POSITIONAL CLONING OF THE GENE RESPONSIBLE FOR DENT'S DISEASE

Simon E. Fisher

A Thesis submitted for the degree of

Doctor of Philosophy

in the University of Oxford

St. Catherine's College

and the

Genetics Laboratory

Department of Biochemistry

Oxford

Michaelmas 1995

Abstract

Positional cloning of the gene responsible for Dent's disease

Simon E. Fisher, St. Catherine's College, Oxford

Submitted for the degree of Doctor of Philosophy

Michaelmas 1995

The hypervariable locus DXS255 in human Xp11.22 has a heterozygosity exceeding 90% and has therefore facilitated the localization of several disease genes which map to the proximal short arm of the X chromosome, including the immune deficiency Wiskott-Aldrich syndrome and the eye disorders retinitis pigmentosa, congenital stationary night blindness and Åland Island eye disease. In addition, a microdeletion involving DXS255 has been identified in patients suffering from Dent's disease, a familial X-linked renal tubular disorder which is characterized by low molecular weight proteinuria, hypercalciuria, nephrocalcinosis, nephrolithiasis (kidney stones) and eventual renal failure. Two YAC contigs were constructed in Xp11.23-p11.22 in order to aid transcript mapping; the first centred on the DXS255 locus, the second mapping distal to the first and linking the genes GATA, TFE3 and SYP to the OATL1 cluster. Eleven novel markers were generated, one of which contains an exon from a novel calcium channel gene. Four putative CpG islands were detected in the region. Analysis of the microdeletion associated with Dent's disease using markers from the DXS255 contig demonstrated that it is confined to a 370kb interval. A YAC overlapping this deletion was hybridized to a kidney-specific cDNA library to isolate coding sequences that might be implicated in the disease ætiology. The clones thus identified detect a 9.5kb transcript which is expressed predominantly in kidney, and originate from a novel gene (CLCN5) falling within the deleted region. Sequence analysis indicates that the 746 residue protein encoded by this gene is a new member of the ClC family of voltage-gated chloride channels. The coding region of CLCN5 is organized into twelve exons, spanning 25-30kb of genomic DNA. Using the information presented in this thesis, other studies have identified deletions and point mutations which disrupt CLCN5 activity in further patients affected with X-linked hypercalciuric nephrolithiasis, confirming the role of this locus in renal tubular dysfunction.

Acknowledgements

I would like to thank Professor John Edwards for giving me space in the laboratory, and Professor Kay Davies for letting me linger while I wrote up. I am grateful to Ian Craig, my supervisor, for support and encouragement during my project, and for co-hosting excellent parties with Sally.

Thanks also to David Hunter, Anil Day, Val Cooper, Zoe Christadoulou, Ben Yudkin, Brian Archer, Sarah Lloyd, Simon Pearce and Raj Thakker.

I am grateful to all members of the Genetics Lab., for providing such a friendly atmosphere in which to work. I would like to thank the following people in particular:

Max, for lots of guidance when I was finding my feet.
Elinor, for supplying me with copious quantities of caffeine.
Saiful, for advice on anything from YACs to phage libraries and for the odd faux pas.
Rudi, for M27β expertise and a wonderful wedding.
Eli, for enthusiastic discussions and contig linking.
Graeme, for a huge amount of encouragement and assistance with the Dent's project.
Inge, for being the lab sequencing Guru, keeping us all organized, and for teaching me the Dutch for 'delicious'.
Don, for waving his hands over my cDNA library lifts, rescuing badly-poured polyacrylamide gels and for introducing me to the joys of early morning Baroque.
Nancy, for help with YAC work, and (hopefully) not minding too much about a spelling mistake.
Jennifer, Karo, Pia and Anita for making G113 an enjoyable working environment.
Julie, for advising me that I should eat more ice-cream and for demonstrating a fool-proof way for winning at pool.
Sonia, for thesis proof-reading, the 'Stand' dance and a joke about a duck. Also for her remarkably bi-functional seminars, which were both informative and entertaining.
Aarti, for everything (advice, encouragement, critical reading, et al.).
Jon (honorary member of the 'Craig' lab), for teaching me all I know about making cocktails and for help with mathematical conundrums (2 + 2 = ?).

Chris, Kath, Mike, Dave & Adam for letting me pretend to be an undergraduate again.

Jayne, for Transatlantic support, and for believing, despite my protests to the contrary, that I spend my whole day looking down a microscope.

Vicky, for technical advice on all dance-related aspects of this thesis and for impressing people at lab parties by knowing what YAC and PCR stand for. Most of all, for sticking with me through all the sadness and euphoria.

Mum and Dad, who have supported and inspired me over the years.

Publications

The work presented in this thesis has been incorporated into the following publications:

Meeting abstracts

Hatchwell, E., Black, G. C. M, Chand, A., Chen, Z.-Y., Fisher, S. E., Hendriks, R. W., Hinds, H., Riley, S., Coleman, M., Monaco, A., Goodfellow, P.N. and Craig, I. (1993) Development of resources for advancing physical mapping of the Xp11.4-cen interval. *Fourth international workshop on X chromosome mapping.*

Fisher, S. E., Hatchwell, E., Chand, A., Ockendon, N., Monaco, A. and Craig, I. (1994) YAC contigs and physical mapping of Xp11.23-p11.22. *Fifth international workshop on X chromosome mapping.*

Chand, A., Fisher, S. E., Hatchwell, E., Ockendon, N., Clark, J., Cooper, C. S., Monaco, A. and Craig, I. (1995) YAC contigs, physical mapping of Xp11.23-p11.22 and detailed analysis of the OATL1 locus. *Sixth international workshop on X chromosome mapping.*

Lloyd, S. E., Pearce, S. H. S., Fisher, S. E., Harding, B., Scheinman, S. J., Goodyer, P., Wrong, O. M., Craig, I. W. and Thakker, R. V. (1995) Hereditary nephrolithiasis is associated with mutations in an X-linked chloride channel gene. *American Society for Bone and Mineral Research (ASBMR) meeting.*

Scheinman, S. J., Lloyd, S. E., Pearce, S. H. S., Fisher, S. E., Salenger, P. V., Hoopes, R. R. Jr., Craig, I. W. and Thakker, R. V. (1995) Analysis of the gene encoding an X-linked voltage-gated chloride channel in idiopathic hypercalciuria. *American Society for Bone and Mineral Research (ASBMR) meeting.*

Journal publications

Fisher, S. E., Black, G. C. M., Lloyd, S. E., Hatchwell, E., Wrong, O., Thakker, R. V. and Craig I. W. (1994) Isolation and partial characterization of a chloride channel which is expressed in kidney and is a candidate for Dent's disease (an X-linked hereditary nephrolithiasis). *Hum. Mol. Genet.* **3:** 2053-2059

Blair, H. J., Ho, M., Monaco, A. P., Fisher, S. E., Craig, I. W. and Boyd, Y. (1995) High-resolution comparative mapping of the proximal region of the mouse X chromosome. *Genomics* **28:** 305-310

Fisher, S. E., Hatchwell, E., Chand, A., Ockendon, N., Monaco, A. P. and Craig, I. W. (1995) Construction of two YAC contigs in human Xp11.23-p11.22, one encompassing the loci OATL1, GATA, TFE3 and SYP, the other linking DXS255 to DXS146. (1995) *Genomics* **29:** 496-502

Fisher, S. E., van Bakel, I., Lloyd, S. E., Pearce, S. H. S., Thakker, R. V. and Craig, I. W. (1995) Cloning and characterization of CLCN5, the human kidney chloride channel gene implicated in Dent disease (an X-linked hereditary nephrolithiasis). *Genomics* **29:** 598-606

Shipley, J. M., Birdsall, S., Clark, J., Crew, J., Gill, S., Linehan, M., Gnarra, J., Fisher, S. E., Craig, I. W., Cooper, C. S. (1995) Mapping the chromosome X breakpoint in two papillary renal cell carcinoma cell lines with a t(X;1) (p11.2; q21.2) and the first report of a female case. *Cell Genet. Cytogenet.* In press

Lloyd, S. E., Pearce, S. H. S, Fisher, S. E., Steinmeyer, K., Schwappach, B., Scheinman, S. J., Harding, B., Bolino, A., Devoto, M., Goodyer, P., Rigden, S. P. A., Wrong, O., Jentsch, T. J. J., Craig, I. W. and Thakker, R. V. (1995) A common molecular basis for three inherited kidney stone diseases. *Submitted*

Table of contents

Chapter 1 – Introduction

Human molecular genetics is a vibrant and rapidly expanding field of research. The substantial advances which have been made in the past ten years or so have had profound implications for studies of human genetic disease. Over 4000 disorders displaying single-gene Mendelian inheritance patterns are known in man (McKusick, 1992). Recently developed techniques have provided a basis for the identification of the primary molecular defects responsible for many of these diseases. They are also giving new insights into the role of genetic factors in complex traits which affect a larger proportion of the population, such as diabetes, asthma, Alzheimer's disease and several different forms of cancer. This research is therefore likely to make a significant impact on the diagnosis, prevention and treatment of disease in years to come. In addition, it is increasing our understanding of the functions of gene products and mechanisms of gene interaction in normal individuals.

1.1 Positional cloning

1.1.1 The advent of positional cloning

Prior to 1986, the isolation of human disease genes relied on pre-existing knowledge regarding the primary biochemical defect responsible. This approach, which uses information regarding the normal protein product or its function has been termed 'functional cloning' (Collins, 1992). However, such information is lacking for the vast majority of single gene disorders that have been described. An alternative approach, known as 'positional cloning', exploits genetic mapping techniques to identify the location on a particular chromosome of the gene responsible. Further refinement of the candidate region can lead to the successful isolation of this gene, and subsequent analysis of the gene product provides insights into its function. As described below, due to the large size of the human genome (3×10^9 bp), the first positional cloning efforts had to contend with a resolution gap between linkage studies/cytogenetics and conventional molecular biology. The recent emergence of 'megabase methods' and a rapid accumulation of genetic and physical mapping data regarding the human genome has since greatly eased the task of disease gene isolation.

1.1.2 Bridging the resolution gap in human genetics

Two main types of study, known as linkage analysis and cytogenetics, can provide preliminary information regarding the localization of a human disease gene to facilitate positional cloning:

i) Linkage analysis

This approach uses genetic markers from different regions of the genome to identify any positive correlations between the segregation of a disease trait and the inheritance of a specific chromosomal region. Only those loci which have detectable polymorphisms (i.e. alleles which can be disinguished) can be exploited for such studies. The general lack of useful markers presented significant problems for linkage mapping in humans for much of this century. However, in 1980, Botstein *et al.* suggested that the naturally occurring variation present at the DNA sequence level could provide an abundant supply of novel polymorphic loci. The first usage of DNA variation in linkage analysis involved the detection of restriction fragment length polymorphisms (RFLPs) resulting from base substitutions which create or destroy cleavage sites for a specific restriction endonuclease (Murray *et al.*, 1982; Gusella *et al.*, 1983). This type of marker has its limitations; few restriction endonucleases will detect such an RFLP at any given locus, due to the low mean heterozygosity of human DNA (~0.001 per base pair) and when detected the majority will be only dimorphic (with the cleavage site either present or absent), so that the marker heterozygosity can never exceed 50% (see Jeffreys *et al.*, 1985a). This reduces the chances of the RFLP being informative in the pedigrees being analysed.

The discovery of hypervariable 'minisatellite' regions of human DNA, showing high heterozygosities as a consequence of their multiallelic variation (see Jeffreys *et al.*, 1985a), greatly increased the efficiency with which DNA polymorphisms could be used to localize disease genes (Reeders *et al.*, 1985). The basis of variation at these loci, also known as VNTRs (variable number of tandem repeats), is a change in copy number of a short (9-64bp) sequence element, iterated in tandem to form arrays of 0.1-20kb long (Jeffreys *et al.*, 1985a; Nakamura *et al.*, 1987). Such polymorphisms can be detected with

any restriction endonuclease that cleaves outside the repeat unit. Jeffreys *et al.* (1985a) identified the presence of many dispersed minisatellite regions in the human genome, all possessing a shared 10-15bp 'core' sequence showing similarity to the crossover hot spot instigator (χ) sequence of *E. coli.* and bacteriophage λ. However, it should be noted that these loci were isolated by hybridization with a minisatellite probe, and it is thus difficult to evaluate the significance of the observed sequence similarity; VNTRs which have been identified independently (such as that contained in the DXS255 locus described below) often lack homology to the core sequence. The hybridization of a probe consisting of tandem repeats of the core to digests of human genomic DNA detects many highly polymorphic minisatellites simultaneously and provides the basis for multilocus 'DNA fingerprinting', a technique which has had profound implications in many fields, including forensics, paternity testing and transplant screening (Jeffreys *et al.*, 1985b).

A further source of variation which has been exploited for genetic mapping involves the length polymorphisms exhibited by microsatellite markers (Weber and May, 1989). These consist of repeats of a di-, tri- or tetranucleotide motif, of which $(dC\text{-}dA)_n\cdot(dG\text{-}dT)_n$ sequences (also referred to as CA repeats) are the most abundant in the human genome (Weber, 1990). Microsatellites are often highly polymorphic and are more evenly and frequently distributed throughout the genome than minisatellites, which tend to cluster in regions adjacent to telomeres (Weissenbach *et al.*, 1992). Microsatellite polymorphisms are observed by using the polymerase chain reaction (PCR; see Section 2.13) to amplify a small target segment containing the repeats, and then sizing the products on polyacrylamide gels. Weissenbach *et al.* (1992) have shown that systematic screening of genomic libraries with a polydinucleotide probe is a highly effective method for generating large numbers of novel genetic markers.

Despite the above advances in human genetic mapping in recent years, the resolution obtained from linkage studies of a disease gene, which relies on the number of informative meioses in the available families, is usually limited to ~1cM (see Collins, 1992). This corresponds to ~1Mb of DNA on average, but varies in different regions of the genome, since genetic distance does not always correlate with physical distance.

The difficulties encountered with genetic mapping of common monogenic diseases are even more significant for rare disorders, where there will be insufficient families for precise localization. Furthermore, in the analysis of complex traits such as diabetes, asthma and many forms of cancer, factors such as incomplete penetrance, phenocopy, genetic heterogeneity or polygenic inheritance provide additional complications for gene mapping.

ii) Cytogenetic studies

The association of a gross cytogenetic abnormality, such as a deletion or translocation, with a particular disease phenotype suggests that the gene responsible is likely to map to a chromosomal region involved in the rearrangement. Such observations are very useful for positional cloning efforts even if they are found in only a few patients suffering from the disorder. In certain cases, a large deletion may result in a contiguous gene syndrome, such as that observed in a male affected with Duchenne muscular dystrophy (DMD), retinitis pigmentosa (RP), chronic granulomatous disease (CGD) and the McLeod erythrocyte phenotype (XK), who was found to have an interstitial deletion of Xp21 (Francke *et al.*, 1985). Another example involves the WAGR syndrome, in which a hereditary predisposition to Wilm's tumour (WT) occurring in association with aniridia (AN2), urogenital abnormalities (UG) and mental retardation, is correlated with constitutional heterozygous deletions of 11p13 (Rose *et al.*, 1990). An unusual type of chromosomal abnormality is that found in fragile X syndrome, the most common form of inherited mental retardation in humans. This disorder is associated with an inducible fragile site, expressed as an isochromatid gap in Xq27.3 on metaphase spreads.

Analysis of patients with cytogenetic abnormalities will typically lead to the identification of markers which flank the region involved (or in the case of deletions, map within it), and therefore define the minimal segment of DNA in which the disease gene may be found. Even when chromosomal rearrangements are not visible in cytogenetic studies, hybridization to Southern blots using probes from a target region (established by linkage, as described above) may detect abnormalities, such as microdeletions, in a subset of patients.

A further area of cytogenetic analysis involves the use of the fluorescence *in situ* hybridization (FISH) technique to determine directly the chromosomal localizations and relative orders of DNA markers (reviewed in Trask, 1991). When conventional FISH is applied to metaphase spreads of human chromosomes, it gives a maximum resolution of ~1Mb. The use of interphase nuclei can increase the resolution to 50-100kb, but is technically demanding and requires large data sampling to provide statistically significant results (see Parra and Windle, 1993).

Thus, for any particular target of positional cloning, the combined results of the kinds of linkage analysis and cytogenetic studies described above will, at best, localize the gene responsible to within ~1Mb of a marker locus. Note that somatic cell hybrid analysis provides an additional source of physical mapping data regarding the orders of loci which delimit a candidate region, but does not, on its own, give estimates of physical distance between them.

iii) Conventional molecular cloning techniques

Size fractionation using conventional agarose electrophoresis can resolve fragments of up to ~50kb. Larger molecules migrate, but are not separated according to size. Similarly, the maximum length of DNA which can be cloned into a cosmid is 40-45kb (Hohn and Collins, 1980). This type of vector is a plasmid that has been modified to include the λ sequence, required for *in vitro* packaging into a bacteriophage λ capsid; the packaging step imposes a size limit on the DNA filling the bacteriophage head. 'Chromosome walking' techniques where overlapping clones are used to obtain sequences progressively further from a starting probe typically proceed with a step of only ~20kb. The iterative screening of large genomic libraries inherent in this technique makes it very laborious to walk even a few hundred kilobases.

iv) The impact of 'megabase methods'

Thus, for studies of the human genome, there is a resolution gap between the ~1Mb measurable using linkage studies/cytogenetics and the 50kb maximum which can be analysed with conventional molecular biological techniques. Prior to the emergence of techniques which bridged this gap, the difficulties involved in constructing bacteriophage/cosmid contigs covering large regions of DNA was a limiting factor for positional cloning strategies. (It should be noted, however, that a small number of disease genes, including that which causes DMD (Monaco *et al.*, 1986) were successfully isolated using clones from such contigs.)

The developments of yeast artificial chromosome (YAC) vectors, which can be used to clone large fragments (typically in the 50-1000kb range, but sometimes up to a few megabases in size) and pulsed field gel electrophoresis (PFGE), which can resolve fragments in a similar size range, have therefore been highly beneficial for human molecular genetic studies. (These techniques are described in detail in Chapter 3.) PFGE mapping can be used to obtain a more accurate estimate of a target region for a disease gene, initially defined by linkage analysis/cytogenetics (e.g. Rommens *et al.*, 1989; Rose *et al.*, 1990). The construction of clone contigs spanning regions of several hundred kilobases is relatively straightforward when using YACs (although some limitations are discussed in Chapter 3) and the process of positional cloning is therefore greatly accelerated.

The availability of 'chromosome jumping' libraries, in which screening with a marker from one end of a genomic fragment of several hundred kiloases yields a novel marker from the opposite end, has also expedited gene hunts, aiding in the isolation of the genes responsible for cystic fibrosis (CF), neurofibromatosis (NF1) and aniridia (AN2), among others (Rommens *et al.*, 1989; Wallace *et al.*, 1990; Ton *et al.*, 1991). However, the use of YACs has generally been favoured over jumping libraries, mainly because the latter does not provide material from intervening regions between chromosome jumps.

Comparable improvements have recently been made in FISH technology, involving hybridization to duplex DNA which has been forcibly stretched (Parra and Windle, 1993). Such methods have been demonstrated to resolve distances from a few kilobases to over a megabase.

1.1.3 Identification of candidate genes from a target region

Only a small proportion (as little as 2%) of the human genome is thought to consist of regions encoding gene products. The remainder is mainly composed of introns (intervening unique sequences within genes, described in detail in Chapter 6), control regions which modulate gene expression, and a large amount of highly reiterated and middle repetitive DNA (including the *Alu* and LINE elements, discussed in Chapter 3). The challenge for any positional cloning exercise is therefore to identify and recover candidate coding regions from within a segment of DNA which may be several hundred kilobases in size. This problem has been approached in several different ways, often using a combination of strategies:

i) Hybridization studies of single copy sequences

A variety of techniques have been used to retreive small DNA fragments from a target region for further analysis. These include phenol-enhanced reassociation (PERT) for specific cloning of fragments absent from a patient with a homozygous or hemizygous deletion (Kunkel *et al.*, 1985), saturation mapping of markers obtained from chromosome-specific genomic libraries (Rommens *et al.*, 1989; Call *et al.*, 1990), isolation of clones by chromosome jumping (Rommens *et al.*, 1989; Wallace *et al.*, 1990; Ton *et al.*, 1991) and subcloning of inserts from phage (Monaco *et al.*, 1986), cosmids (Call *et al.*, 1990) or YACs (Franco *et al.*, 1991). Hybridization with total human DNA allows the screening out of any fragments containing repetitive elements.

The remaining clones can then be assessed for coding potential as follows:

Evolutionary conservation

If a human genomic fragment cross-hybridizes to sequences in the genomic DNA of distantly related species when used to probe a 'zooblot', it suggests that the fragment originates from a gene, since many genes show evolutionary conservation (Monaco *et al.*, 1986). This has been used as an initial assay in many gene hunts.

Northern blots

The subset of the genome which is transcribed in a specific tissue can be recovered as messenger RNA (mRNA). Hybridization of a single copy sequence to a Northen blot containing poly-A enriched mRNA from a series of different tissues and/or cell lines, indicates whether or not the fragment detects a transcript, and also provides useful information on sizes and tissue distributions of any such transcripts, which may be important for cDNA screening and assessment of candidacy (see below). However, it should be noted that, in some cases, a conserved genomic fragment originating from a gene locus may fail to detect signals from Northern blots (Rommens *et al.*, 1989).

Screening of cDNA libraries

Reverse transcription of mRNA molecules to form complementary DNA (cDNA), followed by cloning into a phage or plasmid vector, is used to construct libraries which are greatly enriched for transcribed, processed sequences. These libraries can be screened by hybridization with a single copy genomic fragment, in order to recover a corresponding cDNA clone. If such a clone is successfully isolated, then sequence analysis of the insert can directly identify any open reading frames (ORFs) that may be present, since the cDNA should not be disrupted by introns. (Typically, several overlapping cDNA clones are examined, to verify sequence information and exclude the possibily of cloning artefacts.) An important consideration of any attempt at cDNA screening is the source of the cDNA library, since different tissues have different expression profiles. As described above, probing of Northern blots, prior to cDNA screening, gives an idea of the tissue-specificity of transcripts detected by a fragment.

In addition, it should be noted that knowledge of which tissue types are affected in a patient suffering from a genetic disorder can often suggest which type of cDNA library would be the most appropriate to use for isolation of candidate transcripts.

ii) Identification of CpG islands

The 5' ends of genes are often found to be associated with regions that are rich in hypomethylated CpG residues (Bird, 1986). These regions are known as 'CpG islands' and can be detected by the clustering of cleavage sites for rare-cutting restriction endonucleases. (This is discussed in more detail in Chapter 3.) Several positional cloning efforts have used the presence of a CpG island as a criterion for identifying potential coding regions (e.g. Rommens *et al.*, 1989; Verkerk *et al.*, 1991; Ton *et al.*, 1991). In fragile X syndrome, a CpG island adjacent to the fragile site was found to be hypermethylated in patients, and cosmids containing this island were subsequently used to isolate the associated gene, FMR-1, which is responsible for the disorder (Verkerk *et al.*, 1991 and see below). It should be noted, however, that an estimated 60% of tissue-specific genes are not associated with CpG islands (Gardiner-Garden and Frommer, 1987; Larsen *et al.*, 1992).

iii) Direct isolation of cDNAs using cosmids or YACs

The individual analysis of single copy sequences described above, while often effective, can be laborious and time-consuming. Alternative approaches have recently been developed which exploit the availability of large DNA fragments, cloned in cosmids or YACs, for the direct recovery of cDNAs mapping within the region. One such strategy, involves hybridization of the large fragment, following suppression of repetitive elements, to a cDNA library, and has been used successfully to isolate several disease genes (e.g. Wallace *et al.*, 1990; Buxton *et al.*, 1992; Chen *et al.*, 1992). Similarly, hybridization of cDNAs to immobilized genomic fragments provides a means for direct selection of candidate clones (Parimoo *et al.*, 1991; Lovett *et al.*, 1991). However, as with the analysis of single copy fragments, appropriate choice of cDNA library may be critical for success.

iv) Exon trapping

The 'exon trapping' technique is based on the functional identification of *cis*-acting sequences that are necessary for RNA splicing, and facilitates the isolation of transcripts without prior knowledge of their tissue specificity (Duyk *et al.*, 1990; Buckler *et al.*, 1991). Random segments of genomic DNA from the target region are inserted into an intron present within a mammalian expression vector, which is then transfected into COS cells. If an exon flanked by functional 5' and 3' splice sites is contained in the genomic insert, then *in vivo* transcription and processing will lead to the production of a fusion transcript, derived from the pairing of unrelated vector and genomic splice site signals. These trapped exons can be recovered using RNA-based PCR. The Huntington's Disease Collaborative Research Group (1993) finally tracked down the gene responsible for this disorder, which had eluded isolation for many years, using exon trapping of cosmids from a 500kb region of 4p16.3.

v) Sequence based strategies

As well as being applied to small single copy genomic fragments, sequence based strategies have also been used to identify exons from regions of over 60kb in size which have been entirely sequenced (Legouis, 1991):

Database searches

The use of computer programs to screen nucleotide and protein sequence databases for similarity to a novel genomic sequence may sometimes reveal the presence of an ORF with homology to a known gene. (Homology searches are described in more detail in Chapter 5.) Sinclair *et al.* (1990) employed this approach to identify the coding region of the SRY (sex-determining region Y) gene, which is responsible for testis determination. Since genomic clones can contain introns, this kind of analysis is aided by knowledge of splice site consensus sequences (see Chapter 6).

Exon prediction

Computer programs have recently been developed which use multiple criteria, based on differences in sequence composition between coding and non-coding DNA, and the identification of splice sites, to predict accurately the likely positions of exons in large regions of non-coding DNA (Legouis *et al.*, 1991; Uberbacher and Mural, 1991). For example, Legouis *et al.* (1991) were able to identify coding regions for the Kallman syndrome gene in a 67kb phage contig which had been completely sequenced.

vi) Expansion of triplet repeats

Analysis of the FMR-1 gene in patients with fragile X syndrome revealed that a novel mutational mechanism, involving expansion of a CGG repeat within the gene, was responsible for the disease (Verkerk *et al.*, 1990). Changes in copy number of this repeat are thought to be responsible for the phenomenon of anticipation, a decrease in age of onset and an increase in severity in successive generations. The identification of triplet repeat expansion has significantly aided isolation of the genes responsible for several other diseases that show anticipation, including myotonic dystrophy (Brook *et al.*, 1992) and Huntington's disease (The Huntington's Disease Collaborative Research Group, 1993).

1.1.4 Mutational analysis of candidate genes

The expression pattern and predicted function of a candidate gene for a disorder may provide strong evidence to implicate it in the disease. For example, Kallman syndrome, which is characterized by hypogonadotrophic hypogonadism and absence of a sense of smell, is caused by a defect in neural cell migration, and the candidate gene identified by positional cloning was found to have high homology to proteins involved in neural cell adhesion and axonal pathfinding (Franco *et al.*, 1991; Legouis *et al.*, 1991). However, it should be noted that final proof of a gene's involvement in a disorder is the finding of non-polymorphic mutations in affected individuals.

1.2 Mapping of the human X chromosome

The X chromosome contains ~150Mb of DNA, representing ~5% of the human haploid genome, and is present in two copies in females, but only one in males. The distinctive inheritance pattern of X-linked genes has made this chromosome particularly amenable to genetic study. A mechanism of dosage compensation ensures that the products of most X-linked genes are equally expressed in males and females. In placental mammals, this involves the random inactivation of one X chromosome during early development of female cells, followed by clonal inheritance of this state in the somatic descendants of these cells (Lyon, 1961). The inactivation process is accompanied by condensation; the tightly coiled, hypermethylated inactive X chromosome is visible as a body of heterochromatin in the nucleus, known as a 'Barr body' (Barr and Bertram, 1949). Although most of the inactive X is genetically silent, several genes, interspersed throughout the chromosome, have been found to escape inactivation, including that for steroid sulphatase (STS) (Shapiro *et al.*, 1978), the pseudoautosomal region of distal Xp (Goodfellow *et al.*, 1984) and ubiquitin-activating enzyme E1 (UBE) (Brown and Willard, 1990). In addition, a locus in Xq13, known as XIST, is only transcribed from the inactive X, and is a candidate for the X inactivation centre (XIC) (Brown *et al.*, 1991).

Females who are heterozygous for X-linked genes are usually cellular mosaics due to the process of X-inactivation described above. However, skewing of X-inactivation patterns may sometimes occur in a particular tissue, either by chance, or as a consequence of selection at the cellular level, which can be in favour of or against a particular mutant allele. Some heterozygous females may therefore show symptoms of a recessive X-linked disease, or fail to manifest a dominant X-linked disorder, as a result of such non-random inactivation. For example, asymptomatic females who are carriers for X-linked immunodeficiencies (such as Wiskott-Aldrich syndrome, described below) show skewed X-inactivation patterns in haematopoietic cell lineages due to selection against cells expressing the mutant allele (Gealey *et al.*, 1980). Analysis of X-inactivation status has therefore become a useful tool for identification of carriers in pedigrees segregating such disorders (Goodship *et al.*, 1991).

Consistent skewing of X-inactivation is sometimes found in all tissues of an individual, usually as a consequence of structural abnormalities involving the X chromosome. For example, females who have a deletion from one X chromosome may be only minimally affected, due to preferential inactivation of this defective X chromosome in virtually all cells. By contrast, females with balanced X;autosome translocations tend to show biased inactivation of the intact X, because cells which inactivate the translocated X will be disomic for part of the X, and monosomic for some genes on the translocated autosome (due to spreading of inactivation from the XIC) and will therefore be selected against. Thus, in cases where the translocation interrupts a recessive X-linked disease gene, females will manifest the disorder and the analysis of such breakpoints can aid the isolation of the locus involved.

1.3 The hypervariable locus DXS255 and X-linked genetic disease

1.3.1 Isolation of a highly polymorphic VNTR in Xp11.22

One strategy for the generation of novel X-specific probes involves the construction of cosmid libraries from human X-only/rodent hybrid cell lines, selection of those clones which hybridize to total human DNA, and subsequent identification of any single copy restriction fragments contained in these clones. Using such an approach, Fraser *et al.* (1987) isolated a 2.3kb *Eco*R1 fragment (M27β), which, on hybridization to digests of human genomic DNA, was found to detect RFLPs for several different enzymes at a high frequency. Pedigree analysis indicated that these RFLPs were inherited in a Mendelian fashion (Fraser *et al.*, 1987). Size differences between alleles were found to be consistent for a series of different restriction enzymes, suggesting that the hypervariability of the genomic locus recognized (designated DXS255) was due to the presence of a VNTR-like motif (Fraser *et al.*, 1989). This hypothesis was supported by comparison between the restriction map of the M27 cosmid (from which M27β was isolated) and that of the corresponding genomic region in the parent cell line, which indicated that a 3kb deletion had occurred from a site within a *Hae*III-*Sau*3A fragment during cloning of the cosmid. (Previous studies, such as those by Wyman and White (1980) have shown that regions containing VNTR sequences are refactory to cloning.)

Sequencing of the 334bp *Hae*III-*Sau*3A fragment of M27β revealed the presence of three complete copies and one partial copy of a 26bp repeat (Figure 1.1) and it was proposed that the variability of the DXS255 locus is a consequence of variation in the copy number of these repeats. Each complete 26bp repeat unit contains a 10bp perfect inverted repeat separated by 3 nucleotides (Figure 1.1). This motif has the potential to form a cruciform loop structure by intrastrand basepairing, a feature which has not been found in other VNTR sequences. It has been postulated that the capacity of M27β to form multiple cruciform structures may provide substrates for recombination-promoting enzymes, and that this might contribute to the hypervariability at this locus (Fraser *et al.*, 1989).

DXS255 was initially assigned to Xp11.4-Xcen by mapping against a somatic cell hybrid panel. This localization was further refined to Xp11.22 using *in situ* hybridization to replication banded chromosomes from a normal female 46, XX lymphoblastoid cell line (Fraser *et al.*, 1989).

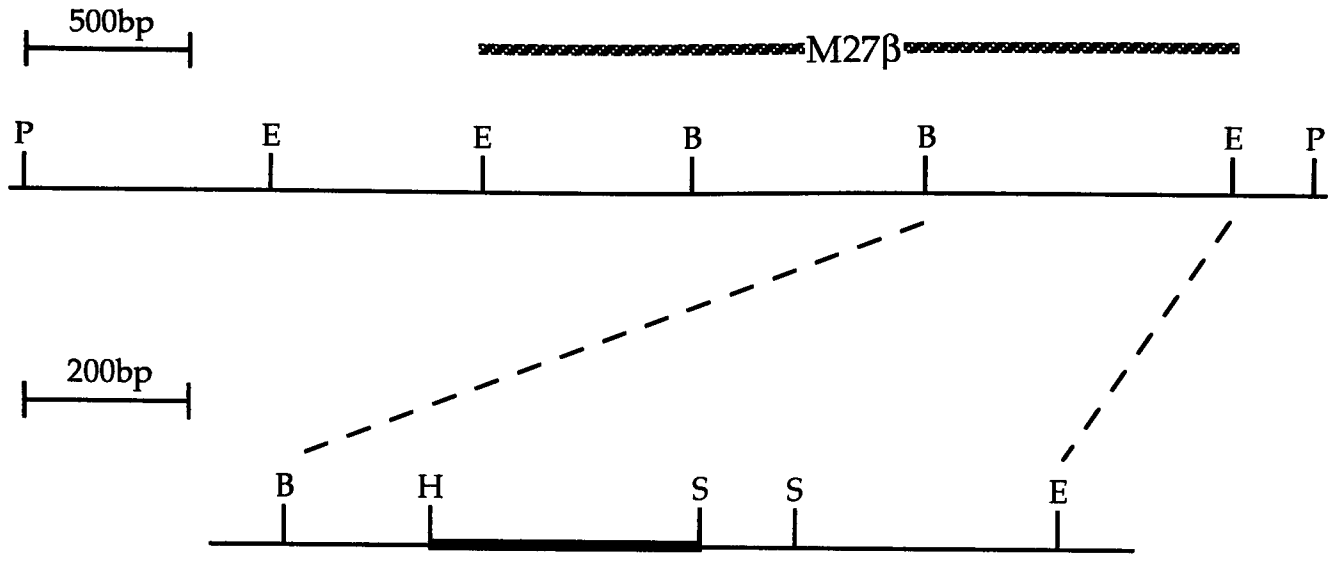
1.3.2 Disease genes closely linked to DXS255

With a heterozygosity value in excess of 90%, the DXS255 locus has facilitated the localization of several different disease genes which map to the proximal short arm of the X chromosome (Figure 1.2):

i) X-linked Retinitis Pigmentosa (XLRP)

Primary retinitis pigmentosa (RP) is a heterogeneous form of retinal degeneration which may be transmitted through autosomal dominant, autosomal recessive, or X-linked recessive modes of inheritance. There are no consistent clinical differences between patients suffering from different subtypes of the disease, but XLRP is usually more severe than its autosomal counterparts and is characterized by an early onset of night blindness (within the first two decades of life), with progression to severe visual impairment by the age of 30 (McKusick OMIM 312600).

a)



b)

```
  1 GATCATGGGTAAGTTCAGGAGTTTCTAGGGTAGGGTGAGGTGGGGAGGAGAGTCCTGAGG
 61 AAAAAGTGGTTATTTATCTGTAGCTACTGAAGTATTTCAGTATTTAGCAAACTGGTATGG
121 CATACAGATGCATATCAACAGAATAGCTGCCCTGAATAAGTGTCCCAGGGGTCAAACTAA
181 GAGTACACTAAAGCATAATGTGGTCCTGGATAGATACTATCCAGGACTCTCCTGGATAGA
241 TACTATCCAGGACTCTCCTGGATAGATACTATCCAGGACTCTCCTGGATAGTTTCTTAGA
301 TGGGAGGGCTACTCCAACTACATTTGCTTCTTGG
```

c)

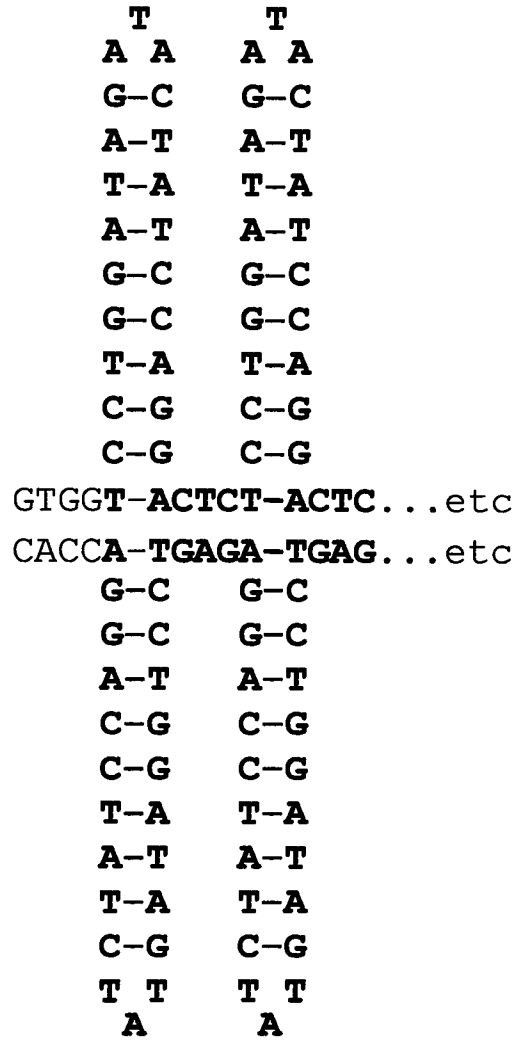


Figure 1.1: Discovery of a VNTR in the cosmid subclone M27β. Adapted from Fraser *et al.* (1989). **a)** Restriction map of cosmid M27 in the region from which M27β was derived (Top) for the enzymes *Bgl*II (B), *Eco*R1 (E) and *Pst*I (P). The fine map for the *Bgl*II/*Eco*R1 fragment is shown below, with positions of sites for *Hae*III (H) and *Sau*3A (S) indicated. The bold region represents the *Hae*III/*Sau*3A fragment from which a ~3kb deletion had occurred during cloning (see text). **b)** Nucleotide sequence of the *Hae*III/*Sau*3A fragment, including three complete copies, and one partial copy of a 26bp repeat, depicted in boldface. The 'core' 26bp repeat is underlined. **c)** Suggested cruciform structure that can be adopted by the repeating motif.

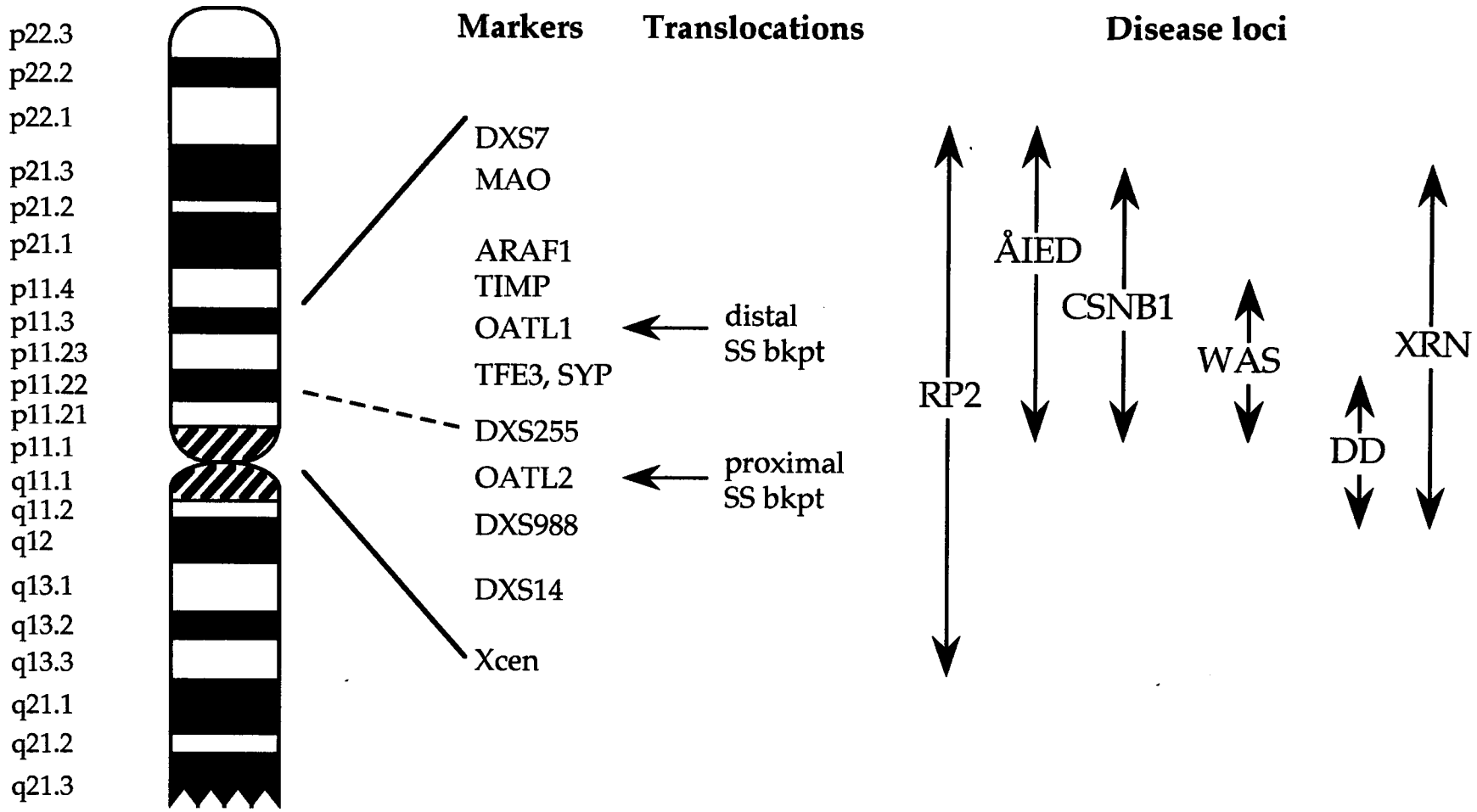


Figure 1.2: Schematic representation of the X chromosome, indicating the relative positions of disease genes which are closely linked to DXS255, as deduced from studies described in the text. Markers which were important for the delimitation of target regions are also shown. **SS**, synovial sarcoma; **RP2**, retinitis pigmentosa 2; **ÅIED**, Åland Island eye disease; **CSNB1**, congenital stationary nightblindness 1; **WAS**, Wiskott-Aldrich syndrome; **DD**, Dent's disease; **XRN**, X-linked recessive nephrolithiasis. Note that recombinants map X-linked thrombocytopaenia to the interval between DXS7 and DXS17 (in Xq22) (de Saint-Basile *et al.*, 1991) and this delimitation is not included in this diagram. The t(X;1) (p11.2; q21.2) translocation associated with papillary renal cell carcinoma has also been excluded from this figure, due to the ambiguities in its mapping, described in the text.

Pedigree analysis initially demonstrated close linkage between XLRP and the locus DXS7 (detected by the probe L1.28) in Xp11.3 (Bhattacharya *et al.*, 1984). Further studies have suggested that XLRP pedigrees can be divided into two distinct groups, one in support of a disease locus mapping proximal to DXS7, the other supporting a location distal to DXS7 (Wirth *et al.*, 1988). It has therefore been proposed that there are two separate loci for XLRP on the short arm of the X chromosome.

The more distal locus, known as RP3 (McKusick 312612), has been localized to Xp21 by analysis of deletions in affected males, including the large interstitial deletion of a patient manifesting DMD, CGD, XK, RP and mental retardation (Francke *et al.*, 1985), which was described in Section 1.1.2. Additional studies have shown that RP3 maps distal to OTC, but proximal to CGD within a 205kb *Sfi*I fragment in Xp21.1 (van Ommen *et al.*, 1986; Ott *et al.*, 1990; Musarella *et al.*, 1991).

By contrast, there are no known cytogenetic abnormalities associated with the more proximal locus on Xp, known as RP2, which might aid in its localization. Family studies using M27β have demonstrated that RP2 is closely linked to DXS255 (Meitinger *et al.*, 1989), but precise delimitation of the position of the disease locus has been hindered by the small number of RP2 pedigrees, and a lack of highly informative markers in the region. A further complication involves the possibility of misclassifying a family segregating RP3, or a form of autosomal dominant RP, as an RP2 pedigree, since there is no reliable method for distinguishing these on the basis of clinical diagnosis. Although it has been reported that some female carriers of RP3 exhibit a metallic luster (known as a tapetal reflex) on fundoscopic examination, there is no evidence for a clear correlation between absence of a tapetal reflex and RP2 (Meitinger *et al.*, 1989). As a result of the above difficulties, little progress has been made in the localization of RP2, which may map anywhere in the DXS7-Xcen interval.

Two other eye diseases have been shown to map to proximal Xp, in similar regions to RP2. A form of **X-linked congenital stationary night blindness (CSNB1)** (McKusick 310500), in which affected males have non-progressive disturbed or absent night vision, reduced visual acuity, congenital nystagmus and myopia , has been localized to the MAO-DXS255 interval by linkage analysis (Gal *et al.*, 1989; Musarella *et al.*, 1989).

In addition, **Åland Island eye disease (ÅIED)** (McKusick 300600), which is characterized by albinism of the fundus, foveal hypoplasia, nystagmus, myopia, astigmatism and protanomalous colour blindness, maps between DXS7 and DXS255 (Alitalo *et al.*, 1991; Schwartz and Rosenberg, 1991; Glass *et al.*, 1993). It is interesting to note that different mutations in the same gene, that encoding rhodopsin, can cause autosomal RP or an autosomal form of CSNB (Dryja *et al.*, 1990, 1993). It therefore seems plausible that RP2, CSNB1 and ÅIED may be allelic disorders due to mutations at a single locus which lies in the MAO-DXS255 interval (Figure 1.2).

ii) Wiskott-Aldrich syndrome (WAS)

The Wiskott-Aldrich syndrome (McKusick 301000) is an X-linked recessive disease, characterized by immunodeficiency, eczema and severe thrombocytopaenia, with platelets of reduced size and function. The disorder has an estimated incidence of 4 per million live births, and affected males have a median survival age of 6 years, with recurrent infections, haemorrhage, and lymphoreticular malignancies as the main causes of early mortality. Carrier females usually have no clinical or immunological abnormalities (but can be diagnosed on the basis of non-random X-inactivation as described below).

Prior to commencement of the work described in this thesis, the genetic defect underlying WAS was unknown. Two heavily O-glycosylated membrane proteins, Gp1b in platelets and sialophorin (CD43) in T-lymphocytes, are structurally unstable in WAS patients, and it has been postulated that an abnormality in the O-glycosylation pathway may be responsible for the disease phenotype (Greer *et al.*, 1990).

The first linkage studies of WAS indicated that it mapped to the pericentromeric region of Xp, between the polymorphic markers DXS7 (Xp11.3) and DXS14 (Xp11.21) (Peacocke and Siminovitch, 1987; Kwan *et al.*, 1988). Further studies demonstrated very close linkage between WAS and the newly identified DXS255 locus at Xp11.22 (de Saint-Basile *et al.*, 1989; Greer *et al.*, 1990). A more precise localization, placing WAS distal to DXS255, but proximal to TIMP (a gene encoding a tissue inhibitor of metalloproteinases in Xp11.3), with an estimated genetic distance of ~3cM between the two markers, was reported by Kwan *et al.* (1991) (Figure 1.2).

Analysis of pedigrees segregating a disorder characterized by thrombocytopaenia, but with no immunodeficiency component, have indicated that this disease, known as **X-linked thrombocytopaenia (XLT)** (McKusick 313900), maps in a similar region of Xp to WAS, suggesting that the two disorders may be caused by different mutations in the same gene (de Saint-Basile *et al.*, 1991). It has been demonstrated that whilst all cells of the haematopoietic lineage display non-random patterns of X-inactivation in carrier females from WAS pedigrees, carrier females from families with X-linked thrombocytopaenia have skewed inactivation in T- and B-lymphocytes, but a random inactivation pattern in granulocytes (de Saint-Basile *et al.*, 1991).

In 1990, Lyon *et al.* reported similarities between the phenotype of Wiskott-Aldrich patients and the X-linked scurfy (*sf*) mouse mutant, which is characterized by scaliness of the skin (possibly due to eczema) and haemotological abnormalities, including thrombocytopaenia, resulting in death of affected males at about 3-4 weeks of age. It has since been shown that *sf* lies in the same ~0.6-3cM interval of the mouse X-chromosome as *Gf-1*, the mouse homologue of GATA1 (Laval and Boyd, 1993). These studies also demonstrated that GATA1 maps between TIMP and DXS255 on the human X-chromosome and is thus localized to the same interval as WAS (Laval and Boyd, 1993). Comparative mapping therefore supported the possibility that scurfy and WAS might be homologous disorders.

iii) Dent's disease

In 1993, after the work described in this thesis had begun, Pook *et al.* demonstrated close linkage between DXS255 and Dent's disease (McKusick 600248), a familial renal tubular disorder which is characterized by low molecular weight proteinuria, hypercalciuria, nephrocalcinosis, nephrolithiasis (kidney stones) and eventual renal failure. In addition, they found that a microdeletion of Xp11.22, encompassing the DXS255 locus, was associated with the disorder in one pedigree (Pook *et al*, 1993). This deletion was shown to be flanked by the genes TFE3 and SYP on one side, and by the marker DXS988 on the other (Figure 1.2). A more detailed description of the clinical features, linkage studies and deletion analysis of Dent's disease is given in Chapter 5.

A similar disorder affecting the renal proximal tubules, referred to as **X-linked recessive nephrolithiasis (XRN)** (McKusick 310468) is also closely linked to DXS255 in proximal Xp (Scheinman *et al.*, 1993). This differs from Dent's disease in that patients do not manifest urinary acidification defects or metabolic bone disease (Fryomyer *et al.*, 1991). In addition haplotype analysis of a large 5 generation family segregating XRN identified two recombination events which localized this disorder to the MAOB-ARAF1 interval in Xp11.4-Xp11.23 (Thakker *et al.*, 1994). These studies suggested the existence of two distinct X-linked genes associated with nephrolithiasis, one in Xp11.22 and the other in Xp11.4-Xp11.23. However, more recent analysis has indicated that the preliminary mapping of XRN distal to ARAF1 was incorrect, due to the misclassification of a mildly affected individual as unaffected (R. V. Thakker, personal communication) and the genetic mapping interval for XRN has therefore been revised to Xp11.4-Xp11.22 (Figure 1.2). This raises the possibility, discussed in Section 5.1.2, that Dent's disease and XRN may be caused by mutations in the same gene in Xp11.22.

1.3.3 An X-inactivation assay using DXS255

As described above (Section 1.2) the detection of non-random X-inactivation patterns in haematopoietic cells is a useful diagnostic tool for identifying asymptomatic female carriers from pedigrees segregating X-linked immunodeficiencies. Skewing of X-inactivation in lymphocytes of WAS carriers was first demonstrated by analysing the expression of glucose-6-phosphate dehydrogenase (G6PD) isoenzymes (Gealey *et al.*, 1980). Few females are heterozygous for G6PD polymorphisms, but direct DNA analysis has since provided a more informative method for the assessment of X-inactivation status. This involves the use of RFLP-detecting probes which also detect methylation differences between the active and inactive X chromosomes (Boyd and Fraser, 1990). Probes from the 5' ends of the phoshoglycerate kinase (PGK) and hypoxanthine phosphoribosyltransferase (HPRT) genes were initially employed for such studies, but their combined heterozygosity has been estimated as less than 50% (Goodship *et al.*, 1991)

The discovery that one of the CCGG sites flanking DXS255 is fully methylated on active X chromosomes, but unmethylated on most inactive X chromosomes, enabled an X-inactivation assay to be developed using this locus (Boyd and Fraser, 1990). In this technique, DNA which has been digested with the restriction enzyme *Pst*I (to distinguish between parental alleles) is further digested with either *Msp*I (which cleaves CCGG sites) or its methylation-sensitive isoschizomer *Hpa*II, and then probed with M27β. If the X-inactivation pattern is unilateral then only one of the two fragments detected in the *Pst*I/*Msp*I digest will be present in the *Pst*I/*Hpa*II digest, and this band originates from the inactive X chromosome (Figure 1.3a). This assay is an improvement over those involving PGK and HPRT due to the high level of heterozygosity of DXS255 (Fraser *et al.*, 1989). It has also been useful for the determination of clonality in female tumour cell lines (Abrahamson *et al.*, 1990).

In 1992, Hendriks *et al.* reported the presence of additional *Msp*I sites within 200bp of the *Msp*I site used in the above X-inactivation assay. Sequence analysis of a 2kb *Bgl*II-*Pst*I fragment containing these sites revealed that they are located within a CpG island that is associated with the 5' end of a LINE-1 repetitive element (Figure 1.4). This CpG island also contains one *Bss*HII site and seven *Hha*I sites. Further analysis demonstrated that, although all the CpG sites were extensively methylated on active X chromosomes, methylation patterns on the inactive X were heterogeneous, varying between different tissues and different samples. However, it was found that at least one *Hha*I site was completely unmethylated on the inactive X in all samples analysed. On the basis of this, a more reliable X-inactivation assay was developed, in which DNA digested with *Pst*I is compared to that digested with *Pst*I and *Hha*I (Figure 1.3b). In this assay, the relative intensities of the two allelic *Pst*I-*Pst*I fragments detected in the *Pst*I-*Hha*I digest reflect the proportions of cells that have either of the two X chromosomes active.

Previous studies have demonstrated a correlation between methylation of CpG islands and transcriptional repression of the genes associated with them (Bird, 1986). The CpG islands of X-linked genes such as G6PD and HPRT are methylated on the inactive X and unmethylated on the active X (Bird, 1986). By contrast the DXS255 CpG island shows the inverse correlation, with methylation on the active chromosome.

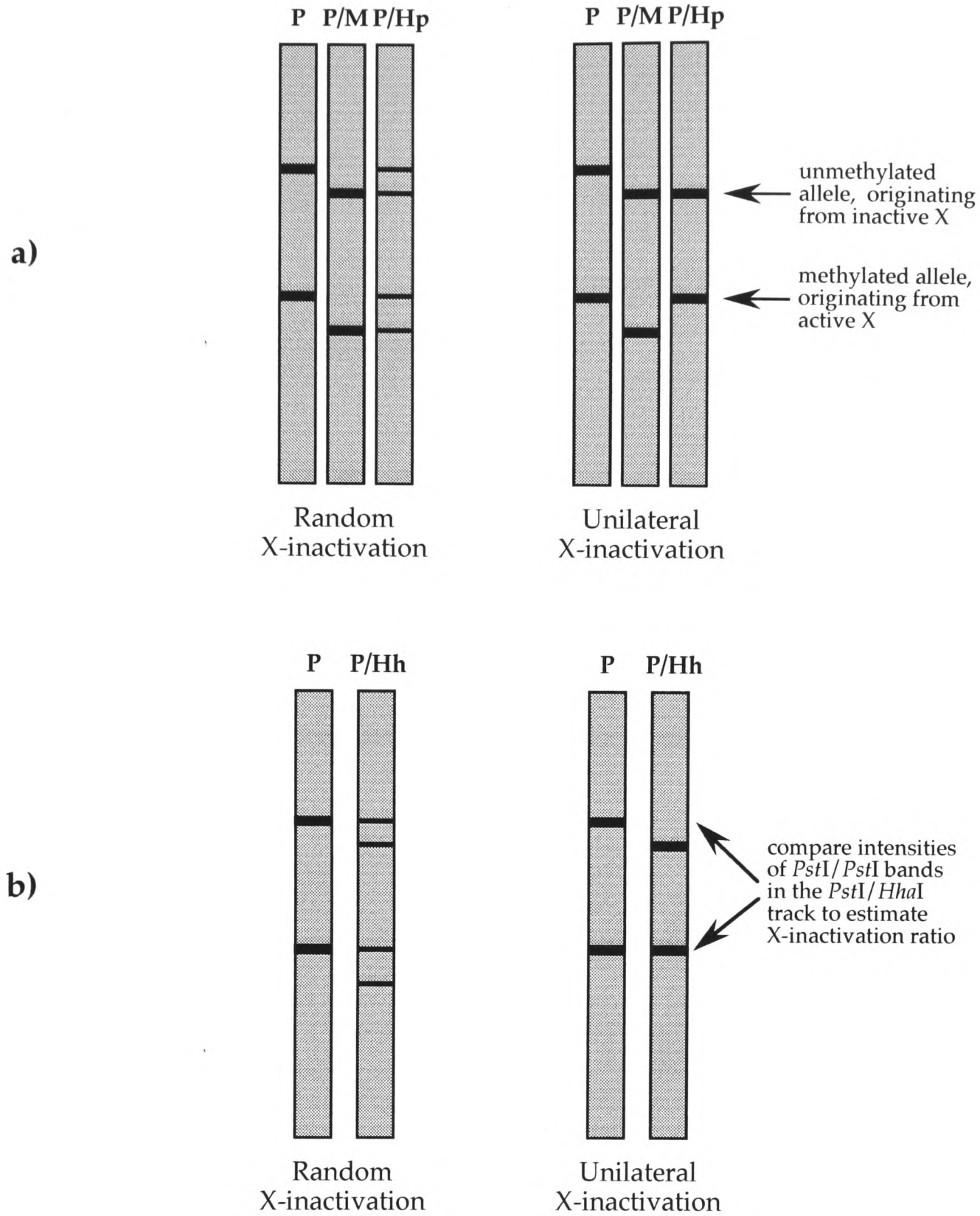


Figure 1.3: Schematic representation of X-inactivation assays developed using the DXS255 locus. Tracks correspond to human female genomic DNA digested to completion with *Pst*I (P), *Pst*I/*Msp*I (P/M), *Pst*I/*Hpa*II (P/Hp) or *Pst*I/*Hha*I (P/Hh), and then probed with M27β.

a) Original assay. A *Pst*I digest is used to indicate whether or not the female is heterozygous for DXS255, and establishes the sizes of the alleles. The X-inactivation pattern may be deduced by comparing the bands seen in the *Pst*I/*Msp*I track to those in the *Pst*I/*Hpa*II track as described in the text. This figure shows the type of result that would be expected with random (50:50) inactivation **(left)** or complete skewing (i.e. 100:0) towards one allele **(right)**, but it should be noted that many cell populations will display intermediate X-inactivation ratios. Furthermore, this diagram represents the ideal situation in which sites are completely methylated on the active X and completely unmethylated on the inactive X. Analysis by Hendriks *et al.* (1992) showed that methylation patterns of CCGG sites associated with the DXS255 locus are heterogenous on the inactive X (see text).

b) Modified assay proposed by Hendriks *et al.* (1992), in which comparison of the intensity of *Pst*I/*Pst*I fragments detected in the *Pst*I/*Hha*I track gives an estimate of the inactivation ratio. This assay relies on observations that at least one *Hha*I site in the CpG island (Figure 1.4) is completely unmethylated on the inactive X in all samples, so that *Pst*I/*Pst*I fragments observed in the *Pst*I/*Hha*I track are exclusively derived from the active X (see text).

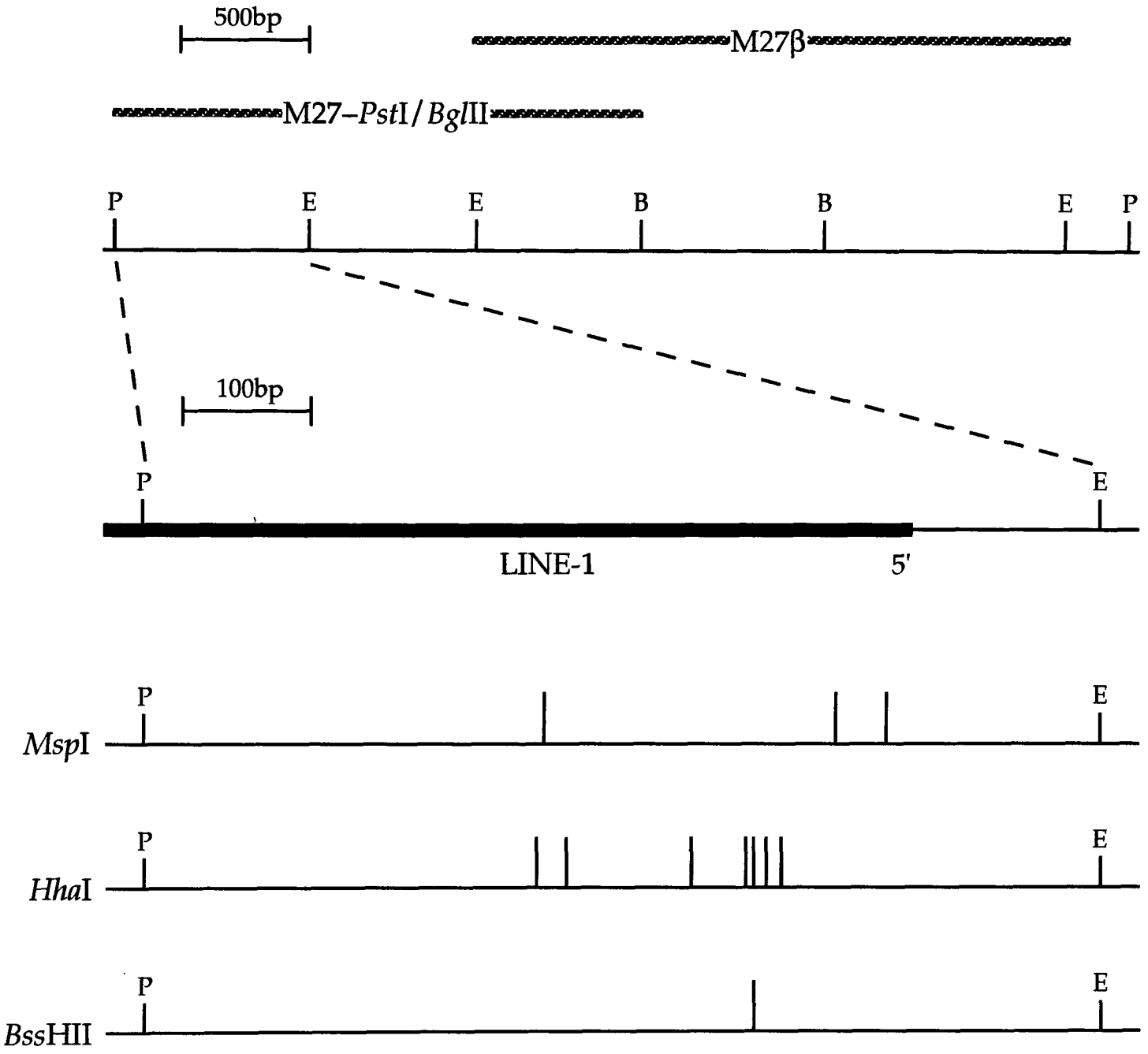


Figure 1.4: A CpG island associated with the 5' end of a LINE-1 element adjacent to the hypervariable DXS255 locus. Restriction map around the DXS255 locus in cosmid M27 **(top)** for the enzymes *Bgl*II (B), *Eco*R1 (E) and *Pst*I (P). The 2.3kb M27β *Eco*R1-*Eco*R1 fragment (which contains the hypervariable VNTR) and the overlapping 2.0kb *Bgl*II/*Pst*I fragments are shown. Sequencing of the latter identified the first 599bp of a LINE-1 element (depicted by a thick line). The positions of *Msp*I, *Hha*I and *Bss*HII sites within the CpG island of this 599bp segment, are also shown **(bottom)**. These sites show differential methylation on active and inactive X chromosomes (see text). Adapted from Hendriks *et al.* (1992).

The finding that the CpG island at this locus is associated with a LINE-1 element suggests a possible explanation for these observations. LINE-1 elements are members of a long interspersed repetitive sequence family in mammalian genomes and are thought to amplify within the genome via a retrotransposition mechanism involving RNA intermediates (see section 3.1.4). When *de novo* insertion of such an element occurs, it may result in the disruption of an important gene, as observed in a case of haemophilia which was caused by LINE-1 insertion into the factor VIII gene (Kazazian *et al.*, 1988). Transcriptional repression of LINE-1 promoters, involving methylation of their CpG islands, is therefore likely to be important in order to maintain them in a quiescent state. This hypothesis is supported by studies of CpG islands in autosomal LINE-1 promoters (Woodcock *et al.*, 1988). Whilst this explains why the DXS255 LINE-1 CpG island is methylated on the active X, it does not account for the general lack of methylation on the inactive X. Hendriks *et al.* (1992) suggested that in the latter case, the inactive status of the chromosome might be sufficient to silence LINE-1 transcription without the need for stringent methylation of the promoter region.

1.3.4 Translocation breakpoints mapping to Xp11.23-p11.22

Independent translocation breakpoints associated with two different tissue-specific tumours have been localized to the Xp11.23-p11.22 interval:

i) Synovial sarcoma (SS)

The soft tissue carcinoma-sarcoma known as synovial sarcoma (McKusick 312820) is characterized by a cytogenetic abnormality involving a t(X;18)(p11.2;q11.2) translocation (Reeves *et al.*, 1989; Gilgenkrantz *et al.*, 1990). Initial evidence suggesting the presence of two distinct translocation breakpoints in Xp11.2 associated with SS was confirmed by FISH studies (Knight *et al.*, 1992; de Leeuw *et al.*, 1993). These breakpoints are coincident with two non-adjacent clusters of homologous sequences known as OATL1 and OATL2, in Xp11.23-p11.22 and p11.21 respectively (Figure 1.2 and see below).

ii) Papillary renal cell carcinoma (RCC)

A t(X;1) (p11.2;q21.2) reciprocal translocation has been identified as the underlying genetic rearrangement in several cases of papillary renal adenocarcinoma (de Jong *et al.*, 1986; Meloni *et al.*, 1993; McKusick 312390). Between 15 and 20% of renal cell carcinomas (the most common cancer of the kidney) fall into the papillary tumour subgroup, and these may be either sporadic or familial in nature. A similar region of the X chromosome has been implicated in two further papillary renal tumours, one associated with a t(X;17) (p11.2;q25) translocation (Tomlinson *et al.*, 1991), the other involving a deletion in Xp11 (Ohjimi *et al.*, 1993). FISH studies revealed that a YAC from the OATL2 cluster (see below) detected signals on der(X) and der(1), suggesting that the breakpoint maps to Xp11.21 (Suijkerbuijk *et al.*, 1993). However, analysis of somatic cell hybrids derived from renal cell tumours supported a more distal location for the breakpoint, placing it between the synaptophysin (SYP) gene and DXS146 in Xp11.23-p11.22 (Sinke *et al.*, 1993). These contrasting results are unlikely to be a consequence of breakpoint heterogeneity (as observed with SS), since a tumour from the same patient was investigated in both studies. Further analysis is therefore necessary in order to identify more precisely the position of this breakpoint in Xp11.2. It is interesting to note that these initial mapping studies placed the breakpoint in a similar region to that implicated in Dent's disease, and thus suggested that disruption of the same locus might be involved in RCC and proximal tubular dysfunction.

1.3.5 An overview of physical mapping in the Xp11.23-p11.22 region

The cloned genes TFE3 (Henthorn *et al.*, 1991), SYP (Ozcelik *et al.*, 1990) and GATA1 (Caiulo *et al.*, 1991), and the anonymous polymorphic marker DXS146 (Kruse *et al.*, 1986) were localized to the Xp11.23-p11.22 region, in the vicinity of DXS255, by a combination of linkage analysis, *in situ* hybridization, somatic cell hybrid studies and pulsed field genomic mapping (Lafreniere *et al.*, 1991b; Cremin *et al.*, 1993; Riley, 1993; Laval and Boyd, 1993). These studies are described in more detail in the introductions to Chapters 3 and 4. Together, they showed that these loci all map to the interval between two homologous clusters (OATL1 and OATL2) containing pseudogene copies of the OAT gene (Lafreniere *et al.*, 1991a), and established Xpter–OATL1–GATA–(SYP, TFE3)–DXS255–DXS146–OATL2–Xcen as the most likely marker order for this region.

1.4 Outline of this study

The work described in this thesis can be divided into two parts:

I) Construction and characterization of YAC contigs in the Xp11.23-p11.22 interval

The original aim of this project was to generate more data on the physical mapping of the region around DXS255 in order to provide a basis for the isolation of disease loci mapping to Xp11.23-p11.22. Chapter 3 describes a bi-directional YAC walk, centred on DXS255, involving the isolation of eight new YAC clones, and the generation of eight novel markers, one of which contains a polymorphic CA repeat. The most proximal YAC of the contig thus constructed provides a physical link with a YAC cluster around DXS146. A rare-cutter restriction map of the YAC contig, containing four putative CpG islands, is also presented. Chapter 4 gives details of a more distal contig encompassing the genes GATA, TFE3 and SYP, which comprises seven new YACs and is linked to the distal OATL1 YAC cluster. In addition, this chapter describes the identification of a putative novel calcium channel gene in cosmids from the region. The value of these physical mapping resources for future studies of Xp11.23-p11.22 is discussed.

II) Isolation and analysis of a novel chloride channel gene implicated in Dent's disease

The discovery of a microdeletion involving DXS255 in patients affected with Dent's disease provided the focus for the remainder of this project. Chapter 5 describes the use of direct cDNA screening with a DXS255 YAC to identify successfully a candidate gene for this disorder. This chapter also presents data from Northern blots, sequence analysis and homology searches, which suggest that the candidate gene (CLCN5) encodes a novel kidney-specific chloride channel. Elucidation of the genomic organization of the CLCN5 coding region is described in Chapter 6, and the application of this information for mutation screening of individuals with hypercalciuric nephrolithiasis is discussed. Possible mechanisms to explain why disruption of CLCN5 should lead to renal tubular dysfunction are suggested in the General Discussion (Chapter 7).

Chapter 2 – Materials and Methods

2.1 Buffers

Church hybridization buffer: 7% (w/v) SDS (Sigma, 99% pure), 0.5M Na_2HPO_4 (pH 7.2), 1% (w/v) BSA (Sigma; stored as a 10% solution, prefiltered through a Sartorius 0.45µm minifilter), 1mM Na_2EDTA (pH 8.0)

Denaturing solution: 1.5M NaCl, 0.5M NaOH

50 x Denhardt's solution: 1% (w/v) Ficoll-400 (Pharmacia), 1% (w/v) BSA (Fraction V, Sigma), 1% (w/v) polyvinylpyrrolidone (Sigma)

Filter stripping solution: 2mM Tris-Cl (pH 7.5), 1% (w/v) SDS, 1mM Na_2EDTA

GTE: 50mM glucose, 25mM Tris-Cl (pH 8.0), 10mM Na_2EDTA

Lambda oligomerization buffer: 2 x SSC, 10mM Na_2EDTA, 10mM Tris-Cl (pH 7.5)

Lambda dilution buffer: 0.1M NaCl, , 10mM $MgSO_4 \cdot 7H_2O$, 35 mM Tris-Cl (pH 7.5), 0.01% (w/v) gelatin

LiDS: 100mM Na_2EDTA, 10mM Tris-Cl (pH 8.0), 1% (w/v) lithium dodecyl sulphate (Sigma)

5 x Ligase buffer: 250mM Tris-Cl (pH 7.6), 10mM $MgCl_2$, 1mM ATP, 1mM DTT (Sigma), 5% (w/v) polyethelene glycol-8000

NDS: 0.5M Na_2EDTA, 10mM Tris-base, 1% (w/v) sodium lauryl sarcosine (Sigma) (pH 9.5)

Neutralizing solution: 1.5M NaCl, 1M Tris-Cl (pH 8.0)

Phosphate buffered saline (PBS): 2.6mM KH_2PO_4, 26mM Na_2PO_4, 145mM NaCl (pH7.4)

Restriction enzyme dilution buffer (REDB): 50mM KCl, 10mM Tris-Cl (pH 7.4), 100µM Na_2EDTA, 1mM DTT, 0.02% (w/v) BSA, 50% (v/v) glycerol

SEB: 1M sorbitol, 20mM Na_2EDTA, 14.4mM β-mercaptoethanol

20 x SSC: 3M NaCl, 0.3M sodium citrate (pH 7.0)

20 x SSPE: 3M NaCl, 0.2M NaH_2PO_4, 20mM Na_2EDTA (pH 7.4)

STEB: 1M sorbitol, 20mM Na_2EDTA, 14.4mM β-mercaptoethanol, 10mM Tris-Cl (pH 7.5)

STET: 5% (w/v) sucrose, 50mM Na_2EDTA, 5% (v/v) Triton X-100 (Sigma), 50mM Tris-Cl (pH 8.0)

STOP (for YAC plugs): 0.5 x TAE, 20mM Na_2EDTA, Orange-G dye (Sigma)

50 x TAE: 2M Tris-acetate, 50mM Na_2EDTA

10 x TBE: 0.9M Tris-borate, 20mM Na_2EDTA (pH 8.3)

TE: 10mM Tris-Cl (pH 7.4), 1mM Na_2EDTA (pH 8.0)

TEN: 150mM NaCl, 10mM Tris-Cl, 10mM Na_2EDTA (pH 8.0)

2.2 Media and antibiotics

LB broth: 1% (w/v) Bacto tryptone (Difco), 0.5% (w/v) Bacto yeast extract, 1% (w/v) NaCl

LB agar plates: LB + 1.5% (w/v) agar

LB/$MgSO_4$ bottom agar: LB + 10mM $MgSO_4$ + 1.5% (w/v) agar

LB/$MgSO_4$ soft top agar: LB + 10mM $MgSO_4$ + 0.7% (w/v) agar

SD + CAT: 0.67% (w/v) Bacto yeast nitrogen base without amino acids (Difco), 2% (w/v) glucose, 1.4% (w/v) Bacto casamino acids (Difco), 55mg/L adenine hemisulphate (Sigma), 55mg/L tyrosine (Sigma)

SD + CAT agar plates: SD + CAT + 2% (w/v) agar

SOC: 2% (w/v) Bacto tryptone (Difco), 0.5% (w/v) Bacto yeast extract (Difco), 10mM NaCl, 2.5mM KCl, 10mM $MgCl_2$, 10mM $MgSO_4 \cdot 7H_2O$, 20mM glucose

2 x TY broth: 1.6% (w/v) Bacto tryptone (Difco), 1% (w/v) Bacto yeast extract, 0.5% (w/v) NaCl

2 x TY agar plates: 2 x TY + 1.5% (w/v) agar

Ampicillin (Sigma): Stock solution 25mg/ml, used in media at 50µg/ml

Kanamycin (Sigma): Stock solution 25mg/ml, used in media at 20µg/ml

X-Gal (Gibco-BRL): Stock solution 2% (w/v) in dimethylformamide, used in media at 0.004%

2.3 Bacterial strains

K802: *gal*K2, *gal*T22, *hsd*R2(r-$_K$, m+$_K$), *lac*Y1, *mcr*A-, *mcr*B-, *met*B1, *mrr*$^+$, *sup*E44

OneShot™ (Invitrogen): F', *end*A1, *rec*A1, *hsd*R17(r-$_K$, m+$_K$),λ-, *sup*E44, *thi*-1, *gyr*A96, *rel*A1, ø80Δ*lac*ΔM15, Δ (lacZYA-argF)U169, *deo*R+

Top10 (Invitrogen): *mcr*A, Δ(*mrr-hsd*RMS-*mcr*BC), ø80ΔlacΔM15, Δ*lac*X74, *deo*R, *rec*A1, *ara* D139, Δ(*ara, leu*) 7679, *gal*U, *gal*K, λ-, *rps*L (Str^r), *end*A1, *nup*G

2.4 Preparation of plasmid DNA

2.4.1 Boiling miniprep

This method was used for analysis of large numbers of clones following subcloning experiments:

1. Resuspend half a streak of cells in 105µl STET.

2. Add 7.5µl TEN containing lysozyme (Sigma) at 10mg/ml. Leave for 5 minutes at room temperature.

3. Add 1µl diethylpyrocarbonate (DEPC, Sigma).

4. Place in a boiling water bath for 1 minute.

5. Spin cells in an Eppendorf centrifuge for 15 minutes at room temperature.

6. Remove pellet of cell debris using a toothpick and discard.

7. Add 11µl 3M sodium acetate (pH 5.2) and 115µl ice-cold isopropanol. Mix and place at -20°C for 30 minutes.

8. Spin in Eppendorf centrifuge for 15 minutes at 4°C. Remove supernatant.

9. Add 500µl of 70% ethanol and leave at room temperature for 5 minutes. Spin for 5 minutes at room temperature. Remove supernatant.

10. Briefly dry the DNA in a vacuum desiccator and dissolve in TE containing DNAse-free RNase (Boehringer Mannheim) at 2.5µg/ml.

11. Incubate at 37°C for 30 minutes. Store at -20°C.

2.4.2 Small scale alkaline lysis

1. Inoculate 10ml of 2 x TY or LB broth, containing the appropriate antibiotic, with a single bacterial colony, or 50µl of a frozen stock.

2. Incubate overnight (or until turbid) in a 37°C shaking incubator.

3. Centrifuge at top speed in an MSE bench centrifuge for 10 minutes.

4. Drain pellet and resuspend in 300µl of GTE containing lysozyme (Sigma) at 5mg/ml.

5. Transfer to a 1.5ml Eppendorf tube and incubate at room temperature for 5 minutes.

6. Add 600µl of freshly made ice-cold 0.2M NaOH/0.1% SDS, mix gently by inverting and stand on ice for 5 minutes.

7. Add 450µl of fresh 5M KAc, pH 4.8 (made by mixing 3ml 5M KAc, 575µl glacial acetic acid and 1425µl water). Mix by vortexing and stand on ice for 20 minutes.

8. Spin in Eppendorf centrifuge for 15 minutes at 4°C to pellet cell debris and bacterial DNA.

9. Remove supernatant. Add 1/2 volume of phenol equilibrated with TE, and vortex. Add 1/2 volume of chloroform (24:1 chloroform:isoamyl alcohol) and vortex again.

10. Separate aqueous and organic phases by spinning in an Eppendorf centrifuge for 5 minutes.

11. Remove aqueous (upper) phase. Add to this two volumes of ice-cold absolute ethanol. Stand at room temperature for 5 minutes.

12. Pellet DNA by spinning in Eppendorf centrifuge for 10 minutes at room temperature. Remove supernatant.

13. Wash pellet in 70% ethanol, spin for 5 minutes at room temperature and discard supernatant.

14. Dry DNA briefly in a vacuum dessicator, and dissolve in 100µl TE containing DNAse-free RNase (Boehringer Mannheim) at 2.5µg/ml.

15. Incubate at 37°C for 30 minutes. Store at -20°C.

2.4.3 Promega plasmid minipreps

These are commercially available kits which give good yields (typically at least 10µg from a 3ml culture) and are quicker than standard alkaline lysis. DNA prepared by this method can be used directly in sequencing reactions without further purification.

1. Obtain a pellet of cells as in steps 1-3 of alkaline lysis (2.4.2).

2. Resuspend cell pellet in 600µl of 'cell resuspension solution' (Promega A711D; 50mM Tris-Cl (pH 7.5), 10mM Na_2EDTA, 100µg/ml RNase A). Divide into three 200µl aliquots and transfer to 1.5ml Eppendorf tubes.

3. To each aliquot add 200µl of 'cell lysis solution' (Promega A712C; 0.2M NaOH, 1% SDS). Mix by inverting tube until a cleared lysate is obtained.

4. Add 200µl of 'neutralization solution' (Promega A713D; 1.32M KAc) to each lysate and mix by inversion. Vortex if necessary.

5. Spin in an Eppendorf centrifuge for 5 minutes at room temperature to pellet cell debris and bacterial DNA.

6. Transfer each supernatant to a new Eppendorf tube. Add 1ml of 'Wizard™ minipreps DNA purification resin' (Promega A767D) and mix by inversion.

7. Attach a 2ml disposable syringe barrel to the Luer-Lok extension of a fresh 'Wizard™ minicolumn' (Promega A712B).

8. Transfer the resin/supernatant mix into the syringe barrel. Insert the syringe plunger and gently push the slurry into the column.

9. Wash the minicolumn as follows: Detach the syringe and remove the plunger. Reattach the syringe to the minicolumn and pipette 2ml of 'column wash solution' (Promega A715D; 0.1M NaCl, 10mM Tris-Cl (pH7.5), 2.5mM Na_2EDTA, 50% (v/v) EtOH) into the syringe. Insert the plunger and push the wash solution through the column.

10. Transfer the minicolumn to a 1.5ml Eppendorf tube and spin for 20 seconds in order to dry the resin.

11. Transfer the minicolumn to a fresh tube and apply 50µl of water or TE. Immediately spin for 20 seconds to elute DNA.

12. Discard minicolumn. The purified DNA is stored at -20°C

2.5 Restriction enzyme digestion

DNA was digested in buffers supplied or recommended by the manufacturers, although compromises were sometimes necessary for double digests. Spermidine (Sigma) was used in human genomic and YAC plug digests, at a concentration determined by the salt concentration of the reaction, as follows:

[Salt]	[Spermidine]
≥50mM	5mM
30-50mM	3mM
10-20mM	2mM
0-10mM	0mM

Spermidine was not usually added to other digests. Precipitation using sodium acetate was avoided when the digests contained spermidine, as this can form a precipitate which is difficult to dissolve.

Digests of mammalian genomic DNA were incubated overnight, using a three-fold excess of restriction enzyme. Other digests were carried out for 1-16 hours, with an excess of enzyme of between 2- and 10-fold.

Prior to agarose gel electrophoresis, genomic digests were precipitated, in order to remove salt from the solution. This eliminates any discrepancies in the speed of electrophoresis between samples, and gives sharper bands in hybridization after Southern blotting.

1. Add NaCl to give a final concentration of 200mM.

2. Add two volumes of ice-cold absolute ethanol.

3. Incubate at -20°C for 15 minutes.

4. Spin in an Eppendorf centrifuge for 15 minutes at 4°C.

5. Discard supernatant, add 0.5-1ml of 70% ethanol to wash pellet and leave at room temperature for 5 minutes.

6. Spin for 5 minutes at room temperature, discard supernatant, and air-dry pellet.

7. Dissolve in the appropriate volume of TE.

2.6 Conventional agarose gel electrophoresis

The Pharmacia GNA100 and Scieplas midi-gel electrophoresis tanks were used for estimating DNA concentration, sizing restriction fragments and visualizing PCR products. For these purposes gels of 0.6-2% (w/v) Type I (Sigma, low EEO) or Type II-A (Sigma, medium EEO) agarose were run rapidly at voltage gradients of 3-8 V/cm in 1 x TBE buffer. Separation of fragments smaller than ~300bp was achieved using 1% (w/v) Type I + 3% (w/v) NuSieve GTG (FMC Bioproducts) gels in 1 x TBE. The Pharmacia GNA200 maxi-gel electrophoresis tank was used when large numbers of samples needed to be analysed. Preparative gels using 0.7-1.5% SeaPlaque (FMC Bioproducts) agarose and 1 x TAE buffer were run when DNA fragments needed to be eluted from a gel matrix. Genomic digests were run overnight in the GNA200 apparatus using 0.6-0.8% (w/v) Type I gels, 1 x TBE, and a voltage gradient of 1.5-2 V/cm.

Loading buffers contained 5 or 10 times the concentration of the appropriate running buffer, along with bromophenol blue and cyanol dyes, and either glycerol (5% (v/v) final concentration) or Ficoll-400 (3% (v/v) final concentration, Pharmacia).

Gels which had been run slowly overnight were stained in water containing ethidium bromide (Sigma) at 0.5µg/ml for 15-30 minutes, with shaking. Ethidium bromide stain was added to other gels prior to pouring, to give a concentration of 0.5µg/ml. Gels were viewed on a UV transilluminator (Ultra Violet Products Inc.) and photographed with a Polaroid Land camera using Polaroid 665 film.

The following size markers (all obtained from Gibco BRL) were used for agarose gels:

λ/*Hind*III (lambda bacteriophage DNA predigested with *Hind*III) gives marker bands of sizes 0.56kb, 2.0kb, 2.3kb, 4.4kb, 6.6kb, 9.4kb and 23.1kb. In addition, comparison of the intensities of marker bands to those of the sample ran on the gel was a routine method for estimating DNA concentration of a digested plasmid or purified restriction fragment.

1kb ladder gives marker bands of sizes 75 bp, 142bp, 154bp, 200bp, 220bp, 298bp, 344bp, 394bp, 506bp, 516bp, 1.0kb, 1.6kb, 2.0kb, 3.0kb and a series of further fragments, each ~1kb larger than the last.

123bp ladder gives marker bands which are multiples of 123bp.

The marker(s) chosen for each gel depended on the expected size ranges of the fragments to be analysed.

2.7 Southern blotting of agarose gels

DNA was transferred from agarose gels to membranes using the technique described by Southern (1975):

1. Following photography, place the gel on a blotting platform containing 0.4M NaOH as transfer solution and using Whatman 3MM paper as a wick.

2. Cut a Hybond-N+ filter (Amersham) to the appropriate size, mark it with a pen and soak it in 0.4M NaOH.

3. Place the filter on the gel, followed by two pieces of Whatman 3MM paper, also soaked in 0.4M NaOH.

4. Add paper towels and a small weight, and allow transfer to proceed for 14-24 hours.

5. Briefly soak the filter in 0.2 x SSC. This can be used immediately for hybridization or wrapped in cling film and stored at -20°C until needed.

Gels containing high molecular weight DNA were exposed to UV for 4 minutes prior to blotting to increase efficiency of transfer, and blotted for a minimum of 18 hours.

2.8 Purification of DNA from gel slices

The Geneclean (BIO 101) protocol was the method of choice for purifying fragments of greater than ~300bp for subsequent use as hybridization probes. Fragments smaller than this were usually labelled directly in agarose (see section 2.9).

1. Run the DNA sample on a SeaPlaque (FMC Bioproducts) agarose gel of appropriate percentage, with 1 x TAE as a buffer, until the desired fragment is well separated from any surrounding fragments.

2. Stain the gel with ethidium bromide and visualise under UV. Excise a slice of agarose containing the required fragment and transfer it to an Eppendorf tube.

3. Spin the gel slice to the bottom of the tube using an Eppendorf centrifuge, and estimated its volume by eye. Add 2-3 volumes of 6M NaI, and mix.

4. Incubate the mixture at 45-55°C for a maximum of 5 minutes, with occasional vortexing. The agarose should have dissolved by the end of this step.

5. Add 5µl of resuspended Glassmilk, mix well and place on ice for 5 minutes.

6. Centrifuge at room temperature for 20 seconds to pellet out the Glassmilk, and discard the supernatant.

7. Add 200µl of 'NEW wash'. Resuspend Glassmilk by vortexing. Spin for 20 seconds, and discard supernatant.

8. Repeat washing step a further two times.

9. After final wash, spin again for 30 seconds, and remove any residual liquid without disturbing pellet.

10. Resuspend in 10µl of water or TE. Incubate at 45-55°C for 3 minutes. Spin for 1 minute. Transfer supernatant to a fresh tube.

11. Resuspend pellet in 5µl of water or TE. Incubate at 45-55°C for 3 minutes. Spin for 1 minute. Add supernatant to that from step 10.

12. The pooled supernatants from steps 10 and 11 contain the purified DNA. The concentration can be estimated by running a sample of it on an agarose gel.

For purification of over 5µg of DNA, a larger amount of glassmilk was used (an additional 2µl for each extra µg of DNA). The amounts of NEW wash, and water/TE used in later steps of the protocol were scaled up accordingly. The incubation on ice in step 5 could also be increased to 15 minutes to improve yields.

2.9 Hybridization protocols

2.9.1 Multiprime labelling of DNA probes

This protocol was as described by Feinberg and Vogelstein (1983, 1984). Probe specific activities were in the region of 10^8 cpm/μg DNA. For hybridizations to genomic DNA, ~5ng of probe was labelled for every ml of Church buffer used in section 2.9.3. If the hybridization was to cloned DNA, then 2ng/ml of Church was usually sufficient. When total genomic DNA was used as a probe, 50ng was labelled with 20μCi of [α^{32}P]-dCTP.

The protocol involves use of an oligomer labelling buffer (OLB), which is made by mixing the following solutions in the ratio (A:B:C) 2:5:3.

Soln A: 1.25M Tris-Cl (pH 8.0), 0.125M $MgCl_2$, 0.5mM dATP, 0.5mM dGTP, 0.5mM dTTP (Pharmacia), 250mM β-mercaptoethanol

Soln B: 2M Hepes (pH6.6) (Sigma)

Soln C: Hexadeoxyribonucleotides [pd(N)6, sodium salt (Pharmacia)], 90 OD units/ml in water

The final multipriming reaction mixture is as follows:

10μl	Oligomer Labelling Buffer (OLB)
2μl	Bovine Serum Albumin (BSA) (Sigma; 10mg/ml in water)
2-5μl	[α^{32}P]-dCTP (Amersham; 3000Ci/mmol, 10μCi/μl)
2μl	Klenow polymerase (Gibco BRL; 1U/μl)
	DNA and water to give total volume of 50μl

1. Mix the DNA and water in a 1.5ml Eppendorf tube.
2. Incubate in a boiling water bath for 10 minutes to denature the DNA.
3. Cool the tube on ice for 1 minute.
4. Add OLB, BSA, label and then enzyme. Mix.
5. Incubate at 37°C for 1-4 hours (or overnight at room temperature).

DNA in agarose (see section 2.8) was labelled in a similar manner to that above, with the following alterations. The excised agarose slice was incubated in a boiling water bath for ~1 minute prior to step 1, in order to melt it so that the appropriate amount could be pipetted into the reaction tube. In step 4, the OLB, BSA, label and enzyme were layered onto the top of the agarose and left to diffuse through, instead of mixing. The labelling reaction was then allowed to proceed for at least 2 hours.

2.9.2 Removal of unincorporated nuleotides by spin dialysis

1. After multipriming, add 50µl of TE. If labelling was done in agarose then add 150µl of TE and melt by boiling for 1 minute.

2. Make a spin column in an Eppendorf tube: use a 21-gauge syringe needle to make a crescent-shaped hole in the bottom; add about 25µl of glass beads (Sigma; 211-300 microns, acid washed and stored in TE); fill the tube with Sepharose (CL6B-200, Sigma; kept as a slurry in TE); close the tube and make a hole in the lid.

3. Spin the column for 3 minutes at 1000rpm in a bench centrifuge at room temperature to dry it.

4. Add labelled DNA mix and spin for 3 minutes as above, collecting the eluted liquid in a fresh Nunc tube beneath the column.

5. Add 100µl of TE to the Eppendorf tube which had contained the labelled mix to wash out any remaining radioactivity. Transfer this 100µl to the column. Spin as in step 4.

6. Unincorporated nucleotides are retained in the column; an estimate of incorporation efficiency is made by comparing the counts from the column with those from the eluted liquid.

2.9.3 Prereassociation for removal of repetitive sequences

For certain probes, it was necessary to remove repetitive sequences, prior to hybridization, by prereassociating with a large excess of unlabelled driver human genomic DNA. In this reaction, the repeated sequences of the genomic DNA anneal with repeats in labelled probe fragments, rendering them double-stranded and therefore inaccessible for hybridization to the filter (Sealey *et al.*, 1985).

The following protocol was used for prereassociation:

1. After spin dialysis (2.9.2), make the volume of the labelled probe up to 247.5µl with TE.

2. Add 15µl of 10mg/ml sonicated placental DNA (200-500bp in size).

3. Boil for 10 minutes to denature the DNA.

4. Cool on ice for 1 minute.

5. Add 37.5µl of 1M Na_2HPO_4 (pH 7.2) and incubate at 65°C for 2 hours.

6. Add the prereassociated probe directly to the hybridization mix as described below (2.9.4 omitting steps 6 and 7).

2.9.4 Hybridization of filters

1. Interleave filters with nylon meshes. Roll into a tube in a tray containing 2 x SSC.

2. Place rolled up tube in hybridization bottle (Hybaid) and add 20ml of 2 x SSC.

3. Rotate the bottle to unroll the tube of filters and meshes onto its inner surface.

4. Replace the 2 x SSC with Church buffer (Church and Gilbert, 1984), preheated to 65°C. 15ml of buffer was used for large bottles and 10ml for medium-sized bottles.

5. Place bottle in Hybaid hybridization oven, with rotation, at 65°C. Leave filters to prehybridize for at least 20 minutes.

6. Incubate purified, labelled DNA probe from section 2.9.2 in a boiling water bath for 10 minutes. (Multiprime labelled λ/*Hind*III fragments can be added prior to boiling for detection of λ/*Hind*III size markers on filters.)

7. Incubate denatured probe on ice for 1 minute.

8. Following prehybridization, pour some of Church buffer from bottle into a universal tube, add denatured probe (or prereassociated probe from 2.9.3), mix and pour back into bottle.

9. Replace bottle in oven and leave to hybridize, with rotation, at 65°C overnight.

2.9.5 Washing, autoradiography and stripping of filters

1. After hybridization, remove filters and meshes and rinse briefly in 400mM Na_2HPO_4 (pH 7.2) at room temperature.

2. Wash filters and meshes for 10 minutes in 400mM Na_2HPO_4, 0.1% (w/v) SDS at 65°C in a shaking water bath.

3. Remove meshes, and rinse in water.

4. Give filters further 10 minute washes in Na_2HPO_4, 0.1% (w/v) SDS, increasing the stringency at each wash by reducing phosphate concentration (typically 100mM, then 40mM), and monitoring filters with a Geiger counter. The stringency of washing required is probe and filter dependent, but a final concentration of 40mM Na_2HPO_4, 0.1% (w/v) SDS is usually sufficient for hybridizations involving single copy probes.

5. When the filters give low enough Geiger readings, air-dry them briefly on Whatman 3MM paper, and seal in Clingfilm. If the filters are dried out completely it becomes difficult to strip them subsequently; this should therefore be avoided.

6. Expose the filters to autoradiographic film in Kodak cassettes with intensifying screens at -70°C. Two speeds of film were available; Kodak X-O-matic is about twice as sensitive as Fuji R-X film, but gives grainier results. Films are developed after overnight exposure, and if necessary subsequent longer exposures can be made.

7. Following autoradiography, remove the probe from the filters by washing for several hours in filter stripping solution at 65°C in a shaking water bath, until the signal has been sufficiently reduced. Filters can then be hybridized again, using a different probe. (In general a filter can be reprobed successfully several times, although signals grow weaker with each new hybridization.)

2.10 General procedures for subcloning into plasmids

The pUC9 vector was used for most subcloning experiments. (See Chapter 6 for details of TA-cloning.)

2.10.1 Ligation

The ratio of vector to insert was determined by the nature of the cloning experiment. A 10-fold molar excess of vector was used in forced directional cloning (i.e. when there were different cohesive ends). For non-directional cloning (identical cohesive ends), a ten-fold molar excess of insert DNA was instead used.

1. Mix appropriately digested vector with the insert (previously purified by the Geneclean method). Make the volume up to 7µl with water.
2. Add 2µl of 5 x ligase buffer and 1µl of T4 DNA ligase (Gibco BRL; 1 unit/µl), giving a final volume of 10µl.
3. Incubate for 2-16 hours at 14°C.
4. Heat inactivate the ligase by incubating at 65°C for 10 minutes.

If the vector/insert mix in step 1 was more than 7µl then a reaction mixture of 15µl or 20µl total volume could be set up. In these cases the amount of buffer used was scaled up appropriately.

2.10.2 Transformation of bacterial cells using heat shock

1. Thaw on ice a vial of frozen OneShot™ (Invitrogen) competent cells (containing a 50µl aliquot).
2. Add 2µl of 0.5M β-mercaptoethanol and mix by tapping gently.
3. Add 2µl of ligated mixture (from 2.10.1) and tap gently.
4. Incubate on ice for 30 minutes.
5. Transfer vial to a 42°C water bath and incubate for 45 seconds without mixing.
6. Place vial on ice for 2 minutes.

7. Add 450µl of SOC (prewarmed to room temperature) and incubate at 37°C for 1 hour while shaking at 225 rpm.

8. Spread 20µl, 50µl and 100µl of the transformed cells on separate LB (or 2 x TY) agar plates containing ampicillin and X-gal.

9. Invert plates and incubate overnight at 37°C.

Untransformed cells are ampicillin sensitive, therefore only cells which have been transformed with plasmid (which contains a gene for ampicillin resistance) will give rise to colonies on the plates. Plasmid vectors also contain a *lac*Z gene which is interrupted by the cloning site. Non-recombinants form blue colonies due to the conversion of the chromogenic substrate X-gal into a blue product by the intact *lac*Z gene, whilst in recombinant clones the disruption of the *lac*Z gene by an insert results in the formation of white colonies.

2.10.3 Frozen storage of transformed cells

Liquid cultures of colony-purified transformed cells were grown up overnight. The culture was aliquoted into Nunc tubes, mixed with 15% sterile glycerol, and frozen in liquid nitrogen. Frozen stocks were be stored at -20 or -70°C.

2.11 Double-stranded sequencing of plasmid DNA

Clones were sequenced by the dideoxy chain-termination technique (Sanger *et al.*, 1977), using Sequenase Version 2.0 enzyme (Tabor and Richardson, 1989) and the accompanying kit (USB). Fragments cloned into pUC9 were sequenced on both strands using the 17-mer M13 universal -40 primer (USB) and the 17-mer M13 reverse primer (Pharmacia), as well as primers derived from the obtained sequence. YAC left end clones generated by plasmid rescue were sequenced with a primer (4L) that was designed from the left arm vector sequence; 5'AAGTACTCTCGGTAGCCAAG3'.

i) Annealing

1. Prepare plasmid DNA using the Promega protocol (2.4.3). Add an equal volume of 0.4M NaOH/4mM Na_2EDTA to 3-5µg of plasmid DNA.

2. Incubate at 37°C for 30 minutes to denature DNA.

3. Add 1/10th volume of 3M sodium acetate (pH 5.2), followed by 2 volumes of ice-cold ethanol, mix and place at -70°C for 15 minutes.

4. Spin in an Eppendorf centrifuge for 15 minutes at 4°C.

5. Discard supernatant and wash pellet with 70% ethanol.

6. Spin in an Eppendorf centrifuge for 5 minutes at room temperature.

7. Discard supernatant and air dry pellet. Dissolve in 7µl of water.

8. Add 2µl of '5 x Sequenase reaction buffer' (0.2M Tris-Cl (pH 7.5), 0.1M $MgCl_2$, 0.25M NaCl) and 1µl of primer (1µM). Incubate at 37°C for 30 minutes.

ii) Labelling

1. Dilute '5 x (dGTP) labelling mix' (7.5µM each of dGTP, dCTP and dTTP) 1:4 in water.

2. Dilute Sequenase enzyme 1:7 in ice-cold 'enzyme dilution buffer' (10mM Tris-Cl (pH 7.5), 5mM DTT, 0.5mg/ml BSA). Keep on ice.

3. To the annealed template/primer add 1µl 0.1M DTT, 0.5µl [$\alpha^{35}S$]-dATP (Amersham; 10µCi/µl), 2µl of diluted labelling mix and 2µl of diluted Sequenase enzyme. Mix. Incubate at room temperature for 2-5 minutes.

iii) Termination

1. Aliquot 2.5µl of each dideoxynucleotide - ddGTP, ddATP, ddTTP and ddCTP (80µM of each dNTP, 8µM of ddNTP, 50mM NaCl) - into separate labelled Eppendorf tubes. Prewarm at 37°C for at least 2 minutes.

2. Add 3.5µl of labelled mix to each of the tubes containing the ddNTPs, mix and incubate at 37°C for 3-5 minutes.

3. Add 4µl of 'Stop solution' (20mM Na_2EDTA, 95% (v/v) formamide, 0.05% (w/v) bromophenol blue, 0.05% (w/v) xylene cyanol FF), mix and place on ice.

4. Samples should be stored at -20°C until they can be run on a polyacrylamide gel.

2.12 Polyacrylamide gel electrophoresis (PAGE)

6% denaturing polyacrylamide gels were used to run sequencing reactions and labelled PCR products from microsatellite polymorphisms:

1. Prepare a 40% (w/v) acrylamide stock solution, containing acrylamide:N,N'-methylenebisacrylamide (BDH) at a ratio of 19:1, and then filtered through Whatman paper. Store in a dark bottle at 4°C.

2. Use this stock to prepare a solution containing 6% (w/v) acrylamide, 4.2% (w/v) urea (BDH), 1 x TBE, which is also filtered and stored in a dark bottle at 4°C.

3. Treat one 55cm glass plate with 'Bind-Silane' (LKB A-174) and the other with 'Repelcote' (BDH), to ensure that the gel attaches to only one plate on their separation following electrophoresis.

4. For each gel, polymerise 40ml of acrylamide solution by adding 0.1% (v/v) TEMED (Sigma) and 1% (v/v) of a 10% (w/v) ammonium persulphate solution, and pour between the two glass plates, separated by 0.2mm spacers.

5. Allow the gel to set for 30-60 minutes, and then place it in an LKB 2010 Macrophor Sequencing apparatus. Clamp an aluminium backing plate onto the back glass plate to reduce temperature differences across the gel.

6. Carefully flush out the wells with 1 x TBE to remove urea, and prerun the gel for 60 minutes at 2kV (~38V/cm) in 1 x TBE buffer.

7. Flush the wells out again prior to loading.

8. Denature samples by incubating them at 94°C for 2 minutes and then place on ice.

9. Use a Macrophor sample syringe with a glass fibre needle (LKB 2010-150) to load ~2µl of sample into each lane. Allow electrophoresis to proceed at 2kV (~38V/cm) until the first blue dye (for short runs) or the second blue dye (for long runs) has run off the bottom of the gel.

10. Remove the gel plates from the apparatus and separate them. The gel should remain attached to the 'Bind-Silane' treated plate.

11. Fix the gel in 10% (v/v) acetic acid/10% (v/v) methanol for 30 minutes, and then dry it at 65°C for at least two hours.

12. Expose 'hyperfilm-βmax' film (Amersham) to the gel at room temperature overnight.

2.13 Polymerase chain reaction (PCR)

PCR is a technique allowing the enzymatic amplification of specific regions of DNA without the need for conventional cloning procedures (Saiki *et al.*, 1985). The process uses oligonucleotides which hybridize to opposite strands of the template and flank the region to be amplified. These primers are orientated towards each other, such that repeated cycles of heat denaturation, primer annealing and extension using DNA polymerase result in exponential increase of product from the target region. The ease of this technique has been greatly increased by the availability of a thermostable DNA polymerase which was isolated from *Thermus aquaticus* (Chien *et al.*, 1976; Saiki *et al.*, 1988).

2.13.1 Oligonucleotide design

The PRIMER computer program (Version 0.5: © 1991, Whithead Institute for Biomedical Research) was used for the selection of oligonucleotides, of 18-23bp in length, from a target sequence. Primers were chosen such that the predicted PCR product would be in the range of 100-600bp. The program allows the specification of the desired *Tm* (melting temperature in °C) of oligonucleotides to be selected. In addition, it screens out any primers which are likely to be self annealing or to form primer-dimers. The program also analyses the target sequence for homology to repetitive elements (such as *Alu*s or LINEs) and will select against oligonucleotides within such regions.

2.13.2 Deprotection and purification of oligonucleotides

Oligonucleotides were synthesized by Dr. Val Cooper at the Dyson Perrins Laboratory, Oxford University using an Applied Biosystems Synthesizer, giving a yield of up to 1mg of detritylated oligos in ~3ml of concentrated ammonia. These needed to be incubated at 55°C for a minimum of 4 hours to remove the base protecting groups.

Two alternative protocols could be used for purification:

i) Standard protocol

1. After deprotection, place at -70°C for 30 minutes.

2. Dry in a vacuum desiccator at 4°C.

3. Resuspend in 500µl of water and extract three times with butan-1-ol. The cleaved protecting groups go into the organic phase, and at each extraction the aqueous phase (containing the oligonucleotide) is reduced in volume.

4. Add 1/10th volume of 3M sodium acetate (pH 5.2) and 4 volumes of absolute ethanol at room temperature.

5. Spin in an Eppendorf centrifuge for 15 minutes at room temperature.

6. Wash pellet with 1ml of 70% ethanol.

7. Dry pellet and resuspend in 500µl of water.

8. Determine the OD_{260} of a 1:200 dilution, and use 1OD = 20µg/ml to calculate the concentration of the purified oligonucleotide.

ii) Quick protocol

1. Following deprotection, aliquot 180µl of unpurified oligonucleotide into a 2ml Eppendorf tube. The remainder can be stored at -20°C, and prepped when necessary.

2. Add 1.8ml of butan-1-ol and vortex. This causes the DNA to precipitate.

3. Spin in an Eppendorf centrifuge for 5 minutes at room temperature.

4. Discard supernatant, dry pellet, and resuspend in 180µl of water.

5. Add 1.8ml of butan-1-ol and vortex.

6. Spin in an Eppendorf centrifuge for 5 minutes at room temperature.

7. Discard supernatant, dry pellet, and resuspend in 40µl of water.

8. Estimate the concentration of the purified oligonucleotide by determining the OD_{260} of a 1:200 dilution as described above.

2.13.3 Conditions for PCR amplification

PCR was carried out in volumes of 20, 50 or 100µl, with a similar volume of paraffin oil layered on the top to prevent evaporation. Reaction mixes consisted of 1µM each primer; 10mM Tris-Cl; 50mM KCl; 1.5mM $MgCl_2$; 200µM dNTPs (Amersham); 0.05U/µl *Taq* polymerase (Boehringer Mannheim or Promega). Between 1-100ng of template DNA was used, depending on its complexity.

A Techne Programmable Dri-Block PHC-1 was used for thermal cycling. All PCRs involved an initial 10 minute denaturing step ('hot start') prior to the addition of dNTPs and *Taq* polymerase. Following this the cycling conditions were as follows; 94°C, 30 seconds (denaturation); *Ta* (annealing temperature) °C, 30 seconds (annealing); 75°C, 18-60 seconds (elongation); 30-36 cycles. A final elongation step of 75°C for 2 minutes was also included. (Note: annealing temperature is calculated by subtracting 5 from the melting temperature of the oligonucleotides i.e. $Ta = Tm - 5$.)

In cases where primers gave additional products as a result of non-specific amplification, the annealing temperature could be increased. Alternatively the concentration of Mg^{2+} in the reaction could be varied (in the range of 0.5-4.0mM). When neither strategy was succesful, it was sometimes necessary to design new primers from the target sequence.

PCR is a very sensitive technique and it is essential to minimize any amplification that may result from contamination of the reaction mix. Levels of contamination were assessed by setting up a control reaction, for each PCR, in which template was absent. Contamination from aerosols when pipetting could be avoided by using a set of PCR-only pipettes for setting up reactions; these were cleaned regularly with absolute ethanol. In addition, disposable pipette tips which contain filters (Rainin) have recently become available. UV irradiation of primers, buffer, water and dNTPS for 5-20 minutes was also effective for minimizing contamination. However, primers containing consecutive thymidine residues were not exposed to UV, because this could result in the formation of thymidine dimers, which would reduce PCR efficiency.

2.13.4 End-labelling of oligonucleotides for microsatellite analysis

One primer from the pair to be used in PCR was radioactively labelled as follows:

1. To 200pmol of oligonucleotide, add 2µl 5 x T4 polynucleotide kinase buffer (Boehringer Mannheim), 2µl of [γ^{32}P]-dATP (Amersham; 3000Ci/mmol, 10µCi/µl) and 1µl of polynucleotide kinase (Boehringer Mannheim; 10U/µl). Make up to 10µl with water. (In general, 20µCi of [γ^{32}P]-dATP is sufficient for 200µl of PCR mix.)

2. Incubate at 37°C for 30 minutes.

3. Add 90µl of water and spin dialyse (2.9.2) using Sephadex G-50 (Sigma).

The end-labelled primer was then used in combination with the 'cold' primer, for PCR. Following amplification, Ficoll loading buffer was added to the samples and they were run on polyacrylamide gels as described in 2.12.

2.14 General techniques for manipulation of Yeast Artificial Chromosomes

2.14.1 Frozen storage of YAC stocks

Liquid cultures of colony-purified cells were grown up overnight. The culture was aliquoted into Nunc tubes, mixed with 50% sterile glycerol, and frozen in liquid nitrogen. Frozen stocks were stored at -70°C.

2.14.2 Preparation of YAC DNA in plugs

1. Inoculate 50ml SD + CAT medium in a 500ml conical flask with a single YAC colony, or 50µl of a frozen stock.

2. Incubate for 36 hours at 30°C, while shaking at 450rpm.

3. Spin at 2000rpm in an MSE bench centrifuge for 10 minutes, to pellet cells.

4. Drain off supernatant and resuspend in 20ml of 50mM Na_2EDTA (pH 8.0).

5. Pellet cells as before. Drain off supernatant.

6. Resuspend in 2ml SEB containing 4mg/ml Novozyme (Novo Nordisk).

7. Melt 1.2% SeaPlaque agarose (FMC Bioproducts) in SEB and cool to 45°C. Add 2ml of this to the 2ml of yeast suspension from step 6 and mix well.

8. Pipette mixture into prechilled plug moulds, and leave to set on ice for 5 minutes. Typically more than 40 plugs can be made from a 50ml culture.

9. Eject plugs into 25ml of STEB containing 4mg/ml of Novozyme (Novo Nordisk).

10. Incubate at 37°C for 2 hours with gentle shaking.

11. Replace with 25ml of LiDS. Incubate at 37°C for 30-60 minutes.

12. Replace with 25ml of fresh LiDS. Incubate at 37°C overnight.

13. Replace with fresh LiDS and store at room temperature.

Concentrated YAC plugs could be made by using 0.5ml of SEB and 0.5ml of 1.2% SeaPlaque agarose in SEB in steps 6 and 7 of the above protocol.

2.14.3 Preparation of lambda oligomers as markers

These give a ladder with steps 48.5kb apart and are therefore ideal as general purpose markers for pulsed field gel electrophoresis. Lambda bacteriophage DNAs form oligomers as they anneal via their 12-base cohesive termini.

1. Add 0.45µg of lambda c1857 Sam7 virion suspension to 7.5ml of PBS. Warm to 37°C.

2. Melt 1.2% SeaPlaque agarose (FMC Bioproducts) in PBS and cool to 45°C. Add 7.5ml to suspension from step 1.

3. Pipette into prechilled plug moulds and leave to set on ice for 5 minutes.

4. Eject plugs into 40ml NDS containing 1mg/ml pronase (Boehringer Mannheim). Incubate at 50°C overnight.

5. Replace with 40ml fresh NDS. Incubate at 50°C overnight.

6. Replace with 50ml filter sterilized 2 x SSC. Stand at room temperature for 30 minutes. Repeat this step an additional three times.

7. Replace with 40ml Lambda oligomerization buffer, and leave at room temperature for two weeks prior to use.

2.14.4 Restriction enzyme digestion of DNA in plugs

Iron ions can form complexes with EDTA which may be nucleolytic; plugs were therefore not allowed to come into contact with ferrous metals. They were handled using glass loops and cut using glass cover-slips. Glass loops were washed in 20% (v/v) ethanol, 0.1% (w/v) SDS prior to use. The protocol for complete digestion was as follows:

1. Wash the appropriate number of YAC plugs (allowing a third of a plug per digest) three times for 30 minutes in TE (made from AnalaR Tris) at 50°C. This reduces the concentration of Na_2EDTA in the plugs from 100mM (which would inhibit enzyme activity) to only 1mM.

2. Cut plugs into thirds. Place each third into an Eppendorf tube containing 400µl of the appropriate 1 x restriction buffer, and allow to equilibrate, on ice, for 30 minutes.

3. Replace this with 60µl of digestion buffer (1 x restriction buffer supplemented with 0.1mg/ml Gelatin (Difco), 1mM DTT (Sigma) and Spermidine (Sigma), as described in 2.5). Stand on ice for 10 minutes.

4. Add 15-20 units of appropriate restriction enzyme and incubate at reaction temperature for 3 hours.

5. Stop the digests by placing on ice, removing the reaction buffer, and replacing with 100µl of STOP buffer. Allow to equilibrate on ice for at least 15 minutes before loading on a pulsed field gel.

6. Digests in STOP can be stored overnight at 4°C prior to loading, if necessary. However, for more long term storage at 4°C, the reaction buffer should be replaced with 400µl of NDS instead of STOP.

YAC partial digests were generated by varying the amount of enzyme in the reaction (within a range of 0.03-15 units, depending on the efficiency of the enzyme), while keeping the incubation time constant (1 hour). When setting up partial digests, reaction mixes were left on ice for 30 minutes after adding enzyme (but prior to incubating at the reaction temperature) to allow diffusion of enzyme through the sample. The various dilutions of restriction enzyme required were made using the generalized restriction enzyme dilution buffer (REDB) described in 2.1.

2.14.5 Pulsed field gel electrophoresis (PFGE)

A rotating plate ('Waltzer') device was used for electrophoresis (Southern *et al.*, 1987), with Type I, low EEO agarose (Sigma) and 0.5 x TAE as running buffer. The apparatus has a capacity of 6 litres, circulated by an Eheim pump, and cooled, indirectly via a glass cooling coil, by an LKB Multitemp II thermostatic circulator.

Pulsed field gels were run using 1.5% agarose, 30-36 hour running time, 16°C running temperature, a field strength of 6V/cm and switch times in the range of 12-60 seconds (which give maximum resolutions of between ~200kb and ~900kb). One exception to this was the sizing gel for the F1001 YAC clone, which used 1% agarose, a 50 hour running time, a 3.6V/cm field strength and a 4 minute switch time, to give a maximum resolution of >1.65Mb.

In addition to the lambda oligomer markers (2.14.3), the chromosomes of *Saccharomyces cerevisiae* (the host background in YAC plugs) could be used to aid estimation of YAC size. *S. cerevisiae* chromosome sizes have been estimated as 245, 280, 360, 450, 590 (doublet), 680, 750, 790, 830, 940, 970 and 1115kb (Jobling, 1991). Sizes of the chromosomes that are larger than this have not been precisely determined but were not needed for this study. YACs and markers were loaded into wells as plug thirds. A mixture of uncut lambda and *Sal*I cut lambda DNA could also be used to give low molecular weight markers (15.0, 33.5 and 48.5kb); this was loaded as a liquid (containing Ficoll/TAE loading buffer) into the well of a submerged gel, while the pump was switched off.

2.14.6 Conventional agarose gel electrophoresis of digested YAC plugs

1. Following digestion (2.14.4), equilibrate YAC plugs with 1ml of TE containing 10mM NaCl and leave at room temperature for 30 minutes.

2. Remove the liquid and add an appropriate amount of Ficoll based loading buffer (e.g. for a plug third, which has a volume of ~30-35µl, add 4µl of 10 x loading buffer).

3. Incubate at 65°C for 20 minutes to melt the plugs; the presence of NaCl reduces denaturation of the DNA.

4. Load melted plugs as a liquid into the submerged wells of a conventional agarose gel. They then solidify in the well, soon after loading, and the gel can be run as normal (2.6).

2.14.7 PCR from DNA embedded in plugs

1. Wash each YAC plug three times for 30 minutes in TE (made from AnalaR Tris) at 50°C, in order to reduce the concentration of Na_2EDTA.

2. Transfer each plug to a 1.5ml Eppendorf tube, and dilute with 400µl of water.

3. Incubate at 65°C for 10 minutes. Vortex briefly.

4. Incubate at 65°C for a further 10 minutes, or until agarose is melted.

5. Use 1µl of this template in a 20µl PCR reaction as described in 2.13.3.

6. The remaining YAC plug template can be stored at 4°C and used directly in PCR reactions when required.

Chapter 3 – A bi-directional YAC walk from the hypervariable locus DXS255 in Xp11.22

3.1 Introduction

3.1.1 Yeast artificial chromosomes (YACs)

In 1987 Burke, Carle and Olsen reported the development of a system for the cloning of large segments of exogenous DNA into yeast, by ligating them to vector sequences that allow their propagation as linear artificial chromosomes. The basic functional units of a yeast chromosome had previously been defined, and could be combined to form vector arms which were less than 10kb in total size (Burke *et al.*, 1987).

Figure 3.1 shows the general strategy for the *in vitro* construction of artificial chromosomes using pYAC4. This circular vector contains pBR322 derived sequences that allow it to be propagated as a single plasmid in a bacterial host, *Escherichia coli*, including a replication origin (**ori**) and a gene conferring ampicillin resistance (**Amp**) for selection. In addition, pYAC4 incorporates three elements that are necessary for the replication and stability of an artificial chromosome in *Saccharomyces cerevisiae*:

- **CEN4** is a segment isolated from yeast chromosome IV which can act as a fully functional centromere in mitotic and meiotic cell divisions.

- **ARS1** is an autonomous replicating sequence, also cloned from yeast chromosome IV, which allows replication of colinear DNA. Cloned ARS1 elements have been shown to replicate once per cell division at a time coincident with the corresponding chromosomal sequence (Hieter *et al.*, 1985).

- Two sequences are included (labelled **TEL**) which can seed formation of functional yeast telomeres *in vivo*. These are derived from the termini of *Tetrahymena* macronuclear ribosomal DNA molecules.

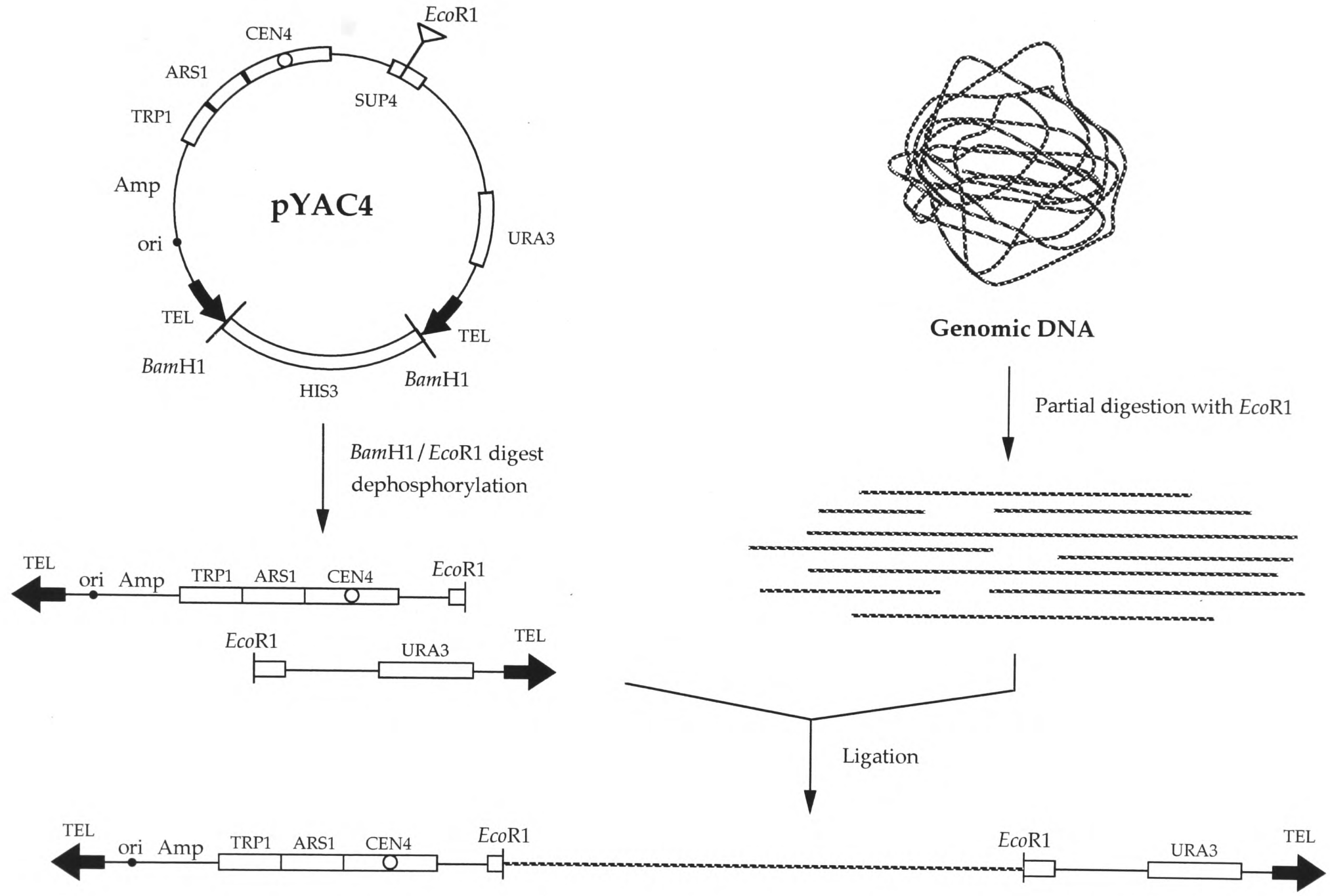


Figure 3.1: Scheme for cloning large inserts into pYAC4 vector, adapted from Burke *et al.* (1987). See text for details.

The *Eco*R1 cloning site of pYAC4 was created within the 14bp intron of **SUP4**. This is an ochre-suppressing allele of a tyrosine tRNA gene, whose interruption is phenotypically detectable (see below). The vector also contains two selectable markers, **TRP1** and **URA3**, on either side of the cloning site, and a **HIS3** gene, separating the two TEL sequences, which is discarded during the cloning process.

The cloning protocol (Figure 3.1) first involves double digestion of pYAC4 with *Bam*H1 and *Eco*R1, to give three parts; a left chromosome arm (including the centromere), a right chromosome arm, and a throwaway fragment containing the HIS3 gene. After dephosphorylation, the chromosome arms are ligated to large insert fragments, derived from the target DNA by partial digestion with *Eco*R1. Transformation is used to introduce the constructs into (*ura3, trp1, ade2-ochre*) yeast spheroplasts. Transformants are selected for complementation of the *ura3* and *trp1* auxotrophic host markers by the URA3 and TRP1 genes on the vector, to ensure the uptake of a construct containing both vector arms.

In *ade2-ochre* yeast mutants, a gene which encodes an enzyme involved in purine biosynthesis is disrupted by an ochre stop codon, resulting in the accumulation of a red intermediate. When such a host is transformed by a YAC containing an uninterrupted cloning site, the intact SUP4 $tRNA^{Tyr}$ gene product is expressed from the YAC and this supresses the ochre mutation, giving rise to white colonies. However, if the transforming construct contains exogenous DNA, the incorporation of a huge insert into the SUP4 intron results in inactivation of the suppressor, and cells therefore exhibit the red (*ade2*) phenoype (Hieter *et al.*, 1985; Burke *et al.*, 1987). It is thus possible to detect recombinant clones.

3.1.2 Human genomic YAC libraries

A significant problem with the first YAC libraries to be constructed was the small average size of clones, ranging from 75kb to only 150kb (Anand *et al.*, 1989). This could be attributed to three main factors:

i) Initial experiments involved extended manipulation of DNA in solution, which is known to cause damage to fragments of high molecular weight (Anand *et al.*, 1989). Such strand damage may also initiate repair mechanisms in yeast, resulting in rearrangements of transformed YACs (Albertsen *et al.*, 1990).

ii) After partial digestion, whilst small fragments make up only a minor proportion of the total weight of human DNA, they constitute a much larger proportion of the number of clonable ends (Anand *et al.*, 1989).

iii) There is an inverse relationship between YAC size and transformation efficiency. For example, it has been shown that a 100kb clone can transform yeast 37 times more efficiently than one of 300kb (Lee *et al.*, 1992).

To provide a >99% probability of finding a random single copy sequence in a human genomic library there must be coverage of five haploid genome equivalents (i.e. 5 x 3,000 Mb) (Burke and Olson, 1991). A YAC library containing a 100kb mean insert size would therefore have to consist of over 150,000 clones. Many difficulties would be encountered with storage and screening of a library of this size. Increasing the average insert size to 500kb (for example) reduces the required clones to 30,000, which is a much more feasible number to manage.

Various techniques have been used to increase average insert sizes when constructing YAC libraries (Table 3.1), including the manipulation of DNA in agarose plugs instead of in solution , and the use of size fractionation to remove small DNA fragments prior to cloning (Anand *et al.*, 1989). The most efficient method of size selection involves excision of the compression band from a preparative pulsed field gel (Section 3.1.3) followed by treatment with agarase to remove the agarose.

Library	Cell line	No. of clones	Average insert size	Construction	Reference
St. Louis	CGM-1 (46, XY)	60,000	275 kb	DNA handled in solution; size selection using sucrose gradients at two steps, one before ligation (>100kb), the other after ligation (>200kb)	Brownstein *et al.*, 1989
ICI	GM1416 (48, XXXX)	30,000	370 kb	DNA handled in plugs; digests size selected by PFGE (>200kb) before ligation	Anand *et al.*, 1989
CEPH	EBV-transformed lymphoblastoid (46, XY)	70,000	470 kb	DNA handled in plugs; two PFGE size selection steps (>300kb) one before ligation, the other after ligation	Albertsen *et al.*, 1990
ICRF	GM1416 (48, XXXX)	16,000	620 kb	DNA handled in plugs; two PFGE size selection steps (>400kb) one before ligation, the other after ligation; polyamines present in all melting steps	Larin *et al.*, 1991
Nussbaum	Xpter-Xq27.3 human/hamster hybrid	3,300	280 kb	DNA handled in plugs; size selection (>300kb) after ligation; [NaCl] of 100mM in all melting steps	Lee *et al.*, 1992
CEPH MegaYAC	EBV-transformed lymphoblastoid (46, XY)	33,000	900 kb	details not published	Chumakov *et al.*, 1992

Table 3.1: Details of human genomic YAC libraries.

However, Larin *et al.* (1991) observed that the average size of clones generated using this procedure was significantly smaller than the size that was selected. They found that the melting of agarose containing high molecular weight DNA (an essential step in the size selection procedure) caused partial degradation of the DNA. It was suggested that this might be due to the presence of metal ion cofactors in commercial agarose, which, when complexed with EDTA, can bind strongly to AT-rich regions of DNA and cause cleavage at 68°C (Larin *et al.*, 1991). The presence of polyamines (spermidine and spermine) in all melting steps was therefore used to protect DNA from degradation, resulting in libraries with significantly larger inserts. Lee *et al.* (1992) overcame this problem in a different way, by increasing NaCl concentration to 100mM during the melting reaction.

It should be noted that, given that the spatial distribution of ARS sequences in the human genome is not yet known, it is possible that libraries with very large inserts may not be fully representative, since such clones might rely on the chance occurrence of internal ARSs for their replication (Anand *et al.*, 1989).

In addition to selecting for larger inserts, strategies involving construction of chromosome specific libraries have been used to reduce the number of clones needed for storage and analysis. Cloning of DNA from sorted human chromosomes into YACs is inefficient due to the limited quantity of source material following sorting. X-specific libraries have been successfully constructed by cloning total DNA from a hamster-human hybrid cell line containing a portion of the X chromosome as its only human DNA, and then using human specific repeats to select clones containing human YACs (Abidi *et al.*, 1990; Lee *et al.*, 1992). (These libraries also have the advantage of low co-cloning rates, as described below.) A third approach employs *Alu*-PCR (see section 3.1.4) of a human monosomic somatic hybrid cell line to generate a chromosome specific probe which is used to select clones from a total genomic library (Chumakov *et al.*, 1992). These can then be maintained as a chromosome-specific sub-library.

The most common artefact found in YAC libraries involves the formation of clones containing two or more non-contiguous fragments of DNA in a single insert (Green *et al.*, 1991). Estimates of the proportion of chimæric YACs in currently available libraries ranges from 11% (Nussbaum library) to 50% (CEPH megaYAC library). It is particularly important to detect such aberrant clones when chromosome walking, since they can result in contigs which contain falsely linked regions of the genome. Lee *et al.* (1992) found that the average size of chimæric YACs was significantly larger than that of the general YAC population in their library, and therefore suggested that selecting for very large average insert sizes when constructing a library may result in an undesirably large proportion of chimæras. Indeed, the very high co-cloning rate of the CEPH megaYACs, which have an average insert size of 900kb, has borne out this prediction (Cohen *et al.*, 1993).

Co-cloning events may occur at two alternative stages during library construction (Figure 3.2):

i) In general, inserts are not treated with phosphatase, in order to ensure minimum exposure of source DNA to enzymes (Burke and Olson, 1991). It is therefore possible to get co-ligation of unrelated fragments *in vitro*.

ii) Following partial digestion, many fragments are likely to have an *Eco*R1 cut at one end, but a double-stranded break at the other due to random shearing. If a yeast cell were to be transformed with two molecules, one spanning from a left-vector-arm telomere to a break, the other spanning from a break to a right-vector-arm telomere, a recombination between them would form an intact chimæric YAC. The presence of dispersed repeat sequences, such as *Alu* motifs, at high frequencies in the human genome and the highly recombinogenic nature of yeast suggest that such an event would be efficient. Several alternative mechanisms also involving *in vivo* recombination (for example, following co-transformation of two intact YACs) could similarly produce aberrant clones.

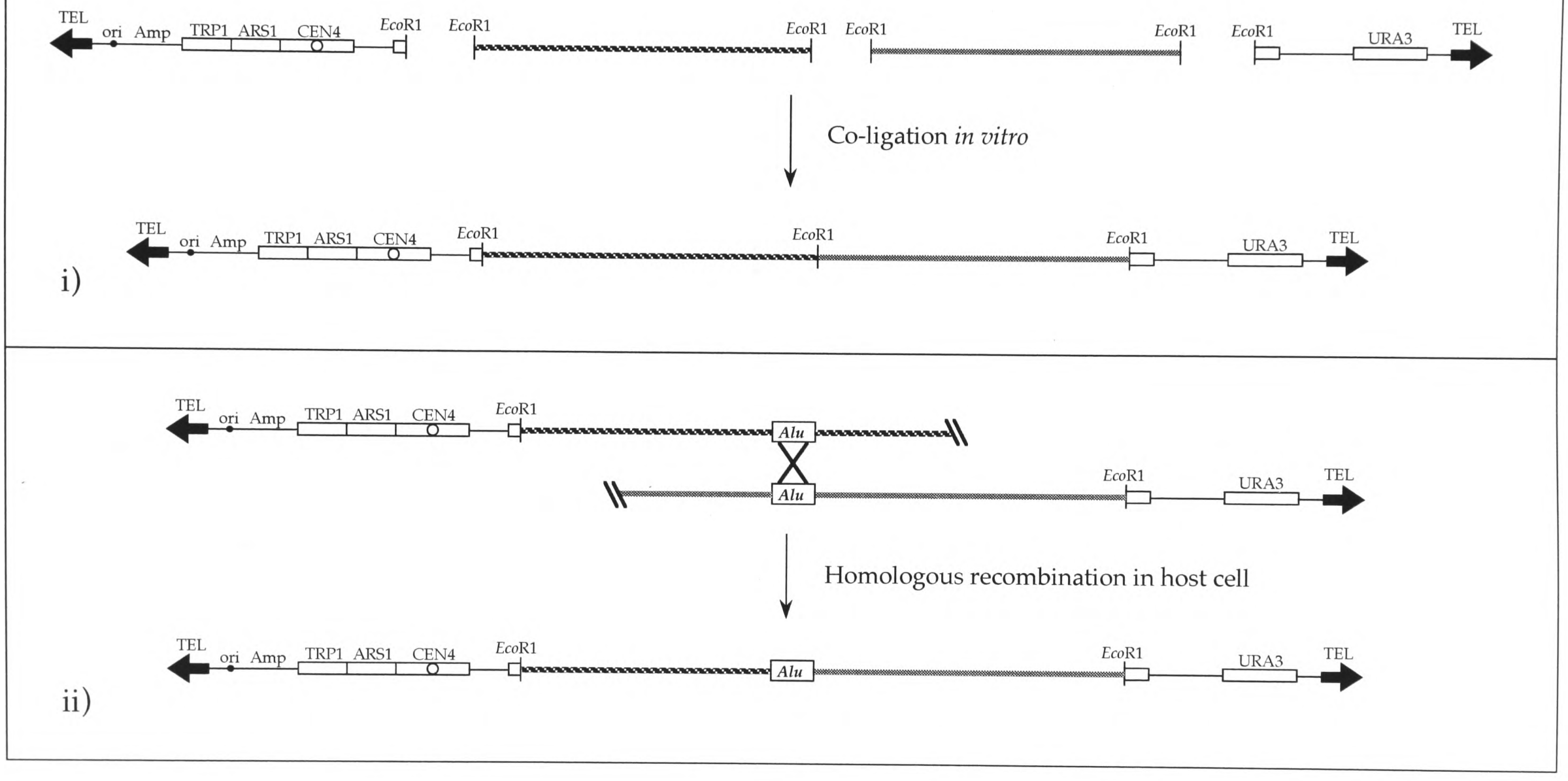


Figure 3.2: Two models to explain formation of chimæric clones. i) *in vitro*. ii) *in vivo*. See text for details.

On detailed analysis of one chimæric clone, Green *et al.* (1991) discovered an *Alu* motif at the junction between the unrelated fragments (see Fig 3.2). Furthermore, it has been found that human X-specific libraries prepared from hamster-human hybrid cell lines have significantly lower co-cloning rates (11-15%) than those made from total human DNA (40-50%) (Abidi *et al.*, 1990; Lee *et al.*, 1992; Cohen *et al.*, 1993). These observations are best explained by a recombination-based model, which would predict a lower proportion of chimæric clones among human YACs of a library derived from a somatic hybrid cell line, since the species-specific repeats in clones containing human fragments would not be present in the excess background of YACs containing hamster DNA.

3.1.3 Analysis of YAC clones

The sizing and mapping of YACs, whose inserts may vary from 100kb to several megabases, requires a system of size fractionation with a comparable range of resolution.

i) Pulsed field gel electrophoresis (PFGE)

Conventional agarose gel electrophoresis can typically separate DNA fragments of up to 50kb. However, resolution above this size becomes progressively worse, because larger molecules show anomolously high electrophoretic mobilities. Attempts to overcome this by reducing gel concentration were found to be impractical; such gels are extremely fragile, require long running times, and only give poor resolution (see Anand, 1986).

It is thought that DNA molecules of less than ~50kb behave as worm-like coils in agarose gel electrophoresis, where the dimensions of the coils are comparable to the size of the pores in the gel (Schwartz and Cantor, 1984). Separation thus results from sieving by the gel matrix. However, extremely long DNA fragments form coils whose dimensions are significantly larger than the average pore size. When they undergo

electrophoresis in a concentrated gel they must stretch out parallel to the electric field, adopting a conformation in which they are threaded through several different pores at once. The front ends of such molecules penetrate the gel together and move at a rate that is independent of fragment size (Southern *et al.*, 1987). They are therefore not separated.

A system for the separation of fragments of 30 to 2000kb was first described in 1984 by Schwartz and Cantor, who successfully fractionated chromosomes from *Saccharomyces cerevisiae*. They achieved this by subjecting DNA to alternatively pulsed, perpendicularly oriented electric fields, one of which was inhomogenous. It was assumed that at each pulse the DNA molecules would have to reorient themselves before moving along the new field direction, and that the time required for reorientation would be dependent on size of the fragment. They proposed that inhomogeneity in the field was necessary for this reorientation. Carle and Olson (1985) also derived an electrophoretic karyotype of yeast, using a similar system of pulsed orthogonal fields, but in this case both were non-uniform. However, an unfortunate consequence of inhomogenous fields is that the migration of fragments varies, depending on where in the gel the sample is loaded, leading to curved tracks which can make analysis difficult (Anand, 1986).

Several systems have since been developed to give straight tracks. Field inversion gel electrophoresis (FIGE), which involves periodically reversing polarity in a standard horizontal apparatus, can result in the co-migration of fragments that differ greatly in size (Anand, 1986). A more reliable system, known as a CHEF (contour-clamped homogenous electric field) employs multiple electrodes arranged along a polygonal contour and clamped to predetermined electric potentials to generate alternating homogenous fields which are 120° apart (Chu *et al.*, 1986). A third method (known as the 'Waltzer') uses a circular gel in a uniform electric field which is rotated through an angle of 110° at each switching interval (Southern *et al.*, 1987).

Analysis with the CHEF and 'Waltzer' systems indicated that field inhomogeneity is not in fact required, but that separation only occurs at obtuse field angles (Chu *et al.*, 1986; Southern *et al.*, 1987). On the basis of this Southern *et al.* suggested an alternative explanation for the behaviour of large DNA fragments in pulsed field gels (Fig. 3.3). Assuming that these molecules are initially in an extended conformation, with their front ends moving together (as described above), it follows that their back ends will trail behind at distances which depend on fragment length. When the current is switched off between pulses the molecules remain in their elongated state. On the application of a new field direction which is at least 90° to the old direction, the molecules lead off with what was the back end of the previous pulse, and the starting point for movement is therefore different for molecules of different length. The result is that a molecule is retarded at each pulse by an amount that is proportional to its length. This theory predicts that if the field direction changes by less than 90°, then the new leading end is likely to be the same as the old front end, and there will be no separation.

The 'Waltzer' apparatus was used for the analysis described in this thesis. It gives uniformly straight tracks, making accurate size estimates possible. In addition, variation of the switching conditions of the system leads to predictable changes in the range of sizes resolved.

ii) Rare-cutter restriction mapping of YACs

In order to make long range restriction maps of the large inserts cloned into YACs, enzymes are needed which cut the DNA infrequently. As described in the general introduction (Section 1.1.3), the bulk of the mammalian genome consists of DNA with low G + C content and a frequency of CpG dinucleotides which is only 20-25% of that expected from the base composition (Bird, 1986). An enzyme that has a >6bp recognition sequence which is GC rich and which includes one or more CpG dinucleotides will thus cut rarely in the genome (Lindsay and Bird, 1987) (Table 3.2).

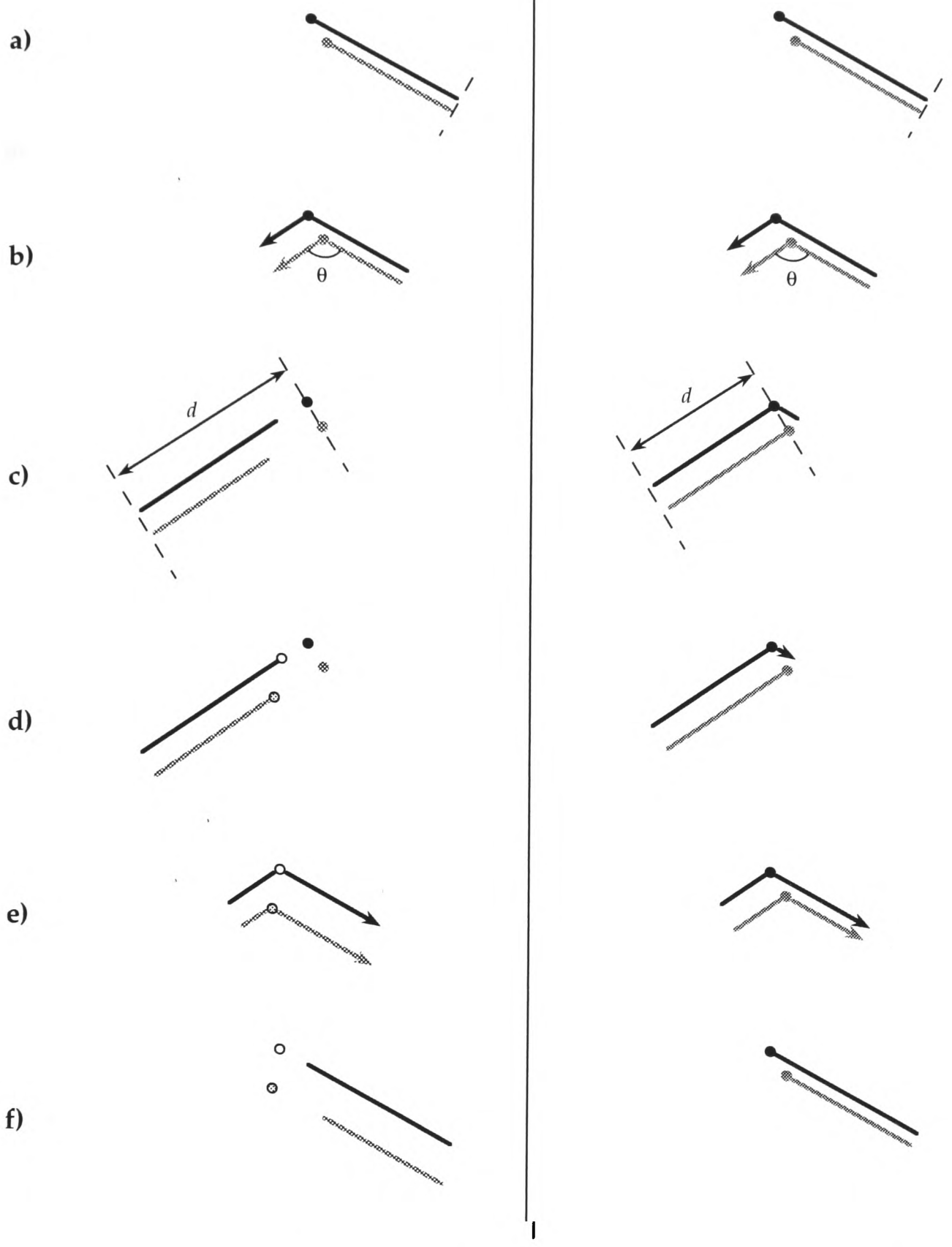


Figure 3.3: Model for separation mechanism of PFGE. θ, the angle between the field directions, must be more than 90°. *d* = the distance moved by the front end in the direction of the field during one complete pulse. Net movement is down the page.

Left hand side: Resolution of two molecules whose lengths (when in an extended conformation) are less than *d*:
a) End of first pulse; leading ends of molecules are together, with their back ends trailing at a distance that is proportional to their length.
b) Start of second pulse; back ends of molecules from a) become leading ends for movement.
● and ● are turning points of each molecule.
c) End of second pulse. Smaller molecule has move further down the gel.
d) Beginning of third pulse; back ends from c) become leading ends for movement. ○ and ● are new turning points.
e) End of third pulse. Separation between molecules has increased since c). Degree of separation is proportional to total number of pulses.

Right hand side: The same two molecules are not resolved when *d* is reduced to less than the length of the small molecule. At end of second pulse, **c),** trailing ends have not passed initial turning point, and on third pulse, **d)** and **e),** molecules pass back through this turning point; i.e. molecules 'hang' in same position. Therefore maximum resolution of a PFG is determined by the pulse time; molecules longer than *d* form a compression band which cannot be resolved even if run time is increased.

Enzyme	Site	Sites per genome		% of sites	Sites per
		Bulk DNA	Island DNA	in islands	island
*Not*I	GCGGCCGC	500	3,600	89	0.12
*Bss*HII	GCGCGC				
*Eag*I	CGGCCG	1.2×10^4	3.5×10^4	74	1.2
*Sst*II	CCGCGG				
*Mlu*I	ACGCGT	2.7×10^4	1.0×10^4	27	0.34

Table 3.2: Calculated distribution of rare-cutter sites in cloned mammalian DNA. Adapted from Lindsay and Bird (1987). Assumptions for calculation were; 40% G + C, 25% expected CpG for bulk DNA; 65% G + C, no deficiency of CpG for island DNA. The genome size was taken to be 3×10^9bp. Islands were assumed to make up 1% of genomic DNA and to have an average length of 1kb.

Furthermore, it has been shown that sites for such enzymes are not distributed randomly in genomic DNA, but tend to be clustered in discrete 'islands' of high G + C content (>65%) in which the CpG frequency is close to that expected (Bird, 1986). The average spacing between islands has been estimated as ~100kb, but there may be considerable variation about this mean (Brown and Bird, 1986). An additional feature of CpG islands is that whilst CpG dinucleotides of bulk DNA are often methylated at cytosine, those found in island DNA are usually non-methylated (Bird, 1986). This means that in, genomic DNA, sites for methylation sensitive rare-cutters will be found almost exclusively in CpG islands. However, calculations suggest that even in cloned DNA, where CpG methylation is completely absent, sites for certain enzymes, such as *Not*I, *Bss*HII, *Eag*I and *Sst*II, will be preferentially located in islands (Lindsay and Bird, 1987) (Table 3.2).

Rare-cutter enzymes therefore cut DNA into fragments of an appropriate size for separation by pulsed field gel electrophoresis. A convenient method for mapping YAC inserts is the technique of indirect end-labelling (Fig. 3.4). Partial digests of the YAC are fractionated on a pulsed field gel and then probed with sequence from either of the two vector arms, revealing a ladder of bands. The size of each fragment detected corresponds to the distance from the appropriate YAC end to a cleavage site for the rare-cutter used. YACs are ideal for indirect end-label mapping since they are linear and any clones can be mapped with just two universal pBR322-derived probes, one from each vector arm (Burke *et al.*, 1987). In addition, the low sequence complexity of yeast DNA (0.5% of that in mammals) allows the detection of partial digest products which may be at very low copy number. Following the construction of a YAC map, the partial digests can be probed with an internal marker from the YAC; the position of the marker within the map is then deduced from analysis of the pattern of bands revealed.

It has been estimated that all housekeeping genes and 40% of tissue-specific genes are associated with CpG islands (Bird, 1987; Gardiner-Garden and Frommer, 1987; Larsen *et al.*, 1992). Rare-cutter mapping of YAC clones can therefore facilitate the identification of transcripts in the region of analysis.

3.1.4 Isolation of novel markers from YAC inserts

One potential drawback of YAC technology is the need to manipulate clones in a more complex background, that of the host yeast chromosomes. It is often desirable to isolate smaller fragments from the human insert of the YAC to facilitate chromosome walking, or further studies of the target region. Subcloning of an entire YAC insert into a plasmid or lambda vector is laborious and relies on the ability to separate the clone from the host chromosomes on a preparative pulsed field gel. However, several alternative techniques have been developed for the isolation of novel markers from YAC inserts.

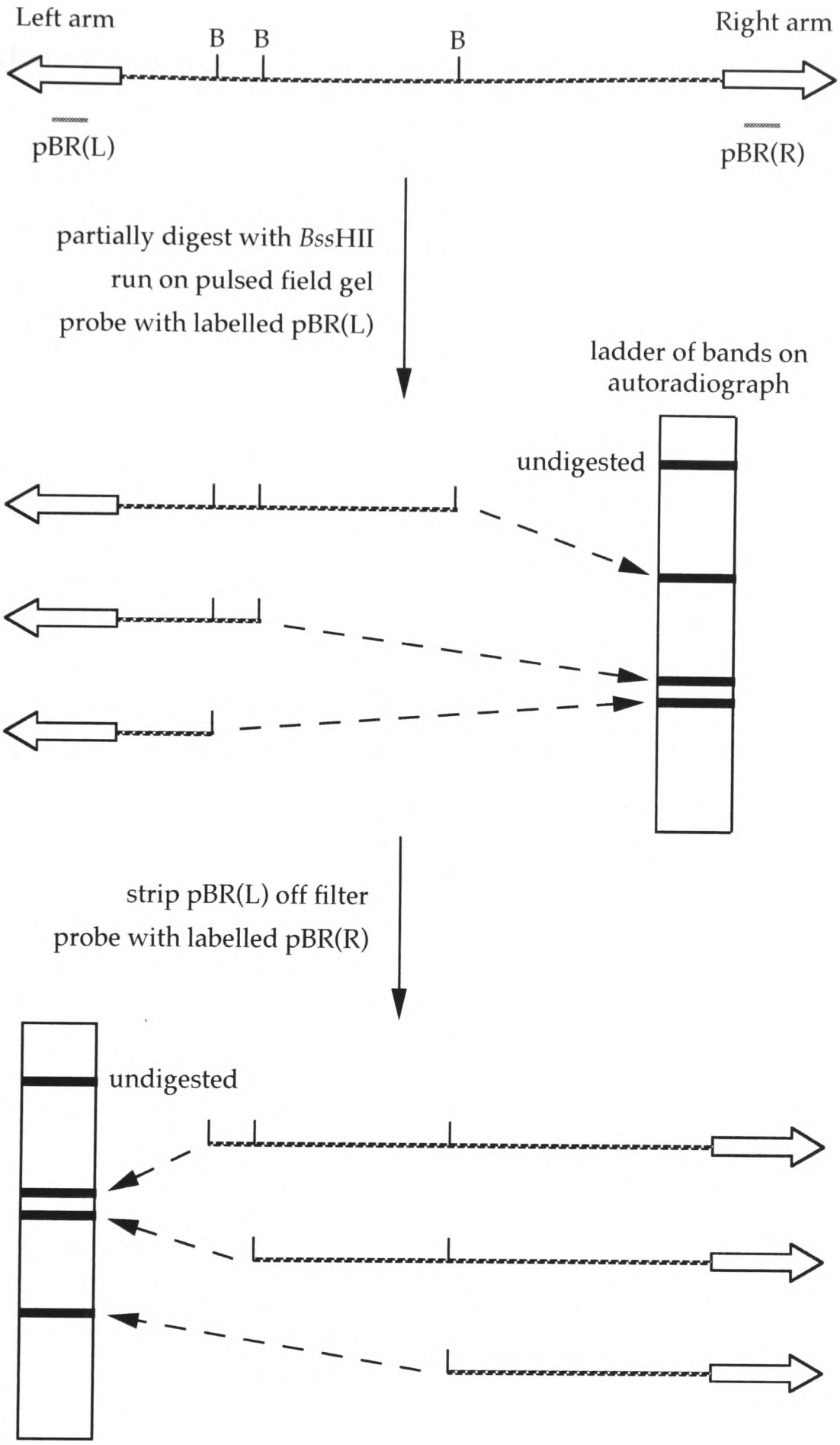


Figure 3.4: Schematic representation of the technique of indirect end-labelling as used for rare-cutter restriction mapping of YACs. Undigested YAC is shown at the top, with the positions of *Bss*HII sites (represented by 'B's) indicated. Partial digests with this enzyme are probed sequentially with pBR(L) and pBR(R) which are fragments from pBR322 that hybridize to the left and right YAC vector arms respectively. Each of these probes detects only those partial digest products which contain the appropriate end, revealing a ladder of bands, from which a map can be derived. The same approach can be used for each of the rare-cutting enzymes.

i) Left end cloning using plasmid rescue

Plasmid rescue (Burke *et al.*, 1987) exploits the fact that the pBR322-derived sequences of the left vector arm of pYAC4 include the origin of replication (ori) and an ampicillin resistance gene (Amp), which are the only parts of the plasmid that are necessary for replication and selection in *E. coli* (see Fig. 3.5). Total yeast DNA containing the YAC of interest is digested with an enzyme, such as *Nde*I, that cuts once in the left vector arm between the telomere and origin of replication. Among the generated products there will be a fragment which starts adjacent to the left TEL and extends to the first *Nde*I site of the insert. When the digested DNA is ligated in a large volume, to favour the formation of monomer circles, and transformed into *E. coli* with selection for ampicillin resistance, the only viable transformants will be those containing the left end clone. Electroporation is used for this step, since it gives high transformation efficiencies (~1 x 10^9 transformants/μg plasmid) and favours large clones (up to ~20kb). The end clone is then treated as a normal plasmid; an *Nde*I/*Eco*R1 double digest will release the insert fragment, which can be used as a hybridization probe to confirm its origin and screen for new YACs. In addition, the end clone can be sequenced using a primer from the left vector arm and the marker can be converted into a sequence tagged site (STS).

This protocol is dependent on the presence of an *Nde*I site in the YAC insert at an appropriate distance (i.e. between ~200bp and ~15kb) from the cloning site. However, the procedure can be modified to use *Xho*I or *Sal*I sites if necessary.

ii) Right end cloning using inverse-PCR

A limitation of pYAC4 is that the right vector arm does not contain sequences to enable its replication and selection in bacteria. Plasmid rescue is therefore unfeasible for right ends. Instead, an inverse-PCR (polymerase chain reaction) technique (Ochman *et al.*, 1988) is used. Whilst conventional PCR uses primers that are oriented towards each other, such that there is amplification of the region between them, inverse-PCR has primers pointing away from each other, and allows the amplification

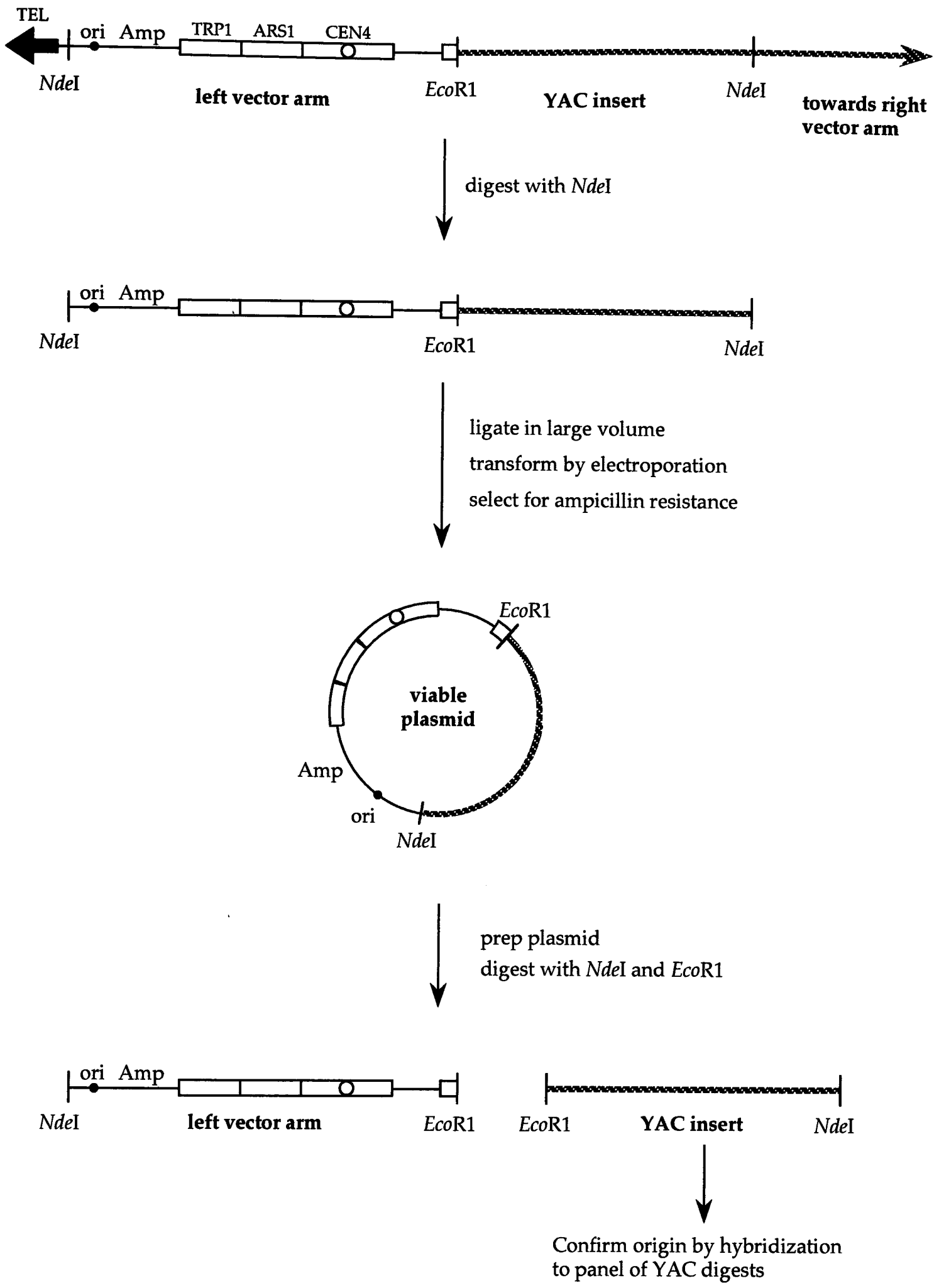


Figure 3.5: Schematic diagram showing the technique of plasmid rescue, used to isolate the left end of a YAC insert. See text for details.

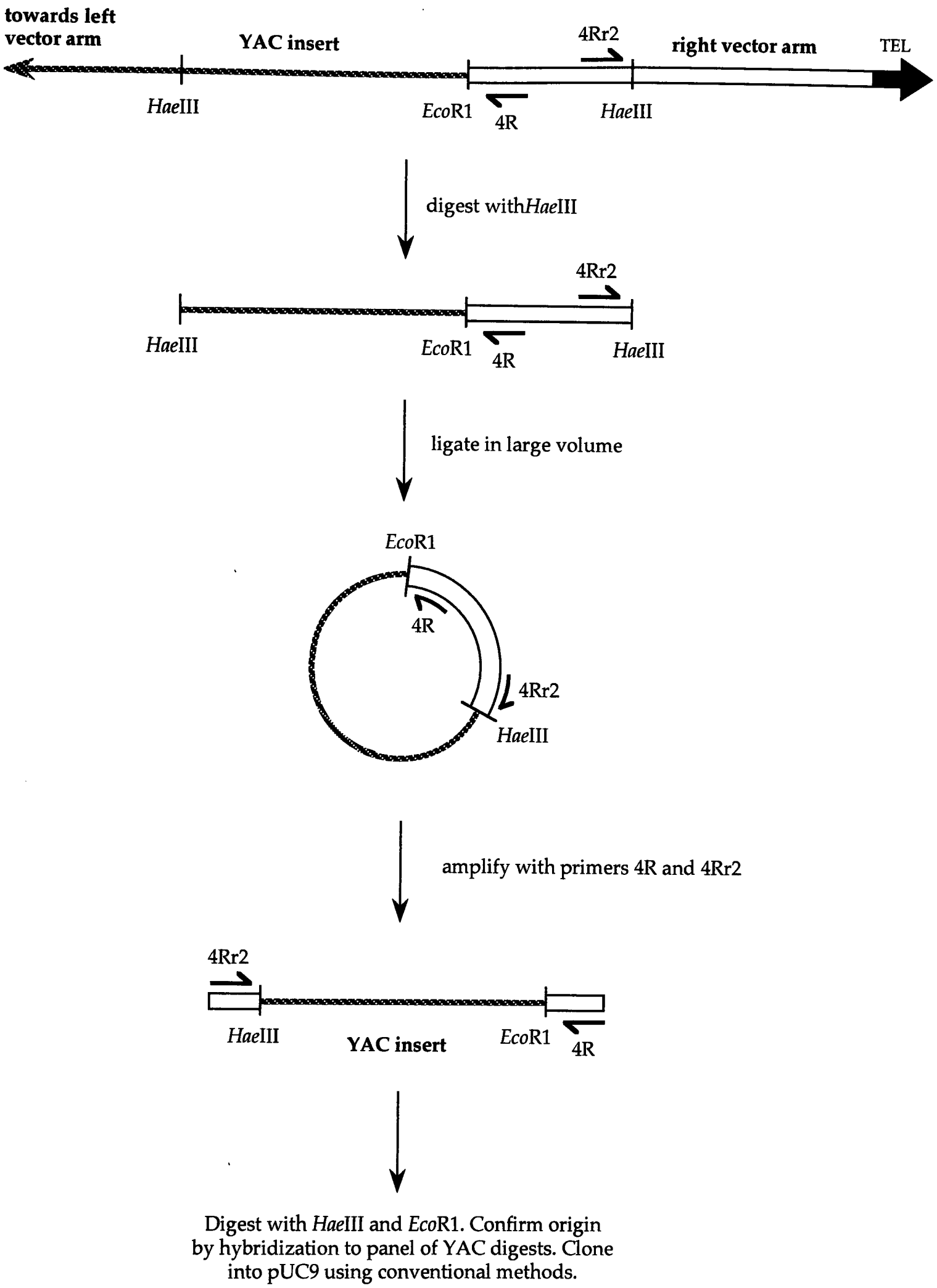


Figure 3.6: Schematic diagram showing the technique of inverse-PCR, used to isolate the right end of a YAC insert. See text for detials.

of flanking DNA of unknown sequence. Figure 3.6 shows the protocol as applied to right end cloning (Silverman *et al.*, 1989). Total yeast DNA containing the YAC is digested to completion with a frequent cutter such as *Hae*III, and fragments are ligated in large volume, to favour circularization. PCR using divergent primers from the right vector arm will result in amplification of human insert sequences from the vector-insert circles. Vector sequences can be trimmed from the ends of the product with an *Eco*R1/*Hae*III double digest, and it can be used as a hybridization probe to confirm its orgin. The product can be conventionally cloned into a plasmid, sequenced, and converted into an STS. As with left end cloning, the procedure relies on an appropriately positioned frequent-cutter site in the YAC insert, and a range of enzymes can be used.

iii) *Alu*-PCR

This technique exploits the discovery that interspersed repetitive DNA sequences make up a significant proportion of all mammalian genomes. The most abundant SINE (short interspersed repeat element) in primates is that known as the *Alu* family. The consensus sequence of *Alu* elements is about 300bp long, and is dimeric in structure, with an A-rich region separating the two units, and an oligo (dA) tail of variable length at its 3' end (Deininger *et al.*, 1981). *Alu* repeats are estimated to be present at a copy number of between 500,000 and 900,000 in the human haploid genome, which suggests an average distance of 3-6kb between copies (Deininger *et al.*, 1981; Korenberg and Rykowski, 1988). However, there may be substantial variation in inter-*Alu* distance, since studies have shown that the repeats are not uniformly distributed through the genomes, but tend to be clustered in R (reverse) bands of metaphase chromosomes (Korenberg and Rykowski, 1988). These GC-rich bands replicate their DNA early, are the main sites of chromosomal exchange processes, and are also enriched for CpG islands and active genes (Craig and Bickmore, 1994).

Homology studies have indicated that the *Alu* family is descended from the 7SL RNA gene (Ullu *et al.*, 1982). Although related mammalian species can have nearly identical SINE families, the intraspecies homogeneity of the repeats is significantly greater than the interspecies homogeneity (Deininger and Daniels, 1986). Analysis of species-specific subfamilies has led to the proposal that an extremely small group of 'master' genes (derived from a parent such as the 7SL RNA) have been responsible for the amplification of repetitive elements, via a retrotransposition mechanism involving an RNA intermediate (Deininger and Daniels, 1986; Deininger *et al.*, 1992).

Alu-PCR was originally developed to enable rapid isolation of human-specific DNA fragments from human/rodent hybrid cell lines (Nelson *et al.*, 1989). A primer corresponding to a highly conserved part of the *Alu* repeat is used for PCR with DNA from the cell line as a template. Products should be obtained from any regions where two adjacent *Alu*s are in inverted orientation relative to one another, provided that they are close enough together (i.e. less than a few kilobases apart) for efficient amplification. As discussed above, the SINES present in rodent DNA, while related to *Alu*, diverge from it sufficiently, so that with the appropriate choice of primers only human DNA is amplified. Furthermore, the conservation of *Alu* sequence within the human species is high enough to ensure that products can be generated from most parts of the genome (provided they are *Alu*-rich). In addition, given that *Alu*s are clustered in regions which are enriched for active genes (and are often found in introns), this technique may preferentially amplify sequences associated with genes.

Alu-PCR can also be used to amplify YAC sequences from yeast background. The protocol followed in this thesis is a modification of the original method, which used primers from the internal part of the *Alu*, and therefore generated products which were repetitive. The primers used here are based on A1, which is an 18 base oligonucleotide corresponding to the 3' end of a frequently observed variant of the *Alu* consensus sequence (Brooks-Wilson *et al.*, 1990). They contain additional residues at their 5' ends to facilitate conventional cloning of products into plasmids.

3.1.5 Aims

The aim of the work described in this chapter was to construct a YAC contig spanning the hypervariable locus DXS255 in Xp11.22, a region which had been only poorly characterized prior to this study, in order to provide a basis for identification of transcripts that might be implicated in closely linked diseases, such as Wiskott-Aldrich syndrome (Cremin *et al.*, 1993) and retinitis pigmentosa 2 (Meitinger *et al.*, 1989). In addition, after this work was begun, it was reported that a form of nephrolithiasis known as Dent's disease mapped in Xp11.22, very close to DXS255 (Pook *et al.*, 1993).

DXS146 is a polymorphic locus recognized by the anonymous single copy probe pTAK8 (Kruse *et al.*, 1986). Linkage analysis and physical mapping using somatic hybrids has localized it to Xp11.22, proximal to DXS255 (Lafreniere *et al.*, 1991b; Cremin *et al.*, 1993). In addition, genomic pulsed field mapping has shown that the distance between DXS255 and DXS146 is in the range of 230-900kb (Riley, 1993). Hybridization of pTAK8 to YAC library filters has previously resulted in the isolation of four DXS146 YACs (A.P. Monaco, unpublished). End cloning has been used to generate novel markers from these YACs (Hatchwell, 1994).

The 'chromosome walking' strategy for contig construction involved the isolation of YACs with DXS255, followed by the generation of novel markers (using techniques described above) flanking the locus, which would then be used to screen YAC libraries for overlapping clones. The objective was to take further 'steps' in each direction, until the contig linked up with the DXS146 YACs on the proximal side, and TFE3/SYP YACs on the distal side (see Chapter 4). The new markers generated by this procedure could be ordered on the basis of YAC analysis, and converted into sequence tagged sites (STSs). An additional aim was to make rare-cutter restriction maps of the YACs isolated; as well as giving an idea of the extent of YAC overlap and positions of novel markers, this allows the identification of CpG islands which may be associated with transcripts in the region.

3.2 Materials and methods

3.2.1 Probes

M27β is a 2.3kb *Eco*R1 fragment which recognizes the hypervariable DXS255 locus (Fraser *et al.*, 1989). The L(G0201) probe is an *Eco*R1/*Nde*I fragment previously isolated from the DXS146 YAC G0201 (Hatchwell, 1994), and was kindly provided by E. Hatchwell.

3.2.2 PCR Screening of YAC libraries (Green and Olson, 1990)

Three libraries were available for screening by PCR:-

i) Clones are isolated from the **ICRF 4X library no.900** (Larin *et al.*, 1991) using a three step process. 41 primary pools are screened with a PCR assay developed for the chosen marker (step 1). Each primary corresponds to four secondary pools, and each secondary consists of DNA pooled from a plate containing a grid of 8 x 12 = 96 YAC clones. PCR screening of the appropriate secondary pools (step 2) identifies the plate containing the clone of interest. A 'rows and columns' PCR (step 3; see Fig. 3.7 for an example), involving 20 samples (corresponding to 8 rows and 12 columns) prepared from the appropriate plate, yields the grid reference of the positive YAC.

ii) The **St. Louis library** (Brownstein *et al.*, 1989) consists of 38 primary pools (A1-19, B1-17 and C1-2). Each primary corresponds to four secondaries, and each secondary to four tertiaries. The tertiary screening identifies the plate containing the clone of interest. The screening process therefore involves four steps, the last of which is a 'rows and columns' PCR as described above.

iii) Isolation of clones from the **ICI library** (Anand *et al.*, 1990) involves screening of 40 primary pools (step 1), followed by screening of nine secondary pools for each primary positive identified (step 2). Again, the final stage is a 'rows and columns' PCR of the appropriate plate (step 3).

Following identification of the putative positive YAC clone, it is streaked out onto an SD agar plate to give single colonies. Plugs prepared from colony-pure liquid cultures are analysed by PCR and hybridization to confirm that they are indeed positive for the marker used to isolate the YAC. Colony purification is important, because YACs from a single grid reference on a library plate may sometimes be in mixed culture. In addition, PFGE sizing and hybridization analysis of several different colony-pure preps of a single YAC is instrumental in detecting any rearrangements that may result from clonal instablility (see Discussion).

The F1001 YAC was kindly provided by A.P. Monaco, who isolated it by hybridization screening of the ICRF library using the M27β probe.

3.2.3 Plasmid rescue for isolation of left ends from YAC inserts

1. Probe *Nde*I, *Sal*I and *Xho*I digests of the YAC with left arm to establish which enzyme will generate the smallest left end fragment. In all cases described in this thesis, *Nde*I was the enzyme of choice.

2. Digest a YAC plug to completion with *Nde*I, and then equilibrate it at room temperature with 1ml of TE.

3. Remove TE and extract digested DNA from agarose plug using the Geneclean procedure (Section 2.8).

4. Make mix up to 73µl with sterile water. Add 20µl of 5 x ligase buffer and 7µl of T4 DNA ligase (BRL; 1 unit/µl).

5. Ligate at 14-16°C overnight.

6. Precipitate DNA by adding 11µl 3M NaOAc, pH5.2, and two volumes EtOH.

7. Spin for 15 minutes at 4°C. Wash pellet in 70% EtOH. It is important to remove all salt in order to ensure that cells do not short in the electroporation step.

8. Resuspend DNA in 4µl of sterile water.

9. Remove an aliquot of electrocompetent cells (Top Ten; Invitrogen) from liquid nitrogen storage and thaw on ice.

10. Add 2µl of DNA to cells, mix and transfer to a prechilled cuvette (Bio-Rad), making sure that no air bubbles are present and that the cells are in contact with both electrodes at the bottom of the cuvette.

11. Place the cuvette in the electroporator (Gene Pulsar; Bio-Rad) and apply a pulse, using the settings: 2.5kV, 200Ω and 25µF.

12. Immediately add 1ml of SOC, prewarmed to room temperature.

13. Incubate in a 37°C shaking incubator for 1 hour, and then plate out on LB containing ampicillin.

14. Plasmid is prepped from transformants and digested with *Nde*I/*Eco*R1 to confirm that the end clone has been isolated.

3.2.4 Inverse-PCR for isolation of right ends from YAC inserts

1. Digest a YAC plug to completion with *Hae*III, and then equilibrate it at room temperature with 1ml of TE.

2. Remove TE and extract digested DNA from agarose plug using the Geneclean procedure (Section 2.8).

3. Make mix up to 73µl with sterile water. Add 20µl of 5 x ligase buffer and 7µl of T4 DNA ligase (BRL; 1 unit/µl).

4. Ligate at 14-16°C overnight.

5. Incubate at 65°C for 20 minutes to inactivate ligase.

6. Use varying amounts of the ligated mix (1µl, 5µl or 10µl) as template in 100µl PCR reactions using the primers 5' AGTCGAACGCCCGATCTCAA 3' and 5' TTCAAGCTC-TACGCCGGA 3' from the right vector arm. Cycling conditions are as follows: 94°C; 5 min (= 'hot start'); 94°C; 1 min; 55°C; 1 min; 75°C; 1 min; 35 cycles; 75°C; 2 min (= end elongation).

7. Any product obtained can be excised from a gel and reamplified.

8. The product is further analysed by hybridization to YAC and human panels to confirm its origin.

All products described in this thesis were isolated using *Hae*III. However an alternative protocol using *Hinc*II instead of *Hae*III can be followed if necessary. PCR primers for the *Hinc*II protocol are 5' AGTCGAACGCCCGATCTCAA 3' and 5' GGAGTCGCATAA-GGGAGAGC 3'.

3.2.5 Generation of internal markers from YACs using *Alu*-PCR

This is essentially a conventional PCR from melted YAC plug template (see section 2.14.7) using two different A1 primers, one modified to include a *Bam*H1 site at its 5' end, the other modified to include a *Sal*I site at its 5' end:-

A1B = 5' TCATGGATCCGCGAGACTCCATCTCAAA 3'
A1S = 5' TCATGTCGACGCGAGACTCCATCTCAAA 3'

1. Several independent 100µl PCR reactions are set up for each YAC with 5µl template and varying magnesium concentrations from 1.5mM to 4.0mM. Sometimes products obtained differ depending on $[Mg^{2+}]$ of the reaction.

2. Cycling conditions are as follows: 94°C; 5 min (= 'hot start'); 94°C; 1 min; 55°C; 1 min; 75°C; 2 min; 35 cycles; 75°C; 2 min (= end elongation).

3. After PCR, run products on a preparative gel, excise band(s) and purify using Geneclean procedure (section 2.8).

4. Purified products can be used directly as hybridization probes to confirm that they originate from YACs.

5. Digest products with *Bam*H1 and *Sal*I, and clone into *Bam*H1/*Sal*I cut pUC9 vector using conventional techniques (section 2.10).

3.3 Results

3.3.1 Isolation and analysis of YACs containing DXS255

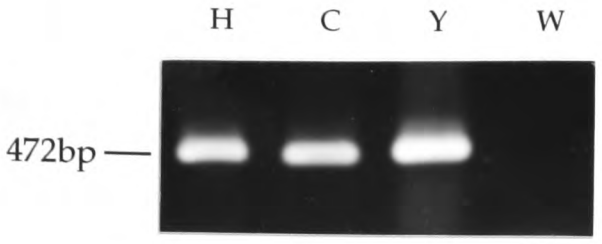
i) Screening of YAC libraries

The probe M̂27β, which recognizes the DXS255 locus, was previously used to isolate a YAC from the ICRF library, clone F1001 (A.P. Monaco, personal communication). A PCR assay was developed using sequence from the 5' end of the DXS255 locus (Hendriks *et al.*, 1992), to facilitate further screening of the ICRF and St.Louis libraries. On PCR screening of primary pools, strong amplification of the expected product was only seen from a single pool; A3 of the St. Louis library (Fig. 3.7). Secondary and tertiary screenings localized the positive YAC to plate A39, and PCR of the rows and columns corresponding to this plate indicated a grid reference of E7 for this clone. Subsequent PCR analysis of several independent colony-pure preps of A39 E7 confirmed that this was the correct clone. In addition, hybridization of M27β to an *Eco*R1 digest of this YAC detected a 2.3kb fragment. It is interesting to note that this corresponds to the smallest allele observed for the DXS255 locus, and comparison to the allele size in the parental cell line indicated that a deletion of ~4.6kb had occurred from the *Eco*R1 fragment during library construction (not shown). This provides further confirmation that the VNTR of DXS255 is refactory to cloning (see Section 1.3.1). Deletion of this region has also been found in the F1001 YAC (not shown). The official identification number of the A39 E7 clone is 6129.

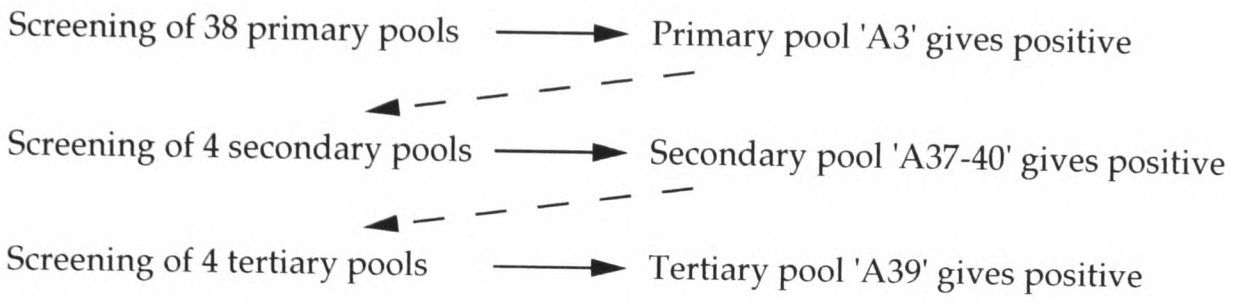
ii) Rare-cutter restriction mapping of DXS255 YACs

Sizing pulsed field gels of undigested DXS255 YACs indicated that F1001 is ~1135kb, while 6129 is ~185kb. No evidence of instability (beyond that described above) was found for either of these clones. Each YAC was mapped using partial digests of the rare-cutting enzymes *Bss*HII, *Eag*I, *Mlu*I, *Not*I, *Sal*I, *Sfi*I and *Sst*II, which were run on pulsed field gels with an appropriate switch time, and then probed sequentially with left and right vector arm (Figures 3.8-3.11; Tables 3.3 and 3.4).

a)



b)



c)

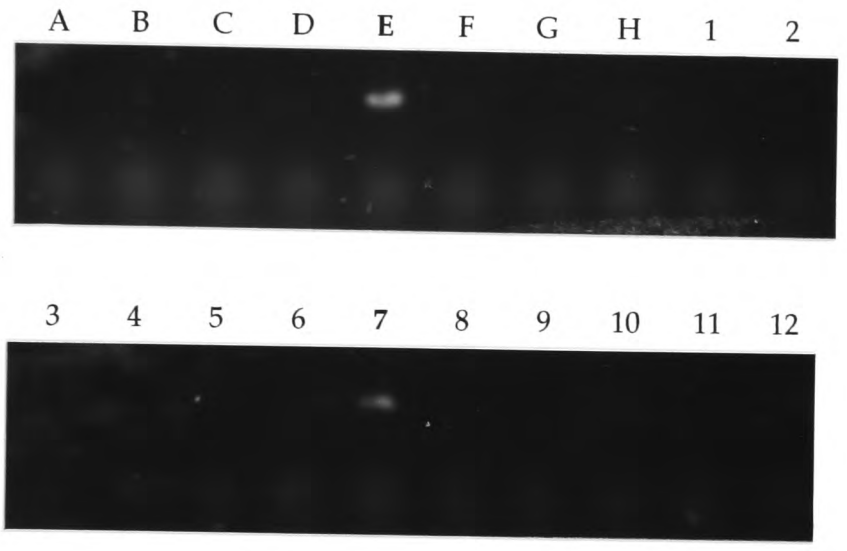


Figure 3.7: Isolation of a DXS255 YAC from the St. Louis library using PCR screening.

a) A PCR assay (MBP) developed from the 5' end of DXS255 (see Table 3.8). The expected 472bp product is amplified from human (H), M27 cosmid (C) and F1001 YAC (Y) templates but not in the water control (W).

b) A three stage PCR screening process to identify which plate of the library contains a YAC which is positive for MBP.

c) The final 'rows and columns' screening of plate A39. PCR of pooled DNA from 8 rows (A-H) and 12 columns (1-12) results in two positives which give a grid reference for the positive YAC. Subsequent experiments are necessary to confirm that this is a true positive.

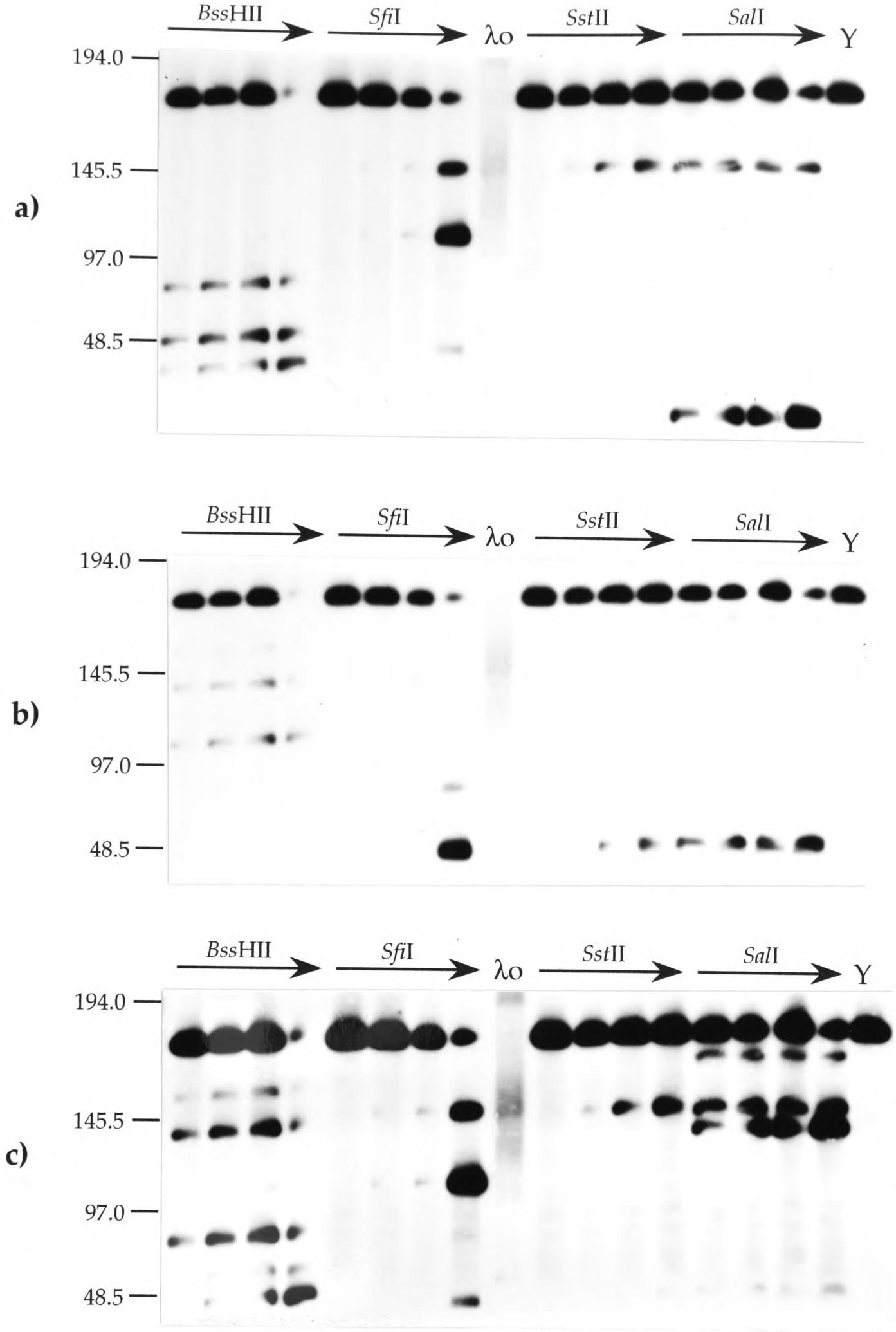


Figure 3.8: Partial digests of the 6129 YAC using different rare cutters, probed with **a)** left vector arm, **b)** right vector arm and **c)** M27β. Direction of arrow represents increasing enzyme concentration (from 0.1-5U) with a 1 hour digestion time. Sizes of lambda oligomer markers (λo) are given in kilobases. The track labelled Y contains undigested YAC. Digests were run on a standard pulsed field gel (section 2.14.5) with a 13 second switch time and a 31 hour run time. Fragment sizes and the rare-cutter restriction map which was derived from them are given in Table 3.3 and Figure 3.9. Note that the 40kb bands seen in *Sfi*I and *Sal*I digests when probing with M27β in fact represent residual signal from the right vector arm hybridization, due to insufficient filter stripping between experiments.

Enzyme	Left vector arm	Right vector arm
*Bss*HII	25, 45, 85	105, 140, 160
*Eag*I	145[a], 160	25, 35, 45
*Sal*I	<15, 145	40, 175
*Sfi*I	105, 145	40, 80
*Sst*II	145	40

Enzyme	M27β
*Bss*HII	40, 60, 85, 140, 160
*Eag*I	145[a], 160
*Sal*I	135, 145, 175
*Sfi*I	105, 145
*Sst*II	145

Table 3.3: Fragment sizes, in kilobases, of bands detected on rare-cutter partial digests of the 6129 YAC clone, when probed with vector (left and right arms), and the internal marker M27β. The 185kb fragments corresponding to undigested YAC are not listed.
[a] very strong 145kb *Eag*I band may mask the presence of a doublet.

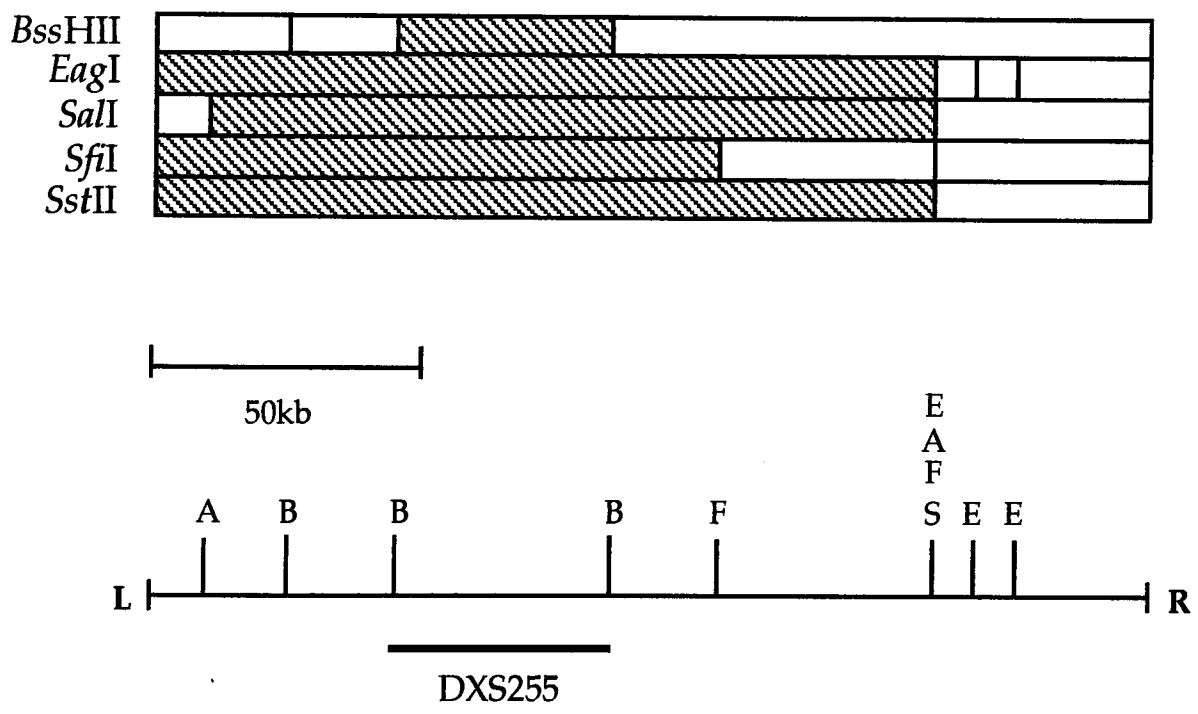


Fig. 3.9: Rare-cutter restriction map of 6129 YAC clone, derived from fragment sizes in Table 3.3, with the position of DXS255 (the locus recognized by M27β) indicated below. Diagonal stripes show fragments on which DXS255 lies. **L**, left arm; **R**, right arm; B, *Bss*HII site; E, *Eag*I; A, *Sal*I; F, *Sfi*I; S, *Sst*II. There are no sites for *Mlu*I or *Not*I in this YAC.

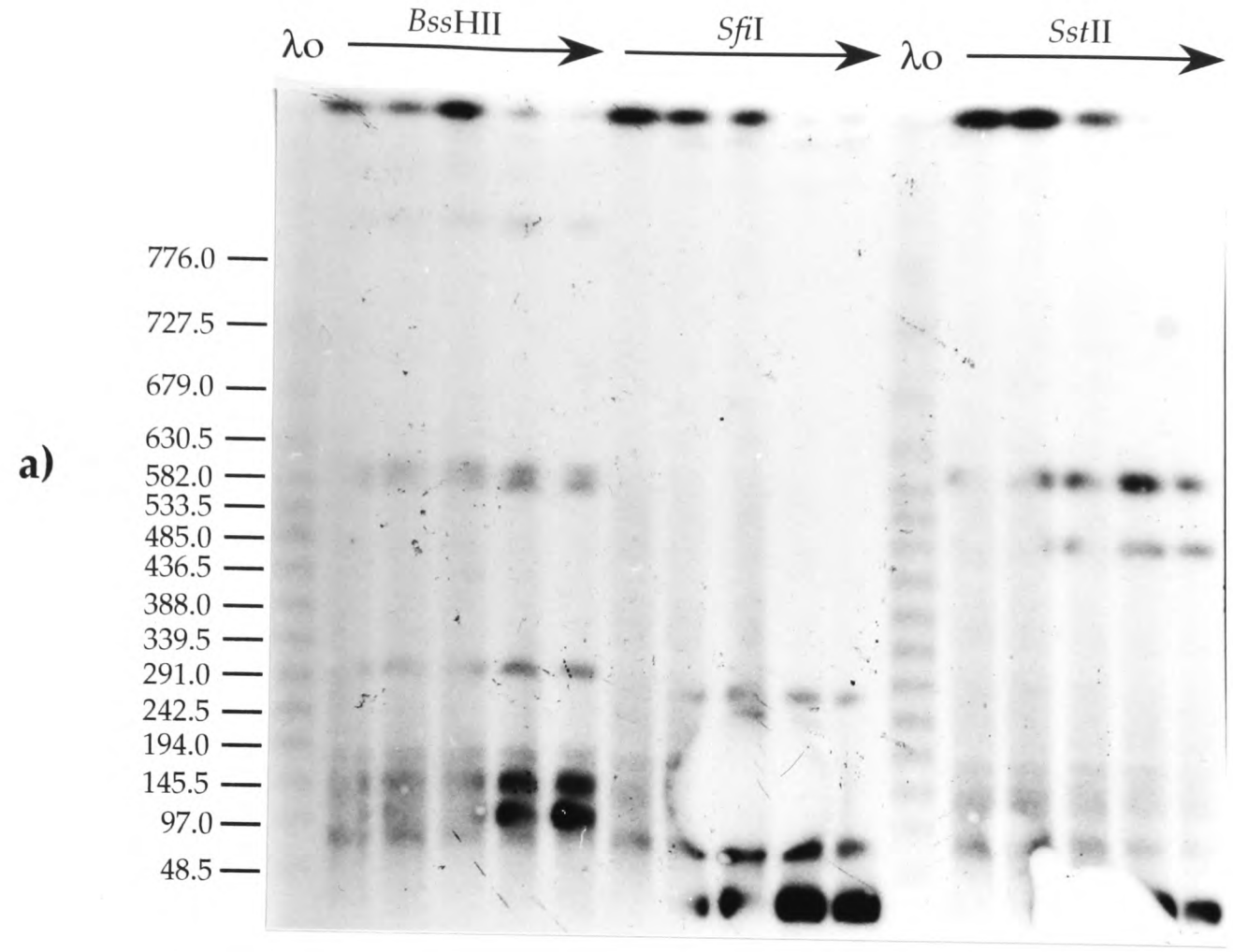


Figure 3.10: Examples of partial digests of the F1001 YAC using different rare cutters, probed with **a)** left vector arm and **b)** right vector arm. Direction of arrow represents increasing enzyme concentration (from 0.1-15U) with a 1 hour digestion time. Sizes of lambda oligomer markers (λo) are given in kilobases. Digests were run on a standard pulsed field gel (section 2.14.5) with a 60 second switch time and a 30 hour run time. Fragment sizes and the rare-cutter restriction map which was derived from them are given in Table 3.4 and Figure 3.11. The absence of signal from sections of the filter in some *Sfi*I and *Sst*II tracks is due to air bubbles during Southern blotting.

Enzyme	Left vector arm	Right vector arm	M27β
*Bss*HII	105, 145, 170, 290, 570, 590, 805	165, 235, 295, 505, 530, 810	40, 60, 145, 170, 290, 570, 590, 815[a]
*Eag*I	20, 30, 40, 355, 590, 605	125, 225, 260, 295	320, 355, 570, 590, 605
*Mlu*I	290	70, 370	290
*Not*I	605, 805	95, 295, 495	605, 805
*Sfi*I	40, 80, 270	165, 235, 840	195, 235, 270
*Sst*II	40, 495, 590	235, 270, 295, 505	460, 495, 555, 590

Table 3.4: Fragment sizes, in kilobases, of bands detected on rare-cutter partial digests of F1001 YAC clone, when probed with vector (left and right arms), and the internal marker M27β. The 1100kb fragments corresponding to undigested YAC are not listed. [a] additional weak bands are present.

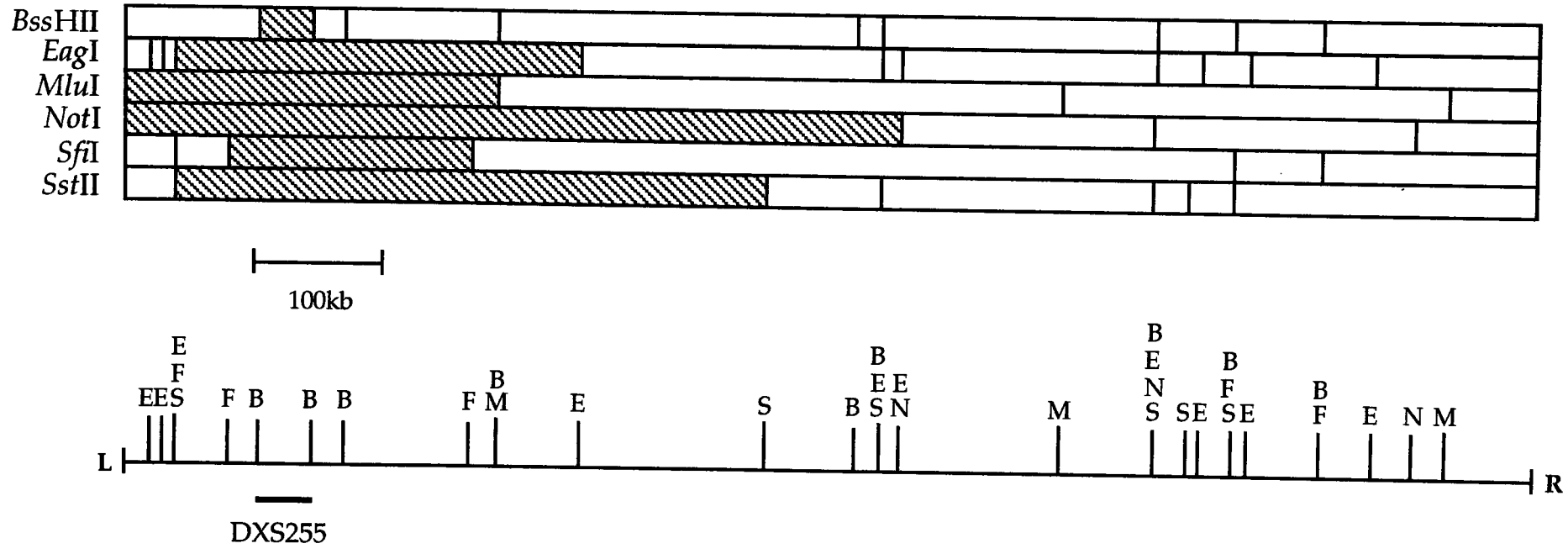


Fig. 3.11: Rare-cutter restriction map of F1001 YAC, derived from fragment sizes in Table 3.4, with the position of DXS255 (the locus recognized by M27β) indicated below. Diagonal stripes show fragments on which DXS255 lies. **L**, left arm; **R**, right arm; B, *Bss*HII site; E, *Eag*I; M, *Mlu*I;N, *Not*I; F, *Sfi*I; S, *Sst*II.

Following this, hybridization with M27β indicated that DXS255 is localized to a 40kb *Bss*HII fragment common to both YACs (Figures 3.8-3.11; Tables 3.3 and 3.4). Previous studies have shown that DXS255 is associated with a CpG island containing a *Bss*HII site (see Section 1.3.3). The locus must therefore map to one end of the 40kb *Bss*HII fragment. Comparison of the rare-cutter restriction maps indicates that 6129 has almost complete overlap with the region adjacent to the left end of F1001, but is oriented in an opposite direction with respect to the X chromosome (see Fig. 3.15). Neither YAC was found to contain the proximal marker DXS146. The genes SYP and TFE3, which map distal to DXS255 (see section 4.1.1) were also absent.

iii) Cloning of left ends from DXS255 YACs

Plasmid rescue was used to isolate the left ends from the insert of each DXS255 YAC (Fig. 3.12), designated L(F1001) and L(6129). Details of these and other left ends are given in Table 3.7. Hybridization of these clones to genomic/hybrid mapping panels indicated that they are both X-specific (Figs. 3.13 and 3.14). Furthermore, each clone was found to be present in both DXS255 YACs (Figs. 3.13 and 3.14). Hybridization to pulsed field partial digests of each YAC enabled the clones to be localized within rare-cutter restriction maps (Fig. 3.15). It was found that L(F1001) and L(6129) are ~180kb apart and flank the DXS255 locus, confirming that the two YACs are oriented in opposite directions with respect to the X chromosome (Fig. 3.15). These novel markers could therefore serve as a starting point for a bi-directional YAC walk.

An autosomal *Alu*-PCR clone has previously been isolated from F1001 (S. Riley, personal communication), suggesting that this is a chimæric YAC. The analysis of left end clones described above indicates that at least 180kb of the region adjacent to the left vector arm of F1001 is from Xp11.22, but does not clarify the extent of the autosomal region in the remainder of this YAC.

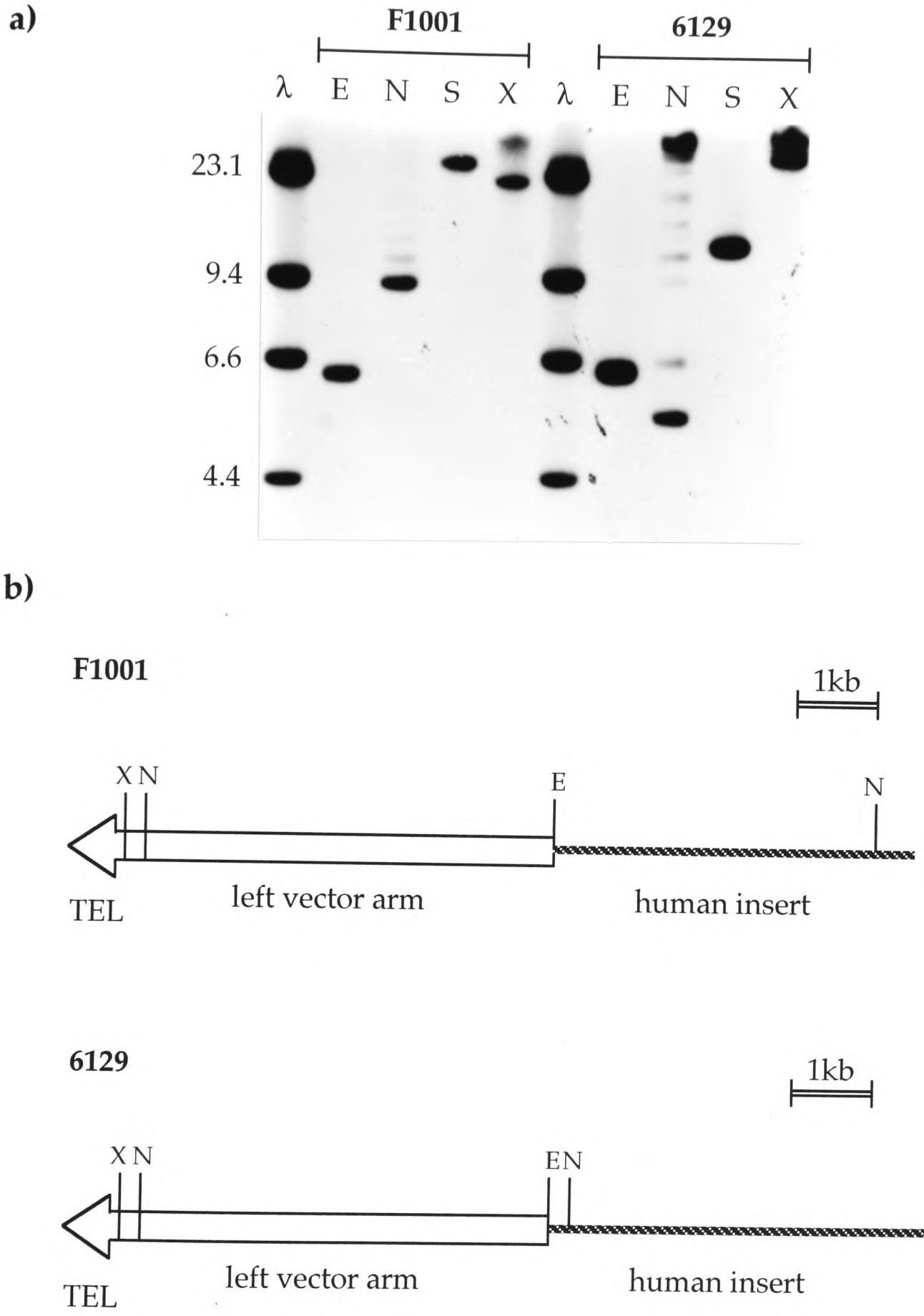


Figure 3.12: Analysis of left ends of DXS255 YACs.

a) Digests of DXS255 YACs, probed with pBR(L), which recognizes the left vector arm of pYAC4. Enzymes used were *Eco*R1 (E), *Nde*I (N), *Sal*I (S) and *Xho*I (X). Sizes of λ markers in kilobases are indicated. pBR(L) detects a constant ~6.0kb *Eco*R1 band in both YACs, corresponding to the complete vector arm. Note that the ladder of fragments above the main band in the *Nde*I tracks is due to incomplete digestion with this enzyme.

b) Restriction maps of left ends of DXS255 YACs, derived from results in a) and the reported sequence of the pYAC4 vector (accession number U01086). These results indicate that plasmid rescue using *Nde*I should be possible for both YACs, giving ~4.0kb of human insert from F1001 and ~200bp of human insert from 6129.

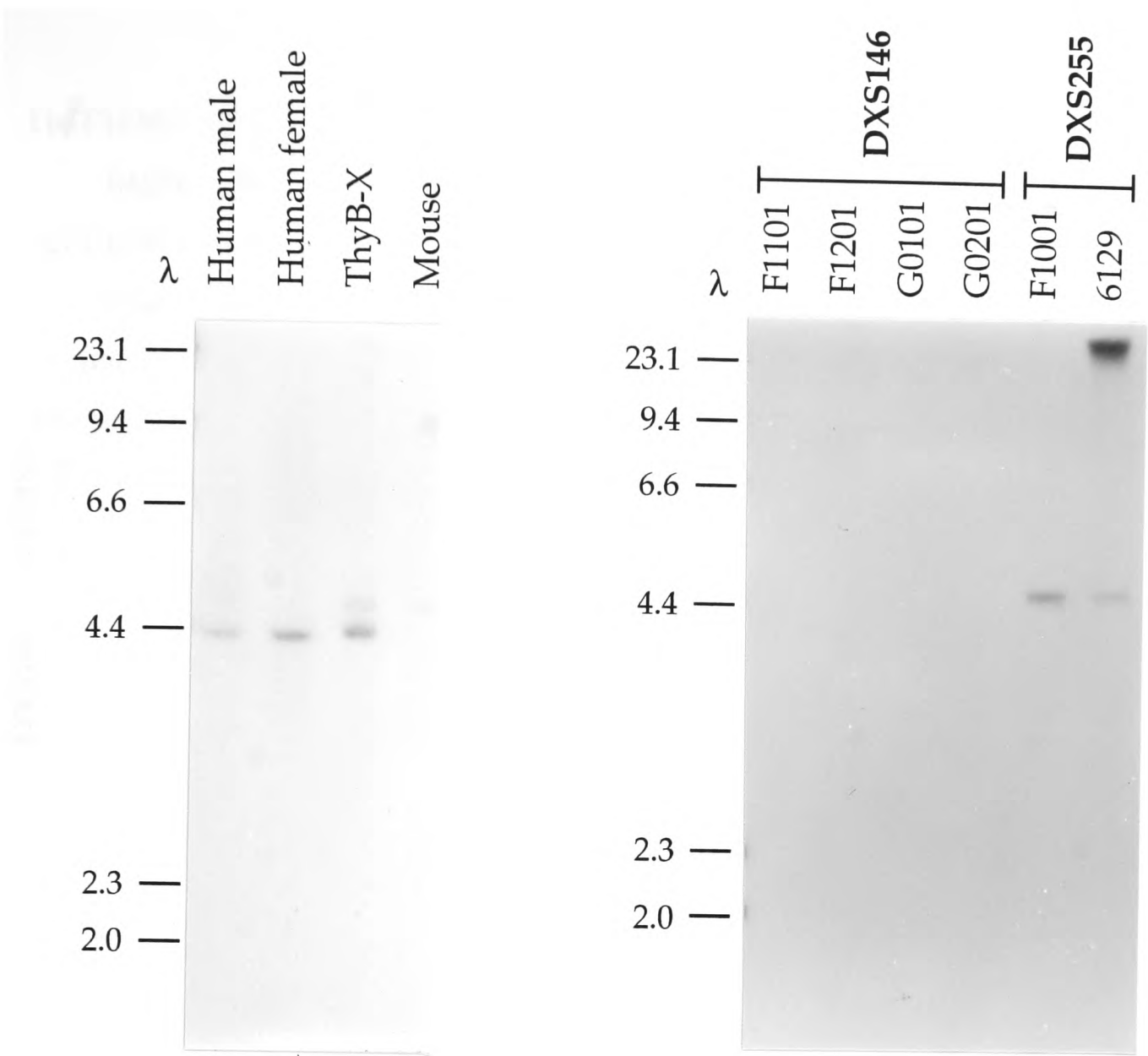


Figure 3.13: Hybridization of the L(6129) probe to *Eco*R1 digests of genomic, hybrid and YAC DNAs. Positions and sizes, in kilobases, of lambda markers (λ) are indicated.
a) L(6129) detects a 4.4kb band in human male and female genomic DNA, and also in the human X-only/mouse hybrid (ThyB-X). This band is absent from mouse genomic DNA. Note, however, that an additional faint fragment of 4.6kb is detected in ThyB-X and mouse DNA. It later became apparent from analysis of cDNAs isolated using the 6129 YAC (as described in Chapter 5) that L(6129) originates from the 3' non-coding region of a human gene (CLCN5), and that this 4.6kb band corresponds to the homologous region in mouse.
b) On probing *Eco*R1 digests of a panel of YACs, L(6129) detects a 4.4kb band in both DXS255 YACs (F1001 and 6129,the YAC of origin). This fragment is absent from all four YACs of the DXS146 cluster. Hybridization to the high molecular weight material in the 6129 track is due to incomplete digestion of this YAC.

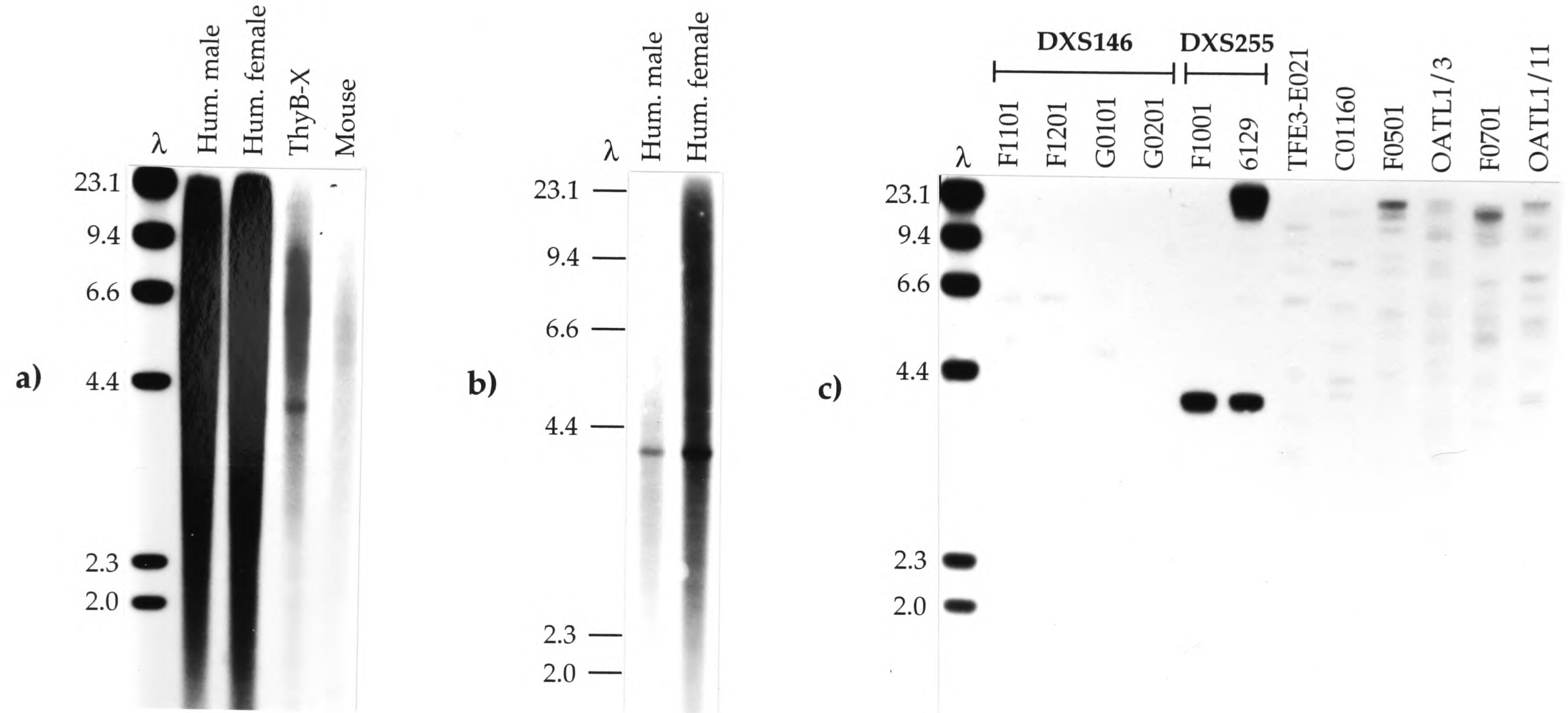


Figure 3.14: Hybridization of the L(F1001) probe to *Eco*R1 digests of genomic, hybrid and YAC DNAs. Positions and sizes, in kilobases, of lambda markers (λ) are indicated.

a) L(F1001) detects a smear when used to probe human genomic DNA, indicating that it contains repetitive sequences. A predominant band of 4.0kb is seen in ThyB-X, but not in mouse genomic DNA.

b) After prereassociation to compete out repetitive elements (Section 2.9.3), the smear in human genomic DNA is reduced, and a 4.0kb fragment is clearly detected.

c) On probing *Eco*R1 digests of a panel of YACs, L(F1001) detects a 4.0kb band in both DXS255 YACs (F1001 and 6129,the YAC of origin). This fragment is not present in any of the YACs from the DXS146 cluster (proximal to DXS255) or from the OATL1-GATA-TFE3-SYP cluster (distal to DXS255; see Chapter 4). Hybridization to the high molecular weight material in the 6129 track is due to incomplete digestion of this YAC.

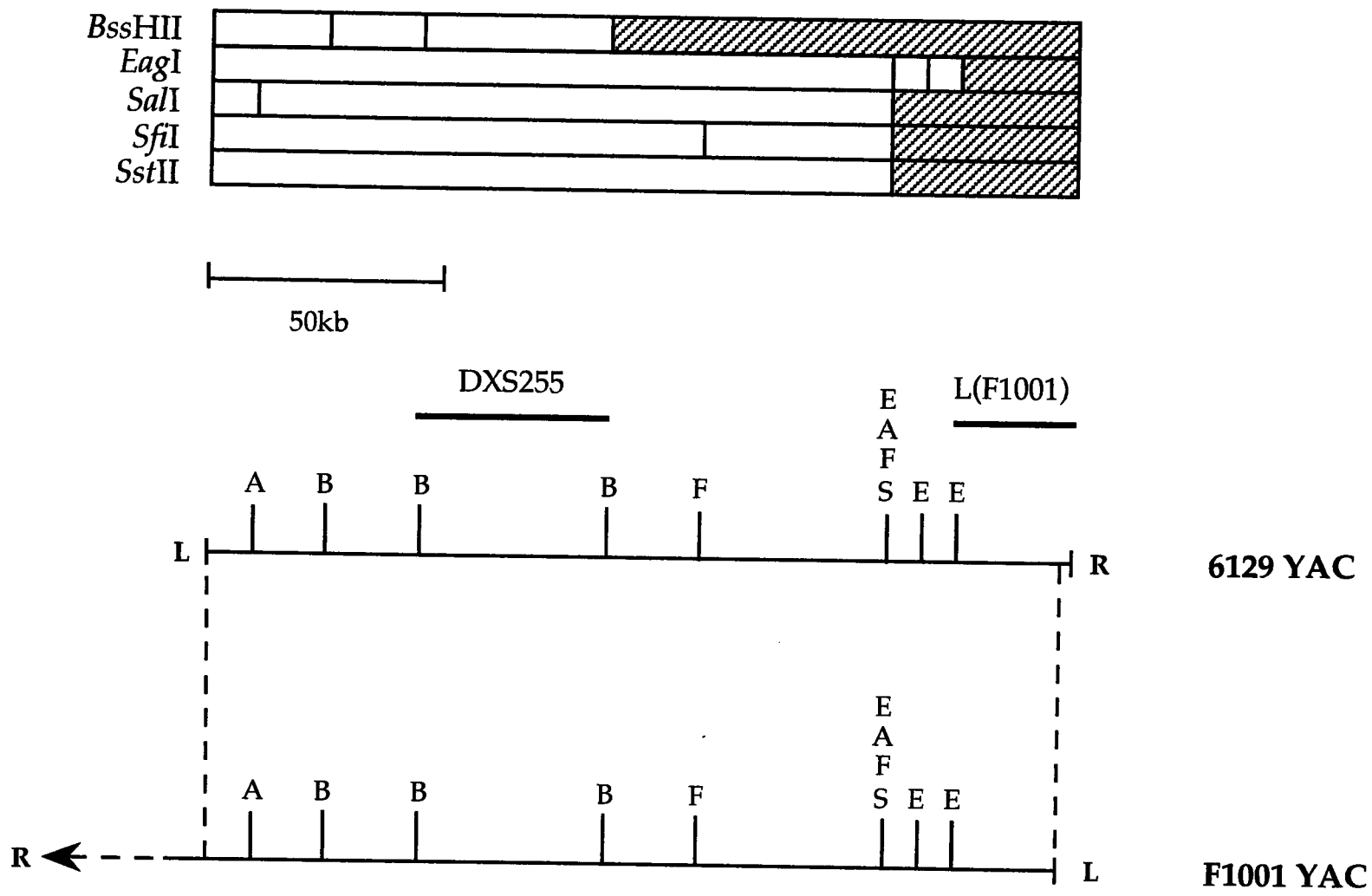


Figure 3.15: Diagram showing overlap and relative orientations of DXS255 YAC cloness, as determined by comparison between rare-cutter maps, and positioning of L(F1001) in the 6129 YAC map. **TOP;** On hybridization to partial digests of 6129, L(F1001) gives an identical pattern to right vector arm (see Fig. 3.9). Shaded areas show fragments on which L(F1001) must therefore lie. **BOTTOM;** Alignment of the 6129 YAC with the left end region of F1001 YAC. Deduced positions of L(F1001) and DXS255 are shown above. Labelling is as in Fig. 3.9. The orientation of these YACs with respect to the X chromosome was unknown at this stage, and is therefore not shown, although it was later established by chromosome walking (see text).

3.3.2 A bi-directional YAC walk using left ends of the DXS255 YACs

i) A YAC isolated using L(6129) links up with a more proximal contig containing DXS146

L(6129) was completely sequenced (see Appendix) and converted into an STS (Table 3.8), which was used to screen for new YACs, resulting in the isolation of clone C0191 from the ICRF library. This YAC was found to be 360kb (Fig. 3.16) and there was no evidence for instability. A previously isolated X-specific probe, L(G0201), which represents an extreme end of a YAC contig spanning the DXS146 locus, detected an *Eco*R1 fragment of appropriate size when hybridized to C0191 (Hatchwell, 1994). This clone therefore represents a link between the DXS255 YACs and the more centromeric YAC cluster around DXS146. In addition, this result establishes the order of markers and the orientation of YACs with respect to the X chromosome.

ii) Two YACs isolated with L(F1001) which map distal to DXS255

On the basis of the above analysis, L(F1001) is on the telomeric side of DXS255. Screening of libraries with an STS developed from L(F1001) (Table 3.8) identified two new YACs, 22AB3 and B0617, which were confirmed as positives by hybridization with the L(F1001) probe. Analysis of undigested YACs on pulsed field gels indicated that both clones are stable (Figure 3.16). The distal markers TFE3 and SYP were found to be absent from these new YACs.

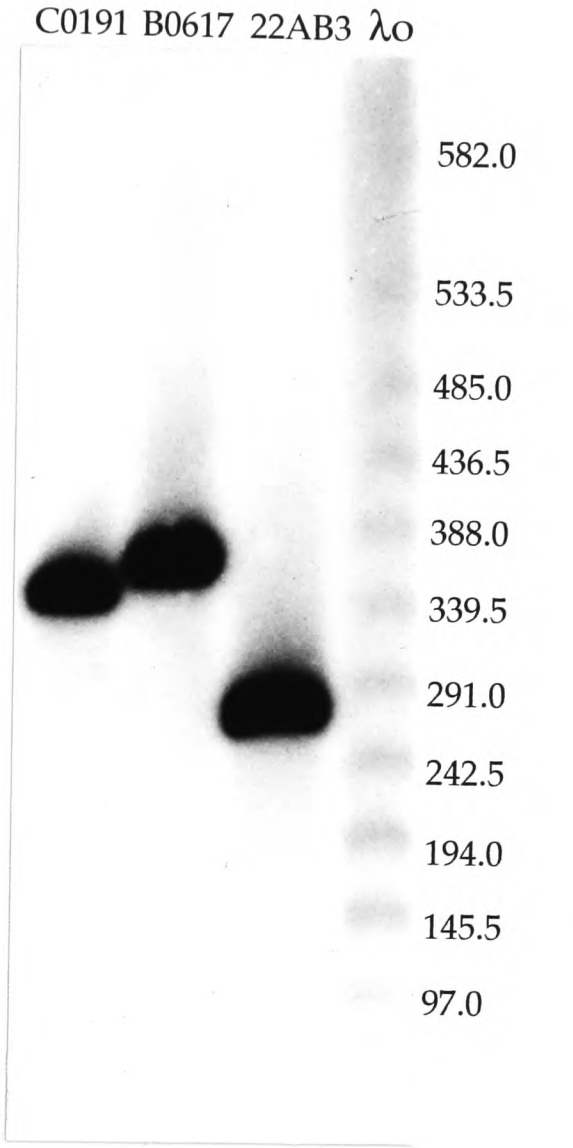


Figure 3.16: Sizing of YAC clones isolated in a bi-directional walk from the DXS255 locus. Plugs from undigested YACs were run for 32 hours on a pulsed field gel with a 50 second switch time. DNA from the gel was transferred to a filter by Southern blotting and probed with radioactively labelled total human DNA. Multiprimed λ/*Hind*III DNA was also included, in order to detect the lambda oligomer markers (λo). Sizes of the latter are given in kilobases. C0191 is on the proximal side of DXS255, whilst B0617 and 22AB3 are on its distal side (see text for details).

iii) Cloning of the left and right ends of YAC B0617

Inverse-PCR was used to isolate the right end of the B0617 insert (Table 3.7). On probing of a genomic/hybrid *Eco*R1 panel, R(B0617) detects two X-specific fragments (Fig. 3.17). These bands are also detected in B0617 and both DXS255 YACs, but not in yeast or any other YAC clones (Fig. 3.17), indicating that B0617 is oriented with its right end towards Xcen. The detection of two *Eco*R1 fragments by this probe may result from hybridization to two homologous regions of the X chromosome. However, it seems more likely that this is a consequence of an internal *Eco*R1 site in R(B0617), since both fragments were present in both DXS255 YACs. This issue could be resolved by digestion of the probe with *Eco*R1, if necessary. R(B0617) was hybridized to partial digests of 6129, and was found to map within a 20kb *Bss*HII-*Sfi*I fragment near the middle of the YAC (Table 3.5, Fig. 3.18). This established that B0617 has 80-100kb of overlap with the DXS255 YACs (Fig. 3.18).

L(B0617) was isolated by plasmid rescue (Table 3.7). A subfragment of this clone gave a repetitive signal on hybridization to a genomic/hybrid panel, but when used to probe YAC *Eco*R1 panels, detected a single band which was present only in the YAC of origin (Fig. 3.19). In order to determine if L(B0617) originated from the X chromosome, it was partially sequenced (see Appendix), and an STS developed (Table 3.8). PCR on a series of somatic cell hybrids (including X chromosome translocation hybrids and a previously characterized panel of irradiation-reduced hybrids) localized L(B0617) to the expected region; Xp11.23-p11.22 (Fig. 3.20). This novel marker therefore represented the extreme distal end of the DXS255–DXS146 contig.

3.3.3 Discovery of a polymorphic dinucleotide repeat region in the L(B0617) clone

Sequencing of L(B0617) revealed the presence of three sets of dinucleotide repeats within a 110bp region (Figure 3.21a and see Appendix). Dinucleotide repeats (also known as microsatellites) are abundant interspersed repetitive elements in eukaryotic genomes, and have been shown to exhibit length polymorphisms (Weber and May, 1989), making them useful markers for linkage analysis (see Section 1.1.2).

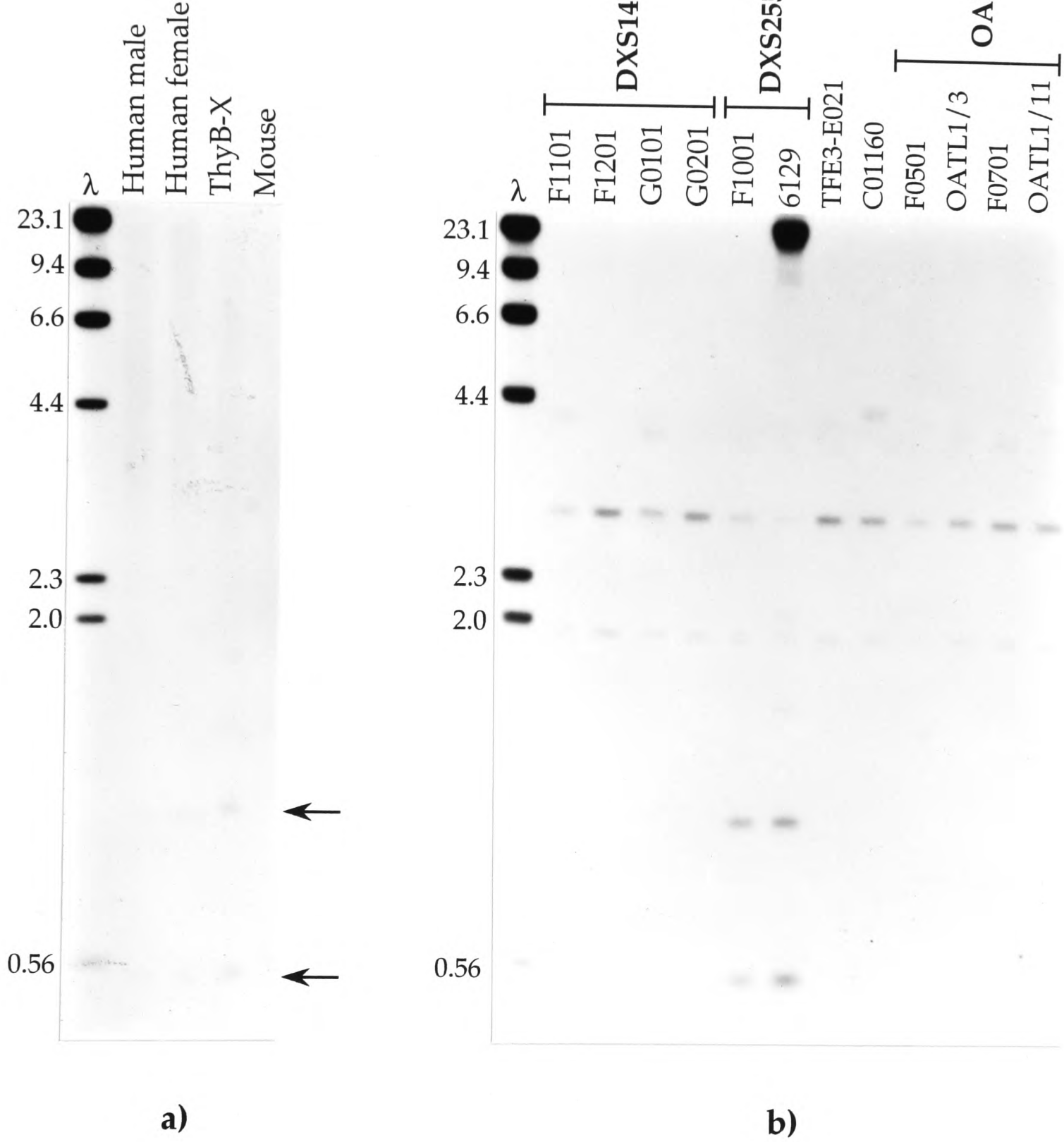


Figure 3.17: Hybridization of the R(B0617) probe to *Eco*R1 digests of genomic, hybrid and YAC DNAs. Sizes of lambda markers (λ) are given in kilobases.

a) R(B0617) detects two bands, of 1.2kb and 500bp, in human male and female genomic DNA, and in ThyB-X, but not in mouse genomic DNA. (These bands, indicated by arrows, are faint on the above photograph, but can be seen more clearly on the original autoradiographs.)

b) On probing *Eco*R1 digests of a panel of YACs, R(B0617) detects the 1.2kb and 500bp bands in both DXS255 YACs (and also in B0617, the YAC of origin, which is not shown here). These fragment are not present in any of the DXS146 YACs or in clones from the OATL1–GATA–TFE3–SYP cluster (see Chapter 4). Hybridization to the high molecular weight material in the 6129 track is due to incomplete digestion of this YAC. Additional fragments are detected in all tracks, because the R(B0617) probe contains pYAC4 right arm vector sequences as a consequence of the inverse-PCR technique which was used to isolate it (see Figure 3.6).

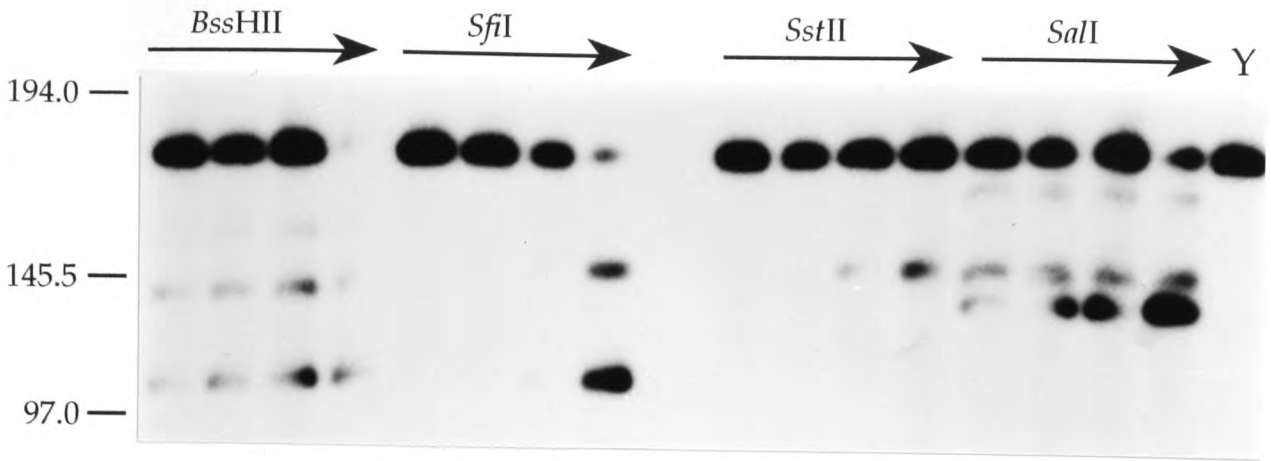


Fig. 3.18a: R(B0617) hybridized to pulsed field partial digests of the 6129 YAC clone. This is the same filter as shown in Fig. 3.8. Direction of arrow represents increasing enzyme concentration. Positions and sizes, in kilobases, of lambda oligomer markers are indicated on the left. The track labelled Y contains undigested YAC.

Enzyme	R(B0617)
*Bss*HII	105, 140, 160
*Eag*I	145, 160
*Sal*I	135, 145, 175
*Sfi*I	105, 145
*Sst*II	145

Table 3.5: Fragment sizes, in kilobases, of bands detected when R(B0617) is used to probe rare-cutter partial digests of the 6129 YAC clone. See Fig. 3.18a.

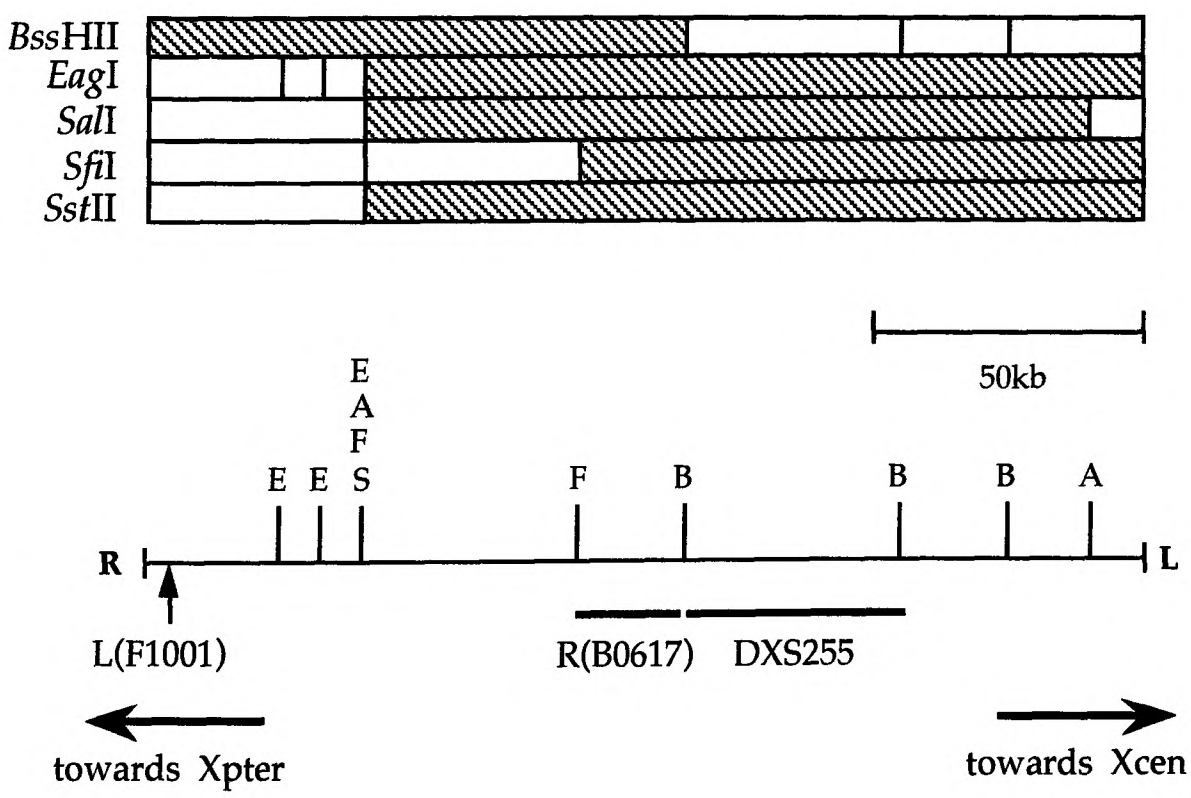


Fig. 3.18b: Position of R(B0617) in rare-cutter restriction map of 6129 YAC clone, derived from fragment sizes in Fig. 3.18a and Table 3.5. Diagonal stripes show fragments on which R(B0617) lies. Sites are given as in Fig. 3.9. Localizations of DXS255 and L(F1001), and orientation of YAC with respect to the X chromosome are indicated.

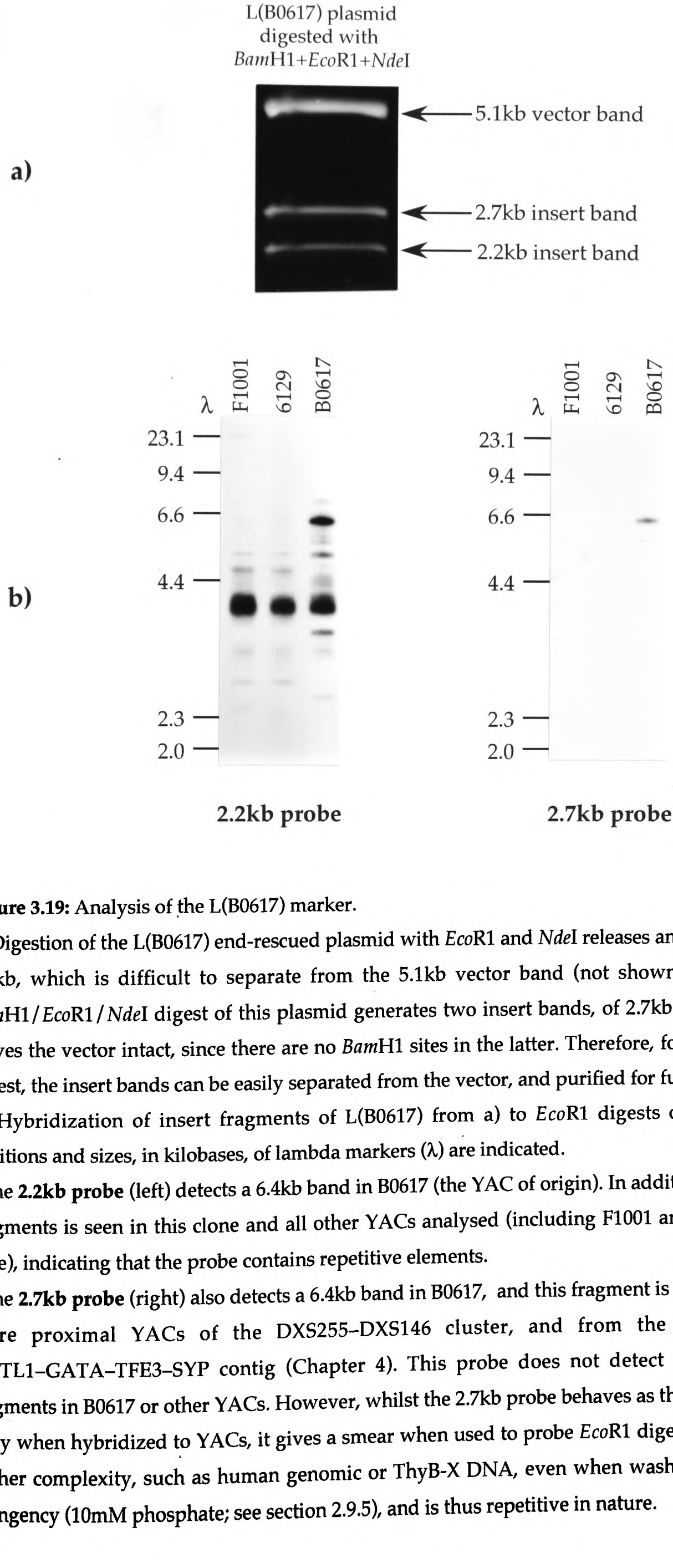


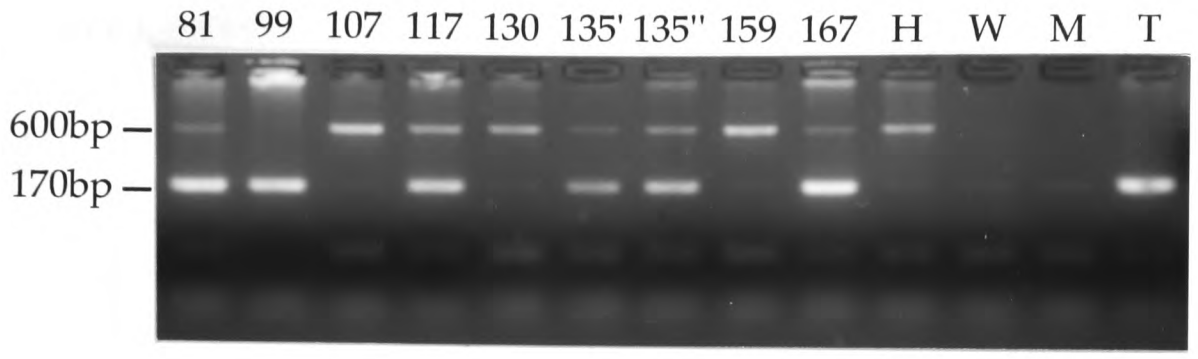
Figure 3.19: Analysis of the L(B0617) marker.

a) Digestion of the L(B0617) end-rescued plasmid with *Eco*R1 and *Nde*I releases an insert band of 4.9kb, which is difficult to separate from the 5.1kb vector band (not shown). However, a *Bam*H1/*Eco*R1/*Nde*I digest of this plasmid generates two insert bands, of 2.7kb and 2.2kb, but leaves the vector intact, since there are no *Bam*H1 sites in the latter. Therefore, following such a digest, the insert bands can be easily separated from the vector, and purified for further analysis.

b) Hybridization of insert fragments of L(B0617) from a) to *Eco*R1 digests of YAC DNAs. Positions and sizes, in kilobases, of lambda markers (λ) are indicated.

- The **2.2kb probe** (left) detects a 6.4kb band in B0617 (the YAC of origin). In addition, a ladder of fragments is seen in this clone and all other YACs analysed (including F1001 and 6129, shown here), indicating that the probe contains repetitive elements.
- The **2.7kb probe** (right) also detects a 6.4kb band in B0617, and this fragment is absent from the more proximal YACs of the DXS255–DXS146 cluster, and from the YACs of the OATL1–GATA–TFE3–SYP contig (Chapter 4). This probe does not detect any additional fragments in B0617 or other YACs. However, whilst the 2.7kb probe behaves as though it is single copy when hybridized to YACs, it gives a smear when used to probe *Eco*R1 digests of DNA of a higher complexity, such as human genomic or ThyB-X DNA, even when washing is to a high stringency (10mM phosphate; see section 2.9.5), and is thus repetitive in nature.

a)



b)

	Hybrid							
Marker	**81**	**99**	**107**	**117**	**130**	**135**	**159**	**167**
117.29	–	+	–	+	–	–	–	+
A1S9T	–	+	–	+	–	–	–	+
DXS426	+	+	–	+	–	+	–	+
L(B0617)	+	+	–	+	–	+	–	+
DXS255	+	–	+	+	–	+	–	+
DXS146	+	–	+	–	–	+	–	–
107.22	+	–	+	–	–	+	–	–

Figure 3.20: Physical mapping of L(B0617) marker using a panel of irradiation-reduced X-chromosome hybrids. These hybrids were formed by fusing human cell lines, whose chromosomes had been fragmented using ionising radiation, with a thymidine kinase (TK) deficient hamster cell line, followed by selection for TK+ human-hamster hybrids (Benham *et al.*, 1989). The hybrids used for this analysis were previously selected and characterized as described by Riley (1993) and Black (1994).

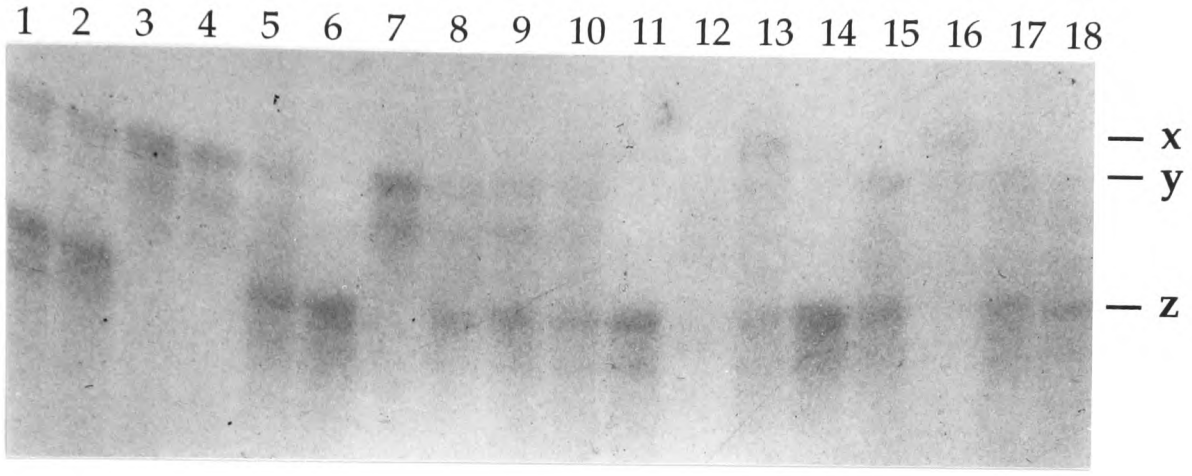
a) Results of L(B0617) PCR assay (Table 3.8) when used to amplify DNA from irradiation-reduced human/hamster hybrids. The 170bp product corresponding to the human locus is amplified from a sub-section of these hybrids (see below). A band of ~600bp is amplified from the hamster background DNA of most hybrids and from hamster genomic DNA (H). The remaining tracks are as follows: W, water control; M, mouse genomic DNA; T, ThyB-X (human X-only/mouse hybrid). Two alternative samples of hybrid 135 (labelled 135' and 135'') were available.

b) Previous characterization of hybrids with proximal Xp markers. +, marker present; –, marker absent. Markers are listed in the order Xpter-Xcen as established by Black (1994). Dark lines represent breakpoints of human material in each hybrid. L(B0617) is expected on the basis of YAC analysis to map between DXS426 (which lies distal to the OATL1 cluster, in Xp11.23) and DXS255, assuming that the B0617 YAC is not chimæric. The PCR results shown in a), summarized here in the shaded boxes, are consistent with this view. 117.29 and 107.22 were previously isolated by *Alu*-PCR of selected hybrids (Black, 1994).

a)

5'-$(TC)_6G(TC)_4$CCCCTTCCTCCT$(TC)_9$GA$(CA)_{13}TA(CA)_2C(CA)_{12}$-3'

b)



c)

	alleles		
samples	**x**	**y**	**z**
1	–	+	+
2	–	+	+
3	–	++	–
4	–	++	–
5	–	+	+
6	–	–	++
7	–	++	–
8	–	+	+
9	–	+	+
10	–	+	+
11	–	–	++
13	+	–	+
14	–	–	++
15	–	+	+
16	++	–	–
17	–	+	+
18	–	+	+

Figure 3.21: A polymorphic region containing dinucleotide repeats within L(B0617).

a) Sequencing of L(B0617) revealed the presence of a TC-rich region containing an imperfect TC repeat sequence with 10 repeats and a perfect TC repeat sequence with 9 repeats. This is followed by an imperfect CA repeat sequence with 27.5 repeat units. Classification of repeats is according to the rules described by Weber (1990). Although this sequence includes three sets of dinucleotide repeats, data from previous studies predicts that it should have a low PIC value (see text).

b) Amplification of genomic DNA from a panel of 18 unrelated females using PCR primers flanking the dinucleotide repeat region of L(B0617) (Table 3.8). End-labelling (2.13.4) and polyacrylamide gel electrophoresis (2.12) was used to visualize alleles. Three alleles (x, y and z) in the range of 170-200bp can be seen. The signal from track 12 is too faint to establish which alleles are present.

c) Summary of results obtained from b) (excluding track 12). Allele frequencies are x = 0.09, y = 0.44, z = 0.47. This marker therefore has a heterozygosity of 0.58 and a PIC value of 0.48 (see text).

Weber (1990) demonstrated that it is possible to predict the genetic informativeness of a given microsatellite marker on the basis of repeat sequence type (classified according to a precise set of rules) and number of repeats. Such analysis was applied to each of the sets of L(B0617) dinucleotide repeats:

i) Sequences with 10 or fewer repeats are usually non-polymorphic. Mutation of dinucleotide repeats is attributed to the process of strand slippage during replication, recombination or repair, and it appears that the rate of strand slippage of dinucleotides increases significantly as the repeat number exceeds 10 (Weber, 1990). The imperfect $(TC)_6G(TC)_4$ repeat and the perfect $(TC)_9$ repeat in the TC-rich region of L(B0617) (Figure 3.21a) are therefore predicted to be non-polymorphic.

ii) Imperfect repeats are defined as two or more runs of uninterrupted repeats (of the same type) separated by no more than three consecutive non-repeat bases (Weber, 1990). $(CA)_{13}TA(CA)_2C(CA)_{12}$ is therefore classified as an imperfect repeat with 27.5 repeat units. However, it has been shown that imperfections tend to reduce the informativeness of the marker relative to that expected on the basis of the total repeat number (Weber, 1990). This may be a consequence of the fact that strand slippage of such sequences would give structures with non-complementary bases. The most reliable prediction of the PIC (polymorphism information content) values (defined below) for such sequences is to take the repeat length as that of the longest run of uninterrupted nucleotides. The informativeness of the L(B0617) CA repeat is thus expected to be comparable to that of a $(CA)_{13}$ perfect repeats, which is predicted to display a PIC value of around 0.35 (Weber, 1990).

The PCR primers of the L(B0617) STS described in Section 3.3.2 and Table 3.8 were originally designed to flank the dinucleotide repeat region, and were therefore used to investigate the polymorphism at this locus (Figure 3.21b-c). Data from 34 X chromosomes in 17 unrelated females indicated that there are at least three alleles for this locus, with sizes in the range of 170-200bp. Allele frequencies derived from this analysis are 0.09, 0.44 and 0.47, and on the basis of this, the marker has a heterozygosity of 0.58 and a PIC value of 0.48.

The PIC value represents the probability that a given offspring of a random mating between a carrier of a rare dominant gene and a non-carrier is informative for linkage between the gene locus and the marker in question, and can be calculated from the equation:

$$1 - (\sum_{i=1}^{n} p_i^2) - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2p_i^2 p_j^2$$

where p_i is the frequency of the ith allele and n is the total number of alleles at the locus. Although the PIC value of L(B0617) is slightly higher than that predicted on the basis of its sequence, this marker is still only moderately informative. Additional studies have demonstrated that L(B0617) polymorphism is inherited in a Mendelian fashion (data not shown).

3.3.4 Extension of DXS146–DXS255 contig towards the telomere

i) Isolation of four YACs using L(B0617)

Four new YAC clones were isolated using the L(B0617) PCR assay; E0250, 16CD8, 27CB12 and 36HB8 (Table 3.6). Three of these YACs were found to be clonally unstable, usually undergoing rearrangements involving ~20kb regions of DNA. For example cultures grown up from three single colonies of a streak of E0250 (the largest of these YACs) yielded clones of 700kb, 680kb or 660kb, all containing the L(B0617) marker, whilst alternative colony-pure preps of 27CB12 contained YACs of either 270kb or 250kb (Figure 3.22). None of these YACs contained SYP, TFE3 or DXS255.

ii) Generation and ordering of new markers from distal YACs

R(27CB12) and R(36HB8) were isolated (Table 3.7) and ordered with respect to other markers according to whether or not they were present in other YACs (Figs. 3.23 and 3.24). L(E0250) was shown to be X-specific and proximal to L(B0617), indicating that YAC E0250 was oriented with its left end towards Xcen (Fig. 3.25 and Table 3.7).

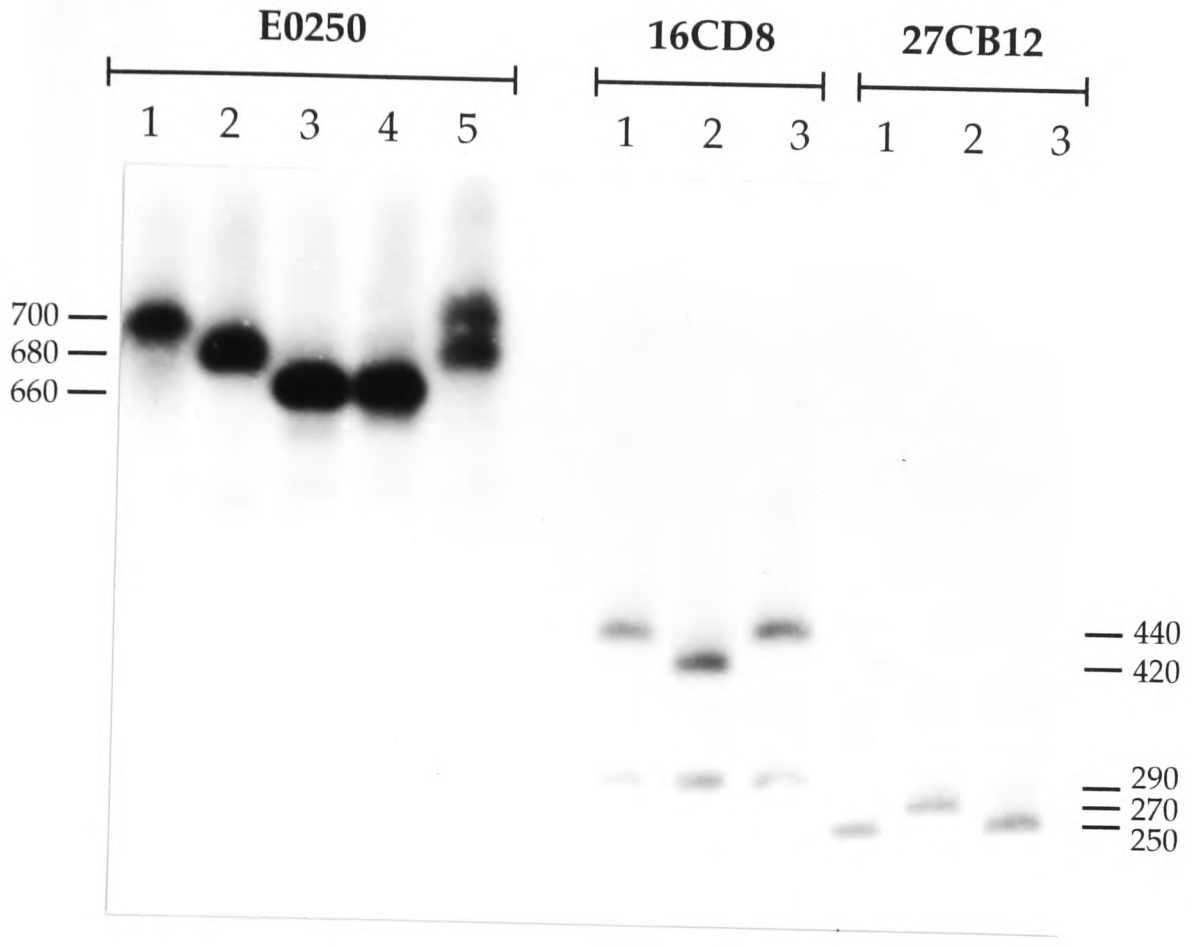
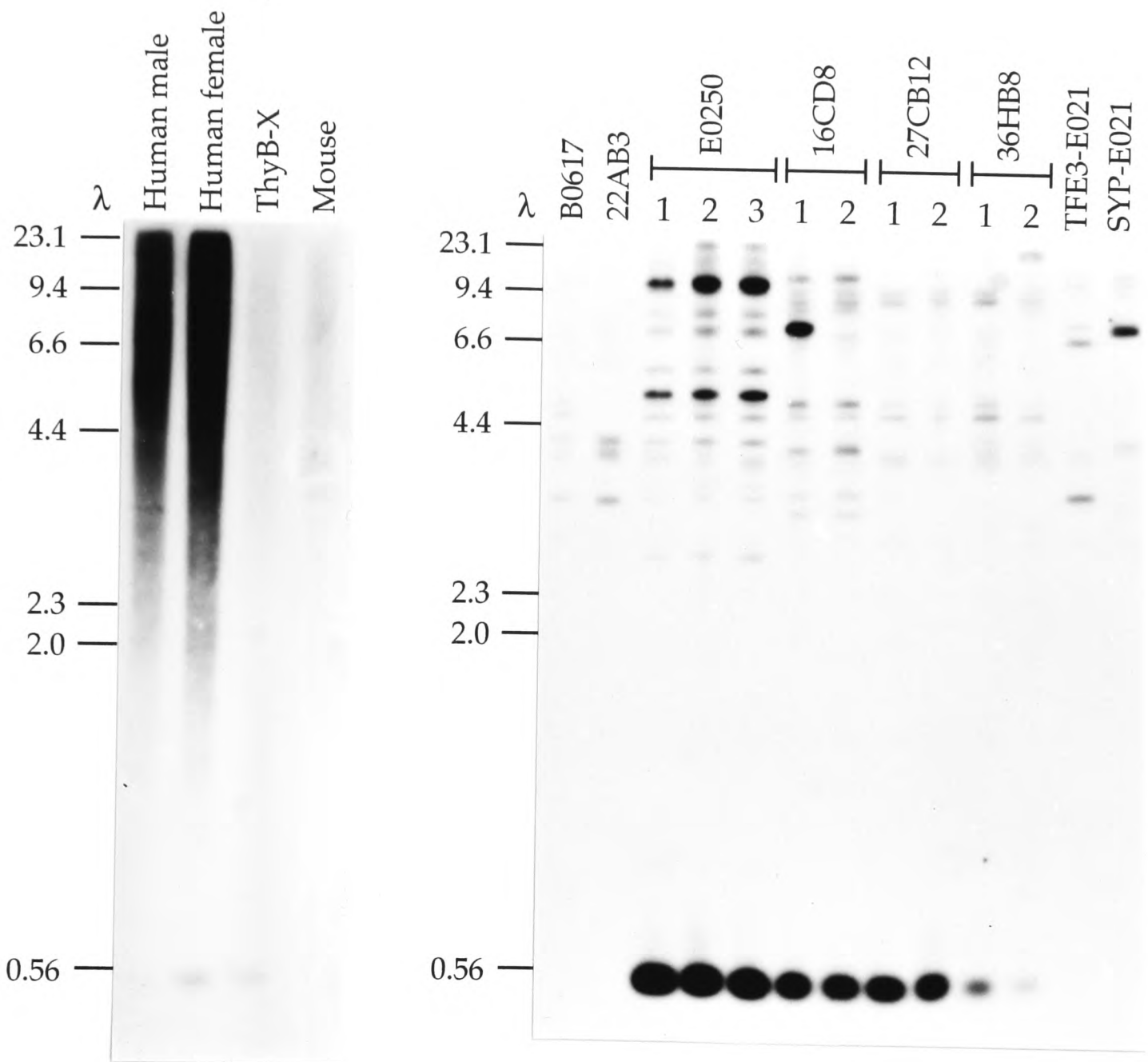


Figure 3.22: Size instability of YAC clones isolated with the L(B0617) marker. Several independent undigested colony-pure preps of each YAC were run on a pulsed field gel with a 60 second switch time and a 32 hour run time. DNA from the gel was transferred to a filter by Southern blotting and probed with radioactively labelled total human DNA. The sizes of the YAC bands thus detected are indicated in kilobases. Whilst preps 1-4 of the E0250 clone each contain a single YAC species, the clone in prep 5 appears to have undergone rearrangement during growth of the culture, resulting in two forms being present in one prep. All three preps of 16CD8 contain two YAC species; in addition, the size of the larger species varies between different preps. It is interesting to note that E0250, 16CD8 and 27CB12 all undergo rearrangements involving gain or loss of ~20kb. Given that these YACs map to the same region of Xp11.22, this supports the view that the observed size variation is a consequence of region-specific instability (see text).



a) b)

Figure 3.23: Hybridization of the R(27CB12) probe to *Eco*R1 digests of genomic, hybrid and YAC DNAs. Positions and sizes, in kilobases, of lambda markers (λ) are indicated.
a) R(27CB12) detects a ~500bp band when used to probe human genomic DNA. This fragment is also present in ThyB-X, but not in mouse genomic DNA. In addition, it hybridizes to a smear of fragments of >~2.0kb in size in the human genomic digests, indicating that it contains repetitive sequences.
b) On probing *Eco*R1 digests of a panel of YACs, R(27CB12) detects the cognate 500bp fragment in 36HB8 and all rearranged forms of clones E0250, 16CD8, 27CB12 (see Figure 3.22; numbers refer to different isolates of the same YAC). This band is absent from the more proximal YACs, such as B0617 and 22AB3. Given that 27CB12 was isolated using L(B0617), this result establishes that this YAC is oriented with its right end towards the telomere, and indicates that the order of markers must be Xpter–R(27CB12)–L(B0617)–L(F1001)–Xcen. The 500bp fragment is also absent from YACs of the OATL1–GATA–TFE3–SYP cluster. The weaker hybridization to fragments of between 23 and 3kb in all tracks is due to the repetitive nature of the probe.

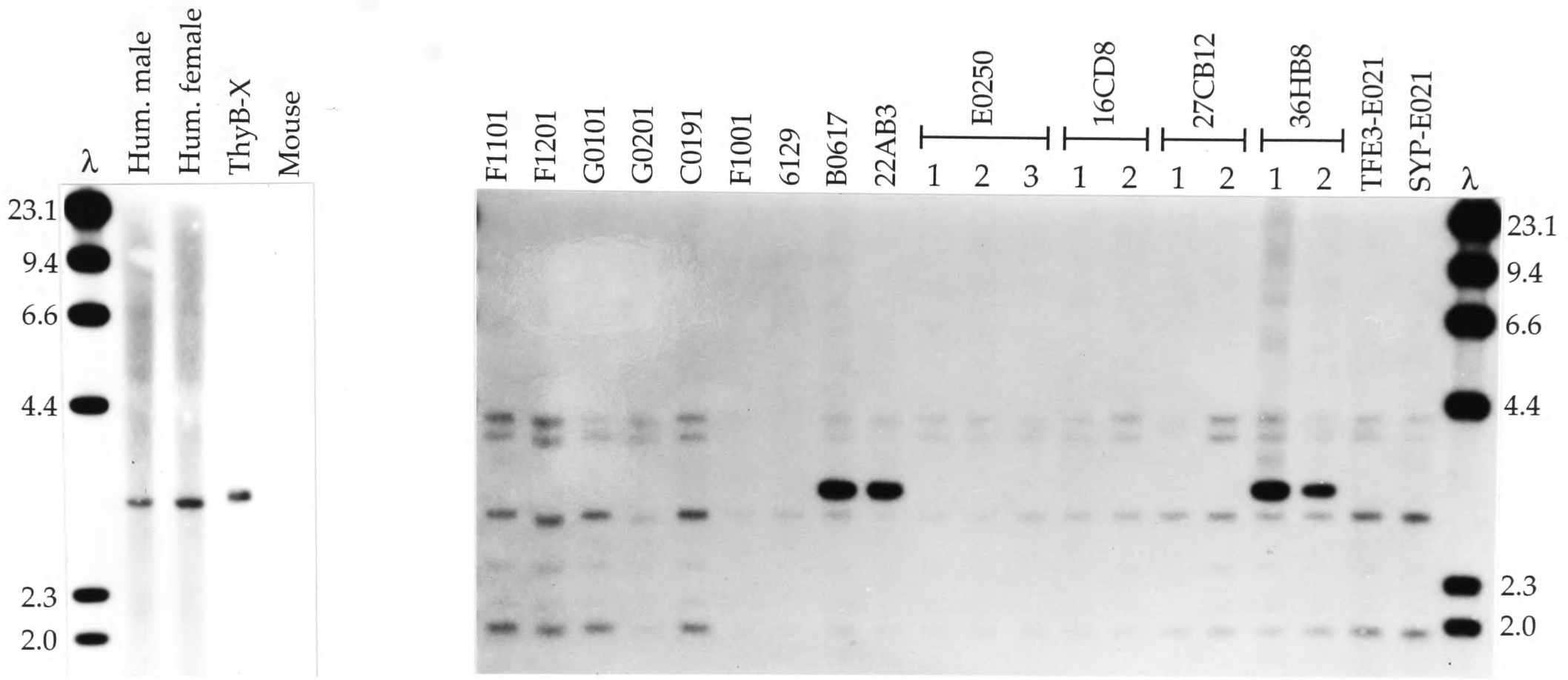


Figure 3.24: Hybridization of R(36HB8) to *Eco*R1 digests of genomic, hybrid and YAC DNAs. Sizes of lambda markers (λ) are given in kilobases.
a) R(36HB8) detects a 3.2kb band when used to probe human genomic DNA. This fragment is also present in ThyB-X, but not in mouse genomic DNA. The fragment appears slightly larger in ThyB-X, but this is due to overloading of the DNA in this track.
b) On probing *Eco*R1 digests of a panel of YACs, R(36HB8) detects the cognate 3.2kb fragment in two colony-pure preps of 36HB8 (the YAC of origin) and in the more proximal YACs, B0617 and 22AB3. Given that 36HB8 was isolated using L(B0617), this result establishes that this YAC is oriented with its right end towards Xcen, giving a marker order of Xpter–R(27CB12)–L(B0617)–R(36HB8)–L(F1001)-Xcen. The 3.2kb fragment is absent from all other Xp11.23-p11.22 YACs described in this thesis. The weaker hybridization to a series of fragments in all digests is due to the presence of pYAC4 right arm vector sequences in the R(36HB8) probe, as a consequence of the inverse-PCR technique which was used to isolate it (see Figure 3.6).

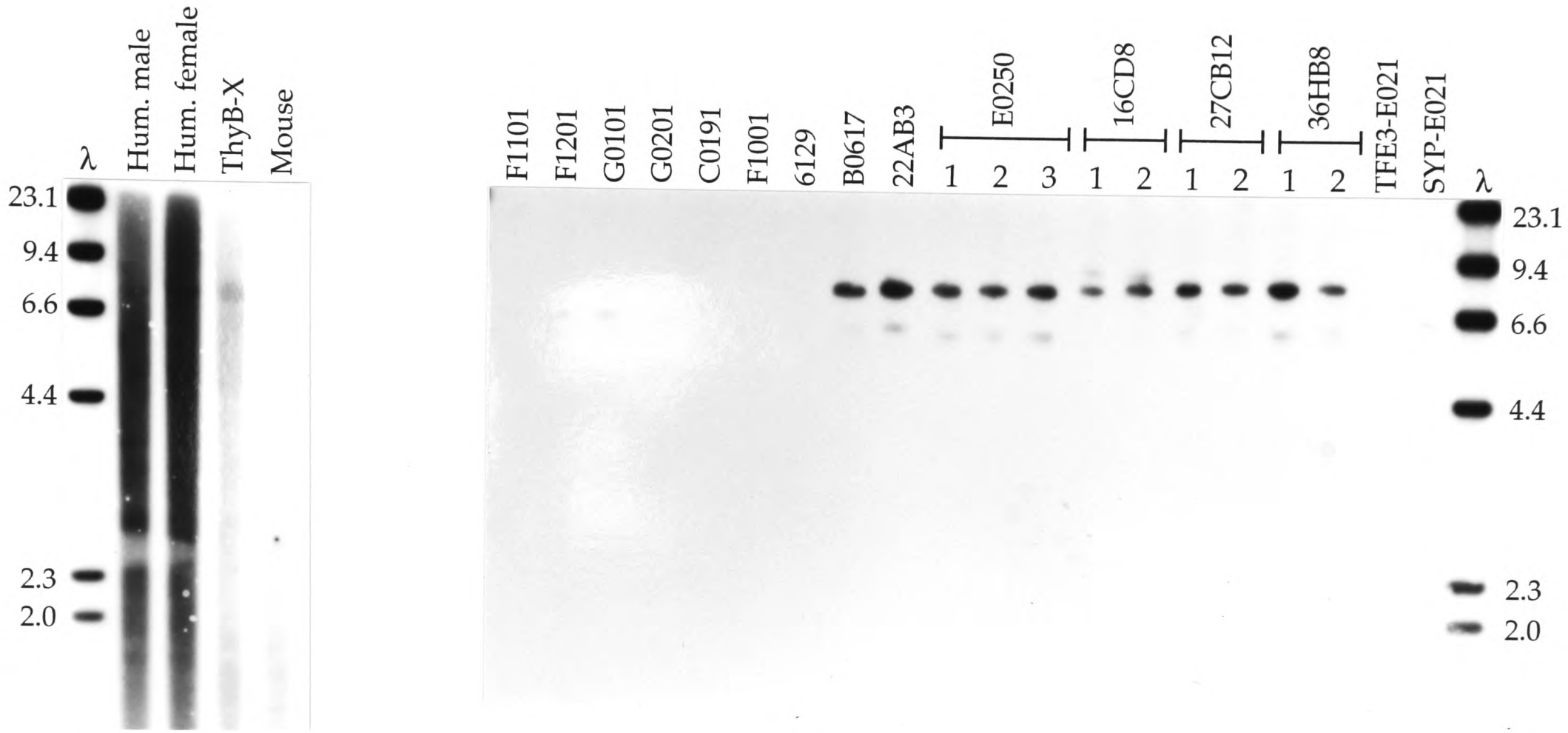


Figure 3.25: Hybridization of L(E0250) to *Eco*R1 digests of genomic, hybrid and YAC DNAs. Sizes of lambda markers (λ) are given in kilobases.
a) L(E0250) detects a smear when used to probe human genomic DNA, suggesting that it contains repetitive elements. An 8.0kb fragment is detected above the background hybridization in ThyB-X. This band is absent from mouse genomic DNA.
b) On probing *Eco*R1 digests of a panel of YACs, L(E0250) detects a predominant band of 8.0kb in all rearranging forms of E0250, 16CD8, 27CB12 and 36HB8, as well as in the more proximal clones, B0617 and 22AB3. Given that E0250 was isolated using L(B0617), this result establishes that this YAC is oriented with its left end towards the centromere of the X-chromosome and gives a marker order of Xpter–R(27CB12)–L(B0617)–L(E0250)–R(36HB8)–L(F1001)-Xcen (see Figure 3.27). The 8.0kb fragment is absent from all other Xp11.23-p11.22 YACs described in this thesis. Weak hybridization to other fragments in some YACs is due to the repetitive nature of the probe.

Attempts to isolate R(E0250) using an inverse-PCR strategy were unsuccessful, due to a lack of appropriate enzyme sites at the right end of the YAC insert (see Section 3.2.4). Instead, *Alu*-PCR was used to generate a novel X-specific marker from this YAC, A(E0250), which is absent from all other clones in the contig, and must therefore be the most distal STS in the DXS255 cluster (Fig. 3.26; Tables 3.7 and 3.8). It is also absent from the OATL1–GATA1–TFE3–SYP cluster (see Chapter 4).

3.3.5 Rare-cutter restriction mapping of YACs from the contig

A subset of YACs from the contig were further analysed using indirect end-label mapping. Partial digests with rare-cutting enzymes *Bss*HII, *Eag*I, *Mlu*I, *Not*I, *Sal*I, *Sfi*I and *Sst*II (or a selection of these) were run on pulsed field gels with an appropriate switch time, and then probed sequentially with left and right vector arm. Maps were thus made for the following clones:

i) C0191 (Figs. 3.28-3.30; Table 3.9)

Following hybridization with vector arms (Fig. 3.28), partial digests of C0191 were probed with the novel markers L(G0201) and L(6129), which had previously been shown to map within the YAC (Fig. 3.29). L(G0201) was thereby localized to a 75kb *Eag*I-*Sal*I fragment in the right-hand half of the YAC, whilst L(6129) was found to lie in a 20kb *Sal*I-*Bss*HII fragment in the left-hand half (Fig. 3.30). The orientation of C0191 relative to the X chromosome could therefore be determined, since L(G0201) was already known to map on the proximal side of L(6129).

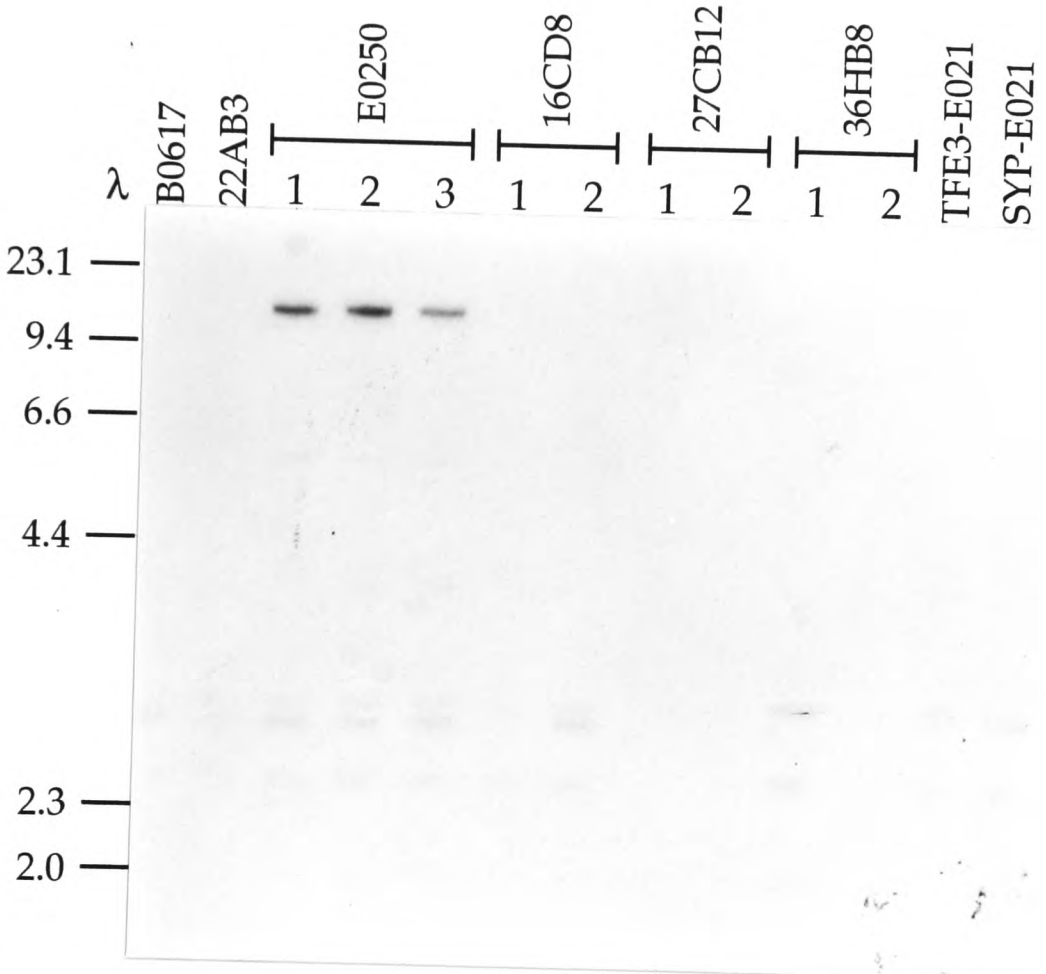


Figure 3.26: Hybridization of the A(E0250) probe to *Eco*R1 digests of YAC DNAs. Positions and sizes, in kilobases, of lambda markers (λ) are indicated. A ~15kb band is detected in three rearranged forms of E0250 (see Figure 3.22). This fragment is not present in any other YACs described in this thesis. A(E0250) hybridizes to a smear of fragments when used to probe *Eco*R1 digests of human genomic or ThyB-X DNA (not shown), indicating that it contains repetitive sequences. PCR analysis using an STS developed from the marker (Table 3.8) has shown that it is X-specific. These results establish that A(E0250) is the most distal X-specific marker from the DXS255–DXS146 contig (see text).

YAC ID	Alternative name	Library	Isolated with	Size/kb	Left end	Right end	Other markers present
E0250	FDTM/1	ICRF	L(B0617)	700/680/660[a]	X		A(E0250), R(27CB12)
16CD8	FDTM/2	ICI	L(B0617)	290/270[a] + 400[b]			R(27CB12), L(E0250)
27CB12	FDTM/3	ICI	L(B0617)	270/250[a]		X	L(E0250)
36HB8	FDTM/4	ICI	L(B0617)	345		X	R(27CB12), L(E0250)
B0617	DTM/1	ICRF	L(F1001)	380	X	X	L(E0250), R(36HB8)
22AB3	DTM/2	ICI	L(F1001)	280			L(E0250), R(36HB8)
F1001	DXS255/1	ICRF	M27β probe	1100	X	aut	R(B0617), CLCN5[c], L(6129)
6129	DXS255/2	St. Louis	MBP assay	185	X		L(F1001), R(B0617), CLCN5[c]
C0191	PTM/1	ICRF	L(6129)	360			CLCN5[c], L(G0201)[d]
G0201	pTAK/4	ICRF	pTAK8	900	X		R(F1101)[d] , L(F081)[d]

Table 3.6: Details of YACs from DXS255-DXS146 cluster in Xp11.22. YACs are listed in the order Xpter-Xcen. L(YAC ID), left end of YAC; R(YAC ID), right end of YAC; A(YAC ID), *Alu*-PCR probe from YAC; X, end clone X-specific; aut, end clone autosomal, indicating chimæric YAC. [a]YAC rearranging. [b]Two forms of same YAC present in all colony pure preps. [c]Localization of CLCN5 is described in Chapter 5. [d]Markers described elsewhere (Hatchwell, 1994). Details of YAC libraries are given in Section 3.2.2.

Marker	Size	Cognate *Eco*R1 band(s)	STS	E0250	16CD8	27CB12	36HB8	B0617	22AB3	6129	F1001	C0191
A(E0250)	1.1kb	15.0kb	+	+	–	–	–	–	–	–	–	–
R(27CB12)	700bp	500bp		+	+	+	+	–	–	–	–	–
L(B0617)	4.9kb	6.4kb[b]	+	+	+	+	+	+	–	–	–	–
L(E0250)	880bp/700bp[a]	8.0kb		+	+	+	+	+	+	–	–	–
R(36HB8)	300bp	3.2kb		–	–	–	+	+	+	–	–	–
L(F1001)	4.0kb	4.0kb[b]	+	–	–	–	–	+	+	+	+	–
R(B0617)	900bp	1.2kb & 500bp		–	–	–	–	+	–	+	+	–
L(6129)	180bp	4.4kb	+	–	–	–	–	–	–	+	+	+

Table 3.7: Details of novel markers from DXS255-DXS146 cluster in Xp11.22. Markers are listed in the order Xpter-Xcen. L(YAC ID), left end of YAC isolated by plasmid rescue; R(YAC ID), right end of YAC isolated using inverse-PCR; A(YAC ID), *Alu*-PCR probe from YAC; +, present in YAC; -, absent from YAC. Those markers which have been converted into STSs are indicated. [a]This left end-clone gives two fragments on double digestion with *Eco*R1/*Nde*I. [b]Probe gives repetitive signal on genomic digests.

Marker name	DXS no.	Primer sequences (5' to 3')	T_a (°C)	Product size (bp)
A(E0250)		GCAGACTCAAAAGGCCACAT TGCATTCACAAAGTTGTGCA	55	215
L(B0617)[a]	6666	CTTCTGGACCTGCAAAGAGG CCCTGAGCAATAGAAGTTAAACC	55	170-200
L(F1001)	6850	TTGTCTCTCTTCACCTTTTGC GGTTGTTTTCGTTTACCCTC	52	106
MBP	255	GCTGGTGCCACGTTATTGA GGGCCACTGGCATTGTAAA	53	472
L(6129)	6851	GACTCTTGAGGGAGTCACAG ACTCATTGTACCTCCCAGC	50	139

Table 3.8: Details of PCR assays developed from markers in DXS255-DXS146 YAC contig. Order of markers is Xpter-Xcen. T_a, annealing temperature. Additional details of PCR amplification conditions are given in Materials and Methods. [a] PCR assay from L(B0617) spans a polymorphic CA repeat and therefore gives products of varying size.

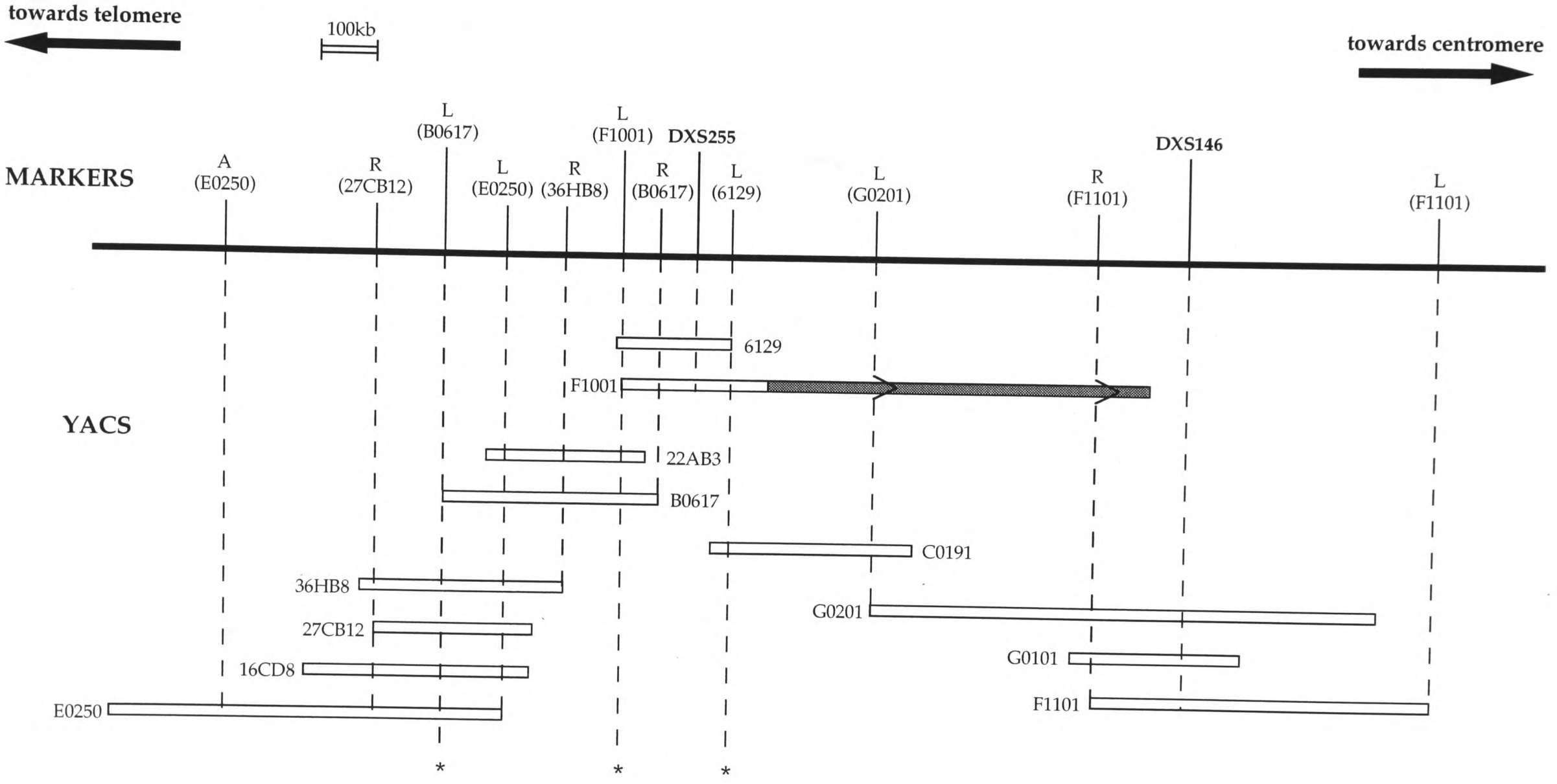


Figure 3.27: YACs and markers in the DXS255-DXS146 cluster.L(YAC ID), left end clone of YAC; R(YAC ID), right end clone of YAC; A(YAC ID), *Alu*-PCR product isolated from YAC. Autosomal regions of chimæric YACs are indicated by shading. Clones are drawn to scale, but the extent of YAC overlap has not been established for all clones in the contig, hence physical distances between markers is not always accurately represented. Asterisks indicate markers used for bi-directional YAC walk. Further details of YACs and markers are given in Tables 3.6-3.8.

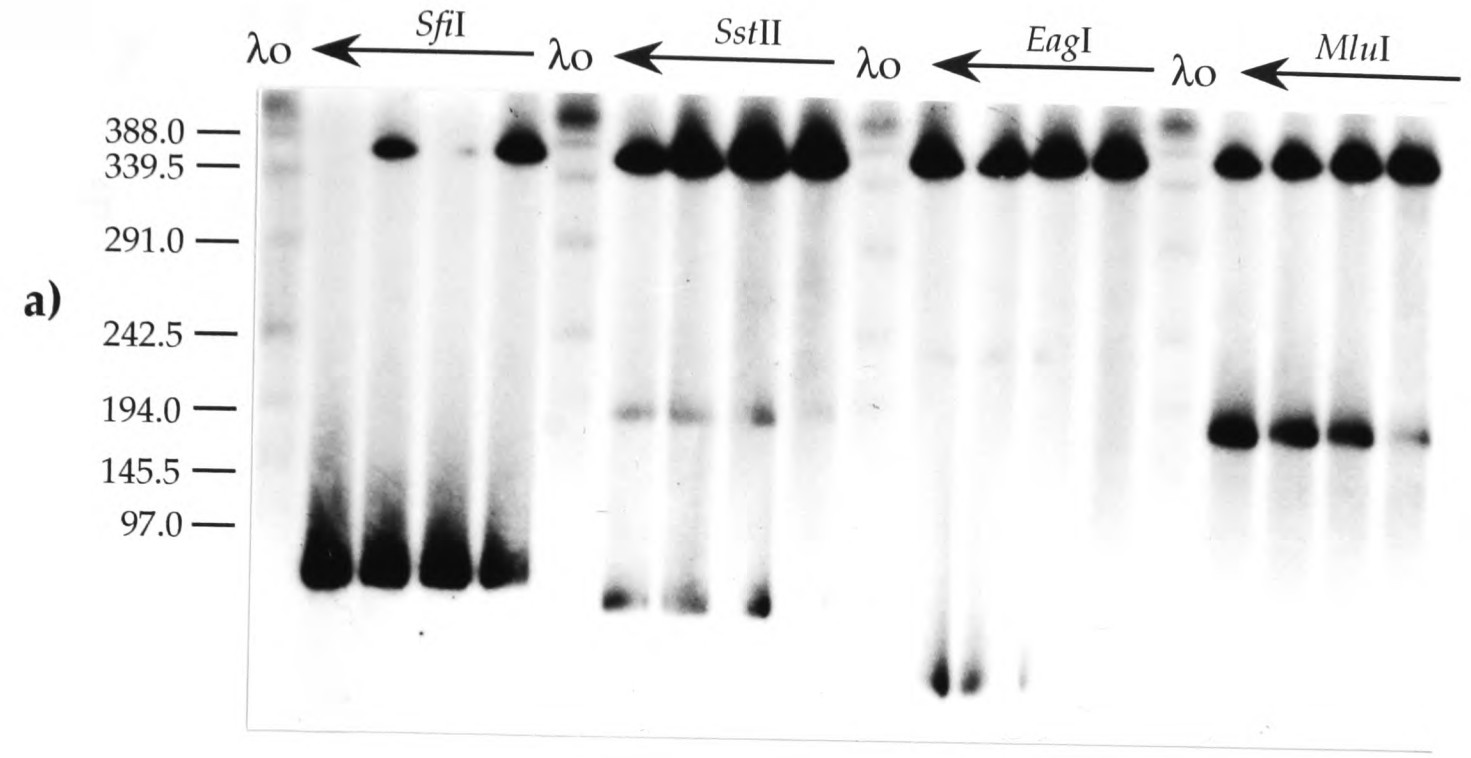


Figure 3.28: Examples of partial digests of the C0191 YAC using different rare cutters, probed with **a)** left vector arm and **b)** right vector arm. Direction of arrow represents increasing enzyme concentration (from 0.3-15U) with a 1 hour digestion time. Sizes of lambda oligomer markers (λo) are given in kilobases. Digests were run on a standard pulsed field gel (section 2.14.5) with a 25 second switch time and a 31 hour run time. Fragment sizes and the rare-cutter restriction map which was derived from them are given in Table 3.9 and Figure 3.30.

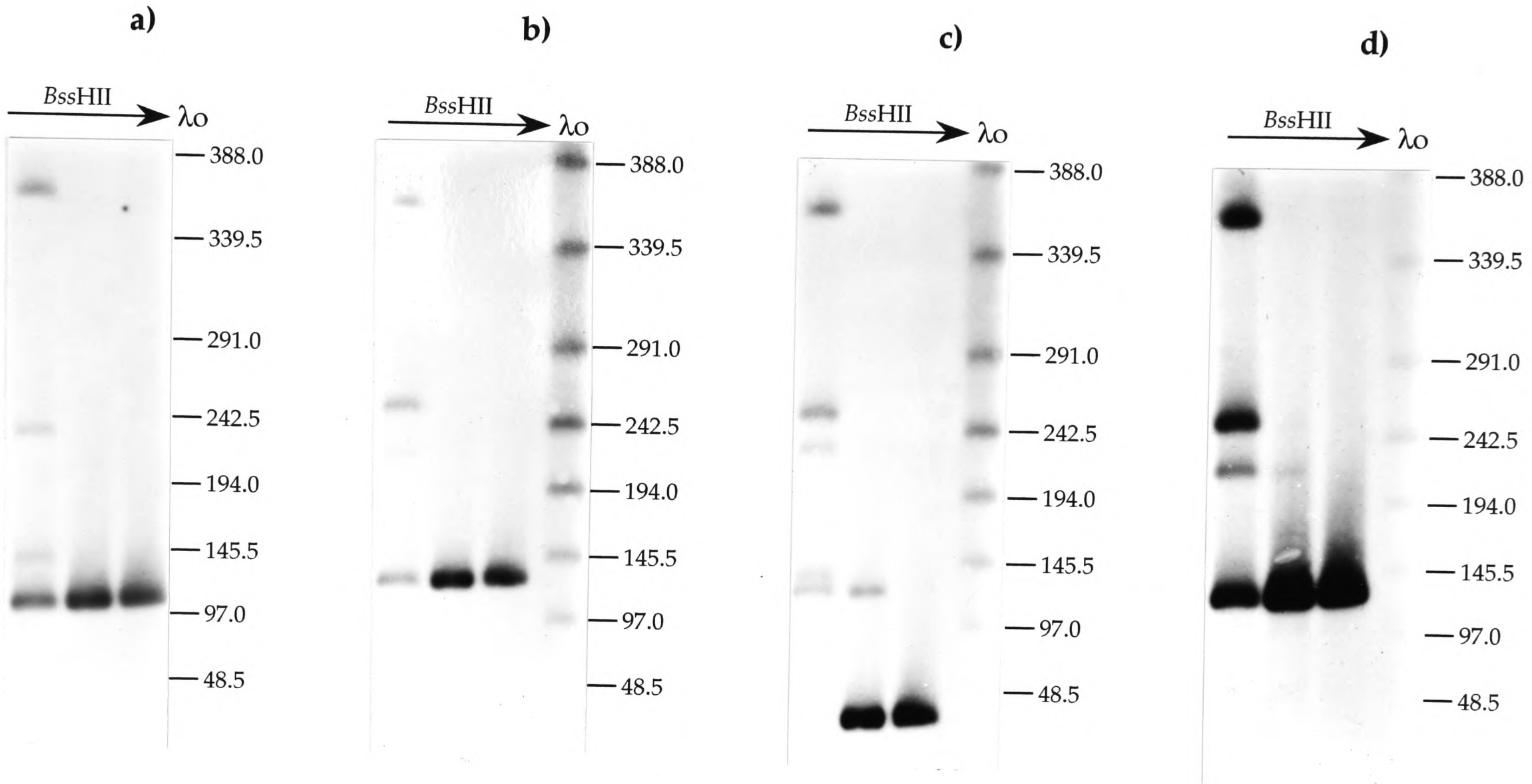


Figure 3.29: Partial digests of the C0191 YAC using *Bss*HII, probed with **a)** left vector arm, **b)** right vector arm, **c)** L(6129) and **d)** L(G0201). Direction of arrow represents increasing enzyme concentration (from 0.03-5U) with a 1 hour digestion time. Sizes of lambda oligomer markers (λo) are given in kilobases. Digests were run on a standard pulsed field gel (section 2.14.5) with a 27 second switch time and a 34 hour run time. Fragment sizes and the rare-cutter restriction map which was derived from them are given in Table 3.9 and Figure 3.30.

Enzyme	Left vector arm	Right vector arm
*Bss*HII	105, 140, 230	130, 220, 255
*Eag*I	<15, 235	125
*Mlu*I	170	<15, 190
*Sal*I	120, 185, 310, 340	20, 50, 175, 240
*Sfi*I	45	145, 170, 305, 315
*Sst*II	30, 185	175, 330

Enzyme	L(6129)	L(G0201)
*Bss*HII	35, 120, 140, 230, 255	130, 220, 255
*Eag*I		125
*Mlu*I		175, 190
*Sal*I	65, 185, 195, 220, 240, 310, 340	125, 155, 175, 195, 220, 240, 310, 340
*Sfi*I		145, 170, 305, 315
*Sst*II		175, 330

Table 3.9: Fragment sizes, in kilobases, of bands detected on rare-cutter partial digests of C0191 YAC clone, when probed with vector (left and right arms), and internal markers (L(6129) and L(G0201)). The 360kb fragments corresponding to undigested YAC are not listed. L(6129) was only hybridized to *Bss*HII and *Sal*I digests, but this was sufficient for localization. See Figs. 3.28 and 3.29.

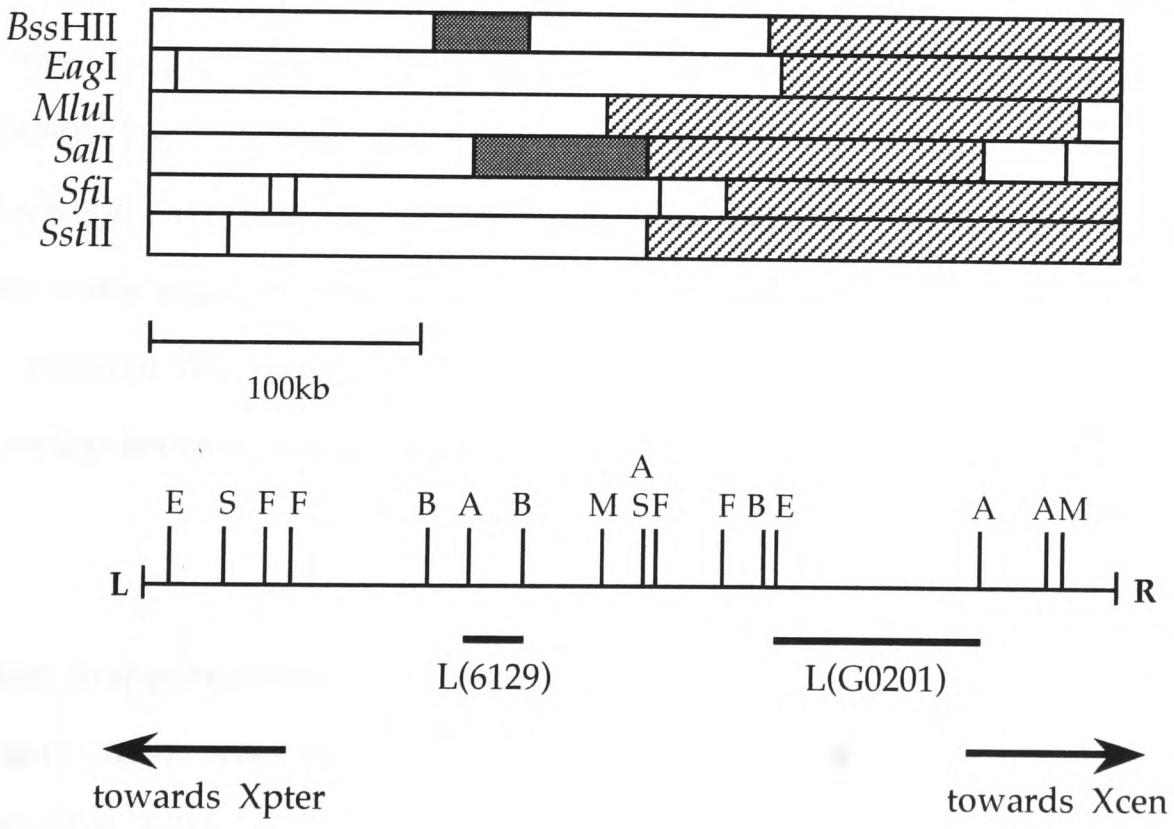


Fig. 3.30: Rare-cutter restriction map of C0191 YAC clone, derived from fragment sizes in Table 3.9, with positions of internal markers indicated below. Shading shows fragments on which L(6129) lies; diagonal stripes show fragments on which L(G0201) lies. **L**, left arm; **R**, right arm; B, *Bss*HII site; E, *Eag*I; M, *Mlu*I; A, *Sal*I; F, *Sfi*I; S, *Sst*II. There are no sites for *Not*I in this YAC. Orientation with respect to the X chromosome is indicated.

ii) B0617 (Figs. 3.31-3.32; Table 3.10)

Partial digests of B0617 were probed with vector arms and then the internal marker L(E0250) (Fig. 3.31). The latter was thus localized to a 105kb *Sfi*I-*Sst*II fragment in the left-hand portion of the YAC. The orientation of this clone relative to the X chromosome had already been determined (see Section 3.3.2iii). Comparison between rare-cutter sites in B0617, F1001 and 6129, and the previous localization of R(B0617) within the DXS255 YAC map, enabled the position of L(F1001) to be deduced (Fig. 3.32).

iii) E0250 (Figs. 3.33-3.34; Table 3.11)

Three size variants of the E0250 clone had been isolated (Section 3.3.4i). However, the 700kb and 680kb species proved to be very unstable, and following the initial preparations of small numbers of plugs for preliminary analysis, only the 660kb clone could be obtained in a stable form. For indirect end-label mapping it is essential that only one of the size variants is present in the plugs used, otherwise ambiguities result that complicate the construction of a coherent map. Partial digests of the 660kb E0250 clone were therefore probed sequentially with left arm, right arm, and the internal marker A(E0250) (Fig. 3.33) and a map derived (Table 3.11;Fig. 3.34). A(E0250) was found to lie in a 70kb *Sfi*I-*Sst*II fragment in the middle of the YAC. In addition, comparison with the rare-cutter sites of B0617 and the previous positioning of L(E0250) within the B0617 map, enabled the position of L(B0617) to be deduced, indicating that there is ~135kb of overlap between E0250 and B0617.

The rare-cutter mapping data and marker positions from the five YACs, E0250, B0617, F1001, 6129 and C0191, were combined to produce a 1.23Mb restriction map of the entire DXS255 contig (Fig. 3.35).

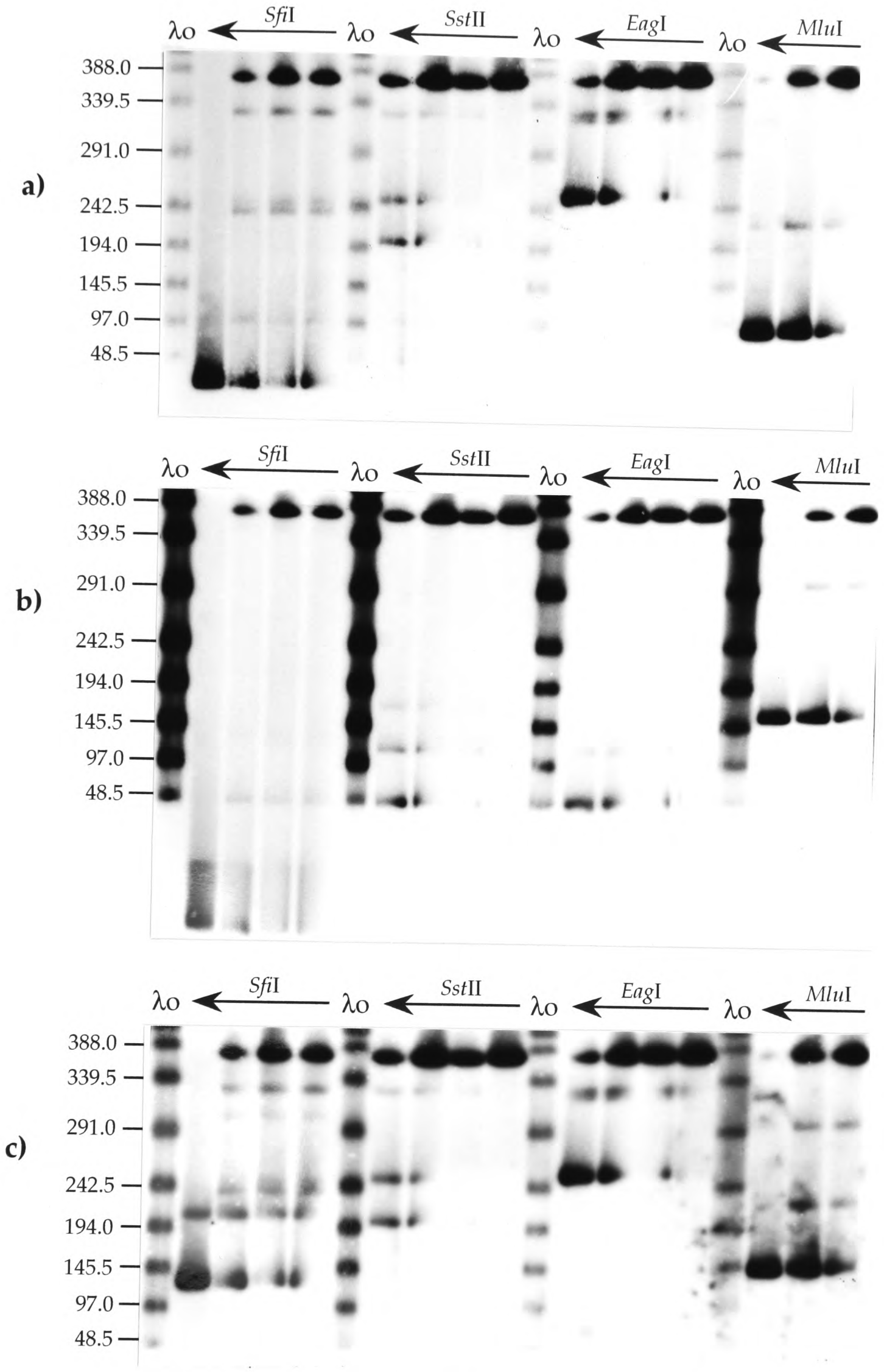


Figure 3.31: Examples of partial digests of the B0617 YAC using different rare cutters, probed with **a)** left vector arm, **b)** right vector arm and **c)** L(E0250). Direction of arrow represents increasing enzyme concentration (from 0.3-15U) with a 1.5 hour digestion time. Sizes of lambda oligomer markers (λo) are given in kilobases. Digests were run on a standard pulsed field gel (section 2.14.5) with a 27 second switch time and a 33 hour run time. Fragment sizes and the rare-cutter restriction map which was derived from them are given in Table 3.10 and Figure 3.32.

Enzyme	Left vector arm	Right vector arm
*Bss*HII	20, 300	80, 360
*Eag*I	260, 325, 330	50, 120
*Mlu*I	85, 220	160, 295
*Sal*I	235, 265, 330	50, 115, 140
*Sfi*I	30, 100, 235, 245, 330	<15, 50, 135, 350
*Sst*II	205, 260, 330	50, 120, 175

Enzyme	L(E0250)
*Bss*HII	280, 300, 360
*Eag*I	260, 325, 330
*Mlu*I	140, 220, 295
*Sal*I	235, 265, 330
*Sfi*I	135, 205, 235, 245, 300, 330, 350[a]
*Sst*II	205, 260, 330

Table 3.10: Fragment sizes, in kilobases, of bands detected on rare-cutter partial digests of B0617 YAC clone, when probed with vector (left and right arms), and the internal marker L(E0250). The 380kb fragments corresponding to undigested YAC are not listed.
[a] additional bands were present but were very faint. See Fig. 3.31.

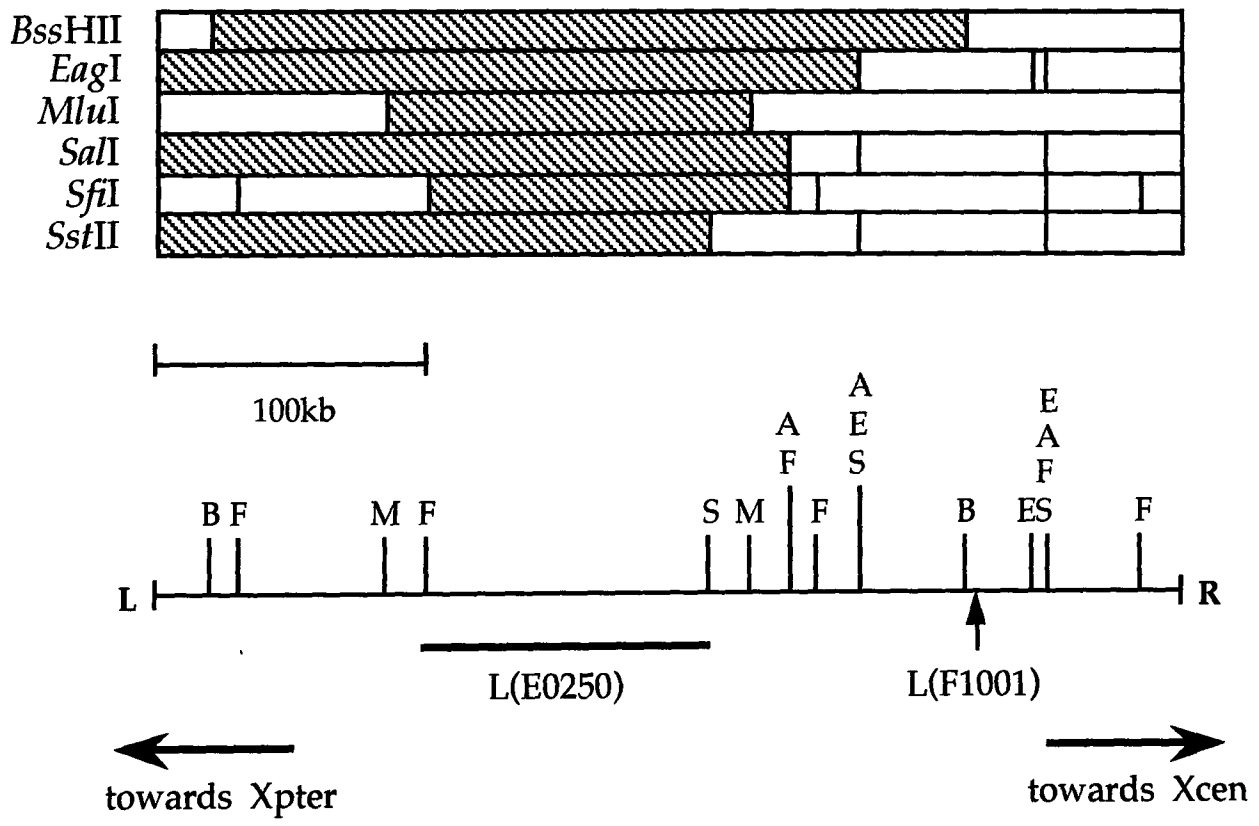


Fig. 3.32: Rare-cutter restriction map of B0617 YAC clone, derived from fragment sizes in Table 3.10, with the position of L(E0250) indicated below. Diagonal stripes show fragments on which L(E0250) lies. Deduced postion of L(F1001) is shown with an arrow. **L**, left arm; **R**, right arm; B, *Bss*HII site; E, *Eag*I; M, *Mlu*I; A, *Sal*I; F, *Sfi*I; S, *Sst*II. There are no sites for *Not*I in this YAC. Orientation with respect to the X chromosome is indicated.

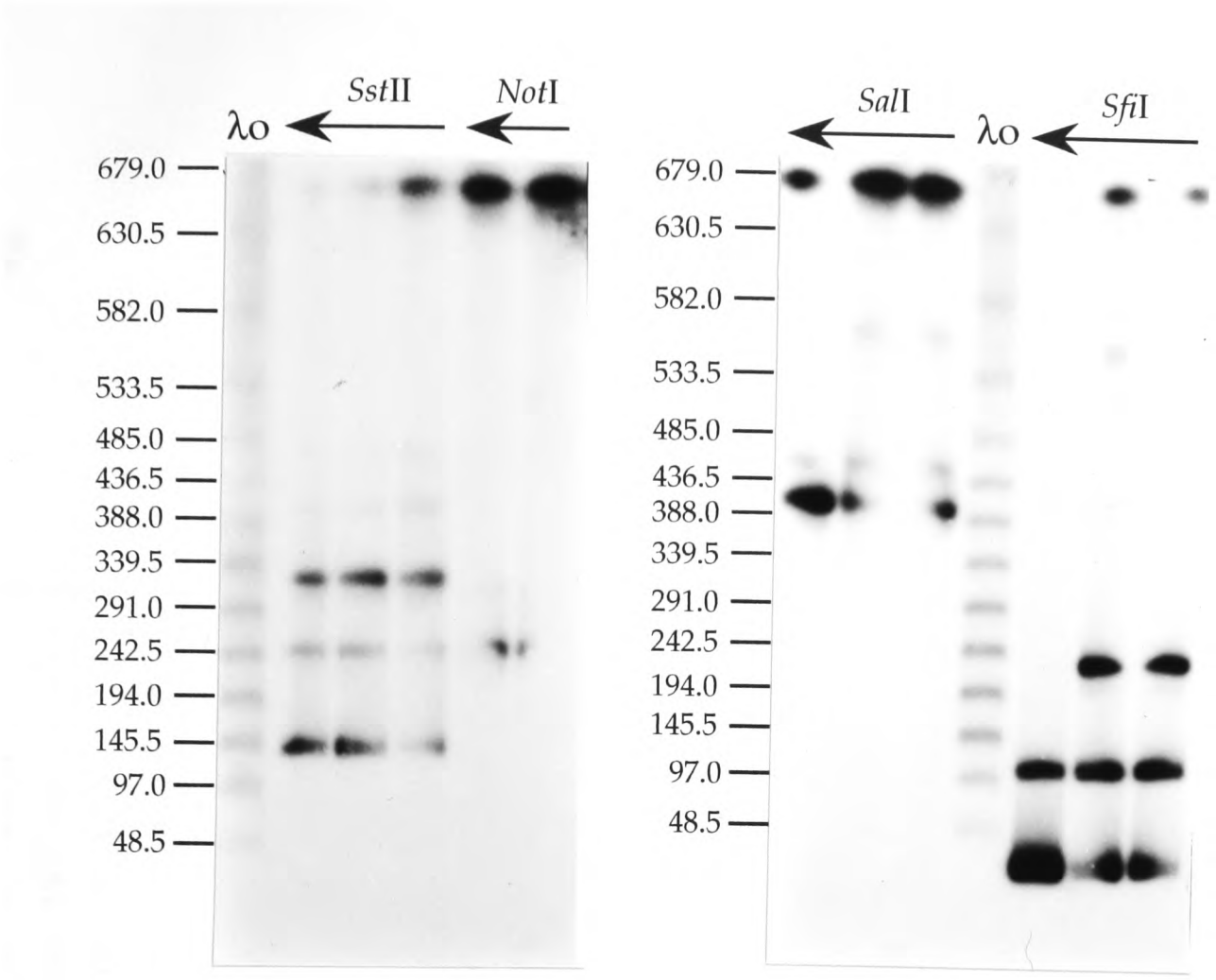


Figure 3.33: Examples of partial digests of the E0250 YAC using different rare cutters, probed with **a)** left vector arm, **b)** right vector arm and **c)** A(E0250). Direction of arrow represents increasing enzyme concentration (in the range of 0.1-15U) with a 1 hour digestion time. Sizes of lambda oligomer markers (λo) are given in kilobases. Digests were run on a standard pulsed field gel (section 2.14.5) with a 50 second switch time and a 32 hour run time. Fragment sizes and the rare-cutter restriction map which was derived from them are given in Table 3.11 and Figure 3.34.

Enzyme	Left vector arm	Right vector arm	A(E0250)
*Mlu*I	50, 615	45	560, 615
*Not*I	250, 310[a]	185, 350, 410	165, 350, 410, 475[a]
*Sal*I	400, 450, 565	95, 210, 260	400, 450, 565
*Sfi*I	30, 105, 235, 390[a], 565[a]	40, 95, 270, 425	150, 330, 425
*Sst*II	140, 250, 320, 405, 480	95, 180, 250, 340	90, 155, 340, 405, 480

Table 3.11: Fragment sizes, in kilobases, of bands detected on rare-cutter partial digests of E0250 YAC clone, when probed with vector (left and right arms), and the internal marker A(E0250). The 660kb fragments corresponding to undigested YAC are not listed. [a] bands are present, but faint.

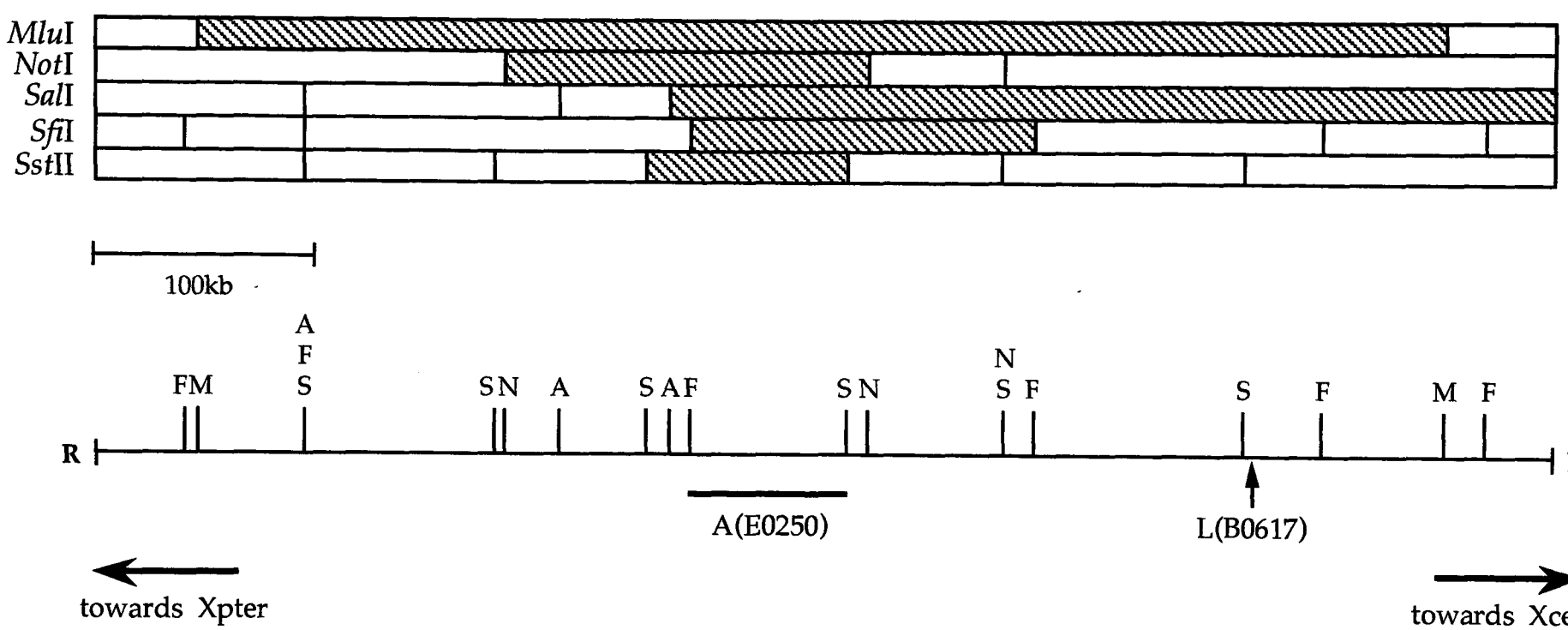


Fig. 3.34: Rare-cutter restriction map of E0250 YAC clone, derived from fragment sizes in Table 3.11, with the position of A(E0250) indicated below. Diagonal stripes show fragments on which A(E0250) lies. **L**, left arm; **R**, right arm; M, *Mlu*I; N, *Not*I; A, *Sal*I; F, *Sfi*I; S, *Sst*II. The deduced position of L(B0617), and orientation with respect to the X chromosome are indicated.

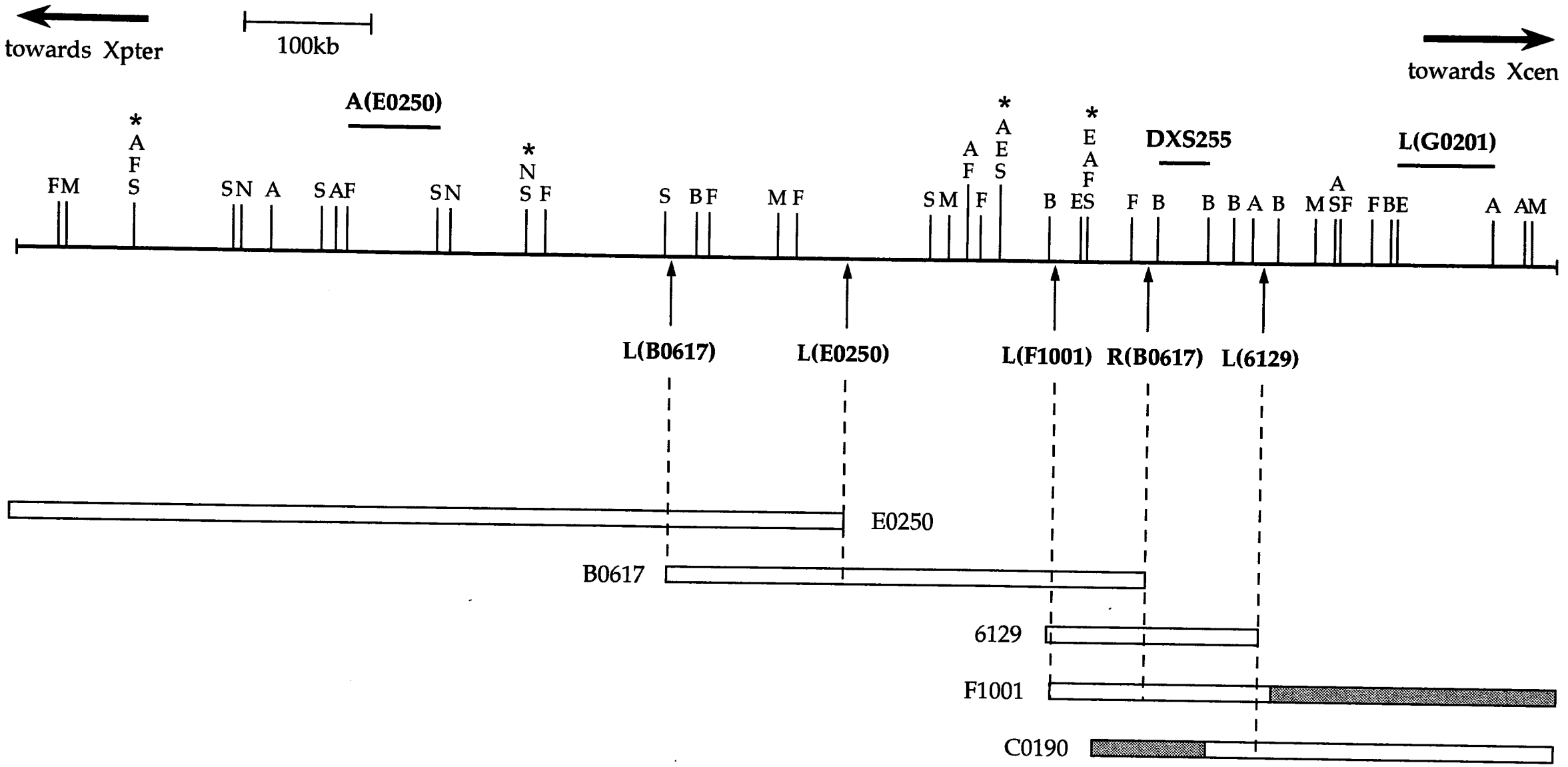


Figure 3.35: Complete rare-cutter restriction map for a 1.23Mb region spanning DXS255 (top), derived from indirect end-label mapping of five overlapping YACs (shown below). B, *Bss*HII; E, *Eag*I; M, *Mlu*I;N, *Not*I;A, *Sal*I; F, *Sfi*I; S, *Sst*II. Asterisks denote putative CpG islands (see text). Positions of novel markers generated in this study are shown by arrows. Localizations of A(E0250), DXS255 and L(G0201) are imprecise and are shown above the map. Shaded parts of YACs are likely to be chimæric/rearranged. Orientation with respect to Xpter-Xcen is indicated. Note that, because E0250 has not yet been mapped with *Bss*HII or *Eag*I, no sites for these enzymes are indicated in the portion of the diagram telomeric to L(B0617).

3.4 Discussion

The contig encompassing DXS255, which was constructed using a bi-directional walking strategy, contains nine YAC clones, isolated from three different libraries. Three steps, one in the proximal direction and two in the distal direction, were required for contig assembly. Eight novel markers were generated by plasmid rescue, inverse-PCR and *Alu*-PCR. The order of these new markers, on the basis of YAC analysis, was found to be Xpter–A(E0250)–R(27CB12)–L(B0617)–L(E0250)–R(36HB8)–L(F1001)–R(B0617)–DXS255–L(6129)–Xcen.

The C0191 clone played a critical role in linking the DXS255 YACs to the four clones of the DXS146 cluster, thereby establishing the orientation of the YACs in each cluster relative to the X chromosome. It was thus possible to identify which of the DXS146 end clones represented the most centromeric marker, and library screenings with this STS have since led to the isolation and characterization of a further four clones, proximal to DXS146 (Hatchwell, 1994). The entire DXS255–DXS146 contig therefore embraces seventeen YACs in total.

Several difficulties may be encountered when attempting to construct a series of overlapping YAC clones which accurately represent a chromosomal region. As stated in the introduction, chimæric clones comprise a significant proportion of YACs in a library, and the detection of such clones is important when chromosome walking. It was therefore essential to confirm X-specificity of each novel marker generated from the contig. In most cases, a novel marker was found to be present in two or more independent Xp11.22 YACs which were already known to overlap. This was taken as strong evidence that the marker was Xp11.22-specific since it is highly unlikely that two YACs will contain chimæric regions from the same part of the genome. Further methods for detection of chimæras are the comparison of rare-cutter restriction maps between different YACs (which suggests here that C0191 may be chimæric), or the use of FISH (fluorescence *in situ* hybridization) to identify the chromosomal band(s) in which the YAC maps.

Previous studies have found evidence of two types of rearrangement in a small proportion of clones in YAC libraries. Albertsen *et al.* (1990) showed that retransformation with a YAC of known size can yield clones of smaller size which have undergone internal deletions. Such rearrangements are clonally stable, and are likely to be induced by nicking of YAC DNA when it is manipulated in library construction, followed by damage repair in the host. Since they are stable, such rearranged clones are difficult to detect, but their frequency has been estimated at only 5% of clones (Albertsen *et al.*, 1990). The second type of rare rearrangement involves clonal instability of a YAC, which is observable as size variation of the clone. As described in this chapter, YACs isolated with the L(B0617) probe had a tendency to undergo rearrangements involving gain or loss of ~20kb regions of DNA, whilst all clones from other regions of the contig were stable. This supports the proposal that clonal instability, which has been estimated to occur in 2% of clones, is sequence-dependent and therefore an inherent property of the region cloned in the YAC (Albertsen *et al.*, 1990). Rare-cutter mapping of the different forms of each unstable YAC clone may help to identify the region of instability, which is likely to map in the overlap between the YACs (see Figure 3.27).

Indirect end-label mapping of five overlapping YACs has enabled the construction of a rare-cutter restriction map spanning the entire E0250-C0191 region around DXS255. Seven of the novel markers generated from the contig were localized within this map and it was thus possible to derive the extent of overlap between YACs. These results suggest that the size of the E0250–C0191 YAC cluster assembled here is 1.23Mb. When this is combined with preliminary data from more proximal YACs around DXS146 (Hatchwell, 1994) an estimate of over 2.1Mb is obtained for the complete DXS255–DXS146 contig.

Given the possibility of rearrangements and chimærism in YAC clones, a degree of caution should be exercised when interpreting restriction maps made from contigs. Several regions of the rare-cutter map presented here are based on data from more than one overlapping YAC, and it is encouraging to note that, in general, there is

agreement between the relative positions of restriction sites and internal markers in the different clones. There are two exceptions to this:

i) The majority of F1001 was already suspected to be chimæric, and map comparisons suggest that the X-specific part of the YAC is restricted to ~200kb at its left end.

ii) The restriction sites in a ~100kb region at the left end of C0191 do not correspond with those established from overlaps between F1001, 6129 and B0617, and probes from this region (DXS255 and R(B0617)) are absent from the clone. It is unclear, at this stage, whether these discrepancies are due to rearrangement or chimærism of C0191, but FISH studies suggest that the latter is unlikely (C. S. Cooper, personal communication).

Indirect end-label mapping of further stable YAC clones from the DXS255–DXS146 contig is necessary in order to verify the current map and extend it towards the centromere. In addition, it will be important to compare the YAC rare-cutter map to genomic long-range pulsed field maps of the region when they are reported. It should be noted, however, that whilst all of the novel markers generated in this study hybridize to single copy sequences in YAC clones, many of them are repetitive when used to probe genomic digests, and it may therefore be difficult to localize them within a genomic map. A further complication involves the fact that CpGs are extensively methylated in genomic DNA (see Section 3.1.3). Since most rare-cutter enzymes are methylation sensitive and contain a high proportion of CpG in their recognition sites, many of the sites present in cloned DNA will not be detected in genomic mapping (Bird, 1986).

One of the markers isolated by end cloning was found to be polymorphic, due to the presence of an imperfect CA repeat. Although L(B0617) only has a heterozygosity of 0.58 and PIC value of 0.48, it maps 385-425kb distal to DXS255 and could therefore be helpful for revising the localizations of diseases such as RP2, CSNB and ÅIED in cases where it is found to be informative. In addition, the YAC clones identified here may represent a useful resource for the generation of further polymorphic markers (Cornélis *et al.*, 1992 and see General Discussion).

A main objective of YAC physical mapping is to facilitate the isolation of transcripts from the region. As described above (Section 3.1.3), one criterion for the identification of CpG islands, which are often associated with genes, is that they are non-methylated (Bird, 1986). Although cloned DNA completely lacks methylation, it is still possible to identify putative CpG islands, as sites in the restriction map which can be cleaved by several different rare-cutters (Lindsay and Bird, 1987). For example, given the calculated percentages of rare-cutter sites in islands presented in Table 3.2, it can be estimated that a site recognized by three enzymes such as *Bss*HII, *Eag*I and *Sst*II, will be located in a CpG island in over 98% of cases. There are three sites in the rare-cutter YAC map presented here which are cleaved by three or more enzymes (Fig. 3.35). In addition, it should be noted that a particularly high proportion (89%) of *Not*I sites are located in CpG islands, and that it is therefore highly likely that the *Not*I/*Sst*II sites found in the E0250 YAC also identify an island (Fig. 3.35). Further analysis of the DNA in the vicinity of these four putative CpG islands may aid the isolation of novel genes from the region distal to DXS255.

A(E0250) represents the most distal novel marker from the contig, and is not present in the OATL1–GATA–TFE3–SYP cluster described in Chapter 4. When E0250 is used in FISH analysis, the predominant signal is seen in Xp11.22, but an additional weaker signal is seen in Xp22.1 (C. S. Cooper, personal communication). Hybridization to the latter may be due to chimærism of E0250, with DNA of Xp11.22 origin fused to that from Xp22. Alternatively, it could be a result of sequences which are present at Xp11.22, but repeated at Xp22.1. At this stage it is not clear which of these two explanations is correct. However, comparison between the intensities of the signals suggest that at least 80% of the clone originates from Xp11.22 (C. S. Cooper, personal communication), and given that A(E0250) lies in the middle of the YAC, it seems highly likely that this marker is also Xp11.22-specific. Screening of YAC libraries using the PCR assay developed from A(E0250) should therefore facilitate further extension of the DXS255–DXS146 contig towards the telomere.

Chapter 4 – Construction of a YAC contig linking the loci SYP, TFE3, GATA and OATL1 in Xp11.23-p11.22

4.1 Introduction

4.1.1 Genes mapping in the OATL1-DXS255 interval

Following the isolation of a gene on the basis of its function, methods such as linkage analysis (if it has a polymorphism), *in situ* hybridization and analysis of somatic cell hybrids can be used to ascertain where the locus maps in the genome. The gene can then serve as a general purpose marker to aid physical characterization of the region in which it maps. In addition, it may be considered as a candidate gene for diseases that have been localized to the same region.

Three genes which were identified by functional cloning have been mapped to Xp11.23-p11.22, in the region distal to DXS255:

i) Synaptophysin (SYP)

Synaptophysin is an abundant integral membrane protein of small synaptic vesicles in brain and endocrine cells, whose precise function is not known (Ozcelik *et al.*, 1990). Somatic cell hybrid analysis localized SYP to the interval between the proximal breakpoint of the SIN176 deletion in Xp11.23 and the DUA translocation breakpoint in Xp11.22 (Fig 4.1). DXS255 and DXS146 have been mapped to this same interval using these hybrids (Lafrenière *et al.*, 1991b). Linkage studies with an *Eco*0109 RFLP detected by SYP suggested that the gene lies in the interval between TIMP (which is distal to OATL1) and DXS255 (Ozcelik *et al.*, 1990).

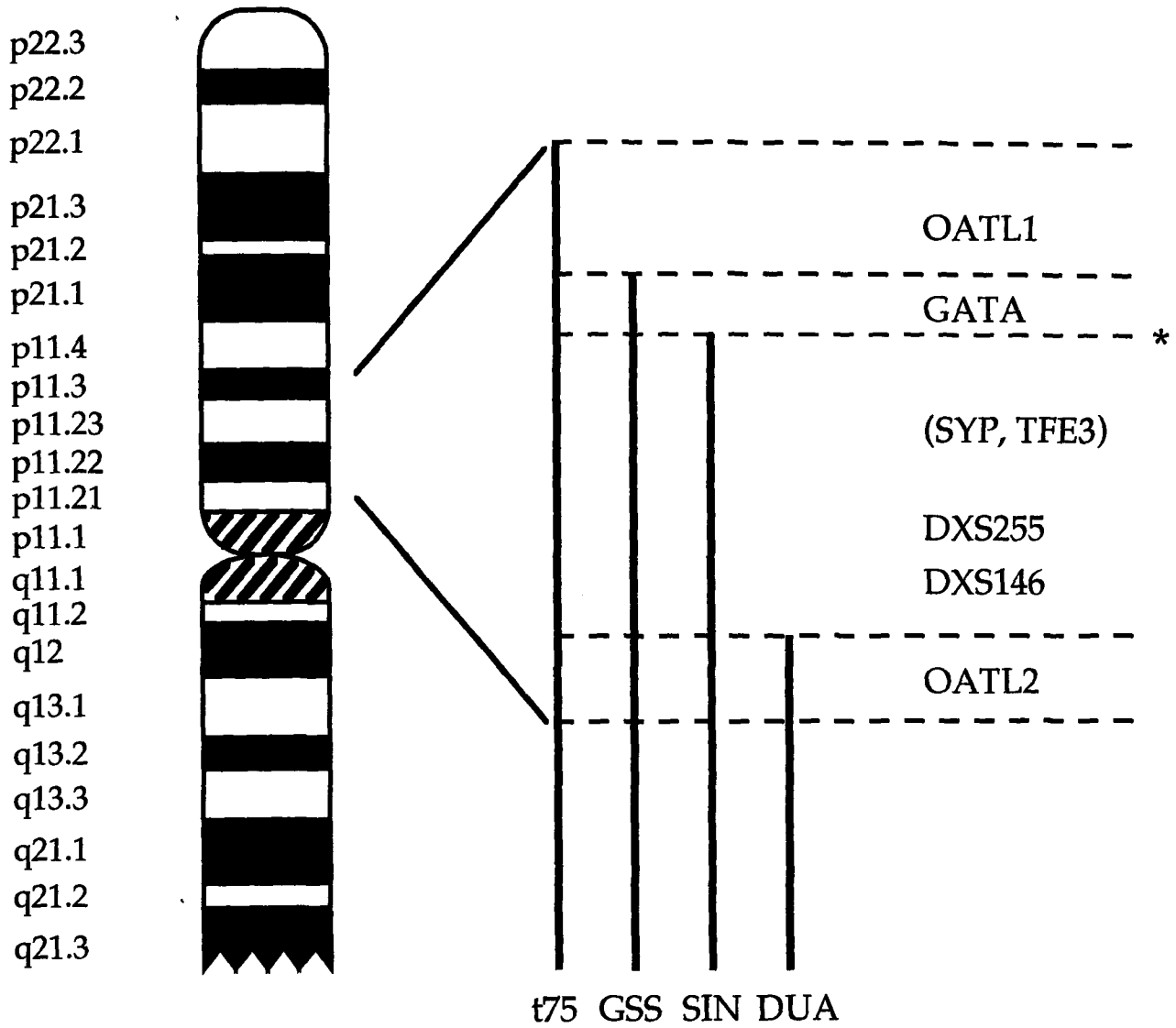


Figure 4.1: Relative positions of markers in Xp11.23-p11.22, prior to construction of YAC contigs, as determined from linkage studies, somatic cell hybrid analysis and pulsed field genomic mapping (see text for details). Translocation breakpoints are **t75**, t75-2ma-1b (p11.3-qter) (Brown and Willard, 1990); **GSS**, Gilgenkrantz synovial sarcoma (p11.23-qter) (Gilgenkrantz *et al.*, 1990)); **SIN**, SIN176 (pter-p22.11::p11.23-qter) (Ingle *et al.*, 1985); **DUA**, DUA-1A (p11.22-qter) (Myerowitz *et al.*, 1985). Only the proximal breakpoint of the SIN deletion is shown, indicated with an asterisk.

ii) TFE3

TFE3 is a member of the helix-loop-helix family of transcription factors which binds to the μE3 motif within the immunoglobulin heavy-chain enhancer and is ubiquitously expressed (Henthorn *et al.*, 1991). Hybridization of TFE3 cDNA to a panel of somatic hybrids established that it mapped to the same SIN176/DUA interval as SYP, DXS255 and DXS146 (Fig 4.1 and Lafreniere *et al.*, 1991b). When an *Rsa*I RFLP detected by TFE3 was used in linkage analysis of 20 pedigrees, it was suggested that the most likely location of the gene was in between DXS255 and DXS146 (Henthorn *et al.*, 1991). However, a genomic pulsed field map of the DXS255-DXS146 interval indicated that this location was incorrect (Riley, 1993); the next most likely position (on the basis of linkage) is distal to DXS255.

iii) GATA1

The transcription factor GATA1 binds to the core DNA sequence (A/T)GATA(A/T) found in the promoters and enhancers of many erythroid specific genes and is essential for normal erythroid development (Zon *et al.*, 1990). The human locus was initially assigned to Xp21-p11 by somatic cell hybrid studies, and this localization was then refined to Xp11.23 using *in situ* hybridization (Caiulo *et al.*, 1991). Further somatic cell hybrid analysis supported this, and showed that GATA1 lies between the Gilgenkrantz synovial sarcoma (GSS) breakpoint, and the SIN176 proximal breakpoint (Fig. 4.1), and is therefore in the OATL1-DXS255 interval (Laval and Boyd, 1993; see Section 4.1.2).

Given that GATA1 is absent from SIN176, whilst SYP and TFE3 are present, the most likely order for the genes in Xp11.23-p11.22, as suggested by the above studies, is Xpter-GATA1-(SYP/TFE3)-DXS255-DXS146-Xcen. All three genes have been considered as candidates for disease genes, such as Wiskott-Aldrich syndrome (WAS) and retinitis pigmentosa 2 (RP2), which map in the interval, but no associated mutations have been found.

4.1.2 A previously characterized YAC contig around OATL1

Ornithine-δ-aminotransferase (OAT) is a mitochondrial matrix enzyme involved in arginine, proline and ornithine metabolism, the loss of which causes gyrate atrophy (progressive degeneration of the choroid and retina). Although the functional gene maps to chromosome 10, clusters of OAT-like (OATL) sequences have been mapped to two non-adjacent intervals on the proximal short arm of the X (Lafreniere *et al.*, 1991a). At least some of these appear to be processed pseudogenes. Whilst the OATL2 sequences lie in Xp11.22-p11.21, proximal to the DUA translocation breakpoint (and hence on the centromeric side of DXS146), the OATL1 sequences map to Xp11.3-p11.23, distal to the SIN176 breakpoint (and therefore on the telomeric side of TFE3 and SYP) (Fig. 4.1). The Gilgenkrantz synovial sarcoma (X;18) translocation breakpoint was shown to map in the cluster of OATL1 sequences (de Leeuw *et al.*, 1993), indicating that OATL1 is distal to GATA (Fig. 4.1). Four YACs have been isolated for OATL1 (A. P. Monaco, personal communication), one of which contains the GSS breakpoint. These were being characterized in detail at the time this work began (Chand, 1994).

4.1.3 Aims

The objective of this part of the work was to use the SYP, TFE3 and GATA1 genes as markers to aid construction of a YAC contig in Xp11.23-p11.22 which would bridge the gap between the OATL1 YACs and the DXS255–DXS146 cluster. At the start of this project, there was a general lack of physical mapping data for the region, beyond the approximate locus assignments described above (Section 4.1.1). Estimates of TIMP-DXS255 distance (~3-4cM) were based on recombination frequencies (Greer *et al*, 1990, Kwan *et al*, 1991), which do not always correlate with physical distance (Section 1.1.2). In addition to providing further information on the organization of the region, a comprehensive YAC contig, containing newly generated markers, would be a useful starting point for the identification of disease genes mapping to the region, like WAS and RP2. The OATL1-DXS255 interval is also thought to contain an (X;1) translocation breakpoint associated with renal papillary adenocarcinoma (de Jong *et al.*, 1986; Meloni

et al., 1993; Sinke *et al.*, 1993) and YACs from the region could facilitate more precise localization of the breakpoint.

The strategy employed was similar to that in Chapter 3. Previously characterized genes were used to isolate YAC clones, novel markers were generated from the ends of these YACs and overlaps between the various clusters were detected by the presence of a common marker. Some of the YACs isolated in this chapter were identified by hybridization of probes to gridded library filters which became available during the course of study, rather than the PCR screening described in Chapter 3. The work was successful in linking OATL1, GATA1, TFE3 and SYP, establishing their order on the X chromosome, and identifying YACs which were later useful in the mapping of the renal carcinoma breakpoint.

4.2 Materials and methods

4.2.1 Probes and PCR assays.

A 0.9kb *Eco*R1 fragment from the hGATA1 cDNA (Table 4.1), subcloned into pUC19, was kindly made available by S. H. Orkin. The TFE3–1.9 probe, which was kindly provided by J.M.Puck, is a 1.9kb *Eco*R1 cDNA fragment derived from the 3' portion of the human TFE3 cDNA, and cloned into pBluescript SK+ (Table 4.1). The complete cDNA sequence and genomic organization of the human SYP gene has been described (Ozcelik *et al.*, 1990). This was used to design primers which would amplify a 551bp fragment (SYP') from the 5' untranslated part of the gene (ending 150bp upstream of the open reading frame) for use as a hybridization probe. The small size of the coding exons of SYP, combined with a lack of information regarding intron size and sequence, hindered attempts to amplify a probe from the more 3' part of this gene.

Marker	Size	Cognate *Eco*R1 band	STS	F0501	27GF2	3578	C01160	4542	B102	12E11	5H12	TFE3-E021	SYP-E021
R(3578)	200bp	1.9kb[b]		+	–	–	–	–	–	–	–	–	–
WASP	600bp	9.0kb	+	+	+	+	–	–	–	–	–	–	–
GATA	900bp	2.7kb	+	–	+	+	+	+	+	–	–	–	–
L(B102)	650bp/350bp[a]	2.4kb	+	–	–	–	–	–	+	+	–	–	–
TFE3	1.9kb	7.9kb	+	–	–	–	–	–	–	+	+	+	+
L(E021)	3.8kb	4.1kb[c]		–	–	–	–	–	–	–	–	+	+
SYP'	551bp	7.7kb	+	–	–	–	–	–	–	–	–	–	+
SAE	3.2kb	3.2kb	+	–	–	–	–	–	–	–	–	–	+

Table 4.1: Details of markers from OATL1–GATA–TFE3–SYP cluster in Xp11.23-p11.22. Markers are listed in the order Xpter-Xcen. L(YAC ID), left end of YAC isolated by plasmid rescue; R(YAC ID), right end of YAC isolated using inverse-PCR. +, present in YAC; –, absent from YAC. Those markers which have been converted into STSs are indicated. [a] This left end-clone gives two fragments on double digestion with *Eco*R1/*Nde*I. [b]Probe gives repetitive signal on genomic digests. [c]The fragment recognized by this marker is autosomal.

Marker name	Primer sequences (5' to 3')	T_a (°C)	Product size (bp)
WASP	GAGAAGACAAGGGCAGAAAGC CCGTAAAGGCGGATGAAGTA	55	600
GATA	AACCGCAAGGCATCTGGAA CTGGCTACAAGAGGAGAAG	53	383
L(B102)	TACAGGCATCCACCACCC AGGCAGGAAAAGCATCTAAGC	55	225
TFE3	AGGCGCACGAACGTTCCATGT CCTTCTCCAGCCTTCTCCTTC	61	544
SYP	CATAGCCATCACTGTTGACC TACCACAGTCTCACCGGCA	55	551
SAE	AAGAGATTTATGATCTGTGGGC CATAGTGGGTGAGAGGGCTG	55	154

Table 4.2: Details of PCR assays developed from markers in OATL1–GATA–TFE3–SYP YAC contig. Order of markers is Xpter-Xcen. T_a, annealing temperature. Additional details of PCR amplification conditions are given in Materials and Methods.

Exon-intron organization has not yet been described for either GATA or TFE3. However, hybridization analysis of various genomic digests with these cDNA probes identified restriction fragments which were unlikely to contain introns (data not shown) and the reported cDNA sequence in these region could therefore be used to design primers which would amplify from the loci from genomic template. Details of all PCR assays which were developed for this part of the study are given in Table 4.2.

4.2.2 Screening of libraries using hybridization to gridded filters

i) YAC libraries

Gridded library filters were available for only a subset of YAC libraries at the time of study:-

- ICRF YAC library filters contain clones from three sources; the **4X library (no. 900)**, made from the GM1416B cell line; the **4Y library (no. 901)**, made from the Oxen49XYYYY cell line (A.P. Monaco, unpublished) and a library **(no. 905)** made from the lymphoblastoid cell line HD1 (M. Ross, unpublished). These libraries were gridded onto two filters containing 144 x 144 squares, with each square consisting of four YAC clones spotted in duplicate (see Fig. 4.6). The 4X library could also be screened by PCR (see Section 3.2.2).

- The Nussbaum X-specific library (Lee *et al.*, 1992), which was constructed from the Micro-21D cell line, is gridded onto two filters containing 24 x 16 squares, with each square consisting of four YAC clones spotted in duplicate (see Fig. 4.6).

ii) Cosmid library

The ICRF X-specific cosmid library (no. 104) was constructed from digests of DNA from 'flow sorted human X-chromosome', ligated into Lawrist4 vector, with DH5 alpha MCR as host (Nizetic *et al.*, 1991). Each filter consists of a grid of 144 x 144 clones representing the entire library. Duplicate gridded filters are screened, to ensure that

legitimate positives can be distinguished from any spots that result from non-specific background hybridization. True positives are identified as those that appear in corresponding positions on both filters.

4.2.3 Preparation of cosmid DNA using modified alkaline lysis

1. Inoculate 10ml of 2 x TY broth, containing 20mg/ml kanamycin, with a single bacterial colony or 50µl of frozen stock.

2. Incubate overnight (or until turbid) in a 37°C shaking incubator.

3. Centrifuge at top speed in an MSE bench centrifuge for 10 minutes.

4. Drain pellet and resuspend in 300µl of GTE containing lysozyme (Sigma) at 5mg/ml.

5. Transfer to a 1.5ml Eppendorf tube and incubate at room temperature for 5 minutes.

6. Add 100µl 4% SDS and 100µl 0.8M NaOH, mix, and stand at room temperature for 5 minutes.

7. Add 200µl of freshly made 5M KAc (3ml 5M KAc, 575µl glacial acetic acid, 1425µl water), mix, and incubate on ice for 10 minutes.

8. Spin in an Eppendorf centrifuge at room temperature for 5 minutes, to pellet cell debris and bacterial DNA.

9. Transfer supernatant to a fresh tube and add 200µl of 2M Tris-Cl (pH 8.9) containing 20µg/ml DNAse-free RNAse.

10. Incubate at 37°C for 30 minutes.

11. Add 600µl of cold (4°C) isopropanol. Mix and incubate at room temperature for 10 minutes.

12. Spin in an Eppendorf centrifuge at room temperature for 10 minutes.

13. Wash pellet of DNA with 1ml of 70% ethanol, spin for 5 minutes, and remove supernatant.

14. Dry DNA and dissolve in 50µl of TE. Store at -20°C.

4.3 Results

4.3.1 Isolation and analysis of YACs containing the GATA gene

Two GATA YACs, 3578 and 4542 (Table 4.3), were obtained by screening of the St. Louis library using a PCR assay which was developed for the gene (Tables 4.1 and 4.2). Hybridization of the GATA cDNA to gridded filters of the ICRF library identified two further clones, C01160 and B102. Pulsed field analysis of undigested clones indicated that C01160 and 3578 were stable and of sizes 120kb and 125kb respectively (Fig. 4.2). However, all preps of 4542 contained two YAC species, one of 300kb, the other of 160kb, and neither hybridized to the GATA cDNA (Fig. 4.2). (The 4542 clone has been confirmed as a GATA positive YAC by the administrators of the St. Louis library.) Furthermore, size variation was seen between alternative preps of the B102 YAC; some contained a YAC of 300kb, in others the predominant species was 200kb (Fig. 4.3a). Hybridization of GATA cDNA to *Eco*R1 digests of these preps suggested that the 200kb YAC had undergone a deletion of the region encompassing the GATA locus (Fig. 4.3b). None of the GATA YACs were found to contain OATL1, TFE3 or SYP.

4.3.2 Linking of the GATA cluster distally to OATL1

The right end of 3578 was isolated and mapped against a panel of YACs. This indicated that it was also present in OATL1.2, the most proximal of the OATL1 cluster (Fig. 4.4), suggesting that there was overlap between the OATL1 and GATA clones. Chromosome walking in a proximal direction from OATL1.2 has confirmed this; a YAC (27GF2) isolated using a novel marker from OATL1.2 was found to also contain GATA and R(3578) (Chand, 1994 and see Fig. 4.4). It is interesting to note that the 27GF2 clone, like the 4542 and B102 YACs, is susceptible to deletions of the region encompassing the GATA locus (Chand, 1994).

YAC ID	Alternative name	Library	Isolated with	Size/kb	Left end	Right end	Other markers present
F0501	OATL1/2	ICRF	OAT cDNA	550	X		L(OATL1/11), A(F0501), R(3578), WASP
27GF2	PTO	ICI	A(F0501)	780/750[c]			R(3578), WASP, GATA
C01160	GATA/1	ICRF	GATA cDNA	120			
B102	GATA/2	ICRF[b]	GATA cDNA	300/200[c]	X	X	
3578	GATA/3	St. Louis	GATA PCR assay	125		X	WASP
4542	GATA/4	St. Louis	GATA PCR assay	300+160[d]			
E021[a]	TFE3-E021	ICRF	TFE3 PCR assay	390	aut		
E021[a]	SYP-E021	ICRF	SYP' probe	375	aut		TFE3, SAE
5H12	TFE3/2	Nussbaum	TFE3 cDNA	120			
12E11	TFE3/3	Nussbaum	TFE3 cDNA	230			L(B102)

Table 4.3: Details of YACs from the OATL1–GATA–TFE3–SYP cluster in Xp11.23-p11.22. YACs are listed in the order Xpter-Xcen. L(YAC ID), left end of YAC; R(YAC ID), right end of YAC; A(YAC ID), *Alu*-PCR probe from YAC; X, end clone X-specific; aut, end clone autosomal indicating chimæric YAC. [a]two different clones of the E021 YAC were isolated (see text). [b]YAC isolated from ICRF library 905. [c]YAC rearranging. [d]two forms of same YAC present in all colony pure preps.

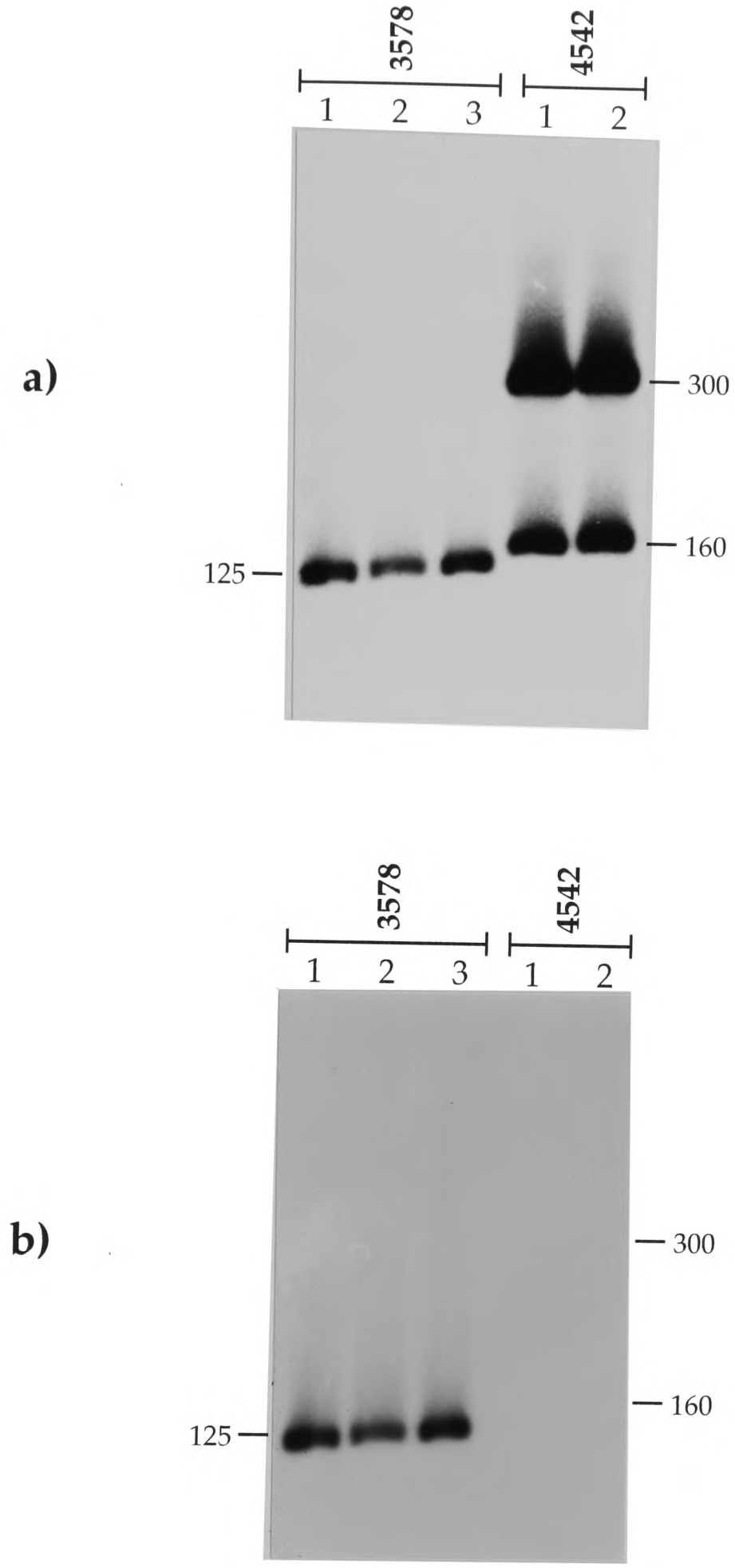
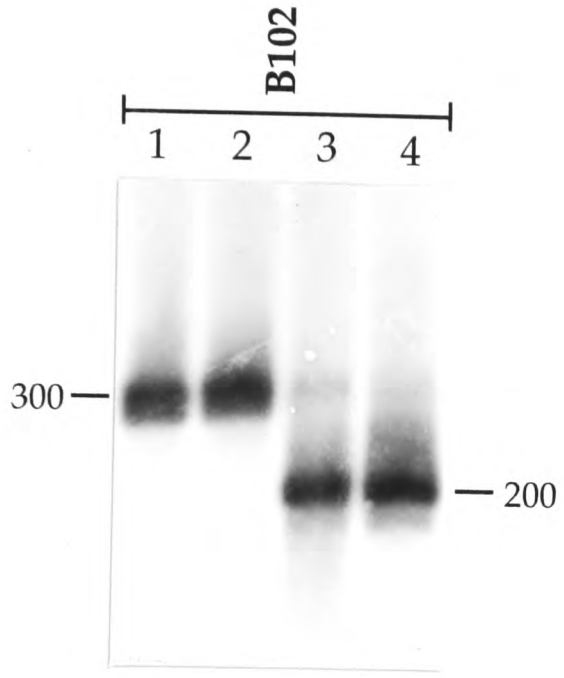
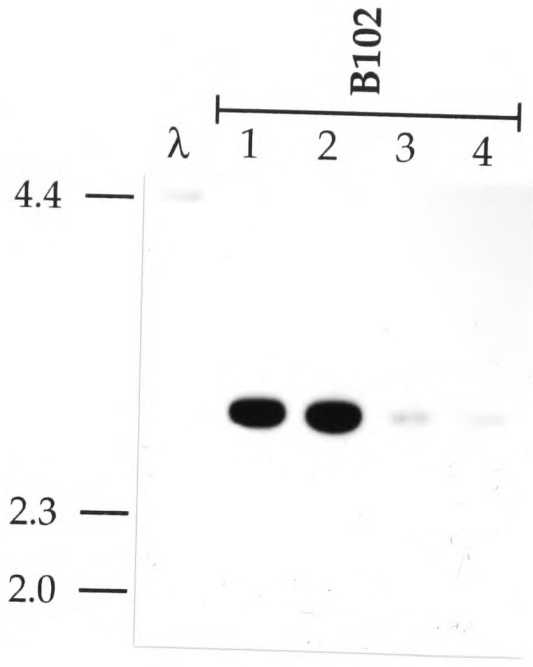


Figure 4.2: Sizing of the St. Louis GATA YACs. Independent, undigested, colony-pure preps of each clone were run on a pulsed field gel with a 60 second switch time and a 30 hour run time. DNA from the gel was transferred to a filter by Southern blotting and probed with **a)** total human DNA and **b)** the GATA cDNA. The sizes of the YAC bands thus detected are indicated in kilobases. The 3578 YAC is clonally stable; the small differences in migration between the tracks are running artefacts. Both preps of the 4542 clone contain two YAC species, neither of which hybrides to the GATA cDNA. See text for discussion.



a)



b)

Figure 4.3: Clonal instability of the B102 GATA YAC, as revealed by pulsed field gel analysis, and probing of YAC digests.

a) Independent undigested colony-pure preps of B102 were run on a pulsed field gel with a 60 second switch time and a 30 hour run time. DNA from the gel was transferred to a filter by Southern blotting and probed with radioactively labelled total human DNA. The sizes of YACs thus detected are indicated in kilobases.

b) *Eco*R1 digests of each B102 prep, probed with the GATA cDNA (Table 4.1). Sizes of lambda markers (λ) are given in kilobases. The cognate 2.7kb band is present in all four preps, but the signal from preps 3 and 4 is of much lower intensity than that from 1 and 2. Probing of this filter with YAC vector sequences reveals bands of equal intensity in all digests (not shown) demonstrating equal loading of all tracks, and confirming that the YAC copy number is the same in each prep.

These results indicate that B102 is susceptible to rearrangement involving the deletion of the GATA locus.

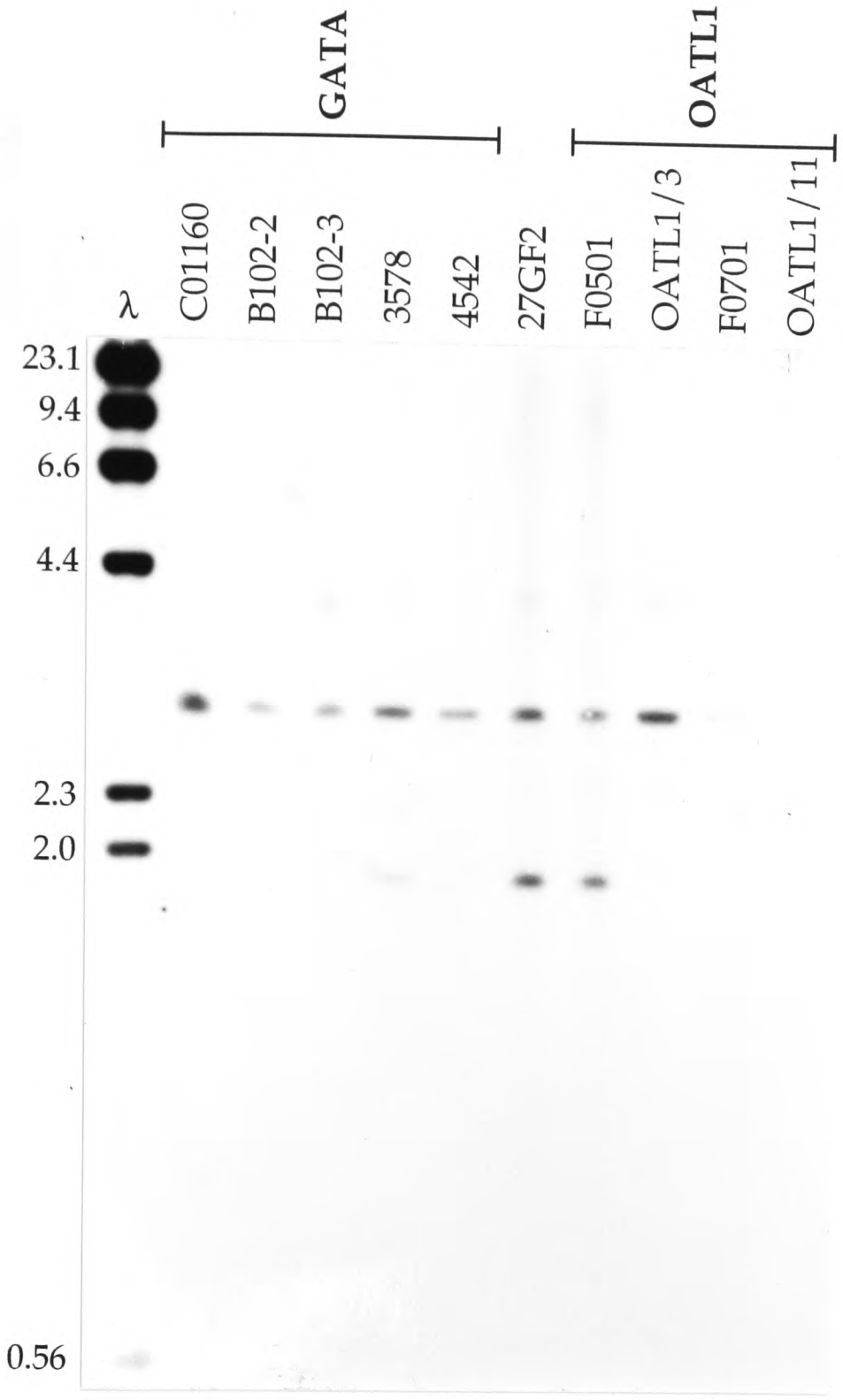


Figure 4.4: Hybridization of the R(3578) probe to *Eco*R1 digests of YACs from the Xp11.23-p11.22 contig, including the GATA and OATL1 clusters. Sizes of lambda markers (λ) are given in kilobases. The B102-2 and B102-3 clones are alternative forms of the same YAC, the latter having undergone an ~100kb deletion (see Figure 4.3). A 1.9kb fragment is detected in YACs 3578, 27GF2 and F0501. This marker therefore provides a link between the OATL1 and GATA clusters and establishes that the 3578 YAC is oriented with its right end towards Xpter. Additional fragments are detected in all tracks, because the R(3578) probe contains pYAC4 right arm vector sequences as a consequence of the inverse-PCR technique which was used to isolate it (see Figure 3.6). On hybridization to *Eco*R1 digests of human genomic DNA, R(3578) detects a smear of fragments (not shown), indicating that it contains repetitive elements.

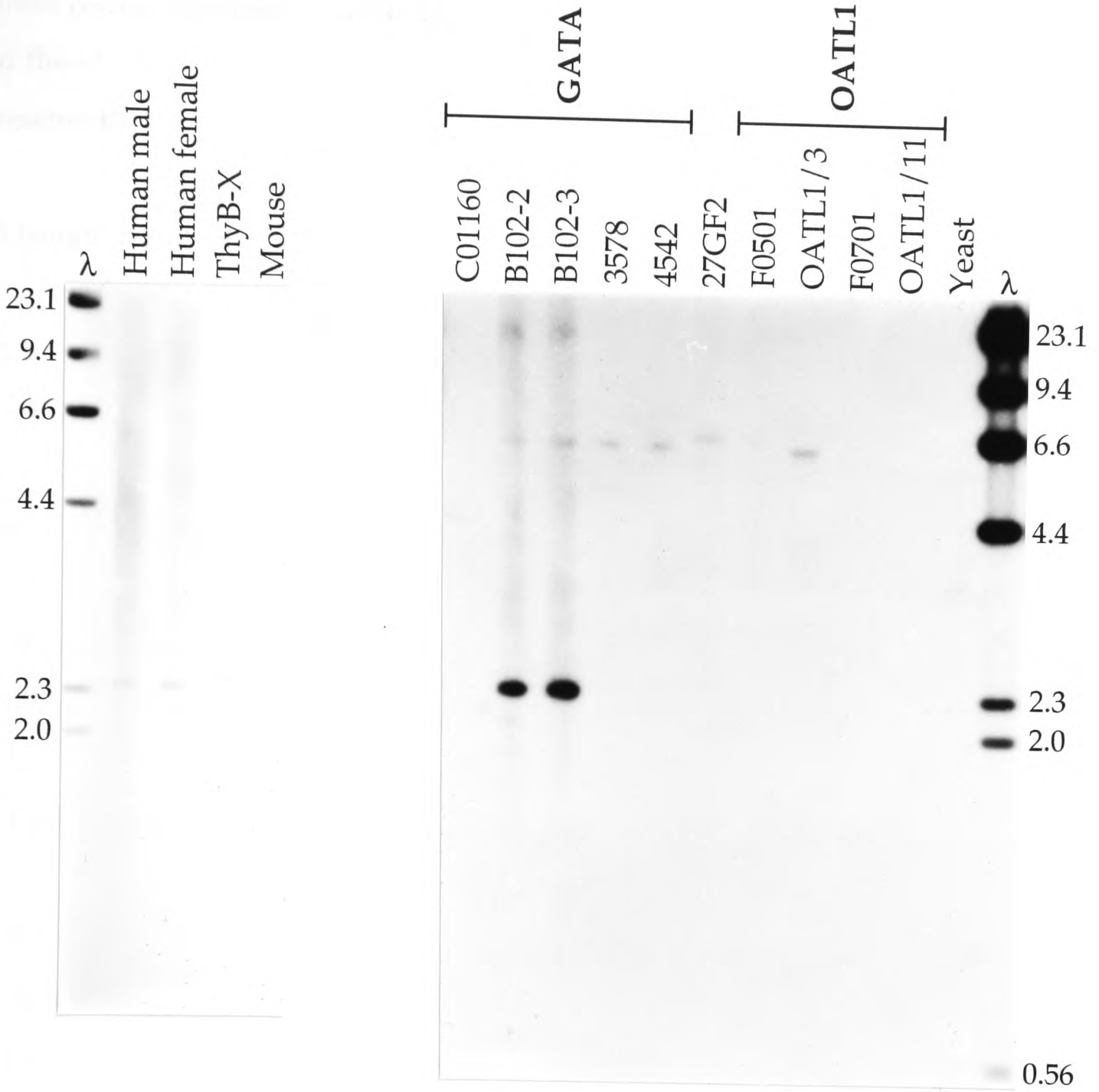


Figure 4.5: Hybridization of the 650bp L(B102) probe (see Table 4.1) to *Eco*R1 digests of genomic, hybrid and YAC DNAs. Sizes of lambda markers (λ) are given in kilobases.
a) L(B102) detects a 2.4kb band in human male and female genomic DNA, and also in ThyB-X, but not in mouse genomic DNA. The band appears to be slightly larger in ThyB-X, but this is due to overloading of the DNA in this track.
b) On probing *Eco*R1 digests of a panel of OATL1–GATA YACs, L(B102) detects a 2.4kb band in clones B102-2 and B102-3 clones, which are alternative forms of the B102 YAC, the latter having undergone an ~100kb deletion (see Figure 4.3). This fragment is absent from all other clones of the OATL1–GATA cluster, indicating that L(B102) is the most proximal marker from these YACs (see text).

Plasmid rescue was used to isolate L(B102), which was found to be X-specific, but absent from the OATL1 cluster and all the other GATA YACs (Fig 4.5). This suggested that it represented the extreme proximal end of the OATL1–GATA contig.

4.3.3 Isolation and analysis of YACs containing the TFE3 gene

A PCR assay developed from the TFE3 locus (Table 4.2) was used to isolate a 390kb YAC, E021, from the ICRF library (Table 4.3). Two additional clones, 5H12 (120kb) and 12E11 (230kb), were identified by screening gridded filters of the Nussbaum X-specific library (Fig. 4.6). All three YACs were found to be stable.

L(E021) was isolated by plasmid rescue and found to be autosomal in origin, indicating that E021 is chimaeric (Fig. 4.7).

4.3.4 Linking of the TFE3 cluster to the more distal OATL1–GATA cluster

When L(B102), the most proximal marker from the OATL1–GATA cluster, was hybridized to filters of the Nussbaum X-specific library, a single clone was identified (Fig 4.6b). This YAC, 12E11, had already been isolated by screening with TFE3. PCR of melted YAC plugs with the L(B102) STS, and hybridization of the L(B102) probe to *Eco*R1 digested clones, confirmed that 12E11 contained both TFE3 and L(B102). This clone therefore links the TFE3 YACs to OATL1 and GATA.

4.3.5 Screening of YAC libraries with the SYP locus

A 551bp probe corresponding to the 5' part of the SYP locus (just upstream of the open reading frame) was prepared by PCR amplification of human template using the primers described in Table 4.2. Digestion of this product with *Taq*I restriction enzyme yielded fragments of the size predicted by analysis of the reported SYP sequence, thereby confirming that it did indeed originate from the SYP locus. Hybridization of this probe, known as SYP', to gridded filters of the ICRF library identified a single positive clone. Interestingly, this 'new' YAC had an identical I.D. number (E021) to that

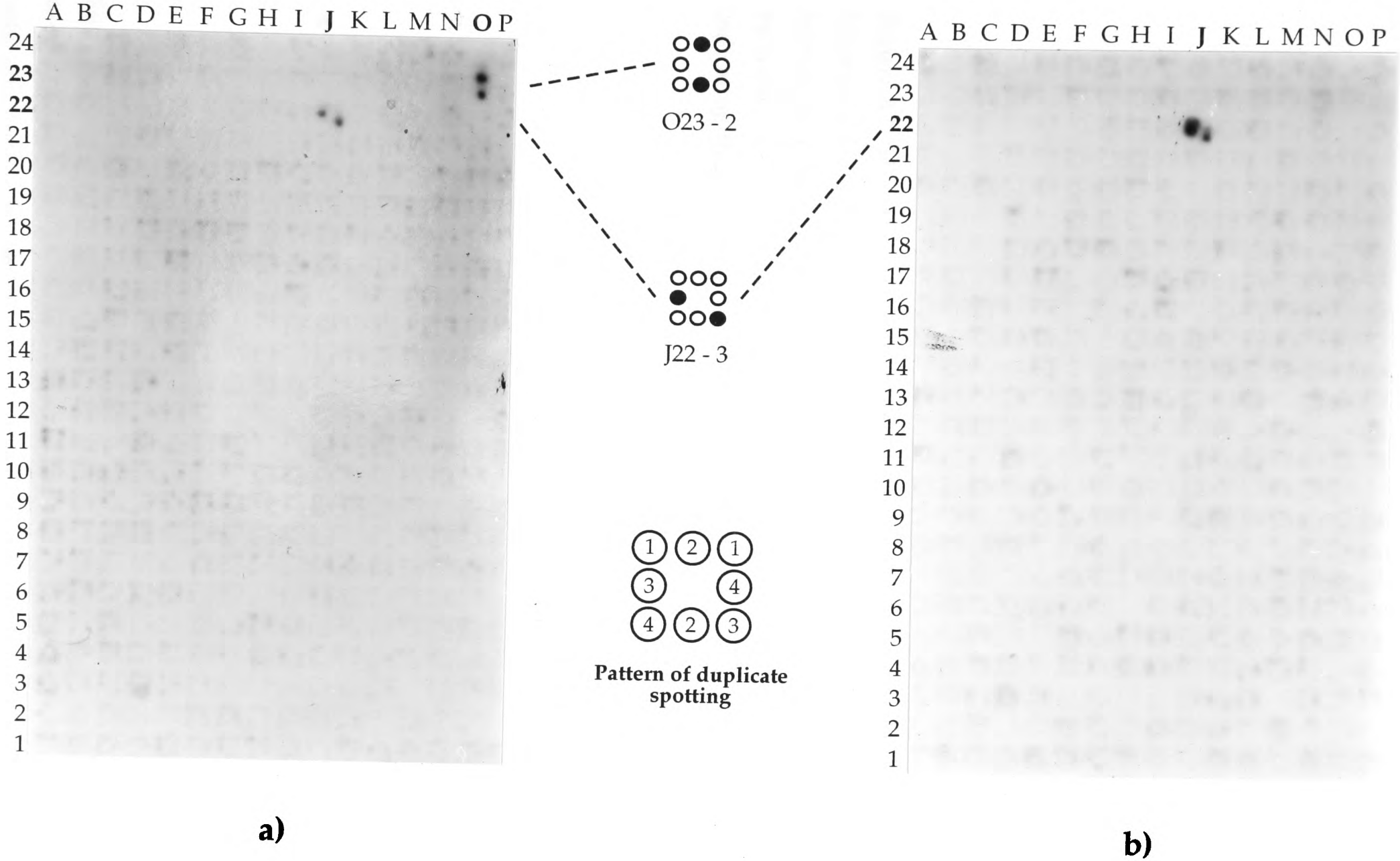


Figure 4.6: Hybridization screening of gridded filters of the Nussbaum X-specific YAC library with **a)** TFE3 and **b)** L(B102). Four clones are spotted in duplicate in each square as shown. Therefore each positive YAC is identified by the grid position of the square and a number corresponding to the pattern of duplicate spots. Official names of the YACs identified are 5H12 (for the clone at O23-2) and 12E11 (for the clone at J22-3). One YAC clone (12E11/J22-3) is identified by both TFE3 and L(B102).

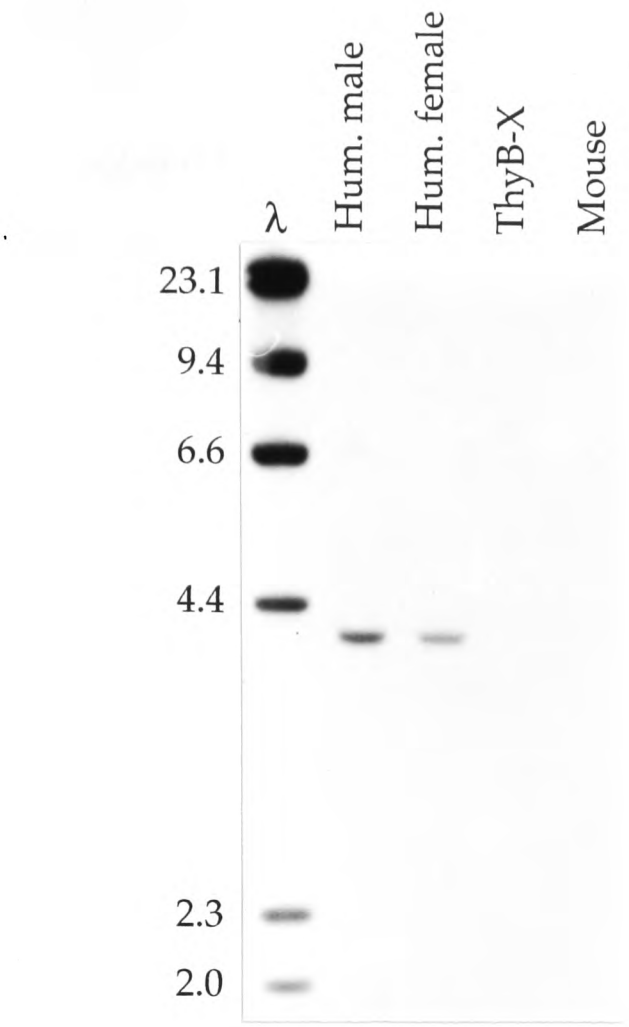


Figure 4.7: Hybridization of L(E021) to *Eco*R1 digests of genomic and hybrid DNAs. Sizes of lambda markers (λ) are given in kilobases. L(E021) detects a 4.1kb fragment in human male and female genomic DNA and also in the YAC of origin (not shown). However, no bands are detected in the human X-only/mouse hybrid (ThyB-X), indicating that the probe is autosomal in origin, and that the E021 YAC is chimæric.

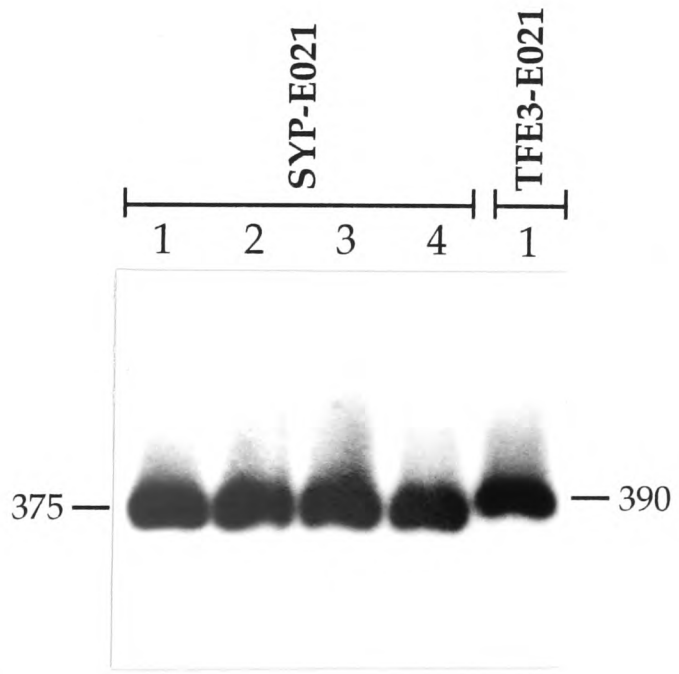


Figure 4.8: Sizing of the SYP-E021 YAC clone and comparison with TFE3-E021. Several independent undigested colony-pure preps of SYP-E021 were run on a pulsed field gel with a 60 second switch time and a 30 hour run time. Only one prep of TFE3-E021 was run on this gel, since it had already been sized and shown to be clonally stable. DNA from the gel was transferred to a filter by Southern blotting and probed with radioactively labelled total human DNA. The sizes of the YAC bands thus detected are indicated in kilobases. These results indicate that SYP-E021 is clonally stable, but smaller than TFE3-E021.

of one of the TFE3 YACs, isolated by PCR screening in Section 4.3.3. However, analysis of this original TFE3-E021 clone had already shown that it was stable and did not contain SYP. The putative SYP-E021 clone was obtained for further analysis. (It should be noted that the group who were distributing clones isolated by hybridization were different from the group who originally gave us access to the library by PCR screening. These different groups had different stocks of the library, therefore the two E021 clones came from different sources.) SYP-E021 contains both SYP and TFE3, and is stable, but is only 375kb and is therefore smaller than TFE3-E021 (Fig. 4.8). Hybridization analysis indicated that the autosomal marker L(E021) is present in SYP-E021 indicating that, like TFE3-E021, this clone is chimæric, involving fusion with the same autosomal region as the latter. This suggested that both clones arose from the same YAC species.

Restriction mapping with *Sal*I provided support for this hypothesis, with sites at the same distance from the left end in both clones (Figs. 4.9-4.10; Table 4.4). Localization of SYP and TFE3 within the SYP-E021 YAC map showed that they both lie within ~125kb of the right end (Fig. 4.10), and this conflicts with pulsed field genomic data which indicate a SYP-TFE3 distance of ~400kb (Derry *et al.*, 1994). It therefore seems likely that SYP-E021 has undergone a deletion of material between the two markers.

When this YAC was used in FISH, the majority of the signal was found on 7q34, indicating that only a small proportion of the clone originates from Xp11.2 (C. S. Cooper, personal communication).

4.3.6 Three cosmids containing the SYP locus

Extensive screening of the St. Louis, ICI and Nussbaum libraries detected no further SYP YACs. It was felt that libraries constructed using an alternative vector-host system might have a better representation of clones covering this region (see Discussion). The SYP' probe was therefore used to screen duplicate gridded filters of the ICRF X-specific cosmid library. Three positives were identified; A1236, G0826 and H0227. These clones were streaked out on selective media to give single colonies. Five independent colony-purified cosmid DNA preps from each clone were digested with *Eco*R1 and run on an

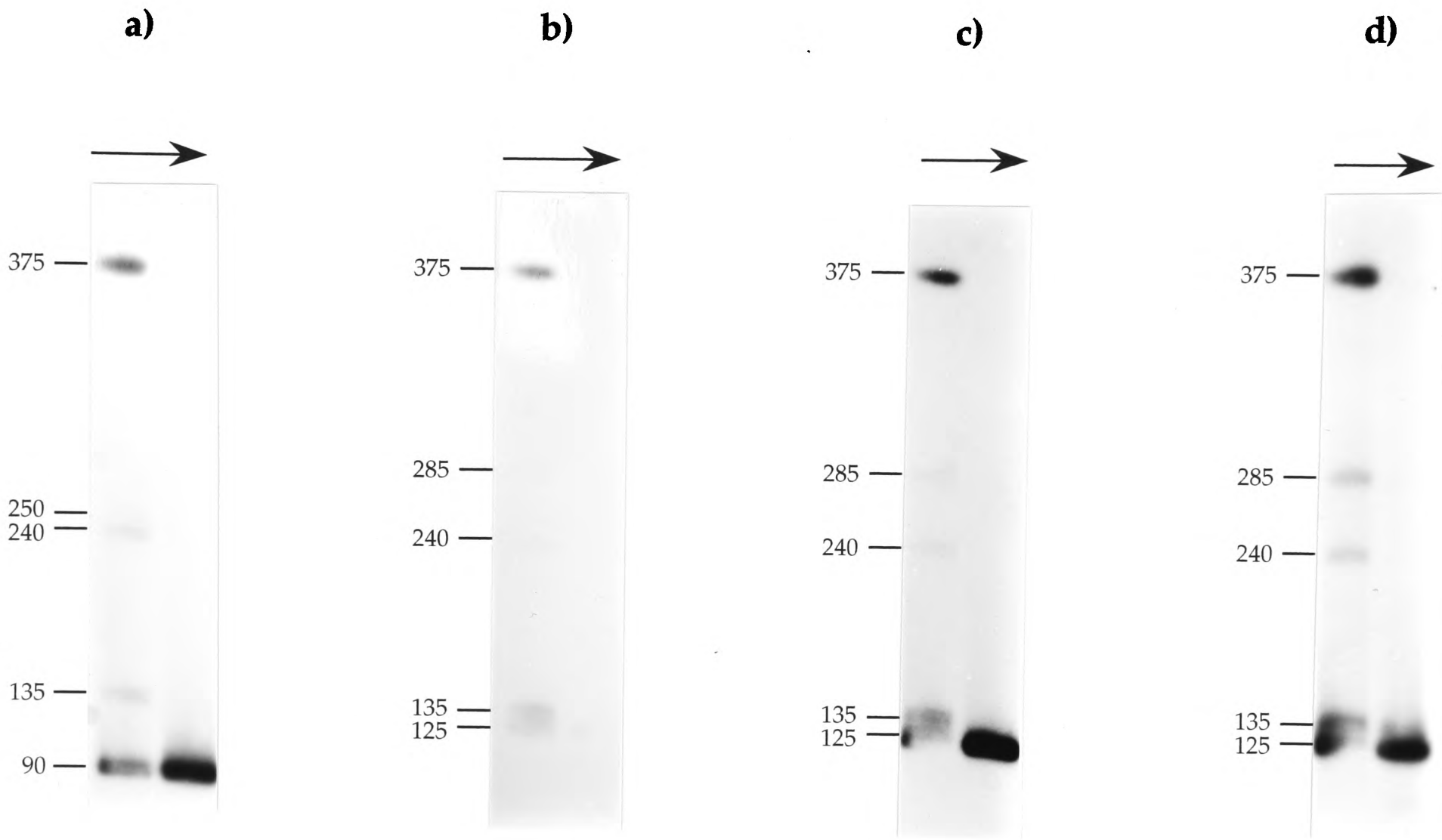


Figure 4.9: Partial digests of the SYP-E021 YAC clone using *Sal*I, probed with **a)** left vector arm, **b)** right vector arm, **c)** SYP and **d)** TFE3. Direction of arrow represents increasing enzyme concentration (from 0.1-15U) with a 1 hour digestion time. Digests were run on standard pulsed field gels (section 2.14.5) with a 27 second switch time and a 34 hour run time. Sizes of bands (summarized in Table 4.4) are given in kilobases. The rare-cutter restriction map which was derived from them is shown in Figure 4.10.

YAC	Left vector arm	Right vector arm	SYP	TFE3
TFE3-E021	90, 135, 240, 250	140, 150, 255, 300	–	140, 150, 255, 300
SYP-E021	90, 135, 240, 250	125, 135, 240, 285	125, 135, 240, 285	125, 135, 240, 285

Table 4.4: Fragment sizes, in kilobases, of bands detected on *Sal*I partial digests of E021 YAC clones, when probed with vector (left and right arms), and the SYP and TFE3 gene markers. The 390 and 375kb fragments corresponding to undigested YACs are not listed. '–' indicates that the SYP marker is not present in TFE3-E021. See Figs. 4.9-4.10.

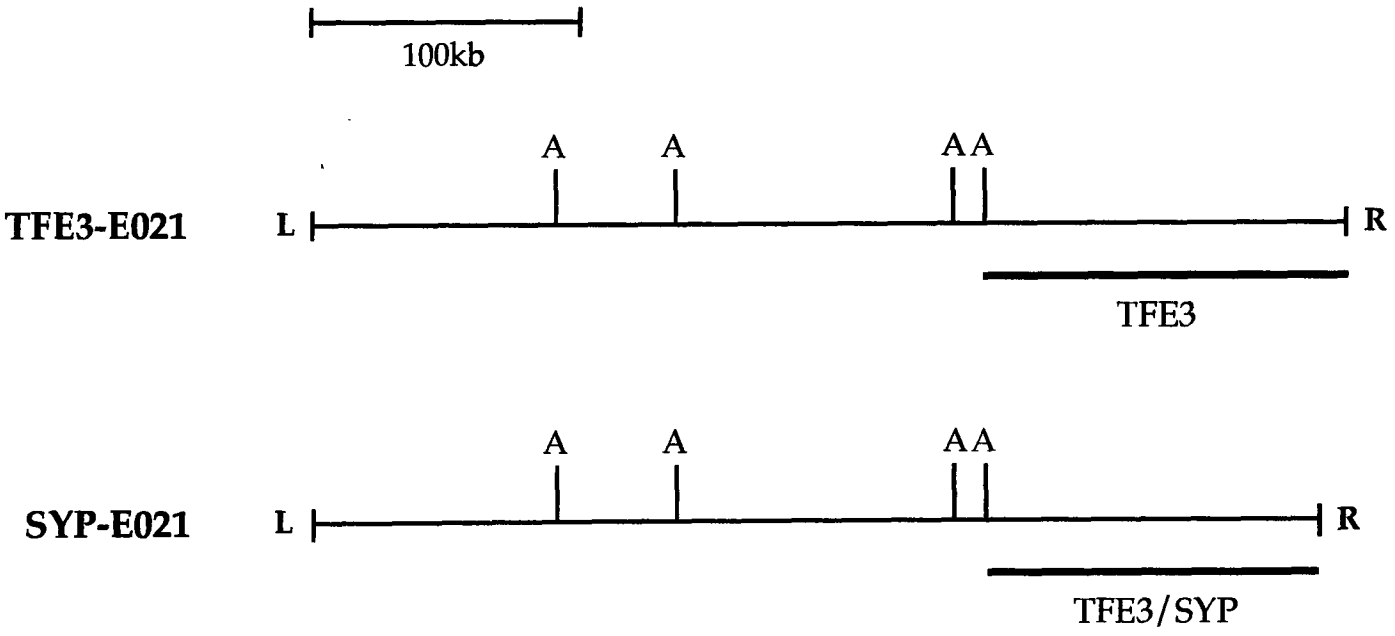


Fig. 4.10: *Sal*I restriction maps of the two E021 YAC clones, derived from fragment sizes in Table 4.4, with the positions of TFE3 and SYP indicated. The clones probably originated from a single larger species, by deletion of different portions from the region between the rightmost *Sal*I site and the right arm. **L**, left arm; **R**, right arm; A, *Sal*I site. Orientation of the YACs with respect to the X chromosome has not been determined.

agarose gel (Fig. 4.11a). A different (but similar) pattern of bands was seen for each prep of the H0227 cosmid, suggesting that this clone is susceptible to rearrangement (Fig. 4.11a). In order to confirm that these cosmids were positive for synaptophysin, a blot of the *Eco*R1 digests was probed with SYP', which detected the cognate 7.7kb band in all three original clones (Fig. 4.11b). This fragment was, however, only present in three of the rearranging H0227 preps.

The *Eco*R1 blot was also probed with a CA oligonucleotide (Pharmacia) in order to detect CA repeats which might be polymorphic, and could therefore represent useful linkage markers in this region. However, no such CA repeats were detected.

This same filter was then hybridized with total human probe, to reveal which of the *Eco*R1 fragments contained repetitive elements (Fig. 4.11c). A ~3.2kb EcoR1 fragment from A1236 (also present in H0227) was not detected and was assumed to be single copy. This fragment (known as SAE) was subcloned into pUC9 and used to probe human genomic and YAC panels. It detected a ~3.2kb band which was X-specific, and present only in the SYP-E021 YAC (Fig. 4.12). Further analysis indicated that, like TFE3 and SYP, it mapped within 125kb of the right arm of SYP-E021. No new YACs were isolated on screening of ICRF, ICI, St. Louis and Nussbaum libraries with SAE.

Partial sequence data from the SAE clone was used to search for homologous sequences in nucleic acid and protein databases. A 135bp open reading frame showing significant similarity to members of a family of L-type (dihydropyridine sensitive) calcium channel genes (Tanabe *et al.*, 1987) was thereby identified at one end of the clone (Fig. 4.13). This ORF is homologous to transmembrane segment IVS6, which is highly conserved amongst members of the calcium channel family. The lack of homology 5' to this 135bp region is likely to be due to the presence of an intron. In support of this, there is a potential 3' (acceptor) splice site (Stephens and Schneider, 1992) adjacent to the boundary at which the similarity ends (Fig. 4.13).

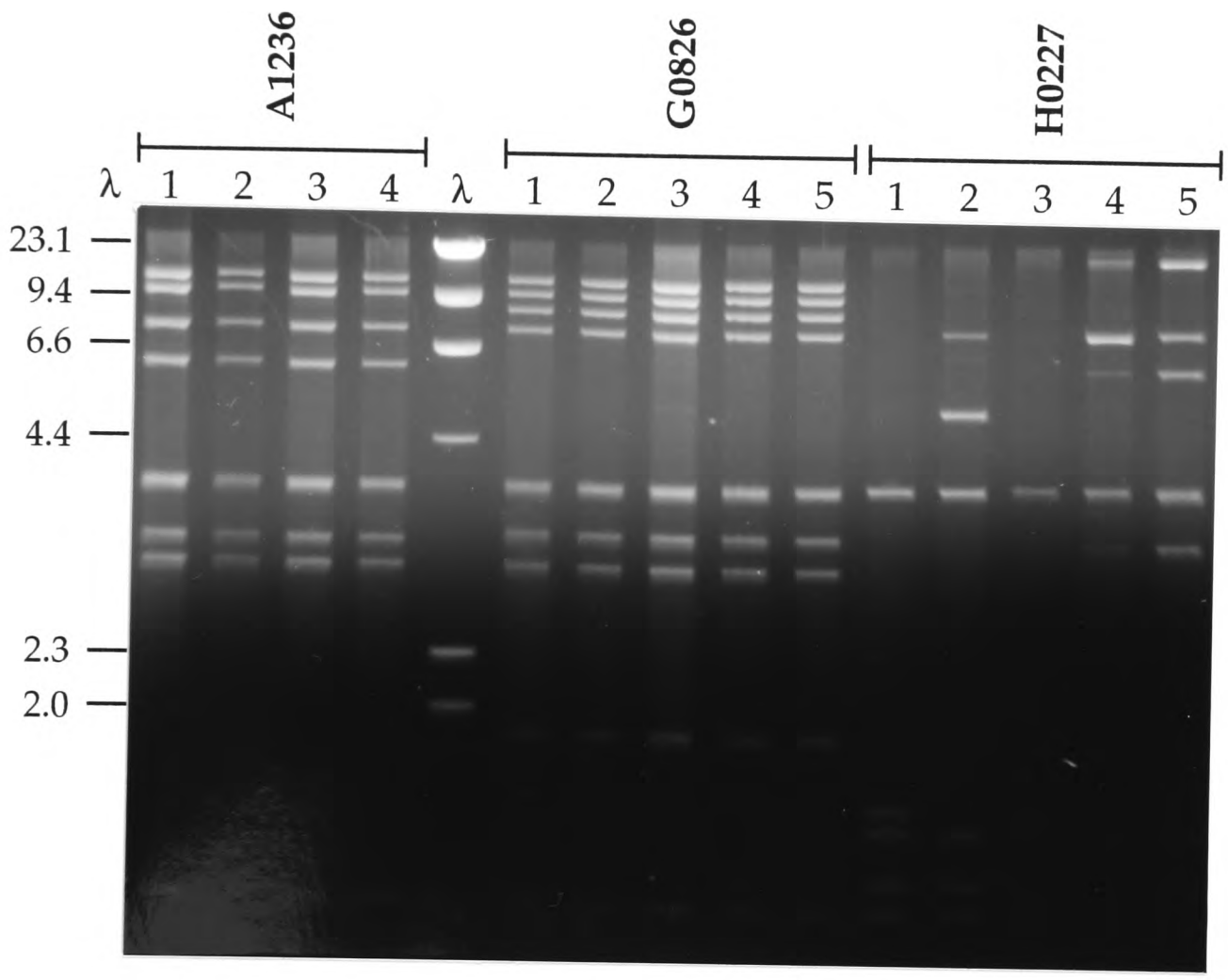
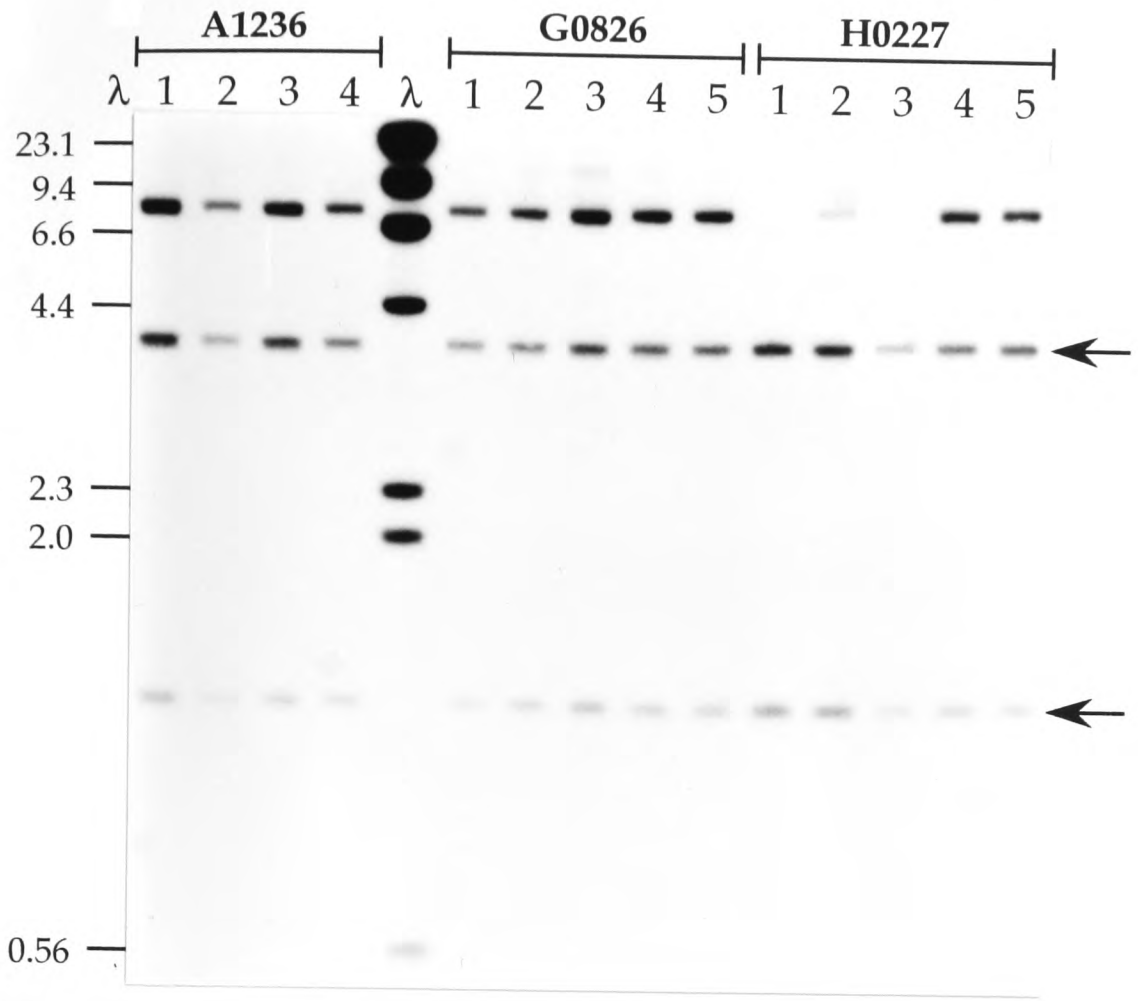


Figure 4.11: Analysis of three cosmids containing the SYP' locus.

a) *Eco*R1 digests of several colony-pure preps of each of the cosmids A1236, G0826 and H0227. Sizes of lambda markers (λ) are given in kilobases. A1236 and G0826 are clonally stable, but different patterns of bands are seen for different preps of H0227, indicating that it has a tendency to rearrange. A band of 7.7kb (the same size as the cognate SYP' *Eco*R1 fragment) is present in preps from each cosmid (see Figure 4.11b)

b)



c)

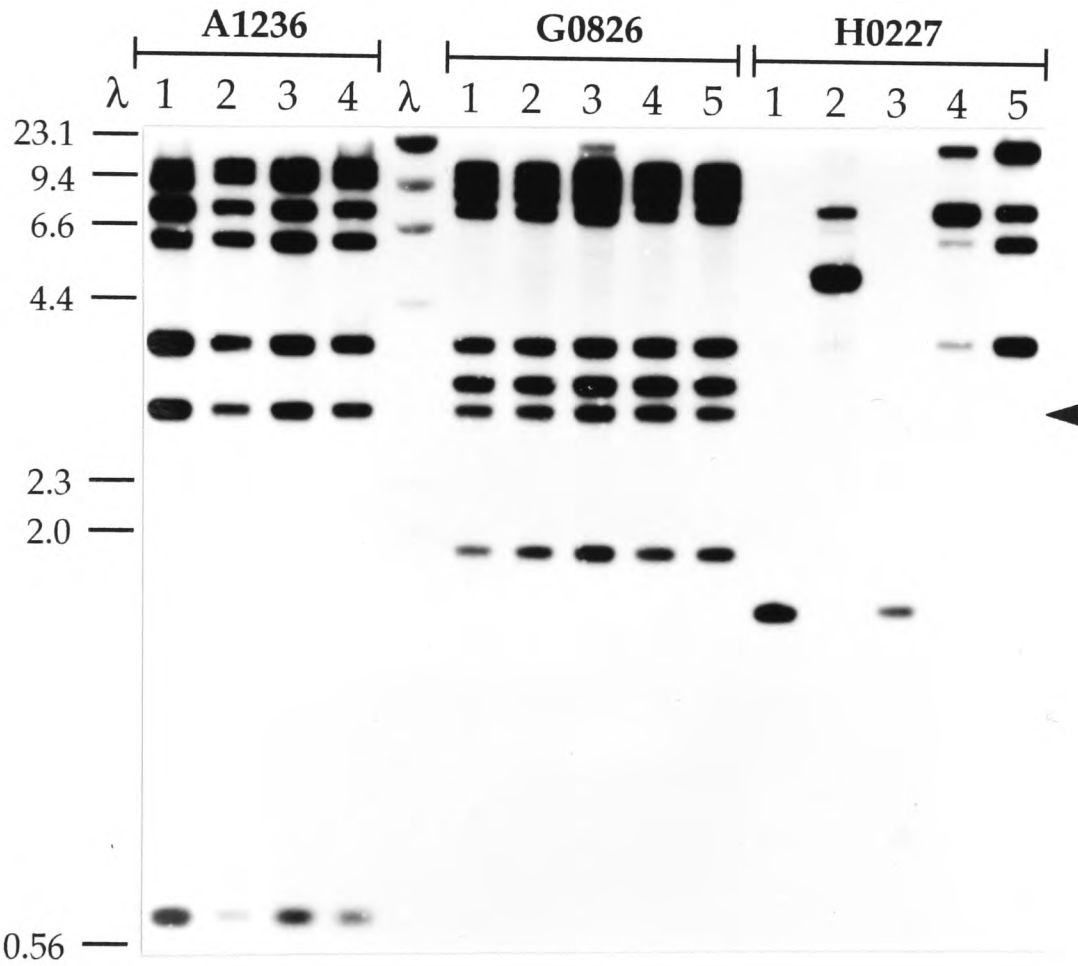
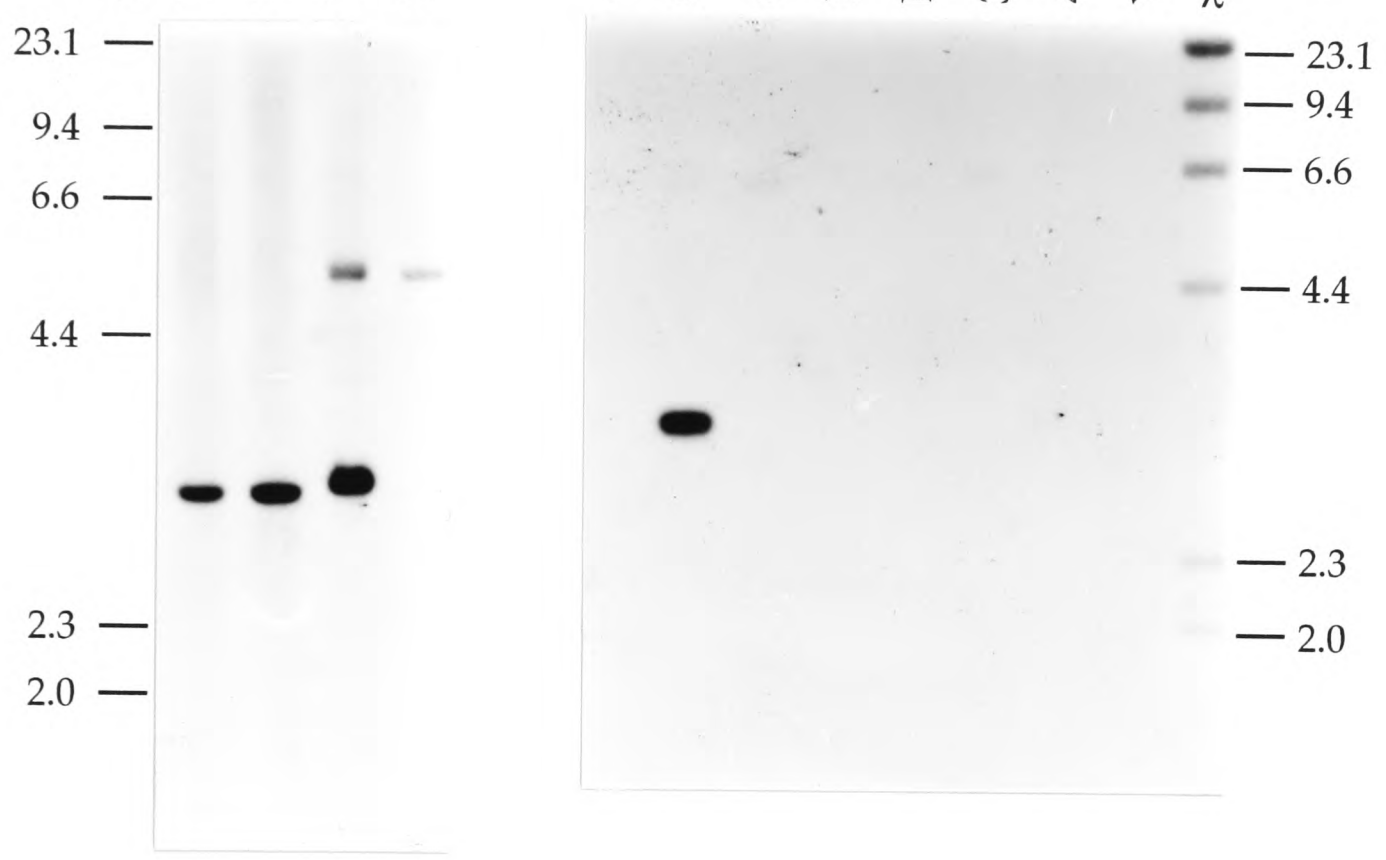


Figure 4.11 (cont): Hybridization studies of A1236, G0826 and H0227, three cosmids containing the SYP' locus. Sizes of lambda markers (λ) are given in kilobases.

b) *Eco*R1-digested colony-pure preps of each cosmid (see Figure 4.11a), probed with SYP'. The cognate 7.7kb fragment is present in all preps of A1236 and G0826, but only in preps 2, 4 and 5 of H0227, which is rearranging. Multiprimed λ/HindIII DNA was included in the hybridization to detect marker bands. This has homology to the cosmid vector and therefore detects vector bands in all tracks (indicated with arrows).

c) The same *Eco*R1-digested cosmid preps, probed with total human DNA. Most of the bands seen in Fig. 4.11a are detected, indicating that they contain repetitive elements. However, a 3.2kb fragment in all preps of A1236 (also present in preps 4 and 5 of H0227) is not detected, suggesting that it is single copy. The position of this fragment (known as SAE) is shown with an arrow.



a) **b)**

Figure 4.12: Hybridization of the SAE probe to *Eco*R1 digests of genomic, hybrid and YAC DNAs. Positions and sizes, in kilobases, of lambda markers (λ) are indicated.

a) SAE detects a 3.2kb band in human male and female genomic DNA, and also in ThyB-X, but not in mouse genomic DNA. The band appears to be slightly larger in ThyB-X, but this is due to overloading of the DNA in this track. In addition, a fragment of 5.5kb is detected in ThyB-X and mouse DNA, suggesting that part of the SAE sequence may be conserved between man and mouse and may therefore originate from a gene. The identification of an open reading frame at one end of SAE supports this view.

b) On probing *Eco*R1 digests of a panel of YACs from the Xp11.23-p11.22 cluster, SAE detects a 3.2kb band in the SYP-E021 clone, but not in TFE3-E021, or any of the GATA YACs. The B102-2 and B102-3 clones are alternative forms of the same YAC, the latter having undergone an ~100kb deletion (see Figure 4.3). SAE is absent from all other YACs described in this thesis (see text).

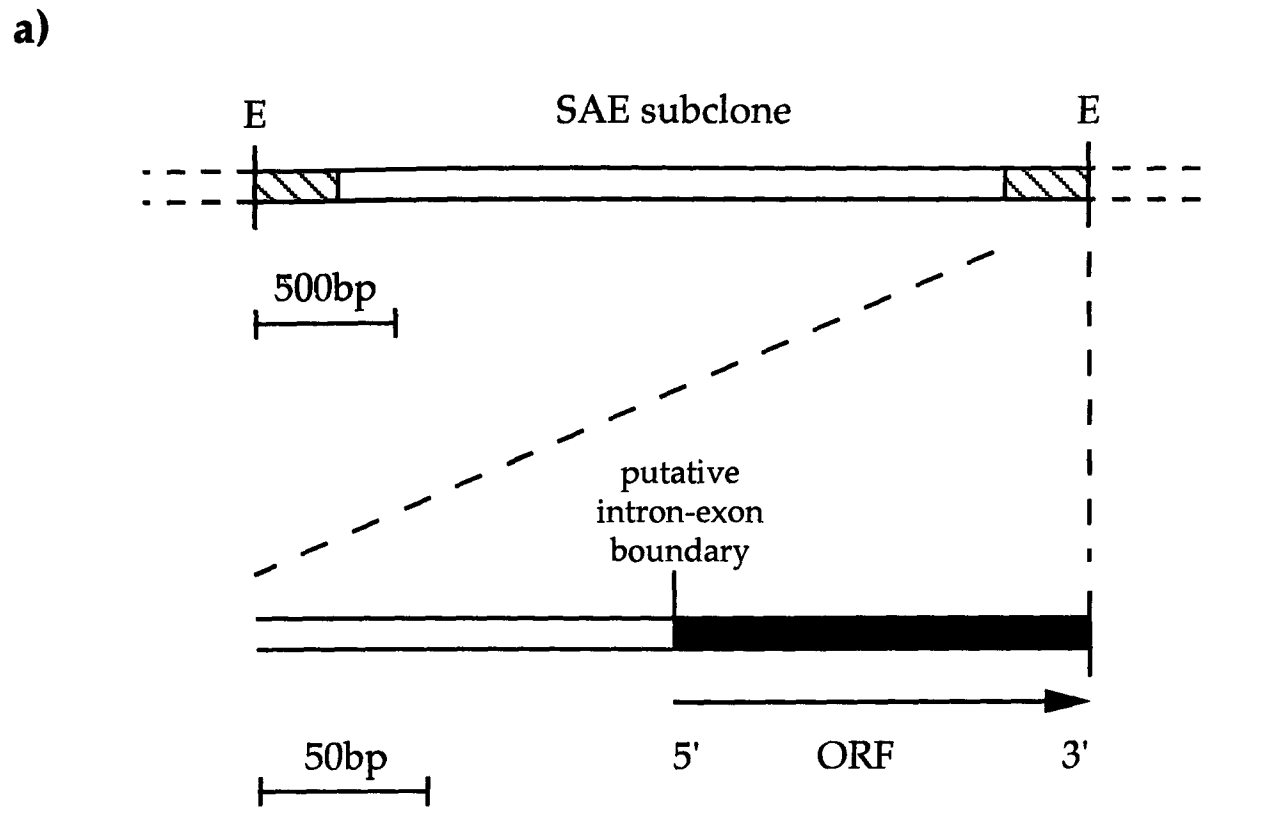


Figure 4.13: An open reading frame (ORF) at one end of SAE shows high homology to calcium channels.

a) Position and orientation of the ORF with respect to SAE subclone. The regions of the subclone which have been sequenced are shaded with diagonal lines in the top part of the diagram. 'E's represent *Eco*R1 sites. The region containing the ORF (shaded with black) is shown, enlarged, beneath this. 5'-3' orientation and the position of the putative intron-exon boundary are indicated.

b) Sequence analysis of the ORF. The DNA sequence of the region from the putative splice site to the *Eco*R1 site is shown. Nucleotides of the splice site which are identical to the consensus (Stephens and Schneider, 1992) are shown in boldface. An arrow indicates the position of the intron-exon boundary. The amino acids which would be encoded by the ORF are given beneath the DNA sequence. Note that the ORF extends up to (and includes) the *Eco*R1 site which was used for subcloning. The amino acid sequence of the corresponding region of a rat kidney calcium channel (Yu *et al.*, 1992) is also shown, aligned beneath the SAE protein sequence, with conserved residues in boldface.

4.3.7 Localization of the Wiskott-Aldrich gene within the contig

A novel gene (WASP) that is mutated in patients with Wiskott-Aldrich syndrome was recently identified by positional cloning from Xp11.23 (Derry *et al.*, 1994). The complete cDNA sequence and genomic structure of the WASP gene have been described (Derry *et al.*, 1994). This information was used to design PCR primers which would amplify regions of the WASP locus from genomic DNA. A ~600bp product from the 5' part of WASP was thereby obtained, and when this was hybridized to a panel of *Eco*R1 digested YACs from the OATL1–GATA–TFE3–SYP contig, it was found to be present in three YACs; F0501, 27GF2 and 3578 (Fig. 4.14). This indicates that WASP maps in the interval between R(3578) and GATA.

4.4 Discussion

The seven YAC clones which were isolated using the GATA, TFE3 and SYP genes could be assembled into a contig, on the basis of overlapping markers, without the need for any chromosome walking (Fig. 4.15). The overall order of genes and novel markers deduced from YAC analysis was found to be Xpter–OATL1–R(3578)–WASP–GATA–L(B102)–TFE3–SYP–Xcen. The R(3578) and WASP markers established a link between the YACs isolated here and the distal OATL cluster.

As previously discussed in Chapter 3, the construction of YAC contigs in a given region may be hindered by problems of chimærism and clonal instability. YAC rearrangment appears to be especially prevalent in clones from the OATL1-SYP part of Xp11.23. For example three out of the five GATA-containing YACs so far isolated from the ICRF, St. Louis and ICI libraries are susceptible to deletions which result in loss of the GATA locus. Instability has also been reported for the OATL1 YACs (Chand, 1994).

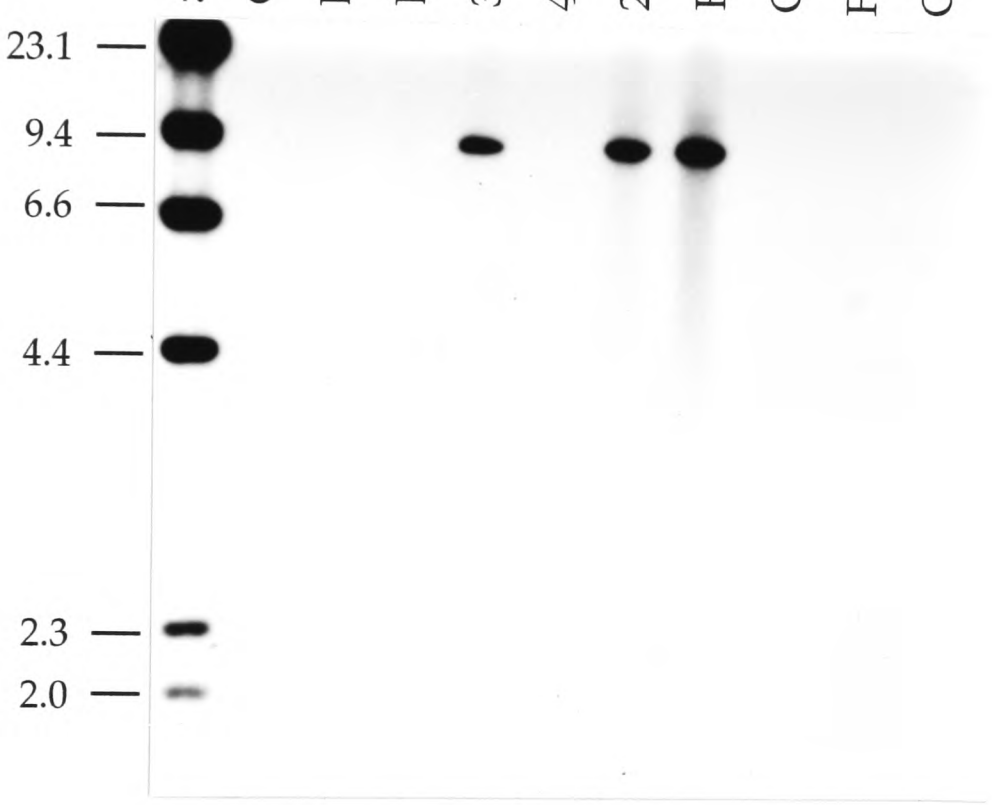


Figure 4.14: Hybridization of a probe from the WASP locus (Tables 4.1 and 4.2) to *Eco*R1 digests of YACs from the Xp11.23-p11.22 contig, including the GATA and OATL1 clusters. Sizes of lambda markers (λ) are given in kilobases. The B102-2 and B102-3 clones are alternative forms of the same YAC, the latter having undergone an ~100kb deletion (see Figure 4.3). A 9.0kb fragment is detected in YACs 3578, 27GF2 and F0501. This places the WASP locus between R(3578) and GATA, as shown in Figure 4.15.

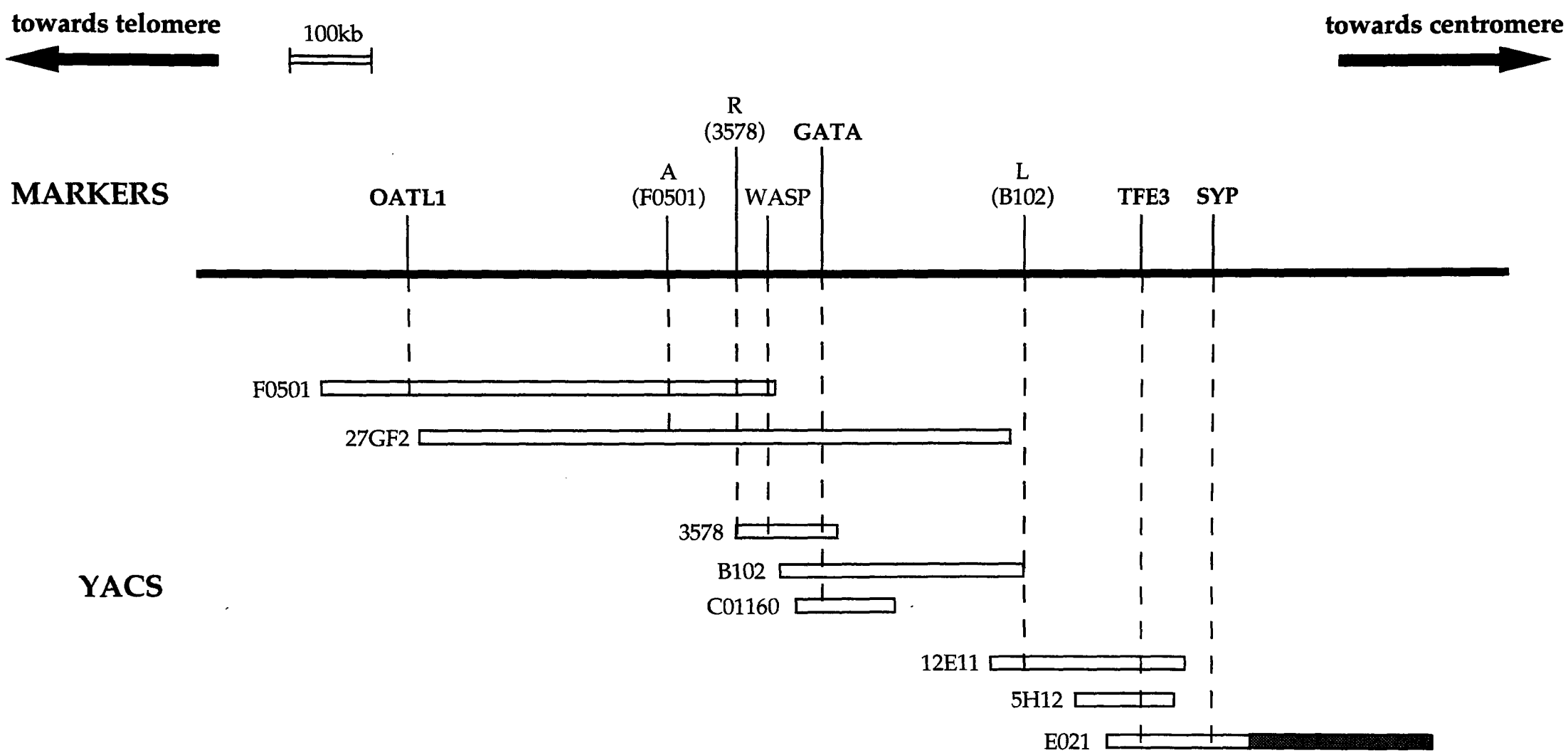


Figure 4.15: YACs and markers in the OATL1–GATA–TFE3–SYP cluster. L(YAC ID), left end clone of YAC; R(YAC ID), right end clone of YAC; A(YAC ID), *Alu*-PCR product isolated from YAC. Autosomal regions of chimæric YACs are indicated by shading. Clones are drawn to scale, but the extent of YAC overlap has not been established for all clones in the contig, hence physical distances between markers is not always accurately represented. Further details of YACs and markers are given in Tables 4.1-4.3.

It is apparent that the region around the SYP gene is particularly unstable in YACs. After screening of four independent libraries by PCR or with a SYP probe, only one positive clone was isolated (SYP-E021), a 375kb YAC, which was found to also contain the TFE3 locus. Further analysis of the clone indicated that it was chimæric and comparison between the rare-cutter restriction map of this YAC and genomic pulsed field maps around SYP and TFE3 suggested that it had rearranged to delete several hundred kilobases of DNA between the loci. A second isolate of E021 was found to contain only the TFE3 locus (TFE3-E021), even though it was larger than the first (~390kb) and another group has reported that E021 is only positive for the SYP locus (Willard *et al.*, 1994). These observations suggest that the clone may have originated as a larger YAC containing SYP and TFE3, and then deleted different portions to yield the different isolates. Derry *et al.* (1994) have reported the isolation of a ~390kb SYP YAC (66G1) from the CEPH library, but FISH analysis has shown that it is chimæric, with the majority of signal (>95%) being found in 3q26 (C. S. Cooper, personal communication).

The low representation of the SYP locus in YAC libraries is likely to result from the local repeat structure of the region, combined with the highly recombinogenic nature of the yeast host. Alternative vector systems, such as cosmids or P1s (Sternberg, 1990), which use bacteria as a host, may therefore provide better opportunities for further analysis of SYP. Preliminary analysis of three cosmids isolated with the 5' part of the SYP locus indicates that this region does indeed display increased stability in this vector system. A novel single copy marker (SAE) was generated from the SYP cosmids and when this was used to screen YAC libraries no new clones could be isolated. This supports the view that the lack of SYP-positive YACs is a consequence of a region-specific instability and not due to some kind of inefficiency in the screening process when using the SYP locus.

The extent of overlap between YACs in the GATA–TFE–SYP cluster has not yet been established, so the size of the contig cannot be precisely determined. Pulsed field genomic mapping has indicated that the GATA-SYP distance is ~550kb (Derry *et al.*, 1994), but as described above, a suspected deletion in the TFE3–SYP linking YAC suggests that the contig does not contain material covering this entire region. The size of the OATL1 YAC cluster has been estimated as ~900kb (A. Chand, personal communication) and combined with data from the sizing of the more proximal non-chimæric YACs, a value of ~1.3-1.4Mb is obtained for the size of the entire OATL1–GATA–TFE3–SYP contig. Analysis of these YACs and the SYP cosmids with the A(E0250) marker isolated from the DXS255–DXS146 contig indicates that there is still an uncloned gap between SYP and DXS255 (see Chapter 3).

One of the initial aims of this project was to provide a basis for the identification of candidate genes for Wiskott-Aldrich syndrome (see Section 1.3.2). Recently, Derry *et al.* (1994), used affinity capture to isolate cDNAs mapping within YAC and cosmid clones from contigs in Xp11.23. Point mutations in one of the genes identified (WASP) were found to be associated with both the Wiskott-Aldrich phenotype and a similar disorder known as X-linked thrombocytopaenia (Villa *et al.*, 1995). WASP is expressed in lymphocytes, spleen and thymus, and encodes a 501 amino acid protein with a high percentage of proline, a potential nuclear localization signal and two highly charged domains; a basic decapeptide sequence (HHHRHHRHRR) and a very acidic C-terminus (Derry *et al.*, 1994). Several different transcription factors have been found to have proline-rich domains and acidic regions which are similar to the C-terminus of WASP. Alternatively, the proline-rich motifs of WASP may bind to SH3 (Src homology 3) domains in intracellular signalling proteins. Thus, although the precise role of WASP has not yet been determined, it is likely to be a key regulator of lymphocyte and platelet function. Although Derry *et al.* position the WASP locus between GATA and TFE3, analysis of the contig presented here indicates that it maps to the OATL1-GATA interval.

Two novel genes have recently been isolated, from chromosomes X and 18, which are disrupted by the t(X;18) (p11.2;q11.2) translocation associated with synovial sarcoma. (Clark *et al.*, 1994). These loci, known as SSX and SYT, were demonstrated to form a chimæric gene, encoding a fusion product, in the majority of SS tumours samples that were analysed (Clark *et al.*, 1994). SSX sequences have been found to map to both the OATL1 and OATL2 clusters, providing an explanation for the observed heterogeneity of the translocation breakpoints (Clark *et al.*, 1994; Chand, 1994).

As described in Chapter 1, a specific reciprocal translocation, t(X;1) (p11.2;q21.2), is associated with several cases of papillary renal cell carcinoma (RCC). The contigs described in this thesis have provided a useful resource for obtaining a more precise localization of the Xp11.2 breakpoint. At present, it appears, on the basis of FISH analysis using YACs and cosmids, that the breakpoint is not associated with the locus implicated in Dent's disease (see Chapter 1), but maps in the ~400kb interval between TFE3 and SYP (C. S. Cooper, personal communication). Further studies of the region will be necessary to aid the identification of genes which may be implicated in the ætiology of this tumour.

Preliminary analysis of the SAE subclone isolated from the SYP cosmids suggests that it contains part of a novel gene encoding the alpha 1 subunit of a calcium channel (Tanabe *et al.*, 1987). More detailed study of the cosmids should facilitate the identification of additional sequences from this putative gene, which in turn will enable the corresponding transcripts to be isolated. As yet, it is not known in which tissues, if any, this gene is expressed, but it is interesting to note that kidney-specific calcium channels of this type have been identified in the rat (Yu *et al.*, 1992). Given that this putative gene maps in the vicinity of SYP, it therefore represents a potential candidate for the X-linked locus which is disrupted in renal cell carcinoma.

Chapter 5 – Isolation and characterization of a candidate gene for Dent's disease (CLCN5)

5.1 Introduction

5.1.1 Dent's disease

i) Description of phenotype

In 1964, Dent and Friedman reported two unrelated male patients suffering from rickets who also showed symptoms of renal tubular disease such as low molecular weight (LMW) proteinuria and hypercalciuria. They initially described this syndrome as 'hypercalcuric rickets'. The progress of these patients was monitored in the years following the original diagnosis; both experienced a gradual decline in kidney function leading to end-stage renal failure, and one developed nephrocalcinosis.

Over the last 31 years, a further 23 affected patients have been discovered in Britain who appear to be manifesting this same disorder, which is now referred to as 'Dent's disease' (Wrong *et al.*, 1994). The syndrome, which is usually familial, affects the two sexes in approximately equal numbers, but is more severe in males, in whom it is characterized by LMW proteinuria, hypercalciuria, nephrocalcinosis, kidney stones and end-stage renal failure. There is a certain degree of variation in the symptoms displayed by affected males, even in different members of the same pedigree, as shown in Table 5.1. For example, whilst all affected males present with LMW proteinuria, and most develop nephrocalcinosis and/or renal failure, metabolic bone disease (rickets/osteomalacia) has only been found in a minority of patients (Wrong *et al.*, 1994).

Urine analysis, renal radiography and ultrasound examination of individuals from pedigrees segregating Dent's disease revealed that renal or urinary abnormalities in females were much milder than in males (Table 5.2). All affected females have LMW proteinuria, and most have hypercalciuria, but only one has so far been found with nephrocalcinosis and renal failure (Wrong *et al.*, 1994).

Patient	Year of birth	LMW Proteinuria	Hypercalciuria	Nephro-calcinosis*	End-stage renal failure*	Renal stones	Rickets*	Amino-acidurea	Hypo-phosphataemia	Acidification defects
A/II/3	1928	+	undet.	63	63	–	–	+	–	undet.
A/III/3	1953	+	+	13	36	–	5	+	+	–
B	1940	+	+	–	48	–	2	+	+	–
C/II/3	1922	+	–	33	65	+	–	+	–	–
C/II/6	1926	+	undet.	56	56	–	–	+	–	+
C/III/2	1950	+	+	34	–	+	–	+	–	+
C/III/6	1954	+	+	29	37	–	15	+	+	+
C/IV/2	1973	+	+	–	–	–	–	+	–	undet.
D/II/1	1951	+	+	13	38	+	10	+	–	–
D/III/2	1991	+	+	–	–	–	–	+	–	undet.
E/II/1	1961	+	+	–	32	+	–	+	+	–
EII/2	1964	+	+	11	–	–	2	+	+	+
F/II/1	1957	+	+	21	–	+	–	+	+	+
G	1944	+	+	43	–	+	–	+	–	+
H	1947	+	+	25	35	+	31	+	–	+
		15/15	12/13	11/15	9/15	7/15	6/15	15/15	6/15	7/12

Table 5.1: Clinical details of males suffering from Dent's disease. Presence (+) or absence (-) of each aspect of the phenotype is indicated. The presence of hypercalciuria and/or acidification defects is undetermined (undet.) for some patients. The proportion of examined male patients who manifest each symptom are given at the bottom of the table. *Age of onset is shown for cases of nephrocalcinosis, end-stage renal failure and rickets. The data presented in this table, which demonstrate the phenotypic variability of the disorder, were obtained from Wrong *et al.* (1994).

Patient	Year of birth	LMW Proteinuria	Hypercalciuria	Nephro-calcinosis*	End-stage renal failure*	Renal stones	Rickets	Amino-acidurea	Hypo-phosphataemia	Acidification defects
A/I/1	1906	+	undet.	59	59	–	–	undet.	undet.	undet.
A/II/5	1930	+	–	–	–	–	–	-	–	undet.
C/II/2	1921	+	mild	–	–	–	–	mild	–	undet.
C/II/4	1925	+	–	–	–	–	–	mild	–	–
C/II/8	1927	+	mild	–	–	–	–	–	–	undet.
C/II/9	1932	+	mild	–	–	–	–	–	–	–
C/III/1	1949	+	+	–	–	–	–	–	–	undet.
D/II/2	1956	+	–	–	–	–	–	–	–	undet.
D/III/3	1991	+	+	–	–	–	–	–	–	undet.
E/I/2	1921	+	-	–	–	–	–	–	–	undet.
		10/10	5/9	1/10	1/10	0/10	0/10	2/9	0/9	0/2

Table 5.2: Clinical details of affected females from Dent's disease. Presence (+) or absence (-) of each aspect of the phenotype is indicated. In certain cases, patients have been classified as only mildly affected with a particular symptom. The presence of hypercalciuria, aminoacidurea, hypophosphataemia and/or acidification defects is undetermined (undet.) for some females. The proportions of examined female patients who manifest each symptom are given at the bottom of the table. *Age of onset is shown for the female affected with nephrocalcinosis and end-stage renal failure. The data presented in this table, which demonstrate the reduced severity of Dent's disease in affected females, were obtained by Wrong *et al.* (1994).

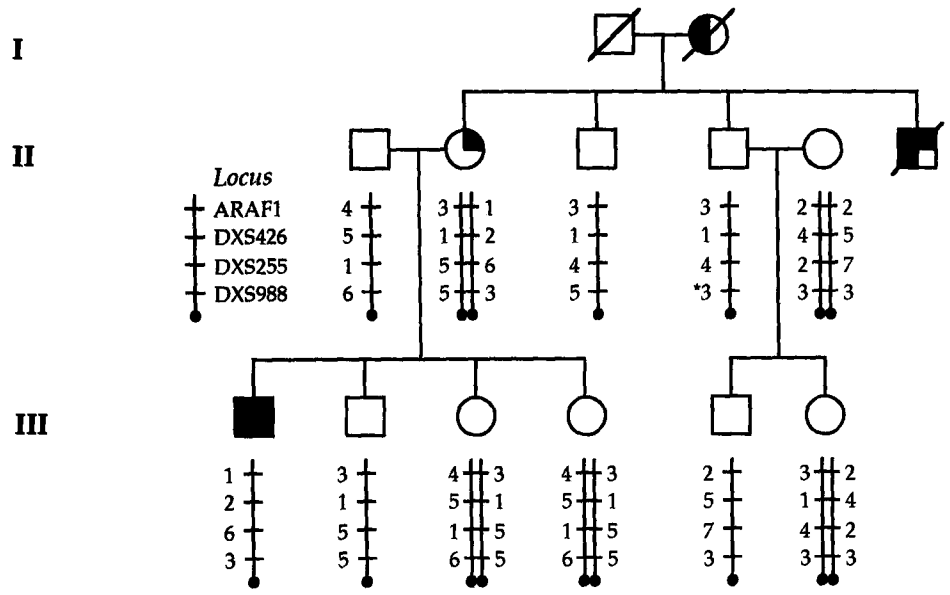
Most of the renal abnormalities seen in Dent's disease (LMW proteinuria, aminoaciduria, hypophosphataemia, acidification defects, rickets and progression to end-stage renal failure) are well recognized features of the Fanconi syndrome, a collective term describing multiple defects of the proximal tubule. However, hypercalciuria is only rarely found as a feature of the Fanconi syndrome. Furthermore, Dent's disease differs in that it is commonly associated with nephrocalcinosis and renal stones, which are almost unknown as complications of other forms of Fanconi syndrome (Wrong *et al.*, 1994).

ii) Linkage analysis

Of the 25 patients studied by Wrong *et al.* (1994), 22 (12 males and 10 females) were from five unrelated, non-consanguineous families, with up to three generations involved. There was no definitive evidence of a familial basis for the disorder in the remaining three affected males. As stated above, male patients are much more severely affected, and it was initially proposed that the inheritance pattern of the disease in the five pedigrees was the result of autosomal dominant transmission with greater penetrance in the male (Wrong *et al.*, 1990). This suggestion was based on anecdotal evidence of renal stones in unexamined males from three of the families, which appeared to support father-to-son inheritance, and therefore exclude X-linkage. However, the incidence of renal stones in the male population is high (~12%) (Smith, 1989), and so the supposed observations of father-to-son inheritance may have arisen by chance. In fact there is an absence of male to male transmission in the examined members of all five pedigrees (Pook *et al.*, 1993), and the alternative hypothesis of X-linkage was therefore considered.

Linkage analysis of pedigrees A and C (both families in which the disease occurred in three generations) using polymorphic markers from the X chromosome (Figure 5.1) suggested that the gene responsible for Dent's disease mapped in the pericentromeric region of Xp, between the markers MAOB and DXS988, and in the vicinity of ARAF1, DXS426 and DXS255 (Pook *et al.*, 1993).

Family A



Family C

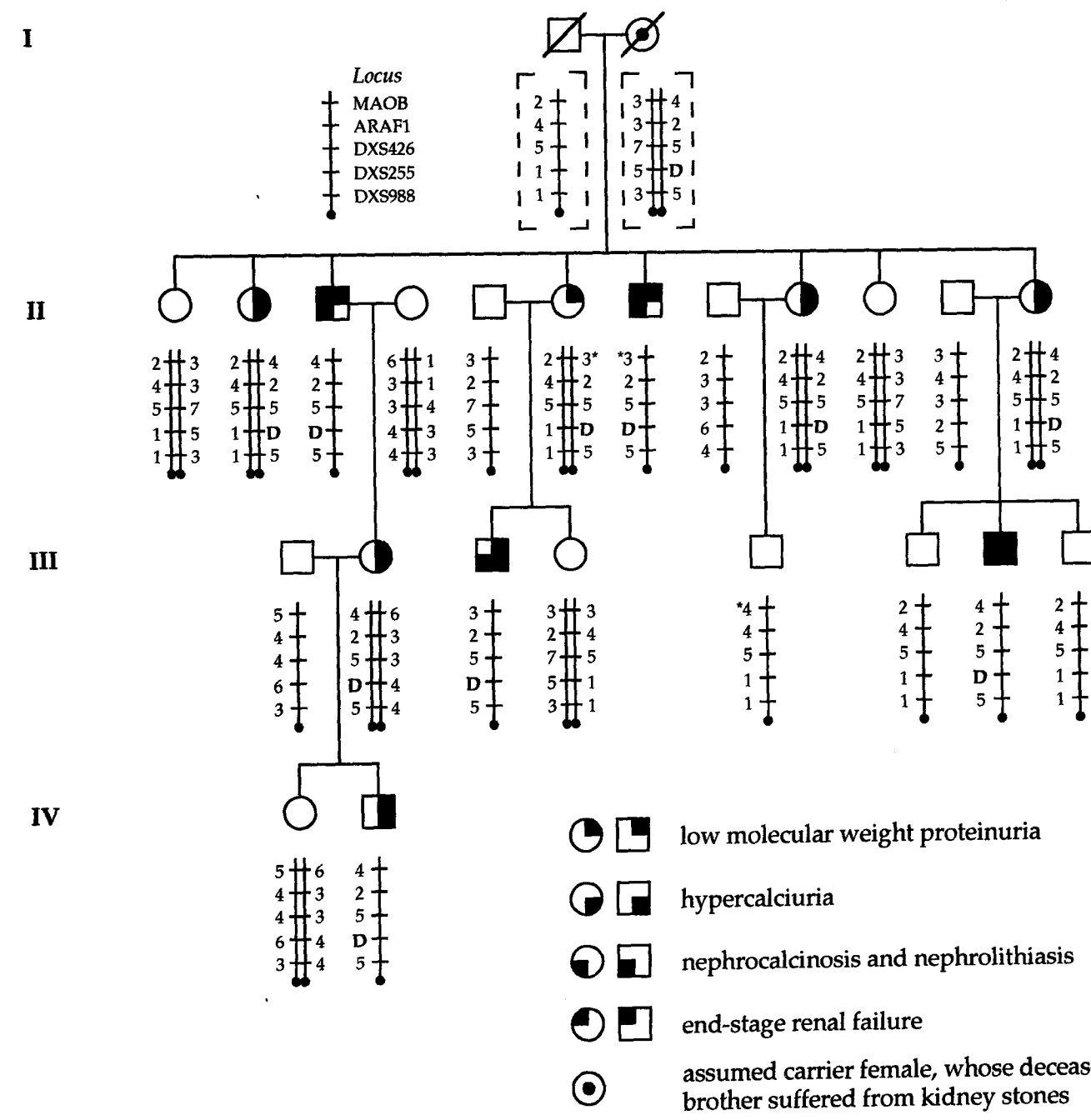


Figure 5.1: Pedigrees from families segregating Dent's disease. A key is given, indicating the symbols used to represent the different aspects of the disease phenotype from which each individual suffers. The relevant results of haplotype analysis with markers from Xp are shown (Pook *et al.*, 1993). The alleles are indicated by numbers, and the microdeletion detected at DXS255 is indicated as '**D**'. Deduced genotypes are shown in brackets and recombinants between Dent's disease and each allele are indicated with an asterisk. The minimum number of total recombinants in these pedigrees is obtained by locating Dent's disease between MAOB and DXS988.

iii) Identification of an associated microdeletion involving DXS255

Analysis of individuals from family C with the probe M27β, which recognizes the hypervariable locus DXS255, revealed a deletion involving this locus in all affected members (Figure 5.1). Further studies indicated that this deletion maps within a ~4cM interval, flanked on the proximal side by DXS988 and on the distal side by the genes TFE3 and SYP. No cytogenetic abnormalities were observed on analysis of high resolution chromosome preparations from deleted individuals. The phenotype of Dent's disease in family C is similar to that observed in the other pedigrees (see Table 5.1). In particular, patients are not mentally retarded, nor do they show evidence of other clinical disorders, suggesting that the deletion does not result in a contiguous gene syndrome (Pook *et al.*, 1993).

5.1.2 Syndromes of similar phenotype, mapping to Xp11.2

i) X-linked recessive nephrolithiasis (XRN)

Fryomyer *et al.* (1991) described a large kindred from New York, consisting of 162 family members from six generations, in which nine males were found to have LMW proteinuria, hypercalciuria, nephrocalcinosis, renal stones and end-stage renal failure. However, these patients lacked bone disease and urinary acidification defects and obligate heterozygote females were reported to be unaffected. This disorder, known as X-linked recessive nephrolithiasis, was therefore initially classified as an independent clinical entity from Dent's disease (Fryomyer *et al.*, 1991). It now seems likely that the two disorders are in fact due to defects in the same gene, for the following reasons:

- As described above, variations in phenotype are found between different male patients manifesting Dent's disease, even within the same pedigree (Table 5.1), suggesting that other factors (genetic or environmental or both) are modifying the phenotype. Only a minority of Dent's patients suffer from bone disease (Wrong *et al.*, 1994). The complete absence of rickets from all affected members of the New York pedigree could partly be due to the higher levels of circulating vitamin D metabolites that are known to be present in US subjects as compared to those in the UK.

- Defective urinary acidification can occur as a secondary consequence of nephrocalcinosis of many different origins, and appears to depend on the severity and duration of renal calcification (Wrong *et al.*, 1994).

- The LMW proteinuria and hypercalciuria of almost all heterozygotic females affected with Dent's disease is very mild in comparison to their male counterparts. It is possible that such renal abnormalities have gone undetected in the obligate carriers of the XRN pedigree. Only one female has been found to be severely affected with Dent's disease, and this could be explained by chance skewing of the X-inactivation pattern in her renal tissue, so that the functional copy of the gene responsible was preferentially inactivated (Wrong *et al.*, 1994).

- Linkage studies of the XRN pedigree have indicated that, like Dent's disease, it maps close to DXS255 in proximal Xp (Scheinman *et al.*, 1993 and see Section 1.3.2),

ii) X-linked recessive hypophosphataemic rickets (XLRH)

Bolino *et al.* (1993) reported an Italian four generation pedigree, in which five males were affected with LMW proteinuria, hypercalciuria, rickets/osteomalacia, hypophosphataemia, nephrocalcinosis and progressive renal failure. They described this as a new form of X-linked recessive hypophosphataemic rickets (XLRH). Although five out of the six (otherwise unaffected) obligate carrier females in this pedigree had hypercalciuria, they attributed this to the segregation of another gene, independent of that causing the XLRH. They suggested, on the basis of an inner ear defect in one patient, that XLRH may be the human equivalent of the *Gyro* phenotype in mouse. The latter is characterized by small stature, rickets, hypophosphataemia, inner ear abnormalities and hyperactivity.

However linkage analysis showed that instead of mapping in Xp22.1-p22.2, the expected region for the *Gyro* homologue, the disorder in this pedigree is localized to Xp11.2 (Bolino *et al.*, 1993). Close linkage was shown to relatively uninformative polymorphisms at the OATL1, SYP, TFE3 and DXS146 loci, but the highly hypervariable minisatellite DXS255 was omitted from these studies. Instead recently identified microsatellite markers were used which have not yet been anchored to the physical map, but are localized to Xp11.2 on the basis of linkage.

Despite the phenotypic similarities and the observation that both disorders map to Xp11.2, Bolino *et al.* suggest that whilst Dent's disease is a generalized proximal tubular defect, the disorder in the Italian family is specific to phosphate reabsorption, and XLRH is therefore a separate entity.

5.1.3 Aims

The goal of this part of the project was to use a positional cloning strategy (Collins, 1992) to isolate kidney-specific transcripts mapping within the microdeletion associated with Dent's disease in Xp11.22 which might therefore be implicated in the ætiology of the disorder. The results of cytogenetic studies, and the phenotypic similarity between deleted and non-deleted patients suffering from Dent's disease suggested that the disorder in family C was due to deletion of a single major locus (section 5.1.1iii). Thus, it might be expected that any kidney-specific gene which was found to map within this deletion would represent a good candidate for the disorder. However, prior to this study, there were insufficient physical mapping data on the extent of the deletion, which was localized to somewhere within a ~4cM interval flanked by TFE3/SYP and DXS988 (Pook *et al.*, 1993). It was therefore important to define the limits of the deletion more precisely.

As described in Chapter 3, a YAC contig spanning DXS255 had already been constructed and characterized, and this provided an excellent starting point for the study. Several novel markers had been generated from the region around DXS255 and localized within rare-cutter restriction maps of the YACs, and these could therefore be used for further analysis of the deletion.

The aim was then to search for coding sequences in those YAC clones which overlapped the deletion. Of the various techniques available for identification of transcripts from a large region cloned in a YAC, the method of choice was hybridization screening using a purified YAC probe (Wallace *et al.*, 1990) of a kidney-specific cDNA library. This was favoured over direct selection of cDNAs using immobilized YACs (Parimoo *et al.*, 1991; Lovett *et al.*, 1991) mainly because other members of the Oxford Genetics Laboratory had previous experience of the hybridization approach. A third strategy considered was to subclone the CpG island (Bird, 1987) regions identified in the YAC contig. However, transcripts isolated using the latter approach are not necessarily expressed in the desired tissue. In addition, since only ~40% of tissue-specific genes are associated with CpG islands (Larsen *et al.*, 1992), it seemed quite possible that the gene responsible for Dent's disease, which would be likely to show a tissue-specific expression pattern, may be undetected.

5.2 Materials and methods

5.2.1 Probes

The isolation of probes L(F1001) and L(6129) is described elsewhere in this thesis. L(G0201) and L(F1101) are markers generated by plasmid rescue from a previously characterized YAC contig around DXS146 (Hatchwell, 1994).

5.2.2 Pulsed field gel purification of the 6129 YAC

Undigested DNA from sixteen concentrated YAC plugs was run on a pulsed field gel (Section 2.14.5) using the following parameters: 1.5% (w/v) agarose (Sigma TypeI)/0.5 x TAE; 16°C; 3.6 V/cm; 13 second switch time; 30 hour run time. The gel was visualized under UV, the YAC fragment excised and purified by the Geneclean (Bio 101) procedure (Section 2.8). This protocol resulted in shearing of the 185kb YAC into small pieces, but this did not matter, since the purpose of purification was to provide material suitable for multiprime labelling.

5.2.3 cDNA library

A human kidney "5' stretch" cDNA library, constructed using mRNA prepared from a 20 year-old Caucasian female, was obtained from Clontech (HL1123n). This library was made by a combination of oligo(dT) and random priming and cloned into the *Eco*R1 site of the Lambda Max1 phage vector. Denaturation of source mRNA with methylmercuric hydroxide was used to release secondary structures, and the library should therefore have a greater representation of 5' sequences than regular libraries. The manufacturers estimated the number of independent clear clones as 1.5×10^6, and found insert sizes to range from 0.6-4.0kb, with an average of 1.5kb. The host bacterial strain is *E. coli* K802. 4×10^5 recombinants were plated out on four 22 x 22 cm plates, and replica plaque lifts prepared from each, as described below.

i) Plating out

Initially the library was plated out at different concentrations on small test plates to determine its titre (~1×10^{10} pfus/ml). The protocol for plating out on 22 x 22 cm plates was modified from the library manufacturer's instructions (Clontech) and is as follows:

1. Streak 5µl of K802 from a 25% glycerol stock onto an LB agar plate containing 0.2% maltose. Incubate at 37°C overnight. This master plate can then be stored at 4°C for up to two weeks.

2. Pick a single, isolated colony from the master plate, and streak onto another LB agar/0.2% maltose plate. Incubate at 37°C overnight. This primary working plate is stored at 4°C.

3. Pick a single, isolated colony from the primary working plate and inoculate 15ml of LB broth containing 10mM $MgSO_4$ and 0.2% maltose. Incubate at 37°C overnight.

4. Serial dilutions are used to prepare a sample of the phage library which will contain 1×10^6 pfu/ml.

5. Add 100µl of this diluted sample to 1.9ml of bacterial culture from step 3. Incubate, with shaking, at 37°C for 15 minutes.

6. Add 28ml of melted LB/$MgSO_4$ soft top agarose (preheated in a 50ml Falcon tube to 45°C), rapidly mix, and pour onto a 22 x 22 cm LB/$MgSO_4$ bottom agarose plate which has been prewarmed in a 37°C incubator for several hours. Swirl plates quickly after pouring to allow for even spreading of the soft top agarose over the entire plate.

7. Leave plates to dry for 30 minutes.

8. Invert plates, and incubate at 37°C for 6-8 hours. Plaques should be in contact with each other, but there should not be confluent lysis.

9. Store plates at 4°C overnight. (Note that there may be further plaque growth for a short while after the plates have been removed from the 37°C incubator, since they may take some time to cool down.)

ii) Protocol for phage lifts

1. Select a Hybond-N+ membrane (Amersham) of the appropriate size and mark the edges using a laundry marker to enable correct orientation of filter once positives have been identified.

2. Place the membrane carefully on to the agar surface, with the orientation marks facing down. If the plate is then inverted, the marks can be seen through the agar, and a waterproof marker is used to copy the marks onto the bottom of the plate.

3. Remove the membrane after 1 minute and place it, plaque side up, on a pad of absorbent filter paper (Whatman) soaked in denaturing solution. Leave for 7 minutes.

4. Transfer membrane, plaque side up to a pad of filter paper soaked in neutralizing solution. Leave for 3 minutes, then repeat with a fresh pad of neutralizing solution.

5. Wash filter in 2 x SSC and transfer to dry filter paper. Leave to air dry.

6. Place membrane, plaque side up, on a pad of filter paper (2-3 pieces thick) soaked in 0.4M NaOH. Leave to fix for 5-30 minutes.

7. Rinse the filter by immersion in 5 x SSC with gentle agitation for 1 minute.

8. Filters can be air dried and stored at -20°C, or used directly for hybridization.

For each plate, a second lift was done, so that duplicate filters could be probed. The second filter was left for 2 minutes (instead of 1) on the agar, and a different coloured waterproof pen was used to copy the orientation marks onto the bottom of the plate. The membrane was then treated as in steps 3 to 8.

5.2.4 Hybridization conditions for screening of library with the 6129 YAC

Purified YAC DNA was multiprime labelled for 6 hours at 37°C with 200µCi [α-^{32}P]dCTP. Prereassociation of probe was carried out in 0.125M Na_2HPO_4 (pH 7.2), containing 1µg/µl of denatured sonicated human placental DNA for 3 hours at 65°C (see 2.9.3). Primary lift filters were prehybridized overnight in Church buffer in the presence of 100µg/ml sonicated human placental DNA. Hybridization was carried out overnight in fresh buffer lacking placental DNA. Filters were washed to a final stringency of 20mM phosphate.

5.2.5 Preparation of DNA from bacteriophage lambda

There are commercially available kits (Promega) which give a fast, reliable method for the purification of lambda DNA from plate or liquid culture lysate. All lambda DNA purified in this thesis was prepped using the liquid lysate method:

i) Production of liquid lysate

1. Using a cut disposable pipette tip, pick a single positive phage plaque from an agar plate. Gently expel the agar plug into a 1.5ml Eppendorf tube containing 100µl of lambda dilution buffer. Place at 4°C overnight.

2. Start a fresh culture of the appropriate *E. coli* host strain (in this case K802) by inoculating a single colony into 5ml of LB medium supplemented with 50µl of 1M $MgSO_4$. Incubate overnight in a shaking incubator at 37°C.

3. Add 500µl of this overnight culture to an Eppendorf tube containing 10-20µl of phage plaque eluate from step 1. Shake at 37°C for 20 minutes.

4. Transfer 50µl of this infected culture to a 10ml culture of LB medium, prewarmed to 37°C and supplemented with 100µl of 1M $MgSO_4$. Shake at 37°C until lysis occurs. This usually takes 6 hours. (Cell debris may be visible in the lysed culture.) Add 50µl chloroform.

5. If the culture has not lysed after 7 hours, add 50µl chloroform anyway, and continue to shake for an extra 15 minutes.

6. Centrifuge the lysate at top speed in a bench centrifuge for 10 minutes to pellet cell debris. Transfer supernatant to a sterile 50ml falcon tube. This lysate can be stored at 4°C for up to six months if desired.

ii) Removal of lambda phage coat

1. Add 40µl of 'nuclease mixture' (Promega; 0.25mg/ml RNase A; 0.25mg/ml DNase I, 150mM NaCl, 50% glycerol) to 10ml of lysate from above.

2. Incubate at 37°C for 15 minutes.

3. Add 4ml of 'phage precipitant' (Promega; 33% PEG-8000; 3.3M NaCl), mix gently and place on ice for 30 minutes.

4. Centrifuge at top speed in a bench centrifuge for 10 minutes.

5. Carefully decant the supernatant. Resuspend the pellet in 500µl of 'phage buffer' (Promega; 150mM NaCl; Tris-Cl, pH 7.4; 10mM $MgSO_4$).

6. Transfer the resuspended phage to a 1.5ml Eppendorf tube. Add 12.5µl of Pronase (Boehringer Mannheim) at 20mg/ml. Incubate at 37°C for 1 hour. (This step ensures there is no carry-over of nuclease which might degrade the lambda DNA.)

7. Spin in an Eppendorf centrifuge for 10 seconds to remove any insoluble particles, and transfer supernatant to a fresh Eppendorf tube.

iii) Purification of lambda DNA

1. Add 1ml of thoroughly mixed 'Wizard™ lambda DNA purification resin' (Promega) and mix by inversion.

2. Attach a 2ml disposable syringe barrel to the Luer-Lok extension of a fresh 'Wizard™ minicolumn' (Promega A712B).

3. Transfer the resin/supernatant mix into the syringe barrel. Insert the syringe plunger and gently push the slurry into the column.

4. Wash the minicolumn as follows. Detach the syringe and remove the plunger. Reattach the syringe to the minicolumn and pipette 2ml of 80% isopropanol into the syringe. Insert the plunger and push the wash solution through the column.

5. Transfer the minicolumn to a 1.5ml Eppendorf tube and spin for 20 seconds in to dry the resin.

6. Transfer the minicolumn to a fresh tube and apply 100µl of water or TE, preheated to 80°C. Immediately spin for 20 seconds to elute DNA.

7. Discard minicolumn. The eluted DNA can be used for analysis without further purification and is stored at -20°C.

5.2.6 Northern analysis

A multiple tissue Northern blot containing poly-A+ RNA from human heart, brain, placenta, lung, liver, skeletal muscle, kidney and pancreas was obtained from Clontech. The protocol for probing was as follows:

1. Warm a solution for hybridization/prehybridization containing 5 x SSPE, 10 x Denhardt's, 2% SDS, 50% formamide (freshly deionized) and 100µg/ml of freshly denatured, sheared salmon sperm DNA to 50°C to dissolve the SDS.

2. Prehybridize the blot in 10ml of this solution for 3-6 hours at 42°C.

3. Replace with 10ml of fresh solution, containing labelled probe, and hybridize overnight at 42°C.

4. Wash blot for 30-40 minutes in several changes of 2 x SSC, 0.05% SDS at room temperature, while shaking continuously.

5. Wash blot for 20 minutes in 0.1 x SSC, 0.1% SDS at 50°C. Repeat this step with fresh 0.1 x SSC, 0.1% SDS.

6. Shake off excess wash solution and seal blot in plastic wrap. Expose to x-ray film as described in section 2.9.4

5.2.7 Computer programs used for sequence analysis

The 'BLAST' algorithm (Altschul *et al.*, 1990) was used to search nucleic acid and protein databases for sequences that were similar to those isolated in this thesis. An example of the output obtained from such searches, and an explanation of how to interpret it, is given in the results section (Figure 5.13).

Several programs from the GCG (Genetics Computer Group) package, originally developed by the University of Wisconsin, were found to be useful in the analysis of cDNA sequences:

TESTCODE helps to identify protein coding sequences by plotting a measure of the non-randomness of sequence composition at every third base, using Fickett's TestCode statistic (Fickett, 1982).

GAP uses an algorithm devised by Needleman and Wunsch (1970) to find the alignment of two complete sequences with the maximum number of matches and the minimum number of gaps. The program reads values from an amino acid comparison matrix with matches equal to 1.5 and mismatches based on the evolutionary distance between amino acids (Dayhoff *et al.*, 1983; Gribskov and Burgess, 1986). A score is obtained for each possible alignment by summing these values and then subtracting an amount based on the number and lengths of any gaps. The *percentage identity* of the optimal alignment is the percentage of amino acids that match between the two sequences. The *percentage similarity* is the percentage of positions where the value for the pair of amino acids is over 0.5 (the similarity threshold). Residues that are across from gaps are ignored for the calculation of these values.

PEPTIDESTRUCTURE makes secondary structure predictions for a peptide sequence. It also calculates a measure of hydropathy, based on the average of a residue specific hydrophilicity index over a window of seven residues (Kyte and Doolittle, 1982). In addition, this program identifies potential glycosylation sites in the protein.

PLOTSTRUCTURE displays the predictions from **PEPTIDESTRUCTURE** in a graphical form.

PILEUP creates a multiple sequence alignment from a group of related sequences using progressive, pairwise alignments. The program begins by scoring the similarity between every possible pair of sequences. These similarity scores are then used to derive clustering relationships between the sequences, which direct the order of the pairwise alignments. A tree representation of these clustering relationships (known as a 'dendrogram') can be plotted, but whilst this gives a general picture of sequence similarity, it should be noted that it is not a phylogenetic reconstruction.

5.3 Results

5.3.1 Characterization of the associated microdeletion using novel YAC markers

In constructing a YAC contig around the DXS255 locus, left and right end cloning of YAC inserts was used to generate novel markers in the region (Chapter 3 and Hatchwell, 1994). DNA from affected individuals of family C, who have an associated microdeletion involving DXS255, was analysed using four of these markers. The presence or absence of the DXS146 locus was also investigated, using the probe pTAK8. Previous analysis indicated an order of Xpter–L(F1001)–DXS255–L(6129)–L(G0201)–R(F1101)–DXS146–Xcen for these markers (Chapter 3 and Hatchwell, 1994). The results demostrated that the microdeletion is confined to the interval between L(F1001) and L(G0201), and includes L(6129) (Fig. 5.2 and S. E. Lloyd, personal communication). This region is cloned in the two overlapping YACs 6129 and C0191, of sizes 185kb and 360kb respectively. Rare-cutter restriction mapping of YAC clones indicates a maximum distance of 370kb between L(F1001) and L(G0201) and thereby establishes the limits of the microdeletion (see Chapter 3).

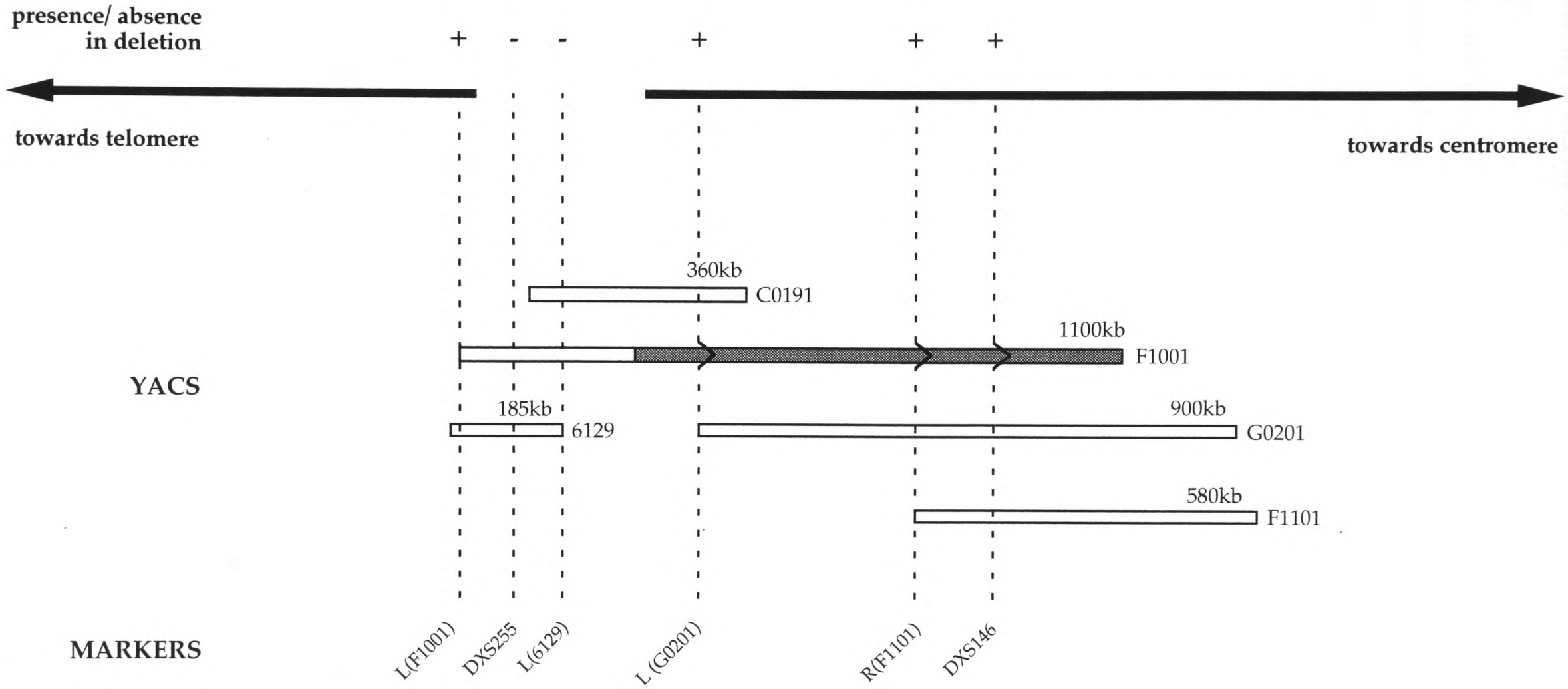


Figure 5.2: Analysis of microdeletion using markers generated from the DXS255–DXS146 YAC contig in Xp11.22 (Chapter 3 and Hatchwell, 1994). The status of each marker with respect to the deletion in family C is indicated (+ = present, - = deleted). YAC sizes and reference numbers are shown. L(YAC name) and R(YAC name) denote novel markers isolated from YACs by left and right end cloning respectively. The autosomal region of the chimæric clone F1001 is shaded. Distances between markers are not shown to scale, but the rare-cutter restriction maps described in Chapter 3 indicate that the flanking markers L(F1001) and L(G0201) are a maximum of 370kb apart.

5.3.2 Screening of kidney cDNA library using the 6129 YAC clone

Previous analysis of 6129 suggests that this is a stable, non-chimæric YAC. Furthermore, it does not co-migrate with any of the host yeast chromosomes on a pulsed field gel and can therefore be gel purified. Investigation of the microdeletion associated with Dent's (section 5.3.1) indicated that 6129 includes a substantial proportion of the deleted region, suggesting that it might contain at least part of the gene responsible for this disorder.

The YAC was separated from the background of yeast chromosomes by preparative pulsed field gel electrophoresis and purified using the Geneclean procedure (Fig. 5.3). Purified YAC DNA was hybridized, following suppression of repetitive elements, to duplicate primary filters of a human cDNA library made from adult kidney (Fig. 5.4). Potential positives were identified as those that gave signals in corresponding positions on both duplicates. Those clones which gave the strongest signals were picked and plated out at a lower density for secondary screening. Following hybridization with the YAC, secondaries were stripped and then probed with total human genomic DNA. The purpose of this step was to screen out any false positives containing repetitive elements, which may have been identified due to insufficient prereassociation of repeat sequences in the YAC probe. A tertiary screening step was used in order to ensure that individual positive plaques could be identified with confidence and therefore picked for further investigation.

This screening procedure resulted in the isolation of two clones (designated RL.3 and RL.6) which mapped back to the 6129 YAC (Figs. 5.5-5.7).

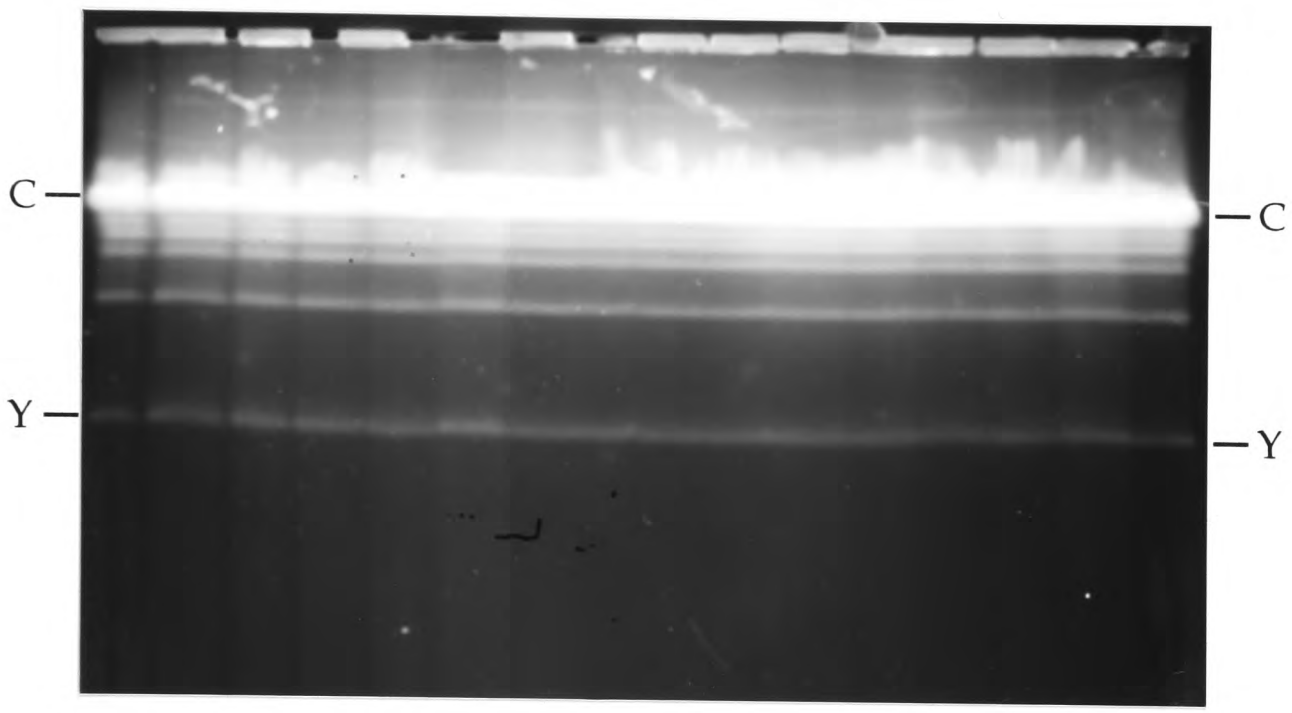
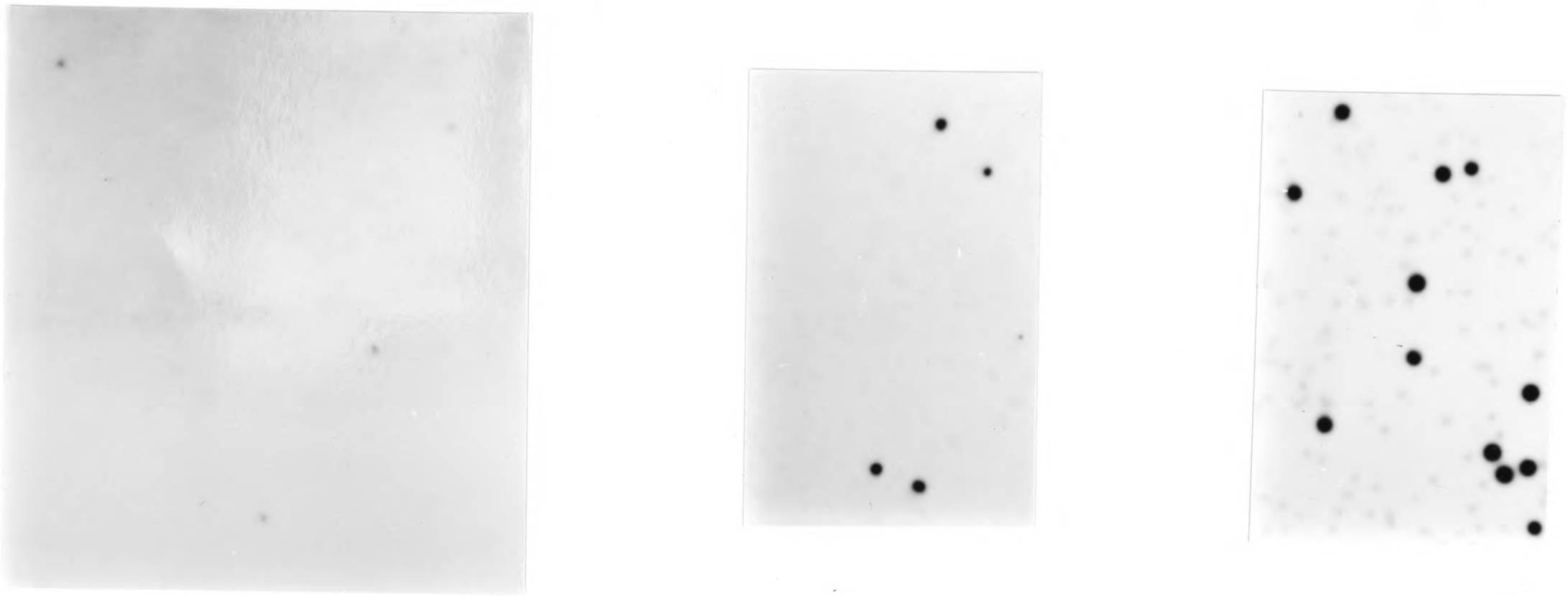


Figure 5.3: Preparative pulsed field gel of the 6129 YAC clone. DNA from undigested concentrated YAC plugs was run for 30 hours, using a 13 second switch time. The gel was stained with ethidium bromide and visualised under UV. The 185kb YAC band (Y) was then excised and purified using the Geneclean procedure (see Materials and Methods). The larger bands are chromosomes of the yeast host, most of which migrate together to form a compression band (C).



a) b) c)

Figure 5.4: Hybridization screening of a kidney-specific cDNA library with purified 6129 YAC. Examples of **a)** primary, **b)** secondary and **c)** tertiary screenings are shown. The faint non-specific hybridization to the background is useful for accurate identification of the positions of positive plaques in secondary and tertiary screenings. The density of plating out in the tertiary step is low enough for a single positive plaque to be picked, without contamination from neighbouring negative plaques. See text for details.

5.3.3 Analysis of cDNA clones RL.3 and RL.6

i) Mapping of clones within microdeletion

RL.3 and RL.6 both contain inserts of ~1.4kb in size. Sequence overlap between the two was suggested by the observation that they cross-hybridized to each other. Both detect an X-specific 4.4kb *Eco*R1 fragment in digests of human genomic DNA and of the 6129 YAC (Figs. 5.5 and 5.6). An additional weak band, of 800bp, is seen on hybridization of RL.3 to YAC DNA (Fig. 5.5). In general, the intensity of signals obtained from YAC digests is considerably higher than those obtained from human genomic DNA. This is due to the 200-fold disparity in the sizes of the yeast and human genomes; only 50ng of yeast DNA is required to give a signal comparable to that obtained from 10μg of human DNA (Brownstein *et al.*, 1989). It is therefore likely that the 800bp fragment *is* present in genomic DNA, but the signal obtained is too weak to be seen on the autoradiographs. The cDNAs also detect cognate bands in digests of the two overlapping YACs, C0191 and F1001 (data not shown).

On probing of pulsed field gel partial digests of 6129, it was found that the genomic fragments detected by RL.3 and RL.6 are localized to a ~5kb region at the extreme left end of the YAC insert, between the cloning site and a *Sal*I site (Fig. 5.7-5.8; Table 5.3). This placed them between 40 and 80kb from DXS255 on its centromeric side, and suggested that they map within the deleted region of family C. Hybridization of RL.3 and RL.6 to a panel of *Eco*R1 genomic digests from Dent's disease patients and normal individuals, confirmed that the 4.4kb genomic fragment detected in male and female normal controls is completely absent in affected males from family C (S. E. Lloyd, personal communication).

ii) Hybridization to a 'zoo-blot'

RL.3 and RL.6 were used to probe genomic digests of distantly related species (a 'zoo-blot'). They both show a high degree of conservation, detecting homologous fragments in primates, marsupials, rodents, reptiles and birds (Fig. 5.9).

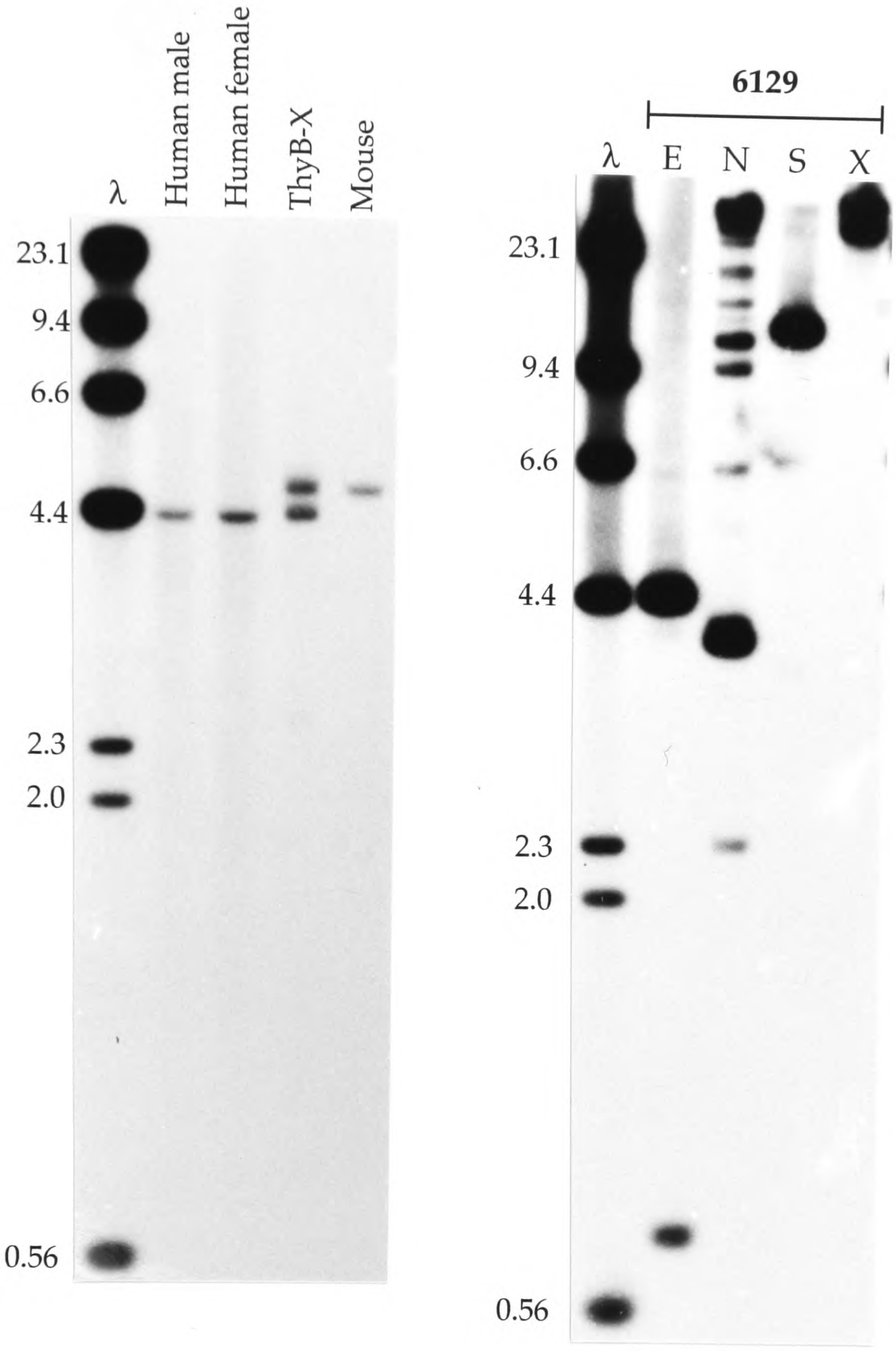


Figure 5.5: Hybridization of the RL.3 cDNA to digests of genomic, hybrid and YAC DNAs. Sizes of lambda markers (λ) are given in kilobases.

a) *Eco*R1 digests of genomic and hybrid DNAs, probed with RL.3. A 4.4kb band is detected in human genomic DNA, and also in the human X-only/mouse hybrid (ThyB-X) but not in mouse genomic DNA. A 4.6kb fragment of similar intensity to the 4.4kb band is seen in ThyB-X and mouse DNA, suggesting that the genomic sequences detected by this probe are highly conserved (see Figure 5.9).

b) *Eco*R1 (E), *Nde*I (N), *Sal*I (S) and *Xho*I (X) digests of the 6129 YAC probed with RL.3. Two *Eco*R1 fragments are detected; a 4.4kb fragment, as identified in a), and also an 800bp fragment. (See text for an explanation of this observation.) The *Sal*I fragment identified is the same size (11kb) as the *Sal*I end-fragment detected by pBR(L) in this YAC (Figure 3.12), suggesting that RL.3 maps within ~5kb of the left arm. This localization was confirmed by further analysis (see Figures 5.7-5.8). The ladder of fragments seen in the *Nde*I track is due to incomplete digestion with this enzyme.

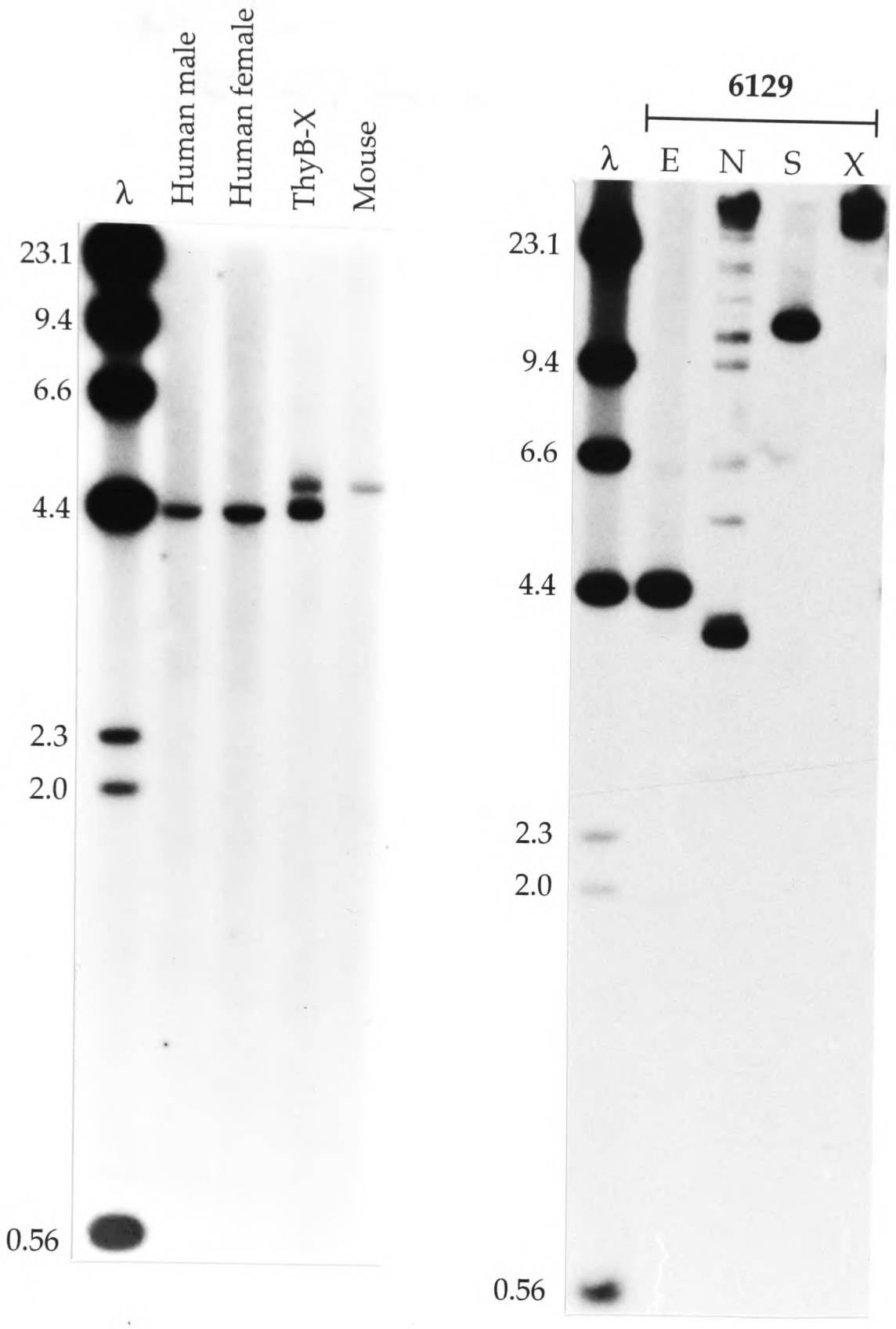


Figure 5.6: Hybridization of the RL.6 cDNA to digests of genomic, hybrid and YAC DNAs. Sizes of lambda markers (λ) are given in kilobases.
a) *Eco*R1 digests of genomic and hybrid DNAs, probed with RL.6. As found with RL.3 (Figure 5.5a) this cDNA detects a 4.4kb X-specific fragment in human genomic DNA, and a 4.6kb fragment in mouse genomic DNA, with both bands present in ThyB-X. However, in this case the mouse band is considerably lower in intensity than the human band, suggesting that sequences detected by RL.6 are less conserved than those detected by RL.3. Sequence analysis provides an explanation for this (Figure 5.11).
b) *Eco*R1 (E), *Nde*I (N), *Sal*I (S) and *Xho*I (X) digests of the 6129 YAC probed with RL.6. The hybridization pattern is very similar to that seen with RL.3 (Figure 5.5b). Note, however, that the 800bp *Eco*R1 fragment is not detected.

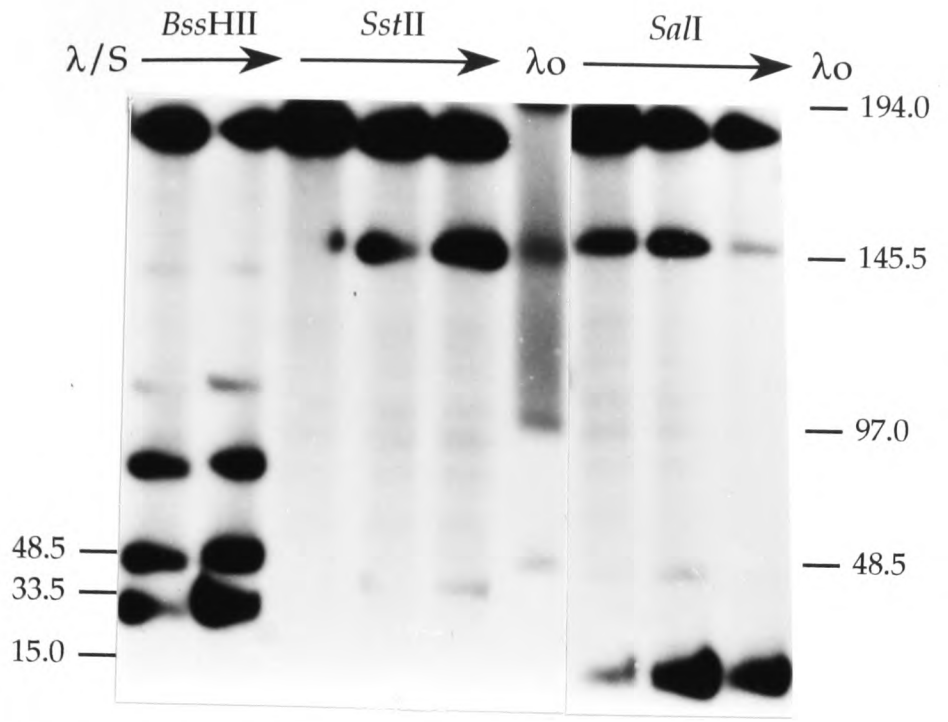


Figure 5.7: Partial digests of the 6129 YAC using different rare-cutters, probed with RL.3 cDNA. Direction of arrow represents increasing enzyme concentration (from 0.1-5U) with a 1 hour digestion time. Sizes of lambda oligomer (λo) and lambda/*Sal*I (λ/S) markers are given in kilobases. PFGE conditions were identical to those given in Fig. 3.8.

Enzyme	RL.3/RL.6
*Bss*HII	25, 45, 85
*Eag*I	145, 160
*Sal*I	<15, 145
*Sfi*I	105, 145
*Sst*II	145

Table 5.3: Fragment sizes, in kilobases, of bands detected on rare-cutter partial digests of 6129 YAC clone, when probed with RL.3 or RL.6 cDNA inserts.

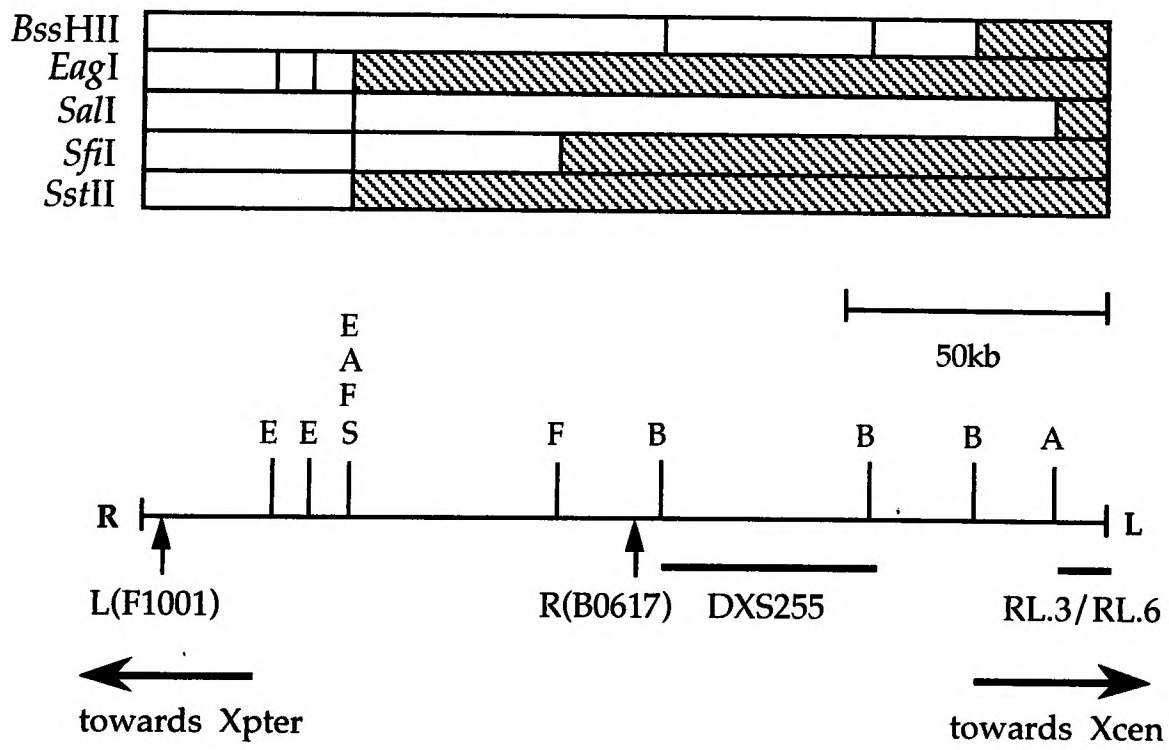


Figure 5.8: Localization of RL.3 and RL.6 within rare-cutter map of 6129 YAC clone, as deduced from fragment sizes in Table 5.3. Diagonal stripes show fragments on which sequences recognized by RL.3/RL.6 lie. Sites are given as in Fig. 3.9. Positions of R(B0617), L(F1001), and DXS255, and orientation of YAC with respect to Xpter-Xcen are indicated.

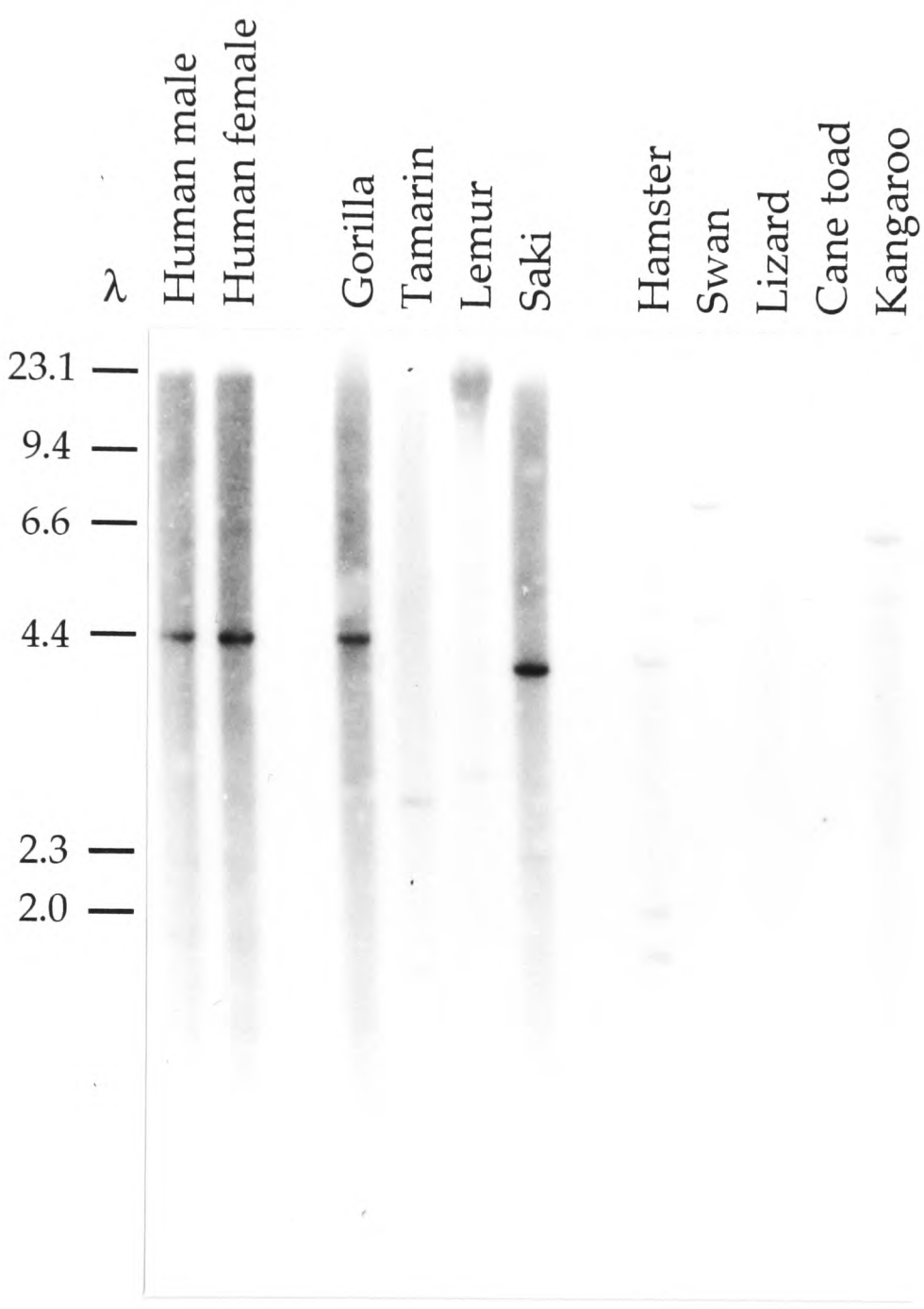


Figure 5.9: Hybridization of the RL.3 cDNA to a 'zoo-blot', containing *Eco*R1 digests of genomic DNA from distantly related species. Positions and sizes, in kilobases, of lambda markers (λ) are indicated.

iii) Pattern of expression and size of transcript

A Northern blot of poly-A enriched mRNA from various human tissues was probed with RL.3 (Fig. 5.10; top). A transcript of approximately 9.5kb was detected, expressed at high levels in kidney and at much lower levels in placenta and skeletal muscle. No expression was seen in heart, brain, lung, liver or pancreas. Subsequent hybridization with β-actin revealed that the loading of the skeletal muscle track was significantly higher than in others (Fig. 5.10; bottom). The relative expression of the ~9.5kb transcript in this tissue may therefore be even weaker than indicated. In addition, the track containing mRNA from pancreas was found to be relatively underloaded, so that possible expression in this tissue may have gone undetected.

iv) Sequence analysis

RL.3 and RL.6 were completely sequenced, and found to have a region of overlap of 878bp (Fig. 5.11). Fickett's analysis predicted the presence of a putative open reading frame (ORF), at the 5' end of RL.3 extending from 1 to 780bp in reading frame +1 and terminating with a TAG (Figs. 5.11-5.12). Sequence overlap was observed between RL.6 and the YAC probe L(6129), indicating that the gene is oriented with its 3' end towards the centromere (Fig. 5.11).

Homology searches using the Blast algorithm revealed that the predicted protein sequence of the 780bp ORF showed significant similarity to all members of the ClC family of voltage-gated chloride channels (Fig 5.13) (Jentsch *et al.*, 1995). This suggested that RL.3 and RL.6 originate from the 3' end of a gene (referred to in the remainder of this thesis as CLCN5) encoding a novel member of this family (termed ClC-5)

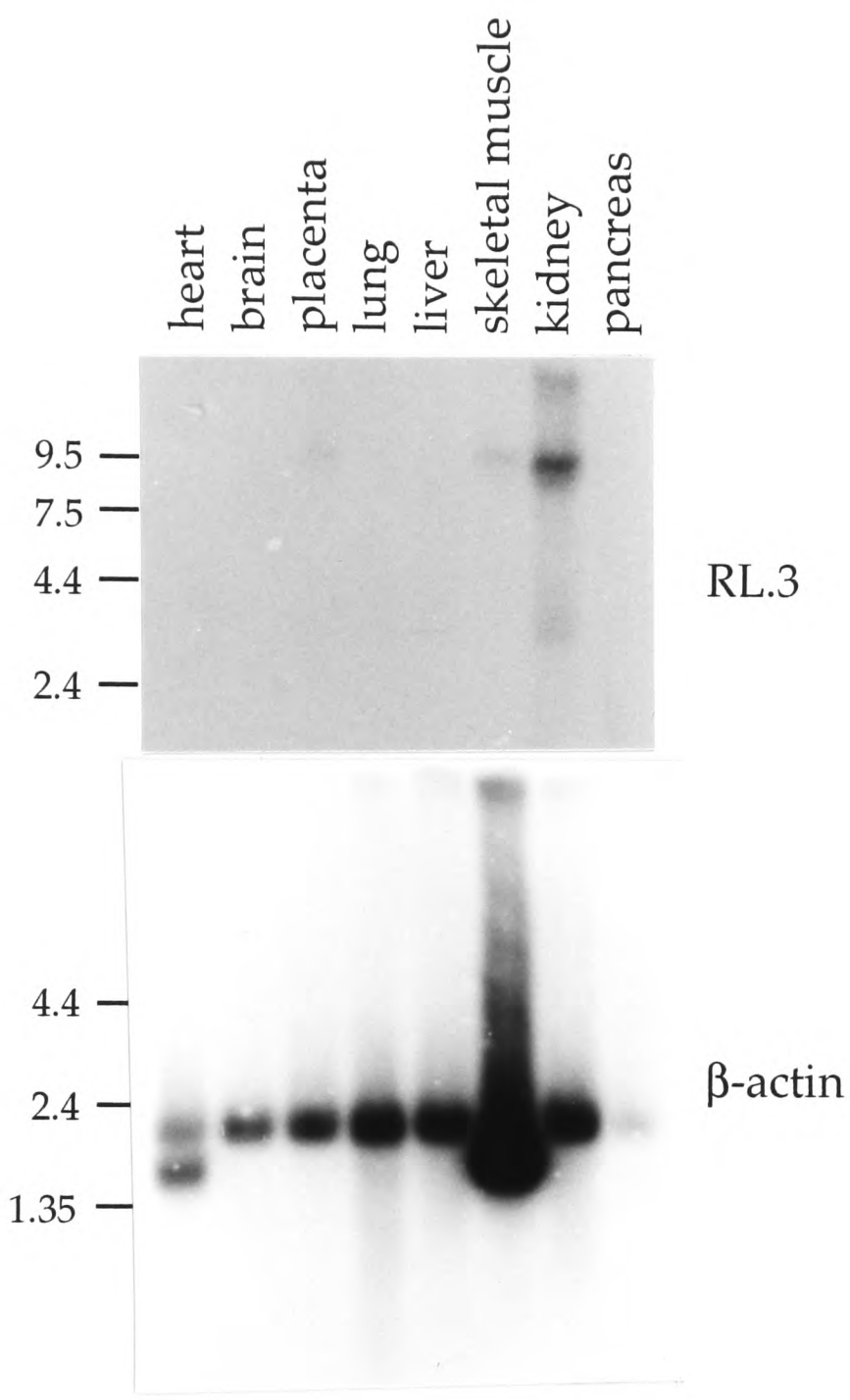


Figure 5.10: Probing of a Northern blot, containing poly-A enriched mRNA from various human tissues, with RL.3 cDNA (top) and β-actin control (bottom). Positions and sizes, in kilobases, of RNA ladder marker fragments are indicated on the left. RL.3 detects a ~9.5kb transcript, expressed at high levels in kidney tissue. See text for discussion.

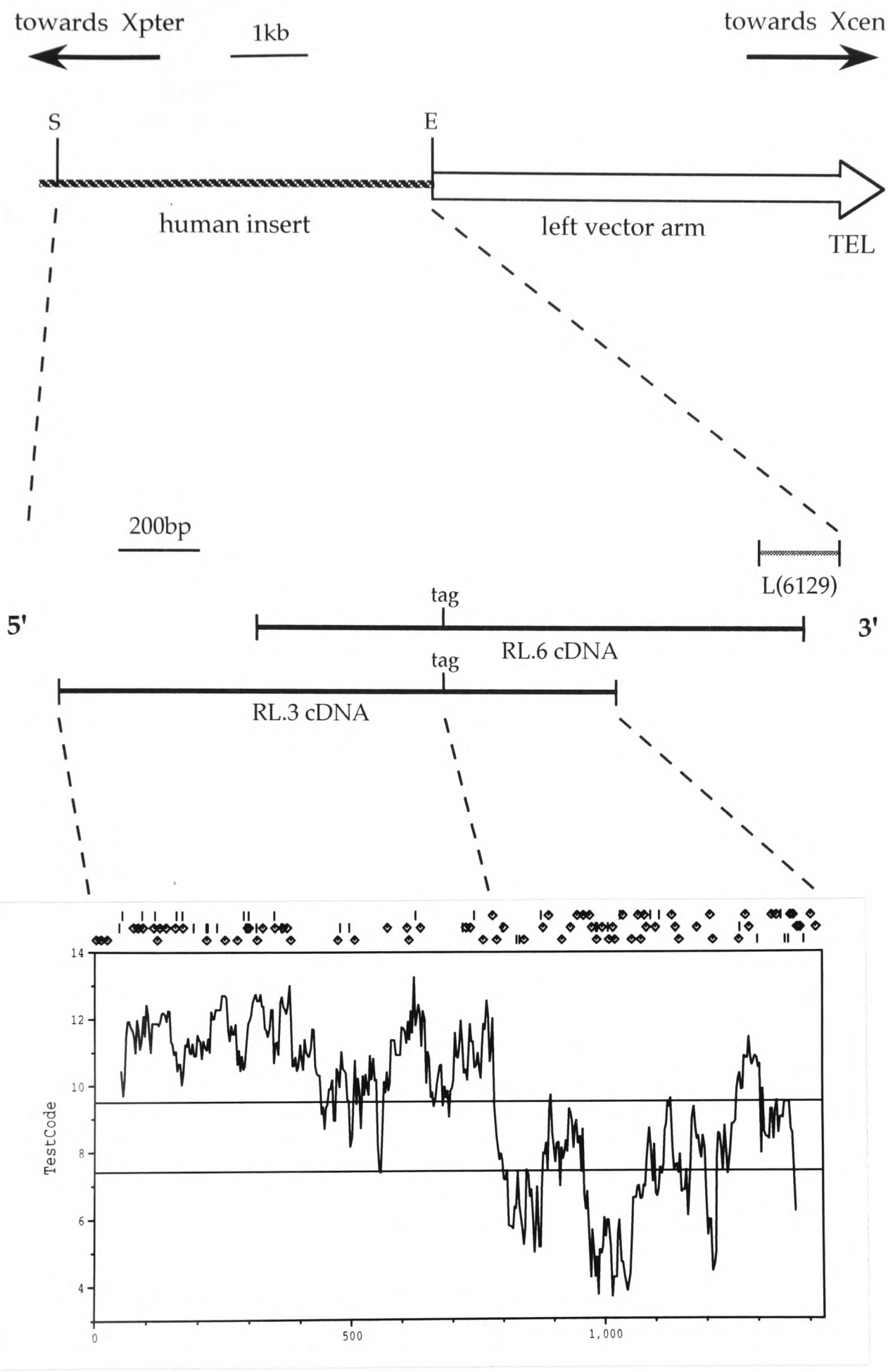


Figure 5.11: Sequence analysis of RL.3 and RL.6 cDNA clones. The ~5kb *Sal*I-*Eco*R1 fragment at the left end of the 6129 YAC is shown **(top)**, with the orientation of the YAC relative to the X chromosome indicated. The overlap between RL.3 and RL.6 cDNAs and their orientation with respect to YAC and chromosome, as inferred from sequence overlap between RL.6 and the genomic clone L(6129), are indicated below this **(middle)**. Fickett's analysis of RL.3 using **TESTCODE** (Section 5.2.7), identifies a 780bp ORF in the +1 reading frame at the 5' end of the cDNA **(bottom)**. The plot is divided into three regions; the top region predicts coding sequence to a 95% level of confidence, the bottom region predicts non-coding sequence to the same confidence level, and the middle region is the 'window of vulnerability' where no significant prediction can be made. The potential start codons (short vertical lines) and stop codons (small diamonds) of each reading frame are indicated above the curve. The position of the stop codon of the 780bp ORF is indicated on the overlapping cDNAs with 'tag'.

```
   1  aatagctggtgtagtcagggagctgattgcatcacccccggcctttatgcaatggttggg
   1   N  S  W  C  S  Q  G  A  D  C  I  T  P  G  L  Y  A  M  V  G

  61  gctgcagcctgcttaggtggggtgactcggatgactgtttctcttgttgtcataatgttt
  21   A  A  A  C  L  G  G  V  T  R  M  T  V  S  L  V  V  I  M  F

 121  gaactgactggtggcttagaatacatcgtgcctctgatggctgcagccatgacaagcaag
  41   E  L  T  G  G  L  E  Y  I  V  P  L  M  A  A  A  M  T  S  K

 181  tgggtggcagatgctcttgggcgggagggcatctatgatgcccacatccgtctcaatgga
  61   W  V  A  D  A  L  G  R  E  G  I  Y  D  A  H  I  R  L  N  G

 241  taccccttcttgaagccaaagaagagtttgctcataagaccctggcaatggatgtgatg
  81   Y  P  F  L  E  A  K  E  E  F  A  H  K  T  L  A  M  D  V  M

 301  aaaccccggagaaatgatcctttgttgactgtccttactcaggacagtatgactgtggaa
 101   K  P  R  R  N  D  P  L  L  T  V  L  T  Q  D  S  M  T  V  E

 361  gatgtagagaccataatcagtgaaaccacttacagtggcttcccagtggtggtatcccgg
 121   D  V  E  T  I  I  S  E  T  T  Y  S  G  F  P  V  V  V  S  R

 421  gagtcccaaagacttgtgggctttgtcctccgaagagatctcattatttcaattgaaaat
 141   E  S  Q  R  L  V  G  F  V  L  R  R  D  L  I  I  S  I  E  N

 481  gctcgaaagaaacaggatggggttgttagcacttccatcatttatttcacggagcattct
 161   A  R  K  K  Q  D  G  V  V  S  T  S  I  I  Y  F  T  E  H  S

 541  cctccattgccaccatacactccacccactctaaagcttcggaacatcctcgatctcagc
 181   P  P  L  P  P  Y  T  P  P  T  L  K  L  R  N  I  L  D  L  S

 601  cccttcactgtgactgaccttacacccatggagatcgtagtggatattttccgaaagctg
 201   P  F  T  V  T  D  L  T  P  M  E  I  V  V  D  I  F  R  K  L

 661  ggactgcggcagtgcctggttacacacaacgggcgattgcttggaatcattaccaaaaag
 221   G  L  R  Q  C  L  V  T  H  N  G  R  L  L  G  I  I  T  K  K

 721  gatgtgttaaagcatatagcacagatggcgaaccaagatcctgattccattctcttcaac
 241   D  V  L  K  H  I  A  Q  M  A  N  Q  D  P  D  S  I  L  F  N

 781  tagaatcatagagttctggatgtaaagcgggaaggacattacagaccatggatatgttgt
 261   @

 841  ttaacggtacccaaaacacattttccatatttggatggtgaagtcacattagtgtgttgt
 901  ctctttcctacaagttaaccagttgcactacataatctctggaaattaattttctcttta
 961  ggagaaattatagttaggcttccatgatgttacattaggaagatatcatgaaagaataaa
1021  taagattgctatggtttaattatatttgctttttaaaagatttttttaacttaaaaagta
1081  gttagccaatatgcaatcactgaaaactatgcaagagaaattccaaccgtcctgacctat
1141  aacctgtaggaaaccgacgaaaaagtcactcttttgggatctaactgttgttactggaag
1201  acgaaggtaaactaaggggctttgcttttcaaaccagagaaaggaaagccagaaggaaaa
1261  gagtaatggtattttctagactgtgaagattcagttcaaatgttatccttgttcctgtta
1321  caatatttagcattattagtttgttatgtgtgtatgtttatgttaattttaatttctgat
1381  tataagacaatgctgctttggttaatctcttctaaaggaattta
```

Figure 5.12: Complete nucleotide and predicted amino acid sequence of the 1.4kb cDNA clone RL.3. The initiation codon of the open reading frame is expected to lie in a more 5' part of the transcript. The stop codon is indicated by an @. Two predicted hydrophobic domains (see Section 5.3.4) are underlined.

```
Query= RL.3 (1425 bases)
 Translating both strands of query sequence in all 6 reading frames

Database:  nbrf
           64,760 sequences; 19,009,044 total letters.

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

A45483 S     A45483 chloride channel, CLC-K1 - Rats     +1    80  8.9e-09    2
S36602 S     S36602 chloride channel protein CLC-1 -... +1    95  1.4e-08    3
S37078 S     S37078 chloride channel protein CLC-1 -... +1    95  1.4e-08    3
S19595 S     S19595 chloride channel protein - rat      +1    95  1.6e-08    4


>S36602 chloride channel protein CLC-1 - human  Length = 988

Plus Strand HSPs:

 Score = 95 (46.9 bits),  Expect = 1.6e-05, P = 1.6e-05
 Identities = 21/45 (46%), Positives = 28/45 (62%), Frame = +1

Query:    64 AACLGGVTRMTVSLVVIMFELTGGLEYIVPLMAAAMTSKWVADAL 198
             AA L G    TVS  VI FELTG + +I+P+M A + +  VA +L
Sbjct:   529 AAALTGAVSHTVSTAVICFELTGQIAHILPMMVAVILANMVAQSL 573

 Score = 61 (30.1 bits),  Expect = 5.3e-05, Poisson P(2) = 5.3e-05
 Identities = 14/39 (35%), Positives = 23/39 (58%), Frame = +1

Query:   346 SMTVEDVETIISETTYSGFPVVVSRESQRLVGFVLRRDL 462
             S T  ++ T++  TT    P+V S++S  L+G V R +L
Sbjct:   619 SYTYGELRTLLQTTTVKTLPLVDSKDSMILLGSVERSEL 657

 Score = 58 (28.7 bits),  Expect = 1.4e-08, Poisson P(3) = 1.4e-08
 Identities = 12/27 (44%), Positives = 17/27 (62%), Frame = +1

Query:    31 ITPGLYAMVGAAACLGGVTRMTVSLVV 111
             I PG YA++GAAA  G V+    + V+
Sbjct:   519 ILPGGYAVIGAAALTGAVSHTVSTAVI 545
```

Figure 5.13: A sample of the output obtained from the original computer homology search using RL.3 cDNA sequence. The BLASTX software tool was used (Gish and States, 1993). This program assigns similarity scores to ungapped, aligned pairs of sequence segments, based on a matrix of similarity/substitution scores for all possible pairs of residues. The matrix, like that used for GAP (Section 5.2.7), is derived from evolutionary distances between amino acids as specified by the PAM (point accepted mutation) amino acid substitution model (Dayhoff *et al.*, 1983); identities and conservative replacements receive positive values, and non-conservative replacements receive negative values. The output lists those database sequences that give "high-scoring segment pairs" (HSPs), along with the score and reading frame of the best HSPs from each sequence. N is the number of HSPs found between query and subject, and the Poisson probability, P(N), gives an estimate of their statistical significance. These results supported the presence of a coding region in frame +1 (Figure 5.11-5.12) and identified similarity with voltage-gated chloride channels. Examples of HSP alignments are shown beneath the sequence list. At the time of searching, CLCN3 and CLCN4, which have the highest homology to CLCN5 (see Section 5.3.4), had not yet been cloned and were therefore absent from all databases.

5.3.4 Cloning and characterization of the CLCN5 coding region

i) Construction of a cDNA contig spanning the CLCN5 ORF

The coding regions of the transcripts encoded by all previously characterized ClC genes exceed 2kb (reviewed in Jentsch *et al.*, 1995). RL.3 only contains 780bp from the 3' end of the CLCN5 ORF. It was therefore necessary to isolate new cDNA clones from the gene, covering the more 5' region of the transcript.

In the initial hybridization of the kidney library with the 6129 YAC, only a third of potential positives identified on primary screening were followed through to secondary/tertiary screening. A possible strategy for isolation of a larger part of the CLCN5 transcript would therefore be to follow through and further analyse these remaining positives. However there were several arguments against adopting such an approach:

- The signal to noise ratio of the primary screening using 6129 was low. It was therefore possible that a number of weak, but legitimate, positives may have gone undetected (for example, if the size of the plaque produced by the positive cDNA was relatively small).

- The preparation of YAC probe was laborious and time-consuming. Purification of YAC DNA from one pulsed field gel (using sixteen concentrated YAC plugs) usually gave enough probe for only two hybridizations. In addition, it was necessary to use large amounts of sonicated human placental DNA for prereassociation in each screening.

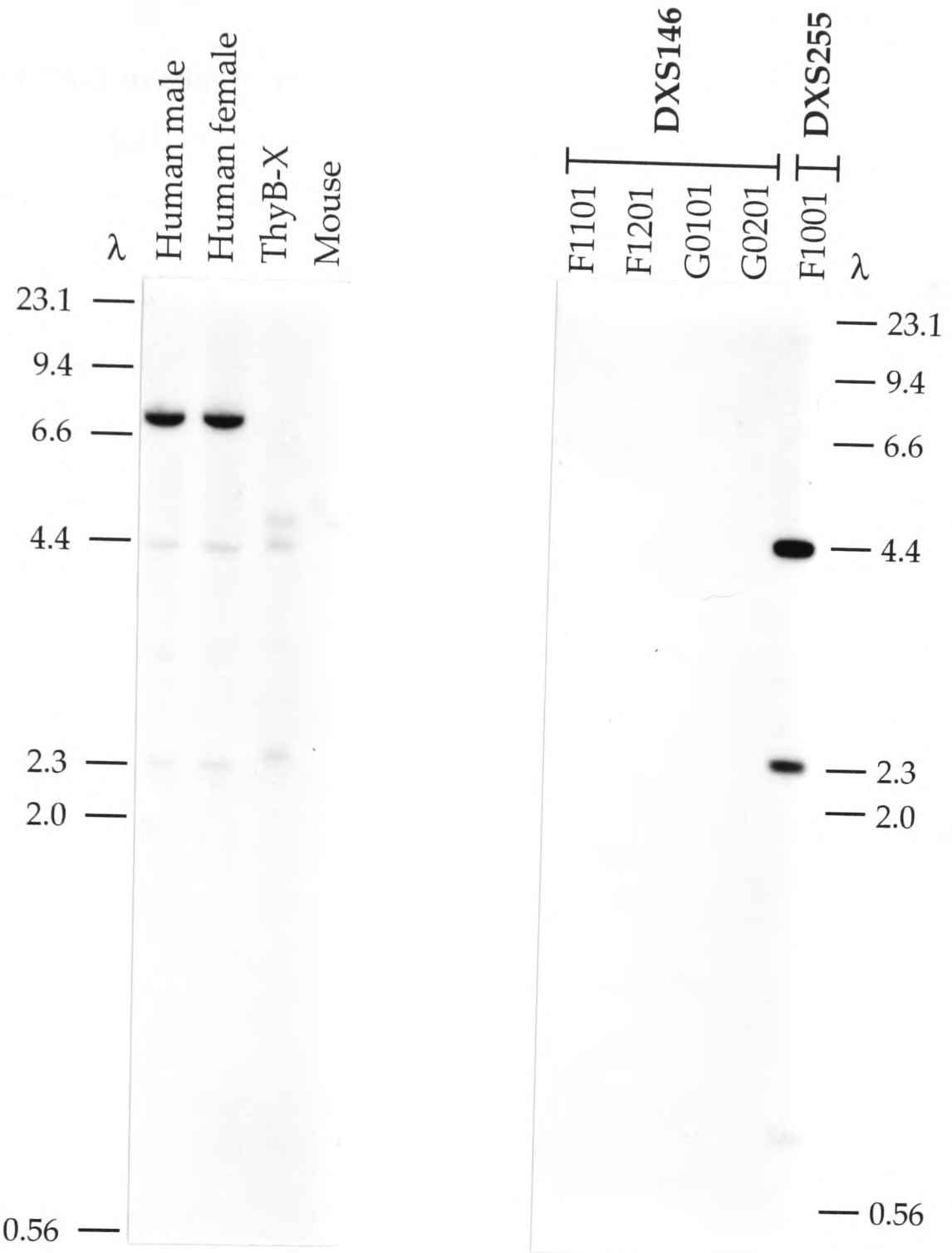
- A significant proportion of 'positives' investigated in the initial screening were found to hybridize to total human at secondary screening, and were therefore discarded.

Instead, an alternative strategy, one of 'cDNA walking', was adopted, involving rescreening of the kidney library with cDNAs which had already been isolated; these probes gave very good signal to noise ratios.

STEP 1:

More than twenty positives were detected when RL.3 was hybridized to primary filters of the library. An almost identical pattern of positives was seen on a subsequent probing of the same filters using RL.6; only one of the positives identified by RL.3 was not detected by RL.6. As described above, RL.6 overlaps with 878bp of the 3' end of RL.3, and the average insert size of the library is reported as ~1.5kb. It thus seemed likely that the single clone which was RL.3-positive but RL.6-negative would contain a significant amount of sequence 5' to that already obtained.

This clone (known as 3A-2) was therefore isolated and found to contain a 1.5kb insert. On hybridization to *Eco*R1 digests of the DXS255 YACs, the insert detected a 2.3kb band, as well as the 800bp and 4.4kb bands previously seen when probing with RL.3 (Fig. 5.14). X-specific *Eco*R1 bands were also faintly detected when 3A-2 was hybridized to a human panel. However, in addition, a 7.4kb fragment of much greater intensity was seen (Fig. 5.14). The latter was not present in the human X-only/mouse hybrid. Sequence analysis of 3A-2 indicated that whilst ~600bp at its 3' end had high homology to ClC genes, its 5' end was identical to human mitochondrial DNA sequences. Since *Eco*R1 digestion of human mitochondrial DNA generates three fragments, one of which is 7.4kb (Anderson *et al.*, 1981), this explains the origin of the non-X-specific band on the human panel. The higher intensity of this band could partly be due to the fact that mitochondrial sequences are present at greater copy number than nuclear sequences. These results suggested that 3A-2 was a chimæric clone involving CLCN5 transcript fused to DNA of mitochondrial origin, and thus probably an artefact of library construction.



a) **b)**

Figure 5.14: Hybridization of the 3A-2 cDNA to *Eco*R1 digests of genomic, hybrid and YAC DNAs. Positions and sizes, in kilobases, of lambda markers (λ) are indicated.
a) 3A-2 detects two faint X-specific bands, of 4.4kb and 2.3kb, in human genomic DNA. These bands appear to be slightly larger in ThyB-X, but this is due to overloading of the DNA in this track. In addition, a 4.6kb band corresponding to the murine homologue of CLCN5 is detected in digests of ThyB-X and mouse DNA (see Figures 5.5 and 5.6). The 7.4kb fragment of high intensity which is seen in the human genomic tracks is likely to be of mitochondrial origin, as described in the text. Detection of this band is a consequence of the chimæric nature of the 3A-2 clone (see text).
b) On probing *Eco*R1 digests of a panel of YACs, 3A-2 detects three bands, of 4.4kb,2.3kb and 800bp, in the overlapping clones F1001, 6129 and C0191, but not in any other YACs from the Xp11.23-p11.22 contigs. A subsection of the YAC panel is shown here.

The 3' end of 3A-2 overlapped with, and was identical in sequence to, 463bp at the 5' end of RL.3 (Table 5.4). However, due to the chimæric nature of the new clone, only an additional 163bp of ClC-like sequence 5' to RL.3 was obtained. It was therefore deemed necessary to analyse more of the primary positives detected by RL.3, in order to find a clone extending further 5'.

Instead of taking all these positives through secondary/tertiary screenings and then sequencing them, a PCR strategy was used to find rapidly which of them would be useful. Eluate from each primary that had been picked was used as template in a PCR reaction involving a forward primer designed from the lambda vector, and a reverse primer designed from the 5' region of RL.3 (Fig. 5.15a). Specific amplification of a ~700bp product was obtained when using the 6H eluate as a template. Digestion of this product with *Eco*R1 gave two fragments, a 193bp vector band and a 499bp band (referred to as 6H-P) containing amplified insert sequences. 6H-P detects X-specific bands of 2.3kb and 800bp on *Eco*R1 digests of human and YAC DNA, as expected from analysis of 3A-2 and RL.3 (Fig. 5.15b-c). The product was cloned using TA-cloning (see Section 6.2.2) and sequencing confirmed that it overlapped with 3A-2 and RL.3 (Table 5.4), and that it contained 298bp of new 5' sequence which was homologous to previously characterized chloride channels.

STEP 2:

Library screening using 6H-P resulted in the isolation of two novel clones, 2B and 7Y, providing an additional 721bp of CLCN5 ORF sequence (Table 5.4). 7Y detected a new X-specific *Eco*R1 band of 7.7kb in YACs and human DNA, as well as the previously detected 2.3kb and 800bp fragments (Fig. 5.16a). However, comparison of the 7Y sequence to that of CLCN4, a highly homologous chloride channel (see below), suggested that the initiation codon had not yet been reached.

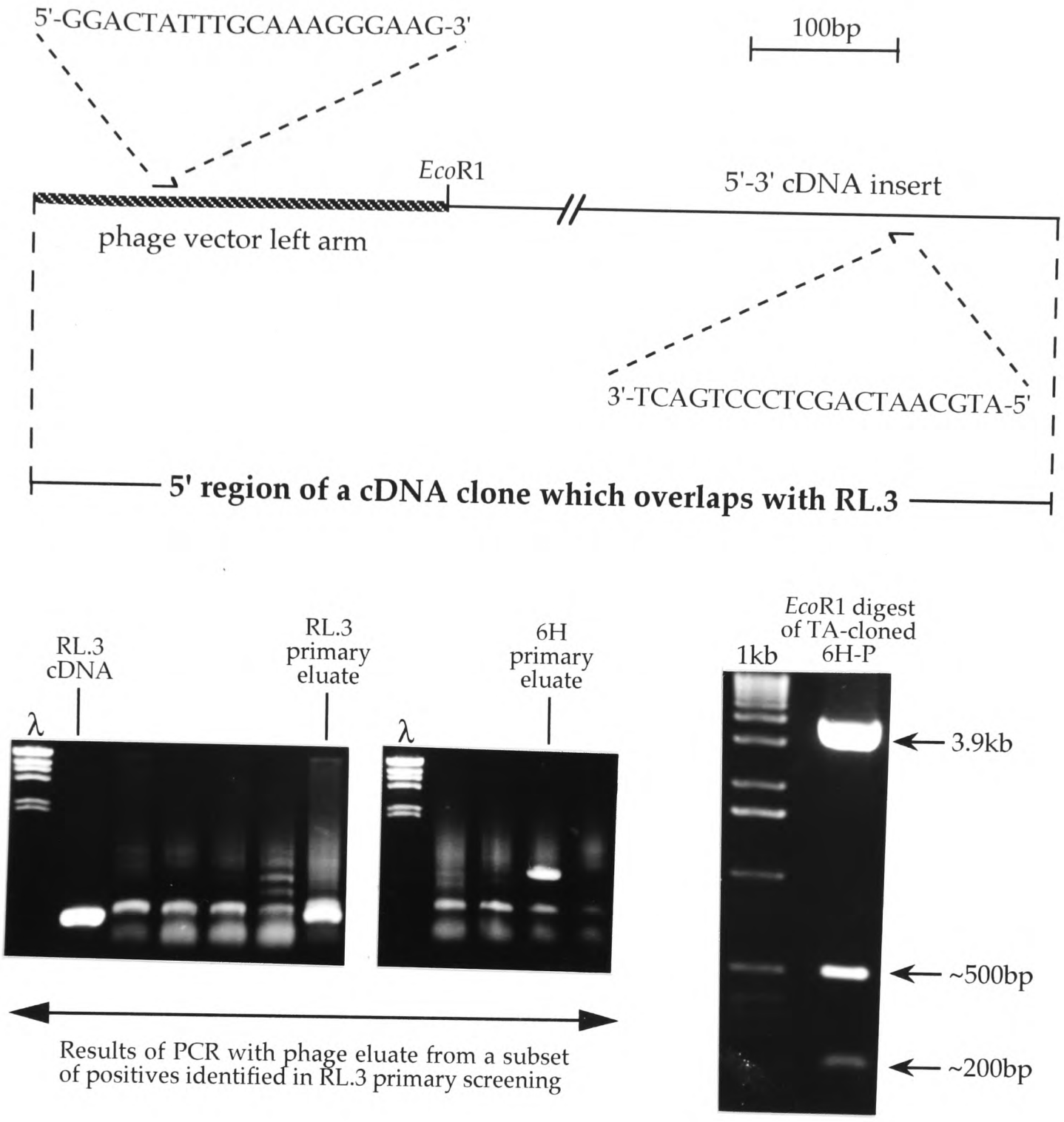
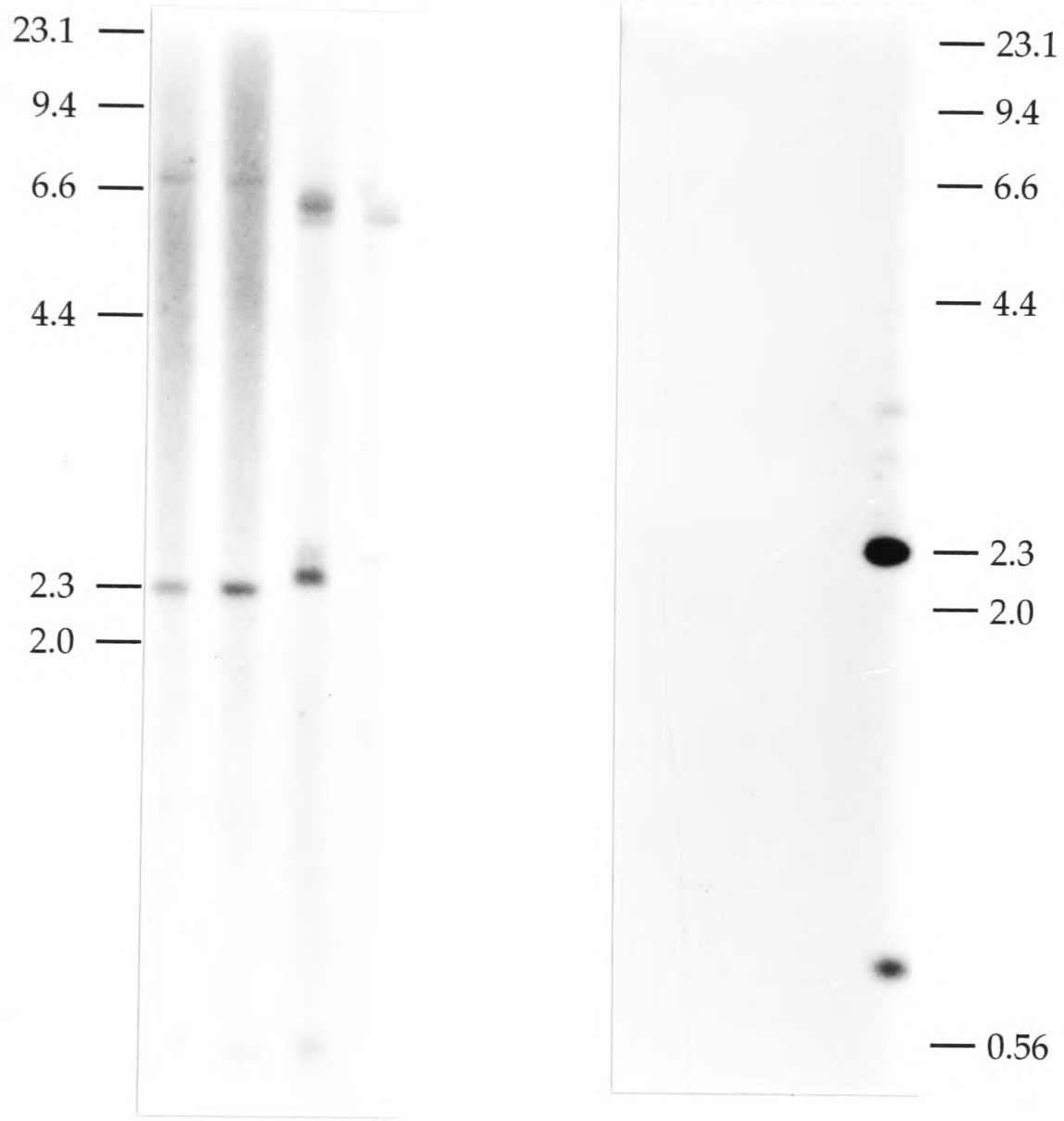


Figure 5.15: Isolation and analysis of the 6H-P clone.
a) A PCR assay was developed using a forward primer from the phage vector left arm*, and a reverse primer from the most 5' region of the RL.3 ORF **(top)**. This was used to test for the presence of sequences 5' to RL.3 in the primary positives which had been identified by hybridization screening with RL.3 (see text). Phage eluate was used directly as template. The expected 225bp product (containing 193bp of phage vector and 32bp of insert) was obtained from the RL.3 primary eluate and purified RL.3 cDNA controls **(bottom left)**. Although all phage eluates gave a non-specific product of ~300bp, one (6H) gave an additional band of ~700bp. This fragment (6H-P) was purified and cloned into a TA-cloning vector (see Section 6.2.2). Digestion of the resulting plasmid with *Eco*R1 **(bottom right)** generated fragments of 3.9kb (TA-cloning vector), ~500bp (6H-P insert) and ~200bp (phage vector sequences). Hybridization analysis (Fig. 5.15b-c) and sequencing of 6H-P confirmed that it originated from the CLCN5 transcript. (This sequence was later verified by comparison to that of other clones in the cDNA contig.) Sizes of lambda (λ) and 1kb ladder (1kb) marker fragments are given in Section 2.6.
* Note that a PCR assay was also developed using a primer from phage vector right arm, because some inserts may have been cloned in a 3'-5' orientation. However, amplification of the primary phage eluates using this assay gave no product (not shown).



b) **c)**

Figure 5.15 (cont): Hybridization of 6H-P to *Eco*R1 digests of genomic and hybrid and YAC DNAs. Positions and sizes, in kilobases, of lambda markers (λ) are indicated.
b) 6H-P detects two X-specific bands, of 2.3kb and 800bp, in human genomic DNA. The bands appear to be slightly larger in ThyB-X, but this is due to overloading of the DNA in this track. In addition, fragments of 6.4kb, 6.0kb and 4.6kb are detected in digests of ThyB-X and mouse DNA . The faint 7.4kb band seen in the human genomic tracks is the residual signal from the previous hybridization using 3A-2 (Figure 5.14).
c) On probing *Eco*R1 digests of a panel of YACs, 6H-P detects 2.3kb and 800bp bands in the overlapping clones F1001, 6129 and C0191, but not in any other YACs from the Xp11.23-p11.22 contigs. A subsection of the YAC panel is shown here.

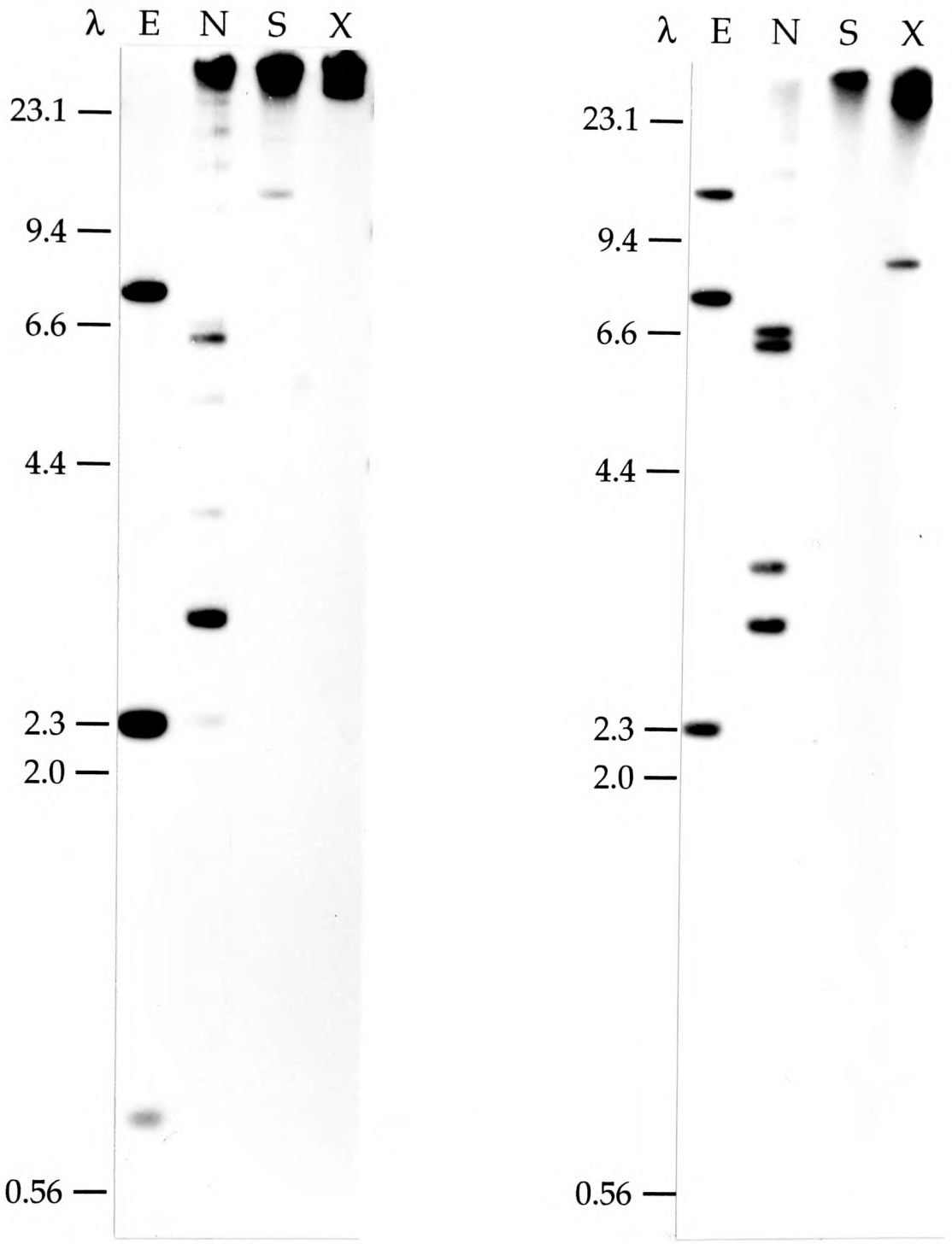
STEP 3:

A 642bp PCR product, amplified from the 5' part of 7Y, was used to rescreen the kidney cDNA library. One of the clones isolated, 2C, was found to have a chimæric insert, involving sequences of human mitochondrial origin at its 5' end, in a similar manner to 3A-2 (Table 5.4). As with 3A-2, hybridization of 2C to a human panel gave a very strong *Eco*R1 band at ~7.4kb which was not X-specific. Analysis of another clone, 3N, suggested that complete coverage of the CLCN5 ORF had been obtained (see below). Both 2C and 3N detect a 12.0kb *Eco*R1 band in YACs containing CLCN5, as well as the more 3' exon-containing fragments (Fig. 5.16b and Table 5.4).

ii) Sequence analysis

Sequencing of the new cDNA clones, combined with sequence data from RL.3 and RL.6 (section 5.3.3) enabled the entire coding sequence of CLCN5 to be determined (Fig. 5.18). The cDNA contig described above gives at least two fold coverage from nucleotides 395 to 3173 of the transcript (Fig. 5.17; Table 5.4), thus allowing verification of the sequence. The open reading frame is predicted to encode a protein of 746 amino acids. The initiation methionine was assigned to the second ATG triplet (nucleotides 292-294) which appears downstream of a stop codon (nuceotides 226-228) in the same reading frame. This start site seems more likely than the first in-frame ATG (at nucleotides 232-234), since its surrounding sequence context fits much more strongly the expectations for a eukaryotic translation initiation site (Figure 5.19) (Kozak, 1987).

The sequence shown in Figure 5.18 extends from the beginning of 3N to the end of RL.3 (i.e. nucleotides 1-3173). Additional data from the 3' untranslated region in RL.6 is given in the Appendix. Although RL.6 contains ~1.15kb of untranslated transcript beyond the end of the ORF, the poly-A tail has not yet been reached. In addition, no polyadenylation signals (AATAAA) have been identified at the 3' end of RL.6. An attempted cDNA walk towards the 3' end of the CLCN5 transcript yielded only one additional clone, which overlapped with the 3' portion of RL.6, but only contained a few nucleotides of extra sequence (data not shown). The size of the 3' untranslated region is therefore undetermined at this stage.



a) **b)**

Figure 5.16: Hybridization analysis of cDNAs from the 5' region of the CLCN5 contig. *Eco*R1 (E), *Nde*I (N), *Sal*I (S) and *Xho*I (X) digests of the 6129 YAC were probed with **a)** 7Y and **b)** 3N. Positions and sizes, in kilobases, of λ markers are indicated. See Table 5.4 and Figure 5.17 for a summary of these results. Further analysis confirmed that the fragments detected by these cDNAs are X-specific and are also present in the YACs F1001 and C0191. The ladder of bands detected in *Nde*I tracks is due to incomplete digestion with this enzyme (see Figure 3.12).

Name	Isolated with	Extent	*Eco*R1 fragments detected				
			12.0kb	7.7kb	2.3kb	0.8kb	4.4kb
3N	7Y-P	1-1481	+	+	+	–	–
2C	7Y-P	399-880	+	+	–	–	–
7Y	6H-P	571-1771	–	+	+	+	–
2B	6H-P	1020-1639	–	–	+	–	–
6H-P	3A	1291-1783	–	–	+	+	–
3A-2	3A	1588-2212	–	–	+	+	+
RL.3	6129 YAC	1750-3173	–	–	–	+	+
RL.6	6129 YAC	2297-3681	–	–	–	–	+

Table 5.4: Details of clones from cDNA contig spanning the complete CLCN5 open reading frame in humans. The extent of each cDNA is given with respect to the composite nucleotide sequence shown in Figure 5.18. The exonic *Eco*R1 fragments detected by each clone on probing digests of human, ThyB-X and YAC DNAs are indicated (see Figs. 5.14-5.16). +, fragment detected; –, fragment not detected.

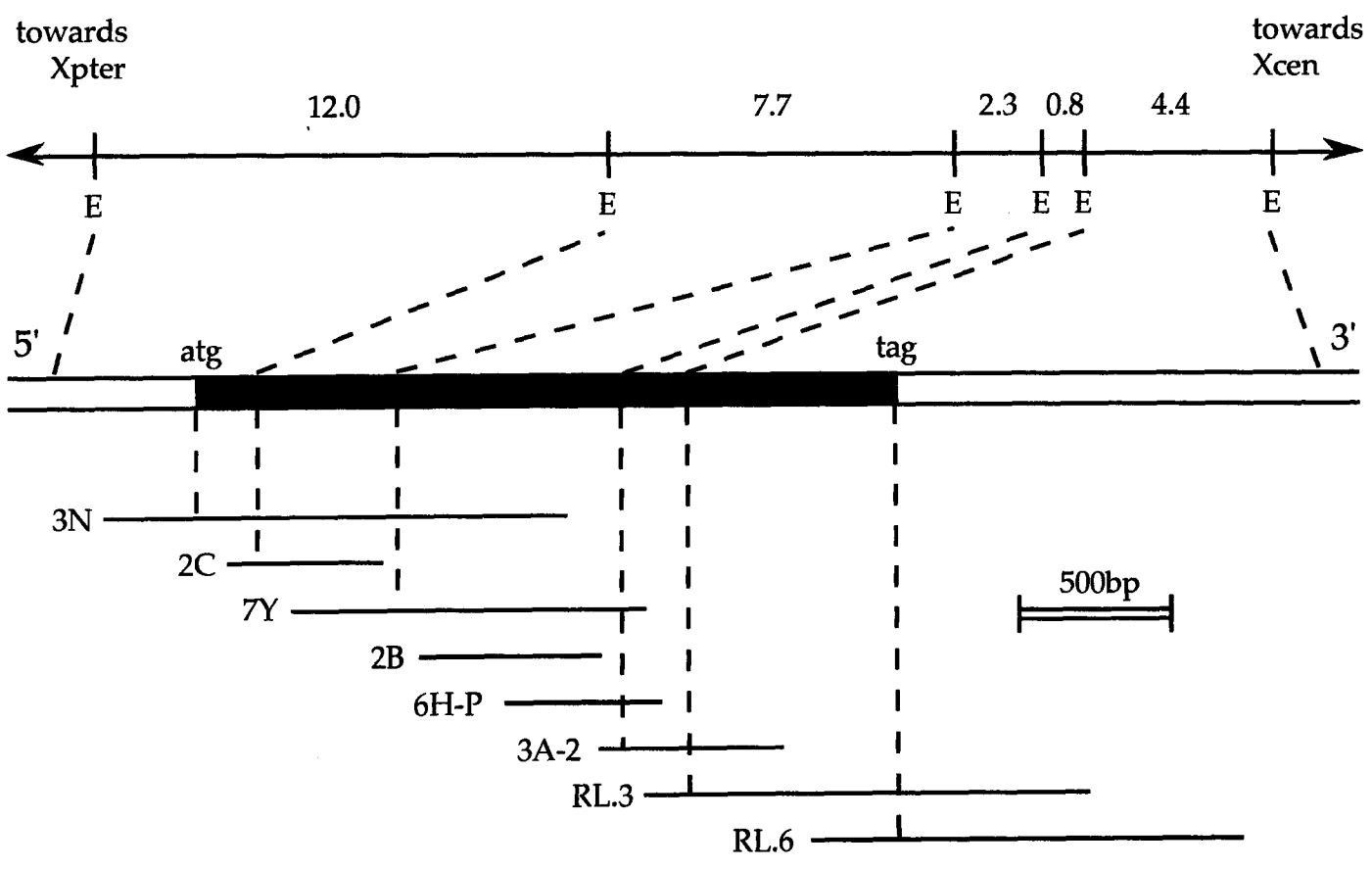


Figure 5.17: Relative overlaps of cDNA clones **(bottom)**, their alignment to the CLCN5 transcript **(middle)** and correlation with *Eco*R1 fragments **(top)**, as deduced from hybridization studies. The scale shown applies only to the cDNA clones and transcript. The initiator methionine of the open reading frame (shaded in black) is represented by "atg" and the stop codon by "tag". No sites for *Eco*R1 are contained in the cDNA contig; all map within introns. Dashed lines indicate how each cDNA relates to the genomic fragments that it detects. Only exon-containing genomic fragments are shown, and so each E may in fact represent more than one *Eco*R1 site. Fragment sizes are given in kilobases. Orientation with respect to the X chromosome is indicated. More detailed analysis of the genomic organization is presented in Chapter 6.

```
1    gatgtgatatggctgcaagtgcctttgacccttttgtctcccttccataaactgaaatacctaagctgctccaacctccttttttgtctt
91   ttgtttcataaatccttttcccattgcacatcaactcctgtctctctttgtactgtcactctcatctgttgctttccattcacactgcctt
181  tagccactcatcattttgtgcctacaccacagaaacctctgaatgtaatggatgttcctaccagaggacaagtcgtacaatggtggagga
                                                    @
271  ataggttcttcaaataggatcatggacttcttggaggagccaatccctggtgtagggacctatgatgatttcaatacaattgattgggtg
1                        M  D  F  L  E  E  P  I  P  G  V  G  T  Y  D  D  F  N  T  I  D  W  V

361  agagagaagtctcgagaccgggataggcaccgagagattaccaataaaagcaaagagtcaacatgggccttaattcacagtgtgagtgat
24    R  E  K  S  R  D  R  D  R  H  R  E  I  T *N  K  S  K  E  S  T  W  A  L  I  H  S  V  S  D

451  gctttttccggctggttgttgatgctccttattgggcttttatcaggttcgttagctggtttgatagacatctctgctcattggatgaca
54    A  F  S  G  W  L  L  M  L  L  I  G  L  L  S  G  S  L  A  G  L  I  D  I  S  A  H  W  M  T
                                                  D1
541  gacttaaaagaaggtatatgcacaggggattctggtttaaccatgaacattgttgctggaactctgagcatgtcacctttgaagagaga
84    D  L  K  E  G  I  C  T  G  G  F  W  F  N  H  E  H  C  C  W  N  S  E  H  V  T  F  E  E  R

631  gacaaatgtccagagtggaatagttggtcccagcttatcatcagcacagatgagggagcctttgcctacatagtcaattatttcatgtac
114   D  K  C  P  E  W  N  S  W  S  Q  L  I  I  S  T  D  E  G  A  F  A  Y  I  V  N  Y  F  M  Y

721  gtcctctgggctctcctatttgccttccttgccgtatctcttgtcaaggtgtttgcgccttatgcctgtggctctggaatccctgagata
144   V  L  W  A  L  L  F  A  F  L  A  V  S  L  V  K  V  F  A  P  Y  A  C  G  S  G  I  P  E  I
                 D2
811  aaaactatcttgagtggtttcattattaggggctatttgggtaagtggactctggttatcaaaaccatcaccttggtgctggcagtgtcg
174   K  T  I  L  S  G  F  I  I  R  G  Y  L  G  K  W  T  L  V  I  K  T  I  T  L  V  L  A  V  S
                                                  D3
901  tctggcttgagcctgggcaaagagggccctctagtgcacgtggcttgctgctgtgggaacatcctgtgccactgcttcaacaaatacagg
204   S  G  L  S  L  G  K  E  G  P  L  V  H  V  A  C  C  C  G  N  I  L  C  H  C  F  N  K  Y  R
                                                  D4
991  aagaatgaagccaagcgcagagaggtcttgtcggctgcagcagcagctggtgtatctgtagcctttggagcacctataggtggagtatta
234   K  N  E  A  K  R  R  E  V  L  S  A  A  A  A  A  A  G  V  S  V  A  F  G  A  P  I  G  G  V  L
                                                  D5
1081 ttcagccttgaagaggtcagctactattttcccctcaaaacattgtggcgttcattctttgctgccttggtggcagcattcactctacgc
264   F  S  L  E  E  V  S  Y  Y  F  P  L  K  T  L  W  R  S  F  F  A  A  L  V  A  A  F  T  L  R
                                                  D6
1171 tccatcaatccatttgggaacagccgcctggtactattttatgtggagtttcacaccccatggcatctctttgagctcgtgccattcatt
294   S  I  N  P  F  G  N  S  R  L  V  L  F  Y  V  E  F  H  T  P  W  H  L  F  E  L  V  P  F  I

1261 ctgctgggcatatttggtggtctgtggggagcactgtttatccgcacaaacattgcctggtgtcggaagcgaaagaccacccagttgggc
324   L  L  G  I  F  G  G  L  W  G  A  L  F  I  R  T  N  I  A  W  C  R  K  R  K (T) (T) Q  L  G
                 D7
1351 aagtatcctgttatagaggtactcgtcgtgacagccatcactgccatcctggctttccccaatgaatacactcggatgagcacaagtgag
354   K  Y  P  V  I  E  V  L  V  V  T  A  I  T  A  I  L  A  F  P  N  E  Y  T  R  M (S) T  S  E
                                 D8
1441 ctcatttctgagctgtttaatgactgtggccttctggactcctccaagctctgtgattatgagaaccgtttcaacacaagcaaaggggt
384   L  I  S  E  L  F  N  D  C  G  L  L  D  S  S  K  L  C  D  Y  E  N  R  F *N  T  S  K  G  G

1531 gaactgcctgacagaccggctggcgtgggagtctacagtgcaatgtggcagctggctttaacactcatactgaaaattgtcattactata
414   E  L  P  D  R  P  A  G  V  G  V  Y  S  A  M  W  Q  L  A  L  T  L  I  L  K  I  V  I  T  I
                                                  D9
1621 ttcacctttggcatgaagatcccttctggcctctttatccctagcatggctgttggtgctatagcaggtcgacttctaggagtaggaatg
444   F  T  F  G  M  K  I  P  S  G  L  F  I  P  S  M  A  V  G  A  I  A  G  R  L  L  G  V  G  M
                                                 D10
1711 gaacagctggcttattaccaccaggaatggaccgtcttcaatagctggtgtagtcagggagctgattgcatcacccccggcctttatgca
474   E  Q  L  A  Y  Y  H  Q  E  W  T  V  F  N  S  W  C  S  Q  G  A  D  C  I  T  P  G  L  Y  A

1801 atggttggggctgcagcctgcttaggtggggtgactcggatgactgtttctcttgttgtcataatgtttgaactgactggtggcttagaa
504   M  V  G  A  A  A  C  L  G  G  V  T  R  M  T  V  S  L  V  V  I  M  F  E  L  T  G  G  L  E
                                 D11
1891 tacatcgtgcctctgatggctgcagccatgacaagcaagtgggtggcagatgctcttgggcgggagggcatctatgatgcccacatccgt
534   Y  I  V  P  L  M  A  A  A  M  T  S  K  W  V  A  D  A  L  G  R  E  G  I  Y  D  A  H  I  R
                       D12
1981 ctcaatggataccccttttcttgaagccaaagaagagtttgctcataagaccctggcaatggatgtgatgaaaccccggagaaatgatcct
554   L  N  G  Y  P  F  L  E  A  K  E  E  F  A  H  K  T  L  A  M  D  V  M  K  P  R  R  N  D  P

2071 ttgttgactgtccttactcaggacagtatgactgtggaagatgtagagaccataatcagtgaaaccacttacagtggcttcccagtggtg
594   L  L  T  V  L  T  Q  D  S  M  T  V  E  D  V  E  T  I  I  S  E  T  T  Y  S  G  F  P  V  V

2161 gtatcccgggagtcccaaagacttgtgggctttgtcctccgaagagatctcattatttcaattgaaaatgctcgaaagaaacaggatggg
624   V  S  R  E  S  Q  R  L  V  G  F  V  L  R  R  D  L  I  I  S  I  E  N  A  R  K  K  Q  D  G

2251 gttgttagcacttccatcatttatttcacggagcattctcctccattgccaccatacactccacccactctaaagcttcggaacatcctc
654   V  V  S  T  S  I  I  Y  F  T  E  H  S  P  P  L  P  P  Y  T  P  P  T  L  K  L  R  N  I  L

2341 gatctcagccccttcactgtgactgaccttacacccatggagatcgtagtggatattttccgaaagctgggactgcggcagtgcctggtt
684   D  L  S  P  F  T  V  T  D  L  T  P  M  E  I  V  V  D  I  F  R  K  L  G  L  R  Q  C  L  V

2431 acacacaacgggcgattgcttggaatcattaccaaaaaggatgtgttaaagcatatagcacagatggcgaaccaagatcctgattccatt
714   T  H  N  G  R  L  L  G  I  I  T  K  K  D  V  L  K  H  I  A  Q  M  A  N  Q  D  P  D  S  I

2521 ctcttcaactagaatcatagagttctggatgtaaagcgggaaggacattacagaccatggatatgttgtttaacggtacccaaaacacat
744   L  F  N  @

2611 tttccatatttggatggtgaagtcacattagtgtgttgtctctttcctacaagttaaccagttgcactacataatctctggaaattaatt
2701 ttctctttaggagaaattatagttaggcttccatgatgttacattaggaagatatcatgaaagaataaataagattgctatggtttaatt
2791 atatttgctttttaaaagatttttttaacttaaaaagtagttagccaatatgcaatcactgaaaactatgcaagagaaattccaaccgtc
2881 ctgacctataacctgtaggaaaccgacgaaaaagtcactcttttgggatctaactgttgttactggaagacgaaggtaaactaaggggct
2971 ttgcttttcaaaccagagaaaggaaagccagaaggaaaagagtaatggtattttctagactgtgaagattcagttcaaatgttatccttg
3061 ttcctgttacaatatttagcattattagtttgttatgtgtgtatgtttatgttaattttaatttctgattataagacaatgctgctttgg
3151 ttaatctcttctaaaggaattta
```

Figure 5.18: Complete nucleotide and predicted amino acid sequence of cDNA clones spanning the ORF of CLCN5. Stop codons are indicated by @. Putative hydrophobic domains D1-D12 are underlined. Potential *N*-linked glycosylation sites are shown with asterisks. (S) and (T) represent consensus phosphorylation sites at Ser-380, Thr-349 and Thr-350 (see text).

Kozak consensus	$G_{44}C_{39}C_{53}R_{97}C_{49}C_{55}\mathbf{A}_{100}\mathbf{T}_{100}\mathbf{G}_{100}G_{46}$
1st ATG (232-234)	T A A T G G **A** **T** **G** T
2nd ATG (292-294)	A G G A T C **A** **T** **G** G

Figure 5.19: Comparison of sequence around potential CLCN5 initiation sites to the Kozak consensus. The latter was derived from analysis of 699 vertebrate mRNAs; the number after each nucleotide indicates the percentage of initiator sequences analysed which contain that nucleotide at that position (Kozak, 1987). Nucleotides in the CLCN5 sites which are identical to the Kozak consensus are underlined. The initiator methionine codons are shown in boldface.

The predicted amino acid sequence encoded by the complete ORF shows significant similarity to all other known members of the ClC family (Fig. 5.21 and Table 5.5). Sequence comparisons indicate that ClC-5 is more closely related to ClC-3 and ClC-4 (to which it shows ~77-78% identity) than to other members of the family (~28-32% identity) (Fig. 5.22). The hydropathy profile of ClC-5 is very similar to that of other proteins (Fig. 5.20), and predicts a topology involving 12 putative transmembrane domains (Fig. 5.18). Homology between ClC-5 and other ClC proteins is particularly high in these putative membrane spans.

There are two potential *N*-linked glycosylation sites in the predicted ClC-5 protein, one at position 38, close to the amino terminus, the other at position 408, in the loop between putative domains D8 and D9. A consensus phosphorylation site for cAMP dependent protein kinase (Pearson and Kemp, 1991) at Ser-380 is also found in the D8-D9 loop (Fig 5.18). In addition, Thr-349 and Thr-350 provide less typical sites for potential phosphorylation by this kinase.

	Human gene				Rat gene				
Locus	**Reference**	**S (%)**	**I (%)**	**Tissue specificity**	**Reference**	**S (%)**	**I (%)**	**Tissue specificity**	**Putative function**
CLCN4	van Slegtenhorst *et al.* (1994)	89.2	78.3	skeletal muscle and brain; lower levels in heart	Jentsch *et al.* (1995)	89.4	78.3	liver and brain; lower levels in heart, muscle, kidney and spleen	unknown
CLCN3	Borsani *et al.* (1995)	88.5	76.8	strongest expression in brain and skeletal muscle; also in heart, placenta, lung, liver, kidney and pancreas	Kawasaki *et al.* (1994)	88.3	76.8	brain; lower levels in liver, pancreas, kidney and adrenal gland	may play a role in neuronal function through regulation of membrane excitability; Kawasaki *et al.*, (1994) reported inhibition of channel by protein kinase C
CLCNKa	Kieferle *et al.* (1994)	51.9	27.7	kidney	two extremely similar rat homologues; rClC-K1 (Uchida *et al.*, 1993) and rClC-K2 (Adachi *et al.*, 1994; Kieferle *et al.*, 1994)	52.4 / 49.5*	28.0/ 27.1*	kidney	rClC-K1 is regulated by dehydration and may be involved in the concentration of urine (Uchida *et al.*, 1993); however, it is unclear which of the rat genes, rClC-K1 or rClC-K2, is the correct homologue of CLCNKa
CLCNKb	Kieferle *et al.* (1994)	52.6	28.1	kidney	as described above, there are two extremely similar rat homologues; rClC-K1 and rClC-K2	52.4 / 49.5*	28.0/ 27.1*	kidney	rClC-K1 may be involved in urine concentration mechanisms; however, it is unclear which of the rat genes, rClC-K1 or rClC-K2, is the correct homologue of CLCNKb
CLCN2	Cid *et al.* (1995)	56.9	30.8	ubiquitous	Thiemann *et al.* (1992)	57.8	32.0	ubiquitous	activated by cell swelling; probably involved in volume regulation in a wide range of tissues
CLCN1	Koch *et al.* (1992)	53.8	30.4	skeletal muscle	Steinmeyer *et al.* (1991a)	52.7	28.7	skeletal muscle	maintains resting chloride conductance of skeletal muscle; mutations are associated with recessive and dominant myotonia (Koch *et al.*, 1992; George *et al.*, 1993)

Table 5.5: Characteristics of mammalian members of the ClC family. Details are shown for human genes and their rat counterparts. **S**, the percentage similarity to CLCN5 and **I**, the percentage identity with CLCN5, are given for the amino acid sequences of each channel. These were determined using the GAP computer program. See section 5.2.7 for a description of this program and definitions of similarity/identity. * similarities and identities to CLCN5 are given for both rat kidney loci (rClC-K1/rClC-K2), since it is unknown which of these is the true homologue of which CLCNK human gene.

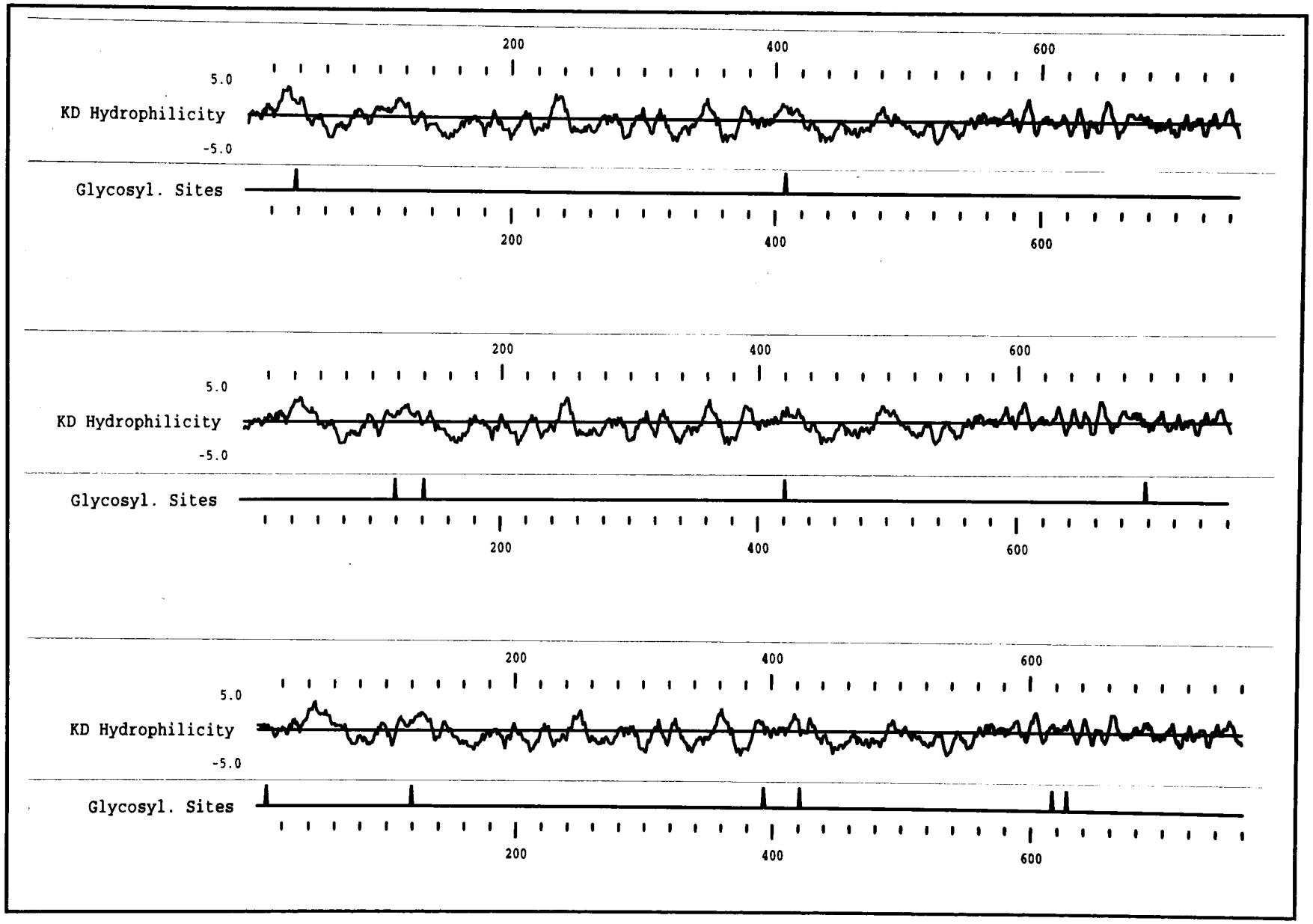


Figure 5.20: Hydropathy plots and positions of potential glycosylation sites for the chloride channels ClC-5 **(top)**, ClC-4 **(middle)** and ClC-3 **(bottom)**. The amino acid sequences of each channel were analysed using the programs PEPTIDESTRUCTURE and PLOTSTRUCTURE of the GCG software package (Section 5.2.7). When the hydropathy curve is in the upper half of the frame, it indicates a hydrophilic region; when in the bottom half, a hydrophobic region. See text for further discussion of hydrophobic domains and glycosylation sites.

```
                                                                                        --------D1--------
5  .............MDFLEEPIPGV..GTYDDFNTIDWVREKSRDRDRHREITNKSKESTWALIHSVSDAFS.GWLLMLLIGLLSGSLAGLIDISAHWMTDLK
4  mvnagamsgsgnlMDFLdEPfPdV..GTYeDFhTIDWlREKSRDtDRHRkITsKSKESiWefIkSllDAwS.GWvvMLLIGLLaGtLAGvIDlavdWMTDLK
3  mtnggsinssthllDlLdEPIPGV..GTYDDFhTIDWVREKckDReRHRrInsKkKESaWemtkSlyDAwS.GWLvvtLtGLaSGaLAGLIDIaAdWMTDLK
Ka ................mEelvGlreGfsgDpvTlqelwgpcphi......rraiqgglewLkqkvfr.lgedWyflmtlGv....LmaLvsyamnfai...

                                                            -------------D2--------------            ----
5  EGICTGGFWFNHEHCCWNSEHVTFEERDKCPEWNSWSQLIISTDEGAFAYIVNYFMYVLWALLFAFLAVSLVKVFAPYACGSGIPEIKTILSGFIIRGYLGK
4  EGvClsaFWysHEqCCWtSnetTFEdRDKCPlWqkWSeLlvnqsEGAsAYIlNYlMYiLWALLFAFLAVSLVrVFAPYrCGSGIPEIKTILSGFIIRGYLGK
3  EGIClsalWyNHEqCCWgSnetTFEERDKCPqWktWaeLIIgqaEGpgsYImNYiMYifWALsFAFLAVSLVKVFAPYACGSGIPEIKTILSGFIIRGYLGK
Ka .Gcvvrahqwly.............................rEigdshllrYlswtvypvalvsfssgfsqsitPssgGSGIPElKTmLaGvIledYLdi

   -----D3---------          --------D4--------          -----------D5------------                 ----
5  WTLVIKTITL..VLALSSGLSLGKEGPLVHVACCCGNILCHCFNKYR....KNEAKRRELLSAAAAAGVSVAFGAPIGGVLFSLEEVSYYFPLKTLWRSFFA
4  WTLlIKTvTL..VLvvSSGLSLGKEGPLVHVACCCGNffsslFsKYs....KNEgKRREvLSAAAAAGVSVAFGAPIGGVLFSLEEVSYYFPLKTLWRSFFA
3  WTLmIKTITL..VLAVaSGLSLGKEGPLVHVACCCGNifsylFpKYs....tNEAKkREvLSAAsAAGVSVAFGAPIGGVLFSLEEVSYYFPLKTLWRSFFA
Ka knfgaKvvgLsctLAtgStLfLGKvGPfVHlsvmiaayLgrvrtttigepenk.sKqnEmLvAAAAvGVatvFaAPfsGVLFSiEvmSshFsvrdyWRgFFA

   ----D6--------                       -----------D7------------                  --------D8--------
5  ALVAAFTLRSINPFGNSRLV...LFYVEFHT..PWHLFELVPFILLGIFGGLWGALFI...RTNIAWCRKRKTTQ..LGKYPVIEVLVVTAITAILAFPNEY
4  ALVAAFTLRSINPFGNSRLV...LFYVEyHT..PWymaELfPFILLGvFGGLWGtLFI...RcNIAWCRrRKTTr..LGKYPVlEViVVTAITAIiAyPNpY
3  ALVAAFvLRSINPFGNSRLV...LFYVEyHT..PWyLFELfPFILLGvFGGLWGAfFI...RaNIAWCRrRKsTk..fGKYPVlEViiVaAITAviAFPNpY
Ka AtcgAFifRllavFnseqetitsLyktsfrvdvPfdLpEiffFvaLGgicGvlscaylfcqRTflsfiktnrysklLatskpvysalaTlllAsityPpgv

                                             *                   --------D9---------   --------D10--------
5  TRMSTSEL.....ISELFNDCGLLDSSKLCDYENRFNTSK.GGELPDRPAGVGVYSAMWQLALTLILKIVITIFTFGMKIPSGLFIPSMAVGAIAGRLLGVG
4  TRqSTSEL.....ISELFNDCGaLeSSqLCDYiNdpNmtrpvddiPDRPAGVGVYtAMWQLALaLIfKIVvTIFTFGMKIPSGLFIPSMAVGAIAGRmvGiG
3  TRlnTSEL.....IkELFtDCgpLeSSsLCDYrNdmNaSKivddiPDRPAGiGVYSAiWQLcLaLIfKIimTvFTFGiKvPSGLFIPSMAiGAIAGRivGia
Ka ghflaSrLsmkqhldsLFdnhswalmtqnssppwpeeldpqhlwwewyhprftifgt...LaffLvmKfwmlIlattipmPaGyFmPifilGAaiGRLLG..

                                   ----------D11-----------     ---------D12---------
5  MEQLAYYHQEWTVFNSWCSQGADCITPGLYAMVGAAACLGGVTRMTVSLVVIMFELTGGLEYIVPLMAAAMTSKWVADALGREGIYDAHIRLNGYPFLEAKE
4  vEQLAYhHhdWyyFrnWCrpgADCvTPGLYAMVGAAACLGGVTRMTVSLVVIMFELTGGLEYIVPLMAAAvTSKWVADAfGkEGIYeAHIhLNGYPFLdvKd
3  vEQLAYYHhdWfiFkeWCevGADCITPGLYAMVGAAACLGGVTRMTVSLVVIvFELTGGLEYIVPLMAAvMTSKWVgDAfGREGIYeAHIRLNGYPFLdAKE
Ka .EaLAvafpEgiVtgg....vtnpImPGgYAlaGAAAfsGaVTh.TiStallaFELTGqivhalPvlmAvlaanaiaqsc.qpsfYDgtIivkklPyL...p

5  EFAHKTLAMDVMKPRRNDPLLTVLTQDSMTVEDVETIISETTYSGFPVVVSRESQRLVGFVLRRDLIISIENARKKQDGVVSTSIIYFTEHSPPLPPYTPPT
4  EFtHrTLAtDVMrPRRgePpLsVLTQDSMTVEDVETlIkETdYnGFPVVVSRdSeRLiGFaqRReLIlaInNARqrQeGiVSnSImYFTEepPeLPansPhp
3  EFtHtTLAaDVMrPRRNDPpLaVLTQDnMTVddiEnmInETsYnGFPVimSkESQRLVGFaLRRDLtIaIEsARKKQeGiVgsSrvcFaqHtPsLPaesPrp
Ka rilgrnigshhvrvehfmnhsittlakdtplEeVvkvvtsTdvteyPlVeStESQiLVGiVqRaqLvqalqae.................pPsraPghqqc

5  LKLRNI....LDLSPFTVTDLTPMEIVVDIFRKLGLRQCLVTHNGRLLGIITKKDVLKHIAQMANQDPDSILFN
4  LKLRrI....LnLSPFTVTDhTPMEtVVDIFRKLGLRQCLVTrsGRLLGIITKKDVLrHmAQMANQDPeSImFN
3  LKLRsI....LDmSPFTVTDhTPMEIVVDIFRKLGLRQCLVTHNGRLLGIITKKDiLrHmAQtANQDPaSImFN
Ka LqdilargcptepvtlTlfseTtlhqaqnlFklLnLqslfVTsrGRavGcvswvemkKaisnltNppapk....
```

Figure 5.21: Comparison of protein sequence encoded by CLCN5 with those of other human chloride channel genes. Alignment is shown for ClC-5, its two closest relatives, ClC-4 (van Slegtenhorst *et al.*, 1994) and ClC-3 (Borsani *et al.*, 1995), and one of the other kidney chloride channels, ClC-Ka (Kieferle *et al.*, 1994). Uppercase letters indicate identity with ClC-5. Letters in boldface represent amino acids which are identical in all four peptide chains. The positions of putative hydrophobic domains are shown. The asterisk indicates an *N*-linked glycosylation site, the position of which is conserved between ClC-5, ClC-4 and ClC-3. This diagram is adapted from an alignment of all known ClC genes obtained using the PILEUP program of the GCG software package.

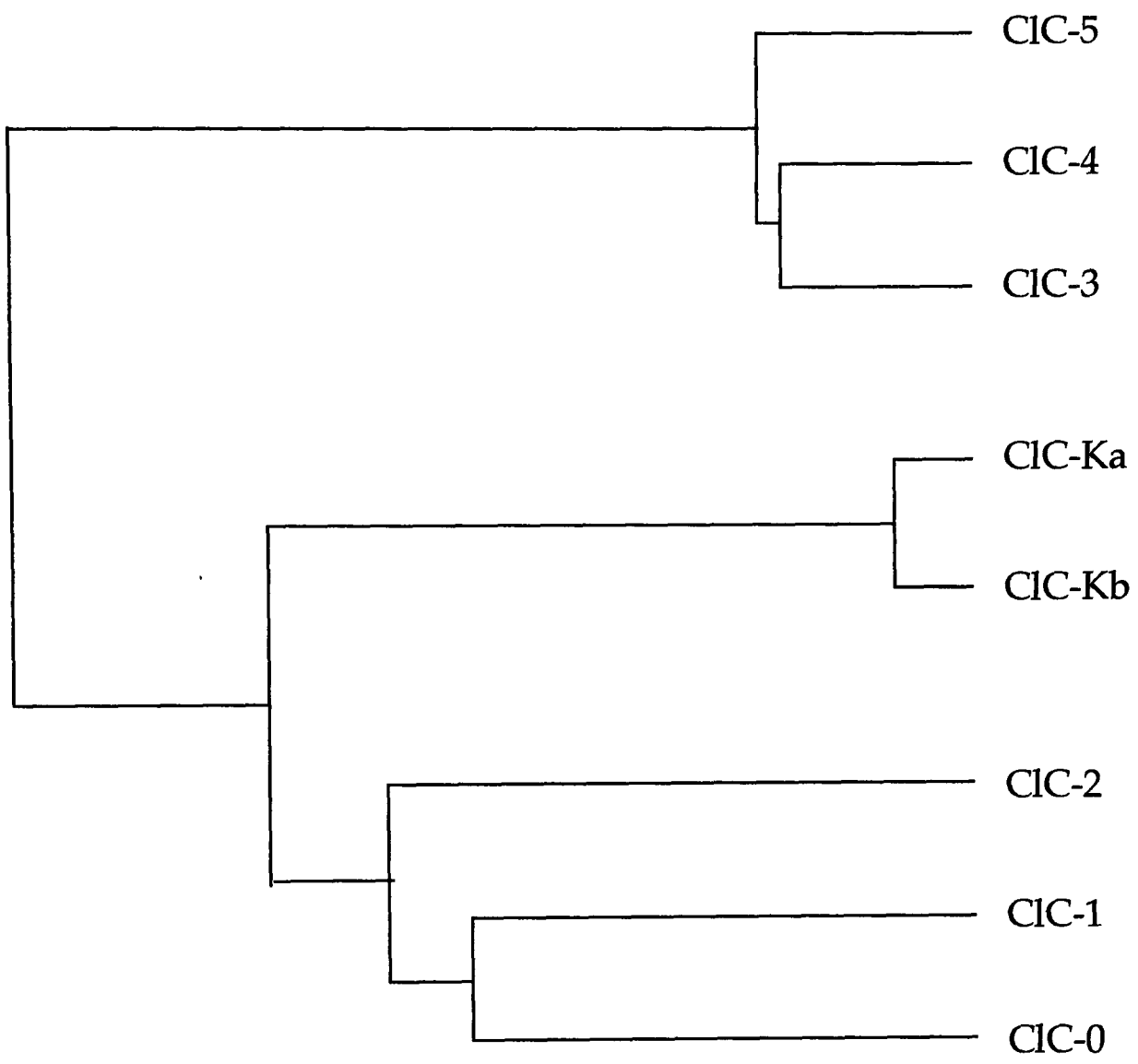


Figure 5.22: Dendrogram showing the relationship between ClC-5 and other members of the ClC-family. This was adapted from the output of the PILEUP program of the GCG software package. The chloride channels shown are from human (Borsani *et al.*, 1995; Cid *et al.*, 1995; Kieferle *et al.*, 1994; Steinmeyer *et al.*, 1991a; van Slegtenhorst *et al.*, 1994), except for ClC-0, the original member of this family, which is from the *Torpedo marmorata* electric organ (Jentsch *et al.*, 1990). It should be noted that this is a representation of clustering on the basis of sequence similarity and not a phylogenetic tree.

5.4 Discussion

A strategy involving the hybridization of the 6129 YAC to a kidney-specific cDNA library was succesful in identifying a strong candidate gene (CLCN5) for Dent's disease. The initial evidence implicating this gene was as follows:

- The genomic sequences detected by the CLCN5 coding region (which span 25-30kb) map entirely within the microdeletion associated with Dent's disease (Section 5.3.4).

- This microdeletion has been shown to be less than 370kb (Section 5.3.1). The age of onset and severity of Dent's disease in deleted individuals does not differ significantly from that found in non-deleted patients, and no other clinical disorders are associated with the deletion (Section 5.1.1). It therefore seems likely that the disease in pedigree C results from the deletion of a single, major locus.

- Northern analysis and homology studies indicate that CLCN5 is likely to encode a kidney-specific chloride channel (Section 5.3.3). This putative role suggests a plausible mechanism whereby loss of ClC-5 activity may result in the clinical phenotype observed in Dent's disease (see General Discussion).

Small deletions and point mutations in the coding region of CLCN5 which are predicted to disrupt the function of this gene have now been identified in patients from all of the pedigrees described in Section 5.1.1 (S. E. Lloyd, personal communication). In addition, analysis of CLCN5 in patients from the New York and Italian pedigrees (Section 5.1.2), employing a series of PCR primers based on the information obtained in this thesis, indicate that the disorders referred to as XRN and XLRH are due to the same primary gene defect as Dent's disease (S. E. Lloyd, personal communication).

Sequence analysis suggests that ClC-5 is a new addition to a rapidly expanding family of voltage-gated chloride channels (Table 5.5). These transmembrane proteins play an important role in various cellular functions such as regulation of cell volume, control of excitability and transepithelial transport (Jentsch *et al.*, 1995). The family was originally defined by the product of the ClC-0 locus, which was isolated from *Torpedo* electric organ using expression cloning (Jentsch *et al.*, 1990). At least six members of this family, in addition to ClC-5, are now known in mammals (Table 5.5); most of these were identified by sequence homology-based strategies. Some are expressed in a tissue-specific manner such as CLCN1 (the major skeletal muscle chloride channel gene) (Steinmeyer *et al.*, 1991) while others, like the swelling-activated CLCN2 gene (Thiemann *et al.*, 1992), are more ubiquitously expressed. Homology studies indicate that CLCN5 belongs to a recently identified distinct branch of the family which also includes CLCN3 (Kawasaki *et al.*, 1994; Borsani et al., 1995), a gene abundantly expressed in brain, and CLCN4 (van Slegtenhorst *et al.*, 1994), which was isolated from Xp22.3.

Two additional members of the family are specific to human kidney, CLCNKa and CLCNKb (along with their rat homologues rClC-K1 and rClC-K2) (Kieferle *et al.*, 1994), but while these are very similar to each other (over 90% amino acid identity), they have only ~28% identity with ClC-5. Although the precise function of these kidney chloride channels has yet to be established, it has been suggested that rClC-K1 is involved in urinary concentration mechanisms (Uchida *et al.*, 1993).

Limitations of the commercial kidney cDNA library meant that several rounds of library rescreening using different parts of the transcript were necessary to assemble the complete coding region of CLCN5, and no single clone could be found to span the entire open reading frame. Knowledge of the sequence of the closely related CLCN4 human transcript was therefore useful in the analysis of new clones. Given the particularly high homology between some members of the ClC family, hybridization to an X-only hybrid and to YACs from the DXS255 contig was used to verify that each novel clone originated from the CLCN5 locus, rather than an alternative ClC gene.

In addition, the cDNA contig provided two fold (or more) coverage of almost all of the coding region, and this, combined with analysis of RT-PCR products in the mutational studies (S. E. Lloyd, personal communication) and PCR products in genomic organization studies (Chapter 6), has confirmed that the sequence presented in Figure 5.18 does represent a truly contiguous transcript from the CLCN5 locus. It is interesting to note that none of the cDNAs isolated during library rescreenings were found to originate from other members of the ClC family, even though a low level of kidney expression has been reported for CLCN4 and CLCN3 (van Slegtenhorst *et al.*, 1994, Borsani *et al.*, 1995). This suggests that in kidney tissue the CLCN5 mRNA is the predominant transcript from this branch of the voltage-gated chloride channels. (The homology to CLCN2, CLCNKa and CLCNKb, which are all expressed at high levels in kidney, is relatively weak; it was therefore unlikely that cDNAs from these genes would be isolated during rescreening.) Although a faint band of approximately 3.5kb (in addition to the main 9.5kb band) could be seen in the kidney track of a Northern probed with the RL.3 cDNA, there is no further evidence as yet for any alternatively spliced transcripts. The observation of clones containing parts of CLCN5 fused to different regions of human mitochondrial sequence is most likely to be due to a cloning artefact during library construction.

Each clone from the cDNA contig detected a different pattern of bands on hybridization to genomic DNA. It was therefore possible to establish that the order of exon-containing genomic *Eco*R1 fragments spanning the CLCN5 coding region (5'-3') is Xpter–12.0kb–7.7kb–2.3kb–0.8kb–4.4kb–Xcen.

Although the entire CLCN5 transcript as detected on Northern blots is ~9.5kb (Section 5.3.3) the region encoding the chloride channel is confined to only ~2.2kb (Section 5.3.4). This size discrepancy, which is partly accounted for by the observation of ~1.15kb of cDNA sequence beyond the 3' end of the ORF, suggests that there is also a large untranslated 5' region. Similar observations have been made for the CLCN4 gene, in which the 2.2kb open reading frame is found at the 3' end of a 7.5kb transcript (van Slegtenhorst *et al.*, 1994). The nature of the 5' untranslated regions of these transcripts is not known at present and their possible role in the control of ClC expression may provide an interesting avenue of further research.

The ClC proteins are structurally unrelated to other channel proteins. Hydropathy analysis of the ClC-5 amino acid sequence identifies twelve putative transmembrane domains (D1-D12) corresponding very closely to those domains predicted in all the other known voltage-gated chloride channels. However, recent studies on glycosylation have indicated that the conventional D1-D12 topological representation of these proteins (Figure 5.23a) may need to be revised. Consensus sites for *N*-linked glycosylation have been identified in the D8 and D9 linker of all ClC channels so far isolated, including ClC-5 (see Figures 5.18 and 5.21). The use of such sites was previously thought to be unlikely, since the conventional model predicts that this loop should be cytoplasmic, but it has recently been shown that the ClC-K proteins *are* indeed glycosylated in the D8-D9 segment, when translated *in vitro* in the presence of pancreatic microsomes (Kieferle *et al.*, 1994). An alternative model for ClC structure placing the D8-D9 loop on the outside was therefore recently proposed (Jentsch *et al.*, 1995), in which D4, the least hydrophobic of the putative domains, does not cross the membrane, and the broad hydrophobic D9-D12 region forms an odd number (3 or 5) of transmembrane spans (Figure 5.23b). Glycosylation of the ClC-3/ClC-4/ClC-5 branch of the family has not yet been studied and further analysis is required to establish which model most accurately represents the topology of ClC-5.

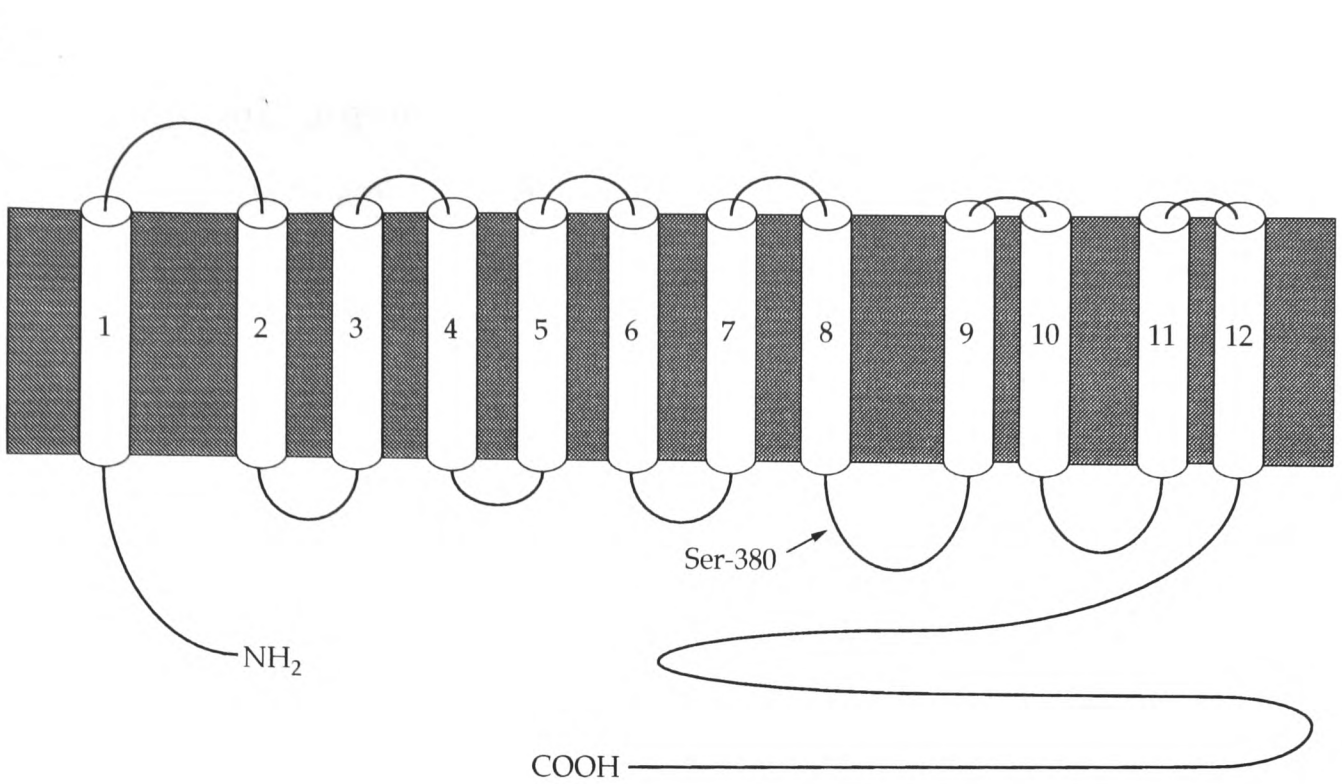


Figure 5.23: Two alternative models for the transmembrane topology of ClC-5, based on hydropathy plots and glycosylation studies (see text).

a) Conventional representation, involving 12 transmembrane domains and an intracellular C-terminus. Renal tubular chloride channels are regulated by cAMP-dependent protein kinase (Bae and Verkman, 1990; Gesek and Friedman, 1992, 1993) and there is a consensus site for phosphorylation by this enzyme at Ser-380 in the D8-D9 loop.

b) Preliminary revised model, adapted from Jentsch *et al.* (1995), which allows for glycosylation within the D8-D9 loop, while maintaining an intracellular C-terminus. In this representation D4, the least hydrophobic of the conventional domains, does not cross the membrane and the D8-D9 loop becomes extracellular. The precise topology of the D9-D12 region is unclear, but involves either 3 or 5 transmembrane spans. In this model the Ser-380 phosphorylation site cannot be used. However, the Thr-349 and Thr-350 residues in the D7-D8 loop are now intracellular, and therefore provide potential (though less typical) sites for phosphorylation by cAMP-dependent protein kinase.

Chapter 6 – Genomic organization of CLCN5, the gene implicated in Dent's disease

6.1 Introduction

6.1.1 The presence of introns in genomic DNA

Prior to the application of recombinant technology to eukaryotic single copy genes, it was generally assumed that messenger RNA molecules, which determine the primary sequences of proteins, were direct copies of gene sequences in the genomic DNA. However, in 1977, several different groups independently reported the presence of large inserts in the DNA coding for eukaryotic single copy structural genes. Comparison between the restriction sites in the rabbit β-globin cDNA clone and those at the corresponding locus in genomic DNA revealed that the coding region in the latter is interrupted by a 600bp segment (Jeffreys and Flavell, 1977). Hybridization of a recombinant phage clone containing the mouse genomic β-globin locus to globin mRNA followed by examination using an electron microscope demonstrated the presence of a similar insert in the mouse gene, observed as a loop in the middle of the DNA (Tilghman *et al.*, 1978). Studies of the chick ovalbumin gene showed that, while the cDNA molecule contained no sites for *Eco*R1 or *Hind*III, it detected several fragments on hybridization to genomic DNA cleaved with these enzymes, indicating that this, too, was a split gene (Breathnach *et al.*, 1977; Doel *et al.*, 1977). For both globin and ovalbumin, it was found that the positions and sizes of inserts were identical in all tissues examined, irrespective of whether the gene was active or inactive.

Since their initial discovery, it has become clear that the presence of such inserts (now known as 'introns') is a widespread phenomenon in eukaryotic genes, and that these introns are transcribed as part of precursor mRNA molecules and then excised by a post-transcriptional splicing mechanism (see below). It has also been shown that a given transcription unit can be spliced in a variety of patterns to give different isoforms of a protein in different cell types (see Rio, 1993). In addition, studies using *Drosophila melanogaster* have indicated that differential splicing mechanisms play an important role in the determination of distinct cell fates during development.

6.1.2 An overview of the spliceosome

The removal of introns from nuclear pre-mRNA is catalysed by a multi-component structure known as the spliceosome and involves two sequential transesterification steps:

i) The 5' splice site is cleaved, and the 5' end of the intron is joined, via a 2'-5' phosphodiester bond, to a conserved adenosine residue at the branchpoint sequence within the intron (Fig. 6.1a).

ii) The 3' splice site is cleaved, the 5' and 3' exons are joined, and the intron is released as a branched lariat RNA molecule (Fig. 6.1b-c).

The spliceosome consists of five small nuclear RNAs (U1, U2, U4, U5 and U6) complexed with numerous proteins, forming a structure of comparable size and complexity to the ribosome (reviewed in Rio, 1993 and Lamond, 1993). However, it is interesting to note that the organelles of plants and fungi contain introns which may be accurately and efficiently excised in the complete absence of any *trans*-acting factors. These self-splicing introns are classified into two types, group I and group II, and the chemical mechanism for excision of the latter is identical to that of spliceosome-catalysed introns (Fig. 6.1). Intramolecular base pairing in self-splicing introns forms conserved secondary and tertiary structures which bring together the appropriate regions involved in transesterification. By contrast, there is little sequence or structural conservation between different nuclear introns. This has led to the suggestion that the snRNAs of the spliceosome form intermolecular base paired structures which are equivalent to the conserved intramolecular domains of group II introns.

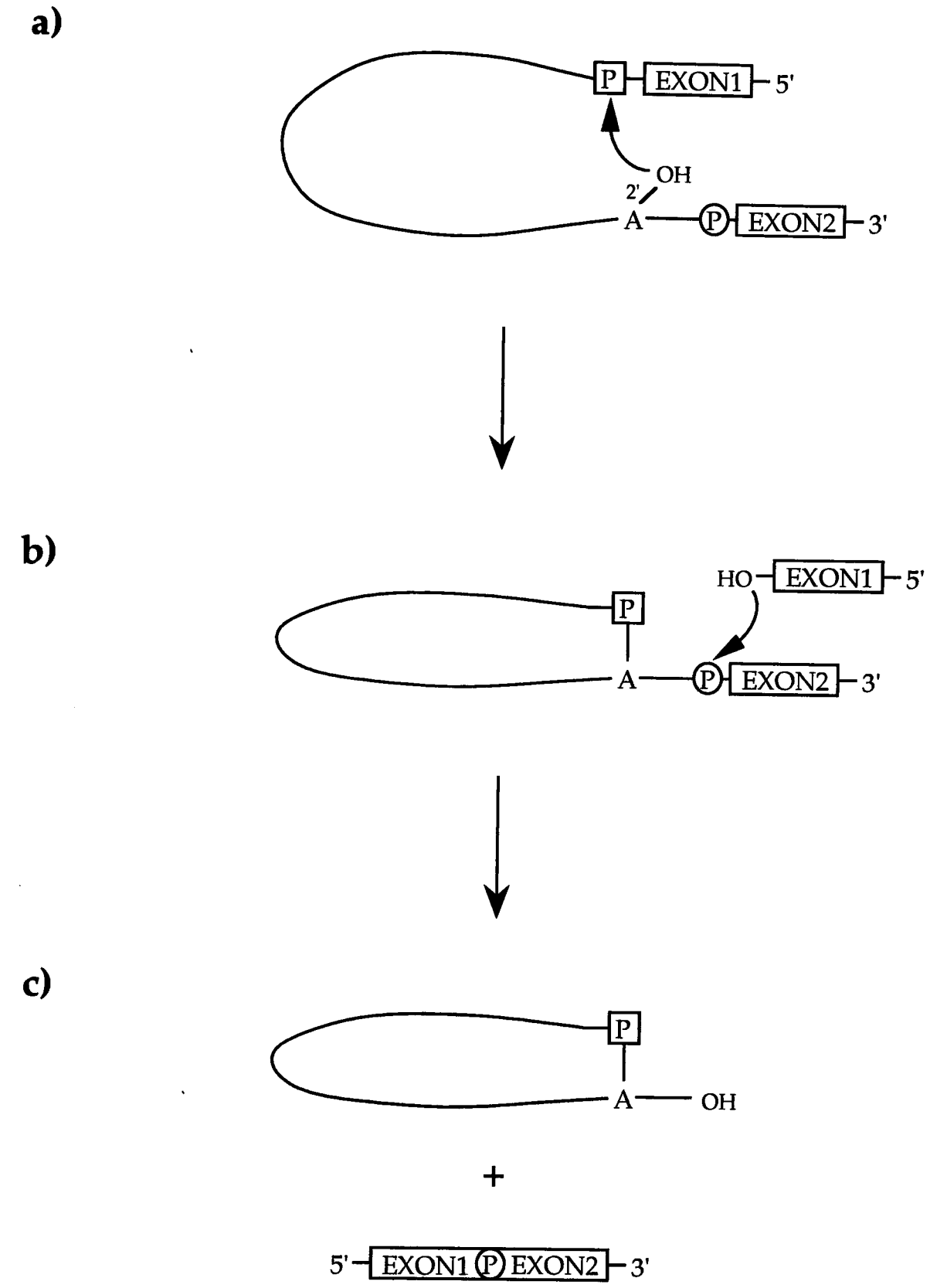


Figure 6.1: The two transesterification steps used for both spliceosome catalysed pre-mRNA introns and group II self-splicing introns.

a) First transesterification step: 2' hydroxyl of an adenosine ribose attacks 5' exon-intron junction.

b) Second transesterification step: 3' hydroxyl of free 5' exon attacks 3' intron-exon junction.

c) Fully excised intron, in 'lariat' form, and spliced mRNA.

Much of the analysis of the spliceosome has therefore focused on RNA-RNA interactions, as deduced from cross-linking experiments, phylogenetic comparisons, and studies using mutagenesis. The current view of spliceosome assembly and activity is summarized in Fig. 6.2.

- The U1 snRNA, which is important in the early steps of splice site recognition and pre-spliceosome assembly, base pairs with both the donor and acceptor sites, and may therefore play a role in bringing together the separate ends of the intron. A second snRNA, known as U2, is also present at this early stage, and binds to the branchpoint within the intron (Fig. 6.2a).

- The spliceosome is then joined by a preformed U4/U5/U6 tri-snRNP particle, in which U4 is tightly coupled with U6. The U5 snRNA base pairs with exon sequences flanking the 5' and 3' splice sites, which were brought together by U1, and the latter is released from the complex (Fig. 6.2b).

- U4 unwinds from U6, exposing sequences in the U6 snRNA which can base pair with the 5' end of U2, just upstream of the branchpoint recognition region. Phylogenetic studies indicate that U6 is the most highly conserved snRNA, and it has been suggested that it may play a direct catalytic role in the splicing mechanism. The structure of the U2/U6 complex closely resembles that of domain 5 and 6 of group IIA self-splicing introns, and the interaction juxtaposes the proposed catalytic residues of U6 with the 5' cleavage site of the intron (Fig. 6.2b).

Several different classes of protein have been identified which are involved in spliceosome activity (reviewed in Rio, 1993; Lamond, 1993; Horowitz and Krainer, 1994). One of these is the 'DEAD box' family of putative ATP-dependent RNA helicases, which may control the RNA-RNA interactions described above. The heterogeneous nuclear ribonucleoproteins (hnRNPs) package RNA in the nucleus and are important components in the early stages of spliceosome assembly.

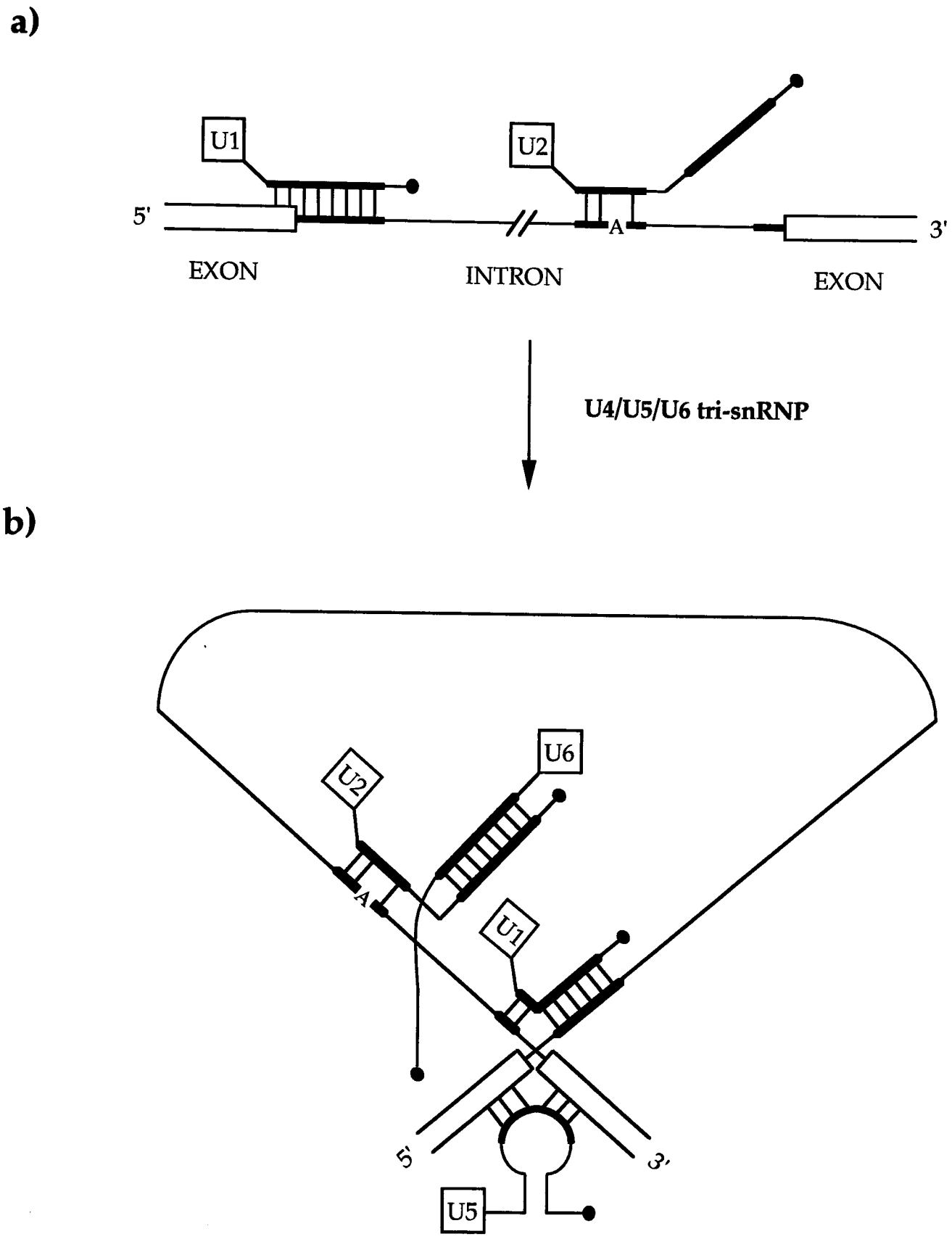


Figure 6.2: Schematic representation of RNA-RNA base-pairing interactions that may take place during splicing of the pre-mRNA. Note that not all the interactions illustrated take place simultaneously. 5' ends of snRNAs are indicated by filled circles. Adapted from Lamond (1993) and Rio (1993). See text for more details.

Members of a recently discovered group of proteins, known as the SR family, are necessary for constitutive splicing and are characterized by an RNA recognition motif (RRM) at the amino terminus and a domain rich in Arginine-Serine dipeptides at the carboxy terminus (Rio, 1993; Horowitz and Krainer, 1994). An example is the U2AF protein, which binds to intron polypyrimidine tracts (usually found adjacent to the 3' splice site) and facilitates the base pairing between U2 snRNA and the branchpoint. Studies of another SR protein, known as SF2/ASF, suggest that it is involved in the selection of splice sites and may bind to purine-rich sequences in the exon downstream of the 3' splice site (see Horowitz and Krainer, 1994).

6.1.3 Definition of splice sites

An essential element of RNA processing is the accurate and efficient identification of the appropriate splice sites. Studies have indicated that complementarity between splice sites and the U1 snRNA is a major determinant of splice site identity (Fig. 6.3). Further analysis has also implicated the U5 and U6 snRNAs, as well as the SR proteins and hnRNP A1 (see above) in splice site definition (reviewed in Horowitz and Krainer, 1994). Consensus sequences for donor and acceptor junctions have been derived from analysis of nearly 1800 human introns (Fig 6.3 and see Fig 6.18) and it has been shown that over 82% of the sequence conservation at these junctions is confined to the intronic side, which reduces constraints on codon choices at splice sites in coding regions (Stephens and Schneider, 1992). It should be noted that similarity to the consensus sequence is not the only determinant in splice site selection. For example, there is considerable variation in human 5' splice site sequences, with an average of seven of the nine bases matching the degenerate consensus (Fig. 6.3). Information analysis shows that this level of sequence similarity is insufficient to ensure correct identification of all 5' splice sites (Stephens and Schneider, 1992) indicating that elements outside the consensus region must influence site recognition. (For example, RNA secondary structure may provide greater accessibility for the spliceosome in specific regions of the pre-mRNA molecule.)

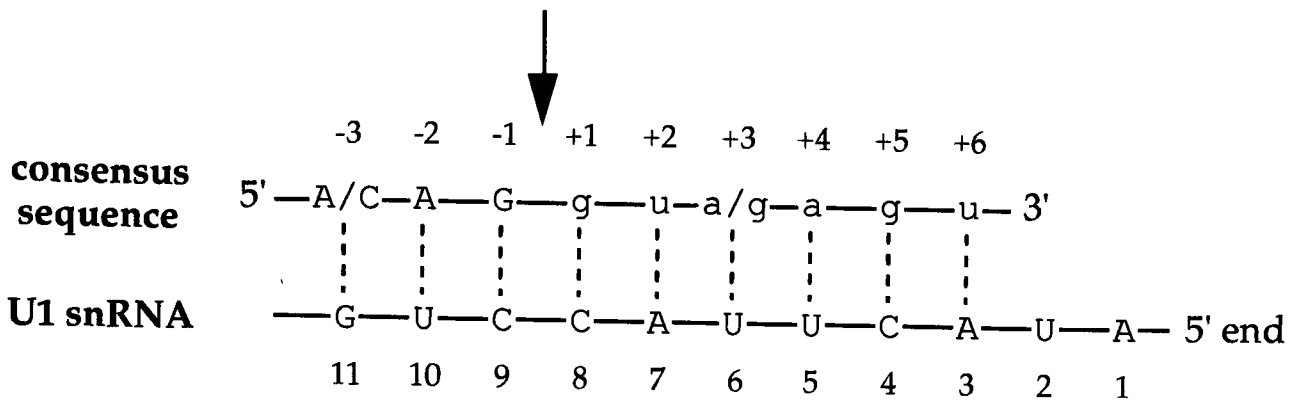


Figure 6.3: The consensus sequence for human 5' splice sites in pre-mRNA as derived from analysis of nearly 1800 introns (Stephens and Schneider, 1992) with the U1 snRNA aligned beneath it. Dotted lines represent proposed RNA-RNA base-pairing interactions in the pre-spliceosome complex. Exonic sequence is in uppercase, intronic in lowercase. The cleavage site is indicated by a vertical arrow. Adapted from Horowitz and Krainer (1994).

6.1.4 Theories on the evolution of introns

Given that nuclear introns do not code for protein, and there is very little conservation of sequence, length or structure between corresponding introns in different species, it seems likely that they are non-functional, and their presence in the genome needs to be explained from an evolutionary perspective. Soon after the discovery of introns, Gilbert (1978) proposed that the split gene organization of eukaryotes increases the ease of rearrangements that shuffle protein coding regions around to form novel mosaic genes, and therefore speeds up evolution. An extension of this 'exon shuffling' hypothesis predicts that each exon should encode a discrete domain of structure of function (Blake, 1978). There is now substantial evidence indicating that exon shuffling has played a major part in the evolution of some proteins; for example, a separate exon containing an epidermal-growth-factor-like domain is found in several otherwise unrelated proteins (Gilbert *et al.*, 1986). Furthermore, the shuffling of exons at the RNA processing level (differential splicing) is an important mechanism which underlies developmental pathways (Rio, 1993). However, it is clear that the long term selective advantages conferred by split genes are not sufficient to explain how they arose in the first place.

There are presently two predominant views on the origin of introns:

i) 'Introns early'

This hypothesis holds that introns are evolutionary relics, present in the common ancestor of prokaryotes and eukaryotes, and that they mark the positions where ancient exons were joined together to form the first modern-sized genes (Darnell and Doolittle, 1986). The lack of introns in modern day prokaryotes is explained by 'streamlining'; such cells use a large proportion of their energy for DNA replication, and there is therefore strong selective pressure in favour of the removal of non-functional sequences which increase the replication load.

This theory has been investigated using data from ancient genes that have a well characterized structure, such as globin, triose phosphate isomerase (TPI) and pyruvate kinase (reviewed in Mattick, 1994). The 'introns early' hypothesis predicts that there should be a connection between exon structure and the secondary or tertiary structure of the protein. Whilst it has been reported that the eleven exons of the proposed ancestral form of TPI would have each encoded a discrete unit of structural compactness (Gilbert *et al.*, 1986), a more comprehensive study of four proteins (including TPI), suggests that there is no correlation between intron position and protein structure (Stoltzfus *et al.*, 1994). An alternative approach involves comparisons between positions of introns in homologous genes from different species (Gilbert *et al.*, 1986). In general it has been found that most introns are not found in common positions between phylogenetically distant homologues (see Mattick, 1994). The few examples of intron conservation between plants, animals and fungi do not necessarily support the 'introns early' hypothesis, since they may have arisen during the early evolution of the eukaryotic cell, prior to the separation of these lineages.

ii) 'Introns late'

Although some introns have indeed been shown to be very ancient, strong evidence against general acceptance of the 'introns early' view has been provided by the analysis of multigene families (reviewed in Rogers, 1989). In several members of the immunoglobulin superfamily, the immunoglobulin-like domain (supposedly a classic example of an ancestral exon) is disrupted near its middle, by introns that can fall in any phase of the reading frame. Discordant introns have also been reported in the serine protease family and the calmodulin superfamily. Furthermore, in several genes which encode proteins containing tandemly repeated domains, introns are found in different positions in different domains.

Proponents of 'introns early' argue that discordant introns can be explained by a combination of differential loss of introns and intron 'slippage' (Kersanach *et al.*, 1993). Although it appears that some introns may have been cleanly removed during evolution, this cannot alone account for differences in position, because in many cases the proposed ancestral form of the gene would consist of a large number of extremely small exons (Mattick, 1994). Intron slippage also seems unlikely, since it would often involve movement across a nonintegral number of codons within highly conserved coding sequence, and separate frameshift mutations at each end of the intron would be needed (see Rogers, 1989).

The above evidence therefore suggests that many introns have inserted themselves into nuclear genes after the divergence of multigene families. This 'introns late' view is supported by the discovery of self-splicing introns in organelle DNA. As stated above, group II introns self-splice via a chemical mechanism that is identical to nuclear introns, and form secondary structures that are strikingly similar to the motifs of the spliceosomal snRNAs. In addition, group II introns have been found which are capable of self-insertion, and some of these encode their own reverse transcriptase, suggesting that they represent mobile genetic elements (Rogers, 1989).

It has thus been postulated that individual classical introns have evolved from group II introns which originated in mitochondrial or chloroplast DNA, and which self-inserted into nuclear DNA. The ability to self-splice would initially ensure that such an intron did not inactivate a gene into which it had inserted, and subsequent mutation would convert it into a classical intron, transferring control of the splicing operation from the autonomous intron to the hosts own mechanism. The conserved elements from the group II intron would then be unnecessary and would rapidly be lost. The RNA elements of the spliceosome itself are also thought to have evolved from group II introns (see Mattick, 1994).

It has been suggested that the lack of introns in protein coding genes of prokaryotes is not a consequence of selective pressure against an excessive DNA replication load, but instead reflects the fact that transcription and translation are intimately coupled in these cells, and the presence of introns would represent a major disruption to protein synthesis (Mattick, 1994). The sequestering of DNA into a nucleus in the first eukaryotic cells resulted in the decoupling of transcription and translation, and this would have reduced the constraints on any self-splicing mobile genetic elements that might insert themselves into coding regions.

It should be noted that the 'introns early' and 'introns late' hypotheses are not necessarily mutually exclusive. There are clear examples of both intron gain and intron loss during the course of evolution, suggesting that there might be a long-term balance between the two (Rogers, 1989). Ancestral genes may indeed have been mosaics, assembled from small open reading frames by recombination and RNA splicing, but it does not follow that all introns present today are remnants of these early events.

6.1.5 Aims

The objective of this part of the thesis was to investigate the genomic organization of the CLCN5 coding region (cloned in the cDNA contig which was assembled in Section 5.3.4). As described in Chapter 5, the finding of point mutations and deletions in this kidney-specific chloride channel has implicated it in three similar forms of X-linked hereditary renal tubular dysfunction. The identification of the exon-intron boundaries of the genomic locus, along with flanking intron sequences, was undertaken in order to facilitate direct PCR screening of genomic DNA from further patients manifesting this type of disorder.

Information on the genomic organization of CLCN5 is of significance for future structural and functional comparisons between the ClC family members, and may increase our understanding of how this family has evolved. In addition, data on the conservation of intron position in different ClC genes are relevant to the 'introns early'/'introns late' debate which was discussed above.

Whilst the region of the CLCN5 transcript covered by the cDNA contig contains no *Eco*R1 sites, hybridization results indicate that it spans five exon-containing *Eco*R1 fragments at the genomic level (Section 5.3.4), which suggests that there are a minimum of five exons for this part of the gene. A strategy involving the comparison of PCR products from cDNA and genomic templates was used to identify introns. Any products which were suspected to contain introns could be cloned via the technique of TA cloning® (Invitrogen), which exploits the nontemplate-dependent activity of *Taq* polymerase that adds a single deoxyadenosine to the 3' ends of all duplex molecules provided by PCR. These 3' A-overhangs are used to insert the PCR product into a vector which contains single 3' T-overhangs at its insertion site. Sequencing of the cloned products would then reveal whether they were legitimate genomic sequences from CLCN5, and precise exon-intron boundaries could be identified by comparison between cDNA and genomic sequences, aided by knowledge of the splice site consensus sequences (Stephens and Schneider, 1992).

6.2 Materials and methods

6.2.1 PCR conditions

Oligonucleotides of 19-22bp in length and 60°C melting temperature were selected from the CLCN5 cDNA sequence. Primer sequences are given in Fig. 6.4. PCR conditions were as described in Section 2.13.3, using Promega *Taq*. Cycling parameters were: 94°C, 5 mins ('hot start'); 94°C, 30 seconds; 55°C, 30 seconds; 74°C, 48 seconds; 30-35 cycles.

6.2.2 TA cloning® of PCR products

The TA cloning® kit (manufactured by Invitrogen) includes the pCR™ II vector, which contains an ampicillin resistance gene and the *lac*Z gene for blue-white colour selection of recombinants (see Section 2.10.2). The insert site is flanked on either side by an *Eco*R1 site; this enzyme can be used to excise the PCR product from the plasmid after cloning. The polylinker also contains sequences that are homologous to the M13 forward and reverse primers, which can therefore be used for sequencing.

1. Excise PCR product from a preparative gel and purify using the Geneclean procedure (Section 2.8).
2. To 6µl of this purified product add 1µl of 10 x ligation buffer, 2µl of pCR™ II vector (25ng/µl) and 1µl of T4 DNA ligase. Incubate for 4-16 hours at 14°C.
3. Use 2µl of ligated mix to transform bacterial cells using heat shock (Section 2.10.2).
4. Select for ampicillin resistant clones.
5. Pick white colonies for plasmid isolation, restriction analysis and sequencing.

Any blue colonies usually result from a T:T mismatch self-ligation of the vector. However, if cloning PCR products of 500bp or less, light blue colonies, rather than white, may appear. These colonies contain inserts which do not disrupt the reading frame of the *lac*Z gene, resulting in a fusion product which retains some activity.

6.3 Results

6.3.1 Identification of exon-intron boundaries using a PCR-based strategy

The **PRIMER** computer programme was used to select a series of overlapping primer pairs, each predicted to give a PCR product of 100-300bp, from the coding region of the CLCN5 transcript. Sequences and positions of the primers which were instrumental in establishing genomic organization are given in Figures 6.4 and 6.5. All oligonucleotides were designed to have the same *Tm* (60°C), so that the forward primer from one pair could be used in concert with the reverse primer of another (although checks for lack of primer-primer complementarity were not made for all possible combinations of oligonucleotides).

Primer pairs were used to amplify from cDNA, 6129 YAC, and human genomic templates. Since PCR is such a sensitive technique, control reactions containing no template were also set up, to show that there was no contamination in any of the components of the PCR mix. Each primer pair was found to give one of three possible patterns of results (see Figs. 6.6 and 6.7):

a) Product of the expected size is amplified from cDNA, human genomic, and YAC templates, but not in the water control. This suggests that there are no introns in the region bounded by the primer pair. If two sets of overlapping primer pairs give this result, then the forward primer from the most 5' pair is used in combination with the reverse primer from the 3' pair, in order to confirm that introns are absent from the interval. For example, this type of procedure was used to establish a lack of introns in the 1093F-1573R region; the intervening primers in such cases have been excluded from Figs. 6.4-6.5, in order to simplify the presentation of the data.

```
   1 tgatgtgatatggctgcaagtgcctttgacccttttgtctcccttccataaactgaaatacctaagctgctccaacctccttttttgtctt
     -------002F--------->>

  91 ttgtttcataaatcctttcccattgcacatcaactcctgtctctctttgtactgtcactctcatctgttgctttccattcacactgcctt
                                                          --------145F--------->>
                                                                 <<-------177R-------

 181 tagccactcatcattttgtgcctacaccacagaaacctctgaatgtaatggatgttcctaccagaggacaagtcgtacaatggtggagga
 271 ataggttcttcaaataggatcatggacttcttggaggagccaatccctggtgtagggacctatgatgatttcaatacaattgattgggtg
                        --------291F------>>
                         <<-------311R-------

                  XhoI
 361 agagagaagtctcgagaccgggataggcaccgagagattaccaataaaagcaaagagtcaacatgggccttaattcacagtgtgagtgat
 451 gctttttccggctggttgttgatgctccttattgggcttttatcaggttcgttagctggtttgatagacatctctgctcattggatgaca
      <<------470R-------
          -------456F----->>

 541 gacttaaaagaaggtatatgcacaggggattctggtttaaccatgaacattgttgctggaactctgagcatgtcacctttgaagagaga
                <<-------570R-------
                         -------561F------->>

 631 gacaaatgtccagagtggaatagttggtcccagcttatcatcagcacagatgagggagcctttgcctacatagtcaattatttcatgtac
                -------641F------->>                    <<-------700R------

 721 gtcctctgggctctcctatttgccttccttgccgtatctcttgtcaaggtgtttgcgccttatgcctgtggctctggaatccctgagata
            <<-------747R-------
                   -------735F------>>

 811 aaaactatcttgagtggtttcattattaggggctatttgggtaagtggactctggttatcaaaaccatcaccttggtgctggcagtgtcg
                                                                        -------889F-

 901 tctggcttgagcctgggcaaagagggccctctagtgcacgtggcttgctgctgtgggaacatcctgtgccactgcttcaacaaatacagg
     ----->>                                                          <<--------990R--------

 991 aagaatgaagccaagcgcagagaggtcttgtcggctgcagcagcagctggtgtatctgtagcctttggagcacctataggtggagtatta
                                                    -------1037F--------->>
                                                           <<-------1067R-------

1081 ttcagccttgaagaggtcagctactattttcccctcaaaacattgtggcgttcattctttgctgccttggtggcagcattcactctacgc
                -------1093F------->>

1171 tccatcaatccatttgggaacagccgcctggtactattttatgtggagtttcacaccccatggcatctctttgagctcgtgccattcatt
       <<------1192R-------

1261 ctgctgggcatatttggtggtctgtggggagcactgtttatccgcacaaacattgcctggtgtcggaagcgaaagaccacccagttgggc
1351 aagtatcctgttatagaggtactcgtcgtgacagccatcactgccatcctggctttccccaatgaatacactcggatgagcacaagtgag
1441 ctcatttctgagctgtttaatgactgtggccttctggactcctccaagctctgtgattatgagaaccgtttcaacacaagcaaaggggt

1531 gaactgcctgacagaccggctggcgtgggagtctacagtgcaatgtggcagctggctttaacactcatactgaaaattgtcattactata
     -1525F---->>              <<-------1573R------

1621 ttcacctttggcatgaagatcccttctggcctctttatccctagcatggctgttggtgctatagcaggtcgacttctaggagtaggaatg
                                                                 --------1696F--

1711 gaacagctggcttattaccaccaggaatggaccgtcttcaatagctggtgtagtcagggagctgattgcatcacccccggcctttatgca
     ----->>                                    <<-------1776R-------

1801 atggttggggctgcagcctgcttaggtggggtgactcggatgactgtttctcttgttgtcataatgtttgaactgactggtggcttagaa
                                          <<-------1859R--------

1891 tacatcgtgcctctgatggctgcagccatgacaagcaagtgggtggcagatgctcttgggcgggagggcatctatgatgcccacatccgt
1981 ctcaatggataccccttttcttgaagccaaagaagagtttgctcataagaccctggcaatggatgtgatgaaaccccggagaaatgatcct
                                                                -------2050F------>>

2071 ttgttgactgtccttactcaggacagtatgactgtggaagatgtagagaccataatcagtgaaaccacttacagtggcttcccagtggtg
2161 gtatcccgggagtcccaaagacttgtgggctttgtcctccgaagagatctcattatttcaattgaaaatgctcgaaagaaacaggatggg
                                <<-------2206R------          --------2224F-------->>

2251 gttgttagcacttccatcatttatttcacggagcattctcctccattgccaccatacactccacccactctaaagcttcggaacatcctc
                                                              <<-------2330R-------

2341 gatctcagccccttcactgtgactgaccttacacccatggagatcgtagtggatattttccgaaagctgggactgcggcagtgcctggtt
2431 acacacaacgggcgattgcttggaatcattaccaaaaaggatgtgttaaagcatatagcacagatggcgaaccaagatcctgattccatt
2521 ctcttcaactagaatcatagagttctggatgtaaagcgggaaggacattacagaccatggatatgttgtttaacggtacccaaaacacat
                          <<-------2561R------
                            -------2544F------>>

2611 tttccatatttggatggtgaagtcacattagtgtgttgtctctttcctacaagttaaccagttgcactacataatctctggaaattaatt
2701 ttctctttaggagaaattatagttaggcttccatgatgttacattaggaagatatcatgaaagaataaataagattgctatggtttaatt
2791 atatttgctttttaaaagatttttttaacttaaaaagtagttagccaatatgcaatcactgaaaactatgcaagagaaattccaaccgtc
2881 ctgacctataacctgtaggaaaccgacgaaaaagtcactcttttgggatctaactgttgttactggaagacgaaggtaaactaaggggct
2971 ttgcttttcaaaccagagaaaggaaagccagaaggaaaagagtaatggtattttctagactgtgaagattcagttcaaatgttatccttg
3061 ttcctgttacaatatttagcattattagtttgttatgtgtgtatgtttatgttaattttaatttctgattataagacaatgctgctttgg
                                                                        <<--------

3151 ttaatctcttctaaaggaattta
     -3161R------
```

Figure 6.4: Oligonucleotides used for PCR analysis of CLCN5 genomic structure, aligned beneath the cDNA sequence of the coding region, as established in Chapter 5. Nucleotides are numbered as in Fig. 5.18. An *Xho*I site at nucleotides 371-376 is indicated.

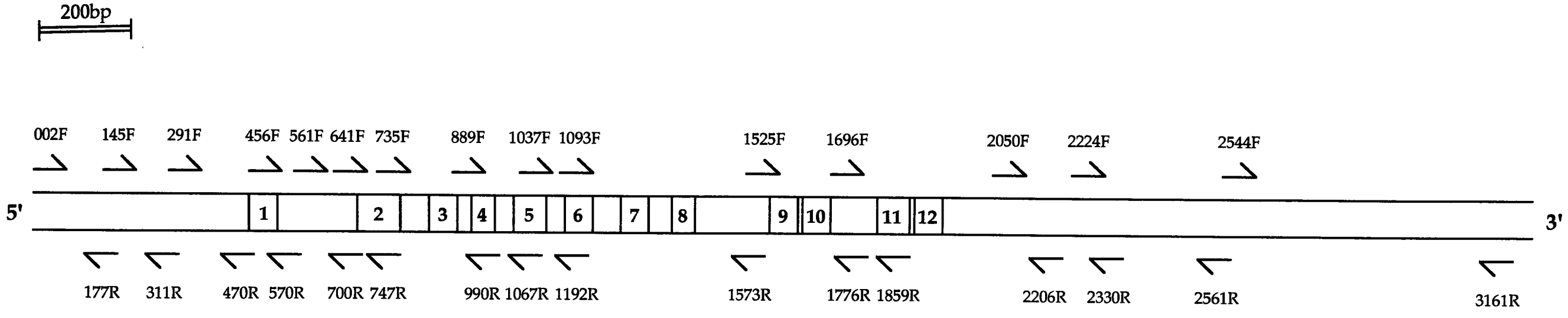


Figure 6.5: Primers used for PCR analysis of CLCN5 exon-intron structure, and their relative positions with respect to the transcript. Forward primers are shown above the transcript, reverse primers below. Regions encoding putative hydrophobic domains are indicated. Note that the primers themselves are not drawn to scale. Primer sequences are given in Figure 6.4. See Table 6.1 for a summary of PCR results obtained using these primer pairs.

b) Although product of the expected size is amplified from cDNA, a larger band results from PCR of YAC and human genomic template, suggesting that there is at least one intron in the region bounded by the primers. The observation of altered product from the YAC, as well as the human template, indicates that it originates from the CLCN5 genomic locus, and is not the result of non-specific amplification from some other region of the genome. To verify this, the putative intron-containing product is cloned, using TA-cloning and sequenced from both ends. Comparison between the sequence of the product, and that of the CLCN5 transcript (as derived from overlapping regions of the cDNA contig described in Section 5.3.4), is then used to identify the likely position of any introns in the former. Knowledge of the consensus sequence for splice sites is useful for the confirmation of intron position. The difference between product size from genomic template versus cDNA template gives an approximate estimate of the size of the intron.

c) Although product of the expected size is amplified from cDNA, no products are amplified from YAC or human genomic template. This may result from the presence of a large (>~4kb) intron in the intervening region between the primers, which is too large to amplify using conventional PCR. Alternatively, the lack of product from genomic template could be due to an intron (of any size) disrupting the binding site of one of the primers. In such cases, further analysis with overlapping sets of primers is required to clarify the situation.

The results of this PCR analysis are summarized in Table 6.1. Eight introns were identified on the basis of products of altered size (Fig. 6.7 and 6.8). Cloning of these products followed by sequencing confirmed that they were indeed amplified from the CLCN5 genomic locus, and enabled the precise exon-intron boundaries to be established (Fig. 6.9 and Table 6.2). Each of these products was amplified, cloned and sequenced in duplicate in order to verify the sequences at the splice sites.

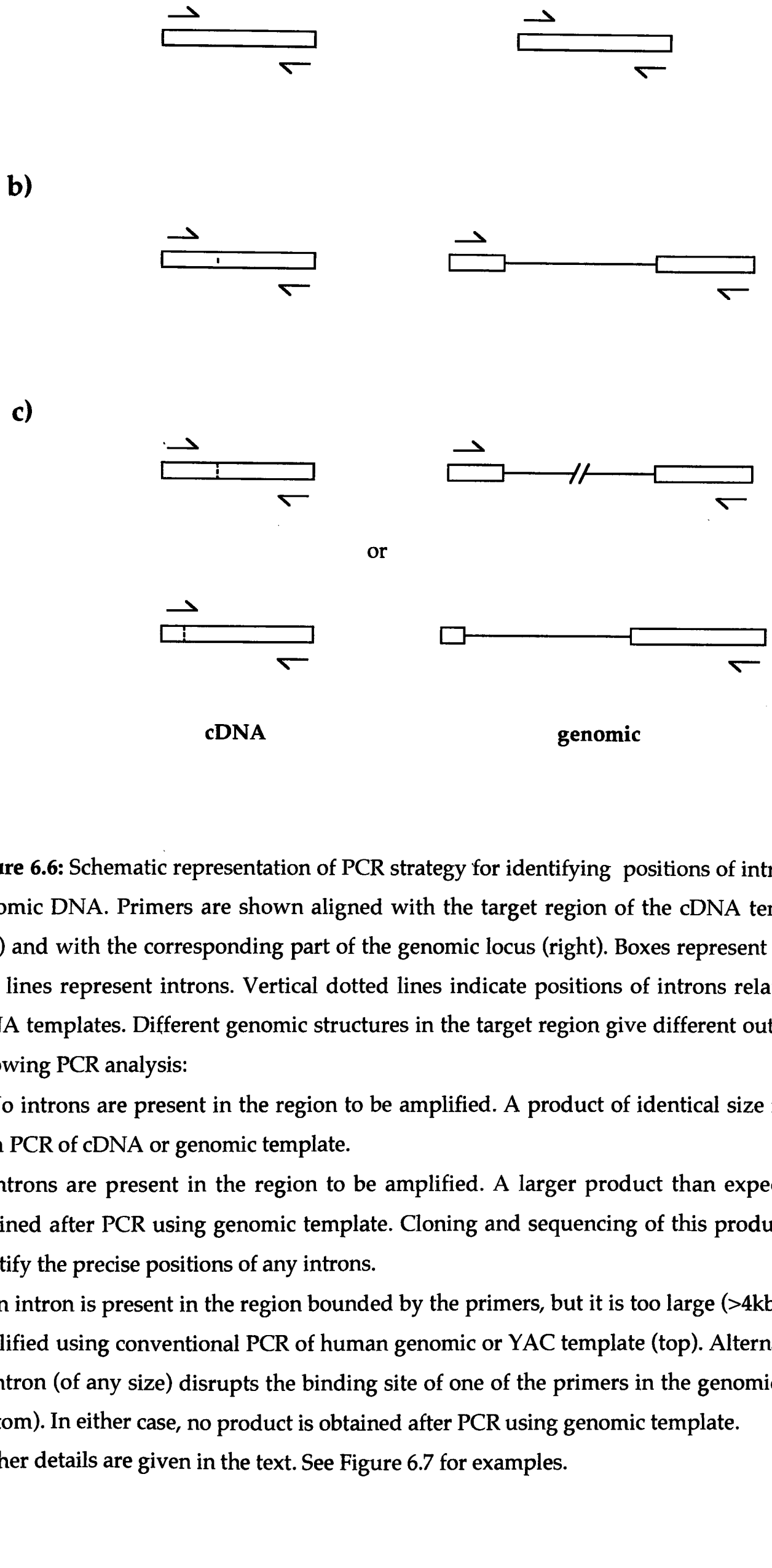


Figure 6.6: Schematic representation of PCR strategy for identifying positions of introns in genomic DNA. Primers are shown aligned with the target region of the cDNA template (left) and with the corresponding part of the genomic locus (right). Boxes represent exons, thin lines represent introns. Vertical dotted lines indicate positions of introns relative to cDNA templates. Different genomic structures in the target region give different outcomes following PCR analysis:

a) No introns are present in the region to be amplified. A product of identical size results from PCR of cDNA or genomic template.

b) Introns are present in the region to be amplified. A larger product than expected is obtained after PCR using genomic template. Cloning and sequencing of this product will identify the precise positions of any introns.

c) An intron is present in the region bounded by the primers, but it is too large (>4kb) to be amplified using conventional PCR of human genomic or YAC template (top). Alternatively an intron (of any size) disrupts the binding site of one of the primers in the genomic locus (bottom). In either case, no product is obtained after PCR using genomic template.

Further details are given in the text. See Figure 6.7 for examples.

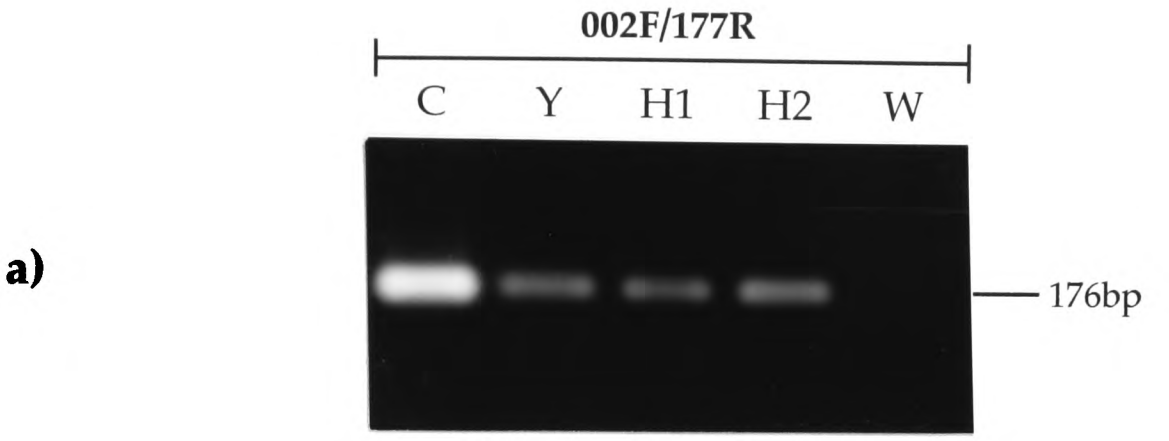


Figure 6.7: Examples of different outcomes of PCR strategy for investigation of genomic organization. Templates are cDNA (C), 6129 YAC (Y), two different human female genomic samples (H1 and H2) and a water control (W).

a) Amplification with 002F and 177R gives product of identical size in C, Y, H1 and H2 tracks, indicating that there are no introns in the 002-177 interval of the transcript.

b) 145F and 311R amplify a product of the expected size in C, but give a larger product in Y, H1 and H2, indicating the presence of a 135bp intron in the region.

c) 456F and 570R amplify a product of the expected size in C, but give no product in Y, H1 and H2, suggesting the presence of a large (>4kb) intron in the interval, or disruption of a primer binding site by an intron. Note that there is in fact a very faint band visible in the genomic tracks, due to contamination in the PCR (see Section 2.13.3), but this fragment is of much lower intensity than that in track C.

Forward primer	Reverse primer	cDNA template	cDNA product	genomic product	PCR pattern	Intron identified	Intron size
002F	177R	3N	176bp	176bp	(a)	–	–
145F	311R	3N	167bp	302bp	(b)	I	135bp
291F	470R	3N	180bp	2.65kb	(b)	II	2.45kb
456F	570R	3N	115bp	–	(c)	III*	–
561F	700R	3N	140bp	–	(c)	IV*	–
641F	747R	3N	107bp	–	(c)	IV*	–
735F	990R	7Y	256bp	1.1kb	(b)	V	850bp
889F	1067R	7Y	179bp	–	(c)	VI*	–
1037F	1192R	7Y	156bp	420bp	(b)	VII	260bp
1093F	1573R	7Y	481bp	481bp	(a)	–	–
1525F	1776R	3A-2	252bp	2.0kb	(b)	VIII	1.75kb
1696F	1859R	3A-2	164bp	1.4kb	(b)	IX	1.25kb
1696F	2206R	3A-2	511bp	1.75kb	(b)	IX	1.25kb
2050F	2330R	RL.3	281bp	430bp	(b)	X	150bp
2224F	2561R	RL.3	338bp	1.6kb	(b)	XI	1.25kb
2544F	3161R	RL.3	618bp	618bp	(a)	–	–

Table 6.1: Details of PCR reactions which were instrumental in establishing the exon-intron organization of CLCN5. See Figure 6.4 for primer sequences and positions with respect to transcript. It was necessary to use different cDNA templates for primer pairs from different regions of the transcript (see cDNA contig from Section 5.3.4). For each primer pair, the sizes of products that were amplified from cDNA and genomic template are given (a '–' denotes an absence of product). The pattern of each PCR result is also indicated, using an (a), (b) or (c) as defined in Figure 6.6. Intron sizes, where given, are estimated from the difference between the size of product obtained from genomic and cDNA template. Sizes of smaller introns (<800bp) are given to the nearest 10bp (except for intron I whose exact size was determined by sequencing). Sizes of larger introns (>800bp) are given to the nearest 50bp. Asterisks denote those introns which were too large to amplify across using conventional PCR on YAC or human genomic template and which were later identified by S. E. Lloyd using an alternative strategy.

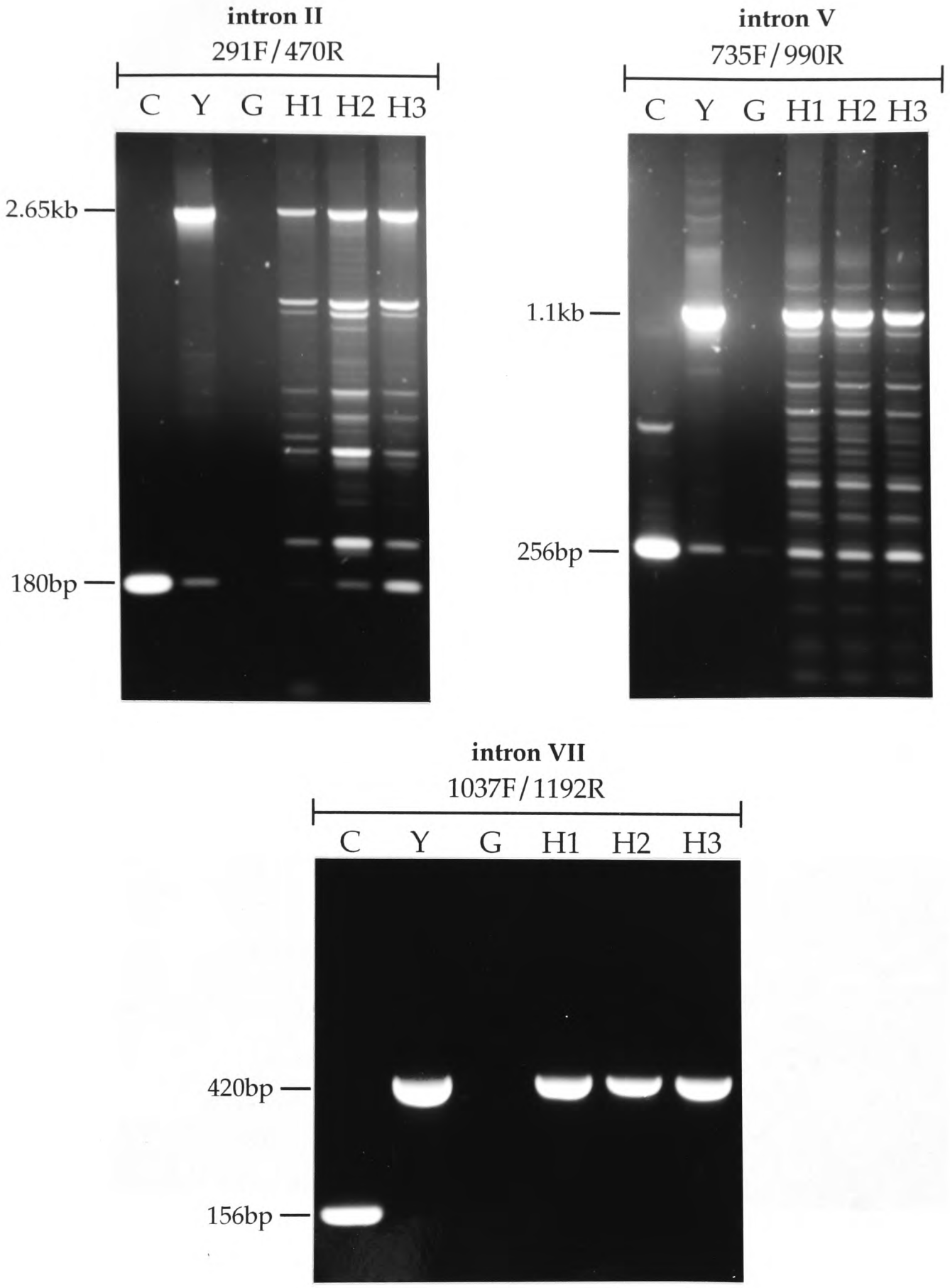


Figure 6.8: Isolation of fragments containing introns using a PCR strategy. The primers used and the intron identified are given above each photograph. Templates were cDNA (C), 6129 YAC (Y), the B102 GATA YAC, described in Chapter 4 (G), and three different human female genomic samples (H1, H2 and H3). The reaction containing the B102 YAC acted as a negative control. Sizes of products are indicated. These were determined by comparison to 1kb ladder marker fragments (not shown). The presence of the cDNA band in YAC and human tracks of 291F/470R and 735F/990R reactions was due to contamination in the PCR (see Section 2.13.3). In addition, when using these primers, a ladder of bands was amplified from human genomic template, but not from the YAC template (which is of lower complexity), in which a single predominant intron-containing product was observed. The appropriate genomic products were TA-cloned and sequenced to identify precise exon-intron boundaries (see text). A summary of these results is given in Table 6.1.

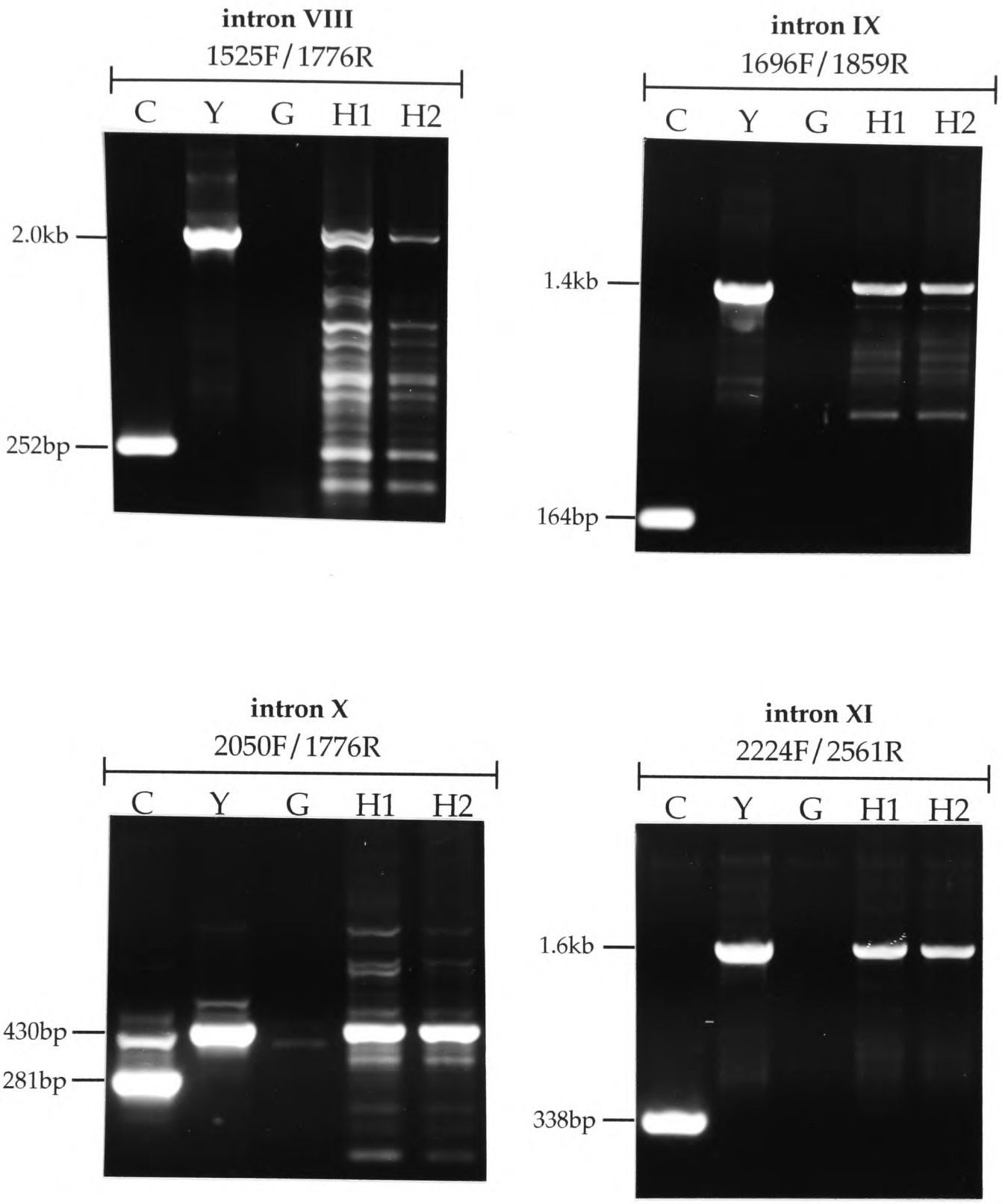


Figure 6.8 (contd): Isolation of fragments containing introns using a PCR strategy. The primers used and the intron identified are given above each photograph. Templates were cDNA (C), 6129 YAC (Y), the B102 GATA YAC, described in Chapter 4 (G), and two different human female genomic samples (H1 and H2). The reaction containing the B102 YAC acted as a negative control. Sizes of products are indicated. These were determined by comparison to 1kb ladder marker fragments (not shown). When using 1525F/1776R primers, a ladder of bands was amplified from human genomic template, but not from the YAC template (which is of lower complexity), in which a single predominant intron-containing product was observed. The appropriate genomic products were TA-cloned and sequenced to identify precise exon-intron boundaries (see text). A summary of these results is given in Table 6.1.

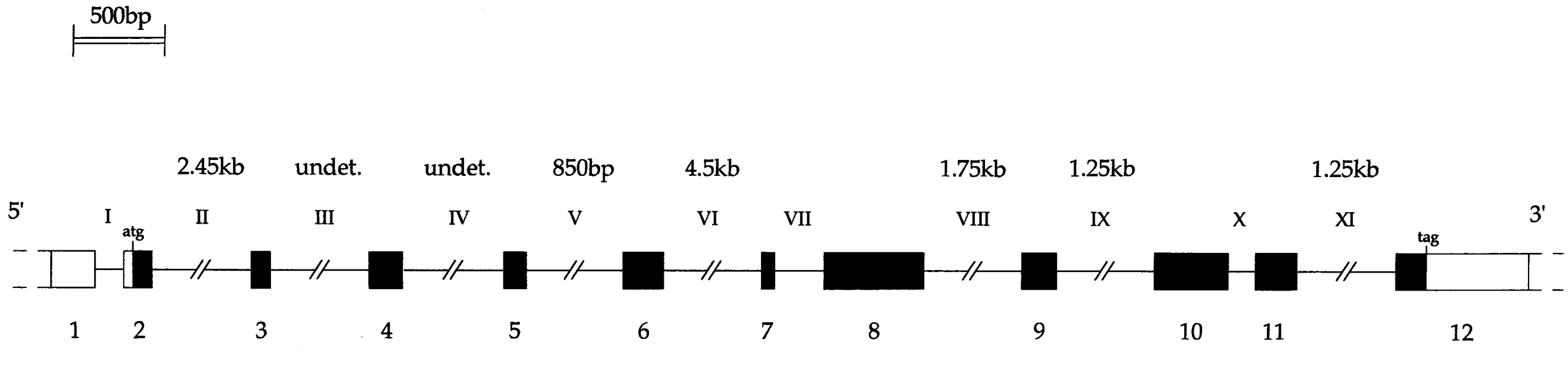


Figure 6.9: Exon-intron structure of the CLCN5 coding region. Intron numbers are indicated above, exon numbers below. The relative sizes of all exons and introns I, VII and X are drawn to scale. Sizes of other introns are indicated, except for III and IV which are undetermined. The initiator methionine is represented by 'atg' and the stop codon by 'tag'. Untranslated regions, which include the whole of exon 1, are indicated by open boxes.

Intron	Position	5'-3' sequence at exon-intron boundary
I	244-245	CCTACCAGgtgactgtatttctttatcaaccccaactgtagtc.....ctgaagtactaatgtctatttcttgttcaatacagAGGACAAG
II	396-397	ACCGAGAGgtaagacaaaagatggcacatgggtaagtgttagg.....gtagagtgtaccaatgttttctcattttcccctagATTACCAA
III	496-497	TTTATCAGgtatggtaaactgttagttttcaaaacaaatctcc.....taactttggcctttccctccctcccnncaaatcagGTTCGTTA
IV	684-685	CAGATGAGgtaacatgtagtgatgttttatgagccattgactt.....taggctaaaacaagcttctttttcccctgttccagGGAGCCTT
V	807-808	TCCCTGAGgtgagtctcttaaaatggtttataaatggttacaa.....agtctatattcaatctttctgtgtttaacctgcagATAAAAAC
VI	1014-1015	GCAGAGAGgtaataatgaatggccttaatagtctcttttttggt.....gtcactaattctgagttttggattttggattttagGTCTTGTC
VII	1095-1096	TTGAAGAGgtaacaacttttcattgtacagcatgtgcatgctt.....tgactgagtttgctttctcaccttctttcttctagGTCAGCTA
VIII	1638-1639	GCATGAAGgtgaggaattcttttgggactcagtggctgcatgc.....tttctcactaaccatctattggtttctctttgcagATCCCTTC
IX	1825-1826	CTGCTTAGgtgagtagtgtttgcattaatttcaagttgctacc.....gcctacctgagtagactgtgtctatttctttgcagGTGGGGTG
X	2224-2225	TTCAATTGgtaaggatttcagaaaggggatagtggaatccact.....catcttcaatttgtttttttccttctgtttgaatagAAAATGCT
XI	2441-2442	CACAACGGgtaagaagtcttgagtgaagtcaaattgaattgtg.....tatttttgtttttttgtattgtgtttgtctttagGCGATTGT

Table 6.2: Sequences of each exon-intron boundary. Position is given with respect to the nucleotide sequence given in Figure 5.18. Uppercase letters are exonic, lowercase intronic. Boundaries for introns III, IV and VI were identified by S. E. Lloyd (see text).

Four primer pairs (456F/570R, 561F/700R, 641F/747R and 889F/1067R) consistently failed to amplify from YAC or human genomic template (Table 6.1). The most likely explanation for this was the presence of large (>~4kb) introns in the regions corresponding to nucleotides 470-735 and 990-1037 of the cDNA sequence. This hypothesis was supported by the use of further primers chosen from these regions, which also gave no product after PCR of genomic template. A CLCN5 cosmid had been isolated by our collaborators at Hammersmith Hospital, and they used cycle sequencing of this cosmid, with primers from the 470-735 and 990-1037 parts of the transcript, to identify the remaining exon-intron boundaries (S. E. Lloyd, personal communication). Three introns were thus found, two (introns III and IV) in the 470-735 interval and one (intron VI) in the 990-1037 region. Note that the presence of intron IV (at 684-685) disrupted amplification from genomic template in both the 561F/700R and 641F/747R PCRs, because it maps in the overlap between the two primer pairs (Table 6.1). Introns III and IV could not be amplified using conventional PCR, even when primers were designed from the exonic sequences just adjacent to the intron and used on cosmid template (which is of a lower complexity than YAC and human genomic template). The sizes of these introns remain undetermined. However, amplification across the third intron (intron VI) was succesful when cosmid template was used, giving a product of ~4.5kb (S. E. Lloyd, personal communication). Hybridization analysis confirmed that this product originated from the CLCN5 locus (see below).

6.3.2 Correlation between exon-intron structure and genomic restriction fragments

By analysing the hybridization patterns of the different cDNA clones from the CLCN5 contig it was possible to derive approximate positions of the four introns which were expected to contain *Eco*R1 sites (Fig. 6.10 and Table 6.3). Given that the positions of all exon-intron boundaries had already been established, it could be deduced that introns III, VI, VIII and either IX or X were likely candidates for these *Eco*R1-containing introns (Table 6.3).

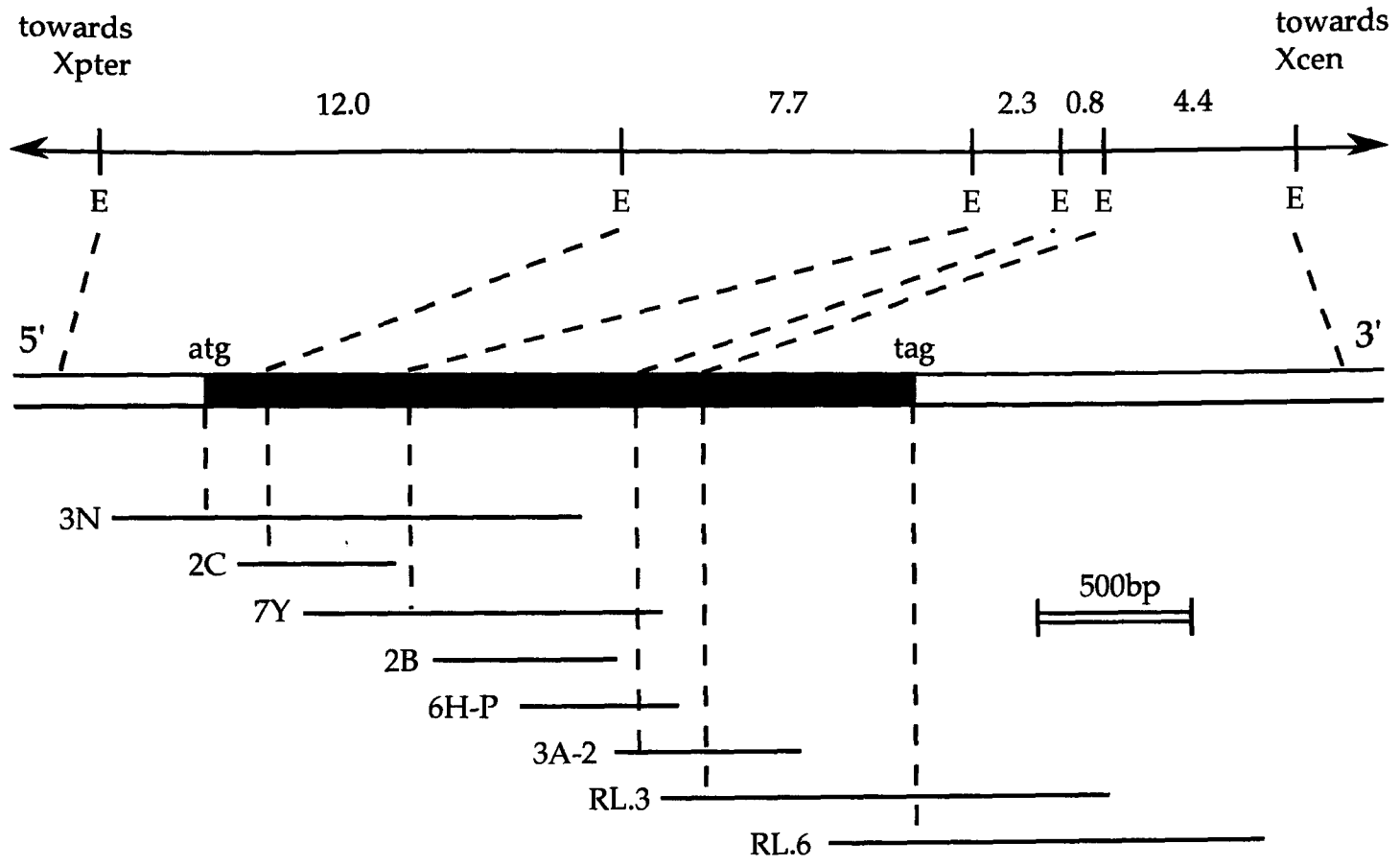


Fig. 6.10: cDNA contig spanning CLCN5 coding region, previously described in Chapter 5. Alignment with *Eco*R1 fragments, as deduced from hybridization analysis, is shown schematically. Dashed lines indicate how each cDNA relates to the genomic fragments that it detects. Only exon-containing genomic fragments are shown, and so each E may in fact represent more than one *Eco*R1 site. Fragment sizes are given in kilobases. See legend for Figure 5.17 for more details.

Exonic *Eco*R1 fragments separated by intron	**Region of cDNA in which intron containing *Eco*R1 site is likely to lie**		**Candidates for introns containing *Eco*R1 sites**
	Start point	**End point**	
12.0kb-7.7kb	399 (start of 2C)	571 (start of 7Y)	III
7.7kb-2.3kb	880 (end of 2C)	1020 (start of 2B)	VI
2.3kb-0.8kb	1588 (start of 3A-2)	1750 (start of 3A)	VIII
0.8kb-4.4kb	1783 (end of 6H-P)	2297 (start of 6F)	IX or X

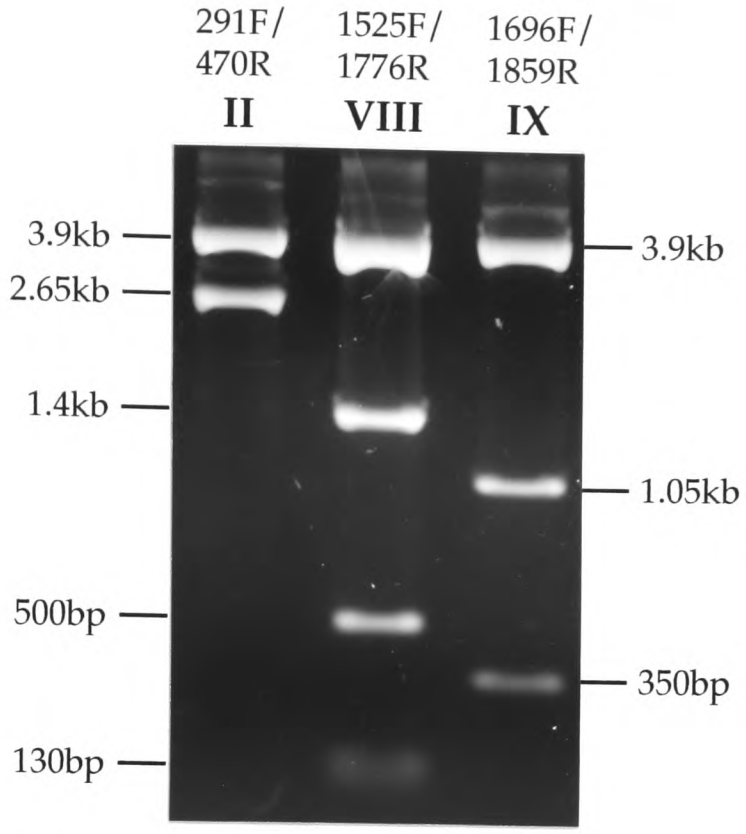
Table 6.3: Predicted positions of introns which contain one or more *Eco*R1 sites, and suggested candidates, based on hybridization patterns of different cDNAs from CLCN5 contig. Start and end points are given with respect to nucleotides from Figure 5.18.

Digestion of the eight cloned introns with *Eco*R1 confirmed that introns VIII and IX did indeed contain sites for this enzyme (Fig. 6.11a). Intron VI, which had been amplified by PCR, but not cloned (see above), was also shown to have internal *Eco*R1 sites by digestion of the PCR product (Fig. 6.11b).

As described above, it had not been possible to amplify across intron III using conventional PCR, and so an alternative strategy was used to confirm that it spanned the 12.0kb-7.7kb *Eco*R1 fragment boundary. A PCR product was amplified from cDNA template using the 291F and 570R primers; this 280bp product contained material from exons 2, 3 and 4 of the transcript. When this cDNA probe was hybridized to an *Eco*R1 digest of YAC DNA, it detected both the 12.0kb and 7.7kb fragments (Fig. 6.12). This result must be due to an *Eco*R1 site in intron III, since it had already been shown that intron II is not cleaved by this enzyme. A 187bp cDNA probe spanning the junction between exons 4 and 5 was made by PCR using the 561F and 747R primers. When this was hybridized to the *Eco*R1 digested YAC it only detected the 7.7kb fragment, confirming that intron IV contains no *Eco*R1 sites.

*Eco*R1 digestion of the 1525F/1776R cloned PCR product which contained intron VIII gave three bands of 1.4kb, 500bp and 130bp (Fig. 6.11a), suggesting that this intron has an internal *Eco*R1 fragment which is not detected by the CLCN5 cDNA. Sequence analysis of the exon-intron junctions indicated that one of the intron VIII *Eco*R1 sites is found at the +6 to +11 positions of the 5' splice site, thus accounting for the 130bp digestion product from '1525F/1776R'. (Note that the *Eco*R1 sites of the pCR II vector flank the cloning site at a distance of 9bp on one side and 8bp on the other.) Given that the 3' end of intron VIII should map within an 800bp genomic *Eco*R1 fragment (Fig. 6.10 and Table 6.3), it follows that the 1.4kb band generated by digestion of '1525F/1776R' is too large to be this end fragment and must therefore correspond to the internal intronic fragment. When the 1.4kb band was excised from a gel, purified and used to probe *Eco*R1 digests of YAC DNA, it detected a 1.4kb fragment, thus confirming this hypothesis (Fig. 6.13).

a)



b)

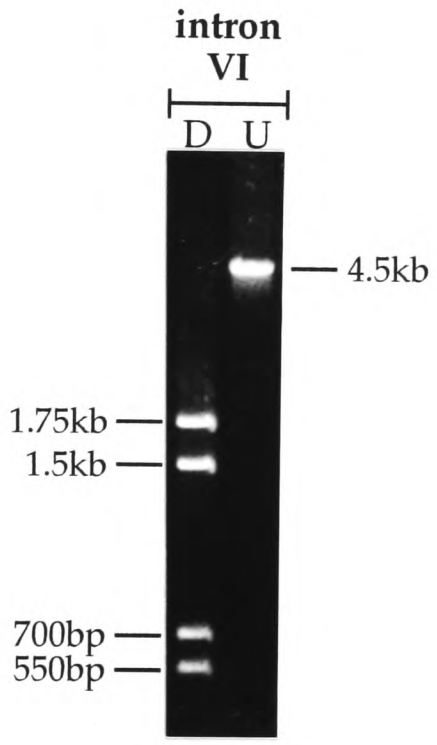


Figure 6.11: Digestion of CLCN5 introns with *Eco*R1. Fragment sizes were derived from comparison to 1kb ladder markers (not shown).

a) *Eco*R1 digests of cloned intron-containing fragments, obtained by PCR of genomic template with CLCN5 primer pairs, followed by TA-cloning. The 3.9kb band in all tracks corresponds to the linearised TA-cloning vector. These results demonstrate that while the 291F/470R product (containing intron II) has no *Eco*R1 sites, 1696F/1859R (containing intron IX) has one, and 1525F/1776R (containing intron VIII) has two. See text for further discussion of these observations.

b) *Eco*R1 digestion of the (uncloned) intron VI PCR product, obtained by PCR of cosmid template with primers adjacent to the exon-intron boundaries. Four fragments are generated, indicating the presence of two *Eco*R1 sites within this intron. D, digested PCR product; U, undigested.

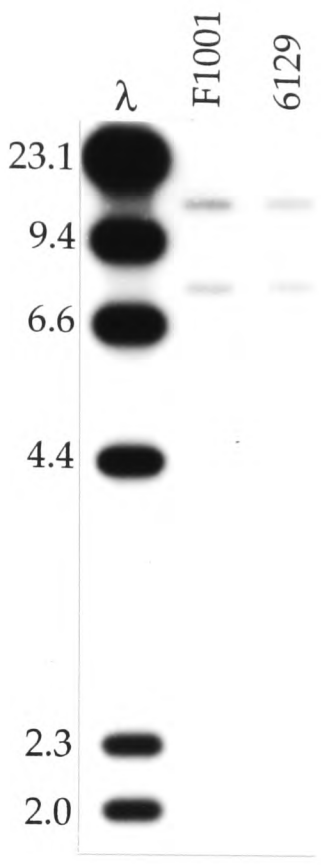


Figure 6.12: Hybridization of a 280bp PCR product spanning exons 2-4 of the CLCN5 cDNA to *Eco*R1 digests of DXS255 YACs. Sizes of lambda markers (λ) are given in kilobases. This cDNA probe detects both the 12.0kb and 7.7kb CLCN5 exonic fragments, thus supporting the presence of an *Eco*R1 site in intron III (see text).

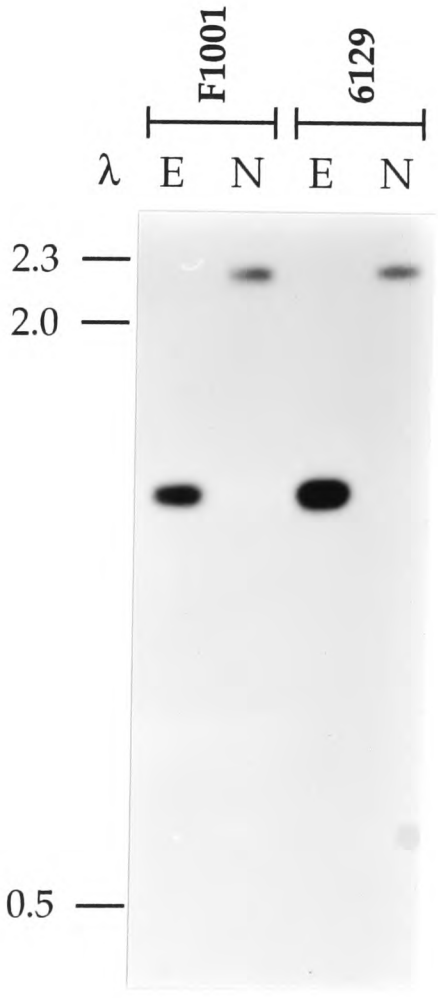


Figure 6.13: Hybridization of the 1.4kb *Eco*R1 subfragment from intron VIII ('1525F/1776R') to *Eco*R1 (E) and *Nde*I (N) digests of the DXS255 YACs. Positions and sizes, in kilobases, of lambda markers (λ) are indicated. A single *Eco*R1 band of 1.4kb is detected, confirming that this probe corresponds to an internal intronic fragment (see text).

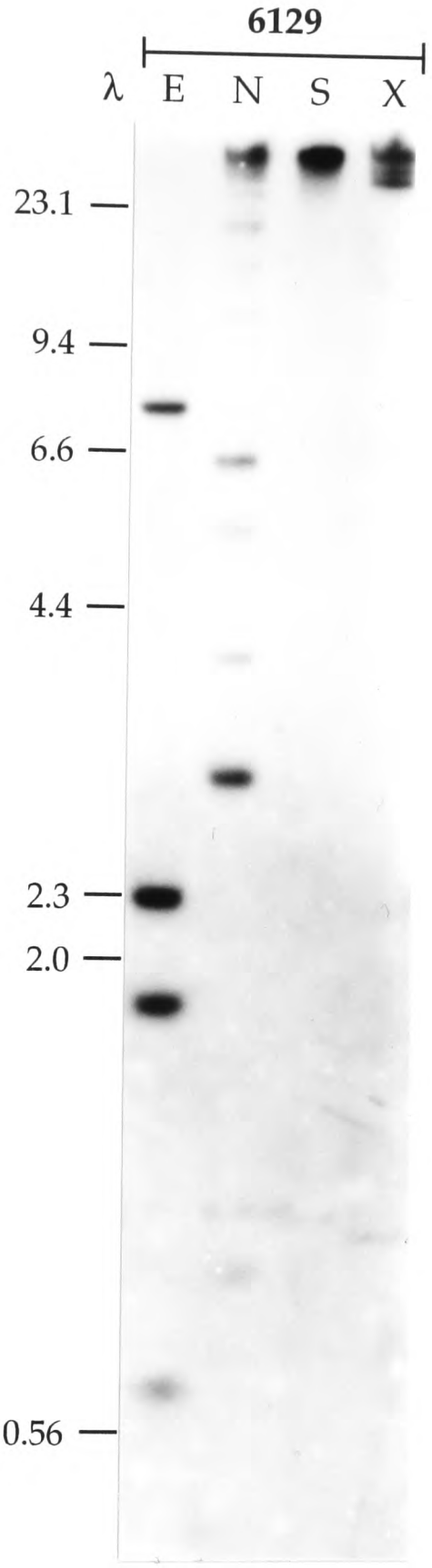


Figure 6.14: Hybridization of a 4.5kb PCR product consisting of intron VI to digests of the 6129 YAC clone. Enzymes are *Eco*R1 (E), *Nde*I (N), *Sal*I (S) and *Xho*I (X). Positions and sizes in kilobases of lambda markers (λ) are indicated. Four *Eco*R1 bands are detected; two exon-intron junction-containing fragments of 7.7 and 2.3kb, and two fragments of 1.75 and 0.7kb which are completely intronic. This result therefore confirms that intron VI contains two internal *Eco*R1 fragments (see text). The ladder of fragments in the *Nde*I track is due to incomplete digestion; analysis of the products obtained with this enzyme are not included in the study of genomic organization presented here.

When the 4.5kb intron VI PCR product (see above) was digested with *Eco*R1, four bands were generated whose sizes were 1.75, 1.5, 0.7 and 0.55kb (Fig. 6.11b), indicating that this intron contains two internal *Eco*R1 fragments. In support of this, hybridization of intron VI to *Eco*R1 digests of YAC DNA detected two intronic fragments of 1.75 and 0.7kb, as well as the two expected exon-containing fragments of 7.7 and 2.3kb (Fig. 6.14). It was not possible to determine the order of the intronic fragments on the basis of this data.

The correlation between the exon-intron organization of CLCN5 and the *Eco*R1 map of the locus is summarized in Figure 6.15.

6.3.3 The extent of the coding region at the genomic level

The 145F/311R primer pair was used to amplify the region containing intron I from genomic template. The 302bp PCR product which was obtained was used to probe partial digests of the 6129 YAC clone (see Section 3.3.1). The pattern of restriction fragments detected (Fig. 6.16; Table 6.4) indicated that this 'intron I' probe mapped within a 20kb *Bss*HII fragment in the left-hand part of the YAC, at least 25kb from the YAC telomere (Fig. 6.17). This localization of the 5' end of the cDNA contig was supported by the identification of an *Xho*I site at nucleotides 371-376 of the mRNA sequence (see Fig. 6.4). Previous analysis of 6129 has indicated that the *Xho*I site which is closest to the left end is ~30kb from the YAC telomere (data not shown). Given that the RL.6 cDNA clone overlaps with L(6129), the 170bp left end clone of this YAC (Section 5.3.3), and bearing in mind that the left vector arm is ~5kb in size it can be deduced that the region encompassed by the CLCN5 cDNA contig (containing the complete ORF), spans at least ~25kb of human genomic sequence in the YAC.

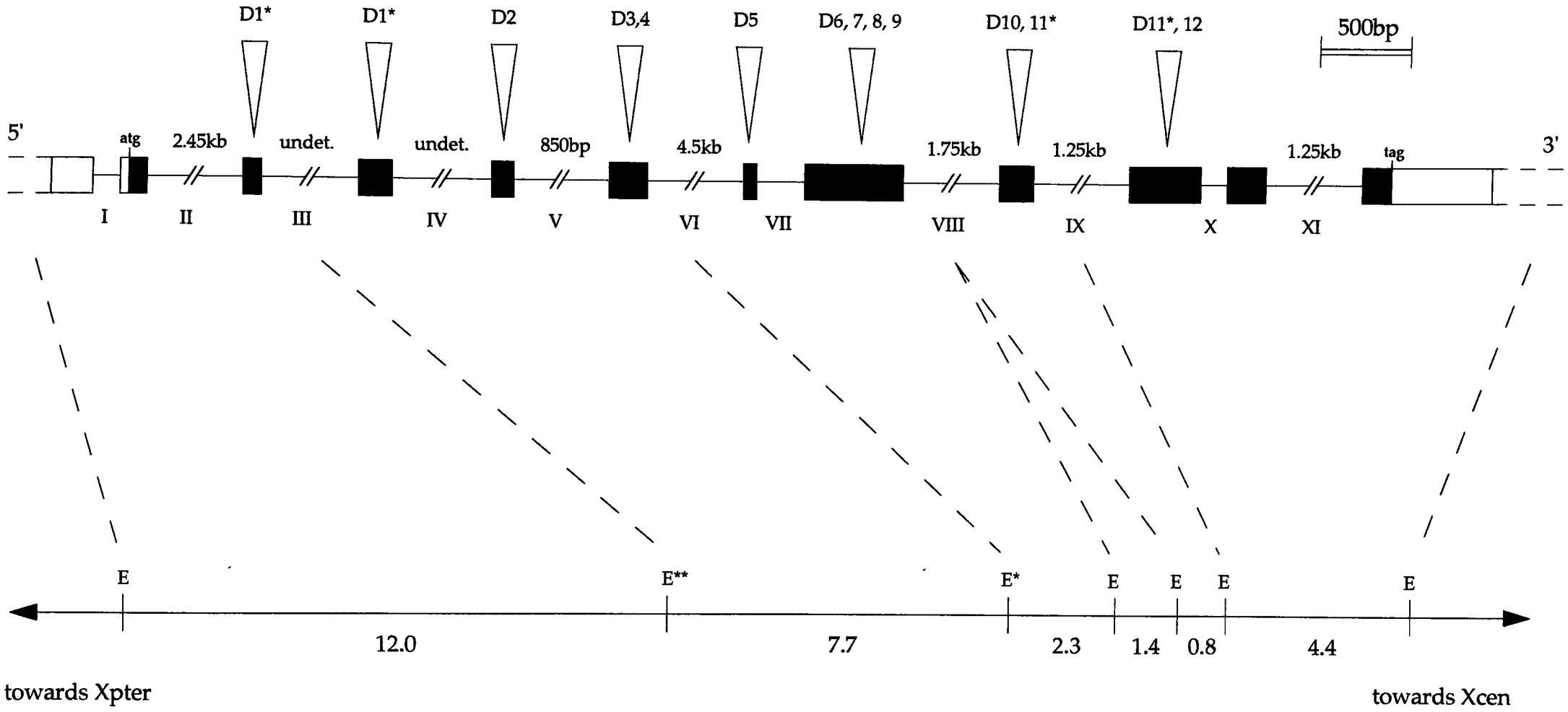


Figure 6.15: Exon-intron organization of CLCN5, and relationship to genomic *Eco*R1 fragments, as deduced from PCR studies and hybridization analyses.
TOP: Exon-intron structure, as given in Fig. 6.9. In addition, the predicted hydrophobic domains contained in each exon are indicated by arrowheads. Asterisks denote domains which are disrupted by introns. The scale which is given applies only to the top part of the figure.
BOTTOM: Correlation with genomic *Eco*R1 fragments. Each E represents a single *Eco*R1 site. Sizes of fragments are given in kilobases. Intron VIII contains a 1.4kb internal *Eco*R1 fragment. Intron VI contains two internal *Eco*R1 fragments of 1.75 and 0.7kb, therefore E* represents more than one *Eco*R1 site. E** may also represent more than one *Eco*R1 site. Orientation with respect to the X chromosome is indicated.

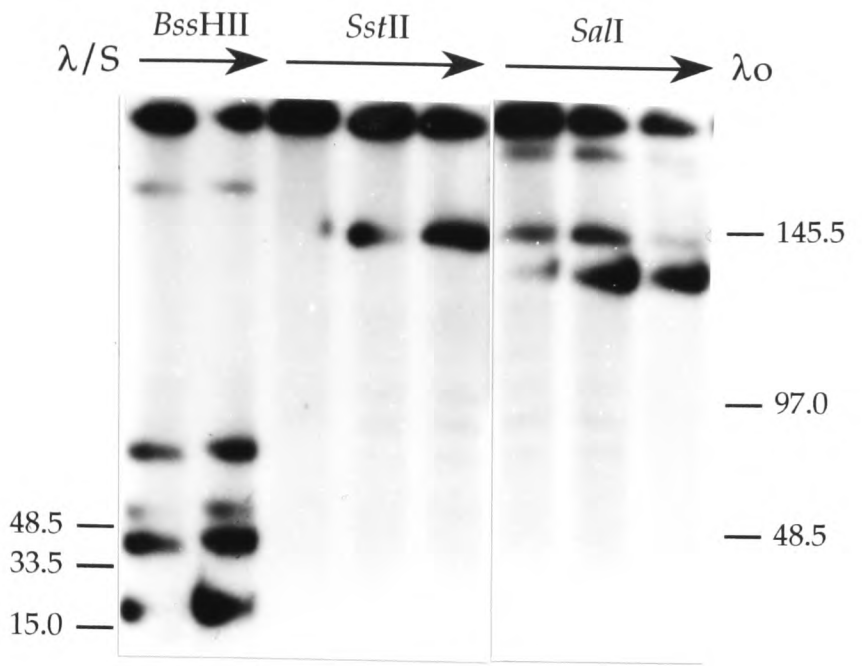


Figure 6.16: Partial digests of the 6129 YAC using different rare-cutters, probed with 'intron I' (see text). Direction of arrow represents increasing enzyme concentration (from 0.1-5U) with a 1 hour digestion time. Sizes of lambda oligomer (λo) and lambda/*Sal*I (λ/S) markers are given in kilobases. This is the same filter as that shown in Fig. 5.7.

Enzyme	Intron I
*Bss*HII	20, 45, 60, 85, 160
*Eag*I	145, 160
*Sal*I	135, 145, 175
*Sfi*I	105, 145
*Sst*II	145

Table 6.4: Fragment sizes, in kilobases, of bands detected on rare-cutter partial digests of 6129 YAC clone, when hybridized with the 'intron I' probe.

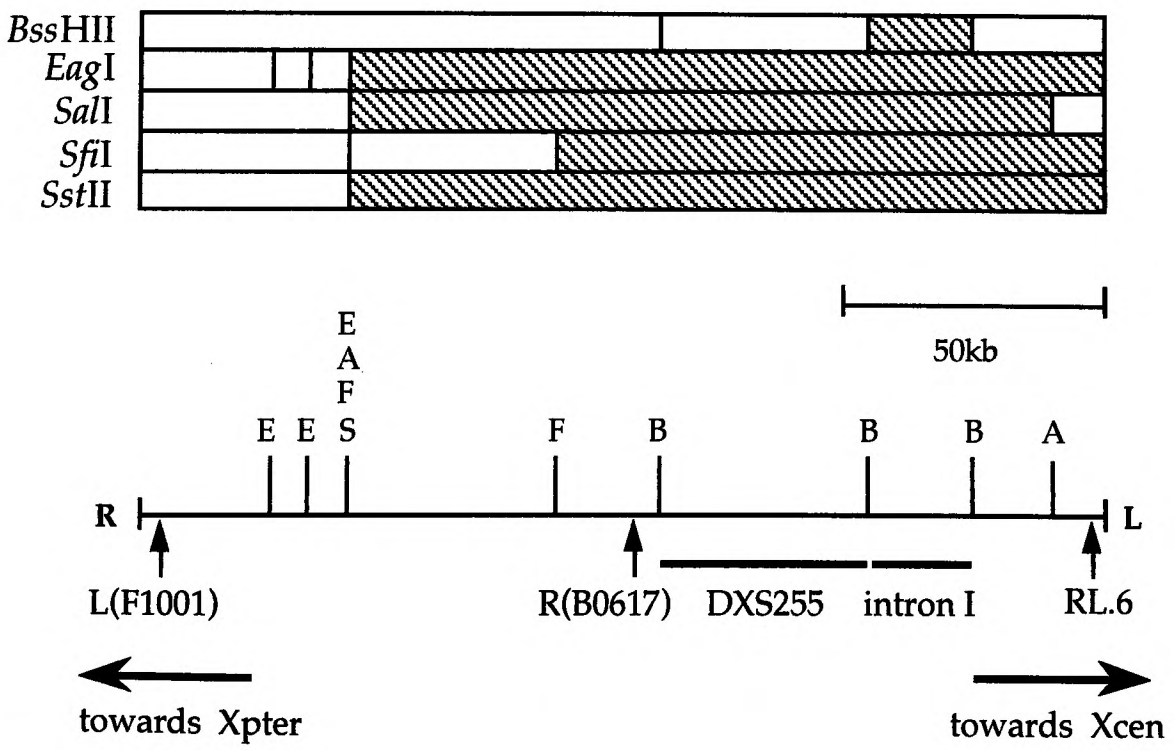


Figure 6.17: Localization of 'intron I' probe within rare-cutter map of 6129 YAC clone, as deduced from fragment sizes in Table 6.4. Diagonal stripes show fragments on which 'intron I' lies. Cleavage sites are given as in Fig. 3.9. Positions of R(B0617), L(F1001), DXS255 and RL.6 and orientation of YAC with respect to Xpter-Xcen are indicated.

6.4 Discussion

The above studies have shown that the coding region of CLCN5, encompassed by clones 3N, 7Y and RL.3 of the cDNA contig described in Chapter 5, is interrupted by eleven introns at the genomic level. It was possible to determine the sizes of nine of these introns using conventional PCR; an alternative strategy will be needed for size estimation of the remaining two, which were too large to be amplified. Hybridization analysis indicated that this region of the CLCN5 gene spans >25kb of genomic DNA. As shown in Fig. 6.18, the sequences of all 5' (donor) and 3' (acceptor) splice sites identified in this locus agree well with the splice site consensus sequences which were previously derived from analysis of nearly 1800 human introns (Stephens and Schneider, 1992).

There is no obvious correlation between the exon-intron structure of the CLCN5 open reading frame and the positions of the predicted hydrophobic domains which it encodes (Fig 6.15). Only two out of the eleven introns were found to disrupt putative domains, but it is unclear whether this observation holds any significance for the 'introns early'/'introns late' evolutionary debate. Even if the positioning of introns relative to the domains *is* non-random, this does not necessarily support the 'introns early' view, since apparent regularities in intron distribution may result from non-random insertion of introns (Rogers, 1989). For example, in serine protease genes, introns appear to have preferentially inserted in regions encoding the variable surface loops of the proteins. Similarly, the inserted introns of the TFIIIA gene tend to map to the loops between domains.

Several different factors could contribute to a non-random distribution of inserted introns (Rogers, 1989). The insertion mechanism may involve a certain degree of sequence specificity, or there might be selection after insertion which would favour the presence of introns in certain positions of the gene. Studies have shown that while *most* of the sequence conservation of splice sites is found in intronic regions, as discussed in section 6.1.3, there are still significant limitations on the coding sequences adjacent to exon-intron boundaries, particularly at the donor site (see Figs. 6.3 and 6.18).

		-3	-2	-1	+1	+2	+3	+4	+5	+6	+7
%	T	12	14	8	<1	**99**	2	8	6	**47**	18
	G	17	12	**78**	**99**	<1	**41**	12	**82**	20	**34**
	C	**37**	15	5	<1	<1	2	9	5	17	22
	A	**34**	**59**	9	<1	<1	**55**	**71**	7	16	26

I	**C**	**A**	**G**	**g**	**t**	**g**	**a**	c	**t**	**g**
II	G	**A**	**G**	**g**	**t**	**a**	**a**	**g**	a	c
III	**C**	**A**	**G**	**g**	**t**	**a**	t	**g**	g	t
IV	G	**A**	**G**	**g**	**t**	**a**	**a**	c	a	t
V	G	**A**	**G**	**g**	**t**	**g**	**a**	**g**	**t**	c
VI	G	**A**	**G**	**g**	**t**	**a**	**a**	t	a	a
VII	G	**A**	**G**	**g**	**t**	**a**	**a**	c	a	a
VIII	**A**	**A**	**G**	**g**	**t**	**g**	**a**	**g**	g	a
IX	T	**A**	**G**	**g**	**t**	**g**	**a**	**g**	**t**	a
X	T	T	**G**	**g**	**t**	**a**	**a**	**g**	g	a
XI	**C**	G	**G**	**g**	**t**	**a**	**a**	**g**	a	a

Figure 6.18a: Conservation of 5' (donor) splice sites identified in the CLCN5 locus. A matrix is shown at the top, giving the frequencies of the different nucleotides at each position, as derived from analysis of >1700 acceptor sites (Stephens and Schneider, 1992). The frequencies of the nucleotides that make up the consensus sequence are shown in boldface. Intron sequences are aligned below the matrix, with nucleotides that agree with the consensus in boldface. Letters in uppercase are exonic, lowercase intronic; the exon-intron junction is indicated with a dotted line. Information analysis has suggested that the recognition region for the donor site runs from -3 to +7 (Stephens and Schneider, 1992). A similar region (-3 to +6) has been implicated in U1 snRNA binding (see text and Figure 6.3). Note that elements outside of this region may also influence splice site selection (see section 6.1.3).

		-26	-25	-24	-23	-22	-21	-20	-19	-18	-17	-16	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	+1	+2
	T	**30**	**30**	**32**	**30**	**32**	**35**	**36**	**33**	**39**	**38**	**39**	**43**	**45**	**45**	**49**	**51**	**49**	**41**	**41**	**43**	**46**	**53**	23	23	<1	<1	11	**34**
%	**G**	16	16	17	16	15	15	16	18	13	14	15	13	13	12	10	10	11	12	10	9	7	6	25	1	<1	**99**	**50**	24
	C	**30**	**32**	**30**	**32**	**30**	**31**	**29**	**33**	**33**	**34**	**33**	**33**	**33**	**35**	**33**	**32**	**32**	**35**	**39**	**38**	**40**	**34**	29	**72**	<1	<1	13	18
	A	24	22	21	22	23	19	19	16	15	14	13	11	9	8	8	7	8	9	10	10	7	7	23	4	**99**	<1	26	24

I	**t**	a	a	**t**	g	**t**	**c**	**t**	a	**t**	**t**	**t**	**c**	**t**	**t**	g	**t**	**t**	**c**	a	a	**t**	a	**c**	**a**	**g**	A	G
II	a	**c**	**c**	a	a	**t**	g	**t**	**t**	**t**	**t**	**c**	**t**	**c**	a	**t**	**t**	**t**	**t**	**c**	**c**	**c**	c	t	**a**	**g**	A	**T**
III	**c**	**c**	**t**	**t**	**t**	**c**	**c**	**c**	**t**	**c**	**c**	**c**	**t**	**c**	**c**	**c**	n	n	**c**	a	a	a	t	**c**	**a**	**g**	**G**	**T**
IV	a	**c**	a	a	g	**c**	**t**	**t**	**c**	**t**	**t**	**t**	**t**	**t**	**c**	**c**	**c**	**c**	**t**	g	**t**	**t**	c	**c**	**a**	**g**	**G**	G
V	**t**	**c**	a	a	**t**	**c**	**t**	**t**	**t**	**c**	**t**	g	**t**	g	**t**	**t**	**t**	a	a	**c**	**c**	**t**	g	**c**	**a**	**g**	A	**T**
VI	**t**	**c**	**t**	g	a	g	**t**	**t**	**t**	**t**	g	g	a	**t**	**t**	**t**	**t**	g	g	a	**t**	**t**	t	t	**a**	**g**	**G**	**T**
VII	**t**	**t**	g	**c**	**t**	**t**	**t**	**c**	**t**	**c**	a	**c**	**c**	**t**	**t**	**c**	**t**	**t**	**t**	**c**	**t**	**t**	c	t	**a**	**g**	**G**	**T**
VIII	a	a	**c**	**c**	a	**t**	**c**	**t**	a	**t**	**t**	g	g	**t**	**t**	**t**	**c**	**t**	**c**	**t**	**t**	**t**	g	**c**	**a**	**g**	A	**T**
IX	a	g	**t**	a	g	a	**c**	**t**	g	**t**	g	**t**	**c**	**t**	a	**t**	**t**	**t**	**c**	**t**	**t**	**t**	g	**c**	**a**	**g**	**G**	**T**
X	**t**	**t**	**t**	g	**t**	**t**	**t**	**t**	**t**	**t**	**c**	**c**	**t**	**t**	**c**	**t**	g	**t**	**t**	**t**	g	a	a	t	**a**	**g**	A	A
XI	**t**	**t**	**t**	**t**	**t**	**t**	g	**t**	a	**t**	**t**	g	**t**	g	**t**	**t**	**t**	g	**t**	**c**	**t**	**t**	t	t	**a**	**g**	**G**	C

Figure 6.18b: Conservation of 3' (acceptor) splice sites identified in the CLCN5 locus. A matrix is shown at the top, giving the frequencies of the different nucleotides at each position, as derived from analysis of >1700 acceptor sites (Stephens and Schneider, 1992). The frequencies of the nucleotides that make up the consensus sequence are shown in boldface. Intron sequences are aligned below the matrix, with nucleotides that agree with the consensus in boldface. Letters in uppercase are exonic, lowercase intronic; the intron-exon junction is indicated with a dotted line. Information analyses and footprint data have suggested that the recognition region for the acceptor site extends from -26 to +2, but that the -4 position has no significance for splice site selection (Stephens and Schneider, 1992). It is possible that the pyrimidine tract may stretch further 5', beyond the -26 position, but the size of the data set for information analysis was not large enough to confirm this.

The sequence context surrounding a newly inserted intron will therefore be expected to influence the efficiency with which it can be spliced out, which will in turn determine whether or not the presence of the intron will be favoured by selection.

It is interesting to note that the exon-intron boundaries for five of the eleven CLCN5 introns occur at a junction between a GAG codon (which would encode a glutamic acid residue) and its 3' neighbour (Fig. 6.19). Given that glutamic acid is a hydrophilic amino acid, it is not suprising that these introns map to the loops between the hydrophobic domains.

Figure 6.19 shows a comparison between the exon-intron structures of CLCN5 and CLCN1, the only other member of the ClC family for which the genomic structure has so far been determined (Lorenz *et al.*, 1994). The positions of introns are given relative to an alignment of the CLCN5/CLCN1 open reading frames obtained using the **PILEUP** computer program (see Chapter 5). Intron I is not included in this comparison, since it lies in the 5' untranslated region of CLCN5. The genomic organizations of these two loci are significantly different; the coding region of CLCN1 is disrupted by twenty two introns and none of these coincide in position with any of the CLCN5 introns.

It is very unlikely that a putative common ancestral chloride channel gene would have contained all 32 of the introns shown in Figure 6.19, and then given rise to the observed intron distributions by differential intron loss. Explanations involving intron slippage are also unsatisfactory, since they would require a change in position of at least 15 bases, in many cases with movement across a non-integral number of codons (see section 6.1.4). It is therefore probable that intron insertion has played a part in the evolution of the genomic structures of these loci, assuming that they both arose from the same ancestral chloride channel gene, thereby supporting the 'introns late' scenario. However, any conclusions from these results should be tentative, since CLCN5 and CLCN1 are from distant branches of the ClC family, and nothing is known, at present, about the genomic structures of the other chloride channel loci. As information accumulates regarding the latter, comparison of intron positions may provide an additional means for investigating the evolution of this gene family.

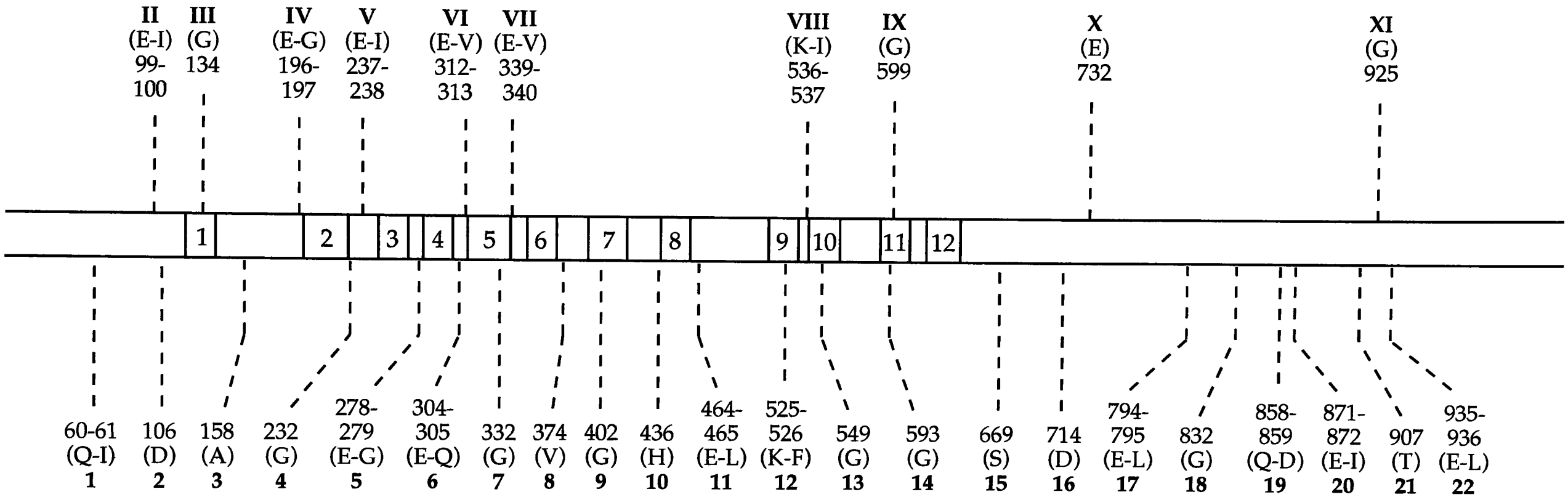


Figure 6.19: Comparison between the genomic structures of CLCN5 (Top) and CLCN1 (Bottom). Positions of introns are given with respect to the amino acids of an alignment of the CLCN5/CLCN1 sequences, obtained using the **PILEUP** computer program. (Numbering therefore relates to this alignment, and not to the amino acid numbers given in Fig. 5.18.) In addition, the symbols for the amino acids disrupted by each intron are shown. Exon-intron boundaries have only been defined for the open reading frame region of CLCN1, so this is the only region included for comparison. The positions of the twelve predicted hydrophobic domains relative to the aligned amino acid sequence are indicated (middle). See text for discussion.

The elucidation of the exon-intron boundaries of CLCN5 has facilitated the identification of mutations associated with Dent's disease in patients from several pedigrees (S. E. Lloyd, personal communication). Two of these mutations (in patients from families E and G) were found to involve single base substitutions in the most highly conserved positions of the intron V donor splice site (Fig. 6.20). Point mutations which affect pre-mRNA splicing are estimated to be responsible for over 15% of human genetic disease (Horowitz and Krainer, 1994). Donor splice site mutations can cause accumulation of unspliced pre-mRNA, retention of incompletely spliced pre-mRNA or a complete absence of transcript (see Parkinson and Thakker, 1992). Alternatively, they may lead to the use of cryptic splice sites, resulting in aberrantly processed mRNA. Another possibility is the use of a different normally occurring donor splice site, with concomitant loss of one or more exons, which is known as 'exon skipping'. RT-PCR analysis has indicated that CLCN5 transcripts of patients from families E and G consist of exon 4 directly spliced to exon 6, demonstrating that the donor splice site mutations are, in these cases, leading to exon skipping. The loss of exon 5 from the transcript is predicted to result in the production of a chloride channel lacking the putative D2 hydrophobic domain and physiological studies have shown that this domain, which is highly conserved among ClC family members, is essential for ClC-5 function (T. J. Jentsch, personal communication).

Nephrolithiasis affects 12% of males and 5% of females by the age of 70 years (Smith, 1989), occurring as a familial disorder in up to 45% of cases (Favus, 1989; Smith, 1989), and is commonly associated with hypercalciuria (Coe *et al.*, 1992; Favus, 1989; Lemann and Gray, 1989). The identification of the exon-intron boundaries of CLCN5 will facilitate the screening of further patients with renal tubular dysfunction for mutations at this locus. Such studies will be neccessary to establish whether or not the disruption of this X-linked gene is a contributing factor to the male predisposition towards kidney stones.

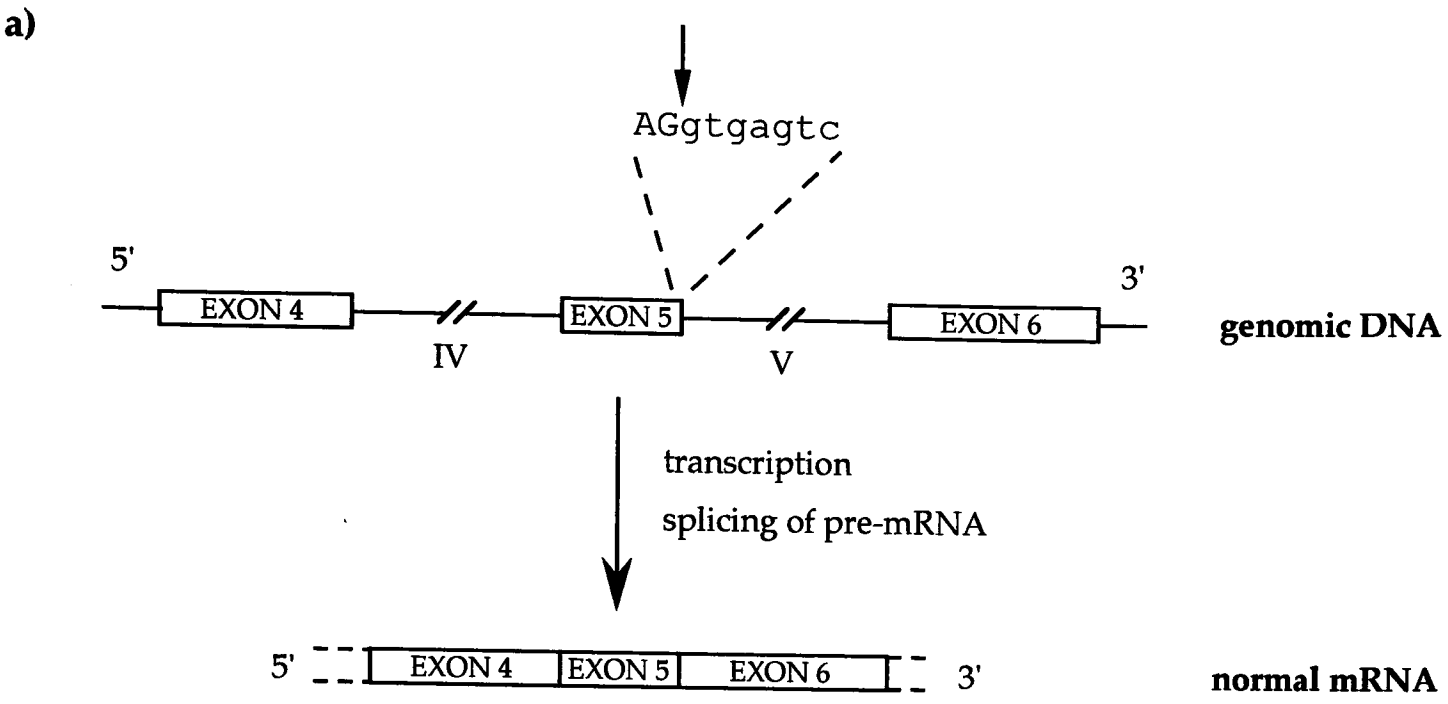


Figure 6.20: Exon skipping resulting from CLCN5 donor splice site mutations in patients with Dent's disease.

a) Splicing of the exon 4-6 region of CLCN5 in normal individuals. The sequence of the wild type donor splice site for intron V is given above the genomic DNA, with exonic sequence in uppercase and intronic in lowercase. The exon-intron junction is also indicated with an arrow.

b) Mutations disrupting the intron V donor splice sites of patients in Families G and E lead to exon skipping and result in the loss of exon 5 from the processed CLCN5 transcript (Lloyd *et al.*, submitted). Mutated nucleotides are shown in boldface.

Chapter 7 – General Discussion

7.1 Mapping of the Xp11.23-p11.22 region - conclusions

In 1989, Olson *et al.* proposed that short segments of single-copy DNA sequence which could be recovered using the PCR technique should constitute the main landmarks of a physical map of the human genome. The use of these 'sequence tagged sites' (STSs) would provide a common language to facilitate the merging of the large amounts of mapping information being generated by diverse methods in many different laboratories worldwide. In addition, the managment of data in the form of STSs would overcome many of the difficulties associated with the storage and distribution of cloned fragments of DNA. It was suggested that the construction of an STS map with an average spacing of no more than 100kb between markers was a reasonable mid-term goal for mapping of the human genome (Olson *et al.*, 1989). In the six years which have followed this proposal, a vast quantity of genetic and physical mapping data has been combined to provide comprehensive maps of the human chromosomes, and the target of a 100kb average inter-STS distance for the entire genome is rapidly being approached.

According to the report of the Sixth International Workshop on X chromosome mapping (Nelson *et al.*, 1995), the X chromosome is now completely covered with an average inter-STS resolution of 500kb, 93% covered with a resolution of 200kb, and over 60% covered with a resolution of 100kb. Furthermore, 90% of the chromosome is represented in YAC contigs with an average clone depth of four-fold coverage. The construction and characterization of the Xp11.23-p11.22 YAC contigs reported in this thesis, combined with work presented by Hatchwell (1994) and Chand (1994), has made a significant contribution to this mapping effort. At the start of this project, information regarding the OATL1–DXS146 interval was sparse; although YAC clusters had been isolated for both of these markers, only a single YAC (the DXS255 clone known as F1001) had been identified from the intervening region.

Fifteen new YAC clones were isolated during the course of this thesis, and seventeen markers, including a polymorphic CA repeat, an exon from a putative novel calcium channel gene and two disease genes (CLCN5 and WASP), have been mapped within them. STSs have been developed for most of these markers, and the remaining few can be converted to STSs with ease, if and when required.

Other laboratories have also been accumulating physical mapping data for Xp11.23-p11.22 in recent years, and several markers which have not been described in this thesis are now known to map between OATL1 and DXS146 (Nelson *et al.*, 1995). These are clustered in the OATL1-GATA-TFE3 region, and include polymorphic CA repeats and newly isolated transcripts. It will be important to localize these additional markers within the YAC contigs of Chapters 3 and 4, in order to aid the integration of the map described here with those presented by other groups. Sequence information from several of these markers is available in computer databases; it should therefore be straightforward to position them within the YACs using a PCR approach.

Further studies will also be necessary in order to bridge the uncloned gap between the OATL1–GATA–TFE3–SYP cluster and the DXS255–DXS146 contig. Given that the most recent estimate of SYP-DXS255 distance, based on pulsed field genomic mapping, is only ~900kb (Meindl *et al.*, 1995), it seems likely that screening of YAC libraries with A(E0250), the most distal marker from the DXS255–DXS146 cluster, will yield bridging clones.

As stressed in Chapters 3 and 4, certain regions of the genome are inherently unstable when cloned into YACs. It is essential to detect these, since the YAC contigs that span them may not provide a faithful representation of genomic DNA. Three such regions of instability were identified in the contigs presented here; one distal to the B0617 YAC clone, another involving the GATA locus, and a third in the vicinity of the SYP gene. Analysis of cosmids containing SYP suggests that the employment of alternative vector systems can circumvent the difficulties encountered when using YACs. It appears likely that a fully representative clone contig of Xp11.23-p11.22 will only be obtained when it incorporates YACs, cosmids and possibly additional types of clone, in vectors such as P1 phage.

The construction of such a contig in the TFE3-SYP interval is needed in order to aid the precise localization of the breakpoint associated with papillary renal cell carcinoma (see Chapter 4). It should be noted that a YAC contig presented by Boycott *et al.* (1995) at the sixth X chromosome workshop is reported to provide complete coverage from OATL1 to DXS146, but has not yet been published in detail. It is therefore unclear at this stage whether this contig has overcome the problems of instability which have been found for this region.

The polymorphic marker L(B0617) (also known as DXS6666) which was isolated in this study has already been used in linkage analysis of families affected with X-linked RP (Chand, 1994). This has led to the identification of a double recombination event in one pedigree which appears to place the RP2 gene proximal to this marker. However, such a conclusion relies on the supposition that this family is segregating RP2, rather than RP3, and at present there are insufficient data to resolve this issue (see Section 1.3.2).

Cornélis *et al.* (1992) have demonstrated that subcloning of YACs, followed by hybridization with a polydinucleotide probe, can be used to isolate polymorphic dinucleotide repeats from a given chromosomal region. They estimate that any 500kb fragment is likely to contain at least one highly polymorphic CA repeat. The OATL1–GATA–TFE3–SYP and DXS255–DXS146 contigs therefore represent a valuable resource for the identification of novel polymorphic markers mapping in Xp11.23-p11.22 which will be useful for future linkage studies.

Several different strategies for identifying coding regions have been exploited in this project:

i) The gene responsible for Dent's disease was successfully isolated by direct screening of a kidney-specific cDNA library using an entire YAC.

ii) Sequence analysis of a single copy clone from a SYP cosmid indicated the presence of an exon from a novel locus showing high homology to calcium channels.

iii) Rare-cutter restriction mapping of a subset of YACs from the DXS255–DXS146 contig detected four CpG islands which are likely to be associated with the 5' ends of genes.

These studies have therefore provided a basis for further analysis of novel transcripts that map in Xp11.23-p11.22.

Adams *et al.* (1995) have recently reported the determination of almost 300,000 partial sequences from randomly isolated cDNA clones. The availability of these expressed sequence tags (ESTs), in combination with data from comprehensive YAC contigs covering a large proportion of the human genome, is likely to change our approaches to gene hunting. Major mapping efforts are now underway in order to place the ESTs within physical maps of human chromosomes. Thus, it may soon be possible to identify transcripts that map within a target region for a particular disease locus simply by screening a computer database of mapped ESTs. Ballabio (1993) and Collins (1995) have suggested that, in years to come, as the density of the transcript map increases, this kind of 'positional candidate' strategy will become the main tool for the isolation of disease genes.

7.2 Determining the role of CLCN5 in kidney function

A pure positional cloning effort results in the isolation of a novel gene in complete absence of knowledge regarding the protein it encodes. Homology searches of databases containing previously identified sequences can then provide insights into the putative function of the predicted gene product. When employed for the analysis of the Dent's disease candidate gene, such studies suggested that it encodes a new member of a rapidly expanding family of voltage-gated chloride channels. In addition to the original member of this family, ClC-0, isolated from the electric ray, and the rat and human chloride channels described in Chapter 5, ClC genes have now been identified in several distantly related organisms, including the rabbit *Oryctolagus cuniculus* (Malinowska *et al.*, 1995), the yeast *Saccharomyces cerevisiae* (Greene *et al.*, 1993; Huang *et al.*, 1994), the wheat glume blotch fungus *Septoria nodorum* (Borsani *et al.*, 1995) and even the bacteria *Escherichia coli* (Fujita *et al.*, 1994).

The rabbit gene, known as ClC-2G, is reported to encode a gastric chloride channel which is activated by cAMP-dependent protein kinase and has high homology to the ubiquitously expressed rat and human ClC-2 proteins (Malinowska *et al.*, 1995). The yeast GEF1 gene was initially identified as the gene that was mutated in a new class of respiration-defective (petite) *S. cerevisiae* mutants, which grow slowly on rich media containing non-fermentable carbon sources (Greene *et al.*, 1993). The mutant phenotype can be suppressed by adding high concentrations of iron to the growth medium. On the basis of this, and the fact that iron uptake is unaffected in *gef1*$^-$ cells, Greene *et al.* (1993) suggested that GEF1 is involved in an intracellular pathway of iron metabolism. However, homology studies indicated that GEF1 (also known as yClC-1) encodes a chloride channel, and it is not clear at this stage how such a transmembrane protein might influence iron metabolism. It has been proposed that yeast chloride channels play a role in the acidification of the vacuole, which appears to be a site of intracellular iron storage (Greene *et al.*, 1993). This is intriguing in light of the possibility, discussed in more detail below, that the function of ClC-5 may be to maintain acidification of the endosomal compartment in the mammalian renal proximal tubule. It should be noted, however, that initial studies have suggested that vacuolar acidification is, in fact, normal in *gef1*$^-$ cells (Greene *et al.*, 1993). The putative *E. coli* ClC gene was identified during a large-scale effort to sequence the entire genome of this bacteria (Fujita *et al.*, 1994).

Interestingly, the human ClC-5 protein shows higher homology to GEF1 (58.4% similarity) and to the *E. coli* chloride channel (57.2% similarity) than to human family members ClC-2, ClC-1, ClC-Ka and ClC-Kb (56.9-51.9% similarity). The sequence studies described in Chapter 5 have shown that ClC-5, ClC-4 and ClC-3 form a distinct branch of the ClC family; it is possible that GEF1 represents the yeast homologue of these channels. The isolation and analysis of chloride channels from additional organisms should shed further light on the evolutionary relationships between members of this family.

Recent studies of CLCN4, which, like CLCN5, is found on the human X chromosome, have shown that its mouse homologue, Clcn4, maps to the X chromosome in *Mus spretus*, but to chromosome 7 in the laboratory mouse *Mus musculus* (Rugarli *et al.*, 1995; Palmer *et al.*, 1995). The gene, which is not observed to escape X-inactivation, and does not have a Y homologue, is therefore the first locus known to fully contravene Ohno's law, that dosage compensated genes on the X chromosome will be conserved as a linkage group in all placental mammals (Rugarli *et al.*, 1995; Palmer *et al.*, 1995). Preliminary analysis of mice with between zero and three copies of Clcn4 revealed no phenotypic abnormalities, suggesting that dosage is unimportant for this gene (Palmer *et al.*, 1995).

CLCN5 is the second member of the voltage-gated chloride channel family to be implicated in the aetiology of a human disorder. Mutations in CLCN1 are responsible for recessive and dominant forms of pure myotonia, a non-dystrophic skeletal muscle disorder (Koch *et al.*, 1992; George *et al.*, 1993) and disruption of the murine homologue has been found in myotonic mice (Steinmeyer *et al.*, 1991b). Detailed functional analysis of mutations causing the dominant form of myotonia indicates that ClC-1 is likely to have a tetrameric structure (Steinmeyer *et al.*, 1994).

Expression studies of human ClC-5 in Xenopus oocytes have confirmed that it functions as an ion channel with a high chloride selectivity (T. J. Jentsch, personal communication). In addition, the chloride conductance observed in oocytes injected with cRNA from wild type CLCN5, was completely abolished, or markedly reduced when mutations identified in Dent's disease, XRN and XLRH patients were introduced into the gene (T. J. Jentsch, personal communication). There is therefore little doubt that disruption of CLCN5 activity is the primary molecular genetic defect responsible for these disorders. However, further studies of this gene are now necessary, in order to elucidate the precise mechanism by which loss of function of a kidney-specific chloride channel may lead to symptoms such as LMW proteinuria, hypercalciuria and nephrolithiasis.

It is only possible, at this stage, to speculate on the putative role of ClC-5 in the normal kidney. Many functions of the proximal tubule, including uptake of luminal protein and recycling of the apical plasma membrane, are dependent on the acidification of its endosomal compartment (see Reeves and Andreoli, 1992). Electrophysiological analysis has shown that endocytic chloride conductance, maintained by a chloride channel, is essential for such acidification, and that this conductance is regulated by cAMP-dependent protein kinase (Bae and Verkman, 1990). It is plausible that disruption of an endocytic chloride channel might lead to LMW proteinuria, damage to the membrane of the proximal tubule and consequent renal dysfunction. In addition, it has been shown that chloride channels are present in the distal convoluted tubule, where changes in chloride conductance, regulated by calcitonin or parathyroid hormone, play an important role in mediating calcium resorption (Gesek and Friedman, 1992, 1993). Furthermore, clinical studies of patients suffering from sporadic hypercalciuric nephrolithiasis have demonstrated a close relationship between dietary chloride intake and the level of calcium in the urine (Muldowney *et al.*, 1994).

Identification of the main sites of CLCN5 expression within the kidney nephron, by *in situ* hybridization or RT-PCR, will help to establish the function of this chloride channel. Future work is likely to involve the construction of a transgenic mouse in which Clcn5 is disrupted, with the aim of providing an animal model for hypercalciuric nephrolithiasis, and characterization of the control regions involved in the expression of the gene.

Questions also remain regarding the phenotypic differences between Dent's disease, XRN and XLRH. Oocyte expression assays of mutations associated with these different forms of kidney disorder yielded no significant differences in activity of the mutant channel (T. J. Jentsch, personal communication). This indicates that modifying genes and environmental factors such as vitamin D intake or levels of calcium in the diet, may have a significant effect on the progression of the disease. Such factors may also explain the variation in phenotype between different affected members of the same pedigree who have identical CLCN5 mutations.

As described in Chapter 6 of this thesis, all exon-intron boundaries within the CLCN5 coding region have now been characterized. This has facilitated the development of genomic PCR assays which will increase the ease of mutational screening at this locus in further patients with renal tubular disease. In the past year several additional families segregating X-linked hypercalciuric nephrolithiasis have come to light (R. V. Thakker, personal communication). Furthermore, disorders associated with renal stones account for up to 1% of hospital admissions (Smith, 1989) and it is possible that CLCN5 may be implicated in a signficant proportion of these. The positional cloning of this gene therefore represents an important step towards the understanding and treatment of kidney dysfunction.

Bilbliography

Abidi, F. E., Wada, M., Little, R. D. and Schlessinger, D. (1990). Yeast artificial chromosomes containing human Xq24-Xq28 DNA: library construction and representation of probe sequences. *Genomics* **7**: 363-376

Abrahamson, G., Fraser, N. J., Boyd, Y., Craig, I. and Wainscoat, J. S. (1990). A highly informative X-chromosome probe, M27β, can be used for the determination of tumor clonality. *Br. J. Haematol.* **74**: 371-372

Adachi, S., Uchida, S., Ito, H., Hata, M., Hiroe, M., Marumo, F. and Sasaki, S. (1994). Two isoforms of a chloride channel predominantly expressed in thick ascending limb of Henle's loop and collecting ducts of rat kidney. *J. Biol. Chem.* **269**: 17677-17683

Adams, M. D. *et al.* (1995). Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377 Suppl.**: 3-174

Albertsen, H. M., Abderrahim, H., Cann, H. M., Dausset, J., Le Paslier, D. and Cohen, D. (1990). Construction and characterization of a yeast artificial chromosome library containing seven haploid human genome equivalents *Proc. Natl. Acad. Sci. USA* **87**: 4256-4260

Alitalo, T., Kruse, T. A., Forsius, H., Eriksson, A. W. and de la Chapelle, A. (1991). Localization of the Åland Island eye disease locus to the pericentromeric region of the X chromosome by linkage analysis. *Am. J. Hum. Genet.* **48**: 31-38

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**: 403-410

Anand, R. (1986). Pulsed field gel electrophoresis: a technique for fractionating large DNA molecules. *Trends in Genetics* **2**: 278-283

Anand, R., Villasante, A. and Tyler-Smith, C. (1989). Construction of yeast artificial chromosome libraries with large inserts using fractionation by pulsed field gel electrophoresis. *Nucleic Acids Res.* **17**: 3425-3433

Anand, R., Riley, J. H., Butler, R., Smith, J. C. and Markham, A. F. (1990). A 3.5 genome equivalent multi access YAC library: construction, characterization, screening and storage. *Nucleic Acids Res.* **18**: 1951-1956

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H. L., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., *et al.* (1981). Sequence and organization of the human mitochondrial genome. *Nature* **290**: 457-465

Bae, H.-R. and Verkman, A. S. (1990). Protein kinase A regulates chloride conductance in endocytic vesicles from proximal tubule. *Nature* **348**: 637-639

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

Barr, M. L. and Bertram, E. G. (1949). A morphological distinction between neurones of the male and female, and the behaviour of the nucleolar satellite during accelerated nucleoprotein synthesis. *Nature* **163**: 676-677

Benham, F., Hart, K., Crolla, J., Bobrow, M., Francavilla, M. and Goodfellow, P. N. (1989). A method for generating hybrids containing nonselected fragments of human chromosomes. *Genomics* **4**: 509-517

Bhattacharya, S. S., Wright, A. F., Clayton, J. F., Price, W. H., Phillips, C. I., McKeown, C. M. E., Jay, M., Bird, A. C., Pearson, P. L., Southern, E. M., *et al.* (1984). Close genetic linkage between X-Linked retinitis pigmentosa and a restriction fragment length polymorphism identified by the recombinant DNA probe L1.28 *Nature* **309**: 253-255

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 in Genetics* **3**: 342-347

Black, G. C. M, D. Phil. Thesis, Oxford University (1994).

Blake, C. C. F. (1978). Do genes-in-pieces imply proteins-in-pieces? *Nature* **273**: 267

Bolino, A., Devoto, M., Enia, G., Zoccali, C., Weissenbach, J. and Romeo, G. (1993). Genetic mapping in the Xp11.22 region of a new form of X-linked hypophosphatemic rickets. *Eur. J. Hum. Genet.* **1**: 269-279

Borsani, G., Rugarli, E. I., Taglialatela, M., Wong, C. and Ballabio, A. (1995). Characterization of a human and murine gene (ClCN3) sharing similarities to voltage-gated chloride channels and to a yeast integral membrane protein. *Genomics* **27**: 131-141

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

Boycott, K. M., Halley, G., Schlessinger, D. and Bech-Hansen, N. T. (1995) *Abstract. Sixth X chromosome workshop. Banff, Alberta.*

Boyd, Y. and Fraser, N. J. (1990). Methylation patterns at the hypervariable X-chromosome locus DXS255 (M27β): correlation with X-inactivation status. *Genomics* **7**: 182-187

Breathnach, R., Mandel, J. L. and Chambon, P. (1977). Ovalbumin gene is split in chicken DNA. *Nature* **270**: 314-319

Brook, J. D., McCurrach, M. E., Harley, H. G., Buckler, A. J., Church, D., Aburatani, H., Hunter, K., Stanton, V. P., Thirion, J.-P., Hudson, T., *et al.* (1992). Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. *Cell* **68**: 799-808

Brooks-Wilson, A. R., Goodfellow, P. N., Povey, S., Nevanlinna, H. A., de Jong, P. J. and Goodfellow, P. J. (1990). Rapid cloning and characterization of new chromosome 10 DNA markers by *Alu* element-mediated PCR. *Genomics* **7**: 614-620

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

Brown, C. J. and Willard, H. F. (1990). Localization of a gene that escapes inactivation to the X chromosome short arm: implications for X inactivation. *Am. J. Hum. Genet.* **46**: 273-279

Brown, C. J., Ballabio, A., Rupert, J. L., Lafreniere, R. G., Grompe, M., Tonlorenzi, R. and Willard, H. F. (1991). A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* **349**: 38-44

Brownstein, B. H., Silverman, G. A., Little, R. D., Burke, D. T., Korsmeyer, S. J., Schlessinger, D. and Olson, M. V. (1989). Isolation of single-copy human genes from a library of yeast artificial chromosome clones. *Science* **224**: 1348-1351

Buckler, A. J., Chang, D. D., Graw, S. L., Brook, J. D., Haber, D. A., Sharp, P. A. and Housman, D. E. (1991). Exon amplification: a strategy to isolate mammalian genes based on RNA splicing. *Proc. Natl. Acad. Sci. USA* **88**: 4005-4009

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

Burke, D. T. and M.V., O. (1991). Preparation of clone libraries in Yeast Artificial Chromosome Vectors *Methods in Enzymol.* **194**: 251-270

Buxton, J., Shelbourne, P., Davies, J., Jones, C., Van Tongeren, T., Aslanidis, C., de Jong, P., Jansen, G., Anvret, M., Riley, B., *et al.* (1992). Detection of an unstable fragment of DNA specific to individuals with myotonic dystrophy. *Nature* **355**: 547-548

Caiulo, A., Nicolis, S., Bianchi, P., Zuffardi, O., Bardoni, B., Maraschio, P., Ottolenghi, S., Camerino, G. and Giglioni, B. (1991). Mapping the gene encoding the human erythroid transcription factor NFE1-GF1 to Xp11.23 *Hum. Genet.* **86**: 388-90

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., *et al.* (1990). Isolation and characterization of a zinc finger polypeptide gene at the human chromosome 11 Wilms' tumor locus. *Cell* **60**: 509-520

Carle, G. F. and Olson, M. V. (1985). An electrophoretic karyotype for yeast. *Proc. Natl. Acad. Sci. USA* **82**: 3756-3760

Chand, A., D. Phil. Thesis, Oxford University (1994).

Chen, Z.-Y., Hendriks, R. W., Jobling, M. A., Powell, J. F., Breakefield, X. O., Sims, K. B. and Craig, I. W. (1992). Isolation and characterization of a candidate gene for Norrie disease. *Nature Genet.* **1**: 204-208

Chien, A., Edgar, D. B. and Trela, J. M. (1976). Deoxyribonucleic acid polymerase from the extreme thermophile *Thermus aquaticus*. *J. Bacteriol.* **127**: 1550

Chu, G., Vollrath, D. and Davis, R. W. (1986). Separation of large DNA molecules by contour-clamped homogenous electric fields. *Science* **234**: 1582-1585

Chumakov, I. M., Le Gall, I., Billault, A., Ougen, P., Soularue, P., Guillou, S., Rigault, P., Bui, H., De Tand, M.-F., Barillot, E., *et al.* (1992). Isolation of chromosome 21-specific yeast artificial chromosomes from a total human genome library. *Nature Genet.* **1**: 222-225

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

Cid, L. P., Montrose-Rafizadeh, C., Smith, D. I., Guggino, W. B. and Cutting, G. R. (1995). Cloning of a putative human voltage-gated chloride channel (ClC-2) cDNA widely expressed in human tissues. *Hum. Mol. Genet.* **4**: 407-413

Clark, J., Roques, P. J., Crew, A. J., Gill, S., Shipley, J., Chan, A. M.-L., Gusterson, B. A. and Cooper, C. S. (1994). Identification of novel genes, SYT and SSX, involved in the t(x;18) (p11.2;q11.2) translocation found in human synovial sarcoma. *Nature Genet.* **7**: 502-508

Coe, F. L., Parks, J. H. and Asplin, J. R. (1992). The pathogenesis and treatment of kidney stones. *N. Engld. J. Med.* **327**: 1141-1152

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

Collins, F. S. (1992). Positional cloning: let's not call it reverse anymore. *Nature Genet.* **1**: 3-6

Collins, F. S. (1995). Positional cloning moves from perditional to traditional. *Nature Genet.* **9**: 347-350

Cornélis, F., Hashimoto, L., Loveridge, J., MacCarthy, A., Buckle, V., Julier, C. and Bell, J. (1992). Identification of a CA repeat at the TCRA locus using yeast artificial chromosomes: a general method for generating highly polymorphic markers at chosen loci. *Genomics* **13**: 820-825

Craig, J. M. and Bickmore, W. A. (1994). The distribution of CpG islands in mammalian chromosomes. *Nature Genet.* **7**: 376-382

Cremin, S. M., Greer, W. L., Bodok-Nutzati, R., Schwartz, M., Peacocke, M. and Siminovitch, K. A. (1993). Linkage of Wiskott-Aldrich syndrome with three marker loci, DXS426, SYP and TFE3, which map to the Xp11.3-p11.22 region *Hum. Genet.* **92**: 250-253

Darnell, J. E. and Doolittle, W. F. (1986). Speculations on the early course of evolution. *Proc. Natl. Acad. Sci. USA* **83**: 1271-1275

Dayhoff, M. O., Barker, W. C. and Hunt, L. T. (1983). Establishing homologies in protein sequences. *Meth. Enzymol.* **91**: 524-545

de Jong, B., Molenaar, I. M., Leeuw, J. A., Idenberg, V. J. S. and Oosterhuis, J. W. (1986). Cytogenetics of a renal adenocarcinoma in a 2-year-old child. *Cancer Genet. Cytogenet.* **21**: 165-169

de Leeuw, B., Berger, W., Sinke, R. J., Suijkerbuijk, R. F., Gilgenkrantz, S., Geraghty, M. T., Valle, D., Monaco, A. P., Lehrach, H., Ropers, H. H., *et al.* (1993). Identification of a yeast artificial chromosome (YAC) spanning the synovial sarcoma-specific t(X;18) (p11.2;q11.2) breakpoint. *Genes Chromosom. Cancer* **6**: 182-189

de Leeuw, B., Balemans, M., Weghuis, D. O., Senica, R., Janz, M., Geraghty, M. T., Gilgenkrantz, S., Ropers, H. H. and Geurts van Kessel, A. (1994). Molecular cloning of the synovial sarcoma-specific translocation (X; 18) (p11.2; q11.2) breakpoint *Hum. Mol. Gen.* **3**: 745-749

de Saint-Basile, G., Arveiler, B., Fraser, N. J., Boyd, Y., Craig, I. W., Griscelli, G. and Fischer, A. (1989). Close linkage of hypervariable marker DXS255 to disease locus of Wiskott-Aldrich syndrome. *Lancet* **2**: 1319-1320

de Saint-Basile, G., Schlegel, N., Caniglia, M., Le Deist, F., Kaplan, C., Lecompte, T., Piller, F., Fischer, A. and Griscelli, C. (1991). X-linked thrombocytopenia and Wiskott-Aldrich syndrome: similar regional assignment but distinct X-inactivation pattern in carriers. *Ann. Hematol.* **63**:

Deininger, P. L., Jolly, D. J., Rubin, C. M., Friedmann, T. and Schmid, C. W. (1981). Base sequence studies of 300 nucleotide renatured repeated human DNA clones. *J. Mol. Biol.* **8**: 4566-4569

Deininger, P. L. and Daniels, G. R. (1986). The recent evolution of mammalian repetitive DNA elements. *Trends Genet.* **2**: 76-80

Deininger, P. L., Batzer, M. A., Hutchison, C. A. I. and Edgell, M. H. (1992). Master genes in mammalian repetitive DNA amplification. *Trends Genet.* **8**: 307-311

Dent, C. E. and Friedman, M. (1964). Hypercalcuric rickets associated with renal tubular damage. *Arch. Dis. Child.* **39**: 240-249

Derry, J. M. J., Ochs, H. D. and Francke, U. (1994). Isolation of a novel gene mutated in Wiskott-Aldrich syndrome. *Cell* **78**: 635-644

Doel, M. T., Houghton, M., Cook, E. A. and Carey, N. H. (1977). The presence of ovalbumin mRNA coding sequences in multiple restriction fragments of chicken DNA. *Nucleic Acids Res.* **4**: 3701-3713

Dryja, T. P., McGee, T. L., Reichel, E., Hahn, L. B., Cowley, G. S., Yandell, D. W., Sandberg, M. A. and Berson, E. L. (1990). A point mutation of the rhodopsin gene in one form of retinitis pigmentosa. *Nature* **343**: 364-366

Dryja, T. P., Berson, E. L., Rao, V. R. and Opsian, D. D. (1993). Heterozygous missense mutation in the rhodopsin gene as a cause of congenital stationary night blindness. *Nature Genet.* **4**: 280-283

Duyk, G. M., Kim, S., Myers, R. M. and Cox, D. R. (1990). Exon trapping: a genetic screen to identify candidate transcribed sequences in cloned mammalian genomic DNA. *Proc. Natl. Acad. Sci. USA* **87**: 8995-8999

Favus, M. J. (1989). Familial forms of hypercalciuria *J. Urol.* **141**: 719-722

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

Feinberg, A. P. and Vogelstein, B. (1984). Addendum: a technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.* **137**: 266-267

Fickett, J. W. (1982). Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res.* **10**: 5303-5318

Francke, U., Ochs, H. D., de Martinville, B., Giacalone, J., Lindgren, V., Disteche, C., Pagon, R. A., Hofker, M. H., van Ommen, G., Pearson, P. L., *et al.* (1985). Minor Xp21 chromosome deletion in a male associated with expression of Duchenne muscular dystrophy, chronic granulomatous disease, retinitis pigmentosa, and McLeod syndrome. *Am. J. Hum. Genet.* **37**: 250-267

Franco, B., Guioli, S., Pragliola, A., Incerti, B., Bardoni, B., Tonlorenzi, R., Carrozzo, R., Maestrini, E., Pieretti, M., Taillon-Miller, P., *et al.* (1991). A gene deleted in Kallman's syndrome shares homology with neural cell adhesion and axonal path-finding molecules. *Nature* **353**: 529-536

Fraser, N., Boyd, Y., Brownlee, G. and Craig, I. W. (1987). Multi-allelic RFLP for M27β, an anonymous single copy genomic clone at Xp11.3-Xcen. *Nucleic Acids Res.* **15**: 9616

Fraser, N. J., Boyd, Y. and Craig, I. (1989). Isolation and characterization of a human variable copy number tandem repeat at Xcen-p11.22 *Genomics* **5**: 144-148

Fryomyer, P. A., Scheinman, S. J., Dunham, P. B., Jones, D. B., Hueber, P. and Schroeder, E. T. (1991). X-linked recessive nephrolithiasis with renal failure. *N. Engld. J. Med.* **325**: 681-686

Fryomyer, P. A., Scheinman, S. J., Dunham, P. B., Jones, D. B., Hueber, P. and Schroeder, E. T. (1991). X-linked recessive nephrolithiasis with renal failure. *N. Engld. J. Med.* **325**: 681-686

Fujita, N., Mori, H., Yura, T. and Ishihama, A. (1994). Systematic sequencing of the *Escherichia coli* genome: analysis of the 2.4-4.1 min (110,917-193,643bp) region. *Nucleic Acids Res.* **22**: 1637-1639

Gal, A., Schinzel, A., Orth, U., Fraser, N. A., Mollica, F., Craig, I. W., Kruse, T., Machler, M., Neugebauer, M. and Bleeker-Wagemakers, L. M. (1989). Gene of X-chromosomal congenital stationary night blindness is closely linked to DXS7 on Xp. *Hum. Genet.* **81**: 315-318

Gardiner-Garden, M. and Frommer, M. (1987). CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**: 261-282

Gealey, W. J., Dwyer, J. M. and Harley, J. B. (1980). Allelic exclusion of glucose-6-phosphate dehydrogenase in platelets and T lymphocytes from a Wiskott-Aldrich syndrome carrier. *Lancet* **1**: 63

George, A. L., Crackower, M. A., Abdalla, J. A., Hudson, A. J. and Ebers, G. C. (1993). Molecular basis of Thomsen's disease (autosomal dominant myotonia congenita). *Nature Genet.* **3**: 305-309

Gesek, F. A. and Friedman, P. A. (1992). On the mechanism of parathyroid hormone stimulation of calcium uptake by mouse distal convoluted tubule cells. *J. Clin. Invest.* **90**: 749-758

Gesek, F. A. and Friedman, P. A. (1993). Calcitonin stimulates calcium transport in distal convoluted tubule cells. *Am. J. Physiol.* **264**: F744-F751

Gilbert, W. (1978). Why genes in pieces? *Nature* **271**: 501

Gilbert, W., Marchionni, M. and McKnight, G. (1986). On the antiquity of introns. *Cell* **46**: 151-154

Gilgenkrantz, S., Chery, M., Teboul, M., Mujica, P., Leotard, B., Gregoire, M. J., Boman, F., Duprez, A. and Hanauer, A. (1990). Sublocalization of the X breakpoint in the translocation (X;18) (p11.2;q11.2) primary change in synovial sarcomas. *Oncogene* **5**: 1063-1066

Gish, W. and States, D. J. (1993). Identification of protein coding regions by database similarity search. *Nature Genet.* **3**: 266-272

Glass, I. A., Good, P., Coleman, M. P., Fullwood, P., Giles, M. G., Lindsay, S., Nemeth, A. H., Davies, K. E., Willshaw, H. A., Fielder, A., *et al.* (1993). Genetic mapping of a cone and rod dysfunction (Åland island eye disease) to the proximal short arm of the human X chromosome. *J. Med. Genet.* **30**: 1044-1050

Goodfellow, P. N., Pym, B., Mohandas, T. and Shapiro, L. J. (1984). The MIC2 locus escapes X-inactivation. *Am. J. Hum. Genet.* **36**: 777-782

Goodship, J., Carter, J., Espanol, T., Boyd, Y., Malcolm, S. and Levinsky, R. J. (1991). Carrier detection in Wiskott-Aldrich syndrome: combined use of M27β for X-inactivation studies and as a linked probe. *Blood* **77**: 2677-2681

Green, D. and Olson, M. V. (1990). Systematic screening of yeast artificial-chromosome libraries by use of the polymerase chain reaction. *Proc. Nat. Acad. Sci. USA* **87**: 1213-1217

Green, E. D., Riethman, H. C., Dutchik, J. E. and Olson, M. V. (1991). Detection and characterization of chimeric yeast artificial-chromosome clones. *Genomics* **11**: 658-669

Greene, J. R., Brown, N. H., Di Domenico, B. J., Kaplan, J. and Eide, D. J. (1993). The GEF1 gene of *Saccharomyces cerevisiae* encodes an integral membrane protein; mutations in which have effects on respiration and iron-limited growth. *Mol. Gen. Genet.* **241**: 542-553

Greer, W. L., Somani, A.-K., Kwong, P. C., Peacocke, M., Rubin, L. A. and Siminovitch, K. A. (1990). Linkage relationships of the Wiskott-Aldrich syndrome to 10 loci in the pericentromeric region of the human X chromosome *Genomics* **6**: 568-571

Gribskov, M. and Burgess, R. R. (1986). Sigma factors from *E. coli, B. subtilis,* phage SP01, and phage T4 are homologous proteins. *Nucleic Acids Res.* **14**: 6745-63

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

Hatchwell, E., D. Phil. Thesis, Oxford University (1994).

Hendriks, R. W., Hinds, H., Chen, Z.-Y. and Craig, I. W. (1992). The hypervariable DXS255 locus contains a LINE-1 repetitive element with a CpG island that is extensively methylated on the active X chromosome. *Genomics* **14**: 598-603

Henthorn, P. S., Stewart, C. C., Kadesch, T. and Puck, J. M. (1991). The gene encoding human TFE3, a transcription factor that binds the immunoglobulin heavy-chain enhancer, maps to Xp11.22. *Genomics* **11**: 374-378

Hieter, P., Mann, C., Snyder, M. and Davis, R. W. (1985). Mitotic stability of yeast chromosomes: a colony color assay that measures nondisjunction and chromosome loss. *Cell* **40**: 381-392

Hohn, B. and Collins, J. (1980). A small cosmid for efficient cloning of large DNA fragments. *Gene* **11**: 291-298

Horowitz, D. S. and Krainer, A. R. (1994). Mechanisms for selecting 5' splice sites in mammalian pre-mRNA splicing. *Trends Genet.* **10**: 100-106

Huang, M. E., Chuat, J. C. and Galibert, F. (1994). A voltage-gated chloride channel in the yeast *Saccharomyces cerevisiae. J. Mol. Biol.* **242**: 595-598

Huntington's Disease Collaborative Research Group (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**: 971-983

Ingle, C., Williamson, R., de la Chapelle, A., Herva, R. R., Haapala, K., Bates, G., Willard, H. F., Pearson, P. and Davies, K. E. (1985). Mapping DNA sequences in a human X-chromosome deletion which extends across the region of the Duchenne muscular dystrophy mutation. *Am. J. Hum. Genet.* **37**: 451-462

Jeffreys, A. J. and Flavell, R. A. (1977). The rabbit beta-globin gene contains a large insert in the coding sequence. *Cell* **12**: 1097-1108

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

Jeffreys, A. J., Wilson, V. and Thein, S. L. (1985b). Individual-specific 'fingerprints' of human DNA. *Nature* **316**: 76-79

Jentsch, T. J., Steinmeyer, K. and Schwarz, G. (1990). Primary structure of *Torpedo marmorata* chloride channel isolated by expression cloning in *Xenopus* oocytes. *Nature* **348**: 510-514

Jentsch, T. J., Gunther, W., Pusch, M. and Schwappach, B. (1995). Properties of voltage-gated chloride channels of the ClC gene family. *Journal of physiology* **482**: 19-25

Jobling, M. A., D. Phil. Thesis, Oxford University (1991).

Kawasaki, M., Uchida, S., Monkawa, T., Miyawaki, A., Mikoshiba, K., Marumo, F. and Sasaki, S. (1994). Cloning and expression of a protein kinase C-regulated chloride channel abundantly expressed in rat brain neuronal cells. *Neuron* **12**: 597-604

Kazazian, H. H., Wong, C., Youssouffian, H., Scott, A. F., Phillips, D. G. and Antonirakis, S. E. (1988). Haemophilia A resulting from *de novo* insertion of L1 represents a novel mechanism for mutation in man. *Nature* **332**: 164-166

Kersanach, R., Brinkmann, H., Liaud, M.-F., Zhang, D.-X., Martin, X. and Cerff, R. (1993). Five identical intron positions in ancient duplicated genes of eubacterial origin. *Nature* **367**: 387-389

Kieferle, S., Fong, P., Bens, M., Vandewalle, A. and Jentsch, T. J. (1994). Two highly homologous members of the ClC chloride channel family in both rat and human kidney. *Proc. Natl. Acad. Sci. USA* **91**: 6943-6947

Knight, J. C., Reeves, B. R., Kearney, L., Monaco, A. P., Lehrach, H. and Cooper, C. S. (1992). Localization of the synovial sarcoma t(x;18) (p11.2;q11.2) breakpoint by fluorescence *in situ* hybridization. *Hum. Mol. Genet.* **1**: 633-637

Koch, M. C., Steinmeyer, K., Lorenz, C., Ricker, K., Wolf, F., Otto, M., Zoll, B., Lehmann-Horn, F., Grzeschik, K.-H. and Jentsch, T. J. (1992). The skeletal muscle chloride channel in dominant and recessive human myotonia. *Science* **257**: 797-800

Korenberg, J. R. and Rykowski, M. C. (1988). Human genome organization: *Alu*, lines and the molecular structure of metaphase chromosome bands. *Cell* **53**: 391-400

Kozak, M. (1987). An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.* **15**: 8125-8148

Kruse, T. A., Ahrens, P., Albertsen, H. M., Jorgensen, B. and Vestergaard, S. R. (1986). An anonyomous single copy X-chromosome clone, pTAK-8, identifies a frequent RFLP at Xp11-q12 *Nucleic Acids Res.* **14**: 1921

Kunkel, L. M., Monaco, A. P., Middlesworth, W., Ochs, H. D. and Latt, S. A. (1985). Specific cloning of DNA fragments absent from the DNA of a male patient with an X chromosome deletion. *Proc. Natl. Acad. Sci. USA* **82**: 4778-4782

Kwan, S.-P., Sandkuyl, L. A., Blaese, M., Kunkel, L. M., Bruns, G., Parmley, R., Skarshaug, S., Page, D. C., Ott, J. and Rosen, F. S. (1988). Genetic mapping of the Wiskott-Aldrich syndrome with two highly-linked polymorphic DNA markers. *Genomics* **3**: 39-43

Kwan, S.-P., Lehner, T., Hagemann, T., Lu, B., Blaese, M., Ochs, H., Wedgewood, R., Ott, J., Craig, I. W. and Rosen, F. S. (1991). Localization of the gene for Wiskott-Aldrich syndrome between two flanking markers, TIMP and DXS255 on Xp11.2-Xp11.3 *Genomics* **10**: 29-33

Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**: 105-132

Lafreniere, R. G., Geraghty, M. T., Valle, D., Shows, T. B. and Willard, H. F. (1991a). Ornithine aminotransferase-related sequences map to two nonadjacent intervals on the X chromosome short arm *Genomics* **10**: 276-279

Lafreniere, R. G., Brown, C. J., Powers, V. E., Carrel, L., Davies, K. E., Barker, D. F. and Willard, H. F. (1991b). Physical mapping of 60 DNA markers in the p21.1-q21.3 region of the human X chromosome. *Genomics* **11**: 352-363

Lamond, A. I. (1993). A glimpse into the spliceosome. *Curr. Opin. Genet. Dev.* **3**: 62-64

Larin, Z., Monaco, A. P. and Lehrach, H. (1991). YAC libraries containing large inserts from mouse and human DNA. *Proc. Natl. Acad. Sci. USA* **88**: 4123-4127

Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. (1992). CpG islands as gene markers in the human genome. *Genomics* **13**: 1095-1107

Laval, S. H. and Boyd, Y. (1993). Partial inversion of gene order within a homologous segment on the X chromosome. *Mammalian genome* **4**: 119-123

Lee, J. T., Murgia, A., Sosnoki, D. M., Olivos, I. M. and Nussbaum, R. L. (1992). Construction and characterization of a yeast artificial chromosome library for Xpter-Xq27.3: a systematic determination of co-cloning rate and X-chromosome representation *Genomics* **12**: 526-533

Legouis, R., Hardelin, J.-P., Levilliers, J., Claverie, J.-M., Compain, S., Wunderle, V., Millasseau, P., Le Paslier, D., Cohen, D., Caterina, D., *et al.* (1991). The candidate gene for the X-linked Kallman syndrome encodes a protein related to adhesion molecules. *Cell* **67**: 423-435

Lemann, J. J. and Gray, R. W. (1989). Idiopathic hypercalciuria. *J. Urol.* **141**: 715-718

Lindsay, S. and Bird, A. P. (1987). Use of restriction enzymes to detect potential gene sequences in mammalian DNA. *Nature* **327**: 336-338

Lorenz, C., Meyer-Kleine, C., Steinmeyer, K., Koch, M. C. and Jentsch, T. J. (1994). Genomic organization of the human muscle chloride channel ClC-1 and analysis of novel mutations leading to Becker-type myotonia. *Hum. Mol. Genet.* **3**: 941-946

Lovett, M., Kere, J. and Hinton, L. M. (1991). Direct selection: a method for the isolation of cDNAs encoded by large genomic regions. *Proc. Natl. Acad. Sci. USA* **88**: 9628-9632

Lyon, M. F. (1961). Gene action in the X-chromosome of the mouse (*Mus musculus* L.) *Nature* **190**: 372

Lyon, M. F., Peters, J., Glenister, P. H., Ball, S. and Wright, E. (1990). The scurfy mouse mutant has previously unrecognized hematological abnormalities and resembles Wiskott-Aldrich syndrome. *Proc. Natl. Acad. Sci. USA* **87**: 2433-2437

Malinowska, D. H., Kupert, E. Y., Bahinski, A., Sherry, A. M. and Cuppoletti, J. (1995). Cloning, functional expression and characterization of a PKA-activated gastric chloride channel. *Am. J. Physiol.* In press.

Mattick, J. S. (1994). Introns: evolution and function. *Curr. Opin. Genet. Dev.* **4**: 823-831

McKusick, V. (1992) Mendelian inheritance in man. Vol. 10, Baltimore and London: John Hopkins University Press.

McKusick, V. (1995) On-line Mendelian inheritance in man (OMIM). John Hopkins University of Medicine.

Meindl, A., de Carvalho, M. R. S., Schindelhauer, D., Herrman, K., Grimm, L., Wehnert, M., Ross, M. T. and Meitinger, T. (1995) *Abstract. Sixth X chromosome workshop. Banff, Alberta.*

Meitinger, T., Fraser, N. A., Lorenz, B., Zrenner, E., Murken, J. and Craig, I. W. (1989). Linkage of X-linked retinitis pigmentosa to the hypervariable DNA marker M27beta (DXS255) *Hum. Genet.* **81**: 283-286

Meloni, A. M., Dobbs, R. M., Pontes, J. E. and Sandberg, A. A. (1993). Translocation (X;1) in papillary cell carcinoma. A new cytogenetic subtype *Cancer Genet. Cytogenet.* **65**: 1-6

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

Muldowney, F. P., Freaney, R. and Barnes, E. (1994). Dietary chloride and urinary calcium in stone disease. *Quart. J. Med.* **87**: 501-509

Murray, J. M., Davies, K. E., Harper, P. S., Meredith, L., Mueller, C. R. and Williamson, R. (1982). Linkage relationship of a cloned DNA sequence on the short arm of the X chromosome to Duchenne muscular dystrophy. *Nature* **300**: 69-71

Musarella, M. A., Weleber, R. G., Murphey, W. H., Young, R. S. L., Anson-Cartwright, L., Mets, M., Kraft, S. P., Polemeno, R., Litt, M. and Worton, R. G. (1989). Assignment of the gene for complete X-linked congenital stationary night blindness (CSNB1) to Xp11.3. *Genomics* **5**: 727-737

Musarella, M. A., Anson, C. C., McDowell, C., Burghes, A. H., Coulson, S. E., Worton, R. G. and Rommens, J. M. (1991). Physical mapping at a potential X-linked retinitis pigmentosa locus (RP3) by pulsed-field gel electrophoresis. *Genomics* **11**: 263-272

Myerowitz, R., Piekarz, R., Neufeld, E. F. and Shows, T. B. (1985). Human β-hexosaminidase α chain: coding sequence and homology with the β chain. *Proc. Natl. Acad. Sci. USA* **82**: 7830-7834

Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E., *et al.* (1987). Variable number of tandem repeats (VNTR) markers for human gene mapping. *Science* **235**: 1616-1622

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol.* **48**: 443-453

Nelson, D. L., Ledbetter, S. A., Corbo, S. A., Victoria, M. F., Ramirez-Solis, R., Webster, T. D., Ledbetter, D. H. and Caskey, C. T. (1989). *Alu* polymerase chain reaction: A method for rapid isolation of human-specific sequences from complex DNA sources. *Proc. Natl. Acad. Sci.* **86**: 6686-6690

Nelson, D. L., Ballabio, A., Cremers, F., Monaco, A. P. and Schlessinger, D. (1995) Report of the sixth international workshop on X chromosome mapping. Banff, Alberta.

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

Ochman, H., Gerber, A. S. and Hartl, D. L. (1988). Genetic applications of an inverse PCR. *Genetics* **120**: 621-623

Ohjimi, Y., Iwasaki, H., Ishiguro, M., Hara, H., Ohgami, A., Kikuchi, M. and Kaneko, Y. (1993). Deletion (X)(p11): another case of renal adenocarcinoma with involvement of Xp11 *Cancer Genet. Cytogenet.* **70**: 77-78

Olson, M., Hood, L., Cantor, C. and Botstein, D. (1989). A common language for physical mapping of the human genome. *Science* **245**: 1434-1435

Ott, J., Bhattacharya, S., Chen, J. D., Denton, M. J., Donald, J., Dubay, C., Farrar, G. J., Fishman, G. A., Frey, D., Gal, A., *et al.* (1990). Localizing multiple X chromosome-linked retinitis pigmentosa loci using multilocus homogeneity tests. *Proc. Natl. Acad. Sci. USA* **87**: 701-704

Ozcelik, T., Lafreniere, R. G., Archer, B. T., Johnston, P. A., Willard, H. F., Francke, U. and Sudhof, T. C. (1990). Synaptophysin: structure of the human gene and assignment to the X chromosome in man and mouse. *Am. J. Hum. Genet.* **47**: 551-561

Palmer, S., Perry, J. and Ashworth, A. (1995). A contravention of Ohno's law in mice. *Nature Genet.* **10**: 472-476

Parimoo, S., Patanjali, S. R., Shukla, H., Chaplin, D. D. and S.M., W. (1991). cDNA selection: efficient PCR approach for the selection of cDNAs encoded in large chromosomal DNA fragments. *Proc. Natl. Acad. Sci. USA* **88**: 9623-9627

Parkinson, D. B. and Thakker, R. V. (1992). A donor splice site mutation in the parathyroid hormone gene is associated with autosomal recessive hypoparathyroidism. *Nature Genet.* **1**: 149-153

Parra, I. and Windle, B. (1993). High resolution visual mapping of stretched DNA by fluorescent hybridization. *Nature Genet.* **5**: 17-21

Peacocke, M. and Siminovitch, K. A. (1987). Linkage of the Wiskott-Aldrich syndrome with polymorphic DNA sequences from the human X chromosome. *Proc. Natl. Acad. Sci. USA* **84**: 3430-3433

Pearson, R. B. and Kemp, B. E. (1991). Protein kinase phosphorylation site sequences and consensus specificity motifs: tabulations. *Meth. Enzymol.* **200**: 62-77

Pook, M. A., Wrong, O., Wooding, C., Norden, A. G. W., Feest, T. G. and Thakker, R. V. (1993). Dent's disease, a renal Fanconi syndrome with nephrocalcinosis and kidney stones, is associated with a microdeletion involving DXS255 and maps to Xp11.22. *Hum. Mol. Gen.* **2**: 2129-2134

Reeders, S. T., Breuning, M. H., Davies, K. E., NIcholls, R. D., Jarman, A. P., Higgs, D. R., Pearson, P. L. and Weatherall, D. J. (1985). A highly polymorphic DNA marker linked to adult polycystic kidney disease on chromosome 16. *Nature* **317**: 542-544

Reeves, B. R., Smith, S., Fisher, C., Warren, W., Knight, J., Martin, C., Chan, A. M., Gusterson, B. A., Westbury, G. and Cooper, C. S. (1989). Characterization of the translocation between chromosomes X and 18 in human synovial sarcomas. *Oncogene* **4**: 373-378

Reeves, W. B. and Andreoli, T. E. (1992). Renal epithelial chloride channels. *Ann. Rev. Physiol.* **54**: 29-50

Riley, S., D. Phil. Thesis, Oxford University (1993).

Rio, D. C. (1993). Splicing of pre-mRNA: mechanism, regulation and role in development. *Curr. Opin. Genet. Dev.* **3**: 574-584

Rogers, J. H. (1989). How were introns inserted into nuclear genes? *Trends Genet.* **5**: 213-216

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

Rose, E. A., Glaser, T., Jones, C., Smith, C. L., Lewis, W. H., Call, K. M., Minden, M., Champagne, E., Bonetta, L., Yeger, H., *et al.* (1990). Complete physical map of the WAGR region of 11p13 localizes a candidate Wilms' tumor gene. *Cell* **60**: 495-508

Rugarli, E. I., Adler, D. A., Borsani, G., Tsuchiya, K., Franco, B., Hauge, X., Disteche, C., Chapman, V. and Ballabio, A. (1995). Different chromosomal localization of the Clcn4 gene in *Mus spretus* and C57BL/6J mice. *Nature Genet.* **10**: 466-471

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

Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A. and Arnheim, N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**: 1350-1354

Sanger, F., Nicklen, S. and Coulsen, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* **74**: 5463-5467

Scheinman, S. J., Pook, M. A., Wooding, C., Pang, J. T., Fryomyer, P. A. and Thakker, R. V. (1993). Mapping the gene causing X-linked recessive nephrolithiasis to Xp11.22 by linkage analysis. *J. Clin. Invest.* **91**: 2351-2357

Schwartz, D. C. and Cantor, C. R. (1984). Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell* **37**: 67-75

Schwartz, M. and Rosenberg, T. (1991). Åland eye disease: linkage data. *Genomics* **10**: 327-332

Sealey, P. G., Whittaker, P. A. and Southern, E. M. (1985). Removal of repeated sequences from hybridisation probes. *Nucleic Acids Res.* **13**: 1905-1922

Shapiro, L. J., Weiss, R., Buxman, M. M., Vidgoff, J., Dimond, R. L., Roller, J. A. and Wells, R. S. (1978). Enzymatic basis of X-linked icthyosis. *Lancet* **2**: 576

Silverman, G. A., Ye, R. D., Pollock, K. M., Sadler, J. E. and Korsmeyer, S. J. (1989). Use of yeast artificial chromosome clones for mapping and walking within human chromosome segment 18q21.3. *Proc. Natl. Acad. Sci. USA* **86**: 7485-7489

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

Sinke, R. J., de Leeuw, B., Janssen, H. A. P., Weghuis, D. O., Suijkerbuijk, R. F., Meloni, A. M., Gilgenkrantz, S., Berger, W., Ropers, H. H., Sandberg, A. A., *et al.* (1993). Localization of X chromosome short arm markers relative to synovial sarcoma and renal adenocarcinoma-associated translocation breakpoints. *Hum. Genet.* **92**: 305-308

Smith, L. H. (1989). The medical aspects of urolithiasis: an overview. *J. Urol.* **141**: 707-710

Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* **98**: 503-517

Southern, E. M., R., A., Brown, W. R. A. and Fletcher, D. S. (1987). A model for the separation of large DNA molecules by crossed field gel electrophoresis. *Nucleic Acids Res.* **15**: 5925-5943

Steinmeyer, K., Klocke, R., Ortland, C., Gronmeier, M., Jockush, H., Grunder, S. and Jentsch, T. J. (1991). Inactivation of a muscle chloride channel by transposon insertion in myotonic mice. *Nature* **354**: 304-308

Steinmeyer, K., Lorenz, C., Pusch, M., Koch, M. C. and Jentsch, T. J. (1994). Multimeric structure of ClC-1 chloride channel revealed by mutations in dominant myotonia congenita (Thomsen). *EMBO J.* **13**: 737-743

Steinmeyer, K., Ortland, C. and Jentsch, T. J. (1991). Primary structure and functional expression of a developmentally regulated skeletal muscle chloride channel. *Nature* **354**: 301-304

Stephens, R. M. and Schneider, T. D. (1992). Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *J. Mol. Biol.* **228**: 1124-1136

Sternberg, N. (1990). A bacteriophage P1 cloning system for the isolation, amplification, and recovery of DNA fragments as large as 100kbp *Proc. Natl. Acad. Sci. USA* **87**: 103-107

Stoltzfus, A., Spencer, D. F., Zuker, M., Logsdon, J. M. J. and Doolittle, W. F. (1994). Testing the exon theory of genes: the evidence from protein structure. *Science* **265**: 202-207

Suijkerbuijk, R. F., Meloni, A. M., Sinke, R. J., de Leeuw, B., Wilbrink, M., Jansenn, H. A. P., Geraghty, M. T., Monaco, A. P., Sandberg, A. A. and Geurts van Kessel, A. (1993). Identification of a yeast artificial chromosome that spans the human papillary renal cell carcinoma-associated t(X;1) breakpoint in Xp11.2. *Cancer Genet. Cytogenet.* **71**: 164-169

Tabor, S. and Richardson, C. C. (1989). Selective inactivation of the exonuclease activity of bacteriophage T7 DNA polymerase by in vitro mutagenesis. *J. Biol. Chem.* **264**: 6447-6458

Tanabe, T., Takeshima, H., Mikami, A., Flockerzi, V., Takahashi, H., Kangawa, K., Kojima, M., Matsuo, H., Hirose, T. and Numa, S. (1987). Primary structure of the receptor for calcium channel blockers from skeletal muscle. *Nature* **328**: 313-318

Thakker, R. V., Pang, J. T., Wooding, C., Scheinman, S. J., Wrong, O. M. and Pook, M. A. Mapping of two hereditary renal tubular disorders associated with kidney stones, and referred to as Dent's disease and X-linked recessive nephrolithiasis, to chromosome Xp11. (1994) *Abstract. Fifth X chromosome workshop. Heidelberg, Germany.*

Thiemann, A., Grunder, S., Pusch, M. and Jentsch, T. J. (1992). A chloride channel widely expressed in epithelial and non-epithelial cells. *Nature* **356**: 57-60

Tilghman, S. M., Tiemeier, D. C., Seidman, J. G., Peterlin, B. M., Sullivan, M., Maizel, J. V. and Leder, P. (1978). Intervening sequence of DNA identified in the structural portion of a mouse beta-globin gene. *Proc. Natl. Acad. Sci. USA* **75**: 725-729

Tomlinson, G., Nisen, P. D., Timmons, C. and Schneider, N. (1991). Cytogenetics of a renal cell carcinoma in a 17 month-old child. Evidence for Xp11.2 as a recurring breakpoint. *Cancer Genet. Cytogenet.* **57**: 11-17

Ton, C. C. T., Hirvonen, H., Miwa, H., Weil, M. M., Monaghan, P., Jordan, T., van Heyningen, V., Hastie, N. D., Meijers-Heijboer, H., Drechsler, M., *et al.* (1991). Positional cloning and characterization of a paired box- and homeobox-containing gene from the Aniridia region. *Cell* **67**: 1059-1074

Trask, B. J. (1991). Fluorescence *in situ* hybridization: applications in cytogenetics and gene mapping. *Trends Genet.* **7**: 149-154

Uberbacher, E. C. and Mural, R. J. (1991). Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. USA* **88**: 11261-11265

Uchida, S., Sasaki, S., Furukawa, T., Hiraoka, M., Imai, T., Hirata, Y. and Marumo, F. (1993). Molecular cloning of a chloride channel that is regulated by dehydration and expressed predominantly in kidney medulla. *J. Biol. Chem.* **268**: 3821-3824

Ullu, E., Murphy, S. and Melli, M. (1982). Human 7S RNA consists of a 140 nucleotide middle repetitive sequence inserted in an *Alu* sequence. *Cell* **29**: 195-202

van Ommen, G. J. B., Verkerk, J. M. H., Hofker, M. H., Monaco, A. P., Kunkel, L. M., Ray, P., Worton, R., Wieringa, B., Bakker, E. and Pearson, P. L. (1986). A physical map of 4 million bp around the Duchenne muscular dystrophy gene on the human X-chromosome. *Cell* **47**: 499-504

van Slegtenhorst, M. A., Bassi, M. T., Borsani, G., Wapenaar, M. C., Ferrero, G. B., de Conciliis, L., Rugarli, E. I., Grillo, A., Franco, B., Zoghbi, H. Y., *et al.* (1994). A gene from the Xp22.3 region shares homology with voltage-gated chloride channels. *Hum. Mol. Gen.* **3**: 547-552

Verkerk, A. J. M. H., Pieretti, M., Sutcliffe, J. S., Fu, Y.-H., Kuhl, D. P. A., Pizzuti, A., Reiner, O., Richards, S., Victoria, M. F., Zhang, F., *et al.* (1991). Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65**: 905-914

Villa, A., Notarangelo, L., Macchi, P., Mantuano, E., Cavagni, G., Brugnoni, D., Strina, D., Patrosso, M. C., Ramenghi, U., Sacco, M. G., *et al.* (1995). X-linked thrombocytopaenia and Wiskott-Aldrich syndrome are allelic diseases with mutations in the WASP gene. *Nature Genet.* **9**: 414-417

Wallace, M. R., Marchuk, D. A., Andersen, L. B., Letcher, R., Odeh, H. M., Saulino, A. M., Fountain, J. W., Brereton, A., Nicholson, J., Mitchell, A. L., *et al.* (1990). Type 1 Neurofibromatosis gene: identification of a large transcript disrupted in three NF1 patients. *Science* **249**: 181-186

Weber, J. L. (1990). Informativeness of human $(dC\text{-}dA)_n \cdot (dG\text{-}dT)_n$ polymorphisms. *Genomics* **7**: 524-530

Weber, J. L. and May, P. E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* **44**: 388-396

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

Willard, H. F., Cremers, F., Mandel, J.-L., Monaco, A. P., Nelson, D. L. and Schlessinger, D. (1994) Report of the fifth international workshop on human X chromosome mapping. *Cytogenet. Cell Genet.* **67**: 295-358

Wirth, B., Denton, M. J., Chen, J., Neugebauer, M., Halliday, F. B., van Schooneveld, M., Donald, J., Bleeker, W. E., Pearson, P. L. and Gal, A. (1988). Two different genes for X-linked retinitis pigmentosa. *Genomics* **2**: 263-266

Woodcock, D. M., Crowther, P. J., Diver, W. P., Graham, M., Bateman, C., Baker, D. J. and Smith, S. S. (1988). RglB facilitated cloning of highly methylated eukaryotic DNA: the human L1 transposon, plant DNA, and DNA methylated *in vitro* with human DNA methyltransferase. *Nucleic Acids Res.* **16**: 4465-4482

Wrong, O. M., Norden, A. G. W. and Feest, T. G. (1990). Dent's disease; a familial renal tubular syndrome with hypercalciuria, tubular proteinuria, rickets, nephrocalcinosis and eventual renal failure. *Quart. J. Med.* **77**: 1086-1087

Wrong, O. M., Norden, A. G. W. and Feest, T. G. (1994). Dent's disease; a familial proximal renal tubular syndrome with low-molecular-weight proteinuria, hypercalciuria, nephrocalcinosis, metabolic bone disease, progressive renal failure and a marked male predominance. *Quart. J. Med.* **87**: 473-493

Wyman, A. and White, R. (1980). A highly polymorphic locus in human DNA. *Proc. Natl. Acad. Sci. USA* **77**: 6754-6758

Yu, A. S., Herbert, S. C., Brenner, B. M. and Lytton, J. (1992). Molecular characterization and nephron distribution of a family of transcripts encoding the pore-forming subunit of calcium channels in the kidney. *Proc. Natl. Acad. Sci. USA* **89**: 10494-10498

Zon, L. I., Tsai, S.-F., Burgess, S., Matsudaira, P., Bruns, G. A. P. and Orkin, S. (1990). The major human erythroid DNA-binding protein (GF-1): Primary sequence and localization of the gene to the X chromosome. *Proc. Natl. Acad. Sci. USA* **87**: 668-672

Appendix

All sequence information is given in the 5' to 3' orientation.

i) Sequence Tagged Sites from novel Xp11.23-11.22 markers generated in this study

Regions corresponding to PCR primers are indicated in boldface.

L(F1001)

ATGAGTTCC**TTGTCTCTCTTCACCTTTTGC**CTTTCCATAACTCTTCAGTGGGGAGAAATTGGTGGGGA

ATGGATTCTGATGGAAGTCATTGAATA**GAGGGTAAACGAAAACAACC**TTTCAAA

L(6129)

The underlined region overlaps with (and is complementary to) the 3' end of the RL.6 cDNA clone (see Chapter 5).

GAATTCTAAG**GACTCTTGAGGGAGTCACAG**CTGCCTATNCTGTNCCAACNCCTTGAATATATTTGGAA

GCATCCCAAGCCTAATTCATGATCATGGGGACAACACACTGGGCCTGTTACACATACTTATC**GCTGGG**

AGGTACAATGAGTGTGCAGTTATTTTAGTACTGGTGGTCTTTACAAA

L(B0617)

GAATTCGCTAGTGGCCAAATAGCTATCA**CTTCTGGACCTGCAAAGAGG**CTGTTTGACTTCACCTTTCT

CTCTCTCTCGTCTCTCTCCCCCTTCCTCCTTCTCTCTCTCTCTCTCTCGACACACACACACACACACA

CACACACATACACACCACACACACACACACACACACACACACNAGGGTCTTATATTCAT**GGTTTAACTTC**

TATTGCTCAGGGNTTACTTATGAGTGGGGTGGGAATC

A(E0250)

TAG**GCAGACTCAAAAGGCCACAT**ATTGTATAGCTCCATTTATATGAAGTGTGCCTAGAATAGCCAAAT

CCATAGAGGCAGCAACTAGATTAGTGGCCTGGTGCTGAAGGTTATAGAAGCATTAAGGGTTGATAGCT

AAGGAGAACTGGGTTTCATTTTCAGGTTATGAAAGTGTTGCAAAGTTGACTGTGGTAATGTT**TGCACA**

ACTTTGTGAATGCAGTAAAATTAGTGAACTGTACACTTTAAGTTAGGGAATTGTATGGTATGTGAATT

ATATCACAATAAAGGTGTTACAAAATAAATCGCT

L(B102)

Letters in lowercase are homologous to *Alu* sequences.

GAATTCttttttttttttttgagacagagtttcgctcttgttgcctaggctgtagtgcagtggTGTGA

TCTTGGCTCATTGCAACCTCCACCTCCCGGGTTCAAGCGATTTTTCTGCCTCAGCCTCCCAAATAGCT

TGGGAT**TACAGGCATCCACCACCC**CGCCTGGCTAATTTTTTGTATTTTTAGTAGAGATAGGGTTTCAC

CATGTTGGCTCGGCTGGTCTTGAACTCCTGACCTCAGGTGATCCACCTGCGTCGGCCTCCCAAATTGC

TGGGATTACAGGCGTGNNCCACACCTGGCCAAAAATGAATTCTTGTTGAGTACTTCCTATGTGTTTAG

CTCTGT**GCTTAGATGCTTTTCCTGCCT**TATATTATTGGAGAAATCCTTCTATTCTCCCTTGTTGGATA

SAE

GAATTCATCAAGGTGATGGGGGCCCAGGATGACCAATCTCTGGTGAGATAATCAAAGTTGTCCATGAT

CACAGCCACA**AAGAGATTTATGATCTGTGGGC**CAGTGAGCAAGGGAGCAGTTAGGTGAAGGGCAGAAC

TCTGTCATGGACGGATGGTNGGAGGCTGGGNNNNCTTATATNGTCATNNGTCTCTNGGTTGATCTTGT

CATCAGAC**CAGCCCTCTCACCCACTATG**AAGATCTG

ii) cDNA sequence from the 3' untranslated region of the CLCN5 trasncript

The following sequence corresponds to the 3' end of RL.6 (i.e. nucleotides 3174-3681 of the cDNA contig presented in Chapter 5). The underlined region overlaps with (and is complementary to) the 3' end of the L(6129) clone:

GAGAGATTGAGCTTTTATAGCTTGCTTGTTGCTGGNACTCTTTTGAAGNGNACAAAGANAAGTTATAG

AGCTTTCTCCTTGTCTGNAAAAANGGTAATTTTCCAGGTGGNATTTGTTAGCTGGTAGACAAGGACTT

TGCNGTGAATTTGTTAGTAGCAAGGAATGATTTCTTCCAGCTTCCTTGAAATGAATAATTAGTAAACT

TCTAAGCAAAGTCAATGACTAGGAGTTTCACATGTTTGTGAGGTCCTAACTAATTCTCTTACCCAGAT

GCCACCTCCATGAATGATGGTGCTTTAGGCTTCTGGAATATGTAAAAACAGCAAAGGGAACCAAGTCT

AGACTGCATACTGGTAAACCAGGGAAGACTCAGAACTGCATATGCATCCTCATGCATTCCTTTGTAAA

GACCACCAGTACTAAAATAACTGGACACTCATTGTACCTCCCAGCGATAAGTATGTGTAACAGGCCAG

TGTGTTGTCCCCATGATCATGAATTAGGCTTGG

iii) Additional sequence information from exon-intron boundaries

Exonic sequence is given in uppercase, intronic in lowercase. See Chapter 6 for details of intron sizes.

Intron I

TTCCTACCAGgtgactgtatttctttatcaaccccaactgtagtcatctgatagtttaagggcccgcc

ttttggaaccaaaaagaatgtattaaaaatgttaagaatttctgaagtactaatgtctatttcttgtt

caatacagAGGACAAGTC

Intron II

GCACCGAGAGgtaagacaaaagatggcacatgggtaagtgttaggaaatacaggggaagaaattgaag

atacatattctttctattcttt.....

.....aaagattatgtgaaaagtagntctaaattggtctctattctcc

agtgatttgtctttgtatttctttgttgaacagnttttaagtgaaaaaatggtgtgtgtagagtgtac

catgttttctcattttcccctagATTACCAATA

Intron V

AATCCCTGAGgtgagtctcttaaaatggtttataaatggttacaatatgaatacttttttggttaaaat

ctcttaatatgtattccagcacatacctcaatttgatgaatgttaaacaggactcagatcctgaggcc

atctaagacctagg.....

.....ggtgcagtttctataatgaggtccagatttttgcttcttagccttgttgac

ttccttagtctatattcaatctttctgtgtttaacctgcagATAAAAACTA

Intron VII

CCTTGAAGAGgtaacaacttttcattgtacagcatgtgcatntttttgtgtcaggaattttgtacattg

cagcnataatttttgtacataatgtacaatatctgagtatatcccagtccctgagtgctctccccaact

tggtgctttaccatgtgactagaaa.....

.....tgactgagtttgctttctcaccttctttcttctagGTCAGC

TACT

Intron IX

GCCTGCTTAGgtgagtagtgtttgcattaatttcaagttgctacccaggtgacatacaacagaagagt

ttattaaacanagttgactgtaggagaatttaaa.....

.....ttcttgtagatgttattttgagaactctgaa

tcgtctccattgtgaccttgctttcttgcaatctctggggtctgnnnncagttgtaggtgagacctca

ttgtttttggtaggtcagntttcctggaaaggcccataattgtaagaatcctgtattttgcctacct

gagtagactgtgtctatttctttgcagGTGGGGTGAC

Intron X

ATTTCAATTGgtaaggatttcagaaaggggatagtggaatccactgtggaactcaataaatatnnctg

aaatnggggaagaca.....

.....aaggggggactggttaggagnnagaaggatagagctagcangtccatctt

caatttgtttttccttctgtttgaatagAAAATGCTCG

Intron XI

CACACAACGGgtaagaagtcttgagtgaagtcaaattgaattgtgggagaaagaggatgcagagatag

aaagaagtagaaagaagtagaaccagt.....

.....gggcacatgtcctactcctgtgatctcatactgaaagg

gcagctagtgatcaaagacaaagttgaaaaggactgaggaggacaagtatcacctttgggaatgctat

tttaactacagatttatttttgtttttttgtattgtgtttgtcttttagGCGATTGCTT