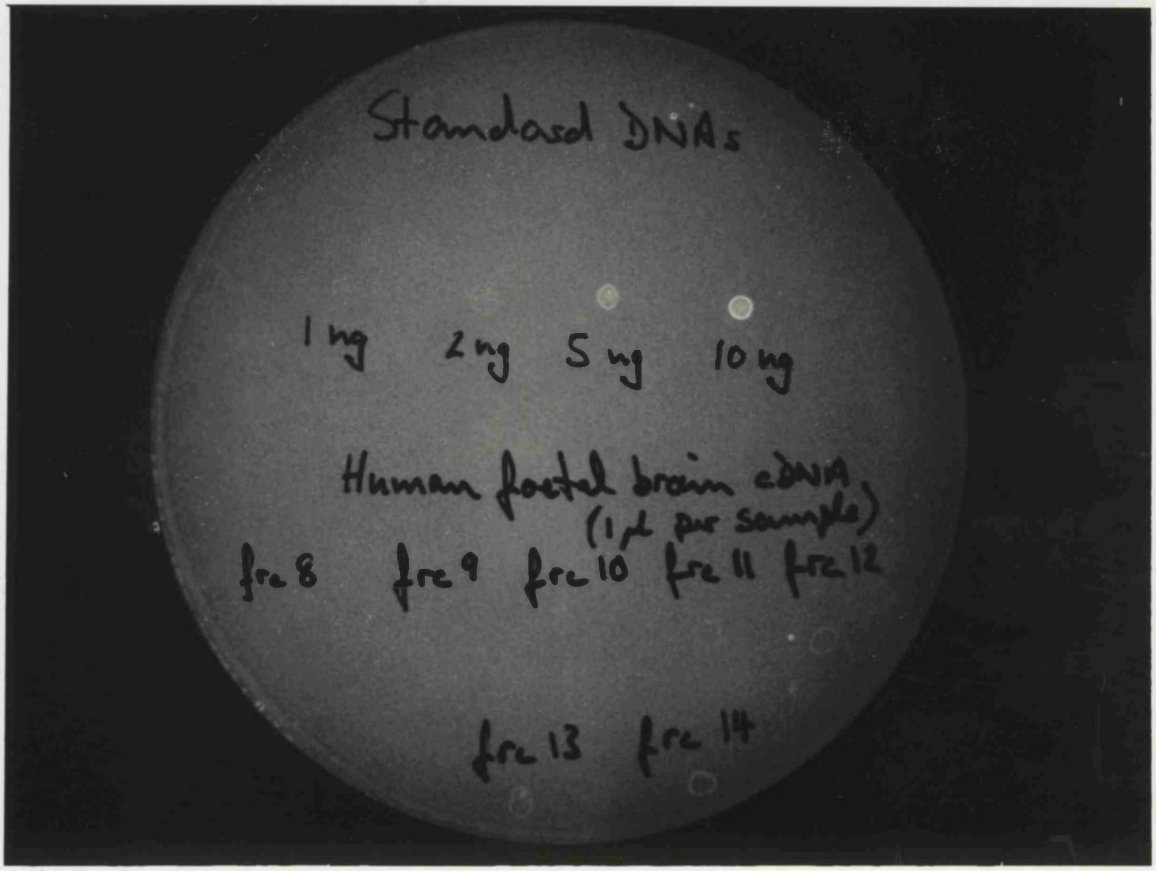


Fig. 3-2 1 µl aliquots of human foetal brain cDNA fractions spotted onto 1% agarose containing 0.5 µg ml^{-1} ethidium bromide. Also spotted are 1 µl aliquots of known conentration DNA standards.

3.2.3 cDNA library arraying into microtitre plates

Initially 20,000 clones generated from the fraction #13 ligation were picked robotically in collaboration with Peter Jones as described for the mouse adult brain cDNA library.

As part of the development of an integrated set of automated procedures for the highly parallel analysis of many thousands of clones (Meier Ewert et al., 1993), a robotic colony picking device was developed in the lab. The advent of extensive comparative mapping projects, encompassing more than half a dozen different organisms, and the increased use of arrayed cDNA libraries have made the process of picking randomly plated colonies into microtitre plates a rate limiting step. There is a clear need for the automation of this process for which there were no commercial appliances at the time. A machine was developed that is able to pick *E.coli* and yeast colonies into microtitre plates at a rate of 3,000 clones per hour. A detailed description has been published in Maier et al. (1994a).

A further 80,000 clones were arrayed into microtitre plate analogs, using the robotic colony picking machine developed in the lab. Approximately 95% of wells contained grown culture after overnight incubation 37℃. These clones were picked into 384-well plates (Q-plates) that have the same footprint of conventional 96-well microtitre plates. The spacing between well centres in 96-well plates is 9 mm, where as that in Q-plates is 4.5 mm. Q-plates therefore contain twice as many rows (labelled A to P) and twice as many columns (numbered 1-24). This format allows conventional microtitre plate accessories such as multichannel pipettes and plate shakers to be used with Q-plates, while reducing the space requirement of any given clone library by a factor of four. Since storage of clones at -70℃ is a limiting factor the reduced space requirement is significant. Q-plates have a total volume of 70 µl per well of which 50 µl are usable, taking into account expansion

of the volume during freezing. Figure 3-3 shows a photograph of a Q-plate.

3.2.4 Quality assessment by hybridisation

High density clone arrays on nylon membranes were generated as described in *Materials and Methods*, using a robotic spotting machine developed in the lab. Clones were arrayed at a density of 20,736 per 22 cm x 22 cm membrane (54 Q-plates).

In order to assess the quality of the cDNA library, several hybridisation experiments were carried out with probes of genes, whose expression levels are known. The number of positively hybridising clones, immediately indicates whether a known sequence is represented in the library at the expected level. Figures 3-4 - 3-8 show hybridisation results with the following probes: total human DNA, mouse β-actin, ribosomal gene, poly-A oligo and glyceraldehyde-3-phosphate dehydrogenase (GAPDH).

The total human DNA probe is expected to hybridise with those cDNA clones containing common repetitive sequence elements (such as Alu sine element), since the complexity of the probe is so high that single copy sequences have insufficient specific activity to detect complementary cDNA clones. The hybridisation shows that approximately 1.5% of clones hybridise with total human DNA. This result is consistent with some findings that suggest between 1 - 2% expressed proteins contain Alu repetitive elements (Makalowski et al., 1994). Although Alu elements form the majority of repetitive elements in the human genome there are likely to be others that give detectable signal with a total human DNA probe. While this hybridisation by no means gives a measure of the number of transcribed repeats, it does indicate that the cDNA library does not contain an unexpectedly large number of repetitive elements, a problem commonly encountered when a significant genomic DNA contamination has occured.

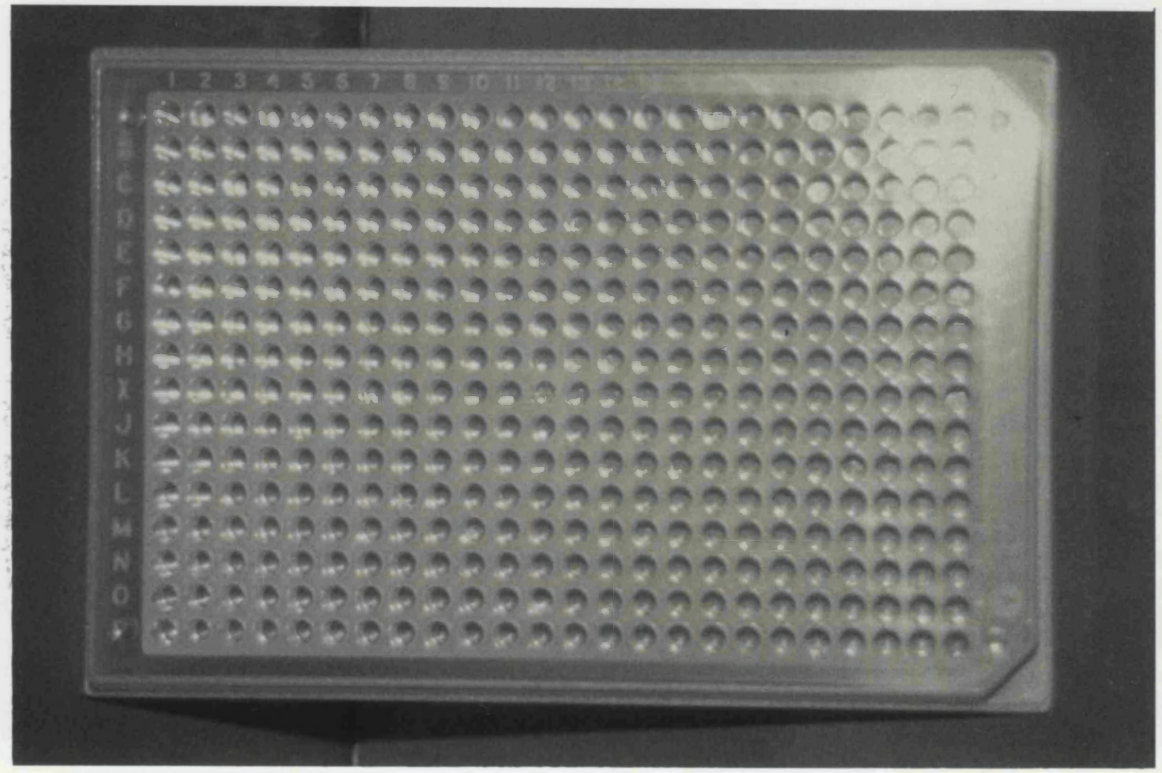


Fig. 3-3 A photograph of a 384-well microtitre plate (Q-plate), showing twice as many rows (A - P) and twice as many columns (1 - 24) as a conventional 96-well microtitre plate. Plates are made in two materials, polystyrene and polypropylene.

Fig. 3-4 An autoradiograph of a hybridisation with a total human DNA probe on a high density nylon filter carrying 21,120 human foetal brain cDNA clones. 150 ng human DNA were labelled with [^{32}P-α] dATP by random priming and hybridised according to Church and Gilbert (1984). Approximately 2% of clones show a positive hybridisation signal.

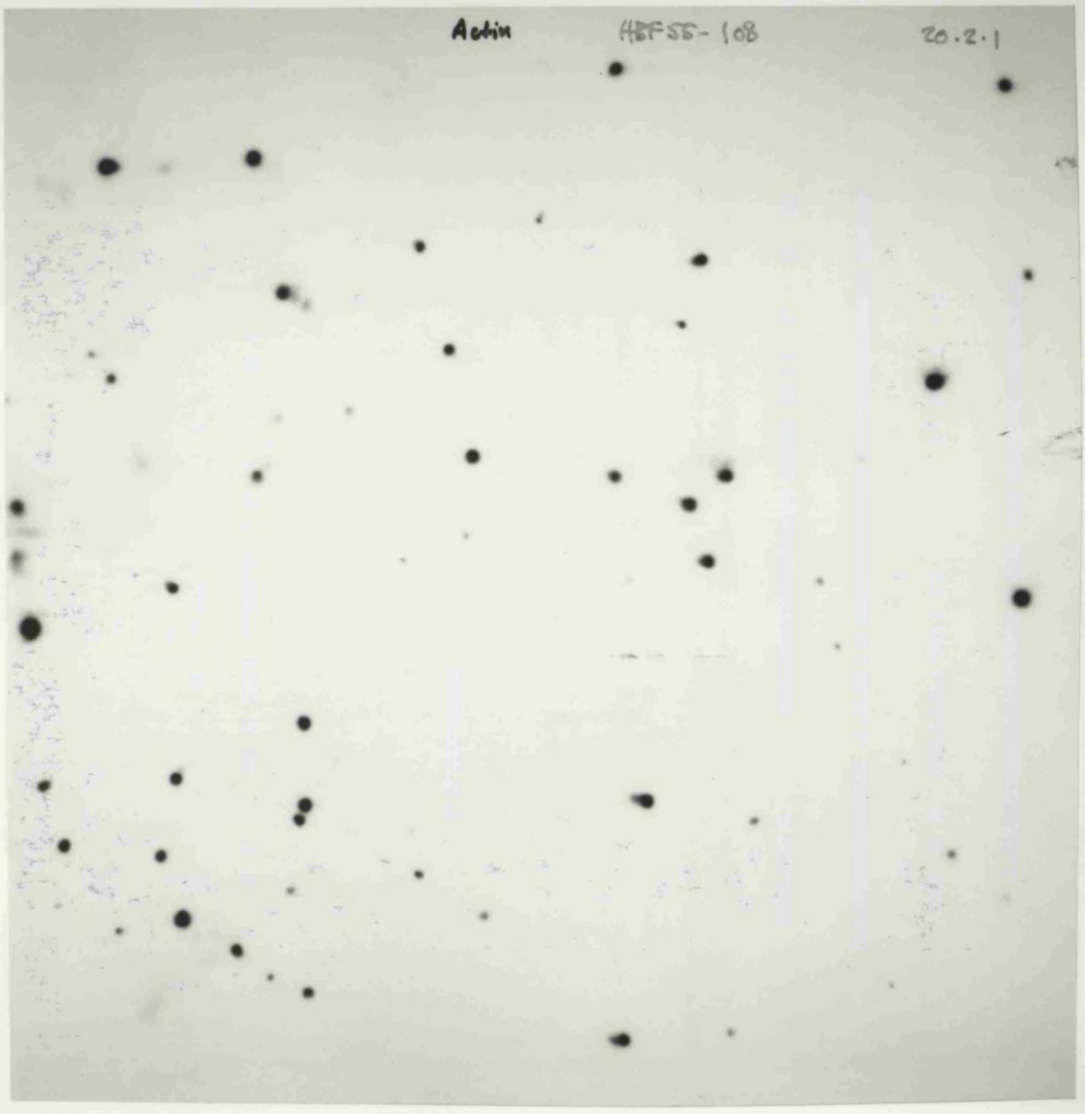


Fig. 3-5 An autoradiograph of a hybridisation with a mouse actin DNA probe on a high density nylon filter carrying 21,120 human foetal brain cDNA clones. 50 ng DNA were labelled with [^{32}P-α] dATP by random priming and hybridised according to Church and Gilbert (1984).

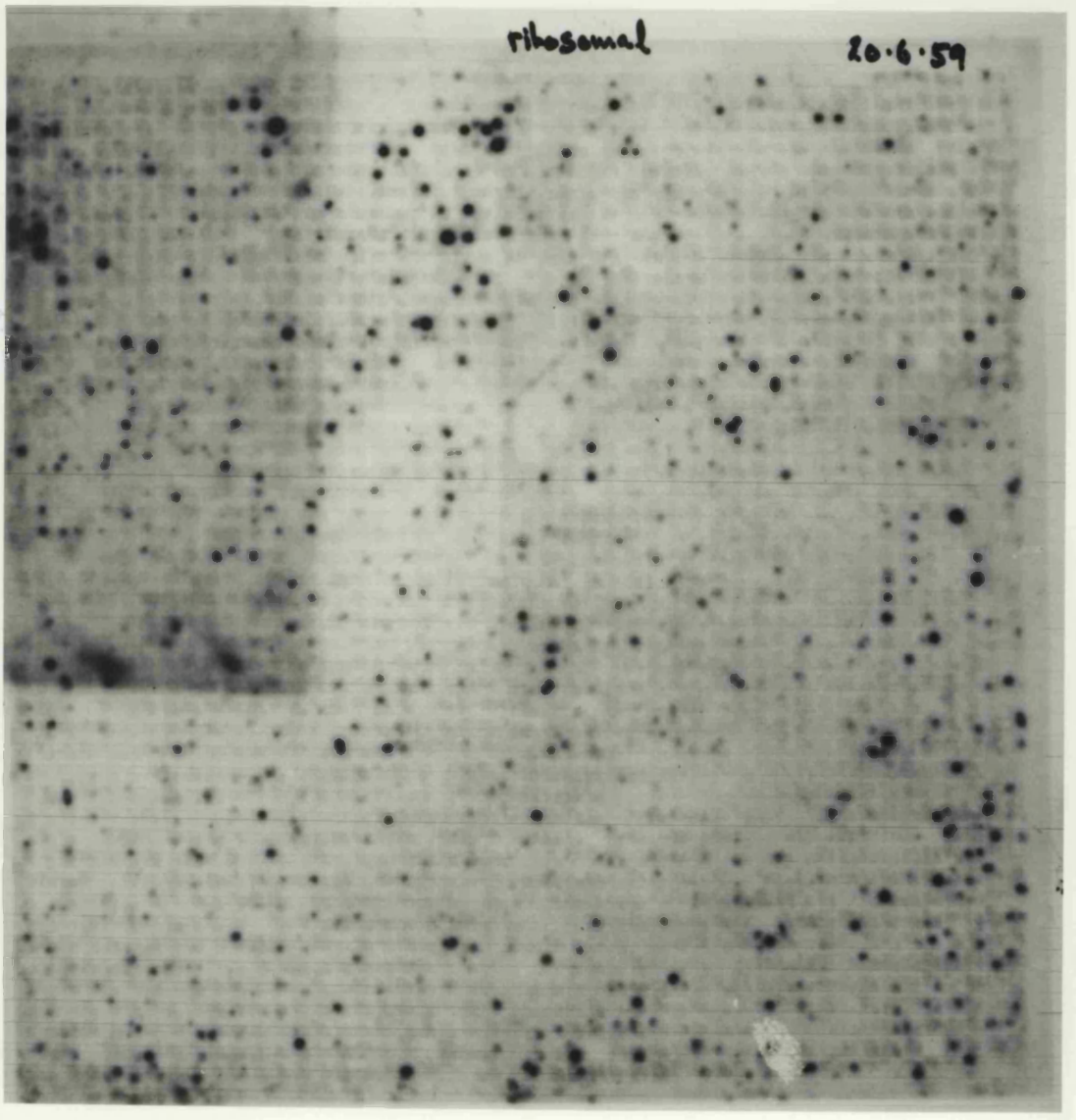


Fig. 3-6 An autoradiograph of a hybridisation with human ribosomal precursor gene DNA probe on a high density nylon filter carrying 21,120 human foetal brain cDNA clones. 50 ng DNA were labelled with [^{32}P-α] dATP by random priming and hybridised according to Church and Gilbert (1984).

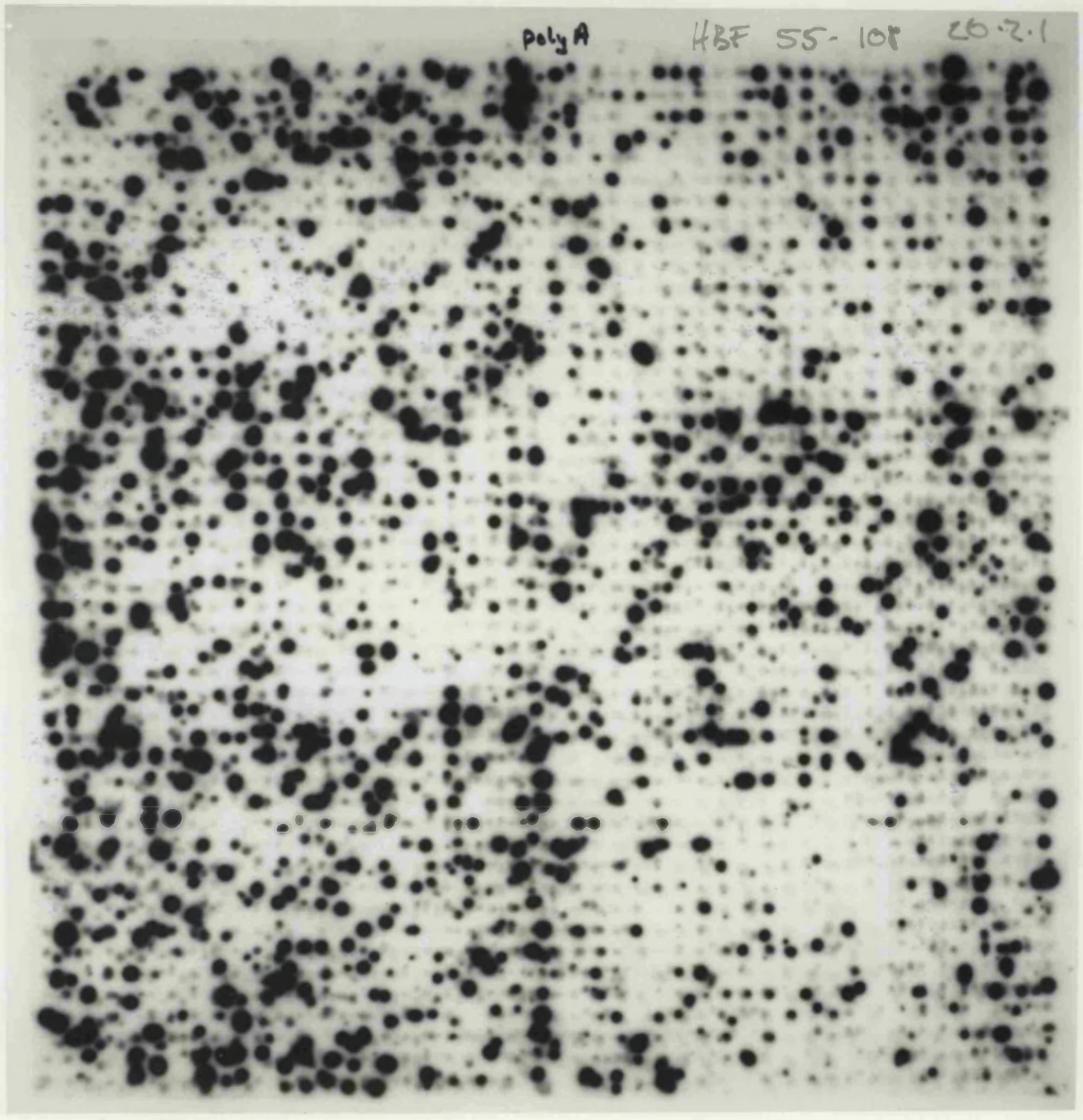


Fig. 3-7 An autoradiograph of a hybridisation with oligo-dA DNA probe on a high density nylon filter carrying 21,120 human foetal brain cDNA clones.The average length of oligomers was 40 bp. 200 ng DNA were labelled with [^{32}P-γ] ATP using T4 polynucleotide kinase and hybridised according to Church and Gilbert (1984).

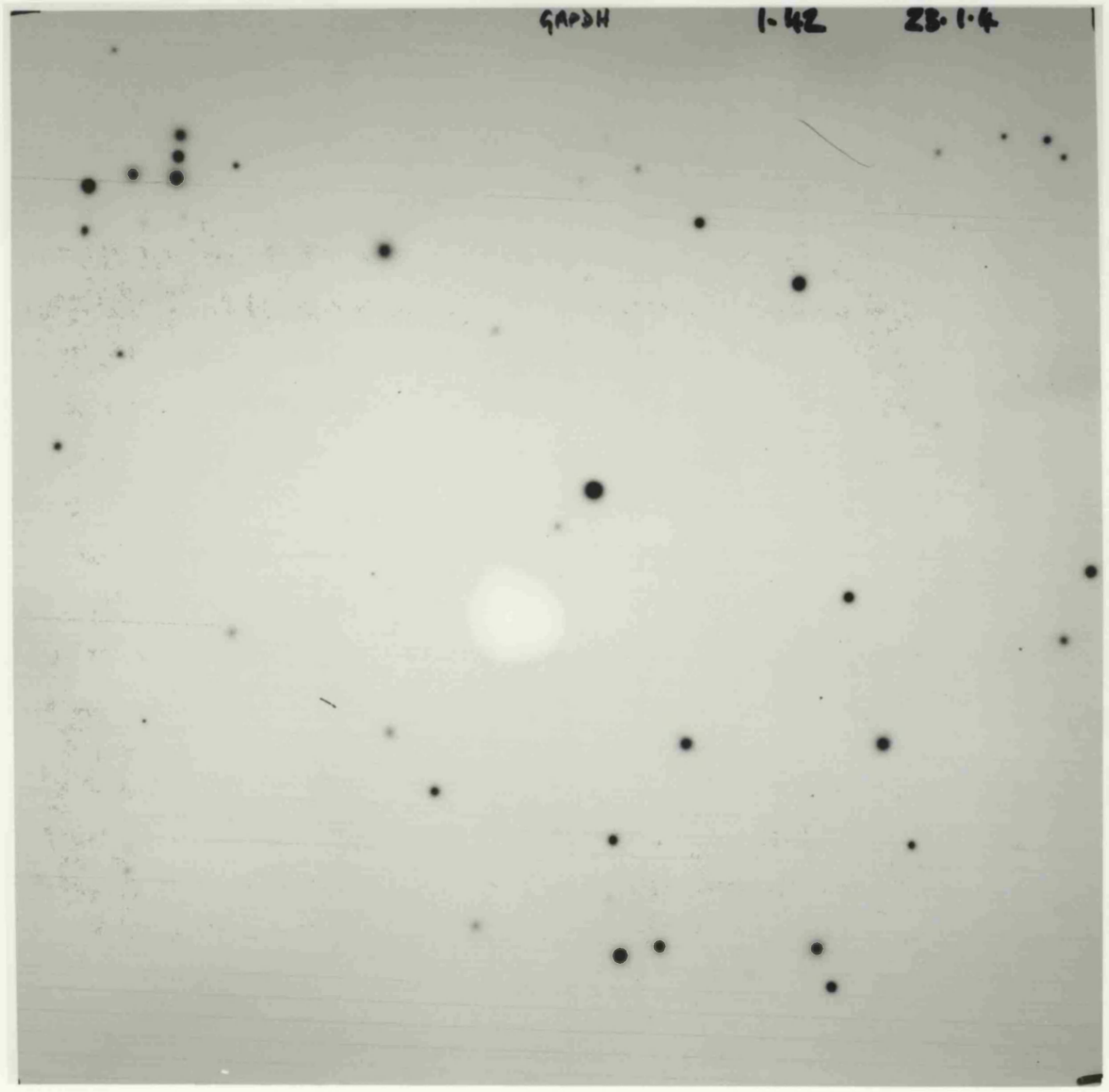


Fig. 3-8 An autoradiograph of a hybridisation with human glyceraldehyde 3-phosphate dehydrogenase (GAPDH) DNA probe on a high density nylon filter carrying 21,120 human foetal brain cDNA clones. 50 ng DNA were labelled with [^{32}P–α] dATP by random priming and hybridised according to Church and Gilbert (1984).

A mouse β-actin cDNA probe identified many homologous human cDNA clones in the library an expected result in view of the high homology between the mouse and human β-actin genes. The homology within the whole of the actin gene family makes it likely that the cDNA clones identified by the hybridisation comprise a mixture of α-actin, β-actin and γ-actin all of which are actively transcribed in brain tissue. No attempts were made to characterise the actin clones further and assign them to sub-family groups. Hybridisations with both the ribosomal RNA (rRNA) precursor gene and the poly-A oligo indicate that while there are many hybridising clones with both probes the numbers are not so high as to throw into doubt the complexity of the library. Two of the most common shortcomings of cDNA libraries are that they either contain a very high proportion of rRNA derived clones due to rRNA priming during first strand cDNA synthesis (a problem generally associated more with cDNA libraries generated by random hexamer priming of mRNA) or a high proportion of clones that contain almost exclusively poly-dA tracts.

The GAPDH probe identified approximately 0.3% of cDNA clones in the library, a result in line with the fact that the glyceraldehyde-3-phosphate dehydrogenase enzyme carries out one of the steps in one of the most active catabolic pathways in brain tissue, glycolysis. A subset of the GAPDH clones was analysed further by PCR amplification of the cDNA inserts. Figure 3-9 shows the PCR products run on an agarose gel. The length of the GAPDH complete coding sequence, extracted from the Genbank database, is 1,268 bp. The gel shows that 71% of the clones are approximately 1,200 bp in length and that none are longer. This suggests that 71% of the GAPDH clones are full length and that none of the clones analysed are derived from unspliced nuclear RNA molecules. While it is tempting to speculate that approximately 70% of 1,200 bp genes are represented as full length in the cDNA library, this assumption cannot be made given variable cloning efficiencies of different DNA sequences.

Fig. 3-9 A 1% agarose gel of PCR amplifications carried out on 34 clones hybridising positively with GAPDH probe. 5 µl of each PCR reaction were loaded in each lane. The size markers used were *BstE*II digested lambda DNA.

The overall assessment of the human foetal brain cDNA library based upon the tests discussed above, is that its representation and complexity are in line with expectation for a library constructed from human brain tissue.

The library has been widely distributed in form of high density filter arrays to 56 labs and 2830 potential positive clones have been identified.

4. Development of oligonucleotide fingerprinting tools

4.1 Waterbath PCR

For the approach of oligonucleotide fingerprinting purified target DNA is required due to the high hybridisation frequency of the probes. For random sequence DNA an oligonucleotide sequence of length *n* will occur every 4^n bases. An average octamer would therefore occur once every 65,536 bases, or once every 32,768 bp of double stranded DNA. The standard technique in the laboratory for immobilising clone DNA on nylon membranes at high density is to perform a crude alkaline lysis on arrayed colonies that have been spotted and grown directly on the membranes (Nizetic et al., 1991a). This procedure attaches the entire cosmid, including vector, and the *E.coli* genome to each spot. This procedure yields sufficient DNA for detection of single copy cosmid sequences by hybridisation, ranging from 'long probe' labelled by random priming to oligonucleotides as short as 11mers. However, hybridisation data with 11mer oligonucleotides generated on cosmids of *S. pombe* have proved very difficult to interpret (Hoheisel et al., 1993). Even once the map order of the cosmids had been established, the oligo data still proved inconclusive. The difficulty in interpretation has been attributed mainly to the high amount of background signal that was present in most hybridisations, assumed to be mainly due to the *E.coli* present at each position in the clone grid. Since for the purpose of characterising cDNA clones by oligonucleotide hybridisation, shorter oligos are required than for mapping cosmids, it becomes imperative to have purified target DNA.

Whatever system of purification is chosen, it must lend itself to a large scale-up to allow the analysis of many tens of thousands of clones which are required to obtain even a moderately representative cDNA

analysis. The polymerase chain reaction (PCR) is a technique that is already automated in terms of the amplification and could in principle be scaled up providing the set-up of the individual reactions can be simplified. PCR had the great advantage of yielding virtually pure samples since the amount of template DNA required is far below the detection limit of any hybridisation system. In addition cDNA clones are of suitable length for reliable amplification using more or less standard amplification conditions. One drawback however, is that long cDNA clones might exceed the maximum length of DNA that can be reliably amplified by PCR (reports on the length of amplification products vary greatly, but most lie below 10,000 bp). There have been reports of amplification over 35 kb (Barnes et al., 1994) but these protocols have not yet been applied to this amplification system Over the past three years several commercial thermocycler machines have become available in which 96 reaction can be carried out in parallel. With a goal of analysing 100,000 cDNA clones by oligonucleotide hybridisation, these machines still do not represent sufficient scale-up since over 1,000 amplification experiments would be required and the set-up is still labour intensive. Some machines are designed to allow amplification in microtitre plates directly, which seemed an attractive possibility especially in view of the fact that the arrayed cDNA clones are stored in microtitre plates (or Q-plate analogues thereof) and much automation had been developed for the microtitre plate format.

4.1.1 Initial tests

Many formats for large scale PCR experiments were considered over a period of months, most of which were concerned with performing many microtitre plate based PCR reactions in parallel. It soon became apparent that the simplest way to scale up PCRs in microtitre plates was to use constant temperature waterbaths to equilibrate the reactions to the required temperatures, since transferring samples form one constant temperature waterbath to another is technically far less demanding than

quickly and accurately changing the temperature of a thermally controlled system.

The most pressing problem to overcome in a waterbath based PCR system is the sealing of the microtitre plate wells, preventing both seepage of water into the reactions as well as cross contamination between wells of a single plate.

4.1.2 Amplification in 96-well plates

Commercially available polycarbonate 96-well microtitre plates were obtained from Techne, which are suitable for thermal cycling. An aluminium clamp was designed and produced in the ICRF workshop (see figure 4-1), so that the wells of the microtitre plate could be sealed with a silicone rubber sheet.

An initial test was carried out in which alternate wells of a microtitre plate were filled with 100 µl water and a dilute solution of bromophenol-blue. The plate was sealed with a silicone rubber gasket, secured in the microtitre plate clamp and submerged in a 94°C waterbath for 90 min. During this period the plate was removed periodically and knocked against the benchtop to facilitate any cross contamination between wells that could occur due to incomplete sealing. At the end of the test the plate was unsealed and the volumes of liquid measured in central and peripheral wells. The volume had reduced by approximately 20% in all wells measured and there was no visible cross contamination of the dye into neighbouring wells (the 20% volume had condensed onto the silicone rubber that was used as gasket).

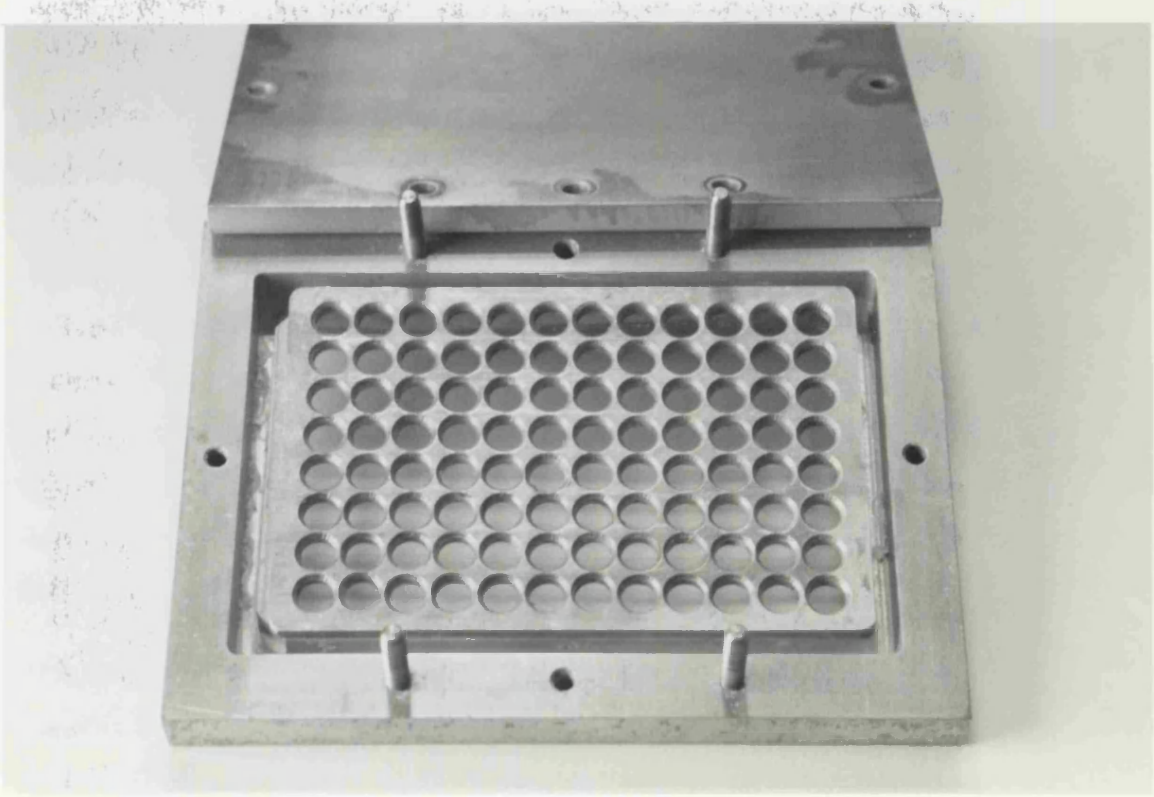


Fig. 4-1 A photograph of an aluminium clamp used to seal the wells of a 96-well microtitre plate. The clamp consists of a top plate that is secured with four wing nuts. A silcone rubber sheet acts as a gasket. Standard 96-well polycarbonate plates fit into the clamp (Techne, cat. #FMW11).

The first attempt at a PCR amplification using the waterbath system was carried out using two cDNA clones that had previously been amplified in a commercial PCR machine. Reactions were set up of both clones in a central and a peripheral well in a total volume of 100 µl containing: 100 pmol of each primer, 50 mM KCl, 10 mM Tris-HCl pH 8.55, 1.5 mM $MgCl_2$, 200 µM dGTP 200 µM dCTP, 200 µM dATP, 200 µM dTTP, 2.5 units TAQ-polymerase. Approximately 5 ng of purified plasmid DNA were used as template in each reaction. Amplification was carried out by cycling the sealed microtitre plate between two waterbaths at 92°C and 72°C for 1 min and 2 min respectively for a total of 30 cycles. Figure 4-2 shows 5 µl of the reactions run on a 1% agarose gel. Amplification had clearly taken place, although there was considerable smearing in the lanes, indicating that non-specific amplification has occurred and there was significant difference in the amplification product between the central and peripheral well for one of the cDNA clones. The central well reaction of the smaller 900 bp cDNA shows a strong band at approximately the same size as the larger 1,500 bp cDNA clones which was amplified in an adjacent well. It seems quite possible that some cross contamination occurred between adjacent wells and that this could account for the larger band in lane 3.

Although amplification had taken place in the previously described test experiment the specificity and yield of the reactions was not satisfactory. In order to take into account the fact that the thermal transfer between the water and the PCR reaction in the microtitre plate wells might be less efficient than in a commercial PCR machine, especially in view of the fact the large aluminium clamp used to seal the plate might act as a significant heat sink, a further test was carried out in which the temperatures of both waterbaths were raised by 2°C to 94°C and 74°C. Previous tests carried out in a commercial PCR machine had indicated that PCR reactions could be successfully inoculated directly from bacterial colonies. In this experiment reactions were inoculated directly from bacterial colonies of mouse adult cDNA clones.

10 reactions were set up as in the previous experiment, except that the reaction volume was 30 µl and all reactions were sealed with mineral oil. Reactions were set up in adjacent wells of one row (B) and 25 cycles were carried out of 1 min incubations in a 94°C and 74°C waterbath respectively.

Figure 4-3 shows 10 µl of the reactions run on a 1% agarose gel. The yield of product is significantly higher than in the first test amplification (compare to figure 4-2), although the specificity of the reaction is still poor. The main problems experienced with waterbath PCR reactions at this stage were that too much non-specific amplification was occurring and that the overall success rate of amplification for random cDNA clones lay at only approximately 30%, depending on the size of the cDNA clones. Figure 4-4 shows the products of a waterbath amplification carried out in the same way as the previously described test, except that the primer concentration was reduced to 10 pmol each. 12 random clones from each of four different size fractions of the mouse adult cDNAs were amplified. Successive size fractions contain cDNAs of decreasing average size and the result seems to indicate that the success rate is correlated with size of the cDNA insert.

The correlation between success rate of amplification and the size of the cDNA inserts suggested that the rate of polymerisation across the cDNA might be limiting. cDNA sequences are thought to form more stable and extensive secondary structure due to their high G+C content and this might account for a slow rate of polymerisation, since lengths of 1 - 2 kb are routinely amplified successfully from genomic DNA. In order to address this possibility, the elongation period (74°C incubation) during the PCR amplification was increased to 4 min. Figure 4-5 shows the product of 12 random cDNAs amplified in a microtitre plate using the following reactions conditions in a total volume of 30 µl: 10 pmol of each primer, 50 mM KCl, 10 mM Tris-HCl pH 8.55, 1.5 mM $MgCl_2$, 200 µM dGTP 200 µM dCTP, 200 µM dATP, 200 µM dTTP, 2.5 units

TAQ-polymerase. The reactions were sealed with mineral oil and 25 cycles of 94°C for 90 sec and 74°C for 4 min carried out.

The specificity and success rate in this experiment were greatly improved, compared to previous tests and represented for the first time a viable amplification system for thousands of cDNA clones by simply performing the waterbath amplification with many microtitre plates in parallel.

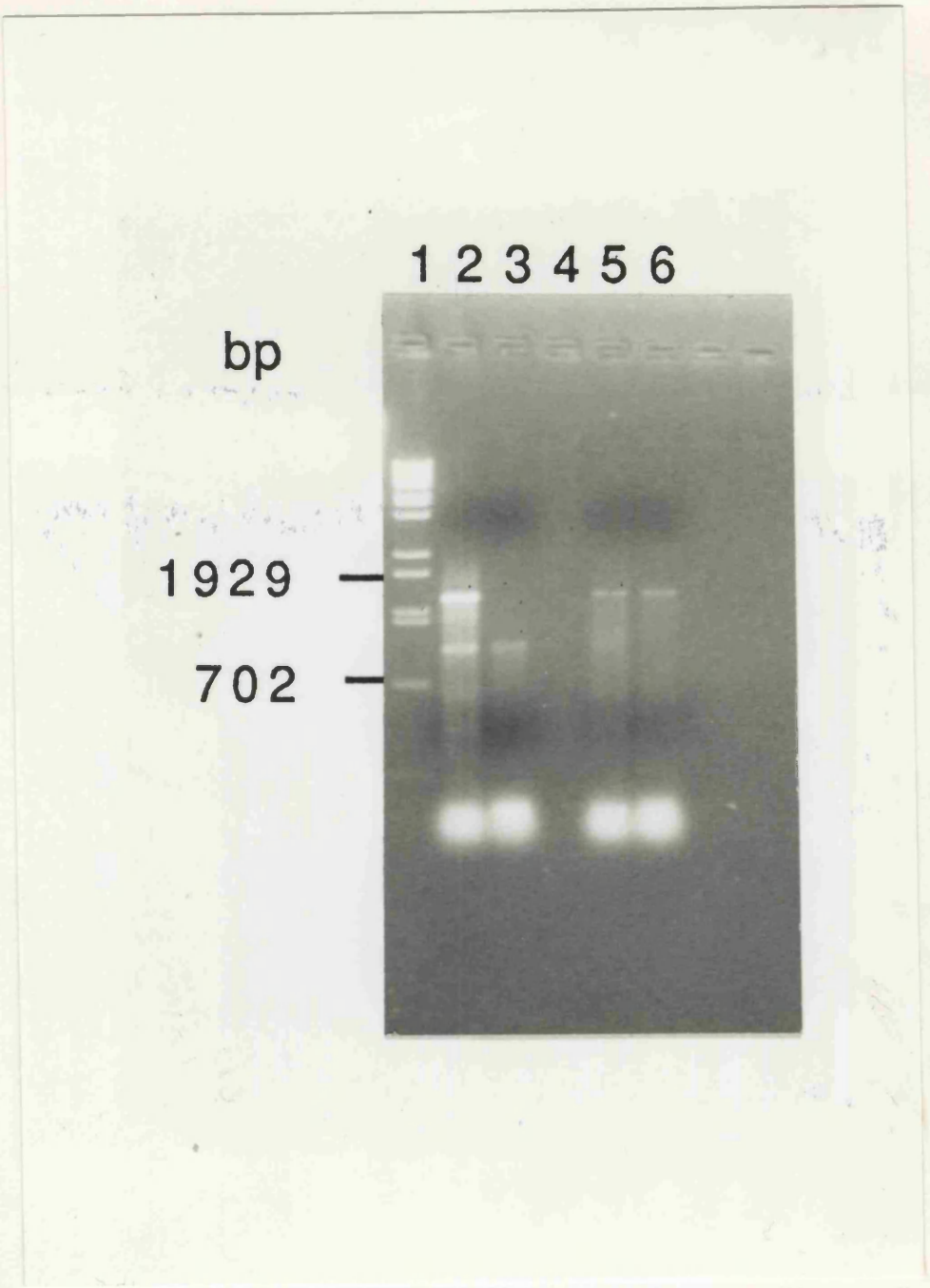


Fig. 4-2 A 1% agarose gel of waterbath PCR amplifications carried out on two cDNA clones in a 96-well polycarbonate plate. Each clone was amplified in a peripheral and central well. Lanes 2 and 5, clone 1, central and peripheral wells respectively. Lanes 3 and 6, clone 2, central and peripheral wells respectively. 10 µl of each 100 µl PCR reaction were loaded in each lane. The size markers used were *BstE*II digested lambda DNA.

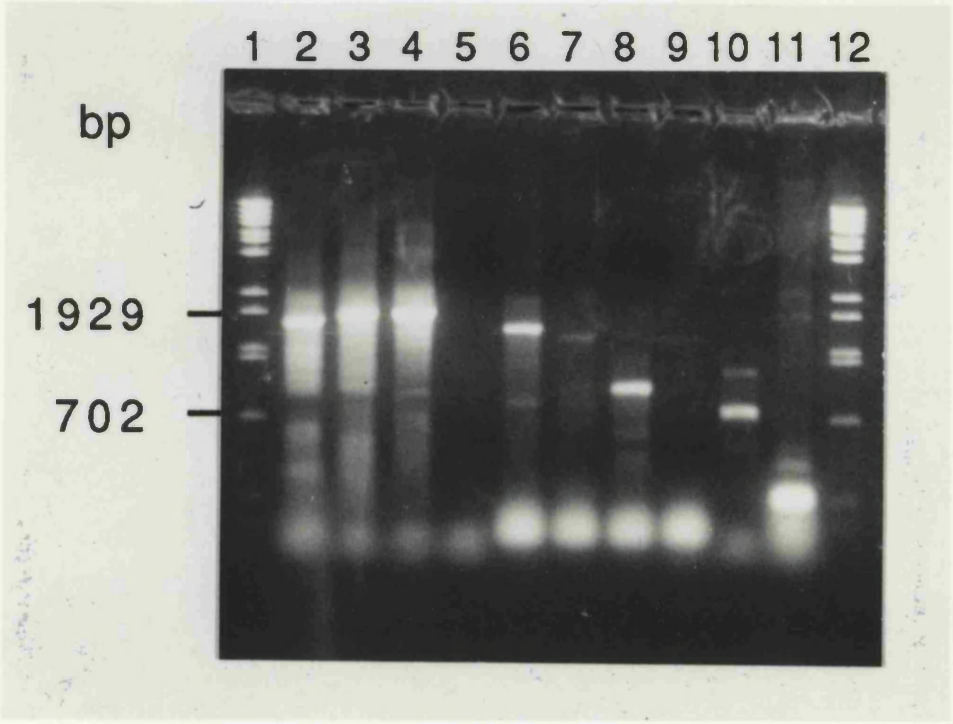


Fig. 4-3 A 1% agarose gel of a waterbath PCR amplifications carried out on mouse adult brain cDNA clones from two different size fractions, in a 96-well polycarbonate plate. In lanes 2 - 6 fraction #8 clones were loaded and in lanes 7 - 11 fraction #12 clones. Each of the reactions was sealed with mineral oil. 10 µl of each 50 µl PCR reaction were loaded in each lane. The size markers used were *BstE*II digested lambda DNA.

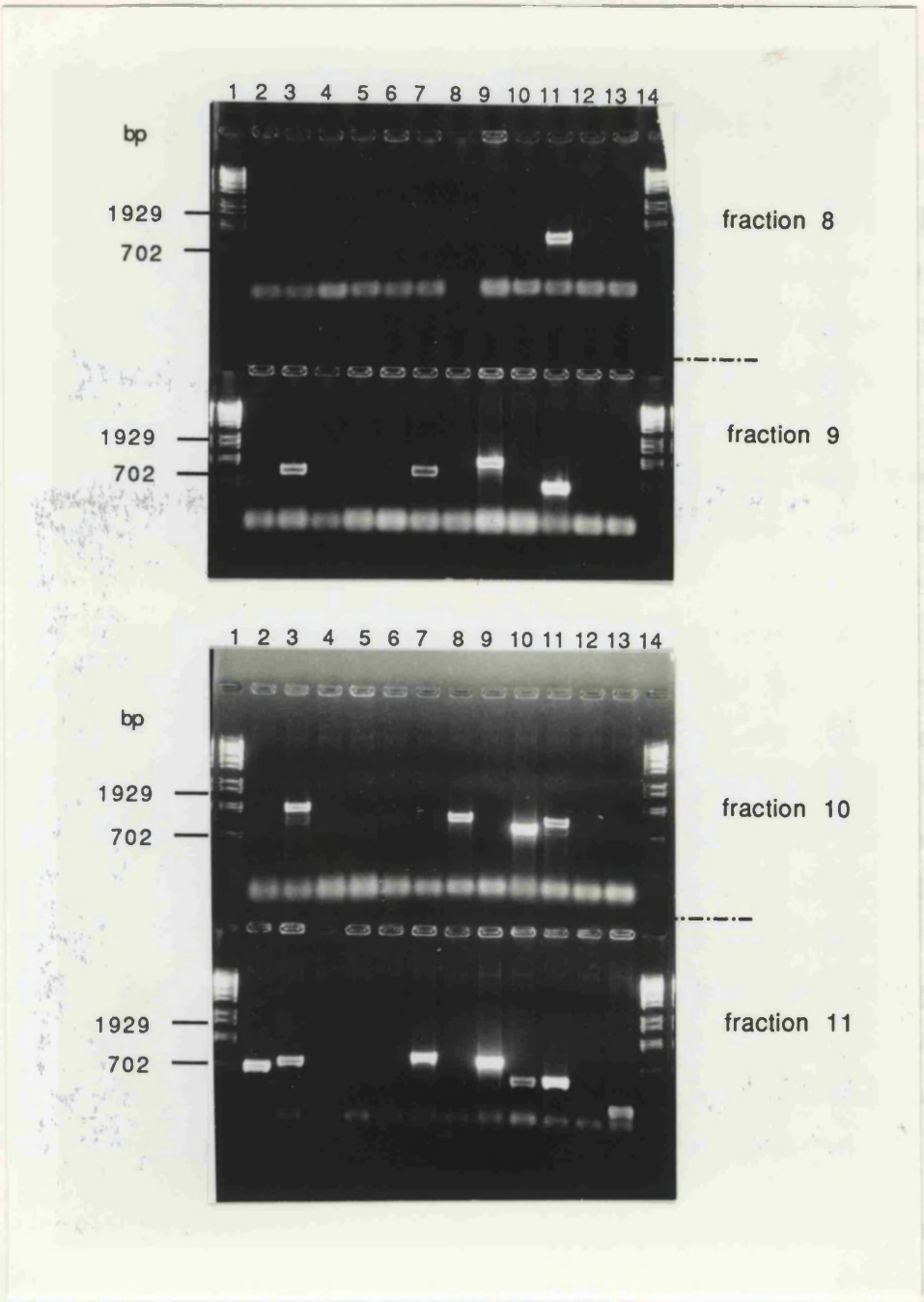


Fig. 4-4 Two 1% agarose gels of waterbath PCR amplifications carried out in a 96-well polycarbonate plate. 25 cycles of 1 min a 94°C and 74°C were carried out. Random mouse adult brain cDNA clones from four different size fractions were amplified. The size markers used were *BstE*II digested lambda DNA.

4.1.3 Amplification in 384-well plates

In order to scale up the waterbath PCR system further, Q-plates (see *cDNA arraying into microtitre plates*) were produced in polypropylene, suitable for cycling at high temperatures. Due to the higher rigidity of the Q-plates, compared to 96-well polycarbonate plates, it was not necessary to use such a hefty clamp as for 96-well plates. In an initial test, reactions were set up as for the previously described 96-well plate experiment (see figure 4-5) in a volume of 40 µl, but without mineral oil since the volume of the well is only 70 µl and hence evaporation is not a significant factor. 12 cDNA previously amplified clones were inoculated simply by stirring a yellow Gilson pipette tip first in the grown microtitre plate well and the in the PCR reaction, without pipetting any volume. The wells were sealed with an adhesive microtitre plate sealing tape (Fass Roll S-695) and the tape secured with a 2 mm aluminium plate, the same size as the Q-plate, held on with two bulldog clamps. 30 cycles of 94°C 1 min and 74°C for 4 min were performed. Figure 4-6 shows 10 µl of each reaction run on a 1% agarose gel. The gel shows that although some amplification has taken place, the yield and specificity is poor and the overall quality of amplification is far less than that achieved with the 96-well plate waterbath system.

Since the Q-plates have a significantly higher thermal mass than the thin walled 96-well polycarbonate plates it is likely that it takes longer for the reactions in the wells to reach the temperature of the waterbath in a Q-plate. The reason that the Q-plates are of higher thermal mass is that in order to be able to form reliably wells of 70 µl volume it is necessary to make the plate by injection moulding rather than a thermo-pressing system that is used for the thin walled 96-well polycarbonate plates. In order to determine the time taken for the temperature inside a Q-plate well to reach that of the waterbath, a thermocouple was inserted through a small hole in the bottom of the plate and sealed with high temperature wax. Figure 4-7 shows a graph of the temperature changes measured for

various transitions in both 74°C and 94°C waterbaths. The data clearly show that in the previous waterbath experiment the 1 min incubation time at 94°C was insufficient to denature completely the double stranded DNA in the Q-plate wells. The data indicate that a 3 min denaturation step in a 94°C waterbath is required to ensure that the reaction has reached the desired temperature of the waterbath. The data also show that 3 min are required for the reactions to cool to the temperature of the 74°C waterbath and suggest therefore, that the normal time for the polymerisation step should be added on to the 3 min.

An alternative sealing process for the Q-plates was developed so that the aluminium plate is no longer required, thus reducing the thermal mass of the plates during waterbath amplification. After many trials with a variety of different approaches, a heat sealing mechanism akin to that used for sealing yoghurt cartons, was developed. A bilaminar plastic film consisting of 0.15 mm polypropylene and 0.30 mm polyester is used with a hot plate (150°C) sealing system in which the polypropylene side is melted onto the surface of the polypropylene Q-plate (melting temperature ~ 130°C) (see figure 4-8). The seal that is formed is totally water tight and permanent. In order to remove the seal the plate is put into the heat sealer, the surface heated to melt the polypropylene, and the seal then quickly torn off.

Figure 4-9 shows a 1% agarose gel of a cDNA clone amplified in a heat sealed Q-plate where 30 µl reactions were cycled 30 times between 94°C for 3 min and 74°C for 6 min. In this experiment home-made TAQ-polymerase was used and the number of units in each reaction were excessive, resulting in some non-specific amplification that shows up as heavy smearing in the gel lanes. The same cDNA clone was amplified in 15 wells evenly distributed across the plate and reasonably even yield was obtained in all wells. Note, the seal had come off at well A12 and that this accounts for the lack of amplification in this well.

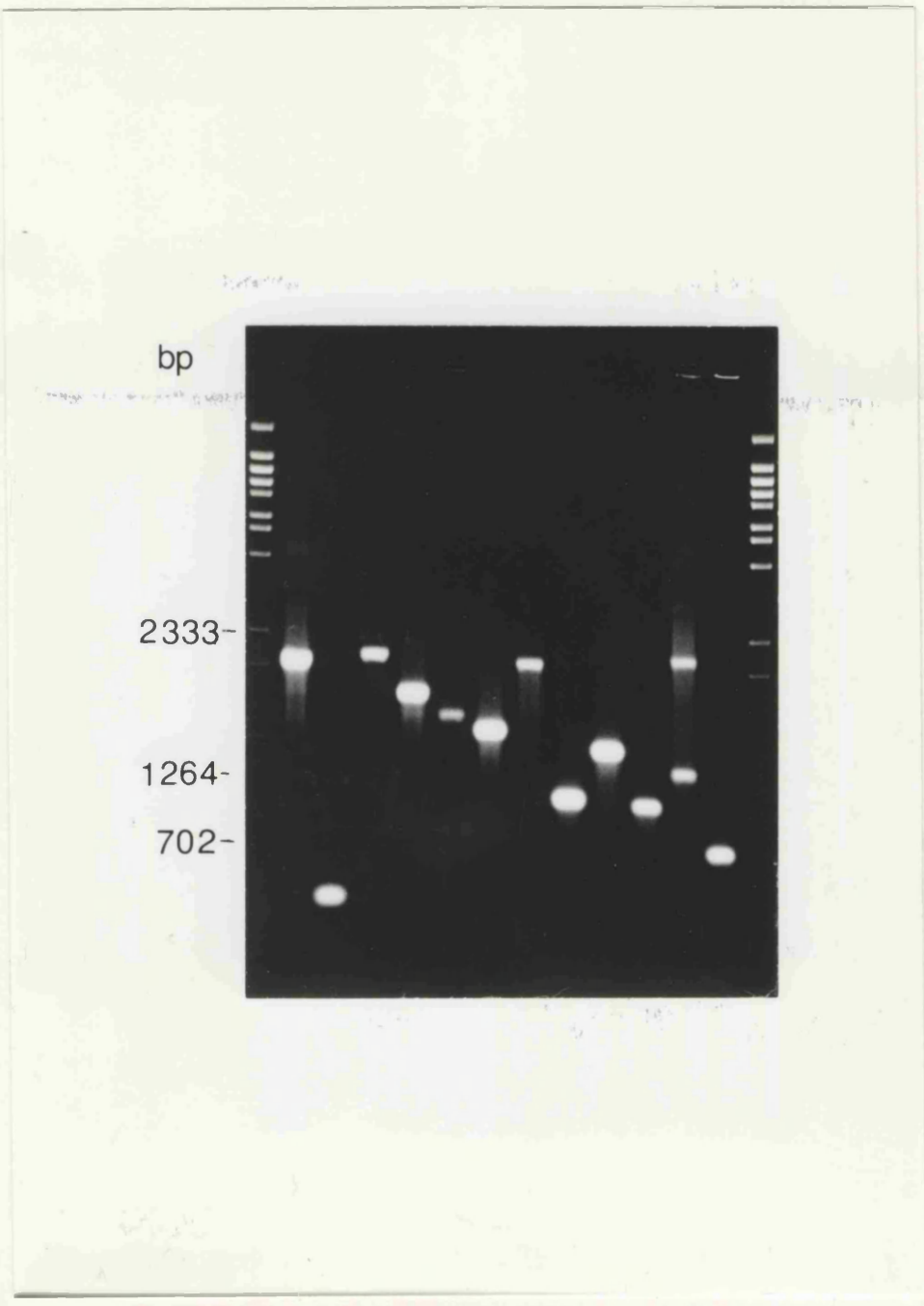


Fig. 4-5 A 1% agarose gel of a waterbath PCR amplification carried out on 12 random cDNA clones in a 96-well polycarbonate plate. 25 cycles of 1 min a 94°C and 74°C were carried out. Each 30 µl reaction was sealed with a drop of mineral oil. 10 µl of each reaction were loaded per lane. The size markers used were *BstE*II digested lambda DNA.

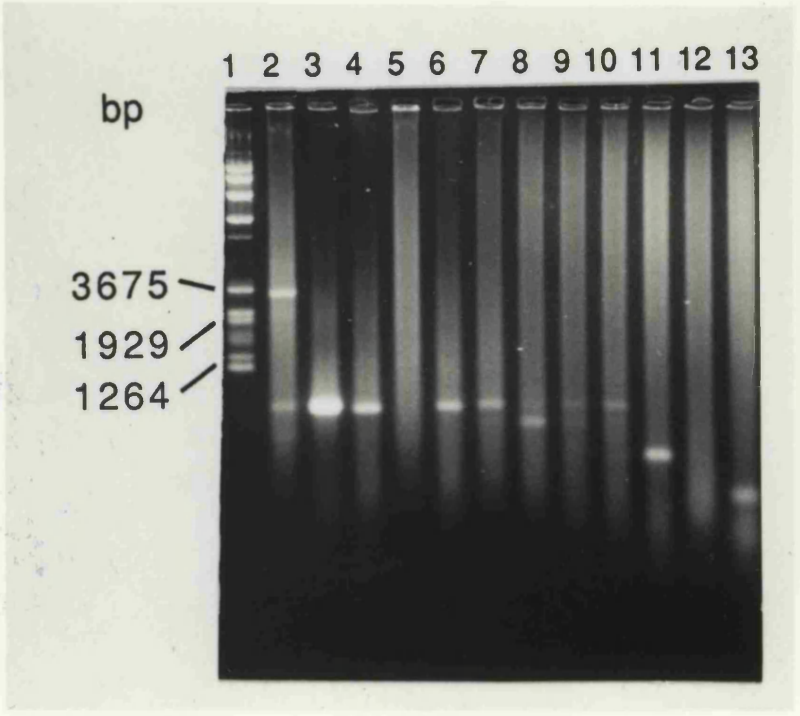


Fig. 4-6 A 1% agarose gel of a waterbath PCR amplification carried out on 12 cDNA clones in a Q-plate. 25 cycles of 1 min 94°C and 4 min 74°C were performed. 5 µl of each 30 µl reaction were loaded and the size marker used was *BstE*II digested lambda DNA.

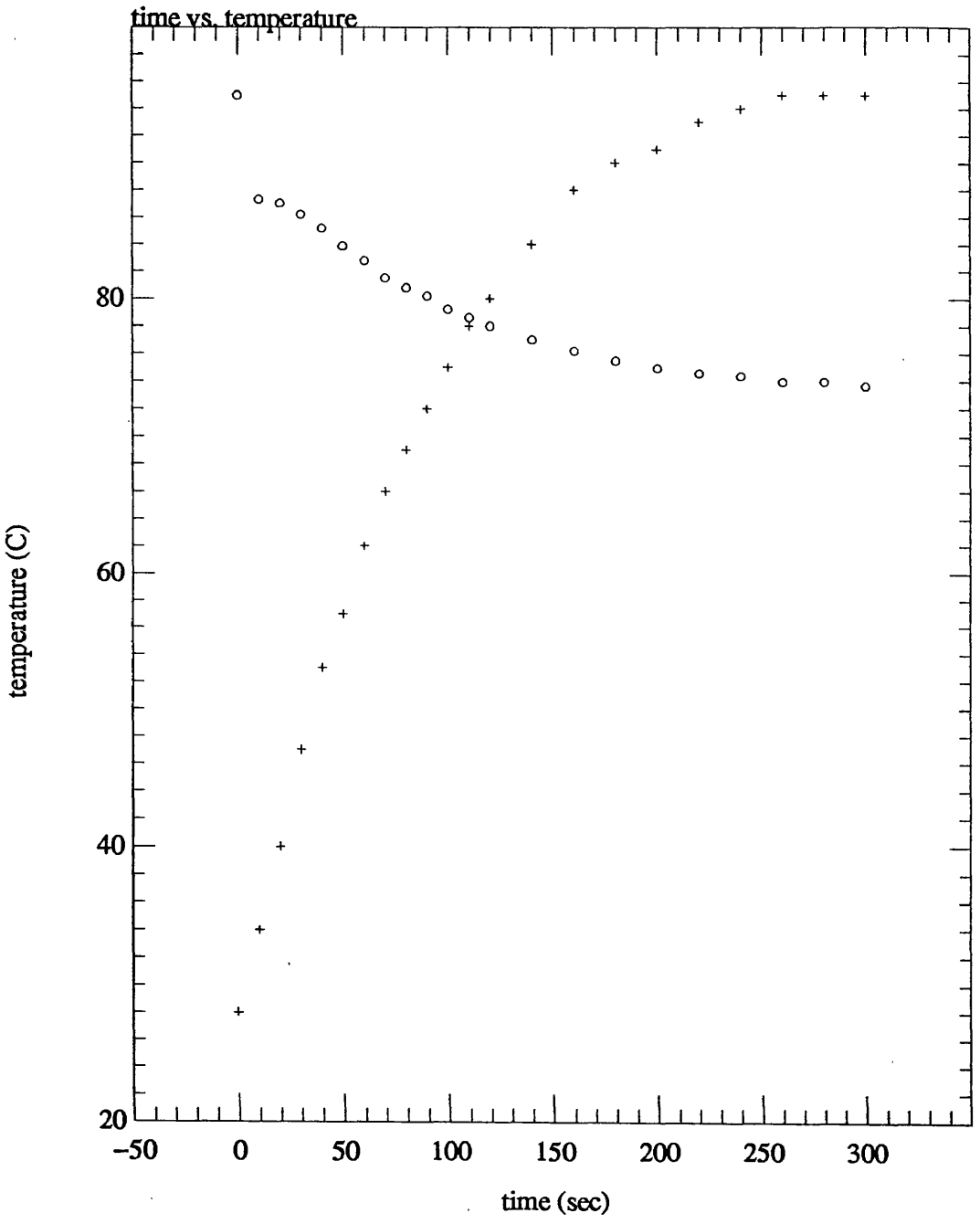


Fig. 4-7 A plot of time versus temperature for readings taken with a thermocouple inserted into a well of a sealed Q-plate. Transitions from room temperature to 94°C (crosses) and from 94°C to 73°C (open circles) are shown.

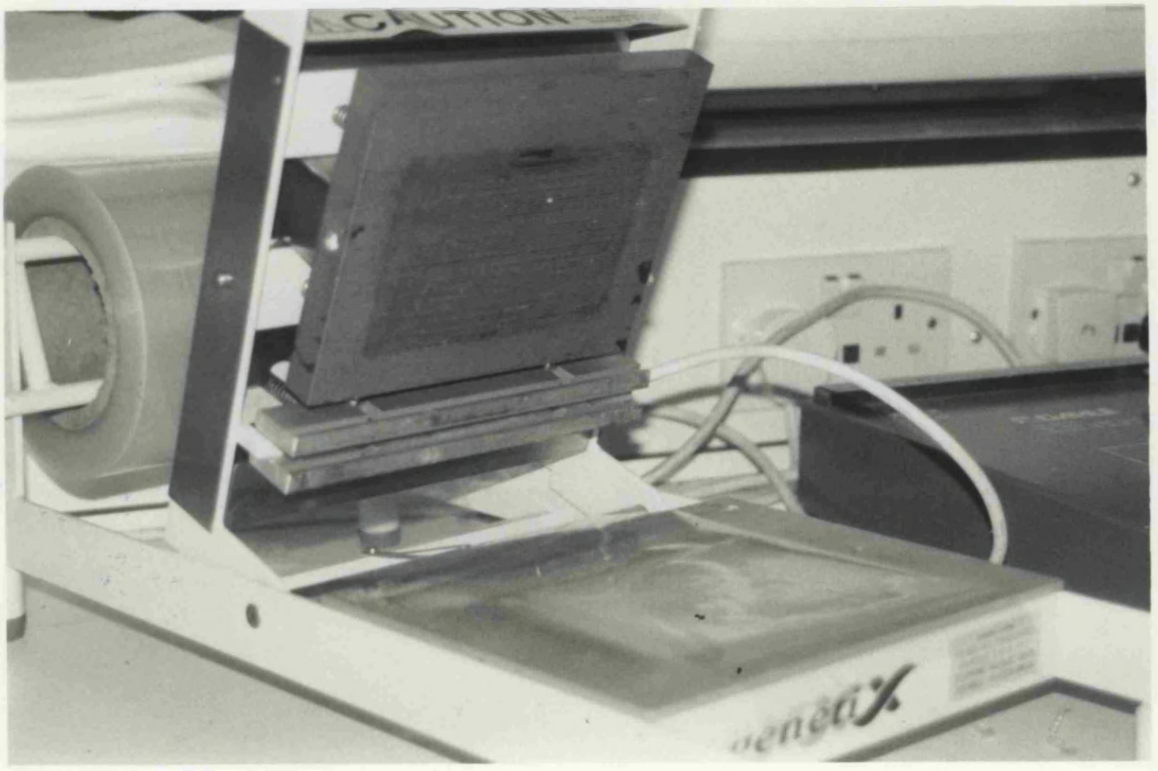


Fig. 4-8 A photograph of the Q-plate heat sealer used to melt a thin film of polypropylene across the wells of a polypropylene Q-plate.

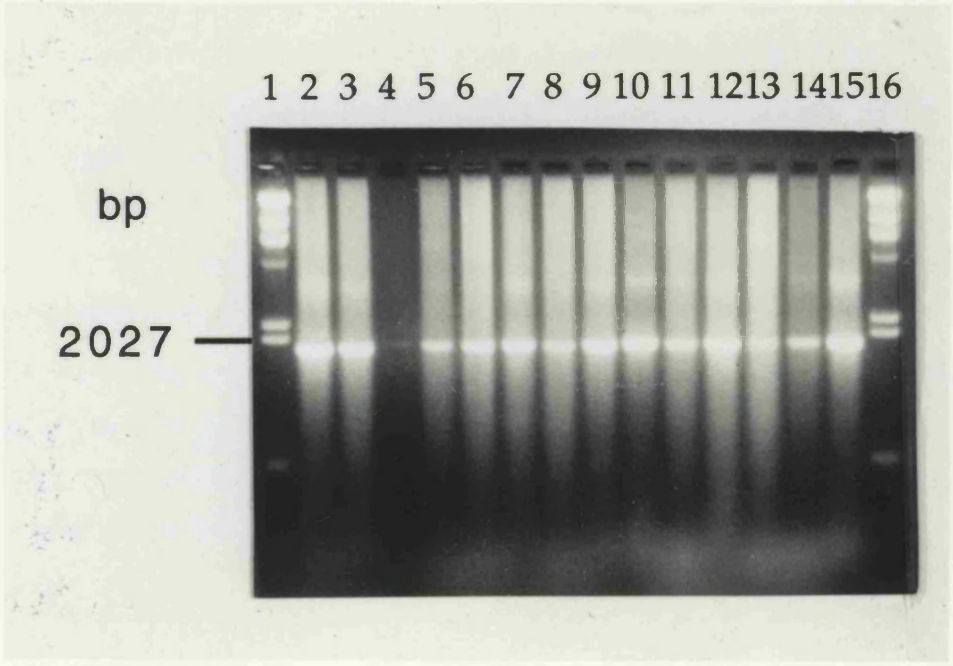


Fig. 4-9 A 1% agarose gel of waterbath PCR amplifications carried out in a heat sealed Q-plate. 30 cycles of 3 min 94°C and 6 min 74°C were carried out. The same cDNA clones were amplified in 14 wells. Lanes: 2) well A1; 3) well A12; 4) well A24; 5) well E6; 6) well E18; 7) well H1; 8) well H12; 9) well H24; 10) well L1; 11) well L12; 12) well L24; 13) well O1; 14) well O12; 15) well O24. 5 µl of each reaction loaded per well. The size marker used was *Hind*III digested lambda DNA.

4.1.4 Development of a phosphate buffer for PCR reactions

One of the future aims in further developments lies in the immobilisation of PCR products on a solid support other than nylon membranes. Work carried out together with Elmar Maier and Igor Ivanov in the lab (Igor Ivanov was a visitor in the lab for 6 months from the Engelhardt Institute, Moscow) addressed the possibility of binding PCR products covalently to derivitised polyacrylamide that is polymerised onto glass plates. One of the standard coupling reagents used extensively for the covalent modification of proteins is a succinimidyl acrylate (Pollak et al., 1980) which can be copolymerised with acrylamide and bisacrylamide to form a succinimidyl ester group which is susceptible to nucleophilic attack. Primary amine groups are suitable for nucleophilic substitutions of the succinimidyl moiety forming a stable amide link. Primary amine groups attached to proteins or in this case incorporated into DNA molecules can be used to form covalent linkages. The details of the experiments will be described by Elmar Maier and Igor Ivanov elsewhere and not discussed herein.

For the method of attachment described above to be successful, primary amine groups have to be incorporated into the PCR products, which is achieved either by the incorporation of a modified nucleotide such as amino-7 dUTP during the amplification, or by use PCR primers that have a primary amine group attached at their 5' ends. Since the standard PCR buffers contain 10 mM Tris and the primary amino groups of the Tris would compete with the desired reaction, an alternative buffering system was sought.

In the paper by Chien et al. (1976) the polymerase activity of *Thermus aquaticus* DNA polymerase (TAQ) was reported for various buffer systems including Tris and potassium phosphate. It showed clearly that TAQ activity was significantly lower in 25 mM potassium phosphate pH

7.4 than in 25 mM Tris-HCl pH 7.8. However, these assays were also carried out in 5 mM magnesium chloride, which provided a clue to the surprising difference in activities observed between the two buffering systems: potassium phosphate is not soluble in aqueous solution at concentrations greater than 10 mM and thus will have precipitated in the assay conditions used in the paper.

Figure 4-10 shows the products of PCR reactions carried out in varying concentrations of potassium phosphate buffer run out on a 1% agarose gel. The buffer was prepared by mixing $K_2HPO_4^=$ and $KH_2PO_4^=$ in the required ratio. The result of this experiment confirms the low activity of TAQ polymerase in 25 mM potassium phosphate, but also demonstrates high activity in potassium phosphate buffer up to a concentration of approximately 5 mM. In a further test PCR amplification of 12 random cDNA clones was compared for a standard PCR buffer (50 mM KCl, 10 mM Tris-HCl pH 8.3, 1.5 mM $MgCl_2$ and 0.01% gelatine) and a 4 mM $HPO_4^=$ buffer (50 mM KCl, 4 mM potassium phosphate pH 8.5, 1.5 mM $MgCl_2$ and 0.01% gelatine). All amplifications (50 µl) were carried out in identical conditions of 93°C for 1 min and 72°C for 3 min in a commercial PCR machine, except for the substitution of the buffer. Figure 4-11 shows 5 µl of the PCR reactions run on a 1% agarose gel. The gel shows that amplifications of all clones has worked to a similar degree in both buffer systems. It is noteworthy however, that the size of the PCR products from identical templates differ significantly in at least two cases (lanes 4 and 6). For both clones the product is longer in the 4 mM $HPO_4^=$ buffer and in the case of lane 6 the difference is greater than 1,500 bp. The cause for these differences remains unsolved although one could speculate that in these two cases secondary structure, affecting amplification, varies between the buffering systems. There was no evidence that two cDNA clones were contained in the reactions. This phenomenon was not investigated any further but would never the less be interesting to pursue.

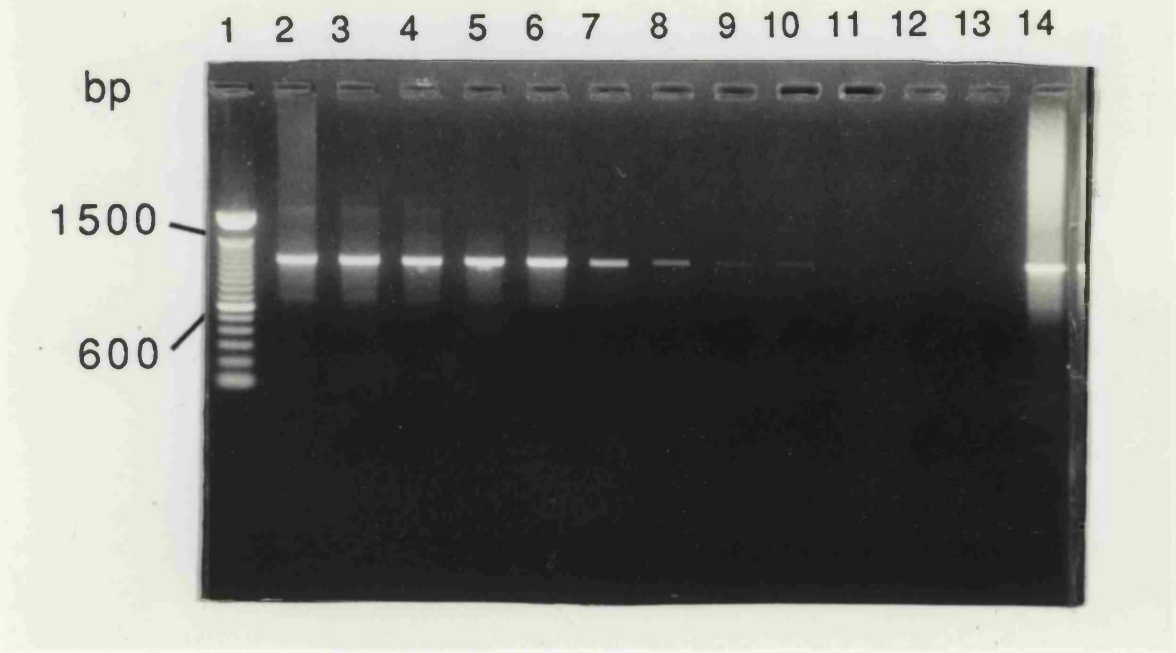


Fig. 4-10 A 1% agarose gel of PCR amplifications carried out in a Cetus 9600 machine, in varying concentrations of potassium phosphate buffer pH 8.3. 30 cycles of 1 min 93°C and 3 min 72°C were performed. Lanes: 1) 100 bp ladder; 2) 1 mM $HPO_4^{=}$; 3) 2 mM $HPO_4^{=}$; 4) 3 mM $HPO_4^{=}$; 5) 4 mM $HPO_4^{=}$; 6) 5 mM $HPO_4^{=}$; 7) 6 mM $HPO_4^{=}$; 8) 7 mM $HPO_4^{=}$; 9) 8 mM $HPO_4^{=}$; 10) 9 mM $HPO_4^{=}$; 11) 10 mM $HPO_4^{=}$; 12) 25 mM $HPO_4^{=}$; 13) 50 mM $HPO_4^{=}$; 14) 10 mM Tris pH 8.55.

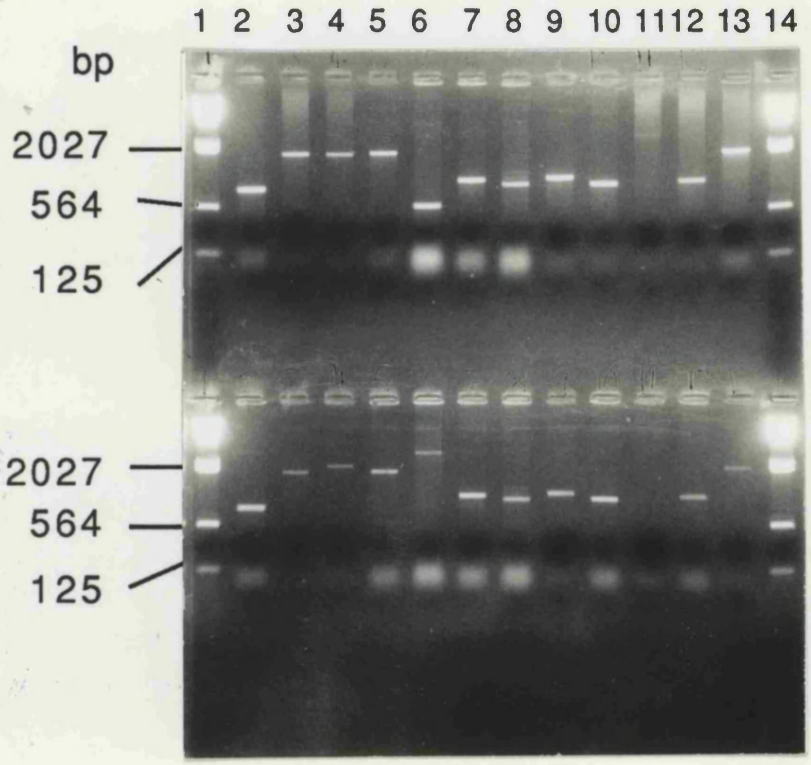


Fig. 4-11 A 1% agarose gel of duplicate PCR reactions carried out on 12 cDNA clones in 10 mM Tris HCl pH 8.55 based buffer (top) and 4 mM potassium phosphate pH 8.5 based reaction buffers (bottom). The size marker used was *Hind*III digested lambda DNA.

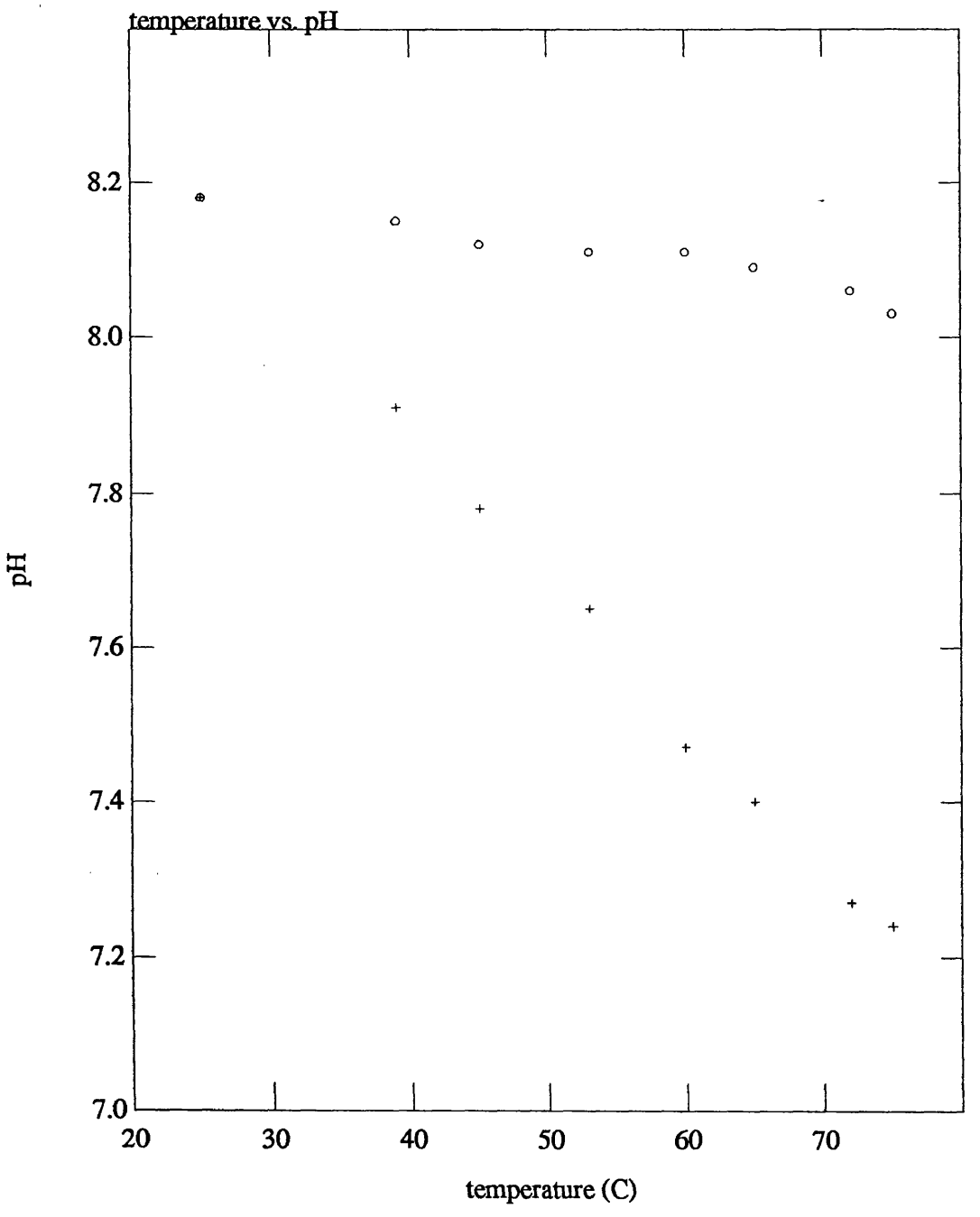


Fig. 4-12 A plot of temperature versus pH for 10 mM Tris HCl pH 8.55 (25°C) (crosses) and 4 mM $HPO_4^{=}$ pH 7.4 (open circles).

Fig. 4-13 A 1% agarose gel of PCR amplifications carried out with decreasing amounts of TAQ polymerase in both Tris HCl pH 8.55 (25°C) (top) based buffer and 4 mM $HPO_4^{=}$ pH 8.5 buffer (bottom). Lanes: 1) 100 bp ladder; 2) 1 µl TAQ; 3) 0.5 µl TAQ; 4) 0.2 µl TAQ; 5) 0.1 µl TAQ; 6) 0.05 µl TAQ; 7) 0.02 µl TAQ; 8) 0.01 µl TAQ; 9) 100 bp ladder.

Considerations at to the differences in the two buffers that might give rise to the varying polymerisation product size, led the realisation that the pH of the $HPO_4^=$ buffer was inappropriate, since the pH of the Tris-HCl buffer varies with temperature whereas that of the $HPO_4^=$ buffer does not, to the same extent. To determine the pH to which the $HPO_4^=$ should be adjusted, the temperature dependence of both the Tris-HCl and the $HPO_4^=$ based PCR buffer were measured and the result is shown in figure 4-12. $HPO_4^=$ buffer was prepared at a pH = 7.4, and the yield of PCR product in this buffer compared to the standard Tris-HCl based buffer determined for a series of TAQ polymerase concentrations. Figure 4-13 shows the PCR products from a known cDNA clone amplified in the two buffers, using decreasing amounts of TAQ polymerase, run on a 1% agarose gel. The result indicates that amplification is at least as efficient in 4 mM $HPO_4^=$ pH 7.4 buffer as in a standard Tris-HCl based buffer.

Observations over many PCR amplification experiments using the waterbath system indicate that the $HPO_4^=$ buffer consistently performs slightly better than a Tris-HCl based buffer in terms of specificity and yield. One hypothesis, that remains untested, is that the TAQ polymerase is marginally more stable over many cycles in a buffer of constant pH. This might be especially noticeable in a waterbath PCR system where slow temperature transitions mean that the TAQ polymerase is subjected for longer periods to temperatures in excess of 90°C.

4.1.5 Isolation of TAQ DNA polymerase expressed in *E.coli*

The large scale amplification of cDNA clones using a PCR system is very expensive when using commercial TAQ polymerase. Approximately 2 units of TAQ polymerase are required for each reaction, adding up to 200,000 units for 100,000 clone cDNA library. The price of commercial

TAQ polymerase is approximately £40 per 250 units, making the cost of amplifying one cDNA library of 100,000 clones in the order of £32,000. Given the fact that large scale PCR amplifications fail to work periodically, for reasons that are not clear, the expense of the TAQ polymerase alone are prohibitive.

A clone was obtained that was prepared exactly the way described in Engelke et al. (1990) and was used to isolate TAQ polymerase over expressed in *E.coli*. The protocol published by Engelke et al. was modified by L. Schalkwyk and myself and is described in detail in *Materials and Methods*. Initially, several single colonies were used to inoculate separate 50 ml cultures. Each culture was induced with IPTG and a 1 ml aliquot assayed for polymerase activity as described in *Materials and Methods*. The culture with the highest activity was centrifuged, the pellet resuspended in 1 ml 2x YT medium + Hogness modified freezing mix and stored at -70°C. This glycerol stock was then used directly to inoculate subsequent 50 ml cultures that in turn were used as inoculum for 1 l preparative cultures. Several attempts, by several people, to isolate TAQ polymerase from re-streaked single colonies were unsuccessful. It was shown that the full length insert is frequently deleted upon culturing and this presumably acccounts for these failures (data not shown). To circumvent this problem, cultures were subsequently not inoculated from single colonies.

The typical yield of TAQ polymerase was approximately 500,000 units per litre of culture. Yields did fluctuate from different stocks of the same clone, presumably due to variable extents of deletion in the early stages of growth. Figure 4-14 shows an autoradiograph of assays carried out on aliquots of samples during several stages of the isolation protocol. The bulk of the polymerase activity was consistently contained in fractions 3, 4 and 5 during the elution of the Biorex column. These three fractions were pooled prior to dyalysis against the storage buffer (buffer D) and then stored at -20°C. Over a period of 20 months no detectable loss of activity occured in any of the preparations. Figure 4-

15 shows a dilution series of four isolates of TAQ polymerase used in a PCR amplification of a known cDNA clones. The isolates perform equally well as commercial TAQ polymerases and were used in all large scale PCR amplification experiments in this project.

The isolation of TAQ polymerase from an overexpressing *E.coli* clone represents a major saving in terms of costs and makes the large scale PCR amplification of entire cDNA libraries financially feasible.

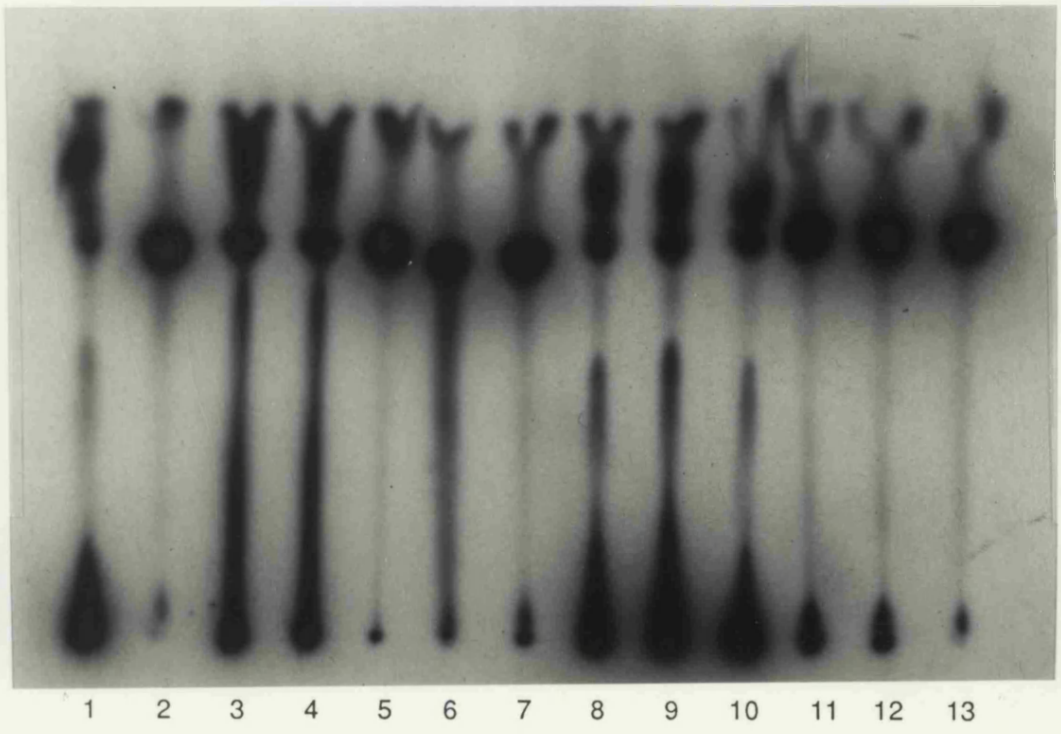


Fig. 4-14 An autoradiograph of TAQ polymerase assays run on a thin layer chromatography plate. 1 µl of each assay was run of a PEI thin layer plate in 0.75 M KH_2PO_4 pH 3.5: 1) 5 U commercial TAQ polymerase (Cetus). 2) Negative control with 1 µl water. 3) 1 µl of crude cell lysate of culture1. 4) 1 µl of crude cell lysate of culture 2. 5) 1 µl of column flow through during loading of lysates. 6) 1 µl of fraction 1 during elution. 7) 1 µl of fraction 2 during elution. 8) 1 µl of fraction 3 during elution. 9) 1 µl of fraction 4 during elution. 10) 1 µl of fraction 5 during elution. 11) 1 µl of fraction 6 during elution. 12) 1 µl of fraction 7 during elution. 13) 1 µl of fraction 8 during elution.

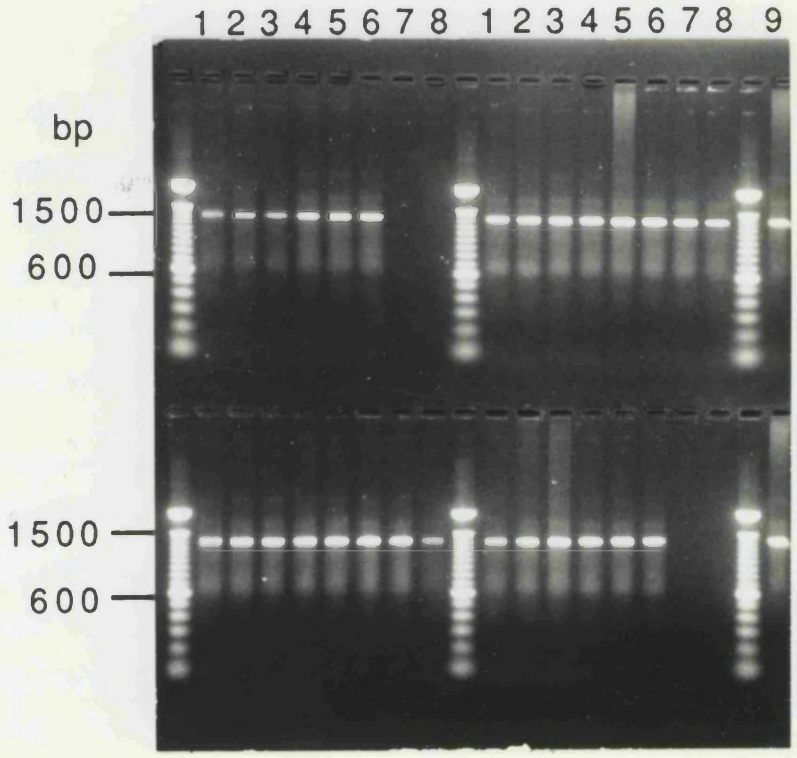


Fig. 4-15 A 1% agarose gel of PCR amplifications carried out on a cDNA clone with a dilution series of four TAQ polymerase isolates. Each reaction was inoculated with ~1 ng cDNA PCR product from a previous amplification. From left to right (lanes 1-8, top and bottom) the amounts of TAQ preperations used per reaction were: 2 µl, 1 µl, 0.5 µl, 0.2 µl, 0.1 µl, 0.05 µl, 0.02 µl and 0.01 µl. Lane 9 shows a control with 2.5 U commercial TAQ polymerase (Cetus). 5 µl of each 30 µl reaction were loaded on the gel. Amplifications were carried out in a Cetus 9600 machine, cycling 30 times between 94°C for 1 min and 73°C for 3 min. The size markers used in this gel were a 100 bp ladder.

4.1.6 Scaling up to 10,000 reactions per experiment

In order to be able to scale the waterbath PCR system up to tens of thousands of clones in parallel, a large waterbath 'PCR robot' was designed and built, allowing up to 120 Q-plates (46,080 reactions) to be cycled simultaneously (Meier Ewert et al., 1993; Maier et al., 1994a). Figure 4-16 shows the PCR robot with two 250 litre waterbaths each temperature controlled by four large heating elements. Two pneumatic slides, controlled by a PC, transfer an aluminium 'basket' that carries the Q-plates between the two waterbaths.

An experiment was set up in which 9,600 clones were amplified in 25 Q-plates using the PCR robot. A bulk PCR mixture for 10,000 reactions was set up (100 µl each primer (1 mM), 30 ml $HPO_4^=$ buffer (10x), 3 ml $MgCl_2$ (150 mM), 1.6 ml dNTPs (10 mM each), 1 ml TAQ polymerase (self-made 20 U μl^{-1}), 264 ml H_2O) and 30 µl dispensed into Q-plate wells using a multichannel pipette. Reactions were inoculated using a standard microtitre plate 96-tip pipette (Handispense) by stirring the tips first in a Q-plate of grown cDNA clones and then in the PCR reactions. The plates were sealed using the heat sealing device described above (figure 4-8) and stored at 4°C while the remaining plates were prepared. 30 cycles of 96°C for 3 min and 73°C for 5 min were performed and 10 µl of a subset of samples run out on a 1% agarose gel. Figure 4-17 shows the PCR products from 2 different plates, in which 75% of the samples show products visible by ethidium bromide staining and 5% show multiple products.

To test the viability of performing more demanding PCR reactions with complex sequence templates, amplification was carried out with ALE1 and ALE3 primers (Cole et al., 1991) on total human DNA, a human / hamster cell line and a human YAC clone (see figure 4-18). The result indicates that complex PCR reactions can also be performed using the PCR robot, an application used in other projects in the lab.

In subsequent experiments a specially designed 384-pin replicating gadget was used to inoculate the PCR reactions. So far, 100,000 cDNA clones have been amplified using this system, of which 32,256 have been used in a large scale fingerprinting experiment.

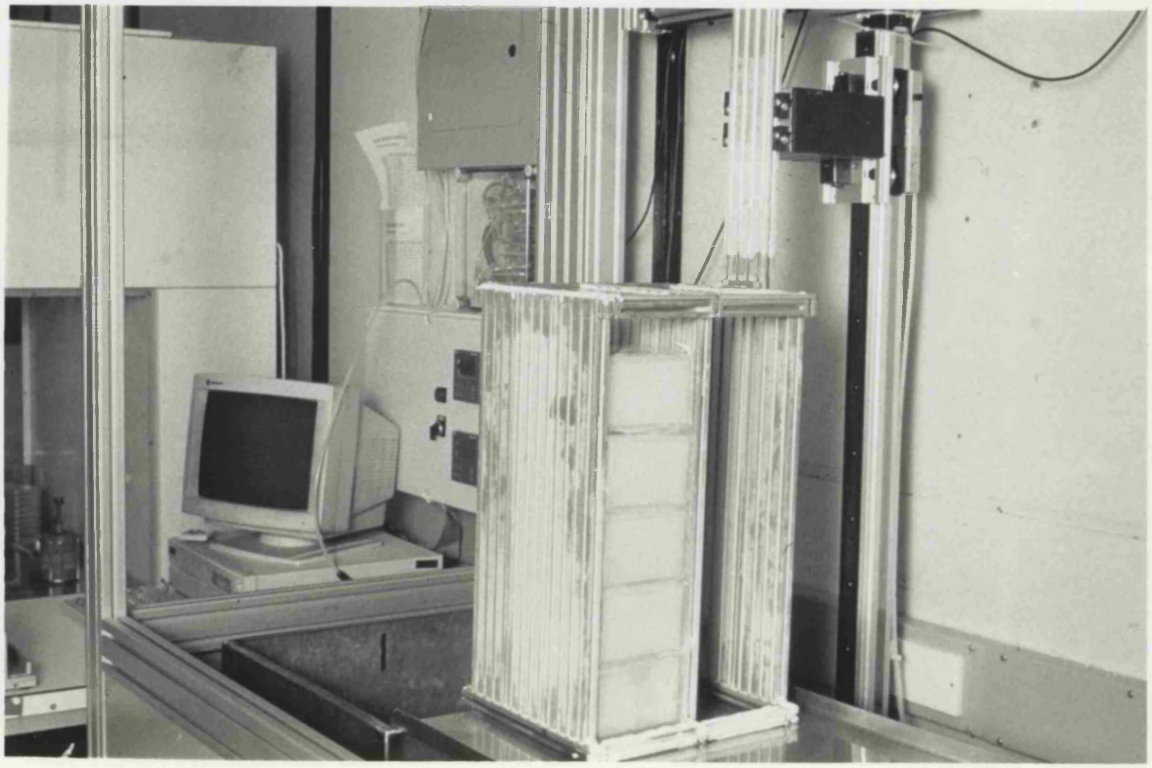


Fig. 4-16 A photograph of the PCR robot showing the plate holding basket loaded with sealed Q-plates. The basket is moved between waterbaths with two pneumatic slides (horizontal and vertical) and is controlled by PC driven software.

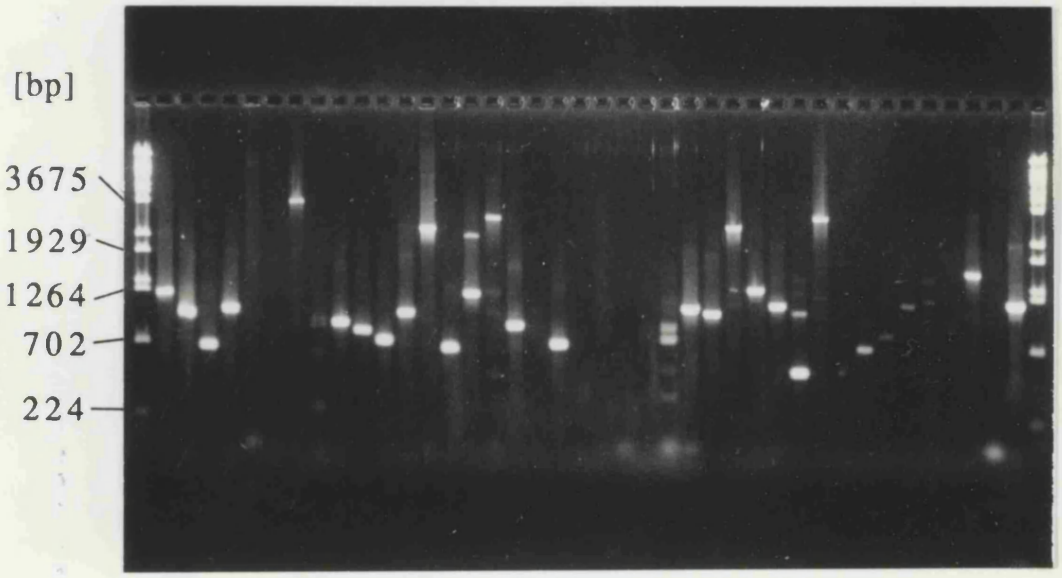


Fig. 4-17 A 1% agarose gel of PCR amplifications carried out in the waterbath PCR robot. The gel shows reactions from 2 different plates from an experiment in which 25 plates were cycled. 5 µl of each 30 µl reaction were loaded in each of the lanes. The size markers used in these gels were *BstE*II digested lambda DNA

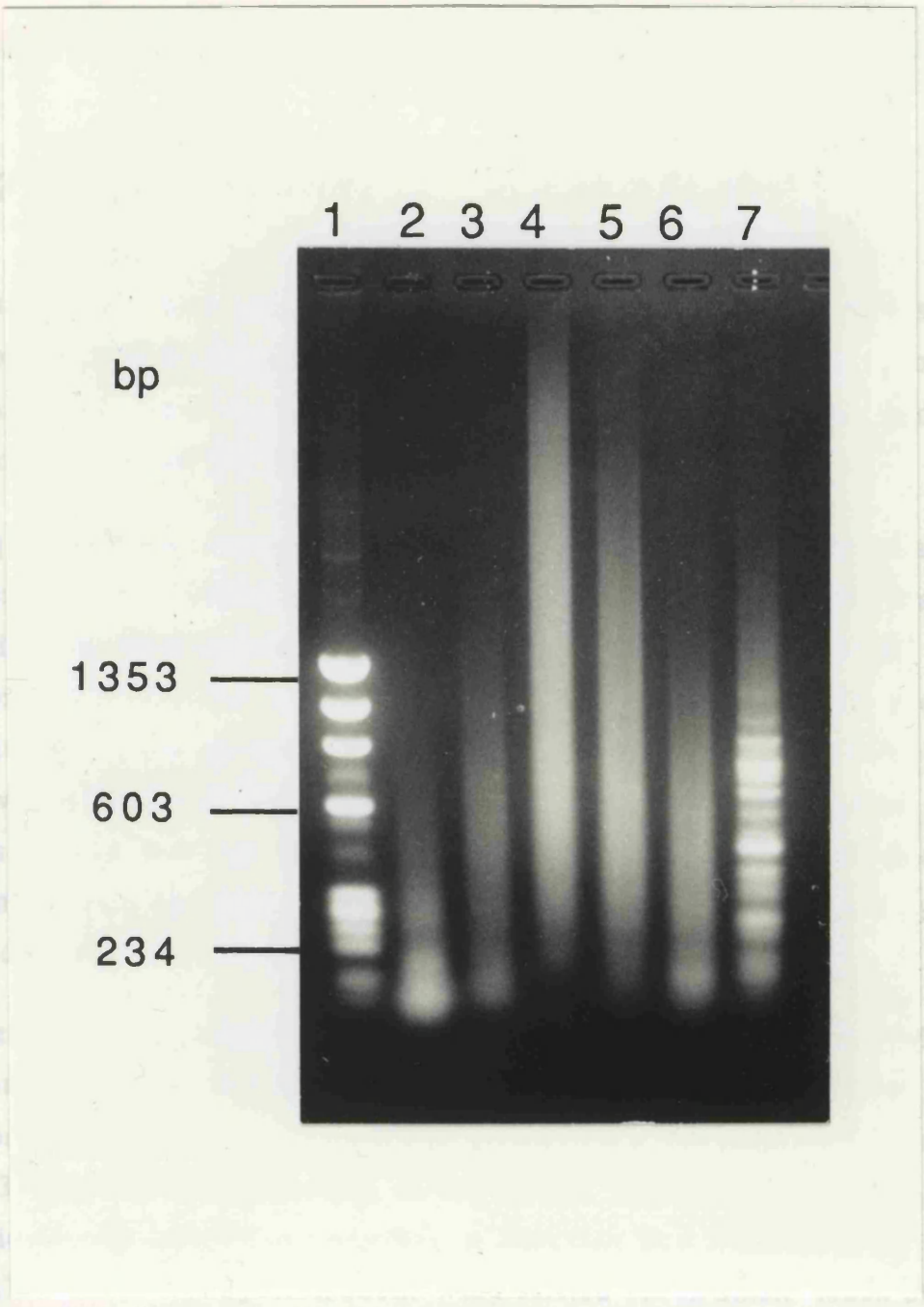


Fig. 4-18 A 1% agarose gel of waterbath PCR amplifications carried out in the PCR robot. Human Alu-specific primers (ALE1 and ALE3) were used to amplify from various templates. Lanes: 1) *Hae*III digested ØX 174 DNA. 2) Negative control with water. 3) Negative control with 20 ng yeast DNA (*S. cerevisae*). 4) 20 ng total human DNA. 5) 20 ng human Xp radiation hybrid DNA on hamster background. 6) Negative control with 20 ng hamster DNA. 7) 20 ng of a human YAC clone.

4.2 Oligonucleotide test hybridisations

A series of small scale test experiments were carried out to test the hybridisation characteristics of octanucleotides, specifically to obtain quantitative data on the detection limit of hybridisations and the discrimination of full / mis-matched hybridisation signals.

Sequenced M13 clones were kindly provided by Stephan Beck, that had been generated during a sequencing project of the human MHC class II region on chromosome 6 (Beck et al., 1992). These clones were approximately 1,000 bp long and there were a total of 1348 used as controls in the oligo hybridisation experiments (there were 1,840 clones of which 1348 had been sequenced at least partially). The human DNA was cloned into the bacteriophage vector mp18, for which specific primers were designed to amplify the inserts using the waterbath system described above.

In the first test experiment four M13 clones were amplified by PCR, spotted manually in a dilution series onto Hybond N+ (Amersham) and hybridised with two octamers that contained a complete match in two M13 clones and mis-matches in the others. The oligonucleotides were radioactively labelled as described in *Materials and Methods* using [^{32}P-γ]ATP. These test hybridisations were carried out in small 'lunch boxes' (18 cm x 12 cm), in a volume of 20 ml, at a 3 nM oligonucleotide concentration using the conditions described in *Materials and Methods* (hybridisation conditions were adapted from (Drmanac et al., 1990b)). Figure 4-19a shows an autoradiograph of a hybridisation in which the clones Bax7.B4 and Bax7.C9 contain a full match, Bax7.D10 a single C-C terminal mis-match and Bax7.F4 no match longer than 4 bp. In this hybridisation the detection limit was around 20 fmol (20 ng of 1,500 bp DNA) and the signal ratio for full match hybridisation to end mis-match signal, a useful measure of discrimination was around 50:1. In a second

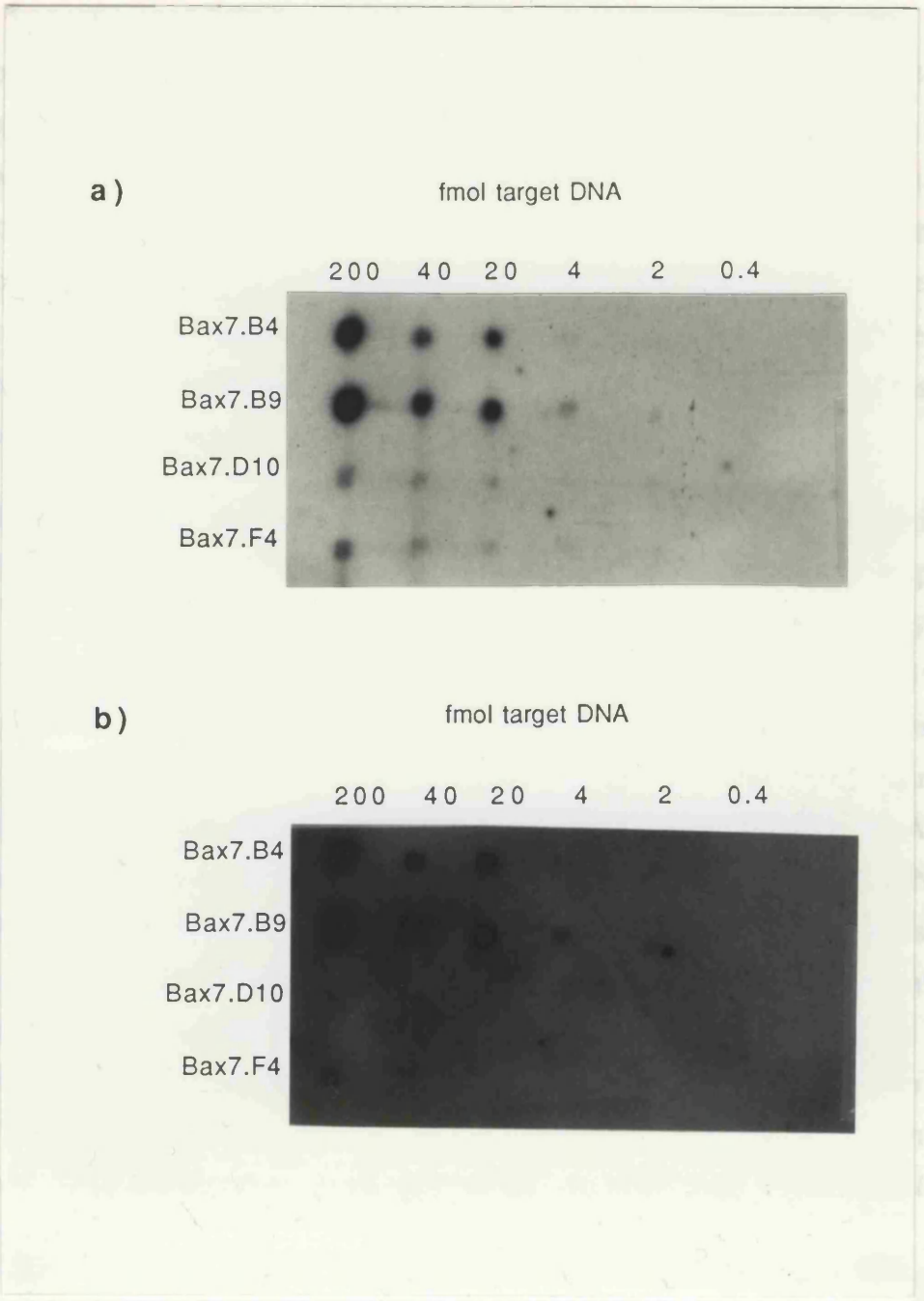


Fig. 4-19 Two oligonucleotide test hybridisations carried out on four sequenced M13 clones. A) Hybridisation carried out with oligo Bax7/8M1 (CAGGCCGA). B) Hybridisation carried out with oligo Bax7/8M3 (CACAGGCC). Both hybridisations were carried out in SSarc buffer at 5°C and 3 nM probe concentration for 3 hours. Washing was carried out also in SSarc buffer at 12°C for 30 min. Oligonucleotide probes were labelled, by a kinasing reaction, with [^{32}P-γ] ATP.

hybridisation with an octamer that contains full matches in Bax7.B4 and Bax7.C9 and no matches longer than 4 bp in either Bax7.D10 or Bax7.F4, shown in figure 4-19b, the detection limit was also around 20 fmol. The spots in these hybridisations were on average 2 mm in diameter (3.14 mm^2), which equates to a detection limit of the order of 6 fmol mm^{-2} for octamer hybridisations. The results of these two hybridisations indicated that sequence specific DNA hybridisations with octamers was possible under the conditions described.

A detection limit of 6 fmol mm^{-2} is sufficient to allow direct spotting of PCR products after amplification using the arraying robots developed in the lab. The yield of PCR amplifications, using the waterbath system, vary between approximately 5 - 50 ng μl^{-1} for those 75% of amplifications that are successful. Spotting pins of 0.4 mm diameter transfer around 200 nl each, onto an area of approximately 0.13 mm^{-2} so that for a typical PCR product a single transfer deposits DNA at a density of between 8 and 80 fmol per spot, for an average cDNA of 1,500 bp. Since this means that at the lower level the DNA density is close to the detection limit, each PCR product was transferred onto the same spot five times, when using the spotting robots. Using the latest generation robots in the lab (Maier et al., 1994a) it is possible to array 20,736 PCR products onto 12 replica membranes (22 cm X 22 cm) in 6 hours. The membranes were processed as described in *Materials and Methods*.

To characterise the oligonucleotide hybridisations a further experiment was carried out to estimate the proportion of target sites filled during an oligonucleotide hybridisation. Three separate membrane strips (Hybond N+, Amersham) were spotted manually with: 1) a serial dilution of known amounts of alkali denatured target DNA alongside a negative control DNA. 2) purified alkali denatured radioactively labelled DNA of known concentration. 3) a dilution series of [^{33}P-γ]ATP. The radioactively labelled DNA was spotted onto two strips of membrane. One was kept untreated while the other was neutralised and then used in

a 'mock' hybridisation, that is, an incubation in hybridisation buffer without probe, followed by the same washing as the actual hybridisations. The membrane containing the serial dilution of target DNA and negative control, was hybridised with an octanucleotide, whose specific activity had been measured, and washed as described above. The percentage DNA binding to the membrane was calculated as the fraction of radioactivity remaining on the filter with the radioacively labelled DNA. After hybridisation and washing the filter containing the target DNA was exposed on one screen together with the strip on which the dilution series of [^{33}P-γ]ATP was spotted. After scanning the screen using a phosphorimager (Molecular Dynamics, Sunnyville CA), the 'Image Quant v3.2' software form Molecular Dynamics was used to quantify the signal on each of the spots. Figure 4-20 shows a plot of the moles of [^{33}P-γ]ATP against the pixel counts measured by the phosphorimager. This plot was used to estimate the moles of probe molecules detected at each hybridising spot. The plot also confirms that the detection of the phoshorimager is linear over 5 orders of magnitude, as claimed by the manufacturers. In this experiment the percent binding of the DNA to the nylon membranes was estimated at 50%. At an oligonucleotide concentration of 5 nM it was estimated that there was approximately one probe molecule for every 7 target molecules bound to each spot (i.e. 14%). This implies that the equilibrium constant (K_{eq}) for the formation of DNA duplex (probe + target <=> duplex), is small for an octamer hybridisation at 5 nM in 1 M Na^+ at 5°C and that the equilibrium lies to the left, assuming that the reaction has reached equilibrium during the hybridisation, which in terms of kinetics is likely in the time given (Wetmur , 1991).

Experience of many oligonucleotide hybridisations indicates that the proportion of target sites filled varies greatly between oligonucleotides, a conclusion derived from the variable hybridisation signals obtained from different oligonucleotides hybridised under identical conditions to the same target DNAs.

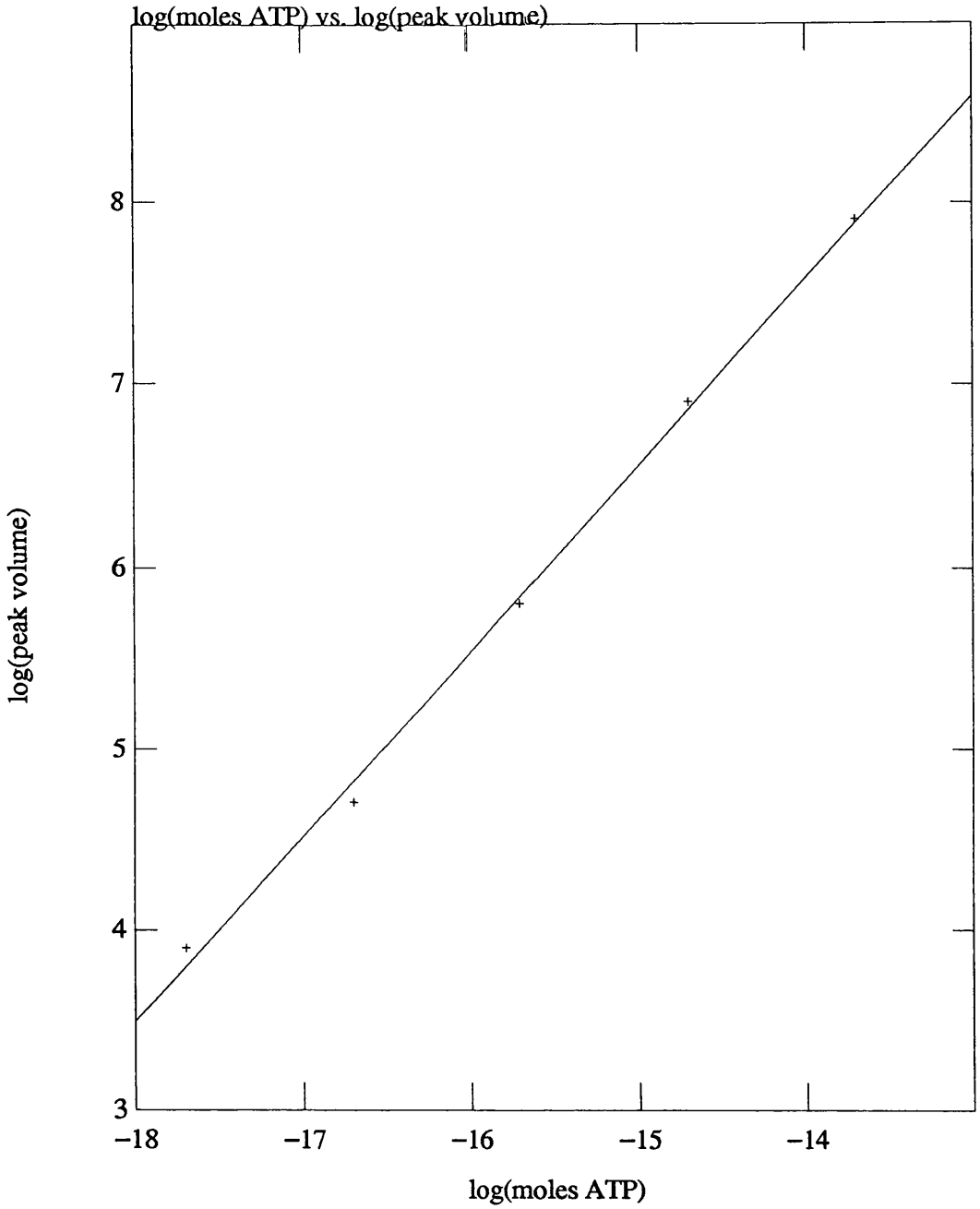


Fig. 4-20: A plot of $\log_{10}$ pixel counts versus $\log_{10}$ moles of [^{33}P−γ] ATP. A serial dilution of [^{33}P−γ] ATP was spotted onto a nylon membrane (Hybond N+, Amersham) and exposed to a phosphor storage screen an room temperature for 1 hour. The image was scanned using a phosphor imager (Molecular Dynamics) and the signal quantitated using the ImageQuant v3.2 software. An ellipse was drawn around each signal and the signal calculated by integration of the volume of the ellipse.

4.3 Software tools for the automated analysis of hybridisation data

In order to analyse hybridisation data of short oligonucleotides hybridised to high density filter grids of cDNA PCR products an automated image analysis system is required. Unlike hybridisations with single copy probes, as used for many physical mapping applications, short oligonucleotide hybridisations give a positive signal rate of the order of several percent of target clones (see *Oligonucleotide selection* later in this chapter), which in the case of our high density clone arrays means more than 1,000 positive signals per hybridisation. The density of the clone arrays also requires a high spatial resolution, for which purpose ^{33}P-ATP was used for labelling the oligonucleotides, since the degree of signal spread during both autoradiography and phosphorimaging is significantly less than with the ^{32}P isotope (the energy of the β-particles emitted is ten fold less for ^{33}P). It is not feasible, or desirable, to score such hybridisation manually, especially since there is a significant variation in hybridisation signals between the positive clones, due to unequal DNA quantities, which requires normalisation. Some time was spent in developing a robust system of data analysis that would perform well with hybridisation data of very variable quality and the resulting package of tools is described below. The conception and realisation of the system described in the following pages was carried out together with Richard Mott. All the computer code for the image analysis was written by Richard Mott using the programming languages of C, AWK and C-shell.

4.3.1 Image capture and quantitation

After hybridisation and washing the membranes were exposed to phosphor storage screens (Molecular Dynamics (MD), Sunnyville, CA)

for 3 - 16 hours at room temperature. The screens were scanned at a resolution of 176 μm and the images captured in 16 bit TIFF format with a phosphorimager (MD) and then transferred onto an optical juke box connected to a network of SUN SPARC stations running SUNOS 4.1.3 and OpenWindows 3. Each image contained approximately 2 x 10^6 pixels and required 4 MB disk space. Figure 4-21 shows an hybridisation image photographed from a SUN monitor.

Software for automated image processing was written in C using the HIPS image analysis library and the Motif X graphics library.

The first and most demanding step in the image analysis is to locate all the spots in the image and assign them to their correct grid positions. Since the membranes distort slightly with use and cannot be reliably exposed to the phosphorstorage screens in exactly the same positions it is necessary to develop a system whereby small offsets and rotations can be accommodated. The cDNA and M13 PCR products were spotted onto nylon membranes (Hybond N+) in blocks of 3 x 3 arrays (i.e. 9 spots) in which the central position of each array was spotted with genomic salmon sperm DNA (20 ng μl^{-1}). The salmon sperm DNA was included because it hybridises with all short oligonucleotides, since its sequence complexity is very high and can therefore act as a beacon. The amount of salmon sperm DNA was adjusted so that the signal is approximately equivalent to that of a positively hybridising PCR product. Figure 4-21 shows a hybridisation image in which the pattern of beacon signals and surrounding experimental clones can be seen. The analysis software exploits the fact that clones were spotted in blocks of 3x3 spots with the centre spot in each block acting as a beacon. The stages in the image-analysis of are as follows:

1. The 16 bit TIFF Phosphorimager file is converted to 16 bit HIPS format.
2. The corners of the clone grid on the image are located manually by clicking on the image with a mouse. From the coordinates of the

corners the angle is calculated, through which the image must be rotated to make the rows and columns parallel to the pixels in the image. The image is then rotated automatically.

3. The salmon sperm DNA beacons at the centre of each block of clones are located by computing the histograms of row and column sums for the image. The centre of each block of 3x3 clones is then defined as the intersection of the corresponding row and column maxima.
4. The non-beacon spots in each block (i.e. the experimental clones) are located by searching for local maxima in the image within a small window around the expected spot locations. If no local maximum can be found for a spot it is treated as missing.
5. For each located spot the local maximum is determined and in addition the minimum in the whole block is also recorded. This helps to take into account the effects of local variations in signal intensity due to uneven hybridisation across the membrane. At this stage the spots are not classified as positive or negative; instead the signal information is carried forward into the joint analysis of the hybridisations.
6. The grid coordinates of all spots are determined and from these, together with the known order of clone spotting of the array, the plate and well coordinates corresponding to each spot are calculated.
7. A text file is written that contains the information of clone identity, grid coordinates, spot intensity and local background for all the spots in the image. If no local maximum was found for a given grid location then a value of zero was assigned and the clone marked as missing ('-1' in the final column) in the text file. Figure 4-22 shows a sample of such a text file.

The complete image analysis of a single hybridisation takes approximately 1-2 minutes of elapsed time on a SUN SPARC server 10-51 (including image conversion, rotation and looking up the clone identities of each located spot).

Provided most of the beacon clones give positive signals this system has proved to be robust and gives good agreement with manual inspection of the same images. If some of the beacons do not hybridise well, which can happen for a variety of reasons, the image analysis still performs well, provided the average signal for the rows and columns of the beacons are significantly higher than of the experimental clones. The name of each image file uniquely identifies each experiment and the experimental details of each hybridisation are stored in a database to which the image name cross references.

The software described above represent a set of tools that enable the collection of quantitative hybridisation data from a large number of experiments, as required for a large scale oligonucleotide fingerprinting project as described in the *Introduction*.

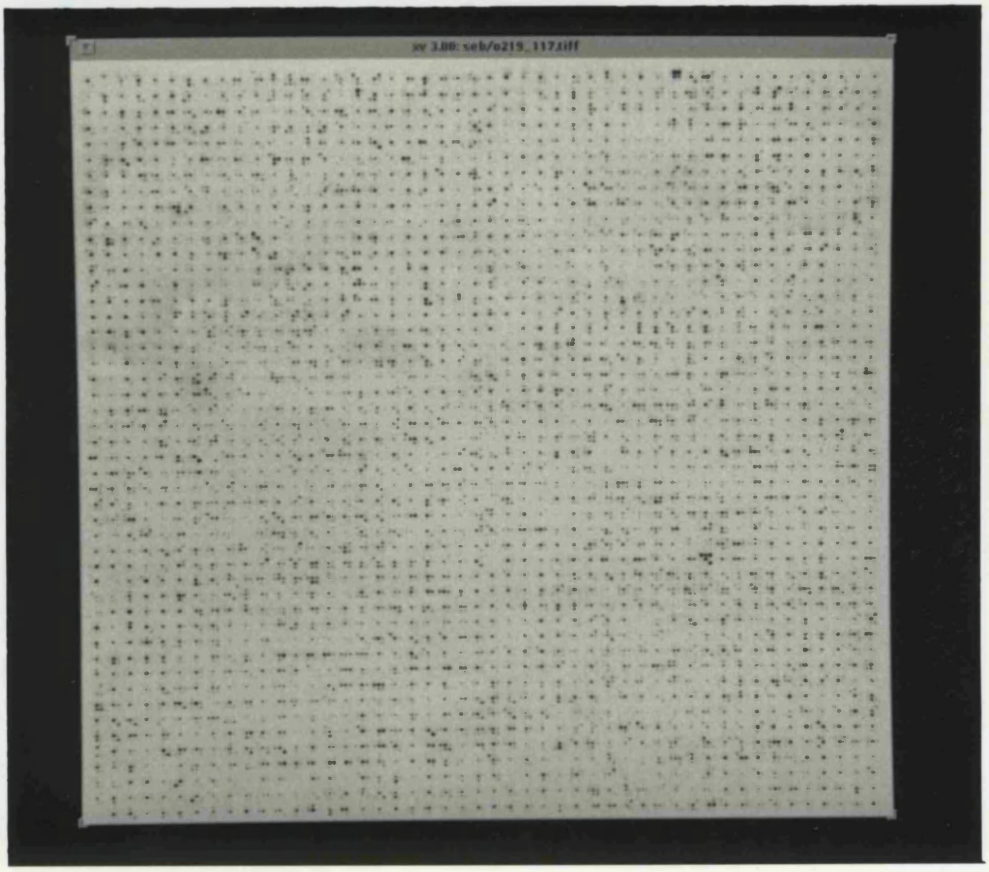


Fig. 4-21 A photograph of a hybridisation image, in this case oligo 219, taken from a Sun workstation monitor. The image shows clearly the beacons (salmon sperm DNA) hybridising positively in the centre of each block of 9 spots and in addition to experimental clones in the remaining positions. Hybridisation and washing were carried out in SSarc at 5°C with a ^{33}P-labelled probe.

ICRFhbf_32024	74	4	684	109	0	4656964	0	-1
ICRFhbf_38024	75	4	693	109	8708401	4656964	4051437	0
ICRFhbf_26023	76	4	705	110	9541921	4481689	5060232	0
ICRFhbf_32023	77	4	711	110	7392961	4481689	2911272	0
ICRFhbf_38023	78	4	717	108	6533136	4481689	2051447	0
ICRFhbf_26022	79	4	729	109	5359225	4268356	1090869	0
ICRFhbf_32022	80	4	734	106	5161984	4268356	893628	0
ICRFhbf_38022	81	4	743	106	5225796	4268356	957440	0
ICRFhbf_26021	82	4	757	111	5745609	4040100	1705509	0
ICRFhbf_32021	83	4	762	109	4844401	4040100	804301	0
ICRFhbf_38021	84	4	767	110	5058001	4040100	1017901	0
ICRFhbf_26020	85	4	782	108	4721929	3810304	911625	0
ICRFhbf_32020	86	4	791	107	4796100	3810304	985796	0
ICRFhbf_38020	87	4	796	107	4761124	3810304	950820	0
ICRFhbf_26019	88	4	808	109	8450649	3841600	4609049	0
ICRFhbf_32019	89	4	815	110	9548100	3841600	5706500	0
ICRFhbf_38019	90	4	822	109	0	3841600	0	-1
ICRFhbf_26018	91	4	835	109	5308416	3976036	1332380	0
ICRFhbf_32018	92	4	841	107	5049009	3976036	1072973	0
ICRFhbf_38018	93	4	848	108	4787344	3976036	811308	0
ICRFhbf_26017	94	4	860	110	4848804	4072324	776480	0
ICRFhbf_32017	95	4	867	109	7806436	4072324	3734112	0
ICRFhbf_38017	96	4	874	109	7059649	4072324	2987325	0
ICRFhbf_26016	97	4	887	110	5400976	4284900	1116076	0
ICRFhbf_32016	98	4	892	109	4928400	4284900	643500	0
ICRFhbf_38016	99	4	898	109	5212089	4284900	927189	0
ICRFhbf_26015	100	4	909	108	5184729	4322241	862488	0
ICRFhbf_32015	101	4	918	109	5541316	4322241	1219075	0
ICRFhbf_38015	102	4	926	108	5808100	4322241	1485859	0
ICRFhbf_26014	103	4	939	112	5112121	4120900	991221	0
ICRFhbf_32014	104	4	946	109	5076009	4120900	955109	0
ICRFhbf_38014	105	4	952	109	6411024	4120900	2290124	0
ICRFhbf_26013	106	4	961	111	5004169	3948169	1056000	0
ICRFhbf_32013	107	4	970	106	5285401	3948169	1337232	0
ICRFhbf_38013	108	4	978	110	5424241	3948169	1476072	0
ICRFhbf_26012	109	4	988	106	4721929	3884841	837088	0
ICRFhbf_32012	110	4	995	107	4892944	3884841	1008103	0
ICRFhbf_38012	111	4	1004	109	5414929	3884841	1530088	0
ICRFhbf_26011	112	4	1012	110	4853209	3952144	901065	0
ICRFhbf_32011	113	4	1022	109	4443664	3952144	491520	0
ICRFhbf_38011	114	4	1029	109	5396329	3952144	1444185	0
ICRFhbf_26010	115	4	1038	112	4343056	3984016	359040	0
ICRFhbf_32010	116	4	1049	109	4713241	3984016	729225	0
ICRFhbf_38010	117	4	1055	108	6817321	3984016	2833305	0

Fig. 4-22 A sample of an output file from the automated image analysis. The text file contains 9 fields: 1) Clone name 2) grid x-coordinate 3) grid y-coordinate 4) pixel x-coordinate 5) pixel y-coordinate 6) local maximum signal 7) local minimum signal 8) difference between local maximum and local minimum 9) status code for spot location (-1 = no local maximum found, 0 = spot located).

5. Scaling up to fingerprinting thousands of clones

As discussed in the *Introduction*, synthetic oligonucleotides can be used as hybridisation probes in order to generate sequence information on the target DNAs. In one application of this technique very short oligonucleotides (e.g. octamers) are hybridised to many thousands of arrayed cDNA clones in parallel. Once sufficient oligonucleotides have been hybridised, providing that the hybridisation data are reproducible, each sequence under investigation will be characterised by a unique set of hybridisation signals. The number of hybridisation events that are required to generate such a unique 'fingerprint' depends on the frequency with which the oligonucleotide probes hybridise to the target DNAs. The realisation of such a strategy for the characterisation of cDNA sequences has required the development of new experimental as well as analytical tools. The previous sections of this chapter have described the technological developments that have been achieved with the aim of assembling a complete set of tools, suitable for fingerprinting of many thousands of cDNA clones by oligonucleotide hybridisation. This section describes the first attempt at a large scale fingerprinting experiment and the assessment of the data obtained, using all the aforementioned systems.

5.1 Selection of Oligonucleotides

The selection of oligonucleotide probes is the one area in which the experimenter can influence significantly the rate of progress during a fingerprinting experiment. The number of hybridisations required correlates in a decreasing function with the hybridisation frequency of the probes used. The frequency of the probes on the other hand is determined mainly by their length, but also by their composition, since

the target DNA sequences are far from random. Another condition of a suitable set of oligonucleotide probes is that they must not be highly correlated, since in such a case the result of one hybridisation would predetermine the probability of correlated probes hybridisation. For example, if two probes have a prior probability of hybridisation (p) equal to 0.1 but differ in sequence only by one base, then a positive hybridisation with one probe increases p for the other oligonucleotide to 0.25. In selecting a set of suitable hybridisation probes a compromise has to be achieved between a high hybridisation frequency and experimentally reliable hybridisation characteristics. Although reliable hybridisation of oligonucleotides as short as hexamers has been reported (Drmanac et al., 1990b), these results are based on the results of very few hybridisations. For this project octamers were chosen as hybridisation probes since experience for dozens such probes existed in the lab. Since the oligonucleotide probes will be critical in the success of any hybridisation fingerprinting experiment considerable thought and effort were invested in the selection of the hybridisation probes.

Using octanucleotides as probes, there are 32,768 ($4^8/2$) sequences from which to select the few hundred required for fingerprinting, so there is a vast excess of possible probe sequences to choose from, as distinct from sequencing by hybridisation (SBH) (Drmanac et al., 1989; Bains and Smith, 1988; Khrapko et al., 1989). An important factor affecting the efficiency of an oligonucleotide-based hybridisation strategy is the choice of probes, which should be selected so that the information each probe imparts will be large and complementary to that of the others. In the ideal case, n probes can generate 2^n different fingerprints and hence theoretically could distinguish between this number of clones. So for example a library of 1000 clones could be covered in only 10 hybridisations. However, in practice less impressive results are obtained because the sequence compositions (and consequently the fingerprints) of the cDNA clones in a library are far from random, and because complementary probes with the optimal hybridisation frequency of 50% cannot be chosen, for experimental reasons (for average cDNAs

of 1,500 bp hexamers are required to achieve a 50% hybridisation frequency and these are difficult to hybridise reliably). Further, it is desirable to exclude probes with certain compositions, such as poly-dA or poly-dT for example.

These problems can be mitigated by restricting the choice of probes to those having high hybridisation frequencies in a databank of known sequences, and which perform well at partitioning the databank (i.e. split it into many partitions of approximately equal sizes). This argument assumes that the sampling properties of the databank sequences will reflect those found in the library to be fingerprinted.

To enable the selection of oligonucleotide probes based on the composition of sequenced genes a set of suitable sequences were extracted from the combined databanks of the EMBL and Genbank databases. Using the GCG package, the string search command was used to obtain a list of all the sequence entries whose names begin with the 3 letters 'hum' (i.e. human sequences) and contain the string 'mRNA' in their definitions. A total of 4,551 sequences were found and closer examination confirmed that all contained sequence data derived from mRNA. This selection of sequences is likely to represent most closely the composition of the cDNA clones in the human foetal brain cDNA library. The sequence names were then used to build a separate database of these sequences using the GCG command 'dataset'.

The following algorithm (designed and coded in C by Richard Mott) was used to select sets of oligonucleotide probes (in fact octamers) for hybridisation, using a subset of 2,000 of the extracted sequences.

Suppose that n probes have been chosen, and that in a databank of M sequences a proportion m_j sequences have a fingerprint $j, (j = 1...n)$. Then a good measure of the partitioning is the information content, or entropy

$$E_n = -\sum_j m_j \log(m_j)$$

since this is a maximum when all partitions are of equal size. Since it is computationally unfeasible to select a set of n probes that maximise E_n because this involves too many trial evaluations, a sequential algorithm was used to select the probes: define $E_0 = 0$ and then at the n^{th} step choose that probe which maximises the increase in entropy

$$\Delta E_n = E_n - E_{n-1}$$

The reason for using ΔE_n to choose the next probe rather than the increase in the number of partitions is to equalise the partition sizes as far as possible. The performance of the algorithm may be compared with the expected number of partitions obtained when the probes are chosen at random: Let p be the average hybridisation frequency of the probes. Group the 2^n possible fingerprints according to the number of positives r, so there are $\binom{n}{r}$ fingerprints with exactly r positives. On the assumption that the M databank sequences are random and independent, the probability (q)that a fingerprint with r positives occurs at least once in the databank is

$$q_r = 1 - (1 - p_r)^M$$

where

$$p_r = p^r (1 - p)^r$$

Hence the expected number of distinct fingerprints is

$$F_n = \sum_r \binom{n}{r} q_r$$

Figure 5-1 shows the performance of the algorithm when applied to two sets of 2000 mRNA sequences from the EMBL and GENBANK databases. Probes were chosen from the most frequent 10% of octamers, excluding oligonucleotides of low complexity. The average hybridisation frequency of this probe set was 0.1. The figure demonstrates that the algorithm is more efficient than picking probes at random from the same probe set, and that although the probes are less

efficient when applied to a new set of sequences, they still perform better than random. The figure also shows the marked departure from the expected progress (equation 5) on a random databank of 2000 sequences with p=0.1, indicating the biased composition of the sequence databases. With a random databank of 2000 sequences it is possible to identify each sequence uniquely with about 50 probes, whereas approx. 150 probes are needed to uniquely identify over 95% of the sequences of the databank (4,551 sequences) used here.

Table 5-1 shows 268 octamer sequences selected from several runs, using the algorithm described above.

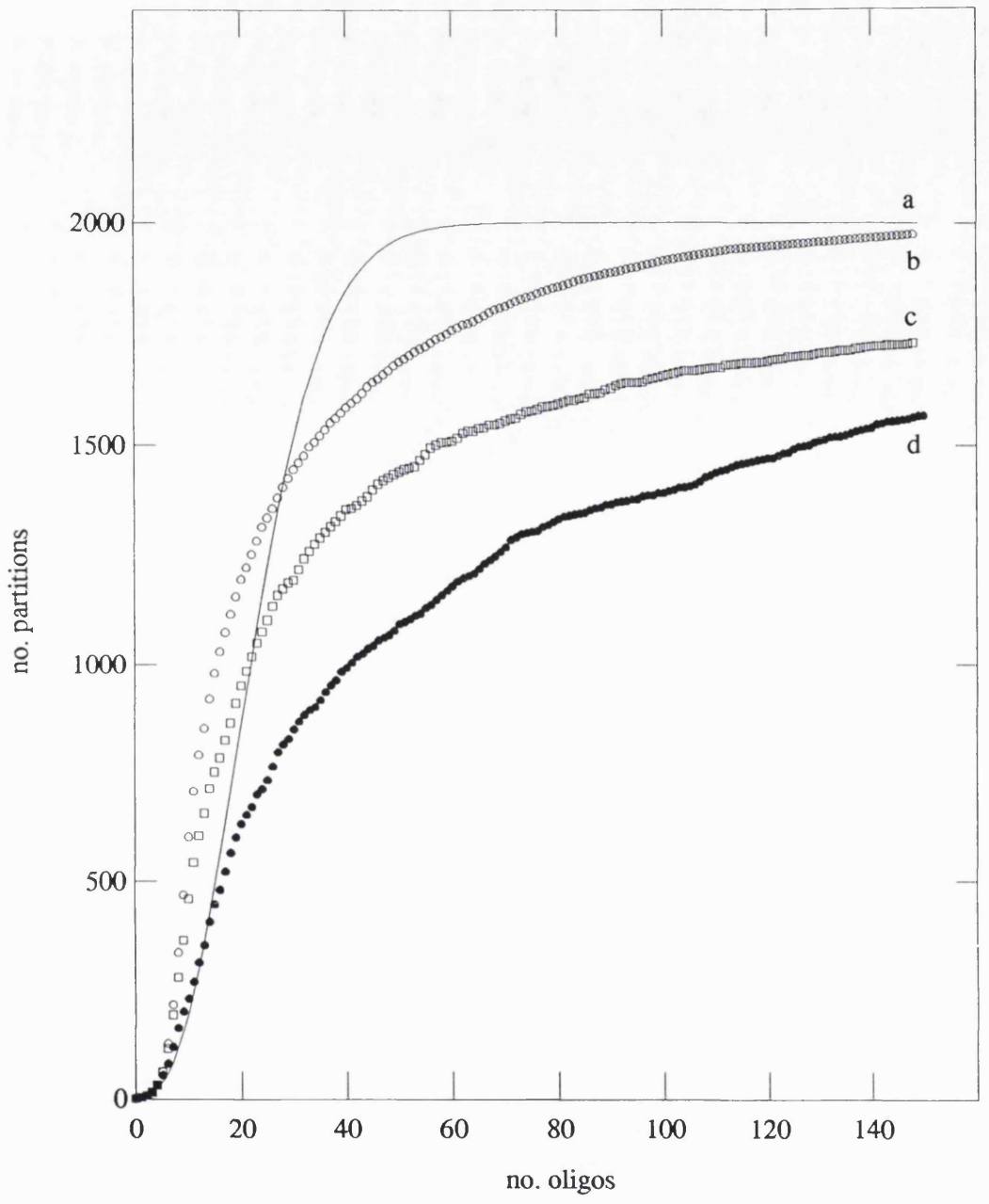


Fig. 5-1 A plot of the increase in the number of partitions in a set of 2,000 expressed sequences plotted against number of probes. The plot shows the progress for: randomly selected probes (d), probes selected according to the partitioning algorithm (see text) (b), the same probes used to partition 2,000 different expressed sequences (c) and the theoretical progress of random selected probes on clones of random sequences (a).

Table 5-1

o1 ggaggagg	o52 tgctggcc	o103 naggacctgn
o2 ccaccacc	o53 tggcagtg	o104 ncctggccan
o3 tggaatgg	o54 cctccttc	o105 ntgctggagn
o4 cctcatct	o55 cacctgga	o106 ngctgctgcn
o5 ggaatgga	o56 cctgcagg	o107 ncctgggctn
o6 tggagtgg	o57 agcagctg	o108 nagcaggagn
o7 cacacaca	o58 gaagacag	o109 ntggaggagn
o8 ccctcatc	o59 agccagaa	o110 ntggagaagn
o9 tgatgatg	o60 gcagaagc	o111 ncctgggcan
o10 ggagtgga	o61 cttctttt	o112 ngctggggcn
o11 cagcctgg	o62 ctgggctg	o113 nccccagccn
o12 ccagcctg	o63 tggagaga	o114 ncctcagccn
o13 caggctgg	o64 tggggcag	o115 ngaaggaggn
o14 ggaggctg	o65 ttgaaaaa	o116 ngctcctggn
o15 cagcctcc	o66 gagaagaa	o117 ncctgggccn
o16 agcctggg	o67 acccccctg	o118 nccaggcccn
o17 ggctgagg	o68 tttgcaga	o119 nctccagccn
o18 ggctggag	o69 ntgggagagn	o120 ntgctggtgn
o19 aggctgag	o70 nggacacctn	o121 ntggagcagn
o20 gaggctga	o71 nagccagggn	o122 ntggccctgn
o21 tgctgctg	o72 natggggaan	o123 nggaggaagn
o22 ggagctgg	o73 nggagacccn	o124 ncctgcaggn
o23 cctgctgg	o74 ngacctgctn	o125 nagctggagn
o24 tgaagaag	o75 nctgctgctn	o126 nctgctggcn
o25 ccctggcc	o76 nggagctggn	o127 nggcagctgn
o26 tggagaag	o77 ncagcctggn	o128 ntggggctgn
o27 cagcctgg	o78 ntgaagaagn	o129 ncctgcagcn
o28 aggaggag	o79 ncctggagan	o130 nggaaggagn
o29 cctcctgg	o80 naggaggagn	o131 ncctggaggn
o30 ttttaaaa	o81 nccctggccn	o132 nctggagaan
o31 cccagccc	o82 ngctgctggn	o133 nttcctggan
o32 agctgctg	o83 naggagaagn	o134 ngaggagaan
o33 cctggagc	o84 ncctggagcn	o135 nccaggaggn
o34 gccacctg	o85 nctggaagan	o136 nctggagcan
o35 aaggaaaa	o86 ncccagcccn	o137 ngagctgggn
o36 cttcctgg	o87 nctcctgctn	o138 nggagcagcn
o37 cctccctg	o88 naggagcagn	o139 ncagcctccn
o38 aggacctg	o89 ncttcctggn	o140 nagctgctgn
o39 agaagaga	o90 nctggaggan	o141 ncctggcccn
o40 ctgcagct	o91 ncagctgccn	o142 ncctggctgn
o41 catcctgg	o92 ngagaagaan	o143 ngctgggccn
o42 tttgtttt	o93 ngaagaggan	o144 ngccctgggn
o43 gctcctgg	o94 nggctggggn	o145 ncctggaagn
o44 ctgggctc	o95 nagctggtgn	o146 ngaggaggan
o45 tggtggtg	o96 nagcagcagn	o147 naaggagaan
o46 tgctggca	o97 ngctgcagcn	o148 ncagccctgn
o47 tgctggag	o98 nctcctggan	o149 ncctgctgcn
o48 ctgggtcc	o99 ncctggcagn	o150 nggaggtggn
o49 gatgagaa	o100 nagcagctgn	o151 naggaagagn
o50 aaagagaa	o101 ntcctcctgn	o152 ncctcctgcn
o51 ctgggaca	o102 nggggctggn	o153 ngcagctggn

o154 ncccagggcn
o155 ntcctggagn
o156 ntggagctgn
o157 nctgggctgn
o158 nggagaagan
o159 ntgctgctgn
o160 nctgcagccn
o161 ncctgctggn
o162 naggaggaan
o163 nggccctggn
o164 ngctggtggn
o165 ntgcagctgn
o166 nccctgctgn
o167 ngcagcagcn
o168 ngaggaagan
o169 ntattactgn
o170 ncagcctggn
o171 nctgctgctn
o172 ntgcagctgn
o173 nccagcagcn
o174 ntggagctgn
o175 nttttaaaan
o176 ncctcctgcn
o177 nggccaaggn
o178 ngcagcagcn
o179 nggagaagan
o180 nctggggccn
o181 nccctgcccn
o182 ntggaggagn
o183 ncagcctgan
o184 ncctggaagn
o185 nctcaccatn
o186 nggctggagn
o187 ngaagaggan
o188 ngccctgggn
o189 nccatctccn
o190 naagagaagn
o191 ntggagaaan
o192 ntgaagaaan

o193 nccctggccn
o194 nctgtgctgn
o195 nttcctcctn
o196 ncctggtcan
o197 nagggaccan
o198 ncctgcaggn
o199 ncttcctggn
o200 ntttgaagan
o201 ngtaccagcn

o232 ntgctgtgtn
o233 nccagaaccn
o234 naatgaggan
o235 ncgtctcctn
o236 ntgctcctgn
o237 ngtggtggtn
o238 ngggctgagn
o239 nttcaccaan
o240 ngaagccccn
nttctggaan
ncctgagccn
nctctggccn
ntgacctggn
ntgatgatgn
ncatcatggn
ntttcagaan
nctgaagatn
nctactgggn
nttcctgcan
nttgcctttn
nctcccacan
nagctcactn
ntgggatggn
nagaagcccn
ngctgggtgn
ncagacaccn
nccttgctn
ncccctgtccn
ngaggcggan
ngaagcagan
ntttctctgn
nccccagccn
nttttaaan
natgagcagn
ngggagagan
ngccaggacn
ncatggcccn

5.2 Generating high density filter arrays for fingerprinting

In the first large scale attempt of this technique, 32,256 cDNA clones were used as hybridisation targets as well as 1,348 sequenced M13 control clones (kindly provided by S. Beck). All the clones were used in waterbath PCR reactions performed directly in 384-well microtitre plates (Q-plates). Amplifications were carried out exactly as described in *Materials and Methods*. All clones were spotted onto two 22 cm x 22 cm filters in arrays of 3 x 3 boxes. The centre position in each block of 3 x 3 clones was spotted with genomic salmon sperm DNA (20 ng μl^{-1}) which acted as beacons during the image analysis (see *Image capture and quantitation*). Each filter contained 16,128 cDNA clones and all the sequenced M13 control clones. The control clones were all spotted in the bottom left hand corner of each block of 3 x 3 clones. Figures 5-2 and 5-3 show hybridisations carried out on the filters using the cDNA PCR primer and the M13 control clone PCR primer respectively. The M13 control clones were only arrayed into wells 1 - 20 of the Q-plates which can clearly be seen in figure 5-3. Filters were always hybridised in pairs and therefore the control clones were hybridised in duplicate. Hybridisations for all probes were carried out with the same hybridisation and washing procedure as described in *Materials and Methods*.

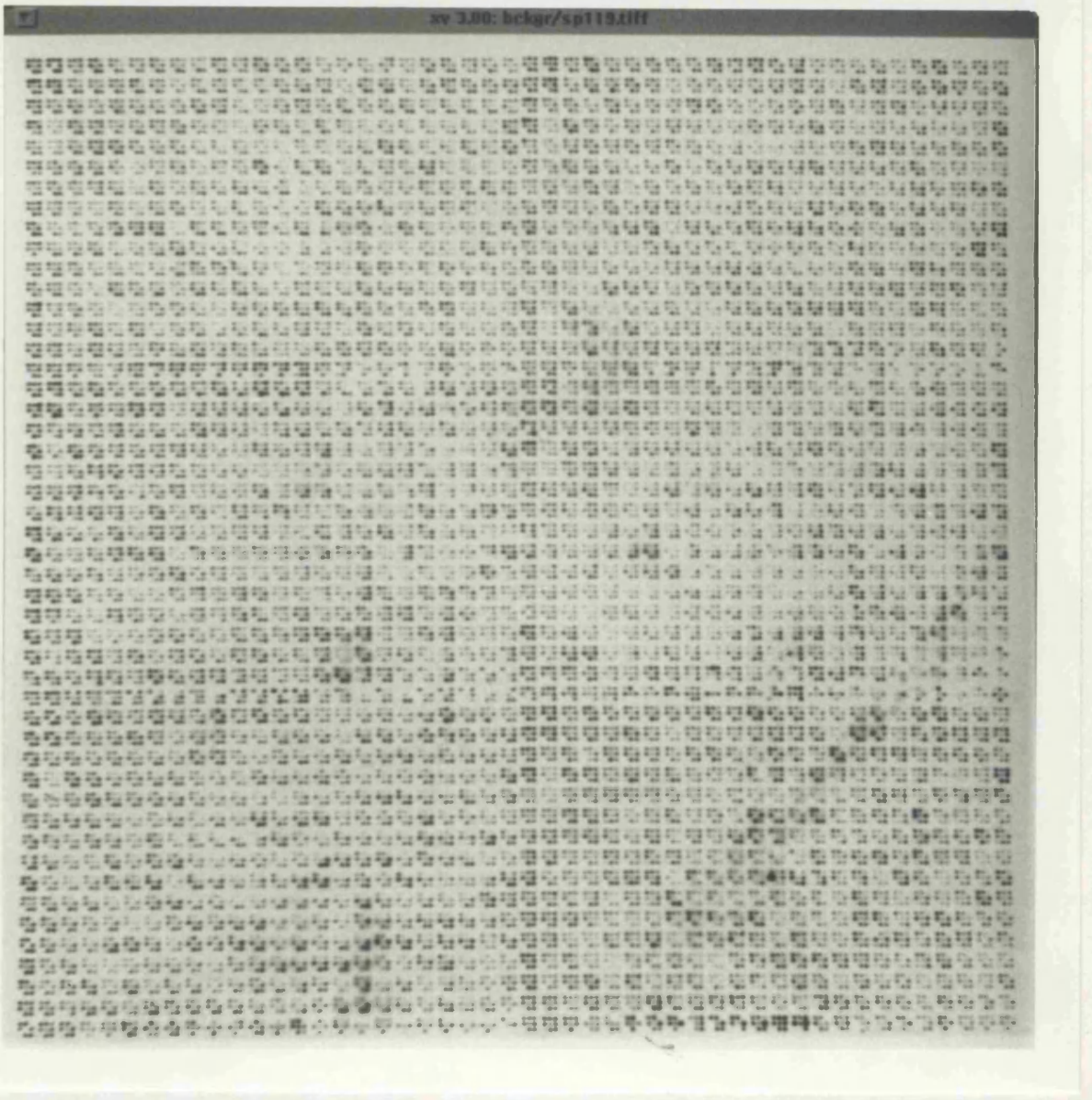


Fig. 5-2 A photograph of a hybridisation carried out on PCR filters using the cDNA PCR primer 5/86 as hybridisation probe. Hybridisation was carried out in SSarc at room temperature at a probe concentration of 1 nM. Washing was also carried out in SSarc at room temperature for 1 hour. The probe was labelled, by a kinase reaction, with [^{33}P-γ]ATP.

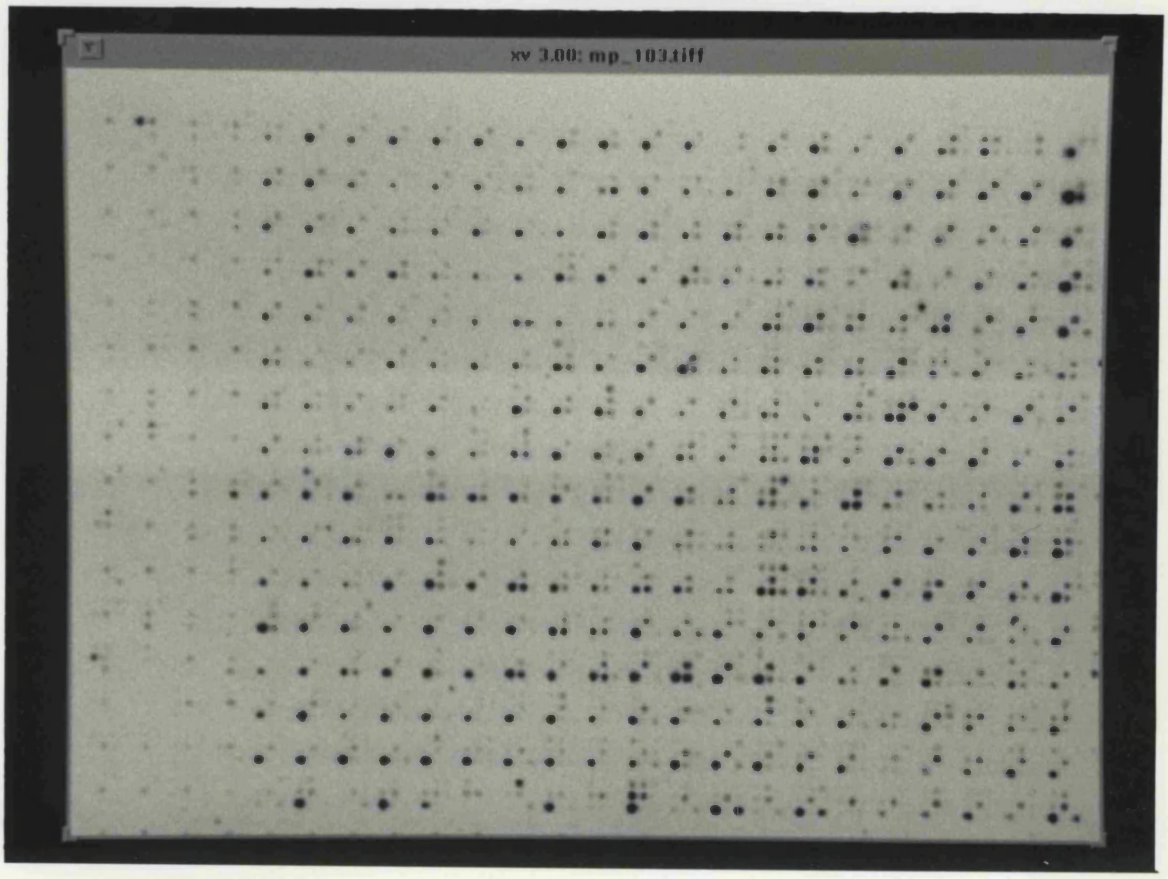


Fig. 5-3: A photograph of a hybridisation carried out on PCR filters using Mp18 control clone PCR primer Mp18RSP as hybridisation probe. Hybridisation was carried out in SSarc at room temperature at a probe concentration of 1 nM. Washing was also carried out in SSarc at room temperature for 1 hour. The probe was labelled, by a kinase reaction, with [^{33}P-γ]ATP.

5.3 Oligonucleotide hybridisations

Although there is substantial evidence that oligonucleotides do not behave equally during hybridisation, no real reliability has been achieved in predicting the hybridisation characteristics of short oligonucleotides. Detailed studies of the hybridisation characteristics of thousands of oligonucleotides (Southern et al., 1994) have confirmed their unpredictability and in some cases defied explanation by standard Watson-Crick base pairing models. Studies by the same group (Maskos and Southern, 1993a) have also shown that the sequence dependence of duplex yield can be reduced by high concentrations of tetramethylammonium chloride and that the yield can be increased up to fifty fold for some oligonucleotides. However, it is not clear from the data shown, whether the discrimination of full matches against mis-matches is improved, using tetramethylammonium chloride. In format 2 it is important that as many oligonucleotides as possible behave with similar characteristics since many thousands are hybridised to a single experimental DNA / RNA in parallel. The approach used here (format 1) uses only one oligonucleotide per hybridisation on thousands of target DNAs and thus the need for uniformity between oligonucleotide hybridisations is not as great. Two studies have been published by two other groups on sequencing by hybridisations trials (Strezoska et al., 1991; Drmanac et al., 1993), in which the hybridisation conditions were based on those published by Drmanac (1990b). So, in the absence of any clear evidence as to which hybridisation protocol is more suitable, hybridisations were carried out according to an adapted version of Drmanac (1990b).

Initially, 68 oligonucleotides (see o1 - o68 in table 5-1) were synthesised (Genosys Biotechnologies, Texas, USA) and hybridised. Ten pairs of hybridisation filters were generated and thus up to ten oligonucleotides were hybridised at any one time. Qualitative analysis of

the hybridisations with oligonucleotides o1 - o68, showed that only 30% of the hybridisations yielded images of sufficient quality suitable for automated image analysis. Most of the poor hybridisation images failed to show any significant signal above background. In a paper published during the course of this project (Drmanac et al., 1993) hybridisation probes were described which consisted of pools of decamers, sharing a core sequence of 8 bases (i.e. NXXXXXXXXN). Initially, this seemed a surprising strategy, since the specific activity of each individual probe sequence is reduced by a factor of 16. If however the duplex yield of a decamer is more than a factor sixteen greater than that of an octamer containing only the core 8 bases, then the overall hybridisation signal will be increased. In a few test experiments the increase in duplex yield between an octamer and a pool of 16 decamers, all containing the octamer as core sequence, ranged from 20 - 50 fold. It was clear from these tests that the increased duplex yield was highly dependent on the sequence of the oligonucleotides and of the target sequence, but that overall an improvement in hybridisation signal was obtained.

For the remaining oligonucleotides (o69 - o268), the octanucleotides selected by the partitioning algorithm (see *Selection of Oligonucleotides*) were synthesised as pools of 16 decamers, as shown in table 5-1. Based purely on a qualitative criterion the success rate of hybridisations with pools of sixteen decamers was around 95%.

A total of 194 oligonucleotides were hybridised onto two filters each. All hybridisation signals were captured via phosphorimaging and stored on optical disks mounted on a network of SUN workstations.

5.4 Data analysis using control clones

In total, 394 hybridisations were image analysed, 194 different probes were used, with 180 being hybridised on at least two filters. The control clones were used to measure the quality of each image and to weight its influence on the overall analysis. Previously each control clone had been partially sequenced (typically about 200 bp out of up to 1000 bp) and ordered into contigs (Beck et al., 1992). Thus for each probe a control clone could be classified as a positive, possible, or negative, according to whether the sequenced part of the clone matched with the probe, or the probe matched within 1000 bp of the clone's starting position in the contig, or the probe was absent from this 1000 bp region. Here a probe match means a match with either the probe or its complement.

5.4.1 Analysis of control clones

All images were analysed using the set of programs described in *Image capture and quantitation* and text files containing the hybridisation signals for all clones written for each image. Initially, the hybridisation signals of the control clones only were analysed. The aim of the analysis of the control clones was twofold. Firstly, to assess the hybridisation behaviour of the control clones and compare that to the expected results based on their known sequences. Secondly, to generate a measure of each hybridisation individually that can be used to appropriately weight the hybridisation in the analysis of the cDNA clones.

The first step in the analysis of the hybridisation signals of the control clones was to obtain a measure of the discrimination of any particular hybridisation. On a particular image let P be the median signal for the

expected positive clones and N that for the negatives. 'Missing' clones were excluded from this analysis. Then define the index of the hybridisation quality as the 'hyb ratio' (R)

$$R = 1 - N / P$$

In a hybridisation with good discrimination the positive clones will in general have much higher signals than the negatives, and R will approach 1. Poorer hybridisations with less discrimination will have R values closer to 0. If N is actually greater than P then R is set to be zero. It was found that the mean number of expected positives per probe was 27, which was sufficient to estimate P accurately, although a few probes had zero or a low number of expected positives, making quality assessment difficult in these cases.

Table 5-2 shows the average hyb ratios, in column 2, for all the oligos used in which column 3 indicates the number of hybridisations from which the average was calculated.

Table 5-2

o1	0.12	2	o126	0.44	2	o177	0.06	2
o9	0.02	2	o127	0.54	2	o178	0.31	2
o10	0.37	2	o128	0.63	2	o179	0.42	2
o17	0.46	4	o129	0.54	2	o180	0.20	2
o75	0.69	2	o130	0.69	2	o181	0.23	2
o76	0.66	2	o131	0.20	2	o182	0.51	2
o77	0.40	2	o132	0.07	2	o183	0.23	2
o78	0.00	1	o133	0.35	2	o184	0.17	2
o79	0.23	2	o134	0.41	2	o185	0.69	2
o80	0.36	2	o135	0.35	2	o186	0.49	3
o81	0.85	1	o136	0.46	2	o187	0.10	2
o82	0.61	2	o137	0.38	2	o188	0.08	2
o83	0.26	2	o138	1.00	2	o189	0.69	2
o84	0.55	1	o139	0.70	2	o190	0.12	2
o85	0.16	2	o140	0.61	2	o191	0.27	2
o86	0.61	2	o141	0.07	2	o192	0.00	1
o90	0.34	2	o142	0.33	2	o193	0.02	2
o91	0.59	2	o143	0.69	2	o194	0.22	2
o92	0.17	2	o144	0.12	1	o195	0.36	2
o93	0.06	2	o145	0.13	2	o196	0.48	2
o94	0.60	2	o146	0.44	2	o197	1.00	2
o95	0.56	2	o147	0.22	2	o198	1.00	2
o96	0.24	2	o148	0.44	2	o199	0.50	2
o97	0.19	2	o149	0.38	2	o200	0.04	2
o98	0.09	1	o150	0.63	2	o202	0.23	2
o99	0.68	2	o151	0.15	2	o203	0.11	2
o101	0.40	2	o152	0.74	2	o204	0.28	2
o102	0.50	2	o153	0.45	2	o205	0.18	2
o103	0.32	2	o154	0.18	2	o206	1.00	2
o104	0.16	1	o155	0.38	2	o207	0.41	2
o105	0.28	2	o156	0.57	2	o208	0.08	2
o106	0.32	1	o157	0.75	4	o209	0.00	2
o107	0.54	2	o158	0.34	2	o210	1.00	2
o108	0.17	1	o159	0.62	2	o211	0.15	2
o109	0.81	2	o160	0.66	2	o212	0.45	1
o110	0.18	2	o161	0.18	2	o213	0.08	2
o111	0.58	2	o162	0.62	2	o214	0.02	2
o112	0.40	2	o163	0.63	2	o215	0.30	2
o113	0.79	2	o164	0.56	2	o216	0.12	2
o114	0.57	2	o165	0.39	3	o217	1.00	2
o115	0.51	3	o166	0.67	2	o218	0.17	2
o116	0.78	2	o167	0.46	2	o219	0.63	2
o117	0.00	2	o168	0.24	2	o220	0.03	2
o118	0.49	2	o169	0.09	2	o221	0.32	2
o119	0.38	2	o170	0.35	2	o222	1.00	2
o120	0.26	2	o171	0.28	2	o223	0.03	2
o121	1.00	2	o172	0.06	2	o224	0.42	2
o122	0.45	2	o173	0.17	2	o225	0.04	2
o123	0.27	2	o174	0.40	2	o226	0.00	2
o124	1.00	2	o175	0.00	2	o227	0.24	2
o125	0.14	2	o176	0.25	2	o228	0.41	2

o229	0.36	2	o243	0.32	2	o257	0.50	2
o230	0.08	2	o244	1.00	2	o258	0.58	4
o231	0.77	2	o245	0.23	2	o259	0.08	4
o232	0.09	2	o246	0.28	2	o260	0.57	4
o233	0.31	2	o247	0.16	2	o261	0.32	2
o234	0.34	2	o248	0.00	2	o262	0.12	2
o235	0.03	2	o249	0.13	2	o263	0.41	2
o236	0.82	2	o250	0.38	2	o264	0.01	4
o237	0.43	2	o251	0.20	2	o265	0.36	4
o238	1.00	2	o252	0.57	2	o266	0.27	4
o239	0.06	2	o253	0.18	2	o267	0.47	4
o240	1.00	2	o254	0.69	2	o268	0.42	2
o241	0.08	2	o255	0.69	2			
o242	1.00	2	o256	0.43	2			

Of the 394 images analysed 311 gave hyb ratios greater than zero. That is there was some degree of discrimination between expected positive and negative clones. For most of the images in which the hyb ratio was less than zero, there were either only very few expected positives, of which most did not hybridise, or the image was of very poor overall quality with little signal above backgound. The hybridisation data do not show any discernible correlation with sequence composition, that might lead to a predictive measure of hybridisation behaviour. The best correlation, which while being significant, was too weak for the purpose of predicting hybridisation characteristics was the G+C content of the oligonucleotides. Figure 5-4 shows a plot of the hyb ratio against G+C content of the probes. The data indicate a positive correlation between hyb ratio and G+C content although it is clear that other factors are involved. In general it was found that observed signal strengths after hybridisation and washing also correlate positively with G+C content of the probe, which is consistent with the finding of Maskos and Southern (1993a) on duplex yield with respect to G+C content.

Figure 5-5 shows a histogram of the hybridisation signals for the positive and negative clones in a hybridisation with oligo o1 (GGAGGAGG). The histogram shows clearly there is substantial discrimination in this particular hybridisation, but also that there is a significant overlap in the distribution of positive and negative signals.

Since almost all the hybridisations were performed in duplicate with respect to the control clones a measure of the reproducibility of the data can be obtained by comparing signals between duplicates. Figure 5-6 shows a plot of the hybridisation signals for the control clones in two duplicate hybridisations of oligo o17 (GGCTGAGG), whose hyb ratios were 0.40 for o17_112.od and 0.44 for o17_102.od. In order to make the scale of both axes equivalent the hybridisation signals were normalised by dividing the signals by the average signal in that

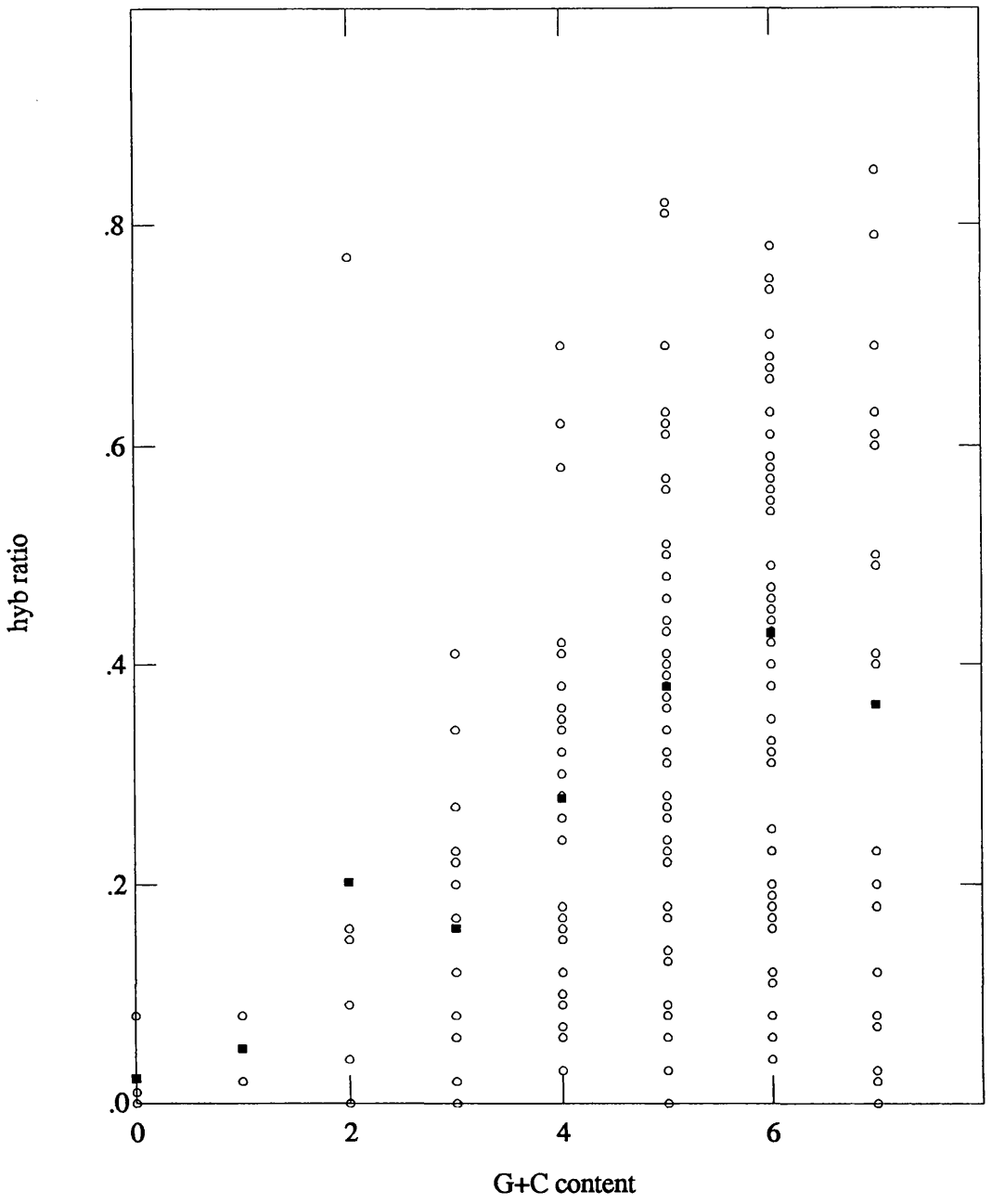


Fig. 5-4 A plot of the observed hyb ratio versus the G+C content of all the oligonucleotide probes used (open circles). For those probes hybridised on duplicate filters the mean hyb ratios are shown. The filled squares show the mean hyb ratio for all oligos with a given G+C content.

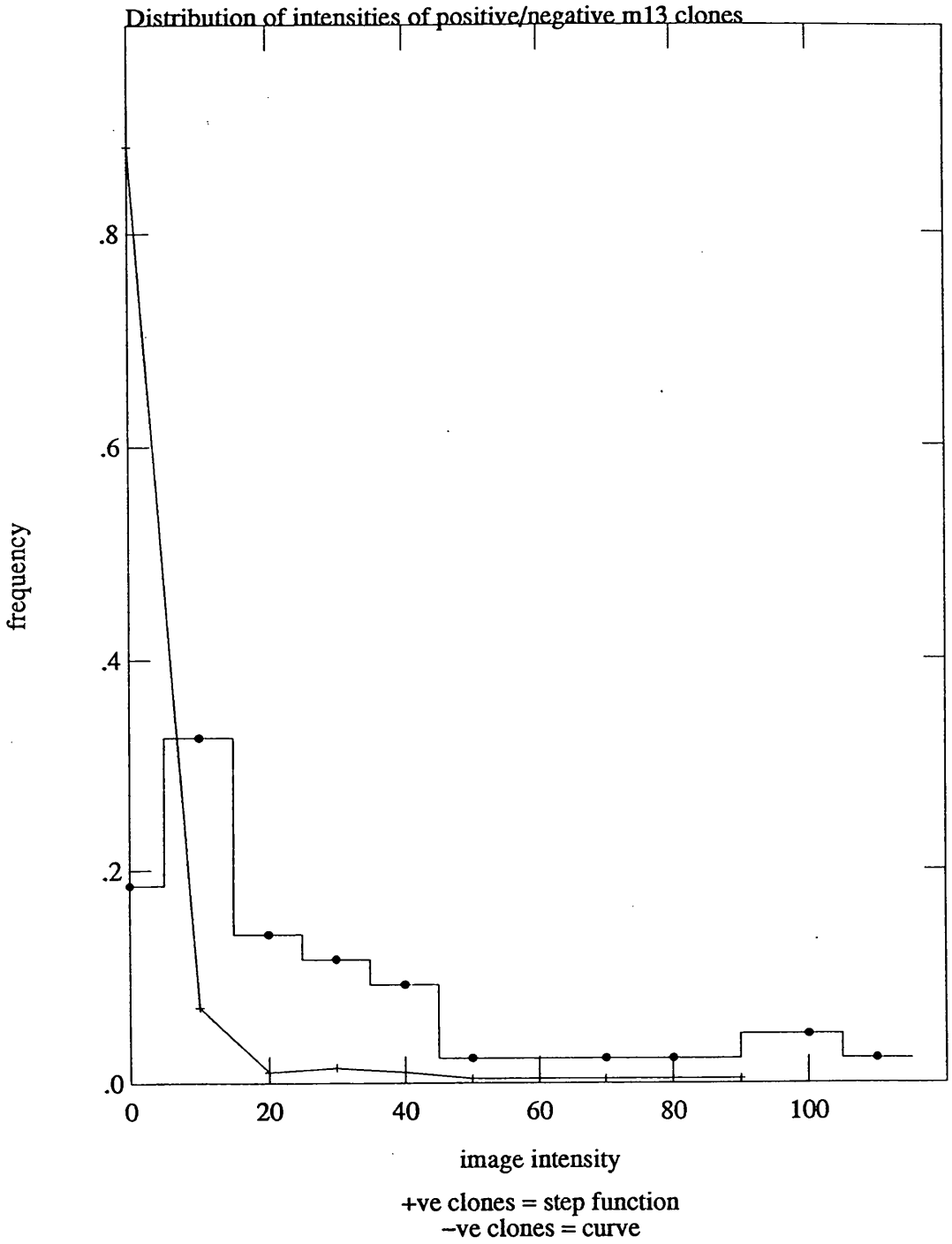


Fig. 5-5 A histogram showing the distribution of signal intensities for positive and negative control clones in a hybridisation carried out with an octamer (o1).

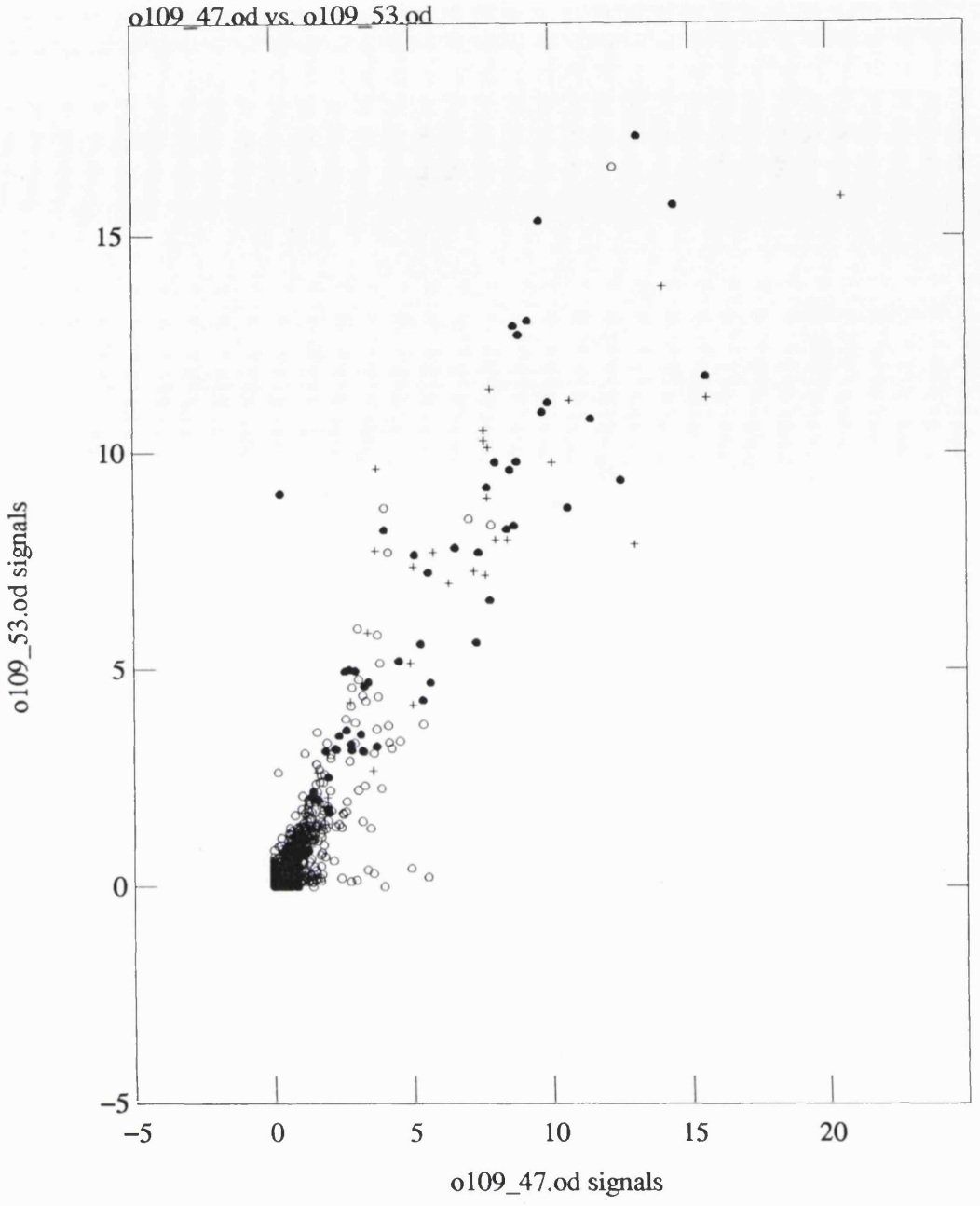


Fig. 5-6 A plot of normalised hybridisation signals for a duplicate hybridisation of oligo 109 (hyb ratio = 0.81). The signal of every control clone in one duplicate is plotted against the signal of the same clone in the second duplicate. Crosses (expected positives), filled circles (possible positives), open circles (expected negatives). Signals were normalised by dividing the hybridisation signal by the average signal for that hybridisation.

hybridisation. This way, absolute signal intensities that are determined by the specific activity of the probe and the length of exposure time are avoided, so that hybridisations can be compared in a meaningful way. The plot clearly shows a good correlation of the signals between duplicates in this case. The plot also shows that the best correlation exists between the expected positives. Also apparent from the plot is that the there are strongly hybridising expected negative clones that are consistent between duplicates. This suggests that some of the 'false positive' signals are not random hybridisation artefacts but reproducible probe / target interactions. The same plot is shown in figure 5-7 for a hybridisation with a poorer hyb ratio (0.12). This figure clearly shows that the reproducibility of the hybridisation signal varies greatly from one oligonucleotide to another.

Figure 5-8 shows a scatter plot of the hyb ratios of duplicate hybridisations of 180 different oligonucleotide probes (i.e. the same probe on two different filters). Those probes which give a hyb ratio > 0.5 in one hybridisation in general will do so in the other, indicating that good probes usually give reproducible results on different filters. On the whole it was found that the hyb ratio is a good indicator of the quality of hybridisations. In the cases where there are no expected positive control clones for an oligonucleotide, an alternative measure could be taken. This measure could be based upon the reproducibility of the duplicate hybridisations, such as a correlation factor for example. The difficulty with calculating a correlation factor for two hybridisation images is that the majority of signals are negative (~95%) and that therefore the correlation between two hybridisations will tend to be high, even if the positive signals are not shared. One possible solution to this problem could be to take the top 10% of signals in each hybridisation and determine the correlation coefficients for the duplicates. The mean of the two correlation coefficients could then be

taken as a measure of the reproducibility of the hybridisation signal. A few hybridisations were analysed manually in detail to check that the image analysis produced hybridisation signals consistent with what can be observed in the image and almost invariably, the signals assigned to a clone by the image analysis were correct.

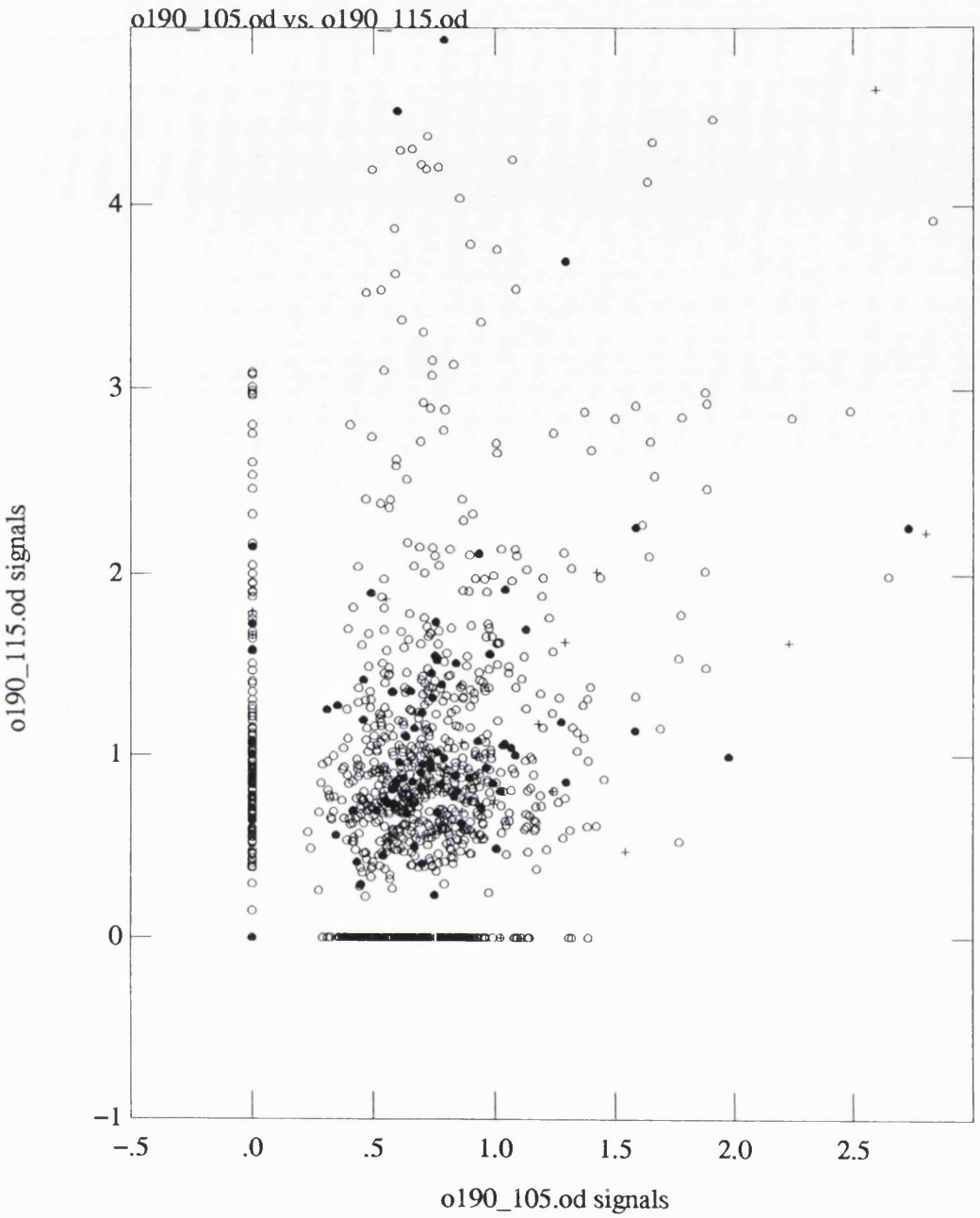


Fig. 5-7 A plot of normalised hybridisation signals for a duplicate hybridisation of oligo 190 (hyb ratio = 0.12). The signal of every control clone in one duplicate is plotted against the signal of the same clone in the second duplicate. Crosses (expected positives), filled circles (possible positives), open circles (expected negatives). Signals were normalised by dividing the hybridisation signal by the average signal for that hybridisation.

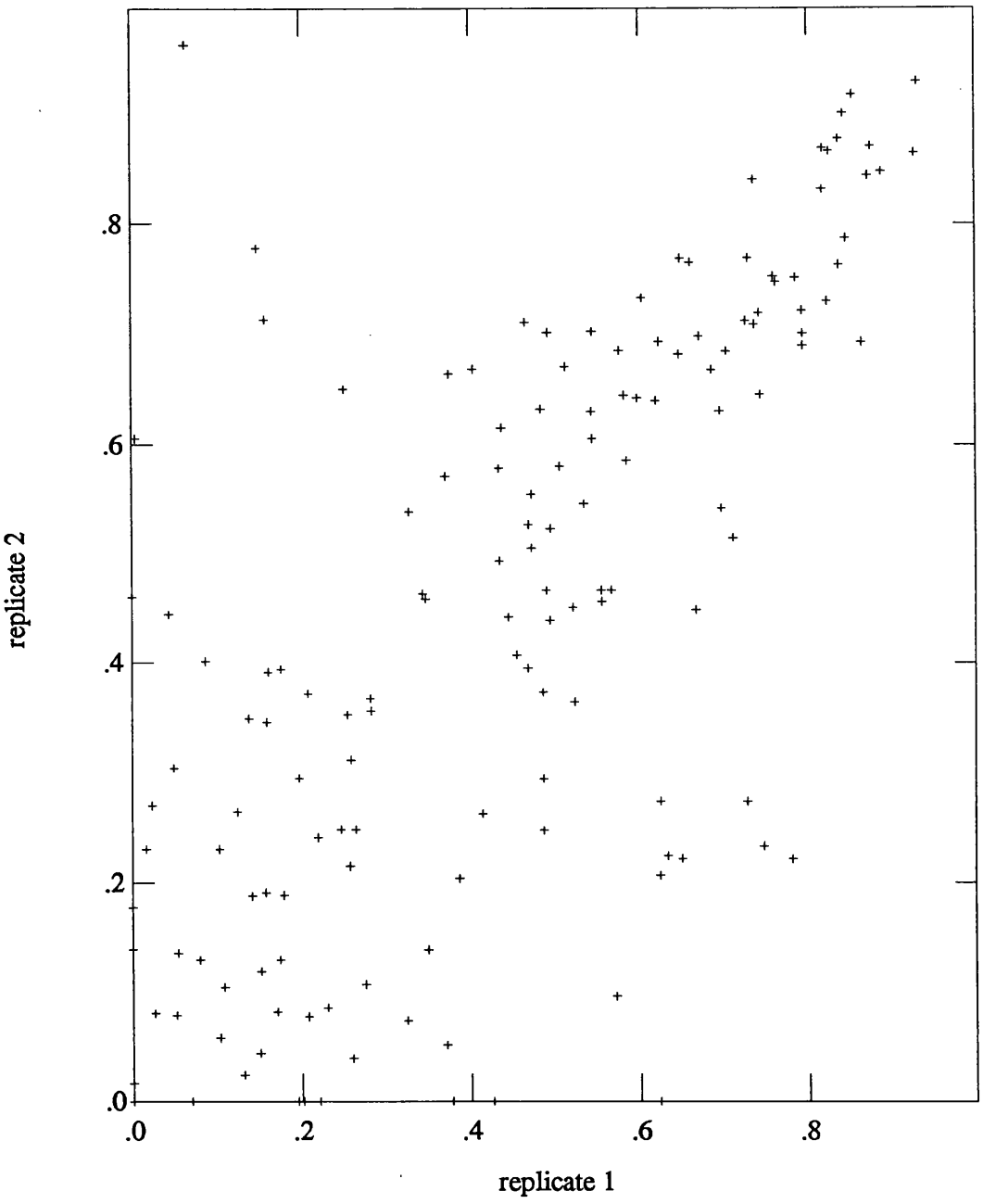


Fig. 5-8 A scatter plot of observed hyb ratio for 150 duplicate hybridisations. Each point represents two hybridisations with the same probe.

5.4.2 Normalisation of hybridisation signals

The absolute hybridisation signals are affected by factors that are very difficult to keep constant over a large number of hybridisations such as the specific activity of the probe and the length of exposure to phosphorstorage screens. Furthermore, the signal strengths vary according to the duplex yield for a given oligonucleotide. To aid the integration of data from different images and to suppress the effects of outliers in the analysis, the hybridisation signals within each image were normalised. There are essentially two methods by which the data can be normalised in a simple way. The signals can be divided by the average signal for each hybridisation. The disadvantage of this transformation is that extreme values, such as are caused occasionally by some hybridisation artefacts in which a few 'specks' of very high signal intensities exist, will greatly distort the data. An alternative normalisation is to rank all the signals in a given hybridisation. This way all hybridisations will have signals that fall into exactly the same range. Extreme signals, while present in the data, will not distort the distribution greatly. For this reason a ranking normalisation was chosen for the analysis of this set of hybridisation signals.

Hybridisation signals were ranked and replaced by their percentiles, so that the brightest signal in each image had intensity 1 and the weakest 0, with the other signals uniformly distributed between. Signals corresponding to "missing" spots were set to zero. One further normalisation is required which is carried out in two separate steps. Since the yield of PCR product in individual reactions varies significantly (by approximately a factor 10) the data analysis must take

into account the amount of DNA that is present on each spot on the filters.

The first step is to identify those spots with so little DNA that hybridisation signals cannot be relied upon. These are essentially those clones that are missing. Due to the sensitivity of the phosphorimaging system however, there are many cases where signals are recorded even though by eye one cannot observe any hybridisation signal, but these clones would be expected to have consistently low ranks in all hybridisations. Clones identified as having consistently low ranks were therefore likely to be missing on all filters and were excluded from the analysis.

In the second step, to normalise for the amount of DNA present on the remaining spots a second ranking was performed: If f_{ch} is the ranked signal for clone c in hybridisation h, then define g_{ch} to be the percentile of the signal f_{ch} of clone c amongst all hybridisations h. In other words f_{ch} is replaced by its percentile g_{ch} over h, so that clones with consistently high/low raw signals were down/up-weighted respectively.

The effectiveness of a ranking transformation depends on the quality of the data generated. On the whole, for good data no ranking would be required and a comparison of the signals relative to the mean would be efficient. Indeed, for good data significant information is lost through a ranking transformation and more data is required to achieve the same confidence as for non-ranked scores. For noisy data however, ranked scores provide greater robustness.

One of the consequences of a ranking transformation is that the sigmoidal distribution of signals, expected for a good hybridisation, is forced into a linear distribution. As a result it is necessary to introduce a cut-off below which the signals are set to zero. For an ideal hybridisation this cut-off should be set at the percentile of the expected

positive hybridisation rate. For noisy data however, such as this set of hybridisation signals, the cut-off may be best set at lower levels (see section on *Data error rates* below).

5.4.3 Data error rates

One of the key advantages of having extensive controls in each hybridisation is that a reasonably accurate estimate of the error rate can be made. In a statistical analysis such as this the error rates can be incorporated as a weighting factor. This means that as long as the error rate for a hybridisation is determined and the weighting procedure works correctly, even a poor result should be able to contribute some information to the overall analysis.

The information contributed by a hybridisation may be characterised by three probabilities: p, the overall acceptance threshold (i.e. so only clones with normalised scores in the upper fraction p in that hybridisation are considered as 'positive' for that experiment), q: the true fraction of positive clones for the probe, and α: the true positive rate, i.e. the probability that a genuine positive will be in the top fraction p. Data from 311 hybridisations were filtered so that all experiments where the proportion of expected positives (q) was less than 0.01 or α was less than 0.3 were omitted. For the 148 remaining hybridisations (comprising 110 different oligonucleotide probes) it was found that $q = 0.041$. The true positive rate α was computed for different values of p, and the resulting power curve plotted. This plot is shown in figure 5-9. In particular, for $p = 0.25$ the true positive rate $\alpha = 0.65$; i.e. on average 65% of the expected true positives for a probe will be in the top 25% of its normalised scores, and at $p = 0.1$, $\alpha = 0.5$.

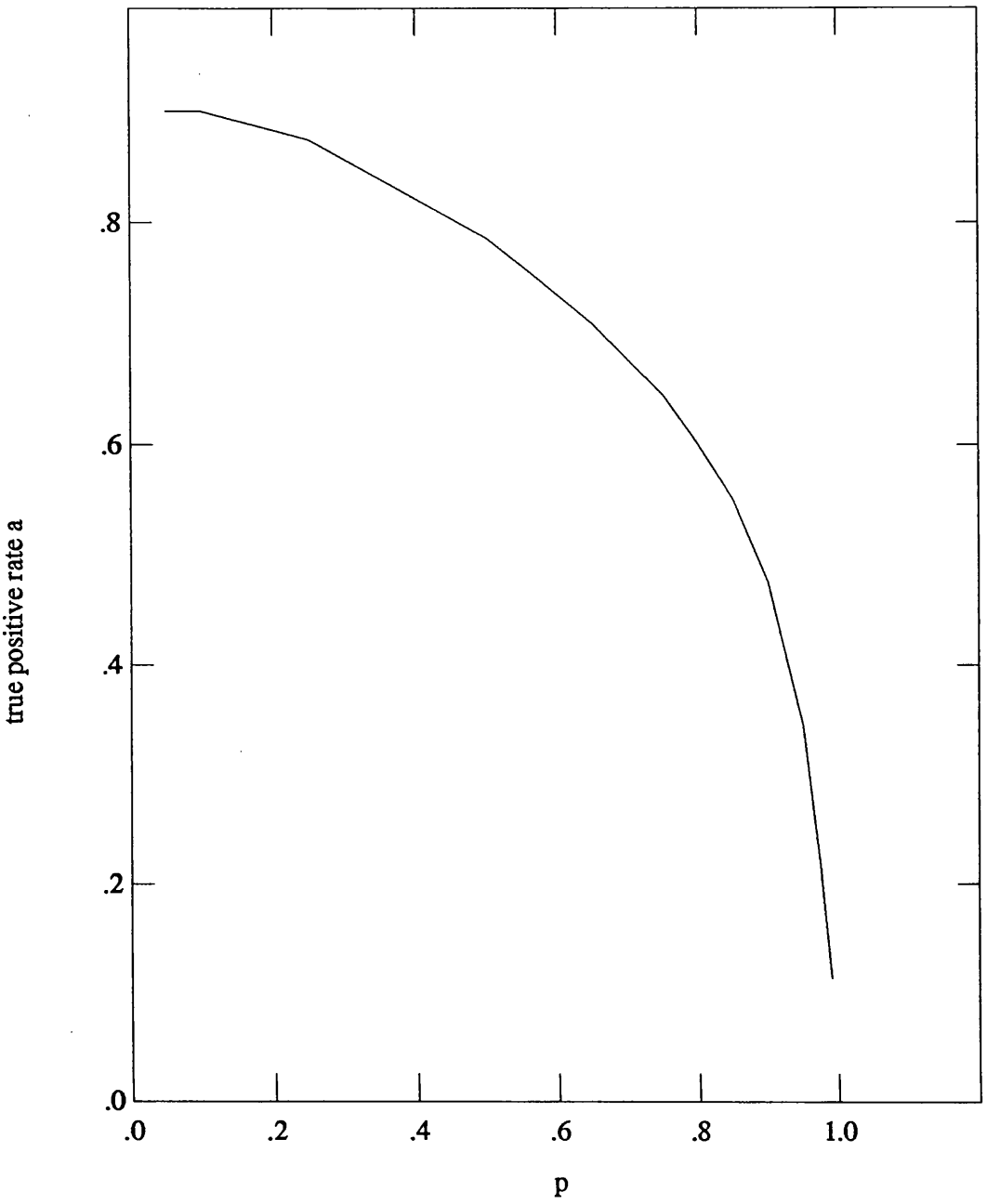


Fig. 5-9 A plot of the averaged observed true positive rate alpha plotted as a function of the positive threshold p. The data is averaged over 146 hybridisations.

For ideal data the cut-off should be set so that p = q, since then the acceptance threshold is set to equal the expected positive rate. At p = q (i.e. 0.041) α = 0.25 giving some indication of how noisy the data are. In other words only 25% of the expected positives (q) are in the top p fraction of the signals. Given these control data it is possible to determine the optimal acceptance threshold for which the ratio of signal to noise is maximised. For the purposes of matching fingerprints of clones it is necessary to perform pairwise comparisons to obtain similarity scores (S_{ab}) between clone *a* and clone *b* of the form of

$$S_{(ab)} = \sum_{h} g_{(ah)}\, g_{(bh)}\, w_{(h)}$$

where $g_{(ah)}$ and $g_{(bh)}$ are the hybridisation scores for clones *a* and *b* respectively for hybridisation *h* and $w_{(h)}$ is a weighting factor for that hybridisation. Since the pairwise comparison is a function of the product of the hybridisation scores the contribution of error (*e*) to the comparisons will proportional to e^2. This increases the tolerance of the system to error.

On the face of it, the error rates seem cripplingly high, since in order to obtain a true positive rate of 65% the false positive rate is at approximately 20%. Since the true positive rate is only 4.1%, this means that there are five times as many false positives as true positives. For the purpose of the data analysis it is not of great importance whether the error rate is determined largely by false positive or false negative data, since it is the overlap of the two distributions that affects the information content of the data. Experimentally however, it is informative to determine the source of errors, since experimental modifications to reduce error rates can be specifically designed for false positive and false negative data. Detailed analyses of individual hybridisations seem to indicate that the major source of error are the expected positive clones that give weak signals (i.e. the false negative

rate). Although there are expected negative clones with signals equal to the expected positives, their numbers are low (see above).

The cause of the very high false negative rate is not clear and still requires further experimental work. One possibility is the formation of secondary structures in the target DNA that prevent access of the probe molecule to the target sites. Secondary structure formation may be particularly high in the conditions of the oligonucleotide hybridisation (1 M Na^+ at 5°C) so it might be advantageous to set the hybridisation up initially in conditions that reduce secondary structure formations, such as high temperature or high pH. After a short time the conditions can be changed to allow duplex formation by dropping the temperature / pH. Another possibility might be to reduce the average length of the target molecules by chemical degradation or sonication, thus reducing the potential for secondary structure formation in each target molecule. Both of these approaches have to be tested experimentally.

Another important finding from this analysis is that under the restrictions imposed on the data (i.e. $q > 0.01$ and $\alpha > 0.3$) only 110 oligonucleotides (57%) contributed any information. The implication is that under the conditions used to generate this dataset approximately twice as many hybridisations as calculated theoretically will be required to generate sufficient information to perform a meaningful fingerprinting experiment. There are two factors that determine whether a hybridisation contributes information under the scheme described here. 1) the number of expected positive clones with a given oligonucleotide 2) the number of true positive clones found in the top 25% of ranked scores. The first is a function of the control clones available, whereas the second is a function of the specific oligonucleotide hybridisation characteristics.

5.4.4 Comparison of observed and expected fingerprints

Given the hybridisation data on the control clones, it is possible to compare the noisy experimental fingerprints with the ones expected from the known sequences of the clones. One way to assess the similarity and to get an idea how good the fingerprint is at distinguishing different clones, is to test how well the theoretical fingerprint of a given clone identifies the experimental fingerprint of the same clone amongst all other fingerprints.

Similarity scores were computed between the observed and the expected fingerprints for every pair of clones, the object being to estimate the probability that a ideal sequence fingerprint can identify a noisy version of itself against a background of other fingerprints. The idealised fingerprint for clone c in hybridisation h is

g'_{ch} = 1 if clone c contains the probe used in hyb h, 0 otherwise

so for perfect experimental data $g_{ch} = g'_{ch}$. Then the comparison between the expected fingerprint for clone c and that observed for clone d is

$$S_{cd} = \sum_h g'_{ch} g_{dh}$$

An acceptance threshold T was set so that those clones d whose S_{cd} scores were in the top $T\%$ of neighbours of c were considered as similar to c. Out of 2000 control clones spotted onto filters, sequence information was available for 1348. Clones with no expected matches with the probe set, or which were deemed as missing on the filters were excluded from the analysis, leaving 1008 clones.

The success rate - defined as the probability that a clone's expected fingerprint matches with that observed - is a function of (a) the number

Q of probes expected positive for the clone, and (b) the true positive rate, which is set by the acceptance threshold T.

Figure 5-10 shows the success rate as a function of Q when $p = 0.25$ and $T = 1\%$ (which for this dataset equates to the top 10 neighbours). The graph shows that provided a clone has more than about 5 expected positive probes the success rate is over 50%, and generally increases with Q. If a less stringent cut-off is used, say $T = 5\%$, then three expected positives are required to obtain a success rate of 50%. The plot also indicates the effect of increased numbers of hybridisations on the probability of a clone identifying itself correctly, since more hybridisations will lead to more expected positive probes per clone. This data suggests that for the small data set that the control clones represent, an increase of a factor of two in the number of hybridisations (i.e. from an average number of 8 positive probes per clone to 16) would yield a success rate close to 100%. This success rate however, is based on finding a clone in the top 1% of all its neighbours, a factor that is dependent on the total number of clones in the analysis. The same criterion applied to the cDNAs (of which there are 32,256 in this data set) would mean that each theoretical fingerprint, such as that of a sequenced gene, would identify itself in the library along with around 300 other clones.

Figure 5-11 shows a plot of the probability that two clones sharing Q expected positive probes would be considered neighbours using $T = 1\%$ (i.e. the probability that the expected fingerprint of one clone matches that observed of another, given that the expected fingerprints contain Q probes in common), and again shows an increase with Q. The data for the higher values of Q are affected by large sampling variation due to a small sample size.

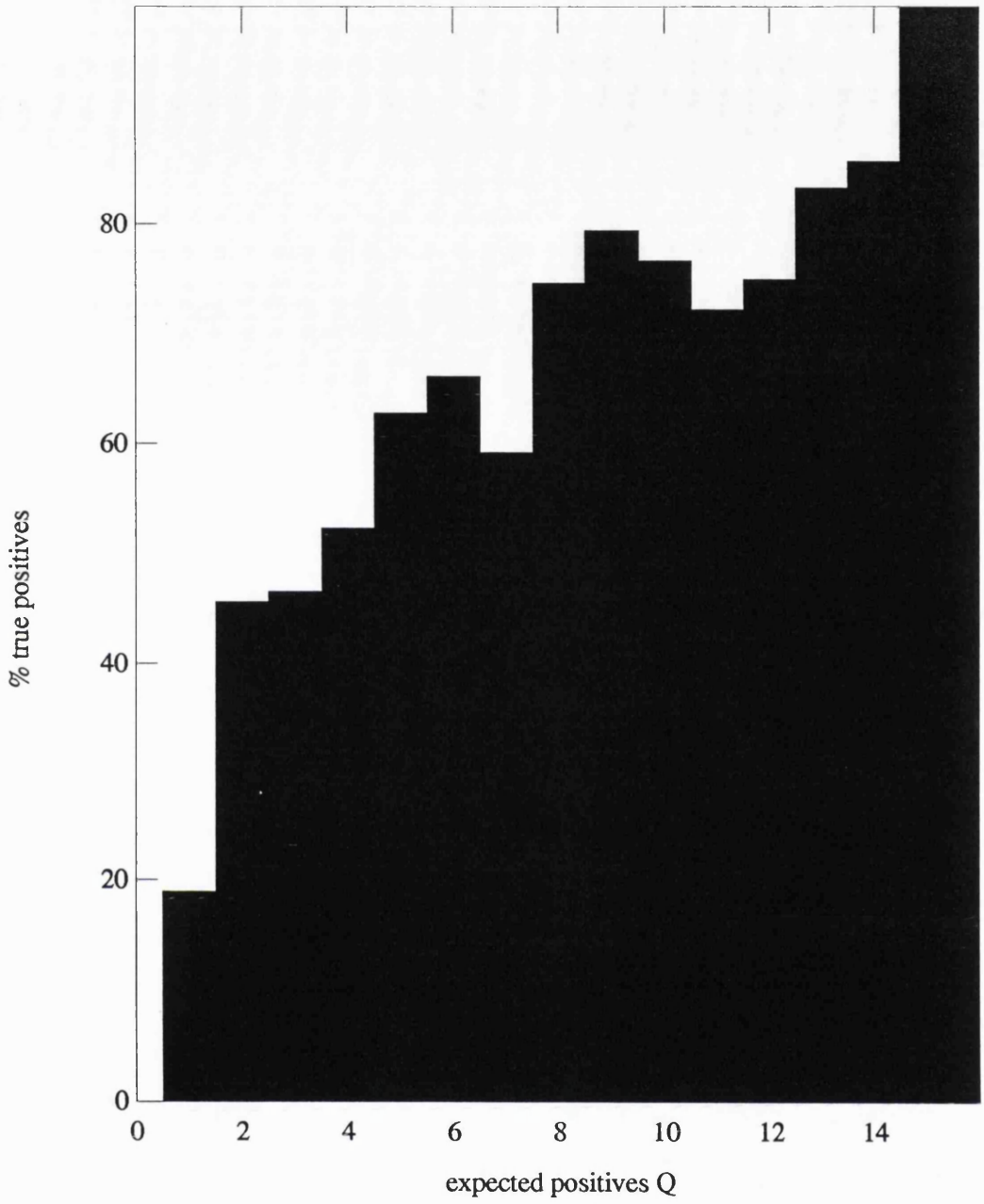


Fig. 5-10 The observed probability that an ideal clone fingerprint will identify a noisy version of itself, as a function of the number of expected positives, Q. Threshold cutoff, T = 1%, p = 0.25.

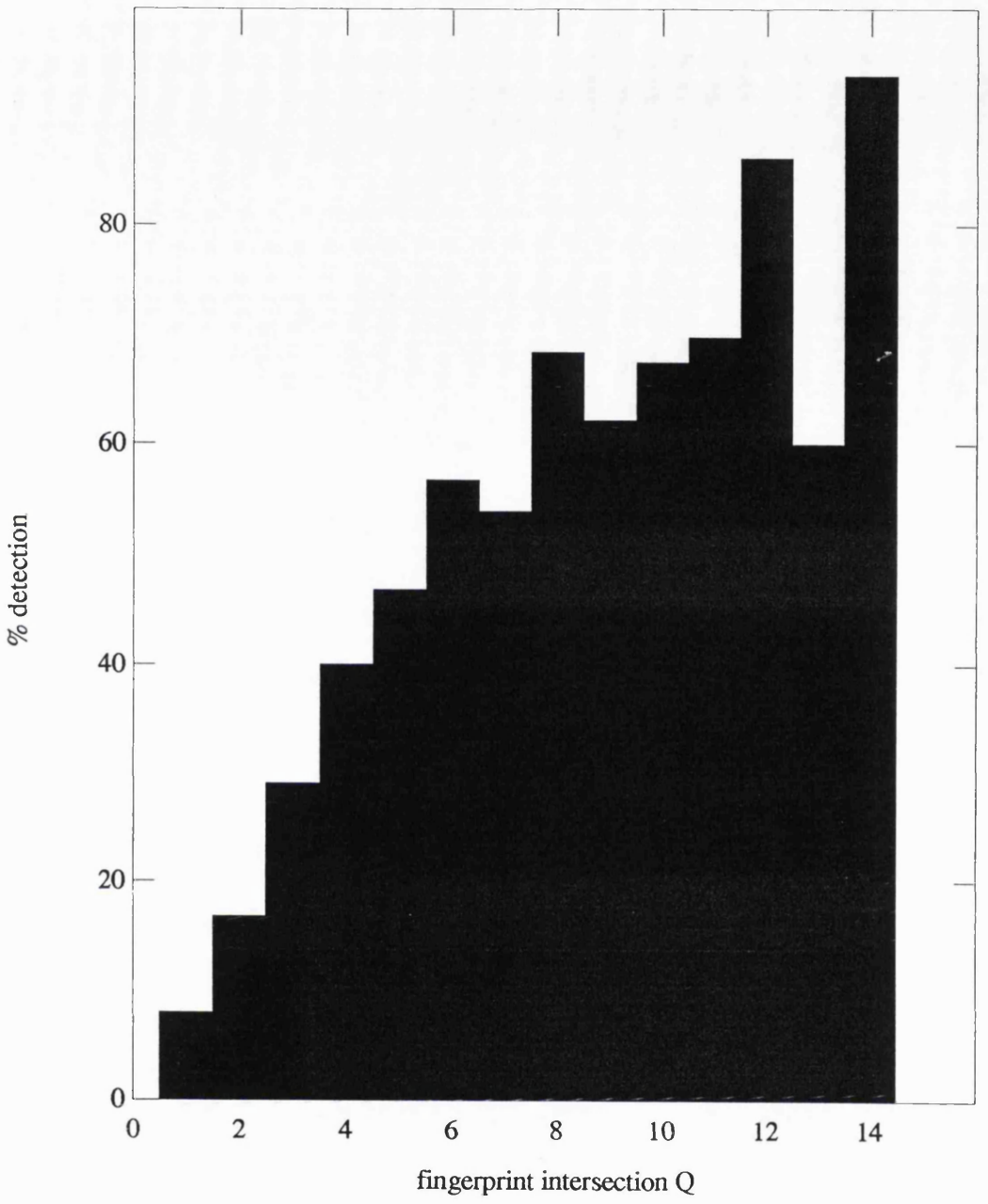


Fig. 5-11 The observed probability that an ideal clone will identify a different clone sharing *Q* expected positives. Threshold cutoff, T = 1%, *p* = 0.25.

The analysis of the control clones presented in the above discussion gives an indication as to the quality of the hybridisation data. The control clones are suitable for the assessment of the hybridisation characteristics of many of the oligonucleotides used. Those oligonucleotides which have no expected positive control clones, cannot be assessed in this case and are difficult to handle. Although useful tests can be performed using the hybridisation data of the control clones, such as determining the error rate and the ability of a fingerprint to identify its true sequence, the data cannot easily be used in predicting the success of the cDNA analysis. Any primary cDNA library is expected to contain, in addition to unique clones, clusters of more or less identical sequences that vary in size between two and hundreds of members, due to the highly variable representation of transcripts according to their expression levels. cDNA libraries that have been generated using oligo-dT primed first strand synthesis will contain clones in which all cluster members share a common 3' end. An oligonucleotide fingerprinting experiment would therefore be expected to generate many non overlapping clusters of fingerprints that share a large number of oligonucleotides. These should be identified by a clustering method. The clustering behaviour of the hybridisation data on the control clones cannot be tested easily since the clones form a contiguous set of overlaps that span approximately 60 kb of the human MHC class II region (Beck et al., 1992).

6. Analysis of cDNA oligonucleotide hybridisation data

In the previous chapter the hybridisation data were discussed in terms of the analysis of the control clones. Much of the software used for this analysis can also be used for analysing the cDNA hybridisation data. As discussed in the previous chapter the fingerprint analysis cannot be performed in the same way since the sequence characteristics and redundancy of the clones is entirely different.

The determination of the hybridisation signals was performed in an identical way to that described for the control clones. That is, for each hybridisation the signals were ranked and then subsequently re-ranked for each clone across all hybridisations. One of the very important aspects of developing automated systems for the analysis of large amounts of data is to check exhaustively all the steps in the analysis. Not only is it important to ensure that each step performs the expected task but it is also vital to have some mechanism whereby the final results of the analysis can be checked against the raw data. For projects such as this, which has generated data of 6,400,000 probe / target interactions, manual data checking is non trivial and can only realistically be performed for a small subset of the data.

In order to compare the automatically generated hybridisation scores with the raw data a small set of clones was scored manually for a subset of oligonucleotides. As part of the initial characterisation of the human foetal brain cDNA library, a glyceraldehyde 3-phosphate dehydrogenase (GAPDH) probe was hybridised to all clones. The clones positively identified in this hybridisation were used as a subset to analyse manually. Using the GCG findpattern command, all the oligonucleotides contained in the human GAPDH gene (as stored in the Genbank

database) were identified. The hybridisation signals for all the GAPDH clones were scored manually on a scale of 0 - 3 for those oligonucleotides expected to be positive in GAPDH mRNA sequence. In total 53 clones were scored across 36 hybridisations. Figure 6-1 shows both the manually scored hybridisation signals and those generated automatically by the software described above, displayed using a package written by Richard Mott. The figure shows that the scores are in broad agreement. In most of the instances in which there is a discrepancy the automated analysis proved to be more accurate. This was apparent when the hybridisation signal for a spot was checked quantitatively in terms of the pixel value, rather than visually in the image. This is presumably due to the fact that the 16 bit hybridisation data contains far more information than can be displayed on the monitor of a workstation.

There are two main functions that an oligonucleotide fingerprint should be able to fulfil. Firstly, it should be possible to cluster all the clones into groups that contain the same fingerprint, and therefore the same sequence. It is probable however, that many clones will not fall into any clusters since they are unique in the library. Secondly, it should be possible to compare the oligonucleotide fingerprints to database sequences and thereby identify clones matching database entries. Error in the data will affect these two functions differently depending on the source of error. If the hybridisation signals in many cases are not sequence specific, as shown by a poor discrimination between expected positive and negative control clones, then matching the fingerprints to database sequences will not be possible. However, provided that the hybridisations are reproducible, albeit not specific in terms of classical Watson-Crick base pairing, then it should still be possible to perform clustering of the clones.

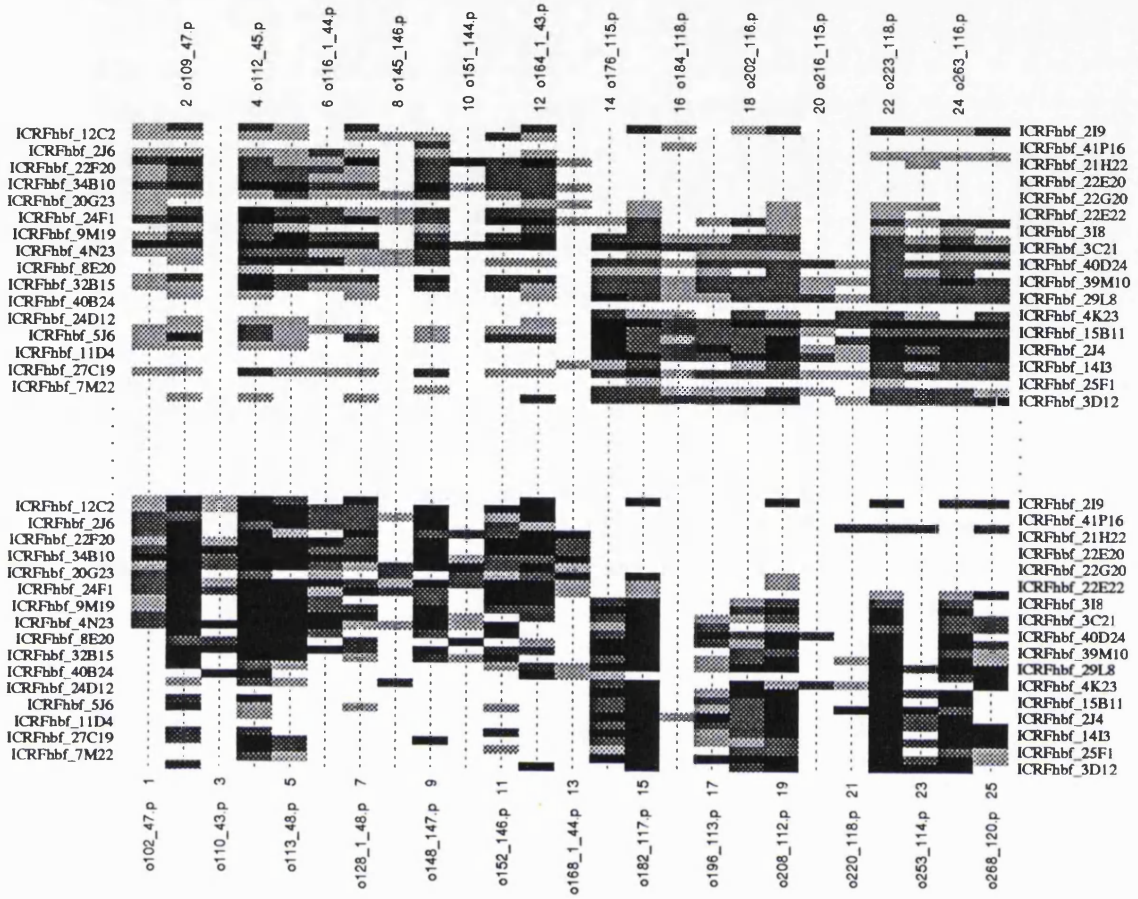


Fig. 6-1 A graphical representation of hybridisation signals of GAPDH cDNA clones (listed vertically along the sides) with expected positive oligonucleotide probes (listed horizontally along top and bottom). Top: manually scored hybridisation signals. Bottom: hybridisation signals calculated by automated image analysis.

In order to cluster the cDNA clones a slightly different analysis was performed than with the control clones. Clustering can be performed in a variety of ways. One of the common algorithms is based on the minimal spanning tree theory, of which there are many implementations. The principal of such algorithms is to find the shortest path connecting all clones in a set. Such an algorithm was applied to the hybridisation data initially, however, few very large clusters resulted, of which subsets seemed correct, but whose overall discrimination between different fingerprints was too poor. These first attempts immediately highlighted one of the limitations of using a minimal spanning tree algorithm, which is if there are two separate clusters that have a single clone in common, then the two clusters will be connected under a standard minimal spanning tree method. With noisy data, in which false connections between clones exist, clusters quickly merge together.

To overcome these difficulties a different more stringent form of clustering was developed and implemented by Richard Mott. For each clone its top 1% of neighbours were determined, using the same pairwise comparison formula as used in the analysis of the control clones. A similarity score is obtained for each pair of clones (*a* and *b*) across all hybridisations *h*

$$S_{(ab)} = \sum_{h} g_{(ah)}\, g_{(bh)}\, w_{(h)}$$

including a weighting factor *w* for each hybridisation. For each clone, all those clones are selected that have a similarity score *S* in the top 1%. In order then to form stringent clusters out of these pre-selected groups, all those clones were selected that show reciprocal relatedness. Consider the top 1% of neighbours for clone *a* (a_1 a_2 a_3 a_4 ... a_n) then select all clones for which *a* is in its top 1% of neighbours. Figure 6-2 shows a histogram of the distribution of reciprocal neighbourhoods for 16,128

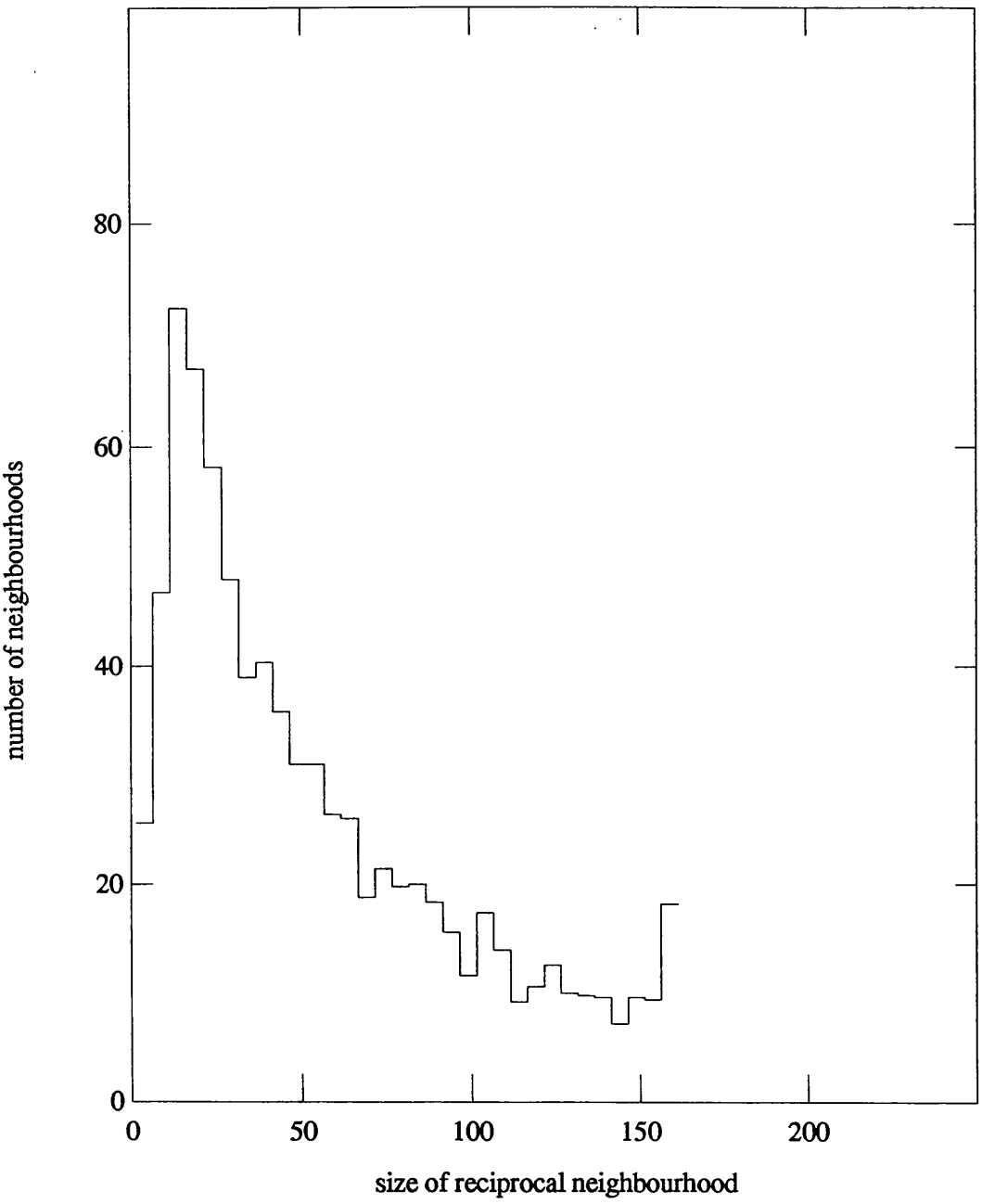


Fig. 6-2 A histogram of the distribution of different size reciprocal neighbourhoods calculated for 16,128 cDNA clones, taking a 1% cutoff.

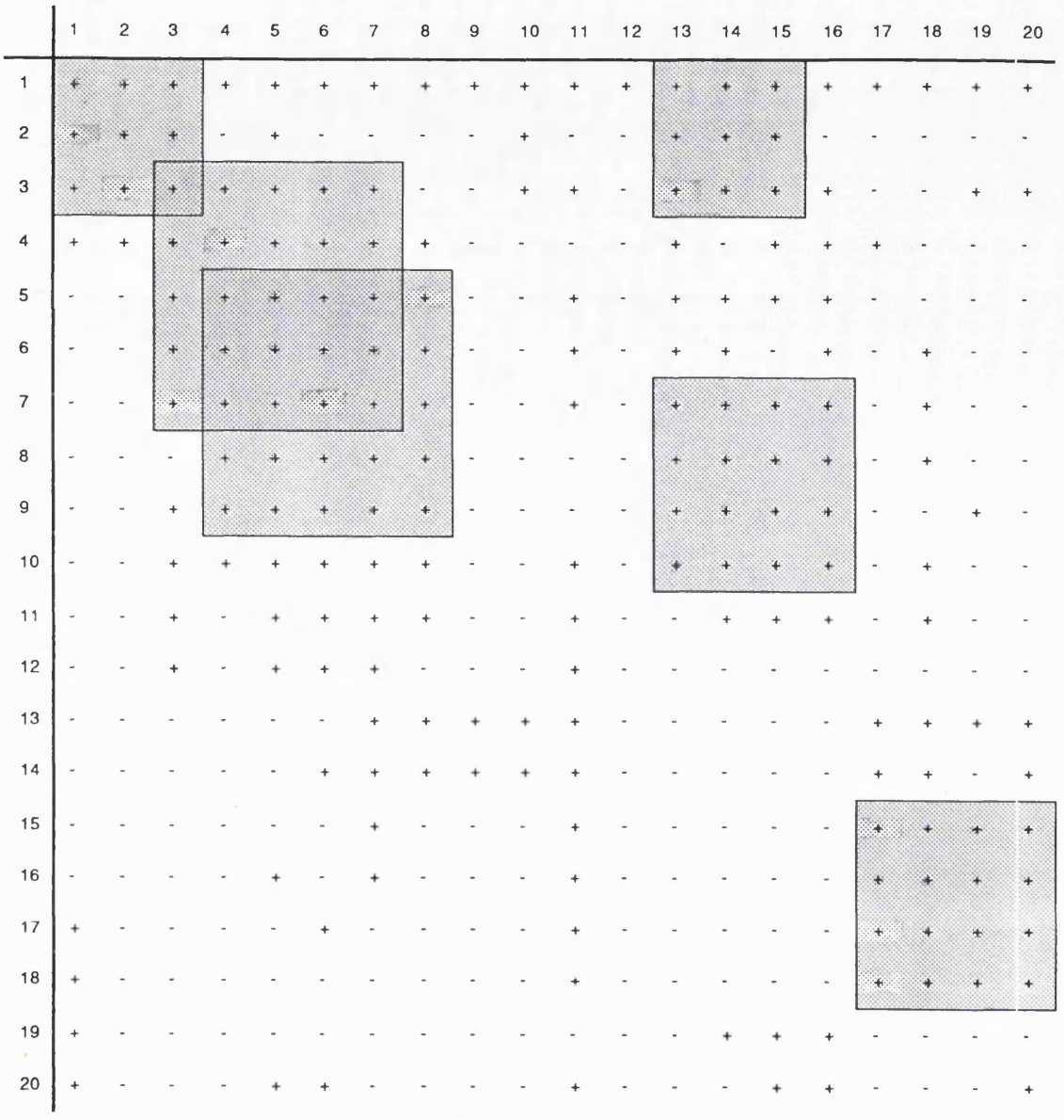


Fig. 6-3 A schematic illustration of the clustering algorithm used on the cDNA data to form stringent clusters. The matrix shows clones for which other clones lie in their top 1% neighbourhood (+). Clusters are formed from the largest possible set of clones such that all are reciprocally related.

cDNA clones. By this method it was found that 75% of clones did not share their top 1% neighbourhood reciprocally with any other clone. That is, 75% of clones remained as singletons. Data from agarose gels indicates that only approximately 75% of cDNAs gave products during the waterbath PCR amplification. The 'missing' clones would be expected not to hybridise with any oligonucleotides and therefore contribute to the singletons. The real proportion of singletons is therefore likely to be closer to 50%, a figure consistant with data that suggest between 26 - 50% of sequences transcribed in mammalian brain tissue are expressed at a level of less than 1 in 100,000 (Milner and Sutcliffe, 1988). The most common neighbourhood size for the remaing data was 10. Figure 6-2 also shows that there is an increase in number of reciprocal neighbourhoods for the maximum size, in this case 161. This is an artefact caused by a subset of aberrant clones that have hybridised with almost all oligonucleotide probes and therefore lie in each other's neighbourhoods. In order to compare the neighbourhood distribution with that expected for random hybridisation signals, the hybridisation scores for each clone were permuted and then processed through exactly the same analysis. Only three neighbourhood sizes were obtained: 59.2% fell into neighbourhoods of 5, 40% into neighbourhoods of 6 and 0.8% into neighbourhoods of 7. This result indicates that the behaviour of the real hybridisation data is far from random under this analysis scheme.

In the next round of selection starting with each member of the top 1% neighbours, those clones are retained that are in the top 1% of neighbours for that member. Finally, select the largest group of clones for which all members are in each other's top 1% of neighbours. Figure 6-3 illustrates this approach in form of a matrix diagram. From a matrix in which + represents membership in the top 1% of neighbours, the largest block is selected in which all scores are +.

Figure 6-4 shows example clusters that were generated from the hybridisation data of 142 oligonucleotides. The figure shows clearly those oligonucleotides that are common to most clones in a cluster. For a small selection of clusters a single member was hybridised to all other clones and the positively hybridising clones compared to the cluster members. Of the clones in cluster 1155, 80% hybridised positively with clone ICRFhbf_21H10.

One of the consequences of generating such stringent clusters is the there are a significant number of clusters that differ only by one clone. These clusters should clearly be grouped together, since they overlap substantially. Just as noisy data has a tendency to create false connections between clusters when using a minimal spanning tree algorithm, the same noisy data will also tend to separate true clusters. In order to identify clusters with substantial overlap, a simulated annealing algorithm was used to generate clusters of clusters (super clusters). The algorithm was used in a previous mapping project and is described in Mott et al. (1993). The results of this are displayed in figure 6-5 and 6-6. Figure 6-5 shows a super cluster in detail, while figure 6-6 shows all the super clusters obtained from the hybridisation data. The fact that most of the data lies on the diagonal, suggests that the clustering has performed well in reducing the noisiness of the hybridisation data. The initial clustering produced 2711 clusters for a subset of 16,128 fingerprints. From these, 569 super clusters were found by the simulated annealing.

In order to compare clones with sequences contained in the Genbank database, a consensus fingerprint was calculated for each cluster. A theoretical fingerprint was generated for all 4,551 expressed human sequences extracted from a database (see *Selection of Oligonucleotides*) and for each consensus fingerprint those sequences with the top 10 similarity scores selected. At present the analysis has not identified any

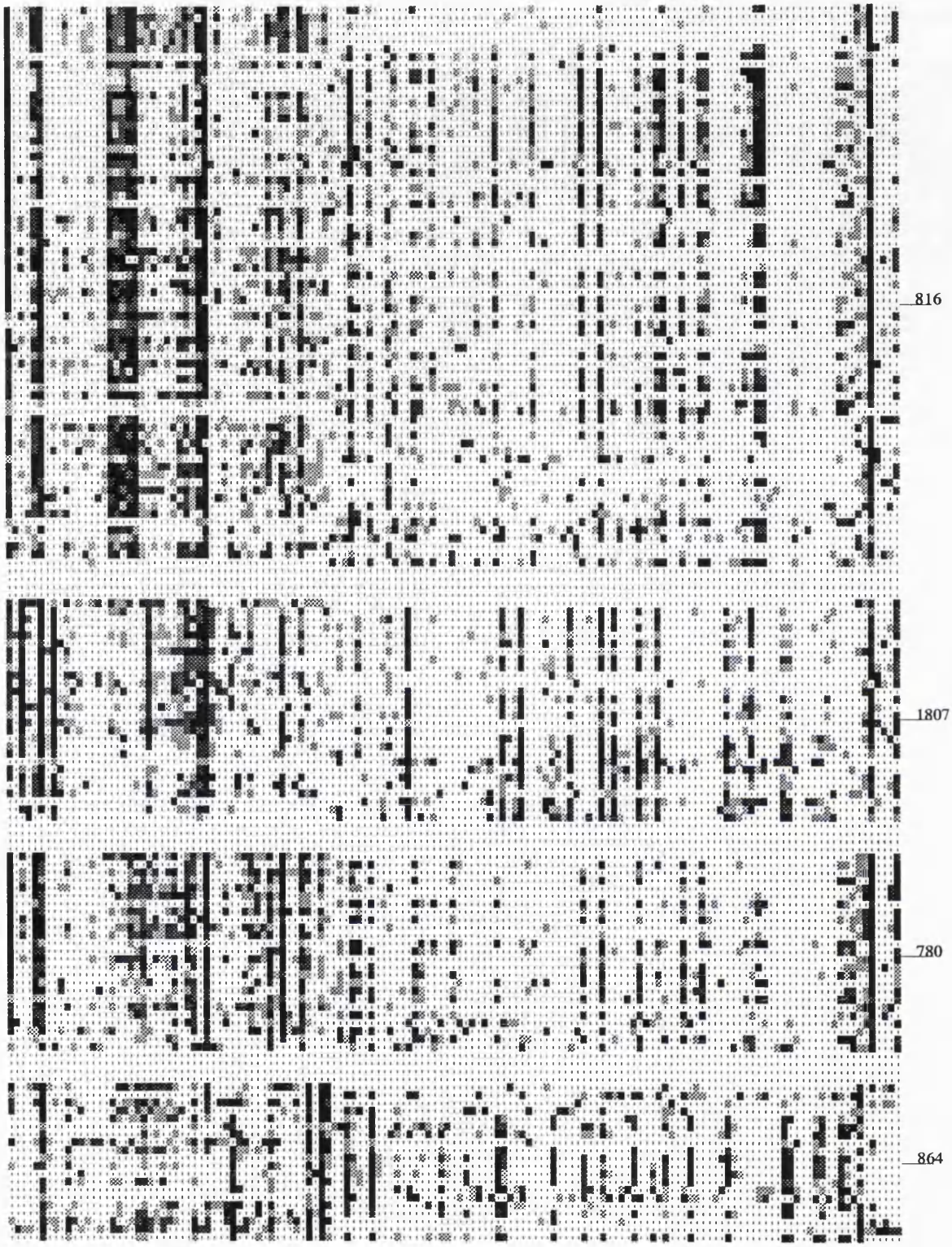


Fig. 6-4 Example clusters of cDNA clones. Clones are listed vertically and oligonucleotide probes are listed horizontally.

Fig. 6-5 A graphical representation of clones in a 'supercluster'.

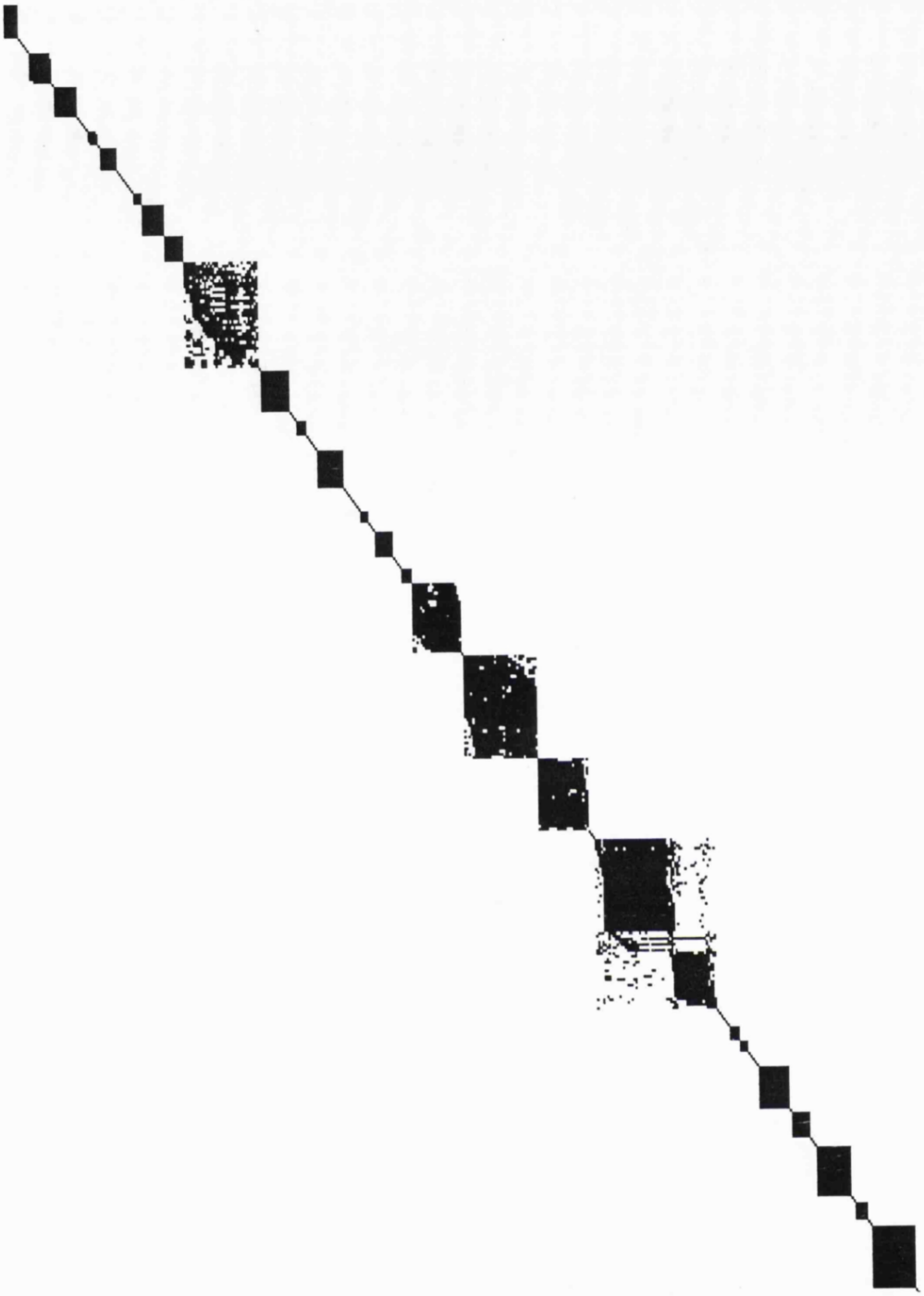


Fig. 6-6 An illustration of some of the superclusters obtained by applying a simulated annealing algorithm to initial clusters formed from the cDNA hybridisation fingerprints.

fingerprints with an overwhelming similarity to a database entry. The cluster (cluster 696) is shown in figure 6-7 (top) and its consensus aligned with database sequences is shown in figure 6-7 (bottom). In this case, a cluster of 13 fingerprints showed the best match with the human pyruvate kinase gene. Ten members of cluster 696 were sequenced from both the 3' and 5' ends. The sequence data were very difficult to read as there were many stops in the sequencing ladder. However, of six clones from which some sequence could be read, four were identical over the first 50 bp. No sequence homology was found between the short sequence read from the clones and the human pyruvate kinase gene. A perfect match, however, was found to an EST (accession number Z21515.GB_EST).

All the members of cluster 696 were amplified by PCR and the products digested with *Alu*I. The digests were run out on a 2.5% agarose gel and are shown in figure 6-8. The gel clearly shows that many bands are shared between most of the clones. It appears that the digest was incomplete since there are bands of varying intensities, the sum of the sizes of which add up to more than the size of the undigested clones. There seem to be two sets of bands that occur in many of the clones. The three largest bands, which are partial digest products, are present in twelve clones and three smaller bands are shared between ten clones. While the sequence data from these clones is inconclusive, the restriction pattern of the cluster members indicates that there are shared sequence features. At least two members of the cluster show few shared bands (clones 3B10 and 16P2, lanes 6 and 9 respectively). Although this might suggest that these clones are included falsely in the cluster, the fingerprints for these two clones are in good agreement with the other cluster members.

The analysis of the clustering output is far from complete, but does give some indication as to the quality of the data and its utility in

characterising cDNA libraries. It seems that at the present level and quality of data it is possible to identify and cluster clones that share a significant amount of sequence. So far, it has not been possible to use the hybridisation data to match database sequences to the oligonucleotide fingerprints. The reasons for this are not clear at present. It is possible however that the fingerprints are reproducible enough from clone to clone to allow clustering, but insufficiently sequence specific to allow database matching. Observations made by Southern et al. (1994) indicate that some oligonucleotide hybridisation interactions are not readily explicable by classic Watson / Crick base pairing rules. Analysis of the control clone data has suggested that the sequence specificity of the data is sufficient to identify the correct clones in the top 1% of its neighbours in 50% of cases, provided there are more than 5 positive oligos (see figure 5-10).

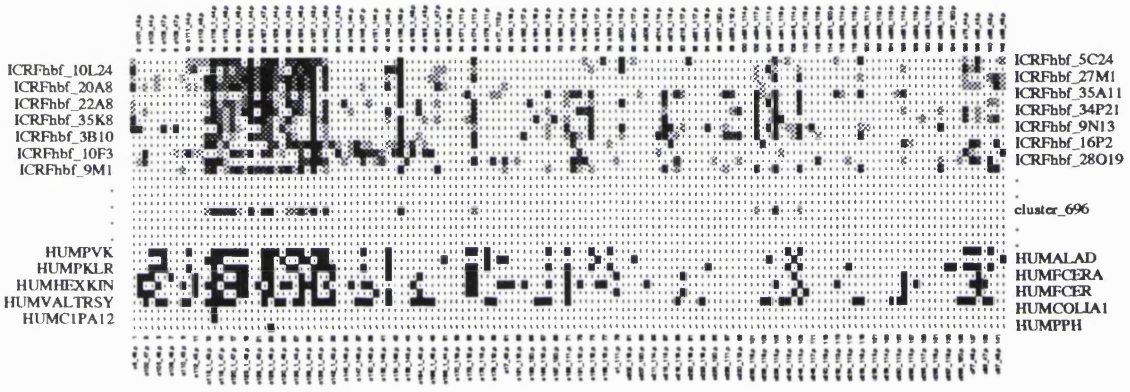


Fig. 6-7 A graphical represention cluster 696 (top) and its consensus fingerprint (middle) aligned with the ten best database matches (bottom).

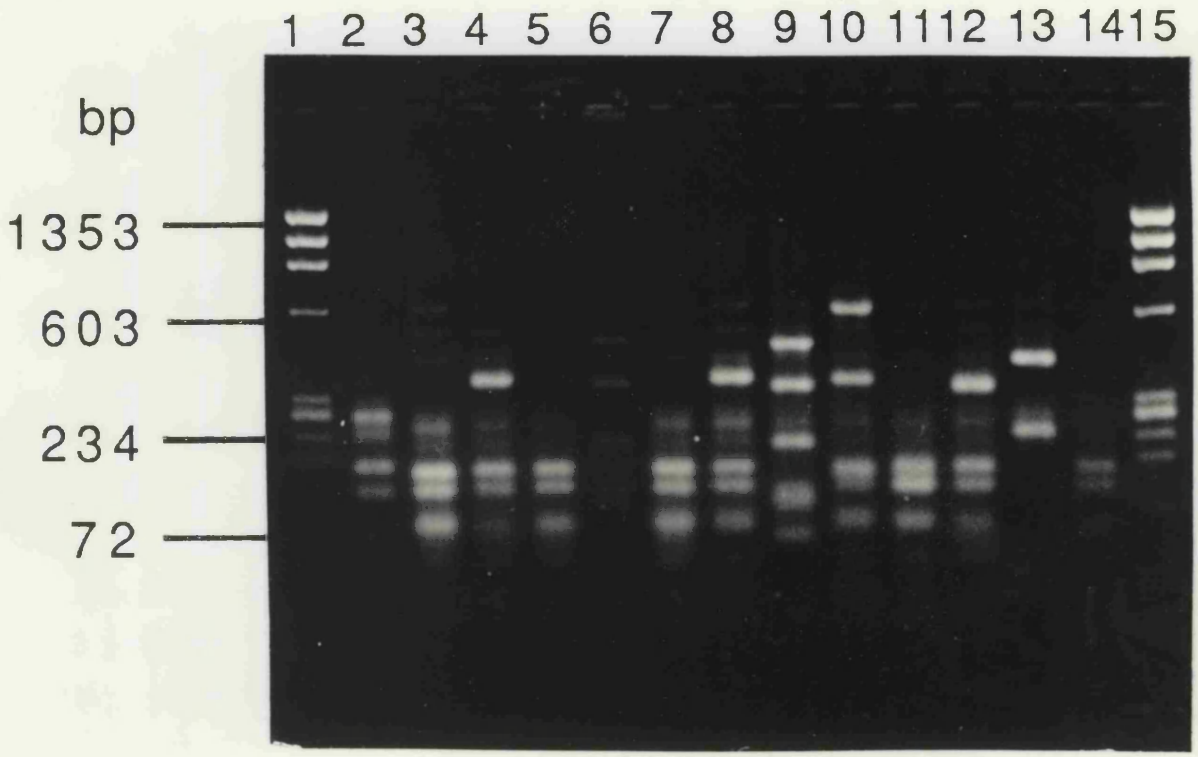


Fig. 6-8 A 2.5% agarose gel of 13 cDNA clones from cluster 696 digested with *Alu*I. Lanes: 1) *Hae*III digested ØX 174 DNA. 2) clone 28O19. 3) clone 9M1. 4) clone 10L24. 5) clone 20A8. 6) clone 3B10. 7) clone 27M1. 8) clone 22A8. 9) clone 16P2. 10) clone 5C24. 11) clone 35A11. 12) clone 9A13. 13) clone 10F3 14) clone 34P21. 15) *Hae*III digested ØX 174 DNA.

7 Conclusions and Prospects

The aim of this thesis was to test the feasibility of assembling a set of experimental and software tools that could be used in large scale oligonucleotide fingerprinting projects. Much of the foundations for such an approach have been layed in the course of this work. The work has involved the development of both experimental procedures and analytical software tools, designed to automate very large scale data analysis.

On the experimental side there has been the development of a large scale PCR system to allow the amplification of entire cDNA libraries. In the course of this work over 100,000 cDNA clones have been amplified with this system, although hybridisation data has so far only been generated on a subset of these. Also, there has been the adaptation of oligonucleotide hybridisations to a method that enables a single person to perform hybridisations on 30 high density clone arrays per day. That can either be one probe on 1,000,000 clones or 30 probes on 36,000 clones per day. A significant amount of automation had to be developed to allow the reliable handling of tens of thousands of clones clones from picking of the colonies, through amplification of the cDNA inserts, to the arraying of PCR products on high density arrays. Much of this work was carried out with a team of both scientists and engineers and was based on the further development of existing technology. Finally, a large amount of work was invested in generating software tools required for the analysis of millions of probe / target ineractions. Indeed, at least 50% of the time spent during the course of this thesis was devoted to the informatics aspects of a large scale molecular biology project.

Incomplete attempts have been made in this work to analyse the hybridisation data generated using the set of tools that were developed.

Only during the very end of the study were significant strides achieved toward the meaningful analysis of the noisy data. An invaluable contribution to the analysis was provided by the control clone data, as it gave an indication of the quality of the data and therefore helped in the choice of suitable strategies. Even noisy data can contirbute useful information provided its significance can be assessed correctly.

From the data generated to date and its preliminary analysis it is clear that while useful information can be obtained, considerable increase in both quality and quantity of data will be required for the approach of oligonucleotide fingerprinting to fulfill its considerable potential. Clustering of homologous sequences appears to be possible with the present dataset, even if some noise still remains in these clusters. Matching the fingerprints to known sequences has not yet been possible and will require the generation of much more data, or data of much higher quality, preferably both.

It seems that an important lesson from this work has been to derive a set of oligonucleotide sequences that function well in a fingerprinting approach. Due to the unpredictability of the hybridisation behaviour of short oligonucleotides it appears that an empirically determined and tested set of sequences must be derived from controlled test hybridisations that can accurately evaluate the hybridisation characteristics of oligonucleotide probes under the exact experimental conditions used.

One of the less demanding applications for which an oligonucleotide fingerprinting approach can be used is as a preselection step in large scale sequencing projects. Fingerprints of the level generated in this thesis can be used to identify clones with high sequence homologies. This information can then be used to sequence initially only those clones that are different, thus reducing considerably the redundancy in sequence generation. Of course, this approach only becomes efficient

when a very large number of clones are to be sequenced, such as large scale 'tag'-sequencing and genomic sequencing projects.

For future oligonucleotide fingerprinting experiments, there are many pointers in this thesis to aspects that can and should still be improved upon. On the experimental side, PCR amplification and oligonucleotide hybridisations ought to be singled out as areas in which improvements to the quality of the hybridisation data can be most readily made. The reliability of PCR amplification could be improved by switching to a different kind of microtitre plate. The large thermal mass of the Q-plates will always be a hinderance in reliable amplification. The overall success rate and yield was in fact greater in the initial waterbath PCR trials using thin walled 96-well plates. A thin walled 384-well plate is currently under design. Previously the problem with a thin walled 384-well plate has been that only small wells could be moulded by the process of thermoforming and that these were unsuitable for clone storage purposes. For PCR purposes however, a small volume of around 15 - 20 µl per well would be sufficient.

The hybridisation data generated to date indicate that there is a very large variation in the hybridisation characteristics of the oligonucleotides used. Conditions used for hybridisations in this pilot study, were standardised for the sake of high throughput. A great deal of improvement in the quality of the hybridisation data is likely to be achievable by modifications to the hybridisation conditions. Hybridisation protocols could be adjusted for individual oligonucleotides according to their previous hybridisation characteristics. Probes with poor signal strength could be hybridised at a higher concentration and/or specific activity and could be washed less extensively. Conversely, oligonucleotides with good signal strengths but poor sequence specificity could be washed for longer periods or at slightly higher temperatures. A modification that should be tested for all

hybridisations is the use of tetra-alkyl ammonium salts to reduce the difference in contribution to the duplex stability between A-T and G-C base pairs and to increase the duplex yield of many oligonucleotides. A significant advance will also be achieved when non-radioactive detection systems can be used routinely. Recent results in the lab suggest that amplified fluorescence could provide an immediate alternative to radiolabelling (Maier et al., 1994b).

Further advances in the quantity of hybridisation data that can be generated will rely on continued development of the automated processes, especially the hybridisations, and on the development of miniaturised clone arrays. The concept of the 'clone chip' has been variously published (Southern et al., 1992; Fodor et al., 1993; Mirzabekov et al., 1994) and points the way to the future of large scale hybridisation approaches. Certainly, technological developments will continue to be the driving force in this area of molecular biology for some years to come.

Experience from this work has shown that a key area for further development lies in the informatics tools that are available for handling very large data sets. Not only are powerful computers and efficient programs required for the analysis of experimental data, but interfaces need to be generated through which biologists can access and interpret the output of complex statistical systems.

Bibliography

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., and et al (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. Science *252*, 1651-1656.

Adams, M.D., Dubnick, M., Kerlavage, A.R., Moreno, R., Kelley, J.M., Utterback, T.R., Nagle, J.W., Fields, C., and Venter, J.C. (1992). Sequence identification of 2,375 human brain genes. Nature *355*, 632-634.

Adams, M.D., Kerlavage, A.R., Fields, C., and Venter, J.C. (1993a). 3,400 new expressed sequence tags identify diversity of transcripts in human brain. Nature Genet. *4*, 256-267.

Adams, M.D., Soares, B.M., Kerlavage, A.R., Fields, C., and Venter, J.C. (1993b). Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. Nature Genet. *4*, 373-380.

Adams, M.D., Kerlavage, A.R., Kelley, J.M., Gocayne, J.D., Fields, C., Fraser, C.M., and Venter, J.C. (1994). A model for high-throughput automated DNA sequencing and analysis core facilities. Nature *368*, 474-475.

Alonso, S., Minty, A., Bourlet, Y., and Buckingham, M. (1986). Comparison of three actin coding sequences in the mouse; evolutionary relationship between the actin genes of warm blooded vertebrates. J. Mol. Evol. *23*, 11-22.

Aposhlan, H.V. and Kornberg, A. (1962). Enzymatic Synthesis of Deoxyribonucleic Acid. J. Biol. Chem. *237*, 519-525.

Bains, W. (1991). Hybridization Methods for DNA Sequencing. Genomics *11*, 294-301.

Bains, W. and Smith, G. (1988). A novel method for nucleic acid sequence determination. J. Theo. Biol *135*, 303-307.

Ballabio, A. (1993). The rise and fall of positional cloning?. Nature Genet. *3*, 277-279.

Barnes, W.M. (1994). PCR amplification of up to 35-kb DNA with high fidelity and high yield from λ bacteriophage templates. Proc. Natl. Acad. Sci. U. S. A. *91*, 2216-2220.

Beck, S., Kelly, A., Radley, E., Khurshid, F., Alderton, R.P., and Trowsdale, J. (1992). DNA Sequence Analysis of 66 kb of the Human MHC Class II Region Encoding a Cluster of Genes for Antigen Processing. J. Mol. Biol. *228*, 433-441.

Bellanne-Chantelot, C., Lacroix, B., Ougen, P., Billault, A., Beaufils, S., Bertrand, S., Georges, I., Glibert, F., Gros, I., Lucotte, G., and et al (1992). Mapping the whole human genome by fingerprinting yeast artificial chromosomes. Cell *70*, 1059-1068.

Boguski, M.S., Lowe, T.M.J., and Tolstoshev, C.M. (1993). dbEST - database for "expressed sequence tags". Nature Genet. *4*, 332-333.

Botstein, D., White, R.L., Skolnick, M., and Davis, R.W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. *32*, 314-331.

Breslauer, K.J., Frank, R., Blöcker, H., and Marky, L.A. (1986). Predicting DNA duplex stability from the base sequence. Proc. Natl. Acad. Sci. U. S. A. *83*, 3746-3750.

Burke, D.T., Carle, G., and Olson, M.V. (1987). Cloning of Large Segments of Exogenous DNA into Yeast by Means of Artificial Chromosome Vectors. Science *236*, 806-812.

Carpenter, E.M., Goddard, J.M., Chisaka, O., Manley, N.R., and Capecchi, M. (1993). Loss of hox-a1 (hox-1.6) function results in the reorganization of the murine hindbrain. Development. *118*, 1063-1075.

Case-Green, S.C. and Southern, E.M. (1994). Studies on the base-pairing properties of deoxyinosine by solid-phase hybridization to oligonucleotides. Nucl. Acids Res. *22*, 131-136.

Chaudhari, N. and Hahn, W.E. (1983). Genetic Expression in the Developing Brain. Science *220*, 924-928.

Chien, A., Edgar, D.B., and Trela, J.M. (1976). Deoxyribonucleic Acid Polymerase from the Extreme Thermophile *Thermus aquaticus*. J. Bact. *127*, 1550-1557.

Chikaraishi, D.M., Deeb, S.S., and Sueoka, N. (1978). Sequence Complexitiy of Nuclear RNAs in Adult Rat Tissues. Cell *13*, 111-120.

Chirgwin, J.M., Przybyla, A.E., MacDonald, R.J., and Rutter, W.J. (1979). Isolation of Biologically Active Ribonucleic Acid from Sources Enriched in Ribonuclease. Biochemistry *18*, 5294-5299.

Chumakov, I., Rigault, P., Guillou, S., Ougen, P., Billaut, A., Guasconi, G., Gervy, P., LeGall, I., Soularue, P., Grinas, L., and et al (1992). Continuum of overlapping clones spanning the entire human chromosome 21q. Nature *359*, 380-387.

Church, G.M. and Gilbert, W. (1984). Genomic sequencing. Proc. Natl. Acad. Sci. U. S. A. *81*, 1991-1995.

Cohen, D., Chumakov, I., and Weissenbach, J. (1993). A first-generation physical map of the human genome. Nature *366*, 698-701.

Cole, C.G., Goodfellow, P.N., Bobrow, M., and Bentley, D.R. (1991). Generation of Novel Sequence Tagged Sites (STSs) from Discrete Chromosomal Regions Using *Alu*-PCR. Genomics *10*, 816-826.

Collins, F.S. (1992). Cystic fibrosis: molecular biology and therapeutic implications. Science *256*, 774-779.

Condie, B.G. and Capecchi, M. (1993). Mice homozygous for a targeted disruption of hoxd-3 (hox-4.1) exhibit anterior transformations of the first and second cervical-vertebrae, the atlas and the axis. Development. *119*, 579-595.

Coulson, A.R., Sulston, J., Brenner, S., and Karn, J. (1986). Towards a physical map of the nematode *Caenorhabditis elegans*. Proc. Natl. Acad. Sci. U. S. A. *83*, 7821

Donehower, L.A., Harvey, M., Slagle, L., McArthur, M., Montgomery, C.A., Butel, J.S., and Bradley, A. (1992). Mice deficient for p53 are developmentally normal but susceptible to spontaneous tumours. Nature *356*, 215-221.

Dower, W.J., Miller, J.F., and Ragsdale, C.W. (1988). High efficiency transformation of *E.coli* by high voltage electroporation. Nucl. Acids Res. *16*, 6127-6145.

Drmanac, R., Labat, I., Brukner, I., and Crkvenjakov, R. (1989). Sequencing of megabase plus DNA by hybridization: theory of the method. Genomics. *4*, 114-128.

Drmanac, R., Lennon, G.G., Drmanac, S., Labat, I., Crkvenjakov, R., and Lehrach, H. (1990a). Partial Sequencing by Oligo-Hybridization: Concept and Applications in Genome Analysis. In The First International Conference on Electrophoresis, Supercomputing, and the Human Genome. C. Cantor and H.A. Lim, eds. (World Scientific), pp. 60-75.

Drmanac, R., Strezoska, Z., Labat, I., Drmanac, S., and Crkvenjakov, R. (1990b). Reliable hybridization of oligonucleotides as short as six nucleotides. DNA Cell Biol. *9*, 527-534.

Drmanac, R., Drmanac, S., Strezoska, Z., Paunesku, T., Labat, I., Zeremski, M., Snoddy, J., Funkhouser, W.K., Koop, B., Hood, L., and Crkvenjakov, R. (1993). DNA Sequence Determination by Hybridization: A Strategy for Efficient Large-Scale Sequencing. Science *260*, 1649-1652.

Dryana, D. and White, R. (1985). The Genetic Linkage Map of the Human X Chromosome. Science *230*, 753-758.

Dumas Milne Edwards, J., Delort, J., and Mallet, J. (1991). Oligodeoxyribonucleotide ligation to single-stranded cDNAs: a new tool for cloning 5' ends of mRNAs and for constructing cDNA libraries by *in vitro* amplification. Nucl. Acids Res. *19/19*, 5227-5232.

Engelke, D.R., Krikos, A., Bruck, M.E., and Ginsburg, D. (1990). Purification of *Thermus aquaticus* DNA Polymerase Expressed in *Escherichia coli*. Anal. Biochem. *191*, 396-400.

Feinberg, A.P. and Vogelstein, B. (1983). A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal. Biochem. *132*, 6-13.

Fodor, S.P.A., Rava, R.P., Huang, X.H.C., Pease, A.C., Holmes, C.P., and Adams, C.L. (1993). Multiplexed biochemical assays with biological chips. Nature *364*, 555-556.

Foote, S., Vollrath, D., Hilton, A., and Page, D.C. (1992). The human Y chromosome: overlapping DNA clones spanning the euchromatic region. Science *258*, 60-66.

Freier, S.M., Kierzek, R., Jaeger, J.A., Sugimoto, N., Caruthers, M.H., Nelson, T., and Turner, D.H. (1986). Improved free-energy parameters for prediction of RNA duplex stability. Proc. Natl. Acad. Sci. U. S. A. *83*, 9373-9377.

Froussard, P. (1992). A random-PCR method (rPCR) to construct whole cDNA library from low amounts of RNA. Nucl. Acids Res. *20*, 2900

Gubbay, J., Collignon, J., Koopman, P., Capel, B., Economou, A., Munsterberg, A., Vivian, N., Goodfellow, P., and Lovell Badge, R. (1990). A gene mapping to the sex-determining region of the mouse Y chromosome is a member of a novel family of embryonically expressed genes. Nature *346*, 245-250.

Gusella, J.F., Wexler, N.S., Conneally, P.M., Naylor, S.L., Anderson, M.A., Tanzi, R.E., Watkins, P.C., Ottina, K., Wallace, M.R., Sakaguchi, A.Y., and et al (1983). A polymorphic DNA marker genetically linked to Huntington's disease. Nature *306*, 234-238.

Gyapay, G., Morisette, J., Vignal, A., Dib, C., Fizames, C., Millasseau, P., Marc, S., Bernardi, G., Lathrop, M., Weissenbach, J. (1994). The 1993-94 Généthon human genetic linkage map. Nature Genet. *7*, 246-339.

Hawkins, J.D. (1988). A survey of intron exon lengths. Nucl. Acids Res. *16*, 9893-9908.

Hedrick, S.M., Cohen, D.I., Nielsen, E.A., and Davis, M.M. (1984). Isolation of cDNA clones encoding T cell-specific membrane-associated proteins. Nature *308*, 149-153.

Hoheisel, J.D., Maier, E., Mott, R., McCarthy, L., Grigoriev, A.V., Schalkwyk, L.C., Nizetic, D., Francis, F., and Lehrach, H. (1993). High resolution cosmid and P1 maps spanning the 14 Mb genome of the fission yeast S. pombe. Cell *73*, 109-120.

Hoheisel, J.D. (1987). Untersuchung Verschiedener Topologisch Induzierter DNS-Skundärstrukturen. Ph.D. thesis, University of Konstanz.

Hollstein, M., Sidransky, D., Vogelstein, B., and Harris, C.C. (1991). p53 mutations in human cancers. Science *253*, 49-53.

Höög, C. (1991). Isolation of a large number of novel mammalian genes by a differential cDNA library screening strategy. Nucl. Acids Res. *19*, 6123-6127.

Hyde, S.C., Gill, D.R., Higgins, C.F., Trezise, A.E.O., MacVinish, L.J., Cuthbert, A.W., Ratcliff, R., Evans, M.J., and Colledge, W.H. (1993). Correction of the ion-transport defect in cystic-fibrosis transgenic mice by gene-therapy. Nature *362*, 250-255.

Jacobs, J.W. and Fodor, S.P.A. (1994). Combinatorial chemistry - applications of light-directed chemical synthesis. Trends Biotech. *12*, 19-26.

Jeffreys, A.J., Wilson, V., and Thein, S.L. (1985). Hypervariable 'minisatellite' regions in human DNA. Nature *314*, 67-73.

Jones, P., Watson, A., Davies, M., and Stubbings, S. (1992). Integration of image analysis and robotics into a fully automated colony picking and plate handling system. Nucl. Acids Res. *20*, 4599-4606.

Kacian, D.L. and Myers, J.C. (1976). Synthesis of extensive, possibly complete, DNA copies of poliovirus RNA in high yields and at high specific activities. Proc. Natl. Acad. Sci. U. S. A. *73*, 2191-2195.

Kahn, A.S., Wilcox, A.S., Polymeropoulos, M.H., Hopkins, J.A., Stevens, T.J., Robinson, M., Orpana, A.K., and Sikela, J.M. (1992). Single pass sequencing and physical and genetic mapping of human brain cDNAs. Nature Genet. *2*, 180-185.

Kaiser, K. (1990). New directions in cDNA cloning. Technique *2*, 1-17.

Kaufmann, Y., Milcarek, C., Berissi, H., and Penman, S. (1977). HeLa cell poly(A)-mRNA codes for a subset of poly(A)+mRNA-directed proteins with an actin as a major product. Proc. Natl. Acad. Sci. U. S. A. *74*, 4801-4805.

Kerem, B., Rommens, J.M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A., Buchwald, M., and Tsui, L.C. (1989). Identification of the cystic fibrosis gene: genetic analysis. Science *245*, 1073-1080.

Khrapko, K.R., Lysov, Y.P., Khorlyn, A.A., Shick, V.V., Florentiev, V.L., and Mirzabekov, A.D. (1989). An oligonucleotide hybridisation approach to DNA sequencing. FEBS-Letters *256*, 118-122.

Khrapko, K.R., Lysov, Y.P., Khorlyn, A.A., Ivanov, I.B., Yershov, G.M., Vasilenko, S.K., Florentiev, V.L., and Mirzabekov, A.D. (1991). A method for DNA sequencing by hybridisation with oligonucleotide matrix. DNA Seq. *1*, 375-388.

Knudson, A.G.J. (1971). Mutation and cancer: statistical study of retinoblastoma. Proc. Natl. Acad. Sci. U. S. A. *68*, 820-823.

Ko, M.S.H. (1991). An `equalized cDNA library' by the reassociation of short double- stranded cDNAs. Nucl. Acids Res. *18*, 5705-5711.

Koopman, P., Gubbay, J., Vivian, N., Goodfellow, P., and Lovell-Badge, R. (1991). Male development of chromosomally female mice transgenic for *Sry*. Nature *351*, 117-121.

Kotewicz, M.L., Sampson, C.M., D'Alessio, J.D., and Gerard, G.F. (1988). Isolation of clones Moloney murine leukemia virus reversetranscriptase lacking ribonuclease H activity. Nucl. Acids Res. *16*, 265-277.

Lee, S.W., Tomasetto, C., and Sager, R. (1991). Positive selection of candidate tumor-supressor genes by subtractive hybridization. Proc. Natl. Acad. Sci. U. S. A. *88*, 2825-2829.

Lehrach, H., Drmanac, R., Hoheisel, J., Larin, Z., Lennon, G., Monaco, A.P., Nizetic, D., Zehetner, G., and Poustka, A. (1990). Hybridisation fingerprinting in genome mapping and sequencing. In Genome Analysis Volume1: Genetic and Physical Mapping. K.E. Davies and S.M. Tilghman, eds. (Cold Spring Harbor: Cold Spring Harbor Laboratory Press), pp. 39-81.

Lennon, G.G. and Lehrach, H. (1991). Hybridization analyses of arrayed cDNA libraries. Trends Genet. *7*, 314-317.

Maier, E., Hoheisel, J.D., McCarthy, L., Mott, R., Grigoriev, A.V., Monaco, A.P., Larin, Z., and Lehrach, H. (1992). Complete coverage of the Schizosaccharomyces pombe genome in yeast artificial chromosomes. Nature Genet. *1*, 273-277.

Maier, E., Meier Ewert, S., Ahmadi, A.R., Curtis, J., and Lehrach, H. (1994a). Appliation of robotic technology to automated sequence fingerprint analysis by oligonucleotide hybridisation. J. Biotech. *35*, 191-203.

Maier, E., Roest Crollius, H., and Lehrach, H. (1994b). Hybridisation techniques on gridded high density DNA and *in-situ* colony filters based on fluorescence detection. Nucl. Acids Res. *22*, 3423-3424.

Makalowski, W., Mitchell, G.A., and Labuda, D. (1994). Alu sequences in the coding regions of mRNA: a source of protein variability. Trends Genet. *10*, 188-193.

Mandel, J-L. (1994). Trinucleotide diseases on the rise. Nature Genet. *7*, 453-455.

Maskos, U. and Southern, E.M. (1992a). Parallel analysis of oligodeoxyribonucleotide (oligonucleotide) interactions .1. analysis of factors influencing oligonucleotide duplex formation. Nucl. Acids Res. *20*, 1675-1678.

Maskos, U. and Southern, E.M. (1992b). Oligonucleotide hybridizations on glass supports - a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesized insitu. Nucl. Acids Res. *20*, 1679-1684.

Maskos, U. and Southern, E.M. (1993a). A novel method for the parallel analysis of multiple mutations in multiple samples. Nucl. Acids Res. *21*, 2269-2270.

Maskos, U. and Southern, E.M. (1993b). A study of oligonucleotide reassociation using large arrays of oligonucleotides synthesized on a glass support. Nucl. Acids Res. *21*, 4663-4669.

Mather, E.L., Alt, F.W., Bothwell, A.L.M., Baltimore, D., and Koshland, M.E. (1981). Expression of J Chain RNA in Cell Lines Representing Different Stages of B Lymphocyte Differentiation. Cell *23*, 369-378.

Meier Ewert, S., Maier, E., Ahmadi, A., Curtis, J., and Lehrach, H. (1993). An automated approach to generating expressed sequence catalogues. Nature *361*, 375-376.

Milcarek, C. (1979). HeLa Cell Cytoplasmic mRNA Contains Three Classes of Sequences: Predominantly Poly(A)-Free, Predominantly Poly(A)-Containing and Bimorphic. Eur. J. Biochem. *102*, 467-476.

Milner, R.J., Bloom, F.E., and Sutcliffe, G.J. (1987). Brain-Specific Genes: Strategies and Issues. Curr. Top. Dev. Biol. *21*, 117-150.

Milner, R.J. and Sutcliffe, G.J. (1983). Gene Expression in Rat Brain. Nucl. Acids Res. *11*, 5497-5520.

Mirzabekov, A.D. (1994). DNA sequencing by hybridisation - a megasequencing method and a diagnostic tool?. Trends Biotech. *12*, 27-32.

Monaco, A.P., Neve, R.L., Colletti Feener, C., Bertelson, C.J., Kurnit, D.M., and Kunkel, L.M. (1986). Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene. Nature *323*, 646-650.

Mott, R., Grigoriev, A.V., Maier, E., Hoheisel, J.D., and Lehrach, H. (1993). Algorithms and software tools for ordering clone libraries: application to the mapping of the genome of *Schizosaccharomyces pombe*. Nucl. Acids Res. *21*, 1965-1974.

Nizetic, D., Drmanac, R., and Lehrach, H. (1991a). An improved bacterial colony lysis procedure enables direct DNA hybridisation using

short (10, 11 bases) oligonucleotides to cosmids. Nucl. Acids Res. *19*, 182

Nizetic, D., Zehetner, G., Monaco, A.P., Gellen, L., Young, B.D., and Lehrach, H. (1991b). Construction, arraying, and high-density screening of large insert libraries of human chromosomes X and 21: their potential use as reference libraries. Proc. Natl. Acad. Sci. U. S. A. *88*, 3233-3237.

Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y., and Matsubara, K. (1992). Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. Nature Genet. *2*, 173-179.

Olson, M.V., Hood, L., Cantor, C.R., and Botstein, D. (1989). A common language for physical mapping of the human genome. Science *245*, 1434

Patanjali, S.R., Parimoo, S., and Weissman, S.M. (1991). Construction of a uniform-abundance (normalized) cDNA library. Proc. Natl. Acad. Sci. U. S. A. *88*, 1943-1947.

Pollak, A., Blumenfeld, H., Wax, M., Baughn, R.L., and Whitesides, G.M. (1980). Enzyme Immobilidation by Condensation Copolymerisation into Cross-Linked Polyacrylamide Gels. J. Am. Chem. Soc. *102*, 6324-6335.

Polymeropoulos, M.H., Xiao, H., Glodek, A., Gorski, M., Adams, M.D., Moreno, R.F., Fitzgerald, M.G., Venter, J.C., and Merril, C.R. (1992). Chromosomal assignment of 46 brain cDNAs. Genomics. *12*, 492-496.

Polymeropoulos, M.H., Xiao, H., Sikela, J.M., Adams, M.D., Venter, J.C., and Merril, C.R. (1993). Chromosomal distribution of 320 genes from a brain cDNA library. Nature Genet. *4*, 381-386.

Poustka, A., Pohl, T., Barlow, D.P., Zehetner, G., Craig, A., Michiels, F., Ehrich, E., Frischauf, A.M., and Lehrach, H. (1986). Molecular approaches to mammalian genetics. Cold. Spring. Harb. Symp. Quant. Biol. *51 Pt 1*, 131-139.

Riordan, J.R., Rommens, J.M., Kerem, B., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., Chou, J.L., and et al (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science *245*, 1066-1073.

Rommens, J.M., Iannuzzi, M.C., Kerem, B., Drumm, M.L., Melmer, G., Dean, M., Rozmahel, R., Cole, J.L., Kennedy, D., Hidaka, N., and et al (1989). Identification of the cystic fibrosis gene: chromosome walking and jumping. Science *245*, 1059-1065.

Rothstein, J.L., Johnson, D., Deloia, J.A., Skowronski, J., Solter, D., and Knowles, B. (1992). Gene-expression during preimplantation mouse development. Genes Dev 1992 *6*, 1190-1201.

Royer Pokora, B., Kunkel, L.M., Monaco, A.P., Goff, S.C., Newburger, P.E., Baehner, R.L., Cole, F.S., Curnutte, J.T., and Orkin, S.H. (1986). Cloning the gene for an inherited human disorder--chronic granulomatous disease--on the basis of its chromosomal location. Nature *322*, 32-38.

Saiki, R.K., Gelfand, S., Stoffel, S., J., Scharf, R., Higuchi, G.T., Mullis, K.B., and Erlich, H.A. (1988). Primer-directed enzymatic amplification of DNA with thermostable DNA polymerase. Science *239*, 487-491.

Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989). Molecular Cloning: A Laboratory Manual (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press).

Sargent, T.D. and Dawid, I.B. (1983). Differential Gene Expression in the Gastrula of *Xenopus laevis*. Science *222*, 135-139.

Sinclair, A.H., Berta, P., Palmer, M.S., Hawkins, J.R., Griffiths, B.L., Smith, M.J., Foster, J.W., Frischauf, A.M., Lovell Badge, R., and Goodfellow, P.N. (1990). A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. Nature *346*, 240-244.

Southern, E.M. (1992). Genome Mapping: cDNA approaches. Curr. Opinion Genet. Dev. *2*, 412-416.

Southern, E.M., Maskos, U., and Elder, J.K. (1992). Analyzing and comparing nucleic-acid sequences by hybridization to arrays of oligonucleotides - evaluation using experimental-models. Genomics *13*, 1008-1017.

Southern, E.M., Casegreen, S.C., Elder, J.K., Johnson, M., Mir, K.U., Wang, L., and Williams, J.C. (1994). Arrays of complementary oligonucleotides for analysing the hybridisation behaviour of nucleic acids. Nucl. Acids Res. *22*, 1368-1373.

Strezoska, Z., Paunesku, T., Radosavljevic, D., Labat, I., Drmanac, R., and Crkvenjakov, R. (1991). DNA sequencing by hybridization: 100 bases read by a non-gel-based method. Proc. Natl. Acad. Sci. U. S. A. *88*, 10089-10093.

Sutcliffe, G.J. (1988). mRNA in the Mammalian Central Nervous System. Ann. Rev. Neurosci. *11*, 157-198.

The Huntington's Disease Collaborative Research Group (1993). A Novel Gene Containing a Trinucleotide Repeat That Is Expanded and Unstable on Huntington's Disease Chromosomes. Cell *72*, 971-983.

Travis, G.H. and Sutcliffe, G.J. (1988). Phenol emulsion-enhanced DNA-driven subtractive cDNA cloning: Isolation of low-abundance monkey cortex-specific mRNAs. Proc. Natl. Acad. Sci. U. S. A. *85*, 1696-1700.

Van Ness, J. and Hahn, W.E. (1980). Sequence complexity of cDNA transcribed from a diverse mRNA population. Nucl. Acids Res. *18*, 4259-4269.

Vollrath, D., Foote, S., Hilton, A., Brown, L.G., Beer Romero, P., Bogan, J.S., and Page, D.C. (1992). The human Y chromosome: a 43-interval map based on naturally occurring deletions. Science *258*, 52-59.

Weber, J.L. and May, P.E. (1989). Abundant Class of Human DNA Polymorphisms Which Can Be Typed Using the Polymerase Chain Reaction. Am. J. Hum. Genet. *44*, 388-396.

Weinberg, R.A. (1991). Tumor Suppressor Genes. Science *254*, 1138-1146.

Weissenbach, J., Gyapay, G., Dip, C., Vignal, A., Morissette, J., Millasseau, P., Vaysseix, G., and Lathrop, M. (1992). A second generation linkage map of the human genome. Nature *359*, 794-801.

Wetmur, J.G. (1991). DNA probes: Applications of the Principles of Nucleic Acid hybridisation. Crit. Rev. Biochem. Mol. Boil. *26*, 227-259.

Wilcox, A.S., Kahn, A.S., Hopkins, J.A., and Sikela, J.M. (1991). Use of 3' untranslated sequences of human cDNAs for rapid

chromosome assignment and conversion to STSs: implications for an expression map of the genome. Nucl. Acids Res. *19/8*, 1837-1843.

Zabner, J., Couture, L.A., Gregory, R.J., Graham, S.M., Smith, A.E., and Welsh, M. (1993). Adenovirus-mediated gene-transfer transiently corrects the chloride transport defect in nasal epithelia of patients with cystic-fibrosis. Cell *75*, 207-216.

Zehetner, G. and Lehrach, H. (1994). The Reference Library System - sharing biological material and experimantal data. Nature *367*, 489-491.

Curriculum Vitae

Name:	Sebastian Meier-Ewert
Date of Birth:	16.04.1968
Place of Birth:	Taunton, England
Nationality:	British / German
Marital status:	Married

School education:

1974-1976:	Mariahilfs Grundschule, Munich, Germany
1976-1979:	Volksschule, Munich (Unterföhring), Germany
1979-1981:	Gymnasium Dr Überreiter, Munich, Germany
1981-1986:	Leighton Park School, Reading, England

Qualifications:

O-levels: English, Maths, Chemistry, Physics, Biology,History, Geography, German, Environmental Science, Woodwork

A-level: Biology, Chemistry, Physics

Further education:

1987-1990:	University College London

Qualification:

First Class Honours in Biochemistry (B.Sc.)

1990-1994: Imperial Cancer Research Fund,
PhD Student

Publications:

Monaco, A. P., Muller, U., Larin, Z., Meier-Ewert, S., Lehrach, H., *Genomics* **11**, 1049-1053 (1991): Isolation of the Human Sex Determining Region from a Y-Enriched Yeast Artificial Chromosome Library.

Cole, C. G., Dunham, I., Coffey, A., Ross, M., Meier-Ewert, S., Borrow, M., Bentley, D. *Genomics* **14**, 256-262 (1992): A Random STS Strategy for Construction of YAC Contigs Spanning Defined Chromosomal Regions.

Meier-Ewert, S., Maier, E., Ahmadi, A., Curtis, J. and Lehrach, H. *Nature* **361**, 375-376 (1993): An automated approach to generating expressed sequence catalogues.

Maier, E., Meier-Ewert, S., Ahmadi, A., Curtis, J., Lehrach, H. *J. Biotechnology* **35**, 191-203 (1994): Application of robotic technology to automated sequence fingerprint analysis by oligonucleotide hybridisation

Rowe, P., Goulding, J., Read, A., Mountford, R., Hanauer, A., Oudet, C., Whyte, M. P., Meier-Ewert, S., Lehrach, H., Davies, K. E., O'Riordan, J. L. H. *Human Genetics* **91**, 571-573 (1993):New Markers for the Linkage Analysis and Physical Mapping of X-linked Hypophosphatemic Rickets.

Larin, Z., Monaco, A. P., Meier-Ewert, S. Lehrach, H. *Methods in Enzymology* **255** Chapter 37, (1993): Construction of Yeast Artificial Chromosome Libraries.

Cox, R. D., Meier-Ewert, S., Ross, M., Larin, Z., Monaco, A. P., Lehrach, H. *Methods in Enzymology* **225**, Chapter 38, 637 - 653 (1993): Genome Mapping and Cloning of Mutations Using Yeast Artificial Chromosomes.

Hoheisel, J. D., Maier, E., Meier-Ewert, S., Lehrach, H. *Annales de Biologie Clinique* **9**, (1993): Relational Genome Analysis Based on Hybridisation Techniques.

Meier-Ewert, S., Rothe, J., Mott, R., Lehrach, H *Identification of Transcribed Sequences* (ed. U. Hochgeschwender) Plenum Press, 253 - 260 (1994): Establishing Catalogues of Expressed Sequences by Oligonucleotide Fingerprinting of cDNA Libraries.

Meier-Ewert, S., Schalkwyk, L., Francis, F., Lehrach, H. *Handbook of Genome Analysis* (ed. N.Spurr) Blackwell Scientific Publications : Long-range physical Mapping *(in the press)*.