Identification of potential gene coding sequences within large cloned DNA arrays: Analysis of zinc finger motif.

Rosalind John.

Department of Biochemistry,

Imperial College of Science, Technology and Medicine,

London,

SW7 2AZ.

Submitted for the Degree of Ph.D.

May 1991

Abstract.

Cosmid clones from mouse/human hybrid cell lines enriched for the short arm of human chromosome 11 are being assembled into cloned DNA arrays to produce a physical map of this region. These were used in the development of two methods to identify coding sequences.

A method was devised to screen cosmid clones for BssHII and EagI sites. These sites frequently occur at " CpG " islands which are associated with particular classes of genes. Their location within the defined map should aid in the identification of these genes.

A second method was used to detect an abundant class of proteins, the zinc finger proteins (ZFPs). This class includes transcription factors, genes involved in developmental processes and genes thought to be involved in tumourigenesis. The involvement of these genes in complex biological processes makes them an important group. Members of this class were located using an oligonucleotide that detects the conserved amino acid motif, HTGEKPY.

4.7% of the human DNA-containing cosmid clones from the human/mouse hybrid cell line hybridised with the zinc finger oligonucleotide compared with 0.5-1% of cosmid clones from a total human DNA library indicating clustering of ZFPs in the human DNA of the cell line. We have isolated 101 cosmids containing potential finger coding regions. Analysis of DNA sequence from strongly hybridising clones has shown the classic C_2H_2 finger protein motif. 69 of these cosmids have been *in situ* mapped, in a collaboration, to 3p, 11p, 19p, 19qter, 20p and 21qprox: some are clustered. Several of these regions are associated with genetic diseases including inherited sucepti.bility to cancers and growth abnormalities.

Cross hybridisation studies have detected two groups of closely related genes. cDNAs have been detected for members of one group. Members from both groups are clustered on chromosome 19.

Acknowledgements.

I would like to thank my supervisor, Dr. Peter Little, for help and guidance during the course of this Ph.D. I am grateful to David Porteous for the E65-9 and EJNAC hybrid cell lines; Afshan Malik and John Cowell, Claire Isacke and Steve Neame, Jenny Dunne, and Andrew Sinclair for cDNA libraries; and Marcel Mannens and Jan Hoovers for the *in situ* work. I would also like to acknowledge members of the sixth floor both past and present for providing a stimulating and enjoyable work environment, with special thanks to Philip Crossley, Kim Harrison-Lavoie and Fiona Mellon-Hemsley. I would like to thank Glynn Millhouse for photographic work. I would like to thank the Science and Engineering Council and Southwark Social Security for their attempts to financially support me. I would also like to thank my mother and my brothers, Peter and Chris, for their encouragement and support, my friends for keeping me sane, and Axel for being there and for putting up with me while this was being written.

List of contents.

List of figures.

List of tables.

List of abbreviations.

AC - adenocarcinoma

ALL - acute lymphoblastic leukemia

AML - acute myeloid leukemia

B-CLL - chronic lymphocytic leukemia (B cell)

bp - nucleotide base pair

BWS - Beckwith-Weiderman syndrome

BSA - bovine serum albumin

Ci - Curie

cM - centimorgan units

CMGT - Chromosome-mediated gene transfer

cR - centirays

cpm - counts per minute

dATP - deoxyadenosine triphosphate

dCTP - deoxycytidine triphosphate

dGTP - deoxyguanosine triphosphate

DNA - deoxyribonucleic acid

DNase - deoxyribonuclease

dNTP - dGTP or dATP or dTTP or dCTP

DTT - dithiothreitol

dTTP - deoxythymidine triphosphate

EDTA - ethylenediamine tetra-acetic acid

EGTA - ethylene glycol-bis (β aminothyl ether) N, N, N', N'-tetra-acetic acid

FIGE - field inversion gel electrophoresis

FLpter - fractional length of the whole chromosome relative to pter.

GL - glioma

H-*ras* - c-Harvey-*ras*-1

HTF - HpaII Tiny Fragments

IPTG - isopropylthio-β-D-galactoside

kb - kilobases

l - litre

*lac*Z - the structural gene for β-galactosidase

M - molar

Mb - megabase pairs

MDS - myelodysplastic syndrome

ML - malignant lymphoma

MPD - myeloproliferative disorder

μl - microlitre

μg - microgram

ml - millilitre

mg - milligram

ng - nanogram

NMR - nuclear magnetic resonance

OD - optical density

PA - pleomorphic adenoma

PCR - polymerase chain reaction

PEG - polyethylene glycol

pers. comm. - personnol communication

PFGE - pulse-field gel electrophoresis

pfu - plaque forming units

pg - picogram

PNK - polynucletide kinase

PV - polycythemia vera

RNA - ribonucleic acid

RNase - ribonuclease

rpm - revolutions per minute

RT - room temperature

RFLPs - restriction fragment length polymorphisms

RREF - rare-cutting restriction enzyme fingerprinting

SCC - small cell carcinoma

SDS - sodium dodecyl sulphate

SSC - standard saline solution

STS - sequence tagged site

T-ALL - acute lymphoblastic leukemia (T cell)

TE - Tris-HCl pH7.5 (10mM)/ EDTA pH8.0 (1mM)

TEMED - tetramethylethylene diamine

VNTR - variable number tandem repeat

WAGR - Wilm's tumour, aniridia, genitourinary tract malformations and mental retardation.

w/v - weight/volume

WT - Wilms' tumour

X - any amino acid residue

Xgal - 5-bromo-4-chloro-3-indolyl-β-D-galactoside

YAC - yeast artificial chromosome

ZF - zinc finger

ZFP - zinc finger protein

Chapter 1.

Introduction.

1.1 Mapping the Human Genome.

The goal of the human genome mapping project is the mapping and sequencing of all the DNA in the human chromosomes. This will provide a background for the analysis of the organisation of human DNA and for basic research into developmental biology and evolution. The localisation of an estimated 100,000 genes will provide the information required in understanding and eventually treating many of the more than 4,000 genetic diseases (McKusick 1988) that afflict mankind, as well as the multifactorial diseases in which genetic predisposition plays an important role.

1.2 Mapping Complex Genomes.

Maps of the human genome are being developed at many different levels. These can be divided into those that order markers with respect to chromosomal position or with respect to the position of a second marker and define distances involved in terms of recombination percentages or percentages of chromosomal length, and those that order markers at a molecular level where the distances are measured in units of physical length, these being the number of nucleotide pairs.

1.2.1 Ordering of markers at low resolution:

There are three ways in which markers can be ordered at a low resolution:

1) Genetic linkage maps.

Genetic linkage maps are based on the co-inheritance of allele combinations across multiple polymorphic loci. When two loci are linked closely on a segment of DNA there is a very small chance of a recombination event taking place between them as compared to loci further apart. Linkage maps are defined in terms of centimorgan units (cM), a map unit of 1cM corresponds to an observed recombination rate of 1%. This cannot be directly related to physical distance as specific regions of DNA vary in their recombination frequency and frequencies also differ between male and female meioses (Donis-Keller *et al.*, 1987). On average, 1cM is usually thought to correspond to 1 megabase pairs (Mb) of DNA in the human genome and 2Mb in the

mouse genome (Barlow and Lehrach, 1987). The markers used are single copy DNA fragments or restriction fragment length polymorphisms (RFLPs). RFLPs are inherited variations in the nucleotide sequence of individuals that can be detected with specific restriction enzymes (Botstein *et al.*, 1980). Averaged over the human genome about one base in 150 is polymorphic, and about one in six random base changes creates or abolishes a restriction site. Thus most 1kb stretches of DNA contain a restriction site which is present in some people and absent in others. Additionally, scattered throughout the human genome are tandemly repeated sequences whose copy number varies between people as a result of unequal recombination. A restriction fragment containing the whole tandemly repeated block, known as a variable number of tandem repeat (VNTR) RFLP (Nakamura *et al.*, 1987), will be a different size in different individuals. Genetic maps based on these polymorphisms are often termed RFLP maps. The genetic map of the human genome, at a resolution of about 10-20cM, is near completion (Donis-Keller *et al.*, 1987).

2) Radiation hybrid mapping.

Radiation hybrid mapping is based on the same theory as linkage mapping but does not rely on family data. Somatic cell hybrids in which regions of human DNA have been isolated in a rodent cell background are subjected to a high dose of x-rays causing breakage into fragments. The fragments are then recovered by fusion to a recipient rodent cell line which non-specifically retains some rodent and human chromosomal pieces from the donor cell line. These are analysed for the presence of certain markers (Goss and Harris, 1975, and Cox *et al.*, 1990). The further apart two markers are the more likely they are to be separated by this process. Distances are estimated in centirays (cR), 1cR being equivalent to 50 kilobases (kb) of DNA. This represents the most precise mapping process with respect the gross ordering of markers as the probability of breakage is equal in all regions of a chromosome. This method is useful in defining specific regions of DNA and has recently been applied to a human chromosome 21q mapping project (Cox *et al.*, 1990 and Burmeister *et al.*, 1991).

3) Cytogenetic maps.

Cytogenetic maps rely on *in situ* hybridisation of DNA probes to prometaphase spreads in relation to chromosomal banding and/or hybridised probes (Gerhard *et al.*, 1981 and Pinkel *et al.*, 198· , Litcher *et al.*, 1988). Patterns of banding are obtained by staining with DNA-specific dyes after treatments that partially removes chromosomal proteins. Two or more probes can be hybridised to a single sample and distinguished by means of differently coloured fluorochromes. This method has been refined to localise fragments to specific regions accurate to within 1% of the chromosome length on metaphase chromosomes by using computer controlled confocal laser microscopy as demonstrated by Lichter *et al.* (1990).

These types of mapping form the essential backbone needed to guide the more precise but smaller scale mapping of specific chromosomal regions.

1.2.2 Molecularly based mapping:

1) Long range restriction enzyme maps.

Pulse-field gel electrophoresis (PFGE) designed by Schwartz and Cantor (1984) is capable of separating fragments of DNA from 50 to 10,000kb. The resolution depends on the time of switching between two fields (the pulse time): longer pulse times will resolve DNA fragments in the higher molecular weight range. By southern blotting and hybridisation, markers can be assigned to the fragments generated by different restriction enzymes to produce a map of markers positioned relative to restriction enzyme sites. Modifications by Chu *et al.* (1986) have allowed the resolution of fragments between 1 and 10Mb. PFGE bridges the gap between the *in situ* hybridisation methods and conventional gel electrophoresis. This type of map exploits restriction enzymes that cleave rarely in the human genome producing large DNA fragments. One of the drawbacks in this type of mapping is that the majority of restriction enzymes that cut rarely within mammalian DNA do so because they recognize a sequence containing the dinucleotide, CpG, which occurs at 1/5th the expected frequency except in regions commonly known as CpG islands (Bird, 1986). This

means that sites for these restriction enzymes are often clustered and can be difficult to map across. This can be overcome by analysing partial digests with these restriction enzymes. Examples of the application of this technique are the 10Mb restriction map around the Duchenne muscular dystrophy locus of Burmeister *et al.* (1988) and the 12Mb map around the cystic fibrosis locus of Fulton *et al.* (1989).

In addition to the ordering of markers, the construction of a long-range physical map of a region allows the direct comparison of physical and genetic distances which can reveal areas of increased and decreased recombination. One example is a region on the short arm of human chromosome 4 which is predicted by recombination data to contain 3.4Mb of DNA but actually contains 300-400kb of DNA with a 10-fold increased recombination frequency (Alitto *et al.*, 1991). These maps are also of importance when investigating chromosomal rearrangements and when comparing maps of chromosomal regions based on different DNA sources.

2) Cloned Maps.

A cloned DNA map consists of DNA fragments from a specific genomic regions cloned in phage and cosmid vectors (35-40,000bp insert size) and, more recently, into yeast artificial chromosome (YAC) vectors which can contain very large fragments of DNA, often greater than 1,000,000bp (Burke *et al.*, 1987). These are ordered by virtue of the number of restriction enzyme sites they contain in common or by hybridisation to DNA probes. A collection of overlapping clones is commonly referred to as a "contig". Technical methods for producing cloned DNA maps were developed during the construction of maps for the *Escherichia coli* genome (Kohara *et al.*, 1987), the *Saccharomyces cerevisiae* genome (Olson *et al.*, 1986) and the *Caenorhabditis elegans* genome (Coulson *et al.*, 1986). These methods are now being applied to the much larger mouse and human genomes to produce comprehensive maps of the individual chromosomes.

3) DNA sequence.

A DNA sequence is the order of nucleotides along a section of DNA. Sequencing the entire mouse or human genomes will require a considerable effort. A short term goal

is the sequencing of relatively small regions to be used as Sequence-Tagged Sites (STS). These are 200-500bp of sequence that uniquely identify a mapped gene or other marker. These sequences can be detected in the presence of all other genomic sequences by amplification by the polymerase chain reaction (Olson *et al.*, 1989). The information for using these "probes" can be completely described as information on a database. This creates a common language to facilitate comparisons between maps generated by the different techniques and by the different laboratories.

1.3 Application of genome mapping methods.

These methods of analysis have been developed and refined during the process of mapping a variety of genomes. The *E.coli* genome which is relatively small in size, 4 x 10^6bp, was first mapped by cloning fragments into phage vectors and mapping these with eight restriction enzymes (Kohara *et al.*, 1987). *S. cerevisiae* has a genome size of 12 x 10^6bp and was initially mapped with two restriction enzymes (Olson *et al.*, 1986). A cloned DNA map of 90% of the *C. elegans* genome, 8 x 10^7bp in size, has been mapped using cosmids as a vector by Coulson *et al.* (1986) and YACs are being used to close the gaps between the cosmid contigs (Coulson *et al.*, 1988).

The mouse and human genomes are significantly larger in size (3 x 10^9 base pairs, NRC, 1988) and the bulk of work to date has concentrated on specific regions. Two of the most intensively studied chromosomal regions in the mouse are those represented by *t haplotypes* and the major histocompatability complex (*H-2*) both on chromosome 17. The *t haplotypes* contain loci that effect embryonic viability, male fertility and meiotic transmission (Frischauf, 1985 and Silver, 1985). The *H-2* complex effects a wide variety of properties, including immune response and complement activity (Klein, 1975, 1986). Many of the mapping projects in the human genome have been centred around regions shown by genetic linkage studies to be associated with a genetic disease. Many of the genes affected in these genetic diseases have been located by intense mapping of these chromosomal regions. The genes affected in cystic fibrosis, Duchenne muscular dystrophy, retinoblastoma, von

Recklinghausen neurofibromatosis, the gene involved in a proportion of chromosome 21-linked Alzheimer's diseases, a gene (DCC) involved in colon cancer progression and one of the genes responsible for Wilm's tumour linked to 11p13 have been detected using these methods (Fearon *et al.*, 1990, Goate *et al.*, 1991, Viskochil *et al.*, 1990, Cawthon *et al.*, 1990 and Wallace *et al.*, 1990, Freind *et al.*, 1986, Monaco *et al.*, 1986 and Koenig *et al.*, 1987, Riordan *et al.*, 1989, Call *et al.*, 1990 and Gessler *et al.*, 1990). These are examples of successes with relatively simple disorders involving a single genetic locus where, except in the case of cystic fibrosis, gross chromosomal rearrangements (either germline or somatic) involving the gene greatly aided the search process. Finding genes responsible for more complex disorders such as schizophrenia, hypertension, epilepsy and susceptibility to cancer or coronary artery disease is likely to be much more difficult. The mapping of the human genome as a whole is likely to be the only approach for identifying the genes involved in these types of disorder.

The cloned DNA map is the most attractive method for mapping the region of the genome of interest as it not only supplies the information on the physical distances between markers but also provides the cloned DNA with which to work. Thus genes are simultaneously ordered and cloned. It then remains to detect the genes within the cloned DNA fragments and analyse their function.

1.4 Isolation of Chromosomal Regions.

The size of the human genome (3 x 10^9bp, NRC, 1988) excludes a direct cloning and mapping approach. In order to produce a cloned DNA map, the genome must be split into more manageable portions, either individual chromosomes or parts of chromosomes. Chromosomes can be sorted by cytofluorimetry to produce relatively pure preparations of individual chromosomes separated on the basis of A-T/G-C content (Gray *et al.*, 1979). In dual systems, the chromosomes are stained with two dyes which fluoresce at different wavelenths. The mitotic chromosomes are forced to flow one by one across a water-filled measuring region and excited sequentially by

two laser beams. The two fluorescence emissions are processed to give a two-dimensional representation and the chromosomes can be separated depending on their fluorescence pattern. A second way of isolating human chromosomal regions is using interspecific hybrids. The somatic cell hybrids can be used as a cloning source to obtain human DNA from a specific region. Two methods were used in the development of the somatic cell hybrids used in the chromosome 11 mapping project:

1) Chromosome-mediated gene transfer (CMGT) is the transfer of subchromosomal lengths of DNA into the appropriate immortalised cells. Whole chromosomes are co-precipitated with $CaPO_4$ and applied to a monolayer of the recipient cells which take up the fragments of the foreign DNA (McBride and Peterson, 1980).

2) Whole cell fusion is the fusion of two different cells to produce one cell with a single nucleus with the chromosome complement of both parent cells. (Barski *et al.*, 1960) in the presence of a chemical fusogen, polyethylene glycol (Pontecorvo 1975). If rodent cells are fused to human cells, the resultant interspecific hybrid spontaneously loses chromosomes (Weiss and Green 1967).

Hybrids are selected on the basis of:

a) The biochemistry of endogenous genes, for example the hypoxanthine, aminopterin, thymidine (HAT) selection method of Szybalski *et al.* (1962).

b) The biochemistry of exogenous genes, for example the *E. coli gpt* gene (Mulligan and Berg, 1980).

c) The presence of cell surface antigens detectable with antibodies (Tunnacliffe *et al.*, 1983).

d) The presence of proto-oncogenes conferring growth advantage in mammalian cells, for example the *ras* gene family (Porteous *et al.*, 1986).

1.5 Chromosome 11.

In this study cosmids are being used to produce a cloned DNA map of the short arm of human chromosome 11 using somatic cell hybrids enriched for this region.

Chromosome 11 is one of the most intensively studied autosomes with over 500 physical markers mapped to it. (Junien and McBride, 1989). Approximately 134Mb in size, a copy of chromosome 11 represents about 2.4% of the human genome (Harris *et al.*, 1986), the short arm being about 0.96%. The availability on hybrid cell lines containing this region isolated in a mouse DNA background and the density of markers in this region, useful in the gross ordering of the over-lapping cloned DNA contigs makes 11p an ideal candidate for a mapping project. The association of 11p with an inherited predisposition for the development of Wilms' tumour (WT) has focussed research in this region. The development of WT often occurs in conjunction with other particular physical abnormalities. When one physical abnormality is found in association with one or more other physical abnormalities it is known as a contiguous gene syndrome. These syndromes can result from the disturbance of several gene lying close together on a chromosome. Two contiguous gene syndromes are linked to 11p: The Wilms' tumour, aniridia, genitourinary tract malformations and mental retardation syndrome (WAGR) and Beckwith-Wiedemann Syndrome (BWS). These consist of a number of seemingly unrelated genetic defects found associated together in a significant proportion of affected individuals. Once these associations were noted, chromosomal abnormalities were observed on 11p indicating the involvement of genes in this region of the genome. Chromosomal abnormalities are only found occasionally in these syndromes but provide the necessary material for defining the limits of the region involved.

1.5.1 The WAGR syndrome.

The WAGR syndrome was first noted in 1964 (Miller *et al.*, 1964) and consists of the association of Wilms' tumour, aniridia, genitourinary tract malformations and mental retardation. Deletions at 11p13 were later found associated to the disease (Riccardi *et al.*, 1978, 1990 and Francke *et al.*, 1979).

1.5.2 Beckwith-Wiedemann syndrome.

Beckwith-Wiedemann syndrome is recognized by macroglossia, giantism, earlobe pits or creases, abdominal wall defects and an increased incidence of the development of tumours, especially adrenal carcinoma, hepatoblastoma, rhabdomyosarcoma and Wilms' tumour (Wiedemann, 1964, Beckwith, 1969 and Sortelo-Avila and Gooch, 1976). Duplications of 11p were first linked to this syndrome in 1983 by Waziri *et al.* (1983) and familial BWS has subsequently been localised to 11p15.5 Koufos *et al.* (1989).

1.5.3 Wilms' tumour.

Wilms' tumour (WT) is an embryonal malignancy of the kidney affecting 1 in 10,000 individuals under five years and is the most frequently detected solid tumour of this age group (Matsunaga, 1981). The development of Wilm's tumour is associated with events at the WAGR locus (Koufos *et al.*, 1984, Orkin *et al.*, 1984, Reeve *et al.*, 1984 and Fearon *et al.*, 1984) and the BWS locus (Reeve *et al.*, 1989 and Koufos *et al.*, 1989) suggesting two predisposition loci on 11p. The existence of a third WT locus is suggested by hereditary studies excluding linkage to 11p by Grundy *et al.* (1988) and Huff *et al.* (1988).

1.5.4 The Candidate 11p13 Wilms' Tumour gene.

Recently a candidate gene for the Wilms' tumour susceptibility gene mapping to the 11p13 region has been isolated which is predominantly expressed in kidney (Call *et al.*, 1990 and Gessler *et al.*, 1990). The gene lies within a 350kb region of overlap shared by various deletions in chromosome 11p13 (Rose *et al.*, 1990). A deletion resulting in uncontrolled cell growth would suggest the involvement of a tumour suppressor gene. The model for tumour suppressor genes was originally put forward by Knudson (1985) and is exemplified by the retinoblastoma gene (*Rb-1*) (Huang *et al.*, 1988) which encodes a protein with nuclear localisation and DNA binding activity (Lee *et al.*, 1987a and b). The introduction of one normal chromosome 11

into a Wilms' tumour cell line can suppress malignancy supporting a similar role for the *Wilms' tumour* (*WT*) gene of tumour suppression (Weissman *et al.*, 1987). The *WT* gene has four copies of the zinc finger motif, a DNA-binding structure originally described in the Xenopus 5S gene transcription factor IIIA (Miller *et al.*, 1985, Brown *et al.*, 1985) indicating a possible DNA-binding activity for the *WT* gene protein product. Evidence that support a role for the gene of Call *et al.* (1990) and Gessler *et al.* (1990) in development of Wilm's tumour comes from *in situ* messenger RNA hybridisation studies by Pritchard-Jones *et al.* (1990). This gene is expressed in the condensed mesenchyme, renal vesicle and glommerula epithelium of the developing kidney, in the related mesonephric glomeruli and in cells approximating these structures in tumours. The other main sites of expression are the genital ridge, foetal gonad and mesothelium. Pritchard-Jones *et al.* (1990) suggest that this gene has a role in kidney development and a wider role in mesenchymal-epithelial transitions. A role in genital development is also implicated in a study by van Heyningen *et al.* (1990) of an individual with aniridia, cryptorchidism and hypospadias sharing a region of deletion of 350-400kb with an individual with similar genital abnormalities and Wilms' tumour but with no aniridia. This region contains the candidate *WT* gene supporting a role in both kidney and genitourinary development. A second gene (*WIT-1*) transcribed in a divergent direction from a single CpG island associated with the candidate *WT* gene of Call *et al.* (1990) and Gessler *et al.* (1990) is also expressed in foetal kidney (Huang *et al.*, 1990). However, only 276bp of open reading frame and the presence of a genomic poly(A) track suggest WIT-1 may represent a processed pseudogene. The authors have suggested a role for this gene similar to that of transcribed elements upstream of the proto-oncogenes, c-*mos* and c-H-*ras* and the *epsilon globin* gene (Wood *et al.*, 1984, Lowndes *et al.*, 1989 and Wu *et al.*, 1989). It has been shown that differential expression of some of these elements are vital in regulating the expression of their associated genes in specific cellular and developmental contexts (Wood *et al.*, 1984 and Lowndes *et al.*, 1989). Thus any

mutations in *WIT-1* would influence the expression of the candidate gene of Call *et al.*, 1990 and Gessler *et al.* (1990).

It is probable that the numerous abnormalities associated with the WAGR syndrome involve other genes in addition to the candidate *WT* gene.

Considerable mapping activity is required to identify all the genes involved in both WAGR and BWS. To this end a cloned DNA map is being constructed in cosmid vectors using hybrid cell lines enriched for this region.

1.5.5 Hybrid cell lines.

E65-9.

E65-9 is a c-Harvey-*ras*-1 (H-*ras*) selected chromosome mediated gene transformant containing 2-3Mb of chromosome 11 DNA from around the H-*ras* gene (Porteous *et* al., 1986) This was used in a preliminary 11p mapping project (Harrison-Lavoie *et al.*, 1989). Porteous *et al.*, 1986 transfected mouse C127 cells with mitotic chromosomes derived from EJ-18-8D, a cloned cell line derivative of a human bladder carcinoma (Hastings and Franks, 1983). The presence of an activated H-*ras* oncogene in immortalised cells induces the formation of clumps or foci as a result of the loss of contact inhibition. Transformants are isolated and analysed. E65-9 was of great value for testing the application of the genome mapping techniques. The transfected DNA was notably rearranged and amplified with respect to the parent EJ-18-8D cell line (Harrison-Lavoie *et al.*, 1989). The process of CMGT is often associated with gross rearrangements of the transferred DNA (Porteous *et al.*, 1986, Pritchard and Goodfellow 1987) and this type of disturbance was not unexpected.

EJNAC4.1.

EJNAC4.1 was produced by the whole cell fusion of human EJ-18-8D cells with mouse C127 cells, the activated H-*ras* oncogene being selected for in immunosupressed mice which develop tumours (Porteous *et al.*, 1989). The parent tumour, EJNAC, contained all the chromosome markers tested distal to 11p12 from

MIC11 to HRAS1, and none of the markers for the long arm of chromosome 11. There is no evidence for genomic rearrangements or interstitial deletions in the process of cell fusion (Bickmore *et al.*, 1989a and b). Cosmid libraries made from this cell line contained 1.7% human clones as judged by hybridisation with total human DNA indicating that over 200Mb of human DNA is present. Regions of human DNA from 3p, 11p, 19p, 19qter, 20p, 21qprox and several centromeres were detected when mitotic spreads of peripheral lymphocytes were hybridised to 100 pooled random cosmids from this library (Hoovers *et al.*, submitted).

EJNAC4.1.5.

EJNAC4.1.5 is a hybrid cell line derived from EJNAC4.1 which has lost the 3p human component. A cosmid library constructed with genomic DNA from this cell line contains 1% human clones indicating the presence of more than 120Mb of human DNA.

1.6 The mapping process.

The work in this project was based within a venture to produce an overlapping cosmid-cloned DNA map of the short arm of chromosome 11. One copy of chromosome 11 represents 2.4% of the human genome which contains 6,000Mb of DNA (NRC, 1988), the short arm approximately being equivalent to 0.96% or 57.6Mb. If each cosmid contains on average 37.5kb then a genomes equivalent would be about 1500 clones. Cosmid libraries were constructed from chromosomal material isolated from the hybrid cell lines. Cosmid clones, isolated on the basis of hybridisation to a total human DNA probe, are analysed by a process known as fingerprinting (Coulson *et al* 1986). This involves restriction endonuclease digestion of very small scale preparations of DNA with HindIII and simultaneous end-labelling of the fragments generated. A second digest is then performed with Sau3A and the resulting mixture of labelled and unlabelled fragments is resolved on polyacrylamide gels to give a characteristic banding pattern for each cosmid. The subsequent data on

autoradiographs is analysed by a computer program developed by Coulson *et al.*, 1986 and Sulston *et al.*, 1988. A pair-wise comparison of cosmids gives a figure for probability of overlap. Random clones are ordered based on this figure which represents the number of bands they have common. The overlapping clones are then checked visually. These techniques allow the construction of large arrays of overlapping cloned DNA sequences known as contigs. The final aim is to clone very large regions of DNA covering the majority of 11p. Certain sequences may not be cloned in a cosmid vector or in a bacterial system leaving gaps in the map. These can be linked by hybridisation of single copy DNA fragments from the ends of contigs to NotI fragments separated by pulse field gel electrophoresis (Schwartz and Cantor, 1984) or by hybridisation of these probes to gridded YAC genomic libraries as for the *C. elegans* physical map (Coulson *et al.*, 1988). A large number of clones are studied in a mapping project and methods have been designed for the manipulation of many samples simultaneously. These manipulations are based on a microtitre system involving the use of 96-well system described in Gibson and Sulston (1987). 96 colonies which contain cloned human sequences are picked into 96-well culture tubes. After growth, DNA is prepared by an automated process using Beckman Biomek 1000 automated Laboratory. This DNA is then used in the fingerprinting reaction. A second screening process is employed at this stage to detect any miss-picked clones which contain mouse DNA using "colony arrays". These are filters with DNA from the clones grown in four 96-well culture tubes. Part of the culture is also stored at -70^{o}C in 25% glycerol in 96-well. These stocks represent the resource from which clones of importance can be obtained. The main advantage of this type of mapping process is the availability of the clones which can be used in studies of specific markers and the surrounding DNA. The acceleration of the cloning of genes, the main goal of this mapping project, is made easier in addition by the existence of the "colony arrays". These can be used in hybridisation studies looking for specific DNA sequences or homologues of known genes. The filters orientate directly back to

the stocks and the clone of interest plus those in the surrounding region are readily available for further studies.

1.7 Locating coding sequences.

The fundamental aim of this project was development of methods to detect coding sequences using the cosmid clones generated by the mapping of the short arm of human chromosome 11. Many approaches have been used to address the problem of locating genes within the bulk of the DNA. The current strategies for isolating genes can be divided into two categories: those that rely on detecting elements possessed by the majority of genes such as CpG islands and exons, and those that are based upon sequence conservation in genes related by ancestry or function.

1.7.1 Techniques based on physical characteristics possessed by the majority of mammalian genes.

1) CpG islands.

A characteristic shared by the many genes is the presence of CpG-islands over the 5' end of the gene. These regions are readily identifiable using restriction endonucleases that recognize DNA sequences containing oneor more copies of the CpG dinucleotide (Brown and Bird 1986). These restriction enzymes are commonly referred to as rare-cutting restriction enzymes and have been used successfully in a variety of ways to clone genes. Three of the most useful diagnostic restriction enzymes are BssHII, EagI and SacII (Lindsay and Bird 1987). Sargent *et al.* (1989) recently cloned 12 novel genes from the major histocompatability complex on the short arm of chromosome 6 using probes from genomic sequences flanking potential CpG islands.

2) Exon Trapping.

This recently developed method of Duyk *et al.* (1990) has been designed to detect the genes containing more than one exon. This method employs a retroviral shuttle vector with a splice donor site cloned next to a gene encoding the α-complementing factor of the *E.coli* β-galactosidase (α-β-GAL) gene which is followed by the cloning site.

When DNA containing a splice acceptor site adjacent to an exon is cloned into the vector and the recombinant retrovirus passes through its life cycle, the intervening DNA including the α-β-GAL gene is spliced out. The RNA is then reverse transcribed and amplified as a circular DNA episome which is used to transformation bacteria. Constructs in which splicing has occurred and which contain the exon are detectable as α-β-GAL$^-$ colonies. This method has yet to be examined on a large scale and involves complex technology.

1.7.2 Techniques based on sequence conservation:

1) Genomic DNA clones as Probes.

Groups have screened cDNA libraries and RNA blots with whole cosmids (Spies *et al.*, 1989) and recently with YACs (Elvin *et al.*, 1990). Hybridisation to genomic southern blots looking for sequences conserved between species has also been successful when a reasonably small area of DNA is examined: this approach was successful in the identification of the Duchenne muscular dystrophy gene (Monaco *et al.*, 1986). A disadvantage of both these methods is the presence of repeat sequences within the genomic DNA which can make results hard to interpret. This can be overcome to some extent by preannealing the radiolabelled probe to sonicated total genomic DNA (Sealey *et al.*, 1985). However, the lack of access to a variety of RNA species from specific tissues and at specific time points can present a considerable drawback in human genetics and many transcripts will be not be detected using these methods.

2) Cloned genes as probes.

Once a gene has been located it can be used as a probe to detect other similar sequences. This may involve cloning the homologue of a gene in a second species examples being the mouse *En-1* and *En-2* genes (Joyner *et al* ., 1985 and Joyner and Martin 1987) and the human *En-2* gene (Poole *et al.*, 1989) which were cloned using a probe from the Drosophila *engrailed* gene and the genes for the retenoic acid receptors, mRAR-α, mRAR-β, mRAR-γ and hRAR-γ (Zelent *et al.*, 1989 and

Krust *et al.*, 1989). It is also possible to use this method to detect related genes (orthologues) within a species, examples being the orthologues of the *ZFY* gene, originally thought to be the testis determining gene (Mardon *et al.*, 1989, Nagamine *et al.*, 1989), two orthologues of the Drosophila *knirps* segmentation gene (Rothe *et al.*, 1989), and the autosomal dystrophin gene (Love *et al.*, 1989).

3) Conserved domains.

Many coding sequences have been located by looking for conserved domains. When sequence data on a number of proteins with a related function has been amassed comparisons can be made and regions with homology detected. These regions can be used directly as probes or oligonucleotides can be designed to detect them. These methods have been successful in cloning a variety of genes most notable a large number of homeobox proteins (Carrasco *et al.*, 1984, Finney *et al.*, 1988, Levine *et al.*, 1984, Muller *et al.*, 1984, McGinnis *et al.*, 1984, Colberg-Poley *et al.*, 1985 are some examples), the related *paired* domain genes, *Pox meso and Pox neuro* (Bopp *et al.*, 1989), the human retenoic acid receptor, *hRAR*, which was cloned on the basis of homology to a region shared by a number of steroid hormone receptors (Petkovich *et al.*, 1987) and members of the *ras* gene superfamily have been isolated using oligonucleotide probes (Shirsat *et al.*, 1990 and Drivas *et al.*, 1990). New members of a family of closely related genes can be cloned using degenerative primers corresponding to consensus sequences to selectively amplify reverse transcribed mRNA. Four new members of the G protein-coupled receptor family (Libert *et al.*, 1989) and four new genes related to the *Wnt-1/int-1* gene of the mouse (Gavin *et al.*, 1990) were cloned in this way.

Recently a group of DNA binding proteins sharing regions of homology known as zinc fingers have been detected in this manner by a number of groups. These are discussed in section 1.9.

When deciding what routes to take in the search for coding sequences both the resources available from the mapping project and the applicability to large numbers of clones was considered. This thesis describes the search for coding sequences by:

1) The detection of genes of unknown function by their association with CpG islands.

2) The detection of genes sharing a region of homology, the zinc finger protein (ZFP) genes.

These are now more fully described:

1.8 CpG Islands.

1.8.1 Introduction.

The association of many genes with non-methylated clusters of CpGs has been described on many occasions and is reviewed in Bird (1986) and Bird (1987). Vertebrate genomes are highly methylated and vertebrate genes are nearly always methylated to some extent (Yisraeli and Szyf 1984). Methylation occurs at position 5 on the cytosine ring to produce 5-methylcytosine using S-adenosylmethionine as a source of the methyl group. Most but not all cytosines followed at the 3' side by a guanosine are methylated. This contrasts with invertebrate genomes which are methylated to a much lesser extent or not at all (Urieli-Shoval *et al.*, 1982).

1.8.2 Deficiency of CpG in mammalian genomes.

When spontaneous deamination of methyl cytosine occurs the product is thymine. When the strand in which the deamination has occurred is the template for the DNA-repair mechanism an adenine will be inserted into the opposite strand producing a point mutation. This sequence occurs at one fifth the expected frequency in vertebrate genomes. The 20% frequency of the CpG dinucleotide represents an equilibrium between the rate of creation of new CpGs and the accelerated rate of CpG loss as a result of methylation, calculated to be 12 times the normal transition rate (Sved and Bird, 1990). There are examples where one third of the point mutations causing genetic diseases involve the mutation of a CpG dinucleotide (Youssoufian *et al.*, 1986 and Nigro *et al.*, 1989). 60-90% of the CpGs dinucleotides

in a mammalian genome are methylated and evenly spread. It has been shown in more than a hundred cases that where the dinucleotide is not methylated it is clustered with other non-methylated CpGs and that these lie in 1-2kb regions over the 5' end of genes and in some cases cover the whole gene an example being the α-globin gene (Bird *et al.*, 1987). These regions are known commonly known as CpG islands or HTF islands (HpaII Tiny Fragments). The presence of these regions associated with genes implies that either these islands are involved in the transcription process or that they occur as a result of transcription activity at an expressed gene. Various studies have established an inverse correlation between transcriptionally active genes and the presence of CpG methylation in examples of housekeeping genes. Where a pseudogene is present, the CpGs have almost entirely been replaced by TpG and CpA as a result of deamination and subsequent non-repair and the remaining CpGs are methylated. An example is the α-globin pseudogene of man (Bird *et al.*, 1987).

1.8.3 CpG islands and gene expression.

There is no compelling evidence for the non-methylated sequences being a signal for gene activity. Tissue-specific genes examined are non-methylated in both expressing and non-expressing tissues (Oliva and Dixon 1989, Bird *et al.*, 1987, Mckeon *et al.*, 1982, Lloyd *et al.*, 1987 and Spanopoulou *et al.*, 1988). Keshet *et al.* (1985) have taken CpG islands that are normally unmethylated, artificially methlylated them and shown the inhibition of expression of the adjacent gene. Genes become methylated and inactivated as cultured cells are passaged and there is now evidence that the *de novo* methylation of CpG islands may have a role in the inactivation of genes in culture. 50% of the islands tested in the mouse cell lines NIH 3T3 and L cells, are methylated and the islands involved are essentially identical (Antequera *et al.*, 1990). These islands are thought to be those present at genes whose expression is not required in culture. The housekeeping genes were not found to become methylated. If methylation prevented transcription the *de novo* methylation of these genes would

render the cell non-viable (the methylation of genes in this category is not expected and has not been observed in the genes examined).

There are two possible ways in which methylation might prevent transcription:

1) Transcription factors cannot bind methylated DNA.

2) A protein preferentially binds methylated DNA blocking access to the transcriptional start site.

Specific protection of methylated CpGs in mammalian nuclei has been demonstrated by comparing the digest pattern obtained with MspI, which cuts CpG islands in naked DNA regardless of methylation state, with the digest pattern obtained with HpaII, which only cleaves non methylated sites, when intact mouse liver or brain nuclei are used as a substrate (Antequera et al., 1989). The pattern obtained with the two restriction enzymes was identical implying that the methylated sites are blocked. A nuclear protein that preferentially binds methylated DNA has been identified (Meehan *et al.*, 1989) using a 30bp molecule containing many copies of the CpG dinucleotide which formed complexes with nuclei extracts only when methylated. One possibility is that methylation acts as a signal for heterochromatin formation. When the nucleus forms after cell division the methyl-binding proteins can bind faster to the methylated DNA than the transcription factors, blocking the binding of the transcription complex and allowing nucleosomes to form.

1.8.4 The role of CpG islands in genome mapping and gene location.

Whatever the function of the CpG islands, the non-abundance of the CpG dinucleotide has provided a valuable tool in the work of those involved in genome mapping. Also, as this sequence is clustered and associated with genes, any restriction enzyme containing one or more CpGs in its recognition sequence can be used to identify these regions within cloned genomic DNA. Many groups involved in identifying genes have used this marker as part of their effort to clone the gene of interest.

As these sequences by their very nature are rare they can also be used in producing long range restriction enzyme maps of a particular region of DNA. Many groups are

using hybrid cell lines to produce cloned DNA maps. The extent of rearrangements or mutations occurring during the construction of these cell lines is not known and it is necessary to validate the integrity of these physical with respect to restriction maps produced by other groups of the same regions. The identification of cosmids containing rare cutting sites will also facilitate in the construction of these cloned DNA maps where gaps can be crossed using these cosmids as start and end points on PFGE analysis.

Lindsay and Bird (1987) tested several restriction enzymes with a recognition sequences containing the CpG dinucleotide. Sites for three of these restriction enzymes, BssHII (GCGCGC), EagI (CGGCCG) and SacII (CCGCGG) were calculated to occur 1.2 times per CpG island in 74% of the total islands present. These three restriction enzymes would be the most useful in the detection of genes. 89% of NotI sites (GCGGCCGC) are predicted to occur in CpG islands as the sequence is more C and G-rich but less than an 8^{th} of islands would be detected by NotI as it recognizes an 8bp sequence. NotI is the more useful of these restriction enzymes in long range restriction mapping.

1.9 Zinc Finger Motif.

1.9.1 Introduction.

The zinc finger proteins are an important group which possess in common an amino acid motif that is associated with a DNA-binding activity. They have been found in many different species and represent a very large family. It is feasible to look for a particular class of proteins using a shared motif as a probe. Newly isolated proteins possessing this motif can be proposed to share the same or a similar function. This family has been shown to represent between 300-500 of the genes in the human genome (Bellefroid *et al.*, 1989).

1.9.2 Protein Motifs.

Despite the apparently almost infinite variation possible in the three dimensional structure of a protein molecule, proteins sharing a similar function often share a similar arrangement of amino acids. This may be a result of divergence from a common ancestors or it may reflect evolutionary convergence.

Geometrically regular motifs are found repeated in proteins examples being the β-pleated sheet and the α-helix which are a consequence of defined hydrogen bonding between different peptide groups. There appear to be a limited number of possible conformations that a protein can take to fulfil its role. Certain basic designs are available for the general folding of a protein molecule. Other designs are required when the protein has a role in catalytic or ligand binding activity. By comparing proteins with similar functions the amino acids playing a key role in the structure and function of the molecule can be defined. Once an invariant arrangement of amino acids or "motif" is found associated with a particular structure or function, uncharacterised proteins can be examined for these motifs to give a possible clue to their function.

1.9.3 DNA-binding Motifs.

A large number of proteins that bind DNA have now been analysed and have been shown to contain particular structures. It is clear from data on the amino acid sequence that the majority of these proteins contain one or more of three distinct structures. These are commonly known as the helix-turn-helix motif, the leucine zipper motif and the zinc finger motif.

1) The Helix-turn-helix motif.

This structure consists of two helices separated by a β-pleated sheet as elucidated by X-ray crytallography and reviewed by Pabo and Sauer, 1984. It is present in a family of proteins known as the homeobox proteins which were isolated from *Drosophila* on the basis of their involvement in early development (Levine and Hoey, 1988) and yeast MATα2 protein (Sauer *et al.*, 1988).

2) The Leucine Zipper motif.

These proteins contain four or five leucine residues that are spaced seven residues apart which form a dimerisation domain adjacent to a DNA-binding domain rich in basic amino acids (Landschulz *et al.*, 1988). Examples include the yeast GCN4 transcriptional activator (Vogt *et al.*, 1987 and Struhl, 1987), the fos, jun and myc oncoproteins (Vogt *et al.*, 1987) and the C/EBP enhancer binding protein (Landschulz *et al.*, 1988). Leucine zippers have been identified in proteins that also contain the helix-loop-helix domain such as the Myc proteins (Prendegast and Ziff 1989).

3) The Zinc Finger motif.

Many classes of protein have been identified complexed with metal ions (Berg, 1986). A major group of DNA-binding proteins co-ordinate with zinc ions. These can be divided into two main classes: those that employ four cysteine residues to co-ordinate a zinc ion, the C_x zinc finger proteins, and those that employ two cysteine and two histidine residues to co-ordinate a zinc ion, the C_2H_2 zinc finger proteins. There are also a smaller number of proteins that employ various combinations of cysteine and histidine residues, examples being the *Drosophila ter* protein, the retroviral nucleic acid binding proteins and one finger of the yeast SWI5 protein which have C_2HC fingers (Baldarelli *et al.*, 1988, Covey, 1986, Stillman *et al.*, 1988, Nagai *et al.*, 1988) and the *E. coli* gene 32 protein which has a CHC_2 motif (Giedroc *et al.*, 1986).

The C_x Class of DNA-binding proteins.

This class essentially comprises of proteins containing a variable number of conserved cysteines. This group is most familiarly represented by the C_2C_2 steroid and thyroid hormone receptor family which have two apparently unrelated fingers encoded by separate exons (Huckaby *et al.*, 1987). The zinc ion requirement for DNA-binding has been demonstrated for the oestrogen receptor (Sabbah *et al.*, 1987). The most striking feature of these proteins is the conservation of cysteine residues which apparently fold to co-ordinate a zinc ion in a region responsible for the DNA-binding specificity (Green *et al.*, 1988). This conserved region has been used to identify other members of this family an example being the human retinoic acid

receptor protein (Petkovich *et al.*, 1987). There are also a number of proteins shown to require zinc ions for DNA-binding that have 6 cysteine residues as exemplified by the GAL4 protein, a yeast transcriptional activator (Laughon and Gesteland, 1984, Johnson, 1987). The DNA-binding specificity of GAL4 has been shown to lie in sequences adjacent to the zinc finger region (Corton and Johnson, 1989). More recently, other cysteine-rich motifs defining families of C_x proteins have been proposed (Freemont *et al.*, 1991 and Opipari *et al.*, 1990) indicating that, as more sequence information is obtained, proteins employing only cysteine residues to co-ordinate zinc will be shown to comprise several different families defined by structurally distinct C_x motifs.

The C_2H_2 Class of DNA-binding proteins.

An increasing number of proteins have been shown to contain repeating units of two cysteines followed by two histidines. This motif has been observed in a variety of organisms from yeast (the *ADR1* gene, Blumberg *et al.*, 1987) and trypanosomes (the *TRS-1* gene, Pays and Murphy 1987) to man (the *Sp1* gene, Kadonaga *et al.*, 1987) and with a variety of functions from transcriptional activation to a purely structural role. The number of fingers varies from two in ADR1 (Blumberg *et al.*, 1987) to 37 in the protein product of the *Xenopus Xfin* gene (Ruiz i Altaba et al., 1987). Again, the four crucial amino acids co-ordinate a zinc ion.

1.9.4 The C_2H_2 Zinc Finger Motif.

The C_2H_2 motif was first observed in the *Xenopus* protein, *Transcription factor IIIA* (*TFIIIA*) by Miller *et al.*, 1985. *TFIIIA* is part of the initiation complex of 5S RNA genes and is found in large quantities stored as a 7S particle with the 5S RNA in the ovaries of immature frogs. The amino acid sequence contains nine repeating units of about 30 amino acids with invariant cysteine and histidine residues that have been shown to co-ordinate around a zinc ion (Diakun *et al.*, 1986) and an invariant phenylalanine and leucine residue, the consensus amino acid sequence being Cys-X_{2-4}-Cys-X_3-Phe-X_5-Leu-X_2-His-X_{3-4}-His-X_6, as shown in figure 1.1(b). The 3

dimensional tertiary structure of the 31st finger of *Xfin*, was determined in the nuclear magnetic resonance (NMR) studies by Lee *et al.* (1989) using a synthetic 25-residue peptide: Y-K-C-G-L-C-E-R-S-F-V-E-K-S-A-L-S-R-H-Q-R-V-H-K-N. The overall fold of the zinc finger consisted of a hairpin structure resulting from an antiparallel β interaction from residue 1 to 10 and a helix from residue 12 to 24 which is α helical at the beginning and 3_{10} helical toward the end, in a globular structure with the exposed face of the helix, where there are four basic amino acids, proposed to contact the DNA in the major grove (Fairall *et al.*, 1986). The amino acids forming a link, between two finger regions were not included in this study. Figure 1.1(a) is a representation of the three-dimensional structure of the C_2H_2 domain adapted from Crossley (1990).

1.9.5 The function of C_2H_2 Zinc finger proteins.

The known functions of members of the ZFP gene family can be divided into four categories:

1) Transcriptional activators: examples being TFIIIA (Engelke *et al* 1980), Krox-20 and Krox-24 (Chavrier *et al.*, 1990), ADR1 (Hartshorne *et al.*, 1986).

2) Transcriptional repressors: examples being the *Drosophila* Krüppel protein involved in zygotic segmentation (Tautz *et al.*, 1987) and the yeast MIG1 protein involved in glucose repression (Nehlin and Ronne 1990).

3) RNA transport proteins: examples being TFIIIA which is stored with 5S RNA in frog oocytes (Pelham and Brown, 1980 and Honda and Roeder, 1980) and p43 which is part of the 42S ribonucleotide storage particle binding exclusively to 5S RNA (Joho *et al.*, 1990)

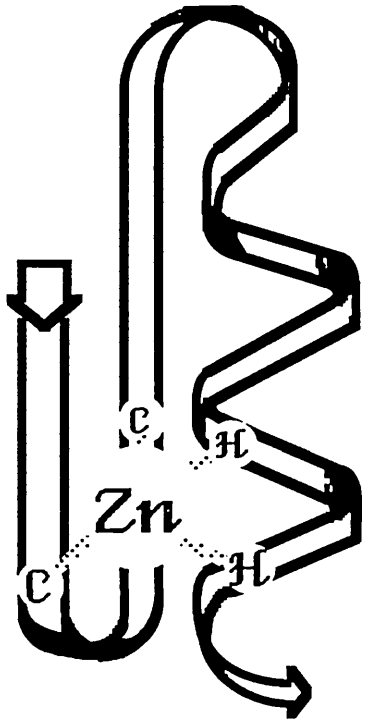
4) Structural support proteins: an example being Suvar(3)7 which has a dose limiting effect on position-effect variegation which is the inactivation in some cells of a gene translocated next to heterochromatin. This protein has a novel arrangement of widely spaced fingers and is thought to help in the packaging of chromatin fibre into heterochromatin (Reuter *et al.*, 1990).

Figure 1.1 Diagrammatic representation of three dimensional structure of the zinc finger domain.

(a) The figure is modified from Crossley (1990). The diagram represents 24 amino acid residues not including the residues, GEKP, which join consecutive domains. The zinc ion (Zn) and co-ordinating cysteine (C) and histidine (H) residues are indicated.

(b) The consensus ZF motif is shown (single letter amino acid code) with the positions of residues involved in DNA-binding specificity indicated by arrows and the sequences comprising the 144-fold redundant ZF oligonucleotide probe shown beneath the H/C link region.

(a)



(b)

1	2	3	4	5	6	7	8	9	10	11	12	13*	14	15	16*	17	18	19	20	21	22	23	24	25	26	27	28
C	x	x	C	x	x	x	F	x	x	x	x	x	L	x	x	H	x	x	x	H	T	G	E	K	P	Y	x

CAC	ACA	GGA	GAA	AAA	CCT	TA
T	T	T	G	G	C	
		G			A	

As regulators of gene expression ZF proteins have also been shown to be involved in development and implicated in tumourogenicity. Genetic mutations have defined several *Drosophila* zinc finger proteins essential for the establishment of body pattern, the genes for which were isolated using classical genetical approaches. Examples of these are *Krüppel*, *snail*, *serendipity beta* and *delta*, and *glass* (Rosenberg *et al.*, 1986, Boulay *et al.*, 1987, Vincent *et al.*, 1985, Moses *et al.*, 1989). Fasano *et al.*, 1991 have isolated a further regulatory gene containing the zinc finger motif, *teashirt*, on the basis of reporter gene expression. Examples of possible tumourogenic genes are *GLI*, a gene amplified in a subset of human brain tumours (Kinzler *et al.*, 1987, 1988), *MOK-2*, a gene encoding a protein almost solely consisting of fingers which is preferentially expressed in transformed mouse cells but, unlike *GLI*, is also expressed in at least two adult tissues (Ernoult-Lange *et al.*, 1990), and *EVI-1*, a gene activated in several retrovirus-transformed myeloid leukemia cell lines where integration of the virus at this site causes the activation of transcription (Morishita e*t al.*, 1988). Also, the candidate gene for susceptibility to Wilms' tumour has four copies of the zinc finger motif (Call *et al.*, 1990 and Gessler *et al.*, 1990). There are now many examples of zinc finger genes with specific patterns of expression: the mouse *Zfp-2* (*mkr2*) gene which is neuron specific (Chowdhury and Gruss 1988), *Egr-1* which is a mouse early growth response gene induced by nerve growth factor (Sukhatme *et al.*, 1988) (also called *NGF1-A*, Milbrandt 1987, *zif-268*, Christy *et al.*, 1988 and *Krox-24*, Lemaire *et al.*, 1988) and *EGR2*, the human homologue of *Krox-20* (Joseph *et al.*, 1988). The specificity of their expression implies a role in the development or functioning of that tissue. Recently putative homologues for the body pattern formation genes of *Drosophila* have been screened for in other species: *xsna* was isolated from Xenopus using *snail* as a probe (Sargent and Bennet, 1990). Both these genes are expressed at sites of involution and of presumptive mesoderm formation in early embryos implying a similar role in dorsoventral pattern determination.

1.9.6 Zinc Finger Gene Families.

The ZF motif may represent a unique solution to a biological problem that has evolved in many different instances. However, there is evidence for a closer relationship between many of the ZF genes analysed. Knochel *et al.* (1989) have identified a common conserved module, the FAX domain, at the N-terminal end of specific *Xenopus* genes which is unrelated to the finger domain and is associated with genes that are differentially expressed in the course of early *Xenopus* embryogenisis and which is also evolutionary conserved. Bellefroid *et al.*, (in press, Biochemistry) have identified a similarly conserved domain, the KRAB domain, which is also associated with developmentally regulated genes and which is different from the FAX domain. It is possible that these indicate the existence of structurally, and probably functionally related subfamilies in ZFP family. The recent duplication of the *Drosophila serendipity beta* and *delta* genes (Vincent *et al.*, 1985) and the divergence of expression patterns of the *hunchback* gene in two Drosophila species (Treier *et al.*, 1989) point a possible evolutionary history in the ZF gene family of duplication events and subsequent functional diversifications. The mouse *ZFa* is an example of a zinc finger gene that has arisen by retroposition (Ashworth *et al.*, 1990) and represents an example of another mechanism by which a large number of related genes could have evolved.

1.9.7 Binding sites.

Two models have been proposed for the *TFIIIA*-DNA interaction:

1) The "wrapping around" model where successive fingers contact the major groove without crossing the minor groove (Berg 1988). In this model one finger would contact 5 bases.

2) The "alternating" model where alternate fingers bind on one face of the DNA in an equivalent manner in the major grove so that successive minor grooves must be crossed (Fairall *et al.*, 1986). In this model one finger would interact with 3 bases.

The large amount of data generated from protection experiments, NMR structure determination and three dimensional structural modeling has yet to resolve this problem. Genetic and biochemical studies have revealed the zinc finger region as being responsible for sequence specific binding (Blumberg *et al.*, 1987, Redemann *et al.*, 1988 and Nardelli *et al.*, 1991) and binding sites for the products of *ADR1*, *SW15*, *MIG1*, *Krüppel*, *Hunchback*, *Suppressor of Hairy Wing*, *zif-268* (also known as *NGF1-A*, *EGR-1* and *Krox-24*), *Krox-20*, *Sp1* genes and the candidate Wilms' Tumour gene have been identified (Eisen *et al.*, 1988, Nagai *et al.*, 1988, Nehlin and Ronne, Stanojevic *et al.*, 1989, Treisman and Desplan, 1989, Spana *et al.*, 1988, Christy and Nathans, 1989, Chavrier *et al.*, 1988a, Kadonaga *et al.*, 1987, Rauscher *et al.*, 1990). It has been suggested that zinc fingers span either 5.5bp based on the periodicity of the recognition sites (Rhodes and Klug, 1986) or 2-3bp based on molecular modelling (Gibson *et al.*, 1988 and Berg, 1990). Nardelli *et al.* (1991) have used *in vitro* directed mutagenisis studies and NMR studies to show that each Krox-20 finger spans 3bp, with amino acids at positions 13 and 16 (figure 1.1(b)) providing base discrimination to either 2 or 3 bp.

Two zinc finger proteins have been shown to bind RNA: TFIIIA (Honda and Roeder, 1980) and p43 (Joho *et al.*, 1990) both bind 5S RNA although, unlike TFIIIA, p43 does not bind the 5S RNA gene. The only sequence homology between these proteins is at residues known to be involved in the putative zinc finger structure although p43 lacks the conserved aromatic amino acid that usually precedes the first cysteine of each finger.

1.9.8 Detecting New ZFP Genes.

After the DNA-binding activity of the zinc-finger domain in the TFIIIA protein was demonstrated (Pelham and Brown, 1980 and Honda and Roeder, 1980), the Cys-X_{2-4}-Cys-X_3-Phe-X_5-Leu-X_2-His-X_{3-4}-His-X_6 motif was recognized in other proteins known to bind DNA, the first example being Krüppel (Rosenberg *et al.*, 1986, Ollo and Maniatis, 1987). The only example as yet of a protein with the C_2H_2

motif that has been shown not to bind DNA is p43 which binds RNA (Joho *et al.*, 1990). It is probable that all the proteins containing this motif will be shown to have a nucleic acid-binding activity, the majority of them binding DNA. The function of these proteins as controllers of transcription has made them an important group to study specifically with respect to developmental genes and genes with a role in malignancy.

In order to isolate other genes, which by sharing homology might also share similar developmental function, groups have used part or all of a *Krüppel* cDNA as a probe under conditions of low stringency to isolate other genes. Schuh *et al.*, 1986, isolated related sequences from *Drosophila* and Chowdhury *et al.*, 1987 isolated two related genes from mouse. Both groups also looked at the distribution of these sequences among other species detecting multiple *Kr*-homologous DNA fragments in all the eukaryotes tested but not in bacteria. The sequences they isolated were not homologues of the *Krüppel* gene but contained the zinc finger motif: Cys-X_{2-4}-Cys-X_3-Phe-X_5-Leu-X_2-His-X_{3-4}-His-X_6. Comparisons of the predicted amino acid sequence of fingers within these genes and to the *Krüppel* gene revealed an additional shared region of homology between the last histidine of one finger and the first cysteine of the second finger, the "H/C link". The H/C link is shown in figure 1.1 (b). This region of six amino acids is shared by all the fingers of the *Krüppel* gene, the *Kr h* gene and the two mouse genes, *mkr*1 and *mkr*2. The consensus sequences for these genes are shown in table 1.1, taken from the papers by Chowdhury *et al.*, 1987 and Schuh *et at.*, 1986. The conservation of these residues is likely to result from constraints imposed by the folding of these proteins.

Although a number of groups have used finger-motif-encoding cDNA fragments in low stringency screens of genomic and cDNA libraries to detect related sequences (Schuh *et al.*, 1986, Chowdhury *et al.*, 1987, Nietfeld *et al.*, 1989, Knochel *et al.*, 1989, and Thiesen, 1990 are examples) the existence of the conserved H/C sequence makes the detection of genes of this class (commonly called *Krüppel*-related genes although their relationship to *Krüppel* is purely incidental) a relatively simple process

using oligonucleotides. A variety of oligonucleotides have been used to detect the zinc finger motif in many species (Bellefroid *et al.*, 1989, Cunliffe *et al.*, 1990, and Ernoult-Lange *et al.*, 1990 and as reported by Bray and Thiesen, 1990). In table 1.2 the presence of the H/C link in 23 members of the zinc finger protein gene family is shown. These genes were not isolated by homology screens with zinc finger gene probes but were all isolated independently. If these are taken to be representative of the nature of this family as a whole, then overall 52% of this family contain at least one copy of the TGEKPY motif and could theoretically be detected with an oligonucleotide to this sequence. A 144-fold redundant 20-base oligonucleotide was used in the work presented in this report. The sequence is shown in figure 1.1 (b). The requirement for the abundant zinc finger genes, and the processes by which these numbers have originated will be clarified as more information is obtained on sequence and function.

1.10 Status of the 11p mapping Project.

E65-9 mapping project.

The mapping of the cosmid clones containing human DNA by the fingerprinting reaction was performed on 112 clones isolated from a HindIII partial E65-9 cosmid library and 1 clone containing part of the H-*ras* gene isolated from a Sau3A partial E65-9 library (Harrison-Lavoie *et al.*, 1989). Two large contigs were constructed, A and B, containing 48 and 16 clones respectively, and three small contigs containing 2-3 clones. The remaining 43 clones were unattached.

Table 1.1 Consensus amino acid sequence of the first 7 "zinc finger" proteins.

(a) Comparison of the amino acid consensus sequences known in 1987 taken from the papers by Chowdhury *et al.*, 1987 and Schuh *et at.*, 1986, and (b) the fingers of TFIIIA and ADR1 with a conserved H/C link sequence similar to the consensus - TGEKPY.

(a)

Protein	Consensus sequence
Kruppel	C XX C XXX F XXXXX L XX H XXX H TGEKPYX
Kr h	C XX C XXX F XXXXX L XX H XXX H TGEKPYX
mkr1	C XE C GKX F XXXSX L IX H QRI H TGEKPYX
mkr2	C XE C GKA F XXXSS L XX H XRI H TGEKPYX
sry	C XX C GKX F SXXXX L XX H MOX H XXXXXXXX
TFIIIA	C XXDG C DKR F TKKXX L KXR H XXXX H XXXXXXYV
ADR1	C XX C XRX F XRXXX L XXR H XXXX H XXXXXYX

(b)

Protein	Amino acid sequence of individual fingers
TFIIIA	C SFAD C GAA Y NKNWK L QA H LCK H TGEKPFP
	C KEEG C EKG F TSLHH L TR H SLT H TGEKNFT
ADR1	C EV C TRA F ARQEH L KR H YRS H TNEKPYP

Table 1.2 Presence or absence of the H/C link in 23 independantly isolated H_2C_2 zinc finger proteins.

The number of copies represents the number of links with the amino acid sequence HTGEKPY, or with six of the seven residues of this consensus link.

Name of sequence	Motif	Reference
ADR1	1	Harshorne et al., 1986.
SWI5	0	Stilman et al., 1988.
MIG1	1	Nehlin and Ronne 1990.
TRS-1	0	Pays and Murphy 1987.
sry β	0	Vincent et al., 1985.
sry δ	0	Vincent et al., 1985.
Kr	4	Rosenberg et al., 1986.
hb	0	Tautz et al., 1987.
snail	1	Boulay et al., 1987.
su(Hw)	3	Parkhurst et al., 1988.
glass	4	Moses et al., 1989.
CF2	2	Shea et al., 1990.
suvar(3)7	0	Reuter et al., 1990.
odd	0	Coulter et al., 1990.
teashirt	0	Fasano et al., 1991.
TFIIIa	2	Miller et al., 1985.
p43	0	Joho et al., 1990.
NGF1-A	2	Milbrandt 1987.
Evi-1	0	Morishita et al., 1988.
ZFY	1	Page et al., 1987.
Gli	2	Kinzler et al., 1988.
Sp1	2	Kadonaga et al., 1988.
H-plk	13	Kato et al., 1990.

EJNAC4.1 and EJNAC4.1.5 mapping project.

Cosmid clones containing human DNA isolated from cosmid libraries of the EJNAC4.1 and EJNAC4.1.5 cell lines are being assembled into contigs as part of the on-going process of physically mapping the short arm of human chromosome 11. At the stage of submission of this work 4428 cosmids have been entered into the data base. 531 contigs have been assembled with a mean occupancy of 3.4 and there are 2445 unattached clones.

1.11 Aims and Objectives.

The fundamental aim of this project was the detection of coding sequences within cloned DNA arrays generated in the physical mapping of the short arm of human chromosome 11. The detection of coding sequences by their association with rare-cutting restriction enzyme sites is described in chapters 3 and 4. The analysis of H_2C_2 zinc finger sequences possessing the TGEKPY motif is described in chapters 5, 6, 7 and 8.

Chapter 2:

Materials and Methods.

2.1 Bacterial Strains and Cloning Vectors.

Plasmids and lambda bacteriophage were propagated in the following *E.coli* strains:

JM101: supE, thi, Δ(lac$^-$proAB$^+$), F' [traD36, proAB$^+$, lacI^q, lacZ ΔM15] (rk$^+$, mk$^+$), mcrA$^+$ (Yanisch-Perron *et al.* 1985).

XL1-Blue: supE44, hsdR17, recA1, endA1, gyrA96, thi, relA1, lambda$^-$, lac$^-$, F'[proAB$^+$, lacI^q, lacZ, ΔM15, Tn10, tet^r] (rk$^-$, rk$^+$) (Bullock *et al.* 1987).

MM294: supE44, hsdR17, endA1, lambda$^-$, pro, thi, F$^-$, (rk$^-$, mk$^+$) mcrA$^+$, mcrB$^+$ (Hanahan, 1983).

NM514: (rk$^-$, mk$^+$) derivative of POP101 (Lecocq and Gathoye, 1973).

Cosmid clones were propagated in *E.coli* strains ED8767 and 1046: supE44, supF58, hsdS- ($r_B^-m_B^-$), recA56, galK2, galT22metB1. (Murray *et al.*, 1977).

Plasmid vectors used for general cloning were pUC12 (Viera and Messing, 1982) and Bluescript pks$^+$ (Short *et al.*, 1988, and the product literature from Stratagene cloning system). Cosmid vectors used were Lorist B (Little and Cross, 1985, Gibson *et al.*, 1987), Lorist X and Lawrist 4 (derived from Lorist B and Lorist 6 by P. de Jong). M13tg130 and M13tg131 (Kieny *et al.*, 1983) were used for the production of single stranded DNA for sequencing.

2.2 Bacterial Media and Standard Solutions.

Liquid cultures of *E.coli* were grown in Luria broth (L-broth) as described in Maniatis *et al.* (1982). L-plates were made by the addition of 15g per litre of Bacto-agar (Difco) to L-broth and autoclaving. M13 and lambda bacteriophage were plated in a 3ml overlay of top-agar (0.7%) or top-agarose (0.7%) respectively onto 80mm Sterilin petri dishes unless otherwise stated. 10μl of 0.1M isopropylthio-β-D-galactoside (IPTG) and 25μl of 0.2% 5-bromo-4-chloro-3-indolyl-β-D-galactoside (Xgal) in dimethylformamide were added to the agar plate or to the overlay of top-agar, where blue-white selection of recombinants with insertional inactivation of the lacZ gene was required. In order to select *E.coli* transformed by antibiotic-resistant extra-chromosomal DNA, 30μg/ml and 100μg/ml of kanamycin or ampicillin was

added to media or plates as appropriate. cDNA libraries in phage vectors were plated in an overlay of top-agarose and clones were stored in SM phage buffer: 10mM TRIS-HCl pH7.5, 10mM MgSO4, 50mM NaCl, 0.001% gelatin.

1M phosphate buffer was made by mixing 1M Na_2HPO_4 and 1M NaH_2PO_4 in a ratio of 23/17 to give a pH of 6.5. 1M "NaPi" was made by adjusting the pH of 500ml of 1M Na_2HPO_4 to 7.2 with orthophosphoric acid (H_3PO_4, BDH) and making up the volume to 1l with dH_2O. All other solutions and buffers were made as described in Maniatis *et al.* (1982).

2.3 Preparation of DNA:

2.3.1 Small Scale Preparation of Cosmid and Plasmid DNA.

1.5ml of bacterial culture was used in the small scale plasmid and cosmid DNA preparations which were performed as described by Little 1987. These preparations are referred to as "minipreps".

2.3.2 Large Scale Preparation of Cosmid and Plasmid DNA.

Plasmid and cosmid DNA was isolated from 250ml overnight cultures as described by Little 1987. These preparations are referred to as caesium-pure DNA preparations.

2.3.3 Small Scale Preparation of Phage DNA.

Bacteriophage DNA was prepared as described by the Amersham cDNA cloning manual.

2.3.4 Preparation of Single stranded M13 phage DNA.

M13 plaques were picked into 2ml of L-broth media supplemented with 20μl of a saturated culture of XL1 cells and incubated at 37^{o}C for 6 hours. DNA was prepared from this culture as described by the Sequenase (USB) Sequencing Manual. The double-stranded replicating form of the phage was prepared simultaneously from the bacterial cell pellet by the "miniprep" method.

2.3.5 Preparation of Chromosomal DNA from Tissue Culture Cells.

Somatic cell hybrids were grown in Dulbecco's modified eagles medium supplemented with 2mM glutamine and 5% foetal calf serum in the presence of streptomycin (200u/ml) and penicillin (50u/ml) at 37°C in 20% CO_2 until confluent, in 10 175cm^2 flasks. Cells were harvested by trypsinisation for 2-3 minutes followed by addition of media and centrifugation to collect the cells. The pellet was resuspended in 5-10mls of 1 x SSC followed by addition of an equal volume of lysis buffer (100mM TRIS/100mM NaCl/10mM EDTA/1% w/v sarkosyl). After 5 minutes, proteinase K was added to a final concentration of 0.1 mg/ml and the mixture incubated for 2 hours at 55°C. This was followed by one phenol, one phenol/chloroform and one chloroform extraction and the DNA was dialysed for 24 hours in TE plus 100mM NaCl and a further 24 hours in TE. A final concentration of 100-200μg/ml was generally obtained and the DNA was stored at 4°C.

2.3.6 Preparation of Large Numbers of Cosmid DNAs using a multiwell system.

Preparation of cosmid DNA on a small scale was performed on 250μl of cultures from large numbers of cosmid clones grown in 1ml tubes as described by Gibson and Sulston (1987) by alkaline lysis using a microtitre system. The average yield for each preparation was 0.5-1μg of DNA which was resuspended in 10μl of TE. These preparations are referred to as "micropreps".

2.3.7 Preparation of DNA fragments from Agarose.

DNA fragments were isolated from agarose gels by excision of bands visualised under UV and treatment as described in the GENECLEAN II Kit BIO 101 Inc. manual to purify the DNA.

2.3.8 Amplification of DNA by Polymerase Chain Reaction (PCR).

PCR was carried out in a total volume of 100μl with 1μg of genomic DNA or 2-10ng of purified cloned DNA, 10-100pmol of each primer, in 10mM Tris-HCl (pH 8.3), 50mM KCl, 1.5mM $MgCl_2$, 0.001% (w/v) gelatin (Sigma), 0.2-1mM each dNTP, with 2.5 units Taq polymerase (Perkin Elmer Cetus) for 35 cycles of 1 minute at 94^{o}C (denaturation reaction), 1 minute at 50-54^{o}C (annealing temperature, dependant on primer used) and 1 minute at 72^{o}C (extension reaction) followed by 7 minutes at 72^{o}C as a final extension reaction, in a Flowgen Coy tempcycler. Primers were synthesized using a Pharmacea LKB Gene Assembler Plus and were used in reactions after deprotection by incubating the support cassette in 35% ammonia (BDH) for 16 hours at 55^{o}C in a screw cap eppendorf, followed by precipitation of the DNA by addition of 2.5 volumes of 96% ammonia and 0.1 volumes of 3M Na Acetate (pH5.5), recovery by centrifugation in the Heraeus Sepatech Biofuge A microfuge, washing the pellet in 80% ethanol and resuspending it in T.E (pH7.5).

2.3.9 Preparation of Phage insert DNA by PCR.

Both purified phage clones and primary mixed picks of positive clones from the lambda gt10 library were amplified directly by PCR in the following way: 1μl of the phage stock was added to a 0.5ml eppendorf tube containing 1mM of each dNTP, 1 x reaction buffer (10 x reaction buffer is 100mM TRIS-HCl pH8.3, 500mM KCl, 15mM $MgCl_2$ and 0.01% (w/v) gelatin), 1.0μM PL1 primer, 1.0μM PL3 primer and 2.5 units Taq polymerase in a volume of 24μl and overlain with 25-35μl Sigma light mineral oil. The reactions were run on 30 cycles of 1 minute at 94^{o}C/1 minute at 50^{o}C/2 minutes at 72^{o}C followed by a final elongation step of 7 minutes at 72^{o}C. The mineral oil was removed by extraction with an equal volume of chloroform and 1/20th of the sample resolved on an agarose gel. Where mixed phage stock was the template the resolved products of the amplification were transferred to biodyne A filters and hybridised to the appropriate radioactive probe to determine the size of the insert.

PL1 = 5' ATGAGTATTTCTTCCAGGGT 3'

PL3 = 5' CAAGTTCAGCCTGGTTAAGT 3'

2.4 Phenol extraction of DNA.

Contaminating proteins were removed from DNA solutions by the addition of an equal volume of neutralised phenol, vortexing and separation by centrifugation in a microfuge (Heraeus Sepatech Biofuge A). The aqueous layer was extracted with an equal volume of 1:1 phenol:chloroform and with an equal volume of chloroform.

2.5 Precipitation of DNA.

DNAs were precipitated from solution either by the addition $1/10^{th}$ volume of 3M NaCl or 3M Na-acetate (pH 5.0) and 2 volumes of 96% ethanol and freezing, and recovered by centrifugation either in a microfuge or in a DuPont RC-5 Superspeed Refrigerated Centrifuge using the Sorvall HB-4 rotor. Pellets were washed in 70% ethanol and air dried before being resuspended in the appropriate volume of TE (pH7.5) by heating in a $55^{o}C$ water bath for 10 minutes.

2.6 Endonuclease Digestion of DNA.

Restriction enzyme digests were obtained from New England Biolabs Inc. or Boehringer Mannheim and performed according to manufacturers instructions. Where difficulties were experienced digesting specific DNA preparations spermadine was added to a final concentration of 1mM.

2.7 Exonuclease Digestion of DNA.

Double-stranded DNA was progressively shortened from both ends using Bal31 nuclease to create a series of approximately 250bp deletions to allow sequencing of DNA not containing convenient restriction sites. The DNA was cloned into the unique EcoR1 of Bluescript pks^{+} and 50μg of the construct linearised with restriction enzymes recognising sites either side of the insert site, these being BamH1

and EcoRV. The linearised DNA was digested in a volume of 200μl using 15 units of Bal31 in a buffer containing 12mM $CaCl_2$, 12mM $MgCl_2$, 0.2M NaCl, 20mM TRIS-HCl (pH8.0) and 1mM EDTA (pH8.0), 41.6μl aliquots being removed at appropriate time points (0, 30, 60, 90, 180 and 300 seconds) into 5μl of 250mM EGTA, a calcium ion-specific chelator. The DNA was digested with EcoR1 by the addition of 114.8μl of dH_2O and 10μl of the restriction enzyme at 10 units/μl to release the shortened products and the insert was separated from the vector by electroporation on a 0.5% LMP agarose gel. The deleted inserts were ligated into M13tg130 digested with EcoRI and HindII and sequenced.

2.8 Ligation of DNA.

Ligations were carried out in 20-30μl volumes either at room temperature for 1 hour (cohesive end ligations) or at 14°C for 14 hours (blunt end ligations). Ratios of between approximately 1:1 to 1:5 vector to insert were used. 1-2 units of T4 DNA ligase from Boehringer Mannheim was used in the reaction with the buffer provided by the manufacturer.

2.9 Transformation of E.coli and Selection for Recombinants.

2.9.1 Preparation of Competent E.coli.

E.coli cells were made competent by the method described in Maniatis *et al* (1982). An overnight culture of the appropriate strain of *E.coli* was diluted 100-fold and grown at 37°C until an OD_{600nm} of 0.4-0.6 units was reached. The cells were pelleted at 2500rpm for 10 minutes at 4°C in a Sorvall RT6000B refrigerated centrifuge (DuPont), resuspended in $1/10^{th}$ volume of ice-cold 50mM $CaCl_2$ and incubated on ice for 30 minutes. After recovering the cells by repeating the centrifugation, they were resuspended in $1/100^{th}$ volume of cold 50mM $CaCl_2$.

2.9.2 Transformation of Competent E.coli.

$1/10^{th}$ and $9/10^{th}$ of a ligation reaction or the appropriate amount of supercoiled DNA was added to 100μl of the competent cells and incubated on ice for 30 minutes followed by heat shocking for 2 minutes at 42^{o}C. In the case of plasmids, 1ml of pre-warmed L-broth was added to the cells and these were incubated at 37^{o}C for 1 hour to allow for the expression of antibiotic resistance before being centrifuged to pellet the cells. Pellets were resuspended in 100μl of L-broth and spread onto pre-dried agar plates containing the appropriate antibiotic. M13 transformations were plated immediately in 3ml top-agar to which 100μl of an overnight culture of cells had been added to provide a bacterial lawn.

2.9.3 Preparation of E.coli for Electroporation.

XL1's were prepared as described by the Bio-Rad Gene Pulser manual and were stored at -70^{o}C in 10% glycerol in water.

2.9.4 Transformation of E.coli by Electroporation.

1-1.5μl of the ligation was added to 40μl of the prepared cells in the cold 0.2cm electroporation cuvette (Biorad). This was subjected to 2.5kV at a capacitance of 25μFD and a resistance of 200 Ohms in the Biorad Gene Pulser before immediate transfer into prewarmed L-broth. The cells were incubated at 37^{o}C for an hour before being plated out.

2.9.5 Identification of Recombinants by Blue/White Selection.

Cloning of insert DNA into the polylinker site of the plasmids pUC12 and Bluescript, and the M13 bacteriophage vectors M13tg130 and M13tg131 disrupts the alpha section of the *E.coli lacZ* gene and therefore recombinants are detected by absence of staining with the chromogenic beta-galactosidase substrate 5-bromo-4-chloro-3-indolyl-β-D-galactoside.

2.9.6 Identification of Recombinants by Hybridisation.

Colony lifts and plaque lifts were performed using Biodyne A (1.2 micron) as described in the Pall Biodyne procedures manual. Filters were baked for 1 hour and fixed by exposure to 0.16kJ/m2 of short wave UV (Church and Gilbert 1984), before being prehybridised and hybridised in the appropriate buffers and radiolabelled probes. After hybridisation the filters were washed in the appropriate solutions and exposed with a Cronex Quanta III intensifying screen to Kodak XAR5 film. The resulting positive spots could be orientated with respect to the original plate by means of location marks made using a needle.

2.10 Radioactive Labelling of DNA.

2.10.1 Labelling Double Stranded DNA using Random Primers and Klenow.

Oligo-labelling of DNA was performed as described by Feinberg and Vogelstein (1983) using a Random Primed DNA Labelling Kit obtained from BCL.The labelled DNA was separated from the unincorporated nucleotide by running it through a column made from Sephadex G-50 which had previously been autoclaved in several volumes of elution buffer (100mM TRIS-HCl pH7.5/ EDTA pH8.0/150mM NaCl). Activities of 1-3 x 10^9cpm/μg DNA were obtained and filters hybridised in 0.5-1 x 10^6 cpm radiolabelled probe per millilitre of hybridisation buffer.

2.10.2 Labelling Single Stranded DNA Using T4 Polynucleotide Kinase.

Oligonucleotides were end-labelled in 1 x PNK buffer (500mM TRIS-HCl pH7.5/100mM $MgCl_2$/50mM DTT/1mM Spermadine). 200ng of the Zinc finger 20 base oligonucleotide was labelled in a total volume of 50μl containing 20 units of T4 Polynucleotide Kinase (Pharmacea) and 200μCi of [γ^{32}P]dATP for 1 hour at 37^{o}C. 10ng of the 10 base PL BSSH and PL EAG oligonucleotides were labelled in a total volume of 10μl containing 2.5 units of T4 Polynucleotide Kinase (Pharmacea) and 40μCi of [γ^{32}P]dATP for 1 hour at 37^{o}C. The labelled oligonucleotides were separated from the unincorporated nucleotide by passing through a Sephadex G-25

column equilibriated in elution buffer. Specific activities of 2-3 x 10^9cpm/μg were obtained.

2.11 Gel Electrophoresis of DNA.

2.11.1 Agarose Gel Electrophoresis.

Agarose gel electrophoresis of DNA was carried out in horizontal slab gels of 0.2-1% (w/v) agarose (FMC Bioproducts) at 7V/cm for the required length of time in 1 x E buffer supplemented with ethidium bromide (0.5μg/ml). To resolve fragments of less than 500bp, DNA was run out on 2-3% FMC Bioproducts NuSieve GTG agarose gels. DNA size markers used were BRL 1kb ladder (0.075, 0.142, 0.154, 0.2, 0.22, 0.298, 0.344, 0.394, 0.506, 0.516. 1.018, 1.635, 2.036, 3.054, 4.072, 5.09, 6.108, 7.126, 8.144, 9.162, 10.18, 11.198 and 12.216kb fragment sizes) and/or lambda digested with HindIII (.564, 2.027, 2.322, 4.361, 6.682, 9.416 and 23.13kb fragment sizes).

2.11.2 Field Inversion Gel Electrophoresis (FIGE).

FIGE (Carle *et al.*, 1986) was carried out in horizontal slab gels of 0.8-1% agarose in 0.5 x TBE for the required length of time at 7V/cm under an inverting electric current directed by a Programmable Power Invertor-100, a microcomputer for running field inversion gels, software version 100.3 (M. J. Devices). Carle *et al.* (1986) achieved separation of fragments between 33 and 15 kb with pulse times of 0.5/0.25 seconds and separation of 48.5, 125 and 170kb with pulse times of 3/1 seconds. A timing was determined to give the best separation of DNA fragments between 48.5kb (uncut lambda) and 28.5 and 20.0kb (lambda cut with NaeI) of 1.25 second forward pulse and a 0.375 second reversed pulse. The buffer was circulated through a LKB Multitemp II Thermostatic Circulator at 10^oC. The gel was soaked in 2μg/ml ethidium bromide in 1 x TBE buffer for 30 minutes and the DNA visualised under UV.

2.11.3 Polyacrylamide Gel Electrophoresis.

Vertical denaturing gels were used to resolve the products of sequencing reactions and fingerprinting reactions with 6% and 4% (w/v) polyacrylamide respectively and 8M urea. The gels (32 x 41cm) were run in 1 x TBE at 80W (40V/cm).

2.12 Transfer of DNA to Solid Support.

2.12.1 Transfer of Gel Electrophoresed DNA.

Southern transfer of DNA resolved on a horizontal agarose gel was performed by modification of the standard procedures in Maniatis *et al.* (1982). After visualisation of the DNA by exposure to short wave UV and photographing alongside a ruler, the gel was soaked in 0.25M HCl for 15-20 minutes to partially depurinate the DNA, and soaked in 0.5M NaOH/1.5M NaCl for 2 times 15 minutes to denature the DNA. DNA was transferred Pall Biodyne A transfer membrane (1.2 micron) which had first been wet in dH_2O followed by denaturation buffer. After 15-20 hours of transfer the filter was neutralised in 50mM "NaPi" (pH 6.5) for 2 minutes, air dried for 30 minutes, baked for 1 hour at 80^oC and immediately fixed by exposure to 0.16kJ/m2 of short wave UV (Church and Gilbert 1984).

2.12.2 Transfer of Bacterial Colonies.

Bacterial colonies were transferred to either Biodyne A filters (1.2 micron, Pall Biodyne) or nitrocellulose filters (Schleicher and Schuell) by placing the filter on the surface of the agar plate for 1 minute, making orientation marks using a sterile needle and transferring the filter reverse side up to a second agar plate which was incubated for 3-6 hours to allow regrowth of the colonies. A second lift was taken from the same plate and also reincubated to provide a duplicate filter. The filters were laid colony side up onto two layers of Whatman 3MM presoaked in denaturation buffer (1.5M NaCl/0.5M NaOH), for 5 minutes, transferred to two layers of Whatman 3MM soaked in a neutralisation buffer (3M Na-acetate, pH 5.5) for 5 minutes, rinsed in 50mM "NaPi" (ph6.5) to remove cell debris and air dried for 30

minutes. The filters were baked at 80°C for 1 hour and in the case of the Biodyne A filters, UV fixed as previously described.

2.12.3 Transfer of Phage Plaques.

Plaque lifts were performed in duplicate as for colony lifts with the additional requirement of pre-cooling the agar plates to 4°C. Plaques were transferred to either Biodyne A filters (1.2 micron, Pall Biodyne) or nitrocellulose filters (Schleicher and Schuell) in duplicate. The filters were laid plaque side up onto two layers of Whatman 3MM presoaked in denaturation buffer (1.5M NaCl/0.5M NaOH), for 1 minute. The filters were transferred to two layers of Whatman 3MM soaked in a neutralisation buffer (3M Na-acetate, pH 5.5) for 5 minutes before being air dried for 30 minutes. The filters were baked at 80°C for 1 hour and in the case of the Biodyne A filters, UV fixed as previously described.

2.13 Hybridisation of Filter Immobilised DNAs.

2.13.1 Hybridisation with Total Genomic DNA in formamide.

Filters were prehybridised for a minimum of 6 hours at 42°C in a solution of 5 x SSC, 0.1% SDS, 2 x Denhardt's solution (100 x Denhardt's is 2% BSA, 2% Ficoll, 2% polyvinyl pyrrolidone, all from Sigma), 100μg/ml single-stranded salmon sperm DNA, 10μg/ml polyA and 10μg/ml polyC (Pharmacea). The radioactive DNA probe of sonicated total genomic DNA was boiled for 10 minutes and added to the hybridisation buffer to give 0.5-1.0 x 10^6cpm/ml and added to the filter. After 48-72 hours of incubation at 42°C the filters were washed for 3 times 20 minutes at room temperature in 1 x SSC/0.1% SDS followed by 2 times 30 minutes at 65°C before being exposed to Kodak XAR5 film at -70°C with a Cronex Quanta III intensifying screen.

2.13.2 Hybridisation with Specific DNA Fragments in Church buffer.

Filters were prehybridised as described by Church and Gilbert (1984) in a buffer of 0.5M "NaPi" (pH7.2)/7% SDS/1mM EDTA for 5 minutes at 65^{o}C. Hybridisation was performed under the same conditions but with the addition of denatured probe to 0.5-2 x 10^{6} cpm/ml for 17 hours. Filters were washed in a solution of 40mM "NaPi" (pH7.2)/1% SDS prewarmed to 65^{o}C for 2 times 5 minutes and for 30 minutes at 65^{o}C. Low stringency washing was done in a similar fashion but at a lower, specified temperature.

2.13.3 Hybridisation with the ZF Oligonucleotide.

Filters were prehybridised for a minimum of 3 hours at 37^{o}C in a buffer containing 10% formamide, 5 x Denhardt's solution, 5 x SSPE, 0.1% SDS and 100μg/ml sonicated and denatured salmon sperm DNA. Hybridisation was performed at 37^{o}C in the same buffer with the addition of a radiolabelled 20 base oligonucleotide probe to give 2 x 10^{6}cpm/ml for 17 hours. Filters were washed in 6 x SSC/0.1% SDS at 37^{o}C for 3 washes of 30 minute duration.

2.13.4 Hybridisation with the BssHII and EagI oligonucleotides.

Filters were prehybridised in church buffer to which 1% BSA had been added, overnight at 24^{o}C. 10ng of the 10-base oligonucleotides kinase labelled to 1-3 x 10^{9}cpm/μg was hybridised to filters in 5mls of church buffer (plus 1% BSA) for 17 hours at 24^{o}C. The filters were rinsed 3 times briefly in cold 6 x SSC followed by 2 half hour washes at 4^{o}C. They were exposed overnight.

Two 10 base oligonucleotides were used:

PL BSSH 3' N N N N G C G C G C 5'

PL EAG 3' N N N N C G G C C G 5'

(N = dGTP or dATP or dTTP or dCTP).

2.13.5 Stripping Biodyne Filters.

One of two methods was used to remove radiolabelled DNA probes from filters:

a) Filters were washed for 1 hour at 65°C in 50% formamide/10mM phosphate buffer (pH6.5).

b) Filters were washed in 200mM NaOH at room temperature for 20 minutes and neutralised in 50mM "NaPi" (pH 6.5).

2.14 DNA sequencing and Analysis.

DNA sequencing was performed by the dideoxy chain termination method using modified bacteriophage T7 DNA polymerase (Sequenase Version 2.0 United States Biochemical) according to the manufacturers instructions. Single-stranded DNA were generated from recombinant bacteriophage M13tg130 or M13tg131. Sequence from double stranded, supercoiled template was generated by denaturing 1-2μg of DNA in 0.1M NaOH for 5 minutes at room temperature followed by the addition of pH 7.5 NH_4-Acetate to 0.18M and immediate precipitation with four volumes of prechilled 95% ethanol. The DNA was recovered by centrifugation, washed in 70% ethanol, resuspended in 7μl of dH_2O and sequenced as normal.

Sequencing data from individual clones was assembled an analysed using the programs of Queen and Korn (1984) contained in the Microgenie software package (Beckman Instruments) on an IBM XT personal computer.

2.15 Cosmid Genomic Libraries.

2.15.1 Partial Digestion of Chromosomal DNA.

Chromosomal DNA was partially digested with either HindIII or Sau3A: A series of test digestions were performed to find the optimum conditions required using a constant time of digestion with a decreasing concentration of restriction enzyme (figure 2.1 (a)).

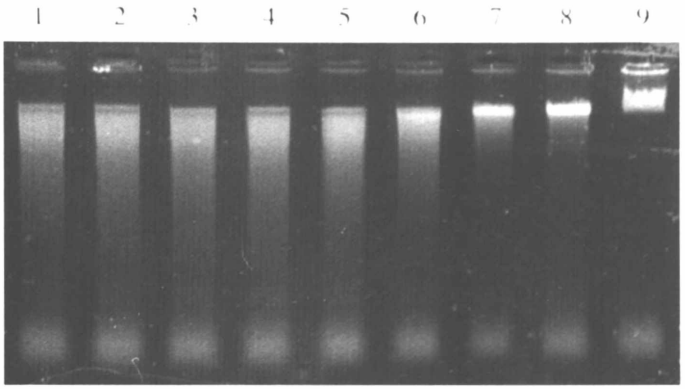
Figure 2.1 Construction of cosmid library.

(a) Nine samples of chromosomal DNA digested with (1) 0.2, (2) 0.1, (3) 0.05, (4) 0.025, (5) 0.0175, (6) 0.00875, (7) 0.0044, (8) 0.0022 and (9) 0.0011 units of HindIII for 30 minutes. The conditions chosen for a scaled-up reaction were (7), (8) and (9), these samples have a majority of fragments of sizes between 50-70kb.

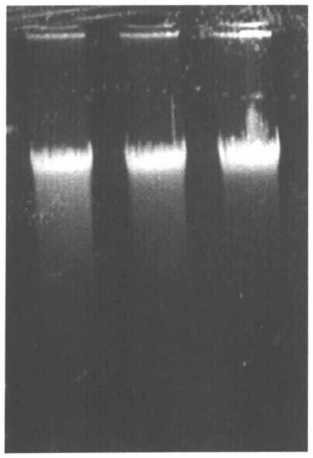
(b) Three 10-fold scaled up samples of chromosomal DNA digested with (1) 0.044, (2) 0.022 and (3) 0.011 units of HindIII.

(c) Fractions (1) #18, (2) #19, (3) #20, (4) #21 and (5) #22 of 30, from sucrose-gradient centrifuged, partially digested chromosomal DNA. Fraction #22 was suitable for cosmid library construction.

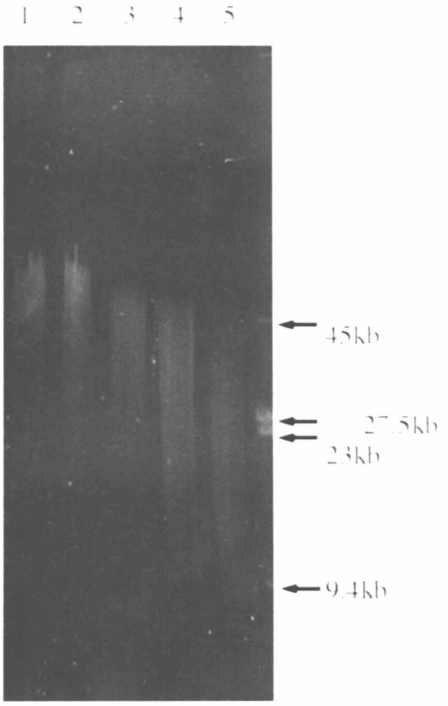
(a)



(b)



(c)



The test DNA digests were resolved on 0.2% agarose gels. Three reactions were set up containing 300-400μg DNA each in a volume of 2ml plus the three concentrations of restriction enzyme judged to digest the DNA to give the majority of fragments between 50-70kb (figure 2.1 (b)). These were incubated for 30 minutes at 37^{o}C and halted by addition of EDTA to 50mM.

2.15.2 Separation of DNA Fragments on a Sucrose Gradient.

Partially digested DNA (300-400μg) was extracted in an equal volume of chloroform and precipitated in 2 volumes of 96% ethanol plus 0.1 volumes 3M Na-Acetate (pH4.5) on ice for 5 minutes. The DNA was pelleted by centrifugation at 1.5K for 10 minutes and washed twice in 1ml of 70% ethanol. The DNA was gently resuspended in 1ml TE and left to fully resuspend at 4^{o}C for 17 hours. It was fractionated on a sucrose gradient which was made as follows: 10% and 40% sucrose solutions containing 20mM TRIS-Cl (pH 8.0), 20mM EDTA (pH 8.0), 1M NaCl, 0.3% sarkosyl with 50g and 200g of sucrose respectively were autoclaved. 19mls of the 40% and 10% sucrose solutions were placed in the appropriate chambers of a Interflon (G. Springham and Co. Ltd.) gradient mixer with small magnetic stirrers to ensure even mixing. The gradients were made in Beckman centrifuge tubes (25 x 89mm) and 1ml of DNA loaded onto the surface. The gradients were centrifuged at 26,000rpm for 16 hours at 10^{o}C in a SW28 rotor. The gradients collected, using a peristaltic pump, as 900μl fractions in 30 eppendorf tubes. 15μl of each fraction was run on a 0.2% agarose gel next to a marker containing 14μl of 40% sucrose, 2.8μl of 5M NaCl, 1μg uncut lambda and 1μg lambda digested with HindIII (figure 2.1 (c)). The appropriate fractions containing DNA sizes between 50-70kb were precipitated in 12 x 75mm Falcon 2005 tubes by the addition of 2 volumes of 96% ethanol plus 2 volumes of 70% ethanol (to prevent the precipitation of the sucrose) overnight at -20^{o}C. The DNA was pelleted by centrifugation for 10 minutes at 2.5K and washed twice in 70% ethanol and each fraction resuspended in 50μl TE. 1-2μl of this was checked against

markers on a 0.2% agarose gel and fractions with the majority of DNA fragments above 50kb used in the ligation.

2.15.3 Dephosphorylation of DNA Fragments.

Partially digested DNA was heated to 68°C to denature the protein and precipitated as described above and resuspended in 1ml TE. A 10μl aliquot was checked on a 0.2% agarose gel and a further 10μl aliquot was put aside. 110μl of the appropriate restriction buffer was added to the remaining DNA plus Calf Intestinal Phosphatase (BCL) to a final concentration of 1 unit to 1μg DNA and the reaction incubated for 30 minutes at 37°C. Nitriloacetic acid (Sigma) was added to a final concentration of 10mM to inactivate the phosphatase, and the reaction incubated for 15 minutes at 68°C and the DNA precipitated. The dephosphorylated DNA was resuspended in TE to give a concentration of 1mg/ml and test ligations performed on 10μl of the DNA before and after dephosphorylation to check the efficiency of the reaction.

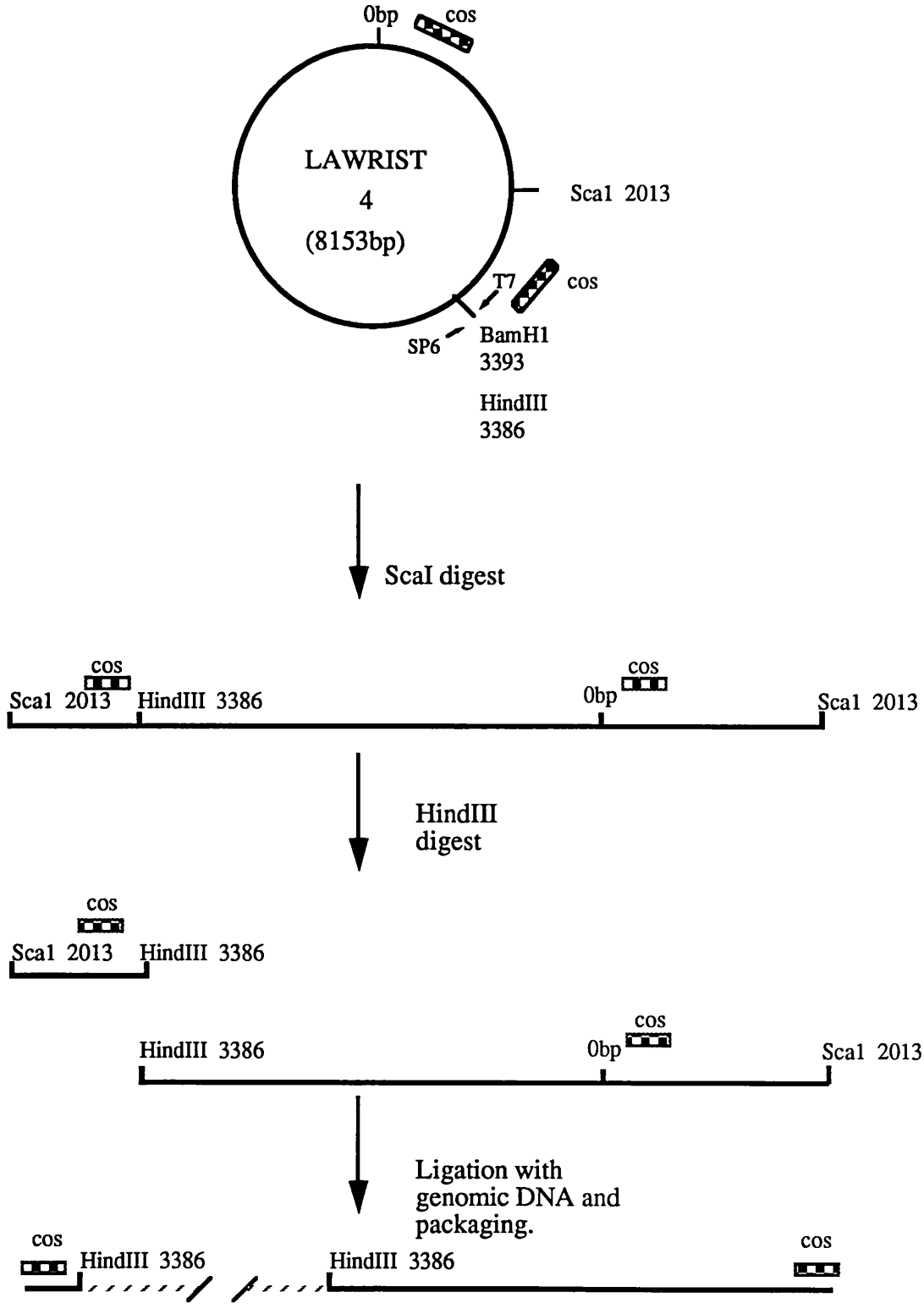
2.15.4 Preparation of Vector Arms.

The cosmid vector used in the construction of both genomic libraries described in this work was Lawrist 4. The vector arms were prepared as follows: 50μg of DNA was digested with 300 units of ScaI in 1ml of 1 x M buffer for 30 minutes and 5μl checked for completion of digestion on an agarose gel. 165 units of calf intestinal phosphatase was added and the DNA incubated at 37°C for a further 30 minutes. The DNA was extracted once with phenol and once with chloroform before being precipitated. The pellet was resuspended in 25μl and a test ligation performed on 0.25μl and directly compared with non-ligated DNA in the same buffers with the addition of 1 unit of ligase to the control immediately prior to loading the samples on the agarose gel.

The remaining DNA was digested with 50-60 units of either HindIII or BamHI, extracted once with phenol and once with chloroform and precipitated.

Figure 2.2 Schematic outline of preparation of cosmid vector "arms" from Lawrist 4.

The vector DNA is represented by the black lines and genomic DNA is represented by the striped lines. The approximate positions of the cos sites are shown as checked boxes. Arrows indicate positions of SP6 and T7 promoters and direction of transcription.


0bp
cos
LAWRIST
4
(8153bp)
ScaI 2013
T7
cos
SP6
BamH1
3393
HindIII
3386
ScaI digest
cos
ScaI 2013
HindIII 3386
cos
0bp
ScaI 2013
HindIII
digest
cos
ScaI 2013
HindIII 3386
HindIII 3386
cos
0bp
ScaI 2013
Ligation with
genomic DNA and
packaging.
cos
HindIII 3386
HindIII 3386
cos

The pellet was resuspended to give a final concentration of 1mg/ml and test ligation performed on 0.25μg and compared, as before, to non-ligated DNA. The preparation of vector DNA diagrammatically shown in figure 2.2.

2.15.5 Ligation of 40-50Kb Fragments into Cosmid Arms.

1-1.5μg of genomic DNA was ligated to 1-1.5μg of Lawrist4 arms prepared as described in 20μl of 1 x ligation buffer with 1 unit of T4 DNA ligase (Boehringer Mannheim) for 17 hours at 14^{o}C. 1μl of the ligation was compared with the equivalent non-ligated reaction on an agarose gel.

2.15.6 in vitro Packaging of Ligation Products.

The Amersham commercial lambda DNA *in vitro* packaging kit was used. 2μl of the ligation (equivalent to 0.1μg of insert DNA) was added to the extracts. The reaction proceeded at room temperature for 2-3 hours. The packaged DNA was stored in 1ml of SM buffer plus 10μl of chloroform at 4^{o}C for 3-6 months.

2.15.7 Screening Cosmid Libraries.

The titre of in vitro packaged DNA was determined by mixing 2μl and 10μl of the reaction with 200μl of plating cells. The cells were prepared by growing a 1/100 dilution of an overnight culture of ED8767 or 1046 to an OD_{600nm} of 0.4-0.6 units, pelleting by centrifugation at 2500rpm for 10 minutes and resuspension in 1/4 volume of 10mM $MgSO_4$ and shaking for 30 minutes at 37^{o}C. The cells plus packaging reaction were incubated for 15 minutes, 1ml of L-broth added and incubated for 1 hour before being pelleted. The pellet was resuspended in 100μl of L-broth and spread on an agar plate containing 30μg/ml kanamycin.

For the isolation of human DNA-containing clones in the genomic mapping project, cosmids were plated out to a density of 1000-2500 per 13cm plate. For the isolation of specific cosmid clones enough packaging reaction plated out to give 0.5-1 million colonies on a square 22cm agar plate. The plates were grown for 10-17 hours at 37^{o}C

and the colonies transferred to Biodyne membranes in duplicate as previously described. In most cases the lifts required a further 3-6 hours incubation before being prepared for hybridisation.

2.15.8 Storage of Plated Cosmid Library.

If storage of the plated out libraries was required, the first lift, having been regrown for 3-6 hours, was placed against a second Biodyne filter with the colonies in the middle, the 2 filters being sandwiched between several layers of Whatman 3MM soaked in L-broth supplemented with 10% glycerol. These could be frozen at -70^{o}C and remain viable for over a year. To reuse a library stored in this way the filter package was slowly thawed, orientation marks made with a sterile needle and the 2 filters peeled apart. A new Biodyne filter was used to replace the replica and the library refrozen. The replica was placed on a L-agar plate supplemented with antibiotic and grown for 3-6 hours.

2.15.9 Isolating Cosmid Clones from a Low Density Plating.

Well isolated positive clones were picked directly from the agar surface by lining up the orientation marks on the autoradiograph with those on the plate over a light box and used to inoculate L-broth supplemented with kanamycin.

2.15.10 Isolating Cosmid Clones from a High Density Screen.

Positives from a high density plating were picked into 1ml of L-broth and further dilutions made in L-broth. 100μl of 10^{-4}, 10^{-5} and 10^{-6} dilutions were replated on 90mm L-agar plates supplemented with 30μg/ml kanamycin. After overnight incubation lifts were performed as described previously and these rescreened with the probe. If necessary this was repeated to obtain an isolated positive colony.

2.15.11 Details of specific cosmid libraries.

Two cosmid libraries were made during the course of this work:

1) HindIII partial EJNAC4.1 cosmid library.

HindIII partial digests of genomic DNA from the EJNAC4.1 cell line were separated on sucrose gradients and ligated into the unique HindIII cloning site of the cosmid vector, Lawrist 4. The percentage of human positives in this library was determined as 1.7%: 5 plates of approximately 1500 colonies each were screened with a total human genomic DNA probe and 129 positives detected. Of the first 799 cosmid clones isolated from this library on the basis of hybridisation of colonies to a total human genomic DNA probe, 140 also hybridised to a total mouse genomic DNA probe: 12 of these were human cosmid clones contaminated with cosmid clones containing mouse DNA and 128 were judged to be co-ligation events. Packaging efficiencies of 1-1.8 x 10^6 colonies per microgram of insert DNA were obtained under optimal conditions.

2) Sau3A partial E65-9 cosmid library.

Sau3A partial digests of genomic DNA from the E65-9 hybrid cell line were dephosphorylated and ligated into the unique BamHI cloning site of Lawrist4. The percentage of human positives for the Sau3A E65-9 cosmid library was not determined but 0.043% of the cosmid clones in a previous cosmid library made from HindIII digested chromosomal DNA from this cell line were human positives (Harrison-Lavoie *et al.*, 1989). The number of co-ligation events was not assessed. Packaging efficiencies of 2-3 x 10^5 colonies per microgram of insert DNA were obtained under optimal conditions.

A third cosmid library made from Sau3A partially digested genomic DNA from human peripheral blood lymphocytes obtained from a patient with acute lymphoblastic leukemia (ALL) was also screened. This library was constructed by Al Ivens from sucrose gradient purified DNA fragments in the cosmid vector LoristB. The packaging efficiency of this library was 5 x 10^5. The library had previously been plated out and the colonies stored as glycerols. When these stocks were titred, 1-5 x 10^9 colonies per millilitre of stock was obtained.

2.16 Fingerprinting Cosmid Clones.

2.16.1 Arraying Human DNA Cosmid Clones.

Cosmid clone colonies found to hybridise with a total human DNA probe were picked into 250μl of kanamycin supplemented L-broth in a Beckman 96 Deep-well titre culture plate and grown overnight. A replicating device, which consists of an arm with a 96-prong hand and two platforms is used to transfer small amounts of culture to biodyne filters. Samples from four plates in 96-well arrays were be impressed onto one filter and these filters were known as "colony arrays". The filters were grown for 3-6 hours before being treated as described previously and hybridised first with total mouse and then with total human genomic DNA probes.

In the initial stages of the project, cosmid clones were analysed by preparation of DNA for each clone by the "microprep" method, HindIII digestion, resolution on 1% agarose gels and transfer to Biodyne A filters. The southern blots were abandoned in favour of the colony arrays as the production of southern blots involved a significantly greater effort and input of resources as well as using a larger area of Biodyne filter (48 clones on a 16 x 18 cm filter compared with 384 clones on an 8 x 12 cm filter). Although more information was obtained on patterns of hybridisation with southern blots, this did not justify the effort required to produce them.

2.16.2 Fingerprinting Reaction.

The fingerprinting reactions were performed as described by Coulson *et al.* (1986).

2.16.3 Generation of Overlapping Contigs.

A genome mapping package has been developed (Sulston *et al.*, 1988) for reading and assembling data from clones analysed by restriction enzyme fragmentation and polyacrylamide gel electrophoresis. The package comprises: data entry, assembly, statistical analysis and modelling. The characteristic banding pattern for each individual clone is digitized and incorporated into a database. Each incoming clone is compared to each clone within the database. A probability of overlap is determined:

the number of bands and the position of the bands on the gel being the important parameters.

2.17 Rare-cutting Restriction Enzyme Fingerprinting (RREF).

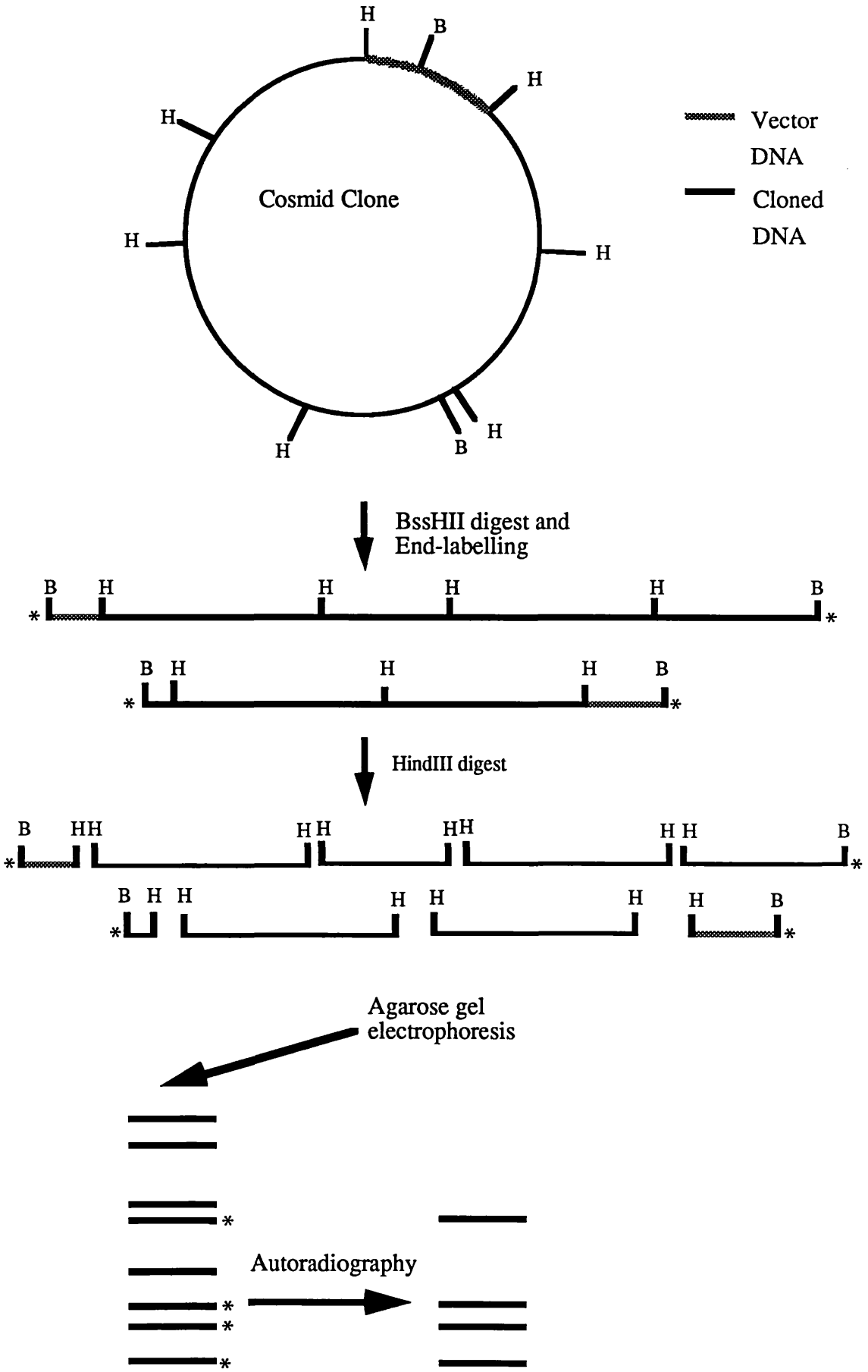
DNA prepared by the microprep method was treated in the following way: For the digestion of 96 samples a reaction mix was made containing 100 units of BssHII, EagI or NotI (Biolabs), 100 units of AMV reverse transcriptase (Life Sciences), 50μg RNase A, 0.3mM dCTP, 2 x M (BssHII) or H (EagI, NotI), 5μCi [α^{32}P] dGTP and dH_2O to a total volume of 200μl. 2μl of this was added to 2μl of DNA in a 96-well corning cell well (round bottomed) plate resting on ice using a Hamilton repetitive dispenser. The wells were capped with 8-cap bands (ICN Flow) briefly spun down and incubated at 50^{o}C (BssHII) or 37^{o}C (EagI, NotI) for 45 minutes. The reverse transcriptase was denatured by incubation at 68^{o}C for 30 minutes. A second reaction mix was made containing for the BssHII digests containing 600 units of HindIII, 1 x M buffer and dH_2O to 600μl. The reaction buffer for the EagI and NotI digests contained 600 units of HindIII in 10mM Tris-HCl, 10mM $MgCl_2$, 1mM dithioerythritol and 16.7mM NaCl (to adjust overall NaCl concentration to 50mM) and dH_2O to 600μl. 6μl of this was added to the digest samples which were incubated for a further 45 minutes at 37^{o}C. 2μl of loading dye was added and the radioactively labelled DNA fragments resolved on 1% agarose gels. The gels were dried down under vacuum at 80^{o}C for 1-2 hours before being exposed to Kodak XAR film for 6-10 hours at room temperature. The RREF is diagrammatically shown in figure 2.3.

2.18 Restriction Mapping Cosmid Clones.

Cosmid clones were restriction mapped using the Amersham lambda Terminase system. The strategy utilizes the 12 bp 5' overhangs at the left and right ends of cosmid molecules linearized with lambda terminase.

Figure 2.3 Schematic outline of rare-cutting restriction enzyme fingerprinting protocol.

The vector DNA is represented by the spotted lines and genomic DNA is represented by the black lines. B = BssHII site and H = HindIII sites.


H
B
H
H
H
Cosmid Clone
H
H
H
B
H
Vector DNA
Cloned DNA
BssHII digest and End-labelling
B H H H H B
B H H H B
HindIII digest
B HH HH HH HH B
B H H H H H H B
Agarose gel electrophoresis
Autoradiography

Oligonucleotides complementary to each overhang (ON-L and ON-R) were 5' end-labelled using [γ-^{32}P] ATP and were used to radiolabel either end of the molecule by hybridisation. 2μg of caesium-purified cosmid DNA was linearised as directed by the Amersham instruction manual in a volume of 20μl. 20μl of 10 x M (HindIII) or H (EcoRI) and 160μl dH_2O was added and an aliquot of 10μl put aside. 1-2 units of restriction enzyme was added to the linearized DNA prewarmed to 37^{o}C and 38μl aliquots removed at one minute intervals into 8μl 100mM EDTA.

The equivalent of 0.1μg of DNA in the partial digests are resolved on 1% agarose gels and 2 time points chosen, one at which the DNA can just be seen to be digested and the second with 1 minute shorter incubation time. These samples were hybridised to the end-labelled oligonucleotides along with the size marker supplied by Amersham and resolved on a 0.8% agarose gel in 1 x E buffer at 4V/cm for 12-17 hours. Where possible the ON-L and ON-R labelled DNAs were run on separate gels as the ON-R oligonucleotide labelled DNA could be run further the larger vector fragment (5027bp for HindIII and 4726bp for EcoRI) being the smallest band size on the gel. The gel was dried down onto a sheet of Whatman DE81 cellulose acetate paper to bind the oligonucleotides at 60^{o}C under vacuum for 1-1$^{1}/_{2}$ hours. The gel was removed from the DE81 by rinsing in 1 x E buffer for 10 minutes and the DE81 redried briefly under vacuum before being exposed to Xray film. The removal of the gel was to reduce background. The order of fragments could be determined.

Figure 2.4 shows the results of terminase mapping four cosmids with HindIII: ZF25, ZF29, ZF30 and ZF39.

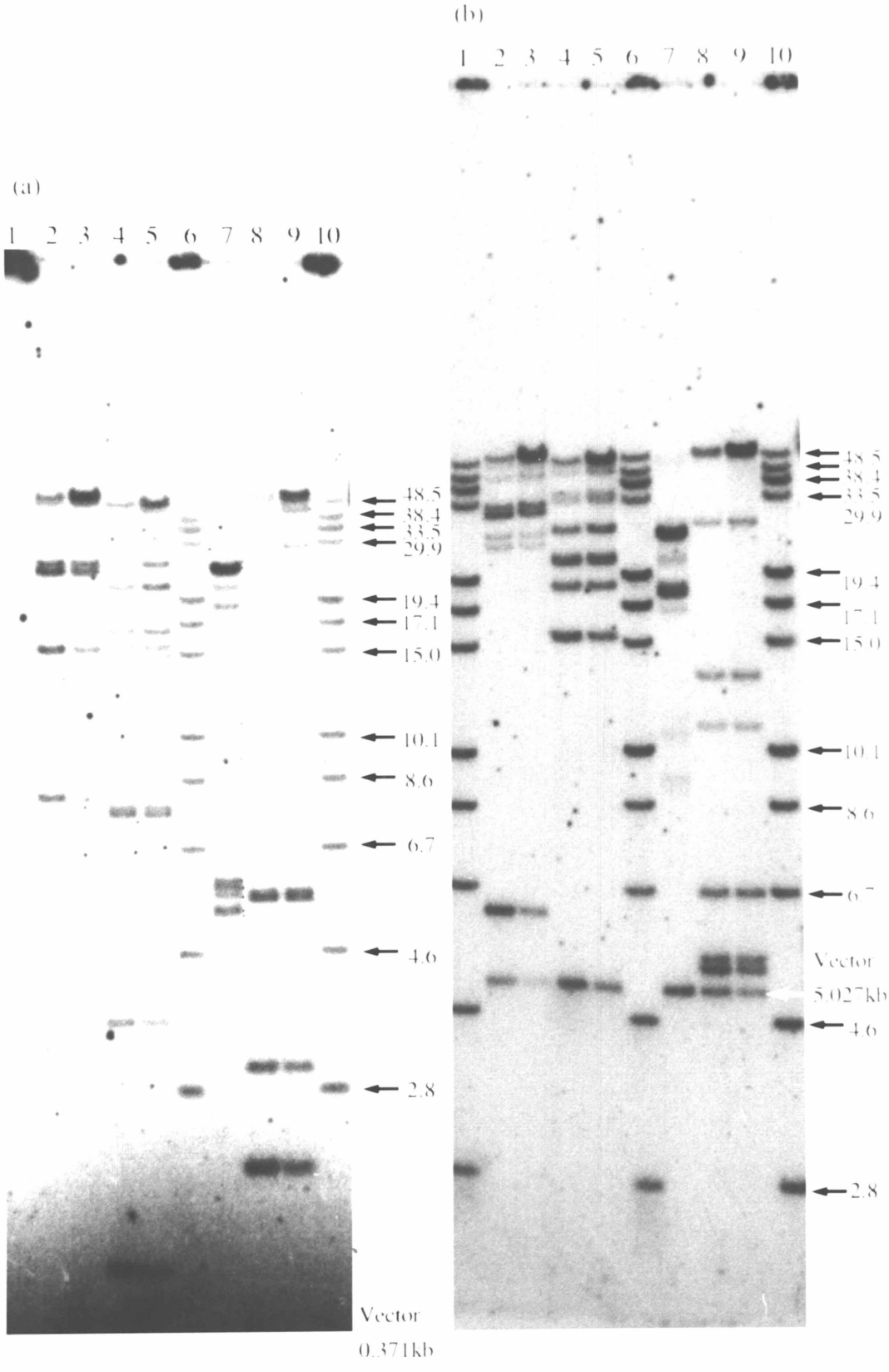
2.19 cDNA Libraries.

2.19.1 Plating Lambda Phage cDNA Libraries.

NM514 or XL1 cells were grown to stationary phase, pelleted by centrifugation at 3K for 10 minutes and resuspended in half a volume of 10mM $MgSO_4$ and stored at 4^{o}C until use.

Figure 2.4 Example of a lambda terminase mapping experiment.

(a) ON-L and (b) ON-R mapping of HindIII sites in four cosmids. Lanes 1, 6 and 10 contain marker DNA, lanes 2 and 3 contain partially digested DNA from ZF25, lanes 4 and 5 contain partially digested DNA from ZF29, lane 7 contains partially digested DNA from ZF30 and lanes 8 and 9 contain partially digested DNA from ZF39. The positions of vector-only labelled fragments are indicated with an arrow.


(a)
1 2 3 4 5 6 7 8 9 10
48.5
38.4
33.5
29.9
19.4
17.1
15.0
10.1
8.6
6.7
4.6
2.8
Vector
0.371kb
(b)
1 2 3 4 5 6 7 8 9 10
48.5
38.4
33.5
29.9
19.4
17.1
15.0
10.1
8.6
6.7
Vector
5.027kb
4.6
2.8

The titre of the library was determined by plating dilutions of the stock. The libraries were plated in one of two ways:

a) The equivalent of 0.5-5 x 10^6pfu was added to 7.5mls of plating cells and incubated at 37°C for 15-20 minutes and divided equally between 10 132mm sterilin plates containing 80-100mls L-agar with the addition of 8mls L-top agarose per plating sample.

b) The equivalent of 0.5-5 x 10^6pfu was added to 3mls of plating cells and incubated at 37°C for 15-20 minutes and plated with the addition of 35mls L-top agarose onto one 22cm^2 sterilin plate containing 400ml l-agar, the plates having previously been dried and warmed to 42°C to prevent premature setting of the agarose.

The plates were grown for 8-14 hours before lifts were performed onto Biodyne A filters.

2.19.2 Plating plasmid cDNA Library.

A plasmid cDNA library was plated out in the same way as for a cosmid library (2.15.7) onto 22cm^2 plates and lifts performed onto Biodyne A membrane in duplicate.

2.19.3 Screening cDNA Libraries.

The cDNA library filters were screened with the 20 base oligonucleotide and with specific DNA fragments as described previously.

2.19.4 Isolating cDNA Clones from Phage Libraries.

Where positive signals were obtained, the appropriate area on the library plate was cored out using the wide end of a pasteur pipette and transferred to 1ml SM buffer plus 20μl chloroform and vortexed briefly. Single plaques were obtained by plating out dilutions of the phage and rescreening to obtain isolated positive plaques.

2.19.5 Details of cDNA libraries

Five cDNA libraries were screened:

1) A human foetal kidney lambda gt10 cDNA library, amplified once, was kindly given by Afshan Malik and John Cowell.

2) A human lymphoblast lambda gt10 library, amplified once, was kindly given by Jenny Dunne.

3) A human fibroblast cDNA made from primary fibroblasts from a skin biopsy of an adult with Gauchers disease (a lipid storage disease) which had been amplified twice, was kindly given by Claire Isacke.

4) A human fiboblast cDNA library made from a primary fibroblast (F1084) culture from an embryonic lung biopsy, amplified once, was kindly given by Claire Isacke and Steve Neame.

5) A pcD2 human fibroblast library made from foreskin fibroblasts, amplified once, was kindly given by Andrew Sinclair. The cloning of cDNA by the method of Okayama and Berg (1982) permits high yield of full length cDNA clones regardless of their size. Modification of the original vector to allow for expression of the inserted clone produced the vector pcD2 which was used in the construction of this library.

2.20 Somatic cell hybrids.

a) E65-9.

Porteous *et al.* (1986) have previously reported the isolation of somatic cell hybrids that contain portions of the short arm of human chromosome 11, in a mouse cell background by chromosome-mediated gene transfer (CMGT). The analysis of one of these cell lines, E65-9, is reported in Harrison-Lavoie *et al.*, 1989. This cell line was made by CMGT of an activated human c-Harvey-*ras*-1 gene (*H-ras*) into non-transformed mouse cell as described in the introduction. This cell line was judged to contain 2-3 x 10^6 bp of human DNA representing 0.043% of its DNA content. CMGT is associated with a severe level of rearrangement of the transferred DNA.

Four lines of evidence suggest that the human DNA in E65-9 is diploid and consists of the same region duplicated four to five times:

1) The human component cannot be detected by *in situ* hybridisation even though Gosden and Porteous (1987) have been able to detect 5 x 10^6 bp of DNA as a single block.

2) The dosage of *H-ras* is equivalent to that seen in other CMGT lines shown to have duplications (Porteous *et al.*, 1986 and Gosden and Porteous, 1987).

3) The L1 "fingerprint" suggests there is about 2 x 10^6 bp of human DNA in E65-9 (Porteous *et al.*, 1986).

4) Analysis of 112 cosmid clones by the fingerprinting method of Coulson *et al.* (1986) generated two large contigs containing 48 and 16 clones of 140 and 125 kb respectively, two small contigs containing 2-3 clones. 43 clones were left unattached, most of these hybridised to some extent to total mouse DNA probes and were judged to be artifactual (Harrison-Lavoie *et al.*, 1989). The random number generated prediction (Sulston *et al.*, 1988), based upon a target genome of 2 x 10^6 bp would be 20 contigs with a mean occupancy of 3-4 clones. If the target size was 4 x 10^5, 2 contigs of 42 occupancy would be predicted which is more consistant with the results obtained.

b) EJNAC4.1 and EJNAC4.1.5.

The founder EJNAC hybrid cell line was generated by the fusion of human EJ bladder carcinoma cells to mouse C127 cells followed by direct selection for tumour growth by injection into immunosuppressed mice (Porteous *et al.*, 1989). Hybridisation studies with 20 gene specific DNA probes identified the presence of chromosomal DNA from human chromosome 11 from 11p14-ter and 11p13 but not 11cen-q13 or 11q23 (Porteous *et al.*, 1989).

No evidence for genomic rearrangements or interstitial deletions has been observed in the process of cell fusion (Bickmore *et al.*, 1989a and b). Two subclones derived from the EJNAC hybrid cell line, EJNAC4.1 and later, EJNAC4.1.5, have been used in the cosmid mapping project and also in the analysis of zinc finger sequences. 1.7% of

cosmid clones from a HindIII partial EJNAC4.1 library and 1% of clones from a HindIII partial EJNAC4.1.5 library were judged to contain human DNA. This indicated that these hybrids contained 200 and 120 megabases of human DNA, respectively.

Chapter 3.

A method to identify cosmids containing rare-cutting restriction enzyme sites.

3.1 Introduction.

The objective of this work was to design a method to detect rare-cutting restriction enzyme sites within isolated DNA fragments which could be applied on a large scale to a genome mapping project. Information on the position of rare-cutting restriction enzyme sites within cloned DNA arrays can be used in three ways:

1) For the identification of genes. Detection of rare-cutting restriction enzymes sites in the cosmids generated in the mapping project will identify CpG islands and the predicted genes associated with these islands, which would not be accessible by conventional methods.

2) In the construction of long range restriction maps. Restriction enzymes that cut rarely in mammalian genomic DNA are commonly used to construct long range restriction maps. Additionally, by locating the sites for these restriction enzymes within a cloned DNA map, the restriction map of the cloned DNA can be compared to a restriction map developed directly from genomic DNA to confirm the accuracy of both mapping procedures and to detect any chromosomal rearrangements that might be present in the hybrid cell line.

3) In the bridging of gaps between contigs. There are often gaps in cloned DNA maps which may occur by chance or may represent regions of DNA that are not readily cloned in *E.coli*. One way to produce a complete map is to link contigs on pulse field gels by means of single, large fragments of DNA generated by rare-cutting restriction enzymes.

A quick an easy method was required for screening the estimated 1500 cosmid clones that would represent the minimal set of a cosmid map of the short arm of human chromosome 11.

The cosmid cloning vector, Lawrist 4, used in the 11p mapping project contains restriction sites for BssHII and EagI. The presence of these sites would represent an internal control for the success of detection of sites with each method tested and these two restriction enzymes were the obvious choice in the design of the methods to detect possible CpG islands.

Three methods were examined:

1) Digestion of cosmid DNA with BssHII and EagI and analysis on field inversion gels.

2) Hybridisation of southern blots of HindIII digested cosmid DNA to oligonucleotides designed to detect BssHII and EagI sites.

3) End-labelling of the BssHII- or EagI-generated fragments followed by HindIII digestion and analysis on normal agarose gels.

The important criteria for success of a method were defined as:

a) Accuracy of detection of all the sites.

b) Use of materials already available from genome mapping project.

c) Ease and speed of manipulations.

3.2 Method 1: Digestion of cosmids with BssHII and EagI and analysis by FIGE.

3.2.1 Experimental Design.

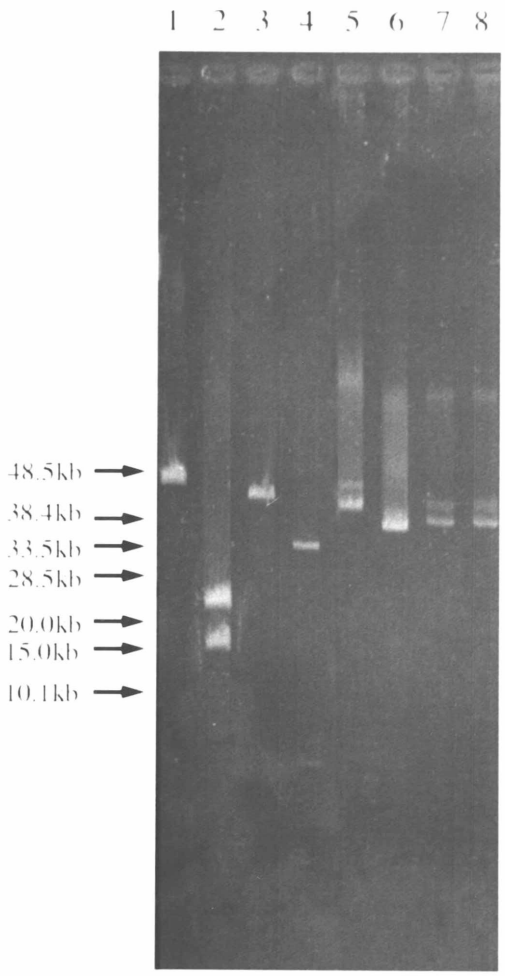
A sample of DNA prepared by the method of Gibson and Sulston (1987) representing one complete preparation (0.5-1μg DNA) was digested with either EagI or BssHII. Since the cosmid vector, Lawrist 4, contains a site for both of these restriction enzymes a single band of 40-50kb would be generated if the cloned genomic DNA contained no sites for the restriction enzyme, and more than one band if the insert DNA did contain a site. To distinguish whether more than one band was generated the fragments were resolved on a 1.0% TBE agarose gel which was subjected to field inversion gel electrophoresis (FIGE). This procedure causes the separation of large fragments of DNA within a relatively short distance (Carle *et al.*, 1986).

3.2.2 Results.

Having determined the most suitable pulse times to separate of fragments between 15-50kb, cosmid clones digested with BssHII or EagI were analysed. One example of these experiments is shown in figure 3.1. DNA from four cosmid clones, C380, C381, C382 and C383, was digested with BssHII and run alongside uncut control samples.

Figure 3.1 Field Inversion Gel Electrophoresis of cosmid DNA.

FIGE of DNA from four cosmid clones, C380, C381, C382 and C383 run after digestion with BssHII (1-4) undigested (5-8). C380 (1) and C382 (3) are digested to give one band indicating the presence of one BssHII site, the site within the vector DNA, and C381 (2) and C383 (4) give two bands indicating the presence of sites within the cloned DNA.


1 2 3 4 5 6 7 8
48.5kb
38.4kb
33.5kb
28.5kb
20.0kb
15.0kb
10.1kb

Digestion of C380 and C382 produced one fragment indicating the presence of one restriction site for BssHII which is the site within the vector DNA. Digestion of C381 and C383 produced two bands indicating the presence of two BssHII sites in the cosmid clone, one within the vector DNA and one within the cloned DNA.

3.2.3 Fulfilment of criteria.

When this method was applied on a large scale, there were difficulties encountered in the three categories judged as important for application of this method to a genome mapping project:

a) The results obtained were not always clear-cut. The method is sensitive to partial digestion which makes the interpretation of some results less easy. Where a site in the cloned DNA is close to the site in the vector DNA or where there are two sites within the cloned DNA that are close together, the relatively small fragments produced would probably not be visible on an agarose gel. Initial problems with experienced with partial digestion and degradation of samples were in part due to unfamiliarity with the DNA preparation protocol resolved with experience, and in part due to partial evaporation of samples when digested with BssHII as a result of the required incubation at $50^{o}C$. The evaporation would probably be reduced by adding a layer of paraffin oil but this would increase the number of manipulations required in the process.

b) The use of resources - To visualise DNA fragments on an agarose gel required the use of an entire small scale preparation of DNA (0.5-1μg). To test with more than one enzyme would require further preparations of DNA. Preparation of DNA from larger cultures would be a possible solution but represents significantly more work than the microprep method and is not using reagents from the mapping project.

c) The number of clones analysed - This was limited by the available space on the gel and the time taken for the run. 32 samples could be run on one gel and fragments were resolved overnight.

This method would not be applicable to a large scale mapping project as it is expensive in both the use of time and of resources. It would be useful in the analysis of smaller numbers of clones and where the amount of DNA was not limited by the preparation protocol.

3.3 Method 2: Hybridisation with oligonucleotides designed to detect BssHII and EagI sites.

3.3.1 Experimental design.

In the early stages of the mapping project, southern blots of the DNA of individual cosmid clones digested with HindIII were used to rescreen the picked clones with total mouse and total human DNA probes to determine the mis-picked clones. These represented a valuable resource for screening with other probes for additional analysis of the clones. Oligonucleotides were designed to detect the recognition sequence of EagI and BssHII restriction enzymes. These were 10 base oligonucleotides with the 6 base site for the restriction enzyme plus 4 degenerate positions:

BssHII oligonucleotide: PL BSSH 3' GCGCGCNNNN 5'.

EagI oligonucleotide: PL EAG 3' CGGCCGNNNN 5'.

These oligonucleotides were used to screen the southern blots for the presence of EagI and BssHII sites. The sites for these restriction enzymes in the vector HindIII fragment of 5.4kb would be detected by the oligonucleotides to act as an internal control.

3.3.2 Results.

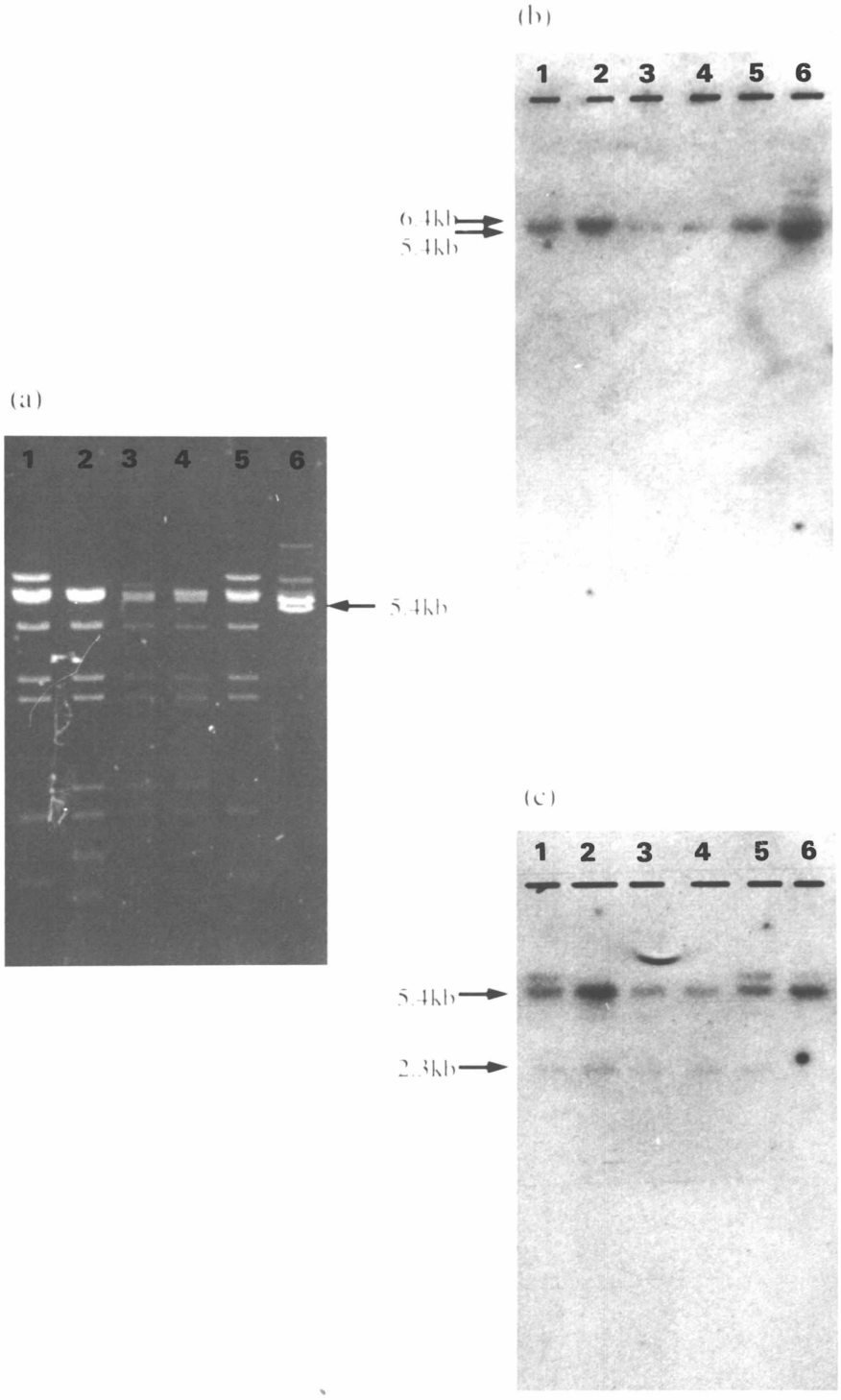
To investigate the feasibility of this method a contig from the E65-9 mapping project (Harrison-Lavoie *et al.*, 1989) was used in trial hybridisations. The contig consisted of 6 overlapping cosmids: C464, C440, C460, A449, C457, and C397 which comprise part of contig B (figure 4.1). A HindIII digest of these clones is shown in figure 3.2 (a).

Figure 3.2 Hybridisation experiments with rare-cutting restriction enzyme oligonucleotides.

(a) HindIII digests of a six overlapping cosmid clones from E65-9 contig B: C464 (1), C440 (2), C460 (3), A449 (4), C457 (5) and C397 (6).

(b) Hybridisation pattern of a southern transfer blot of this contig obtained with a PL BSSH oligonucleotide probe.

(c) Hybridisation pattern of a southern transfer blot of this contig obtained with a PL EAG oligonucleotide probe.


(a)
1 2 3 4 5 6
5.4kb
(b)
1 2 3 4 5 6
6.4kb
5.4kb
(c)
1 2 3 4 5 6
5.4kb
2.3kb

The southern blot prepared from this agarose gel was hybridised to the radiolabelled oligonucleotides. Figures 3.2(b) and (c) show the results obtained with PL BSSH and PL EAG respectively.

Hybridisation to the vector band (5.4kb) on the southern blots was observed with both probes showing that the known restriction sites in the vector detected by the two oligonucleotides. In addition, a 2.3kb band was detected with the EagI oligonucleotide in five of the samples: C464, C440, C460, A449, and C457. The five cosmids all have a 2.3kb HindIII band in common. On double digestion of these clones with HindIII and EagI, the 5.4kb vector band was cut to give two new bands of 3.65 and 1.75kb as expected, and the 2.3kb HindIII fragment was cut to give bands of 1.95 and 0.65kb which confirmed the existence of an EagI site in this fragment. Only one band was visible on hybridisation with the BssHII oligonucleotide. Double digests with HindIII and BssHII performed on the samples gave the expected 4.14 and 1.25kb bands from digestion of the vector. Two other new bands were observed of 3.7 and 2.5kb which result from the digestion of a 6.4kb HindIII fragment shared by all 6 cosmids. Hybridisation of the PL BSSHII oligonucleotide to this fragment would be expected but it was probable that the hybridisation of this band is obscured by the hybridisation signal to the vector fragment. No false hybridisations or cross hybridisations were observed: The 2.3kb fragment hybridised only to the EagI oligonucleotide and not to the BssHII oligonucleotide even though the sites for these two restriction enzymes are similar. Only a small number of clones were analysed when testing this method and its accuracy when applied to a large number of clones was not determined.

3.3.3 Fulfilment of criteria.

a) Accuracy of detection of sites - This method was sufficiently accurate in the detection of EagI and BssHII sites on southern blots. However, in cases where the fragment of the insert DNA containing the rare site is of a similar size to the vector band, only an increased intensity of hybridisation would be observed which would

mean a certain number of sites would be missed. Estivill and Williamson (1987) have used a similar method to detect NotI and EagI sites in cosmid clones on DNA southern blots. They used an 8 base NotI oligonucleotide to successfully detect NotI sites and EagI sites. However, they also detected a number of false positives. Melmer *et al.* (1990) improved the sensitivity and specificity of the procedure by varying the hybridisation and washing temperatures and specifically detected NotI sites with no false positives. They also used 8 base oligonucleotides to specifically detect a proportion of ApaI, BssHII, SacII and EagI sites and were able to distinguish between an 8 base match and a 7 base match.

b) Use of materials - the filters are a reusable resource that can be used with several different oligonucleotides designed to detect other rare cutting restriction enzyme sites. This method was designed and tested for use with southern blots of the cosmid clone DNA. The application of this method with the colony arrays, which replaced the southern blots as a means of determining hybridisation with total human and total mouse genomic probes, was not explored.

c) Ease of manipulation - little additional work would be required to screen filters of immobilised cosmid clone DNA generated by the mapping project.

Under the conditions tested this method fulfilled 2 of the 3 criteria required for screening large numbers of clones. As the southern blots generated by the mapping project were superseded by colony arrays, the method of oligonucleotide hybridisation required refining. To screen the colonies arrays with oligonucleotides to EagI and BssHII sites a distinction would have to be made between hybridisation of the probe to the vector copy of rare-cutting restriction site and any additional sites within the insert DNA. The vector site cannot be blocked during the prehybridisation reaction with a second oligonucleotide designed to detect the vector sequence in and around the vector EagI and BssHII sites. Both vector strands would have to be blocked as the oligonucleotides are detecting a palindrome sequence and oligonucleotides designed to do this would preferentially hybridise together. Relying on differing intensities of hybridisations was predicted to be difficult as there would

be background from hybridisation to *E.coli* DNA rare-cutting restriction sites and the possible differential growth abilities of individual clones would give ambiguous results.

Oligonucleotides could be designed to detect rare-cutting restriction enzyme sites not contained in the vector, NotI (GC/GGCCGC) and SacII (C/CGCGG) being two candidates. The analysis of NotI sites by hybridisation of oligonucleotides to colonies of cosmids has recently been reported by Melmer and Buchwald (1990). Their original screening with a 8 base oligonucleotide designed to detect the NotI site failed to give signal above background. They then used a 12bp oligonucleotide with the recognition sequence for the NotI restriction enzyme extended by two degenerative nucleotides at each end. With a longer oligonucleotide they were able to raise the temperature of hybridisation to give a greater specificity of signal with respect to the background of bacterial DNA. 7 out of 13 colonies identified through screening with the 12 base oligonucleotide contained NotI sites.

In this examples the number of false positives detected is significant. In addition, if the 12 base NotI oligonucleotide was used with the colony arrays only 12% of the total number of CpG islands would be detected (Lindsay and Bird 1987). This method could be applied to the colony arrays with a 12 base oligonucleotide designed to detect the SacII recognition sequence which would detect 74% of the total number of sites but with 6 of the bases in such a probe being degenerate and no second probe and the level of false positives would probably be higher than that predicted by the work of Melmer and Buchwald (1990) would be unacceptable in the application of this method to the 11p project.

3.4 Method 3: Rare-cutting Restriction Enzyme Fingerprinting (RREF).

3.4.1 Experimental design.

The third method tested involved the adaptation of the fingerprinting method of Coulson and Sulston (1986) to detect the presence of rare cutting sites within the cloned DNA. Coulson and Sulston's protocol was designed to efficiently analyse large

numbers of clones and was being applied in the 11p mapping project. This meant that the major components were readily available. This method could be applied to large numbers of the clones using DNA prepared as part of the mapping project in a 96-well system.

The adaptations from the original method are as follows:

1) A rare cutting restriction enzyme (BssHII, EagI or NotI) is used in the first step in place of HindIII.

2) 5μCi [α^{32}P] dGTP (and cold dCTP) is used compared with 20μCi [α^{32}P] dATP (and cold dGTP) as an excess of label obscured smaller fragments when the digests were run on agarose gels.

3) The final incubation time was reduced from 2 hours to 45 minutes and the volume of the digest increased to allow for the addition of more units of HindIII. This reduces the possibility for degradation of the DNA is cases where the preparation of DNA was less pure. This degradation was associated more with the BssHII digests as they require an incubation temperature of 50^{o}C which increases the reaction rate of the hydrolysis of the DNA by any contaminating deoxyribonucleases.

4) The final reactions were resolved in 1% agarose gels. Two rows of 24 samples were run on each 22 x 22cm gel which were dried down under vacuum at 80^{o}C before being exposed to Kodak film for 8-17 hours at room temperature.

3.4.2 Results.

Examples of applications of the rare-cutting restriction fingerprinting method are shown in figure 3.3. Four cosmids, A421, C423, C436 and A449 from the E65-9 mapping project (Harrison-Lavoie *et al.*, 1989) are shown digested with BssHII (a) and EagI (b) using the rare-cutting restriction fingerprinting protocol. With BssHII two constant bands of 4.14 and 1.25kb are always observed which are the end-labelled vector fragments. Only the two vector bands are observed with C421 and C436 indicating the absence of a BssHII site in the insert DNA.

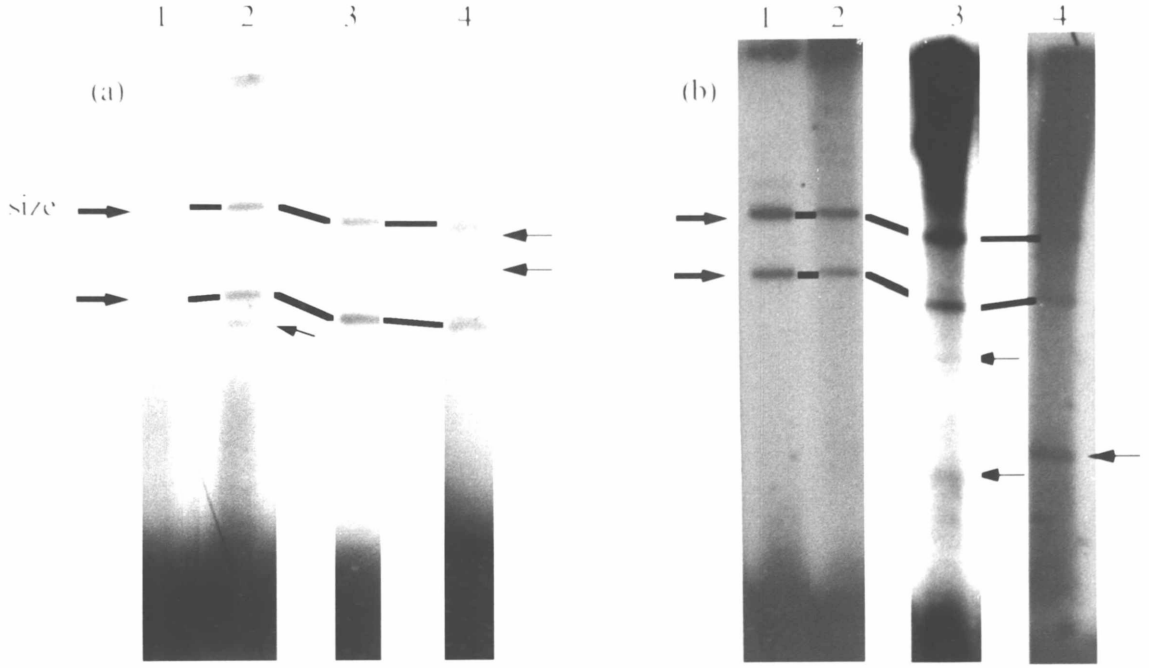
Figure 3.3 Examples of detection of BssHII and EagI sites by end-labelling.

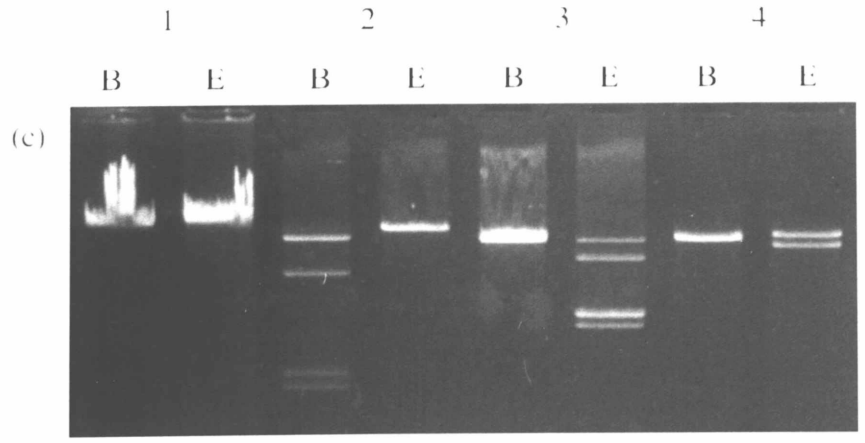
(a) Four cosmid clones after rare-cutting restriction enzyme fingerprinting (RREF) with BssHII, electrophoresis and autoradiography. Arrows to the left of the figure indicate vector-generated fragments of 4.14 and 1.25kb.

(b) Four cosmid clones after RREF with EagI, electrophoresis and autoradiography. Arrows to the left of the figure indicate vector-generated fragments of 4.14 and 1.25kb.

(c) Four cosmid clones digested with BssHII (B) or EagI (E) and resolved on a 1% agarose gel.

1 = A421, 2 = C423, 3 = C436 and 4 = A449. Insert-generated end-labelled fragments are marked with arrows to the right of the figures. As samples were run on different agarose gels they have been resolved to slightly different extents and the black lines mark the common vector bands.


1
2
3
4
(a)
size
(b)


1
2
3
4
B
E
B
E
B
E
B
E
(c)

Bands additional to the vector bands are seen with C423 and C449 indicating the presence of BssHII restriction sites in the cloned DNA. With EagI two constant bands of 3.65 and 1.74kb are observed which are the end-labelled vector fragments. A421 and C423 appear not to contain an EagI site in the cloned DNA. Digestion of C436 and A449 produces bands additional to the vector bands indicating the presence of EagI sites within the cloned DNA. To confirm the reliability of this method, large scale DNA preparations of these clones purified on caesium gradients were digested with BssHII or EagI and resolved on a 1% agarose gel (figure 3.3 (c)). This analysis confirmed that the cloned genomic DNA of A421 contained no site for either EagI or BssHII, C423 contained three sites for BssHII and no sites for EagI and that C436 contained no sites for BssHII and three sites for EagI. In the case of A449, the analysis by single digestion indicated the presence of one EagI site within the cloned DNA as predicted by the rare-cutting restriction enzyme fingerprinting (RREF) but indicated that the predicted BssHII was absent. This clone had previously been shown to contain a BssHII site when testing method 1: The BssHII site was present in a 6.3kb HindIII fragment as determined by comparison of a HindIII digest of DNA for this clone with a HindIII/BssHII double digest. There were two explanations for this site not being detected in the analysis by single digestion: either the BssHII site in the insert DNA was long way from the BssHII site in the vector and the two bands of about 20kb produced by digestion with BssHII were not resolved on the gel, or the site was very close to the site within the vector producing a band too small to be seen. Data from the BssHII/HindIII analysis indicated that the cloned BssHII site is a minimum of this (that being the sum of the smaller BssHII-HindIII genomic DNA fragment of 2.5kb and the smaller BssHII-HindIII vector DNA fragment of 1.25kb) from the vector BssHII site and a 3.75kb fragment would be visible on the 1% agarose gel. It was probable that digestion of this clone with BssHII site generated two fragments similar in size which were not separated on the 1% agarose gel. Analysis of the BssHII digest by FIGE (method 1) might have separated the two fragments but this was not determined. This site was not detected in the

oligonucleotide hybridisation experiment (section 3.3.2). The analysis of this one clone emphasises the sensitivity and accuracy of the RREF method.

3.4.3 Fulfilment of criteria.

The method fulfilled the criteria set in the following ways:

a) Ambiguous results - results were obtained for 60% of the clones on the first analysis of the DNA when testing this method. Low yield or impure DNA preparations generally accounted for the rest. When applied to the clones used in the 11p mapping project the DNA would have previously been used in the fingerprinting reaction and would be of known quality. This method is insensitive to partial digestion: if this occurs in the first step the final result is not effected as only digested ends are substrates for the reverse transcriptase and the partial digestion of DNA in the final step is prevented by addition of HindIII greatly in excess to that required.

b) Use of materials - as 0.05-0.1μg of DNA was required, equivalent to $1/10^{th}$ of a small scale "microprep" preparation, several reactions using different restriction enzymes could be performed from one preparation of DNA from the genome mapping project. 2 units of the restriction enzyme were required per digest reducing the cost of analysing a large number of clones with these relatively expensive restriction endonucleases.

c) Ease of manipulation - the processing of the clones by RREF closely follows the fingerprinting protocol used in the 11p mapping project and all the procedures were designed for ease of use with large numbers of clones. The main rate-limiting step for analysis of clones by the fingerprinting method is the running of samples on acrylamide gels which require significant time and effort to set up and run. With the rare-cutting restriction enzyme fingerprinting analysis, 56 samples can be resolved on one 22 x 22cm agarose gel. As several gels can be run in one day over 200 samples can be analysed in 24 hours with relative ease.

3.5 Summary.

Of the three methods compared, the rare-cutting restriction end-labelling protocol would be the most applicable in the final stages of the mapping project. It requires a comparatively minimal effort for the amount of data that was generated. A description of the application of this method to two sets of cosmid clones is related in the next chapter.

Chapter 4.

Identification of rare-cutting restriction enzyme sites.

4.1 Introduction.

Having developed a method that should allow the identification of rare-cutting restriction enzyme sites in screens of large numbers of cloned DNA fragments, two sets of clones were analysed. The first set analysed consisted of 112 cosmid clones which were being used in the construction of a cloned DNA map of the region of human DNA contained within the hybrid cell line, E65-9 (Harrison-Lavoie *et al.*, 1989). The second set analysed comprised 101 clones which were a subset of the human DNA clones isolated from EJNAC4.1 and EJNAC4.1.5 cosmid libraries which hybridise to an oligonucleotide designed to detect zinc finger protein genes via a conserved motif.

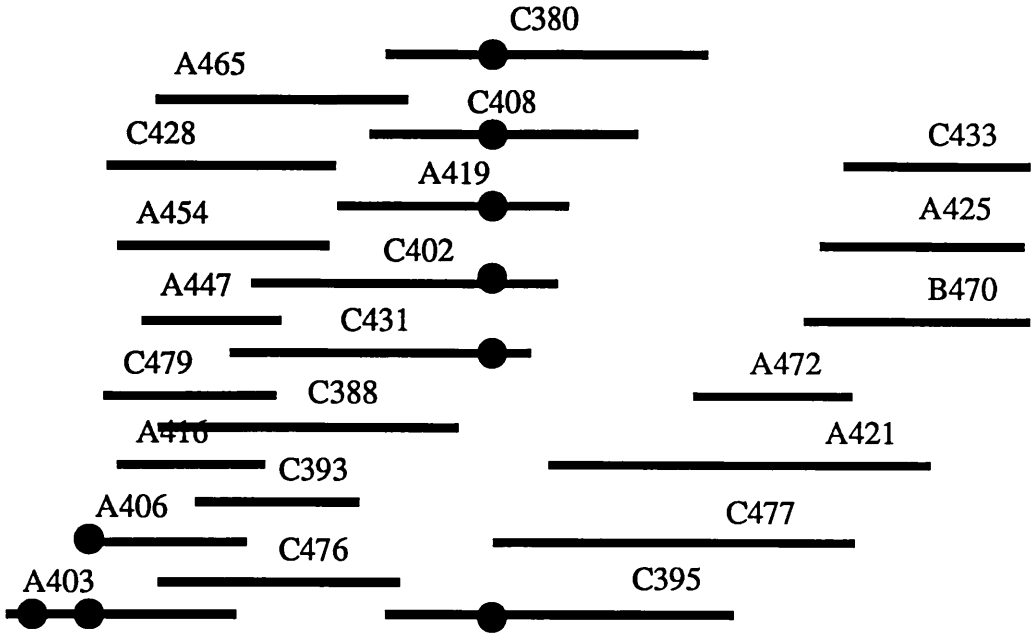
4.2 Results of analysis clones from HindIII partial E65-9 cosmid library.

112 cosmid clones had been isolated from a HindIII partial E65-9 Lorist 2 library on the basis of hybridisation with a total human DNA probe and were being analysed and assembled into contigs using the fingerprinting method of Coulson *et al.* (1988). This analysis resulted in the construction of two major contigs, A and B, of 48 and 16 overlapping clones respectively, plus three small contigs of three and two clones, and 43 unattached clones. Contigs A and B are shown in figure 4.1. The degree of overlap as shown in the figure is defined by the number of bands that each clone has in common and is not taken to represent physical distance. Many of the cosmid clones in these contigs are "buried". A buried clone is one which is entirely contained within a second, usually larger, clone. These clones are not shown in the figure. The set of clones overlapping to the smallest degree but containing all the DNA in a contig is defined as the minimal set. It is this minimal set which would be used in analysis by the rare-cutting restriction fingerprinting method at the end of a large scale genome mapping project.

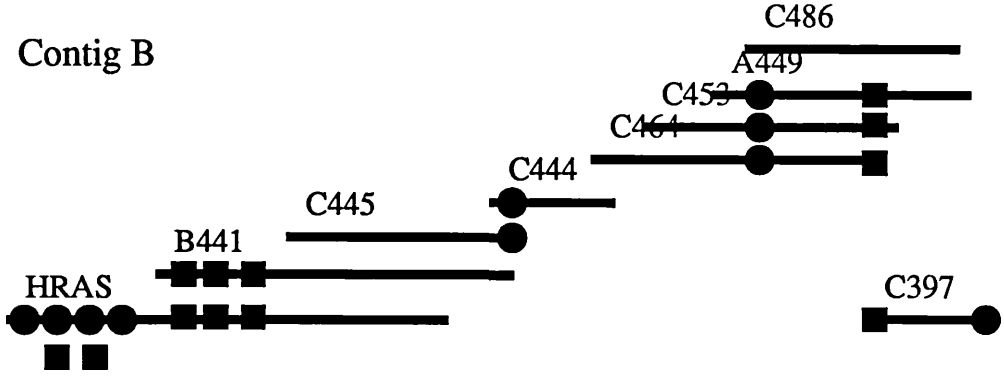
Figure 4.1 Structure of two large E65-9 contigs, A and B, with approximate sites for BssHII and EagI sites.

The two major contigs assembled from the E65-9 mapping project, A and B, are shown with the location of BssHII sites marked by filled squares and EagI sites by filled circles. The positions of sites is not precise and is determined by comparing data on overlapping clones.

Contig A



Contig B



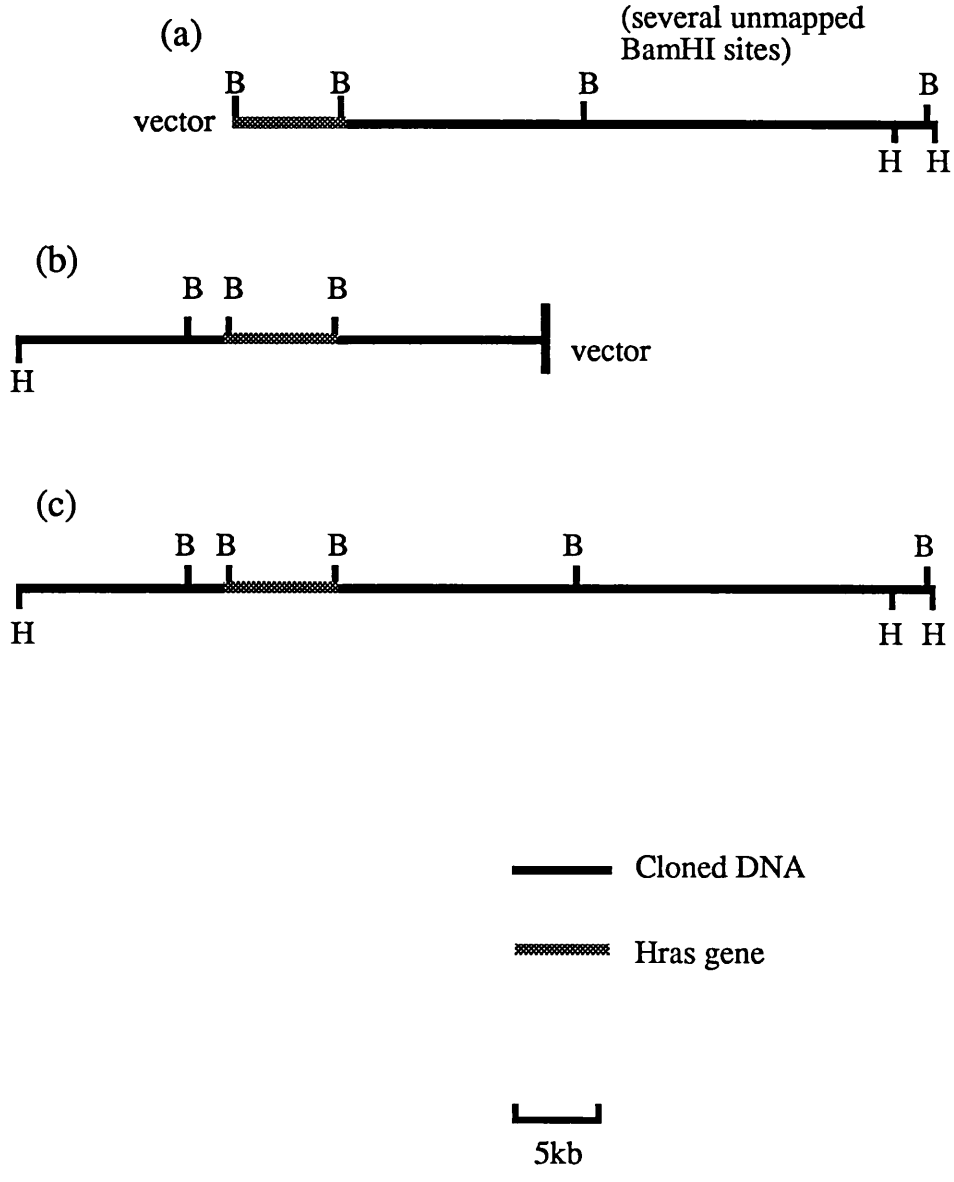
The 112 E65-9 cosmids were analysed by the RREF method for BssHII and EagI sites, 11 (10%) were found to have one or more BssHII sites in the insert DNA and 27 (24%) clones had one or more EagI sites. 16 (14%) were found to contain sites for both restriction enzymes. When the data from the RREF analysis was compared with the contig data a good correlation was found between clones that overlapped and those that contained sites which enabled the approximate positioning of these sites within the contigs as shown in figure 4.1.

4.3 Isolation of c-Harvey-ras-1 (H-ras) from a human genomic, Sau3A partial, cosmid library.

A cosmid clone containing the H-*ras* gene was not isolated from the library constructed from HindIII partial DNA from E65-9. As the hybrid cell line was selected on the basis of the presence of an activated H-*ras* gene it was important to show that this gene was present and to link it to the contig data. One possibility was that the gene was contained on a HindIII fragment much larger than 45kb and could not be cloned in a HindIII partial cosmid library. An ALL Sau3A partial (described in section 2.15.11), LoristB library made from DNA prepared from the white blood cells of a patient with acute lymphoblastic leukemia was screened. This had previously been plated out and stored at -70^{o}C. Replicas were grown and screened with a 6.6kb BamHI DNA fragment containing the H-*ras* gene which was excised from the pEJ6.6 plasmid (Shih and Weinberg, 1982). 29 positive colonies were detected. 10 of these were purified and shown to contain the 6.6kb BamHI fragment containing the H-*ras* gene. Restriction analysis of two cosmids with a minimal degree of overlap, RAS15-10 and RAS5-3, was performed using the restriction endonucleases HindIII, BamHI and EcoRV. These cosmids have one common BamHI fragment, that which contains the H-*ras* gene, and no common HindIII fragments. The restriction map showed that the H-*ras* gene was contained on a HindIII fragment of about 56kb and would not be cloned in a HindIII partial cosmid library. Figure 4.2 shows the approximate restriction map of the region.

Figure 4.2 Partial restriction map of c-Harvey-ras-1 locus in the ALL genome.

(a) = partial restriction map of RAS5-3, (b) = partial restriction map of RAS15-10, and (c) = deduced restriction map of the region around the c-Harvey-*ras*-1 locus in the ALL genome. H = HindIII, B = BamHI and E = EcoRV.


(a)
(several unmapped
BamHI sites)
B
B
B
B
vector
H H
(b)
B B B
H
vector
(c)
B B B B B
H
H H
Cloned DNA
Hras gene
5kb

4.4 Isolation of H-ras from a Sau3A partial E65-9 cosmid library.

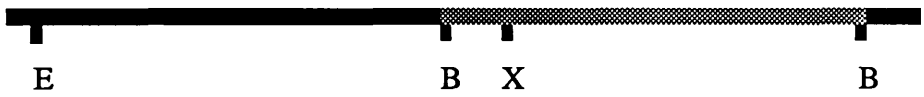
A Sau3A library was constructed from DNA from the E65-9 cell line in the cosmid vector Lawrist 4 as described in chapter 2. $2.2x10^5$ clones were screened and one clone containing the H-*ras* gene was isolated (HRAS). This clone was partially restriction mapped using a combination of single and double digests with BamHI, EcoRI, XhoI and HindIII and analysis of these digests on southern blots. It did not contain a 6.6kb BamHI fragment and shared only one 2.2kb BamHI fragment in common with RAS15-10 and no BamHI fragments in common with RAS5-3. The restriction map is shown in figure 4.3 and compared to the published genomic restriction map of the c-Harvey-*ras*-1 locus in EJ-18-8D, the parent cell line, taken from Morten *et al.* (1987). Restriction mapping indicated that only the 5' end of the gene had been cloned in HRAS. Additionally, a molecular rearrangement with respect to the parent cell line was detected in the region adjacent to the 5' BamHI site of the gene. This was demonstrated by the presence of a HindIII site 2.7 kb away from the BamHI site, 5' to the gene. This HindIII site is not present in the published restriction map of the DNA of the parent cell line or in the restriction map obtained for this region from the RAS15-10 and RAS5-3 clones. HRAS contains several HindIII sites none of which it shares with the two clones isolated from the ALL library. Many hybrid cell lines made by CMGT have been observed with rearranged and amplified regions of human DNA (Morten *et al.*, 1987). E65-9 was a cell line that exhibited both these changes (Harrison-Lavoie *et al.*, 1989). It is apparent from this analysis that the region adjacent to the H-*ras* gene was rearranged with respect to the parent cell line, EJ-18-8D, and with respect to ALL genomic DNA.

The clone containing the H-*ras* gene isolated from E65-9, HRAS, was entered into the fingerprint analysis and linked to the end of contig B (figure 4.1).

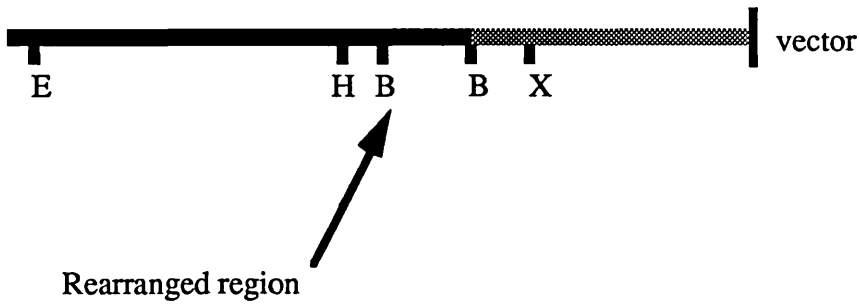
Figure 4.3 Partial restriction map of c-Harvey-ras-1 locus in the E65-9 genome.

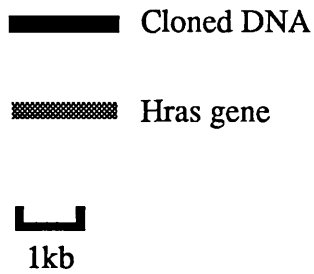
(b) = partial restriction map of c-Harvey-*ras*-1 locus in E65-9 determined from analysis of a cosmid clone, HRAS, and (o) = restriction map of c-Harvey-*ras*-1 locus in EJ-18-8D taken from Morten *et al.* (1987). H = HindIII, B = BamHI and X = XhoI.

(a)


E
B
X
B

(b)


vector
E
H
B
B
X
Rearranged region


Cloned DNA
Hras gene
1kb

This linkage was confirmed by comparing HindIII digests of HRAS with B441 and C437, two E65-9 clones that were shown by fingerprinting to overlap HRAS. HRAS and B441 shared all HindIII fragments in common except the fragment that contains the H-*ras* gene.

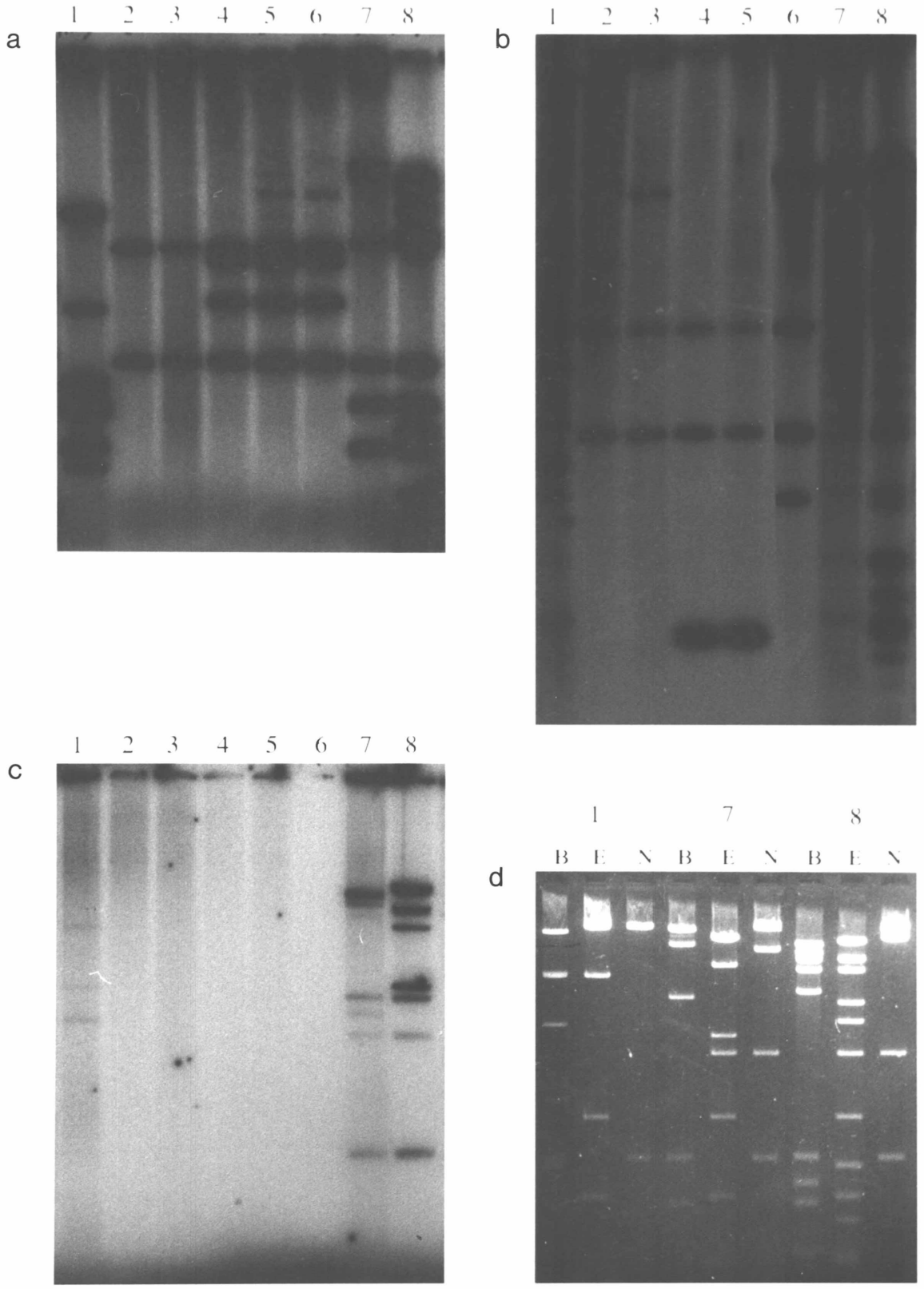
4.5 Rare-cutting restriction analysis of H-ras contig.

The results of rare-cutting restriction analysis on the minimal set of clones of contig B including the separately isolated clones containing the H-*ras* gene from both the Sau3A partial E65-9 library (HRAS) and the Sau3A partial ALL library (RAS15-10 and RAS5-3), are shown in figure 4.4. The approximate positions of the rare-cutting restriction enzyme sites are shown on the assembled contig in figure 4.1. The H-*ras* gene has previously been shown to be associated with a large number of these sites (Capon *et al.*, 1983) and several bands were expected in the RREF analysis. HRAS gave 8 bands with BssHII, 9 bands with EagI and 3 bands with NotI. The RAS15-10 cosmid from the ALL cell line gave 7 bands with BssHII, 10 bands with EagI and 7 bands with NotI and RAS5-3 gave 13 bands with BssHII, 20 bands with EagI and 7 bands with NotI. Caesium-purified DNA was digested with the three restriction enzymes (figure 4.4 (c)) to confirm the presence of these sites. HRAS was shown to contain 5 BssHII sites, 4-5 EagI sites and 2 NotI sites. The cosmid clone RAS15-10 has 4 BssHII, 6 EagI sites and 4 NotI sites. The cosmid clone RAS5-3 has 8 BssHII sites, 10 EagI sites and 4 NotI sites. The results obtained with the single DNA digests agreed with the RREF analysis. In some cases the number of bands obtained in the rare end-labelling process was less than expected. This might result from: (a) Two end-labelled fragments of a similar size that were not resolved on the agarose gel and appeared as one band on the autoradiograph, (b) A small end-labelled fragment being obscured by the signal from the unincorporated radiolabelled nucleotide, or (c) The occurrence of two of the rare-cutting restriction sites on one HindIII fragment producing three bands instead of the expected four bands that would result if the two sites were on different HindIII fragments.

Figure 4.4 RREF of the minimal set of clones forming the E65-9 c-Harvey-ras-1 contig.

Cosmid clones RREF with (a) BssHII, (b) EagI and (c) NotI. The two vector-generated fragments are indicated with an arrow (the clones containing DNA from the c-Harvey-*ras*-1 gene in lanes 1, 2 and 3 were isolated from Sau3A partial libraries and the second vector fragment is not of a fixed size is not marked). Figure (d) is an agarose gel of DNAs from the cosmid clones containing regions around the c-Harvey-*ras*-1 gene digested with BssHII (B), EagI (E) and NotI (N).

Lane 1 = HRAS, a cosmid clone from the Sau3A partial E65-9 library containing DNA from the c-Harvey-*ras*-1 locus, 2 and 3 = RAS15-10 and RAS5-3, cosmid clones from the Sau3A partial ALL library containing DNA from the c-Harvey-*ras*-1 locus, 4 = C445, 5 = C444, 6 = C464, 7 = A449 and 8 = C397. HRAS, C445, C444, C464, A449 and C397 form the minimal set of overlapping clones representing contig B as shown in figure 4.1.


a
1 2 3 4 5 6 7 8
b
1 2 3 4 5 6 7 8
c
1 2 3 4 5 6 7 8
d
1 7 8
B E N B E N B E N

4.6 A CpG island is present in the rearranged DNA adjacent to the H-ras gene.

The distribution of rare-cutting restriction enzyme sites within the ALL clones differed significantly from those within the E65-9 clone indicating that the rearrangement of DNA 5' to the H-*ras* gene in E65-9 removed part of the H-*ras* CpG island. There also appeared to be novel rare-cutting restriction enzyme sites in the E65-9 clone indicating the presence of a second CpG island. Further work would be required to assess the importance of the new sites with respect to a CpG island and with respect to the H-*ras* gene. It is possible that the putative CpG island is associated with a second gene or that it contains regulatory elements that might influence the transcription of the H-*ras* gene.

4.7 Results of analysis of ZF clones from EJNAC4.1 and EJNAC4.1.5 cosmid libraries.

As part of the screen for coding sequences, cosmid clones were isolated on the basis of hybridisation to a specific probe designed to detect the zinc finger motif. This work is described in chapters 5-8. These cosmid clones were analysed using the rare-cutting restriction system for the presence of BssHII, EagI and NotI sites. The information obtained from this analysis would be useful in the construction of large scale restriction maps around these putative genes. The mapping could be used confirm the information obtained from *in situ* hybridisation to prometaphase spreads that many of the cosmids containing these sequences were clustered together, and also aid the assembly of contigs in these regions. This data would also be valuable in examining the relationship of CpG islands, which are mainly associated with housekeeping genes as described in chapter 1, and zinc finger genes which are thought to be involved in specific cellular processes more likely to be spatially or temporally specific in their expression.

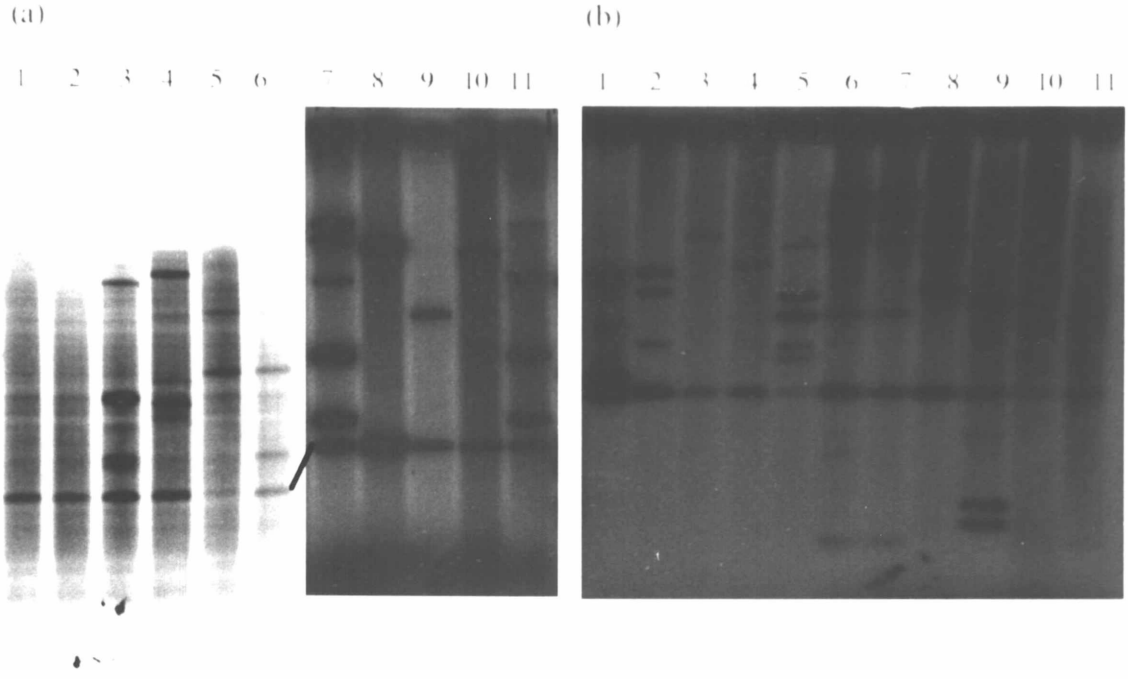
101 clones were analysed by the RREF method, the DNA being prepared by the microtitre method of Gibson and Sulston (1987). The analysis was performed with BssHII, EagI and NotI. The reactions with BssHII and EagI were monitored by the

internal control of endlabelling the vector sites. There was no control for the NotI reaction as there are no internal sites for this restriction enzyme in the vector. Results were obtained for 91% of the BssHII reactions and 90% of the EagI reactions and a success rate of 90.5% was assumed for the NotI analysis. Figure 4.5 shows the typical results obtained with RREF applied to 11 clones with BssHII (a), EagI (b) and NotI (c): ZF93, ZF94, ZF95, ZF96, ZF97, ZF98, ZF99, ZF100, ZF101, ZF102 and ZF103. With BssHII and EagI one of the vector fragments gave a constant band of 1.25kb and 1.74kb respectively. The second vector fragment of a constant size was not observed with these clones as they were isolated from a Sau3A partial EJNAC4.1.5 library and the insert DNA adjacent to this vector fragment will not be cleaved away by HindIII resulting in a fragment of a variable size. Table 4.1 lists the overall results of the analysis.

Lindsay and Bird (1987) predict that NotI sites will occur 10 times less frequently than BssHII and EagI sites and as expected, fewer ZF clones contained NotI sites than BssHII or EagI sites. Of the 101 clones analysed ZF13, ZF23, ZF31, ZF32, ZF38, ZF57, ZF84, ZF93, ZF94, ZF95, ZF98, ZF99, and ZF103 gave bands when rare-cutting restriction fingerprinted with NotI. Figure (c) shows the bands obtained for 6 of these cosmids. To confirm the accuracy of these results, caesium-purified DNA of all these clones was digested with NotI alone and resolved on 0.7% agarose gels. Confirmation of the presence of a NotI site was obtained for 12 of the 13 clones with digestion of ZF32 DNA producing 2 bands as predicted. The digestion of ZF93, ZF94, ZF95, ZF98, ZF99 and ZF103 with NotI are shown in figure (d). ZF95 could only be digested partially or not at all with NotI even when different DNA preparations were used. Rare-cutting restriction fingerprinting analysis of this clone with EagI gave two strong bands resulting from end-labelling of the vector site plus a third weaker band. The presence of sites for EagI and NotI in this clone could not be confirmed.

Figure 4.5 RREF of 11 ZF cosmid clones.

RREF with (a) BssHII, (b) EagI and (c) NotI of DNA from 11 putative ZF gene-containing cosmid clones: ZF93 (1), ZF94 (2), ZF95 (3), ZF96 (4), ZF97 (5), ZF98 (6), ZF99 (7), ZF100 (8), ZF101 (9), ZF102 (10) and ZF103 (11). Figure (d) is an agarose gel of NotI digested (C) and undigested DNA (U) of six of the clones predicted to contain NotI sites: ZF93 (1), ZF94 (2), ZF95 (3), ZF98 (6), ZF99 (7) and ZF103 (11).


(a)
1 2 3 4 5 6 7 8 9 10 11
(b)
1 2 3 4 5 6 7 8 9 10 11

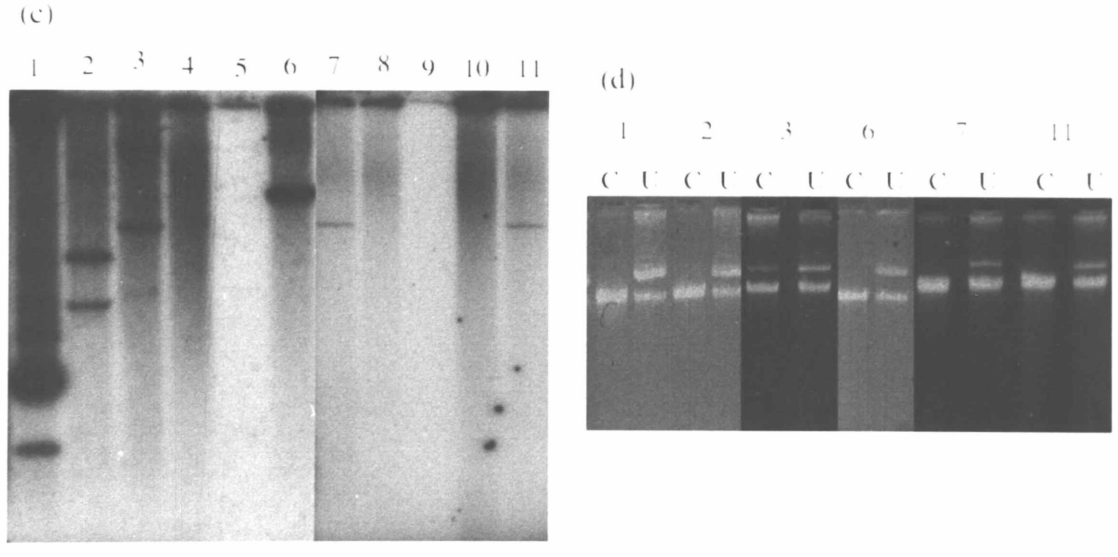

(c)
1 2 3 4 5 6 7 8 9 10 11
(d)
1 2 3 6 7 11
C U C U C U C U C U C U

Table 4.1 Summary of rare-cutting restriction enzyme sites in ZF cosmids.

The information obtained from rare-cutting restriction enzyme fingerprinting analysis of 101 ZF cosmids is summarised in the table. The minimum number of sites within the cloned DNA is shown. ND indicates that no data was obtained for the clone.

CLONE	BSSHII	EAGI	NOTI
ZF1	0	2	
ZF2	1	0	
ZF3	0	1	
ZF4	1	0	
ZF6	1	3	
ZF8	1	1	
ZF9	0	2	
ZF10	0	1	
ZF11	0	0	
ZF12	0	0	
ZF13	1	1	1
ZF14	0	0	
ZF15	0	0	
ZF16	0	2	
ZF17	1	1	
ZF18	1	0	
ZF19	0	0	
ZF20	1	0	
ZF21	0	ND	
ZF22	0	0	
ZF23	0	2	1
ZF24	0	0	
ZF25	0	2	
ZF26	0	0	
ZF27	0	1	
ZF28	0	3	
ZF29	1	7	
ZF30	0	3	
ZF31	0	ND	1
ZF32	ND	ND	2
ZF33	ND	1	
ZF34	1	0	
ZF35	0	0	
ZF36	0	1	
ZF37	0	0	
ZF38	2	3	1
ZF39	0	3	
ZF40	1	ND	
ZF41	0	4	
ZF42	0	ND	
ZF43	0	0	
ZF44	0	1	
ZF45	2	0	
ZF46	0	0	
ZF47	ND	1	
ZF48	1	0	
ZF49	0	1	
ZF50	0	0	
ZF51	0	0	
ZF52	2	1	
ZF53	0	2	
ZF54	1	ND	
ZF55	ND	ND	
ZF56	4	0	
ZF57	4	0	1
ZF58	ND	0	

CLONE	BSSHII	EAGI	NOTI
ZF59	1	3	
ZF60	0	0	
ZF61	0	1	
ZF62	0	1	
ZF63	0	1	
ZF64	1	0	
ZF65	1	0	
ZF66	ND	0	
ZF67	2	1	
ZF68	0	0	
ZF69	0	0	
ZF70	1	0	
ZF71	1	2	
ZF72	ND	1	
ZF73	ND	2	
ZF74	ND	0	
ZF75	0	0	
ZF76	0	0	
ZF77	0	4	
ZF78	1	ND	
ZF79	0	1	
ZF80	0	1	
ZF81	0	1	
ZF82	0	0	
ZF83	0	0	
ZF84	0	2	1
ZF85	0	0	
ZF86	0	1	
ZF87	0	ND	
ZF88	ND	2	
ZF89	0	ND	
ZF90	0	ND	
ZF91	0	0	
ZF92	ND	1	
ZF93	0	2	1
ZF94	0	1	1
ZF95	2	(1)	(1)
ZF96	1	0	
ZF97	1	2	
ZF98	1	2	1
ZF99	2	2	1
ZF100	0	2	
ZF101	0	1	
ZF102	0	0	
ZF103	2	3	1

4.8 The ZF clones are enriched for BssHII, EagI and NotI sites.

67 cosmids were shown to contain at least one site for one of the restriction enzymes tested. There was no apparent correlation between the intensity of hybridisation (thought to be related to the number of ZF motifs present) and the presence or absence of sites. In mammalian genomes sites for BssHII and EagI occur on average every 100kb and sites for NotI occur every 1,000kb (Sambrook *et al.*, 1989). In the cosmids isolated on the basis of hybridisation to the zinc finger oligonucleotide, BssHII sites occurred on average every 77.6kb, EagI sites every 37kb and NotI sites every 260kb, assuming the average insert size for a cosmid is 37.5kb (table 4.2). There was a greater than expected number of rare-cutting restriction enzyme sites in this class of cosmid clone.

Molecular analysis of approximately 1500kb of mouse DNA showed that CpG islands were located an average of about 100kb apart (Brown and Bird, 1986) and Lindsay and Bird (1987) predicted that 74% of BssHII and EagI sites, and 89% of NotI sites occur within CpG islands and are therefore associated with genes. If 89% of the clones containing a NotI site plus all the clones containing both BssHII and EagI sites plus 74% of the clones containing either a BssHII or an EagI site are taken as definitely containing a CpG island than this data suggests that a minimum of 55% of the clones contain CpG islands and would be predicted to contain genes. If the putative islands are associated with the ZF sequences then this is strong evidence that these sequences are genes. Additionally, the association of CpG islands with genes containing the ZF motif may reflect their role as general "housekeeping" genes or these may be other examples of genes with a specific pattern of expression, like the human α-globin gene (Bird *et al.*, 1987) which are associated with CpG islands. One way to determine this would be to locate the putative coding sequences within these cosmids and using these to screen RNA blots or cDNA libraries.

Table 4.2 Number of rare-cutting restriction enzyme sites detected in the ZF cosmid clones compared to the number expected.

An average insert size for each cosmid of 37.5kb is assumed. The expected number of sites is taken from data in Sambrook *et al.*, 1989.

Enzyme	Cosmids Checked	Total DNA screened (kb)	Sites detected	Sites expected
BssHII	89	3337.5	43	33.37
EagI	90	3375	92	33.75
NotI	89.5 (av.)	3356.25	13	3.35

Chapter 5.

Detecting the ZF motif in cosmid clones from two hybrid cell lines: EJNAC4.1 and EJNAC4.1.5.

5.1 Introduction.

The use of oligonucleotides as hybridisation probes designed to detect a particular motif is an alternative route to identifying uncharacterised genes from a particular region of the genome. In this study cosmid clones isolated from the EJNAC4.1 and EJNAC4.1.5 hybrid cell lines were screened with a redundant oligonucleotide designed to detect genes that encode zinc finger proteins which contain the H/C link, TGEKPY, observed by Chowdhury *et al.*, 1987 and Schuh et at., 1986 in the *Krüppel* gene, the *Kr h* gene and the two mouse genes, *mkr1* and *mkr2* (table 1.1). Clones containing the zinc finger motif from the regions of human DNA in the hybrid cell lines would be identified in this way. The involvement of proteins containing the ZF motif in complex developmental processes makes their isolation and analysis an important priority. There is a reasonable likelihood that any gene with the zinc finger motif will bind DNA and could therefore be involved in the control of gene expression. A number of human genetic diseases which result from disturbances of the normal pattern of development such as growth retardation or uncontrolled cell growth might involve zinc finger genes, as demonstrated by the isolation of a zinc finger gene proposed to be involved in the development of Wilm's tumour (Call *et al.*, 1990 and Gessler *et al.*, 1990). The physical mapping of these genes with respect to the regions implicated in disorders of this type is important.

The oligonucleotide (designed by Ashworth and Little) used in this study was a 144-fold redundant 20 base oligonucleotide with an exclusion of the CpG dinucleotide, previously shown to be under-represented in mammalian DNA (Bird 1986), which was designed to detect the H/C link found in many genes of the zinc finger family (figure 1.1).

0.5% of the clones in a mouse lymphoblast cDNA library (Crossley, pers. comm.) and 0.3% of the clones in a human foetal kidney cDNA library (Little, pers. comm.) were detected by the zinc finger (ZF) oligonucleotide. Variable of intensities of hybridisation signal were observed which were reproducible on isolation of the clones. Sequence analysis of random cDNA clones shows that this was due to

variability in the number of finger motifs contained within the clone. These findings are in broad agreement with Bellefroid *et al.* (1989) who reported the partial sequence of randomly isolated finger protein cDNA clones. 0.5% of cosmid clones from a human genomic cosmid library hybridised to the ZF oligonucleotide when filters of the plated library were screened and 1% of a colony array of 384 random clones from this library hybridised with the ZF oligonucleotide. This represents an estimated minimum of 250, and a maximum of 1000 zinc finger genes in the human genome, assuming there are 50,000-100,000 genes in the human genome. This data is compatible with that of Bellefroid *et al.* (1989) who predicted that there are at least 300 genes containing the H/C zinc finger link.

5.2 Screening cosmid clones isolated from the EJNAC4.1 and 4.1.5 somatic hybrids.

Two somatic cell hybrid cell lines, EJNAC4.1 and EJNAC4.1.5, which were enriched for the short arm of human chromosome 11, were being used in the mapping project. In addition to 11p, both EJNAC4.1 and EJNAC4.1.5 contain regions from other human chromosomes these being 19p, 19qter, 20p, 21qprox and some centromeric regions. EJNAC4.1 additionally contains part of the short arm of human chromosome 3. The screening of cosmid libraries constructed from these hybrid cell lines with a total human DNA probe identifies clones containing genomic DNA from these specific human chromosomal regions which are then analysed to produce a cloned DNA map of these regions. The systematic screening of these clones with the ZF oligonucleotide should identify genes containing the zinc finger motif within these regions.

5.3 The sensitivity of colony arrays versus southern blots.

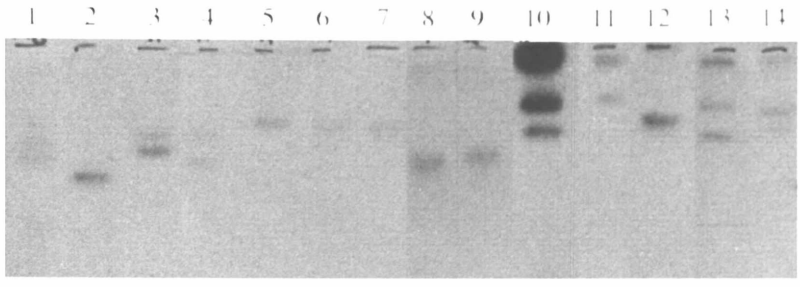
Southern blot analysis of cosmid clones isolated from the EJNAC4.1 library with total human DNA and total mouse genomic DNA probes allowed the identification of mis-picked clones.

Figure 5.1 Comparison of ZF oligonucleotide hybridisation signal of southern blotted DNA with colony array DNA.

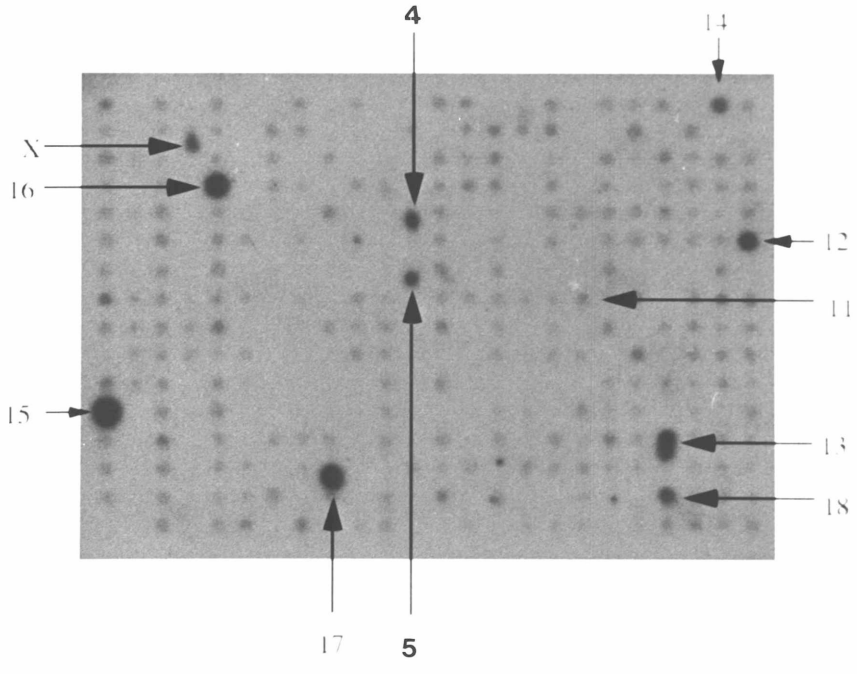
(a) Results of probing of southerns blots of DNAs from ZF12 (1), ZF13 (2), ZF14 (3), ZFm-1 (4), ZFm-2 (5), ZF15 (6), ZF16 (7), ZF17 (8), ZF18 (9), ZF19 (10), ZF20 (11), ZF21 (12), ZF22 (13) and ZFm-3 (14) with the ZF oligonucleotide.

(b) Results of probing of a colony array of 384 cosmid clones with the ZF oligonucleotide. Numbering as for (a) and with additional positives: ZF29 (15),ZF30 (16), ZF31 (17) and ZF65 (18). X = a background mark.

(a)



(b)

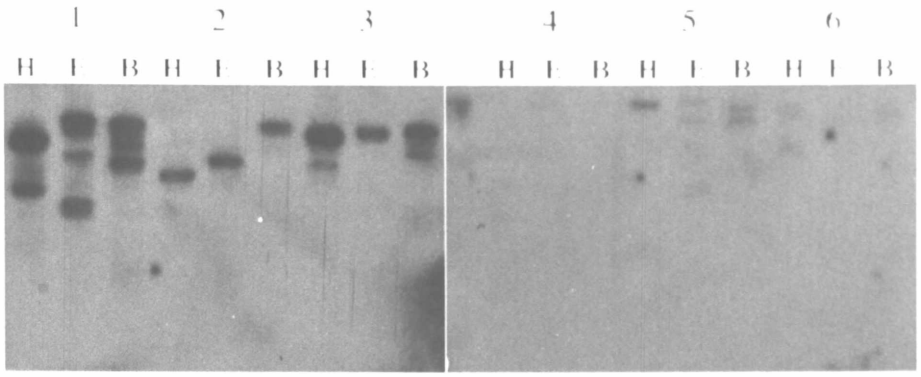


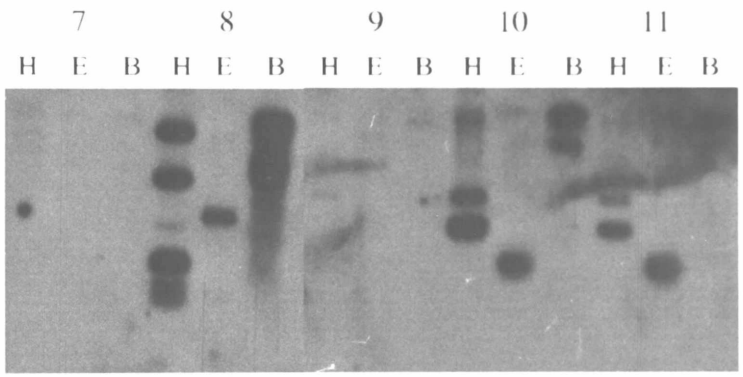
The southern blots were then used to identify clones which hybridised to the ZF oligonucleotide. 799 human DNA-containing clones were analysed on southern blots. 26 gave a signal with the ZF oligonucleotide. Figure 5.1 (a) shows the hybridisation signal obtained for the 14 of the clones ZF12, ZF13, ZF14, ZFm-1, ZF15, ZFm-2, ZF16, ZF17, ZF18, ZF19, ZF20, ZF21, ZF22 and ZFm-3. "Minipreped" DNA from these 26 cosmid clones was digested with three restriction enzymes, transferred to nylon filters and hybridised to the oligonucleotide probe, and with total human DNA and total mouse genomic DNA probes. 24 were confirmed as hybridising to the ZF oligonucleotide. Three of these, ZFm-1, ZFm-2 and ZFm-3, did not hybridise to a total human DNA probe but did hybridise to a total mouse DNA probe indicating that they were clones of mouse DNA. There was a significant variation in the intensity and pattern of hybridisation obtained with these clones. Figure 5.2 shows the results obtained from analysis of the 11 human clones, (ZF12, ZF13, ZF14, ZF15, ZF16, ZF17, ZF18, ZF19, ZF20, ZF21 and ZF22). Some clones had a single band that hybridised to the ZF oligonucleotide after digestion with all three restriction enzymes (the examples shown being ZF13 and ZF22) suggesting the presence of one discrete ZF sequence. Other clones had several fragments that hybridise to the oligonucleotide (the example shown being ZF12) suggesting blocks of ZF sequence separated by non-ZF sequence, possibly representing several exons of one large gene or representing more than one ZF gene within the cosmid insert.

Colony arrays of the cosmid clones replaced the role of "microprep" southern blots as the preferred method for distinguishing mis-picked clones which contain cloned mouse DNA. As these were available for screening with the ZF oligonucleotide it was important to determine how sensitive they would be in the detection of cosmids containing the ZF sequence. To this end an equivalent screen of the 384 cosmid clones on southern blots and colony arrays was compared. Figure 5.1 (b) shows the results obtained from screening a colony array of 384 cosmids clones with the ZF oligonucleotide.

Figure 5.2 Comparison of pattern of hybridisation with ZF oligonucleotide with 11 cosmids.

Southern blot of DNAs from ZF12 (1), ZF13 (2), ZF14 (3), ZF15 (4), ZF16 (5), ZF17 (6), ZF18 (7), ZF19 (8), ZF20 (9), ZF21 (10) and ZF22 (11) digested with HindIII (H), EcoRI (E) and BamHI (B) hybridised with the ZF oligonucleotide.


1 2 3 4 5 6
H E B H E B H E B H E B H E B H E B


7 8 9 10 11
H E B H E B H E B H E B H E B

This blot includes DNA from the 11 human clones, ZF12-ZF22, and the three mouse clones, ZFm-1, ZFm-2 and ZFm-3 that hybridised with the ZF oligonucleotide on the southern blots. There was a significant background of hybridisation when the ZF oligonucleotide was hybridised to the colony array. 6 of the 14 clones that gave a signal with the ZF oligonucleotide gave a signal above background on the colony array, these being ZFm-2, ZF15, ZF20, ZF21, ZF22 and ZFm-3. ZF12, ZFm-1, ZF16, ZF17 and ZF19 did were not detected as the clone had not grown on the colony array (note absence of background hybridisation). When a second colony array of the same 384 clones was screened, ZF12 and ZF19 were detected. ZF13, ZF14, ZFm-1, ZF16, ZF17 and ZF18 showed no hybridisation signal above background on either colony array. The clones that failed to give a signal above background on the colony array were weakly hybridising clones except for ZF13 which was a medium hybridising clone. Four additional clones were detected on the colony arrays which had been missed on the screen of the southern blots as a result of loss of samples during the preparation of southern blots, these being ZF29, ZF30, ZF31 and ZF65. There was a total of 18 possible ZF clones out of the 384 clones analysed: 14 of these were detected on southern blots and 12 were detected on the colony arrays.

5.4 Analysis of cosmid clones hybridising to the ZF oligonucleotide.

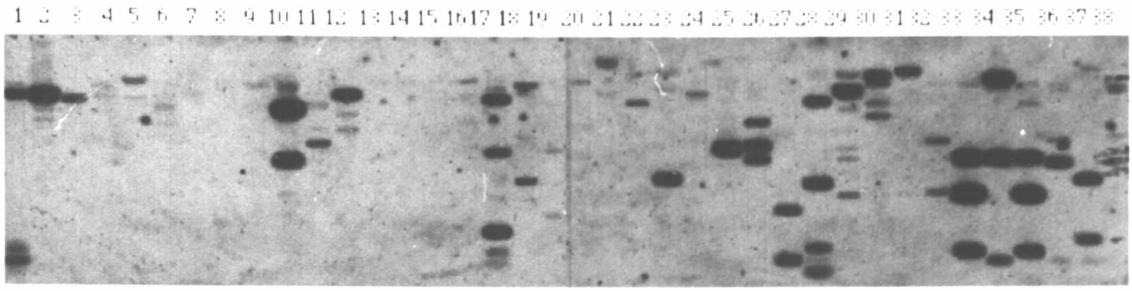
Having determined the level of sensitivity of screening colony arrays, the cosmid clones detected by the ZF oligonucleotide were processed as follows: After a primary screen of filters generated from the mapping project, the clones hybridising to the oligonucleotide were analysed on southern blots by sequential hybridisation with the ZF oligonucleotide, with a total mouse DNA probe and with a total human DNA probe. An example of one analysis is shown in figure 5.3 for southern blot A which contains DNA from 38 ZF cosmid clones digested with HindIII.

This analysis allowed a comparison of the sizes of hybridising HindIII fragments and intensities of signal with the ZF oligonucleotide and with the total human probe to identify overlapping cosmids.

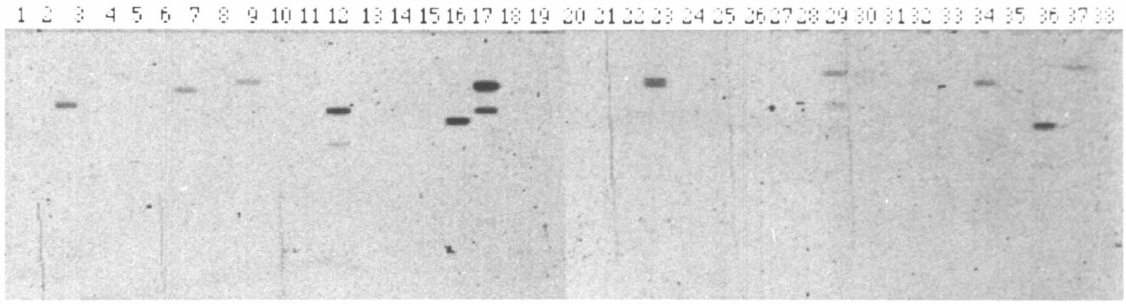
Figure 5.3 Probing Southern blot A of 38 cosmid clones.

Pattern of hybridisation obtained with a southern blot of DNAs from ZF1 (1), ZF2 (2), ZF3 (3), ZF4 (4), ZF6 (5), ZF8 (6), ZF9 (7), ZF10 (8), ZF11 (9), ZF12 (10), ZF13 (11), ZF14 (12), ZF15 (13), ZF16 (14), ZF17 (15), ZF18 (16), ZF19 (17), ZF20 (18), ZF21 (19), ZF22 (20), ZF23 (21), ZF24 (22), ZF25 (23), ZF26 (24), ZF27 (25), ZF28 (26), ZF29 (27), ZF30 (28), ZF31 (29), ZF32 (30), ZF33 (31), ZF34 (32), ZF35 (33), ZF36 (34), ZF37 (35), ZF38 (36), ZF39 (37) and ZF40 (38) digested with HindIII with (a) the ZF oligonucleotide, (b) a mouse genomic DNA probe and (c) a human genomic DNA probe.

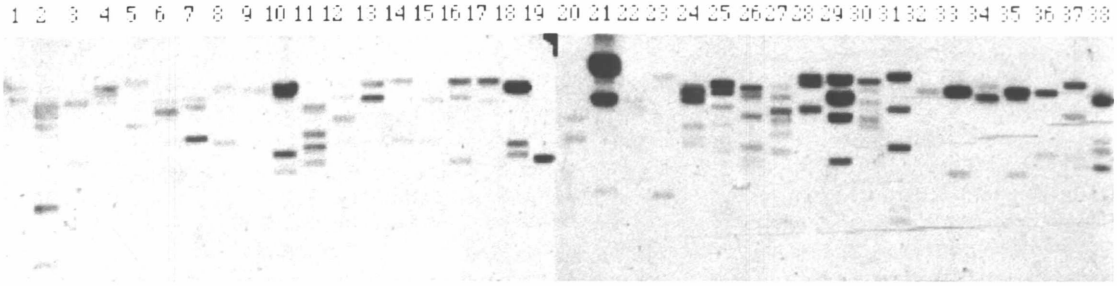
(a)



(b)



(c)



ZF fragments not containing human repeat sequences were identified which could be used as probes for screening cDNA libraries and genomic southern blots. In some cases fragments hybridised to the total mouse DNA probe. These were likely to result from co-ligation of mouse and human DNA fragments although it was possible that some represented sequences conserved between mouse and human genomes or junction fragments between mouse and human material that exist in the hybrid cell lines.

5.5 Sequencing fragments hybridising to the ZF oligonucleotide.

To confirm the presence of the ZF motif in the fragments of DNA hybridising to the ZF oligonucleotide, cosmid clones were digested with a restriction enzyme which had been determined on southern blots to generate small fragments that hybridise to the ZF oligonucleotide. These fragments were partially or totally sequenced. Table 5.1 lists the results obtained from sequencing fragments from 20 of the 38 ZF cosmid clones analysed on southern blot A. Sequence information from 11 of the subcloned fragments confirmed the presence of the ZF motif. Only partial sequence information was obtained on the other clones. It was probable that the ZF motif was present in all the fragments and would be seen when the complete sequence was obtained. Figure 5.4 shows the amino acid sequence predicted from the DNA sequence of the 11 fragments.

Both the H/C link and the Cys-X_{2-4}-Cys-X_3-Phe-X_5-Leu-X_2-His-X_{3-4}-His-X_6 amino acid sequence characteristic of ZFP genes were observed in all 11 fragments. The results of this analysis indicate that all the DNA fragments that hybridise to give a medium to strong signal with the ZF oligonucleotide will contain the ZF motif in many copies.

Table 5.1 Summary of sequence information obtained on fragments from ZF cosmids.

All the fragments completely or partially sequenced hybridised with medium to strong intensity with the ZF oligonucleotide. The size of the fragment, the number of bases sequenced and the number of ZF motifs detected is shown. - = no sequence homology to ZF oligonucleotide, + = 1 copy of ZF motif, ++ = 2 copies...

Parent cosmid clone.	Size (bp)	Sequence (bp)	Motifs
ZF1	257	257	+++
ZF2	5400	400	-
"	2300	260	-
"	2700	250	-
ZF3	?	485	+++
ZF4	?	355	-
ZF6	?	200	-
ZF10	?	230	-
ZF11	1600	160	-
ZF12	600	295	++
ZF13	900	317	-
ZF18	2300	230	-
ZF19	1100	160	++
ZF21	2800	410	-
ZF25	?	305	++++
ZF29	173	173	++
ZF30	87	87	+
"	1000	500	-
ZF32	800	267	+
ZF34	450	440	+++
"	200	200	++
ZF39	329	329	++++
ZF40	500	262	-

Figure 5.4 Predicted amino acid sequence of 11 DNA fragments hybridising with ZF oligonucleotide.

Amino acid sequence (single letter code) of fragments from ZF1 (1), ZF3 (2), ZF12 (3), ZF19 (4), ZF25 (5), ZF29 (6), ZF30 (7), ZF32 (8), ZF34 (9 and 10) and ZF39 (11). The consensus ZF sequence is shown below (12). * indicates a stop codon.

(1) H N I V H T G D K P Y K

C K D C G K I F K W S S N L T I H Q R I H S G E K P Y K

C E E C G K A F K Q S S K L N E H M R A H T G E K F Y K

C E E C G K A

(2)

C E I C E K K F Y A M A H V R K H M V A H M K D M P F T

C E T C G K S F K C S M S L M G H S L Q H S G E K S F R

C R N C D E G F S Y N Y Q L H T H M S I H V A H K R S C

(3) H T G E K P Y E

C K E C G K P F S F L T G F R V H M R M H T G E K P Y K

C K D C G N A F I W R A S L Q Y H V K K V H

(4) H K K S H T E E K P Y K

Y E E C G K G F N G P Q P L T I H K I I H T G E K P Y K

C E E C G K A

(5) H T G E K P Y K

C E E C G Q A F I S S S T L N G H K R I H T R E K P Y K

C E E C G K A F S Q S S T L T R H K R L H T G E K P Y K

C G E C G K A L K S P Q L L

(6) F Y R F S Y L I K H K T S H T G E K F Y K

C E E C G K G F N W S S A L T K H K R I H T G E K P Y K

C E E C G K A

(7) F N L S S H L T T H K I I H T G E K P Y K

C E E C G K A

(8) H T G E K P Y H

C N R C G K T F C D C T D F S Q H

(9) C G T S F I W S S Y L I Q H K K T H T G E K P Y E

C D K C G K V F R N R S A L T K H E R T H T G I K P Y E

C N K C G

(10) H F T A E K P Y D

Y N E C G T F F H L E L L P Y S T * E K L I L E K K T Y E

C D K C G K V F R N R S A L T K H E R T H T G I K P Y E

C N K C G

(11) K A F Y * F L Y L T K H K I I H K G D K L Y K

C E E C G K A F S I F S T L T K H K I I H T E E K P Y K

C E E C G K A F S L S L Q F T A H K I I Y T G E K P Y K

C E K C G K P F N Q S S T L T T H * I I H A G E K P Y K

C E K C G K A

(12)

C X X C X X X F X X X X X L X X H X X X H T G E K P Y X

The sequence analysis of fragments hybridising weakly to the ZF oligonucleotide was not attempted. It is possible that these weakly hybridising fragments contain regions of ZF genes with only one or two copies of the H/C link present or that they contain copies of the H/C link significantly diverged from the HTGEKPY motif or that they contain sequences not related to ZF genes. More extensive sequence analysis of a number of the fragments in this category would be required to determine the nature of the sequence hybridising weakly to the ZF oligonucleotide.

5.6 There are more human ZF sequences in the hybrid cell lines than expected.

To date 101 cosmid clones containing human DNA have been isolated that hybridise to the ZF oligonucleotide. 45 were isolated from screening 1017 cosmid clones from the EJNAC4.1 library and 56 were isolated from screening 1166 clones from the EJNAC4.1.5 library. This represents 4.4% and 4.8% of the human DNA-containing clones respectively. The number of these sequences in the human DNA component of these cell lines was much greater than expected from studies of a total human DNA cosmid library where 0.5-1% of clones hybridise to the ZF oligonucleotide.

In the EJNAC4.1 cosmid library 1.7% of the clones hybridise to a total human DNA probe. This represents approximately 200Mb (1.7% of 1.2 x 10^{10}bp) of human DNA roughly equivalent to 5300 non-overlapping cosmids with an average insert size of 37.5kb. 1017 cosmid clones have been screened or about 0.2 genome equivalents. EJNAC4.1.5 (1% human cosmid clones) by the same calculation contains 120Mb of human DNA, and 1166 clones or 0.36 genome equivalents have been screened. As complete genome equivalents of the human DNA have not been screened from either cell line the majority of the cosmids identified would not be predicted to overlap. A comparison of the fragment sizes generated by HindIII digestion and the hybridisation patterns obtained with the ZF oligonucleotide and the total human DNA probes indicates that the majority of clones isolated are unique and do not overlap. There is a greater than four-fold higher number of zinc finger sequences in the hybrid cell lines than would be predicted. Either these sequences are distributed

evenly throughout the regions of human DNA within the hybrid cell lines or, more likely, there are one or more large clusters of ZF sequences at specific positions.

5.7 Linkage of ZF cosmids by Fingerprint analysis.

Over half the 4428 cosmid clones from the EJNAC4.1 and EJNAC4.1.5 libraries in the database have been placed in contigs. These include 33 of the 92 ZF cosmids that have been fingerprinted. Currently there are 20 contigs of between 2-4 occupancy which contain one ZF cosmid linked to clones not containing ZF sequence and 5 contigs that contain more than one ZF cosmid. The cosmids in 5 contigs containing more than one cosmid clone share common ZF sequence and are shown in figure 5.5. The fingerprint analysis and the analysis of HindIII fragment sizes and hybridisation patterns indicates that 94 of the cosmid clones contain unique ZF sequence.

Figure 5.5 Computer generated ZF cosmid contigs.

Five contigs are shown, A-E. The number of bands in common is represented by the degree of overlap of the lines. Clones numbers refer to the database code number. Contig A contains 6 comids, two of which are buried.

Contig A Z43 = ZF1, e1307 = ZF62, and E1320 = ZF63.

Contig B Z756 = ZF35 and Z816 = ZF37.

Contig C e2167 = ZF75 and e2229 = ZF76.

Contig D e2316 = ZF81 and e2318 = ZF82.

Contig E e2476 = ZF85 and e2673 = ZF101.

A

e2543
e1307
E1320
Z43

B

Z816
Z756

C

e2176
e2229

D

e2316
e2318

E

e2476
e2673

Chapter 6.

Detection of two groups of related ZF sequences.

6.1 Introduction.

There are probably over 300 genes containing the HTGEKPY motif in the human genome. The conservation of the zinc finger motif in such a large group of proteins would indicate that either this motif is a unique solution to a particular biological function that has evolved in many instances and that the group represents an example of convergent evolution, or that there has been divergence from a common ancestral gene which has been duplicated many times over. If the ZF genes originated from duplications of common ancestoral genes, it is possible that some of these genes will be more closely related than others, having resulted from more recent duplication events. It also is possible that ZF genes that are more closely related in sequence will share a similar function. If this is the case, the ZF gene family could be divided on the basis of cross-hybridisation, into subgroups related by function.

6.2 Cross-hybridisation studies

101 cosmid clones identified as hybridising to the ZF oligonucleotide were used in cross-hybridisation studies. Fragments shown by sequence analysis to contain the ZF motif were used as probes against southern blots, DNA dot blots and colony arrays of the ZF cosmids. The blots were initially washed at 50^{o}C in Church buffer to detect any cross-hybridising sequences and the temperature then raised by 5^{o}C increments to 65^{o}C.

6.3 Identification of a group of cross-hybridising sequences.

In the first of the cross-hybridisation screens a 173bp HindIII fragment from ZF29, ZF29p1, was used as a probe. The complete predicted amino acid sequence of this fragment is shown in figure 5.4 (5). At the initial washing temperature of 55^{o}C, sequences contained in 19 of the 101 ZF cosmid clones cross-hybridised with ZF29p1. This pattern of hybridisation was unchanged at 60^{o}C (figure 6.1 (a)). At 65^{o}C (Fig 6.1 (b)), the stringent temperature, the probe detects the 173bp parent fragment from ZF29 and less strongly, a 5.7kb fragment from a second cosmid clone,

ZF19. The pattern of hybridisation at 60°C closely resembled that obtained with the ZF oligonucleotide (figure 5.3 (a)) indicating that the probe is cross-hybridising with sequences containing the ZF motif.

A reciprocal cross-hybridisation was performed with a fragment from ZF1, a cosmid that cross-hybridised with ZF29p1, to determine the degree of relatedness. Hybridisation with the 257bp fragment from ZF1, ZF1p2, gave the same pattern of cross-hybridisation as with the ZF29p1 probe, at 60°C. Figure 6.1 (c) shows a DNA blot of the first 63 cosmids and a colony array of the remaining 38 of the 101 ZF cosmids analysed, hybridised to the ZF1p2 probe and washed at 60°C. The complete predicted amino acid sequence of this fragment is shown in figure 5.4 (1). Of the 101 ZF cosmids analysed, 16 cross-hybridised strongly with both these probes. An additional three cosmids, ZF12, ZF27 and ZF97 hybridised to these probes less strongly. Cosmids cross-hybridising with these probes were defined as members of "Group 1".

6.4 A second group of related sequences.

The ZF motif is present in the "group 1" probes used to screen the ZF cosmids. It was possible that the probes were detecting the general presence of the ZF motif rather than a subset of more related sequences. If this was the case all fragments strongly hybridising with the ZF oligonucleotide would hybridise to these probes. A comparison of the pattern of hybridisation obtained with the "group 1" probes (figure 6.1(a)) with that obtained with the ZF oligonucleotide (figure 5.3(a)) indicated that this was not the case. There were strong ZF positives which were not members of the first group.

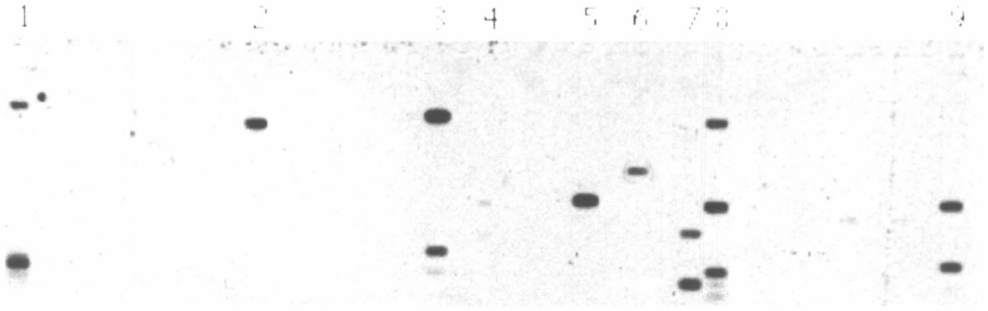
To test whether any of the strong ZF clones not in "group 1" were related by sequence, a 450bp fragment from ZF34, ZF34p2 (predicted amino acid sequence of part of this fragment shown in figure 5.4 (9)), was used in a low stringency screen.

Figure 6.1 Examples of "group 1" cross-hybridisation experiments.

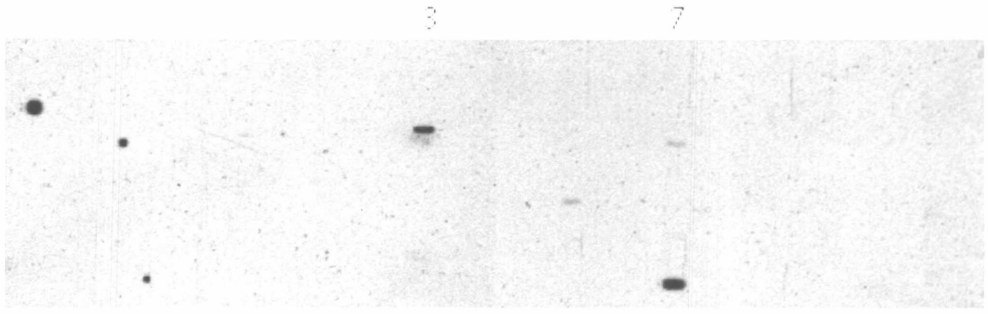
Southern blot A (see figure 5.3) hybridised with a "group 1" probe, ZF29p1, and washed at 60°C (a) and at 65°C (b). And (c) DNA dot blot of first 38 ZF cosmids and colony array of a further 63 hybridised with a second "group 1" probe, ZF1p2 and washed at 60°C. White arrows indicate clones that contained fragments used as probes.

Clones cross-hybridising with probes are numbered: 1 = ZF1, 2 = ZF12, 3 = ZF19, 4 = ZF21, 5 = ZF25, 6 = ZF27, 7 = ZF29, 8 = ZF30, 9 = ZF39, 10 = ZF62, 11 = ZF63, 12 = ZF78, 13 = ZF86, 14 = ZF89, 15 = ZF90, 16 = ZF92, 17 = ZF93, 18 = ZF97 and 19 = ZF102.

(a)



(b)



(c)

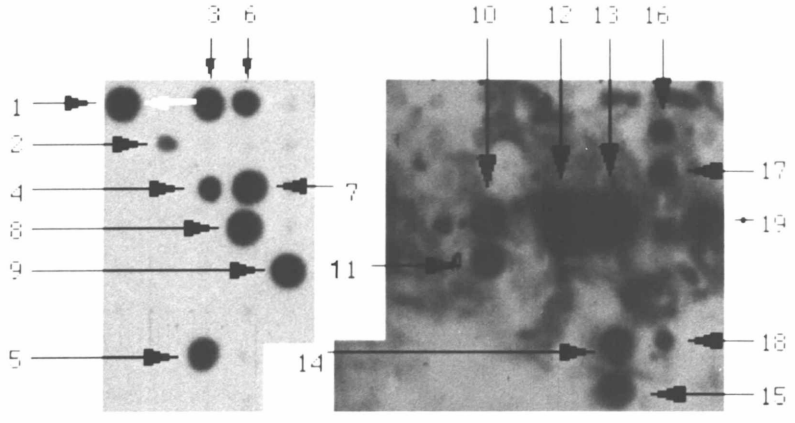
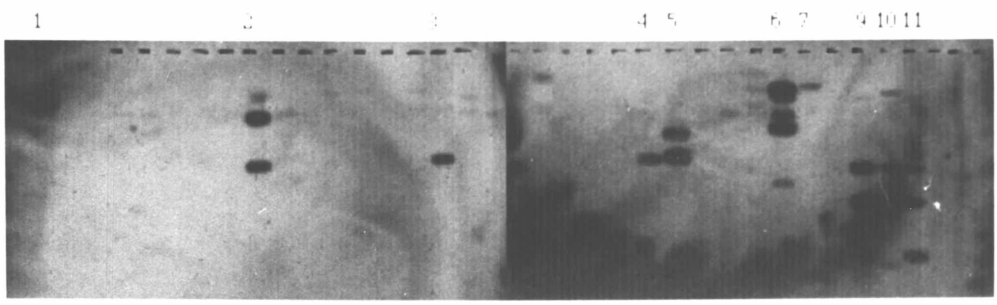


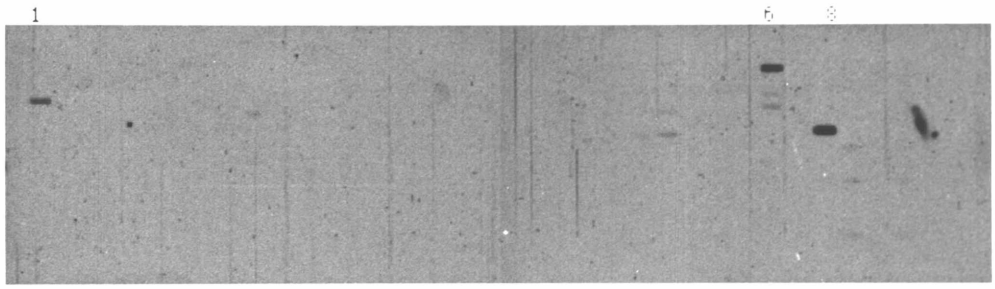
Figure 6.2 Examples of "group 2" cross-hybridisation experiments.

Southern blot A (see figure 5.3) hybridised with (a) a "group 2" probe, ZF32p4, and washed at 60°C, and then (b) a second "group 2" probe, ZF34p2, and washed at 60°C. Figure (c) is a DNA dot blot of first 61 ZF cosmids and colony array of a further 40 hybridised with a ZF32p4 probe and washed at 60°C. Clones cross-hybridising with probes are indicated with numbers 1 = ZF2, 2 = ZF12, 3 = ZF19, 4 = ZF27, 5 = ZF28, 6 = ZF32, 7 = ZF33, 8 = ZF34, 9 = ZF35, 10 = ZF36, 11 = ZF37, 12 = ZF38, 13 = ZF44, 14 = ZF46, 15 = ZF47, 16 = ZF48, 17 ZF52, 18 = ZF53, 19 = ZF55, 20 = ZF56, 21 = ZF59, 22 = ZF60, 23 = ZF61, 24 = ZF73, 25 = ZF89, 26 = ZF90, 27 = ZF93 and 28 = ZF96.

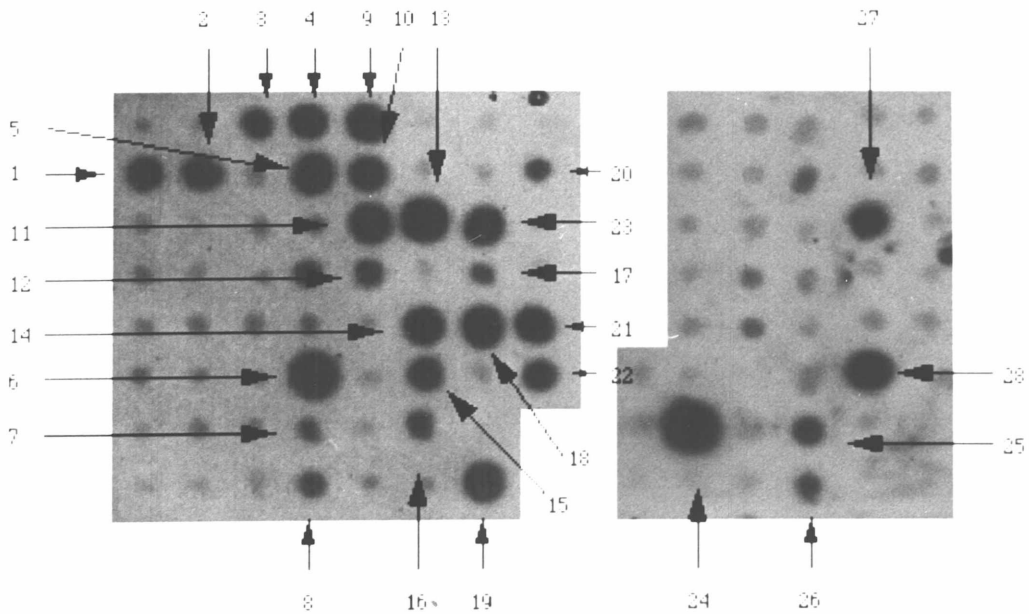
(a)


1
4 5
6 7
9 10 11

(b)


1
6
8

(c)


2
3
4
9
10
13
37
5
1
11
12
14
6
7
20
23
17
21
22
38
35
18
15
8
16
19
24
26

This fragment hybridised strongly to the ZF oligonucleotide and sequence analysis had identified the multiple copies of the ZF motif. ZF34p2 cross-hybridised with fragments in three other cosmids, ZF2, ZF32 and ZF46, at 60^{o}C (Figure 6.2 (b)). Again, a reciprocal cross-hybridisation was performed, with a fragment from ZF32 to determine the degree of relatedness. The 800bp fragment from ZF32, ZF32p4 (predicted amino acid sequence from part of this fragment is shown in figure 5.4 (7)), detected the same HindIII fragments in ZF2 and ZF46 as the ZF34p2 probe (figure 6.2 (a)) but did not detect any fragments from ZF34 on the southern blot. On the DNA dot blot it hybridised weakly to ZF34. In addition to ZF2 and ZF46, the ZF32p2 probe detected 16 other cosmids strongly and a further 9 with a weak signal. Fig (c) is a DNA blot of cosmid clones ZF1-63 plus a colony array of cosmid clones ZF64-101 hybridised with the ZF32p4 probe and washed at 60^{o}C. The cross-hybridisation data is summarised in table 6.1. The cosmids cross-hybridising with ZF34p4 are defined as members of "group 2". Cosmid clones that hybridised with probes from both groups were assigned to the group giving the more intense hybridisation signal. In total there are 17 cosmids in "group 1" and 24 cosmids in "group 2". The assignment of cosmids into these two groups is based on whether they cross-hybridise to the fragments used as probes, at a lowered stringency. It is possible that further analysis will detect other clones that contain related sequences that cross-hybridise. Fragments from two clones, ZF13 and ZF40, which are not members of either group, were used in similar cross-hybridisation studies but did not detect any fragments in other clones that hybridised to the ZF oligonucleotide. Systematic studies with other cosmid clones may reveal the existence of other groups.

6.5 The cosmids in "group 1" and "group 2" are a large proportion of the human ZF sequences in EJNAC4.1 and EJNAC4.1.5.

Restriction analysis and hybridisation studies of the cosmid clones in "group 1" and "group 2" indicated that a different cross-hybridising sequence was contained in almost every cosmid.

Table 6.1 Summary of "group 1" and "group 2" hybridisation experiments and intensities of hybridisation with ZF oligonucleotide.

The clones cross-hybridising with "group 1" and/or "group 2" probes are indicated by +. Where (+) is indicated the hybridisation signal was weak. The hybridisation signal; obtained with the radiolabelled ZF oligonucleotide is indicated as +++ for a clone hybridising strongly, ++ for a clone hybridising with medium intensity, + for a clone hybridising with weak intensity and (+) for a clone hybridising very weakly.

CLONE	"Group 1"	"Group 2"	Intensity with ZF.
ZF1	+		+++
ZF2		+	+++
ZF3			++
ZF4			+
ZF6			+
ZF8			+
ZF9			+
ZF10			(+)
ZF11			+
ZF12	(+)	+	+++
ZF13			++
ZF14			+++
ZF15			(+)
ZF16			(+)
ZF17			(+)
ZF18			+
ZF19	+	(+)	+++
ZF20			++
ZF21	+		++
ZF22			+
ZF23			+
ZF24			++
ZF25	+		+++
ZF26			+
ZF27	(+)	+	+++
ZF28		+	+++
ZF29	+		+++
ZF30	+		+++
ZF31			+
ZF32		+	+++
ZF33		(+)	++
ZF34		(+)	+++
ZF35		+	+++
ZF36		+	+++
ZF37		+	+++
ZF38		(+)	+++
ZF39	+		+++
ZF40			+
ZF41			++
ZF42			++
ZF43			+
ZF44		+	+++
ZF45			+
ZF46		+	++
ZF47		+	+++
ZF48		(+)	++
ZF49			++
ZF50			+
ZF51			++
ZF52		(+)	++
ZF53		+	+++
ZF54			++
ZF55		+	+++
ZF56		(+)	+
ZF57			+
ZF58			+

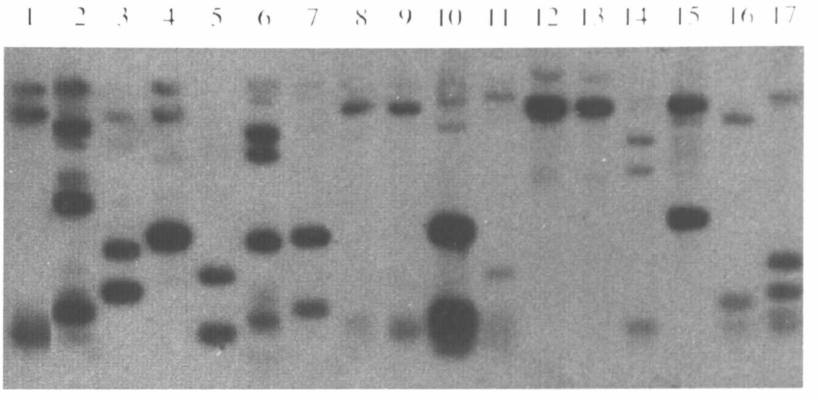
CLONE	Group 1	Group 2	Intensity with ZF.
ZF59		+	+++
ZF60		+	+++
ZF61		+	+++
ZF62	+		++
ZF63	+		+++
ZF64			(+)
ZF65			(+)
ZF66			+
ZF67			+++
ZF68			+++
ZF69			+++
ZF70			(+)
ZF71			+++
ZF72			++
ZF73		+	+++
ZF74			+++
ZF75			+
ZF76			++
ZF77			(+)
ZF78	+		+++
ZF79			+++
ZF80			++
ZF81			++
ZF82			+
ZF83			+++
ZF84			+++
ZF85			+
ZF86	+		(+)
ZF87			+
ZF88			+
ZF89	+	(+)	++
ZF90	+	(+)	++
ZF91			++
ZF92	+		+
ZF93	+	(+)	+
ZF94			(+)
ZF95			+
ZF96		+	++
ZF97	(+)		++
ZF98			+
ZF99			+
ZF100			(+)
ZF101			+
ZF102	+		++
ZF103			++

Figure 6.3 ZF hybridisation pattern for "group 1" and "group 2" cosmids.

(a) Southern blot of HindIII digests of 17 "group 1" cosmids hybridised with the ZF oligonucleotide. Numbering is as for figure 6.1: 1 = ZF1, 3 = ZF19, 4 = ZF21, 5 = ZF25, 7 = ZF29, 8 = ZF30, 9 = ZF39, 10 = ZF62, 11 = ZF63, 12 = ZF78, 13 = ZF86, 14 = ZF89, 15 = ZF90, 16 = ZF92, 17 = ZF93, 18 = ZF97 and 19 = ZF102.

(b) Southern blot of HindIII digests of 23 of the 24 "group 2" cosmids hybridised with the ZF oligonucleotide. Numbering is as for figure 6.2: 1 = ZF2, 2 = ZF12, 4 = ZF27, 5 = ZF28, 6 = ZF32, 7 = ZF33, 8 = ZF34, 9 = ZF35, 10 = ZF36, 11 = ZF37, 12 = ZF38, 13 = ZF44, 14 = ZF46, 15 = ZF47, 16 = ZF48, 17 ZF52, 18 = ZF53, 19 = ZF55, 20 = ZF56, 21 = ZF59, 22 = ZF60, 23 = ZF61, 24 = ZF73 and 28 = ZF96.

(a)


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

(b)

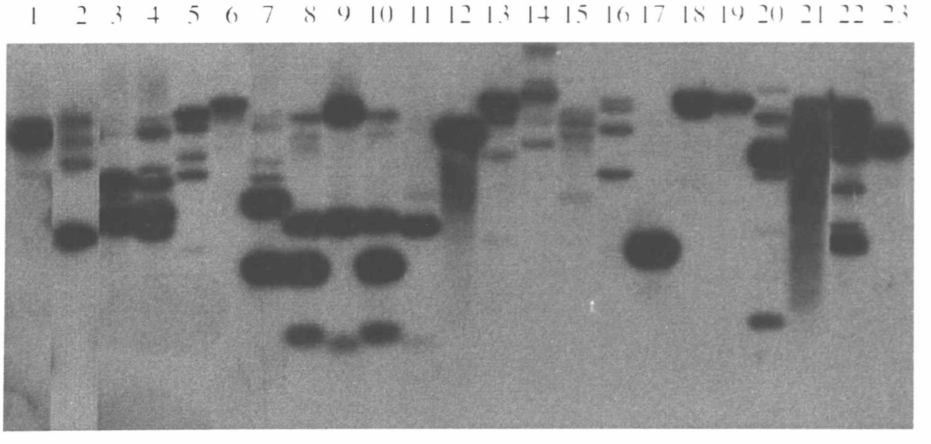

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

Figure 6.3 (a) shows the pattern of hybridisation obtained when a southern blot of the 17 "group 1" clones digested with HindIII was hybridised to the ZF oligonucleotide. Clones ZF1, ZF61 and ZF62 and clones ZF89 and ZF90 share fragments in common that hybridise with the oligonucleotide but the majority of these clones have a unique pattern of hybridisation and do not obviously overlap. This group of cosmid clones represents nearly 17% of the ZF cosmid clones obtained from the EJNAC4.1 and 4.1.5 hybrid cell lines.

Fig 6.3 (b) shows the pattern of hybridisation obtained when a southern blot of 23 of the 24 "group 2" clones (ZF53 is not shown) digested with HindIII was hybridised to the ZF oligonucleotide. Again, the majority of the clones do not overlap and there is variation in the size of fragments hybridising with the ZF oligonucleotide and in the intensity of signal. This group of cosmid clones represents nearly 24% of the ZF cosmid clones obtained from the EJNAC4.1 and 4.1.5 hybrid cell lines.

60 of the 101 ZF cosmids do not belong to "group 1" or "group 2". Of these 9 hybridise strongly to the oligonucleotide, 15 hybridise with a medium strength signal, 26 hybridise weakly and 10 hybridise very weakly.

6.6 Organisation of ZF sequences within 6 cosmids.

The data generated from the screening of cosmid clones from the hybrid cell lines indicated that the ZF sequences were present at a higher than expected frequency in the hybrid cell lines and might be clustered within the regions of human DNA present. If this was the case, and if related sequences were clustered, it was possible that some of the cosmid clones in the groups will share common sequences and fall into contigs. It was also possible that some of these cosmid clones contained more than one putative ZF gene.

To partially address this, the organisation of the ZF sequences within 6 cosmids from "group 1" and "group 2" was determined. The clones were terminase mapped with HindIII and EcoRI and the positions of the ZF sequence and human repeat sequence determined from southern blot hybridisation data. These cosmids were ZF1, ZF12,

ZF19, ZF25, ZF29 and ZF39 as shown in figure 6.4. In three of these cosmids, ZF19, ZF25 and ZF39, the ZF finger sequence mapped to one discrete region of the cosmid of 0.9kb, 1.2kb and 1.75kb respectively making it unlikely that more than one putative ZF gene was present within each of these clones. In the cases of ZF1, ZF12 and ZF29 there were two regions containing ZF sequence separated by a minimum of 30kb, 24kb and 6kb that also cross-hybridised with probes from "group 1" and "group 2". It was possible that these cosmids contained two ZF sequences of one gene separated by a large non-ZF region, or in two exons, or that two related genes were contained in one cosmid.

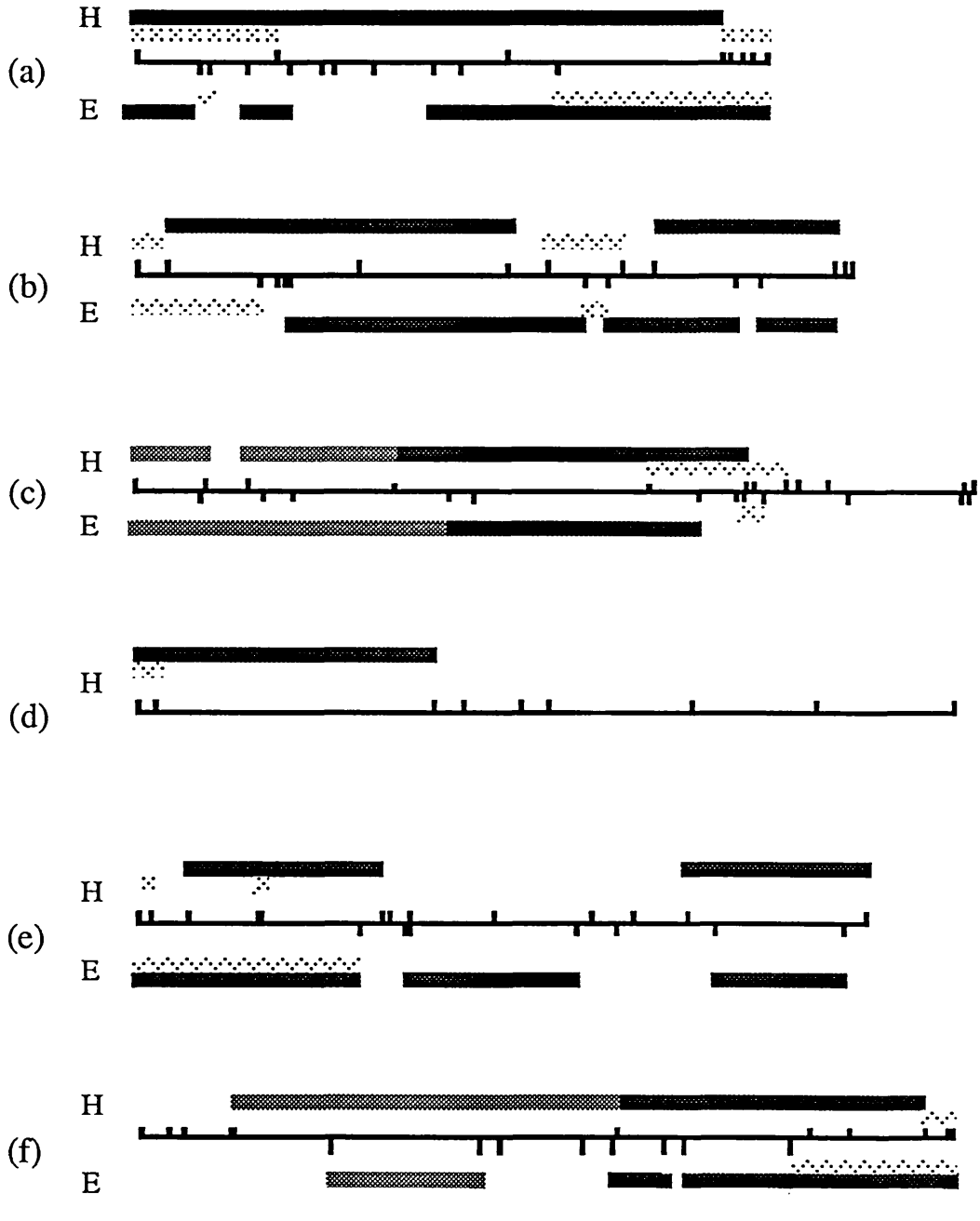
ZF25, ZF29 and ZF39 have been *in situ* mapped to the same region on chromosome 19. The terminase mapping data indicates that the cosmids do not share any common HindIII fragments. As the mapping project proceeds data generated by the fingerprinting process will begin to link clones. This project is still at an early stage and these three clones have not yet been linked.

6.7 Summary.

Cross-hybridisation studies have defined two groups of cosmids accounting for nearly half the human ZF sequences isolated from the hybrid cell lines. The overall increased frequency of the ZF sequence in the hybrid cell lines indicates that zinc finger sequences are concentrated in the regions of human DNA in EJNAC4.1 and 4.1.5. This is preliminary evidence that related sequences are clustered.

Figure 6.4 Restriction maps of ZF1, ZF12, ZF19, ZF25, ZF29 and ZF39.

Restriction maps of 6 cosmid clones are shown as determined from lambda terminase mapping with HindIII (H) and EcoRI (E), sites are shown respectively, above and below the line representing the cosmid clone. (a) = ZF1, (b) = ZF12, (c) = ZF19, (d) = ZF25, (e) = ZF29 and (f) = ZF39. Black boxes indicate HindIII or EcoRI fragments that contain human repeat sequences determined by hybridisation studies with a genomic human DNA probe and spotty boxes indicate HindIII or EcoRI fragments that hybridise with the ZF oligonucleotide. ZF19 and ZF39 contain fragments that hybridise with a genomic mouse DNA probe represented by the grey boxes.


(a)
H
E
(b)
H
E
(c)
H
E
(d)
H
(e)
H
E
(f)
H
E
ZF oligonucleotide hybridising fragment
Human hybridising fragment
Mouse hybridising fragment
2Kb

Chapter 7.

"Group 1" ZF sequences are expressed.

7.1 Introduction.

To determine whether the ZF sequences isolated from the EJNAC4.1 and EJNAC4.1.5 cosmid libraries were expressed, a cDNA library made from foetal kidney mRNA was analysed.

The body of evidence generated on the ZF gene family suggests that these genes are likely to be expressed at low levels, in specific tissues and/or at specific time points. There is a high degree of complexity of clones that hybridise to the ZF oligonucleotide in cDNA libraries from Xenopus, mouse and human tissues (Koster *et al.*, 1988, Schuh *et al.*, 1986, Chowdhury *et al.*, 1987, Chavrier *et al.*, 1988 and Bellefroid *et al.*, 1989). Data indicate that there are probably more than 300 genes containing the zinc finger motif in the human genome. Analysis of 250 cDNAs hybridising to the ZF oligonucleotide from a human foetal kidney cDNA library showed a wide variety in the size of the cDNA fragments and in the intensity of hybridisation to the oligonucleotide (Cardinal, pers. comm.). It was probable that a large number of different ZF genes were represented in this cDNA library, each at a low level. The kidney is a complex organ and a cDNA library made from embryonic tissue would be predicted to contain a large number of genes, including transcription factors, involved in its development.

Two groups closely related ZF sequences were identified by cross-hybridisation studies. By directly screening a complex cDNA library, plated at high density, with a probe from one of these groups at low stringency, cDNAs might be detected that were members of the groups expressed in that tissue. Instead of using one probe to isolate one gene, one probe would be used to isolate several related genes belonging to a specific group. There was an increased probability of detection of cDNAs homologous to the "group 1" or "group 2" ZF sequences isolated from the hybrid cell lines as genomic sequence for several members of the groups had been isolated.

7.2 Screening a human foetal kidney cDNA library.

7.2.1 Identification of cross-hybridising sequences.

The first cDNA library screened was made from mRNA isolated from human foetal kidney tissue. 0.3% of the cDNA clones in this library hybridised with the ZF oligonucleotide (pers comm. Little). 3.4 x 10^7 cDNA clones from this cDNA library were screened with ZF29p1, a 173bp "group 1" fragment which contained two copies of the ZF motif and which cross-hybridised with all the members of this group at 60oC. At low stringency, 784 clones hybridised with this probe. An example is shown in figure 7.1 (c). This represented 1 in 43000 of the clones in this library, equivalent to 1 in 130 of the clones that also hybridised to the ZF oligonucleotide. The "group 1" sequences in this library were not abundant. As a control, a DNA blot with the first 38 ZF cosmid clones on was washed simultaneously with these filters. The stringency of washing was then raised to 65oC. At this temperature no positives were identified.

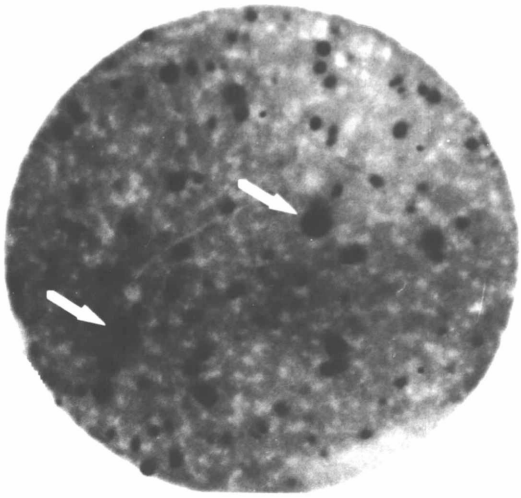
7.2.2 Analysis of 5 "group 1" cDNAs.

To determine the complexity of the "group 1" cDNA clones detected in this library, 5 of the most strongly hybridising phage were purified. The inserts were subcloned into bluescript and partial sequence was obtained from either end of the clones. Figure 7.2 shows the predicted amino acid sequence for these clones. Neither a polyA tail nor translation start signal was identified in any of the clones. There were repeating units of the ZF motif in all the sequence obtained and it was probable that, as these clones all hybridised strongly to the ZF oligonucleotide, full sequencing would reveal that they entirely comprised of repeating units of the motif. None of the clones shared regions of homology and it was unlikely that they were cDNAs derived from the same gene. Table 7.1 summarises the information obtained on these clones.

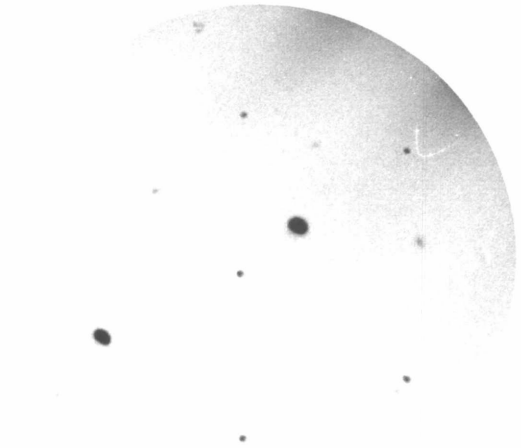
Figure 7.1 Screening a human foetal kidney cDNA library with "group 1" probes.

One round filter containing DNA from 4.25 x 10^6 plaques hybridised with a "group 1" probe, ZF1p2, and washed at 60^oC (a) and then at 65^oC (b), and hybridised with a second "group 1" probe, ZF29p1, and washed at 60^oC (c). The positions of two cDNA clones hybridising with the ZF1p2 probe at 65^oC are indicated by the white arrows.

(a)



(b)



(c)

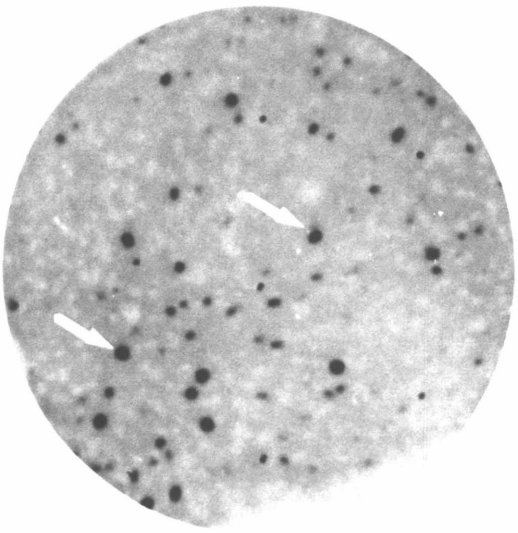


Figure 7.2 Predicted amino acid sequence for "group 1" cDNAs.

Amino acid sequence (single letter code) deduced from partial sequencing of 5 cDNAs isolated on the basis of cross-hybridisation with a "group 1" probe, ZF29p1, at 60^{o}C. (1) = 5' end of ZF29c2, (2) = 3' end of ZF29c2, (3) = 5' end of ZF29c3, (4) = 5' end of ZF29c4, (5) = 3' end of ZF29c4, (6) = 3' end of ZF29c5a and (7) = complete sequence of ZF29c5b. The amino acid sequence of the genomic fragment used as a probe is shown beneath (8).

(1) 5' H T G E K P Y K

C K E C G K A F K R S S N L T E H R I I H T G E K P Y K

C E E C G K A F N V L T P Y N T

(2) I H T G E N A Y K

C K E C G K A F N Q S S T L T R H K I I H A E G E K P Y I

C E H C G R A F N Q S S N L 3'

(3) 5' Q S L V N A L Q P L G K H K R I H T G E K P Y K

C E E C G K G Y K D S L S S P L

(4) 5' I H T G E K P Y K

C K E C G K A F K R S S N L T E H R I I H T G E K P Y K

(5) H L I H T K I I H T G E N A Y K

C K E C G K A F N Q S S T L T R H K I I H A G E K P Y I

C E H C G R A F N Q S S N L T K H K R I 3'

(6) F R L S S H L T T H K I I H T G E K P Y K

C E E C G K A F N K S S H L T R H K S I H T G E K P Y Q

C E K C G K A S N Q S S N L T E H K N I H T E E K P Y K

C E E C G K A F N Q F S N L T T H K R I 3'

(7) 5' I H I G E K S Y I

C E E C G K A C N Q F T N L T T H K I I Y T R D K L Y K

R E E C S K A F N L S S H I T T H T I I H T G E N P Y K

R E E C D K A F N Q S S T L T T H K I I H T R E K L N E

Y K E C G K A F N Q S S H L T R H K I I H T G E K P Y K

C E E C G K A F R Q S S H L T T H K I I H T G E K P Y K

C E E C G K A F N K S S H F T R H K S I H T G E K P Y Q

C E K C G K A S N Q S S N L T E H K N I H T E E K P Y K

C E E C G K A F N Q F S N L T R H K R I 3'

(8) A F Y R F S Y L I K H K T S H T G E K F Y K

C E E C G K G F N W S S A L T K H K R I H T G E K P Y K

C E E C G K A

Table 7.1 Comparison of sequence of 5 "group 1" cDNAs with a genomic "group 1" sequence.

The amount of sequence for each cDNA is shown, most of the clones were partially sequenced from either end and the amount of sequence is given separately. The degree of homology is given as a percentage of shared nucleotides and the highest degree of homology and the longest region of homology are shown separately.

Name	Sequence (bp)	Motifs	Homology to ZF29p1 (Longest)	(Highest)
ZF29c2	150+140 of 600	4.5	85% (89bp)	85% (89bp)
ZF29c3	130 of 150	1	91% (78bp)	97% (62bp)
ZF29c4	40+190 of 550	4	87% (112bp)	96% (49bp)
ZF29c5a	305 of 650	3	86% (138bp)	88% (74bp)
ZF29c5b	670 of 670	8	86% (136bp)	88% (91bp)

It was possible that cDNAs isolated on the basis of cross-hybridisation with the ZF29p1 probe might be cognate sequences for "group 1" cosmid clones already isolated from the hybrid cell lines. To determine if this was the case the clones were radiolabelled and hybridised to DNA blots of the ZF cosmids. At low stringency they give a "group 1" pattern of cross-hybridisation. At 65^{o}C one of the clones, ZF29c5b, cross-hybridised with a 1kb HindIII fragment from cosmid ZF30 (figure 7.3 (a) and (b)). The only ZF motif sequence obtained from this cosmid was from a 90bp HindIII fragment which hybridised only at 60^{o}C to the cDNA. When the two sequences were compared there were 6 differences in a stretch of 13bp (figure 7.3 (c)). It was possible that the fragment from the cosmid represented a different part of the gene not present in the cDNA or it might be part of a second, highly related gene in the same cosmid. 250bp of sequence was obtained from either end of the 1kb fragment ZF30 HindIII fragment but this did not contain any copies of the ZF motif or homology to ZF29c5b. More rigourous sequencing would be required to determine the relationship between ZF30 and the cDNA.

7.2.3 Analysis of the ZF1 cDNA.

7.2.3.1 Isolation of ZF1 cDNA clones.

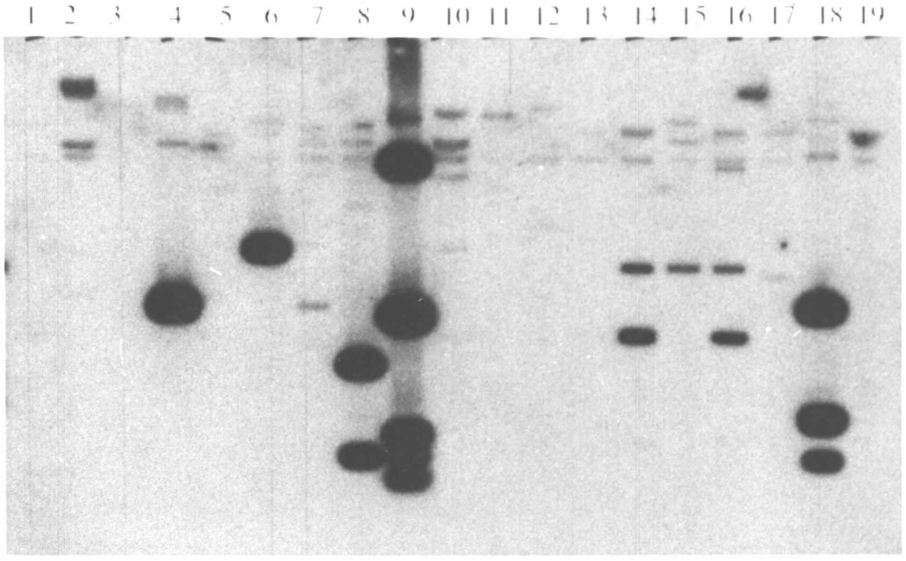
Having determined that several "group 1" genes were expressed in embryonic foetal kidney, the cDNA library filters were successively hybridised with genomic fragments from other members of this group. These fragments all detected the same pattern of positives at the lower stringency. One of these fragments also detected 18 positives at 65^{o}C. This was a 259bp HindIII fragment from cosmid ZF1. Figure 7.1 (a) shows the hybridisation of one round filter to the ZF1p2 probe washed at 60^{o}C. Figure 7.1 (b) shows the same filter washed at the 65^{o}C. Two positives were detected with ZF1p2 on this filter.

Figure 7.3 Hybridisation of ZF29c5b to Southern A.

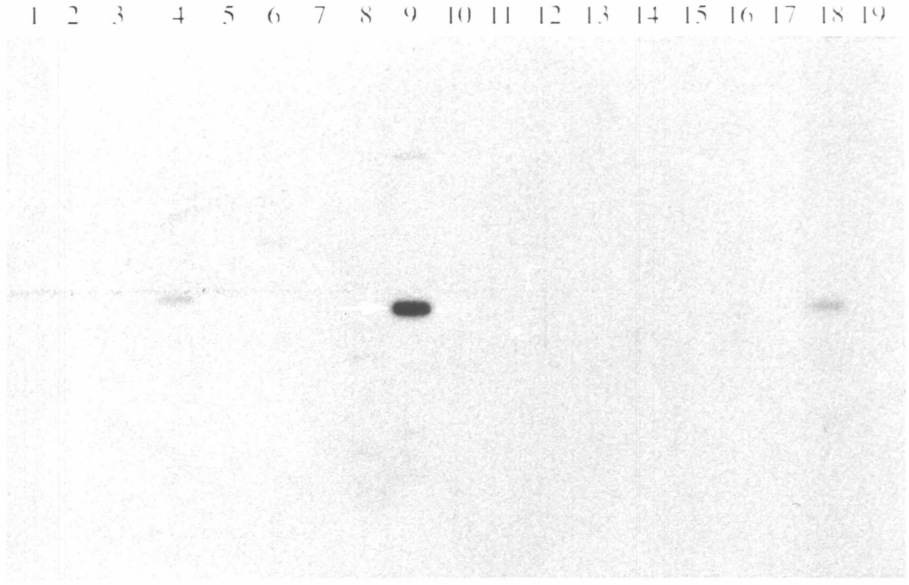
Clones ZF22-40 from southern blot A (see figure 5.3) hybridised with a ZF29c5b probe and washed at 60^{o}C (a) and then at 65^{o}C (b). Arrow points to 1kb fragment from ZF30. ZF22 (1), ZF23 (2), ZF24 (3), ZF25 (4), ZF26 (5), ZF27 (6), ZF28 (7), ZF29 (8), ZF30 (9), ZF31 (10), ZF32 (11), ZF33 (12), ZF34 (13), ZF35 (14), ZF36 (15), ZF37 (16), ZF38 (17), ZF39 (18) and ZF40 (19).

Figure (c) is a comparison of nucleotide sequence of the region of cDNA ZF29c5b with homology to a 87bp HindIII fragment from genomic clone ZF30. Differences in sequence are underlined

(a)


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

(b)


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

(c)

AAGCTTTTAACCTATCTTCACACCTTACTACACATAAGATAATTCATACTGGAGAGAAACCCTACAAA
AAGCTTTCAACCAGTCCTCACACCTTACCAGACATAAGATAATTCATACTGGAGAGAAACCCTACAAA
A F N Q S S H L T R H K I I H T G E K P Y K

TGTGAAGAATGTGGCAAAGCTT
TGTGAAGAATGTGGCAAAGCTT
C E E C G K A

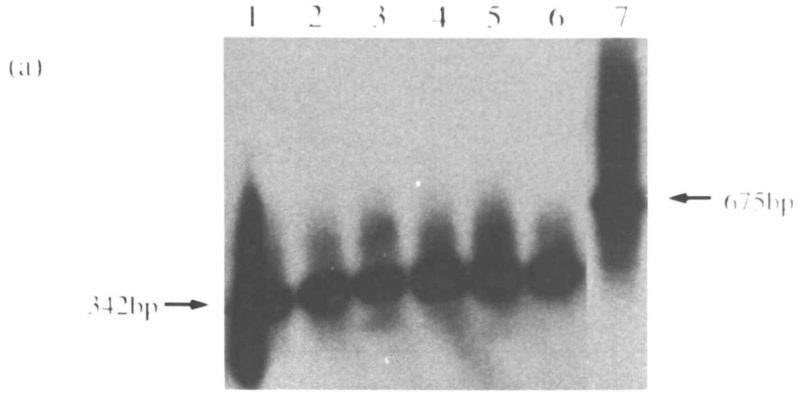
6 cDNA clones hybridising with ZF1p2 at 65°C, ZF1c1, ZF1c3, ZF1c4, ZF1c5, ZF1c9 and ZF1c14, were purified and DNA was prepared from the phage and digested with EcoRI to excise the insert DNA. DNA was also prepared by PCR amplification using lambda gt10 primers. The products of both the EcoRI digestion and the amplifications were approximately 350bp in size indicating there was no EcoRI site within the insert. All 6 clones were confirmed as hybridising with the genomic probe, ZF1p2, at 65°C: Figure 7.4 (a) shows a southern blot of the PCR products of the 6 positive clones and of a seventh clone, ZF29c5b (a cDNA isolated in a low stringency screen with the ZF29p1 probe), hybridised with a ZF1p2 probe and washed at 60°C. All the clones hybridised to ZF1p2 at 60°C indicating that they were all "group 1" sequences. The blot was then washed at 65°C (fig 7.4 (b)). At 65°C ZF29c5 no-longer hybridised with ZF1p2 indicating that ZF1p2 was more homologous to the ZF1 cDNAs than to a randomly isolated "group 1" cDNA. In a reverse hybridisation, the 342bp cDNA from ZF1c1 hybridised exclusively to the 259bp genomic HindIII fragment used to isolate it at 65°C.

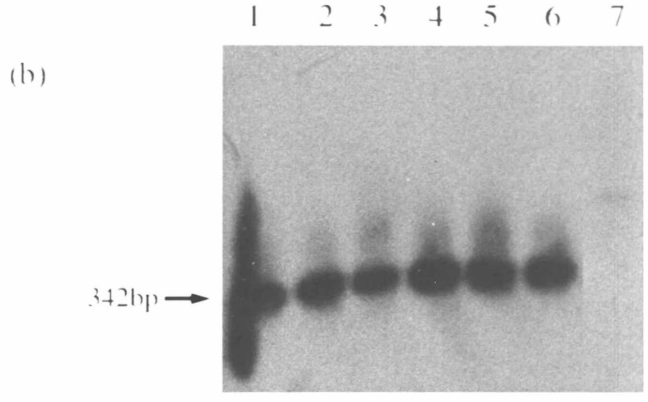
7.2.3.2 Sequence analysis of ZF1c1.

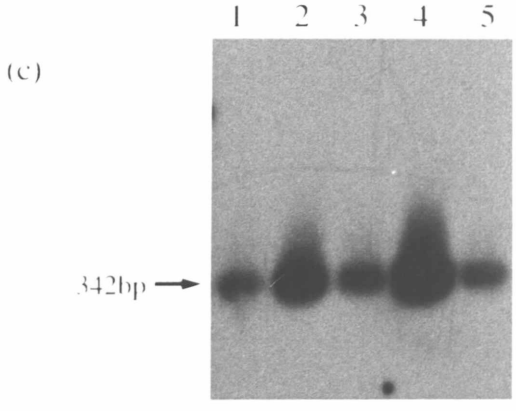
Sequence analysis of the 5' end of four of the cDNA clones showed that they started at the same position indicating these were amplified copies of the same cloning event. One of the cDNAs, ZF1c1, was completely sequenced (fig (a)). The clone was 342bp in size and would code for 4 copies of the ZF motif. There was no 5' start signal or polyA tail present in this sequence indicating that the full length cDNA was not cloned. In order to find a more full length clone, 5 primary phage picks which had not previously been analysed, were directly amplified using lambda gt10 PCR primers. The products of this amplification were analysed on a southern blot. A band of about 350bp was seen with a ZF1p2 probe (figure 7.4 (c)). 11 of the 18 clones isolated from this library were the same size and probably resulted from the amplification of a single cloning event.

Figure 7.4 Southern blot analysis of PCR products of ZF1 cDNA clones.

Southern blot of PCR products of 7 purified phage cDNAs ZF1c1 (1), ZF1c3 (2), ZF1c4 (3), ZF1c5 (4), ZF1c9 (5), ZF1c12 (6) and ZF29c5b (7) hybridised with a ZF1p2 radiolabelled probe and washed first at $60^{o}C$ (a) and then at $65^{o}C$ (b). Figure (c) is a southern blot of PCR products of 5 primary phage cDNAs isolations: ZF1c13 (1), ZF1c14 (2), ZF1c15 (3), ZF1c16 (4) and ZF1c17 (5), hybridised with ZF1p2 radiolabelled probe and washed at $65^{o}C$.


1 2 3 4 5 6 7
(a)
675bp
342bp


1 2 3 4 5 6 7
(b)
342bp


1 2 3 4 5
(c)
342bp

The absence of internal EcoRI sites and the occurrence of the EcoRI cloning sites within finger coding regions at either end of this clone might indicate that this was an internal fragment of a larger cDNA that was digested by EcoRI during the construction of the cDNA library. The identification and analysis of a cDNA from a second library would be required to determine this. It is unlikely that a transcript would be detected on a northern blot as this transcript represents 1 in 1.9 x 10^6 of the cDNA molecules in the cDNA library.

7.2.3.3 ZF1c1 has 96.5% homology to genomic probe.

When the sequence of the of the genomic fragment was compared to the equivalent region in the cDNA there were 9 mismatched base pairs (figure 7.5). Either this cDNA was not the cognate sequence for the genomic DNA used to isolate it but was a closely related sequence not distinguishable by hybridisation or the loci was highly polymorphic within the human population.

7.3 "Group 1" genes are expressed in several tissues.

7.3.1 Screening a human lymphoblast and a human fibroblast cDNA library.

The four "group 1" probes were used to screen two other cDNA libraries to determine whether members of this group were expressed in other tissues. A human lymphoblast cDNA library (<0.32% positives with ZF oligonucleotide) and a human fibroblast cDNA library (% positives with ZF oligonucleotide not determined) were screened. The results of the "group 1" screens are summarized in table 7.2.

cDNAs cross-hybridising with genomic "group 1" probes were detected in both these libraries. The same pattern of hybridisation was detected by all four clones at the lower stringency of washing, as seen with the human foetal kidney cDNA library. An example of this is shown in figure 7.6 (a) of the human fibroblast library hybridised with a ZF1p2 probe washed at 60°C and (b) with a ZF39p3 probe washed at 60°C. When washed at 65°C, 9 positives were detected by the ZF39p3 probe in the fibroblast library 7.6 (c).

Figure 7.5 Sequence of ZF1c1.

Nucleotide sequence and predicted amino acid sequence of the cDNA ZF1c1 and comparison of nucleotide sequence of a region of ZF1c1 with 95.7% homology to the genomic fragment used to isolate it, ZF1p2. The genomic sequence is shown above the cDNA sequence and the differences in sequence are highlighted underlined. * indicates a stop codon.

```
                                                    GAATTCATACTTGAGAGAAACCCTATAAA
                                                      I  H  T  *  E  K  P  Y  K

                    AAGCTTTTTACTATTCCTCAGGCCTTACTCAACATAACATAGTTCATACTGGAGACAAACCCTACAAA
TGTGAAGAATGTGGCAAAGCTTTTTACTGTTTCTCAGGCCTTACTCAACATAACATAGTTCATACTGGAGACAACCCCTACAAA
 C  E  E  C  G  K  A  F  Y  C  F  S  G  L  T  Q  H  N  I  V  H  T  G  D  N  P  Y  K

TGTAAAGATTGTGGCAAAATTTTTAAGTGGTCTTCGAACCTTACTATACATCAGAGAATTCATAGTGGAGAGAAACCCTACAAA
TGTAAAGATTGTGGCAAAATTTTTAAGTGGTCTTCAGACCTTACTATACATCAGATAATTCATAGTGGAGAGAAACCCTACAAA
 C  K  D  C  G  K  I  F  K  W  S  S  D  L  T  I  H  Q  I  I  H  S  G  E  K  P  Y  K

TGTGAAGAATGTGGCAAAGCCTTTAAGCAATCTTCAAAACTGAATGAACATATGAGAGCTCATACTGGAGAGAAATTCTACAAA
TGTGAAGAATGTGGCAAAGCCTTTAAACAATCCTCAAAACTGAATGAACATATGAGAGCTCATACTGGAGAGAAATCCTACAAA
 C  E  E  C  G  K  A  F  K  Q  S  S  K  L  N  E  H  M  R  A  H  T  G  E  K  S  Y  K

TGTGAAGAATGTGGCAAAGCTT
TGCAAAGAATGTGGCAAAGCTTTTAAACAACCTTCAGGCCTTACTCTACATAAGAGAATTC
 C  K  E  C  G  K  A  F  K  Q  P  S  G  L  T  L  H  K  R  I
```

Table 7.2 Screening cDNA libraries with group1 fragments.

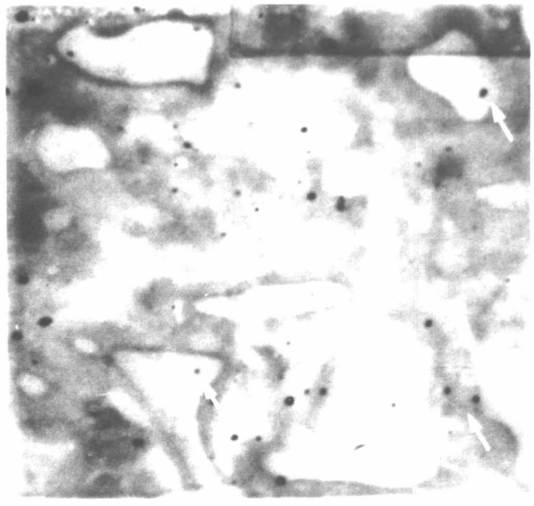
The number of clones from each cDNA library hybridising to four "group 1" probes at low ($60^{o}C$) and high ($65^{o}C$) stringency are given.

Tissue	Clones screened	All 60°C	ZF01 65°C	ZF29 65°C	ZF30 65°C	ZF39 65°C
Foetal kidney	3.4×10^7	784	18	0	0	0
Lymphoblast	3.8×10^6	1260	0	0	0	0
Fibroblast	7.0×10^6	120	-	0	0	9

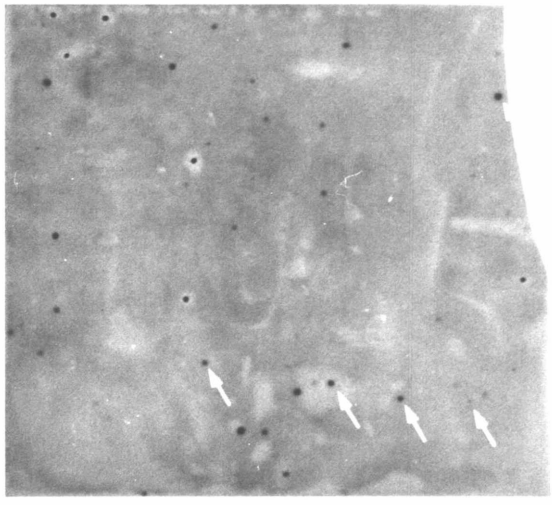
Figure 7.6 Screening a human fibroblast cDNA library with "group 1" probes.

One square filter containing DNA from 4.2 x 10^6 plaques hybridised with a "group 1" probe, ZF1p2, and washed at 60°C (a) and the same filter hybridised with a second "group 1" probe, ZF38p3, and washed at 60°C (b), and then at 65°C (c). The five cDNA clones hybridising with ZF39p3 at 65°C are indicated by the white arrows.

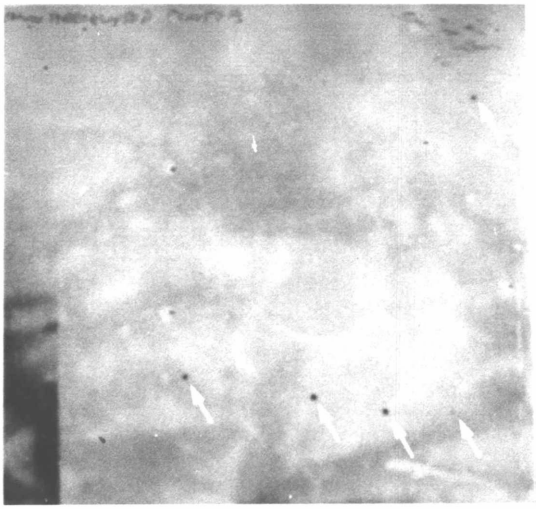
(a)



(b)



(c)



7.3.2 Analysis of the ZF39 cDNA.

7.3.2.1 Isolation of ZF39 cDNA clones.

The 9 cDNA clones hybridising with the ZF39p3 probe at 65°C, ZF39c1 to ZF39c9, contained inserts of 1.2kb and did not have internal EcoRI sites. All were confirmed as hybridising to the ZF39p3 probe at 65°C on a southern blot. In a reverse hybridisation experiment, a southern blot of ZF29 and ZF39 DNA digested with HindIII and EcoRI was hybridised with a ZF39c1 probe and washed at 60°C (figure 7.7 (b)) and then at 65°C (figure 7.7 (c)). The cDNA hybridised with three HindIII fragments from ZF39 of 1.2, 0.35, and 0.15kb and one EcoRI fragment of 9.55kb at 65°C. These are fragments that also hybridise to the ZF oligonucleotide (figure 7.7 (a)). The hybridisation to a 0.15kb fragment was not seen on this exposure but was seen when the blot was exposed for longer.

7.3.2.2 Sequence analysis of ZF39c1.

The cDNA clones ZF39c1, ZF39c2, ZF39c3, ZF39c4 and ZF39c7 all had identical 5' start points as judged by sequence analysis. This analysis indicated that only one copy of this sequence had been isolated and that the 9 cDNA clones were amplifications of a single cloning event.

The cDNA was fully sequenced on both strands by generating Bal31 deletions of ZF39c1. Figure 7.8(a) shows the regions of sequence obtained from these overlapping subclones. The sequence of 1214bp of ZF39c1 contained 14 repeating units. The predicted amino acid sequence is shown in figure 7.8(b). Each of these units resembled a ZF motif. However, 11 of the 14 repeats were disrupted by either a stop codon, a frame shift or a change in one of the predicted Zn ion co-ordinating residues. Three units appeared to be normal fingers, these being "fingers" 3, 6, and 9. Of the other repeating units:

1) Ten lacked one or more of the cysteine or histidine residues at the expected positions, these being repeats 1, 2, 4, 7, 8, 10, 11, 12, 13, and 14.

Figure 7.7 Hybridisation pattern of ZF39.

Southern blot of DNA from two "group 1" cosmid clones, ZF29 (1) and ZF39 (2), digested with HindIII (H) or EcoRI (E) hybridised with (a) the ZF oligonucleotide, and with a ZF39c1 cDNA probe washed at 60°C (b) and at 65°C (c).

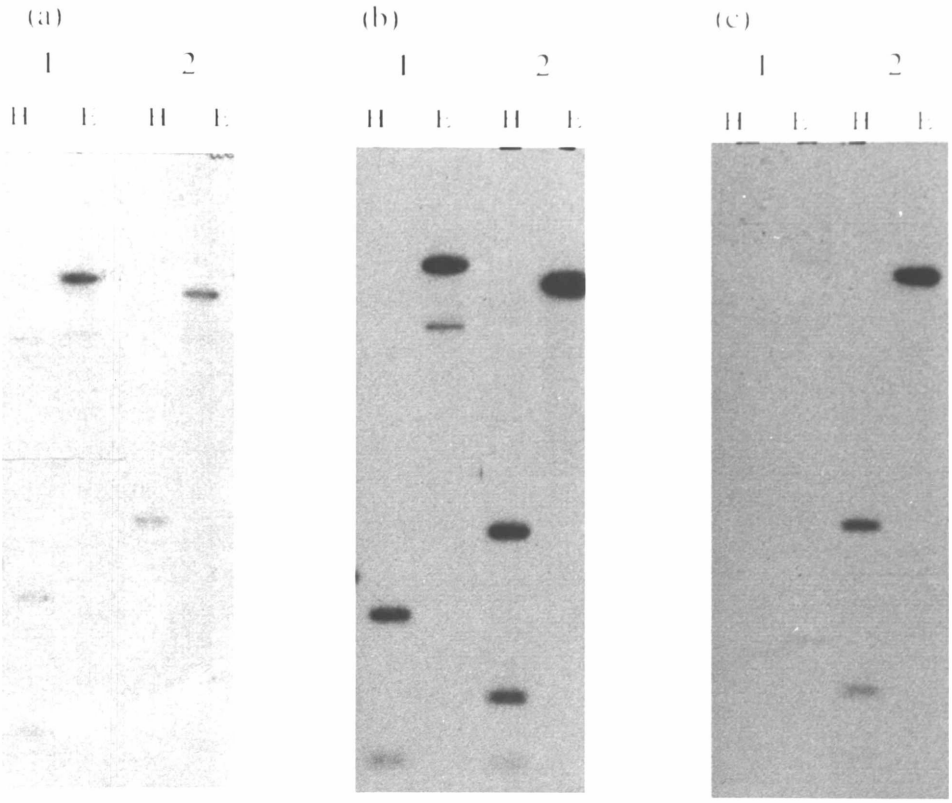
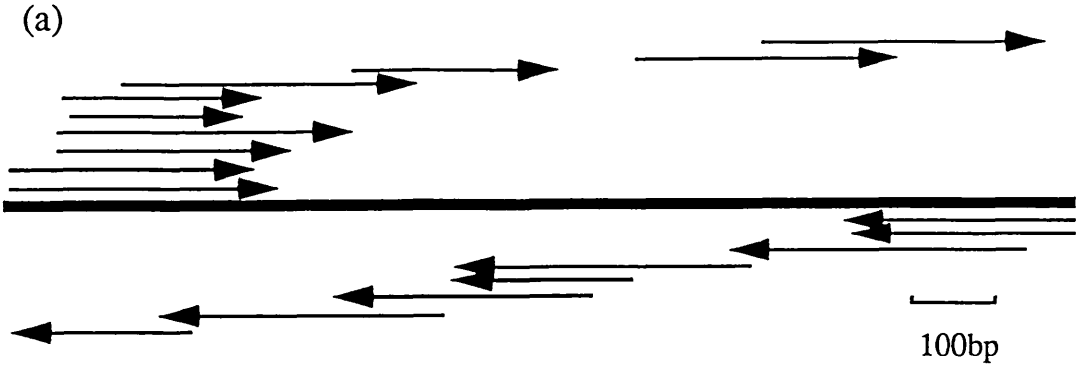

(a)
1
2
H
E
H
E
(b)
1
2
H
E
H
E
(c)
1
2
H
E
H
E

Figure 7.8 Sequence analysis of ZF39c1.

(a) Sequencing strategy for the ZF39c1 cDNA. The arrow lines above and below the main line indicate the extent of sequence information from clones generated by Bal31 digestion from the left (above the line) and the right (below the line). (b) Nucleotide sequence and predicted amino acid sequence (single letter code) of ZF39c1. The frame best fitting a ZF motif sequence is shown. Two frame shifts are present and amino acids from both reading frames in these regions are shown. * = stop codon. Sequence of 339bp HindIII fragment from cosmid ZF39 is shown above the cDNA sequence and differences in sequence are underlined. The polymorphic HindIII site is highlighted.

(a)



(b)

```
                                                    GAATTCATACCATAGAGAAATCCTACAAA
                                                       I  H  T  I  E  K  S  Y  K

TATGAAGAATGTGACAAAGCTTTTAACCACTTCTCAACCCTGCCTACACGTAAGATAATTCATACTGGAAGGAAACCCTACAAA
 Y  E  E  C  D  K  A  F  N  H  F  S  T  L  P  T  R  K  I  I  H  T  G  R  K  P  Y  K

                    AAGCTTTTTACTGATTCTTATACCTTACTAAACATAAAATAATTCATAAAGGAGATAAATTATACAAA
TATGAGGAATGTCTCAAAGCTTTTTACTGATTCTTATACCTTACTAAACATAAAATAATTCATAAAGGAGATAAATTATACAAA
 Y  E  E  C  L  K  A  F  Y  *  F  L  Y  L  T  K  H  K  I  I  H  K  G  D  K  L  Y  K

TGTGAAGAATGTGGCAAAGCCTTTAGTATATTCTCAACCCTTACTAAACATAAGATAATTCATACTGGAGAGAAAACCTACAAA
TGTGAAGAATGTGGCAAAGCTTTTAACAAATCCTCATCCATTAGTAAACATAAGATAATTCATACTGGAGAGAAAACCTACAAA
 C  E  E  C  G  K  A  F  N  K  S  S  S  I  S  K  H  K  I  I  H  T  G  E  K  T  Y  K

TGTGAGGAATGTGGCAAAGCCTTTAGCCTGTCCCTCCAATTTACTGCACATAAGATAATTTATACTGGAGAGAAGCCCTACAAA
TGTGAGGAATGTGGCAAAGCCTTTAGCCTGTCCCTCCAATTTACTGCACATAAGATAATTTATACTGGAGAGAAGCCCTACAAA
 C  E  E  C  G  K  A  F  S  L  S  L  Q  F  T  A  H  K  I  I  Y  T  G  E  K  P  Y  K

TGTGAAAAATGTGGCAAACCTTTTAACCAATCCTCAACCCTTACTACACATTAGATAATTCATGCTGGAGAGAAACCCTACAAA
TGTGAAAAATGTGGCAAACCTTTTAACCAATCCTCAACCCTTACTACACATTAGATAATTCATGCTGGAGAGAAACCCTACAAA
 C  E  K  C  G  K  P  F  N  Q  S  S  T  L  T  T  H  *  I  I  H  A  G  E  K  P  Y  K

TGTGAAAAATGTGGCAAAGCTT
TGTGAAAAATGTGGCAAAGCTTTTAACCAATTTTCAAACCTTACTAAACATAAGATAACTCATACTGGAGAAAAATCTTACAAA
 C  E  K  C  G  K  A  F  N  Q  F  S  N  L  T  K  H  K  I  T  H  T  G  E  K  S  Y  K

TGTGAAGAATGTGGCAAAGCTTTATCCAGTCCTCAACTCCTAGTAAACATAATTAATGATGGAGAGAAACCATACAAC
 C  E  E  C  G  K  A  L  S  S  P  Q  L  L  V  N  I  I  N  D  G  E  K  P  Y  N

TGTGAAGAATGTGGCAAAGCTTTTAACCAGTCCTCAAACTTTATTGAACAAAATAATTCATACAGGAGAGAAACCCTACAAA
 C  E  E  C  G  K  A  F  N  Q  S  S  N  F  I  E  Q  N  N  S  Y
                   L  L  T  S  P  Q  T  L  L  N  K  I  I  H  T  G  E  K  P  Y  K

TGTGAAGAATGTGACAAAGCCTTTAACCAGTCCTCAATTTTTACTAAACATAAGAAAATTCATACTGGAGAGAAACCCTATGAT
 C  E  E  C  D  K  A  F  N  Q  S  S  I  F  T  K  H  K  K  I  H  T  G  E  K  P  Y  D

TGTGAAAAATATGGCAAAGGCTTTAACTAGTCCTCAGTTCTTAACACACATACGATAATTCTTACTGCAGAGAAACTCTACAAA
 C  E  K  Y  G  K  G  F  N  *  S  S  V  L  N  T  H  T  I  I  L  T  A  E  K  L  Y  K

CCAGTAAGATGTGACAGTGCTTCTGACAACATCTCGAAACTTTTCTAATCATAAAAGAAATCATATTGGTGAGAAATCCTAGAAA
 P  V  R  C  D  S  A  S  D  N  I  S  K  L  F  *
          V  T  V  L  L  T  T  S  R  N  F  S  N  H  K  R  N  H  I  G  E  K  S  *  K

TGTGGAGAATGTAACAAAGTATTTAAATGGTTGTCACACTTGATTATAGGTAATATTCATATTGGAAAAATTTCCTACAAG
 C  G  E  C  N  K  V  F  K  W  L  S  H  L  I  I  G  N  I  H  I  G  K  I  S  Y  K

TAAGAACAATGTGGCAAAGTTTTTAACTAATACACCTTATTGCACAGAAAATCATTTATATTTGAGAAAAATTGTAGAAATATAGAC
 *  E  Q  C  G  K  V  F  N  *  Y  T  L  L  H  R  K  S  F  I  F  E  K  N  C  R  N  I  D

TGTGAAAAAGACGTCAATATCTGCTCACATCTTACTAAACACCAGAGAGTTCATGCTTAATAAAAGCATGATAAG
 C  E  K  D  V  N  I  C  S  H  L  T  K  H  Q  R  V  H  A  *  *  K  H  D  K

TGCAATTACTGCCAAAAGGAATTC
 C  N  Y  C  Q  K  E  F
```

2) Six of the repeats contained one or more stop codons (2, 5, 10, 11, 13 and 14).

3) Five of the repeats had aberrant spacing (7, 11, 12, 13, and 14).

4) Repeats 8 and 11 contained 2 frame shifts. The amino acid sequence around these frame shifts is shown in figure 7.8(b).

Additionally, no polyA tail and no polyA addition signal (ATAAA) was detected.

7.3.2.3 The cDNA, ZF39c1, has 99.4% homology to genomic sequence, ZF39p3.

A sequence comparison was made between the cDNA and the genomic DNA used to isolate it (figure 7.8 (c)). There are 7 differences in nucleotide sequence between the cDNA and the genomic DNA which lie within a stretch of 19 bases. These differences would correspond to 6 differences in amino acid sequence but would not alter the key amino acids of the ZF motif. Both the cDNA and the genomic sequence included two stop codons. This region was rigourously sequenced in both strands in both the genomic and the cDNA.

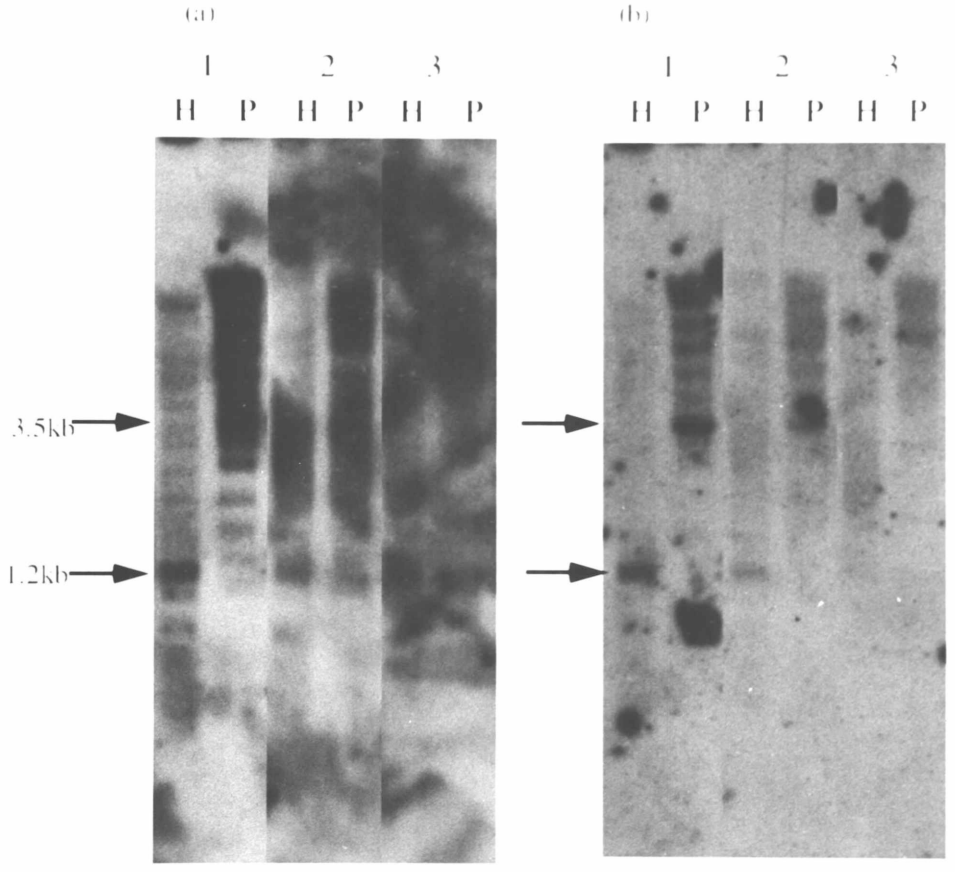
It is unlikely that a functional protein could be translated from this sequence as the first stop codon is 142bp from the 5' end of the fragment. Since no Met codon was present at the 5' end of the sequence, the actual 5' end probably lies further upstream but it is unlikely that the 1072bp of sequence 3' to the first stop signal represents the long untranslated 3' region of a much larger transcript. There is no indication of a polyadenylation signal or a poly(A) tail. It is more likely that this sequence is a pseudogene not making a functional gene product.

7.3.2.4 ZF39c1 is the cognate sequence for ZF39.

A ZF39c1 probe was hybridised to a genomic southern blot of DNA from human peripheral lymphocytes, the EJNAC4.1 cell line and mouse C57 black liver, digested with HindIII and PstI. Figure 7.9 (a) is this southern blot washed at 60^{o}C and (b) at 65^{o}C. Multiple bands hybridised to the ZF39c1 cDNA probe at 60^{o}C.

Figure 7.9 Genomic southern blot analysis of ZF39c1.

Southern blot of genomic DNAs prepared from human white blood cells (1), EJNAC4.1 tissue culture cells (2) and C57 black mouse liver (3) and digested with HindIII (H) and PstI (P), hybridised to a ZF39c1 probe and washed at 60^{o}C (a) and then at 65^{o}C (b).


(a)
(b)
1
2
3
H
P
H
P
H
P
3.5kb
1.2kb

Similar results have been obtained by several groups with ZF gene probes isolated from different organisms. A typical example is the results obtained by Chowdhury *et al.* (1987) where southern blotting experiments with *mkr1*, a mouse gene isolated in a low stringency screen with a fragment from the *Drosophila Krüppel* gene revealed a complex series of bands under low stringency conditions in human, calf, hamster, chicken and mouse genomic DNA. At 65^{o}C, the ZF39c1 probe hybridised with a 1.2kb HindIII fragment and a 3.5kb PstI fragment in both the human genomic and the hybrid cell line genomic DNA. The hybridisation to a 1.2kb HindIII fragment was expected as the cDNA hybridised to a HindIII fragment of this size in the ZF39 cosmid clone. The cDNA probe weakly detected a 7kb PstI fragment in the mouse DNA tract indicating that this sequence was also conserved in the mouse genome.

1 x 10^{7} clones from the ALL human genomic cosmid library were screened with a ZF39c1 probe. Two cosmids were isolated, ZF39B and ZF39C, which appeared to have three HindIII fragments in common with ZF39 of 4.4, 2.2 and 1.2kb (figure 7.10(a)). A southern blot of these cosmids was hybridised with the ZF39c1 probe. After washing at 65^{o}C two bands were seen with the three cosmids (figure 7.10(b)). The common 4.4 and 2.2kb HindIII fragments and the hybridisation of the 1.2 kb with the cDNA indicated that the same genomic locus had been identified from the ALL library as was present in the EJNAC4.1 cosmid. There appeared to be a polymorphism in the ZF sequence detected by digestion with HindIII as the second band hybridising to the cDNA probe in the ALL cosmids was 0.25kb but in the original EJNAC4.1 cosmid there was a 0.35kb fragment.

A series of PCR experiments were performed to verify the polymorphic HindIII site. Primers were designed based on sequence information on the cDNA. The primers were designed to prime from position 115 to position 504 to amplify a 389b fragment (figure 7.11 (a)). This fragment included three HindIII sites in the cDNA at positions 131, 215 and 467. Digestion of the PCR product from the cDNA with HindIII would produce fragments of 252, 84, 37 and 16bp.

Figure 7.10 DNA around ZF39c1 locus isolated from ALL cosmid library.

(a) Agarose gel electrophoresis of two cosmid clones from an ALL library, ZF39B (1) and ZF39C (2) and a cosmid clone, ZF39 (3), from the EJNAC4.1 library, digested with HindIII.

(b) Southern blot of this agarose gel hybridised with a ZF39c1 probe and washed at $65^{o}C$.

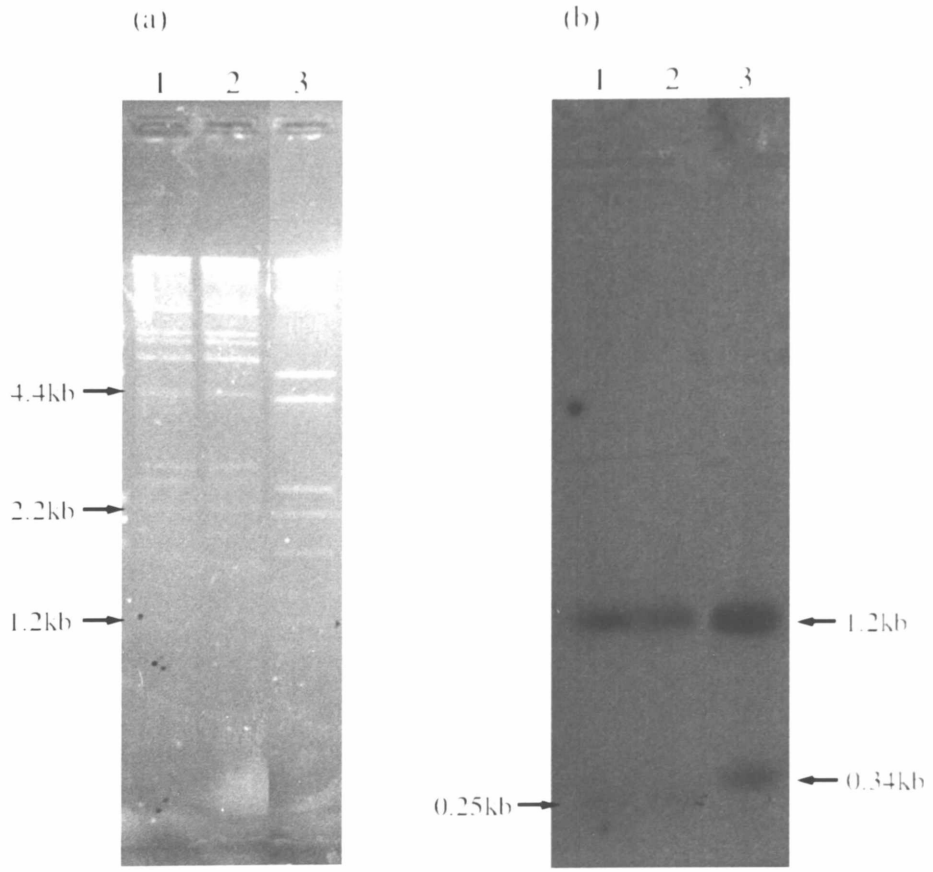
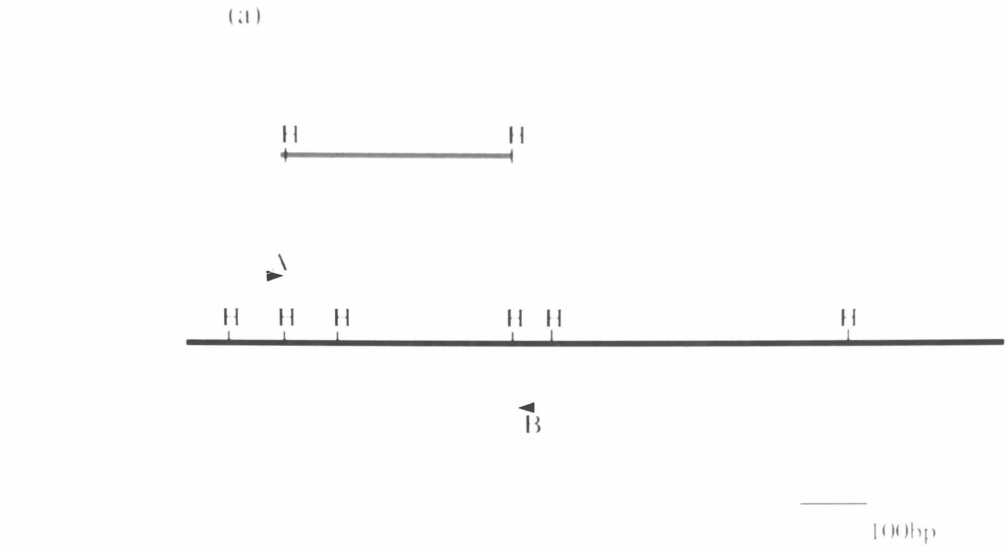
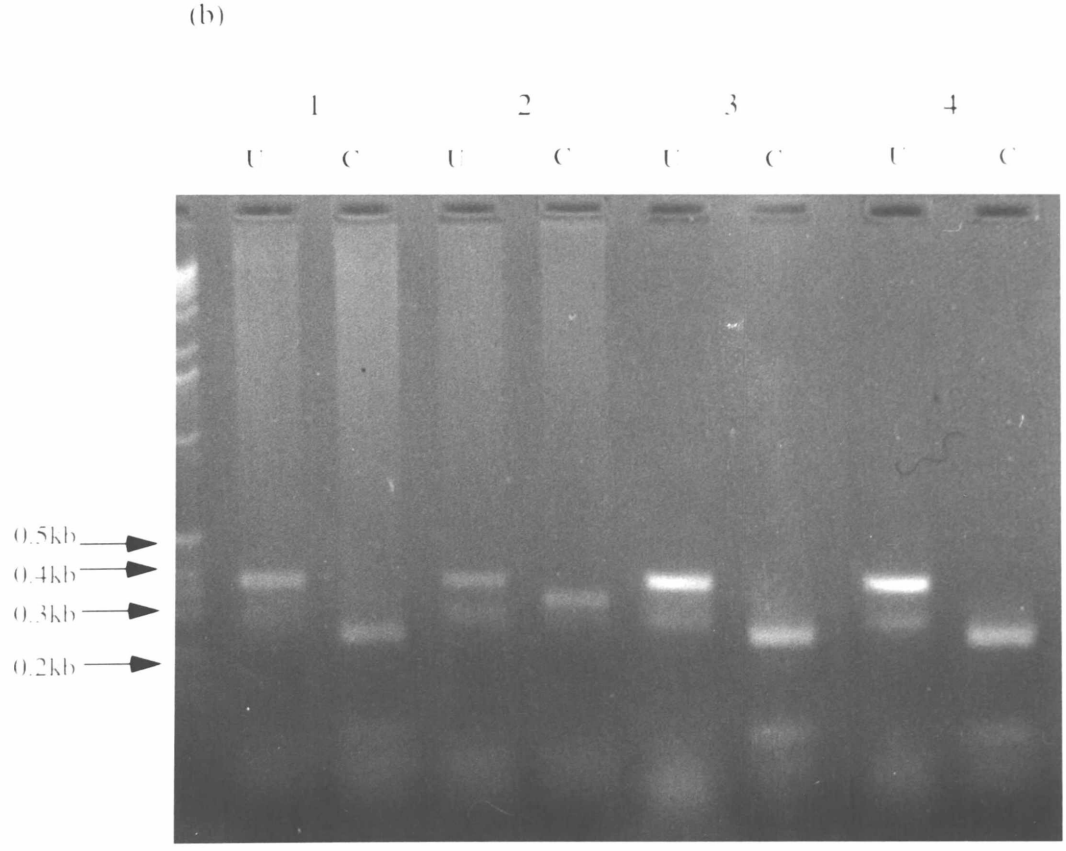

(a)
1
2
3
4.4kb
2.2kb
1.2kb
(b)
1
2
3
1.2kb
0.34kb
0.25kb

Figure 7.11 PCR of ZF39c1 locus.

(a) Position of primers (A and B) with respect to the cDNA, ZF39c1. The black line represents the length of the cDNA and the grey line represents the 329bp HindIII fragment from ZF39 used as a probe to isolate the cDNA (ZF39p3). HindIII sites are marked by H and the black arrow heads indicate the positions of the primers.

(b) PCR analysis of the ZF39c1 locus. The templates were DNA prepared from human white blood cells in lane 1 and from EJNAC4.1 tissue culture cells in lane 2, and DNAs from the ALL cosmid clone ZF39B and the cDNA ZF39c1 in lanes 3 and 4.


(a)
H
H
A
H
H
H
H
H
H
B
100bp


(b)
1
2
3
4
U
C
U
C
U
C
U
C
0.5kb
0.4kb
0.3kb
0.2kb

The sequence from the EJNAC4.1 cosmid lacks the HindIII site at position 215 and the PCR product amplified from this region digested with HindIII would produce fragments of 336, 37 and 16bp. The design of the primers had to take into account the presence of 14 repeats of the ZF motif so primers were made for non-overlapping regions of the greatest variation. DNA was amplified under the conditions described in chapter 2 with 100pmol of each primer for 35 cycles of 1 minute at $94^{o}C$, 1 minute at $52^{o}C$ and 1 minute at $72^{o}C$ followed by a 7 minute $72^{o}C$ extension reaction. $1/5^{th}$ of the reaction was digested with HindIII and run alongside the uncut reaction on a 2% NuSieve agarose gel (figure 7.12 (b)). The size of the uncut products (U) for the cDNA template, the ALL cosmid template, and the genomic DNA templates of human DNA and EJNAC4.1 DNA was 389bp. On digestion with HindIII (C lanes) the PCR product amplified from EJNAC4.1 DNA was 336bp whereas the products of the other amplifications were 252bp and 84bp identifying the polymorphic HindIII restriction site.

All the evidence indicated that the cDNA arises from the same locus as the EJNAC4.1 cosmid, ZF39, and that this locus is highly polymorphic.

7.3.2.5 Screening other fibroblast cDNA libraries:

Two other fibroblast cDNA libraries were screened with the ZF39c1 clone. Table 7.3 summarises the results obtained from all the fibroblast cDNA library screens. A cDNA was not detected in either library. It was possible that insufficient clones were screened: It was present at 1 in 7.7×10^{5} in the first library screened but this library had been amplified twice which may have affected the representation of sequences in this library. It was also possible that the sequence was not present in these libraries. The original library may have been made from fibroblasts at a different stage of development or may have been contaminated with other tissue(s).

The presence of stop codons in the sequence of both the cDNA and the genomic DNA would indicate that this was a transcribed pseudogene.

Table 7.3 Results from screening three human fibroblast cDNA libraries with a ZF39p3 probe.

The number of clones from each cDNA library hybridising to a ZF39p3 probe at low ($60^{o}C$) and high ($65^{o}C$) stringency are given.

Source of mRNA	Pfu screened	60°C	65°C
Adult skin fibroblasts.	7.0×10^6	120	9
Adult foreskin fibroblasts.	4.7 x 106	239	0
Embryonic fibroblasts.	5.0×10^6	116	0

Further work on this sequence would require direct analysis of mRNA. mRNA could be directly amplified using the primers already available. PCR of mRNA from fibroblast cells at different stages of development and from a wide selection of other tissues should be sensitive enough to detect a very low level of transcript.

7.4 Summary.

The aim of these experiments was to determine whether cloned genomic sequences identified as containing the ZF motif were transcribed. Data from screens of various cDNA libraries demonstrated that expressed ZF genes are numerous and varied. The isolation of a cDNA for a genomic sequence when there are no clues as to the specificity of expression is a difficult approach to take. Two points are clear from the work with the "group 1" clones:

1) Data from screening cDNA libraries for with the ZF oligonucleotide has shown that although sequences containing the ZF motif are generally abundant in all the tissues examined, individual sequences are expressed at a very low level. These findings are supported by the results from the "group 1" screens. In the HFK cDNA library 1 in 1.9 x 10^6 clones of one sequence, ZF1c1, and 1 in 6.7 x 10^5 clones of a second sequence, ZF29c5b, were present. In the fibroblast cDNA library 1 in 7.8 x 10^5 clones were of the ZF39c1 sequence. If all the ZF genes within the cosmid clones used in this analysis were expressed at a similarly low level, transcripts would probably not be detectable on northern blots.

2) In the cross-hybridisation studies very closely related sequences could not be distinguished by hybridisation studies. In the case of ZF1c1, the cDNA had 96.5% homology to a 247bp genomic fragment and hybridised with it at 65^{o}C.

As a transcript detected on an RNA blot might not be the cognate sequence and as these sequences are expressed at a very low level further work to determine which tissues are expressing these genes would require PCR analysis of mRNA.

The direct screening of cDNA libraries showed that several different "group 1" genes were expressed in human foetal kidney, in human lympohoblasts and in human fibroblasts. Cognate sequences for one EJNAC4.1 cosmid, ZF39, were detected in the

fibroblast cDNA library. Candidate cognate sequences were detected for two other EJNAC4.1 cosmids, ZF1 and ZF30, in the foetal kidney cDNA library. There were differences in sequence between all these clones. For ZF39 and the cDNA ZF39c1, the loci was shown to be polymorphic, one polymorphism detectable as a RFLP. The disruptions in the ZF39 sequence indicate that it is probably a pseudogene. Pseudogenes generally arise from mutations in a second copy of a gene, possibly a duplicated gene or a retroposon (a sequence that arises from reverse transcription and integration of a mRNA) that are "sheltered" by the presence of the functioning copy of the gene. The polymorphism between the loci obtained from the different sources would not be unexpected as an inactivated gene would be free to accumulate changes in bases by mutation and would be predicted to be more polymorphic than a functional gene.

Members of "group 1" must be expressed in many different tissues all three of the cDNA libraries screened contain these sequences. There are many "group 1" genomic sequences in the hybrid cell lines (the majority of them represented in the cosmids clones isolated from these cell lines). Further screening with other "group 1" sequences would probably detect cognate cDNAs for many of these sequences if sufficient libraries were screened. Random isolation of "group 1" cDNAs and hybridisation to DNA blots of the "group 1" cosmids from the hybrid cell lines would also be a way to match sequences. The significance of the polymorphisms between cognate loci and the sequence conservation between different loci presents several interesting problems. Large gene families generally evolve to meet a transcriptional demand as seen with the small nuclear RNA U2 genes family with 10 to 20 tandemly repeated arrays located at a single chromosomal site at 17q21-q22 (Lindgren *et al* ., 1985) or a functional demand as with the 153-member EF-hand family (Moncrief *et al*., 1990). Additional work is required to determine the evolutionary history and current function of the "group 1" gene family.

The random screening of cDNA libraries has shown that at least some of the "group 1" sequences are expressed in a variety of tissues. Taking into account the association

of CpG islands with these sequences, it is probable that many of the ZF sequences in the other cosmids isolated from the hybrid cell lines are functional genes. As shown in this work, the detection of cDNAs for these genes will not be a trivial matter. With related genes there is a significant problem with cross-hybridisation and with all these genes it would be fortunate to detect transcribed sequences by randomly screening northerns or cDNA libraries. One possible route to detecting cDNAs would be to define specific tissues to be examined based on the possible roles of the ZF genes. Clues to the function of a particular gene might come from sequence comparison with previously isolated ZF genes or from the chromosomal location of these genes with respect to regions known to be linked to those heritable traits and diseases compatible with the general role of ZF genes as transcription factors.

Chapter 8.

Localisation of ZF sequences.

8.1 Introduction.

It was important to determine a more precise location for the cosmid clones containing the ZF motif isolated from in the hybrid cell lines, EJNAC4.1 and 4.1.5. As the 11p mapping project was at an early stage the position of these clones could not as yet be determined within this map. One way of mapping sequences to regions within the human genome is using panels of mouse-human somatic cell hybrids. These are hybrid cell lines with known regions of human DNA present. The position of a clone can be determined by hybridisation of a unique fragment to southern blots of genomic DNA from these lines. This approach was used to localise *EGR2* to chromosome 10 (Joseph *et al.*, 1991) and *PRDII-BF1* to chromosome 6 (Gaynor *et al.*, 1988). *EGR2* and *PRDII-BF1* were then subchromosomally localised to 10q21-22 and 6p22.3-p24 by *in situ* hybridisation of fluorescently labelled probes to chromosomal spreads. Since there were specific chromosomal regions within the EJNAC4.1 and EJNAC4.1.5 hybrid cell lines the ZF cosmids were directly mapped by *in situ* hybridisation. 69 of the ZF cosmids were mapped by Dr Jan Hoovers and Dr Marcel Mannens, Institute of Human Genetics, Amsterdam. The cosmids were localised by measuring the relative position of the probe on the chromosome, after chromosome identification using morphological criteria, simultaneous hybridisation with centromere-specific probes or simultaneous Q-banding. Map position of the cosmids are expressed as the fractional length of the whole chromosome relative to pter (FLpter).

8.2 In situ localisation of ZF cosmids.

8.2.1 Results.

The results from this mapping are given in table 8.1. The approximate position of clones mapping to 3p, 11p, 19p, 19qter, 20p and 21qprox are shown in figure 8.1.

Table 8.1 In situ map position of 69 ZF cosmids.

The chromosome to which each cosmid was mapped is shown with the 95% confidence interval (c.i.) for the relative distance of the cosmids to pter (FLpter) and the sublocalisation of the cosmid to a chromosomal band. 95% c.i. is based on 10-20 measurements. Where no FLpter values are given, localisation was based on banding studies. Cosmids from EJNAC4.1 are marked with *. Membership of "group 1" or "group 2" is indicated by 1 or 2 respectively.

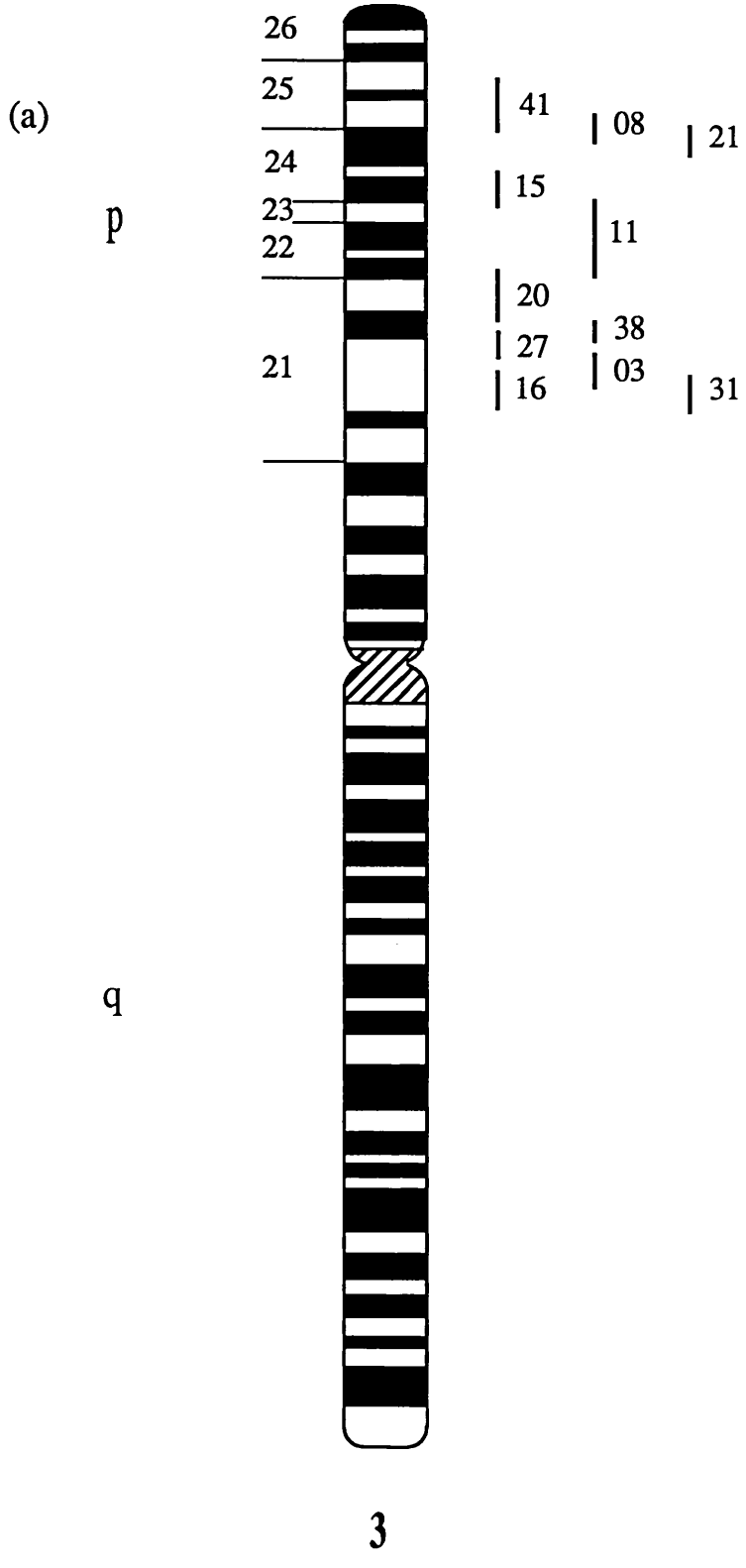
NAME	CHROMOSOME	FLpter 95% c.i.	BAND	GROUP
ZF41 *	3	0.049-0.087	p24.3-25.3	
ZF08 *	3	0.078-0.089	p24.3-p25.1	
ZF21 *	3	0.081-0.097	p24.3-p25.1	1
ZF15 *	3	0.118-0.139	p23-p24.2	
ZF11 *	3	0.136-0.188	p22.1-p24.1	
ZF20 *	3	0.185-0.215	p21.32-p22.1	
ZF38 *	3	0.213-0.231	p21.31-p21.33	2
ZF27 *	3	0.224-0.239	p21.31-p21.33	2
ZF03 *	3	0.239-0.262	p21.31	
ZF16 *	3	0.252-0.273	p21.31	
ZF31 *	3	0.255-0.281	p21.1-p21.31	
ZF50	8	---	centromere	
ZF70	11	0.047-0.062	p15.4-p15.5	
ZF83	11	0.054-0.088	p15.3-p15.4	
ZF94	11	0.100-0.118	p15.1-p15.3	
ZF04 *	11	0.113-0.144	p15.1-p15.2	
ZF40 *	11	0.132-0.151	p15.1	
ZF65	11	0.148-0.167	p14.3-p15.1	
ZF57	11	0.168-0.187	p14.3-p15.1	
ZF17 *	11	0.215-0.231	p13-p14.1	
ZF13 *	11	0.262-0.285	p12-p13	
ZF28 *	19	0.106-0.115	p13.3	2
ZF59	19	0.019-0.050	p13.3	2
ZF24 *	19	---	p13.3	
ZF58	19	---	p13.3	
ZF85	19	---	p13.3	
ZF101	19	---	p13.3	
ZF55	19	0.169-0.187	p13.12-p13.2	2
ZF53	19	0.174-0.218	p13.12-p13.2	2
ZF23 *	19	0.185-0.215	p13.12-p13.2	
ZF35 *	19	0.194-0.236	p13.12-p13.2	2
ZF60	19	0.180-0.245	p13.12-p13.2	2
ZF36 *	19	0.186-0.254	p13.12-p13.2	2
ZF37 *	19	0.228-0.327	p13.12-p13.2	2
ZF02 *	19	0.262-0.305	p13.12-p13.2	2
ZF10 *	19	0.248-0.279	p13.12-p13.2	
ZF12 *	19	0.220-0.266	p13.12-p13.2	2
ZF79	19	---	p13.12-p13.2	
ZF54	19	0.344-0.404	p12-p13.11	
ZF39 *	19	0.357-0.431	p12-p13.11	1
ZF29 *	19	0.351-0.381	p12-p13.11	1
ZF25 *	19	0.354-0.397	p12-p13.11	1
ZF76	19	---	p12-p13.11	
ZF82	19	---	p12-p13.11	
ZF86	19	---	p12-p13.11	1
ZF14 *	19	0.481-0.555	q11-q12	
ZF19 *	19	0.912-0.958	q13.4	1
ZF33 *	19	0.929-0.961	q13.4	2
ZF32 *	19	0.933-0.966	q13.4	2
ZF09 *	19	0.910-0.948	q13.4	
ZF47	19	0.948-0.972	q13.4	2
ZF46	19	---	q13.4	2
ZF34 *	19	0.953-0.962	q13.4	2
ZF51	19	0.934-0.953	q13.4	

ZF81	19	---	q13.4	
ZF84	19	---	q13.4	
ZF87	19	---	q13.4	
ZF98	19	---	q13.4	
ZF99	19	---	q13.4	
ZF103	19	---	q13.4	
ZF43	20	0.068-0.131	p12-p13	
ZF56	20	0.245-0.313	p11.22-p12.1	2
ZF75	20	---	p11.21-p11.23	
ZF06 *	20	0.394-0.414	p11.1-p11.21	
ZF18 *	21	---	q11.2-q21.1	
ZF52	21	---	q11.2-q21.1	
ZF01 *	-	---	several centromeres	1
ZF30 *	-	---	several centromeres	1
ZF100	-	---	heterochromatic regions	

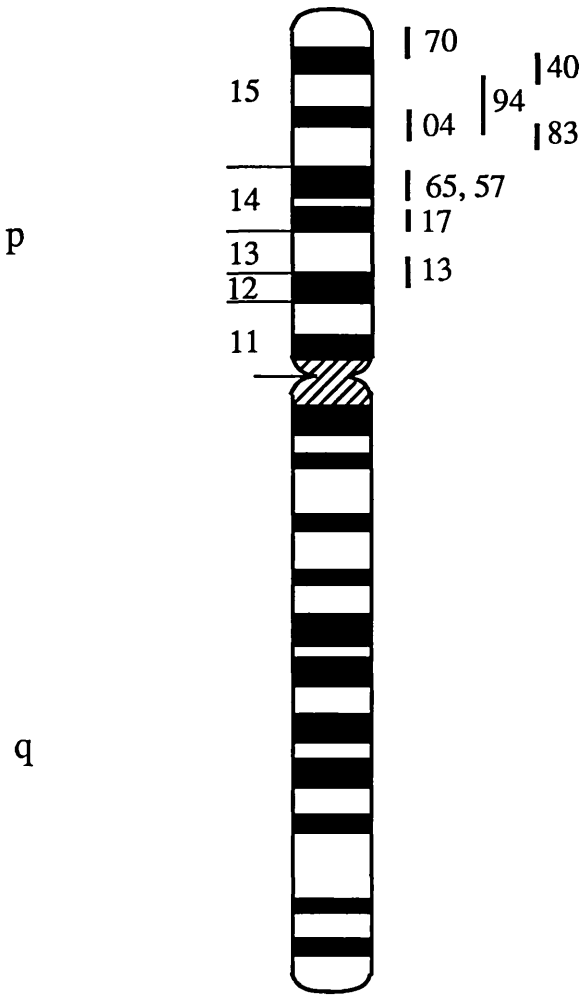
Non localised members of "group 1": ZF62, ZF63, ZF78, ZF89, ZF90, ZF92, ZF93, ZF97.
Non localised members of "group 2": ZF44, ZF48, ZF61, ZF73, ZF96.

Figure 8.1 Regional localisation of in situ mapped ZF cosmid clones.

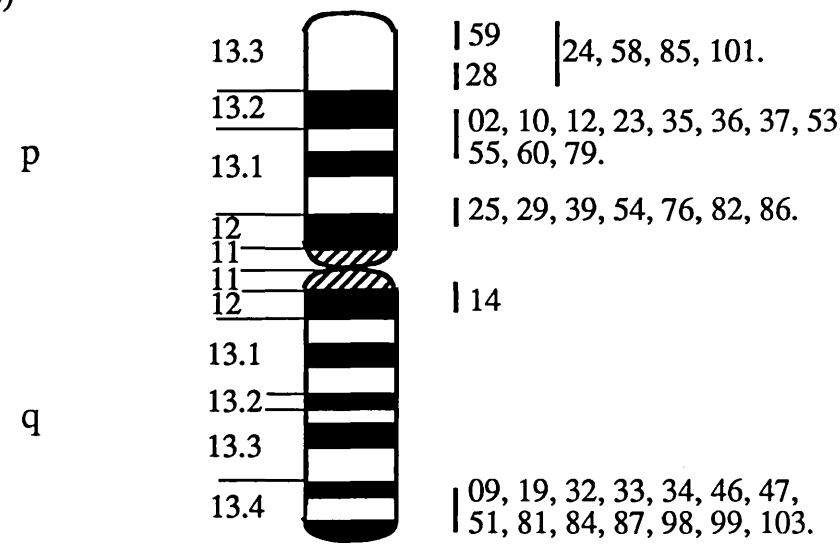
The positions of the mapped clones are indicated by the vertical lines to the left of the ideograms. (a) chromosome 3, (b) chromosome 11, (c) chromosome 19, (d) chromosome 20, and (e) chromosome 21.


(a)
p
q
26
25
24
23
22
21
41
08
21
15
11
20
38
27
03
16
31
3

(b)

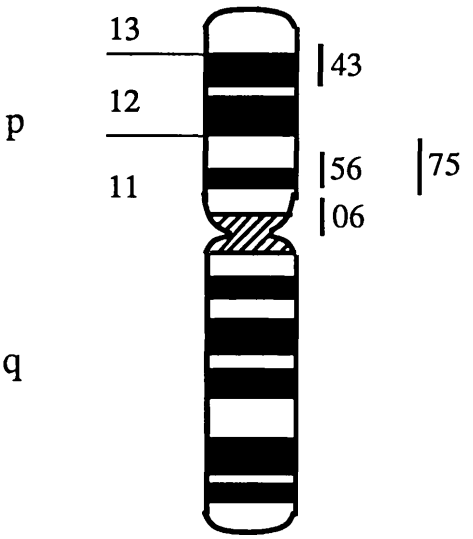


(c)


p
q
13.3
13.2
13.1
12
11
11
12
13.1
13.2
13.3
13.4
59
28
24, 58, 85, 101.
02, 10, 12, 23, 35, 36, 37, 53, 55, 60, 79.
25, 29, 39, 54, 76, 82, 86.
14
09, 19, 32, 33, 34, 46, 47, 51, 81, 84, 87, 98, 99, 103.

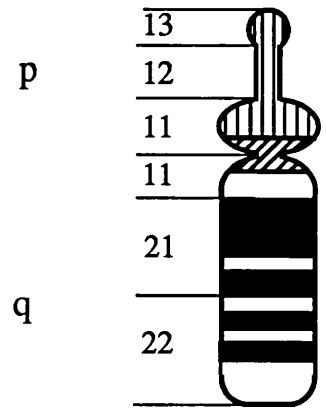
19

(d)



20

(e)



18, 52.

21

8.2.2 Nature of the human DNA in the hybrid cell lines.

Chromosomal regions other than 11p were present in the EJNAC4.1 and 4.1.5 cell lines. To define these regions, pools of 100 random cosmids from each cell line were simultaneously *in situ* hybridised to human chromosomes. The cosmids from EJNAC4.1 identified regions of human DNA from chromosomes 3p, 11p, 19p. 19pter, 20p, 21qprox and several centromeric regions also showed fluorescent signals. The hybridisation with clones from EJNAC4.1.5 identified the same regions except 3p. The loss of 3p was consistent with reduced frequency of human cosmid clones from a EJNAC4.1.5 cosmid library compared with a EJNAC4.1 cosmid library and with the absence of ZF cosmids mapping to 3p from the EJNAC4.1.5 cosmid library.

8.2 3 ZF sequences are clustered on human chromosome 19.

8.2.3.1 Predicted density of ZF sequences in the human genome.

Approximately 0.2 genome equivalents of EJNAC4.1 and 0.36 genome equivalents of EJNAC4.1.5 were screened with the ZF oligonucleotide (see section 5.6). 94 of the 101 ZF cosmids isolated in this screen contained unique ZF regions as judged by analysis of HindIII fragments and analysis of fingerprinting data from the 11p fingerprint mapping project.

On the basis that 0.5-1% cosmid clones from a total human DNA library hybridised to the ZF oligonucleotide and the average insert size for a cosmid is 37.5kb, there are an estimated 400-800 ZF sequences in the human genome. For the purpose of the following calculations the ZF sequences were assumed to be spread evenly within the human genome, one every 5.6Mb.

8.2.3.2 Density of ZF sequences on 11p and 3p.

The short arm of human chromosome 11 contains approximately 57.6Mb of DNA (Harris *et al.*, 1988) and would be predicted to contain 10 ZF sequences. 9 of the ZF cosmids map to this region of the genome. A major part of the short arm of chromosome 3 is present in EJNAC4.1 which is approximately equivalent to 94.5Mb

of DNA (chromosome 3 is 210Mb or 3.5% of the human genome, Harris *et al.*, 1988 and Naylor and Bishop, 1989). 11 of the ZF cosmids map here as compared to the predicted 17. Not all the sequences with homology to the ZF oligonucleotide from these regions have been isolated and some of the sequences mapped overlap but the coverage of 11p and 3p appears to be within the predicted limits.

8.2.3.3 Density of ZF sequences on chromosome 19.

The whole of chromosome 19 represents about 1.08% of the human genome (Harris *et al.*, 1988). An entire copy of this chromosome would represent 32.4Mb of DNA and would be expected to be contain 6 ZF sequences. From the data generated from the *in situ* hybridisation of 100 random cosmids it is clear that only part of 19p and the telomere of 19q was present and therefore less than 6 sequences would be expected. Of the 69 cosmids *in situ* mapped, 39 map to chromosome 19. Assuming there is an equal chance of mapping any particular sequence to *in situ* spreads this would indicate that 56% of the human ZF sequences in these hybrid cell lines originate from chromosome 19 even though the chromosome 19 component represents less than one quarter of the human DNA in these cell lines.

If this information is taken with the 4.4-fold and 4.8-fold increase in the number of the human DNA-containing clones from EJNAC4.1 and EJNAC4.1.5, respectively, hybridising with the ZF oligonucleotide, it indicates the clustering of these sequences within the human component of these hybrid cell lines with significant clustering on chromosome 19.

8.2.3.4 The clones mapped to chromosome 19 appear to fall into 3 clusters.

Cluster A at 19p13.12-p13.2 contains 11 cosmids.

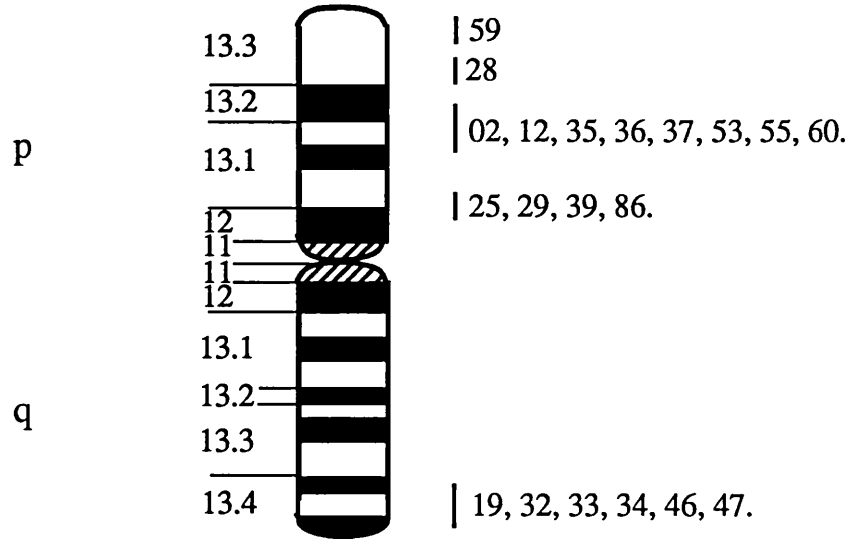
Cluster B at 19p12-p13.11 contains 7 cosmids.

Cluster C at 19q13.4 contains 14 cosmids.

Of the 32 cosmids in the three clusters, 18 are members of "group 1" or "group 2". Figure 8.2 shows the positioning of these clones on chromosome 19.

Figure 8.2 Regional localisation of in situ mapped "group 1" and "group 2" clones to chromosome 19.

Position of clones are indicated by lines to the left of the ideogram of chromosome 19.


p
q
13.3
13.2
13.1
12
11
11
12
13.1
13.2
13.3
13.4
59
28
02, 12, 35, 36, 37, 53, 55, 60.
25, 29, 39, 86.
19, 32, 33, 34, 46, 47.

19

To see if any of these clustered cosmids shared common sequence, HindIII digests were performed. Figure 8.3 (a) shows these digests resolved on agarose gels and (b) the results of hybridisation with the ZF oligonucleotide.

10 cosmids from cluster A, 4 cosmids from cluster B and 9 cosmids from cluster C were analysed. Of these, ZF35 and ZF37 overlapped completely and ZF36 appeared to share a 1.65kb fragment with ZF35 and ZF37 which hybridised with the ZF oligonucleotide. The others cosmids analysed did appear to overlap.

8.2.4 Positions of ZF sequences relative to known loci.

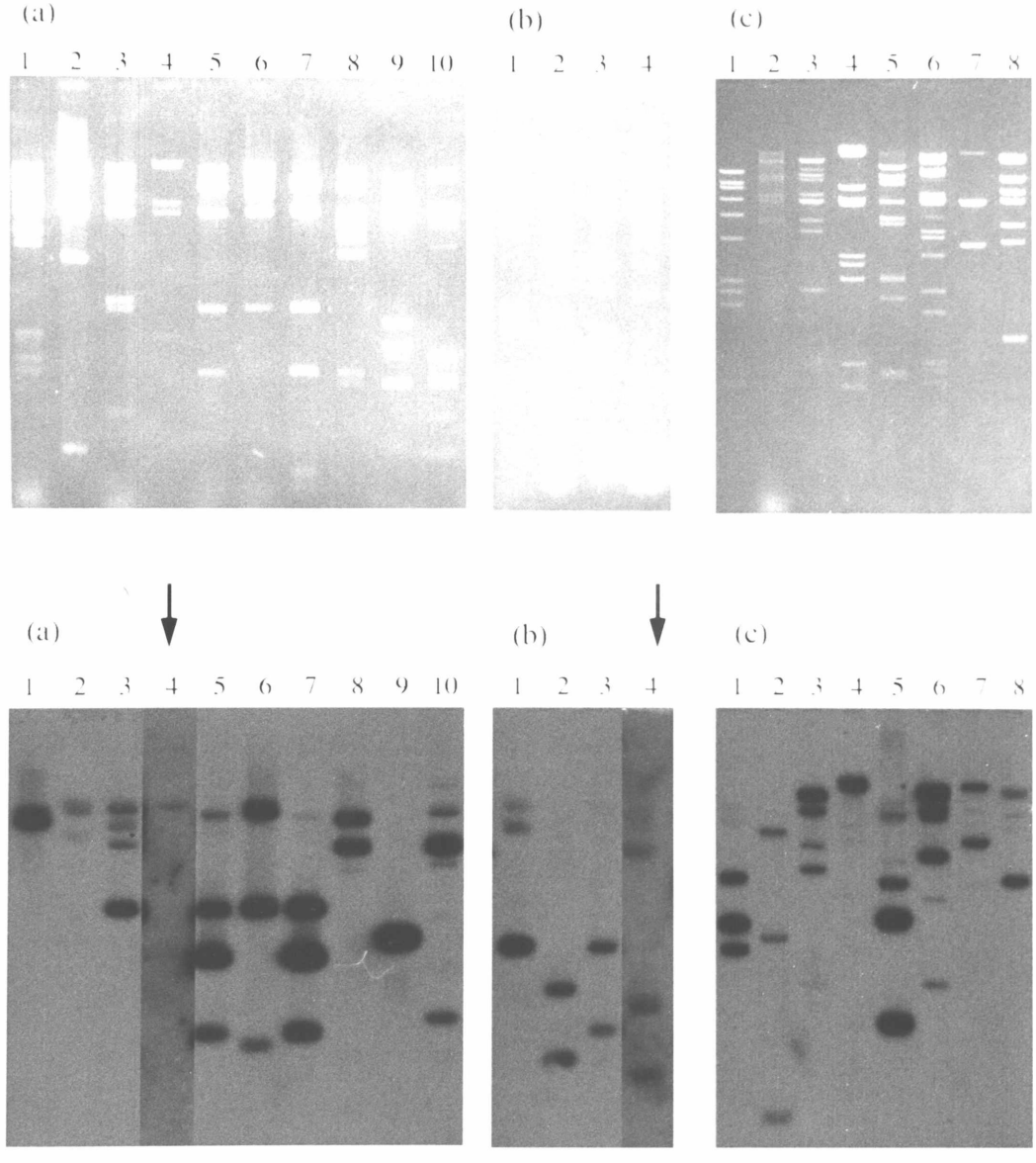
The involvement of members of the ZF protein family with complex developmental processes has previously been discussed. This involvement makes any member of this family a candidate gene for the numerous human diseases involving growth abnormalities and/or neoplasia. ZF genes are important candidates for recessive oncogenes: a recessive oncogene is one in which both alleles must be inactivated for tumours to form. The retinoblastoma gene is an example of a recessive oncogene which has been shown to code for a DNA-binding protein. Allelle loss at 11p13 in Wilm's tumour would point to this gene also being a recessive oncogene. A strong candidate gene for the 11p13 locus codes for a protein with four ZF motifs which has DNA-binding activity (Rauscher *et al.*, 1990). Although there are many classes of protein that can be defined as tumour supressor genes that do not code for DNA-binding proteins, an example is the von Recklinghausen neurofibromatosis candidate gene, a putative GTPase-activating protein (Xu *et al.*, 1990), it is probable that a many will have a DNA-binding activity.

Diseases are mapped to chromosomal regions either by genetic linkage studies generally using RFLP markers (Donis-Keller *et al.*, 1987) or by the association of structural changes in chromosomes with the disease. The mapping of any gene to chromosomal regions implicated in a disease is preliminary evidence for candidacy.

Figure 8.3 Cosmid clusters.

Agarose gel (a) and southern blot hybridised with ZF oligonucleotide (b) of HindIII digested DNAs of a proportion of the cosmid clones mapping to chromosome 19 clusters.

Cluster A contains ZF2 (1), ZF10 (2), ZF12 (3), ZF23 (4), ZF35 (5), ZF36 (6), ZF37 (7), ZF53 (8), ZF55 (9) and ZF60 (10); cluster B contains ZF25 (1), ZF29 (2), ZF39 (3) and ZF54 (4): cluster C contains ZF9 (1), ZF19 (2), ZF32 (3), ZF33 (4), ZF35 (5), ZF46 (6), ZF47 (7) and ZF51 (8). Tracks indicated with arrows were exposed to X-ray film for a longer time.


(a)
1 2 3 4 5 6 7 8 9 10
(b)
1 2 3 4
(c)
1 2 3 4 5 6 7 8
(a)
1 2 3 4 5 6 7 8 9 10
(b)
1 2 3 4
(c)
1 2 3 4 5 6 7 8

The mapping of a gene coding for a putative DNA-binding protein to a region implicated in a developmental disorder is significant evidence for candidacy.

8.2.4.1 Developmental diseases associated with 3p and 20p.

Von Hippel-Lindau Disease (VHL) is a an autosomal dominant trait characterised by a predisposition to develop retinal angiomas, cerebellar and spinal hemangioblastomas, pheochromcytomas, renal cell carcinomas, and benign tumours of the epididymis and pancreas (Melmon and Rosen, 1964). The VHL gene has recently been localised to a small region of chromosome 3 between RAF1 (3p25) and D3S18 (3p26) by genetic linkage studies of 25 families with the disease (Hosoe *et al.*, 1990). Recent data are consistent with the concept that the VHL gene is a recessive oncogene (Tory *et al.*, 1989).

Alagille-Watson Syndrome (AWS) is an autosomal-dominant inherited disease with reduced penetrance, variable expressivity and is characterised by growth retardation, pulmonary stenosis, intrahepatic hypoplasia and skeletal abnormalities. Schnittger *et al.* (1989) have proposed that AWS is a contiguous gene syndrome and have provisionally assigned a 20p11.23-p12.1 localisation.

8.2.4.2 Neoplasias linked to structural changes of 3p, 11p, 19p, 19qter 20p and 21qter.

Table 8.2 lists chromosomal rearrangements in tumours studied using direct preparations or short term cultures associated with 3p, 11p, 19p and 19qter, 20p and 21qprox, regions to which the ZF cosmids have been mapped. The data is taken from the report of the committee on structural chromosome changes in neoplasia (Trent *et al.*, 1989).

Table 8.2 Chromosomal changes associated with neoplasia and candidate ZF cosmids

The positions of the chromosomal changes is taken from Trent *et al.*, 1989. AC - adenocarcinoma, AML - acute myeloid leukemia, B-CLL - chronic lymphocytic leukemia (B cell), GL - glioma, MDS - myelodysplastic syndrome, ML - malignant lymphoma, MPD - myeloproliferative disorder, PA - pleomorphic adenoma, PV - polycythemia vera, SCC - small cell carcinoma, T-ALL - acute lymphoblastic leukemia (T cell), WT - Wilms' tumor.

Chromosomal Location	Disease	Candidate cosmids
3p21-13	AC ovarian	
3p21	ML	ZF3 ZF16 ZF20 ZF27
3p21-14	AC kidney	ZF31 ZF38
3p21-14	PA	
3p23-14	AC lung	All 3p mapped clones
3p23-14	SCC lung	
11p15	AML, MPD	ZF4 ZF40 ZF50 ZF57
11p15	AML	ZF65 ZF70 ZF83 ZF94
11p13	WT	ZF13 ZF17
11p13	T-ALL	
19p13	ALL	ZF2 ZF10 ZF12 ZF23 ZF24 ZF25
19p13	AML, ALL	ZF28 ZF29 ZF35 ZF36 ZF37 ZF39
19q13	GL	ZF53 ZF54 ZF55 ZF58 ZF59 ZF60
19q13	B-CLL	ZF76 ZF79 ZF82 ZF85 ZF86 ZF101
21p11-q11	AML, MDS	ZF18 ZF52

8.2 5 Other ZF genes mapped.

Members of the human ZF gene family have been mapped to almost every human chromosome (Page *et al.*, 1987, Muller and Schempp, 1989, Ruppert *et al.*, 1988, Sukhatme *et al*, 1988, McCance *et al.*, 1988, Ashworth *et al.*, 1989, Kato *et al.*, 1990, Lania *et al.*, 1990, Gaynor *et al* 1991 and Huebeur *et al.*, 1991). The published sequence of these genes does not correspond to the sequences determined from the EJNAC4.1 and EJNAC4.1.5 clones. Table 8.3 shows the consensus sequences for the 148 motifs analysed by Gibson *et al.* (1989) and for 45 "group 1" motifs. The consensus sequences for the five published human ZF sequences, *HKR4* (chromosome 8), *ZNF7* (chromosome 8q24), *H-plk* (chromosome 7), *Kox8* (chromosome 7q/22q11) and *Kox27* (chromosome 7q/22q11) which are the most similar published sequences to the "group 1" consensus are also shown (Ruppert *et al.*, 1988, Lania *et al.*, 1990, Kato *et al.*, 1990 and Huebner *et al.*, 1991). The consensus is taken to be that where a residue appear in over 50% of the motifs. A zinc finger motif is defined by 6 conserved amino acid residues, two cysteines, two histidines, a phenylalanine and a leucine. An additional 6 residues comprise the conserved H/C link. A consensus sequence of the 148 ZF motifs analysed by Gibson *et al.* (1989) has three other conserved residues, lysine, serine, and arginine. The remaining 13 residues vary and are likely to be less important in the structure of the ZF domain. In the 44 "group 1" ZF motifs analysed, 11 of these 13 residues are conserved in over 50% of the motifs, indicating that these sequences are very closely related. The remaining two positions are those proposed by Nardelli *et al.* (1991) to be responsible for DNA-binding. All the 11 consensus "group 1" residues are present in the *H-plk* consensus sequence, 8 in *Kox 27*, 6 in *Kox8*, 5 are conserved in *ZNF7*, and 4 in *HKR4*. *H-plk* would almost certainly be detected by a "group 1" probe and it is possible that these other ZF genes might also be detected.

Table 8.3 Consensus sequences for human ZF genes.

Consensus sequences (greater than 50% conservation) for *Kox8*, *Kox27*, *HKR4*, *ZNF7* and *H-plk* (Theisen *et al.*, 1989, Ruppert *et al.*, 1988, Lania *et al.*, 1990 and Kato *et al.*, 1990) and consensus of 148 ZF taken from Gibson *et al.*, 1988, compared to consensus sequence for 44 ZF sequences belonging to "group 1".

Name	Consensus sequence
Gibson's 148	**C XX C XKX F XXXSX L XX H XRX H TGEKPXX**
"group 1"	**C EE C GKA F NQSSX L TX H KII H TGEKPYK**
H-plk	**C EE C GKA F NQSSX L TX H KXI H TGEKPYK**
Kox27	**C EE C GKA F XXSSX L TX H KXX H XXXKPYK**
Kox8	**C XE C GKX F XXSSX L XX H KII H TGEKPYK**
ZNF7	**C XE C GKA F XXSSX L XX H QRI H TGEKPYX**
HKR4	**C XX C GKA F XXSSX L XX H QRI H TGEKPYX**

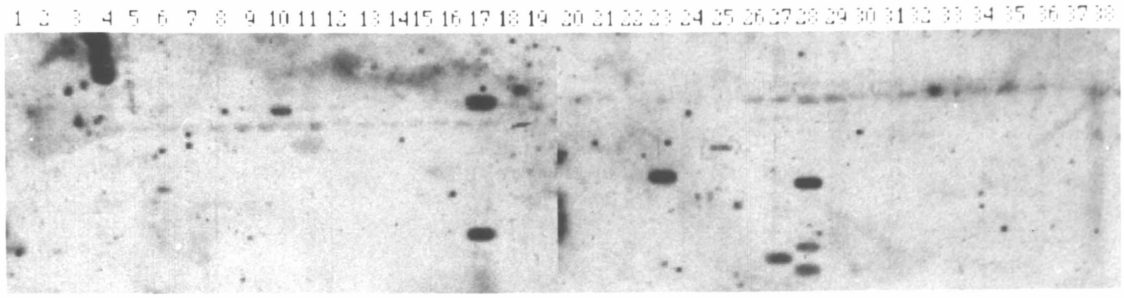
8.2.6 Related groups of ZF genes are clustered in other species.

Several groups have detected cross-hybridising sequences in other species. Crossley (1990) detected groups of closely related genes the mouse which contain many repeating units of the ZF motif, members of one group all mapped to the t-complex on chromosomes 17 and several members of a second group were mapped to the *fv-1* locus on chromosome 4 suggesting clustering of related genes in the mouse genome. Having shown that related ZF sequences were clustered within the human genome, it was possible that clusters ZF genes in the mouse had homologues in the human genome and might be related to those isolated from the EJNAC4.1 and EJNAC4.1.5 hybrid cell lines. One mouse ZF cDNA, ZFec4, was mapped to within 1cM of the Tam-1 locus (tosyl arginine methyl esterase-1) on chromosome 7, which lies between two blocks of synteny with human chromosomes 11 [PTH (11pter-p15.4) - LDHA (11p15.1-p14)] and 19 [D19S19 (19cen-q13.2) - LHB (19q13.31)] (Lallet et al., 1989). These chromosomal regions were present in the EJNAC4.1 and EJNAC4.1.5 hybrid cell lines and it was possible that ZFec4 might be related to either the "group 1" or "group 2" sequences isolated. To test this, ZFec4 was hybridised to southern blot A and washed at increasing stringency from 55-63^{o}C. At 55^{o}C ZFec4 hybridised strongly to clones ZF19, ZF25, ZF29, and ZF30 (figure 8.4(a)) which were members of "group 1". At 58.5^{o}C ZFec4 hybridised to the 5.8kb HindIII fragment of ZF19 (figure 8.4(b)) and at 60^{o}C there was no cross-hybridisation. ZF19 had been localised to 19q13.4 which could be within the region of shared synteny with mouse chromosome 7. ZFec4 appeared to be a murine "group 1" gene. It was possible that ZF19 contained the ZF gene whose murine homologue was ZFec4. It would be difficult to show that these are homologues based on sequence data as all the members of this group are closely related, at least within their finger regions. Although the clustering of ZFec4-related genes was not shown, genomic southern data suggested ZFec4 belonged to a large family of sequence-related genes.

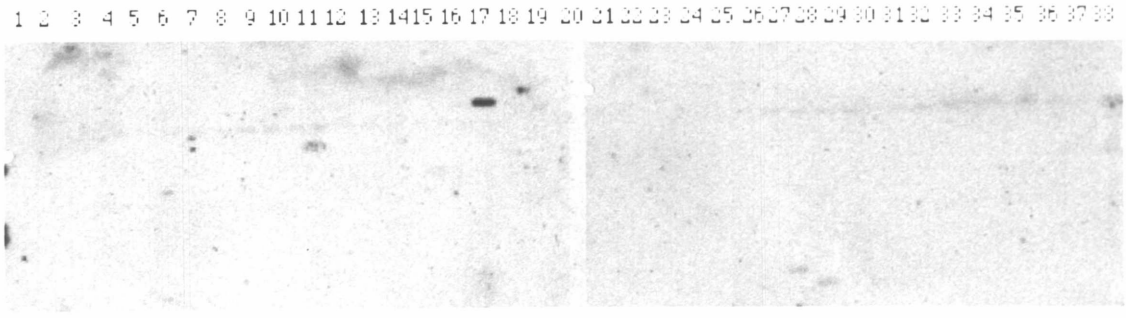
Figure 8.4 Cross-hybridisation of a mouse cDNA to human ZF cosmid clones.

Southern blot A (figure 5.3) hybridised with ZFec4, a mouse cDNA clone, washed at (a) 55^{o}C and then at (b) 58.5^{o}C. ZF1 (1), ZF2 (2), ZF3 (3), ZF4 (4), ZF6 (5), ZF8 (6), ZF9 (7), ZF10 (8), ZF11 (9), ZF12 (10), ZF13 (11), ZF14 (12), ZF15 (13), ZF16 (14), ZF17 (15), ZF18 (16), ZF19 (17), ZF20 (18), ZF21 (19), ZF22 (20), ZF23 (21), ZF24 (22), ZF25 (23), ZF26 (24), ZF27 (25), ZF28 (26), ZF29 (27), ZF30 (28), ZF31 (29), ZF32 (30), ZF33 (31), ZF34 (32), ZF35 (33), ZF36 (34), ZF37 (35), ZF38 (36), ZF39 (37) and ZF40 (38).

(a)


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38

(b)


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38

8.3 Summary.

The difficulty with detecting genes involved in heritable traits and diseases is that the vast majority of such genes are known only by their phenotype: their protein product is not known. One route to locating the chromosomal region in which these genes lie is through the use of a genetic linkage map. By studying the inheritance of randomly selected RFLPs in families, human diseases can be linked to specific markers whose chromosomal location is known. A second route to showing involvement of a specific chromosomal region in a human disease is by looking for structural changes in that chromosome. The association of recurring chromosomal alterations with neoplasia pinpoints the sites of genes important in regulatory processes such as cellular oncogenes. Once the chromosomal location associated with a particular disease is known this region must be scanned for coding sequences to determine which is involved. The primary purpose for mapping the ZF cosmids was to determine their possible role in relation to disease loci. Having mapped many of the ZF cosmids the next stage will be to use these sequences to determine in which tissues and at what time points these ZF genes are expressed and what relationship these sequences have to particular diseases. These sequences can also be used to analyse chromosomal rearrangements associated with particular diseases. If transcripts are not detected, for whatever reason, and cannot be used to follow the expression of these genes in normal and diseased tissues then the association of deletions of these sequences can be investigated instead. Even if the ZF sequences mapping to chromosomal regions associated with disease are not directly involved in the disease they are still useful tools for more precise mapping of these regions.

Incidental to the chromosomal location of ZF sequences to loci implicated in disease, a large number of ZF sequences have been shown to be clustered on chromosome 19. As some of these sequnces cross-hybridise and are closely related in sequence, it is tempting to speculate that these clusters result from duplication events. This could be investigated by comparing the organisation of these sequences within the clustered regions. As the cloned map of the human regions within the EJNAC4.1 and

EJNAC4.1.5 cell lines progresses, contigs of the regions containing clusters of ZF sequence will be built up and may provide answers to the origin of these sequences.

Chapter 9.

Discussion.

9.1 Introduction.

The search for genes within the large regions of cloned DNA generated by genome mapping projects is an important priority. Two approaches were examined: the detection of coding sequences by their association with CpG islands and the detection of a specific class of genes with a conserved motif, the zinc finger gene family.

9.2 CpG islands.

Of the three methods examined for detecting rare-cutting restriction enzyme sites, the rare-cutting restriction enzyme fingerprinting protocol (RREF) was the most suited to a large scale genome mapping project based on "colony arrays" and the fingerprinting protocol. The method developed to screen cosmids for rare-cutting restriction enzyme sites was applied to the two sets of cosmid clones, the E65-9 clones and the ZF clones.

CpG islands are associated with genes (Bird *et al.*, 1986) and finding cosmid clones with sites for BssHII, EagI and NotI will assist in the location of these genes. The identification of clones containing the recognition site for rare-cutting enzymes will also be useful in comparing the cloned DNA map with long range restriction enzyme maps of the same region.

RREF of the ZF cosmid clones indicated that more than 55% of these clones contained CpG islands. These clones were isolated on the basis of hybridisation with the zinc finger oligonucleotide which was designed to detect the nucleotide sequence coding for HTGEKPY. There was no association of sites for the rare-cutting restriction enzymes with the strength of hybridisation to the oligonucleotide. It is probable that 55% of the ZF sequences in the EJNAC hybrid cell lines are associated with CpG islands. This could be taken as evidence that these are housekeeping genes with a general cellular function rather than a role in a specific developmental event: CpG islands are associated with all polymerase II housekeeping genes characterised to date though they are also associated with several tissue-specific genes such as α-globin. This is contradicted by the known roles of several ZF genes in specific

developmental processes. Either the association of CpG islands with genes does not reflect a housekeeping role or the particular association of CpG islands with ZF genes does not reflect their housekeeping role. It is also possible that the information on the general roles of ZF genes is biased by the ways in which they have been identified. Most of the ZF genes whose functions are at least partly understood were identified as developmental mutations. This can be addressed by identifying the genes associated with the islands and investigating their expression. If all ZF genes are associated with CpG islands then the ZF cosmids not apparently containing rare-cutting restriction sites may contain islands not containing sites for the enzymes tested, or the ZF gene-associated site might not be cloned in the cosmid. Sequences detected by the oligonucleotide may be ZF pseudogenes or non-ZF sequences with homology to the ZF oligonucleotide. Published work would indicate that CpG islands can be associated with ZF genes expressed in a tissue-specific manner (*Zfa*, Ashworth *et al.*, 1990)

9.3 Zinc Finger genes are clustered.

101 human cosmid clones were detected with the ZF oligonucleotide. Sequence analysis would suggest that most of the sequences detected would code for proteins with the appropriate amino acid residues to fit the consensus sequence of the C_2H_2 zinc finger motif. 45 were isolated from screening 1017 clones from the EJNAC4.1 library and 56 were isolated from screening 1166 clones from the EJNAC4.1.5 library. The higher than expected frequency of this sequence within the hybrid cell lines resulted mainly from the clustering of these sequences to three regions on chromosome 19; nearly half the sequences mapped by *in situ* were located to this chromosome.

9.4 There are at least two closely related groups of human ZF sequences.

Two groups of ZF sequences were defined by cross-hybridisation studies. The conservation of the ZF motif region in the "group 1" sequences was over 75% in 23

of the 28 predicted amino acid residues. The two least conserved residues (<25%) were those predicted by Nardelli *et al.* (1991) to be responsible for DNA-binding specificity indicating that proteins translated from these sequences would bind different DNA sequences.

9.4.1 Related ZF sequences are clustered.

Both groups were highly abundant in the hybrid cell lines. 5 of 8 "group 1" sequences and 12 of the 19 "group 2" sequences that were *in situ* mapped, were located in three clusters on chromosome 19p. The abundance of these groups within the human genome was not determined though multiple bands on southern blots of human DNA revealed under low stringency conditions with a "group 1" probe suggested that these sequences were not confined to the human chromosomal regions within the hybrid cell lines. Miller *et al.* (1985) suggest that the high degree of conservation of the H/C link region in C_2H_2 zinc finger protein is a result of these proteins emerging from a common ancestoral finger motif, multifinger proteins arising from gene duplications and/or conversions. The observation that there are multiple cross-hybridising sequences clustering at chromosome 19 loci would be consistent with the existence of subfamilies of related ZF genes arising from duplication events.

9.4.2 A "group 1" ZF family is present in mice.

The "group 1" sequences appear to be present in the mouse genome. The detection of a band on a southern blot of mouse DNA with a "group 1" cDNA probe under stringent conditions would indicate that this sequence is also conserved in the mouse genome. A mouse cDNA, ZFec4, cross-hybridised with "group 1" sequences.

9.4.3 "Group 1" sequences are expressed.

The detection of "group 1" sequences within cDNA libraries of three human tissue indicates that at least some of the sequences isolated from the hybrid cell lines are expressed. As three out of the three tissues analysed appear to be expressing "group

1" genes, it is likely that these genes are also expressed in a wide variety of other tissues. The widespread occurrence of the group as a whole did not appear to result from the expression of the same genes in the tissues from which the cDNA libraries were made; three cDNA isolated were present only in the libraries from which they were isolated.

9.4.4 Some "group 1" sequences are expressed pseudogenes.

Sequence analysis revealed mutations involving key residues in zinc finger motifs and stop codons in both genome fragments and in cDNAs. This indicates that there are "group 1" ZF pseudogenes, some of which are expressed. Pseudogenes are DNA sequences that resemble genes but do not make a functional gene product, they are often homologues of normal genes where mutations have generally resulted in loss of expression. In cases where the genes are transcribed the sequence indicates a truncated or non-functional gene product. Examples of expressed pseudogenes include the glucocerebrosidase pseudogene (Gross *et al.*, 1990) & the interferon pseudogene (Goeddel *et al.*, 1981).

. Pseudogenes may occur as a result of mutations in a second copy of a gene: pseudogenes with introns presumably arise as a result of tandem duplication of genes and "processed" pseudogenes, which lack introns, are presumably are written into the genome from processed mRNA by reverse transcriptase. An example of a processed ZF gene or "retroposon" is *ZFa* (Ashworth *et al.*, 1990). Although a crucial cysteine residue in the third finger of *ZFa* has been mutated to a tyrosine residue the maintenance of this sequence for several million years would indicate that this does not impair the function of the protein and that this may not be a pseudogene. A second example of an expressed ZF gene with apparently deleterious mutations in the ZF motif is the human proviral mRNA transcripts of the HERV-R (ERV3) human provirus, *H-plk*, which contains stop codons in the 13^{th} and 15^{th} fingers and a histidine to glutamine mutation in the 2^{nd} finger and a cysteine to tyrosine mutation in the 3^{rd} finger (Kato *et al.*, 1990). The 1281bp

sequence before the stop codon would code for 427 amino acid residues including 11 ZF domains indicating that this protein may be translated and may have a function. Despite the apparently deleterious mutations in some of the "group 1" sequences they cannot be excluded from having a biological function.

9.4.5 Clusters of related sequences indicate they may have evolved by gene duplication.

Multiple copies of a gene are favoured by evolution when there is a high transcriptional demand. As each gene in a gene family partially shelters the other gene copies from the full force of natural selection, individual members of such a family will accumulate changes and large gene families will therefore include pseudogenes.

The sequence similarity and clustering of the two "group 1" and "group 2" ZF subgroups suggests that they may have evolved from the duplication of ancestral genes. If this is the case then it is possible that they are related in function. However, since sequence was obtained for ZF regions only, it is also possible that these sequences are dissimilar outside the ZF domain and these are not related sequences. The preliminary analysis has not identified possible functions and their abundance may merely be a reflection of the biological requirement for numerous transcription factors. If they are related functionally, their abundance would suggest that they are involved processes requiring a large number of related factors with subtlety different roles. RNA processing is one such process: TFIIIA and p43 are ZF proteins known to bind RNA (Honda and Roeder, 1980 and Joho et al., 1990). DNA methylation might also involve a large number of different DNA-binding proteins, as might the packaging of DNA into heterochromatin, a role that has been suggested for the Drosophila Suvar(3)7 gene product (Reuter et al., 1990).

In order to gain an insight into both the evolutionary relationship and the functions of "group 1" and "group 2" ZF genes, full length cDNAs for several members of each group must be isolated and the homology in these regions determined.

9.5 In situ mapping of ZF genes identifies candidates for genetic diseases and neoplasias.

It is important to know the location of ZF genes because they represent candidate genes for genetic diseases involving developmental and growth abnormalities. The importance of transcription regulators in transformation and malignancy is well known. Four ZF genes, *Evi-1*, *GLI*, *MOK-2* and the 11p13 ZF gene, have suggested roles in these processes (Morishita *et al.*, 1988, Kinzler *et al.*, *1987* and Kinzler *et al.*, 1988, Ernoult-Lange *et al.*, 1990, Call *et al.*, 1990 and Gessler *et al.*, 1990). The association of recurring chromosomal rearrangements with neoplasias identifies the sites of genes important in the growth regulatory process. ZF sequences have been in situ mapped to the chromosomal regions linked to the genetic diseases Von Hippel-Lindau disease (3p25-p26) and Alagille-Watson syndrome (20p11.23-p12.1). To investigate a possible role for any of these ZF sequences will require the analysis of the expression of these genes and the detection of subtle differences in affected individuals and their families.

9.6 Future work.

Two major challenges have been raised by this work: to investigate mapped ZF sequences with respect to disease loci and to determine the relationship of the related "group 1" and "group 2" ZF sequences.

Initially, the importance of each ZF sequence with respect to the disease loci must be determined. The ZF sequences mapping to chromosomal regions associated with neoplasia can be directly investigated by looking for recurrent deletions or rearrangements of these sequences in tumours. Genetic linkage studies could be used to follow the inheritance of the ZF sequences mapping to chromosomal regions associated with genetic diseases, once an RFLP for this region has been found. Both these lines of investigation will require the isolation of unique sequences in the region containing the ZF motif. Once a ZF sequence was shown to be closely

associated with a particular abnormality, the isolation of a full length cDNA would be the next priority. It would then be a question of determining an involvement of the gene.

There appears to be two groups of abundant, closely related and clustered ZF sequences in the human genome and probably at least one of these groups in the mouse genome. The most interesting aspect of the existance of apparently clustered, closely related ZF sequences is how they arose and what was the requirement for such large numbers.

The evolutionary relationship could be initially investigated by isolating full length cDNAs and comparing the 5' and 3' non-finger regions, the organisation of the ZF domains and the intron-exon structures of different members of the two families. This should identify whether these sequences are related. It would also be of interest to isolate cosmids for the remaining "group 1" and "group 2" sequences from a total human cosmid library and mapping by *in situ* to determine if all the sequences lie in clusters. Pulse field gel analysis could be used to determine how close these genes are within the chromosomal regions to which they have been mapped and their chromosomal organisation. They may be clustered as a result of duplication events or they may be clustered for a functional reason. The next stage would be to determine in what tissues and at what stages these genes are expressed and if there is a common expression pattern or perhaps a developmentally overlapping pattern of expression of clustered genes.

To investigate what role the EJNAC4.1 and EJNAC4.1.5 ZF genes play, whether they are involved in transcriptional regulation and if so, what sequences they bind, would first require expressing these sequences in *E. coli* and determining a DNA-binding activity. The identification of key amino acids responsible for specificity of DNA binding by Nardelli *et al.* (1991) might allow the prediction of DNA sequences to which the ZF domain might bind. These could then be tested in gel retardation assays. Gel retardation assays could also be used to determine whether they bind RNA and affinity column techniques applied to identify the RNA species.

The binding of a recombinant protein to a particular DNA sequence *in vitro* does not necessarily mean that there is an *in vivo* relationship and this could be tested in transient transfection assays.

Once binding sites are identified for the ZF proteins, the next step would be to identify genes associated with these sites with which they might interact. The possible biological roles of the zinc finger proteins are numerous and it will require extensive DNA-binding studies and expression studies to identify them.

Literature Cited.

Allitto, B. A., MacDonald, M. A., Bucan, M., Richards, J., Romano, D., Whaley, W. L., Falcone, B., Ianazzi, J., Wexler, N. S., Wasmuth, J. J., Collins, F. S., Lehrach, H., Haines, J. L. and Gusella, J. F. (1991). Increased recombination adjacent to the Huntington disease-linked D4S10 marker. Genomics 9:104-112.

Antequera, F., Macleod, D. and Bird, A. (1989). Specific protection of methylated CpGs in mammalian nuclei. Cell 58:509-517.

Antequera, F., Boyes, J. and Bird, A. (1990). High levels of *de novo* methylation and altered chromatin structure at CpG islands in cell lines. Cell 62:503-514.

Ashworth, A., Williams, B. P., Buchberg, A. M., Goodfellow, P. N., Soloman, E., Potter, J. and Willison, K. R. (1989). Chromosomal localisation of zinc finger protein genes in man and mouse. Genomics 4:323-327.

Ashworth, A., Skene, B., Swift, S. and Lovall-Badge, R. (1990). *ZFa* is an expressed retroposon derived from an alternative transcript of the *ZFx* gene. EMBO J. 9:1529-1534.

Baldarelli, R. M., Mahoney, P. A., Salas, F., Gustavson, E., Boyer, P. D., Chang, M-F., Roak, M. and Lengyel, J. A. (1988). Transcripts of the Drosophila blastoderm-specific locus, *terminus*, are concentrated posteriorly and encode a potential DNA-binding finger. Dev. Biol. 125:85-95.

Barlow, P. D. and Lehrach, H. (1987). Genetics by gel electrophoresis: the impact of pulse field gel electropohoresis on mammalian genetics. Trends Genet. 3:167-171.

Barski, G., Sorieul, S. and Cornefert, F. (1960). Production dans cultures *in vitro* de deux souches cellulaires en association, de cellules de caratere "hybrid". C. R. Acad. Sci. 251:1825-1831.

Beckwith, J. B. (1969). Macroglossia, omphalocoele, adrenal cytomegaly, gigantism and hyperplastic viseromegaly. Birth Defects 5:188-196.

Bellefroid, E. J., Lecocq, P. J.., Benhida, A., Poncelet, D. A., Belayew, A. and Martial, J. A. (1989). The human genome contains hundreds of genes coding for zinc finger proteins of the *Krüppel* type. DNA 8:377-387.

Berg, J. M. (1986). Potential metal-binding domains in nucleic acid binding proteins. Science 232:485-486.

Berg, J. (1988). Proposed structure for the zinc-binding domains from transcriptional factor IIIA and related proteins. PNAS 85:99-102.

Bickmore, W. A., Maule, J. C., van Heyningen, V. and Porteous, D. J. (1989a). Long range structure of H*ras*1-selected transgenomes. Somat. Cell Mol. Genet. 15:229-235.

Bickmore, W. A., Porteous, D. J., Christie, S., Seawright, A., Fletcher, J. M., Maule, J. C., Couillin, P., Junien, C., Hastie, N. D. and van Heyningen, V. (1989b). CpG islands surround a DNA segment located between translocation breakpoints associated with genitourinary displasia and aniridia. Genomics 5:685-693.

Bird, A. P. (1986). CpG-rich islands and the function of DNA methylation. Nature 321: 209-213.

Bird, A. P. (1987). CpG islands as gene markers in the vertebrate nucleus. Trends Genet. 3:342-347.

Bird, A. P., Taggart, M. H., Nicholls, R. D. and Higgs, D. R. (1987). Nonmethylated CpG-rich islands at the human α-globin locus: implications for evolution of the α-globin psuedogene. EMBO 6:999-1004.

Blumberg, H., Eisen, A., Sledziewski, A., Bader, D. and Young, E.T. (1987). Two zinc fingers of a yeast regulatory protein shown by genetic evidence to be essential for its function. Nature 328: 443-445.

Bopp, D., Burri, M., Baumgartner, S., Frgerio, G. and Noll, M. (1986). Conservation of a large protein domain in the segmentation gene *paired* and in functionally related genes of *Drosophila*. Cell 47:1033-1040.

Bopp, D., Jamet, E., Baumgartner, S., Burri, M. and Noll, M. (1989). Isolation of two tissue-specific *Drosophila* paired box genes, *Pox meso and Pox neuro*. EMBO J. 8:3447-3457.

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

Boulay, J.L., Dennefeld, C. and Alberga, A. (1987). The *Drosophila* developmental gene *snail* encodes a protein with nucleic acid binding fingers. Nature 330:395-398.

Bray, P. and Thiesen, H-J. (1990). Putting the finger on DNA. New Biol. 2:325-327.

Brown, R.S., Sander, C. and Argos, P. (1985). The primary structure of transcription factor TFIIIA has 12 consecutive repeats. FEBS Lett. 186: 271-274.

Brown, R. A. and Bird A, P. (1986). Long range restriction site mapping of mammalian genomic DNA. Nature 322:477-481.

Bullock, W. O., Fernandez, J. M. and Short, J. M. (1987). A high efficiency plasmid transforming recA *Escherichia coli* strain with beta-galactosidase selection. Biotech. 5:376-379.

Burke, D. T., Carle, G. F. and Olson, M. V. (1987). Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science 236:806-812.

Burmeister, M., Monaco, A. P., Gillard, E. F., van Ommen, G. J. B., Affara, N. A., Fergerson-Smith, M. A., Kunkel, L. M. and Lerhrach, H. (1988). A 10-megabase physical map of human Xp21, including the Duchenne muscular dystrophy gene. Genomics 2:189-202.

Burmeister, M., Kim, S., Roydon Price, E., De Lange, T, Tantravahi, U., Myers, R. M. and Cox, D. R. (1991). A map of the distal region of the long arm of human chromosome 21 constructed by radiation hybrid mapping and pulse field gel electrophoresis. Genomics 9:19-30.

Busch, S. J. and Sassone-Corsi, P. (1990). Dimers, leucine zippers and DNA binding domains. Trends Genet. 6:36-40.

Call, K. M., Glaser, T., Ito, C. Y., Buckler, A. J., Pelletier, J., Haber, D.A., Rose, E. A., Kral, A., Yeger, H., Lewis,W. H., Jones, C. and Houseman, D. E. (1990). Isolation and characterisation of a zinc finger polypeptide gene at the human chromosome 11 Wilms' tumor locus. Cell 60:509-520.

Capon, D. J., Chen, E. Y., Levinson, A. D., Seeberg, P. H. and Goeddel, D. V. (1983). Complete nucleotide sequences of the T24 human bladder carcinoma oncogene and its normal homologue. Nature 302:33-37.

Carle, G. F., Frank, M. and Olson, M. V. (1986). Electrophoretic separation of large DNA molecules by the periodic inversion of the electric field. Science 232:65-68.

Carrasco, A. E., McGinnis, W., Gehring, W. J. and DeRobertis, E. M. (1984). Cloning of a *X. laevis* gene expressed during early embryogeninsis coding for a peptide region homologous to *Drosophila* homeotic genes. Cell 37:409-414.

Cawthon, R. M., Weiss, R., Xu, G., Viskochil, D., Culver, M., Stevens, J., Robertson, M., Dunn, D., Gesteland, R., O'Connell, P. and White, R. (1990). A major segment of the neurofibromatosis type 1 gene: cDNA sequence, genomic structure, and point mutaions. Cell 62:193-201.

Chavrier, P., Zerial, M., Lemaire, P., Almendral, J., Bravo, R. and Charnay, P. (1988a). A gene encoding a protein with zinc fingers is activated during G0/G1 transition in cultured cells. EMBO J. 7: 29-35.

Chavrier, P., Lemaire, P., Revelant, O., Bravo, R. and Charnay, P. (1988b). Characterisation of a mouse multigene family that encodes zinc finger structures. Mol. Cell. Biol. 8:1319-1326.

Chavrier, P., Janssen-Timmen, U., Mattei, M-G., Zerial, M., Bravo, R. and Charnay, P. (1989). Structure, chromosomal location, and expression of the mouse zinc finger gene *Krox-20*: multiple gene products and coregulation with the proto-oncogene *c-fos*. Mol. Cell. Biol. 9:787-797.

Chavrier, P., Vesque, C., Galliot, B., Vigneron, M., Dolle, P., Duboule, D and Charnay, P. (1990). The segment-specific gene *Krox-20* encodes a transcription factor with binding sites in the promoter region of the *Hox-1.4* gene. EMBO J., 9:1209-1218.

Chowdhury, K., Dressler, G., Breier, G., Deutsch, U. and Gruss, P. (1988). The primary structure of the murine multifinger gene *mKr2* and its specific expression in developing and adult neurons. EMBO J. 7:1345-1353.

Christy, B. A., Lau, L. F. and Nathans, D. (1988). A gene activated in mouse 3T3 cells by serum growth factors encodes a protein with "zinc finger" sequences. PNAS 85:7857-7861.

Christy, B. and Nathans, D. (1989). DNA-binding site of the growth factor-inducible protein Zif268. PNAS 86:8737-8741.

Chu, G., Vollrath, D. and Davis, R. W. (1986). Separation of large DNA molecules by contour-clamped homogenous electric fields. Science 234:1582-1586.

Church, G.M. and Gilbert, W. (1984). Genomic Sequencing. PNAS USA 81:1991-1995.

Colberg-Poley, A. M., Voss, S. D., Chowdhury, K. and Gruss, P. (1985). Structural analysis of murine genes containing homeobox sequences and their expression in embryonal carcinoma cells. Nature 314:713-718.

Corton, J. C. and Johnson, S. A. (1989). Altering DNA-binding specificity of GAL4 requires sequences adjacent to the finger. Nature 340:724-727.

Coulson, A., Sulston, J., Brenner, S. and Karn, J. (1986). Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. PNAS 83: 7826-7830.

Coulson, A., Waterston, R., Kiff, J., Sulston, J. and Kohara, Y. (1988). Genome linking with yeast artificial chromosomes. Nature 335:184-186.

Coulter, D. E., Swaykus, E. A., Beran-Koehn, M. A., Goldberg, D., Wieschaus, E. and Schedl, P. (1990). Molecular analysis of *odd-skipped*, a zinc finger encoding segmentation gene with a novel pair-rule expression pattern. EMBO J. 8:3795-3804.

Covey, S. N. (1986). Amino acid sequence homology in *gag* region of reverse transcribing elements and the coat protein gene of cauliflower mosaic virus. Nuc. Acids Res. 2:623-633.

Cox, D. R., Burmeister, M., Price, E. R., Kim, S. and Myers, R.M. (1990). Radiation hybrid mapping: A somatic cell genetic method for constructing high resolution maps of mammalian chromosomes. Science 250:245-250.

Cunliffe, V., Koopman, P., McLaren, A. and Trowsdale, J. (1990). A mouse zinc finger gene which is transiently expressed during spermatogenesis. EMBO J. 9:197-205.

Diakin, G. P., Fairall, L. and Klug, A. (1986). EXAFS study of the zinc finger binding sites in the protein Transcription Factor IIIA. Nature 324:698-699.

Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K., Keith, T. P., Bowden, D. W., Smith, D. R., Lander, E. S., et al. (1987). A genetic linkage map of the human genome. Cell 51:319-337.

Drivas, G. T., Shih, A., Coutavas, E., Rush, M. G. and D'Eustachio, P. (1990). Characterisation of four novel *ras*-like genes expressed in a human teratocarcinoma cell line. Mol. Cell. Biol. 10:1793-1798.

Eisen, A., Taylor, W. E., Blumberg, H. and Young, E. T. (1988). The yeast regulatory protein ADR1 binds in a zinc dependant manner to the upstream activating sequence of *ADH2*. Mol. Cell. Biol. 8:4552-4556.

Elvin, P., Slynn, G., Black, D., Graham, A., Butler, R., Riley, J., Anand, A. and Markham, A. F. (1990). Isolation of cDNA clones using yeast artificial chromosome probes. Nuc. Acids Res. 18:3913-3917.

Engelke, D. R., Ng, S-Y., Shastry, B. S. and Roeder, R. G. (1980). Specific interaction of a purified transcription factor with an internal control region of 5S RNA genes. Cell 19:717-728.

Ernoult-Lange, M., Kress, M. and Hamer, D. (1990). A gene that encodes a protein solely of zinc finger domains is preferentially expressed in transformed mouse cells. Mol. Cell. Biol. 10:418-421.

Estiville, X. and Williamson, R. (1987). A rapid method to identify cosmids containing rare restriction sites. Nuc. Acids Res. 15:1415-1425.

Evans, R.M. and Hollenberg, S.M. (1988). Zinc fingers: Guilt by association. Cell 52:1-3.

Fairall, L., Rhodes, D. and Klug, A. (1986). Mapping of the sites of protection on a 5S RNA gene by the *Xenopus* transcription factor IIIA. J. Mol. Biol. 192:577-591.

Fasano, L., Roder, L., Core, N., Alexandre, E., Vola, C., Jacq, B. and Kerridge, S. (1991). The gene *teashirt* is required for the development of *Drosophila* embryonic trunk segments and encodes a protein with widely spaced zinc finger motif. Cell 64:63-79.

Fearon, E. R., Vogelstein, B. and Feinberg, A. P. (1984). Somatic deletion and duplication of genes on chromosome 11 in Wilms' tumours. Nature 309:176-178.

Fearon, E. R., Cho, K. R., Nigro, J. M., Kern, S. E., Simons, J. W., Ruppert, J. M., Hamilton, S. R., Preissinger, A. C., Thomas, G., Kinzler, K. W. and Vogelstein, B. (1990). Identification of a chromosomal 18q gene that is altered in colorectal cancers. Science 247:49-56.

Fekete, G., Gessler, M., Bruns, G, A., Huerre-Jeanpierre, C., Junien, C., et-al (1990). Role for the Wilms tumor gene in genital development? PNAS 87:5383-5386.

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

Finney, M., Ruvkun, G. and Horvitz, H. (1988). The *C. elegans* cell lineage and differentiation gene *unc-86* encodes a protein with a homeodomain and extended similarity to transcription factors. Cell 55:757-769.

Francke, U., Holmes, L. B., Atkins, L. and Riccardi, V. M. (1979). Aniridia-Wilms' tumour association: evidence for specific deletion of 11p13. Cytogenet. Cell Genet. 24:185-192.

Freemont, P. S., Hanson, I. M. and Trowsdale, J. (1991). A novel cysteine-rich sequence motif. Cell 64:483-484.

Friend, S. H., Bernards, R., Rogeli, S., Weinberg, R. A., Raraport, J. M., Albert, D. M. and Dryja, T. P. (1986). A human DNA seqment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. Nature 323:643-646.

Frischauf, A-M. (1985). The T/t complex of the mouse. Trends Genet. 1:100-103.

Fulton, T. R., Bowcock, A. M., Smith, D. R., Daneshvar, L., Green, P., Cavalli, S. L. L. and Donis-Keller, H. (1989). A 12 megabase restriction map at the cystic fibrosis locus. Nuc. Acids Res. 17:271-284.

Gaynor, R. B., Muchardt, C., Diep, A., Mohandas, T. K., Sparkes, R. S. and Lusis, A. J. (1991). Localisation of the zinc finger DNA-binding protein HIV-EP1/MBP-1/PRDII-BF1 to human chromosome 6p22.3-p24.

Gerhard, D. S., Kawasaky, E. S., Bancroft, C. F. and Szabo, P. (1981). Localisation of a unique gene by direct hybridisation *in situ*. PNAS 78:3755-3759.

Gessler, M., Poustka, A., Cavanee, W., Neve, R. L., Orkin, S. H. and Bruns, G. A. P. (1990). Homozygous deletion in Wilms' tumours of a zinc-finger gene identified by chromosome jumping. Nature 343:774-778.

Gibson, T. J. and Sulston J. E. (1987a). Preparation of large numbers of plasmid DNA samples in microtiter plates by the alkaline lysis method. Gene Anal. Techn. 4: 41-44.

Gibson, T., Rosenthal, A. and Waterson, R. (1987b). Lorist 6, a cosmid vector vector with BamHI, NotI, ScaI and HindIII cloning sites and altered neomycin phosphotransferase gene expression. Gene 53:283-286.

Gibson, T. J., Postma, J. P. M., Brown, R. and Argos, P. (1988). A model for the tertiary structure of the 28 residue DNA-binding motif ("zinc finger") common to many eukaryotic transcriptional regulatory proteins. Prot. Eng. 2:209-218.

Giadroc, D. P., Keating, K. M., Williams, K. R., Konisberg, W. H. and Coleman, J. E. (1986). Gene 32 protein, the single stranded DNA-binding protein from bacteriophage T4, is a zinc metallo protein. PNAS 83:8452-8456.

Goate, A., Chartier-Harlin, M-C., Mullan, M., Brown, J., Crawford, F., Fidani, L., Giuffra, L., Haynes, A., Irving, N., James, L., Mant, R., Newton, P., Rooke, K., Roques, P., Talbot, C., Pericak-Vance, M., Roses, A., Williamson, R., Rossnor, M., Owen, M. and Hardy, J. (1991). Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. Nature 349:704-706.

Goeddel, D. V., Leung, D. W., Dull, T. J., Gross, M., Lawn, R. M., McCandliss, R., Seeburg, P. H., Ulrich, A., Yelverton, E. and Gray, P. W. (1981). The structure of eight distinct cloned human leukocyte interferon cDNAs. Nature 290:20-26.

Gosden, C., Bard, J., Buckler, A., Pelletier, J., Housman, D., et-al. The candidate Wilms' tumour gene is involved in genitourinary development (1990). Nature 346 :194-197.

Goss, S. J. and Harris, H. (1975). New method for mapping genes in human chromosomes. Nature 255:680-684.

Green, S., Kumar, V., Theulaz, I., Wahli, W. and Chambon, P. (1988). The N-terminal DNA-binding 'zinc-finger' of the oestrogen and glucocorticoid receptors determines target gene specificity. EMBO J. 7:3037-3044.

Gross, S. E., West, C. and Beutler, E. (1990). High level transcription of the glucocerebrosidase pseudogene in normal subjects and patients with Gaucher disease. J. Clin. Invest. 86:1137-1141.

Grundy, P., Koufos, A., Morgan, K., Li, F. P., Meadows, A. T. and Cavenee, W. K. (1988). Familial predisposition to Wilms' tumor does not map to the short arm of chromosome 11. Nature 336:374-376.

Hanahan, D. (1983). Studies on transformation of *Escherichia coli* with plasmids. J. Mol. Biol. 166:557-580.

Hard, T., Kellenbach, E., Boelens, R., Maler, B. E., Dahlman, K., Freedman, L. P. Carlstedt-Duke, J., Yamamoto, K. R., Gustaffson, J-A. and Kapstein, R. (1990). Solution structure of the glucocorticoid receptor DNA-binding domain. Science 249:157-160.

Harris, P., Boyd, E., Young, B. D. and Ferguson-Smith (1986). Determination of human DNA content of human chromosomes by flow cytometry. Cytogenet. Cell. Genet. 41:14-21.

Harrison-Lavoie, K., John, R., Porteous, D. and Little, P. (1989). A cosmid clone map derived from a small region of human chromosome 11. Genomics 5: 501-509.

Hastings, R. J. and Franks, L. M. (1983). Cellular heterogeneity in a tissue culture cell line derived from a human bladder carcinoma. Brit. J. Cancer 47:233-244.

Hartshorne, T. A., Blumberg, H. and Young, E. T. (1986). Sequencetor with 5S RNA and altered levels of the factor during cell differentiation. Cell 22 119-126.

Herr, W., Sturm, R. A., Clerc, R. G., Corcoran, L. M., Baltimore, D., Sharp, P. A., Ingraham, H. A., Rosenfeld, M. G., Finney, M., Ruvkun, G. and Horivitz, H. R. (1988). The POU domain: a large conserved region in the mammalian *pit-1*, *oct-1*, *oct-2*, and *Caenorhabditis elegans unc-86* gene products. Genes Dev. 2:1513-1516.

Honda, B. M., and Roeder, R. G., (1980). Association of a 5S gene transcription factor with 5S RNA and altered levels of the factor during cell differentiation. Cell 22:119-126.

Hoovers, J. M. N., Mannens, M., John, R., Bliek, J., van Heyningen, V., Poteous, D. J., Leschot, N. J., Westerveld, A. and Little, P. F. R. High resolution localisation of 69 potential human zinc finger protein genes: a number are clustered. (Submitted to genomics)

Hosoe, S., Brauch, H., Latif, F., Glenn, G., Daniel, L., Bale, S., Choyke, P., Gorin, M., Oldfield, E., Berman, A., Goodman, J., Orcutt, M. L., Hampsch, K., Delisio, J., Modi, W., McBride, W., Anglard, P., Weiss, G., Walther, M. M., Linehan, W. M., Lerman, M. I. and Zbar, B. (1990). Localisation of the Von Hipperl-Lindau disease gene to a small region of chromosome 3. Genomics 8:634-640.

Huang, H. J. S., Yee, J. K., Shew, J. Y., Chen, P. L., Bookstein, R., Friedmann, T., Lee, E. Y. H. P. and Lee, W. H. (1988). Supression of the neoplastic phenotype by replacement of the *Rb* gene in human cancer cells. Oncogene 3:345-348.

Huckaby, C. S., Conneely, O. M., Beattie, W. G., Dobson, A. D. W., Tsai, M-J. and O'Malley, B. W. (1987). Structure of the chromosomal chicken progesterone receptor gene. PNAS 84:8380-8384.

Huebner, K., Druck, T., Croce, C. M. and Thiesen, H-J. (1991). Twenty-seven non-overlapping zinc finger cDNAs from human T cells map to nine different chromosomes with apparent clustering. Am. J. Hum. Genet. 48:726-740.

Huff, V., Compton, D. A., Chao, L., Strong, L. C., Geiser, C. F. and Saunders, G. F. (1986). Lack of linkage of familial Wilms' tumour to chromosomal band 11p13. Nature 336:377-378.

Johnson, M. (1987). Genetic evidence that zinc is an essential co-factor in the DNA binding domain of GAL4 protein. Nature 328:353-355.

Joho, K. E., Darby, M. K., Crawford, E. T and Brown, D. D. (1990). A finger protein structurally simalar to TFIIIA that binds exclusively to 5S RNA in *Xenopus*. Cell:293-300.

Joseph, L. J., Le Beau, M. M., Jamieson Jr, G. A., Acharya, S., Shows, T. B., Rowley, J. D. and Sukhatme, V. P. (1988). Molecular cloning, sequencing, and mapping of *EGR2*, a human early growth response gene encoding a protein with "zinc-binding finger" structure. PNAS. 85:7164-7168.

Junien, C. and McBride, O. W. (1989). Report of the committee on the genetic constitution of chromosome 11. Cytogenet. Cell Genet. 51:226-258.

Joyner, A., Kornberg, T., Coleman, K., Cox, D. and Martin, G. (1985). Expression during embryogenisis of a mouse gene with sequence homology to the *Drosophila engrailed* gene. Cell 43:29-37.

Joyner, A. and Martin, G. (1987). *En-1* and *En-2*, two mouse genes with sequence homology to the Drosophila *engrailed* gene: expression during embryogenesis. Genes Dev. 1:29-38.

Kadonaga, J.T., Carner, K.R., Masiar, F.R., and Tijan, R. (1987). Isolation of cDNA encoding transcription factor Sp1 and functional analysis of DNA binding domain. Cell 51:1079-1090.

Kadonaga, J. T., Courney, A. J., Ladika, J. and Tijan, R. Distinct regions of Sp1 modulate DNA binding activity and transcriptional activation. Science 242:1566-

Kato, N., Shimotohno, K., VanLeeuwen, D and Cohen, M. (1990). Human proviral mRNAs down regulated in choriocarcinoma encode a zinc finger protein related to Krüppel. Mol. Cell. Biol. 10:4401-4405.

Keshet, I., Yisraeli, J. and Cedar, H. (1985). Effect of regional DNA methylation on gene expression. PNAS 82:2560-2564.

Kieny, M. P., Lathe, R., Lecocq, J. P. (1983). New versatile cloning and sequencing vectors based on bacteriophage M13. Gene 26:91-99.

Kinzler, K. W., Bigner, S. H., Bigner, D. D., Trent, J. M., Law, M. L., O'Brien, S. J., Wong, A. J. and Vogelstein. Identification of an amplified, highly expressed gene in a human glioma. Science 236:70-73.

Kinzler, K.W., Ruppert, J.M., Bigner, S.H. and Vogelstein, B. (1988). The *GLI* gene is a member of the Krüppel family of zinc finger proteins. Nature (London) 332: 371-374.

Klein, J. (1975). "Biology of the mouse histocompatability-2 complex." Springer-Verlag, New York.

Klein, J. (1986). "Natural history of the major histocompatability complex." Whiley, New York.

Knochel, W., Koster, M., El-Baradi, T., Nietfeld, W., Bouwmeester, T. and Pieler, T (1989). Evolutionary conserved modules associated with zinc fingers in *Xenopus laevis*. PNAS. 86:6097-6100.

Knudson, A. G. (1985). Heriditary cancer, oncogenes, and antioncogenes. Cancer Res. 45:1437-1443.

Koenig, M., Hoffman, E. P., Bertelson, C. J. Monaco, A. P., Feener, C. and Kunkel, L. M. (1987). Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organisation of the DMD gene in normal and affected individuals. Cell:50:509-517.

Kohara, Y., Akiyama, K. and Isono, K. (1987). The physical map of the whole *E. coli* chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. Cell:50:495-508.

Koufos, A., Hansen, M. F., Lampkin, B. C., Workman, M. L., Copeland, N. G., Jenkins, N. A. and Cavenee, W. K. (1984). Loss of alleles on human chromosome 11 during genesis of Wilms' tumour. Nature 309:170-172.

Koufos, A., Grundy, P., Morgan, K., Aleck, A. A., Hadro, T., Lampkin, B. C., Kalbakji, A. and Cavenee, W. K. (1989). Familial Wiedemann-Beckwith syndrome and a second Wilms' tumor locus both map to 11p15.5. Am. J. Hum. Genet. 44:711-719.

Krust, A., Kastner, P., Petkovich, M., Zelent, A. and Chambon, P. (1989). A third human retenoic acid receptor, hRAR-γ. PNAS 86:5310-5314.

Lalley, P. A., Davison, M. T., Graves, J. A., O'Brien, S. J., Womack, J. E., Roderick, T. H., Crean-Goldberg, N., Hiilyard, A. l., Dolittle, D. P. and Rogers, J. A. (1989). Report of the committee on comparative mapping. Cytogenet. Cell. Genet. 51:503-532.

Landschulz, W. H., Johnson, P. F. and McKnight, S. L. (1988). The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. Science 240:1759-1764.

Lania, L., Donti, E., Pannuti, A., Pascucci, A., Pengue, G., Feliciello, I., La Mantia, G., Lanfrancone, L. and Pelicci, P-G. (1990). cDNA isolation, expression analysis, and chromosomal localisation of two human zinc finger genes. Genomics 6:333-340.

Laughon, A. and Gesteland, R. F. (1984). The primary structure of the *Saccharomyces cerevisiae GAL4* gene. Mol. Cell. Biol 4:260-267.

Lecocq, J. P. and Gathoye, A. M. (1973). A bacterial mutation augmenting the frequency of lysogenation by lambda phage. Arch. Int. Biochem. Physiol. 81:803.

Lee, M. S., Gippert, G. P., Soman, K. V., Case, D. A. and Wright, P. E. (1989). Three-dimensional solution structure of a single zinc finger DNA-binding domain. Science 245:635-637.

Lee, W. H., Bookstein, R. Hong, F., Young, L. J., Shew, Y. and Lee, E. Y. (1989a). Human retinoblastoma susceptibility gene: Cloning, identification and sequence. Science 235:1394-1399.

Lee, W. H., Shew, J. Y., Hong, F. D., Sery, T. W., Donoso, L. A., Young, L. J., Bookstein, R. and Lee, E. Y. (1989b). The retinoblastoma susceptibility gene encodes a nuclear phosphprotein associated with DNA binding activity. Nature 329:642-647.

Lemaire, P., Revelant, O., Bravo, R. and Charnay, P. (1988). Two genes encoding potential transcription factors with identical DNA-binding domains are activated by growth factors in cultured cells. PNAS 85:4691-4695.

Levine, M., Rubin, G. and Tijian, R. (1984). Human DNA sequences homologous to a protein coding region conserved between homeotic genes of *Drosophila*. Cell 38:667-673.

Levine, M. and Hoey, T (1988). Homeobox proteins as sequence specific transcription factors. Cell 55:537-540.

Libert, F., Parmentier, M., Lefort, A., Dinsart, C., Van Sande, J., Maenhaut, C., Simons, M-J., Dumont, J. E. and Vassart, G. (1989). Selective amplification and cloning of four new members of the G protein-coupled receptor family. Science 244:569-572.

Lichter, P., Chang, C-J., Call, K., Hermanson, G., Evans, G. A., Housman, D. and Ward, D. C. (1990). High-resolution mapping of human chromosome 11 by *in situ* hybridisation with cosmid clones. Science 247:64-69.

Little, P.F.R. and Cross, S. H. (1985). A cosmid vector that facilitates restriction enzyme mapping. PNAS 82:3159-3163.

Little, P.F.R. (1987). "Choice and use of cosmid vectors". In Gene Cloning (D.M. Glover, Ed.), Vol. 3 pp. 19-42. IRL Press, Oxford.

Lingren, V., Ares, M., Weiner, A. M. and Francke, U. (1985). Human genes for U2 small nuclear RNA map to a major adenovirus 12 modification site on chromosome 17. Nature 314:115-116.

Lloyd, J., Brownson, C., Tweedie, WS., Charlton, J. and Edwards, Y. (1987). Human muscle carbonic anhydrase: gene structure and DNA methylation patterns in fetal and adult tissues. Genes Dev. 1:594-602.

Love, D. R., Hill, D. F., Dickson, G., Spurr, N. K., Byth, B. C., Marsden, R. F., Walsh, F. S., Edwards, Y. H. and Davies K. E. (1989). An autosomal transcript in skeletal muscle with homology to dystrophin. Nature 339:55-58.

Lowndes, N. F., Paul, J., Wu, J., Allan, M. (1989). *c-Ha-ras* gene possesses a bidirectional promotor. Mol. Cell. Biol. 9:3758-3770.

Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982). "Molecular cloning : A Laboratory Manual". Cold Spring Harbour Laboratory, Cold Spring Harbour, NY.

Mardon, G., Mosher, R., Disteche, C. M., Nishioka, Y., McLaren, A. and Page, D. C. (1989). Duplication, deletion, and polymorphism in the sex-determining region of the mouse Y chromosome. Science 243:78-80.

Matsunaga, E. (1981). Genetics of Wilms' tumor. Hum. Genet. 57:231-246.

McBride, O. W. and Peterson, J. L. (1980). Chromosome-mediated gene transfer in mammalian cells. Ann. Rev. Genet. 14:321-345.

McCarrey, J. R., and Thomas, K. (1987). Human testis specific *PGK* gene lacks introns and possesses characteristicss of a processed gene. Nature 326:501-505.

McGinnis, W., Levine, M. S., Hafen, E., Kuroiwa, A. and Gehring, W. J. (1984). A conserved DNA sequence in homeotic genes of the Drosophila antennapedia and bithorax complexes. Nature 308:428-433.

McGinnis, W., Hart, C.P.,Gehring, W. J. and Ruddle, F. H. (1984). Molecular cloning and chromosome mapping of a mouse DNA sequence homologous to homeotic genes of Drosophila. Cell 38:675-680.

McKeon, C., Ohkubo, H., Pastan, I. and de Crombrugghe, B. (1982). Unusual methylation pattern of the *α2(I)* collagen gene. Cell 29:203-210.

McKusick, V. A. (1988). "Mendelian inheritance in man." 8th ed. Johns Hopkins University Press, Baltimore.

Meehan, R. R., Lewis, J. D., McKay, S., Kleiner, E. L. and Bird, A. P. (1989). Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. Cell 58:499-507.

Melmer G., Sood, R., Rommens, J., Rego, D., Tsui, L-C. and Buchwald, M. (1990). Isolation of clones on chromosome that contain recognition sites for rare-cutting enzymes by oligonucletide hybridisation. Genomics 7:173-181.

Melmer G. and Buchwald, M. (1990). Use of short oligonucleotides to screen cosmid libraries for clones containing G/C-rich sequences. DNA Cell Biol. 9:377-385.

Melmon, K. L., and Rosen, S. (1964). Lindau's disease. Review of the literature and study of a large kindred. Amer. J. Med. 36:595-617.

Milbrandt, J. (1987). A nerve growth factor-induced gene encodes a possible transcriptional regulatory factor. Science 238:797-799.

Miller, J., Mclachlan, A.D. and Klug, A. (1985). Repetative zinc-binding domains in the protein transcription factor TFIIIA from *Xenopus* oocytes. EMBO J. 4: 1609-1614.

Miller, R. W., Fraumeni, J. F. and Maninning, M. D. (1964). Association of Wilms' tumor with aniridia, hemihypertrophy and other congenital anomalies. N. Engl. J. Med. 270: 922-927.

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

Morishita, K., Parker, D.S., Mucenski, M.L., Jenkins, N.A., Copeland, N.G. and Ihle, J.N. (1988). Retroviral activation of a novel gene encoding a zinc finger protein in IL-3-dependant myeloid leukemia cell lines. Cell 54:831-840.

Morten, J. E. N., Hirst, M. C. and Porteous D. J. (1987). The c-Harvey-*ras*-1 oncogene in chromosome mediated gene transfer. Antican. Res. 7:573-588.

Moses, K., Ellis, M. C. and Rubin, G. M. (1989). The *glass* gene encodes a zinc-finger protein required by *Drosophila* photoreceptor cells. Nature 340:531-534.

Muller, M. M., Carrasco, A. E. and DeRobertis, E. M. (1984). A homeobox-containing gene expressed during oogenisis in Xenopus. Cell 39:157-162.

Muller, G. and Schempp, W., (1989). The human ZYX locus maps to Xp21.3 by *in situ* hybridisation. Hum. Genet. 82:82-84.

Mulligan, R. C. and Berg, P. (1980). Expression of bacterial gene in mammalian cells: *E. coli* gene codes for xanthine-guanine phosphoribosyl transferase corrects defect in human Lesch-Nyhan cells. Science 209:1422-1427.

MuKusick, V. A. (1988). "Mendelian inheritance in man". 8^{th} edition. Johns Hopkins University Press, Baltimore.

Murray, N. E., Brammar, W. J. and Murray, K. (1977). Lambdoid phages that simplify the recovery of *in vitro* recombinants. Mol. Gen. Genet. 150:53-61.

Nagai, K., Nakaseko, Y., Nasmyth, K. and Rhodes, D. (1988). Zinc-finger motifs expressed in *E. coli* and folded *in vitro* direct sequence specific binding to DNA. Nature 332:284-286.

Nagamine, C. M., Chan, K., Kozak, C. A. and Lau, Y-F. (1989). Chromosome mapping and expression of a putative testis-determining gene in mouse. Science 243:80-83.

Nakamura, Y., Leppert, M., O'Connell, P., Wolffe, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M. Kumlin, E. and White, R. (1987). Variable number of tandem repeat (VNTR) markers for human gene mapping. Science 235:1616-1622.

Nardelli, J., Gibson, T. J., Vesque, C. and Charnay, P. (1991). Base sequence discrimination by zinc-finger DNA-binding domains. Nature 349:175-178.

National Research Council (1988). "Mapping and sequencing the human genome". National Academy Press, Washington, DC.

Nehlin, J. O. and Ronne, H. (1990). Yeast MIG1 repressor is related to the mammalian early growth response and Wilm's tumour finger proteins. EMBO 9:2891-2898.

Nietfeld, W., El-Baradi, T., Mentzel, H. and Pieler, T. (1989). Second-order repeats in *Xenopus laevis* finger proteins. J. Mol. Biol. 208:639-659.

Nigro, S. M., Baker, S. J., Preisinger, A. C., Jessop, J. M., Hostetter, R., Cleary, K., Bogner, S. H., Davidson, N., Baylin, S., Devilee, P., Glover, T., Collins,

F., Weston, A., Modali, R., Harris, C. C. and Vogelstein, B. (1989). Mutations in the *p53* gene occur in diverse human tumour types. Nature 342:705-709.

Okayama, H. and Berg, P. (1982). High efficiency cloning of full length cDNA. Mol. Cell. Biol 2: 161-170.

Okayama, H., Kawaichi, M., Brownstein, F. L., Yokota, T. and Arai, K. (1987). High efficiency cloning of full length cDNA; Construction and screening of cDNA expression libraries for mammalian cells. Methods Enzymol. 154:3-28.

Oliva, R. and Dixon, G. H. (1989). Chicken protamine genes are intronless. J. Biol. Chem. 264:12472-12481.

Ollo, R., and Maniatis, T. (1987). *Drosophila Krüppel* gene product produced in a baculovirus expression system is a nuclear phosphoprotein that binds to DNA. PNAS 84:5700-5704.

Olson, M. V., Dutchik, J. E., Graham, M. Y., Brodeur, G. M., Helms, C., Frank, M., MacCollin, M., Scheinman, R. and Frank, T. (1986). Random clone strategy for genomic restriction mapping in yeast. PNAS 83:7826-7830.

Olson, M., Hood, L., Cantor, C. and Botstein, D. (1989). A common language for physical mapping of the human genome. Science 245:1434-1453.

Opipari, A. W., Boguski, M. S. and Dixit, V. M. (1990). The *A20* cDNA induced by tumor necrosis factor α encodes a novel type of zinc finger protein. J. Biol Chem. 265:14705-14708.

Orkin, S. H., Goldman, D. S. and Sallan, S. E. (1984). Development of homozygosity for chromosome 11p markers in Wilms' tumour. Nature 309:172-174.

Pabo, C. O. and Sauer, S. T. (1984). Protein-DNA recognition. Ann. Rev. Biochem. 53:293-321.

Page, D. C., Mosher, R., Simpson, E. M., Fisher, E. M. C., Mardon, G., Pollack, J., McGillivray, B., de la Chapelle, A. and Brown, L. G. (1987). The sex determining region of the human Y chromosome encodes a finger protein. Cell 51:1091-1104.

Parkhurst, S. M., Harrison, D. A., Remington, M. P., Spana, C., Kelley, R. L., Coyne, R. S. and Corces, V. C. (1988). The *Drosophila su(Hw)* gene, which controls the phenotypic effect of the gypsy transposable element, encodes a putative DNA binding protein. Genes Dev. 2:1205-1215.

Pays, E. and Murphy, N. B. (1987). DNA-binding fingers encoded by a Trypanosome retroposon. J. Mol. Biol. 197:147-148.

Pelham, H. R. B. and Brown, D. D. (1980). A specific transcription factor that can bind either the 5s RNA gene or 5S RNA. PNAS. 77:4170-4174.

Petkovich, M., Brand, N. J., Krust, A. and Chambon, P. (1987). A human retenoic acid receptor which belongs to the family of nuclear receptors. Nature 330:444-450.

Pinkel, D., Straume, T. and Gray, J. T. (1986). Cytogenetic analysis using quantitative, high sensitivity, fluorescence hybridisation. PNAS 83:2934-2938.

Poole, S. J., Law, M. L., Kao, F-T. and Lau, Y-F. (1989). Isolation and chromosomal localisation of the human *En-2* gene. Genomics 4:225-231.

Pontecorvo, G. (1975). Production of mammalian somatic cell hybrids by means of polyehylene glycol treatment. Somat. Cell. Genet. 1:397-400.

Porteous, D. J.,Morten, J. E. N., Cranston, G., Fletcher, J. M., Mitchell, A., van Heyningen, V., Fantes, J. A., Boyd, P. A. and Hastie, N. D. (1986). Molecular and physical arrangements of human DNA in *HRAS1*-selected, chromosome-mediated transfectants. Mol. Cell. Biol. 6:2223-2232.

Porteous, D., Wilkinson, M., Fletcher, J., and van Heyningen, V. (1989). Human-mouse hybrids carrying fragments of single human chromosomes selected by tumour growth. Genomics 5:680-684.

Prendergast, G. C. and Ziff, E. B. (1989). DNA-binding motif. Nature 341:392.

Pritchard, C. A. and Goodfellow, P. N. (1987). Investigation of chromosome-mediated gene transfer using the HPRT region of the human X chromosome as a model. Genes Dev. 1:172-178.

Pritchard-Jones, K., Fleming, S., Davidson, D., Bickmore, W., Porteous, D., Gosden, C., Bard, J., Buckler, A., Pelletier, J., Housman, D., van Heyningen, V. and Hastie, N. (1990). The candidate Wilms' tumour gene is involved in genitourinary development. Nature 346:194-197.

Rauscher III, F, J., Morris, J. F., Tournay, O. E., Cook, D. M. and Curran, T. (1990). Binding of the Wilms' tumor locus zinc finger gene to the EGR-1 consensus sequence. Science 250:1259-1262.

Redemann, N., Gaul, U. and Jackle, H. (1988). Disruption of putative Cys-zinc interaction eliminates the biological activity of the *Krüppel* finger protein. Nature 332:90-92.

Reeve, A. E., Housiaux, P. J., Gardener, R. J., Chewing, W. E., Grindley, R. M. and Millow, L. J. (1984). Loss of Harvey *ras* allele in sporadic Wilms' tumour. Nature 309:174-176.

Reeve, A. E., Sih, S. A., Raizes, A. M. and Feinberg, A. P. (1989). Loss of allelic heterozygosity at a second locus on chromosome 11 in sporadic Wilms' tumor cells. Mol. Cell. Biol. 9:1799-1803.

Reuter, G., Giarre, M., Farah, J. Gausz, J., Spierer, A. and Spierer, P. (1990). Dependance of position-effect variegation in *Drosophila* on dose of a gene encoding an unusual zinc finger protein. Nature 344:219-223.

Rhodes, D. and Klug, A. (1988). "Zinc Fingers": A novel motif for nucleic acid binding. In "Nucleic Acids and Molecular Biology, Vol. 2" (ed. by Eckstein and Lilley) pp149-166. Springer-Verlag.

Riccardi, V. M., Sujansky, E., Smith, A. C. and Francke, U. (1978). Chromosomal imbalance in the aniridia-Wilm's tumor association: 11p interstitial deletion. Pediatrics 61:604-610.

Riccardi, V. M., Hittner, H. M., Smith, A. C. Francke, U., Yunis, J. J., Ledbetter, D. and Borges (1980). The aniridia-Wilm's tumour association: The critical role of chromosome band 11p13. Cancer Genet. Cytogenet. 2:131-137.

Richardson, C.C. (1981). "Bacteriophage T4 polynucleotide kinase". In The Enzymes, 3rd Edition (ed. P.D. Boyer), vol. 14, p. 299. Academic Press, New York.

Riordan, J. R., Rommens, J. M., Kerem, B-S, Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., Chou, J-L., Drumm, M. L., Iannuzzi, M. C., Collins, F. C. and Tsui, L-C. (1989). Identification of the cystic fibrosis gene: cloning and characterisation of complementary DNA. Science 245:1066-1073.

Rose, E. A., Glaser, T., Jones, C., Smith, C. L., Lewis, W. H., Call, K. M., Minden, M., Champagne, E., Lewis, W., Bonetta, L., Yeger, H. and Housman, D. E. (1990). Complete physical map of the WAGR region of 11p13 localises a candidate Wilms' tumor gene. Cell 60:495-508.

Rosenberg, U.B., Schroeder, C., Priess, A., Kienlin, A., Cote, S., Reide, I. and Jackle, H. (1986). Stuctural homology of the product of the *Drosophila Krüppel* gene with *Xenopus* transcription factor TFIIIA. Nature 319: 336-339.

Rothe, M., Nauber, U. and Jackle, H. (1989) Three hormone receptor-like *Drosophila* genes encode an identicle DNA binding finger. EMBO J. 8:3087-3094.

Royer-Pokara, B., Kunkel, L. M., Monaco, A. P., Goff, S. C., Newburger, P. E., Baehner, R. L., Cole, F. S., Curnatte, J. T. and Orkin, S. H. (1984). Cloning the gene for an inherited human disorder - chronic granulomatous disease - on the basis of its chromosomal location. Nature 322:32-38.

Ruiz i Altaba, A., Perry-O'Keefe, H. and Melton, D.A. (1987). *Xfin*: an embryonic gene encoding a multifingered protein in *Xenopus*. EMBO J. 6: 3065-3070.

Ruppert, J. M., Kinzler, K. W., Wong, A. J., Bigner, S H., Kao, F-T., Law, M. L., Seuanez, H. N., O'Brien, S. J. and Vogelstein, B. (1988). The *GLI*-Kruppel family of human genes. Mol. Cell. Biol. 8:3104-3113.

Russel, M., Kidd, S. and Kelley, M. R. (1986). An improved filamentous helper phage for generating single-stranded plasmid DNA. Gene 45: 333-338.

Sabbah, M., Redeuilh, G., Secco, C. and Baulieu, E-E. (1987). The binding activity of the estrogen receptor for DNA and heat shock protein (Mr 90,000) is dependant on receptor-bound metal. JBC 262:8631-8635.

Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989). "Molecular cloning: A Laboratory Manual". 2nd edition Cold Spring Harbour Laboratory, Cold Spring Harbour, NY.

Sargent, C. A., Dunham, I. and Cambell, R. D. (1989). Identification of multiple HTF-island associated in the human major histocompatability complex class III region. EMBO J. 8:2305-2312.

Sargent, M. G. and Bennett, M. F. (1990). Identification in *Xenopus* of a structural homologue of the *Drosophila* gene *snail*. Dev. 109:967-973.

Schnittger, S., Hoefers, C., Breermann, F., Heidemann, and Hansmann, I. (1989). Alagille-Watson Syndrome is assigned to 20(p1.1-p1.2) and provisionally to the region p11.23-p12.1. Cytogenet. Cell. Genet. 51:1074.

Shih, C. and Weinberg, R. A. (1982). Isolation of a transforming sequence from a human bladder carcinoma cell line. Cell 29:161-169.

Sauer, R. T., Smith D. L. and Johnson, A. D. (1988). Flexibility of the yeast α2 repressor enables it to occupy the ends of its operator, leaving the centre free. Genes Dev. 2:807-816.

Schuh, R., Aicher, W., Gaul, U., Cote, S., Preiss, A., Maier, D., Seifert, E., Nauber, U., Schroder, C., Kemler, R. and Jackle, H. (1986). A conserved family of nuclear proteins containing structural elements of the finger protein encided by *Krüppel*, a *Drosophila* segmentation gene. Cell 47:1025-1032.

Schwartz, D. C. and Cantor, C. R. (1984). Separation of yeast chromosome-sized DNAs by pulse field gradient gel elecrophoresis. Cell 37:67-75.

Sealey, P. G., Whittaker, P. A. and Southern, E. M. (1985). Removal of repeated sequences from hybridisation probes. Nucl. Acid Res. 13:1905-1922.

Shirsat, N. V., Pignolo, R. J., Kreider, B. L. and Rovera, G. (1990). A member of the *ras* gene superfamily is expressed specifically in T, B, and myeloid hemopoietic cells. Oncogene 5:769-772.

Short, J. M., Fernandez, J. M., Sorge, J. A. and Huse, W. D. (1988). A bacteriophage lambda expression vector with *in vivo* excision properties. Nuc. Acids Res. 16:7583-7600.

Silver, L. M. (1985). Mouse t-haplotypes. Ann, Rev. Genet. 19:179-208.

Sortelo-Avila, C. and Gooch, W. M. (1976). Neoplasms associated with the Beckwith-Wiedemann syndrome. Perspect. Pediatr. Pathol. 3:255-272.

Spana, C., Harrison, D. A. and Corces, V. G. (1988). The *Drosophila melanogaster supressor of Hairy-wing* protein binds to specific sequences of the gypsy retrotransposition. Genes Dev. 2:1414-1423.

Spanopoulou, E., Giguere, V. and Grosfeld, F. (1988). Transcriptional unit of the murine *Thy-1* gene: different distribution of transcription initiation sites in brain. Mol. Cell. Biol. 8: 3847-3856.

Spies, T., Blanck, G., Bresnahan, M., Sands, J. and Strominger, J. L. (1989). A new cluster of genes within the human major histocompatability complex. Science 243:214-217.

Stanojevic, D., Hoey, T. and Levine, M. (1989). Sequence-specific DNA-binding activities of the gap proteins encoded by *hunchback* and *Krüppel* in *Drosophila*. Nature 341:331-335.

Stephens, J. C., Cavahaugh, M. L., Gradie, M. I., Mador, M. l. and Kidd, K. (1990). Mapping the human genome: Current status. Science 250:237-244.

Stillman, D. J., Bankier, A.T., Seddon, A., Groenhout, E.G. and Nasmyth, K.A. (1988). Characterisation of a transcription factor involved in mother cell specific transcription of the yeast *HO* gene. EMBO J. 7: 485-494.

Sulston, J., Mallet, F., Staden, R., Durbin, R., Horsnell, T. and Coulson, A. (1988). Software for genome mapping by finger printing techniques. Cabios 4:125-132.

Sukhatme, V. P., Cao, X., Chang, L. C., Tsai-Morris, C., Stamenkovich, D., Ferreira, P. C. P., Cohen, D. R., Edwards, S. A., Shows, T. B., Curran, T., Le Beau, M. M. and Adamson, E. D. (1988). A zinc finger-encoding gene coregulated with c-*fos* during growth and differentiation, and after cellular depolarisation. Cell 53: 37-43.

Sved, J. and Bird, A. (1990). The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. PNAS 87:4692-4696.

Szybalski, W., Szybalska, E. H. and Ragni, G. (1962). Genetic studies with human cell lines. Nat. Cancer Inst. Monogr. 7:75-89.

Tautz, D., Lehmann, R., Schnurch, H., Schuh, R., Seifert, E., Kienlin, A., Jones, K. and Jackle, H. (1987). Finger protein of novel structure encoded by *hunchback*, a second member of the gap class of *Drosophila* segmentation genes. Nature 327: 383-389.

Thiesen, H-J. (1990). Multiple genes encoding zinc finger domains are expressed in human T-cells. New Biol. 2:363-374.

Treisman, J. and Desplan, C. (1989). The products of the *Drosophila* gap genes *hunchback* and *Krüppel* bind to the *hunchback* promoters. Nature 341:335-337.

Treier, M., Pfeifle, C. and Tautz, D. (1989). Comparison of the gap segmentation gene *hunchback* between *Drosophila melanogaster* and *Drosophila virilis* reveals novel modes of evolutionary change. EMBO J. 8:1517-1525.

Trent, J. M., Kaneko, Y. and Mitelman, F. (1989). Report of the comittee on structural chromosome changes in neoplasia. Cytogenet. Cell. Genet. 51:533-562.

Tunnacliffe, A., Jones, C. and Goodfellow, P. (1983). Somatic cell genetics, immunogenetics and gene mapping. Immuno. Today 4:230-233.

Urieli-Shoval, S., Gruenbaum, Y., Sedaty, J. and Razin, A. (1982). The absense of detectable methylated bases in *Drosophila melanogaster* DNA. FEBS Lett. 146:148-152.

van Heyningen, V., Bickmore, W. A., Seawright, A., Fletcher, J. M., Maule, J., Fekete, G., Gessler, M., Bruns, G. A. P., Huerre-Jeanpierre, C., Juniens, C., Williams, B. R. G. and Hastie, N. D. (1990). Role for the Wilms' tumor gene in genital development? PNAS 87:5383-5386.

Viera, J. and Messing, J. (1982). The pUC plasmid, an M13mp7-derived system for insertional mutagenisis and sequencing with synthetic universal primers. Gene 19:259-268.

Vincent, A., Colot, H. V. and Rosbash, M. (1985). Sequence and structure of the *serendipity* locus of *Drosophila melanogaster*: a densely transcribed region including a blastoderm-specific gene. J. Mol. Biol. 186:146-166.

Viskochil, D., Buchberg, A. M., Xu, G., Cawthon, R. M., Stevens, J., Wolff, R. K., Culver, M., Carey, J. C., Copeland, N. G., Jenkins, N. A., White, R. and O'Connell, P. (1990). Deletions and a translocation interrupt a cloned gene at the neurofibromatosis type 1 locus. Cell 62:187-192, 193-201.

Vrana, K. E., Churchill, M. E. A., Tullius, T. D. and Brown, D. D. (1988). Mapping functional regions of the transcription factor TFIIIA. Mol. Cell. Biol. 8:1684-1696.

Wallace, M. R., Marchuk, D. A., Anderson, L. B., Letcher, R., Odeh, H. M., Saulino, A. M., Fountain, J. W., Brereton, A., Nicholson, J., Mitchell, A. L., Brownstein, B. H. and Collins, F, S. (1990). Type 1 neurofibromatosis gene: Identification of a large transcript disrupted in three NF1 patients. Science 249: 181-186.

Waziri, M., Patil, S. R., Hanson J. W. and Bartley, J. A. (1983). Abnormality of chromosome 11 in patients with features of Beckwith-Wiedemann syndrome. J. Pediatr. 102:873-876.

Weiss, M. and Green, H. (1967). Human-mouse hybrid cell lines containing partial complements of human chromosomes and functioning human genes. PNAS 58:1104-1111.

Weissman, B. E., Saxon, P. J., Pasquale, S. R., Jones, G. R., Geiser, A. G. and Stanbridge, E. J. (1987). Introduction of a normal human chromosome 11 into Wilm's tumour cell line controls its tumourigenic expression. Science 236:175-236.

Wiedemann, H. R. (1964). Complexe malformatif familial avec hernie ombilicale et macroglossia - un syndrome nouveau? J. Genet. Hum. 13:223-232.

Wood, T. G., McGeady, M. L., Baroudy, B. M., Blair, D. G. and Vande Woude, G. F. (1984). Mouse c-*mos* oncogene activation is prevented by upstream sequences. PNAS 81:7817-7821.

Wu, J., Grindlay, J., Bushel, P., Mendelsohn, L. and Allan, M. (1989). Negative regulation of the human epsilon-globin gene by transcriptional interference: Role of an Alu repeatative element. Mol. Cell. Biol. 10:1209-119.

Xu, G., O'Connell, P., Viskochil, D., Cawthon, R., Robertson, M., Culver, M., Dunn, D., Stevens, J., Gesteland, R., White, R. and Weiss, R. (1990). The neurofibromatosis type 1 gene encodes aprotein related to GAP. Cell 62:599-608.

Yanisch-Perron, C., Vieira, J. and Messing, J. (1985). Improved M13 phage cloning vectors and host strains: Nucleotide sequences of the M13mp18 and pUC19 vectors. Gene 33:103-119.

Yisraeli, J and Szyf, M. (1984). Gene methylation patterns and expression. In "DNA Methylation: Biochemistry and Biological Significance" (Razin, A., Cedar, H. and Riggs, A.G.., eds), pp353-358, Springer-Verlag.

Youssoufian, H., Kazazian Jr., H. H., Phillips, D. G., Aronis, S., Tsiftis, G., Brown, V. A. and Antonarakis, S. E. (1986). Recurrent mutations in haemophilia A give evidence for CpG mutation hotspots. Nature 342:380-382.

Zelent, A., Krust, A., Petkovich, M., Kastner, P. and Chambon, P. (1989). Cloning of murine α and β retenoic acid receptors and a novel receptor γ, predominantly expressed in skin. Nature 339:714-717.

Appendix-Published Material.

GENOMICS 5, 501–509 (1989)

A Cosmid Clone Map Derived from a Small Region of Human Chromosome 11

K. J. HARRISON-LAVOIE, R. M. JOHN, D. J. PORTEOUS,* AND P. F. R. LITTLE

*Department of Biochemistry, Imperial College, London SW7 2AZ, and *M. R. C. Human Genetics Unit, Western General Hospital, Edinburgh EH4, United Kingdom*

Received March 2, 1989; revised May 24, 1989

We have used cosmid "fingerprinting" to construct an overlapping DNA clone "map" of the human DNA in a mouse/human hybrid cell line, E65-9, that contains about 4×10^6 bp, including the H-*Ras* gene, as its human component. We have additionally used ^{32}P-labeled RNA probes to establish linkage of particular sets of clones, and the final map comprises about 300,000 bp and is contained in three nonoverlapping segments. The reasons for failure to close the gaps by direct probing are discussed. We have developed techniques to search for rare cutting restriction enzyme cleavage sites in large numbers of cloned DNAs and have positioned sites for *Eag*I and *Bss*HII on our clone map. The methods we used are capable of considerable scale-up and are currently being applied to the short arm of human chromosome 11.

INTRODUCTION

Many problems in mammalian molecular genetics involve the identification of genes within small regions of DNA flanked by appropriate markers. The general problem of isolating such small regions free from the bulk of the genome can sometimes be overcome by using the techniques of somatic cell hybridization. Porteous *et al.* (1986) have previously reported the isolation of somatic cell hybrids that contain portions of the short arm of human chromosome 11, in a mouse cell background. In this paper we report the analysis of one of these cell lines, E65-9, that contains about 2–3 $\times 10^6$ bp of human DNA, including the H-*Ras* gene. The cell line was made by chromosome-mediated gene transfer (CMGT) of an activated human H-*Ras* gene into nontransformed mouse cells. The transgenome in such a cell line makes an ideal small target for detailed DNA analysis.

The ultimate aim of programs to construct detailed maps of the genome of humans and other species is to facilitate the isolation and identification of genes and DNA sequences from defined regions of the genome. Three types of physical "map" are being constructed; DNA sequence, overlapping DNA clones, and restriction sites. The most attractive feature of the cloned DNA map is that it not only provides positional information (the physical relationship of one DNA clone to another) but it also generates immediately usable cloned DNAs for further work. For this single reason we argue that the construction of overlapping cloned DNA maps of genomes is inherently attractive.

Several groups have reported on the technical methods by which this may be achieved for genomes the size of *Escherichia coli* (Kohara *et al.*, 1987), yeast (Olson *et al.*, 1986), and the nematode, *Caenorahbditis elegans* (Coulson *et al.*, 1986). Each method is different in detail but all three have an underlying similarity. Random cloned DNAs are characterized by restriction enzyme fragment patterns or site maps and computerized matching routines are then used to compare individual clone patterns against each other. Statistically significant similarity of pattern is taken to imply that the clones contain overlapping DNA sequences. These techniques allow the construction of large arrays of overlapping cloned DNA sequences, known as "contigs"; as the project proceeds the contigs increase in size and decrease in numbers. Using only random clone analysis strategies, it is not possible to construct a finished map of a large region of DNA and alternative methods of analysis, that are more directed, are required toward the end of a mapping project of this type (Coulson *et al.*, 1988).

There are three candidate cloning vectors for clone mapping projects: phage λ, cosmid, and yeast artificial chromosome (YAC) (Burke *et al.*, 1987). The last vector has the disadvantage, compared to the first two systems, that there is no general and simple method to isolate convenient amounts of pure recombinant DNA from large numbers of YAC-containing clones. Both cosmids and λ have simple methods for DNA isolation.

0888-7543/89 $3.00

Cosmids have 2.5- to 3-fold greater capacity for insert DNA than λ vectors and this has made them appealing analytic tools that have been extensively used for the *C. elegans* mapping project (Coulson *et al.*, 1986).

Some cosmid recombinants are unstable and prone to rearrangement. More critically, cosmids exhibit nonrandom cloning statistics, a feature they share with λ vector systems. We have previously reported on the construction of cosmid vectors that are based on the phage λ origin of replication (Little and Cross, 1985) and that contain *E. coli* RNA polymerase terminators flanking the cloning sites. We have shown that these have enhanced stability by reduced transcriptional interference of the origin of DNA replication (Gibson *et al.*, 1987). We accept that there are certain DNA sequences that are likely to be difficult or impossible to maintain in *E. coli* at high copy number. "Gaps" between contigs may indicate the presence of such sequences or may be due to chance events. In either case such gaps could be bridged by utilizing other physical techniques to establish order and linkage (Coulson *et al.*, 1988).

We have chosen to use cosmid fingerprinting to analyze the human DNA in the cell line E65-9. In outline our method of analysis comprises the following steps: the construction of cosmid libraries from E65-9 DNA, the isolation of human DNA clones by hybridization of clones with total human DNA, and the use of the fingerprinting technique of Coulson *et al.* (1986) to identify cosmid overlaps. This involves cutting of individual cosmid DNAs with *Hin*dIII, labeling of the DNA ends with reverse transcriptase, cleavage with *Sau*3AI, acrylamide gel electrophoresis, and detection of labeled fragments by autoradiography. Fragment patterns are then analyzed by a computerized matching procedure, detailed in Coulson *et al.* (1986) and Sulston *et al.* (1988), to identify overlapping DNA sequences and enable construction of contigs. In this project we also screen the cosmids for *Bss*HII and *Eag*I (*Xma*III) sites by a simple labeling technique. These sites have been reported by Bird (1986) to occur commonly in CpG islands associated with a particular class of genes and can be used to identify potential gene-containing clones. Individual clones are used in direct probing experiments to identify cosmid clones that do not overlap to sufficient extent to allow linkage in the contig analyses. We make use of SP6 and T7 RNA polymerase-specific promoters contained on the Lorist cosmid vectors (Cross and Little, 1986), to generate ^{32}P-labeled RNA probes that are specific to the ends of cloned human inserts. The methods that we use are simple, reproducible, and applicable to significantly larger projects.

Our analyses are designed to test the feasibility of scale-up to cell lines that contain a substantial portion of the short arm of chromosome 11 and also establish the structure of the transgenome with a longer term goal of increasing our understanding of the experimental behavior of CMGT.

MATERIALS AND METHODS

Cosmid Library Construction

Libraries were constructed in the λ origin cosmid Lorist 2 using *Hin*dIII partial digestion of E65-9 DNA and inserting this into the *Hin*dIII site of the vector. All procedures for DNA isolation, partial digestion, size fractionation, ligations, and library plating were as detailed in Little (1987). Subsequent to the library construction reported here, we have found that sucrose gradient purification of 40-kb insert DNA is not satisfactory. Significant amounts of smaller DNA cosediments and we detect artifactual ligation of nonadjacent DNA fragments. We find that phosphatasing of partial digest DNA is more satisfactory. The choice of cloning enzyme was made because it facilitates identification of short overlapping DNA regions in cosmids. The fingerprint of any two cosmids that overlap must contain at least two *Hin*dIII *Sau*3AI fragments in common. This is not the case for *Sau*3AI partials inserted into a *Bam*HI site. We also find it very useful to be able to observe all the human DNA fragments by single cleavage with *Hin*dIII and know that any pair of overlapping clones must have at least one fragment of identical size. We discuss reservations to this strategy below.

Cosmid DNA Preparation

Cosmid DNA was prepared by two different protocols. For isolation of cosmid DNA for restriction analyses and some Southern blotting we used the "miniprep" procedure of Little (1987), isolating DNA from 1.5 ml of saturated culture and resuspending in a final volume of 50 μl.

For fingerprint and some other procedures we used the "microprep" protocol isolating DNA from 0.25 ml saturated cultures following the method of Gibson and Sulston (1987). In outline, cell pellets were lysed by a modified alkaline–SDS procedure followed by lithium chloride precipitation of debris and DNA precipitation with propan-2-ol. DNA pellets were resuspended in 6–10 μl to give a final concentration of about 100 μg/ml. DNA made by this procedure is suitable for a wide variety of analyses and this procedure allows the preparation of DNA in multiple-well trays, making extensive use of multitip and automatic micropipettors to facilitate manipulations. Routinely, it is possible for a single person to make 400 DNA samples at one time.

Contig Construction

Fingerprint procedures. Cosmid fingerprints were made as described in Coulson *et al.* (1986) and Coulson and Sulston (1988). Cosmid microprep DNA (1.5 to 2

μl) was cut with *Hin*dIII, labeled with [^{32}P]dATP by reverse transcriptase, and cleaved with *Sau*3AI. The final reaction volume was 10 μl. Of this reaction mix, 3 μl was run on a 33-cm-wide 4% acrylamide/urea gel at 74 W constant power for 80–90 min. Running buffer was as described in Coulson and Sulston (1988). Marker DNA was λ CI857 cut with *Sau*3AI and labeled with [^{35}S]dATP using reverse transcriptase.

Visual inspection of gels is difficult if gels "smile" or suffer from other running artifacts. We have found the BRL Model S2 gel sequencing apparatus is satisfactory. It is also important that gels are bonded to one of the glass support plates and taken through fixation, drying down, and exposure to X-ray film in this state (Coulson and Sulston, 1988).

Analysis system. Fragment patterns on the autoradiographs were automatically digitized using the scanning and digitizing system of Sulston *et al.* (1988, 1989). Contigs were assembled using the set of programs outlined in Coulson *et al.* (1987) and Sulston *et al.* (1988), running on a Vax 8600 computer. All contigs were assembled using interactive graphics assembly routines contained in the suite of programs and visual observation of autoradiographs.

Rare Cut Sites

We have used two procedures to identify *Eag*I and *Bss*HII sites in cosmids. The first method used followed the same principles as those in fingerprinting (Coulson *et al.*, 1986) except that a rare cutting enzyme (*Bss*HII or *Eag*I) was substituted for *Hin*dIII in the first reaction and *Hin*dIII used in the second reaction instead of *Sau*3AI. A reaction mix for a typical experiment contained 30 μCi of [^{32}P]dATP, 2 μl of 10 × medium (*Bss*HII)– or very high-salt (*Eag*I) restriction endonuclease buffer (Maniatis *et al.*, 1982), 30 μg of boiled RNAase, 25 μ*M* ddGTP, 60 units *Bss*HII or *Eag*I endonuclease and 60 units of reverse transcriptase in a final volume of 120 μl. Of this mixture, 2 μl was placed in each well of a 96-well Corning Cell Well plate using a repetitive dispenser and 1.5 μl of cosmid clone DNA, prepared by the microprep method, was added. Incubation was for 45 min at 50°C (*Bss*HII) or 37°C (*Eag*I). To prevent evaporation, the wells were sealed with Micronic 8 cap bands. The reverse transcriptase was inactivated at 68°C for 30 min and after cooling on ice, 2 units of *Hin*dIII in 2 μl of medium (*Bss*HII)– or no-salt (*Eag*I) restriction buffer was added (to adjust the final salt concentration to medium–salt restriction buffer in both cases) and the samples were incubated at 37°C for 120 min. The digests were run on 1% agarose gels at 10 V/cm for 60 min and dried down under vacuum before being autoradiographed for 20–40 min at −70°C.

The second method required only a single digest with the appropriate enzyme of 10 μl of a microprep clone (equivalent to about 0.5 μg of DNA). The DNA was analysed by FIGE (Carle *et al.*, 1986). The digest was run on a 1% agarose gel under an inverting field of 10 V/cm, 1.25 s forward and 0.4 s reversed, for 6–8 h, and the DNA was visualized by standard procedures.

Where it was necessary the presence or absence of an *Eag*I or *Bss*HII site was confirmed by double-digest analysis, with *Hin*dIII, of purified cosmid DNA.

Probing

Human cosmids were identified in the cosmid library by hybridization to ^{32}P-labeled total human DNA, sheared to about 400 bp by sonication, using oligolabeling (Feinberg and Vogelstein, 1983) to generate specific activities of $>10^8$/μg. Similarly labeled mouse DNA was used in control hybridizations.

T7 RNA polymerase-generated probes were made from cosmid miniprep DNA as described in Little (1987). DNA from an entire miniprep was spermidine-precipitated by addition of spermidine hydrochloride to 10 m*M*, left on ice for 15 min, spun down, and resuspended in 10 μl of 0.5*M* NaCl at room temperature. Distilled water (100 μl) was added and 75 μl of this used in a transcription reaction. We find this step to be essential for efficient incorporation of [^{32}P]UTP. We use 50 μCi of label per reaction and achieve 30–90% incorporation of label into RNA. We found it impossible to generate satisfactory SP6 probes from miniprep DNA, even after spermidine precipitation, and used 20 μg of cesium chloride-purified DNA for this. T7 synthesis is always more efficient than SP6 and we use the former enzyme where possible.

Probe link-up experiments were carried out on DNA samples immobilized on Pall Biodyne filters using the manufacturer's recommended conditions and a BRL "dot blot" manifold to position samples. One to two microliters of either mini- or micro-prep DNA was immobilized per spot. Hybridization conditions were as described in Little (1987).

RESULTS

Library Screen

In preliminary experiments we screened the E65-9 library at high density (>10,000 colonies per filter) for human clones, using the protocols described under Materials and Methods, and detected 0.043% as first round positives. E65-9, its parent mouse cell C127 and all CMGT transformants derived from it, are subtetraploid, with a modal chromosome number of 85, equivalent to about 10^{10} bp of DNA. We would expect E65-9 to contain about 4.3×10^6 bp of human DNA which could be contained in about 100 cosmids. However, three lines of evidence, taken with the mapping data we present below, suggest that the human DNA

in E65-9 is diploid, with each block consisting of four to five repeats of near identical DNA containing the H-*Ras* gene and flanking sequences. The transgenome in E65-9 cannot be detected by *in situ* hybridization. Gosden and Porteous (1987) have shown that it is possible to detect 5×10^6 bp of DNA in single block. The dosage of H-*Ras* in E65-9 is equivalent to that of other CMGT lines for which Porteous *et al.* (1986) and Gosden and Porteous (1987) have direct evidence of transgenome duplication. Finally the L1 "fingerprint" suggests only about 2×10^6 bp of human DNA in E65-9 (Porteous *et al.*, 1986).

We used two different methods to collect human clones free of contaminating mouse colonies. Initially we screened at high density, picked positives from single filters into pools, and rescreened each pool at low density. These enriched pools gave 4% positives that were easy to pick directly. This protocol was not successful since we encountered clones that had become dominant in each pool. We found that it was more efficient, although wasteful of filters, to screen at colony densities that would allow direct picking of pure colonies. This becomes more attractive with cell lines that yield higher frequencies of human positives.

We were able to isolate 112 clones that were thought to contain human DNA by using these methods and these were subjected to further analyses. Using the random sampling statistic of Clarke and Carbon (1976), we can calculate that based upon a target size of about 2×10^6 bp, 112 clones represents 90% coverage.

Fingerprint Analysis

The clones were all analyzed by the fingerprinting technique of Coulson *et al.* (1986) and the computerized analysis of probability of overlap and visual inspection of autoradiographs was used to construct contigs. The data initially were reduced from a substantial number of small contigs to 6 larger contigs with 43 unattached clones (that is, clones not contained in contigs). A surprising feature of the analysis was that one contig contained 42 cosmids with a high degree of overlap. The random number generated prediction (Sulston *et al.*, 1988, program Rancl5) for this analysis, based upon a target genome of 2×10^6 bp, would be 20 contigs of mean occupancy 3–4. We believe that this occurs because the human transgenome in E65-9 contains several identical or near identical repeats of DNA, resulting in clones from each repeat mapping into the same contig. The same random number prediction for a genome size of 500,000 bp is 4 contigs with mean occupancy of 21 and for 400,000 bp, 2 contigs of 42 occupancy. This latter case is equivalent to 10 copies of the human DNA per tetraploid set, or 2 five-fold amplified blocks. We discuss this problem below.

Probing–Hybridization Link-up

We used the SP6 or T7 promoters in the cosmid vector to generate RNA probes from specific cosmids. Cross and Little (1986) have shown that the probes are specific to the ends of the DNA inserted into the vector and their use in hybridization experiments allows the identification of clones that overlap by only a small amount of DNA. These would not be detected by the fingerprint analyses. We have confirmed this using selected cosmids, of known location and relationship, cut with *Hin*dIII and analyzed by Southern blotting (data not presented).

It is labor intensive to carry out Southern analyses on large clone arrays so we used a dot blot procedure to simplify analysis. Figure 1 shows the results of one experiment. The first probe was made from the T7 promoter of clone C445 located at the end of contig 8. There is strong self-hybridization as well as hybridization to clones C484, C468, C442, C483, C437, and C444. Clones C484 and C468 are contained in contig 7: clones C442, C483, C437 are all buried in the probe clone. C444 is unattached. C471 is a weak false positive and C427 is a mouse/human hybrid clone probably generated by a ligation artifact. In a separate experiment (Fig. 1B) we used an SP6 transcript of contig 6 clone C464 which self-hybridizes and also detects clones C444 and a further unattached clone. This implies that C444 bridges contigs 4 and 6. The reciprocal experiment, probing with C444, detects the appropriate cosmids (data not presented). The fingerprint data on C444 shows that it has nine bands, two of which overlap with C445 in contig 6 and two with C464 in contig 4. The analytical probability (0.1 and 0.05 respectively) of this overlap was too high for it to be identified by fingerprint data alone. The two clones in contig 7 also match into the overlap of contigs 4 and 6 but we have not entered these onto the computer printout due to clone instability problems. Experiments identical in construction to these enabled us to link several other clones into contigs. The hybridization data, along with a map that incorporates the final structure of the contigs, are presented in Fig. 2. In some cases (clone A421 in contig A, for example) the patterns are not logically consistent and we would move this clone to the right such that it no longer overlaps with A419 and C402. The fingerprint alignment is sometimes inaccurate to the extent of a couple of fragments in the band pattern.

We would stress that in most projects it would be neither possible nor necessary to probe as frequently as we have; our primary concern was to see if we could establish a contig structure for the whole of the human DNA in E65-9 rather than a consensus of the repeat unit. This led us to carry out more probings than would normally be required when mapping a less complex

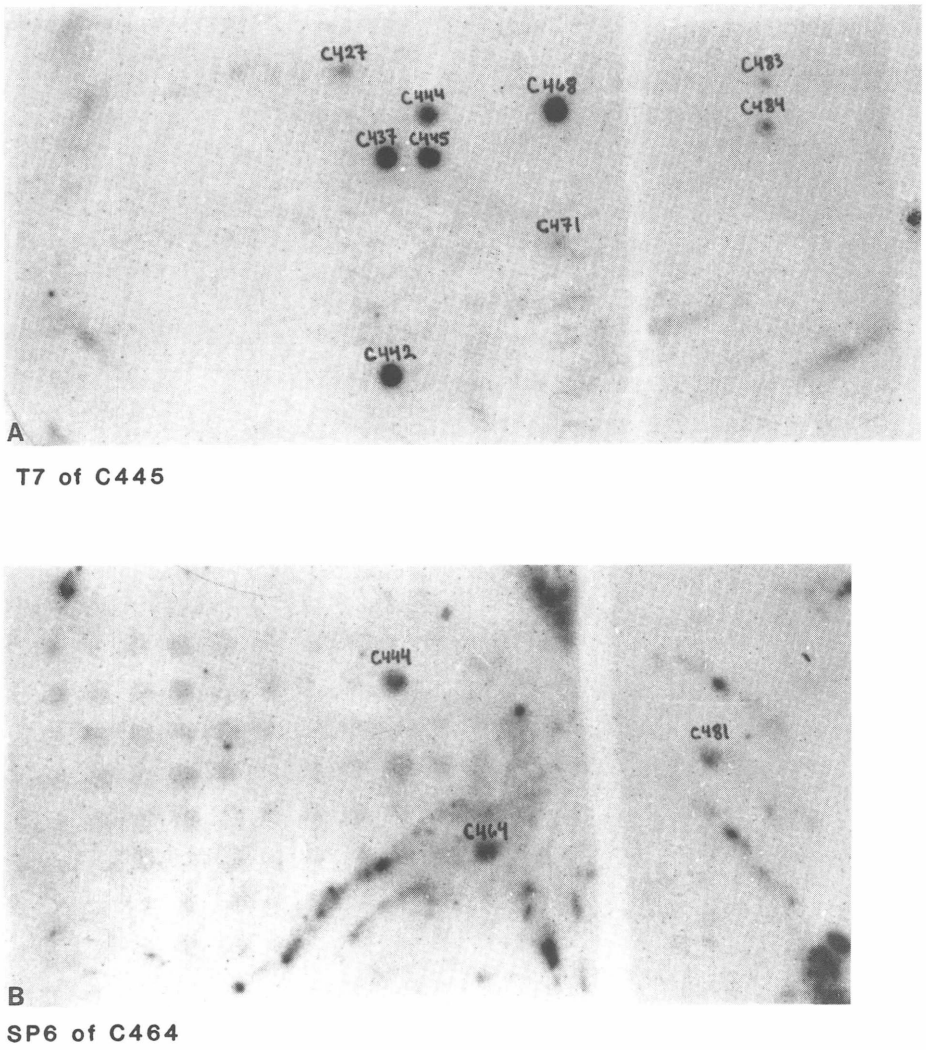


FIG. 1. Probing of the human cosmids from E65-9 with (**A**) a T7 RNA polymerase-generated transcript of cosmid C445 and (**B**) an SP6 RNA polymerase-generated transcript of cosmid C464. Note that both probes hybridize to clone C444, allowing the link-up of two contigs that had not previously been linked in the fingerprint analysis. Map locations of all three cosmids are given in Fig. 2.

transgenome. The final map structure is discussed below.

The major technical problem to this approach is the presence of repetitive sequences in the DNA immediately adjacent to the promoter. We have estimated (Cross and Little, 1986) that 1 in 6 to 1 in 10 probes contain highly repeated sequences. We use dot blots of total human DNA as controls for this but in many cases the internal logic of the known location of a clone with respect to other members of the contig identifies anomalous cross-hybrization.

Isolation of the H-Ras Gene

Extensive hybridization analysis of the 112 clones in the database using the 6.6-kb *Bam*HI fragment that contains the whole of the H-*Ras* gene (Goldfarb *et al.*, 1982) failed to identify a candidate clone. Similar screens of the *Hin*dIII E65-9 library also failed. We have isolated H-*Ras* from a normal human *Sau*3AI partial digest cosmid library on two overlapping clones and have shown by restriction analysis that it is contained on a *Hin*dIII fragment of 40–45 kb (data not presented); this would be too large to be cloned in a *Hin*dIII partial digest library. However, the H-*Ras* gene in E65-9 is rearranged and contained on a 12-kb *Hin*dIII fragment (data not presented). We screened 2.2×10^5 clones from a *Sau*3AI partial digest library of E65-9 DNA, cloned into LoristX, and isolated a single positive clone containing part of this fragment. Fingerprint analysis shows that the clone is located on the end of contig B. Preliminary restriction mapping (data not presented) indicates that we have cloned the 5′ half of the gene with an apparent rearrangement to

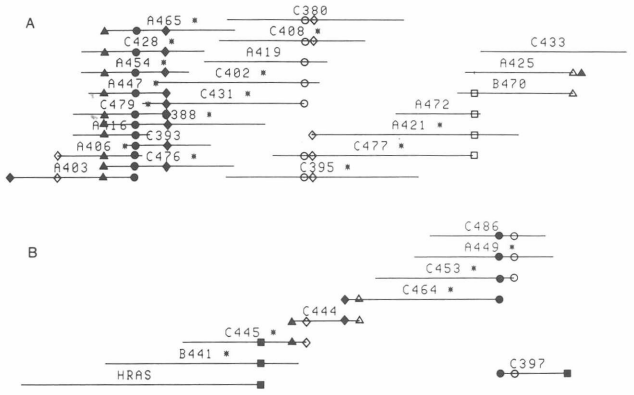


FIG. 2. Summary data for all probing experiments. Symbols should be read in a vertical direction only and the presence of a symbol indicates cross-hybridization. Note that contigs **A** and **B** do not overlap and are drawn above each other for reasons of convenience. Clones marked by * contain buried clones which are not displayed for reasons of space.

the 5′ side of the 5′ *Bam*HI site. The nature of the rearrangement is not yet clear but is detected by the presence of a new *Hin*dIII site 12 kb from the normal 3′ *Hin*dIII site. This is discussed below.

Rare Cutting Restriction Sites

Parallel with the fingerprint analyses we developed methods to detect *Eag*I and *Bss*HII sites in the clones. In practice we would only carry out these experiments within the framework of a reasonably extensive contig framework and only analyze that set of cosmids, which we call minimal set cosmids, that fully define a contig.

We used two methods to detect *Eag*I and *Bss*HII sites in the human DNA, both detailed under Materials and Methods. The first uses digestion of all of a cosmid microprep with either of the enzymes, followed by field inversion gel electrophoresis (FIGE) analysis (Carle *et al.*, 1986). There is a single site for both enzymes in the Lorist cosmid vector and cleavage of the human DNA generates at least two fragments that can normally be resolved on the FIGE system (data not presented). We find this method to be slightly unsatisfactory since the yields of DNA from a microprep are only high enough to allow detection of large DNA fragments and it is possible to miss clustered sites. We also find that the cosmids are quite variable in size and it can be difficult to distinguish partial digestion and circular products.

The alternative stategy is to cut the cosmid clones with *Bss*HII or *Eag*I, label the ends by filling in with reverse transcriptase, cleave the DNA with *Hin*dIII and analyze the products on 1% agarose gels followed by autoradiography (Fig. 3). The internal sites within the vector act as controls for successful cleavage (double-digest fragments of 4139 and 1247 bp are generated by *Bss*HII plus *Hin*dIII and of 3648 and 1738 bp by *Eag*I plus *Hin*dIII). If any site for either enzyme exists within the human DNA, then a new radioactive fragment or fragments will be seen on the autoradiograph.

Figure 3 shows an example of the results of the labeling protocol. Fragment patterns of clones labeled for *Bss*HII sites indicates clone A423 has a *Bss*HII site in its insert; clones A421, C422, and A424 have no sites in their inserts. Figure 3 shows similar data for *Eag*I: C457 and A449 have an *Eag*I site and A448 and C458 have none.

We have screened all 112 clones in the database for *Eag*I and *Bss*HII sites. The insert DNA of 27 (24%) clones have only an *Eag*I site(s), 11 (10%) have only *Bss*HII, and 16 (9%) have both.

We find that partial digestion by the rare cutting enzyme is a common problem but the end-labeling method is insensitive to this, since uncleaved sites will not be labeled. It is difficult to separate unincorporated label from the DNA on agarose gels and this can hide small fragments. For example, subsequent analysis of purified DNA of clone C423 revealed other *Bss*HII sites which generate fragments that are small and obscured in the labeling analysis. Soaking the gel or fixation in 5% trichloracetic acid and extensive washing does not seem to improve the background. We also find a low frequency of false positives in this assay. We do not know how these are generated and they may represent

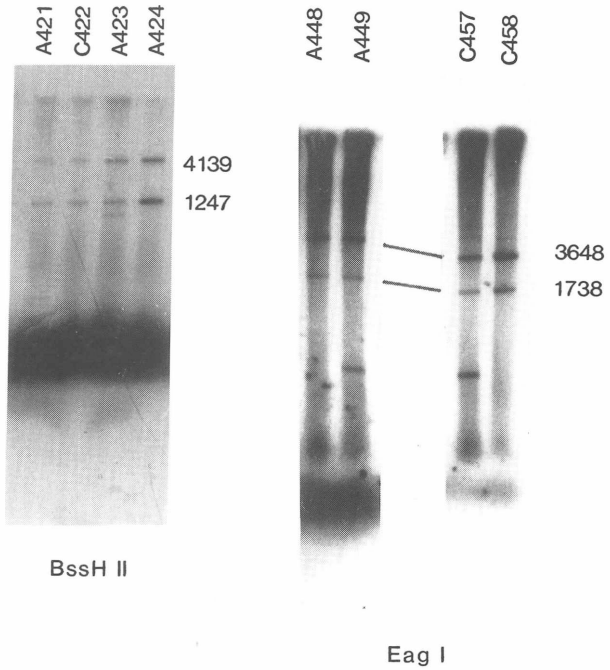


FIG. 3. Examples of the detection of *Bss*HII and *Eag*I sites by end-labeling. Autoradiographs of the final agarose gel analysis are displayed: vector-generated bands are marked with their size and the additional band in A423, A449, and C457 indicates a site for the appropriate enzyme is present in the human insert DNA.

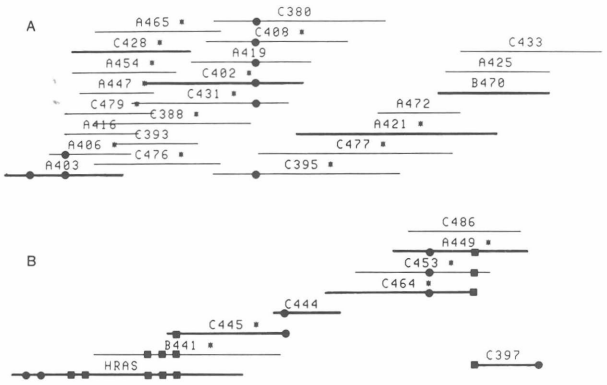


FIG. 4. Summary of the contig structures from E65-9. Clones marked by * contain buried clones. Locations of *Eag*I sites are marked by ● and *Bss*HII by ■. Note that location of these sites within a clone is not accurate: the symbol indicates a region of DNA, defined with respect to its overlapping neighbors, that must contain a site. Clones that are drawn bold correspond to the "minimal set" clones displayed in Fig. 5.

F episomal DNA in the host ED8767, reported by Fuller (1985).

The locations of the sites we have detected are given in Fig. 4. We find that there is logical consistency in the presence or absence of sites in most clones, which is defined by the position, the extent, and the direction of overlap of individual clones in the contig.

We find three *Eag*I sites within contig A and no *Bss*HII sites: contig B contains five *Eag*I sites and six *Bss*HII sites. The fragment of the H-*Ras* gene that we have cloned contains a number of otherwise rare cutting restriction sites: a 2.1-kb *Bam*HI fragment immediately to the 5′ side of the gene contains *Bss*HII, *Eag*I, *Nru*I, *Xho*I, and two *Sac*II sites, suggesting that this may contain a CpG island. This fragment does not appear to be rearranged compared to the normal and the E65-9 DNA must differ further to the 5′ side.

Final Structure

Figure 4 summarizes most of the information we have about the DNA from this region. Using a combination of probing and detailed analysis of fingerprints, we reduced the data to a decreasing number of contigs. Finally, we were able to construct two large contigs, A and B, containing 48 and 16 clones, respectively, and two small contigs containing two to three clones each and 43 unattached clones. These are predominantly mouse or mouse/human clones that we suspect are artifactual (source of this is discussed under Materials and Methods, under library construction) and our true figure for human clones is about 85. In one case, C433, such a clone is located at the end of contig A and it may represent the junction DNA of the transgenome.

It is impossible to estimate the amount of DNA contained within the contigs from the fingerprint data directly since overlap in this analysis is defined by the number of fragments in common, not by physical size. We can assess this by size analysis of *Hin*dIII-digested minimal set cosmids displayed in Fig. 5. Contig A contains about 140 kb and contig B about 125 kb of DNA. The remaining 2 contigs contribute about 40 kb and we do not know the size of the gap to the 5′ side of the H-*Ras* gene in contig B. The minimum total human DNA in E65-9 must, from these figures, be 305 kb.

DISCUSSION

In this paper we have detailed the methods used to construct a set of overlapping cosmid DNA clones that contain about 300,000 bp of DNA from the region of the H-*Ras* gene on the short arm of human chromosome 11. Our primary reasons for doing this was to test out

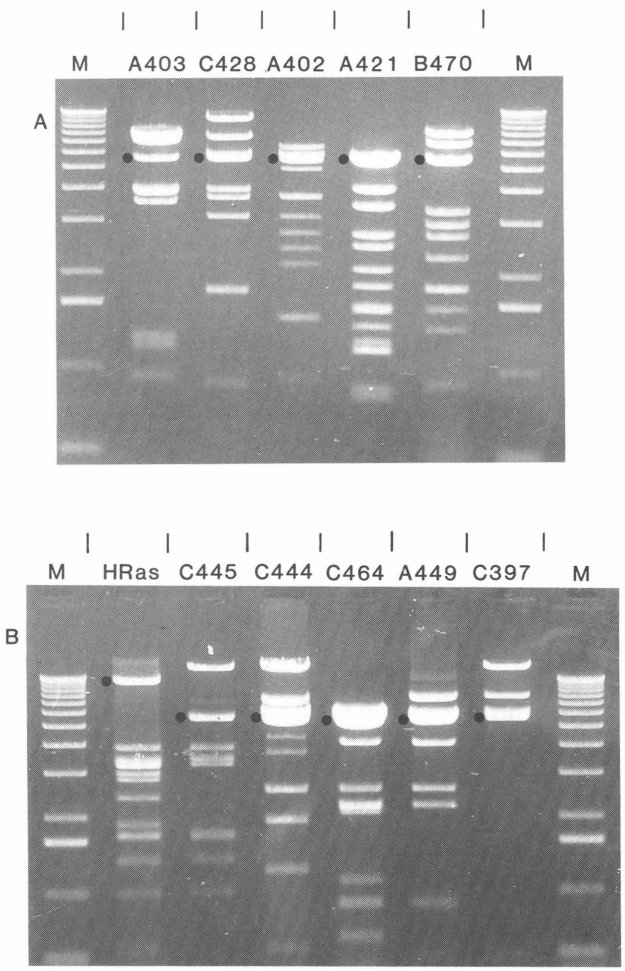


FIG. 5. *Hin*dIII digests of the "minimal set" clones that must completely define contig **A** or **B**. Vector or vector-containing fragments are marked ●. The contig size is calculated by summing the size of the smallest set of fragments which uniquely defines every fragment in all cosmids.

the methods, to find their limitations, and finally to derive a strategy that was capable of application to larger regions of the genome. We feel that it is necessary to carry out the limited pilot project to establish detailed protocols, to identify common sources of experimental difficulty, and to refine simple and reproducible methods of application of fingerprint mapping to our experimental system. We believe that the methods we describe here are capable of scale-up and we are currently attempting similar analyses of a cell line, EJNAC (D.J.P. and V. van Heyningen, unpubl.), that contains most of the short arm of chromosome 11, corresponding to about 2×10^7 bp of DNA.

The limitations of this technique are of three kinds, problems associated with the library construction, with the contig analysis, and finally with the cell lines to be used as targets for analysis.

Library construction is sensitive to two problems with restriction site distribution. The sites may be distributed such that only one specific partial digestion fragment may be the correct size to fit in a cosmid vector. This would reduce the frequency of the fragment appearing in the library. It is also possible that the sites are such that no partial fragments of the correct size can be generated, making the DNA unclonable.

The manipulative advantage that partial *Hin*dIII cloning provides is significant and this has to be weighed against the problems we have encountered. The contig analysis is itself sensitive to large *Hin*dIII fragments, since these generate clones with few fingerprint fragments, which can be difficult to link with statistically significant overlaps. The advantage of *Sau*3AI partial digest libraries is that some of the DNA of large *Hin*dIII fragments is clonable whereas with partial *Hin*dIII cloning, none is. We believe that a combination of cloning enzymes or physical shearing may be necessary, with the bulk of the data generated from *Hin*dIII clones and specific probing being carried out against *Sau*3AI libraries.

It is possible that the difficulty we had in isolating a single H-*Ras* cosmid was due to some feature of the DNA adjacent to the gene. We know that the H-*Ras* gene in E65-9 is amplified and would expect to isolate more than the single cosmid from screening 2.2×10^5 clones. There may be sequences near H-*Ras* that cause deleterious effects on *E. coli* resulting in reduced frequency or complete absence of the gene in the library. We have discussed the potential nature of these sequences elsewhere (Gibson *et al.*, 1987). In our analysis, we encountered problems with clone C444 and two clones that overlap with it, C468 and C484. C444 is somewhat unstable and deletes upon large scale growth: this can be seen in Fig. 5 where two submolar fragments are seen in the *Hin*dIII-digested DNA of this clone.

We believe it will be necessary to screen other vector systems to recover sequences that may not readily clone in cosmid vectors. A random clone analysis approach must inevitably result in gaps within the contigs because of the statistics of sampling. "Unclonable" or rarely cloned sequences may be present in these gaps and other methds of nonrandom study, such as direct DNA analysis by Southern blotting and PFGE or analysis of YAC clones, will be required (Coulson *et al.*, 1988). These alternative methods either lack, or have different constraints, on randomness and this will bypass the problematic sequences. The difficulty remains should the "unclonable" sequence be required in a cosmid clone.

The cell line that we have chosen to use is rearranged with respect to the normal human genome and we have made no attempt to address this problem in this paper. We accept that any mapping technique using clone-based analysis is insensitive to repeats in human DNA larger than the vector capacity. The E65-9 cell line is, we hope, an extreme example of this problem. Porteous *et al.* (1986) have identified rearrangements in several chromosome-mediated gene transfer (CMGT) cell lines that contain substantial portions of chromosome 11, and this seems to be a common feature of the procedure. We will have to carry out more detailed Southern blotting experiments to identify the extent of rearrangement in E65-9, and this will be of particular concern over the relative positions of the potential CpG island and the H-*Ras* gene.

The map we have generated has one area of potential biological interest in the relationship of the possible CpG island and the H-*Ras* gene. Our provisional understanding of this is that the H-*Ras* gene in E65-9 has undergone a rearrangement of its 5′ sequences, involving the introduction of DNA that contains a *Hin*dIII site. We do not know if this has an effect on H-*Ras* gene expression or not. The novel sequences are of human origin and must presumably be located close to H-*Ras* in normal DNA. Morton *et al.* (1987) have shown 5′ rearrangement of H-*Ras* in a cell line E65-7 that was generated at the same time and by the same method as E65-9.

The H-*Ras* gene is known to rearrange in a wide variety of tumors: loss of heterozygosity, aquisition of homozygosity, and hemizygosity have all been observed (reviewed by Ponder, 1988). It will be interesting to extend these studies to flanking markers and genes, with particular emphasis on the sequences we have shown to be involved in rearrangement.

ACKNOWLEDGMENTS

We are very pleased to acknowledge the support, encouragement, and help of John Sulston and Alan Coulson (L.M.B. Cambridge). Without their involvement, this project could not have been initiated. This work was supported by a grant to P.F.R.L. from the United Kingdom Cancer Research Campaign.

REFERENCES

1. BIRD, A. P. (1986). CpG-rich islands and the function of DNA methylation. *Nature (London)* **321:** 209–213.
2. BURKE, D. T., CARLE, G. F., AND OLSON, M. V. (1987). Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* **236:** 806–812.
3. CARLE, G. F., FRANK, M., AND OLSON, M. V. (1986). Electrophoretic separations of large DNA molecules by periodic inversion of electric field. *Science* **232:** 65–68.
4. CLARKE, L., AND CARBON, J. (1976). A colony bank containing synthetic ColE1 hybrid plasmids representative of the entire *E. coli* genome. *Cell* **9:** 91–99.
5. COULSON, A., AND SULSTON, J. (1988). Genome mapping by restriction fingerprinting. *In*"Genome Mapping, A Practical Approach" (K. E. Davies, Ed.), pp. 19–39, IRL Press, Oxford.
6. COULSON, A., SULSTON, J., BRENNER, S., AND KARN, J. (1986). Towards a physical map of the genome of the nematode *Caenorahbditis elegans. Proc. Natl. Acad. Sci. USA* **83:** 7821–7825.
7. COULSON, A., WATERSTONE, R., KIFF, J., SULSTON, J., AND KOHARA, Y. (1988). Genome linking with yeast artificial chromosomes. *Nature (London)* **35:** 184–186.
8. CROSS, S. H., AND LITTLE, P. F. R. (1986). A cosmid vector for systematic chromosome walking. *Gene* **49:** 9–22.
9. FEINBERG, A. P., AND VOGELSTEIN, B. (1983). A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.* **132:** 6–13.
10. FULLER, L. (1985). *E. coli*K12 strain ED8767 contains the F plasmid. *BioTechniques* **3:** 454–455.
11. GIBSON, T. J., COULSON, A. R., SULSTON, J. E., AND LITTLE, P. F. R. (1987). Lorist 2—A cosmid with transcriptional terminators insulating vector genes from interference by promoters within the insert; effect on DNA yield and clone frequencies. *Gene* **53,** 275–281.
12. GIBSON, T. J., AND SULSTON, J. E. (1987). Preparation of large numbers of plasmid DNA samples in microtiter plates by the alkaline lysis method. *Gene Anal. Techn.* **4:** 41–44.
13. GOLDFARB, M., SHIMIZU, K., PERUCHO, M., AND WIGLER, M. (1982). Isolation and preliminary characterisation of a human transforming gene from T24 bladder carcinoma cells. *Nature (London)* **296:** 404–409.
14. GOSDEN, J. R., AND PORTEOUS, D. J. (1987). HRAS1-selected, chromosome mediated gene transfer: *In situ* hybridisation with combined biotin and tritium labels localises the oncogene and reveals duplications of the human transgenome. *Cytogenet. Cell Genet.* **45:** 44–51.
15. KOHARA, Y., AKIYAMA, K., AND ISONO, K. (1987). The physical map of the whole *E. coli* chromosome: Application of a new strategy for rapid anaylsis and sorting of a large genomic library. *Cell* **50:** 495–508.
16. LITTLE, P. F. R. (1987). Choice and use of cosmid vectors. *In* "Gene Cloning" (D. M. Glover, Ed.), Vol. 3, pp. 19–42. IRL Press, Oxford.
17. LITTLE, P. F. R., AND CROSS, S. H. (1985). A cosmid vector that facilitates restriction enzyme mapping. *Proc. Natl. Acad. Sci. USA* **82:** 3159–3163.
18. MANIATIS, T., FRITSCH, E. F., AND SAMBROOK, J. (1982). "Molecular Cloning: A Laboratory Manual." Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
19. MORTON, J. E. N., HIRST, M. C., AND PORTEOUS, D. J. (1987). The c-Harvey-*ras*-1 oncogene in chromosome mediated gene transfer. *Anticancer Res.* **7:** 573–588.
20. OLSON, M. V., DUTCHIK, J. E., GRAHAM, M. Y., BRODEUR, G. M., HELMS, C., FRANK, M., MACCOLLIN, M., SCHEINMAN, R., AND FRANK, T. (1986). Random clone strategy for genomic restriction mapping in yeast. *Proc. Natl. Acad. Sci. USA* **83:** 7826–7830.
21. PONDER, B. (1988). Gene loss in human tumours. *Nature (London)* **335:** 400–401.
22. PORTEOUS, D. J., MORTEN, J. E. N., CRANSTON, G., FLETCHER, J. M., MITCHELL, A., VAN HEYNINGEN, V., FANTES, J. A., BOYD, P. A., AND HASTIE, N. D. (1986). Molecular and physical arrangements of human DNA in HRAS1-selected, chromosome-mediated transfectants. *Mol. Cell. Biol.* **6:** 2223–2232.
23. SULSTON, J., MALLETT, F., STADEN, R., DURBIN, R., HORSNELL, T., AND COULSON, A. (1988). Software for genome mapping by fingerprint techniques. *CABIOS* **4:** 125–132.
24. SULSTON, J., MALLET, F., DURBIN, R., AND HORSNELL, T. (1989). Image analysis of restriction enzyme fingerprint autoradiographs. *CABIOS* **5:** 101–106.