Chromosome 1 Map, Sequence and Variation

by

Simon Gray Gregory

Thesis submitted for the

degree of Doctor of Philosophy

The Open University

17th February 2003

The Wellcome Trust Sanger Institute

Wellcome Trust Genome Campus

Hinxton

Cambridge, UK

This thesis is dedicated to my wife, Deborah,
and my daughter, Olivia, for their unwavering
support, understanding, strength and love.

Also to my parents for their constant
encouragement and for instilling me
with a belief in my own abilities.

Abstract

The construction of well characterised sequence-ready physical maps has been central to the generation of high quality genomic sequence by the Human Genome Project. The technological advances that made possible a clone based sequencing approach to large genomes have included the use of large insert bacterial clones and the development of high throughput fingerprinting techniques.

The first part of this thesis is devoted to development and application of these improvements in technology. The adaptation of fluorescent technologies and their application to existing fingerprinting methods described in this work has resulted in a fingerprinting technique which improves upon levels of data accuracy, increases throughput and incorporates of increased levels of safety and automation. The initial application of this and other restriction digest fingerprinting methods to the assembly of large insert P1-artificial chromosome clones (PACs) was also evaluated. PACs were used to construct a 1.4 Mb contig across a region of chromosome 13q12 that includes the breast cancer susceptibility gene *BRCA2*. These experimental and technical developments were then utilised within a hierarchical mapping strategy to construct a 13 Mb contig of human chromosome 1pcen – 1p13.

The finished sequence generated by the clone based sequencing strategy provides the basis for the elucidation of genic features and the motifs that influence their regulation within the human genome sequence. Detailed analysis of the finished genomic sequence from 1pcen – 1p13 is described. These analyses include the characterisation of base composition and

determination of repeat content within the region, as well as identification of known and novel genes by manual annotation.

The majority of differences between individuals can be attributed to allelic sequence variation. The characterisation of sequence differences and comprehension of how they may affect the expression and function of genes will be crucial for the study of molecular alterations in human disease. A subset of highly similar genes within 1pcen – 1p13, in addition to seven other genes of interest, were investigated by developing and assessing assays to determine sequence variation. The particular challenges of investigating gene families where sequences are nearly identical were explored, and enable better resolution of new and previously available data. The consequences that these sequence changes may have upon gene function is also discussed, and this provides an example of the ways in which knowledge of genomic sequence can be analysed to support new areas of structural and functional research.

Acknowledgements

I would like to thank my supervisor, David Bentley, for fulfilling a promise that he made more than 10 years ago, by seeing me through to the submission of my doctorate. I have appreciated your advice, friendship and tolerance, during not only my PhD but throughout our working relationship. I would also like to thank my second supervisor, Richard Wooster, for his guidance and many informal 'thesis discussions'. Thanks are also extended to my long suffering office neighbour, Mark Ross, who has endured me barging into to his office with a multitude of questions and yet proffering an answer to *almost* every one.

Traditionally, a Sanger Institute PhD could not be completed without invaluable contributions from a multitude of individuals (all of whom I have aggravated at some stage). In particular, I wish to start by thanking Carol Scott for providing me with unflagging informatics support which extends beyond the work described in this thesis. I would also like to acknowledge Cordelia Langford, Nigel Carter, Laurent Baron, Andy Smith and Panos Deloukas for their help with the generation and assaying of the small insert library. Thanks to the awe-inspiring chromosome 1 sequencing team (Kirsten McLay, Becky Hall, Karen Barlow and the Sequencing Centre) for pushing my 1pcen – 1p13 clones through the sequencing pipeline and for carrying me on the C1 project! Assistance has also been gratefully received from Dave Beare, Stephen Keenan, James Gilbert, Jose Oliver and Jen Ashurst who helped me to analyse the 1pcen – 1p13 sequence. I have appreciated the help from Liz Sheridan, Jackie Bye and Liz Huckle with cDNA library screening and sequencing. Thanks also to the ExoSeq project (Andrew Bentley, Wilb Dunham, Sarah Hunt and Sarah Lindsay) for helping me to generate so much data so quickly. My thanks also go to Robert Steward, Jane Loveland and Tim Hubbard for opening my eyes to the wonderful world of proteins.

On a personal note, I would like to thank my many friends and colleagues at the Sanger whose help in large or small ways has made my molecular biological rites of passage more bearable. To Alison 'Pod' Coffey, Ian 'Roo' Barrett and Tamsin 'Mate' van de Ven thanks for all the proof reading, years of encouragement (Pod) and endless clichés to help me realise that I was *actually* making progress. To all the chromosome 1 mapping group members, past and present, many thanks for being a great bunch to work with and for allowing me to be a very 'hands off' group leader, particularly in the past 18 months. Thanks to Jane Rogers for your support, especially during the final stages of the thesis, and for accepting the same tired old excuses as to why the C1 project wasn't progressing as quickly you'd hoped. Thanks Sean and Lisa for being good friends over the years. To Gareth, thanks for treading the long grass before me; without your calming influence, good humour and friendship I could have finished the thesis a lot earlier, but been the worse for it!

"He was calm; however, he had to be supported during the journey through the long corridors, since he planted his feet unsteadily, like a child who has just learned to walk, or as if he were about to fall through like a man who has dreamt that he is walking on water only to have a sudden doubt: but is this possible?" (Vladimir Nabokov; 1899–1977)

Thanks to my mother and father for equipping me with the tools to make this arduous journey and to my wife, Deborah, and daughter, Olivia, without whose confidence and support I would ever have arrived.

Table of Contents page

List of figures:

Chapter 5

Chapter 6

List of tables:

Chapter 6

Chapter 7

Glossary of Abbreviations

1ace	1 chromosome version of ACeDB
ACeDB	*A C. elegans* database
AMPD2	adenosone monophosphate deaminase 2
Alu-PCR	*Alu*-element-mediated polymerase chain reaction
ATP (dATP, ddATP)	adenosine 5'-triphosphate (deoxy-, dideoxy-)
BAC	bacterial artificial chromosome
BLAST	basic local alignment search tool
BLIXEM	BLAST In an X-windows Embedded Multiple Alignment
β-ME	β-mercaptoethanol
bp	base pair
BSA	bovine serum albumin
^{o}C	degrees Celsius
CaM	calmodulin
cDNA	complementary deoxyribonucleic acid
chr	chromosome
CEPH	Centre d'Etude du Polymorphisme Humain
(c)M	(centi)Morgan
cm	centimetre
CDD	CONSERVED domain database
CpG	cytidyl phosphoguanosine dinucleotide
cR	centiRays
CTP (dCTP, ddCTP)	cytidine 5'-triphosphate (deoxy-, dideoxy-)
dbEST	database of expressed sequence tags
DNA	deoxyribonucleic acid
dNTP	2'-deoxyribonucleoside 5'-triphosphate
EDTA	ethylenediamine tetra-acetic acid
EMBL	European Molecular Biology Laboratory
EST	expressed sequence tag
FISH	fluorescence *in situ* hybridisation
FP	forward primer

FPC	FingerPrinted Contigs
g	gram
GDAP2	ganglioside-induced differentiation-associated protein 2
G banding	Geimsa banding
GDB	Genome Database
GSC	Genome Sequencing Centre, St Louis
GST (M)(T)(P)	glutathione S-transferase (mu) (theta) (pi)
GTP (dGTP, ddGTP)	guanine 5'-triphosphate (deoxy-, dideoxy-)
HapMap	haplotype block map
HGMP	Human Genome Mapping Resource Centre
HGNC	Human Genome Nomenclature Committee
HGP	Human Genome Project
H-W	Hardy-Weinberg
IHGSC	International Human Genome Sequencing Consortium
INSNPMWG	International SNP Map Working Group
kb	kilobase pairs
l	litre
LD	linkage disequilibrium
LINE	long interspersed nuclear element
LOH	loss of heterozygosity
M	molar
Mb	megabase pairs
MDS	myelodysplastic syndromes
μg	microgram
μl	microlitre
μM	micromolar
min(s)	minute(s)
mg	milligram
ml	millilitre
mm	millimetre
mM	millimolar
NCBI	National Centre for Biotechnology Information

NGFB	nerve growth factor – beta
NFE2L2	nuclear factor erythroid 2-like 2
NRAS	neuroblastoma RAS viral oncogene homolog
ng	nanogram
nm	nanometre
O/N	overnight
OD	optical density
OMIM	On-line Mendelian Inheritance in Man
ORF	open reading frame
PAC	P1-derived artificial chromosome
(e)PCR	(electronic) polymerase chain reaction
PDB	Protein Data Bank
PFAM	Protein Family
PFGE	pulsed-field gel electrophoresis
PNRC2	proline-rich nuclear receptor co-regulatory protein 2
poly(dT)	poly-deoxyribothymidyl oligonucleotide
R banding	Reverse Geimsa banding
RH	radiation hybrid
RFLP	restriction fragment length polymorphism
RNA (mRNA, rRNA, tRNA)	ribonucleic acid (messenger-, ribosomal-, transfer-)
RP	reverse primer
Rnase A	ribonuclease A
rpm	revolutions per minute
RT	room temperature
RT-PCR	reverse transcription polymerase chain reaction
SCL	stem cell leukaemia
SDS	sodium dodecyl sulphate
sec(s)	second(s)
seq	sequence
SIL	small insert library
SINE	short interspersed nuclear element
snoRNA	small nucleolar RNA

SNP	single nucleotide polymorphism
STS	sequence tagged site
TEMED	N,N,N',N'-tetramethylethylenediamine
TrEMBL	Translated EMBL
TSS	transcription start site
TSC	The SNP Consortium
TIGR	The Institute of Genome Research
Tris	tris(hydroxylmethyl)aminomethane
U	unit
UCSC	University of California Santa Cruz
UNR	upstream of NRAS, gene
UTR	untranslated region
uv	ultraviolet
V	volt
v/v	volume/volume
VNTR	variable number of tandem repeats
W	watt
w/v	weight/volume
Wash U.	Washington University
WG(S)	whole genome (shotgun)
XLA	X-linked agammaglobulinaemia
YAC	yeast artificial chromosome

Publications:

Parts of the work presented in this thesis have appeared previously in the following publications which are bound at the back of this thesis:

Gregory, S.G., Howell, G.R. and Bentley, D.R. Genome Mapping by Fluorescent Fingerprinting. *Genome Research* (1997) **7**: 1162-1168.

The International Human Genome Mapping Consortium. A physical map of the human genome Nature. 2001 Feb 15;409 (6822):934-41.

Bentley D.R., Deloukas P., Dunham A., French L., **Gregory S.G.**, Humphray S.J., Mungall A.J., Ross M.T., Carter N.P., Dunham I., *et al.* The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. *Nature*. 2001 Feb 15;409(6822):942-3.

Parts of the work presented in this thesis have appeared previously in the following publications which are not bound at the back of this thesis:

The International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409, 860-921.

Chapter 1

Introduction

In 1920, German botanist Hans Winkler first used the term ‘genome’, reputedly by the fusion of GENe and chromosOME, in order to describe the complex notion of the entire set of chromosomes and all the genes contained within an organism. A great deal of progress has since been made in the elucidation of the complex molecular interactions that underlie cellular functioning and the syntenic relationship between organisms at a nucleotide level.

The basis for these advances was the characterisation of the structure of DNA by Watson and Crick in 1953 (Watson *et al*., 1953), and the realisation that DNA could be decoded to provide a guide to genetic inheritance. This underpinned the concept of genetics and gave scientists the possibility to explore and quantify the nature and extent of the biological information passed on from one generation to the next. The characterisation of biological inheritance permitted the elucidation of what it was that was being encoded and how it could determine biochemical function. Finally, extending from elucidation of the mechanisms behind inheritance of monogenic diseases, scientists are beginning to grasp how sequence is also involved in complex interactions, occasionally under the influence of environmental factors, to contribute to many (but still not all) diseases. Whilst the generation of the complete sequence of the human genome may provide the starting point for the characterisation of all human disease, clinical diagnosis and classification by specialists remains central to the appropriate treatment of individuals suffering genetic disease.

The speed at which the vast amount of human sequence data was generated can be attributed to the evolution of strategies and techniques developed to sequence organisms such, as bacteria (Kohara *et al*.,1987), yeast (Olson *et al*., 1986) and the nematode worm

(Coulson *et al*., 1986). The availability of such an evolutionary diverse collection of sequences, with the addition of mouse (MGSC 2002) and other complex multi-cellular organisms, has also enabled comparisons to be made at a nucleotide level. These inter-species sequence comparisons, in conjunction with direct experimentation and computer based prediction programs, is facilitating the identification of evolutionarily conserved sequences, such as genes, and, to a lesser extent, the motifs that regulate them (Pennacchio *et al*., 2001).

The elucidation of the molecular complexity of gene structure and determining their regulation is, however, only the first step in understanding the intricate networks in which genes interact. Though sequence analysis and homology matching may assist to define a gene on the nucleotide level, determining the structure of the gene's product, the protein, and its function, is a difficult paradigm to resolve. Whilst traditional methods, for example X-ray crystallography, are capable of characterising the structure of a protein their speed of application precludes them from large scale protein analysis. However, *in silico* modelling using previously determined domains or three dimensional structures may provide a means of inferring function and help characterise the networks in which they are involved (Skolnick *et al*., 2000).

The identification of nucleotide differences (polymorphisms) between individuals, by comparison of high quality sequence, will assist our understanding of phenotypic variation. The localisation of single nucleotide polymorphisms (SNPs), particularly within functionally important sequences, such as genes, will contribute to our understanding of the aetiology and susceptibility to human disease and responsiveness to biochemical treatment. The identification of SNPs may also provide the opportunity to

partition the human genome into ancestral segments that have undergone minimal evolutionary recombination (haplotype blocks). Haplotype blocks, identified by linkage disequilibrium mapping, can then be used as a means of identifying multiple genes associated with complex phenotypes within unrelated individuals, where family-based studies are impossible because the complexity of the factors which contribute to the phenotype obscure any familial component. It is through approaches such as this that the true impact of the human genome sequence on human health and disease may bear the most fruit.

1.1 Mapping and sequencing model genomes

The first genomes that were characterised were relatively small by current standards, for example bacteriophage ϕX174 (5 kb; Sanger *et al*., 1977, Sanger *et al*., 1978) and bacteriophage λ (48 kb; Sanger *et al*., 1982), but they provided the underlying techniques and strategies that are being used for the more complex organisms currently being studied. Chain termination sequencing, developed by Sanger *et al* was a synthetic method, in which the nested sets of labelled fragments which constituted the sequence ladder were generated *in vitro* by a DNA polymerase reaction. The method was highly sensitive and robust. It was therefore amenable to biochemical optimisation to produce long, accurate sequence reads; and also to automation, which was necessary for large-scale application of the technique. In these respects it differed from the method of Maxam and Gilbert (1977), which necessitated production of all the labelled material prior to chemical degradation to form the sequence ladders of nested fragments. As a result, the synthetic method has remained the technique by which the majority of genomic sequence

from a variety of complex organisms is presently being generated, see figure 1.1. Neither method was capable of generating single reads of greater than 2-300 nucleotides, limited in part by the sequence ladder production itself, and partly by the ability to separate the sequence ladder by gel electrophoresis at single-base resolution (even today sequencing read-lengths approaching 1kb are rare). Assembly of larger tracts of DNA therefore required the development of methods to re-assemble a single sequence from multiple individual reads. Two approaches were adopted for this; first, the use of maps of restriction fragments, where multiple enzymes with sequence-specific cleavage activity are used, singly or in combination, to order and orientate segments of the sequence, which could be individually selected for sequencing; second, the use of the information gained from each individual sequence read to order and orient each segment relative to overlapping neighbours. This required the development of advanced computer programs to make the task possible on all but the smallest scale. In a further modification, the random shotgun strategy used by Anderson *et al*., (1981a) to elucidate the mitochondrial genome involved using a random fragmentation process, by partial DNAse I digestion (Anderson *et al*., 1981b). This removed the dependence on sequence-specific restriction enzymes, while still relying on sequence-based assembly of contiguous tracts of overlapping reads.

The random shotgun approach provided the basis of the strategies used to assemble sequences of large inserts cloned in plasmids, lambda phage and cosmid vectors, and also the later bacterial artificial chromosome (BAC) and P1-derived artificial chromosome (PAC) clones. The same strategy was adopted to sequence the 1.8 Mb genome of the bacterium *Haemophilus influenzae* (*H. influenzae*) (Fleischmann *et al*., 1995). Whilst the whole genome shotgun sequencing approach has proven itself to be a successful strategy

for the rapid assembly of smaller genomes, there are, however, doubts as to whether this strategy is suitable for assembling the sequence of complex organisms (as discussed in section 1.3).

The generation of a physical map, in which the genome is divided into bacterial clone units of 40 -200 kb and assembled into contiguous stretches (contigs) of overlapping clones, is a process analogous to the sequence contig assembly process. In contrast to sequence assembly, however, the information used to compare individual clones and identify overlaps, for the *C. elegans* (Coulson *et al*., 1986) and *S. cerevisiae* (Olson *et al*., 1986) genome projects, was a one-dimensional fingerprint, prepared by separating restriction fragments from a limit digest of each cloned DNA by electrophoresis. Overlaps between clones were detected on the basis of partially (or completely) shared fingerprint patterns. An alternative approach to identify overlapping relationships between clones was to test clones for the presence of characterised markers. Overlaps between clones could be identified on the basis that they shared a single copy sequence. The presence of the sequence was identified using a specific hybridisation probe or PCR assay.

Given a physical map of overlapping clones, individual clones can then be selected from the map to provide maximum genomic coverage with minimal redundancy. These clones permit specific regions to be targeted for further investigation, and in particular for determination of the complete DNA sequence separately from the other clones. Because the source of the genomic sequence is limited to an individual clone, problems encountered with sequence assemblies are greatly reduced compared to the corresponding whole genome assemblies.

At the time of their inception, the physical maps of the *Caenorhabditis elegans* (*C. elegans*) (Coulson *et al*., 1986) and *Saccharomyces cerevisiae* (*S. cerevisiae*) (Olson *et al*., 1986) genomes were constructed to enhance the molecular genetics of the respective organisms by facilitating the cloning of known genes and to serve as an archive for genomic information. However, the data associated with the construction of the clonal physical maps – even with good alignment to the genetic map – carried only a tiny proportion of information present within the genome. Consequently, a minimum tile path of the 30 kb cosmid and 15 kb lambda clones, used to build the physical maps of the *C. elegans* and *S. cerevisiae*, respectively, were subcloned into M13 phage vectors (1.3-2 kb insert size) and sequenced on a per clone basis. The physical maps of the two genomes, and subsequently of *Escherichia coli* (*E. coli*) (Kohara *et al*., 1987), *Arabidopsis thaliana* (*A. thaliana*) (Arabidopsis Genome Initiative, The, 2000), *Drosophila melanogaster* (*D. melanogaster*) (Hoskins *et al*., 2000) and human (McPherson *et al*., 2000), used restriction enzyme fragments in various ways to overlap clonal units for the construction of genome wide physical maps.

For the *C. elegans* project, polyacrylamide gel electrophoresis was used to resolve DNA fragments that had been generated by digesting cosmid DNA with two restriction enzymes, *Hin*d III and Sau 3AI. Restriction fragments, of which the *Hin*d III ends had been labelled with a radioactive molecule, were then detected by exposure to autoradiograph film. Digitised cosmid specific ‘fingerprints’ were analysed by pair-wise comparison to establish contigs of overlapping bacterial clones within a seven fold genomic fingerprint data set (Coulson *et al*., 1986). The mapping of *S. cerevisiae* used a similar contig construction strategy but alternatively, fingerprints were generated by a single restriction digest of lambda DNA and the clone fragments separated on an agarose

gel prior to reassembly into contigs (Olson *et al*., 1986). Whilst both methods generated large tracts of genomic contig coverage, gaps remained in the physical maps. To compensate for the regions lacking cloned representation in the *C. elegans* map, fosmids, maintained as single copy within the host cell (Kim *et al*., 1992), and large insert yeast artificial chromosome (YACs), (Burke *et al*., 1987), were incorporated. Fosmids proved to be most useful generating bridging coverage in central, gene rich regions of chromosomes, whilst YACs tended to generate *de novo* map coverage on the more repeat-dense chromosome arms (A Coulson *pers comm*.).

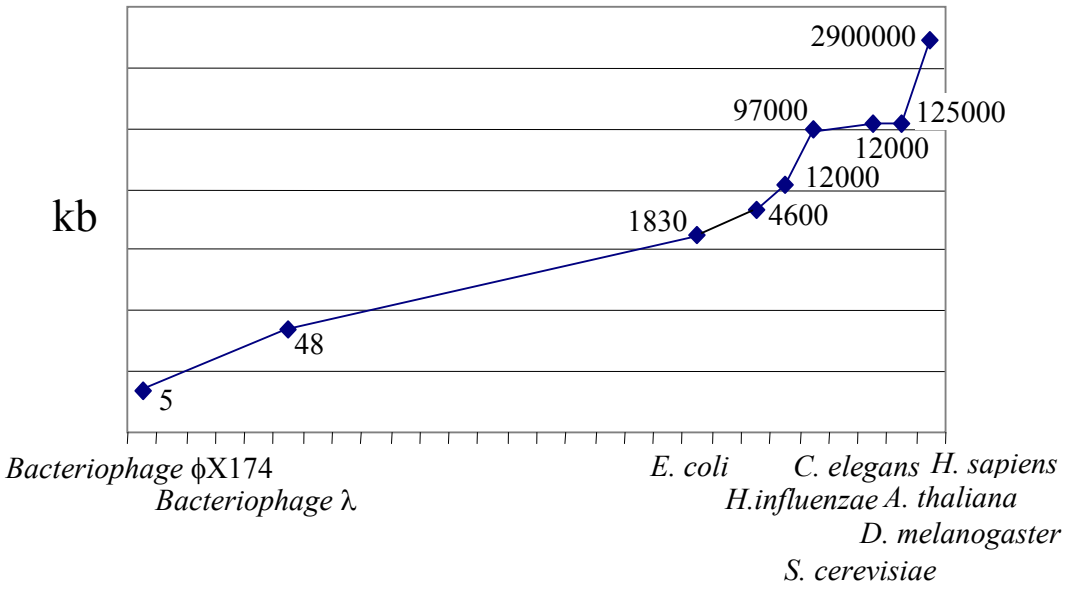


Figure 1.1 A plot of the near logarithmic increase in the complexity of genomic sequencing from the first full genomic sequence of Bacteriophage ϕX174 in 1977 to the anticipated completion in the human genome in April 2003.

1.2 Mapping and sequencing the human genome

The human genome is contained within 22 autosomes (numbered from 1 – 22, largely according to size) and two sex chromosomes, X and Y (female XX and male XY). Chromosomes are punctuated with centromere structures that are either located close to chromosome ends (acrocentric), towards a chromosome end (submetacentric) or centrally between ends (metacentric). The initial size estimate of the genome, 3200 Mb, was based largely upon cytometric measurements (Morton 1991) and has since been revised to 29000 Mb in light of the higher resolution human draft sequence analysis (IHGSC, 2001) supported by observed sizes of completed chromosome sequences, which suggested the earlier figures were over-estimates (Dunham *et al*., 1999, Hattori *et al*., 2000, Deloukas P *et al*., 2001). The construction of a map of the human genome was an important step towards understanding and characterising the sequence contained within it as it provided a means by which all the features could be ordered and partitioned, and the task of detailed characterisation and sequencing could be divided up into manageable segments.

1.2.1 Cytogenetic mapping

The treatment of metaphase chromosome spreads with trypsin digestion and Giemsa staining creates differential chromosome banding patterns. The generation of light (R-bands) and dark (G-bands) bands by Giemsa staining is reliant upon nucleotide content and the staining pattern therefore reflects the base composition and correlates other properties of the different regions (see table 1.1). However, the maximum genome-wide resolution was limited to an 850 genome-wide banding pattern (Bickmore *et al*., 1989). The recognition of characteristic banding patterns of chromosomal regions provided the

basis for much of the early characterisation of chromosome aberrations (duplications, deletions and translocations) which were associated with clinical phenotypes (Pinkel *et al*., 1988, Tkachuk *et al*., 1990, Dauwerse *et al*., 1990).

The ability to hybridise labelled probes containing specific sequences and to detect their location on metaphase chromosome by autoradiographic or fluorescent detection techniques (fluorescence *in situ* hybridisation (FISH) (Pinkel *et al*., 1986)) revolutionised cytogenetic mapping. Initially, the location of the probe relative to the metaphase banding pattern provided an approximate map position for the sequence represented by the probe. Pairs of markers, differentially labelled, could be simultaneously placed relative to the cytogenetic banding, and also ordered with respect to each other. The use of pairs of differentially labelled markers in combination with a third reference marker enabled FISH to be applied to chromosomal DNA in a less condensed state (in interphase nuclei). Although no banding pattern can be obtained in interphase DNA, the decondensed state of the chromatin relative to metaphase chromosomes means that increased levels of resolution could be obtained, as probes were better separated. An inter-probe distance of 1- 5 Mb can be resolved using metaphase FISH, 0.1 – 1.0 Mb by interphase FISH (Wilke *et al*., 1994), and 5 kb by FISH using mechanical pre-treatment to extend DNA into fibres (Heiskanen *et al*., 1994).

Table 1.1: Comparison of G-bands and R-bands

G-bands	R-bands
Dark staining Giemsa bands	Light Staining Giemsa bands
AT rich	GC rich
Late replicating	Early replicating
Early condensation	Late condensation
DNase insensitive	DNase sensitive
SINE poor, LINE rich	SINE rich, LINE poor
Gene poor	Gene rich
Less frequent recombination	More frequent recombination

Adapted from Bernardi (1989)

1.2.2 Genetic mapping

Genetics maps utilise the likelihood of recombination between adjacent markers during meiosis to calculate inter-marker genetic distances, and from this to infer a physical distance. The closer two landmarks are together on a chromosome, the less likelihood there is of a recombination event occurring between, with the opposite being true for markers that are further apart. The calculation of distance, and therefore the metric upon which the genetic map is based, is the length of the chromosomal segment that, on average, undergoes one exchange with a sister chromatid during meiosis, the Morgan (M). Therefore, a 1% recombination frequency is equivalent to 1 centimorgan (cM), and, since the human genome covers 3000 cM and contains approximately 30000 Mb, 1cM is approximately equivalent to 1 Mb. However, recombination is known to be non-random which can lead to a level of inaccuracy (Dib *et al*., 1996) in inferring physical distances from measurements of genetic recombination.

The inherent limitation of primary genetic maps was the lack of availability of polymorphic markers between which genetic distances could be calculated. This was ameliorated in part by the suggested use of restriction fragment polymorphisms (RFLPs), identified by Kan and Dozy (1978), for the construction of a genome wide genetic linkage map (Botstein *et al*., 1980).The first such map (Donis-Keller *et al*., 1987) was limited in its usefulness, however, due to RFLPs having a maximum heterozygosity of 50% and the low level of resolution of the 403 characterised polymorphic markers, including 393 RFLPs, covering the genome. The identification of hypervariable regions, which showed multi-allelic variation (Wyman and White 1980), provided a new source of markers for genetic mapping. The variable regions contained short 11 to 60 bp variable

number tandem repeats which showed allelic variation. However, these minisatellite markers (Jeffreys *et al*., 1985) and variable number tandem repeats (VNTRs) (Nakamura *et al*., 1987) were shown to cluster at chromosome arms and were not inherently stable (Royle *et al*., 1988). The identification of microsatellite markers (containing di-, tri- or tetra nucleotide repeats) greatly facilitated the generation of genetic maps. They were proven to be widely distributed throughout the genome, showed allelic variation (Litt and Luty 1989, Weber and May, 1989), were amenable to PCR amplification (Saiki *et al*., 1988) by sequence-tagged-site screening (Olson *et al*., 1989). In a relatively short period of time a number of genetic maps were published with increasing marker density and resolutions, culminating in the most recent deCODE genetic map which contains 5136 markers genotyped across 1257 meioses (Kong *et al*., 2002), table 1.2.

Table 1.2: A comparison of marker content within genetic maps.

No. of Markers	Reference
100	Hudson *et al*., 1992
813	Weissenbach *et al.,* 1992
2066	Gyapay *et al*., 1994
5840	Murray *et al*., 1994
5264	Dib *et al*., 1996
5136	Kong *et al*., 2002

1.2.3 Radiation hybrid mapping

The utilisation of somatic cell hybrids to maintain human genomic fragments, such as whole chromosomes or chromosomal regions, permits the generation of another form of mapping resource to be generated, the radiation hybrid map. The modification of a technique that fragmented human chromosomes by irradiation and which were then rescued by fusion to rodent cells (Goss and Harris 1975) prompted Cox *et al*., (1990) to

propose that radiation hybrid (RH) mapping could be applied to the construction of long range maps of mammalian chromosomes.

The premise of the technique is similar to that of the genetic map, i.e. the more closely related two markers are related within the genome the less likelihood there is of a radiation induced break in between them in a reference panel of cell lines, and hence the less likely is their segregation to different chromosomal locations based on association of the markers to different sets of fragments. As the presence of two markers within a radiation fragment gives no indication to their physical distance a panel of radiation hybrids was required. By estimating the frequency of breakage, and thus the distance between two markers, it is possible to determine their order. The unit of map distance is the centiRay (cR) and represents 1% probability of breakage between two markers for given a radiation dose. Unlike the level of information garnered from a genetic marker, that may or may not be informative within a varying number of meioses, the radiation hybrid marker is either positive or negative for a DNA fragment, effectively digitising PCR results. Any amplifiable single copy sequence can therefore be placed in a radiation hybrid map. The radiation hybrid mapping technique has been used for the construction of high resolution gene maps (Schuler *et al*., 1996, Deloukas *et al*., 1998) and has also been used to supplement the construction of chromosome physical maps (Mungall *et al*., 1996, Deloukas *et al*., 2001, chapter 4).

1.2.4 Physical mapping

The generation of a physical map relies upon the construction of an ordered and orientated set of clone based contigs. The term "contig" was coined by Staden (Staden 1980) to refer to a contiguous set of overlapping segments which together represent a consensus region. These segments can be sequence, or clones, whose overlapping relationship is defined by information in common to each pair of overlapping segments. The overlaps are identified by performing a pair wise comparison of the dataset associated with each segment. Similarities that are statistically significant indicate the presence, and sometimes the extent, of overlap. Bacterial clone contigs are the most convenient route for the sequence generation of larger genomes. They presenting a means of coordinating physical mapping and, because of the way in which they are constructed, provide an optimal set of clones (the tile path) for sequencing.

1.2.4.1 YAC Maps

The main benefit of using YACs for the constructing of a physical map is that the insert size (up to 2 Mb) results in coverage of large regions of the genome with relatively few clones. Green and Olson (1990) utilised YACs to construct a physical map across the cystic fibrosis region on human chromosome 7 by overlapping YACs by STS content data. Chromosome specific (Chumakov *et al*., 1992, Foote *et al*., 1992) and genome wide YAC maps have also been published (Chumakov *et al*., 1995, Hudson *et al*., 1995). Though STS content mapping is the most frequent method used to generate YAC contigs, techniques such as repeat mediated fingerprinting, either by *Alu*-PCR (Coffey *et al*., 1992) or by repeat content hybridisation (Cohen *et al*., l993), have also been used.

The advantages of using YACs are, however, offset by the relative difficulty of constructing YAC libraries and of analysing the cloned DNA, compared to the use of bacterial cloning systems. Many YAC clones have also been found to be chimeric, that is, to contain fragments derived from non-contiguous parts of genomic DNA being cloned (Green *et al*., 1991, Bates *et al*., 1992, Slim *et al*., 1993). Rather than being used as a primary sequence resource, YACs became more generally used to support the construction of detailed landmark maps, and to underpin sequence ready bacterial clone maps (Collins *et al*., 1995; Bouffard *et al*., 1997). Recently, YACs have been used to facilitate gap closure in the bacterial clone maps by linking contigs (Coulson *et al*., 1995). The links are identified by STS content mapping. In these cases the YACs have been sequenced directly.

1.2.4.2 Bacterial Clone Maps

In contrast to YACs, bacterial clone libraries are easier to make and the cloned DNA is more easily manipulated. Chimerism is low (Shizuya *et al*., 1992, Ioannou *et al.*, 1994), and the supercoiled recombinant DNA can be purified readily from the host DNA. An important factor influencing the construction of bacterial clone contigs is the available genomic resources. Whilst the *C. elegans* and *S. cerevisiae* maps utilised total genomic 30 kb cosmid and 15 kb lambda libraries, in 7- and 5-fold coverage respectively, current bacterial clone contig construction utilises large insert P1-derived artificial chromosome (PAC) (Iaonnou *et al.*, 1994) and bacterial artificial chromosome (BAC) (Shizuya *et al*., 1992) libraries. Each BAC or PAC clone typically contains an insert of 100 – 300 kb and maps have been constructed from a >15 fold genomic clone coverage.

1.3 Generating human genomic sequence

The elucidation of the all genic and other features contained within the human genome is reliant upon the generation of high quality sequence. Two different strategies were adopted to produce human genomic sequence; the first , utilised by the International Human Genome Sequencing Consortium (IHGSC), is a hierarchical approach utilising bacterial clones from a well characterised sequence ready map as the basis for generating genomic sequence; the second strategy is a whole genome shotgun (WGS) approach, adopted by Celera, which relies upon the assembly of a consensus sequence from randomly derived genomic sequence reads (Venter *et al*., 2001).

The IHGSC used an approach in which a sequence ready bacterial clone map was constructed by utilising high density panel of markers (15 markers / Mb) (Olson 1993, Bentley *et al*., 2000) derived from genetic and radiation hybrid maps. Large insert bacterial clones, primarily PACs (Ioannou *et al*., 1994) and BACs (Shizuya *et al*., 1992), were identified by hybridisation and overlapped by restriction digest fingerprinting (Gregory *et al*., 1997, Marra *et al*., 1997), with STS content data (Olson *et al*., 1989) supporting ambiguous overlaps. DNA generated from each one of a minimally overlapping set of clones from the physical map (tile path) was fragmented into 1.4 - 2.2 kb units, subcloned into M13 or plasmid vectors (Bankier *et al*., 1987) and then sequenced by modified chain termination sequencing (see chapter 5). One of the benefits of generating sequence on a clone by clone basis is that if the consensus sequence generated by the shotgun phase is not contiguous, the bacterial clone from which the sequence was derived can be then used for directed finishing experiments. The production of finished sequence, which is >99.99% accurate and without the presence of

gaps, is one of the main differences between the WGS and hierarchical approaches. Whilst the private WGS strategy was declared to be complete following the assembly of the shotgun sequence, The public hierarchical approach, provided an initial draft covering 90% of the genome in 2000; and its production of highly accurate sequence, which will serve as a long-term reference, is now 87% done and due to be completed in April 2003. The results of the private WGS strategy were also announced in 2000 following the assembly of the shotgun sequence, incorporating a representative set of sequences from the public domain draft; however no subsequent finishing was undertaken on this product.

The elucidation of complete genomic sequence by the generation of whole genome shotgun data, as used by Celera, is a relatively simple approach that was first used for the characterisation of the mitochondrial genome (Anderson *et al*., 1981). In 2001, Venter *et al.,* (2001) published their applied strategy to the elucidation of the human genome by assembling sequence data from complete inserts or ends of 1 – 2 kb and 50 kb subcloned plasmids, respectively, and by the incorporation of BAC end sequences. The publication of the human genome sequence in, on average, 100kb scaffolds proved that a considerable amount of human sequence could be generated and assembled in a relatively short period of time. However, some doubts remain as to the success of the approach as a sole means of assembling a complex genome (Waterston *et al*., 2002, Myers *et al*., 2002, Green *et al*., 2002). Without the incorporation of the publicly available sequence data, and the inherent map information associated with it, highly repetitive motifs and low copy duplications introduce errors into the assembly of the sequence.

1.4 Interpreting the human genome landscape

The complete characterisation of the human genome will not be restricted to experimental or *in silico* identification of coding features and the elements affecting their expression. A full description of the long range sequence composition will be necessary to, amongst other things, facilitate an understanding of the coordinated regulation of genes, identify the under lying reasons for repeat mediated genomic rearrangements and to provide a basis for the identification of variation between populations.

1.4.1 Sequence composition

The human genome has previously been shown to contain considerable variation in its nucleotide composition. Separation of mammalian genomic DNA on caesium gradients indicated compositional heterogeneity (Thiery *et al*., 1976) whilst fractionation of human DNA on Cs_2SO_4-Ag^+ gradients showed a broad range of GC based separation (Bernardi *et al*., 1985). Regional partitioning of genomic sequence based on GC content can be used as a low resolution means of determining biological properties such as cytogenetic banding (Hurst and Eyre-Walker 2000), repeat composition and gene density (Zoubak *et al*., 1996, Gardiner *et al*., 1996) and structure (Oliver *et al*., 1996). Analysis of the human draft sequence which, at the time of publication covered an estimated 94% of the genome, revealed GC content could vary by as much as 59.3% to 33%, from the genome average of 41%, within a 300 kb region (IHGSC, 2001). It was also estimated that 98% of large insert clones mapping to the darkest Giemsa staining bands contained below average GC content, whilst 80% of the clones mapping to light bands contained higher than average GC content.

The fractionation of human DNA on Cs_2SO_4-Ag^+ led Bernardi *et al.,* (1985) to suggest that DNA fractions could be classified based on their relatively homogeneous GC content (isochores). GC-poor isochore family members, L1 and L2, contained <38% and 38-42%, GC respectively, whilst heavy (GC rich) family members contain 42-47%, 47-52% and >52%, GC respectively. However, the debate continues as to whether isochores provide a useful means of partitioning the sequence contained within the human genome (IHGSC, 2001, Oliver http://genomebiology.com/)

1.4.2 CpG island identification

CpG islands are short stretches of hypomethylated DNA that are rich in GC nucleotides and have a CpG:GpC ratio approaching or exceeding 1:1. Their occurrence within human genomic sequence is one fifth of the expected 4% frequency, by multiplying the relative fractions of G's and C's within the genome (0.21 X 0.21)) (IHGSC, 2001) due to the spontaneous deamination of the methyl-C residue which give rise to a thymine (Coulondre *et al*., 1978). By contrast, deamination of non-methylated cytosine produces uracil residues which are recognised and repaired as cytosine by the cell. The identification of CpG islands within human sequence is significant because they are often associated with the 5' ends of genes (Bird *et al*., 1986, Gardiner-Garden and Frommer 1987). It has been estimated that 56% of human genes are associated with CpG islands (Antequera and Bird 1993), including constitutively expressed genes, and approximately 40% of genes with tissue specific patterns of expression (Larsen *et al*., 1992). Analysis of the draft human sequence identified 28,890 putative CpG islands (putative as no information on the methylation status was obtained) within sequence from which repeats had been removed by RepeatMasker (Smit and Green, unpublished) (IHGSC, 2001).

Whilst variation in the length (up to 36,619 bp), and density (2.9 - 22 CpG islands / Mb) of putative CpG islands was observed, their estimated number closely matched previous estimates (Antequera and Bird 1993).

1.4.3 Repeat content

Analysis of the human draft sequence revealed that repetitive motifs account for at least 50% of the genomic content (IHGSC, 2001). These repeat motifs can be roughly divided into five different classes,

1) transposon-derived (interspersed) repeats
2) inactive retroposed copies of genes (processed pseudogenes)
3) simple sequence repeats (e.g. $(A)_n$, $(CA)_n$, $(CGG)_n$)
4) segmental duplications (10 – 300 kb duplicated within the genome)
5) tandemly repeated sequences (centromeres / telomeres / RNA gene clusters).

The transposon-derived interspersed repeats account for more than 90% of all repeats currently identified (IHGSC, 2001). This class of repeats, which includes short-interspersed-elements (SINES), long interspersed elements (LINES), LTR retrotransposons and DNA transposons, accounts for 13%, 20%, 8% and 3% of the draft sequence, respectively. In general the SINE and LINE elements, as has previously been reported (Soriano *et al*., 1983), show an inversely proportional distribution by genomic GC content. LINE elements are found at a higher density in AT-rich, GC-poor, regions whilst SINEs are found in AT-poor, GC-rich regions. The prevalence of LINE elements in AT-rich regions, where gene density is lower, is logical from the perspective that these genomic parasites would not present a mutational burden to their host. Whereas SINEs,

that utilise LINE transposon machinery for replication (Jurka *et al*., 1997), do not co-localise with LINEs and are in fact found in GC-rich regions at a four fold higher density than in GC-poor region. It has been proposed that SINEs may somehow target GC-rich regions for insertion or that they are co-distributed similarly to LINEs, in gene-poor regions, but their distribution is subsequently reshaped by evolutionary forces (IHGSC, 2001) or that they are randomly integrated but are then fixed preferentially in GC-rich DNA (Smit *et al*., 1999).

1.5 Gene identification

It is estimated that 5% of the human genome contains coding sequence (IHGSC, 2001), with as little 1 – 2% encoding for protein (Green *et al*., 2001). It is therefore important that the structures of genes are clearly and accurately defined above the background of apparent non-coding and repetitive sequence. It is known that human genes encode RNAs that, for example, facilitate the expression of protein-coding genes by the excision of introns by the spliceosome (small nuclear ribonucleoproteins -snRNA), participate in translational machinery (ribosomal RNA - rRNA and transfer RNA - tRNA), or act as the template for the transcription of messenger RNA (mRNA) which is subsequently translated into a protein product. The majority of the protein-coding genes will most likely be identified within human genomic sequence by *in silico* gene prediction with the support of cDNA and EST sequence alignment. However, techniques that were developed to identify genes from the transcript rather than the sequence will be central to the elucidation of human genes that are difficult to characterise, and to assist

identification of genes within other organisms for which large amounts of sequence data will not be available.

Prior to the availability of large tracts of human genomic sequence, the identification of protein-coding genes was largely reliant upon the isolation of a transcript from tissue specific mRNA or cDNA clones. These early studies used mRNA enrichment coupled with biological assays of *in vitro* translated products, or cDNAs library hybridisation, using plasmids containing coding fragments, or oligonucleotide mixtures based on peptide sequence, to obtain clones for sequencing to characterise the genes of interest. This highly targeted approach has successfully identified a number of genes, including the rabbit (Rabbitts *et al*., 1976) and human α- β- γ- globins (Little *et al*., 1978), human interferon, IFN (Taniguchi *et al*., 1980) and factor IX (Choo *et al*., 1982).

Genomic fragments, in the form of cloned DNA (cosmids and YACs), have also been used to identify cDNA clones by direct screening of cDNA libraries. This technique proved successful for the identification of the neurofibromatosis type 1 (Wallace *et al*., 1990) and Menkes genes (Chelly *et al*., 1993). Whilst the identification of candidate genes by genomic fragment hybridisation is successful, it relies upon the correct hybridisation kinetics between the exons in the genomic DNA and transcribed sequence. Therefore, genomic hybridisation may not be the most productive method of identifying the gene of interest where the gene or exons may be very short in the genomic sequence.

Direct selection, which can be used to enrich cDNA libraries for genes encoded by large genomic regions, can result in a 1000-fold amplification of target cDNAs. The technique has successfully been applied to the characterisation of genes within a 300 kb region around the G6PD on Xq28 (Sedlacek *et al*., 1993) and a 6.5 Mb region on Xq21, which

led to the identification of the X-linked agammaglobulinemia gene (XLA) (Vetrie *et al*., 1993).

Another technique of gene discovery, which utilises the genomic sequence of a gene is exon trapping. Though the technique has successfully been used in several positional cloning projects (Walker *et al*., 1993, Trofatter *et al*., 1993, The Huntington's Disease Collaborative Research Group 1993) its technical complexity precludes it from large-scale application.

The technique of amplifying multiple genes by oligo dT priming (Verma *et al.,* 1972, Wickens *et al*., 1978) or with anchored oligo(dT) priming (Khan *et al*., 1991), together with advances in sequencing technologies, led to the development of a strategy to generate single pass sequence of all human cDNAs. It was suggested that this approach would obviate the need to map and sequence the entire genome (Brenner *et al*., 1990). Large-scale sequencing projects were initiated to generate gene fragments by single pass sequencing from the 5' and 3' ends of cDNA clones (Adams *et al*., 1991, Okubo *et al*., 1992, Adams *et al*., 1993, Sudo *et al*., 1994, Hillier *et al*., 1996). One of the difficulties encountered with this strategy was the representation of genes by multiple expressed sequence tags (ESTs). Consequently, a database (UniGene) was established that rationalised EST data by clustering the sequences into a non-redundant data set (Boguski and Schuler 1995, Schuler *et al*., 1996) (http://www.ncbi.nlm.nih.gov/UniGene/clust.cgi?ORG=Hs). UniGene currently contains (31st October 2002) 3,911,348 EST sequences which have been condensed into 110743 clusters. Almost half of the clusters, 50476, contain a single EST entry, whilst the average cluster contains 5 – 8 ESTs, with the largest cluster containing 16385-32768

ESTs. Another problem encountered with the generation of EST sequence was the contamination of the cDNA libraries with genomic DNA. Poly(A) sequences within the genomic DNA act as templates for the oligo(dT) primer and false cDNA sequence is generated. Additionally, some transcripts have proven to be incomplete, for example, large transcripts that are sometimes not represented if reverse transcription of the mRNA has terminated prematurely. A further problem with the cDNA based approach is that some transcripts may be absent if the gene is not expressed in the tissues used to construct any of the cDNA libraries used for EST sequencing.

Whilst the integration of ESTs within existing radiation hybrid and genetic maps has facilitated the identification of putative candidates for a number of diseases by positional cloning (APOE; Pericak-Vance *et al*., 1991, Strittmatter *et al*., 1993, RET; Mulligan *et al*., 1993) it is their localisation within the human genome sequence that ESTs will prove their greatest worth. Alignment of ESTs to genomic sequence have facilitated the correct annotation of gene structure by assisting to define exon – intron boundaries and helped to identify splice variants, which are present in at least 35% of all human genes (IHGSC, 2001).

Detailed sequence analysis of previously characterised genes has permitted conclusions to be drawn their about generalised genomic structure. Regulatory elements, such as those residing at the 5' ends of genes, which act as a template upon which transcription factors assemble prior to the initiation of RNA synthesis, have been inherently difficult to identify. This is due to the absence of consistently shared motifs within the genomic sequence. However, the combined occurrence of specific elements, CpG islands and TATA boxes flanked by regions of C-G, permits *in silico* prediction of at least a fraction

of the core promoter region to be made. Whilst Eponine (Down and Hubbard 2002) and PromoterInspector (Scherf *et al*., 2000) currently have approximately 50% sensitivity (detecting known TSSs) and 75% specificity (predictions supported by known TSSs) their accuracy is likely to increase when trained upon larger data sets of elucidated promoters.

Unlike prokaryotic organisms, in which genes are located as a single tract of DNA, complex organisms, from yeast onwards, contain genes that are usually segmented by introns of non-coding sequence (Tilghman *et al*., 1978; Gilbert *et al*., 1978). The spliceosome recognises sequence motifs within the intron which leads to their excision, usually a GT di-nucleotide at the 5' end of an intron (splice donor), an AG di-nucleotide at the 3' end (splice acceptor) and an internal branch point (Moor and Sharp 1993).

Localising the site of translation initiation within genomic sequence, unlike the TSS, has been facilitated by the identification of a consensus motif within the genomic sequence. The translation start site, usually an ATG, is typically found in a consensus, GCC$^{A}/_{G}$CC**ATG**G, which includes the two bases which exert the strongest effect, a G at the first base after the translation start, ATG, and a purine (preferably an A) which is located three nucleotides upstream (Kozak *et al*., 1987). The identification of stop codons (TGA, TAA or TAG) and polyadenylation signals (Kessler *et al*., 1986), most commonly as AATAAA (Beaudoing *et al*., 2000), have helped define the 3' ends of genes within genomic sequence.

1.6 Computational Genomics

1.6.1 *In silico* gene prediction

The *in silico* prediction of genes within genomic sequence utilises characteristic sequence motifs associated with genes mentioned above. Whilst initial computer programs were developed to identify single exons, for example GRAIL (Uberbacher *et al*., 1996), and HEXON (Solovyev *et al*., 1994), more recently developed programs attempt to identify complete gene structures. These programs, GENSCAN (Burge *et al*., 1997) and FGENESH (Solovyev *et al*., 1995) predict individual exons based on codon usage and sequence signals and assemble these putative exons into candidate gene structures. The greatest problem associated with using *in silico* gene prediction programs to identify coding structures within genomic sequence is the number of over predictions that are generated. Whilst all genes may be identified by setting low prediction thresholds, the gene structures will be generated with low specificity. Guigo *et al* (Guigo *et al*., 2000) assayed the accuracy of gene prediction methods using an artificially generated data set. They found that GENSCAN accurately predicted 90% of coding nucleotides with 70% of the exons being predicted correctly. However, there was also significant over prediction of 30% based upon predictions in simulated intergenic sequences. The study concluded that it is not currently viable to use computational methods alone to accurately identify the exonic structure of every gene in the human genome (Guigo *et al*., 2000).

The estimated total number of genes contained within the human genome has varied considerably according to the technique that was used to evaluate it and the data that was available at the time. Estimates have been variously based upon average genome and

gene size, the number of observed CpG islands (and the proportion associated with genes), redundant and non-redundant EST sequences and, latterly, upon chromosome specific totals, comparative analyse and genome-wide sequence analysis (table 1.3).

Table 1.3: Genes in the human genome

Data set	Gene Number	Date	Reference or source
Hypothetical	100,000	1992	Gilbert *et al*., 1992
CpG Islands	80,000	1993	Antequera *et al*., 1993
EST clusters	60,000-70,000	1994	Fields *et al*., 1994
Unigene clusters	92,000	1996	Schuler *et al*., 1996
Gene sequences	140,000	1999	*IncyteGenomics
Chr. 22 sequence	43,000-61,000	1999	Dunham *et al*., 1999
Chrs 22, 21 sequence	44,000	2000	Hattori *et al*., 2000
Tetraodon seq.	28,000-34,000	2000	Roest-Crollius *et al*., 2000
ESTs in dbEST	120,000	2000	Liang *et al*., 2000
EST and mRNA	35,000	2000	Ewing, B*., et al.*, 2000
Draft Sequence	31,000	2001	IHGSC, 2001

*press release available at http://incyte.com/company/news/1999/genes.shtml

1.6.2 Sequence Analysis

Whilst computer programs have been written to predict a number of features associated with coding sequences (table 1.4) the alignment of experimental data is critical to the validation of predicted structures.

Table 1.4: Prediction programs used to identify gene features

Program	Description	Reference
RepeatMasker	Repeat Sequence prediction	Smit and Green, unpublished
CPGFIND	CpG island prediction	Micklem, unpublished
PromoterInspector	Promoter Prediction	Scherf *et al*., 2000
Eponine	TSS prediction	Down and Hubbard, 2002
Hexon	Exon prediction	Solovyev *et al*., 1994
Grail	Exon prediction	Uberbacher *et al*., 1991
GENSCAN	Gene prediction	Burge *et al*., 1997
FGENESH	Gene prediction	Solovyev *et al*., 1995

Programs such as CLUSTALW (Thompson *et al*., 1994), which is used to align multiple nucleotide or protein sequences, and DOTTER, which utilises pair-wise local sequence alignment strategy (Sonnhammer *et al*., 1995), are useful for inferring structural and functional conservation by sequence homology. Sequence homology searching can also be performed using SSAHA (Ning *et al*., 2001), Exonerate (Slater, unpublished) and BLAT (Kent *et al*., 2002) which permits homology sequences to be identified within gigabases of DNA. BLAST (Basic Local Alignment Search Tool), which measures the local similarity between two sequences (Altschul *et al*., 1990, 1997), is the primary method for identifying protein and DNA sequence similarities prior to incorporation of the features into project specific ACeDB databases (Durbin and Thierry-Meig 1994) (http://www.acedb.org/) or genome browsers. One of the major advantages of using BLAST for sequence alignment is the flexibility with which nucleotide and amino acid sequences can be aligned (table 1.5)

Table 1.5: Sequence queries available using BLAST alignment

Program	Query	Database	Comparison
blastn	DNA	DNA	DNA level
blastp	Protein	Protein	Protein level
blastx	DNA	Protein	Protein level
tblastn	Protein	DNA	Protein level
tblastx	DNA	DNA	Protein level

From Brenner (1998)

Whilst BLAST alignment of human mRNA, EST and protein sequences to predicted coding structures provides a primary level of support, predicted features can also be supported by alignments with sequence from other organisms for comparison (comparative sequence analysis). The identification of sequences that are conserved between species is important because sequences that contain elements that are potentially

functional are more likely to retain their sequence than non-functional segments, under the constraints of natural selection during evolution. The evolutionary distance between species is an important consideration. Sequence comparisons between closely related species may facilitate the identification of gene structures and regulatory elements but, if the evolutionary distance between the species is relatively small, these sequences may be obscured by non-functional sequence conservation. Therefore a variety of species, including more distantly related species may be required to identify potential functional sequences using the comparative approach.

The identification of conserved sequences by comparative analysis has focused on the identification of non-coding regions (Hardison *et al*., 1993; Koop *et al*., 1994; *et al*., 1997; Hardison *et al*., 1997) and protein coding regions (Makalowski *et al*., 1996; Ansari-Lari *et al*., 1998; Jang *et al*., 1999) between human and mouse genomes. The alignment of sequence from multiple organisms has also been used to identify upstream regions that may affect gene expression. Gottgens *et al* (Gottgens *et al*., 2000) used the alignment of human, mouse and chicken sequences to identify a novel neural enhancer element in their elucidation of the human stem cell leukaemia (SCL) gene region. Whilst comparative sequence analysis may not identify all control regions associated with a gene, conserved regions may be identified that would be candidates for further experimental investigation (Pennacchio *et al*., 2001). Large scale sequencing and comparative analyses is progressing on a number of different organisms (table 1.6).

Table 1.6: A list of the large scale comparative organisms sequencing projects

Organism	**Genome Size (Mb)**	**Reference**
E. coli	4.6	Blattner *et al.*, 1997
S cerevisiae	12	Goffeau *et al.*, 1996
C. elegans	97	The C. elegans sequencing consortium, 1998
D. melanogaster	120	Adams *et al.*, 2000
M. musculus	2600	The Mouse Genome Sequencing Consortium, 2002*
D. rerio	1600	on going
F. rubripes	400	on going
R. rattus	2800	on going
T. nigroviridis	350	on going

* draft sequence publication

The availability of large tracts of human genomic sequence has necessitated the development of databases (genome browsers) that provide a framework upon which the enormous amount of data associated with the human genome can be stored and displayed. The two main databases, Ensembl, developed at the Sanger Institute and the European Bioinformatics Institute (http://www.ensembl.org/Homo_sapiens/), and the University of California Santa Cruz (UCSC) genome browser. (http://genome.cse.ucsc.edu/) contain information pertaining to physical maps, chromosome specific sequence assemblies, aligned mRNAs and ESTs, cross species homologies, SNPs and repeat elements. The development of generic genome browsers, as such as those hosted by Ensembl, makes possible the rapid identification of homologous sequences between comparative organisms and in doing so assist to identify conserved features that may be of some functional significance.

1.7 Allelic variation

Most differences between individuals, at the nucleotide level, can be attributed to allelic sequence variation. The characterisation of sequence differences and comprehension of how these genomic variations affect the expression and function of genes will be crucial for the study of molecular alterations in human disease. Whilst sequence variation has previously been used for genome-wide linkage and positional cloning studies (leading to the identification of many disease causing genes (see 1.2.2)), the association of single nucleotide polymorphisms (SNPs) with genes, either by mapping or as causal sequence variants, promises to be a valuable method in the future for identifying genes involved in complex diseases.

Approximately 90% of the allelic differences existing within the human genome can be attributed to SNPs, the remainder being insertions or deletions (Collins 1998b). A comparison of any two diploid genomes is estimated to identify one SNP per 1.3 kb which has an allele frequency of > 1% (ISNPMWG, 2001). The prevalence of SNPs in the genome, their existence as bi-allelic variants and their stability through inheritance makes them amenable to large-scale high through-put analyses. SNPs will, therefore, be applied to several research areas, including 1) large-scale genome analysis of linkage disequilibrium and haplotype patterns, 2) genetic analysis of simple and complex disease states, and 3) genetics and diversity of human populations.

1.7.1 SNP discovery

De novo candidate SNPs were initially identified by the alignment of STSs and ESTs to available genomic sequence (Wang *et al*., 1998, Picoult-Newberg *et al*., 1999, Irizarry *et al*., 2000, Deutsch *et al*., 2001). The clone based strategy used by the Human Genome Project for the large-scale production of human genomic sequence contributed to a dramatic increase in the SNP numbers by allowing identification of novel SNPs within sequence overlaps between minimum tile path clones (Taillon-Miller *et al*., 1998, Dawson *et al*., 2001). A directed approach to SNP discovery was initiated by sequencing DNA from population specific individuals (Mullikin *et al*., 2000, Altshuler *et al*., 2000). Two to five fold redundant shotgun sequence coverage was generated from 1.5 kb small insert library clones and the resultant sequences were aligned to each other in clusters. As the Human Genome Project progressed, these assemblies and additional shotgun sequence data were aligned to available genomic sequence to identify more SNPs. The total number of SNPs identified using the strategies outlined above culminated in the International SNP Map Working Group (ISNPMWG) constructing a SNP map of the human genome which contained 1.42 million candidate SNPs (ISNPMWG 2001). A proportion of the candidate SNPs were validated experimentally during the project, confirming that >90% were real SNPs. The SNPs identified by The SNP Consortium (TSC) (http://snp.cshl.org, Marshall *et al*., 1999) and the HGP had generated a SNP density of 1 SNP per ~1.9 kb of available sequence.

Akin to the rationalisation that was required to establish a unique set of ESTs, a database was established to generate a non-redundant collection of candidate SNPs (dbSNP) (Sherry *et al*., 2001, http://www.ncbi.nlm.nih.gov/SNP/index.html). dbSNP currently

contains 4.8 million entries which have been condensed into a non-redundant set of 3.0 million SNPs, 522,072 of which have been validated to date (build 110, 13th January 2003). Localisation of these unique SNPs within a recent the human sequence assembly (build 30) yields a SNP density of 1 per 1.2 kb. Additional validation of SNPs was carried out on a subset of TSC and HGP candidate SNPs by Marth *et al.*, (2001) by screening 1200 SNPs across 30 individuals from 3 difference populations. Results indicated that 80% of the SNPs were polymorphic in the populations tested, and that 50% had allele frequencies of greater than 20%. Data generated by Kruglyak and Nickerson (Kruglyak and Nickerson 2001) suggests that the number of non-redundant SNPs currently present within dbSNP, in which the minor allele is present in > 1%, comprise 11 - 12% of all single nucleotide sequence variants. The identification of >95% of available SNPs will require analysis of 96 haploid genomes, many more than the number from which the candidate SNPs have been derived so far.

There are many different platforms that have been developed for SNP analysis but which are based upon four basic allele-specific assays types, 1) hybridisation with allele-specific probes, 2) oligonucleotide ligation, 3) single nucleotide primer extension and 4) enzymatic cleavage. Many of the techniques have been developed further and automated in commercial systems. The range of formats used include colorimetric microtitre-plate based assays (Taqman by Applied BioSystems or Invader assay by Third Wave Technologies) or fluorometric methods of detecting SNP alleles that have been separated by gel electrophoresis (Applied Biosystems), fluorometric assay of target hybridised to oligonucleotides immobilised in a microarray chip format (Affymetrix), or immobilised via beads on the ends of arrays light sensing glass fibres (Illumina).

1.7.2 Utilising SNPs

SNPs may be utilised for population genetic studies in order to identify an association between a SNP allele and a specific phenotype. Ultimately the goal of such a study is to identify the causal variant, the mechanism by which the variant has its functional effect. The functional variant will have maximal predictive value in future individual tests, and the gene involved may encode a target protein or mRNA for possible therapeutic intervention. Functional variants may be assayed for using two approaches. The direct approach requires prior availability of a candidate functional variant (e.g. a SNP which alters the encoded protein sequence in a non-conservative way, thus affecting function). The variant is then tested by genotyping a population of defined phenotype and comparing the frequency of one allele with the frequency in a population of matched controls (a case-control study). In the absence of a candidate functional variant, the indirect approach can be taken, in which available SNPs within specific genes (candidate gene association studies) or throughout the genome (genome-wide association studies) can be used to test the same populations.

An aid to the indirect approach is to the identification of allele specific sequence variants and generation of a map of common combinations of specific alleles (or haplotype patterns) that have been largely conserved during the recent population expansions. Among other factors, it is believed that the regions of conserved local haplotype patterns have been maintained by the absence of ancestral recombination within each region. Identification of these conserved segments is facilitated by the availability of SNPs identified by the ISNPMWG. Pairs of alleles can be statistically quantified to determine whether recombination has occurred between them, in which case they are said to be in

equilibrium, or if the alleles share evolutionary co-segregation and are therefore in linkage disequilibrium (LD). The generation of an LD map does not require the analysis of related individuals (by comparison to the genetic map), only that they share a common evolutionary history (although inclusion of pedigrees allows direct determination of the phase between SNP alleles and facilitates definition of long range haplotypes). It is hoped that the generation of a map of common haplotype patterns (HapMap) will facilitate the identification of common diseases by indirect association studies as described above (Couzin *et al*., 2002, Harris *et al*., 2002).

The availability of genome sequence with annotated gene structures provides the means to search for candidate functional variants. Build 110 in dbSNP (13th January 2003) contains 60541 SNPs that have been localised to exons, untranslated regions or non-coding regions adjacent to genes (introns or flanking sequence), table 1.7.

Table 1.7: SNP totals contained within or adjacent to coding features.

SNP Count	FUNCTIONAL CLASSIFICATION
20851	Gene region
5228	Synonymous
5220	Non-synonymous
13462	untranslated region
15651	Intron
129	Splice site

SNPS localising within coding a feature (cSNPs) have the greatest potential to affect the structure and function of the gene. The characterisation of allelic variants enables conclusions to be drawn as to whether a specific allele may have an effect upon the amino acid sequence. Slightly more than half of the SNPs localising to coding sequences result in a synonymous change (no change in the amino acid sequence because of codon redundancy) whilst the remaining SNPs result in a non-synonymous change. Non-

synonymous changes are further classified as to whether the resultant amino acid has similar biological properties to the 'normal' allele, in which case the change is conservative, or if the biological properties of the amino acid are different, then the change is non-conservative. Whilst the molecular significance cSNPs have upon protein structure and function has previously been reported (Chasman *et al*., 2001, Sunyaev *et al*., 2001, Wang and Moult 2001), the effects that SNPs have in non-coding sequence such as splice junctions (Pan *et al*., 2002, Khan *et al*., 2002), folding of mRNAs (Shen *et al*., 1999) and promoter function (Knight *et al*., 1999, Hijikata *et al*., 2000) have also been described.

The identification of SNPs that show an allelic influence on the functioning of proteins, particularly of drug metabolising enzymes, promises a bright future for the optimisation of clinical therapeutics. Associating inherited variations with pharmacological responsiveness provides a basis for the possible development of personalised medicine which will improve the efficacy of drug treatments and decrease the side effects experienced by the individual (Roses *et al*., 2000, 2002, Pfost *et al*., 2000).

1.8 Chromosome 1

Chromosome 1, the largest human chromosome, is estimated to be 263 Mb in length (Morton 1991), which represents 8% of a 3200 Mb human genome. The chromosome is submetacentric and contains a large block of heterochromatin adjacent to the centromere on the long arm which, together with tandemly repeat sequences contained within the centromere and telomeres, reduces the euchromatic size of the chromosome to 214 Mb

(IHGSC, 2001). The chromosome has an average GC content of 43%, compared to the genome average of 41%, with the 40 Mb telomeric region of the short arm containing 47.1% GC (IHGSC, 2001). Draft sequence analysis indicates that chromosome 1 contains slightly more (~11 / Mb) than the genome average of 10.5 CpG islands / Mb, whilst the estimated gene content is noticeably higher (~15 genes / Mb) when compared to the genome average of ~ 11.5 genes / Mb (IHGSC, 2001).

A hierarchical strategy was used to map and sequence chromosome 1, as outlined in section 1.3. To date (15th of February 2003), the sequence ready map of chromosome 1 is contained within 9 bacterial clone contigs from which 2244 minimum tile path clones have been selected for sequencing (including 5 cosmids and 4 YACs). Currently, 95% of the chromosome is contained within finished sequence clones. The chromosome 1 mapping and sequencing project has directly facilitated the elucidation of a number of genetic disease genes, table 1.8.

Table 1.8: Genes that are associated with disease that have been elucidated as a result of the Sanger Institute chromosome 1 mapping and sequencing project.

Gene	Disease	Publication
CACP	Camptodactyly-arthropathy-coxa vara-pericarditis	Marcelino *et al*., 1999
SLC19A2	Thiamine-responsive megaloblastic anaemia	Labay *et al*., 1999
HPC1	Prostate cancer	Carpten *et al*., 2002
LMNA	Partial lipodystrophy	Shackleton *et al*., 2000
CIAS1	Muckle-Wells syndrome	Hoffman *et al*., 2001
IRF6	Van der Woude syndrome	Kondo *et al*., 2002
TBCE	HRD/Autosomal recessive Kenny–Caffey syndrome	Parvari *et al et al*., 2002

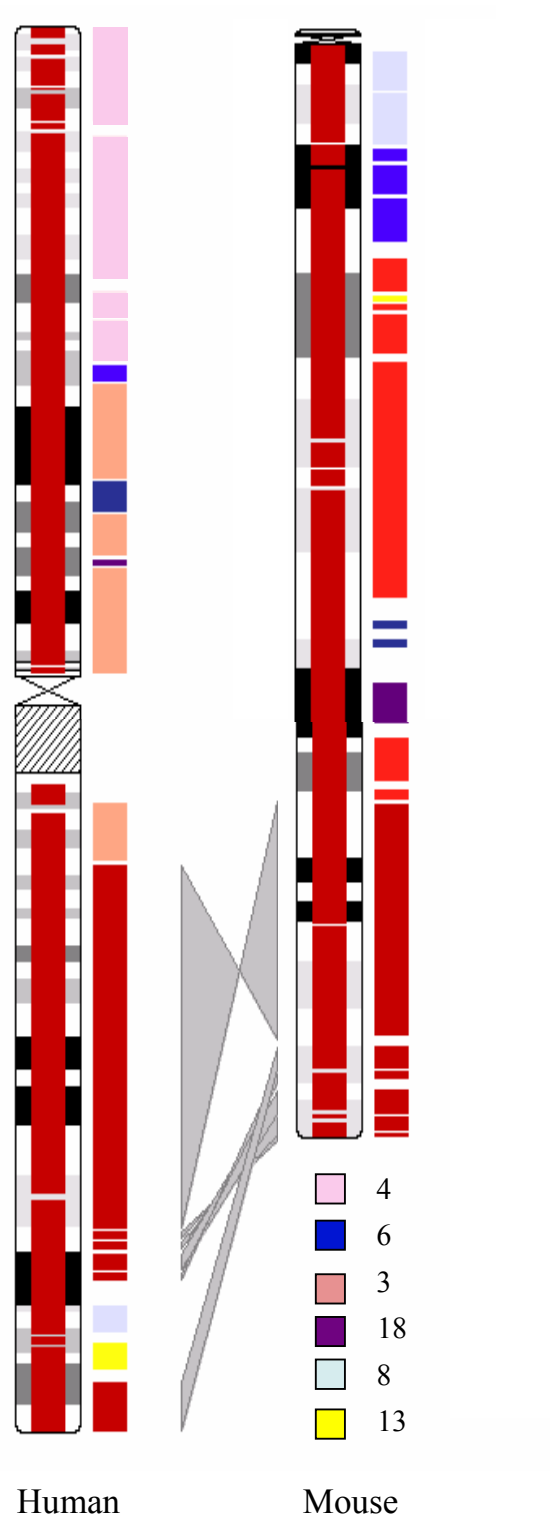
Included among the 157 genetic diseases localised to chromosome 1 within OMIM (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM) (16th of January 2003) are: Alzheimer disease-4 (Levy-Lahad *et al*., 1995), Chediak-Higashi syndrome (Kritzler *et*

al., 1964), Fanconi anemia (Loetscher *et al*., 1987), Gaucher disease (Hsia *et al*., 1959) and Usher syndrome, type 2 (Kimberling *et al*., 1990)

Alterations of chromosome 1 are amongst the most common chromosomal abnormalities in human neoplasia. These abnormalities include translocations and other structural rearrangements involving chromosome 1 and other regions of the genome, as well as region specific deletions and amplifications. In general, regions on 1p (the short arm) seem to be most frequently lost, indicating the possible presence of tumour suppressor genes, whereas regions on 1q (the long arm) are more often amplified. Loss of 1p has been detected in meningiomas and oligodendrogliomas (Bello *et al*., 2000, 2002), endometrial hyperplasia (Kiechle *et al*., 2000) and primary gastrointestinal tumors (El-Rifai *et al*., 2000). Gains of 1q have previously been implicated in neoplasia such as, hepatoblastoma (Nagata *et al*., 1999), lymphoma (Rao *et al*., 1999) and sarcoma (Forus *et al*., 1995).

The generation of continuous stretches of genomic sequence has permitted higher resolution synteny maps to be drawn than those based on mapping of orthologous genes or genetic markers. The construction of a physical map (Gregory *et al*., 2002) and the draft sequence (MGSC 2002) has enabled accurate localisation of syntenic boundaries between mouse and human genomes. The availability of more genomic sequence from a broader range of organisms will improve upon our understanding of chromosomal evolution. Figure 1.2 is a representation of the alignment of human and mouse chromosome 1 (also shown are the mouse syntenic blocks contained within human chromosome 1).

A number of physical maps have previously been constructed on chromosome 1. The clones used in their construction have evolved with the availability of new cloning systems, i.e. earlier physical maps were constructed using YACs and cosmids whilst later maps have utilised PACs and BACs. Comprehensive lists of chromosome 1 physical maps have been published within chromosome 1 workshop reviews (Gregory *et al*., 1998, White *et al*., 1999, Schutte *et al*., 2001). Though constructed primarily for the elucidation of disease genes maps have also been generated in difficult to clone regions, such as the telomere on 1q (Xiang *et al*., 2001).



(From Gregory *et al.*, 2002)

Figure 1.2 The alignment of syntenic region between human and mouse chromosomes 1. Bacterial clone coverage on each chromosome is represented by red boxes on the ideograms of each chromosome, whilst chromosomes syntenic to human and mouse chromosome 1 are individually coloured and listed.

1.9 Aims of this thesis

The construction of physical maps, prior to the commencement of this thesis, primarily relied upon the assembly of YAC contigs by STS content hybridisation and, to a lesser extent, cosmid assembly into contigs by radioactive restriction digest fingerprinting. If the goal to generate sequence of the human genome was to be realised, techniques would require development that would enable physical maps to be constructed rapidly and safely. The first part of this thesis describes the development of a fluorescence based restriction digest fingerprinting technique that could be applied to the generation of sequence ready maps. The second aim of this thesis was to apply this fingerprinting, in combination with large-insert bacterial clone library screening, to the generation of a 12 Mb contig within 1pcen – 1p13. Minimum tile path clones from the contig would then selected for sequencing and to assist in elucidating disease causing genes that had been localised to the interval (table 1.9)

. **Table 1.9: Disease loci mapping to 1pcen -1p13**

Disease	**Reference**	**Localisation**
Radiation induced meningioma	Zattara-Cannoni *et al.*, 2001	1p11
Autosomal recessive tachycardia	Lahat *et al.*, 2001	1p11 - 1p13.3
Acute megakaryoblastic leukaemia	Mercher *et al.*, 2001	1p13
Hypothyriodism	Dracopoli *et al.*, 1986	1p13
Achromatopsia	Kohl *et al.*, 2002	1p13
Adrenal Hyperplasia II	Zachmann *et al.*, 1979	1p13.1
Colorectal Cancer	Nitta *et al.*, 1987	1p13.2

The third aim was to use the sequence data generated from the first two sections to characterise the genomic landscape of the interval and to identify as many coding features as possible by *in silico* prediction and experimental support. A family of genes, annotated during the course of this section, and seven other genes of medical interest, form the basis for the final part of the thesis.

The aim was to assemble exon specific sequences from 47 unrelated individuals, of the 12 target genes, to confirm the presence of known, or identify novel, single nucleotide polymorphisms. SNPs identified by this strategy would then be categorised according their occurrence in coding sequence and, where found, further analysis carried out to predict what possible affect they may have upon the structure and function of the resultant protein.

Chapter 2

Materials and Methods

Materials

2.1 Chemical reagents

2.2 Enzymes and commercially prepared kits

2.3 Nucleotides

2.4 Solutions

- 2.4.1 Buffers
- 2.4.2 Electrophoresis and Filter preparation solutions
- 2.4.3 Media
- 2.4.4 DNA labelling and hybridisation solutions
- 2.4.5 General DNA preparation solutions

2.5 Size markers

2.6 Hybridisation membranes and X-ray and photographic film

2.7 Sources of genomic DNA

2.8 Bacterial clone libraries

- 2.8.1 Cosmid libraries
- 2.8.2 PAC and BAC libraries
- 2.8.3 cDNA libraries

2.9 Primer sequences

2.10 World Wide Web addresses

Methods

2.11 Isolation of bacterial clone DNA

- 2.11.1 Miniprep of cosmid
- 2.11.2 Microprep of cosmid, PAC and BAC DNA for restriction digest fingerprinting
- 2.11.3 Filterprep of PAC and BAC DNA for restriction digest fingerprinting

2.12 Bacterial clone fingerprinting

- 2.12.1 Radioactive fingerprinting
- 2.12.2 Fluorescent fingerprinting
- 2.12.3 Hind III fingerprinting

Materials

2.1 Chemical reagents

All common chemicals were purchased from Sigma Chemical Co., BDH Chemical Ltd., and Difco Laboratories unless specified below or in the text.

Amersham Pharmacia Biotech	Dextran sulphate, Na^+ salt
Bio-Rad Laboratories	β-mercaptoethanol
Gibco BRL Life Technologies	Foetal bovine serum
	ultraPURE™ Ammonium sulphate, enzyme grade
	ultraPURE™ agarose
Roche Applied Science	Restriction Buffer B
Stratagene®	Perfect Match® (1 U/µl)
	Taq Extender

2.2 Enzymes and commercially prepared kits

All restriction endonucleases were purchased from New England Biolabs, unless listed.

Amersham Pharmacia Biotech	T4 DNA ligase (1 U/µl)
	*Sau*3A1
Bio101Inc	Geneclean II
Gibco BRL Life Technologies	M-MLV reverse transcriptase
New England Biolabs	T4 DNA ligase
PE Applied Biosystems	Amplitaq™
	Sequenase
	TaqFS
Qiagen	DNA gel purification
Roche Applied Science	Klenow enzyme (sequencing grade, 5 U/µl)
	T4 Polynucleotide kinase
Sigma	Ribonuclease A
USB	Shrimp Alkaline Phosphatase
	Exonuclease I

2.3 Nucleotides

Amersham Pharmacia Biotech	Redivue™[α-^{32}P]-dCTP (AA 005) aqueous solution (370 Mbq/ml, 10 mCi/ml)
	Redivue™[γ-^{32}P]-dATP (AG 1001) aqueous solution (370 Mbq/ml, 10 mCi/ml)
	[α-^{35}S]-dATP (Q11135) (370 Mbq/ml, 400 Ci/mmol)
PE Applied Biosystems	Fluorescently labelled (TET, HEX, NED) dideoxyadenosine triphosphate (ddA)
	Fluorescently labelled (ROX) dideoxythymidine triphosphate (ddT)
Amersham Pharmacia Biotech	2'-deoxynucleoside 5'-triphosphates (dATP, dTTP, dGTP, dCTP)
	dideoxyguanine 5' -triphosphate (ddGTP)
	Random hexanucleotides pd$(N)_6$, 5'-PO_4, Na^+ salt

2.4 Solutions

Solutions used in the present study are listed below, alphabetically within each section. Final concentrations of reagents are given for most solutions. Amounts and/or volumes used in preparing solutions are given in some cases. Unless otherwise specified, solutions were made up in autoclaved nanopure water.

2.4.1 Buffers

10x Ligase buffer	500 mM Tris-HCl (pH 7.4)
	100 mM dithiothreitol
	100 mM $MgCl_2$
10x PCR buffer	670 mM Tris-HCl (pH7.4)
	166 mM $(NH4)_2SO_4$
	67 mM $MgCl_2$
1x TE	10 mM Tris-HCl (pH 7.4)
	1 mM EDTA

1x $T_{0.1}E$	10 mM Tris-HCl (pH 8.0) 0.1 mM EDTA
SAP reaction buffer	20 mM Tris-HCl (pH 8.0) 10 mM $MgCl_2$

2.4.2 Electrophoresis and hybridisation solutions

6x Buffer II	0.25% bromophenol blue 0.25% xylene cyanol 15% ficoll
Denaturation solution	0.5 M NaOH 1.5 M NaCl
Formamide dyes	80% v/v deionised formamide 0.1% w/v bromophenol blue 0.1% w/v xylene cyanol 1 mM EDTA 50 mM Tris-borate (pH 8.3) (*i.e* 0.56x TBE)
Formamide dyes mix	0.0075% w/v SDS 3.75 mM EDTA 1.6x formamide dyes
6x Glycerol dyes	30% v/v glycerol 0.1% w/v bromophenol blue 0.1% w/v xylene cyanol 5 mM EDTA (pH 7.5)
Neutralisation solution	1.5 M NaCl 1 M Tris-HCl (pH 7.4)
20x SSC	3 M NaCl 0.3 M Trisodium citrate

10x TAE	400 mM Tris-acetate 20 mM EDTA (pH8.0)
10x TBE	890 mM Tris base 890 mM Borate 20 mM EDTA (pH 8.0)

2.4.3 Media

All media were made up in nanopure water and either autoclaved or filter-sterilised prior to use.

For agar used for bacterial growth 15 mg/ml bacto-agar was added to the appropriate media. Antibiotics were added to media as appropriate (see Table 2.1) to the following final concentrations: Ampicillin (sodium salt dissolved in 1 M sodium bicarbonate, stored at -20^{o}C), 100 μg/ml; Kanamycin (purchased as a solution, stored at 4 ^{o}C), 30 μg/ml; Chloramphenicol (stored at 4 ^{o}C), 12.5 μg/ml (all supplied by Sigma).

Table 2.1: Clones and appropriate antibiotics.

Clone type	**Library**	**Antibiotic**
Cosmid	LL22NC01	Kanamycin
PAC	RPCI1,3,4,5, 6	Kanamycin
BAC	RPCI-11, 13	Chloramphenicol
SIL clones	FS library	Ampicillin

LB	10 mg/ml bacto-tryptone 5 mg/ml yeast extract 10 mg/ml NaCl (pH 7.4)
2X TY	15 mg/ml bacto-tryptone 10 mg/ml yeast extract 5 mg/ml NaCl (pH 7.4)

2.4.4 DNA labelling and hybridisation solutions

100x Denhardt's	20 mg/ml Ficoll 400-DL
	20 mg/ml polyvinylpyrrolidine 40
	20 mg/ml BSA (pentax fraction V)
Hybridisation buffer	6x SSC
	1% w/v N-lauroyl-sarcosine
	10x Denhardt's
	50 mM Tris-HCl (pH 7.4)
	10% w/v dextran sulphate
OLB3	240 mM Tris-HCl (pH 8.0)
	75 mM β-mercaptoethanol
	0.1 mM dATP
	0.1 mM dGTP
	0.1 mM dTTP
	1 M HEPES (pH 6.6)
	0.1 mg/ml hexadeoxyribonucleotides (2.1 OD units/ml)

2.4.5 General DNA preparation solutions

GTE	50 mM glucose
	1 mM EDTA
	25 mM Tris-HCl (pH 8.0)
3 M K^+/5 M Ac^-	300 mM potassium acetate (pH 4.8)
	11.5 ml glacial acetic acid
	28.5 ml H_2O

2.5 Size markers

1 kb ladder (1 mg/ml) (Gibco BRL Life Technologies)
Contains 1 to 12 repeats of a 1,018 bp fragment and vector fragments from 75 to 1,636 bp to produce the following sized fragments in bp: 75, 142, 154, 200, 220, 298, 344, 394, 516/506, 1,018, 1,635, 2,036, 3,054, 4,072, 5,090, 6,108, 7,125, 8,144, 9,162, 10,180, 11,198, 12,216.

Lambda DNA/*Hin*d III (Gibco BRL Life Technologies)
Contains *Hin*d III restricted dsDNA fragments of the following sizes (kb): 23.13, 9.416, 6.557, 4.361, 2.322, 2.027, 0.564, 0.125

Analytical marker DNA wide range (Promega)
Provides an evenly spaced distribution of DNA fragments from 0.702 kb to 29.95 kb

DNA molecular weight marker V (Boehringer - Mannheim)

2.6 Hybridisation membranes and X-ray and photographic film

Amersham	Hybond-N™ Nylon (78 mm x 119 mm) (used for high-density clone gridding)
Polaroid	Polaroid 667 Professional film
X-ray film	Fuji RX medical X-ray film

2.7 Sources of genomic DNA

Human placental DNA for pre-reassociation (ready-sheared) was purchased from Sigma Chemical Co. Human placental DNA for PCR was purchased from Sigma Chemical Co. DNA from whole chromosome 1 hybrid cell line, GM13139, and chromosome 1p specific cell line, GM11526A, was kindly provided by Richard Wooster

2.8 Bacterial clone libraries

2.8.1 Cosmid libraries

Cosmid clones used for the development of the fluorescent fingerprinting were from the Lawrence Livermore flow-sorted 22 chromosome cosmid library (LL22NC01) (prefixed 'cE') were kindly provided by Ian Dunham and the chromosome 22 mapping group.

2.8.2 PAC and BAC libraries

The RPCI-1, RPCI-3, RPCI-4, RPCI-5 (prefixed 'dJ'), and the RPCI-11 (prefixed 'bA') BAC libraries were used as a source of human derived PAC clones and BAC clones respectively in this thesis. These libraries were all kindly provided by Pieter de Jong and Joe Catanese (see http://bacpac.med.buffalo.edu/), and imported and maintained by the Sanger Institute Mapping Core Group.

2.8.3 cDNA libraries

A range of up to 9 different cDNA libraries were used in this study (see Table 2.2). cDNA libraries were imported and maintained by Jacqueline Bye and Susan Rhodes. Each library contains 500,000 cDNA clones, divided into 25 pools of 20,000 clones. Five pools were combined to form a superpool containing 100,000 clones. Prior to their use in PCR, each superpool was diluted 1:100 and 1:1000 in $T_{0.1}E$.

Table 2.2: cDNA libraries used.

cDNA lib. code	cDNA library description	Vector	Source/ Reference
1. U*	(Monocyte NOT activated-from a patient with promonocytic leukaemia) (U937+)	pCDM8	Simmons (1993)
2. AK	Adult kidney	pcDNA3.1	Invitrogen
3. AB	Adult brain	pcDNA3.1-Uni	Invitrogen
4. HeLa	Cervical carcinoma cell	pcDNA3.1-Uni	Invitrogen
5. SK	Neuroblastoma cells	pCDNA1	Invitrogen
6. T	Testis	pCDM8	Clontech
7. FLU	Fetal lung	pCDNA1	Invitrogen
8. HPB*	T cell from a patient with acute lymphocytic leukaemia (HPBALL)	pH3M	Simmons (1993)
9. AH	Adult heart	pcDNA3-Uni	Invitrogen

* Generously provided by Dr Simmons, Oxford (Simmons, D., et al., 1993).

2.9 Primer sequences

All primers were synthesised in house by Dave Fraser or externally by Genset. Table 2.3 lists the universal primer. Tables 2.4 and 2.5 list the primer pairs used in this thesis for cDNA library screening and exon amplification, the sequence and size in base pairs (bp) of each PCR product. The 110 STSs used to screen large insert bacterial clone libraries (and the resultant 878 positive PAC clones) can be found in 1ace within STS pools pool_1pc-1p13_STS1 – 5. Where appropriate, the clones, or genes from which the STSs were derived are also listed.

Table 2.3: Vector-specific primer used in vectorette PCR.

Primer Name	Primer Sequence
224*	CGAATCGTAACCGTTCGTACGAGAATCGCT
BPHI*	CAAGGAGAGGACGCTGTCTGTCGAAGGTAAGGAACGGACGACAG AAGGGAGAG
BPHII*	CTCTCCCTTCTCGAATCGTAACCGTTCGTACGAGAATCGCTGTCCT CTCCTTG

* (Riley, J., et al., 1990)

Table 2.4: STSs designed for cDNA screening and Link PCR product synthesis. S=Sense, A=Anti-sense; a 60^{o}C annealing temperature was used for all PCR reactions. stSG # corresponds EMBL accession numbers assigned to individual STSs.

Gene	Exon	stSG #	Primer 1 (S)	Primer 2 (A)	bp
bA483I13.C1.1.mRNA	e2	452926	GGTATCTGCCGACCCTTTGT	GAGTAGGCAGTAGCTTGAGT	147
bA483I13.C1.3.mRNA	e2	452927	GCAGTCTGGAGATTGGTGGA	TGCATCATGACTTTCAAGCG	102
	e5	452928	GGGATTATTGATTGTGGCAA	CGGCATAAGGTACAATGCCT	100
	e7	452929	GGTTTCACCTCAAACATCAT	ACATCTCTTTATAACACAGG	164
bA475E11.C1.2.mRNA	e1 5'UTR	452930	TGACGGCTGAAGAAACAGTG	CTCCAGGGCCAGCATACTAA	104
	e4	452931	GCTTTTGACTTTGCCTCGTC	GCTTCCTATCAGCAGGGATG	128
	e7	452932	CAGCACTCAACCAGCAATGT	TCCAGGATTACGAGGAGTGC	149
	e10	452933	ATGAGACTCCTAAGCAGCCG	GGGCAGCACTTTGACGTATT	137
	e12	452934	ATACCTGGAGTGGCTGGATG	GTTGTGCCAACAACACGAAC	100
	e14 3'UTR	452935	CGAAGAGGCCCCTTATTACC	GGAGTGCACACCAACAACTG	166
bA475E11.C1.1	e1 5'UTR	452936	AGGCTATGCATAGTGAGACT	GCTTGACTTAGAAGCGTCTC	155
	e5	452937	ACTGCAGGGACACCTTGAAC	CCAACGATTGTTGATTCGTG	105
	e6	452938	TTGGAGATGCTGCTTGAAGA	AGAGAAGGTGGAGGCCAAGT	117
	e10	452939	GAAGCAAAACGTGGAGAAAA	TTGAATCTGAGTGTGGTGCC	92
	e13	452940	CTTCCAAATCCAGCCCTACA	ATGGGTTGCTACCAACTTGC	127
bA297O4.C1.1.mRNA	e1 5'UTR	452941	GCCACTATTGGGAGACCAAG	GTAGAGCCAGAGGTTCGACG	124
dJ831G13.C1.1.mRNA	e1	452942	CTTTGCTATTTTCGCCTTCG	CTGAAGGGATAGCCAAATGC	123
	e3	452943	CCTTCACCTTCTTCTGGCTG	CTTCCTCTCCATGGCACACT	120
	e4	452944	GCTCAGTGCTTCTTGTGCAG	TGGCTGCTAGGAACCAGTCT	146
bA180N18A.C1.2.mRNA	e1	452945	CCCACTGATCGTGAACAACA	CTCTGAGTCTTTGCGCTGGT	154
dJ773N10.C1.1.mRNA	e2	452946	CAGCCTGCATCTTCCTCTTT	AGACCTTCTCCAGCTCCTCC	125
dJ1003J2.C1.1	e2	452947	CCCTGAATGAGAAGGAGCTG	CACGGACTCAGTGACATGCT	148

	e4	452948	CTGTCTCTTTGTGGGGCTGT	ACAGGACATTCACTCCAGGG	102
	e7	452949	ACCCAGGTCTTCTTTGCCTT	GCCAACACTGACGTGAAGAA	134
	e9	452950	CCTTCATCGCCTTCACTGAG	GTGAACATCTCCTTGGGCAC	169
	e12	452951	CGCTACCTGTATTTCCCCAA	CCCTTCTTGTAGGACACGGA	149
bA470L19.C1.2.mRNA	e1	452952	CACCAAGCATTCCATACGTG	GAACCCAATGGGGATTCTTT	150
bA284N8.C1.2/.3	e2	452953	ATGCTCGGCTGTCTTCAAGT	AATGGTGAGTCATTCTGGGC	128
bA165H20.C1.3.mRNA	e3	452954	TGTTACTTCACCAACTGGGC	TGGTGATCTCGTTGTTCTGC	127
	e5	452955	TTCCACTCCTGAGAACCACC	CTGCACCAGGACAGTGAAGA	151
	e7	452956	CTATGACCTCCATGGCTCCT	ACATTGAGGTAGGCGTTGCT	96
	e10	452957	AACAACTTTGGAGGTGCCAT	TTGTACTCTGCAGGCCCAGA	124
dJ1125M8.C1.1mRNA	e2	452958	GCGGATAACTACCCTTTTGG	AAATAACACCCAGGCCCTCT	128
	e4	452959	ATGCAGGCAGGTACCAGAAA	TTCTTAAATCGAGGCACCAA	91
	e6 3'UTR	452960	ACGTTACTGTGGCCCTCTTG	ACAGAAACCCACAGACCCAG	154
dJ1125M8.C1.2	e1	452961	GAATGGAGGAGCAGGGTGTA	TCCAGGTAGTTGGTGAAGGG	121
	e3	452962	ACAACAGGTTCAATCCCAGC	TGTCATAGCCCAGGAACACA	105
	e5	452963	ACCCGCCAGTATTGTGGAGA	GGCAATCTGCCAGTACAGTT	107
	e8	452964	AACAATGGCTACTGCAGGCT	CTAGGCAGAGAAGGCAAAGC	153
bA552M11.C1.4.1/.2	e2 .2	452965	ACCACGTGGGATTTGATGTT	GGATGCCAAATTAAGAGCCA	137
	e3/4 .1	452966	CCTTTTGTGCTGGGGTTCTA	GCTGGAGGATCTGAGTGAGG	121
	e5	452967	TGAGGCTTGAATCCATTTCC	CTCTGGCCAGGAAAAGACTG	175
bA552M11.C1.5	e1/2	452968	TGCTTCCTTCCAGTCATGTG	TGGTCAGGCAGGACATAGTG	120
	e3/4	452969	CAGGCAACAAAACCAGAAGC	CCCAAACCCGTGATCAGTAT	103
	e5	452970	ATCATTTGCAGCCAGGTAGC	GTCCCCAATCCAGATTCTCC	154
dJ836N10.C1.1	e2/3	452971	GGAAGAACAAGGAAAAGGGC	CTCAATGCTTCCCCTCACTG	176
	e4	452972	AAAAGCCAGAGCTTCCTGAC	TGTGGTCCCTTTCTTCTTGT	120
dJ1073O3.C1.3	e1	452973	CTGGGCTGAAAACTGCTTGT	GTTGGGCTCAAGAAGTCCAT	134
	e3	452974	GACCTGGTGTGCTCAGGATT	TTCCCATTGATCATACCCGT	144
dJ1037B23.C1.1.mRNA	e2	452975	TCCCTCTTCTGCTAATCCCC	ACCTCAGCTGGGATATCTGG	122
	e4	452976	AGCGTGGACTTGGGAGAGAT	GTGATGTCCATCGCCTTGAG	107
	e6	452977	TAGGAGTCTGTCGTGTGGGG	TTACCTCCACCAAGGAGTGC	114
	e8 3'UTR	452978	AAACAGTGTTGTGCAGTCGC	CATCACCTTGGGAGACACAA	144
dJ1156J9.C1.1	e1 5'UTR	452979	ACCTTGGAGCGGGATCTTAT	TGCCAGGGAATTGTTGTATG	127
dJ929G5.C1.1.mRNA	e2	452980	TCCTGTTGAAGAGTGGCTCC	TCCAGAATAAGTGGATTCGG	157
	e4	452981	GTTTGTGTTTCGTGCCCTTT	TATTGCACAATGCCCTGGTA	120
	e6	452982	CAGTAACAATGCCACTGGCC	CTTCTTTACTCGCCGTTCAT	118
	e8	452983	GCAATATGACAAGGACCGCT	TACGAGGCTGAAGTCCAAGC	121
bA12L8.C1.1.mRNA	e2	452984	CATCCTCATTGCACTGGTTG	TGCACGTGCTTATGGATCTC	159
dJ655J12.C1.2.mRNA	e1	452985	AAGACAAGGAAGAGCACCTG	GAGTCCTTGAAGTGGTCGGA	120
	e2	452986	GCTCTGTTCAGGAAAATGCC	TGATCATCAGTGAGCCAAGC	165
dJ655J12.C1.3.mRNA	e2	452987	AGTCCTCCCTGAACTGTTGC	TGGGCATGAGATAAAACACG	106
dJ686J16.C1.1.mRNA	e1/2	452988	GGGCAGTGTCCAATTTATGG	GGAAGGAGGACTGATGGTGA	116
bA39H13.C1.1.mRNA	e1	452989	AAAAACCCAGCTGGACAATG	TCAGCAAGATTCCTCGGTCT	142
	e2	452990	ATCTGGAAGCAGAGCCAGTA	CCATTTCAGAGCTTCTGTGC	90
bA42I21.C1.1.mRNA	e1	452991	CACATGCGTCGGCTTAAATG	CCTCCACGATCGATGTTTCT	95
	e2	452992	CCCTCGCTGGGAAAGACATA	TGCTGGGGGAAAAGATTACT	139
dJ776P7.C1.1	e1	452993	GGCACTCTATTCGCACGTCT	CTCCATCATCCCAGGACACT	99
	e1/2	452994	GCTGAGAGGATTATGGAGGC	CTGAACTCTGCCCCTTACCA	101
	e4	452995	CTGTCCTCCCACTGGAATGT	TTCCGAGGTGAAGGAGAAAG	145
dJ832K2.C1.1.mRNA	e1	452996	AAAAACTCCAGGACCTCCGT	ACCTGCAGCCTCAGTTTCAC	171
	e6/7	452997	CCGCATAATACCACCCTTTT	CAGCTGTTTCGTTTGCATCT	131
dJ832K2.C1.2.mRNA	e2	452998	CCTCCAAACACAGGCTCTCT	CATGATGTACCTGCCAGCTC	126
dJ832K2.C1.3.mRNA	e2	452999	CCTTCAAGAAGCCCATAAGC	CAACATTGGAGTGGAGAGCA	146

	e5	453000	AGGCAAGGATAACGCAGAGA	CTTAGGTTCTGGTTGGTGGG	133
	e8	453001	ATCCCTTCAGCACTCACTCC	TCTTTGGGTTTTTCTTTGCC	98
bA224F24.C1.1.mRNA	e1	453002	TTAGAGGCCAATGCTTCTCC	AGCGAGGGTTCCCATATCTT	95
	e4	453003	ACGGCAGCAAAAGCAATTAT	TTCTTTTCATTTTCCCGTCG	125
	e6	453004	GGGCTTTAACAATCCTCAGC	CTGGTTAACTGCTGCCAGGT	124
	e8	453005	TTGCCTGCTTGATGTATGAC	CTCTGAAGTTGGCATGGCTT	131
	e11	453006	AAGAAGATCTCGTCCCACCC	GACAGAGTGAGGGCAGAAGG	105
	e15	453007	AGCAACCGAGAACCTTCAGA	AGAGACTCATGTTGGGGCTG	149
dJ794L19.C1.1.mRNA	e1	453008	GGCGGCTAAAATGAGTGAAA	ATAGACAGGTCCAGCCCCTT	141
	e3	453009	AGGATGTTTCCTGCCATGAG	TTTTATTGTCCACAGGCACA	94
	e5	453010	CTCGAGTTCATGTGATTCGC	AGGCCGTAACTGTGGTGAAC	123
	e8	453011	TACCAACTCCTCCCTCGTTG	CATGTGTGGTGATGAGGAGC	137
dJ834N19.C1.1.mRNA	e1	453012	TACCGGTCAGACTCCAGGTC	AGGTCCTCTTCTTTGCCTCC	93
	e3	453013	CAGTAACTGAGGAGGGCCAC	GGCTGCGATAGAAAGCAAAG	143
dJ834N19.C1.2.mRNA	e2	453014	AGACGAGGTCTTGCCACATT	GCATGGTGGCTTATGCTGTA	108
dJ599G15.C1.5.mRNA	e1	453015	TCGCTAGCCATTATCCAACC	CCTGTCCTTGTAGTGGGCAT	129
dJ599G15.C1.6.mRNA	e1	453016	ACTCTTCAGGAGCCACATGC	TCTACTGGAAGAGCACCAGC	95
dJ1042I8.C1.4.mRNA	e2	453017	ATGCTGGCCACAATCTACCT	GATCACTCCCCACAGCACTT	127
	e3	453018	TGGTCCAGTGAGAAAGCAGA	CCGGCCATTTGAGTTACAAG	125
	e5	453019	GGAACGAGAGCTGATCCAGT	AGCTGTTCTCGGAAGTCCTG	138
	e7	453020	GGAGGAATGTGCCATCACTT	GAGCATCCTGCCATTCATCT	152
	e9	453021	AGGAAGCTGCAGGAGTCTGA	CCAAGAAAGTGCCTTCACAA	124

Table 2.5: Exon specific primer pairs designed to pharmacogenomic gene targets. An annealing temperature of 60^{o}C was used for PCR reactions except where listed in chapter 6.

Target	Chr	Exon	stSG #	Primer 1	Primer 2	bp
GSTM4	1	1	452701	TCCGGACCTTGCTCCCTGAA	CCTAGTCTACACTGCACTGC	528
		2	452702	TTCCCTCCTTAGGGCTATCT	TGTCATAGTCAGGAGCTGCC	724
		3/4	452703	TCACTTCTTCTTCCCCACGG	ACACAGACTCACTCTGAGCAT	566
		5	452704	TGGCTGGATTGGGGTGCTAT	GGTGCTATTACATCCCACTACA	527
		6/7	452705	GTGTAATAAATGCTGGTATG	AGGACTGACCCTCTCATTCAA	701
		8	452706	CCTCAGCACTTGAGCCCACG	AGCAAATAAGACAAGACTAT	689
GSTM2	1	1	452707	AGCCCCATGAGCGCGCTCCA	GGGGAGCCCCATCTCCTCCT	408
		2	452708	GCGGTGGGACGGGGGTGCGT	CCCATCATAATTACCCAGAC	534
		3/4	452709	GCCCCCGTCTGGGTAATTAT	CGCATCTTGCAACCAAATCT	719
		5	452710	TCGGCTTGGCTGGGCTGTGAG	TGTATTTTCTTTCACTCGTCA	614
		6/7	452711	CAGCTGGGGCCATGCACAAA	GGCGTGAGCCACCGCACATG	508
GSTM1	1	1	452712	CCTGGGAGGCGGGAGGAAGT	GACTTTGTCTGCACCAGGGA	448
		2	452713	TGGGACGAGGGCGCAGGGGA	AAGCCCTGAGGGACACCCGT	535
		3	452714	GCCCCCTGTCTAATTGGGAC	TCACATGAACGAATGCAGGT	484
		4/5	452715	TGTCCACCTGCATTCGTTCA	AGATGCAGCTCACTGGGGAC	468
		6/7	452716	CTCTGCCTTCTGATCAGTTT	GATAATTCTGTTACCTTACT	709
GSTM5	1	1	452717	ACTGGGAGGCGGGAGGGGGC	GACTTTCTCTGCACCAGGCC	448
		2	452718	TAAGCGAGGGTCCTCTGGTG	AACCCAATTAGACAGGGTGT	548
		3/4	452719	CTGGGGCGGGATGCTGGACA	TGACCAGCTCCATGTGGTTA	878

		5	452720	AATGTGCGGGGGGAAGGTGA	GGTAGCAGATCATGACCAGT	409
		6/7	452721	GAAGAGCATCTCATTCTGAT	GTATAATGTGCTGGGCATGA	663
		8	452722	GCCTGCAGCAAAGCTACTTG	ACAGTCCTGAGTCAAGGGAG	621
GSTM3	1	1	452723	CGGCCCTGTGGAGCCGCGGA	CAGCGGTTCAGCGACTGCGC	520
		2/3	452724	CCCCGGGCCGGGGAACGTTA	GCAAGGATGGATATACTTGAA	622
		4/5	452725	TGTTCACTGCCCTGCAAGTGT	GCAGCAGAATGGAACAGAGA	463
		6/7	452726	GTGCTGGTGCCTCTTCTTTC	CCCATTAGGCAAAAGCCGGG	567
		8	452727	GGTTGGGGTCGTTATAAGAT	TCTCCTACCCCGTGGTCACA	748
GSTP1	11	1	158595	GGTCCTCTTCCTGCTGTCTG	CCCCTGAAAGCCGCTAAC	465
		2	158596	GTTAGCGGCTTTCAGGGG	GAGGGAACAGGGAAACAGGT	384
		3	158597	GCCCCAGTGTTGTGTGAAAT	AGGTCTCCGTCCTGGAACTT	385
		3/4	158591	TCTCGTACTTCTCCCTCCCC	GATTTAAACAAAAGGGCTCCG	399
		5	158592	ACATCCTCTTCCCCTCCTCC	AGGTTGTGTCTTGTCCCAGG	399
		6	158593	AAGGATGGACAGGCAGAATG	CATCCCCTAGGTCTGCTCTG	399
		7	158594	GCTTCCAGATGGACACAGGT	TGCTGGAGGAGCTGTTTTCT	498
GSTT1	22	1	140015	CTCCAAACCAGACCAGCAAT	CTAAAGAGTGTCCCAGGCGT	348
		2	140017	TGGAATAGCAGGAAGGCAAG	GTCTTTGCCAACCAGGAGTG	351
		3	452728	GACTATGTATGAAATACCCA	CTGGTGCCTGAACACCTTTG	276
		4	452729	GCTCAGCATCACTAATCATT	GATTTGGGGACCACAGATCT	406
		5	140020	GGGGGTTGTCTTTTGCATAG	CCTGCTTATGCTGCCACAC	659
CYP1A1	15	1a	452730	CTGACACTCTAGATATTGGCT	GTCAGAGGCAATGGAGAAAC	556
		1b	452731	ATGGTCAGAGCATGTCCTTC	CCCAGGCCCTGATGCCATCT	533
		2/3/4	452732	CCTGTGGACTTTCCCTACCT	CAGTGGCTCCATGGGGCCTT	564
		5/6	452733	TTGCCCTGAGCCTGACTGAG	GGTAGACAGAGTCTAGGCCT	637
		3' UTRa	452734	TTGAGAGCCCTGAGGCCTAG	CAGGACAGCAATAAGGGTCT	620
		3' UTRb	452735	CAGCAAGTTAGAACTAGCCA	GGCTACACCTCTTCACTGCT	697
CYP1A2	15	1a	452736	ATCTCAACCCTCAGCCTGGT	CTCATCGCTACTCTCAGGGA	700
		1b	452737	GCCCTCAACACCTTCTCCAT	TCTGAGGTGTCCAGAGCCTT	536
		2/3	452738	ACCTTGGAAGTGCCAGAGGT	TCAAGGCTTCTCCTGGCTAT	753
		4	452739	GACAGTCTTACATAAGAGTG	CAATAGGGTCATGCTTGTGA	511
		5	452740	TGCTGAAGTTAAAGAACAGG	GATTATAGGCTTGAGCCACT	484
		6	452741	CCATCTCCTCTGTTCCTCTT	GCCTCCTAAAATGCTGGGAT	464
CYP2A6	19	½	452742	GGCAGTATAAAGGCAAACCA	GAACACTGAGACCTTCGTGT	722
		¾	452743	TGTCTCCATTCCCGCGTTCA	GCAGTTGGCAGGTTGTGGTA	626
		5	452744	CAGCCTCGTTTAAATACCTG	GGATTACAGGCTGTTAGCCA	588
		6	452745	GAGCGAGTCTGGTAGATCTA	CCTGTCTCTGGACAGCAAGT	404
		7	452746	TGGCACAACCTGGTTAACAG	CAGGGTCTAGAAAGCTTCTA	467
		8	452747	CCTGTTTCAGAGATGTGAAC	GGTAGATTCTAACAGGAACT	443
		9	452748	TGCACTGAGAGTGGGCTTCA	ATTAGGTGAGCGTGCAATGG	538
CYP3A4	7	1	452755	ACTGCAGGCAGAGCACAGGT	GGCATGATCTCAGCTCACTG	613
		2	452756	GACCATGAAGACTTCAGCTG	GAGGTTCCTGAGAGTTAGCA	519
		3	452757	GACATCAGAAAGACAAAGAG	TCCCATTGCAATACTCTACA	510
		4	452758	CATATGAAGACTTGAGTGGC	GCACATAAAGCTGGTGAAAT	549
		5/6	452759	CCTCCACAACTGATGTAGGA	GGAATATCAGCTCCATGGCA	589
		7	452760	AGTGTCTCCATCACACCCAG	ATGATGACAGGGTTTGTGAC	592

		8	452761	GTCATAGAAGGAATGGCTTC	GCTGTCTCTGACTCATTCTC	476
		9	452762	CCATCTCACATGATAGCCAG	TCTAGCATGCCAGGTTTGCT	512
		10	452763	GCCCACATTCTCGAAGACCT	TTCAGAGCCTTCCTACATAG	523
		11	452764	ACATCTCAGTGGGCCACTGA	GGACATAACTGATGACCTTC	654
		12	452765	CAGCCCACAAAAGTATCCTG	GGCCTAATTGATTCTTTGGC	419
		13a	452766	CCTCAACACTGAAGGAGTGT	GTGCAGGAAAGCATCTGATA	669
		13b	452767	CACATGTTTTCTCTGGAGTA	GTGCTTTTAGGCTTATTGCT	791
NFE2L2	2	1	452768	AGAGCGCTGCCCTTATTTGC	TCCTGGCTCTGGCCAGACGT	589
		2	452769	CACTTCCCACCATCAACAGT	GTGTTTCCTTAAACCTGCCA	419
		3	452770	GTGCATCAAAGTGATCTCTG	CTTCGTTTATTGCCCAGCTG	577
		4	452771	GGGTCATGACTGGTTAGTAA	TCAGAGTTCCCAGATCAGAC	527
		5a	452772	GAGATAAGCCTGAAGATAAT	TCTTCCACTTCAGAATCACT	649
		5b	452773	TGTGGCATCACCAGAACACT	GCTGCAGGGAGTATTCACTA	695
		3' UTR	452774	GCATGCTACGTGATGAAGAT	TTATTTCTCTGTAACCCTGG	756

2.10 World Wide Web addresses

ACeDB	http://www.acedb.org/
ACT	http://www.sanger.ac.uk/Software/ACT
Baylor College of Medicine Search Launcher	http://searchlauncher.bcm.tmc.edu/
British Columbia Genome Sequence Centre	http://www.bcgsc.bc.ca/
Chromosome 1 Mapping Project - Sanger Institute	http://www.sanger.ac.uk/HGP/Chr1/
CHLC	http://chlc.org/
Coriell	http://locus.umdnj.edu/ccr
dbSNP	http://www.ncbi.nlm.nih.gov/SNP /index.html
DOTTER	www.cgr.ki.se/cgr/groups/sonn/ Dotter.html
EMBL	http://www.ebi.ac.uk/
Ensembl	http://www.ensembl.org/
GDB	http://gdbwww.gdb.org/
Généthon	http://www.genethon.fr/php/index.php
HGNC	http://www.gene.ucl.ac.uk/ nomenclature/
INTERPRO	http://www.ebi.ac.uk/interpro/scan.html
Incyte	http://incyte.com/company/news/1999 /genes.shtml
Locus Link	http://www.ncbi.nlm.nih.gov/LocusLink/
National Centre for Biotechnology Information	http://www.ncbi.nlm.nih.gov/

OMIM	http://www3.ncbi.nlm.nih.gov/Omim/
Primer 3	http://www.sanger.ac.uk/cgi-bin/primer3.cgi
PIX	http://www.hgmp.mrc.ac.uk/Registered/Webapp/pix/
PSI-BLAST	http://www.ncbi.nlm.nih.gov/BLAST/
RepeatMasker	http://ftp.genome.washington.edu/RM/webrepeatmaskerhelp.html
The Institute for Genome Research	http://www.tigr.org/
The SNP Consortium	http://snp.cshl.org
The Wellcome Trust Sanger Institute	http://www.sanger.ac.uk/
Unigene	http://www.ncbi.nlm.nih.gov/UniGene/
University of California, Santa Cruz genome browser	http://genome.cse.ucsc.edu/
Washington University Center Genome Sequencing Center	http://genome.wustl.edu/
Whitehead Institute	http://www-genome.wi.mit.edu/

Methods

2.11 Isolation of bacterial clone DNA

2.11.1 Miniprep of cosmid

1. Ten ml of 2 X TY containing 30 μg/ml of appropriate antibiotic (kanamycin for cosmids and PACs, chloramphenicol for BACs) were inoculated with a scraping from the frozen glycerol stock of the chosen bacterial clone and incubated overnight at 37^{o}C with shaking.
2. The cells were collected by centrifugation at 4,000 rpm for 10 minutes at room temperature in a Beckman J6-MC, re-suspended in 200 μl of GTE in a 1.5 ml eppendorf tube, and left on ice for 5 minutes.
3. 400 μl of freshly prepared 0.2 M NaOH/1% SDS were added to the cells, mixed by gentle inversion, and the sample left on ice for another 5 minutes.
4. 300 μl of 3 M K^{+}/5 M Ac^{-} (pH 4.8) were added, mixed by gentle inversion and left on ice for 10 minutes. The sample was centrifuged for 10 minutes at 14,000 rpm in an eppendorf microfuge.

5. The supernatant was transferred to a fresh tube and mixed with 600 µl of cold isopropanol and left on ice for at least 10 minutes. The tube was subjected to centrifugation for 15 minutes at 14,000 rpm in an Eppendorf microfuge at 4°C to pellet the DNA, the supernatant removed and the pellet re-suspended in 200 µl of $T_{0.1}E$.
6. 200 µl of 50:50 (v/v) phenol/chloroform were added to the sample, which was vortexed and briefly centrifuged. 20 µl of 3 M sodium acetate (pH 5.2) and 200 µl of isopropanol were added to the aqueous layer, and the sample placed at -20°C for at least 10 minutes. The tube was subjected to centrifugation at 14,000 rpm in an Eppendorf microfuge for 15 minutes at 4°C to pellet the DNA. The pellet was washed with 70% ethanol and re-suspended in 50 µl of $T_{0.1}E$.
7. 1 µl of 10 mg/ml RNase was added and the sample incubated at 37°C for 1 hour prior to storage at -20°C.

2.11.2 Microprep of cosmid, PAC and BAC DNA for restriction digest fingerprinting

1. 500 µl of 2 X TY containing 30 µg/ml of appropriate antibiotic (kanamycin for cosmids and PACs, chloramphenicol for BACs) were added to a 96-well deep-well microtitre plate (COSTAR).
2. Each well was inoculated from a glycerol stock with either a 96-well inoculating tool, or a sterile cocktail stick. A plate sealer (Dyntax) was placed on top of the plate to seal the wells, and the cultures grown for 18 hours at 37°C with gentle shaking.
3. For each well, 250 µl of the overnight growth were transferred to a clean microtitre plate. The cells were collected by centrifugation (Sorvall RT7, Du Pont Company Sorvall, Delaware US) at 1550 g for 4 minutes.
4. For each well, the supernatant was removed and the pellet re-suspended in 25 µl of GTE, by vortexing gently (a cocktail stick was used for resuspending pellets still attached to the plate).
5. 25 µl of GTE were added to each well and gently mixed. 25 µl of freshly prepared 0.2 M NaOH/1% SDS were added, mixed and left to stand for 5 minutes at RT.
6. 25 µl of 3 M K^+/5 M Ac^- (pH5.0) were added, mixed and left at RT for 5 minutes. A plate sealer was placed on top of the plate and the plate was vortexed gently for 10 seconds, prior to centrifugation of the samples in a microtitre plate in a Sorvall RT7 centrifuge at 1800 g for 10 minutes at 4°C (Sorvall RT7, Du Pont Company Sorvall, Delaware US).

7. 75 µl of the supernatant from each well were transferred to a fresh microtitre plate containing 150 µl of 96% ethanol (or 100 µl of isopropanol) and left at -20 °C for 30 minutes.
8. The microtitre plate containing the samples was centrifuged at 1800 g in a Sorvall RT7 centrifuge for 10 minutes at 4°C. The supernatant was discarded and the samples drained on a tissue.
9. 25 µl of ddH_20 were added to each well and the pellet resuspended by tapping or with a sterile toothpick. 25 µl of 4.4 M lithium chloride were then added to each well mixed by tapping and left for 60 minutes at 4°C.
10. After centrifuging the samples in a Sorvall RT7 centrifuge at 1800 g for 10 minutes at 4°C, the supernatant from each well was added to 100 µl of 96% ethanol in a fresh plate and left for 60 minutes. The samples were then spun in a Sorvall RT7 at 1800g for 10 minutes at 4°C, the supernatant discarded and 200 µl of 70% ethanol added. The pellets were then air dried until transparent.
11. 10 µl of $T_{0.1}E$ were added to each well and mixed by careful pipetting up and down. The resuspended DNA was then used directly for fingerprinting or stored at -20 °C.

2.11.3 Filterprep of cosmid, PAC and BAC DNA for restriction digest fingerprinting

1. Steps 1 – 6 from the microprep procedure (*2.11.2*) were followed before progressing to step 2 of the filterprep procedure.
2. A microtitre plate containing 100 µl of isopropanol was taped to the bottom of 2 µm filter-bottomed plate (Millipore cat. no. MAGVN2250). The total well volume of the sample was transferred to the filter-bottomed plate and the sample was filtered by centrifugation in a Sorvall RT7 at 1550 g for 2 minutes at 20°C.
3. The filter-bottomed plate was removed and the microtitre plate was left at RT for 30 minutes, before samples were centrifuged in a Sorvall RT7 at 1500g for 20 minutes at 20°C.
4. The supernatant was tipped from the microtitre plate and the DNA dried by inverting the plate and placing it on clean tissue paper, ensuring no disruption of the pellet.
5. 100 µl of 70% ethanol were added to the dried DNA to wash the pellet, mixed gently, and the DNA precipitated by centrifugation in a Sorvall RT7 at 1500g for 10 minutes at 20°C. For restriction digest fingerprinting the wash was repeated.
6. 5 µl of freshly prepared $T_{0.1}E$ / 1 µg/ml RNase were added and mixed gently to resuspend the DNA. Samples were stored at -20°C.

2.12 Bacterial clone fingerprinting

2.12.1 Radioactive fingerprinting

1. For each well of a 96-well microtitre plate of sample DNA, a premix containing 1x NEB2 buffer (as supplied by New England Biolabs), 0.72 U *Hin*d III, 1.3 U *Sau*3AI, 0.4 U Reverse Transcriptase, 0.07 µl [alpha-^{32}P]dATP (3000Ci/mmol), 0.04 µl 10 mM ddG was prepared in a 1.5 ml microfuge tube.
2. 2 µl of premix were added to the sample DNA using a Hamilton repeat dispenser. The reaction was mixed by gentle agitation and the samples were centrifuged in a Sorvall RT7 at 150 *g* for 10 seconds.
3. The reaction was incubated for 1 hour at 37^{o}C. The reaction was then stopped by the addition of 2 µl formamide dye.
4. The sample DNA was denatured at 80^{o}C for 10 minutes and loaded in groups of 6 onto a 4% polyacrylamide gel, leaving the first well and every subsequent seventh well of empty (see Section 2.14.2). Marker DNA (see Section 2.13.1) was denatured by boiling for 5 minutes and 2 µl were loaded in the first well and every seventh well. Fragments were resolved by running the gel at 74 W for 1.5 hours (or until the bromophenol blue dye front reached the bottom of the gel).
5. Following electrophoresis, the back plate was removed and the gel was fixed in a 10 % glacial acetic acid solution for 10 minutes then washed in water for 25 minutes. The gel was dried onto the front plate by incubation at 80^{o}C for 45 minutes in an oven. Autoradiography was for 72 hours at RT.
6. The autoradiograph was scanned using a flat bed scanner (Amersham) and the digitised version imported to IMAGE.

2.12.2 Fluorescent fingerprinting

1. For one 96-well microtitre plate of sample DNAs, three digest premixes were prepared, one for each fluorescent label, in three 1.5 ml microfuge tubes labelled TET, HEX and NED. Each premix contained 25.5 µl $T_{0.1}E$, 24.5 µl NEB2 buffer, 5.0 µl *Hin*d III (20 U/µL), 8.0 µl Taq FS, (32 U/µl) and 3.0 µl *Sau*3AI (30 U/µl) and 4.0 µl of the appropriate ddA-dye. Each premix was mixed prior to being aliquoted.
2. 2 µl of the TET premix were added to wells A1-H4 of the microtitre plate containing sample DNAs using a Hamilton repeat dispenser. Similarly, 2 µl of the HEX premix were added to wells A5-H8 and 2 µl of the NED premix were added to wells A9-H12. The plate was covered with a plate sealer, the reaction mixed by gentle agitation on a vortex. In order to ensure the sample was in the bottom of the wells the plate was centrifuged at 150 *g* for 10 seconds (Sorvall RT7, Du Pont Company Sorvall, Delaware US).
3. The reaction was incubated for 1 hour at 37^{o}C.
4. To precipitate the DNA, 7 µl of 0.3 M sodium acetate (pH 5.2) and 40 µl 96% ethanol were added to each well. For multiplexing the samples, rows 5 and 9 were added to row 1, rows 6 and 10 were added to row 2, rows 7 and 11 were added to row 3, and rows 8 and 12 were added to row 4 respectively, using a multichannel pipette.
5. The samples were incubated at RT for 30 minutes in the dark.
6. The samples in the microtitre plate were centrifuged in a Sorvall RT7 at 1550 *g* for 20 minutes at 20^{o}C to pellet the DNA.
7. The supernatants were discarded and the pellets dried by tapping the plate face down onto tissue paper.
8. The pellets were washed by adding 100 µl of 70% ethanol to each well, mixed gently tapping the plate, and the samples centrifuged in a Sorvall RT7 at 1550 *g* for 10 minutes at 20^{o}C.
9. The supernatants were discarded and the pellet dried as in step 7.
10. The fingerprinted DNAs were re-suspended in 5 µL $T_{0.1}E$.
11. Prior to loading, 2 µl of the marker DNA (see Section 2.13.2) were added to each sample using a Hamilton repeat dispenser. The samples were denatured for 10 minutes at 80^{o}C. 1.25 µl of each sample were loaded on a 5% denaturing acrylamide gel and resolved on an ABI377 Automated DNA sequencer using a 0.2 mm, 12 cm well-to-read 4.5% denaturing polyacrylamide gel (prepared by Sanger Institute Gel Production Team). Data were collected using the ABI Prism Collection Software v1.1.
12. After data collection, the gel image was transferred to a UNIX workstation for entry into IMAGE.

2.12.3 *Hin*d III and *Eco* RI fingerprinting

1. For one 96-well microtitre plate of sample DNA, a premix containing 231 µl H_20, 99 µl Boehringer buffer B (50 mM NaCl, 10 mM Tris-HCl, 10 mM MgCl2, 1 mM DTE, pH 7.5), 55 µl *Hin*d III 10U/µl (or *Eco* RI 10U/µl), was prepared in a 1.5 ml microfuge tube, and mixed using a vortex. 4 µl of the premix were added to each well of a 96-well microtitre plate containing previously prepared DNA (see Section 2.11.2), and the plate covered with a plate sealer (Dynex).
2. The reactions were mixed gently on a vortex and incubated at 37 ^oC for 2 hours.
3. The reactions were terminated by the addition of 2 µl of buffer II and either loaded straight away or stored at 4 ^oC.
4. 0.8 µl of the marker (see Section 2.13.3) were added to the first well and then every sixth well of a freshly prepared 1% agarose/1x TAE gel (see Section 2.14.1 for preparation). 1 µl of each sample was loaded (i.e. wells 2-5, 7-10 *etc*.) between the marker lanes. Fragments were resolved by electrophoresis through the gel at 4^oC in a cold room for 15 hours at 90 volts.
5. Following electrophoresis, the gel was cut down so the length was 19-20 cm and stained in Vista Green (mix 5 ml 1 M Tris HCL, 0.5 ml 0.1 M EDTA, 50 µl Vista Green, make up to 500 ml with H_20) for 30-45 minutes on a shaker. The gel was washed with ddH_20 to remove excessive stain.
6. The gels were scanned on a FluorImager SI. The parameters were set to 530 nm for emission filter, the pixel size was 100 microns, detection sensitivity was normal, digital resolution was at 16 bits, dye was single label, excitation filter was 488 nm, Em filter 1530 nm and PMT voltage was 800.
7. The gel image was transferred to a UNIX workstation for entry into IMAGE.

2.13 Marker preparation

2.13.1 Radioactive fingerprinting marker

1. 171 µl of $T_{0.1}E$, 25 µl of 10 X NEB2 buffer (as supplied by New England Biolobas), 16.5 µl of lambda DNA (500 µg/ml) and 5 µl of *Sau*3AI (50 U/µl) was added to a 1.5 ml Epperndorf microfuge tube and incubated at 37^{o}C for 1 hour
2. To a 43.5 µl aliquot of digested lambda DNA (from step 1) in a 1.5 ml Epperndorf microfuge tube was added 2.0 µl 10 mM dGTP, 2.0 µl 10 mM ddTTP, 4.0 µl [alpha-^{35}S] dATP (3000Ci/mmol), 1.0 µl AMV reverse transcriptase (10 U/µl) and incubated at 37^{o}C for 1 hour.
3. The reaction was stopped by the addition of 53 µl of 1:15 dilution of formamide dye.
4. The marker was stored at -20^{o}C.

2.13.2 Fluorescent fingerprinting marker

1. 70 µl $T_{0.1}E$, 10 µL NEB2, 6 µl lambda DNA (500 ng/µl), 6 µl *Bsa*J 1 (2.5 U/µl), 4 µl TaqFS (32 U/µl), 4 µl of 10 mM ddC-ROX were added to a 1.5 ml microfuge tube and incubated for 1 hour at 60^{o}C.
2. 100 µl of 0.3 M sodium acetate (pH 5.2) and 400 µl 96% ethanol were added to the reaction mix and incubated at RT in the dark for 15 minutes, then at -20^{o}C for 20 minutes. The Eppendorf microfuge tube was subjected to centrifugation in a bench top centrifuge at maximum speed for 20 minutes to pellet the DNA.
3. The supernatant was discarded and the DNA pellet dried by tapping the tube gently onto tissue paper. The pellet was washed by adding 200 µl 70% ethanol and spun in a bench top centrifuge at maximum speed for 5 minutes, the supernatant discarded and the pellet dried as described in step 2.
4. The DNA was re-suspended in 120 µl $T_{0.1}E$ and 120 µl blue dextran formamide dye.
5. The marker was stored at -20^{o}C.

2.13.3 For *Hin*d III fingerprinting

1. 19.2 µl $T_{0.1}E$, 1.5 µl Analytical Marker DNA wide range, 0.2 µl Molecular Weight Marker V and 4.2 µl 6x loading dye were added to a 1.5 ml microfuge tube.
2. The marker was stored at -20^{o}C.

2.14 Gel preparation and electrophoresis

2.14.1 Agarose gel preparation and electrophoresis

1. Agarose gels were prepared in 1x TBE (or 1x TAE, for *Hin*d III fingerprinting) containing 250 ng/µl ethidium bromide and the appropriate percentage of agarose according to the size of fragments being separated: 2.5 % agarose gels were used for electrophoresis of fragments below 1 kb; 1.0% agarose gels were used for analysis of larger fragments. Electrophoresis was performed at 50 - 90 V for 15 - 45 minutes depending on the separation required.

2.14.2 Polyacrylamide gel preparation for radioactive fingerprinting

1. 42.0 g of urea were dissolved in 10 ml 10x TBE and 35 ml ddH_20 by warming to 37^oC, and stirring.
2. A large glass plate (back plate – Gibco BRL) was washed on both sides and one side was treated with 2 % dimethyldichlorosilane, and left to dry.
3. A small glass plate (front plate – Gibco BRL) was washed on both sides with detergent and water and one side was treated with freshly prepared bonding solution (3 ml 96 % ethanol, 50 µl 10 % Acetic acid, 5 µl methacryloxypropyl-trimethoxysilane; Sigma-Aldrich) and left to dry.
4. The front and back plates were taped together along three edges (treated sides facing inwards) separated by 4 mm spacers.
5. 10 ml 40% acrylamide, 800 µl 10 % ammonium persulphate and 80 µl TEMED (KODAK) were added to the dissolved urea solution, mixed and poured in between the glass plates using a 50 ml syringe. A 4 mm, 60 well comb (IBI) was placed in the top of the gel (the edge not taped) and the glass plates clamped with bulldog clips. The gel was left to set for up to 3 hours.

2.15 Construction of small insert library

2.15.1 Library preparation

1. 1 µl 4 mg/ml PMSF was added per every 100 µl of flow-sorted chromosome preparations (previously treated with Proteinase K and sarkosyl) for a 40 µg/ml final concentration and incubated at room temperature for 40 minutes.
2. 20 µl 5 M NaCl was added to a concentration of 0.2 M. Two volumes of absolute ethanol were added, and the DNA precipitated overnight at -20^{o}C.
3. DNA was pelleted in a microfuge at maximum speed (14,000 rpm) at RT for 15 minutes.
4. The supernatant was removed with a P1000 Gilson pipette and the remainder removed with Gilson P200 pipette, avoiding the pellet. 1 ml 70% ethanol was added and the eppendorf tube centrifuged again at maximum speed for 7 minutes and the supernatant removed as before. Pellets were air-dried for approximately 10 minutes.
5. 17.0 µl of sterile TE were added directly to the pellet to give approximately 8.0 ng / µl of flow sorted DNA. DNA was mixed by very gentle flicking and allowed to dissolve for at least 2 hours.
6. 2.0 µl 10x buffer 2 (as supplied by New England Biolabs), 1.0 µl HindIII (20 U/µl) and 2.0 µl ddH_20 were added to 15 µl of flow-sorted DNA which was digested for 3 hours at 37^{o}C.
7. The *Hin*d III was inactivated by incubation of the digest at 65^{o}C for 20 minutes.
8. 2.5 µl (15 ng) of digested flow-sorted DNA, 1 µl (5 ng) of phosphatased pBluescript II (SK+) vector DNA (5 ng/µl) (kindly supplied by Dr. Mark T. Ross), 1 µl of 10x ligase buffer (NEB), 0.5 µl of T4 DNA ligase (NEB; 400 U/µl) and 6.5 µl of water (a total reaction volume of 10 µl) were incubated for 16 hours at 16^{o}C
9. The ligase was inactivated by incubation at 65^{o}C for 10 minutes.
10. 10 µl of TE were added and the ligation stored at -70^{o}C.
11. 1 µl of each of the samples were electroporated into *E. coli* XL1-blue electrocompetent cells (see 2.15.2)

2.15.2 Electroporation and library plating.

1. LB agar plates (containing 50 µg/ml ampicillin) were poured and left to dry before 40 µl of Xgal and 4 µl of 200 mg/ml of IPTG were spread on the surface of the plates and left for 2 hours at 37^{o}C to dry.

2. 40 µl of XL1 blue cells (previously thawed at room temperature and stored on ice) and 1 µl of *Hin*d III digested DNA were added to a 0.1 cm cuvette and mixed. After cells were electroporated at 25 µFD, 1.8 KV and 200 Ω, 1ml of LB media was added, mixed then the whole sample added to a 50 ml Falcon tube and incubated in a shaking incubator for 45 minutes. at 37^{o}C.
3. 50 µl were plated on the previously prepared selection agar plates and grown overnight at 37^{o}C and then stored at 4^{o}C for 1 hour to accentuate the blue non-recombinant cells. Recombinant cells were picked into flat bottom microtitre plates containing 150 µl LB broth containing 50 µg/ml ampicillin and grown overnight at 37^{o}C . Glycerol was added to the cultures to a final concentration of 7.5% v/v, aliquotted and then stored at -70 ^{o}C.

2.16 Applications using the polymerase chain reaction

2.16.1 Primer design

Primers were designed manually using the following guidelines:

1. Primer 3 (http://www.sanger.ac.uk/cgi-bin/primer3.cgi) was used to design primers sequences that were 20 bp in length, were possible beginning and ending with a C or G, and with an optimal annealing temperature of 60^{o}C.
2. Where possible sequences were chosen to avoid areas of simple sequence showing non-representative use of the bases and obvious repetitive sequence i.e., runs of single nucleotide (e.g. TTTT) or double nucleotide (CGCGC) motifs.
3. Sequences were chosen to exclude palindromes which will form inhibitory secondary structure, especially at the 3' ends (e.g. GACGTC).
4. As far as possible, sequences were chosen with a GC content of at least 50%.
5. Sequences were chosen to avoid complementarity between pairs of primers, especially at the 3’ end, which could result in primers annealing to each other and forming primer dimers.
6. If possible, sequences were chosen which would generate products of at least 100 bp in length.

216.2 Oligonucleotide preparation

All oligonucleotides used were synthesised in house by David Fraser or supplied as working dilutions from Genset. The concentration of the primer in ng/µl was determined by measuring the absorbance at 260 nm (Abs_{260}) and multiplying this by 33 and any necessary dilution factor.

2.16.3 Amplification of DNA by PCR

1. 1-3 ng/µl of genomic DNA were amplified in a reaction volume of 15 to 50 µl as required. Reactions contained approximately 1.3 µM of each oligonucleotide primer, 67 mM Tris-HCl (pH 8.8), 16.6 mM $(NH_4)_2S0_4$, 6.7 mM $MgCl_2$, 0.5 mM of each deoxyribonucleoside triphosphate (dATP, dCTP, dGTP, dTTP), 1.5 U of Amplitaq™ (Cetus Inc.); 10 mM β-mercaptoethanol and 170 µg/ml of BSA (Sigma Chemical Co., A-4628) were added to the reactions from freshly made stock solutions as the reactions were set up.

2. Unless specified otherwise, cycling conditions were as follows: all reactions were preceeded by an initial denaturing step of 5 minutes at 94°C, followed by 35 cycles of: 94°C for 30 seconds, (primer-specific annealing temperature) for 30 seconds, 60°C for 30 seconds, 72°C for 30; followed by a final extension step of 5 minutes at 72°C. Primer-specific annealing temperatures are given for each primer pair in the text or in Tables 2.3 – 2.9.

3. PCR products were separated on 2.5% agarose minigels as described in Section 2.14.1 and visualised by ethidium bromide staining.

2.17 Radiolabelling of DNA probes

2.17.1 Radiolabelling of PCR products by PCR

PCR products were radiolabelled essentially as described in Bentley et al. (1992).

1. 5 - 10 µl of PCR product were separated on a 2.5% agarose minigel and visualised by ethidium bromide staining.
2. The gel was rinsed in deionised water to remove excess buffer. The desired band was excised from the gel and placed in 100 µl of $T_{0.1}E$ at 4°C overnight.

3. 2 µl of the $T_{0.1}E$ were used as template in the PCR-labelling reaction containing 40 ng of each primer, 1 µl of 10x PCR buffer, 0.5 µl of [α-^{32}P]-dCTP (3,000 Ci/mmol), 0.5 U of *Taq* polymerase (Cetus) and 0.375 mM each of dATP, dTTP and dGTP. Reactions were performed in a 0.5 ml microfuge tube and overlaid with mineral oil (Sigma) in a DNA thermal cycler (Perkin Elmer, USA).
4. PCR cycling conditions were as follows: 94°C for 5 minutes; followed by 20 cycles of: 93°C for 30 seconds, 55°C for 30 seconds, and 72°C for 30 seconds; followed by 72°C for 5 minutes.
5. Probes were pre-reassociated (as described in Section 2.17.2) prior to use if necessary. All probes were boiled for 5 minutes and snap-chilled on ice prior to use.

2.17.2 Pre-reassociation of radiolabelled probes

1. Radiolabelled probe was mixed with 125 µl of 20x SSC and 250 µl of the sheared 10 mg/ml human placental DNA (Sigma) in a final volume of 500 µl.
2. The mix was boiled for 5 minutes, snap-chilled in ice-water, and then added directly to the hybridisation reaction.

2.18 Hybridisation of radiolabelled DNA probes

2.18.1 Hybridisation of DNA probes derived from STSs

1. Filters were prehybridised flat in sandwich boxes for 3 hours in 10-25 ml of hybridisation buffer at 65°C with gentle shaking.
2. Radiolabelled probe was added and hybridised to the filters.
3. Filters were washed twice at RT in 2x SSC for 5 minutes, twice at 65°C in 0.5 x SSC, 1.0% Sarkosyl for 30 minutes. Filters were rinsed at RT in 0.2x SSC prior to draining the excess liquid, wrapping in Saran Wrap (Dow Chemical Co.) and exposing to autoradiograph film.

2.18.2 Stripping radiolabelled probes from hybridisation filters

1. Filters were washed in 0.4 M NaOH for 30 minutes at 42°C followed by 30 minutes in 0.2 M Tris-HCl (pH 7.4), 0.1x SSC, and 1.0% w/v Sarkosyl at 42°C with gentle shaking. Successful removal of radiolabelled probe was assessed by autoradiography.

2.19 Restriction endonuclease digestion

1. 4 µl (approximately 150 ng) of prepared cosmid DNA (described in Section 2.11.1) were digested with *Hin*d III using commercial buffers according to manufacturers' instructions in a final volume of 10 µl.
2. 5 µl of each digest were checked for complete digestion by electrophoresis on a 1% agarose minigel and visualised by ethidium bromide staining.

2.20 Clone library screening

2.20.1 cDNA library screening by PCR

The strategy used to screen the cDNA libraries by PCR is illustrated in Figure 2.1.

1. Nine different cDNA libraries were subdivided into 25 subpools of 20,000 clones, which were then combined to produce 5 superpools of 100,000 clones by J. Bye and S. Rhodes. Details of the cDNA libraries are given in Table 2.2.
2. Aliquots of the superpools of each library were arranged in a microtitre plate to facilitate subsequent manipulations with a multi-channel pipetting device.
3. In the primary screen, 5 µl of each superpool were used as template in a 15 µl final volume PCR using buffer and PCR conditions as described in Section 2.16.3.
4. PCR products were loaded on 20 cm x 20 cm 2.5% agarose horizontal slab gels using an 8-way multi-channel pipetting device, separated by electrophoresis and visualised by ethidium bromide staining.
5. In the secondary screen, 5 µl of each of the 5 subpools of 20,000 clones corresponding to the superpool that were positive in the first round, were screened by PCR with the same primer pair as used in step 2. PCR products were separated by electrophoresis through 2.5% agarose minigels and visualised by ethidium bromide staining.

2.20.2 Vectorette PCR on cDNA libraries

1. Vectorette PCR was performed on the superpools of the cDNA libraries (figure 2.1). PCR was performed using 5 µl of the diluted superpools as template in a 15 µl final volume using buffer conditions as described in Section 2.16.3. Primer combinations were as follows: 224 and specific primer A, 224 and specific primer B.
2. PCR was performed in a DNA thermocyle (Omingene) using hot start. Cycling conditions were as follows: an initial denaturation step of 5 minutes at 95^{o}C, followed by 17 cycles of: 94^{o}C for 5 seconds, 65^{o}C for 30 seconds and 72^{o}C for 3 minutes, followed by 18 cycles of: 94^{o}C for 5 seconds, 65^{o}C for 30 seconds and 72^{o}C for 3 minutes, followed by 72^{o}C for 5 minutes. The PCR was paused after 4 minutes of the initial denaturation and 2 µl of Taq premix (containing 0.12 µl Amplitaq, 0.12 µl TaqExtender, 0.12 µl Perfect Match, 0.5 µl 40% sucrose + cresol red, 1.14 µl $T_{0.1}E$) were added to each reaction (pipetting underneath the oil).
3. Products were separated by electrophoresis through 2.5% agarose gels and visualised by ethidium bromide staining. Products were gel-purified using gel extraction kits and following manufacturer's instructions from either GeneClean^{TM} or Qiagen^{TM} prior to sequencing directly.

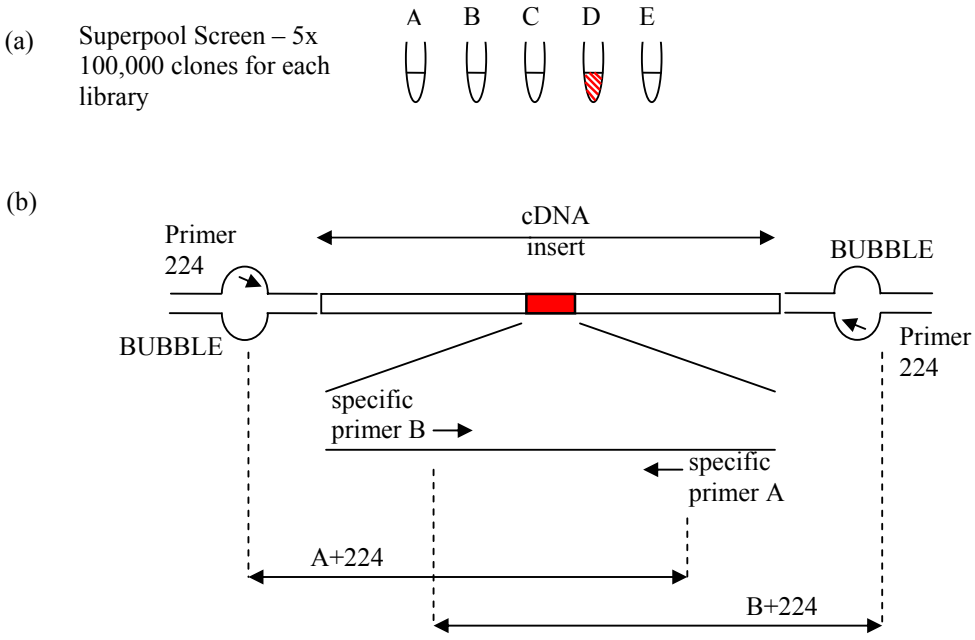


Figure 2.1: Strategy for vectorette PCR screening of cDNA libraries. a) Superpools representing 100,000 clones (A-E) were screened by PCR and positive super pools recorded (e.g. D, shown in red). (b)Rescue of the insert of the cDNA of interest by vectorette PCR. Insert of cDNA is shown as white rectangle; ligated 'bubble' is shown in yellow. Original position of primers for pool screening is indicated by a red rectangle. A combination of the 'bubble' primer (224) and insert specific primers (e.g A and B) were to generate two products (e.g. 224+A, 224+B) representing the entire insert of the cDNA of interest.

2.21 Exon Amplification

1. Exon-specific primers were diluted from 1 mg / ml to a stock 100 ng / µl and a working dilution of 8 ng / µl in $T_{0.1}E$.
2. 1 µl 10 X NEB buffer, 1 µl dNTPs (5mM), 0.33 µl BSA, 0.14 µl ßME, 0.12 Amplitaq and 1.41 µl of T0.1E were added to 1 µl of individual CEPH DNA (5 ng / µl) aliquotted in a 384-well PCR plate.
3. Following PCR samples were separated on an agarose gel as described in 2.14.3.
4. Residual primers and dNTPs were removed by the addition of 2 µl of a premix containing 0.66 µl of shrimp alkaline phosphatase, 0.66 µl of reaction buffer, 0.66 µl of double distilled H_2O and 0.066 µl of exonuclease I. The 384-well microtitre plate

containing the exon amplified PCR products was incubated at 37°C for 30 minutes, and then 80°C for 15 minutes.

5. Exon specific PCR products were then sequenced by the Research and Development Group or The Sanger Institute Sequencing facility.

2.22 Mapping and sequence analysis software and databases

2.22.1 IMAGE

Gel images from radioactive, fluorescent and *Hin*d III gels were processed using IMAGE. Band patterns for each lane were extracted and the data manually edited and normalised by marker locking prior to data entry into another program, fingerprinted contigs (FPC) (Soderlund *et al*., 1997), for analysis. Within IMAGE several procedures were run on each gel in turn:

Lane tracking – a grid was superimposed and manually edited on each gel image to ensure it exactly matched the lanes on the gel.

Band calling – an analysis module then identified labelled restriction digest fragments within each lane. Manual editing ensured the correct bands are chosen.

Marker locking – A set of DNA fragments of known length or migration distance was loaded as a marker to facilitate normalisation of restriction fragments and permit accurate clone to clone comparisons to be made (see Section 2.13 for specific marker patterns for each method of fingerprinting used). Manual editing of each marker lane ensured the observed experimental marker matched the previously established standard marker pattern.

Normalisation – once marker lanes were locked onto standard marker positions, the bands within the sample lanes were normalised to account for differences in gel to gel run conditions. IMAGE finally generated a 'Bands' file for each gel containing normalised migration distances for all selected bands in each clone lane.

2.22.2 FPC

FPC, which was used for all contig construction described in this thesis, utilises the bands file output from IMAGE as the digitised set restriction fragments for each clones. Fingerprint patterns for each clone were compared to those of all clones in the database. The relationship between two clones was reported as a probability of coincidence, i.e. the probability that bands

in common between two clones overlap by chance. Two variables can be set to filter the reported overlaps:

Cut off – a match between two clones will only be reported if the probability of coincidence is less than or equal to the cut off. When analysing matches between larger insert PAC and BAC clones the tolerance was 1e –08.

Tolerance – two bands of equal size are considered as their migration distances differ by less than tolerance. For the analysis carried out in this thesis the tolerance was set to 7.

Overlapping clones were identified by automatic analysis and contigs were constructed manually using the available editing tools within FPC. Clones were overlapped by iterative pair-wise analysis by determining the number of bands in common between clones. Subsequent clones were added positioned within the initial clone assembly based on the number of bands they shared with the existing clones in the contig. Marker data was imported from 1ace and integrated into the FPC contigs to identify clone overlaps that could not be identified using pre-determined analysis parameters. A minimum set of clones for sequencing was chosen based on a combination of shared bands and shared marker data.

Contig Sizing –one unit in the contig display represents one fingerprint band which permits for estimates to be made for contig sizes. For each fingerprinting method, a kilobase/band figure was derived. For radioactive cosmid fingerprinting, one band was the equivalent of 3.3 kb, based on the fact the average number of fingerprint bands for bacterial clones in comparison to finished sequence. For fluorescent fingerprinting of PACs and BACs an average figure for each clone type, based on the number of bands observed in clones that had generated finished sequence, was 4.0 kb/band. *For Hin*d III *fingerprinting of PACs and BACs*, 4.4 kb/band was used based on the length of finished genomic sequence.

2.22.3 1ace

All mapping and sequencing data generated in this thesis were stored in 1ace, a chromosome-specific implementation of ACeDB (Richard Durbin and Jean Thierry-Mieg, 1991).

ACeDB uses a flat file format with data being accessed using a series of windows to represent various data types. All windows are linked in a hypertext fashion, so that clicking on an object will display further information about that object. For example, clicking on a region of a chromosome map will highlight landmarks mapping to that part of the chromosome; clicking on a landmark will display information about that landmark including landmark-clone associations, etc.

All PAC, BAC and cosmid library filters and polygrids are represented graphically in 1ace and data were entered directly. Data were then saved in the database establishing landmark-to-clone associations which can be displayed as text windows relating to either the landmark or the clone. Data can also be entered via text windows or via an internal web page. PCR library pool screening and colony PCR results were entered via the text windows.

In addition to the data generated by the 1 chromosome mapping group, 1ace also contains displays published 1 chromosome maps, which have been used as part of the project. This greatly facilitates integration of maps from different sources. Genomic sequence data is also displayed in ACeDB along with the collated results from the computational sequence analysis performed by the Sanger Institute Human Sequence Analysis Group. 1ace can be accessed by following the instructions at http://www.sanger.ac.uk/HGP/Chr1.

2.22.4 Blixem

Individual matches identified as a result of similarity searches using the BLAST algorithm, or matches between sequences of cDNA clones or PCR products amplified from genomic DNA generated as part of the project, were viewed in more detail using Blixem. Blixem, (Blast matches In an X-windows Embedded Multiple alignment) is an interactive browser of pairwise Blast matches displayed as a multiple alignment. Either protein or DNA matches can be viewed in this way at either the amino acid or nucleotide level. Blixem contains two main displays: the bottom display panel shows the actual alignment of the matches to the genomic DNA sequence, and the top display shows the relative position of the sequence being viewed within the context of the larger region of genomic DNA. A program “pfetch” retrieves the record from an external database (*e.g.* EMBL, SWISSPROT).

2.22.5 RepeatMasker

Human repeat sequences were masked using RepeatMasker, a program that screens DNA sequence for interspersed repeats and low complexity DNA sequence (Smit, AFA & Green, P RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html). The output of the program is a detailed annotation of the repeats that are present in the query sequence. Sequence comparisons are performed by the program cross_match, an implementation of the Smith-Waterman-Gotoh algorithm developed by P. Green. The interspersed repeat databases screened by RepeatMasker are based on the repeat databases (Repbase Update) copyrighted by the Genetic Information Research Institute.

Chapter 3

Using large insert clones to construct contigs:

The development of fluorescent fingerprinting

3.1 Introduction

The construction of sequence-ready maps of overlapping sets of bacterial clones has been one of the central components of large-scale genomic sequencing. Map construction is reliant upon techniques that are able to accurately assemble bacterial clone contigs at a depth that will facilitate the identification of a set of minimally overlapping clones. Restriction digest fingerprinting has proved to be a robust method that can be used as a means of producing accurate fragment patterns to allow the overlapping relationship between genomic clones to be determined accurately, as well being a technique that is amenable to increases in scale and speed for generation of clone data.

The approach of overlapping sets of random cloned DNA was first used to assemble contigs of cosmids or bacteriophage lambda clones of the *Caenorhabditis elegans* (Coulson *et al*., 1986) and *Saccharomyces cerevisae* (Olson *et al*., 1986, Riles *et al*., 1993) genomes, respectively. The basic premise of these projects was to generate overlaps between clones that share a statistically significant number of restriction fragments, where the shared fragments representing the region of overlap. Restriction digest fingerprinting has also been applied to the construction of sequence-ready maps within regions of the human genome (Carrano *et al*., 1989, Heding *et al*., 1992, Taylor *et al*., 1996). The inception of large insert bacterial clones has greatly facilitated the construction of maps of larger chromosomes and genomes. Large insert clones BACs and PACs are maintained in single to low copy number, (like fosmids), which means that a smaller percentage of the clones are likely to be unstable or to lose any cloned material. Additionally, a smaller

fraction of the genome is likely to be unstable in the host *E. coli*. PAC and BAC clones can be manipulated in the same way as cosmids and fosmids and, because of their larger insert size, generate genomic coverage more rapidly. As sequencing projects increase in scale (e.g. from *C. elegans* to human) there is a need to adapt existing fingerprinting techniques to the analyses of genome scale clone resources, to improve upon through-put and to incorporate increased levels of safety and automation.

This chapter outlines the evaluation of the first application of restriction digest fingerprinting to large insert clones by the construction of a 1.4 Mb contig across a region of chromosome 13q12 that includes the breast cancer susceptibility gene *BRCA2*. The chapter also details the development and testing of fluorescent fingerprinting, describes an assessment of the technique in comparison to other fingerprinting methods and discusses its application to the large scale generation of sequence-ready bacterial clone maps.

3.2 Large insert clones

3.2.1 Application of restriction enzyme fingerprinting to large insert clones

Linkage analysis in a series of breast cancer families localised the putative BRCA2 gene to a 3cM interval on 13q12-13. The work described in this chapter therefore contributed to the search for the BRCA2 gene, as well as providing a first opportunity to test the suitability of large insert bacterial clones (specifically PACs) as a resource for map assembly. The clones were identified from a whole genomic library by hybridising the *Alu*-PCR products of five yeast artificial chromosomes (YACs) (Burke *et al*., 1987) from a suggested CEPH YAC tiling path (Cohen *et al*., 1993) (figure 3.1a) and nine genetic markers (from the region) to a total genomic PAC library (see chapter 2.8.2) (initial library screening was done in collaboration with Richard Wooster). All clones were restriction digest fingerprinted (Coulson *et al*., 1986) using the established radioactive *Hin*d III/*Sau*3A I procedure and assembled into contigs using image editing (http://www.sanger.ac.uk/Software/Image) and data manipulation and analysis programs (Soderlund *et al*., 1997). Two separate walking procedures were employed to confirm overlaps within contigs and to bridge the remaining five gaps (in conjunction with Richard Wooster). Clones at the ends of contigs were used either as a template for SP6 or T7 primers vector specific primers (Ragoussis and Olavesen, 1997) or for the production of inter-*Alu* PCR fragments which were hybridised back to the library and to clones previously identified. The identification of new overlaps and the extent of overlap between clones by restriction fingerprinting was confirmed by the hybridisation results. Figure 3.1

displays the complete contig of the critical region containing 49 PACs and the SP6/T7 and *Alu* -PCR hybridisation probes and results. A minimum set of fourteen clones was chosen from the contig, representing the central region D13S260 to sls234, based on hybridisation and fingerprinting overlaps. The absolute size of the region represented by this minimum clone set was later determined to be 1.4 Mb based on complete genomic sequence data

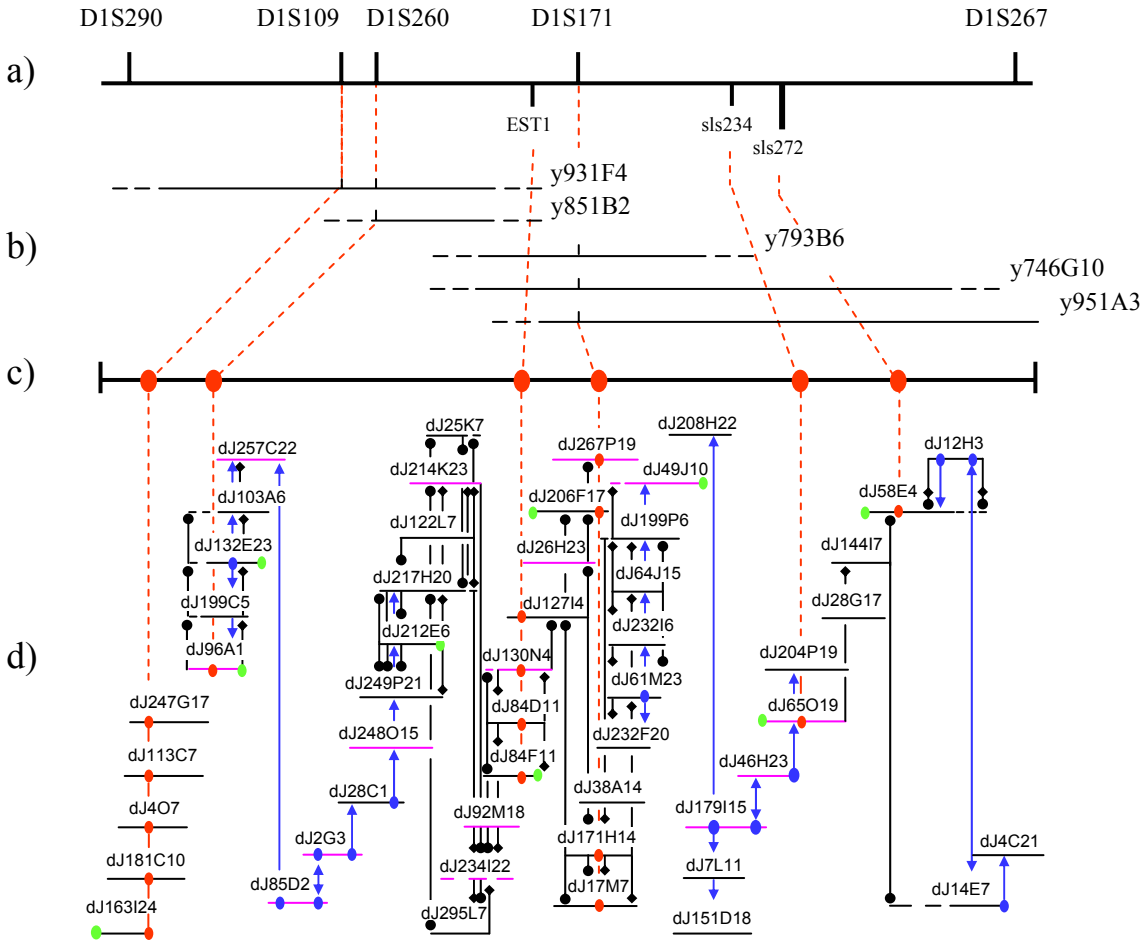


Figure 3.1: A representation of the mapping of BRCA2 region.

a) Genetic markers used in the identification of the putative region; b) YAC minimum set; c) Extent of PAC contig coverage; d) Fingerprint contig across the BRCA2 region; clones isolated by ● *Alu*-PCR hybridisation; ● clones fluorescent *in situ* hybridised; ⬇ T7 🌡 SP6 end probes; ▬ sequencing minimum set (234I22 sequenced as a PCR product)

3.2.2 Validation of PAC inserts

Given that the PACs were a new cloning system it was important to examine the integrity of the large insert PAC clones. Contiguous sequence from the 1.4 Mb contig provided an opportunity to compare the fingerprint patterns obtained empirically with those expected from the finished sequence. An additional check was possible by cross-comparison of fingerprints of regions in common between overlapping clones Four hundred and ninety two ordered *Hin*d III fragments were produced from a 'virtual' digestion of the 1.4 Mb of submitted sequence. The virtual fragments were compared to corresponding bands of the *Hin*d III fingerprints from 37 clones completely overlapping the chosen minimum set (figure 3.2, for example). The resolution of agarose gels limited the accurate sizing of bands to those over 500 bases (figure 3.2b). Occasionally bands of similar size, migrating in the same position of the gel, were not distinguishable. Figure 3.3 displays the band content of the *BRCA2* clones from the contig as compared to the virtual bands derived from the sequence. Bands of less than 500 bp, for which the size could not be accurately determined, accounted for 17% (82/492) of sequence fragments whilst missing bands (see

figure 3.3, red boxes) accounted for 2% (22/1091) of all fingerprint bands. These missing bands do not indicate the presence of large rearrangements within the PACs as no additional bands that could have resulted from the rearrangement were observed in overlapping clones. The absence of bands at the same location of overlapping clones (blue arrow in figure 3.3) may indicate a genuine restriction fragment length polymorphism within the PAC library. The missing bands within 199C5, 257C22 and 103A6 were large enough to be accurately sized on the agarose gel and were contained within overlapping clones.

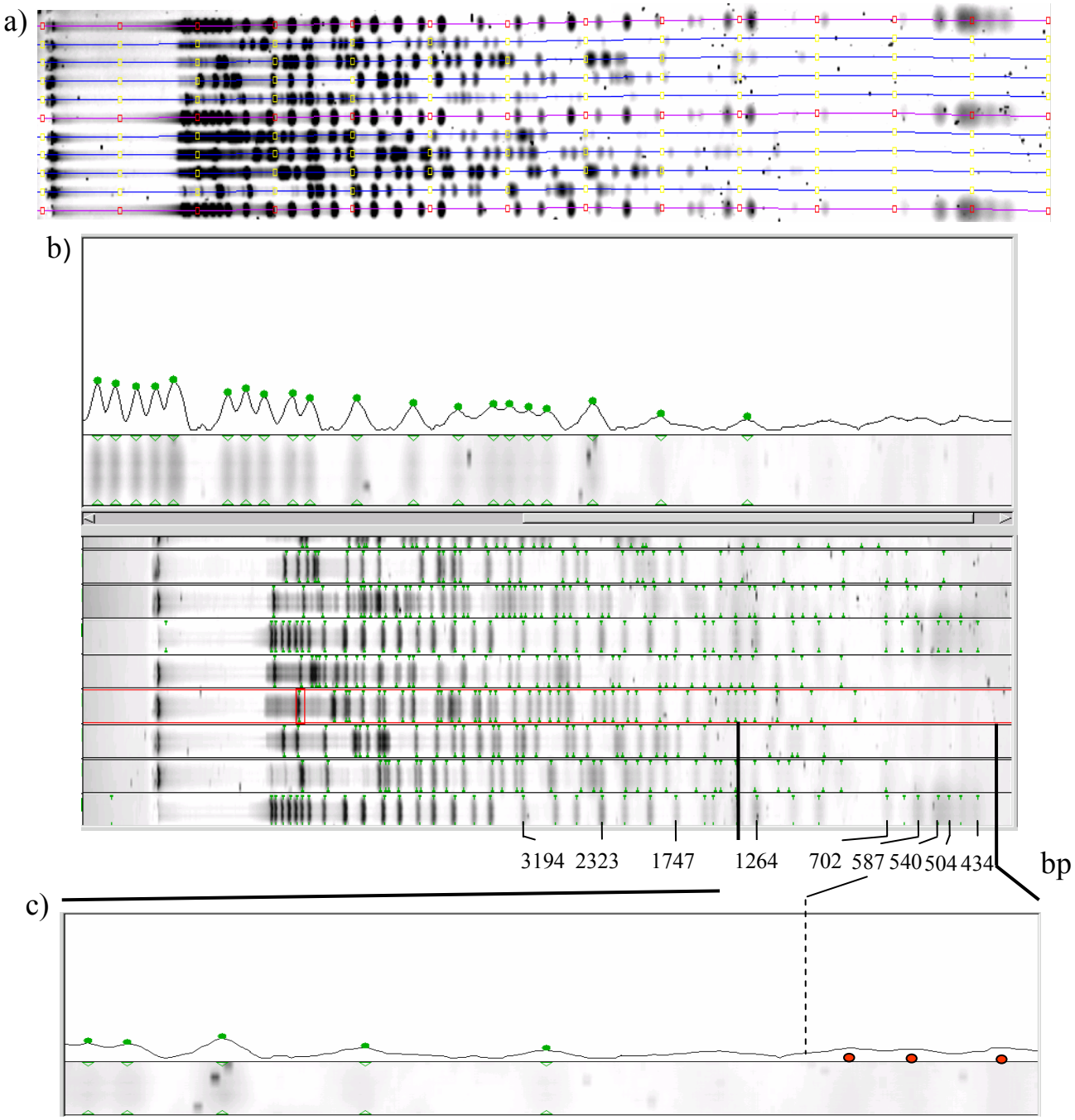


Figure 3.2: An agarose gel fingerprint of large insert bacterial clones in IMAGE. a) depiction of lane lock of 8 clones (yellow boxes) and 3 marker lanes (red boxes). b) data entry phase with edited bands indicated by green pips, also represented are marker lanes

with a sub-set of sized fragments in base pairs. c) zoomed region of clone 3 indicating restriction fragments below 500 bp in size (red dots).

Figure 3.3: A comparison of *Hin*d III fingerprint fragments and genomic sequence for the BRCA2 contig, top left to bottom right. Black boxes indicate when a band contained within the genomic sequence is accounted for by a fingerprint band, grey boxes where a band is ambiguous on the agarose gel, clear boxes where a band derived from the sequence

falls outside of the resolution of the agarose gel, red boxes where the band was absent from the restriction fingerprint. The blue arrow indicates a possible restriction fragment length polymorphism.

3.3 Fluorescent fingerprinting

The development of a hierarchical clone by clone approach by the Human Genome Project (HGP) for the large-scale generation of human genomic sequence (see chapter 4) required the development of a restriction fingerprinting technique that was high throughput, accurate and safe. For these reasons the application of fluorescent sequencing technologies to restriction fingerprinting was investigated.

The first experiments utilized fluorescent dideoxy molecules, used in terminator sequencing, to end label fingerprint fragments. The aim was to emulate the dual restriction digest and end labelling technique developed by Coulson *et al.,* (1986) for the construction of a physical map of the C. elegans genome. In this technique, a radioactive reporter molecule, [α -^{32}P] deoxy-adenosine-tri-phosphate (dATP), is complementarily conjugated via a DNA polymerase (AMV RT) with the first base in the 5' overhang of *Hin*d III restriction digest. A dideoxy guanine-tri-phosphate (ddGTP) is then incorporated at the second base to ensure irreversible completion of the labelling reaction, prior to a second restriction digest with *Sau*3A I. Fingerprint fragments were resolved on a denaturing polyacrylamide gel and migration values of the fragments determined by comparison to a known marker run interposed between fingerprints on the gel. The marker is produced by digesting lambda DNA with *Sau*3A I, filling the first position of the 5' overhang with

dGTP, labelling with radioactive [α -^{33}S] dATP and utilizing ddTTP to prevent exonuclease removal of the reporter molecule.

3.3.1 Fluorescent labelling of cosmid DNA

The first fluorescent fingerprint experiment was designed to mimic the radioactive protocol. DNA minipreps of two chromosome 22 cosmid clones (see chapter 2.4.3) were digested with *Hin*dIII and Sequenase used to incorporate the unlabelled dATP and fluorophore conjugated ddGTP. A phenol/chloroform extraction step was introduced to remove latent fluorescent dye (soluble in phenol) from the labelling mix prior to gel running and data capture on an ABI 373 using ABI Prism sequencing software. The results from this initial experiment showed a stochastic relationship between DNA concentration and signal strength of the fingerprint fragment peaks. Though the data collected from the gel run identified only smaller restriction fragments it determined that fragments had been successfully labelled and identified a noticeable salt front and a large peak corresponding with unincorporated dye (figure 3.4).

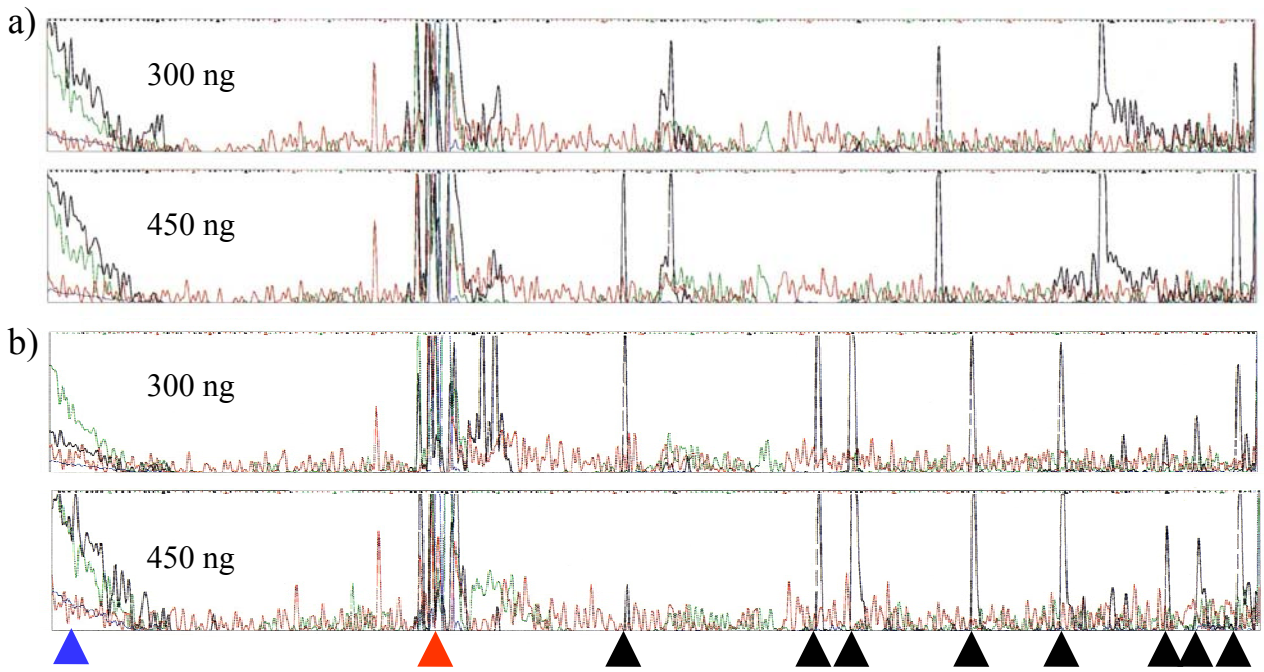


Figure 3.4: The result of the first fluorescent fingerprinting experiment. Two cosmid clones (a and b) of known concentration (shown), were digested with *Hin*dIII and *Sau*3A, labelled with ddGTP dye-terminator run on an ABI 373 sequencing machine. Labelled fragments of different DNA concentrations were collected using Prism sequencing software. The blue arrow denotes salt/buffer front, the red arrow denotes latent dye-terminator present in fingerprint reaction and black arrows identify labelled restriction fragments.

Having proven that it was possible to generate and detect fluorescently labelled fingerprint fragments, the next aim was to repeat the cosmid labelling experiment and to replicate the labelling of the radioactive lambda marker which would enable a comparison to the radioactive cosmid fingerprint to be made and therefore determine if all cosmid restriction

fragments were being collected (figure 3.5). Lambda DNA was digested with *Sau*3A I and labelled by filling the 5' overhangs with dATP, dGTP and fluorescent dideoxy thymine-tri-phosphate (ddTTP) using Sequenase. This protocol necessitated incorporation of the labelled dNTP at the third position of the 5' overhang which would be spectrally distinct to nucleotides used to label cosmids in the first and second positions.

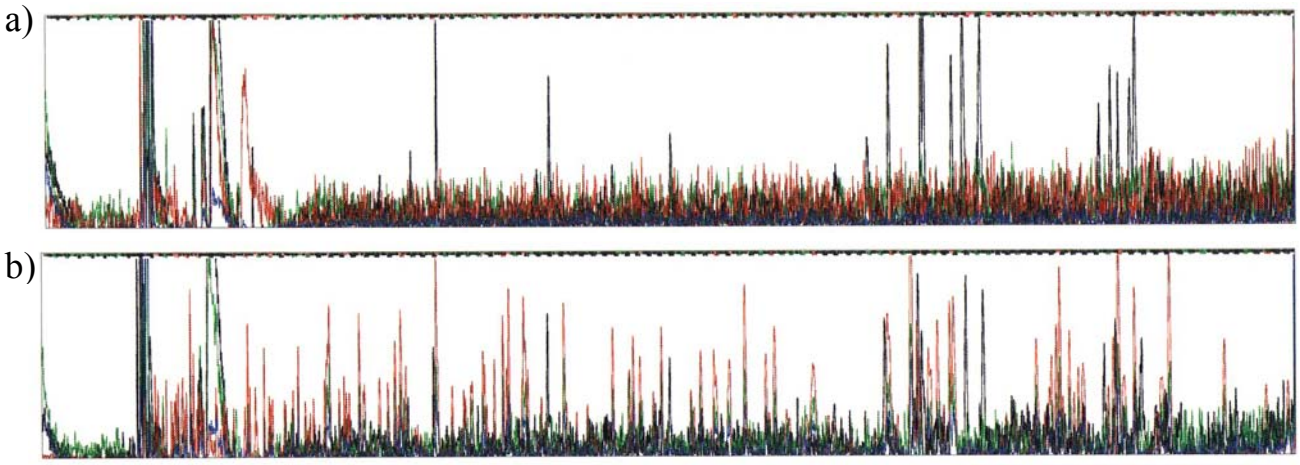


Figure 3.5: Fluorescent fingerprint data collection using an extended run time. a) fragments were labelled as previously but the run time was extended. b) Replications of the labelling of radioactive lambda marker using a spectrally distinct fluorophore to label a Sau3A I digest of lambda DNA.

3.3.2 Residual dye removal

A transition was made to ABI Genescan software which would allow us to separate fingerprint and marker traces from within a single lane on the gel. The presence of latent dye remaining in the fingerprint reaction could affect data production. Small fingerprint

fragments migrating adjacent to the dye front may be masked by dye signal or the dye may affect the resolution of separating fragments. Comparisons were made between techniques to remove the unincorporated dye, namely phenol/chloroform extraction, molecular sieving using Sepharose beads (figure 3.6a and b) and ammonium acetate/ethanol precipitation (figure 3.6c). It was noted that all dye removal techniques resulted in a reduction of fragment signal. Phenol/chloroform extraction resulted in the largest loss of signal (60-70%), ammonium acetate (30-40%) and P10 Sepharose beads (20-30%). Though the P10 beads resulted in the least amount of signal loss, ammonium acetate extraction could more easily be incorporated into the fingerprinting procedure without the introduction of an additional extraction step.

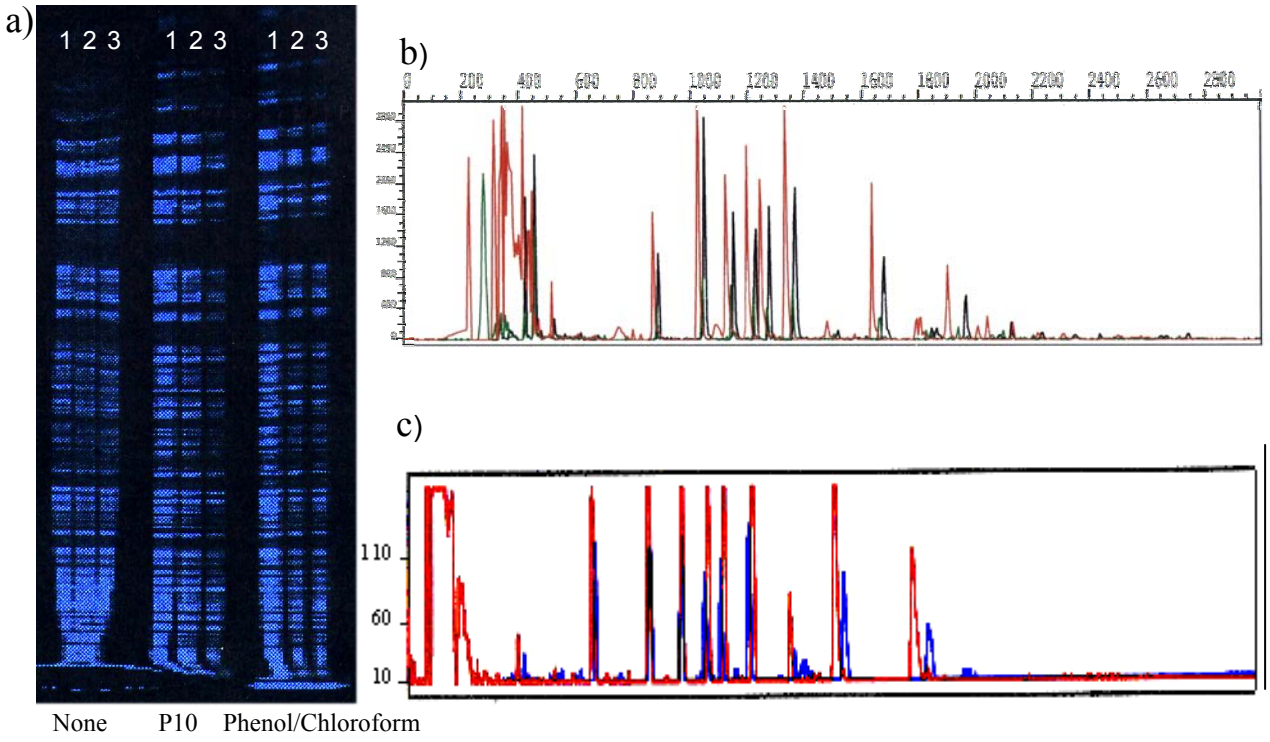


Figure 3.6: A comparison of labelled fragments when investigating removal. a) *Sau*3A I digest of 100 ng (1), 200 ng (2), and 500 ng (3) of lambda DNA. DNA digested with Sau3A I was labelled with ddTTP dye-terminator and data displayed and analysed using Genscan software. b) signal intensities of 500ng sample tracks displaying differences between dye removal techniques. Red traces correspond to fingerprint reaction with no dye removal, black traces using P10 Sepharose beads and green traces phenol/chloroform extraction. c) a comparison of ammonium acetate dye extraction (blue) and fingerprint reaction from which no dye was removed (red).

3.3.3 First Position Labelling

Previously, it was shown that fluorophore labelled dNTPs could be incorporated in positions 1 and 2 of *Sau*3A I digested lambda and position 2 of a *Hin*d III restriction digests of cosmid DNA (figure 3.7a). However, incorporation of a dideoxy-nucleotide reporter molecule in the first position of bacterial clone fingerprints would negate the effect of poor fragment labelling through incomplete dinucleotide incorporation which would prevent efficient labelling with a dideoxy-nucleotide in 2nd or 3rd positions. To test the efficiency of first position labelling, lambda DNA at different concentrations was digested with *Hin*d III and labelled with fluorescent ddATP, figure 3.7. The labelling of restriction fragments with a fluorophore conjugated ddA in the first position of *Hin*d III digest, without the requirement of deoxy-nucleotides, means that DNA could be digested in parallel with *Sau*3A I without these fragments being labelled (figure 3.8a). Results indicated that labelling of relatively low concentrations of DNA with very low background could be achieved using first position labelling. Custom synthesis of ddATP spectrally distinct derivatives, on which one of three differently fluorescing fluorophores were attached (figure 3.7c), permitted multiplexing of three different fingerprint samples within one lane on a gel (yellow, blue and green). Identical sizes of the conjugated fluorophores meant no correction was required during data collection to permit fragment comparison. Lambda DNA digested with *Hin*d III/*Sau*3A I was again used as the control for testing the efficiency of the three ddATP fluorophores.

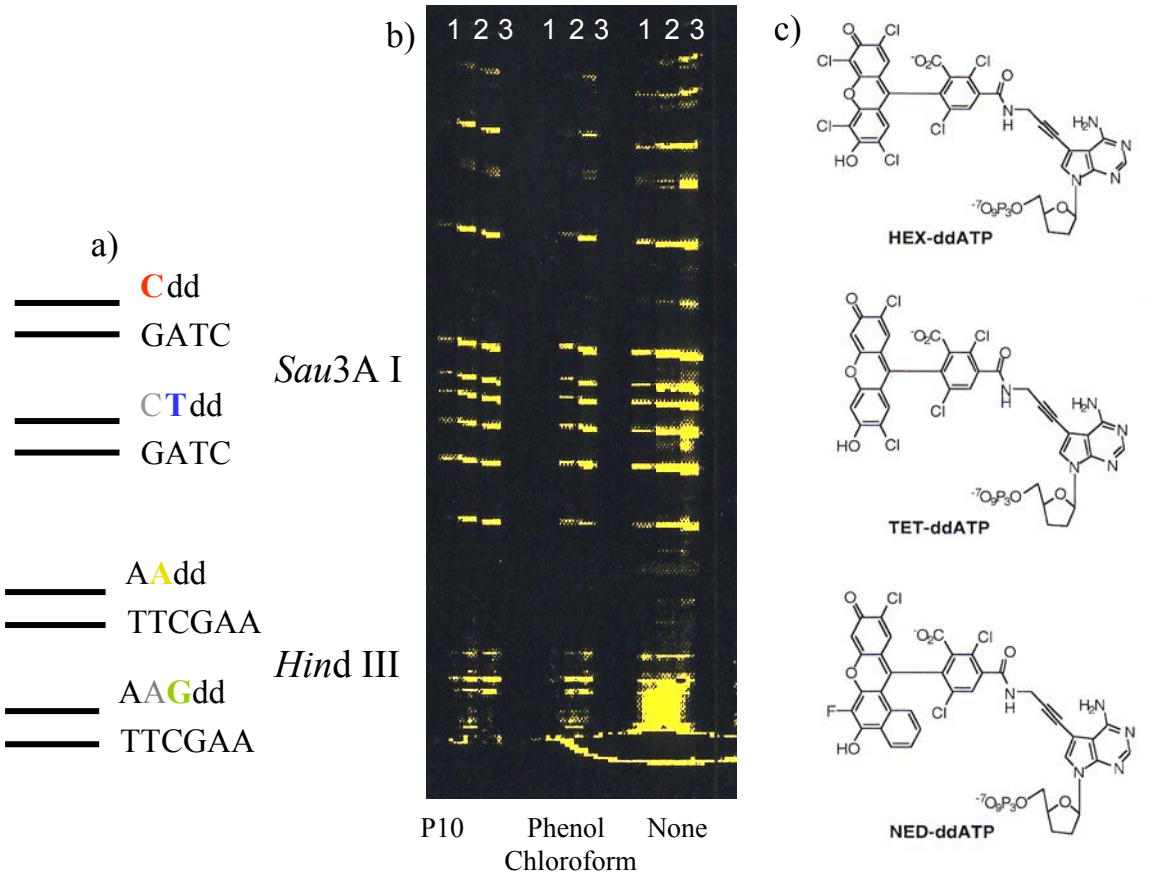


Figure 3.7: Labelling with spectrally distinct fluorophores. a) Labelled nucleotides of spectrally distinct fluorophores attached to the 5' over-hangs of *Sau*3A I and *Hin*d III restriction fragments in positions 1, 2 and 3. Also shown are unlabelled nucleotides (grey) b) First position labelling of lambda DNA. 100ng (1), 200ng (2), and 500ng (3) of DNA are digested with *Hin*d III, labelled using ddATP dye-terminators dye prior to subsequent digestion with *Sau*3A I. Signal strength of labelled fragments is compared between dye extraction techniques, P10 beads and phenol/chloroform. P10 dye extraction of 100 ng of lambda DNA give clear signal above background indicating that first position labelling provides a clear alternative to using dye-terminators and dNTPs for fill-in labelling. c) Spectrally distinct dye-terminators molecules which were custom conjugated onto ddATP and used for separate labelling and multiplexing of fingerprint samples. d) Titration of ddATP dye-terminators using chromosome 22 cosmid DNA.

3.3.4 One-step reaction

The next modification to be made to the fingerprinting protocol was to initiate a one step dual enzyme digest and labelling reaction based on the protocol published by Tang *et al*., (1996). This adaptation would reduce the time required for a fingerprint reaction from $3^{1/2}$ hours to1 hour prior to latent dye extraction. Initial results looked very promising with only a small decrease in the signal of the one-step reaction compared to that of the two-step (figure 3.8). The *Hin*d III/*Sau*3A I combination used to generate fluorescent fingerprints were suitable for the one-step reaction as digest with *Hin*d III results in a T in the first position available to incorporate a fluorescent ddATP whilst a *Sau*3A I digest would only incorporate a ddATP in the second position of in the presence of a dGTP Fingerprints were generated in the presence and absence of unlabelled ddGTP in the reaction mix to determine whether *Sau*3A I fragments were being labelled in the one-step reaction in the presence of residual dGTP (figure 3.9a) or misincorporation of ddATP at first position. No discernable partial or background labelling was noted in either experiment.

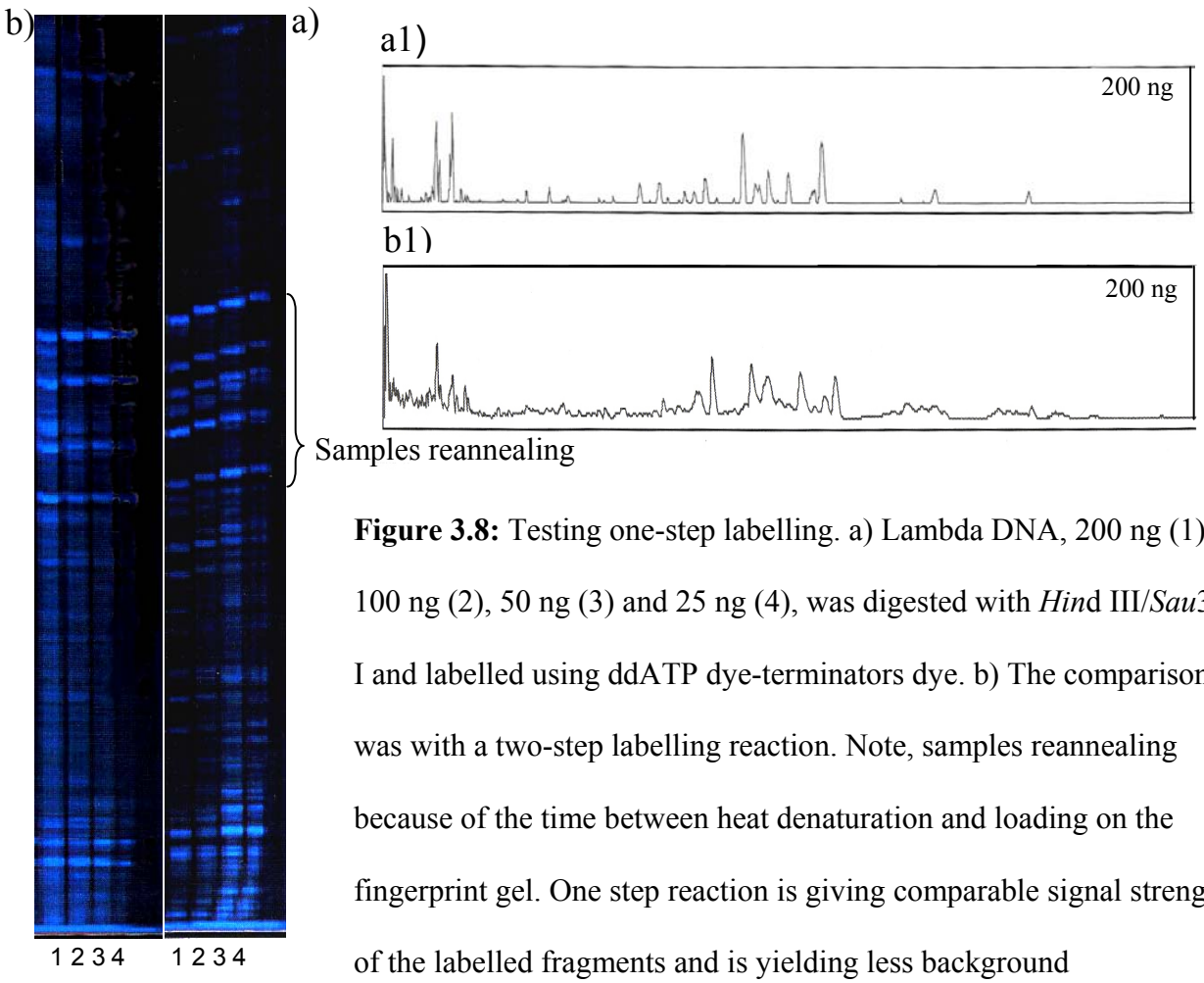


Figure 3.8: Testing one-step labelling. a) Lambda DNA, 200 ng (1), 100 ng (2), 50 ng (3) and 25 ng (4), was digested with *Hin*d III/*Sau*3A I and labelled using ddATP dye-terminators dye. b) The comparison was with a two-step labelling reaction. Note, samples reannealing because of the time between heat denaturation and loading on the fingerprint gel. One step reaction is giving comparable signal strength of the labelled fragments and is yielding less background

3.3.5 DNA prep modifications

The protocol used for isolating cloned bacterial DNA for radioactive and initial fluorescent fingerprinting experiments was based upon the microprep procedure published by Gibson and Sulston (1987). Modifications to this procedure were made to improve DNA yield by utilizing Corning 2 μm filter bottom plates to remove the supernatant from large molecular weight protein precipitated by alkaline-lysis. Adding RNAse to T0.1E before final DNA resuspension replaced the need for RNA removal by lithium chrloride precipitation and in doing so removed an additional DNA precipitation step (figure 3.9b). A moderate increase in signal was observed from the protocol using DNA extracted by the filter prep method. This is presumed to result from a higher DNA yield of the F prep compared to the M prep, attributable to the increased volume of supernatant transferred during alkaline lysis and to the reduction in the number of precipitation steps required before final DNA isolation.

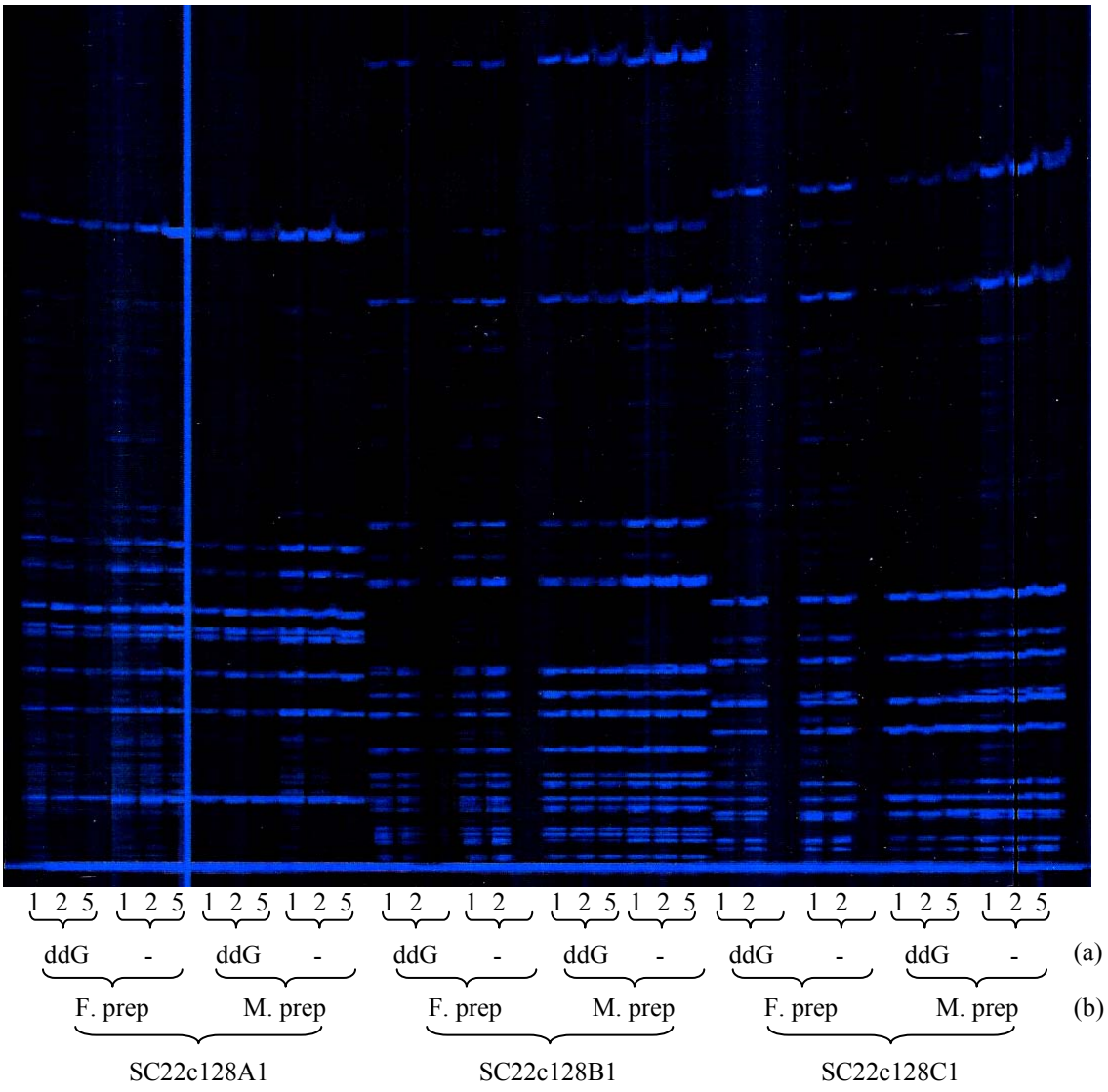


Figure 3.9: Testing one-step labelling and DNA prep protocols. a) Labelling of cosmid DNA (SC22c128A1, SC22c128B1, SC22c128C1) in a one step fingerprinting reaction in the presence and absence of unlabelled 1mM ddG. 1 µl, 2 µl and 5 µl of fingerprint reaction were loaded per lane. No discernable difference was observed in the signal strength or level of background labelling (*Sau*3A I labelling) between one-step and two-step labelling reactions. b) A comparison of filterprep (F) and microprep (M) protocols for the isolation of cosmid DNA samples (see methods). A moderate increase in signal strength above background was observed in the filterprep samples.

3.3.6 New size standard

The separation of radioactive fingerprints is achieved by running samples and lambda *Sau*3A I marker on a 40cm denaturing polyacrylamide gel until the loading dye reaches the bottom of the gel. This form of fragment separation generates an exponential-like distribution of fragment sizes. Lambda *Sau*3A I marker in this context provides even distribution along the length of the gel. Fluorescent fingerprints are also separated on a denaturing polyacrylamide gel but, as fragment data is collected in real time, separation is more linear with increased separation of larger fragments.

Fluorescently labelled lambda *Sau*3A I marker was unsuitable for fingerprint fragment size determination because of uneven distribution of bands along the length of the gel. Uniform distribution of marker bands is necessary to accurately assign migration values to fingerprint bands for overlap analyses. An important additional feature of an appropriate restriction enzyme would be the generation of a G in the first position of the overhangs of the digested fragments, as this would permit incorporation of a red ddC fluorophore allowing a marker to be run in the same lane as green, yellow and blue fluorophore labelled ddA fingerprints. Another important factor in selecting a replacement enzyme is

that its optimum temperature should be compatible with the conditions under which Taq FS polymerase end labels fingerprint fragments.

Two enzymes that fulfil the criteria of having a G in the first position for a labelled ddC nucleotide to anneal, *Bsa*J I and *Taq*α I, were labelled and run with ddT labelled lambda/*Sau*3A I to permit overlap comparison, figure 3.10. *Bsa*J I gave a more uniform distribution of marker fragments along the length of the gel, particularly in the 100 bp to 2000 bp region of the gel in which the majority of fingerprint fragments would migrate. Overlapping the *Sau*3A I with *Bsa*J I marker lanes clearly indicates the distribution differences between the two enzymes (figure 3.10).

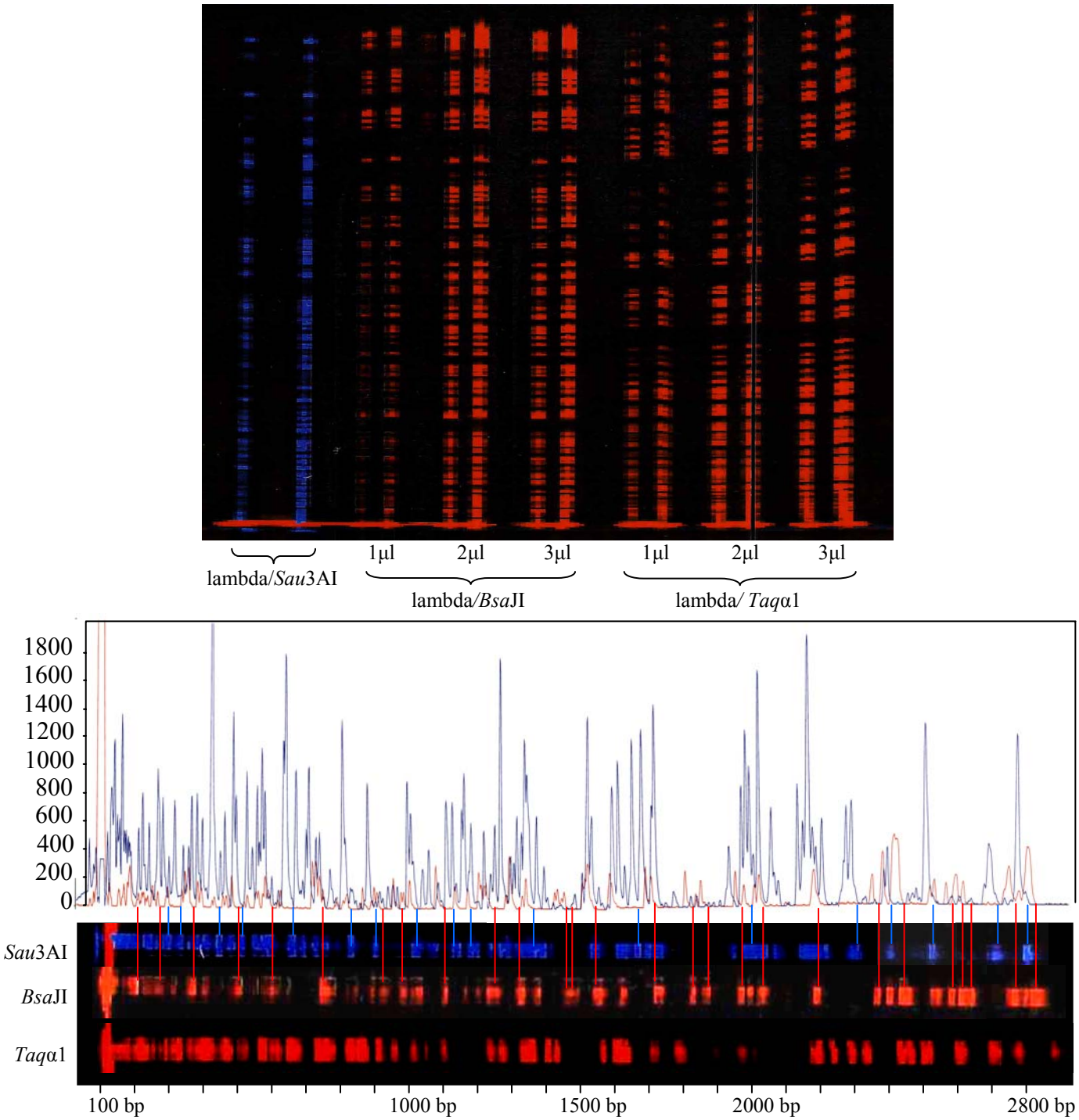


Figure 3.10: A comparison of labelled lambda digest fragments using *Sau*3A I, *Bsa*J I and *Taq*α 1 restriction enzymes. Lambda DNA labelled with fluorescent ddTTP (*Sau*3A I) and ddCTP (BsaJ I and *Taq*α 1) and loaded in different aliquots. b) BsaJ I generates a more uniform distribution of fragments in the 30 bp to 2.8 kb size range of in comparison to *Sau*3A I (which is used for generating the marker standard for radioactive fingerprints) and *Taq*α 1.

3.3.7 Data collection and processing

During protocol development data was collected on an ABI 373 and visualised using Genscan software version 2.0.2. Though the software was useful during the testing phase of the technique it was not designed to interpret the complexity of fingerprint or marker bands produced by the fluorescent fingerprinting. Instead, unprocessed gel images were collected using Prism Collection software (v1.1) running on an ABI 377 and downloaded onto a UNIX workstation. Data was edited using modified version of IMAGE before transfer into Fingerprinted Contigs (FPC) (Soderlund *et al*., 1997) for analysis. This eliminated all of the gel processing, autoradiography and scanning that is associated with radioactive fingerprinting.

3.3.8 Reproducibility

Labelling the marker with a different fluorophore to the fingerprint fragments permits the incorporation of a marker in each fingerprint lane. This facilitates greater accuracy of mobility determination of fingerprint fragments within the 30 bp to 2.8 kb size range than was previously possible. To test the accuracy of fragment analyses and reproducibility of the labelling procedure, a PAC clone, dJ163I24, was fingerprinted 10 times with each of the three ddATP dyes and compared (figure 3.11). All fingerprints were highly

reproducible. The migration values of seven fingerprint bands, representing an even distribution along the length of the gel, were compared between clones. In all cases examined, the variation of migration of bands in 30 independent fingerprints was ±0.26% (maximum observed: ± 11/4500 intervals), which is less than the tolerance used in the analysis.

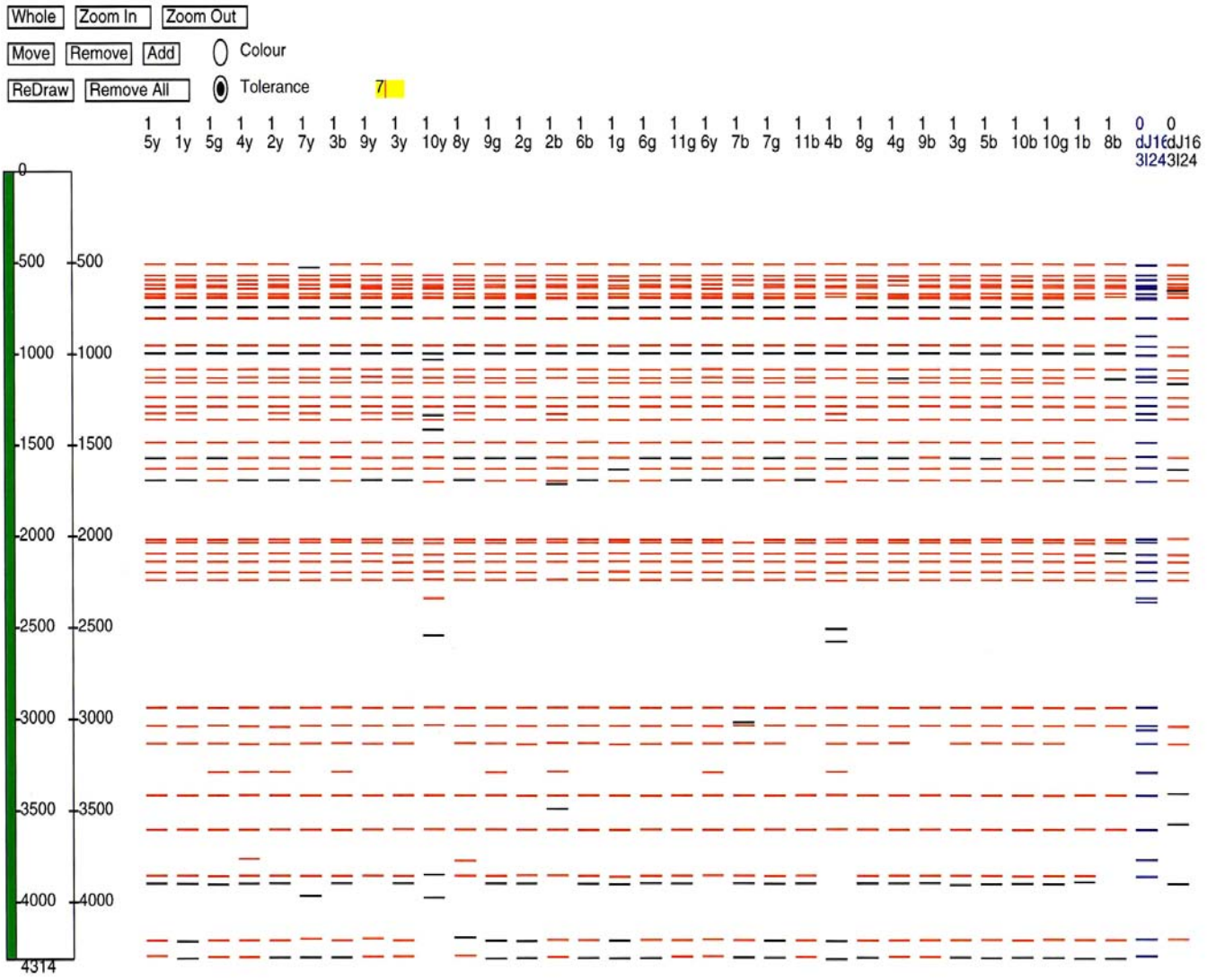


Figure 3.11: An FPC display of band labelling uniformity. 10 PAC, dJ163I24, duplicates were fingerprinted with one of each of the three ddATP fluorophores to test for labelling consistency, clone number followed by suffix b = ddATP-TET, y = ddATP-NED, g =

ddATP-HEX. Horizontal red lines correspond to fingerprint bands that are within the matching tolerance, as discussed in chapter 2.22.2, of the original fingerprint (blue bands).

3.3.9 Validation of fluorescent fingerprinting

Once the labeling and reproducibility of the fluorescent technique was established a comparison of radioactive, fluorescent and non-isotopic fingerprinting was performed using 49 PAC clones from the 1.4 Mb contig constructed as described previously (see figure 3.1). The aim was independently to assemble clones from the contig using the different fingerprinting techniques and compare the results with overlaps defined by 14 clones representing a minimum set of clones for which complete sequence was now available (figure 3.12a). Vertical arrows in (figure 3.12b) represent the overlaps that were not found between clones in each of the four methods. Figure 3.12c represents the contigs as constructed by the fluorescent fingerprinting method. The vertical arrows coincide with those in figure 3.12b, denoting overlaps that were not detected by fingerprinting alone. There was excellent agreement between the contigs assembled using either the radioactive or fluorescent *Hin*d III / *Sau*3A I data. Minor variations were observed where small overlaps between clones fell just outside the probability cut-off used. (These overlaps were originally detected by end-probe hybridisation, and later confirmed by the sequence).

There was also good agreement between the *Hin*d III / *Sau*3A I results and those of the *Hin*d III and *Eco*R I fingerprints resolved on agarose gels, but matches in the latter protocols required a higher statistical probability of overlap, i.e. smaller overlaps were not detected. This can be attributed in part to the fact that each agarose fingerprint pattern is divided into 2000 1 mm intervals along the length of a 20 cm lane as opposed to the 3400 and 4500 intervals of the radioactive and fluorescent fingerprints respectively. A smaller number of intervals implies that there is a greater probability that bands will occur randomly in the same bin and contribute to background matches. However, it should be pointed out that the two approaches are both useful in different ways. The *Hin*d III /*Sau*3A I method can detect smaller overlaps but samples 256 bp / 4096 bp (14%) of the insert DNA of a clone, and provides no information on the remainder of the sequence. By contrast the use of complete *Eco*R I, *Hin*d III or similar digests resolved on agarose gels, although achieving lower resolution, does display bands representing all the DNA of the cloned insert. This data is therefore more appropriate for verifying the integrity of the sequence compared to all genomic clones covering the region. It also confers the ability to look for possible deletions or other rearrangements that result in detectable size alterations in individual restriction fragments.

Where all four methods failed to detect overlaps between adjacent clones (figure 3.12, 1a – c), the overlapping bands were only contained within two clones. This problem would be resolved with greater contig depth. Details of analysis of overlaps which were not detected by individual methods are given in the figure 3.12 legend.

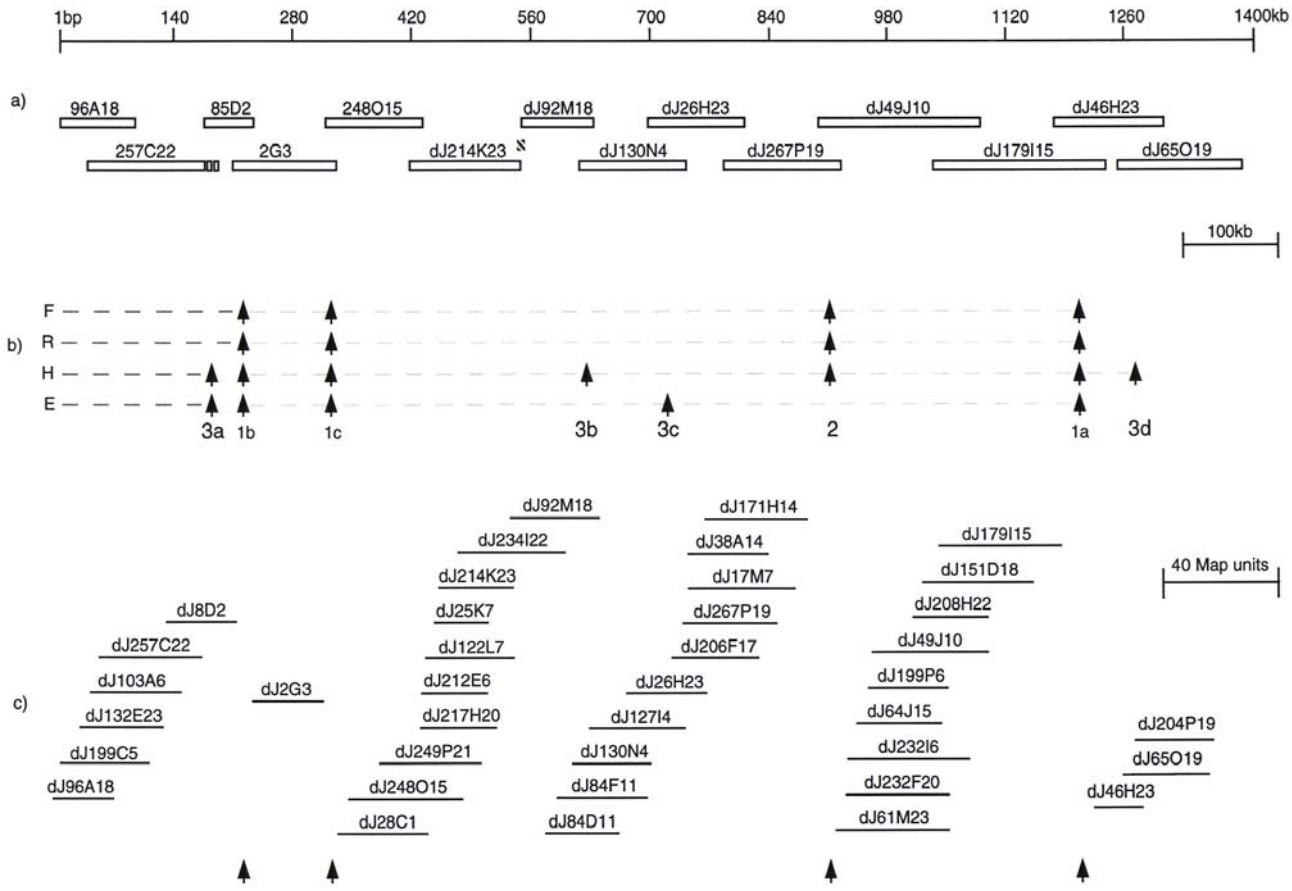


Figure 3.12: A comparison between fingerprinting methods of using 14 clones comprising a minimum tiling path. a) Finished sequence data were used to define the overlaps between

clones and absolute length of each of the clones, represented as clear boxes and scale in kb (except dJ257C22, where data were incomplete). b) Vertical arrows denote overlaps that were not detected by each method: (F) fluorescence, (R) radioactive; (H) *Hin*d III, and (E) *Eco*R I. As expected, there is variation based on recognition sites between the overlaps detected using *Hin*d III (b and d) and those established with *Eco*R I, c). Details of the analysis of undetected overlaps are as follows: (Overlap 1a, 62 kb) Nine *Hin*d III and 11 *Eco*R I fragments are present in this overlap on the basis of DNA sequence. Statistical probability on the basis of these bands shared between dJ179I15 and dJ46H23 is 1 × 10 4 (4) fluorescence, 1 radioactive, 4 *Hin*d III, and 5 *Eco*R I (whose probability cutoffs for all matches are 4, 5, 7, and 7, respectively). (Overlap 1b, 25 kb) Clone dJ2G3 is unattached. Three *Hin*d III or six *Eco*R I fragments shared with clone dJ85D2 are insufficient to score the overlap. (Overlap *1c,* 11 kb) Clone dJ2G3 is unattached. Four *Hin*d III or three *Eco*R I fragments shared with clone dJ248O15 are insufficient to detect the overlap. (Overlap 2, 24 kb) Overlap between dJ267P19 and dJ49J10 was not found by any of the *Hin*d III methods. Six *Hin*d III fragments are not sufficient to detect overlap. Arrows 3a-3d denote where only the agarose methods failed to find overlap. (Overlap 3a) Overlap between dJ257C22 and dJ85D2 could not be determined because of incomplete overlap data and therefore was not analyzed. (Overlap 3b, 18 kb) Four *Hin*d III fragments shared between dJ130N4 and dJ92M18 are not sufficient to detect overlap. (Overlap 3c, 44 kb) Six *Eco*RI fragments shared between dJ130N4 and dJ267P19 are not sufficient to detect overlap. (Overlap 3d, 55 kb) 16 *Hin*d III fragments between dJ46H23 and dJ65O19 should have been significant. The lack of overlap may be attributed to the migration of similar-sized fragments in the same location of the agarose gel and therefore not be representative of the

cloned insert. (*c*) Although all but 1 of the 49 clones (dJ2G3) fell into contigs, only 4 of the contigs that lie within the sequence interval are represented. Arrows denote where the overlaps between clones were not found.

3.4 Discussion

The construction of a 1.4 Mb contig on human chromosome 13q provided the first opportunity to test the application of restriction digest fingerprinting of large insert clones (PACs) for the construction of a physical map and their utilisation as a sequencing resource. Results indicated that they are stable, provide coverage to this extent, and few/no discrepancies between clones suggested little rearrangement. The existing *Hin*d III/*Sau*3A I fingerprinting proved suitable for the generation of PAC fingerprints therefore permitting integration of other types of smaller insert bacterial clones (Dunham *et al*., 1999, Deloukas *et al*., 2000).

The advent of large-insert bacterial clone libraries of PACs and BACs for the construction of high-resolution maps has facilitated sequencing and gene identification within chromosomal regions (Kudoh *et al*., 1997, Hubert *et al*., 1997, Guru *et al*., 1997). Construction of a bacterial clone map in either a regional or a chromosome wide study relies on initial identification of the clones by using hybridisation- or PCR-based markers to screen available libraries. The existence of overlaps between clones may be identified on

the basis of shared marker content, however, as a marker represents one point in the genomic DNA, no measurement of extent of overlap is obtained. In most regions of the human genome marker density is too low to achieve closure of all gaps by marker content alone. Alternative methods to define overlaps between bacterial clones include the hybridisation of end probes to clone arrays (Evans and Lewis, 1989), high resolution fluorescence *in situ* hybridisation (FISH) using DNA fibres (Wang *et al*., 1996) and matching of ends to complete insert sequences (Kupfer *et al*., 1995, Roach *et al*., 1995). End-probe hybridisations do not define the extent of overlap, also a limitation of whole cosmid-cosmid hybridisations (Xie *et al*., 1993), and all hybridisation-based methods employing undefined sequences can yield false positive signals because of cross hybridisation of repeat motifs. The end-sequencing strategy provides precise information on extent of overlap but relies on extensive prior investment in sequencing all clones in the library and also does not permit assembly of contigs prior to sequencing of the inserts. Fibre-FISH, using bacterial clones as probes and yeast artificial chromosomes (YACs) as a target, provides good estimates of both the extent of overlaps and size of gaps between clones but require the existence of a well-define YAC map across the region of interest and has yet to be scaled up. Fingerprinting, however, can detect overlaps where there is no marker available and gives an indication of the extent of each overlap.

Other approaches to fluorescent based fingerprinting have included ligation of fluorochrome-labeled oligonucleotide adaptors to *Eco*R I digests of cosmids (Lamerdin and Carrano 1993) and incorporation of fluorescent dideoxy terminators in 5' overhangs produced by endonuclease digestion (Ding *et al*., 2001). The former technique uses

agarose gels to resolve labeled fragments has been superceded by the VISTRA-green-stained agarose gel fingerprints (Marra *et al*., 1997), in this chapter, and the latter technique results in a complex fragment pattern that is not necessarily amenable to the generation of large scale physical maps.

The fluorescent fingerprinting technique discussed in this chapter utilises three spectrally distinct dye-ddA terminators to label the restriction fragments of three different *Hin*d III-*Sau*3AI endonuclease digests. The combination of three fingerprint samples plus a marker standard, labeled with a dye-terminator in a fourth colour, facilitates multiplexing and enhances the accuracy of data generation and overlap analyses. Improvements in data collection software, permitting the 64 multiplexed samples per gel, resulted in a two fold increase in data collection, increasing throughput to 192 clones per gel run. A requirement for the scale up in the construction of sequence-ready maps is the need to develop non-isotopic forms of contig construction. Fluorescent fingerprinting assists the progression towards large scale mapping by increasing throughput by multiplexing, enhancing the accuracy of the data generated and providing improved safety. Fluorescent fingerprinting has been successfully applied to the construction of a physical map of a region of mouse chromosome X (Mallon *et al*., 2000) as well as large scale sequence-ready bacterial clone maps of human chromosomes 1, 6, 20 (Deloukas *et al*., 2001) and X (Bentley *et al*., 2000) as discussed in the next chapter.

Chapter 4

Construction of a sequence-ready bacterial clone contig of 1pcen – 1p13

4.1 Introduction

Alterations of human chromosome 1 are among the most common form of chromosomal abnormality detected in human disease. Chromosomal aberrations localised to chromosome 1pcen-1p13 have been associated with human neoplasia such as acute megakaryoblastic leukaemia (Mercher *et al*., 2001), radiation induced meningioma (Zattara-Cannoni *et al*., 2001), colorectal Cancer (Nitta *et al*., 1987), as well as other diseases such as non-goitrous hypothyroidism (Dracopoli *et al*., 1986) and autosomal recessive ventricular tachycardia (Lahat *et al*., 2001).

Previously, two attempts have been made to construct physical maps within 1pcen-1p13 (Carrier *et al*., 1996, Brintnell *et al*., 1997). These maps, primarily consisting of YACs, were constructed for the purposes of characterising the genic environment of the nerve growth factor gene (NGFB) (Carrier *et al*., 1996) and for the elucidation of putative candidate genes within a smallest region of overlapping loss of heterozygosity (LOH) in breast cancer studies (Brintnell *et al*., 1997). Though the YACs used in the construction of these maps provided a means of generating coverage of large genomic regions with relatively few clones, the comparative difficulty of constructing the libraries and of analysing the cloned DNA (including shotgun sequencing) means that YACs as the primary resource are not well-suited for the construction of sequence ready maps. Instead, YACs have provided a means of linking bacterial clone contigs where genomic sequence is not represented within the bacterial clone libraries.

In contrast to YACs, bacterial clone libraries are easier to make and the cloned DNA is more easily manipulated. The use of bacterial clones for construction of a high resolution sequence ready map within 1pcen-1p13 would not only facilitate the identification of disease genes but would also provide the basis for a detailed characterisation of the genomic landscape of the interval. This chapter describes the construction of a sequence ready bacterial clone map in 1pcen-1p13 and compares previously published genetic, radiation hybrid and physical maps within the interval.

4.2 Construction of sequence-ready map of 1pcen – 1p13

A hierarchical strategy was used to construct the map as follows (see figure 4.1). At the start of the project, 3642 markers were publicly available which had been localised to the region by genetic or RH mapping. To provide additional markers for map construction a small insert library (SIL) derived from flow sorted chromosomes was constructed (see section 4.2.1). SIL clones were picked at random and sequenced to generate novel STSs. A subset of these novel STSs were placed in the region by RH mapping. The combined set of markers was used to initiate contig coverage by screening large-insert bacterial clones. Bacterial clone coverage was supplemented with contigs generated by McPherson *et al.,* (2000) from a whole genome fingerprint database. Contigs from this database were selected for inclusion in the 1pcen-1p13 map if they overlapped with existing clone contigs on the basis of shared fingerprints and/or marker content. The map was completed by iterative rounds of walking, using sequences at or near the end of each contig to re-screen the available libraries for clones.

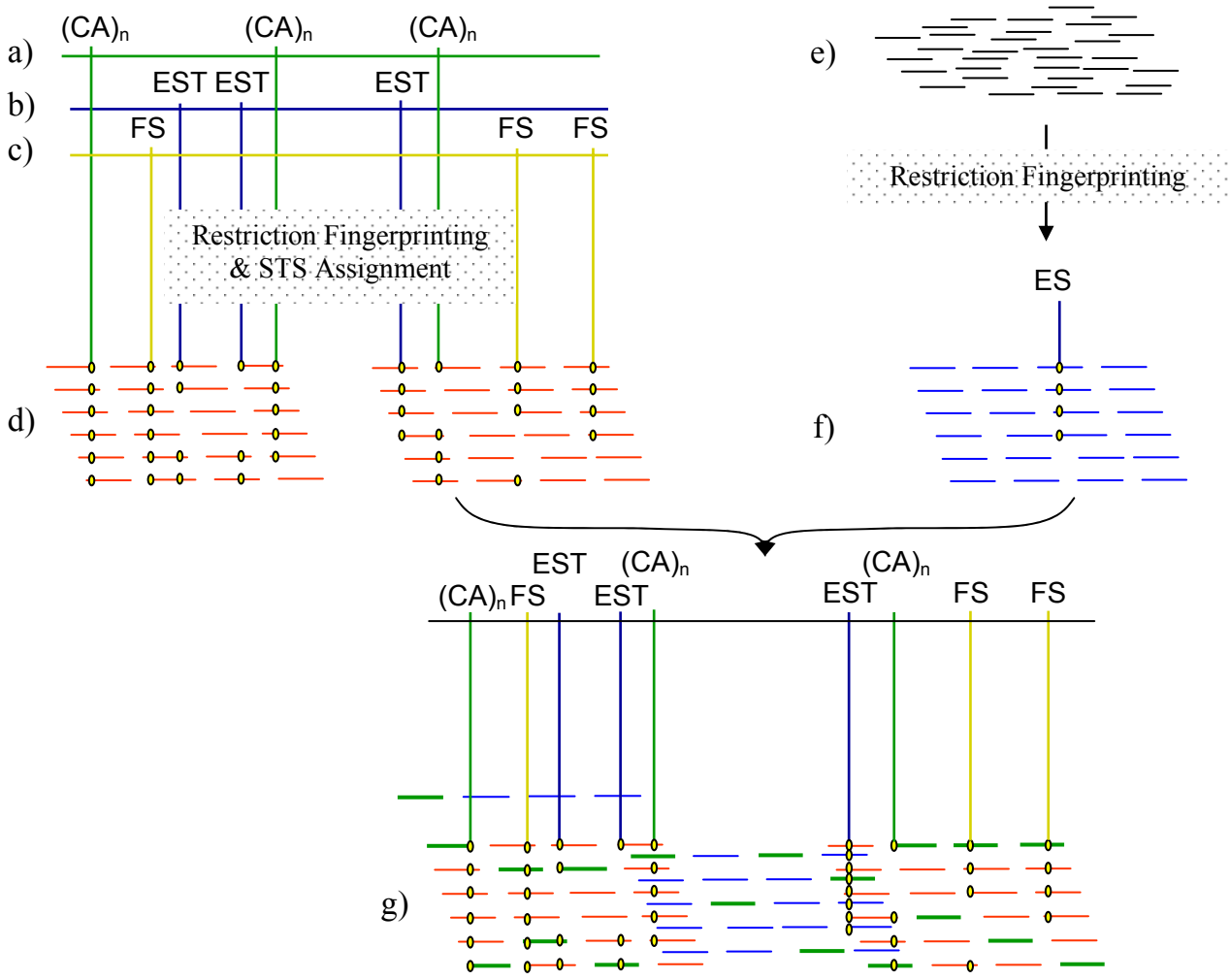


Figure 4.1: A representation of the two strategies used to construct a sequence-ready bacterial clone map of 1pc – 1p13. The hierarchical method utilises polymorphic genetic markers CA)n a), expressed sequence tags (EST) b) and markers generated by sequencing small insert libraries (FS) c) which are radiation hybrid mapped and screened across genomic libraries. Restriction fingerprinting and STS assignment is performed in parallel prior to data assimilation in FPC (orange lines, d). The whole genome fingerprinting approach utilises restriction digest fingerprinting of a 15-fold redundant genomic library e), including some marker data, to establish bacterial clone coverage f). Data from both these techniques is combined to generate contiguous map coverage g) proving a resource for the selection of a minimum tile path clones, bold green lines.

4.2.1 Small insert library construction

A bivariate flow karyotype of human DNA (figure 4.2a) was generated to facilitate purification of chromosome 1 DNA from other chromosomes (flow sorting was performed by Nigel Carter). Purified DNA was then completely digested with *Hin*d III prior to cloning into pBluescriptII vector and electroporated into *E. coli* XL1 blue electrocompetent cells (figure 4.2b). Two hundred test recombinants were picked from Xgal indicator plates, miniprepped (Birnboim *et al*., 1979) and DNA separated by electrophoresis on a 1% agarose gel as undigested and digested to ascertain average insert sizes (figure 4.2c). Of the two hundred test recombinants, 96% contained inserts and the average size was estimated to be ~5.6 kb. Successful generation of sequence data from a high percentage of 400 test SIL clones (data not shown) prompted a further 7296 SIL clones to be picked and sequenced (by others). A 36% failure rate (2611) at the prepping stage reduced the number of sequence templates to 4685.

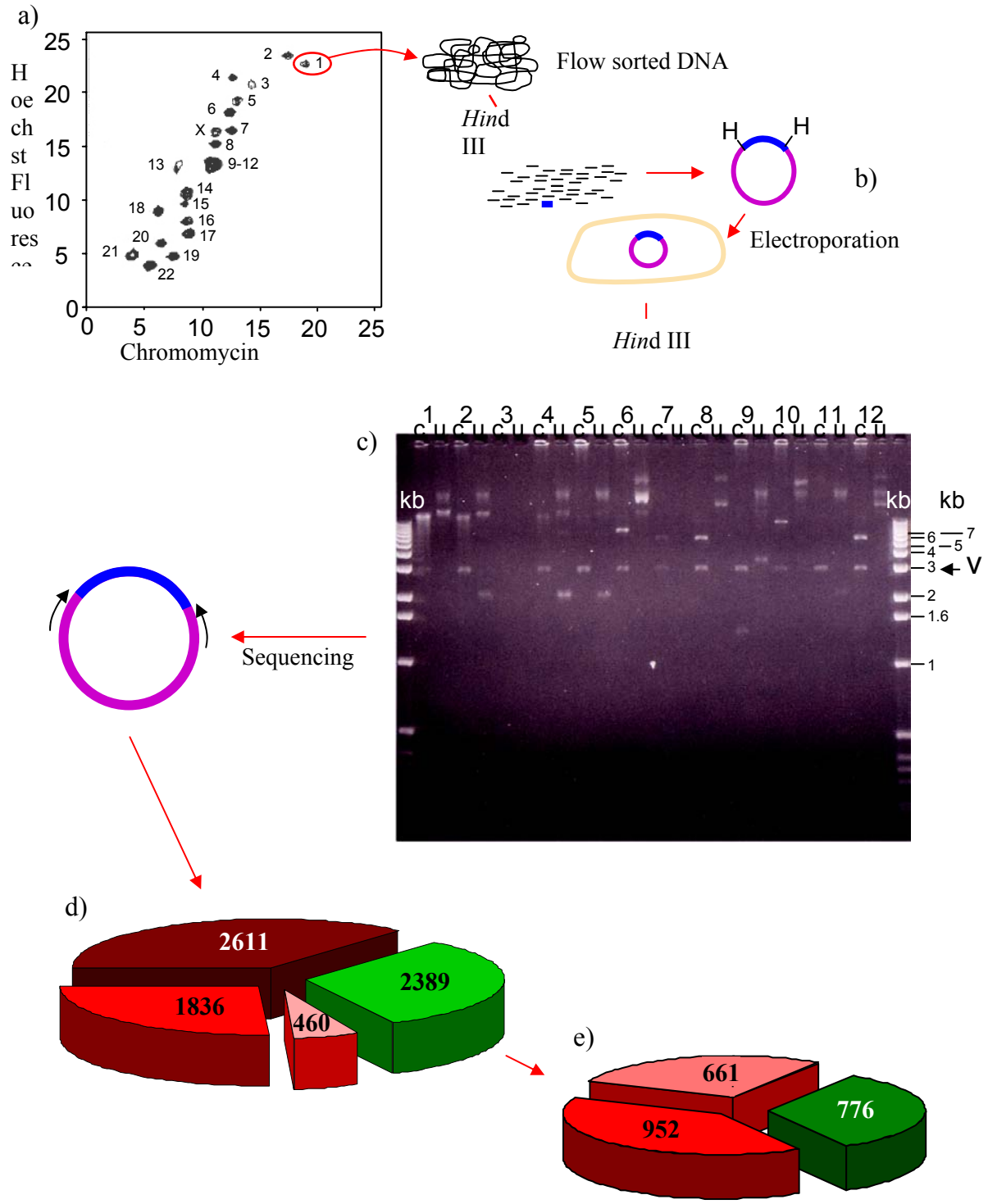

a)
Chromomycin
Flow sorted DNA
Hind III
b)
Electroporation
Hind III
c)
kb
V
Sequencing
d)
2611
2389
1836
460
e)
661
776
952

Figure 4.2: The construction of a chromosome 1 specific small insert library. a) A bivariate flow karyotype was generated to flow sort chromosome 1 DNA. b) DNA was completely digested with *Hin*dIII, cloned into pBluescriptII vector and electroporated into *E. coli* XL1 blue electrocompetent cells. c) Subclones were separated by electrophoresis cut (c) and uncut (u) on an agarose gel and insert sizes determined by use of 1 kb ladder (kb). The vector band (v) was used to correlate the insert size with the undigested band. d) Prior to primer design there were small insert library clone failures at prepping (dark red), sequencing (including sequence of < 80 bp (red)) and STS design stages (pink). Further failures e) were due to repeats contained within either primer (dark red) or during experimentation (pink). The total number of RH mapped flow sorted markers with unambiguous placement on the physical map is represented in figure 4.2e) (dark green).

Figure 4.3 illustrates the size distribution of the 4685 SIL clones that were sequenced. Of these clones, 49% had insert sequences that were inappropriate for primer design. This set comprised 8% that failed to produce sequence; 31% with sequences <80 bp; and 10% with sequence that did not satisfy primer design parameters (e.g. unequal or suboptimal GC content preventing primer design, PCR product size shorter than an acceptable minimum length). The remaining 2389 (51%) small insert library SIL clones sequences (31% of the original number of picked SIL clones) produced sequence from which primer pairs were could be designed (figure 4.2d).

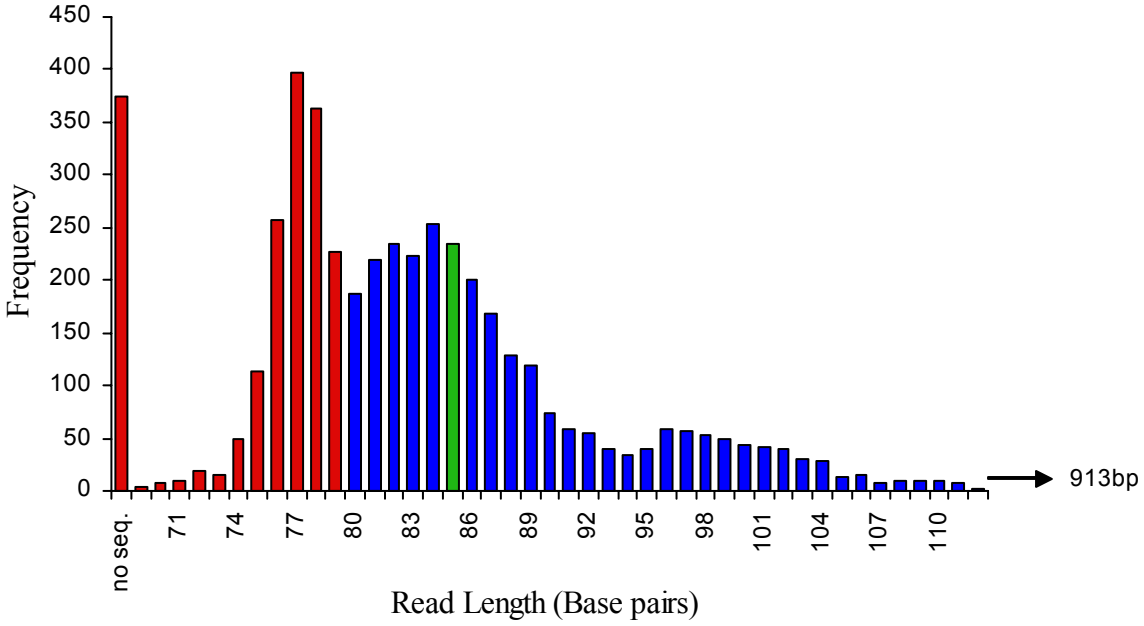


Figure 4.3: Sequence length and frequency of SILs passing STS design stage. Sequences too short for primer design (red), statistically significant proportion (91%) of STSs that generated sequence that were suitable for primer design (blue) and the mean read length, 85 base pairs (green).

4.2.2 Radiation Hybrid Mapping of SIL markers

All 2389 primer pairs were synthesised and tested by RH mapping (by others). Of these, 952 were rejected as at least one of a primer pair wholly or partly contained repeat sequence and a further 439 were experimental failures. These failures included mapping of a marker to multiple chromosomal locations, failure to map to a human chromosome and poor specificity (indicated by production of multiple PCR products generated when different annealing temperatures were used during PCR). The 776 flow sorted STS markers that were successfully

radiation hybrid mapped and placed uniquely on the physical map represent 11% of the flow sorted small insert library clones originally picked for sequencing. As a result of this work, the total number of unique markers in the chromosome 1 RH map was increased from 3642 to 4418.

4.2.3 Bacterial clone contig construction

The combined set of 204 markers localised to the 1pcen-1p13 interval by RH mapping were used to isolate large-insert bacterial clones for contig construction. Initially, each PCR assay was tested by amplification of genomic DNA and DNA from a chromosome 1p hybrid in order to establish optimal conditions for specificity (figure 4.4a). PCR products from the assays were then radiolabelled by PCR and hybridised in pools of 20 – 30 to genomic arrays of PAC clones (fig 4.4b). Positive PAC clones were picked into microtitre plates and re-arrayed on a bacterial clone polygrid filter. Probes were then hybridised individually to the polygrid filter. A total of 110 publicly available and flow-sorted markers were screened in pools across RPCI 1, 3 – 5 PAC libraries identifying 878 PAC clones (the remaining 86 available markers were PAC screened by others as part of the chromosome 1 mapping project). Hybridisation data was entered into a chromosome specific ACeDB database (Durbin and Thierry-Meig 1994), 1ace. In parallel, positive bacterial clones identified by the pool hybridisation were subjected to restriction digest fingerprinting (Gregory *et al*., 1996) (figure 4.4c) and marker hybridisation data from 1ace was assimilated within FPC (Soderlund *et al*., 1997) following contig assembly. Minimum tile path clones representing bacterial clone coverage of island contigs generated by marker hybridisation, were selected for sequencing.

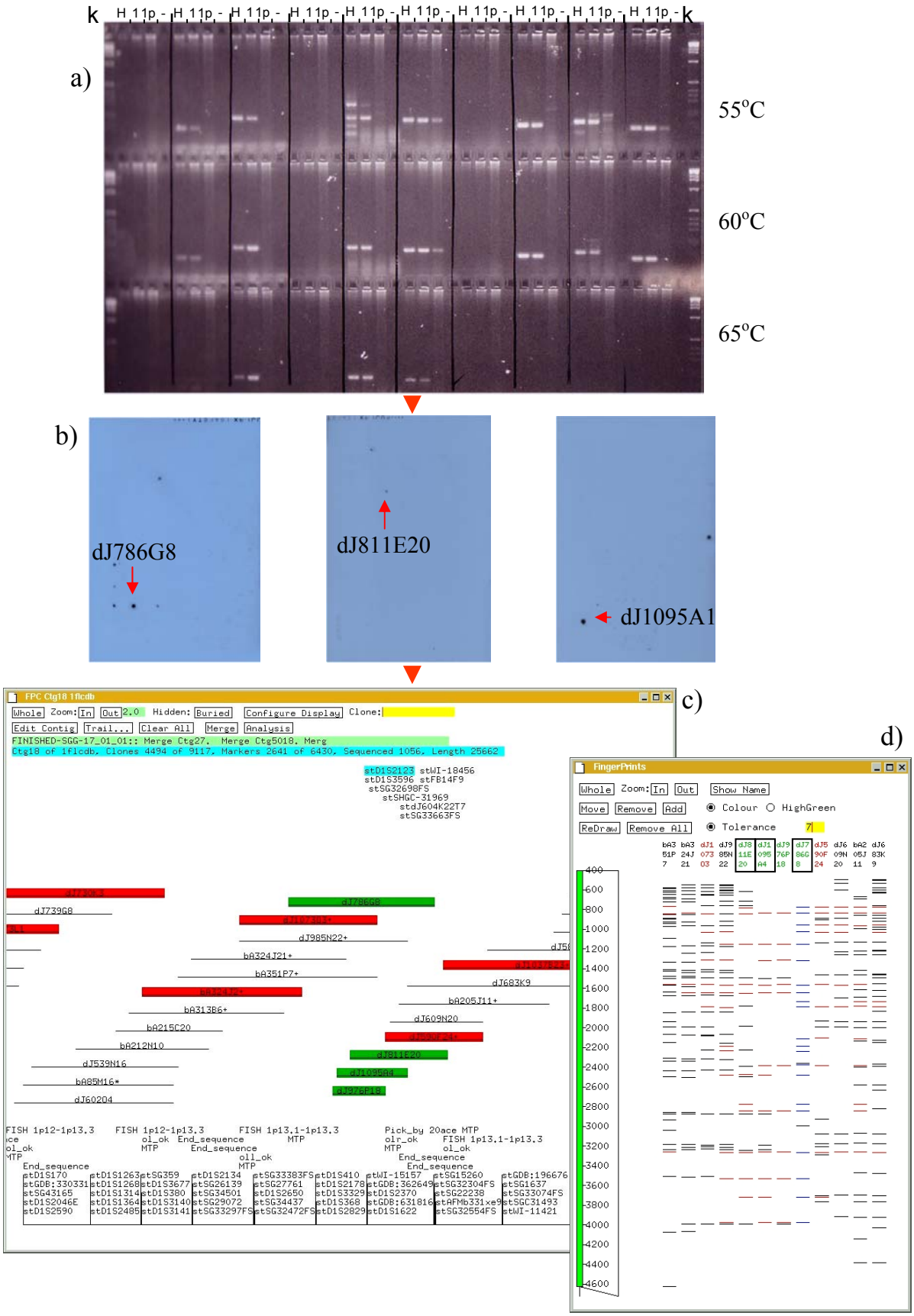

k H 11p - H 11p - H 11p - H 11p - H 11p - H 11p - H 11p - H 11p - H 11p - k
a)
55°C
60°C
65°C
b)
dJ786G8
dJ811E20
dJ1095A1
c)
d)
FPC Ctg18 1flcdb
Whole Zoom: In Out 2.0 Hidden: Buried Configure Display Clone:
Edit Contig Trail... Clear All Merge Analysis
FINISHED-SGG-17_01_01:: Merge Ctg27. Merge Ctg5018. Merg
Ctg18 of 1flcdb. Clones 4494 of 9117. Markers 2641 of 6430. Sequenced 1056. Length 25662
stD1S2123 stWI-18456
stD1S3596 stFB14F9
stSG32698FS
stSHGC-31969
stdJ604K22T7
stSG33663FS
dJ739G8
dJ786G8
dJ985N22+
bA324J21+
bA351P7+
dJ683K9
bA205J11+
bA313B6+
dJ609N20
bA215C20
bA212N10
dJ539N16
bA85M16*
dJ602O4
dJ811E20
dJ1095B4
FISH 1p12-1p13.3
FISH 1p12-1p13.3
FISH 1p13.1-1p13.3
Pick_by 20ace MTP
FISH 1p13.1-1p13.3
End_sequence
stD1S170 stD1S1263 stSG359 stD1S2134 stSG33383FS stD1S410 stWI-15157 stSG15260 stGDB:196676
stGDB:330331 stD1S1268 stD3677 stSG26139 stSG27761 stD1S2178 stGDB:362649 stSG32304FS stSG1637
stSG43165 stD1S1314 stD1S380 stSG34501 stD1S2650 stD1S3329 stD1S2370 stSG22238 stSG33074FS
stD1S2046E stD1S1364 stD1S3140 stSG29072 stSG34437 stD1S368 stGDB:631816 stAFMb331xe9 stSGC31493
stD1S2590 stD1S2485 stD1S3141 stSG33297FS stSG32472FS stD1S2829 stD1S1622 stSG32554FS stWI-11421
FingerPrints
Whole Zoom: In Out Show Name
Move Remove Add Colour HighGreen
ReDraw Remove All Tolerance 7
bA3 51P 7 bA3 24J 21 dJ1 073 03 dJ9 85N 22 dJ8 11E 20 dJ1 095 A4 dJ9 76P 18 dJ7 86G 8 dJ5 90F 24 dJ6 09N 20 bA2 05J 11 dJ6 83K 9
400 600 800 1000 1200 1400 1600 1800 2000 2200 2400 2600 2800 3000 3200 3400 3600 3800 4000 4200 4400 4600

Figure 4.4: Generation of sequence ready bacterial coverage using the hierarchal strategy. a) An agarose gel showing PCR products generated at annealing temperatures of 55^{o}C, 60^{o}C and 65^{o}C. The PCR reaction used total human DNA (H), chromosome 1 (1), chromosome 1p (1p) and a negative control (-) as template. b) Autoradiographs of an STS pool hybridisation to genomic PAC library filters, positives are indicated by red arrows. c) PACs identified by hybridisation were restriction digest fingerprinting, assembled into contigs and assimilated with STS data within FPC. d) The digitised fingerprints (black boxes) of 3 clones identified by pool screening and assembled within FPC.

Bacterial clone coverage of 1pcen-1p13 within the whole genome (WG) fingerprint database (McPherson *et al*., 2001) was identified either by using the chromosome 1 markers associated with assembled contigs (where they were available), or by adding a representative set of experimental or virtual *Hin*d III fingerprints of PAC clones isolated by the hierarchical strategy (Figure 4.5a). Placement of PACs with associated RH mapped STS data within the stringently assembled WG fingerprint database facilitated the localisation and joining of WG BAC contigs. A three-fold redundant tiling path of BAC clones was picked from the WG database (Figure 4.5b) and incorporated into the chromosome 1 specific fluorescent fingerprint database (as part of this thesis and the chromosome 1 project). Once WG BACs were assimilated into the 1pcen-1p13 PAC contigs a more optimal series of BAC and PAC clones were chosen (with more minimal overlaps) as the sequencing tile path (Figure 4.5c). The use of both hierarchical and WG fingerprint approaches resulted in the 1pcen-1p13 interval being covered by 2 bacterial clone contigs. The two bacterial clone contigs were fortuitously linked by 2.3 kb of unfinished sequence from BAC clone bA722J12, produced by

RIKEN Genomic Sciences Center as part of their chromosome 18 sequencing project. The clone appears to have been erroneously chosen by RIKEN as the unfinished sequence is placed uniquely on chromosome 1 by high BLAST score alignment.

Figure 4.5. Can be viewed in a separate PDF

Figure 4.5: The assimilation of PAC contigs into whole genome and chromosome specific fingerprint databases. a) PAC clones representing bacterial clone coverage from the hierarchical mapping technique were added to the WG fingerprint database by either experimental or virtual restriction digest fingerprinting (light blue clones). b) a three fold redundant set of BACs were selected from the WG fingerprint database and incorporated into the chromosome 1 specific database by fluorescent fingerprinting. c) Minimum tile path clones are boxed in red, alignment between the 3 maps is shown by selected clones (circled).

The final 13 Mb bacterial clone contig covering human 1pcen-1p13.2 (figure 4.6a) includes 1130 bacterial clones (480 BACs, 648 PACs and 2 cosmids) and contains 250 markers by hybridisation. Two hundred and four of these markers (157 publicly available and 47 flow sorted) have associated RH mapping data. The final map contained an additional 46 markers without RH information; but these markers were added to the contig map by hybridisation. All markers have a unique chromosomal placement, i.e. either all or the majority of bacterial clones identified by hybridisation have fingerprints that assemble within the 1pcen-1p13.2 contig and to no other location the chromosome. A minimum tile path of 136 minimum clones (figure 4.6b), including 2 previously sequenced cosmids (AC000031 and AC000032), was selected for sequencing (by others) within the Sanger Institute. Figure 4.6c represents an FPC display of a 3 Mb region of the 13 Mb contig which illustrates the extensive coverage of independent map information which anchors clones across the 1pcen-1p13.2 contig. The 3 Mb interval is defined by two framework markers, D1S221 and D1S2746 (markers are denoted by yellow boxes, figure 4.6c). Also shown are markers with RH mapping data (light blue boxes) that were used to identify large insert bacterial clones by hybridisation to genomic library filters.

Figure 4.6 can be viewed in a separate PDF

Figure 4.6: A representation of PAC and BAC contig coverage of 1pcen – 1p13. a) An ideogram of the region of bacterial clone contig coverage generated by this study as represented at 850 band resolution of human chromosome 1. b) An AceDB display of the minimum tile path of 136 clones from the 13 Mb contig. Represented are the sequencing statuses of minimum tile path clones (as of 14th October 2002), pre-shotgun (clear boxes), pre-finished (green boxes), finished sequence (red boxes) and submitted sequence (black boxes). c) a 3 Mb region of the final contig depicted in FPC reflecting the density of framework markers (yellow boxes), RH mapped markers (light blue boxes) and non-RH mapped markers (clear) positioned by hybridisation within the bacterial clone contig. Minimum tile path clones are highlighted in red.

4.3 Evaluation of SIL marker distribution in chromosome 1

The chromosomal distribution of 776 RH mapped STSs, originating from the flow sorted small insert library, was evaluated based on the fingerprint map (figure 4.4). The histogram in figure 4.7b depicts the number of STSs in 5 Mb intervals along the length of chromosome 1. The interval size was calculated directly from the fingerprint map using 4 bands / kb (see section 2.22.2). The average number of markers (14.7 / 5 Mb) varies appreciably next to the centromere and telomeres. This may be due to difficulties in cloning bias encountered when cloning chromosomal centromeric and telomeric repeat sequences (Doggett *et al*., 1995). The overrepresentation of flow-sorted markers in some intervals may relate to sequence content. For example, a region of genomic sequence with a particular base composition (e.g. AT-rich)

may be particularly amenable to digestion with *Hin*d III, and thus yield a high fraction of short (easily cloned) *Hin*d III region compared to a GC-rich region. This bias would be reflected in the representation of inserts in the SIL, as the SIL was prepared from HindIII-digested DNA.

The distribution of the 776 flow sorted markers in the fingerprint-based map was compared to the radiation hybrid map of chromosome 1 (figure 4.7c). There is very good agreement between the two maps, with 12 STSs (1.5%) showing incongruent placement (blue dots). These discrepancies may result from low copy repeats (duplicates not being identified within radiation hybrid vectors) or by experimental error where the wrong STS has been hybridised to the gridded arrays of bacterial clones. The density of markers that have associated RH mapping data and unique chromosome hybridisation data within 1pcen-1p13, was increased from 12 / Mb to 17 / Mb upon addition of flow-sorted markers.

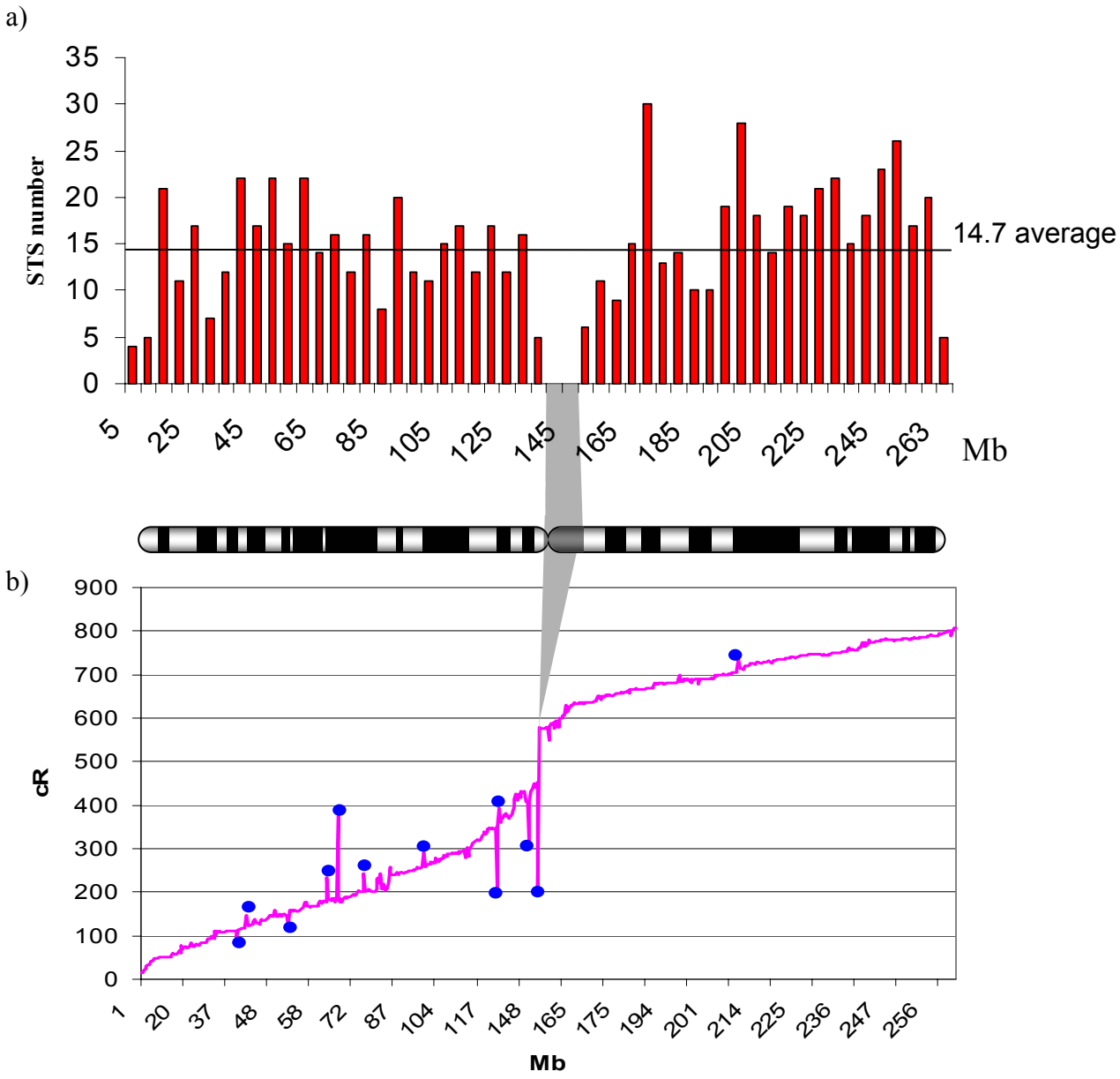


Figure 4.7: Chromosomal distribution of flow sorted markers and comparison to of radiation hybrid and physical maps. a) The chromosomal distribution of 776 RH mapped STSs originating from the flow sorted small insert library along the length of the chromosome in 5 Mb map intervals. b) A comparison of the radiation hybrid map of 778 flow sorted markers and their localisation to bacterial clone contigs by hybridisation, outliers are denoted by blue dots.

4.4 Comparison of Published Maps:

4.4.1 Physical maps

Two physical maps, consisting primarily of YACs, have previously been published within 1pcen-1p13 (Carrier *et al*., 1996, Brintnell *et al*., 1997). Carrier *et al.,* (1996) used genetic markers and STSs designed from genes localised to 1p13 to screen CEPH YAC libraries. Overlaps between the eight non-chimeric YACs used to build the 3 Mb contig around the NGFB locus were determined by long-range restriction mapping and by STS content data. Brintnell *et al.,* (1997) used a similar approach to generate YAC coverage (from CEPH and ICI YAC libraries) across an interval of 1p13.1 but also included CIT library BACs within the contig. Though the two contigs overlap, based on their shared content of marker D1S252, the Brintnell *et al.,* (1997) map extends distally by 1.7 Mb. Both maps, with the exception of one marker, were in broad agreement with the physical map described here (within the limits of localising a marker or gene by hybridisation to a YAC or YAC restriction fragments). One conflicting marker, D1S3347, was positioned within 1p13 by Brintnell *et al.,* (1997) but has subsequently been localised to 1p35.3 by RH mapping. BLAST analysis of D1S3347 places the marker in 1p35.1 and 1p12 via a high BLAST score and 95% sequence alignment (the parent sequence contains ambiguous bases, thus the <100% BLAST score). D1S3347 (synonymous with WI-8708) was derived from an EST. Subsequent alignment of the EST to genomic sequence (by Ensembl analysis) indicates the marker is derived from the 3' UTR of an mRNA isolated from small cell carcinoma lung tissue by the IMAGE consortium. The three exon gene is localised to 1p35.1 whilst a processed mRNA, inserted into the genomic

sequence in the reverse orientation, is contained within 1p12 and thus explains the discrepant placement of the marker in the Brintnell *et al.,* (1997) study.

4.4.2 Genetic map

The generation of a bacterial clone contig within 1pcen-1p13 permitted a comparison to be made with the genetic map of chromosome 1 (Dib *et al*., 1996). The order and distance between 27 genetic markers placed uniquely in the bacterial clone contig by hybridisation was determined. There was very good agreement between the two maps in relation to marker order (figure 4.8), with the higher resolution bacterial clone map facilitating the separation of markers that had previously been placed within the same genetic interval. Two genetic markers showed discrepant placement; the first, D1S2852 (152.2 cM, 8.2 Mb) localises to its current position in the physical map by hybridisation and sequence alignment; the second, D1S418 (152.2 cM, 11.3 Mb), was positioned by weak hybridisation to two PAC clones within the contig. BLAST analysis of the sequence from which D1S418 was derived (which contains ambiguous bases) localises the marker within PAC clone dJ671G15 placing it in the correct physical map location. Incorrect placement of these markers within the Généthon genetic map may be attributed to the level of resolution provided by the number of meioses used to construct the map.

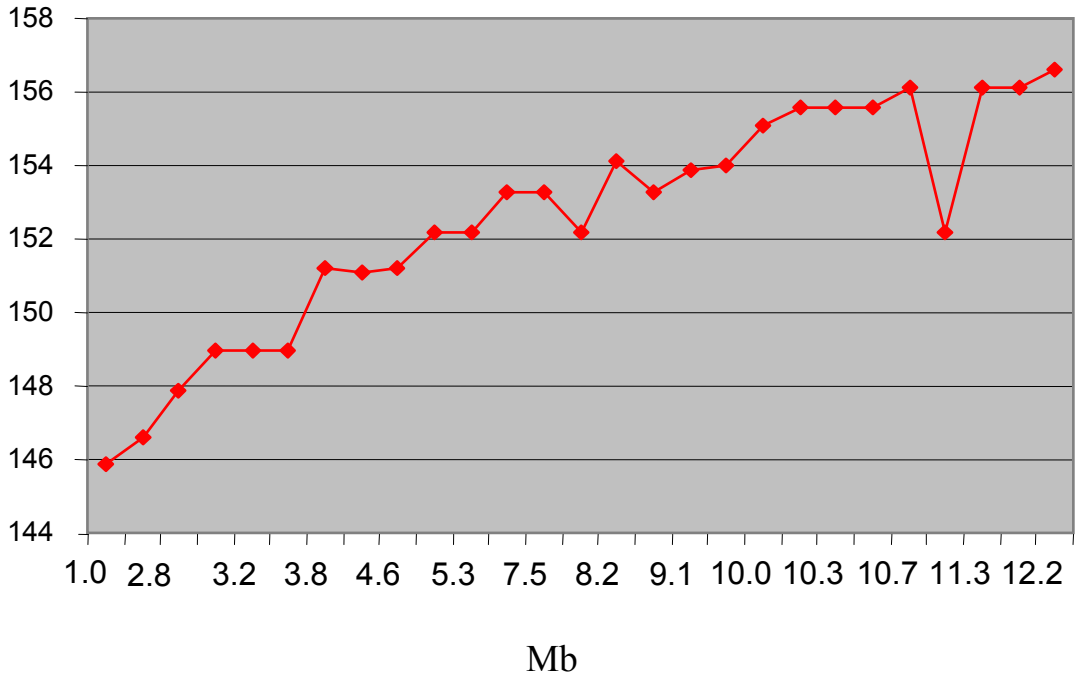


Figure 4.8: The distribution of genetic mapped markers positioned within the 1pc – 1p13 contig by hybridisation. The physical map is plotted on the X-axis in megabase (Mb) and the Y-axis in centimorgans (cM).

4.4.3 Radiation hybrid map

To investigate the order and resolution of radiation hybrid markers on the physical map, markers with unique hybridisation positions on the fingerprint map were plotted against their RH map locations. Figure 4.9 shows a good correlation for the most part between physical and RH maps but indicates a number of markers that appear to be poorly resolved on the RH map. Twenty seven markers show discrepant placement of >15 cR from their physical map position, the resolution of the chromosome 1 RH map (Panos Deloukas personal communication). One marker was subsequently placed correctly by BLAST analysis and a

second marker incorrectly placed because one of the primer pairs was designed within a repeat. The remaining 25 are placed on the physical map by unambiguous hybridisation and or electronic PCR (ePCR) (Schuler *et al*., 1997).

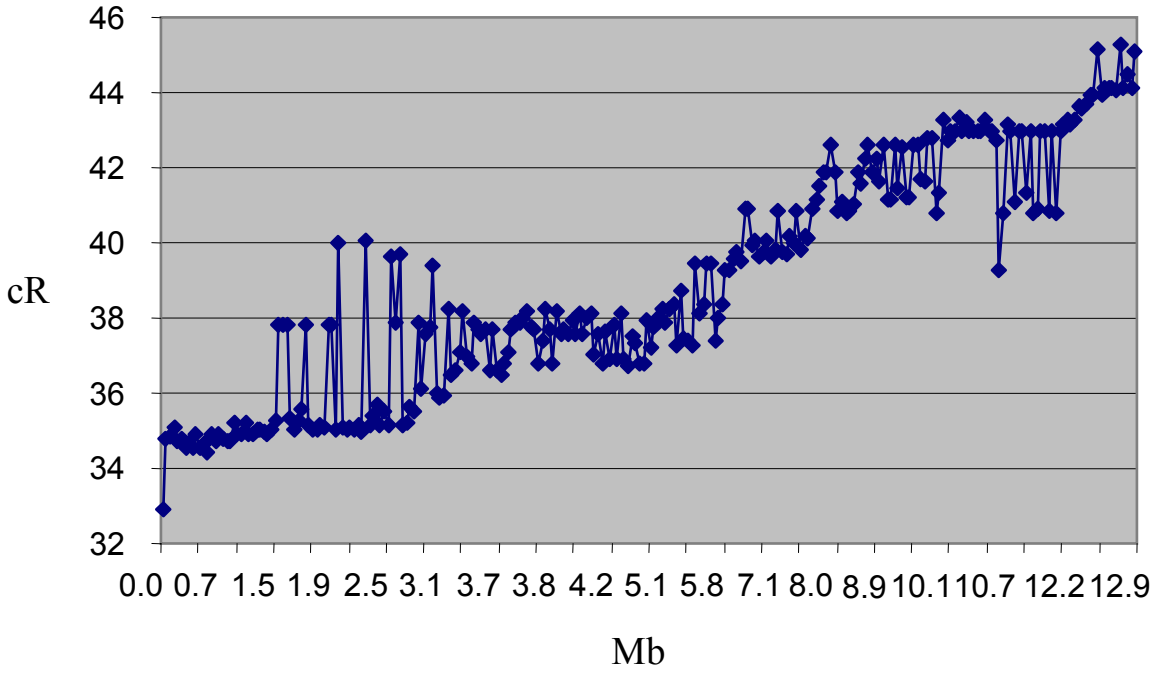


Figure 4.9: The distribution of radiation hybrid mapped markers positioned within the 1pc – 1p13 contig by hybridisation. The physical map is plotted on the X-axis in megabase (Mb) and the Y-axis in centi-ray (cR).

4.4.4 A comparison of three maps

The comparative accuracy of the genetic and radiation hybrid maps can be evaluated by plotting a set of markers contained within all three maps from the analyses above (figure 4.10). The genetic map shows good agreement with the physical map (discrepant markers being accounted for above) with markers placed within the same genetic interval being

resolved. The RH map provides higher resolution and is also mostly in good agreement with the other map and provides both more markers and a higher resolution than the genetic map. However it also shows more discrepancies in marker order (see chapter 7).

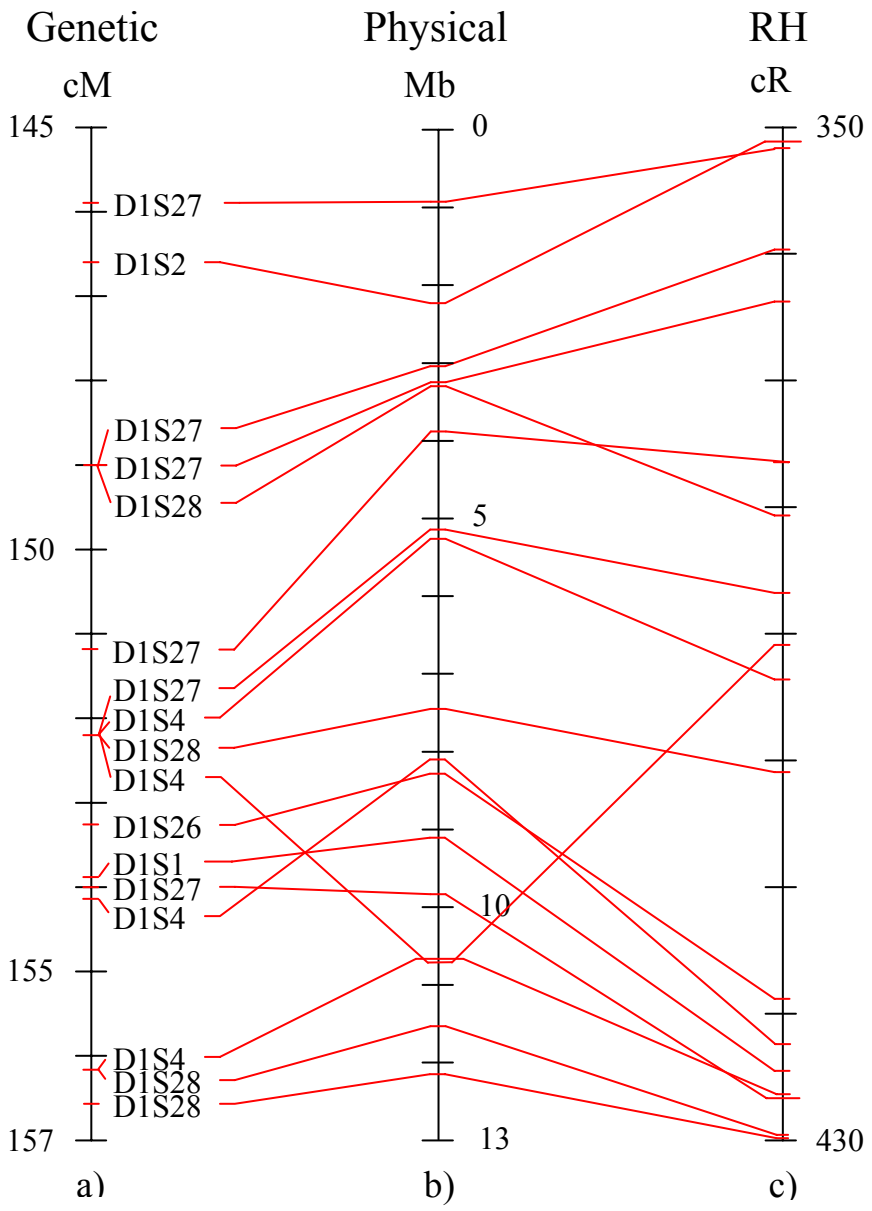


Figure 4.10: A comparison of marker distribution between genetic a), physical b) and radiation hybrid c) maps of 1pcen – 1p13. Only markers that were present on all three maps have been represented.

4.5 Discussion

Two different strategies were used to construct a 13 Mb bacterial clone contig of human chromosome 1pcen-1p13. Bacterial clone coverage was initiated by utilising publicly available markers, localised to 1pcen - 1p13 by RH mapping, to screen large insert PAC libraries. The non-random distribution of these markers (a large proportion of which had been derived for positional cloning projects or from ESTs, thus enriching marker density in disease regions or gene-rich regions, respectively) required the generation of flow sorted STSs. A total of 776 markers derived from a SIL, were successfully placed on the chromosome 1 specific radiation hybrid map and 47 of them mapped within the region 1pcen-1p13 markers. This supplemented the publicly available markers in the region (157) and increased the density of markers from 12 per Mb to 17 per Mb, above the target density of 15 per Mb (Olson *et al*., 1993, Bentley *et al*., 2000). PACs identified by genomic library screening of all markers within the interval were fingerprinted and assembled into contigs. Selected PACs from these island contigs were incorporated within the whole genome BAC fingerprint database by *Hin*d III fingerprinting and overlap analysis. Clones from this database were retrieved and incorporated into the local study. At the end of the project, a three-fold redundant tile path of 480 BAC clones, from the whole genome fingerprint database, in addition to the 648 PACs, identified by library screening, constitute the 13 Mb contig.

The generation of contiguous physical map coverage has permitted a high resolution comparison to be made of genetic and radiation hybrid maps within the interval. The algorithms used to construct the RH map assume random distribution of breaks along the

length of the chromosome. Experimental data (Deloukas *et al*., 1998) has shown that the high retention rate at the centromere, coupled with variation of DNA fragment sizes (in comparison to the rest of the chromosome), results in an overestimation of the cR distances between markers adjacent to the centromere. These anomalies explain why there are a relatively high percentage of markers that show discrepancies between the two maps and why the usual size estimate of 1 cR_{3000} to 250 kb (Deloukas *et al*., 1998) does not hold within 1pcen-1p13. According to the RH map the 105 cR interval should be 26 Mb where as it has been shown to be ~13 Mb.

The relative uniformity of the physical map metric facilitated the resolution of markers that had previously been localised to the same interval on the genetic map. Apparent differences in recombination rate across the interval, as shown by the step-wise comparison of genetic and physical maps in figure 4.8, cannot be determined from the resolution of the bacterial clone contig and would have to be resolved by placement of the markers within contiguous genomic sequence. The physical map is not a uniform metric as it relies upon the existence of *Hin*d III sites for the generation and overlapping of fingerprint bands; a clone with a large number of bands may contain the same amount of genomic sequence as a shorter clone with fewer bands.

It has been proven that there is a large variation in the rate of recombination along the length of a chromosome (IHGSC, 2001). Recombination rates at telomeres are greater than within chromosome arms which are in turn greater than regions adjacent to the centromere.

Therefore, it was not unexpected that a shorter genetic distance, 8.7cM defines a larger physical distance, 13 Mb, within 1pcen-1p13.

The bacterial clone map has provided a means by which a more accurate estimate of the physical size of the interval can be made, overcoming inherent levels of lower resolution within genetic and radiation hybrid mapping. High resolution sequence analysis of the 136 minimally overlapping bacterial clones selected for sequencing from the contig constructed here forms the basis of the next chapter.

Chapter 5

Sequence analysis of 1pcen – 1p13.2

5.1 Introduction

The production and analysis of human genomic sequence facilitates the systematic identification of genes and other functional units within the human genome. Accurate annotation of genes and characterisation of regulatory elements will not only help to identify disease genes but further our understanding of the biological systems in which all genes are involved.

Unfinished (draft) and finished genomic sequence has provided the foundation for these analyses. Analysis of draft sequence data can provide a powerful insight into a genomic landscape but because of its inherent limitations (incomplete genomic coverage, uncertain contig orientation and standard of trace data) it is preferable to use contiguous finished sequence. Finished sequence contains higher quality data: The criteria established in early genome sequencing projects (The *C. elegans* Sequencing Consortium 1998) and extended to the human genome required that all sequence be finished with an accuracy of >99.99%, leaving no gaps. This provides the best possible starting point to permit exact annotation of genes and characterisation of the genomic landscape. Furthermore, an accurate reference sequence allows for the identification of genetic variation, such as single nucleotide polymorphisms (SNPs) by comparison of additional high quality sequence traces to the finished sequence (see chapter 6). The availability of finished sequence also enables comparative analyses with other genomes to be performed. Such analyses can subsequently assist in the determination of gene structures at the sequence level and provide some insight into common evolutionary origins.

Finished sequence from 127 of the 136 PAC, BAC and cosmid minimum tile path clones, selected from the bacterial clone contig constructed in the previous chapter, enabled such a genomic analysis of 1pcen – 1p13 to be carried out. This chapter describes a detailed characterisation of repeat sequences (including high resolution GC and isochore analysis) and the localisation and annotation of known genes and identification of novel transcripts.

5.2 Sequence Composition Analysis

Eight sequence contigs, containing 127 minimum tile path clones, represented 95% coverage of 1pcen – 1p13 (11.8 Mb / 12.4 Mb) (figure 5.1a). This long range sequence continuity permitted a detailed investigation of the genomic landscape to be made, including GC profile, repeat content and CpG island identification (sequence analysis performed by James Gilbert).

5.2.1 G-Banding

Chromosome banding, produced by Giemsa staining of metaphase chromosomes, provides a means of partitioning regions of individual chromosomes for low-resolution cytogenetic mapping. Giemsa preferentially binds to AT rich regions of DNA therefore producing characteristic patterns of dark-staining or G(iemsa) bands (AT rich - GC poor) and light-staining or R(everse) bands (GC rich) (Francke *et al*., 1994). The characterisation of GC content within an interval is important feature to determine, as variation of GC between

regions has been associated with differences in biological properties such as repeat composition, gene density and structure. *In situ* analysis of 54 clones from the mapped contig (see table 5.1) confirmed the localisation of the contig to 1pcen – p13, relative to the 850 cytogenetic G-banding pattern previously reported (reviewed in Bickmore *et al*., 1989, Francke *et al*., 1994) (see fig 5.1c). Examination of the genomic sequence within 1pcen – 1p13 indicates a correlation between variations in GC content across the interval (figure 5.1b) and the G and R bands (figure 5.1c). The relative position of the light bands, 1p13.1 and 1p11.2, show a good correlation with regions in the sequence of GC content higher than the genome average of 41% (blue dotted line). Conversely, the location of dark bands 1p13.2 and 1p12 correlate with regions of below-average GC content. The designation of 1p11.1 as a grey band (containing an intermediate GC content) seems to be born out by comparison with GC within the finished sequence.

Table 5.1: Fluorescence *in situ* hybridisation data of selected bacterial clones from 1pcen – p13. Data associated with clones listed in the first column can be placed on the map via accession clones in the third column.

Clone	FISH	Acc Clone	Acc number
RP11-401O13	1p13.2	RP11-356N1	AL390036
RP11-258P6	1p13.2-1p21.1	RP11-256E16	AL160171
RP4-667F15	1p12-1p13.3	RP4-667F15	AL138933
RP4-641D22	1p13.1-1p13.3	RP11-352P4	AL356389
RP5-831G13	1p13.3-1p21.1	RP5-831G13	AL355145
RP4-6768I12	1p13.3-1p21.1	RP5-1160K1	AL355310
RP11-195M16	1p13.3-1p21.3	RP11-195M16	AL450468
RP4-742A5	1p13.2-1p21.1	RP4-742A5	AL355817
RP4-773N10	1p13.1-1p13.3	RP4-773N10	AL160006
RP5-1003J2	1p12	RP5-1003J2	AL137790
RP5-1074L1	1p13.3-1p21.1	RP5-1074L1	AL355488
RP11-498A13	1p13.2-1p21.1	RP11-498A13	AL354713
RP11-96K19	1p13.2-1p21.1	RP11-96K19	AL360270

RP5-1019F20	1p13.1-1p13.3	RP11-96K19	AL360270
RP5-1180E21	1p13.1-1p13.3	RP5-1180E21	AL355816
RP4-758H6	1p13.3	RP5-1180E21	AL355816
RP5-836N10	1p13.1-1p13.3	RP5-836N10	AL391063
RP4-773A18	1p13.2-1p21.1	RP4-773A18	AL049557
RP11-534M8	1p13.2-1p21.1	RP11-88H9	AL512665
RP4-671G15	1p13.1-1p13.3	RP4-671G15	AL354760
RP4-580L15	1p13.1-1p13.3	RP4-580L15	AL158844
RP11-31F15	1p13.1-1p13.3	RP11-31F15	AL390242
RP4-658C17	1p11.1	RP4-658C17	AL139016
RP4-730K3	1p12-1p13.3	RP4-730K3	AL133517
RP5-1073O3	1p13.1-1p13.3	RP5-1073O3	AL137856
RP5-1037B23	1p13.1-1p13.3	RP5-1037B23	AL162594
RP4-543J13	1p13.1-1p13.3	RP4-543J13	AL121999
RP4-591B8	1p13.1	RP4-591B8	AL035410
RP5-1156J9	1p12-1p13.2	RP5-1156J9	AL133382
RP5-1000E10	1p12-1p13.3	RP5-1000E10	AL096773
RP5-1165D20	1p13.1-1p13.3	RP11-350E19	AL358372
RP4-666F24	1p13.1-1p13.3	RP4-666F24	AL109660
RP4-662B22	1p12-1p13.3	RP4-662B22	AL049825
RP5-940J24	1p13.1-1p13.3	RP5-940J24	AL157950
RP11-12L8	1p12-1p13.3	RP11-12L8	AL357137
RP5-1185H19	1p13.1-1p13.3	RP5-1185H19	AL121982
RP4-787H6	1p12-1p13.2	RP4-787H6	AL355538
RP5-1086K13	1p12-1p13.2	RP5-1086K13	AL390066
RP4-655N15	1p13.1-1p13.3	RP4-655N15	AL135798
RP4-753F5	1p13.1-1p13.3	RP4-753F5	AL157904
RP4-570D9	1p12-1p13.3	RP4-570D9	AL139248
RP11-188D8	1p12-1p13.2	RP11-188D8	AL358072
RP4-675C20	1p13.2	RP4-675C20	AL157902
RP11-172A5	1p11.1-1p13.1	RP4-675C20	AL157902
RP4-757N13	1p13.1-1p13.3	RP4-757N13	AL122007
RP4-776P7	1p13.1-1p13.3	RP4-776P7	AL121993
RP5-832K2	1pcen-1p12	RP5-832K2	AL139345
RP4-730H16	1p13.1-1p13.3	RP4-730H16	AL122006
RP5-876G11	1p11.1-1p13.1	RP11-94F13	AL606843
RP4-712E4	1p11.1	RP4-712E4	AL139420
RP5-920G3	1p12-1p13.3	RP5-920G3	AL121995
RP4-599G15	1p12-1p13.2	RP4-599G15	AL109966
RP4-656M7	1p11.1-1p13.1	RP4-656M7	AL139251
RP5-1042I8	1p11.1-1p13.2	RP5-1042I8	AL359752

5.2.2 Isochores

Another means of determining the different components of GC content from the interval is by isochore analysis. It has been demonstrated that human nuclear DNA can be resolved into a number of different components based on GC content when ultra-centrifuged in Cs_2SO_4-Ag^+ gradients. These studies led Bernardi *et al*., (1985) to propose that the separated components, termed isochores, consist of long regions within which GC content is relatively homogeneous. The individual isochores were subsequently classified according to their relative GC content, i.e. light (GC poor) family members L1 and L2 contain <38% and 38-42%, GC respectively, whilst heavy (GC rich) family members, H1, H2 and H3 contain 42-47%, 47-52% and >52%, GC respectively. To determine the isochore content within 1pcen – 1p13, sequence contigs were analysed (by Jose Oliver) (Bernaola-Galvan *et al*., 1996) and the results plotted against the GC and cytogenetic landscape of the interval (figure 5.1d). Plotting of the isochore family members shows that there is a very good correlation of variation in GC content and provides an additional level of resolution in comparison to the G-banding. Though the resolution of isochore analysis is less than that of a GC profile it allows for defined regional assessments of GC analysis across the interval. Chromosome bands 1p13.2, 1p13,1, 1p12, 1p11.2 and 1p11.1 gave average GC contents based on isochore analysis that was in accordance with their banding pattern as determined by Giemsa staining i.e. L2 (39.1%), H1 (43.9%), L1 (39.7%), H1 (42.4%) and L1 (39.7%) respectively. The majority of the 11.8 Mb of finished sequence of 1pcen – 1p13.2 is contained within GC poor L family isochores (57.3%) whilst H1 isochores (32.6%) make up the majority of the H family isochore coverage, with H2 and H3 isochores contributing 8.9% and 1.2% respectively. The average percentage GC for the entire interval is 41.5% (fluctuating between 30% to 58), which is marginally above the genome average of 41% (IHGSC, 2001).

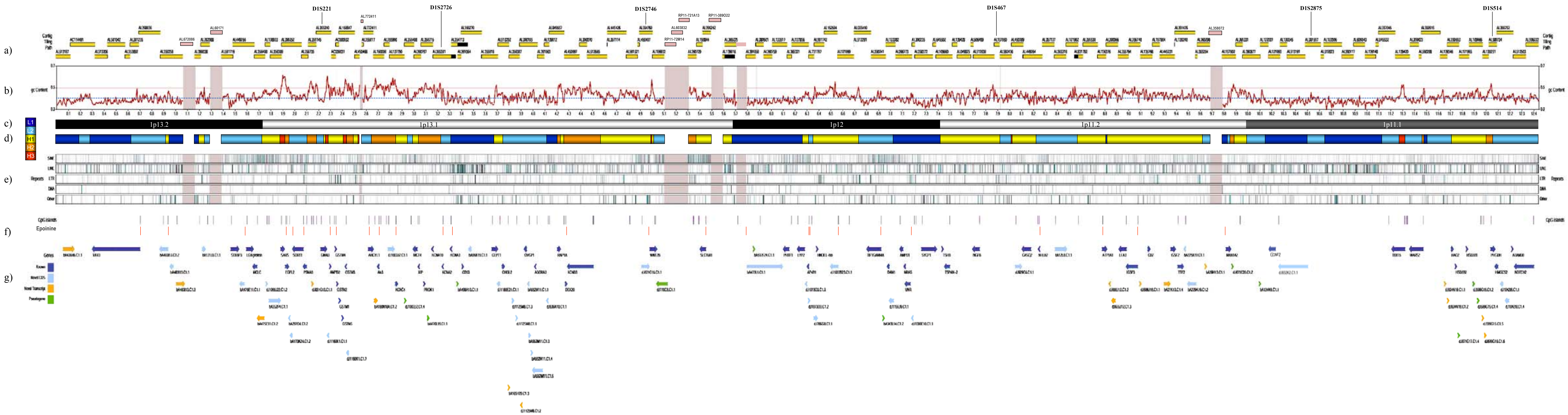

D1S221
D1S2726
D1S2746
D1S467
D1S2875
D1S514
a)
b)
c)
d)
e)
f)
g)
1p13.2
1p13.1
1p12
1p11.2
1p11.1
L1
L2
H1
H2
H3
Epoinine

Figure 5.1: The genomic characterisation of human chromosome 1pc – 1p13. Figure 5.1a) represents the framework markers and 136 minimum tile path clones from the interval (finished clones with accession numbers are yellow, unfinished clones with accessions or clone names are light pink). Variations in GC profile, b), across the region are represented as a red line with the genomic average of 41% drawn as a dotted blue line. G (dark) and R (light) banding cytogenetic patterns of from Giemsa staining are illustrated in c). The results of isochore analysis are depicted in d), dark blue band = L1 isochore, light blue = L2, yellow = H1, orange = H2 and red = H3. Transposon derived repeats, short interspersed elements (SINEs), long interspersed elements (LINEs), long terminal repeat retrotransposons (LTRs), DNA transposons (DNA) and others are represented in e). Putative promoter and transcription start sites are represented by CpG islands and Eponine predictions, respectively, in f). Genes within the interval, and their classification, are represented in g). The direction of transcription and size of the gene within genomic sequence is indicated by the direction and length of the arrow drawn above the gene name.

5.2.3. Repeats

It is estimated that repeat sequences account for approximately 50% of the human genome (IHGC 2001). Therefore, assessment of repeat type and distribution is an important factor when characterising the genomic landscape within 1pcen – 1p13. Transposon-derived repeats, which account for approximately 90% of repeats in the human genome (IHGSC, 2001), were plotted by analyzing the sequence content within a 8000bp sliding window, sampled every 4000bp, with RepeatMasker (Smit and Green, unpublished,

http://repeatmasker.genome.washington.edu) (figure 5.1e). Repeat content was divided into short interspersed elements (SINEs, including Alu repeats), long interspersed elements (LINEs), long terminal repeat retrotransposons (LTRs), DNA transposons (DNA) and others. Whilst LTRs and DNA transposons exhibit a fairly uniform distribution across 1pcen – 1p13, LINE and SINE repeats share an inversely related distribution. LINE elements conform to their reported higher distribution within AT rich, dark band regions (Smit *et al*., 1999) whilst SINE elements show a higher density in GC rich light bands. Interestingly, 1p13.1, a GC rich light band, contains a LINE 'island' which corresponds with an L1 isochore at approximately 3.5 Mb of the finished sequence link. The total repeat content within 1pcen – 1p13 of the various transposon-derived repeats is represented in table 5.2.

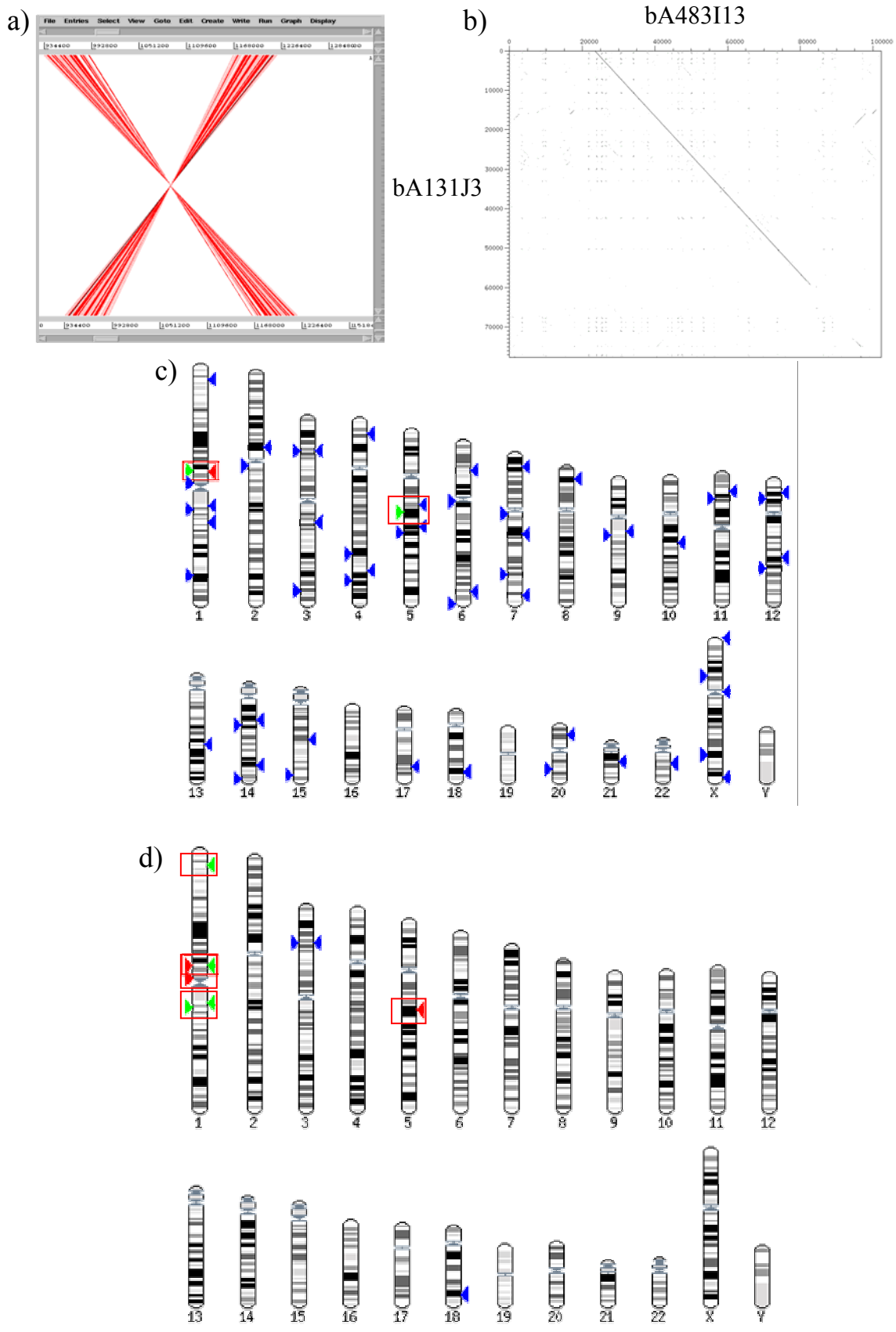
Table 5.2: The breakdown of repeat content within 1pcen – 1p13.2.

Repeat	**Mb**	**Percentage**
Alu	1.09	9.27
MIR	0.35	2.96
MIR3	0.05	0.45
Total SINE	**1.49**	**12.68**
L1	1.98	16.88
L2	0.55	4.71
L3	0.06	0.48
Total LINE	**2.59**	**22.08**
Total DNA	**0.34**	**2.91**
Total LTR	**0.79**	**6.75**
RNA	**0.00**	**0.03**
Unclassified	**0.04**	**0.33**
Total	**5.26**	**44.77%**

5.2.4. Low copy repeats

An estimated 3.3% of the human genome is duplicated in segments of greater that 1 kb with 90-99.5% sequence identity (IHGSC, 2001), with intrachromosomal duplications accounting for almost two thirds of these (i.e. 2% of the genome). To identify low copy repeats within 1pcen – 1p13 the 11.8Mb of finished sequence was initially analysed for repeats using RepeatMasker (Smit and Green, unpublished, http://ftp.genome.washington.edu/RM/RepeatMasker.html) to remove previously characterised common repeats and then compared to itself by BLAST analysis. ACT (http://www.sanger.ac.uk/Software/ACT), which was used to view intrachromosomal duplications after self-matches were removed (figure 5.2a), indicated two segmental duplications 59 kb in size and located 79 kb apart, with one copy of the repeat being inverted with respect to the second. Dotter (Sonnhammer *et al*., 1995) analysis of the repeats (figure 5.2b) indicates the size and level of sequence homology (99%) of the low copy repeat shared between bA483I13 (AL359258) and (AL390038). BLAST analysis of the one duplicated region within Ensembl (http://www.ensembl.org/) indicated that the region was also involved in an interchromosomal duplication. Results indicated a match between the two closely linked regions in1p12 (red boxes figure 5.2c) described above and an additional locus in 5q14.3 with a BLAST alignment of 99% and score of 15000. It was noted that an transcript was contained within each of the three segmentally duplicated regions. BLAST analysis of the mRNA (AK057395) within Ensembl identified an additional five high BLAST scoring loci, BLAST alignments of > 83% and scores of > 880, containing homologs to the duplicated mRNA (figure 5.2d).

The conservation in type and position of SINE repeats within the gene structures annotated from AK057395 suggests that the low copy repeat duplication arose since the divergence of *Homo sapiens* from mouse. The occurrence of Alu elements within the human genome coincides with the radiation of primates in the past 65 million years (Deininger *et al*., 1986), therefore the duplication of this region must have occurred within this period of time (figure 5.2e).


a)
b)
bA483I13
bA131J3
c)
d)

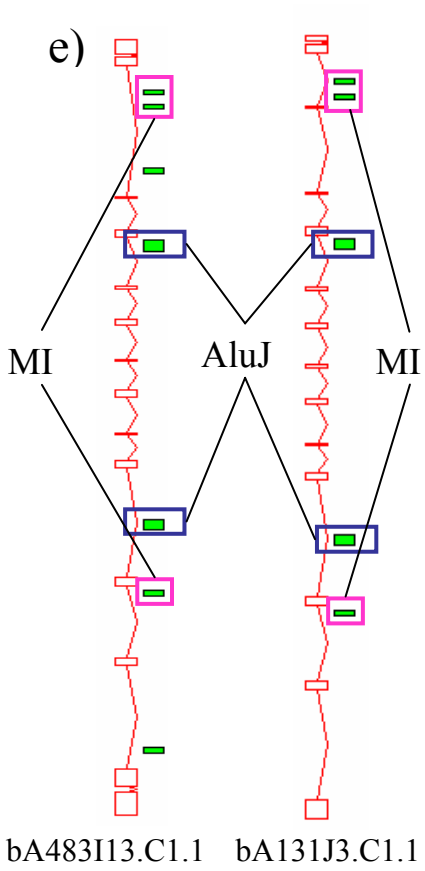


Figure 5.2: Low copy repeat detected within 1pc – 1p13. a) The relationship between two 59Kb inverted repeats, containing two novel genes, show an inverted relationship within ACT. b) Dotter displays the level of sequence homology between the low copy repeat regions. c) BLAST analysis of one of the low copy genomic repeat sequences within Ensembl identifies two regions of homology (red boxes). The adjacent region containing the repeat sequence is also identified (green arrow – chromosome 1) in addition to a homologous region on chromosome 5q14.3 (green arrow, red boxes, d). BLAST analysis of the mRNA from which the two chromosome 1 genes were derived shows high BLAST homology to two other regions of chromosome 1. e) Comparison of the repeat sequences contained within the original duplicated regions in 1pc – 1p13 reveals maintenance of SINE repeat family types

5.2.5. CpG islands

CpG islands are characteristic regions of GC-rich DNA that contain unmethylated CpG dinuclotides and are predicted to lay at the 5' ends of approximately 56% of human genes (Antequera and Bird, 1993). The occurrence of putative CpG islands (i.e. predicted by base composition, but without experimental testing of their methylation state) adjacent to the 5' ends of genes has been used as a means of identifying the sites of transcription initiation and therefore as an *in silico* assay to determine the completeness of gene annotation and, to a lesser extent, a method of estimating gene density. CpG islands within 1pcen – 1p13 were predicted (courtesy of Gos Micklem) by searching for DNA sequences of >400bp in length, >50% GC content and having an expected / observed CpG count of >0.6. A total of 94 CpG islands were predicted within 1pcen – 1p13. The distribution of CpG islands within isochores followed an expected association with GC content, with GC-poor L1 and L2 isochores containing 14 (0.57 CpG / Mb) and 20 (0.61 CpG / Mb) respectively, whilst GC-rich H1, H2 and H3 isochores contained 33 (1.01 CpG / Mb), 18 (2.02 CpG / Mb) and 9 (7.67 CpG / Mb) respectively. Detailed *in silico* annotation and experimental analysis of the region (see section 5.3) identified 102 full length gene structures, 58 of which (57%) were located adjacent to a putative CpG island. The percentage association of CpG islands to genes within the interval is very close to the predicted genome average of 56% (Antequerra and Bird, 1993). An interesting feature of the localisation of putative CpG islands within the interval is the apparent sharing of a CpG island by a pair of genes orientated on opposite strands of DNA suggesting the presence of a possible bi-directional promoter sequence (see section 5.4.2.1).

5.2.6 Eponine

Eponine (Down and Hubbard, 2002) was used to predict promoter regions associated with genes in the 1pcen – 1p13 region. The program is designed for detecting transcription start sites (TSSs) in human genomic sequence by identifying promoter core motifs within a 600 bp window located at the 5' ends of genes. Eponine is reported as having a >50% sensitivity of detecting annotated mRNA start sites based on human chromosome 22 data used in its design. A total of 70 TSS predictions were predicted within Ensembl (build 30) for clones making up the sequence link objects within 1pcen – 1p13. Of this number, 26 (37%) were associated with the 5' ends of complete genes (see section 5.4) and 17 (65%) corresponded with CpG islands, figure 5.1f.

5.3 Gene Identification

Having determined GC and CpG island content, *ab initi*o gene prediction programs and sequence homology matching were used to identify coding features within the finished sequence. RepeatMasker was used to filter out transposon-derived repeats prior to alignment of all known protein and nucleotide sequences (EST and cDNA) via BLASTX and BLASTN, respectively (Altschul *et al*., 1990). In parallel, gene prediction, using FGenesH (Solovyev *et al*., 1995) and GENSCAN (Burge *et al*., 1997), and exon prediction, using Hexon (Solovyev *et al*., 1994) and GRAIL (Uberbacher *et al*., 1996), was carried out to elucidate putative genes

or exons. The results of genomic sequence analysis were then assimilated and visualised within ACeDB (figure 5.3) (Durbin and Thierry-Meig 1994).

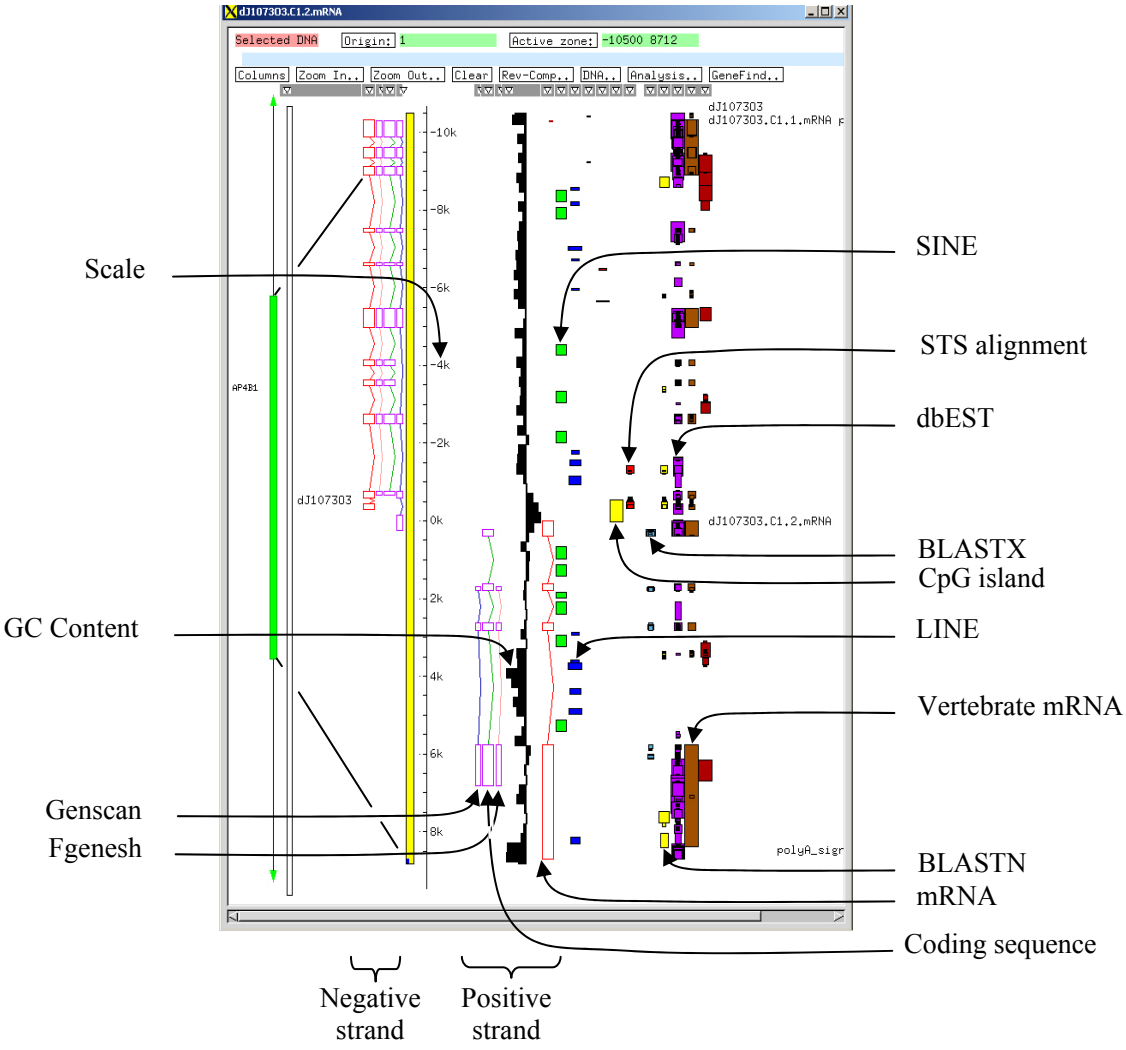


Figure 5.3: An ACeDB display of two annotated genes, including coding sequences, on opposite strands of DNA. Represented are LINE and SINE repeats as well as GC content and CpG island predictions. Vertebrate mRNAs, ESTs and STSs are positioned on genomic sequence by BLAST alignment.

Putative coding features (figure 5.1g), identified by *in silicoo* analysis and experimental support, were manually annotated and classified according to the level of coding support and completeness of gene structure. The features were divided into four categories: 'known' genes, for which an identical cDNA or protein sequence has been aligned to genomic sequence; 'novel' genes, those that contain an open reading frame (ORF), are identical to two or more splicing human ESTs, and/or have homology to genes or proteins from other species; 'novel' transcripts, similar to novel genes but an ORF cannot be determined; and 'pseudogenes', sequences that are homologous to known genes but with a disputed ORF. Manual annotation of these features involved overlaying correct gene structures onto the genomic sequence by accurately locating exon / intron boundaries of mRNAs and splicing ESTs, reviewing and resolving conflicts, and, where there was sufficient supporting data available, identifying 5' and 3' termini of genes.

5.3.1. Known genes

A total of sixty-seven known genes were localised to the interval by BLASTN matching of mRNAs at 100% alignment to genomic sequence. Table 5.3 includes the names of known genes, the accession number associated with the full length mRNA and the form of mRNA submission The majority of the genes have official human genome nomenclature committee (HGNC) names (http://www.gene.ucl.ac.uk/nomenclature/), whilst italicised genes are those that have Locus Link entries associated submitted mRNAs (http://www.ncbi.nlm.nih.gov/LocusLink/) but for which an official gene name has not been assigned. Names in parentheses are original gene names as represented on figure 5.1g.

Table 5.3: Known genes localising to 1pcen – 1p13. Italic symbols denote interim gene name.

Gene Name	Gene	Acc. #	Reference
Vav 3 oncogene	VAV3	AF067817	Trenkle *et al*., 2000
Syntaxin binding protein 3	STXBP3	D63506	Gengyo-Ando *et al*., 1996
LGN protein	*LGN*	U54999	Mochizuki *et al*., 1996
Mid-1-related chloride channel 1	*MCLC*	BC002939	Direct Submission
Seryl-tRNA synthetase	SARS	BC000716	Direct Submission
EGF LAG seven-pass G-type receptor 2	EGFL2	AF234887	Direct Submission
Sortilin 1	SORT1	X98248	Petersen *et al*., 1997
Proteasome (prosome, macropain) subunit, alpha type, 5	PSMA5	X61970	DeMartino *et al*., 1991
Guanine nucleotide binding protein (G protein), alpha inhibiting activity polypeptide 3	GNAI3	M27543	Sparkes *et al*., 1987
Adenosine monophosphate deaminase 2 (isoform L)	AMPD2	U16272	Van den Bergh *et al*., 1995
Glutathione S-transferase M4	GSTM4	BC015513	Direct Submission
Glutathione S-transferase M2	GSTM2	M63509	Vorachek *et al*., 1991
Glutathione S-transferase M1	GSTM1	J03817	Seidegard *et al*., 1988
Glutathione S-transferase M5	GSTM5	L02321	Takahashi *et al*., 1993
Glutathione S-transferase M3	GSTM3	BC000088	Direct Submission
S-adenosylhomocysteine hydrolase-like 1	AHCYL1	AF315687	Dekker *et al*., 2002
Aristaless-like homeobox 3	Alx3	AF008203	Direct Submission
Potassium voltage-gated channel, Shaw-related subfamily, member 4	KCNC4	M64676	Vega-Saenz de Miera *et al*., 1992
Solute carrier family 16 (monocarboxylic acid transporters), member 4	SLC16A4 (MCT4)	U59185	Direct Submission
Hepatitis B virus x interacting protein	HBXIP (XIP)	XM_059235	Direct Submission
Prokineticin 1	PROK1	AF333024	Direct Submission
Potassium voltage-gated channel, shaker-related subfamily, member 10	KCNA10	U96110	Orias *et al*., 1997
Potassium voltage-gated channel, shaker-related subfamily, member 2	KCNA2	L02752	Ramashwami *et al*., 1990
Potassium voltage-gated channel, shaker-related subfamily, member 3	KCNA3	M85217	Attali *et al*., 1992
CD53 antigen	CD53	M37033	Angelisova *et al*., 1990
Choline/ethanolaminephosphotransferase	*CEPT1*	AF068302	Henneberry *et al*., 1999
Chitinase 3-like 2	CHI3L2	U49835	Hu *et al*., 1992
Oviductal glycoprotein 1	OVGP1	U09550	Direct Submission
Adenosine A3 receptor	ADORA3	L22607	Salvatore *et al*., 1993
RAP1A, member of RAS oncogene family	RAP1A	M22995	Kitayama *et al*., 1989
DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 20	DDX20	AF171063	Charroux *et al*., 1999
Potassium voltage-gated channel, Shal-related subfamily, member 3	KCND3	AF120491	Isbrandt *et al*., 2000
Wingless-type MMTV integration site family, member 2B	WNT2B	AB045116	Direct Submission

Solute carrier family 16 (monocarboxylic acid transporters), member 1	SLC16A1	AL162079	Direct Submission
Putative homeodomain transcription factor 1	PHTF1	AJ011863	Raich *et al*., 1999
Protein tyrosine phosphatase, non-receptor type 22 (lymphoid)	PTPN22 (LYP2)	AF077031	Direct Submission
Adaptor-related protein complex 4, beta 1 subunit	AP4B1	AF092094	Dell'Angelica *et al*., 1999
HNOEL-iso protein	*HNOEL-iso*	AF201945	Direct Submission
Tripartite motif-containing 33	TRIM33 (TIF1GAMMA)	AF220137	Reymond *et al*., 2001
Breast carcinoma amplified sequence 2	BCAS2 (DAM1)	AB020623	Nagasaki *et al*., 1999
Adenosine monophosphate Deaminase 1 (isoform M)	AMPD1	M60092	Sabina *et al*., 1992
Neuroblastoma RAS viral (v-ras) oncogene homolog	NRAS	X02751	Hall *et al*., 1985
NRAS-related gene	*UNR*	AB020692	Nagase *et al*., 1998
Synaptonemal complex protein 1	SYCP1	D67035	Kondoh *et al*., 1997
Thyroid stimulating hormone, beta	TSHB	M23671	Direct Submission
Tetraspan 2	*TSPAN-2*	BC021675	Direct Submission
Nerve growth factor, beta polypeptide	NGFB	X52599	Direct Submission
Calsequestrin 2 (cardiac muscle)	CASQ2	D55655	Direct Submission
Nescient helix loop helix 2	NHLH2	M97508	Brown et al 1992
ATPase, Na+/K+ transporting, alpha 1 polypeptide	ATP1A1	BC003077	Direct Submission
CD58 antigen, (lymphocyte function-associated antigen 3)	CD58 (LFA3)	Y00636	Wallner *et al*., 1987
Immunoglobulin superfamily, member 3	IGSF3	AF031174	Saupe *et al*., 1998
CD2 antigen (p50), sheep red blood cell receptor	CD2	M16445	Seed *et al*., 1987
Immunoglobulin superfamily, member 2	IGSF2	Z33642	Direct Submission
Transcription termination factor, RNA polymerase II	TTF2	AF080255	Direct Submission
Mannosidase, alpha, class 1A, member 2	MAN1A2	AF027156	Tremblay *et al*., 1998
Ganglioside induced differentiation associated protein 2	GDAP2	AK000149	Direct Submission
WD repeat domain 3	WDR3	AF083217	Claudio *et al*., 1999
T-box 15	TBX15	AK096396	Direct Submission
Tryptophanyl tRNA synthetase 2 (mitochondrial)	WARS2	AJ242739	Direct Submission
Hydroxyacid oxidase 2 (long chain)	HAO2	AF231917	Jones *et al*., 2000
Hydroxy-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 2	HSD3B2	M77144	Lachance *et al*., 1991
Hydroxy-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 1	HSD3B1	S45679	Dumont et al 1992
Phosphoglycerate dehydrogenase	PHGDH	BC011262	Direct Submission
3-hydroxy-3-methylglutaryl-Coenzyme A synthase 2 (mitochondrial)	HMGCS2	X83618	Direct Submission
A disintegrin and metalloproteinase domain 30	ADAM30	AF171933	Direct Submission
Notch homolog 2 (Drosophila)	NOTCH2	AF315356	Direct Submission

5.3.2. Novel genes

Novel genes (complete gene structures containing ORFs) were annotated from supporting evidence such as splicing EST and mRNA alignment or by the addition of *de novo* cDNA sequence. The cDNA clones, from which the *de novo* cDNA sequence was generated, were identified by pooled cDNA library screening with 41 primer pairs designed to exons contained within putative gene structures. The cDNA libraries, each of which represented 500,000 clones from nine different tissue types, were initially divided into twenty-five pools containing 20,000 cDNA clones and then recombined into superpools containing 100,000 clones (kindly provided by Jackie Bye). Superpools that were positive from initial exon specific cDNA library screening were then used to generate PCR products that linked between exons (link PCR) which were subsequently sequenced and aligned to the genomic structure of the gene. Validation of possible gene structures by the alignment of sequence from splicing ESTs, mRNAs or the *de novo* cDNA sequence resulted in the identification 35 novel genes.

Table 5.4 represents a summary of cDNA library screening and link PCR results from novel genes within 1pcen – 1p13. Where possible, primers were designed to satisfy previously established criteria (see section 2.6.1). Of the 96 cDNA primary pool screens, 71% (68) identified at least one cDNA library (see section 2.8.3). Libraries that yielded a PCR product (bold denoting a strong band on an agarose gel) are listed next to each primer, with the red lettered library being used as template for the link PCR experiment. Primer combinations used in the link PCR depended on the type of validation or extension required for each putative coding feature. A vectorette primer, 224, was used in combination with sense or anti-sense primers to extend the putative genes to 3’ or 5’ UTR respectively (figure 5.4).

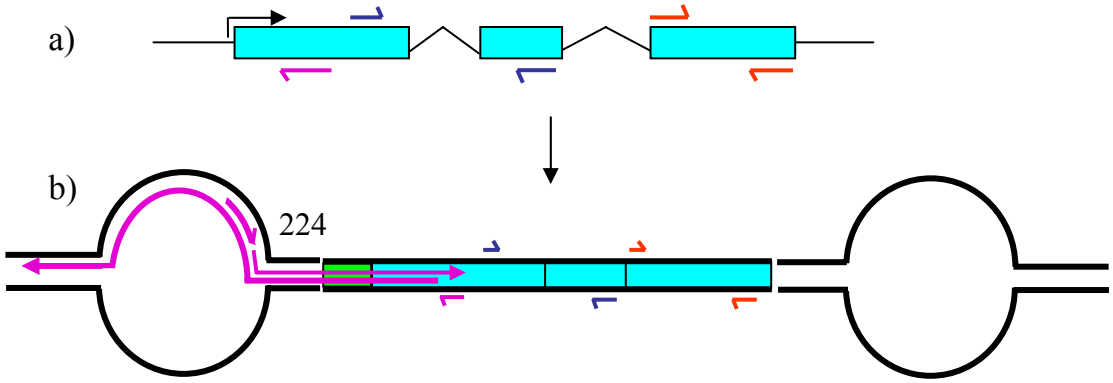


Figure 5.4: Primer combinations used to validate putative gene structures. a) Primers pairs, of the same colour, are designed to the annotated gene structure. Blue primers are designed between exons, red primers within an exon and the pink primer is designed to be used in conjunction with vectorette primer 224. Black arrow indicates the direction of transcription. b) cDNA clone with ligated vectorette arms. Exon specific primer (pink) anneals and elongates through the non-complementary vectorette arm before the 224 vector primer can anneal and elongate in the reverse direction. A normal PCR reaction from these initial templates then follows. Novel 5' cDNA sequence is represented in green.

Primer combinations within genes were also used to validate gene structures.
35% of the 93 vectorette and link PCR reactions resulted in the generation of single strong PCR product when run on a 2.5% agarose gel (Y in the Link' column of table 5.4) which was subsequently purified and sequenced (by others). A further 25% of PCR reactions generated a faint single band (R in the 'Link' column) which requires re-amplification prior to sequencing. Finally, 40% of the gene validation experiments (M within the Link column of table 5.4) resulted in the generation of multiple Link PCR products; these require refinement

of primer design (because of possible mispriming events) or an increase of the PCR Tm to increase the specificity of primer annealing. 81% (25) of the sequenced products yielded sequence which was subsequently aligned to the interval by BLAST analysis. Attempts were made to generate experimental data for 17 of the final total of 35 novel genes by cDNA screening. Sequence from link or extension PCR was generated from 12 (71%) of these possible gene structures, including the 6 genes with 5' or 3' extensions. Unsuccessful attempts were also made to extend five known genes (dark blue - table 5.4) in the 5' direction so as to increase the size of 5' UTR.

Table 5.4: cDNA primary pool and link PCR screening results. Gene structures are coloured according to their final category as drawn in figure 5.2. Columns correspond to gene name, the exon from which primer pairs were designed, the EMBL accession number associated with the primer pairs, whether a PCR product was generated from cDNA superpool screening (Y = yes, N = no), the cDNA library that yielded a PCR product, whether a vectorette (before /) or link PCR product (after /) was generated (Y = single strong band, R = faint single band, M = multiple bands) and whether the product was successfully sequenced. For key to cDNA library codes, see methods table 2.2

Gene	Exon	stSG	1^0	cDNA Library	Link	Seq
bA483I13.C1.1.mRNA	e2	452926	Y	**TD**	**Y/**	**N**
bA483I13.C1.3.mRNA	e2	452927	N			
	e5	452928	Y	**HeLa B,C,E**	**R/**	
	e7	452929	N			
bA475E11.C1.2.mRNA	e1 5'UTR	452930	Y	FLU A, HP B-E, **SK C**		
	e4	452931	Y	**FLU A**, T A, HPB B,**C**,D,E, **SK C**	**Y/**	**N**
	e7	452932	Y	AH E		
	e10	452933	Y	AK B,E, **AH A**,**D, He La E**, **T C,** HP B-E, **SK A-E**		
	e12	452934	Y	AK B, **AH A,D,** He La A,**E, T C,** U E, HPB B-E, SK A-E		

	e14 3'UTR	452935	Y	AK B,E, AH **A,B**,C,**D,E**, He La A,**E**, T **A,C,E,** HPB B-E, **SK A-E**		
bA475E11.C1.1.mRNA	e1 5'UTR	452936	N			
	e5	452937	N			
	e6	452938	N			
	e10	452939	N			
	e13	452940	N			
bA297O4.C1.1.mRNA	e1 5'UTR	452941	N			
dJ831G13.C1.1.mRNA	e1	452942	Y	**T D**	**R/M**	
	e3	452943	Y	FLU D, **T** B,**D**		
	e4	452944	Y	**T D**	**R/Y**	**N**
bA180N18A.C1.2.mRNA	e1	452945	N			
dJ773N10.C1.1.mRNA	e2	452946	Y	**SK C-E**	**R/**	
dJ1003J2.C1.1.mRNA	e2	452947	Y	**FLU D**	**R/N**	
	e4	452948	Y	**FLU D**		
	e7	452949	N			
	e9	452950	Y	FLU C, T E, **SK D**	**N/N**	
	e12	452951	N			
bA470L19.C1.2.mRNA	e1	452952	N			
bA284N8.C1.1.1/.2.mRNA	e2	452953	Y	**SK B**	**Y/**	**N**
bA165H20.C1.3.mRNA	e3	452954	Y	**AH A,E**	**Y/Y**	**Y**
	e5	452955	N			
	e7	452956	Y	**FLU D, AH A-E**	**M/Y**	**Y**
	e10	452957	N			
dJ1125M8.C1.1.mRNA	e2	452958	Y	**FLU A-C,**D,**E, T E**	**M/N**	
	e4	452959	Y	**AK C,D**, **FLU A-E, AH E**	**/Y**	**Y**
	e6 3'UTR	452960	Y	**FLU A**,B,**D,E, AH C,E**	**M/Y**	**Y**
dJ1125M8.C1.2.mRNA	e1	452961	Y	AK A,B,D, AH C, HP B	**M/**	
	e3	452962	N			
	e5	452963	N			
	e8	452964	N			
bA552M11.C1.4.1/.2.mRNA	e2 .2	452965	N			
	e3/4 .1	452966	Y	**AK A-D, FLU A, U C, SK B**	**M/Y**	**N**
	e5	452967	Y	**AK A,B,** FLU A, U C	**M/N**	
bA552M11.C1.5.mRNA	e1/2	452968	Y	**T A-E**, SK A-C	**Y/Y**	**Y**
	e3/4	452969	Y	**FLU A,C**, **T A-E**	**/Y**	**Y**
	e5	452970	Y	**FLU A-C**, **T A-E**	**R/Y**	**Y**
dJ836N10.C1.1.mRNA	e2/3	452971	Y	**T E**	**R/Y**	**Y**
	e4	452972	Y	**T E**	**R/Y**	**Y**
dJ1073O3.C1.3.mRNA	e1	452973	Y	**AK A,C**, T E	**Y/Y**	**Y**
	e3	452974	Y	**AK A,C**, SK A	**/Y**	**Y**
dJ1037B23.C1.1.mRNA	e2	452975	N			
	e4	452976	Y	T A, **SK C**	**R/Y**	**Y**
	e6	452977	Y	**SK C**	**/Y**	**Y**

	e8 3'UTR	452978	Y	AH C-E, T E, U C, **SK C**,D	**M/Y**	**Y**
dJ1156J9.C1.1.mRNA	e1 5'UTR	452979	Y	**AH D, T C, SK B**	**Y/**	**Y**
dJ929G5.C1.1.mRNA	e2	452980	Y	**He La** C,D,**E,** T C	**M/Y**	**Y**
	e4	452981	Y	**AK D,E, FLU D,** AH A,**B,C, He La A-E, T B,C, U A-E, HPB A-E, SK A**	**/Y**	**Y**
	e6	452982	Y	**AK D,E, FLU D,** AH A,**B,C**,D, **AB A,B,D,E, He La A,B,D,E, T B--D, U A-E, HPB A-E, SK A**	**/Y**	**Y**
	e8	452983	Y	**AK D,E, FLU D,** AH **A-C**,D, **AB A,B,D,E, He La A,B,D,E, T B-D,**E, **U A-E, HPB A-E, SK A**	**M/N**	
bA12L8.C1.1.mRNA	e2	452984	Y	**FLU C,** AH C, T E, HPB E, SK E	**R/**	
dJ655J12.C1.2.mRNA	e1	452985	N			
	e2	452986	N			
dJ655J12.C1.3.mRNA	e2	452987	Y	**AKA,C-E**	**M/**	
dJ686J16.C1.1.mRNA	e1/2	452988	Y	T E	**M/**	
bA39H13.C1.1.mRNA	e1	452989	Y	T E	**R/R**	
	e2	452990	Y	**T E**	**R/R**	
bA42I21.C1.1.mRNA	e1	452991	Y	**AK D,E**, AH C	**N/**	
	e2	452992	Y	AH C, AB A, T E, HPB A		
dJ776P7.C1.1.mRNA	e1	452993	Y	**AK A,B,E**, **T** A,C,**D,E**	**R/N**	
	e1/2	452994	N			
	e4	452995	Y	**T D,E,**	**Y/**	**Y**
dJ832K2.C1.1.mRNA	e1	452996	Y	AH C, **T C,E**	**M/Y**	**Y**
	e6/7	452997	N			
dJ832K2.C1.2.mRNA	e2	452998	Y	AH C, **T C,E**, SK C,E	**M/Y**	**Y**
dJ832K2.C1.3.mRNA	e2	452999	Y	**FLU C,E**, AH E, **T D**	**M/N**	
	e5	453000	Y	AH E, **T D,E,** SK C	**/N**	
	e8	453001	Y	FLU B, **T D**	**Y/R**	**Y**
bA224F24.C1.1.mRNA	e1	453002	Y	**AK E**	**Y/N**	**Y**
	e4	453003	Y	**AK E**	**/Y**	**N**
	e6	453004	N			
	e8	453005	N			
	e11	453006	Y	**AK E,** T D,E	**Y/Y**	**Y**
	e15	453007	N			
dJ794L19.C1.1.mRNA	e1	453008	Y	AH E	**M/R**	
	e3	453009	Y	T E	**/R**	
	e5	453010	Y	AH B,E	**/N**	
	e8	453011	Y	AH E	**M/N**	
dJ834N19.C1.1.mRNA	e1	453012	N			
	e3	453013	Y	AK C,D, He La E, T A,E, **HPB B,D,E**	**R/R**	

dJ834N19.C1.2.mRNA	e2	453014	Y	AH C-E, T E	**R/R**	
dJ599G15.C1.5.mRNA	e1	453015	N			
dJ599G15.C1.6.mRNA	e1	453016	Y	FLU D, T E, HPB B	**R/**	
dJ1042I8.C1.4.mRNA	e2	453017	Y	**AK C, FLU A,D, AH B,D,E, He La A-E, T C,E, U A,B,D, HPB A,B,D, SK A,C-E**	**M/N**	
	e3	453018	Y	AK E, AH C,E, T E, SK E		
	e5	453019	Y	**AK C, FLU B,D,E, AH A,B,D,E, He La A,D,E, T C,** HPB A,**B**,C,D	**/N**	
	e7	453020	Y	T D,E		
	e9	453021	Y	AH C, T E, U D, HPB A	**M/N**	

5.3.2.1 Splicing ESTs support the structure of a gene

A proportion of the total number of novel genes identified within 1pcen – 1p13 were initially annotated as incomplete gene structures based upon *in silico* gene prediction and BLAST alignment of splicing ESTs to genomic sequence. Figure 5.5 outlines an example of how experimental support was generated for a putative gene feature, bA552M11.C1.5.mRNA, originally annotated from *in silico* prediction (figure 5.5a) and EST alignment (figure 5.5b). Primers were designed, where possible, to predicted exons (arrows – figure 5.5c) and screened across the 45 cDNA library superpools (figure 5.5d) to experimentally establish the full length gene structure. PCR results indicated (figure 5.5d - red arrows) that testis cDNA library superpools A-E were all positive when tested with each of the three primer pairs, whilst primer pair SG452969 provided an additional positive result for foetal liver A and C, and primer pair SG452970 provided a positive result for foetal liver A – C. PCR from primer SG452969 generated an additional faint band of an unexpected size at approximately 400bp (figure 5.5d - blue arrow). BLAST analysis indicated that the sense primer of SG452969 localised to 24 positions in the genome at a high BLAST score (maximally within 1pcen –

1p13) which may have resulted in the secondary product being generated from a mispriming event, unlike the other primers which were unique in the genome by BLAST analysis. The absence of a positive control band for primers SG452968 (1908bp) and SG452969 (848bp) could be attributed to the inability of the PCR reaction to produce a product from the genomic positive control by primer pairs which have been designed in separate exons.

Link PCR was performed using a combination of primer pairs (figure 5.5d) to both validate and extend the putative coding structure. Testis super pool A was used as the template for generating link PCR products for each of the three primer pairs. A primer designed to the vectorette bubble ligated to cDNA subclones, 224, was used to prime from the 5' end of the cDNA clone. Products from the PCR reaction were excised from agarose and sequenced (by others) and aligned to the genomic sequence. cDNA sequence supported the annotated gene structure, elongated the gene to incorporate a new 5' exon (spanning an existing known gene, ADORA3) and identified a novel splice variant. Subsequent BLAST analysis of the new gene identified an Image 5' cDNA clone, BI463020, (brown gene structure figure 5.5f) which, when aligned to genomic sequence, supported the annotation of the gene and the coding region (figure 5.5f green gene structure).

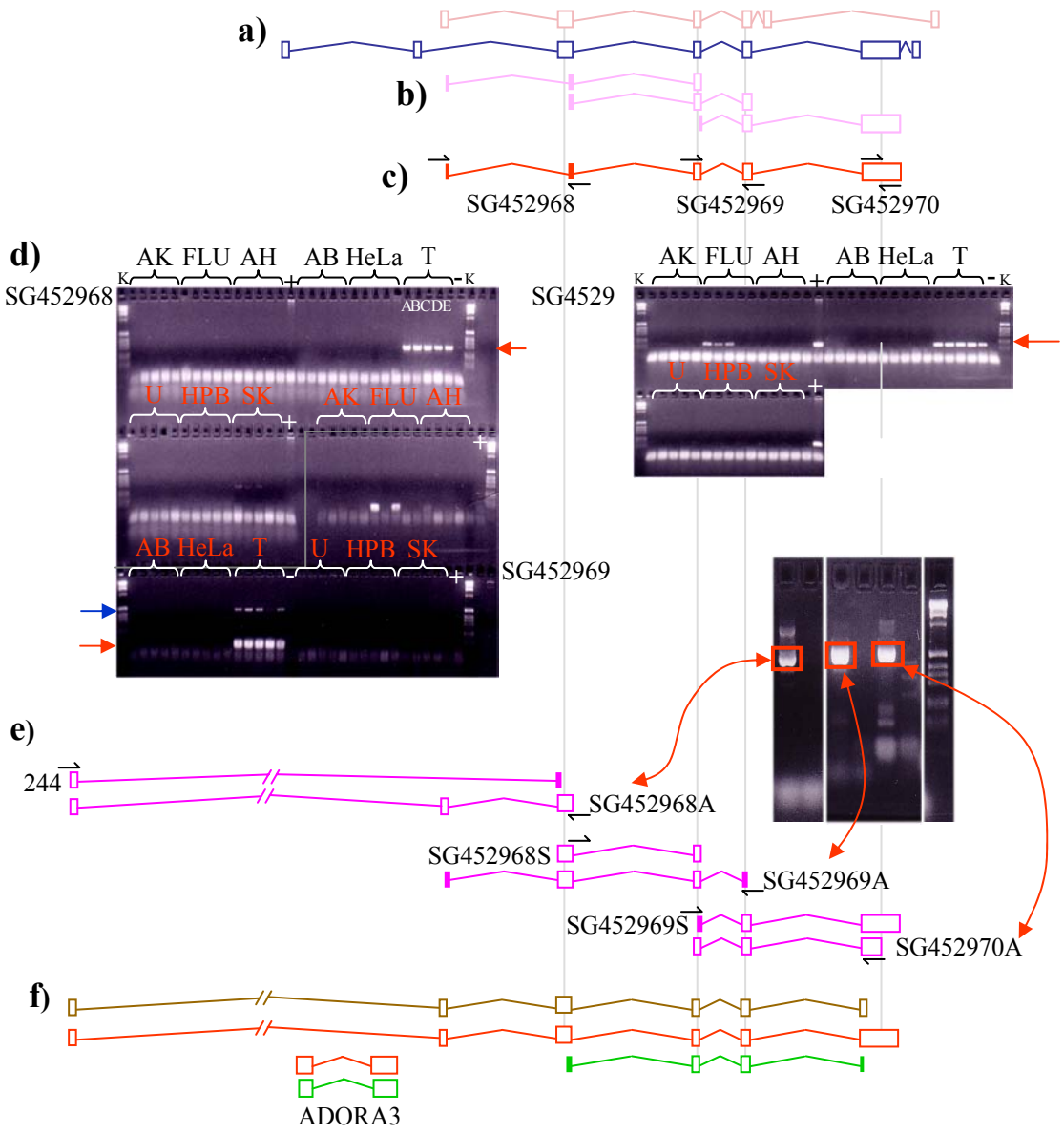


Figure 5.5: The annotation of a novel gene from *de novo* prediction and splicing EST alignment. a) *In silico* prediction of a novel gene is represented by Fgenesh (light pink) and GENSCAN (dark blue) structures. b) Alignment of splicing ESTs (pink) supports the presence of a gene. c) Annotation of a putative gene (red) enabled primer pairs to be designed (black arrows and accession numbers). d) PCR products from cDNA library screening, including negative (-) and positive controls (+), are run on a 2% agarose gel. e) Selected

cDNA libraries, from d), were then used as a template for the generation of vector and link PCR products using primer combinations (black arrows), which were then sequenced. f) The final gene structure is represented with the full length gene drawn in red, the coding in green and a newly submitted cDNA clone, BI463020, brown structure.

5.3.2.2. mRNA support of novel coding features

Genomic alignment of incomplete mRNAs derived either from human or other species can facilitate the identification of novel genes by providing experimental support for *in silico* predictions and splicing ESTs. Initial analysis of *in silico* prediction (figure 5.6a), EST (figure 5.6b) and mRNA alignment (figure 5.6c) to three overlapping sequence clones (RP11-224F24, RP5-832K2 and RP4-776P7) resulted in the annotation of four gene structures within 230 kb of each other on the same DNA strand (figure 5.6d). Three of the four putative coding features were based on overlapping splicing ESTs and the fourth by alignment of a novel incomplete mouse mRNA. Primers were designed, where possible, to predicted exons and, as previously described, screened across nine different cDNA libraries. Four of the twelve primer pairs (figure 5.6e) and table 5.4, 452996 – 453007) failed to generate products from cDNA library screening. Link PCR between putative coding structures was attempted because of the likelihood that they contributed to a single gene due to the orientation and proximity of these genes within a GC / gene poor band in which gene density is reportedly lower (IHGSC, 2001). Link PCR sequence derived from cDNA clones from within super pool testis C (figure 5.6f) – 452998 and 453001 (not shown)) facilitated the joining of gene features 3 and 4. BLAST analysis of GENSCAN and Fgenesh exon predictions that were not

supported by an mRNA or splicing EST identified a recently deposited partial human mRNA which, when aligned to genomic sequence, spanned features 2, 3 and 4 and overlapped the mouse mRNA used to annotate feature 1 by 90bp. The putative structure now had experimental support from a novel mouse mRNA (AK016477), which did not have a previously described translation stop site, and a novel human mRNA (AL833485) (figure 5.6g) which lacked a translation start site, splicing ESTs and novel cDNA sequence. A second iteration of BLAST searching identified a recently submitted human mRNA (AK091816) (figure 5.6g) which supported the structure annotated from mouse mRNA homology and which overlapped the downstream human mRNA. The full length gene, dJ832K2.C1.1.1, contains 49 exons spanning 230 kb, is adjacent to a predicted CpG island and contains a polyA signal and polyA site. BLAST analysis of the sequence contained within the 6.7 kb ORF, or the translated protein derived from it, failed to show homology to any known gene, (figure 5.6h).

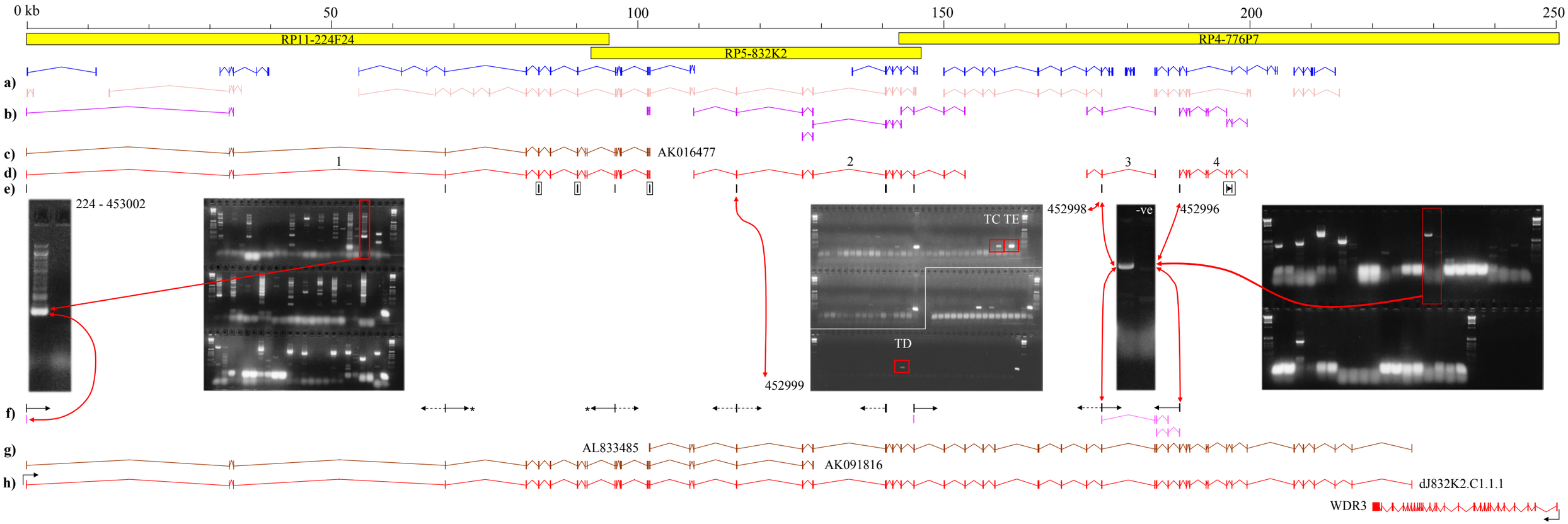

0 kb
50
100
150
200
250
RP11-224F24
RP5-832K2
RP4-776P7
a)
b)
c)
d)
e)
AK016477
1
2
3
4
224 - 453002
452998
-ve
452996
TC TE
TD
452999
f)
g)
AL833485
AK091816
h)
dJ832K2.C1.1.1
WDR3

Figure 5.6: The annotation of a novel gene from *de novo* prediction, splicing EST and homologous mRNA alignment. a) *In silico* prediction of four distinct novel genes is represented by Fgenesh (light pink) and GENSCAN (dark blue) structures within three overlapping sequence clones (yellow boxes). b) Alignment of splicing ESTs and, c), a mouse mRNA to the genomic sequence supports *in silico* predictions and the presence of the four genes. d) Putative genes were annotated and, e),exon flanking primer pairs were designed and screened across cDNA libraries (the center gel is an example of cDNA library screening with two primer pairs). Primers that failed to produce a PCR product from a cDNA library are boxed. f) Selected cDNA libraries were then used as a template for the generation of vector and link PCR products. Arrows indicate where primer combinations successfully generated PCR products, dotted lines indicate link PCR failure and the arrow with an asterisk corresponds to the generation of a link PCR product but which subsequently failed to sequence. Vertical pink lines with no link to adjacent exons indicate where exon specific sequence was generated from 244 vector priming. g) Alignment of novel mRNA sequence supports the final gene structure, h), which was shown to over lap at its 3' end with an adjacent gene, WDR3.

5.3.3 Novel transcripts

cDNA library screening was also used to identify ORFs within gene structures which had been initially annotated as novel transcripts, i.e. genes which are similar to novel genes but for which an ORF cannot be identified. Following experimental analysis 16 gene structures remained in the novel transcript category. Three quarters of the final number of putative genes

(12 / 16) identified a cDNA clone within the library pools (table 5.4), but only 1 yielded any sequence from link PCR, but did not identify an ORF within the putative structure.

5.3.4 Pseudogenes

A total of 11 pseudogenes were identified within 1pcen – 1p13. These gene structures were the result of either insertion of a processed mRNA or an unspliced feature that has an interrupted open reading frame. Figure 5.7 is an example of a processed pseudogene in which the original coding structure is present in another region of the genome, in this case elsewhere on human chromosome 1. The example shown relates to the marker used by Brintnell *et al*., (1997) (see chapter 4.4.1) to construct a YAC map within 1pcen – 1p13. As previously described, D1S3347 was derived from the 3' end of a gene, proline-rich nuclear receptor co-regulatory protein 2 (PNRC2) (pink box figure 5.7). The full length gene (figure 5.7a) is derived from 1p35.3, whilst the processed mRNA (figure 5.7b) has been incorporated in the genomic DNA in 1p12 (figure 5.7c). Evidence of mRNA processing can be found by the presence of a polyA tail incorporated into the genomic sequence (figure 5.7d). Whilst it may be possible that the insertion of a processed mRNA into the genomic sequence may result in continued expression of the gene, in this instance the open reading frame of the PNRC2 transcript is disrupted by the occurrence of multiple *de novo* stop codons.

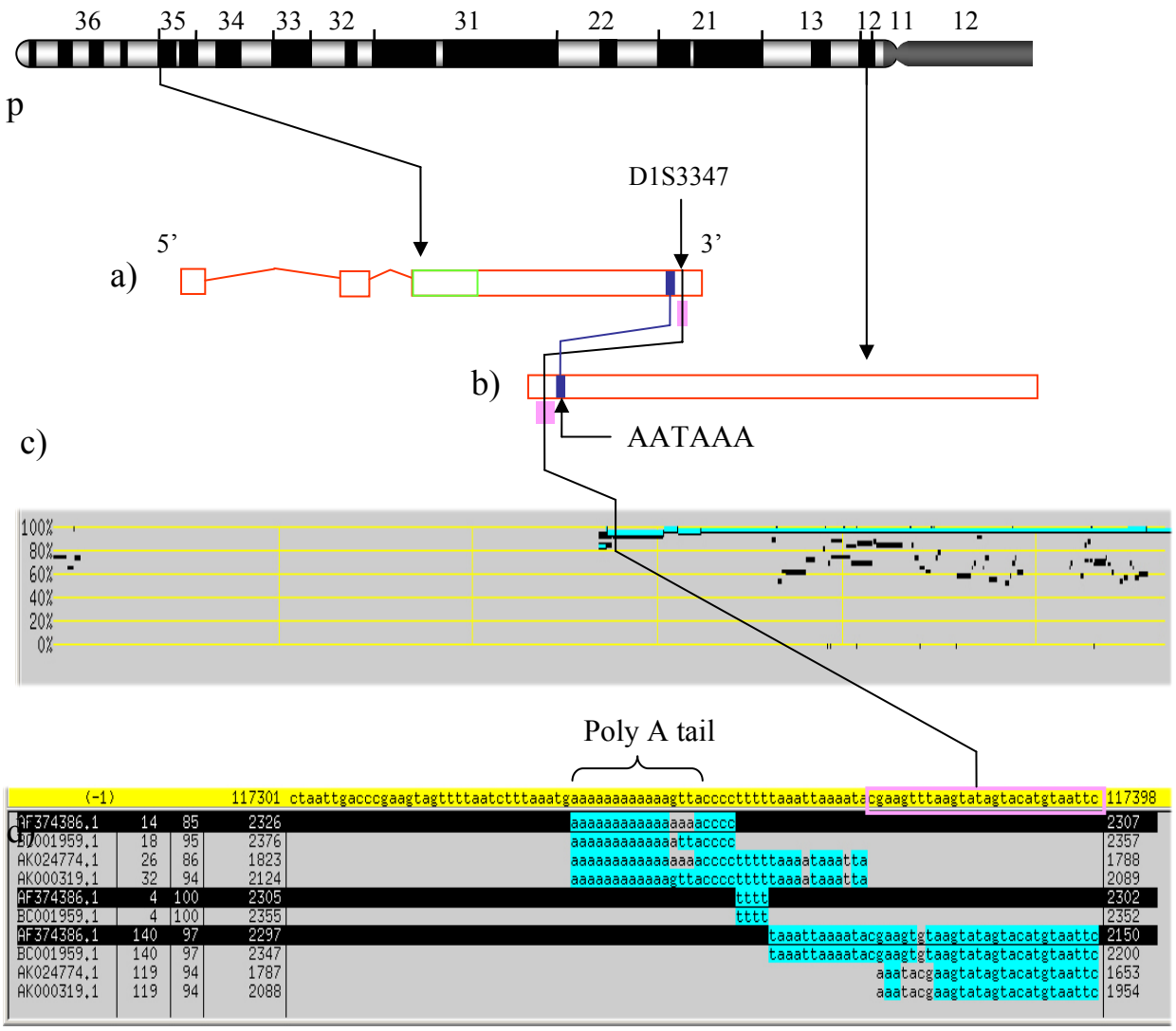


Figure 5.7: The characterisation of a processed pseudogene to 1p12, the original of which localised to 1p35. Arrows indicate where the functional gene, a), and pseudogene, b), are located on chromosome 1. D1S3347 was derived from the 3' UTR of the functional gene. Evidence that the pseudogene is processed originates from the identification of a polyA signal and the presence of a polyA tail, which are added during pre-mRNA processing. Figures 5.7c) and d) show the genomic alignment of cDNA AF374386 within BLIXEM which indicates the presence of a pseudogene.

5.4. Gene assessment

As previously mentioned, CpG islands can be used as a means of localising transcription start sites of genes. The 5' ends of 56% of genes within the interval are located next to a predicted CpG island, the same as the previously reported percentage (Antequera and Bird, 1993). The identification of a polyadenylation (polyA) signal at the 3' end of a gene can be used to assess the completeness of gene annotation as the consensus sequence, usually AATAAA, is found adjacent to the termination of transcription.

Analysis of the coding features annotated within 1pcen – 1p13 indicated that 76% of the 102 genes possessed a polyA signal within 50 bases of their 3' ends. Sixteen percent of the polyA signals contained an alternative ATTAAA motif (the second most common polyA signal) which was slightly higher than the previously reported number, 14.9%, contained within 12 genes (Beaudoing *et al.,* 2000). Another important feature to annotate is the site of polyadenylation. Genomic alignment of sequence from the mRNA which is adjacent to a 3' polyA tail permits the localisation of the site at which the pre-mRNA is cleaved prior to the addition of the poly A tail. The generation of mRNAs sequences from cDNA libraries by oligo dT priming can, however, lead to aberrant mRNAs sequences being produced. These aberrant sequences may be generated by contaminant genomic DNA acting as the template for oligo dT priming from polyA tracts in genomic sequence, or oligo dT primers may anneal to polyA tracts within the mRNA and result in a truncated coding structure.

Figure 5.8 shows the alignment of an mRNA, AF119043, submitted as a full length coding sequence of the transcriptional intermediary factor 1 gamma gene. The alignment of AF119043 (figure 5.8c and d) at the 3' UTR of the gene (figure 5.8a) indicates that the mRNA was more likely to have been generated by oligo dT priming from a tract of polyAs present in cDNA sequence which is also present in the genomic sequence. Evidence that the complete 3' UTR may not be present within the mRNA is yielded by the alignment of more ESTs 1.4kb further 3' to the position of AF119043 (figure 5.8b); these 3' ESTs contain a polyA signal. Figure 5.8e indicates the alignment of a cluster of 3' ESTs to genomic sequence which indicates where the 3' UTR of TIF1G terminates. The actual full length of the gene will, however, require further experimental support, by cDNA library screening for example, as there is not complete coverage of the 3' UTR in overlapping ESTs.

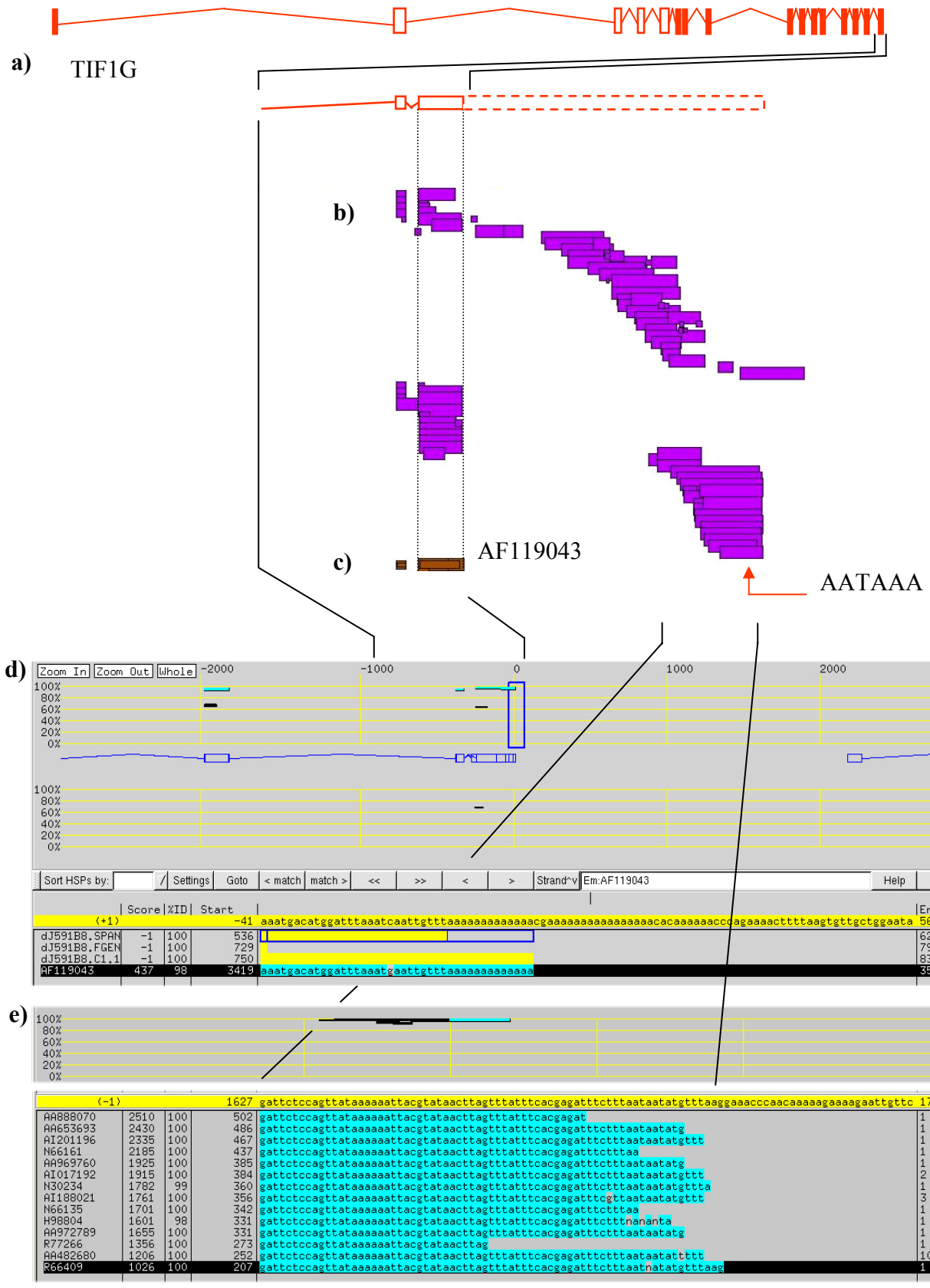

a)
TIF1G
b)
c)
AF119043
AATAAA
d)
Zoom In
Zoom Out
Whole
-2000
-1000
0
1000
2000
Sort HSPs by:
Settings
Goto
< match
match >
<<
>>
<
>
Strand^v
Em:AF119043
Help
Score
%ID
Start
End
dJ591B8.SPAN
dJ591B8.FGEN
dJ591B8.C1.1
AF119043
e)

Figure 5.8: Incomplete polyA primed mRNA. a) The annotated structure of the transcriptional intermediary factor 1 gamma using the alignment of an mRNA, AF119043 (c and d) submitted as a full length transcript. Alignment of a cluster of 3' ESTs to the genomic sequence (b and e) indicates the full length gene (including a polyA signal) extends beyond the submitted end (red dotted box, a).

The site of polyadenylation was identified within 16% of full length genes. The majority of mRNA sequence used in this study to characterise the 3' end genes, were submitted to the public databases without polyA tails therefore precluding identification of the polyadenylation site.

5.4.1 Alternative splicing

The diversity of protein coding sequence within complex organisms may be attributed in part to the widespread occurrence of gene processing mechanisms. Examples include multiple transcription start sites, pre-mRNA editing and post-translational modifications, and alternative pre-mRNA splicing all of which may be important sources of protein diversity. Alternative splicing is a highly regulated process that is capable of producing many different proteins from a single gene. It is estimated that 35 – 59% of all human genes are subject to alternative splicing (Modrek and Lee 2002) but this is likely to be an underestimate because the identification of splice variation is dependent upon EST alignment. The average alternative isoform / gene ratio detected so far ranges from 2.6 on human chromosome 22 to 3.2 on human chromosome 19, with approximately 70% of splice variants affecting amino

acid sequence of the encoded protein (IHGSC, 2001). Only 16 genes (15%) within 1pcen – 1p13 show evidence of splice variants. This fraction may be expected to increase after further iteration of EST alignment to the genomic sequence. Figure 5.9 is an example of a gene, adenosine monophosphate deaminase 2 (AMPD2), which has 4 different transcripts. On the basis of translation of each transcript isoform to predict open reading frames which start at the first AUG codon, each isoform would be expected to encode a distinct polypeptide. AMPD2 regulates the intracellular production of adenosine by competing with cytostolic 5' nucleotidase in a mechanism which regulates contractile binding in mammalian skeletal muscle. Four different splice variants (figure 5.9a-d) were annotated using previously characterised mRNA sequence. Alignment of spliced ESTs not only supported the four known gene structures but also identified a previously uncharacterised putative splice variant. This new gene structure may be a novel AMPD2 functional variant as it provides evidence for an ORF that is different from the previous four (i.e. it lacks the amino acids encoded by the second exon).

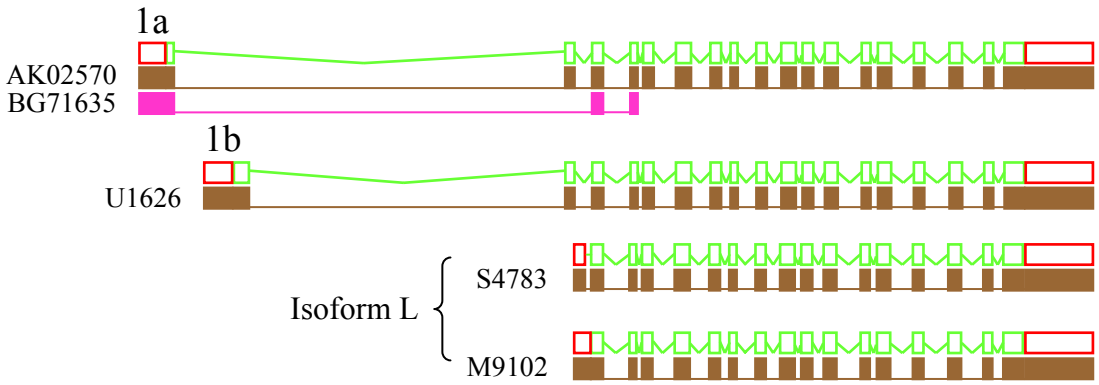


Figure 5.9: Splice variants of adenosine monophosphate deaminase 2 (AMPD2). Four splice variants were annotated (coding green boxes, UTR red boxes) by alignment of known mRNAs (brown boxes) to genomic sequence. Alignment of novel EST, BG716359 (pink box), identified a fifth potential splice variant. This variant would encode an altered protein which lacked the 43 amino acids encoded by exon 2.

5.4.2. Genic features

Detailed annotation of coding structures within a contiguous genomic sequence has provided the opportunity to investigate the context in which genes are positioned within the genome. An interesting feature to arise from the analysis of 1pcen – 1p13 is the head to head, and head to tail juxtaposition of genes which raises queries about the possible functional consequences about such gene localisation.

5.4.2.1. Putative bidirectional promoters

Within the annotated sequence of 1pcen-13, a pair of genes was observed to be orientated head to head on opposite strands of DNA. The genes, WDR3 – GDAP2 (figure 5.10a) were located in a bidirectional fashion and in each instance the 5' UTR of each gene was contained within the same CpG island. It has previously been reported that at least twenty loci have pairs of genes juxtaposed in a head to head orientation with many of these being implicated in DNA repair mechanisms (Shimada *et al*., 1989, Platzer *et al*., 1997, Xu *et al*., 1997, Connelly *et al*., 1998, Galgoczy *et al*., 2001). Genes that encode proteins involved in systems such as DNA replication, cell cycle regulation and metabolic pathways, which are commonly associated with CpG islands (Gardiner-Gardner and Frommer 1987), have also been found in this particular bidirectional orientation (Adachi *et al*., 2002). If there is a functional consequence for these genes to be related in this fashion, it may be that a common promoter element is utilised for co-ordinated expression.

The bidirectional gene pair includes two known genes, WDR3 and GDAP2. WDR3 is a member of a widely expressed family of proteins which are characterised by a gly-his and trp-asp (GH-WD) repeat and believed to facilitate the formation of heterotrimeric or multiprotein complexes. WD family members are involved in a variety of cellular processes including cell cycle progression, signal transduction, apoptosis and gene regulation. GDAP2 (ganglioside-induced differentiation-associated protein 2) was identified as one of 10 different mRNAs highly expressed in a neuroblastoma cell line which had been transfected with a GD3 synthase cDNA construct (Liu *et al*., 1999). Again, the GDAP genes are expressed in most

tissues and, like WDR3, have an inferred involvement in signal transduction (Liu *et al*., 1999). The commonality of promoter elements suggests that they may share common function and have some form of coordinated cellular expression. Experimental evidence for co-ordinate expression may be obtained by cloning the putative promoter region in a reporter construct, for example a luciferase promoter assay.

5.4.2.2. Overlapping genes

Only two genes were identified as possessing overlapping pre-mRNA structures. The 3' UTR of UNR (gene upstream of NRAS) and the 5' UTR of NRAS (neuroblastoma RAS viral oncogene homolog) were shown to overlap by 415bp (figure 5.10b). UNR contains four different 5' splice variants, and differential use of polyA signals would also allow for multiple 3' ends, whilst NRAS has a polyA signal and polyA site but no additional isoforms were identified. It is difficult to ascertain a possible functional or regulatory relationship between UNR and NRAS as UNR does not have a primary protein structure or sequence homology to any known gene. However, coordinated regulation is inferred on the basis of the same spatial relationship being maintained in species from which NRAS has been isolated and, to a lesser extent, that both genes were expressed in all tissues examined (Jeffers *et al*., 1990). The role of NRAS, as an oncogene playing a role in cellular proliferation, differentiation and transformation, and conceivably may be differentially regulated by splice variants of UNR.

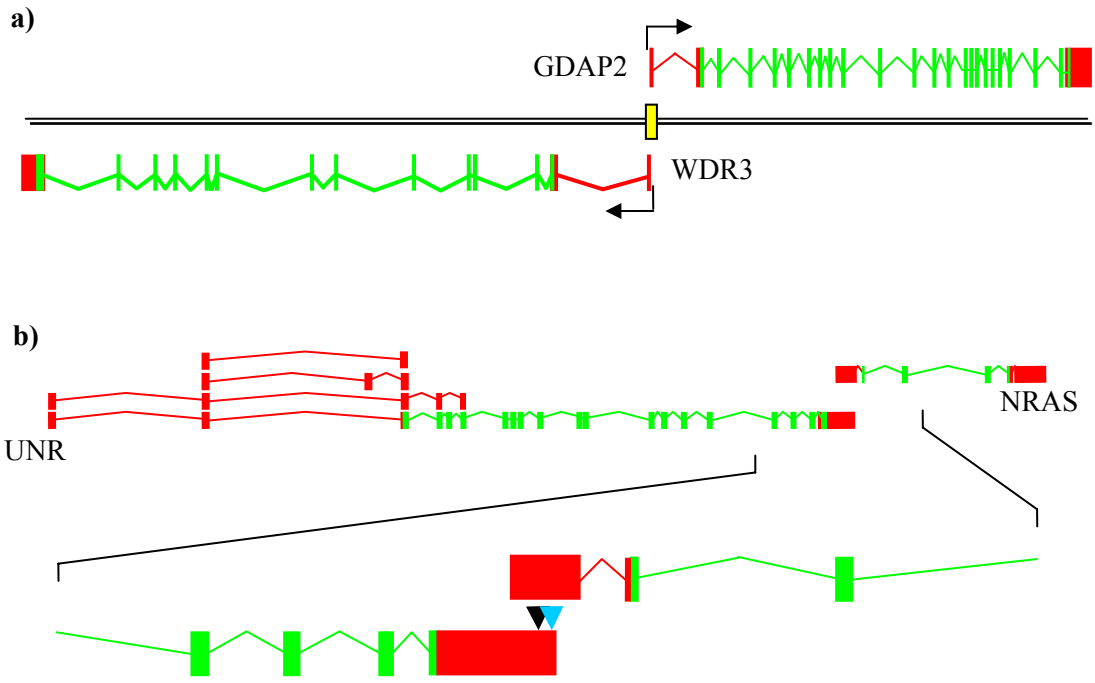


Figure 5.10: Genes in genomic context. Head to head orientation of a pair of genes (a) that share a CpG (yellow box) between the first non-coding exon (red box, coding green box) of each gene. Figure 5.10b depicts two genes, UNR and NRAS, whose 3' and 5' non-coding sequences partly overlap, respectively. PolyA signal (black arrow) and polyA site (blue arrow) are represented in the 3' UTR of UNR.

5.5 Inferring function by protein homology

Greater than one third of all full length coding features identified within 1pcen – 1p13 were novel genes. The possible function of these genes can be inferred by homology, at both nucleotide and amino acid sequence level, with previously characterised genes from either

human or other species. These types of analyses may facilitate the association of a novel coding feature to an existing gene family or may assist in predicting gene function by identifying the individual protein domains within the gene.

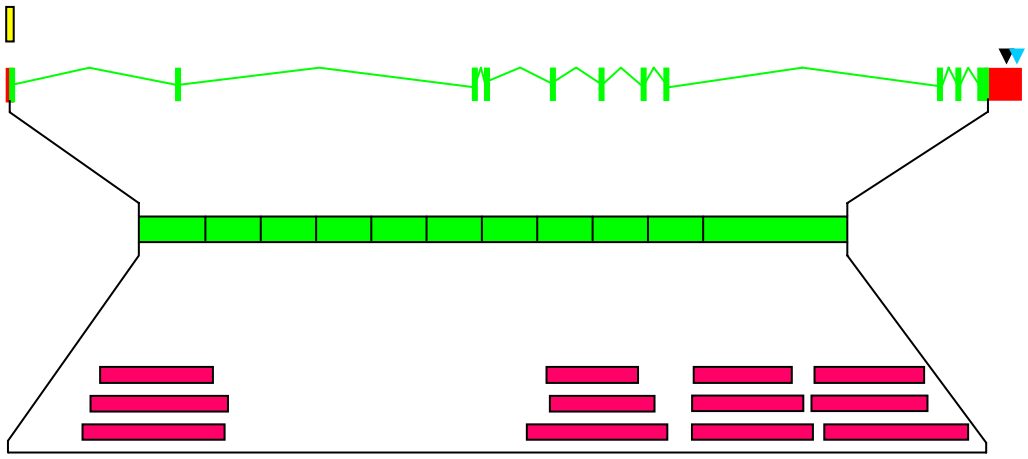
5.5.1. Identifying function through sequence homology

To investigate the means of identifying gene function through DNA and protein homology gene, bA12L8.C1.1 (figure 5.11a), was analysed within PIX (http://www.hgmp.mrc.ac.uk/Registered/Webapp/pix/) and PSI-BLAST (http://www.ncbi.nlm.nih.gov/BLAST/). bA12L8.C1.1 was annotated from a full length uncharacterised IMAGE mRNA and was supported by sequence from cDNA library screening. Analysis of the translated mRNA within PIX (which uses a suite of programs to characterise features within the protein) predicted four transmembrane domains (figure 5.11b). The four helical structures were each predicted by three different transmembrane programs, TMHMM (Sonnhammer *et al*., 1998), TMPRED (Persson and Argos 1994) and TMAP (Milpetz *et al*., 1995), which cumulatively supports the presence of the structure within the sequence. A depiction of the possible *in situ* protein structure is represented by figure 5.11c. PSI-BLAST http://www.ncbi.nlm.nih.gov/BLAST/) was used to identify sequence homology to the putative transmembrane protein. Analysis of the novel protein identified a 74% homology to a sugar transporter domain within the conserved domain database (CDD) originating from PFAM (figure 5.11d, e). In parallel, BLAST alignment of the protein sequence showed a 99% homology to a hypothetical human protein, NP_060890 (not shown). The protein was derived from the direct submission of a previously described

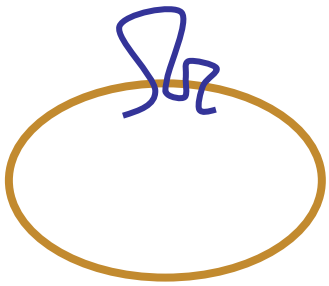
mRNA (but not aligned by annotation here) which is purported to show homology to the Drosophila Orct gene and mammalian carnitine transporters, a family of genes which are involved in transmembrane organic ion transportation.

a)

b)



c)



d)

e)

```
Query:  132  LSDRFGRKKVYLTGFALDILFAIANGFSP-SYEFFAVTRFLVGMMNGGMSLVAFVLLNEC  190
Sbjct:  69   LGDRFGRKKSLLIGLVLFVIGSLLSGAAPGAFYLLIVGRVLVGFGVGGESVLVPMYISEI  128

Query:  191  VGTAYWALAGSIGGLFFAVGIAQYALLGYFIR-----SWRTLAILVNLQGTVVFLLSLFI  245
Sbjct:  129  APKALRGALGSLYQLGITIGILVAAIIGLGLNKTSNWGWRIPLGLQLVPAILLLIGLLFL  188

Query:  246  PESPRWLYSQGRLSEAEEALYLIAKRNRKLKCTFSLTHPANRSCRETGSFLDLFRYRVLL  305
Sbjct:  189  PESPRWLVLKGKLEEARAVLAKLRGVPDVDQEIQEIKASLEAGVSSEKAELELFRGRT-R  247

Query:  306  GHTLILMFIWFVCSLV------YYGLTL--SAGDLGGSIYANLALSGLIEIPSYPLCIYL  357
Sbjct:  248  QRLLMGVMLQIFQQLTGINAIFYYSPTIFKSVGLSDSVALLVTIIVGVVNFVATFIAIFL  307

Query:  358  INQKWFGRKRTLSAFLCLGGLACLIVMFLPEKKDTGVFAVVNSHSLSLLGKLTISA----  413
Sbjct:  308  V--DRFGRRPL----LLLGAAGMAICFLILGVALLLLNKPAAGIVAIVFILLFIAFFALG  361

Query:  414  AFNIVYIYTSELYPTVIRNVGLGTCSMFSRVGGIIAPFI  452
Sbjct:  362  WGPVPWVLLSELFPTGVRPKALSLATAANWLANFIIGFL  400
```

Figure 5.11: Putative assignment of structure and function of a novel gene. Figure 5.11a is the full length gene structure (including CpG island (yellow box), polyA signal (black arrow) and site (blue arrow)) of novel gene, bA12L8.C1.1. b) PIX analysis of the coding sequence identified four transmembrane domains (red boxes) giving rise to a putative cellular conformation, c). PSI-BLAST analysis of the novel gene identified putative functional domains (a sugar transporter within the conserved domain database, d) and sequence alignment of the highest percentage homolog by BLAST alignment, to a hypothetical protein, NP_060890, subject in e). Homologous residues between the two proteins are shown in red and conservative residues in blue.

5.5.2. Identifying function by structural homology

Another means of characterising the function of a novel protein is by utilising a sequence-to-structure-to-function analysis. Using this method, the function of a protein can be inferred from a homologous protein whose 3-D structure has already been elucidated. This analysis was used to predict the function of a novel gene bA483I13.C1.2. A structural homologue of the novel gene was identified by BLAST alignment of the translated protein with previously characterised motifs within Swiss-Model (Peitsch *et al*., 1993). The novel protein showed the highest matching probability with the previously elucidated structure of 1VRK (Mirzoeva *et al*., 1999) which is the peptide binding complex formed between calmodulin (CaM) and RS20, the CaM recognition site peptide from vertebrate smooth muscle cells. The X-ray crystallographic structure of 1VKR permitted the 3D structure of the novel protein to be predicted by amino acid sequence alignment. To determine the putative tertiary structure of

the novel protein it was first read into DeepView (Guex *et al*., 1997) and then aligned to the template protein, 1VRK (figure 5.12a). The sequence alignment was edited to reduce the energy state of each group contained within the new structure by the introduction of gaps between amino acid residues (figure 5.12b). Adjustment of the amino acid sequence ensured that the side chains of the new structure were not in conflict, thus stabilising the new conformation. The two protein structures were then superimposed (figure 5.12c). Residues of the predicted novel protein structure are coloured according to their energy state, i.e. best fitting residues are blue and the least are red. Superimposition of the putative 3-dimensional structure (figure 5.12d) provided visual confirmation of the structural similarity between the proteins. Whilst this type of analyses provides some evidence for the function of the novel protein (by structural homology to a previously characterised protein) experimental support would be required to more accurately define novel protein function, particularly in light of structural differences between 1VRK and bA483I13.C1.1 as denoted by the red asterisk in figure 5.12d.

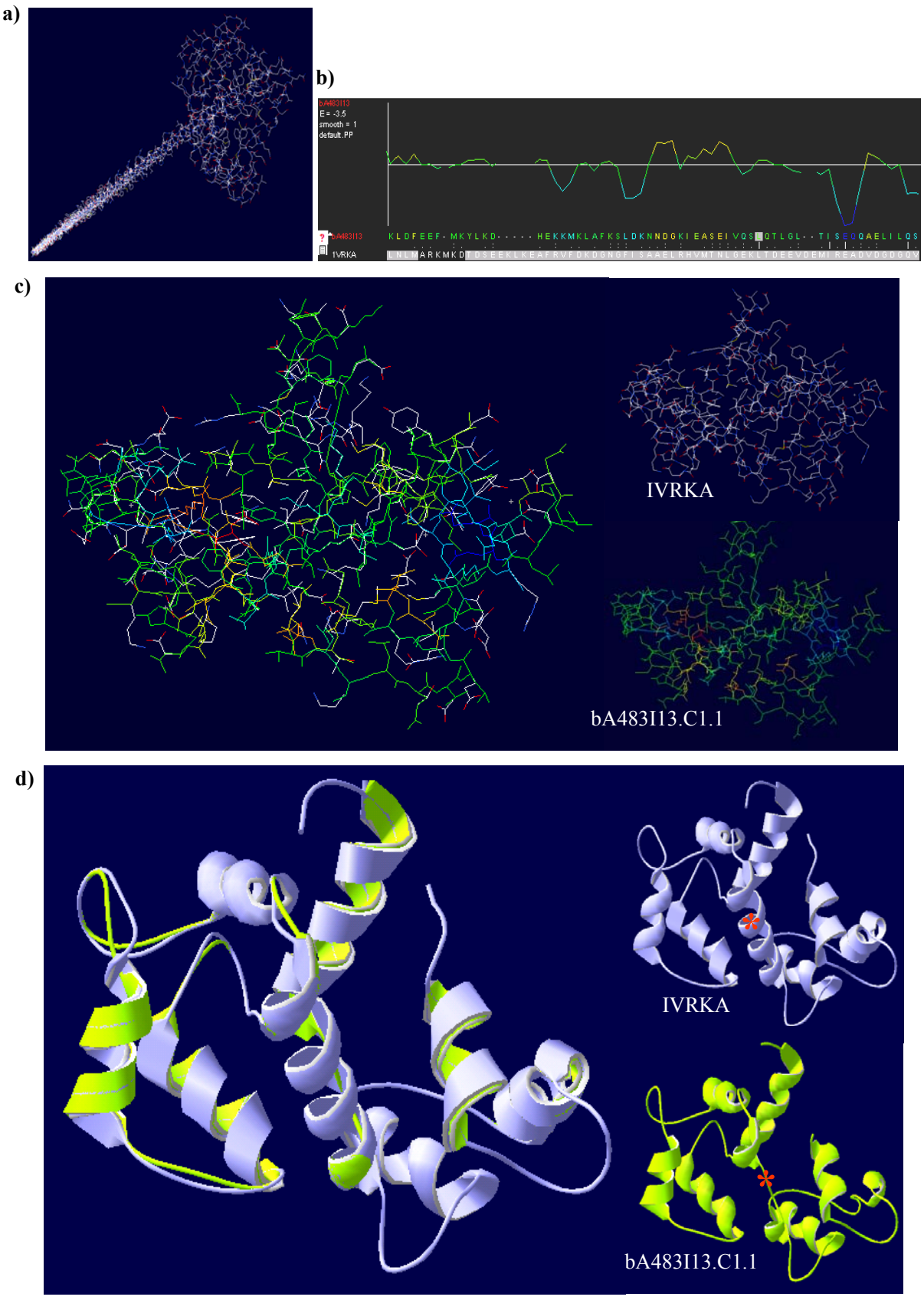

a)
b)
c)
IVRKA
bA483I13.C1.1
d)
IVRKA
bA483I13.C1.1

Figure 5.12: Identification of putative functional domain of a novel protein. Structural alignment was initiated by threading the novel protein (5.12a, rod structure) on characterised 3D protein. b) The novel protein is edited to reduce threading energy of new structure. c) The putative novel 3D structure is superimposed with the energy of amino acid sequences coloured red (high) to blue (low). d) Superimposition of ribbon conformations of novel and known protein structures. The asterisk denotes differences between two protein structures.

5.6. Discussion

A comprehensive characterisation of the genomic landscape contained within a 12.4Mb region of 1pcen – 1p13 is described. A correlation is made between the DNA profile of the interval with giemsa staining and isochore partitioning, repeat content and gene distribution. The detailed annotation of eight contiguous finished sequence links (see appendix, table 5.5) representing 95% coverage of the interval, has determined the genomic structure of 102 full length genes, including 67 known and 35 novel protein coding genes. In addition, 16 novel transcripts and 11 pseudogenes have also been identified.

Genes are typically associated with functional elements within genomic sequence whose presence can help determine whether the full length structure of a gene has been correctly elucidated. The detection of these elements aids the determination of full length gene annotation. Whilst some of these elements are relatively easy to identify, for example exon - intron boundaries by mRNA/EST alignment, translation start and stop sites and polyadenylation signals by motif recognition, others such as promoters, are more difficult to characterise. Promoters and other regulatory elements residing at the 5’ ends of genes which

act as a template upon which transcription factors assemble prior to the initiation of RNA synthesis, are inherently difficult to identify using *in silico* methods because they do not contain consistently shared sequence motifs. There are, however, some characteristic features associated with, or contained within, the predicted structure of a promoter. These features can be used to aid the identification of promoters and, in doing so, help to localise the 5' end of a gene.

A degree of sequence motif conservation within the core of the promoter permits *in silico* prediction of transcription start sites (TSS) of human genes. Eponine (Down and Hubbard, 2002), the TSS prediction program used herein, uses TATA box motifs flanked by regions of C-G enrichment in conjunction with a predicted CpG island to identify the TSS of a gene. This program has a reported sensitivity (being able to detect a known mRNA start) of 54% and a selectivity (the proportion of predictions that are confirmed by a known mRNA start) of 74%. A representation of the constraint distributions and sequence motifs Eponine uses to identify the gene TSS is represented in figure 5.13. As previously mentioned, predicted CpG islands can also be used to help identify the full length transcript and promoter region as they are associated with approximately 56% of 5' ends of genes (Antequera and Bird, 1993). Two thirds of the 102 known or novel genes from 1pcen – 1p13 were associated with either a CpG island (57%) or an Eponine prediction (9%), while 17% of genes within the region were associated with both.

Figure 5.13 represents a 'generic' gene structure with a predicted TSS and CpG island at the 5' end of the gene and red (untranslated) and green boxes (translated) represent the

transcribed length of the gene. Splice donor, GT, and splice acceptor, AG, consensus sequences (contained within 99.9% of all introns (Levine *et al*., 2001)) used by the spliceosome to excise introns during pre-mRNA processing are highlighted, as are consensus poladenylation signal (AATAAA) and polyadenylation sites. Processing of the pre-mRNA (figure 5.13) results in the removal of introns, the addition of a 5' guanine cap (dark blue box) and a polyA tail. The optimal consensus sequence at the site of translation initiation, GCC$^{A}/_{G}$CC**AUG**G (Kozak, 1987), which includes the two bases which exert the strongest effect, a G at the first base after the translation start, AUG, and a purine (preferably A) three nucleotides upstream, are also shown. Whilst the model of a single promoter effecting the transcription of a single gene product is relatively easy to discern, there are examples of tissue specific promoter regulation and coordinated expression of genes sharing a promoter that complicates our understanding of how promoters function.

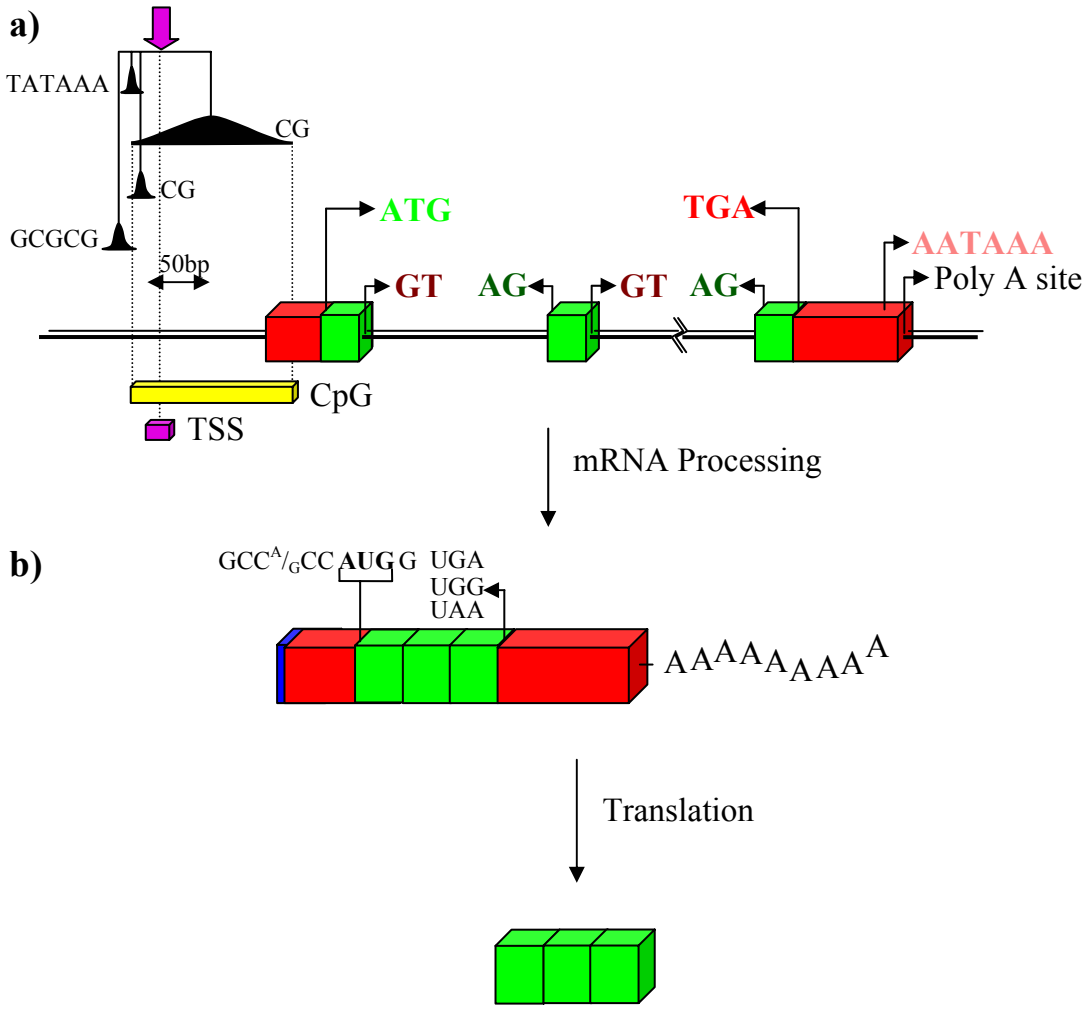


Figure 5.13: Generic structure of a gene. Figure 5.13a represents the regulatory elements (predicted CpG island and transcription start site consensus motifs (purple box)) at the 5' end of the gene, translation start (green lettering) and stop sites (red lettering), polyA signal (pink lettering) and sites and exon splice donor (dark green lettering) and acceptor sites (dark red lettering). b) Processing of pre-mRNA results in the addition of guanine 5' cap and addition of a polyA tail. Also represented is the Kozak consensus sequence and translation start and stop sites. Translation of the gene subsequently follows.

The annotation of a full length coding feature within the context of genomic sequence is only the starting point when trying to fully characterise a novel gene. Whilst the regulation and tissue distribution gene expression poses a difficult question, the function of the encoded protein product is perhaps a more difficult problem to resolve. There are two different approaches that are used to assign a putative function to a novel protein. The sequence-to-function approach utilizes a two dimensional pair wise sequence alignment or motif alignment to suggest protein function in a novel gene on the basis that structural homology reflects functional similarity. This method was used when predicting the transmembrane sugar transporting role function to the novel gene bA12L8.C1.1. The second approach is to utilize a sequence-to-structure-to-function paradigm. This technique is believed to be more powerful because the development of the structure is more in accord with how the protein functions. However, the folds of a protein alone cannot determine its function as proteins with similar folds may have completely different utility. It is the combination of dimensional structure, active sites and protein – protein complexes which will provide a more precise aid in predicting novel gene function. The limitation of the *in silico* methods described above is that they require a degree of homology to be identified (whether at the level of primary or tertiary structure) with a protein of known structure or function. The library of these known models will need to be increased by experimental methods of structure determination and complex formation to broaden the applicability of this approach.

The structure and function of a gene and its cognate protein are inherently determined by the genomic sequence from which it is transcribed and elements which affect its regulation. Alterations within these coding elements by as little as a single base change may result in

conformational and functional consequences. The next chapter details the development of assays designed to identify these changes within a number of genes, including a gene family localised to 1pcen – 1p13, and discusses the potential affects these changes may have.

5.7. Appendix

Table 5.5: Minimum tile path clones and accessions from the 1pcen – 1p13.2 contig (October 2002). Boxes denote contiguous blocks of finished sequence.

1	Link_bA436H6	AL513187	bA436H6
2		AC114491	AC114491
3		AL513206	bA382F13
4		AL591042	bA480L11
5		AL353892	dJ547O1
6		AL391235	bA320L5
7		AL390036	bA356N1
8		AL359258	bA483I13
9	In Finishing	AL672086	dJ673H23
10	Link_bA131J3	AL390038	bA131J3
11		AL392088	dJ964H19
12	In Auto-analysis	AL160171	bA256E16
13	Link_bA293A10	AL591719	bA293A10
14		AL449266	bA475E11
15		AL356488	dJ1065J22
16		AL138933	dJ667F15
17		AL356389	bA352P4
18		AL390252	bA297O4
19		AL356735	bA173K24
20		AL355145	dJ831G13
21		AL355310	dJ1160K1
22		AC000031	cgtm1
23		AC000032	cgtm12
24		AL158847	dJ735C1
25		AL450468	bA195M16
26	Re-submission	AL772411	bA180N18B
27	Link_dJ742A5	AL355817	dJ742A5
28		AL772412	bA180N18A
29		AL160006	dJ773N10
30		AL355990	dJ1028L10
31		AL137790	dJ1003J2
32		AL355488	dJ1074L1
33		AL390797	bA225L12
34		AL358215	bA470L19
64	Link_bA512F24	AL391058	bA512F24
65		AL389921	bA228G5
66		AL390759	bA473L1
67		AL133517	dJ730K3
68		AL365321	bA324J2
69		AL137856	dJ1073O3
70		AL731797	dJ786G8
71		AL591742	dJ590F24
72		AL162594	dJ1037B23
73		AL121999	dJ543J13
74		AL512291	dJ802H15
75		AL035410	dJ591B8
76		AL390241	bA343L14
77		AL133382	dJ1156J9
78		AL096773	dJ1000E10
79		AL390235	dJ1146M22
80		AL358372	bA350E19
81		AL645502	bA109G4
82		AL109660	dJ666F24
83		AL139428	dJ977F20
84		AL049825	dJ662B22
85		AL606499	dJ1034E9
86		AL512638	dJ663N10
87		AL157950	dJ940J24
88		AP001393	bA722J12
89		AL592436	bA710N8
90		AL450389	dJ929G5
91		AL449264	bA485H8
92		AL357137	bA12L8
93		AL365318	bA159M11
94		AL121982	dJ1185H19
95		AL831782	dJ636P16
96		AL355538	dJ787H6
97		AL136376	dJ655J12

35		AL365361	bA284N8
36		AL354713	bA498A13
37		AL391064	bA392B1
38		AL360270	bA96K19
39		AL355816	dJ1180E21
40		AL513202	bA165H20
41		AL356387	dJ1125M8
42		AL390195	bA552M11
43		AL391063	dJ836N10
44		AL139012	dJ1091G18
45		AL049557	dJ773A18
46		AL450997	dJ1086I18
47		AL390070	bA99M15
48		AL512665	bA88H9
49		AL357114	bA57I1
50		AL445426	bA62J10
51		AL591521	dJ965F6
52		AL450407	dJ1160J2
53		AL354760	dJ671G15
54		AL109932	dJ770C6
55	In Finishing		bA72M14
56	In Auto-analysis	AL603832	bA426L16
57	Pre-Sequencing		bA721A13
58	Link_dJ522D1	AL390729	dJ522D1
59		AL158844	dJ580L15
60		AL390242	bA31F15
61	In Finishing	AL357055	bA389O22
62	Link_dJ658C17	AL139016	dJ658C17
63	Re-submission	AL365225	bA179A5
98		AL390066	dJ1086K13
99		AL355794	dJ781D12
100		AL356748	dJ686J16
101		AL135798	dJ655N15
102		AL157904	dJ753F5
103		AL445231	bA27K13
104		AL139248	dJ570D9
105		AL391476	bA229A19
106		AL365264	bA287H7
107		AL360298	bA39H13
108	In Finishing	AL358072	bA188D8
109	Link_dJ675C20	AL157902	dJ675C20
110		AL365331	bA42I21
111		AL390877	bA134N8
112		AL122007	dJ757N13
113		AL121993	dJ776P7
114		AL139345	dJ832K2
115		AL513191	bA224F24
116		AL391557	bA506J19
117		AL512823	dJ881A21
118		AL122006	dJ730H16
119		AL390117	bA116P22
120		AL606843	bA94F13
121		AL139148	dJ630J13
122		AL845532	bA183H8
123		AL357045	dJ794L19
124		AL139420	dJ712E4
125		AL359823	dJ610L12
126		AL590288	bA212F6
127		AL359915	bA418J17
128		AL139346	dJ834N19
129		AL359553	dJ871G17
130		AL121995	dJ920G3
131		AL109966	dJ599G15
132		AL139251	dJ656M7
133		AL589734	dJ683H9
134		AL359752	dJ1042I8
135		AL512503	bA323K8
136		AL596222	bA114O18

Table 5.6: Primer pairs designed for the validation of predicted gene structures by cDNA library screening. Red background denotes primer pairs for which no PCR product was generated from cDNA library screening.

Gene	Exon	stSG	Primer 1 (S)	Primer 2 (A)	Size	1pool
bA483I13.C1.1.mRNA	e2	452926	GGTATCTGCCGACCCTTTGT	GAGTAGGCAGTAGCTTGAGT	147	Y
bA483I13.C1.3.mRNA	e2	452927	GCAGTCTGGAGATTGGTGGA	TGCATCATGACTTTCAAGCG	102	N
	e5	452928	GGGATTATTGATTGTGGCAA	CGGCATAAGGTACAATGCCT	100	Y
	e7	452929	GGTTTCACCTCAAACATCAT	ACATCTCTTTATAACACAGG	164	N
bA475E11.C1.2.mRNA	e1 5'UTR	452930	TGACGGCTGAAGAAACAGTG	CTCCAGGGCCAGCATACTAA	104	Y
	e4	452931	GCTTTTGACTTTGCCTCGTC	GCTTCCTATCAGCAGGGATG	128	Y
	e7	452932	CAGCACTCAACCAGCAATGT	TCCAGGATTACGAGGAGTGC	149	Y
	e10	452933	ATGAGACTCCTAAGCAGCCG	GGGCAGCACTTTGACGTATT	137	Y
	e12	452934	ATACCTGGAGTGGCTGGATG	GTTGTGCCAACAACACGAAC	100	Y
	e14 3'UTR	452935	CGAAGAGGCCCCTTATTACC	GGAGTGCACACCAACAACTG	166	Y
bA475E11.C1.1	e1 5'UTR	452936	AGGCTATGCATAGTGAGACT	GCTTGACTTAGAAGCGTCTC	155	N
	e5	452937	ACTGCAGGGACACCTTGAAC	CCAACGATTGTTGATTCGTG	105	N
	e6	452938	TTGGAGATGCTGCTTGAAGA	AGAGAAGGTGGAGGCCAAGT	117	N
	e10	452939	GAAGCAAAACGTGGAGAAAA	TTGAATCTGAGTGTGGTGCC	92	N
	e13	452940	CTTCCAAATCCAGCCCTACA	ATGGGTTGCTACCAACTTGC	127	N
bA297O4.C1.1.mRNA	e1 5'UTR	452941	GCCACTATTGGGAGACCAAG	GTAGAGCCAGAGGTTCGACG	124	N
dJ831G13.C1.1.mRNA	e1	452942	CTTTGCTATTTTCGCCTTCG	CTGAAGGGATAGCCAAATGC	123	Y
	e3	452943	CCTTCACCTTCTTCTGGCTG	CTTCCTCTCCATGGCACACT	120	Y
	e4	452944	GCTCAGTGCTTCTTGTGCAG	TGGCTGCTAGGAACCAGTCT	146	Y
bA180N18A.C1.2.mRNA	e1	452945	CCCACTGATCGTGAACAACA	CTCTGAGTCTTTGCGCTGGT	154	N
dJ773N10.C1.1.mRNA	e2	452946	CAGCCTGCATCTTCCTCTTT	AGACCTTCTCCAGCTCCTCC	125	Y
dJ1003J2.C1.1	e2	452947	CCCTGAATGAGAAGGAGCTG	CACGGACTCAGTGACATGCT	148	Y
	e4	452948	CTGTCTCTTTGTGGGGCTGT	ACAGGACATTCACTCCAGGG	102	Y
	e7	452949	ACCCAGGTCTTCTTTGCCTT	GCCAACACTGACGTGAAGAA	134	N
	e9	452950	CCTTCATCGCCTTCACTGAG	GTGAACATCTCCTTGGGCAC	169	Y
	e12	452951	CGCTACCTGTATTTCCCCAA	CCCTTCTTGTAGGACACGGA	149	N
bA470L19.C1.2.mRNA	e1	452952	CACCAAGCATTCCATACGTG	GAACCCAATGGGGATTCTTT	150	N
bA284N8.C1.2/.3	e2	452953	ATGCTCGGCTGTCTTCAAGT	AATGGTGAGTCATTCTGGGC	128	Y
bA165H20.C1.3.mRNA	e3	452954	TGTTACTTCACCAACTGGGC	TGGTGATCTCGTTGTTCTGC	127	Y
	e5	452955	TTCCACTCCTGAGAACCACC	CTGCACCAGGACAGTGAAGA	151	N
	e7	452956	CTATGACCTCCATGGCTCCT	ACATTGAGGTAGGCGTTGCT	96	Y
	e10	452957	AACAACTTTGGAGGTGCCAT	TTGTACTCTGCAGGCCCAGA	124	N
dJ1125M8.C1.1mRNA	e2	452958	GCGGATAACTACCCTTTTGG	AAATAACACCCAGGCCCTCT	128	Y
	e4	452959	ATGCAGGCAGGTACCAGAAA	TTCTTAAATCGAGGCACCAA	91	Y
	e6 3'UTR	452960	ACGTTACTGTGGCCCTCTTG	ACAGAAACCCACAGACCCAG	154	Y
dJ1125M8.C1.2	e1	452961	GAATGGAGGAGCAGGGTGTA	TCCAGGTAGTTGGTGAAGGG	121	Y
	e3	452962	ACAACAGGTTCAATCCCAGC	TGTCATAGCCCAGGAACACA	105	N
	e5	452963	ACCCGCCAGTATTGTGGAGA	GGCAATCTGCCAGTACAGTT	107	N
	e8	452964	AACAATGGCTACTGCAGGCT	CTAGGCAGAGAAGGCAAAGC	153	N
bA552M11.C1.4.1/.2	e2 .2	452965	ACCACGTGGGATTTGATGTT	GGATGCCAAATTAAGAGCCA	137	N
	e3/4 .1	452966	CCTTTTGTGCTGGGGTTCTA	GCTGGAGGATCTGAGTGAGG	121	Y

	e5	452967	TGAGGCTTGAATCCATTTCC	CTCTGGCCAGGAAAAGACTG	175	Y
bA552M11.C1.5	e1/2	452968	TGCTTCCTTCCAGTCATGTG	TGGTCAGGCAGGACATAGTG	120	Y
	e3/4	452969	CAGGCAACAAAACCAGAAGC	CCCAAACCCGTGATCAGTAT	103	Y
	e5	452970	ATCATTTGCAGCCAGGTAGC	GTCCCCAATCCAGATTCTCC	154	Y
dJ836N10.C1.1	e2/3	452971	GGAAGAACAAGGAAAAGGGC	CTCAATGCTTCCCCTCACTG	176	Y
	e4	452972	AAAAGCCAGAGCTTCCTGAC	TGTGGTCCCTTTCTTCTTGT	120	Y
dJ1073O3.C1.3	e1	452973	CTGGGCTGAAAACTGCTTGT	GTTGGGCTCAAGAAGTCCAT	134	Y
	e3	452974	GACCTGGTGTGCTCAGGATT	TTCCCATTGATCATACCCGT	144	Y
dJ1037B23.C1.1.mRNA	e2	452975	TCCCTCTTCTGCTAATCCCC	ACCTCAGCTGGGATATCTGG	122	N
	e4	452976	AGCGTGGACTTGGGAGAGAT	GTGATGTCCATCGCCTTGAG	107	Y
	e6	452977	TAGGAGTCTGTCGTGTGGGG	TTACCTCCACCAAGGAGTGC	114	Y
	e8 3'UTR	452978	AAACAGTGTTGTGCAGTCGC	CATCACCTTGGGAGACACAA	144	Y
dJ1156J9.C1.1	e1 5'UTR	452979	ACCTTGGAGCGGGATCTTAT	TGCCAGGGAATTGTTGTATG	127	Y
dJ929G5.C1.1.mRNA	e2	452980	TCCTGTTGAAGAGTGGCTCC	TCCAGAATAAGTGGATTCGG	157	Y
	e4	452981	GTTTGTGTTTCGTGCCCTTT	TATTGCACAATGCCCTGGTA	120	Y
	e6	452982	CAGTAACAATGCCACTGGCC	CTTCTTTACTCGCCGTTCAT	118	Y
	e8	452983	GCAATATGACAAGGACCGCT	TACGAGGCTGAAGTCCAAGC	121	Y
bA12L8.C1.1.mRNA	e2	452984	CATCCTCATTGCACTGGTTG	TGCACGTGCTTATGGATCTC	159	Y
dJ655J12.C1.2.mRNA	e1	452985	AAGACAAGGAAGAGCACCTG	GAGTCCTTGAAGTGGTCGGA	120	N
	e2	452986	GCTCTGTTCAGGAAAATGCC	TGATCATCAGTGAGCCAAGC	165	N
dJ655J12.C1.3.mRNA	e2	452987	AGTCCTCCCTGAACTGTTGC	TGGGCATGAGATAAAACACG	106	Y
dJ686J16.C1.1.mRNA	e1/2	452988	GGGCAGTGTCCAATTTATGG	GGAAGGAGGACTGATGGTGA	116	Y
bA39H13.C1.1.mRNA	e1	452989	AAAAACCCAGCTGGACAATG	TCAGCAAGATTCCTCGGTCT	142	Y
	e2	452990	ATCTGGAAGCAGAGCCAGTA	CCATTTCAGAGCTTCTGTGC	90	Y
bA42I21.C1.1.mRNA	e1	452991	CACATGCGTCGGCTTAAATG	CCTCCACGATCGATGTTTCT	95	Y
	e2	452992	CCCTCGCTGGGAAAGACATA	TGCTGGGGGAAAAGATTACT	139	Y
dJ776P7.C1.1	e1	452993	GGCACTCTATTCGCACGTCT	CTCCATCATCCCAGGACACT	99	Y
	e1/2	452994	GCTGAGAGGATTATGGAGGC	CTGAACTCTGCCCCTTACCA	101	N
	e4	452995	CTGTCCTCCCACTGGAATGT	TTCCGAGGTGAAGGAGAAAG	145	Y
dJ832K2.C1.1.mRNA	e1	452996	AAAAACTCCAGGACCTCCGT	ACCTGCAGCCTCAGTTTCAC	171	Y
	e6/7	452997	CCGCATAATACCACCCTTTT	CAGCTGTTTCGTTTGCATCT	131	N
dJ832K2.C1.2.mRNA	e2	452998	CCTCCAAACACAGGCTCTCT	CATGATGTACCTGCCAGCTC	126	Y
dJ832K2.C1.3.mRNA	e2	452999	CCTTCAAGAAGCCCATAAGC	CAACATTGGAGTGGAGAGCA	146	Y
	e5	453000	AGGCAAGGATAACGCAGAGA	CTTAGGTTCTGGTTGGTGGG	133	Y
	e8	453001	ATCCCTTCAGCACTCACTCC	TCTTTGGGTTTTTCTTTGCC	98	Y
bA224F24.C1.1.mRNA	e1	453002	TTAGAGGCCAATGCTTCTCC	AGCGAGGGTTCCCATATCTT	95	Y
	e4	453003	ACGGCAGCAAAAGCAATTAT	TTCTTTTCATTTTCCCGTCG	125	Y
	e6	453004	GGGCTTTAACAATCCTCAGC	CTGGTTAACTGCTGCCAGGT	124	N
	e8	453005	TTGCCTGCTTGATGTATGAC	CTCTGAAGTTGGCATGGCTT	131	N
	e11	453006	AAGAAGATCTCGTCCCACCC	GACAGAGTGAGGGCAGAAGG	105	Y
	e15	453007	AGCAACCGAGAACCTTCAGA	AGAGACTCATGTTGGGGCTG	149	N
dJ794L19.C1.1.mRNA	e1	453008	GGCGGCTAAAATGAGTGAAA	ATAGACAGGTCCAGCCCCTT	141	Y
	e3	453009	AGGATGTTTCCTGCCATGAG	TTTTATTGTCCACAGGCACA	94	Y
	e5	453010	CTCGAGTTCATGTGATTCGC	AGGCCGTAACTGTGGTGAAC	123	Y
	e8	453011	TACCAACTCCTCCCTCGTTG	CATGTGTGGTGATGAGGAGC	137	Y
dJ834N19.C1.1.mRNA	e1	453012	TACCGGTCAGACTCCAGGTC	AGGTCCTCTTCTTTGCCTCC	93	N
	e3	453013	CAGTAACTGAGGAGGGCCAC	GGCTGCGATAGAAAGCAAAG	143	Y
dJ834N19.C1.2.mRNA	e2	453014	AGACGAGGTCTTGCCACATT	GCATGGTGGCTTATGCTGTA	108	Y

dJ599G15.C1.5.mRNA	e1	453015	TCGCTAGCCATTATCCAACC	CCTGTCCTTGTAGTGGGCAT	129	N
dJ599G15.C1.6.mRNA	e1	453016	ACTCTTCAGGAGCCACATGC	TCTACTGGAAGAGCACCAGC	95	Y
dJ1042I8.C1.4.mRNA	e2	453017	ATGCTGGCCACAATCTACCT	GATCACTCCCCACAGCACTT	127	Y
	e3	453018	TGGTCCAGTGAGAAAGCAGA	CCGGCCATTTGAGTTACAAG	125	Y
	e5	453019	GGAACGAGAGCTGATCCAGT	AGCTGTTCTCGGAAGTCCTG	138	Y
	e7	453020	GGAGGAATGTGCCATCACTT	GAGCATCCTGCCATTCATCT	152	Y
	e9	453021	AGGAAGCTGCAGGAGTCTGA	CCAAGAAAGTGCCTTCACAA	124	Y

Chapter 6

The identification and analysis of single nucleotide polymorphisms

6.1 Introduction

Genetic variation, locus specific differences in sequence between individuals, has arisen within the human population primarily through unique mutational events some time in the past. The genetic variants (polymorphisms) in present-day chromosomes reflect these historical mutational events, and analysis of them can therefore be used as a means of determining evolutionary relationships between populations and individuals.

Single nucleotide polymorphisms (SNPs) are the most common form of genetic variation within the human genome. On this basis, comparison of any two human genomic sequences would identify approximately 2.5 million single base pair variations, at a frequency of 1 SNP every 1.25 kb (Li *et al*., 1991, ISNPMG, 2001). A subset of these variants will cause an alteration in an encoded function, for example with respect to transcription, translation, or the structure or catalysis of a protein. More over, any of these functional changes which are deleterious to health represent a genetic contribution to a human disease state. These functional variants, and other polymorphisms which are linked to them, can be investigated to characterise genetic contributions to human disease. From a disease perspective, the sequence polymorphisms of an individual can contribute to the severity, age of onset and susceptibility to disease, or even the physiological response to the drug treatment.

The association between a polymorphism and a disease phenotype can be studied by direct or indirect approaches (reviewed by Collins, 1998). In the direct approach, a functional variant allele is tested directly for association with the phenotype under study, for example a case-control study. The frequency of a variant allele is compared in

the case population and in an equivalent group of matched controls. A skew in the frequency of an allele is indicative of an association, which can then be confirmed by additional population genetic studies and functional studies. This approach depends on prior identification of putative functional variants; given the range and possible low frequency of many of these, a systematic study is required to provide the candidate functional variants for such studies. The alternative, indirect approach is carried out by testing variants (usually SNPs) selected by genomic position, for association with the phenotype in the association study. This approach depends on adequate association (i.e. linkage disequilibrium) between the SNP(s) used in the test, and the unknown functional variant, which is likely to be nearby in the genome. An association detected by the indirect approach therefore results in identification of candidate regions within the genome to target the search for the functional variant(s) which then require confirmation in the same way described above.

While large-scale studies will lead to a well-characterised panel of SNPs for genome-wide indirect association studies to search for new genes and variants involved in human disease, a more targeted small scale approach can be carried out by selecting candidate genes for SNP discovery and association studies. The identification of SNPs within coding sequence (cSNPs) in genes of medical importance, and associating coding and structural changes with gene function, will be essential to our understanding and treatment of human disease. Additional studies of gene regulatory regions may also be required to find other functional variants which alter transcriptional processes, but search for cSNPs represents a valuable first step in targeted SNP searches in candidate genes. This chapter outlines an approach to the identification of cSNPs within a family of genes localised to 1pcen – 1p13 (GSTM1-5) which has inferred association with

increased cancer susceptibility. Seven other genes of medical interest, the majority of which are involved in drug metabolism, are also included in this investigation. Diseases associated with genetic variation within these additional loci include gastric (Tsukino *et al*., 2002) and lung (Nakachi *et al*., 1991) cancer as well as susceptibility to treatment by chemotherapy (Allan *et al*., 2001).

6.2 Gene Annotation

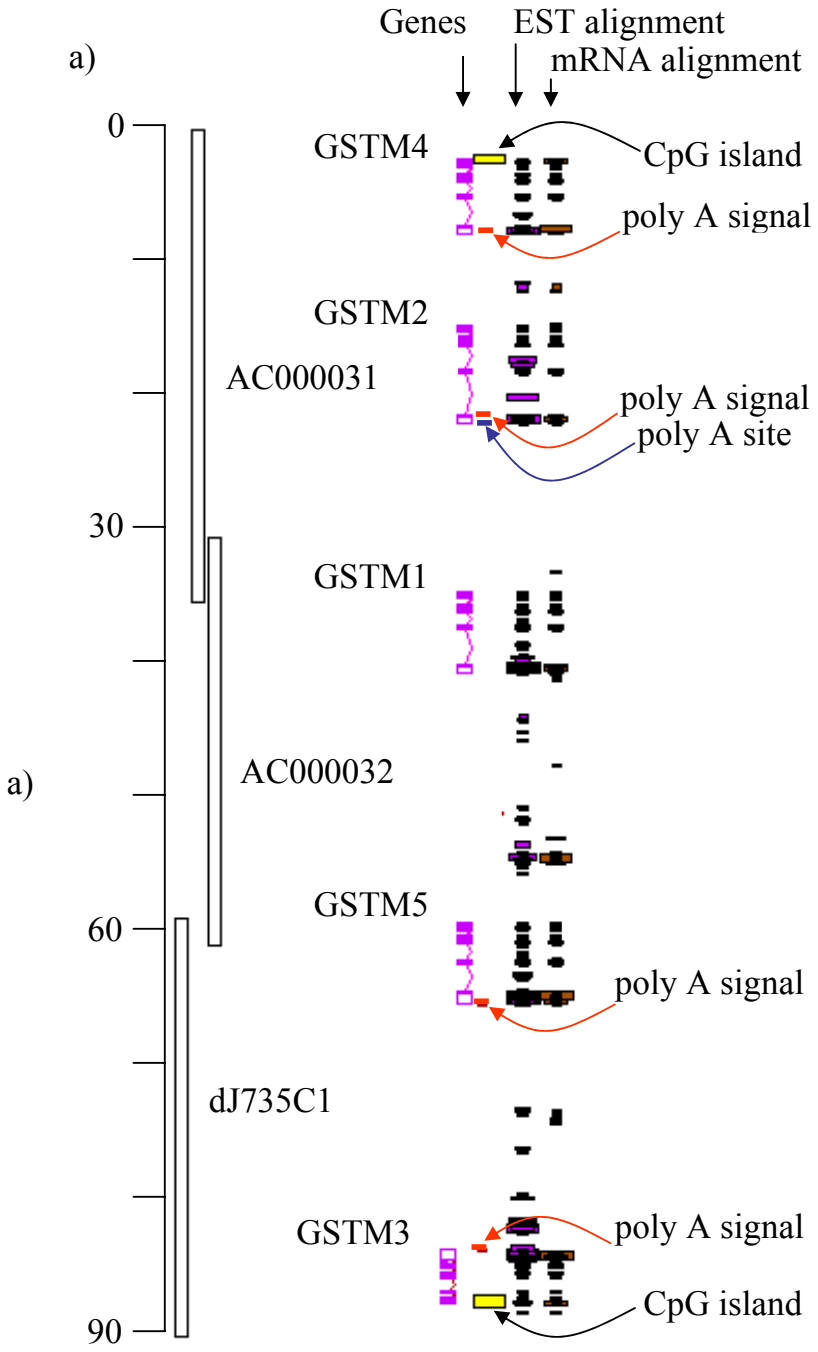
Correct annotation of coding features in their genomic context is central to the accurate design of primers for the identification of single nucleotide polymorphisms that are located within genes. The genes, GSTM 1 - 5, localised to chromosome 1p13 by mRNA / EST BLAST alignment and were annotated during the course of the full genic characterisation of 1pc – 1p13, as described in the previous chapter. The remaining genes considered in this study were either previously annotated as part of a chromosome specific sequencing project, GSTT1 on chromosome 22 (by others), or by *de novo* annotation of clone based sequence within a project specific ACeDB database. These genes, GSTP1, CYP1A1, CYP1A2, CYP2A6, CYP3A4 and NFE2L2, were localised to chromosomes 11, 15, 15, 19, 7 and 2, respectively, by BLAST alignment and the full length structure is manually annotated within clone based sequence.

6.3 Identifying SNPs within Gene Families

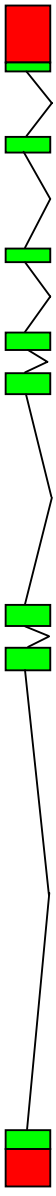
To date, 69% (19626/28574) of known and predicted proteins show homology to existing proteins within Interpro (http://www.ebi.ac.uk/proteome/), suggesting a large

proportion of genes in the human genome are members of a gene family. It is therefore an important consideration when designing primers to identify cSNPs that the assay will identify SNPs from the correct gene and not from a closely related family member. GSTM 1 – 5 were used as a model to assess the difficulties of designing and generating sequence from exon specific primers from closely related genes.

Glutathione S-transferases are a functionally diverse multi-gene family of soluble enzymes which are involved in the detoxification of a wide variety of chemicals via conjugation and reduction of glutathione (Booth *et al*., 1961, Mannervik and Danielson, 1988). Mammalian GSTs are divided into five main classes, alpha (A), mu (M), pi (P), theta (T) and Zeta (Z) and exist as dimeric proteins, with only subunits within the same class forming homodimers. In general, members of the same class share more that 40 – 50% sequence identity but less than 25-30% sequence identity with GSTs in other classes (Hayes and Pulford, 1995). Within the GSTM family, four of the five members (GSTM1, 2, 4 and 5) share a > 70% genomic sequence homology whilst the fifth gene, GSTM3, which is transcribed in the opposite direction, shares a low genomic sequence homology but a high percentage identity with the other four genes at the coding nucleotide (>77%) and amino acid level (>78%). All five genes are evenly distributed within a 90kb interval spanned by one PAC clone (dJ735C1) and two previously sequenced cosmids ctgm1, (AC000031) and cgtm12 (AC000032) (Xu *et al*., 1998), as displayed in figure 6.1a. The structure of all five genes is assumed to be complete as polyA signals were identified within all genes, whilst a CpG islands have been localised to the 5' end of GSTM4 and GSTM3. The consensus GSTM gene structure contains 8 exons, the majority of which are identical in size between family members. The intron size is also conserved with introns 3, 5 and 7, showing most variation (figure 6.1b).


a)
Genes
EST alignment
mRNA alignment
0
GSTM4
CpG island
poly A signal
GSTM2
AC000031
poly A signal
poly A site
30
GSTM1
a)
AC000032
60
GSTM5
poly A signal
dJ735C1
GSTM3
poly A signal
CpG island
90

b)



	M4	M2	M1	M5	-M3
3'UTR	263	14	51	94	171
Exon1	**36**	**36**	**36**	**36**	**48**
Intron1	287	273	260	269	331
Exon2	**76**	**76**	**76**	**76**	**76**
Intron2	427	424	427	425	339
Exon3	**65**	**65**	**65**	**65**	**65**
Intron3	310	300	310	300	1062
Exon4	**82**	**82**	**82**	**82**	**82**
Intron4	100	99	95	95	93
Exon5	**101**	**101**	**101**	**101**	**101**
Intron5	941	1742	945	1185	339
Exon6	**96**	**96**	**96**	**96**	**96**
Intron6	90	90	87	87	88
Exon7	**111**	**111**	**111**	**111**	**111**
Intron7	2054	3163	2641	2096	287
Exon8	**90**	**90**	**90**	**90**	**99**
3'UTR	536	540	539	930	534

Figure 6.1: An ACeDB display of GSTM 1 – 5 and a generic GSTM gene structure. a) The annotated structures of GSTM 1 – 5 contained within two overlapping cosmids (AC000031 and AC000032) and one PAC (dJ735C1) clone. Supporting EST and mRNA sequence, in addition to poly A signals, sites and CpG islands are also shown. b) A generic GSTM gene structure with base pair sizes of exons and introns. Red boxes and green boxes represent untranslated and translated sequence, respectively.

To facilitate the identification of primer pairs that would uniquely flank GSTM exons the genomic sequence of each of the GSTM loci, including approximately 1000 bp preceding the 5‘ UTR and approximately 500 bp beyond the 3’ UTR, was aligned. Sequence was exported from 1ace in Fasta format and analysed within ClustalW (Higgins *et al*., 1994) prior to editing in GeneDoc (Nicholas *et al*., 1997) (figure 6.2).

Figure 6.2 (inserted at end of chapter 6)**:** A genomic sequence alignment of GSTM 1 – 5**:** Bases that were in common between all 5 nucleotide sequences were coloured red, 4 sequences green, 3 sequences blue and 2 or less black. 5’ and 3’ untranslated regions have been shaded yellow whilst coding sequence has been shaded grey and the corresponding exon denoted. PolyA signals at the 3’and sites, where detected, were boxed in black and red, respectively. Coloured shading (yellow, orange, dark green, pink, blue and light green) indicate the positions where primers could be designed within the genomic sequence relative to coding features.

6.4 Primer Design

Attempts were made to design exon flanking primers for each of the 91 exons contained within the 12 target genes. Primers were designed as outlined in materials and methods (section 2.16.1) but with an optimal PCR product size of 600bp. In some instances, because of sequence and product size constraints, primer pairs were designed across multiple exons (e.g. GSTM4 exons 3/4 and 6/7) or overlapped to ensure complete exon coverage (e.g. NFE2L2 exon 5a, b), as listed in table 6.1. All 5’ and 3’ UTR primer pairs were designed to incorporate as much of the untranslated region as possible. Three exons did not have working assays associated with them; primers from exon 8 of GSTM1 and GSTM2 failed primer design due to the high percentage of sequence conservation within the introns of the two genes, whilst the sense primer of exon 2 from GSTM2 failed during primer synthesis due to the enforced high GC content (80%) of the primer.

Table 6.1: A summary of the primers designed to the exons of 12 genes for the detection of coding polymorphisms. EMBL accession number associated with the primer pairs and estimated PCR product sizes are also listed.

Gene	Chr	Exon	stSG #	Size (bp)
GSTM4	1	e1	452701	528
		e2	452702	724
		e3/4	452703	566
		e5	452704	527
		e6/7	452705	701
		e8	452706	689
GSTM2	1	e1	452707	408
		e2	452708	534
		e3/4	452709	719
		e5	452710	614
		e6/7	452711	508
		e8		
GSTM1	1	e1	452712	448
		e2	452713	535
		e3	452714	484
		e4/5	452715	468
		e6/7	452716	709
		e8		
GSTM5	1	e1	452717	448
		e2	452718	548
		e3/4	452719	878
		e5	452720	409
		e6/7	452721	663
		e8	452722	621
GSTM3	1	e1	452723	520
		e2/3	452724	622
		e4/5	452725	463
		e6/7	452726	567
		e8	452727	748
GSTP1	11	e1	158595	465
		e2	158596	384
		e3	158597	385
		e3/4	158591	399
		e5	158592	399
		e6	158593	399
		e7	158594	498
GSTT1	22	e1	140015	348
		e2	140017	351
		e3	452728	276
		e4	452729	406
		e5	140020	659

Gene	Chr	Exon	stSG #	Size (bp)
CYP1A1	15	e1a	452730	556
		e1b	452731	533
		e2/3/4	452732	564
		e5/6	452733	637
		3' UTRa	452734	620
		3' UTRb	452735	697
CYP1A2	15	e1a	452736	700
		e1b	452737	536
		e2/3	452738	753
		e4	452739	511
		e5	452740	484
		e6	452741	464
CYP2A6	19	e1/2	452742	722
		e3/4	452743	626
		e5	452744	588
		e6	452745	404
		e7	452746	467
		e8	452747	443
		e9	452748	538
CYP3A4	7	e1	452755	613
		e2	452756	519
		e3	452757	510
		e4	452758	549
		e5/6	452759	589
		e7	452760	592
		e8	452761	476
		e9	452762	512
		e10	452763	523
		e11	452764	654
		e12	452765	419
		e13a	452766	669
		e13b	452767	791
NFE2L2	2	e1	452768	589
		e2	452769	419
		e3	452770	577
		e4	452771	527
		e5a	452772	649
		e5b	452773	695
		3’ UTR	452774	756

6.5 DNA screening

DNA from the Coriell CEPH/Utah reference family collection (http://locus.umdnj.edu/ccr) was used as template for the generation of exon specific PCR products. A summary of the 47 DNAs used, which represent the founders of the collection, are listed in table 6.2. The pedigree from which one of the DNA originated, CEPH/Utah pedigree 1333 (red underline), is shown in figure 6.3. All individuals used in this study are Caucasian and of Northern European extraction.

Table 6.2: A summary of the CEPH/Utah DNA used for the generation of exon sequence. The maternal grandmother DNA listed in red was not used in this study.

Family ID	Paternal Grandfather	Paternal Grandmother	Maternal Grandfather	Maternal Grandmother
1331	NA07007	NA07340	NA07016	NA07050
1333	NA07049	NA07002	NA07017	NA07341
1341	NA07034	NA07055	NA06993	NA06985
1346	NA12043	NA12044	NA12045	NA12046
1347	NA11879	NA11880	NA11881	NA11882
1362	NA11992	NA11993	NA11994	NA11995
1408	NA12154	NA12236	NA12155	NA12156
1416	NA12248	NA12249	NA12250	NA12251
1423	NA11917	NA11918	NA11919	NA11920
1334	NA12144	NA12145	NA12146	NA12239
1340	NA06994	NA07000	NA07022	NA07056
1420	NA12003	NA12004	NA12005	NA12006

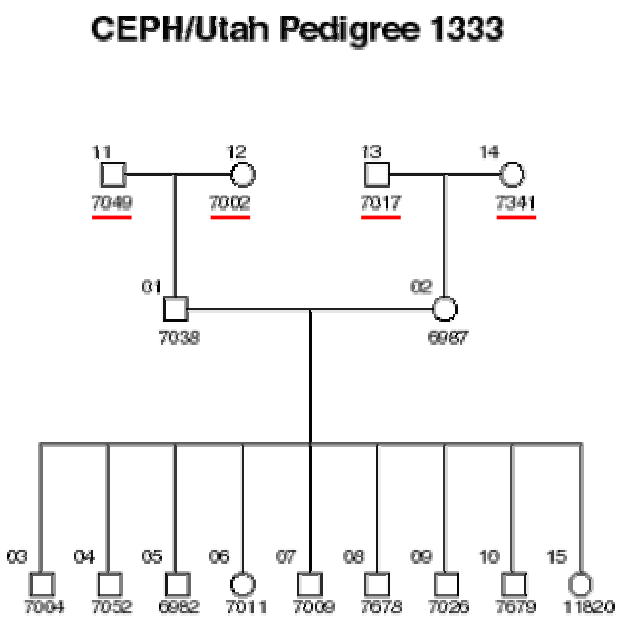


Figure 6.3: A CEPH pedigree. Pedigree represents four of the DNAs used (red underlines) in the *de novo* generation of cSNPs from the CEPH/Utah family collection.

The plate format of the exon specific – CEPH DNA PCR reactions was designed so that all primer pairs were screened across four CEPH DNAs within one 384 well microtitre plate. An aliquot of PCR products from each well of the first, DNAs 1-4, and second plates, DNAs 5-8, were separated by gel electrophoresis to determine the success rate of PCR reactions prior to sequencing (by others). Figure 6.4 shows an example of the resultant PCR products separated by gel electrophoresis, generated at a Tm of 60^{o}C from CEPH DNAs 5-8, of primer pairs designed to exons 1, 2, 3 and 4 of GSTM4. Product sizes of the three exon specific primers were shown to correspond with their estimated size from genomic sequence, 528 bp, 744 bp and 526 bp respectively, but variation in intensity and multiplicity of bands was observed.

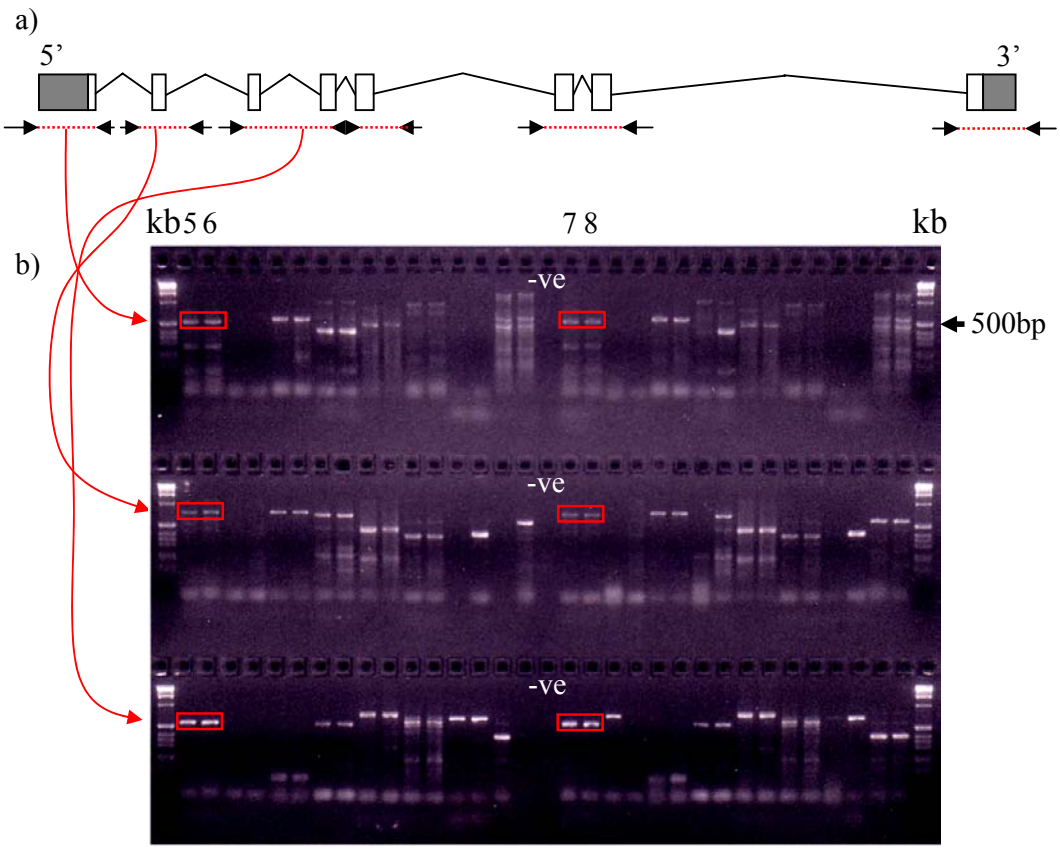


Figure 6.4: Screening of CEPH DNAs with exon specific primer pairs designed to GSTM4. a) Exon specific primer pairs were designed to an annotated gene structure of GSTM4. b) Red arrows indicate the amplification of a 500 bp product from CEPH DNAs 5 – 8 from exons1, 2, 3 and 4 at 60^{o}C. A size marker was included to size PCR products (kb).

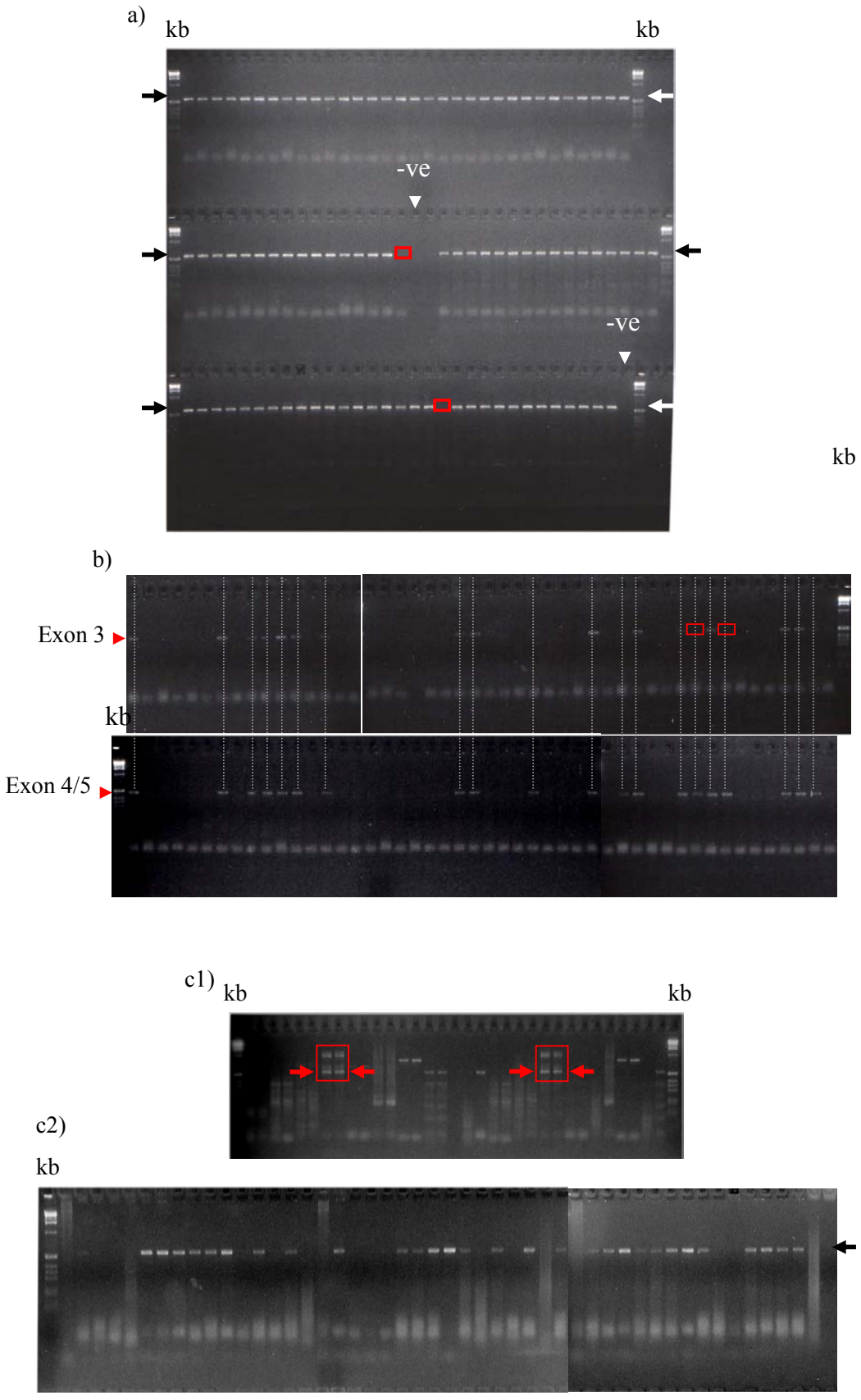
Initial collation of PCR results from electrophoresis of the first two plates (CEPH DNAs 1-8) indicated that well locations C7 – G7 from the primer microtitre plate (exon 5 – GSTP1, 3' UTRa – CYP1A1, exon 7 – CYP2A6, exon 4 – CYP3A4 and exon 3 – NFE2L2, respectively) consistently failed to generate a PCR product. This error was systematic (caused by blocked pins on a liquid handling robot and not through any primer design or DNA template problem) and the missing PCR reactions were repeated. In this instance, each of the 47 CEPH DNAs was tested with the missing primer pairs, in addition to two working GSTM1 primers. Repetition of 158592, 452734, 452746, 452758 and 452770 resulted in three of the five PCR reactions yielding strong single bands following electrophoresis, whilst 158592 generated multiple bands and 452734 failed to generate any product. Products from the repeated round of PCR reactions were sequenced (by others) with those generated from the initial experiment (figure 6.5a).

GSTM1 is known to be a null allele within 50% of the population (Board *et al*., 1990); therefore parallel generation of PCR products from all DNAs would permit an estimation of the percentage of CEPH individuals that miss both copies of the gene. Whilst the intensity of the PCR products when assessed by agarose gel electrophoresis was faint (figure 6.5b), a reproducible relationship could be observed between the DNAs that produced a band from exon 3 and exon 4 / 5 specific primers, thereby confirming the presence of the gene in these individuals. Eighteen of the forty seven DNAs (38%) generated a PCR product for the aforementioned primers whilst a further two DNAs yielded products for one of the primer pairs. The lower observed occurrence of null alleles within the sample set (38% observed compared to 50% expected) may be due to a population bias for the occurrence of the null status of the gene within the

Northern European Caucasian population or may be due to the failure of the faint PCR products to be detected from the electrophoresis gels.

To investigate whether increasing the annealing temperature of primers during the PCR reaction would obviate the production of multiple products, six pairs of primers which generated multiple bands in the first experiment (158592, 452712, 452717, 452731, 452764 and 452767) were tested across each of the 47 CEPH DNAs at 65^{o}C. Two of the six re-tested primer pairs yielded a strong single band following electrophoresis, whilst the remainder failed to generate any product. One of the two successful primer pairs that worked at the more stringent annealing temperature, 453731 covering the 3' end of CYP1A1 exon one, is shown in figure 6.5c. The use of a higher annealing temperatures in PCR reactions as a 'clean-up step' has subsequently been adopted by the Sanger Institute project involved with the large scale production of exon specific sequences.

Figure 6.5 (see over)**:** PCR products from exon primers using CEPH DNA as template. a) PCR primers from exon 4 of CYP3A4 (452758) and exon 3 of NFE2L2 (452770) screened across 47 CEPH DNAs at an annealing temperature of 60^{o}C. Included are size markers (kb), expected PCR product sizes (black arrows) and negative controls ($T_{0.1}E$). DNAs for which no PCR product was generated at boxed in red. b) PCR products generated by primers designed to exons 3 and 4 / 5 and screened across 47 CEPH DNAs generated at an anneling temperature of 60^{o}C. Dotted grey lines join CEPH DNAs that are thought to contain the GSTM1 gene. DNAs for which no PCR product was generated at boxed in red. c1) Initial screening of CEPH DNAs produces multiple bands at 60^{o}C (red box). Increasing the stringency of primer annealing temperature to 65^{o}C c2) results in the generation of a single PCR product (black arrows).


a)
kb
kb
-ve
-ve
kb
b)
Exon 3
kb
Exon 4/5
c1)
kb
kb
c2)
kb

Analysis of the PCR reactions from the 77 primer pairs, covering 91 coding exons of the 12 target genes listed in table 6.3, revealed that 71 (92%) of primers generated bands from at least one of the eight CEPH DNAs when products were separated by gel electrophoresis. The number of working PCR assays was further reduced when primers that produced multiple bands, without the presence of a single band in one of the eight DNAs, were subtracted. These 62 assays (81% of the total) contained two primer pairs (CYP1A1 – exon 1b and GSTM5 – exon 1) for which the products were generated at a primer annealing temperature of 65^{o}C.

Table 6.3: A summary of exon specific PCR reaction results using CEPH DNAs 1 – 8 as a template. GSTM1 – 5 coloured shading corresponding to primers in figure 6.2, the presence 'Y' (yellow shaded if there were multiple bands) or absence 'N' of PCR products upon gel electrophoresis is listed under the CEPH DNA columns. Pink shading indicates products generated at an annealing temperature of 65^{o}C. Grey box, CEPH DNA 3, indicates possible degradation due to high failure rate.

		CEPH DNAs										CEPH DNAs							
Target	**Exon**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Target**	**Exon**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
GSTM4	e1	Y	Y	N	N	Y	Y	Y	Y	**CYP1A1**	e1a	Y	Y	Y	Y	Y	Y	Y	Y
	e2	N	N	N	Y	Y	Y	Y	Y		e1b	Y	Y	Y	Y	Y	Y	Y	Y
	e3/4	N	N	Y	Y	Y	Y	Y	Y		e2/3/4	Y	Y	Y	Y	Y	Y	Y	Y
	e5	N	N	N	Y	Y	Y	N	Y		e5/6	N	N	N	N	N	N	N	N
	e6/7	N	N	N	Y	Y	Y	Y	Y		3' UTRa	N	N	N	N	N	N	N	N
	e8	N	N	N	Y	Y	Y	Y	Y		3' UTRb	Y	Y	Y	Y	Y	Y	Y	Y
GSTM2	e1	N	N	N	Y	Y	Y	Y	Y	**CYP1A2**	e1a	Y	Y	N	Y	Y	Y	Y	Y
	e2										e1b	Y	Y	N	Y	Y	Y	Y	Y
	e3/4	N	N	N	Y	Y	Y	Y	Y		e2/3	Y	Y	N	Y	Y	Y	Y	Y
	e5	N	N	N	Y	Y	Y	Y	N		e4	Y	N	Y	N	Y	N	Y	N
	e6/7	N	N	Y	Y	Y	Y	Y	Y		e5	Y	Y	N	Y	Y	Y	Y	Y
	e8										e6	Y	Y	Y	Y	Y	Y	Y	Y
GSTM1	e1	Y	N	N	N	N	N	Y	Y	**CYP2A6**	e1/2	Y	Y	Y	Y	Y	Y	Y	Y

	e2	Y	N	N	N	Y	N	Y	N		e3/4	Y	Y	N	Y	Y	Y	Y	Y
	e3	Y	N	N	N	N	N	Y	N		e5	Y	Y	N	Y	Y	Y	Y	Y
	e4/5	Y	N	N	N	N	N	Y	N		e6	Y	Y	Y	Y	Y	Y	Y	Y
	e6/7	Y	N	N	N	N	N	Y	N		e7	Y	Y	Y	Y	Y	Y	Y	Y
	e8										e8	Y	Y	N	Y	Y	Y	Y	Y
											e9	Y	Y	Y	Y	Y	Y	Y	Y
GSTM5	e1	N	N	Y	Y	N	N	Y	Y										
	e2	Y	Y	Y	Y	Y	Y	Y	Y	**CYP3A4**	e1	N	N	N	N	N	N	N	N
	e3/4	Y	Y	N	Y	Y	Y	Y	Y		e2	Y	Y	Y	Y	Y	Y	Y	Y
	e5	Y	Y	N	N	Y	Y	N	N		e3	Y	Y	Y	Y	Y	Y	Y	Y
	e6/7	Y	Y	N	Y	Y	Y	Y	Y		e4	Y	Y	Y	Y	Y	Y	Y	Y
	e8	Y	Y	N	Y	Y	Y	Y	Y		e5/6	Y	Y	N	Y	Y	Y	Y	Y
											e7	Y	Y	N	Y	Y	Y	Y	Y
GSTM3	e1	Y	Y	Y	Y	Y	Y	Y	Y		e8	Y	Y	Y	Y	Y	Y	Y	Y
	e2/3	Y	Y	N	Y	Y	Y	Y	Y		e9	Y	Y	Y	Y	Y	Y	Y	Y
	e4/5	Y	N	N	N	Y	N	N	N		e10	Y	N	Y	N	Y	N	Y	N
	e6/7	Y	Y	Y	Y	Y	Y	Y	Y		e11	N	N	N	N	N	N	N	N
	e8	Y	Y	N	Y	Y	Y	Y	Y		e12	N	Y	Y	Y	Y	Y	N	Y
											e13a	Y	Y	N	Y	Y	Y	Y	Y
GSTP1	e1	N	N	N	N	N	N	N	N		e13b	Y	Y	Y	Y	Y	Y	N	Y
	e2	N	N	N	N	N	N	Y	Y										
	e3	Y	Y	Y	Y	Y	Y	Y	Y	**NFE2L2**	e1	N	Y	N	N	N	Y	N	Y
	e3/4	N	N	N	N	N	N	N	N		e2	Y	Y	N	Y	Y	Y	Y	Y
	e5	Y	Y	N	N	Y	Y	N	N		e3	Y	Y	Y	Y	Y	Y	Y	Y
	e6	Y	Y	N	N	Y	Y	Y	Y		e4	Y	Y	N	Y	Y	Y	Y	Y
	e7	Y	Y	N	Y	Y	Y	Y	Y		e5a	Y	Y	Y	Y	Y	Y	Y	Y
											e5b	Y	Y	N	Y	Y	Y	Y	Y
GSTT1	e1	Y	Y	N	Y	Y	Y	N	Y		3' UTR	Y	Y	Y	Y	Y	Y	Y	Y
	e2	Y	Y	N	Y	Y	Y	N	Y										
	e3	Y	N	Y	N	Y	N	N	Y										
	e4	Y	Y	Y	Y	Y	Y	N	Y										
	e5	Y	Y	Y	Y	Y	Y	N	Y										

6.6 Sequence generation and assembly

Dye-terminator sequencing, using exon specific primers and PCR products is described in section 6.5, was carried out by Sarah Lindsay and the high throughput sequencing facility at the Sanger Institute. PCR products were sequenced and then run on ABI 3700 and 3730 sequencing machines. Bases within sequence reads were 'called' using Phred (Ewing *et al*., 1998) and assembled within Phrap (Ewing *et al*., 1998) utilising, amongst other criteria, Phred assigned base-quality values (Q-values). Sequence reads were assembled and imported into gene specific directories by Sarah Hunt. Vector masked and quality clipped sequences reads, together with the consensus sequence, were viewed within gene specific Gap4 (Bonfield *et al*., 1995) databases.

Genomic coverage of the assembled exon specific sequence contigs was determined within the respective ACeDB databases by exact alignment of 10 – 15 good quality bases (light sequences within the assembled contig, figure 6.6) from either end of a sequence assembly. The number of reads contained within an assembly was determined to be the total number of reads contained within a sequence contig.

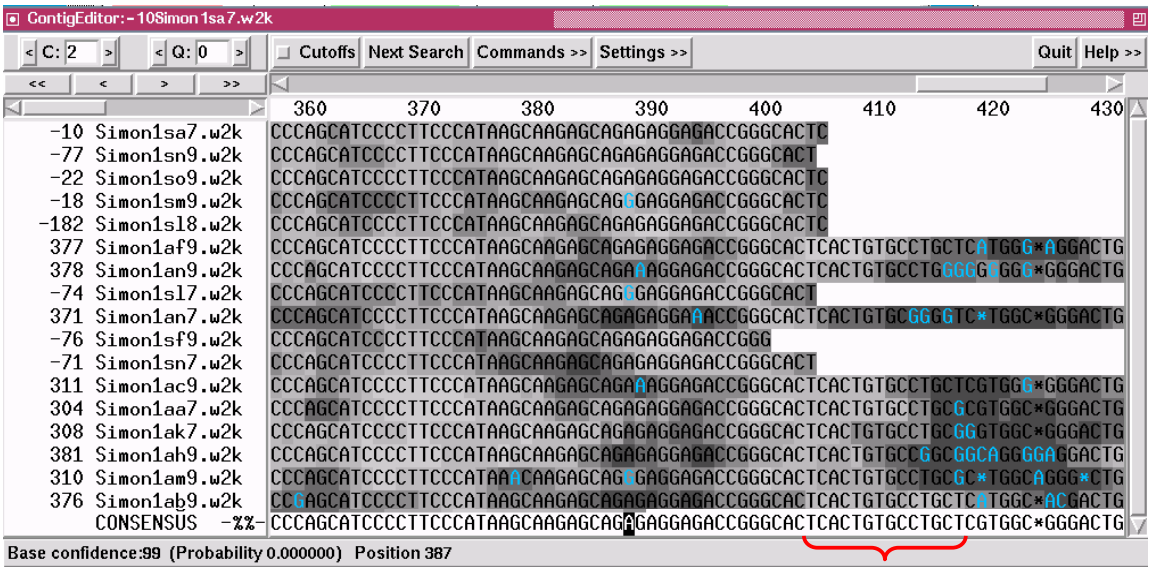
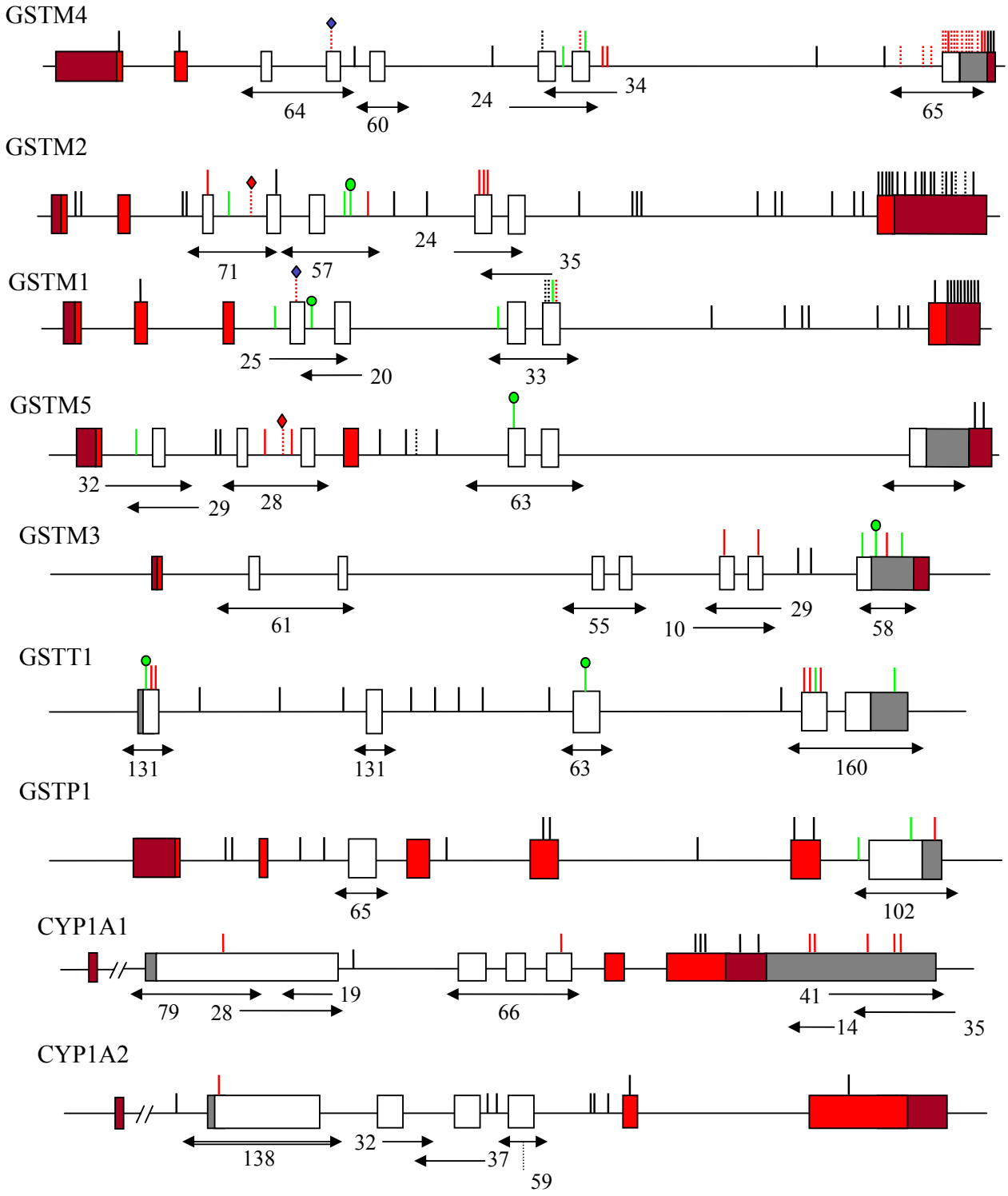


Figure 6.6: Assembly of *de novo* exon specific sequences shown in Gap4. Red brackets indicate the consensus sequence that was used to determine the extent of sequence coverage by alignment within 1ace.

6.7 Exon coverage of sequence contigs

A minimum of ten good quality sequence reads, including both sense and anti-sense reads, was used as the criteria for determining whether *de novo* sequence had successfully been generated to identify known and novel SNPs within an exon. Fifty six of the 77 initial primer pairs (73%) provided data which could be assembled into sequence contigs using the above criteria. This represents 79% of primer pairs excluding those for which PCR assays failed to produce a product. These 56 sequence contigs account for 67 of the 91 exons (74%) from the 12 target genes. Sequences derived from 17 (30%) of the PCRs were assembled separately as groups of sense or anti-sense reads. These were paired up ad counted as a single contig when calculating coverage figures.


GSTM4
64
60
24
34
65
GSTM2
24
71
57
35
GSTM1
25
20
33
GSTM5
32
29
28
63
GSTM3
61
55
10
29
58
GSTT1
131
131
63
160
GSTP1
65
102
CYP1A1
79
28
19
66
41
14
35
CYP1A2
138
32
37
59

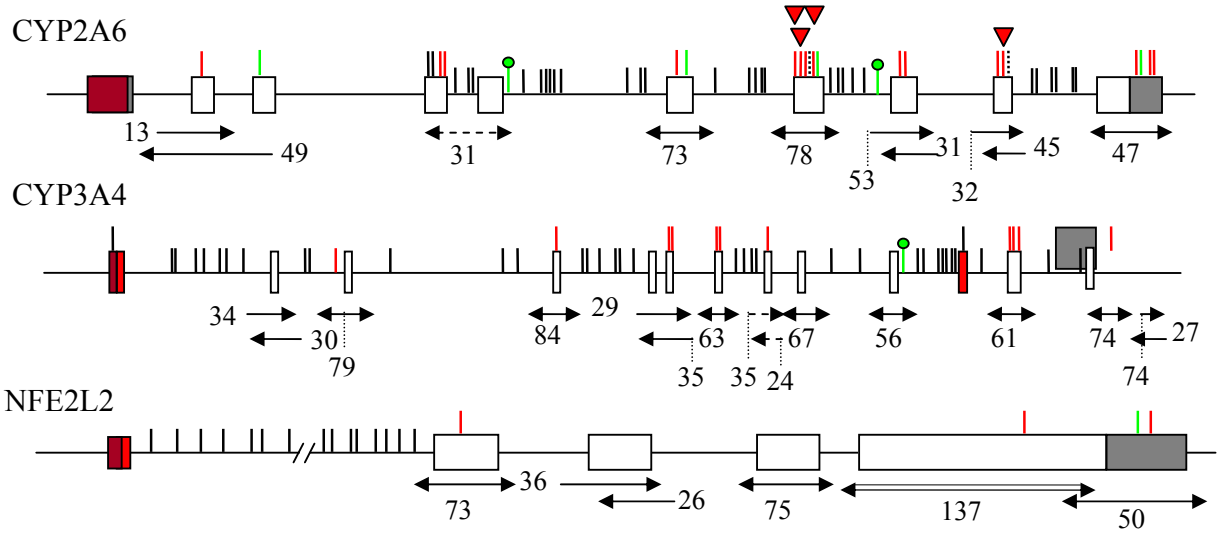


Figure 6.7: A summary representation of the *de novo* sequence coverage of exons from the target genes. Clear and grey boxes represent *de novo* sequence coverage of exons and untranslated region, respectively (not drawn to scale); exons and UTR without sequence coverage are red and dark red, respectively. Double headed arrows indicate sequence from both primer pairs assembled into one contig (number indicates the total number of reads assembled); single headed arrows specify assembly of sequence from one primer; double lined arrows are where four primers have been assembled. Black vertical lines represent known SNPs not covered by sequence coverage; black dotted vertical lines indicate where the flanking sequence from the known SNP did not align to genomic sequence; red vertical lines are known SNPs not identified by this study; vertical red dotted lines are known SNPs with multiple loci within genomic sequence (red and blue diamonds denote where extended genomic flanking sequence alignment was required to localise the SNP); green vertical lines are known SNPs found within *de novo* sequence; green vertical lines with round heads are novel SNPs identified here.

6.7.1 Validation and localisation of known SNPs

One hundred and three (35%) of the two hundred and ninety one known SNPs, which are derived from dbSNP (http://www.ncbi.nlm.nih.gov/SNP/index.html URL), and localised to the 12 genes of interest within Ensembl (http://www.ensembl.org/), are covered by the *de novo* sequence contigs (figure 6.8a). To assess whether these SNPs were present within annotated gene structures the known SNPs, and their flanking sequence, were aligned to the genomic sequence within respective ACeDB databases. This alignment identified 17 SNPs which are primarily found within GSTM4, but also localised to multiple genes and 8 SNPs for which the submitted sequence could not be found when carrying out the alignment using 15-20 bp either side of the SNP, including each allele in the search.

The remaining 76 known SNPs were localised along with their flanking sequence within the *de novo* sequence contigs by sequence matching within respective Gap4 databases. Each read of the assembled sequence contigs were then checked manually for the presence or absence of the relevant SNP allele, in either homozygous or heterozygous form (figure 6.8c). An important consideration when identifying known SNPs within the sequence assembly was the orientation of the sequence contig relative to the forward or reverse read data associated with the submitted SNP. The majority, 59 (78%), of the 76 uniquely localised known SNPs were not found within the CEPH DNAs tested here. Of the remaining 17 SNPs (22%), 8 were determined to be located within coding sequence, and 9 were located within non-coding sequence.

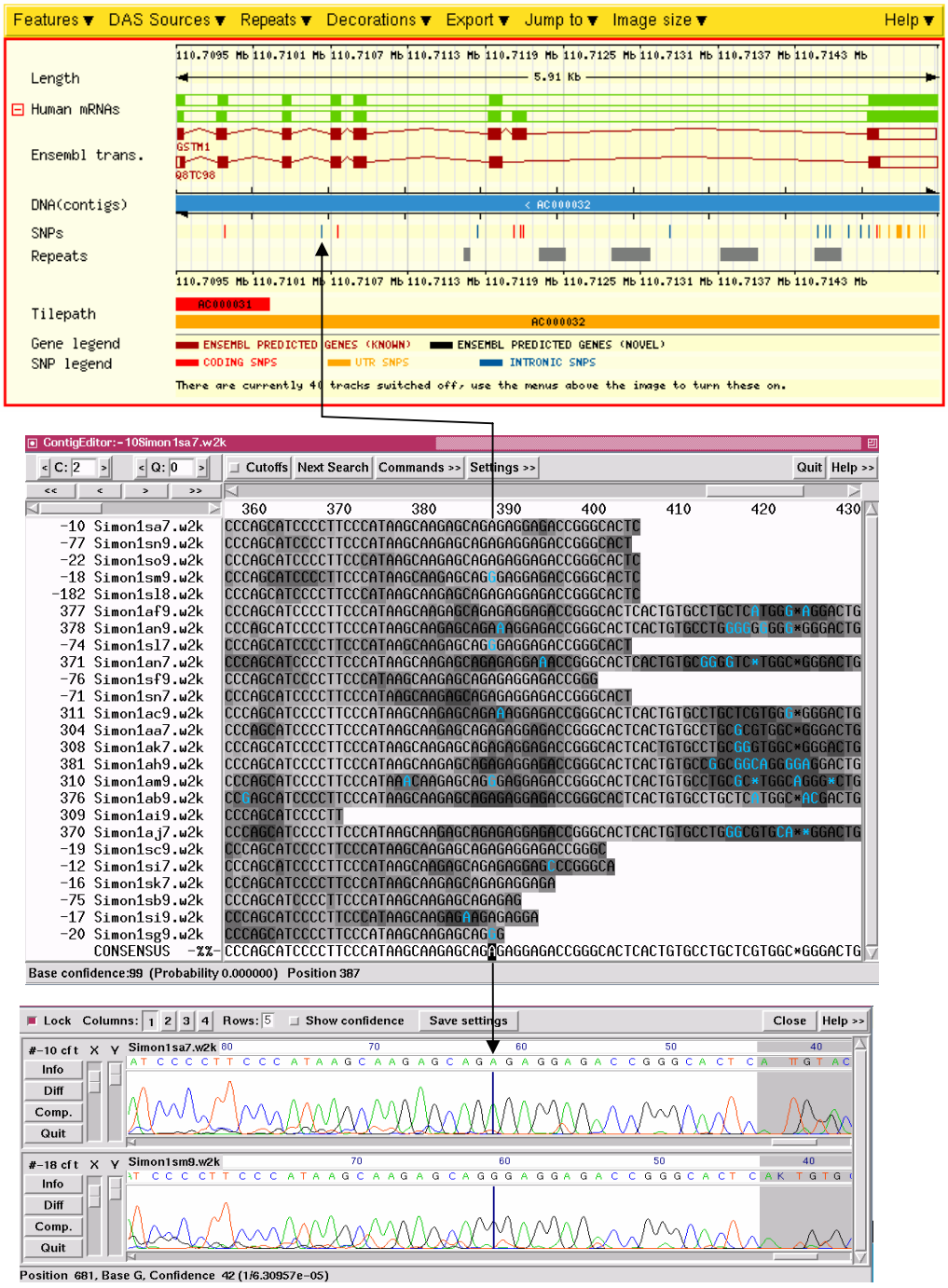


Figure 6.8: Identification of a known SNP which is within intron 3 of GSTM1. Represented is the known SNP within Ensembl, a), the specific nucleotide position within a sequence assembly displayed in a Gap4 database, b), and the homozygous sequence reads for each allele.

6.7.2 Identification of novel SNPs

To more readily identify novel SNPs within assembled Gap4 sequence contigs the quality value of the reads was converted into grey-scale and the ambiguous base calls, and potential novel SNPs, were coloured blue (figure 6.8b). Ambiguous bases were then checked manually within all traces. A putative novel SNP was identified when the base containing the homozygous minor allele had a quality value of 30 (Q-30) or the heterozygote, containing both alleles, was surrounded by bases with > Q-30 quality values. Nine novel SNPs, three of which were located within coding sequence, were identified within the 12 target genes. Additional support for a novel SNP was obtained by checking the reverse read for each read pair. A breakdown of known and novel SNP numbers are listed in table 4 whilst a summary of the placement of all SNPs within annotated gene structures can be found in figure 6.7.

Table 6.4: A summary of the known and novel SNPs associated with 12 target genes.

Gene	Known SNPs	With Seq Coverage	Uniquely Assigned	Align to Genomic	Not Observed	Observed	Coding	Novel	Coding
GSTM4	32	23	8	7	5	2	1	0	-
GSTM2	43	8	7	7	5	2	-	1	-
GSTM1	26	7	5	3	-	3	1	1	-
GSTM5	12	4	4	3	2	1	-	1	1
GSTM3	7	5	4	4	3	1	1	1	-
GSTP1	13	3	3	3	1	2	1	0	
GSTT1	16	7	7	7	5	2	1	2	2
CYP1A1	13	7	7	7	7	0	-	0	-
CYP1A2	9	1	1	1	1	0	-	0	-
CYP2A6	54	23	23	19	16	3	3	2	-
CYP3A4	47	11	11	11	11	0	-	1	-
NFE2L2	19	4	4	4	3	1	-	0	-
TOTAL	**291**	**103**	**84**	**76**	**59**	**17**	**8**	**9**	**3**

6.8 SNP Analysis

Under conditions of random mutation it is expected that half as many nucleotide transitions occur (purine→ purine or pyrimidine→pyrimidine) as transversions (pyrimidine←→ purine), table 6.5 (Dawson *et al*., 2001). In addition, as the strand on which the change has occurred is unknown, no distinction can be made between A↔G and C↔T and between C↔A and G↔T. Analysis of 28 SNPs contained within this data set revealed that 19 transitions (68%) occur approximately twice as often as 9 transversions (32%) (table 6.6). The observed contradiction to the expected frequency has previously been reported (Horton *et al*., 1998; Dawson *et al*., 2001; Halushka *et al*., 1999, Deutsch *et al*., 2001). A possible reason for the occurrence of an increased number of C↔T (G↔A) transversions may be due to the deamination of 5-methylcytosine that occurs frequently at CpG dinucleotides.

Table 6.5: Expected occurrence of transitions and transversions in genomic sequence.

	A	**T**	**C**	**G**
A	-			
T	Transversion	-		
C	Transversion	Transition	-	
G	Transition	Transversion	Transversion	-

Table 6.6: The observed number of transitions and transversions of known and novel SNPs within the 12 target genes.

Variation	**Number**	**Percentage**
C↔T (G↔A)	19	68.00%
Transitions	**19**	**68.00%**
C↔A (T↔G)	6	21.00%
C↔G	3	11.00%
A↔T	-	-
Transversions	**9**	**32.00%**

Of the 28 detected novel and known SNPs, 12 were localised to introns, 5 to 3' untranslated regions, whilst 11 occurred within exons (table 6.7). Seven of the 11 coding SNPs resulted in synonymous changes, i.e. maintaining amino acid type by codon usage, whilst the remaining four SNPs resulted in an amino acid change. Each of the four amino acid changes were caused by substitutions in the first base of the codon, two of which resulted in a conservative change (maintaining a similar amino acid type, for example valine to an isoleucine change maintains an aliphatic, hydrophobic amino acid) and the remaining two changes were non-conservative changes.

Table 6.7: Exonic SNP analysis.

Gene	Exon	Position	Allele	change	Codon	change	Encoded	a. a.	Synon	Cons
GSTM4	7	78	C	↔ T	TTT	↔ TTC	phe	↔ phe	yes	-
GSTM1	7	72	C	↔ T	GAC	↔ GAT	asp	↔ asp	yes	-
GSTM5	6	22	T	↔ C	TTG	↔ CTG	leu	↔ leu	yes	-
GSTM3	8	91	G	↔ A	GTA	↔ ATA	val	↔ iso	no	yes
GSTT1	1	13	C	↔ T	CTG	↔ TTG	leu	↔ leu	yes	-
GSTT1	3	110	A	↔ C	ACG	↔ CCG	thr	↔ pro	no	no
GSTT1	4	155	G	↔ A	GTA	↔ ATA	val	↔ iso	no	yes
GSTP1	7	111	A	↔ C	AGT	↔ ACT	ser	↔ ser	yes	-
CYP2A6	2	37	T	↔ C	TTG	↔ CTG	leu	↔ leu	yes	-
CYP2A6	5	117	C	↔ T	CGC	↔ CGT	arg	↔ arg	yes	-
CYP2A6	6	100	C	↔ T	CGC	↔ TGC	arg	↔ cys	no	no

6.8.1 Validating SNPs within highly homologous genes

One of the difficulties encountered whilst identifying valid SNPs associated with GSTM 1- 5 paralogues was to determine whether a SNP was valid (i.e. two allelic variants at the same locus) or merely the alignment of two locus specific base pair variants, one from each of two closely related gene family members. A comparison between the

known SNPs localised to GSTM 1 – 5 and the number which were uniquely placed (columns 2 and 4 in table 6.4) indicates the scale of the problem that was encountered, i.e. 65% of GSTM1 SNPs that had *de novo* sequence coverage localised to other GSTM loci. Three criteria were used to assist with the elucidation of a valid candidate SNP from locus specific variants of paralogous genes; 1) the SNP must be found within *de novo* sequence traces, 2) it must be in Hardy-Weinberg (H-W) equilibrium (i.e. allele 1 = p and allele 2 = q, $p^2 + 2pq + q^2 = 1$) and 3) the alternative allele should not be present as a base variant in a paralogous gene. Valid SNPs were divided into 'known', i.e. those that have a dbSNP entry, and 'novel', i.e. those identified in this study. 'Suspect candidate' SNPs fulfilled criteria 1) and 2) but the alternate SNP allele was be present within a homologous gene. 'Rejected candidates' are not observed in sequence traces (due to BLAST alignment of the flanking sequence of a SNP from a closely related gene family member), the alternative allele was present within paralogues and it was not in H-W equilibrium. The observed occurrence and genic location of known, novel, suspect and rejected GSTM 1 – 5 SNPs is shown in table 8.

Table 6.8: Summary of categorised GSTM 1 – 5 SNPs.

Status	Exon	Intron	3' UTR	Total
Known	1	4	2	**7**
Novel	-	3	-	**3**
Suspect	3	1	-	**4**
Rejected	6	3	9	**18**

6.8.2 Validating SNPs

6.8.2.1 Known SNPs

The presence of a valid known SNP was easily detected once its location within the *de novo* sequence contig had been identified using flanking sequence matches within the respective Gap4 database. The alternative minor allele, even when present as a heterozygote, was clearly discernable within good quality sequence. An addition example of a known SNP, dbSNP: 737497, is shown in figure 6.9. The SNP was found within the *de novo* sequence as a homozygote in major and minor alleles and localises to intron 3 within GSTM1. The T233C SNP was submitted to dbSNP by The SNP Consortium (TSC). The orientation of the flanking sequence had been designated as a forward read within dbSNP. However, integration of the cosmids (from which the SNP was derived) within the minimum tile path (chapter 4) indicated the submission, and therefore the SNP alleles, was in the wrong orientation. The allele frequency of the major and minor alleles, with Q-values >30, was 0.75 and 0.25, respectively, from 24 chromosomes sampled. The SNP was also in H-W equilibrium.

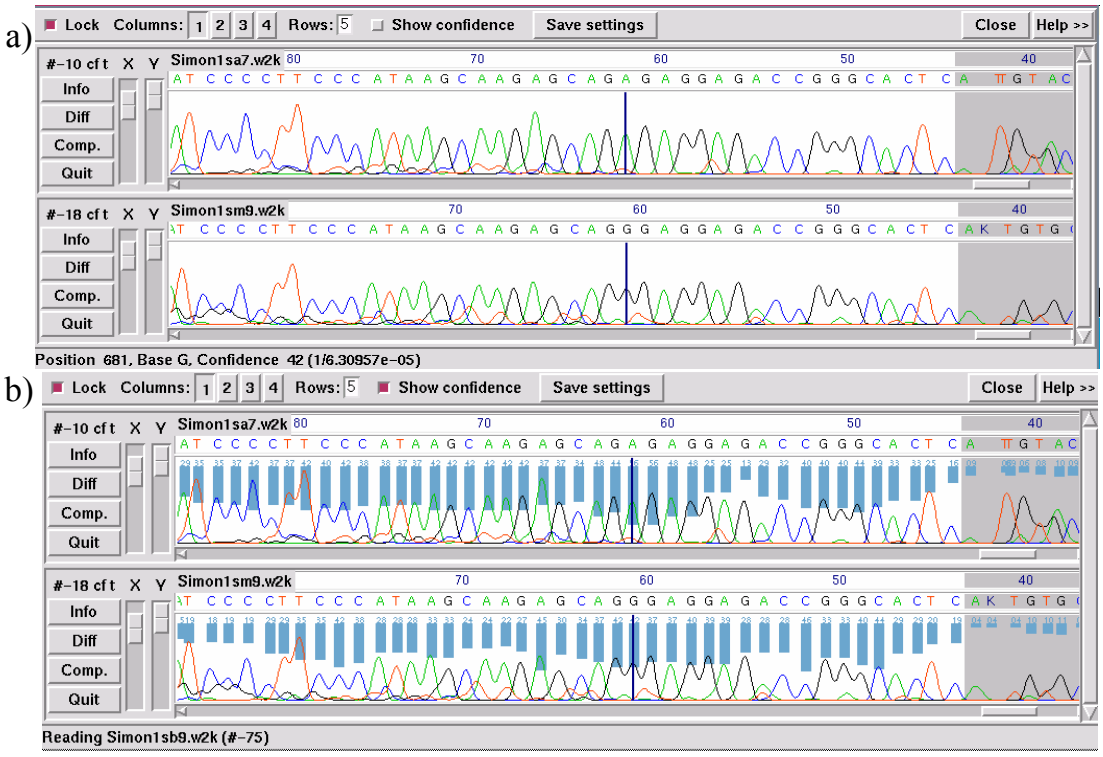


Figure 6.9: The identification of a known T/C SNP, dbSNP: 737497. a) Assembled GAP4 traces used to identify this SNP are displayed in the reverse orientation. The top trace is the homozygous major allele and the bottom trace is the homozygous minor allele. b) Q-values associated with the SNP nucleotides.

6.8.2.2 Novel SNPs

Novel SNPs were validated using support from the base quality values, either of the homozygote minor allele or of the surrounding bases for a heterozygote allele. Where possible, SNPs were verified within reverse reads. All novel SNPs that were identified as part of this study were in HW equilibrium. Figure 6.10 is an example of a novel A152C SNP localised to intron 5 of GSTM2 and identified in the opposite orientation. Allele frequencies of the major (0.96) and minor (0.4) alleles were determined from a sample of 48 chromosomes.

a)

b)

Figure 6.10: A novel A/C SNP identified within intron 3 of GSTM2. a) Assembled Gap4 traces for the homozygous major allele, top trace, and the heterozygous minor allele, bottom trace (in the reverse orientation). b) Assembled traces (from 6.10a) with associated Q.

6.8.2.3 Suspect candidate SNPs

Within Ensembl, dbSNP entries 1056806 and 506008 were localised to exon 7 of GSTM1 and 4 in positions 72 and 78, respectively. These two SNPs were also localised within GSTM1 and GSTM4 in dbSNP. Alignment of the SNPs within the coding sequence, figure 6.11, revealed that the minor allele, T at position 72 and C at position 78, was present within other gene family members. Therefore, it was uncertain if these

suspect SNPs were valid, or were incorrectly aligned to the genomic sequence. Errors associated with SNPs that have been identified by sequence based detection may arise from; re-sequencing without careful selection of primers; lesser quality shotgun sequencing and subsequent alignment of the genomic read; the alignment of the SNP derived reference sequence to more than one genomic position. The positions of dbSNP:1056806 and dbSNP:506008 were localised within the assembled sequence contigs of exon 7 in GSTM1 and GSTM4. Analysis of the sequence traces at positions 72 and 78 revealed that major and minor alleles could be identified therefore satisfying criteria 1), however detailed analysis of the flanking sequence for each SNP uniquely assigned 1056806 to GSTM1 and 506008 to GSTM4. Having determined that, within the data set described here, the two SNPs belonged to different genes, the allele frequencies and H-W equilibrium were calculated (therefore satisfying criteria 2). The allele frequencies of dbSNP:1056806 were 0.58 and 0.42, whilst the frequencies of dbSNP:506008 were 0.85 and 0.15. Both loci were determined to be in H-W equilibrium.

Having proven, within this data set, the SNPs were locus specific, there remains the possibility the minor allele may have arisen from the production of sequence from a mispriming event of a gene family member. Whilst there is a high degree of sequence conservation within exon 7 there are single base variants between loci which should be present in the sequence traces if the primers had misprimed. The bases of the alternative alleles, located at positions 28, 68, 80 and 85 of figure 6.11, should be present as heterozygous alleles in all traces. Analysis of the sequence failed to show any bi-allelic reads at these positions therefore supporting the unique assignment of the two submitted SNPs


Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
Exon 6
Exon 7
Exon 8
28
68
72
78
80
85
GSTM1
GSTM5
GSTM2
GSTM4
GSTM3

Figure 6.11: The alignment of coding sequence for the GSTM1 – 5 genes using ClustalW and representation within GeneDoc. Exons are alternatively coloured light blue and grey. Red boxes denote the position of SNPs dbSNP:1056806 and dbSNP:506008 at positions 72 GSTM1 and position 78 GSTM4, respectively. Dotted red boxes indicate nucleotide variants between genes which permitted that unique assignment of the two SNPs.

6.8.2.4 Rejected candidate SNPs

The identification of SNPs by mRNA BLAST alignment to genomic sequence can erroneously localise SNPs within highly homologous genes. Seventeen GSTM SNPs with *de novo* sequence coverage were rejected as candidate SNPs within this study because they could not be uniquely placed by genomic alignment within a chromosome specific ACeDB. Thirteen of these SNPs could not be uniquely assigned due to the failure of two of the five paralogous genes to generate 3'UTR sequence (to which the majority of ambiguous matches aligned). Detailed analysis of one of the remaining four rejected candidates, dbSNP: 3211191, which localised to exon 4 of GSTM1 and GSTM4 (blue diamond in figure 6.7), indicated that exact sequence alignment of 20 bp either side of the SNP failed to align uniquely assign it to either individual GSTM gene. There was 100% sequence homology extending for 58 bases 5' and 41 bases 3' of the SNP into intron 3 and intron 5, respectively. The observation of a single base difference between the two genes, in the present study, allowed assignment of the read and therefore the SNP to GSTM4. A second SNP that failed to localise to a unique position by sequence matching was dbSNP: 402505. The SNP was located within intron 3 of GSTM2 and 5. It had 100% homology both 5' and 3' of the SNP but nucleotide

differences at 155 bases 5' and 46 bases 3' of the SNP, from the present study, permitted assignment of the SNP to GSTM5. The remaining two of the seventeen unlocalised SNPs were rejected on based upon similar arguments to those examples described above.

6.8.3 Effect of sequence variation upon gene structure

A novel SNP, identified within the *de novo* sequence of GSTT1, provided the opportunity to examine the effects allelic variation may have upon the translated amino acid sequence and protein structure. The protein encoded by GSTT1, like the GSTM family members, is a phase 2 enzyme that detoxifies carcinogenic metabolites, for example halomethanes, by conjugation of glutathione which changes the polarity of the metabolite making them more readily excreted (Mannervik *et al*., 1988).

It has previously been shown that the GSTT1 gene is absent from 38% of the population (Pemble *et al*., 1994). Analysis of the set of DNAs examined in this study indicated that only 1 of 8 CEPH DNAs tested, NA07017, failed to generate a PCR product from any primer pair of the 5 exons from the gene. The lower observed null percentage (13%) compared to that previously reported may relate to the small sample set tested.

The novel A110C SNP, within exon 3 of GSTT1, results in a first base substitution of amino acid 104 causing a non-synonymous, non-conservative change of a threonine to a proline (Thr104-Pro). Threonine is an aliphatic amino acid that has a neutral side chain whereas proline is an amino acid with a secondary group. Alignment of the protein domain containing the single nucleotide polymorphism was performed within PFAM

(http://www.sanger.ac.uk/Software/Pfam/). Of the 751 proteins that have conserved homology to the domain encoding the Thr104 – Pro nucleotide variant none contain a proline at position 104. This suggests that the protein conformation resulting from the non-synonymous, non-conservative SNP is not usually assumed by GSTT1 or any closely related protein domain. A homologous protein with an elucidated 3-D structure was identified within PDB by sequence alignment to investigate what effect the SNP may have upon the conformation of GSTT1. 3LJR, whose 3-D structure as a dimer conjugated with glutathione substrate was elucidated by X-ray diffraction, is a glutathione S-transferase family member, GSTT2, that shows a 55% sequence homology with GSTT1 (figure 6.12). The substitution Thr104 to Pro in GSTT1 corresponds to Asp104 to Pro in 3LJR. Viewing the elucidated 3-D structure within ICMLite (http://www.molsoft.com/products/icmlite.htm, Abagyan *et al*., 1994) (with assistance from Robert Steward) shows Asp104 is at the C-terminal of an α-helix and shares a hydrogen bond with Trp101. This hydrogen bond cannot exist with a Pro substitution, since the Pro residue is cyclic with no free NH group. The loss of this hydrogen bond by an introduction of a Pro may destabilise the 3D conformation of the protein (Karvonen *et al*., 1998).

The substitution of a proline at amino acid position 104 may have a significant affect upon the function of the protein as Asp 104 (in 3LJR) conjugates with the glutathione substrate within the other subunit of the homodimer. The structural affect the proline substitution may also be important as an adjacent residue, Arg 107, interacts with the glutathione substrate on the same subunit (figure 6.13). Therefore a change in geometry caused by Thr104 – Pro may effect the conformation of active site binding residues (thereby effecting protein function). Further experimental studies would be required to

determine whether the introduction of a Pro at amino acid position 104 would affect GSTT1 enzyme function when conjugating glutathione as a homodimer.

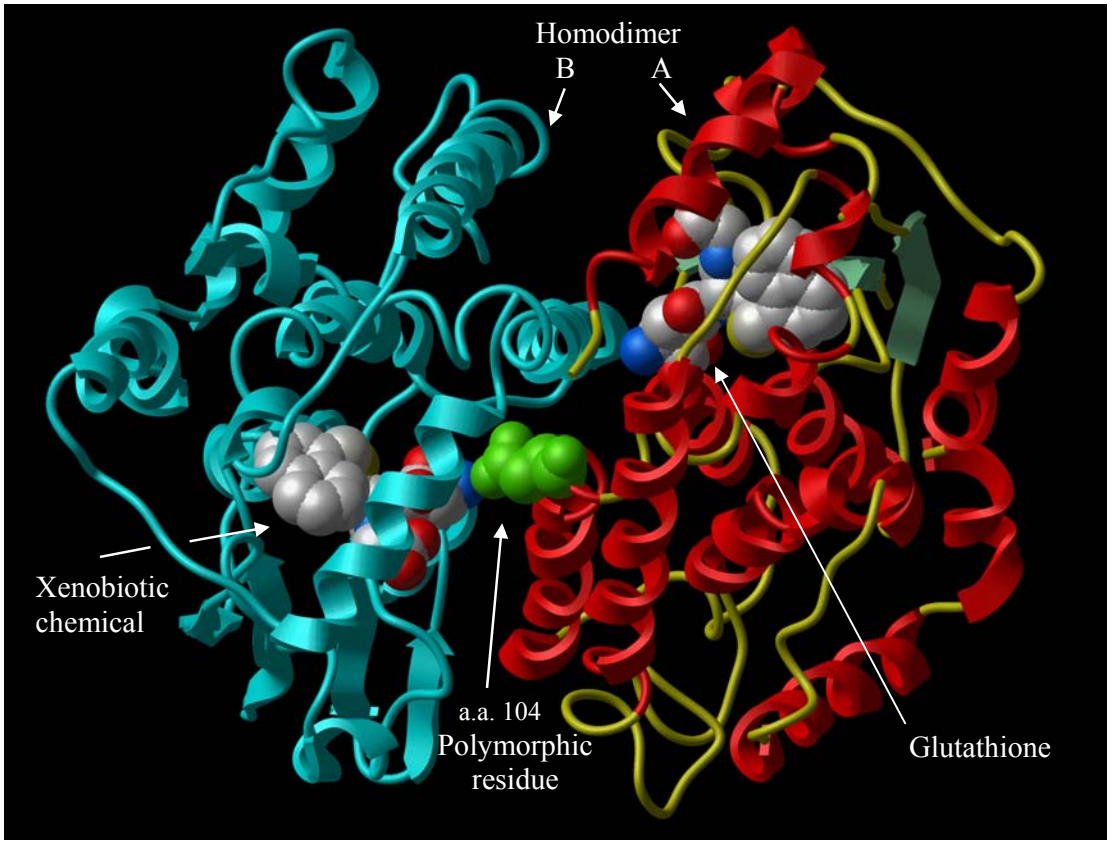


Figure 6.12: A 3-D representation within ICMLite of the homodimeric GST model, 3LJR. The model was used to interpret the effect that a novel cSNP within GSTT1 may have upon the structure of the protein. Shown are the two GST units that form the dimer (yellow/red and green) conjugated with glutathione and xenobiotic biochemicals. The polymorphic Asp104Pro residue is drawn in green at the C-terminal of an α-helix.

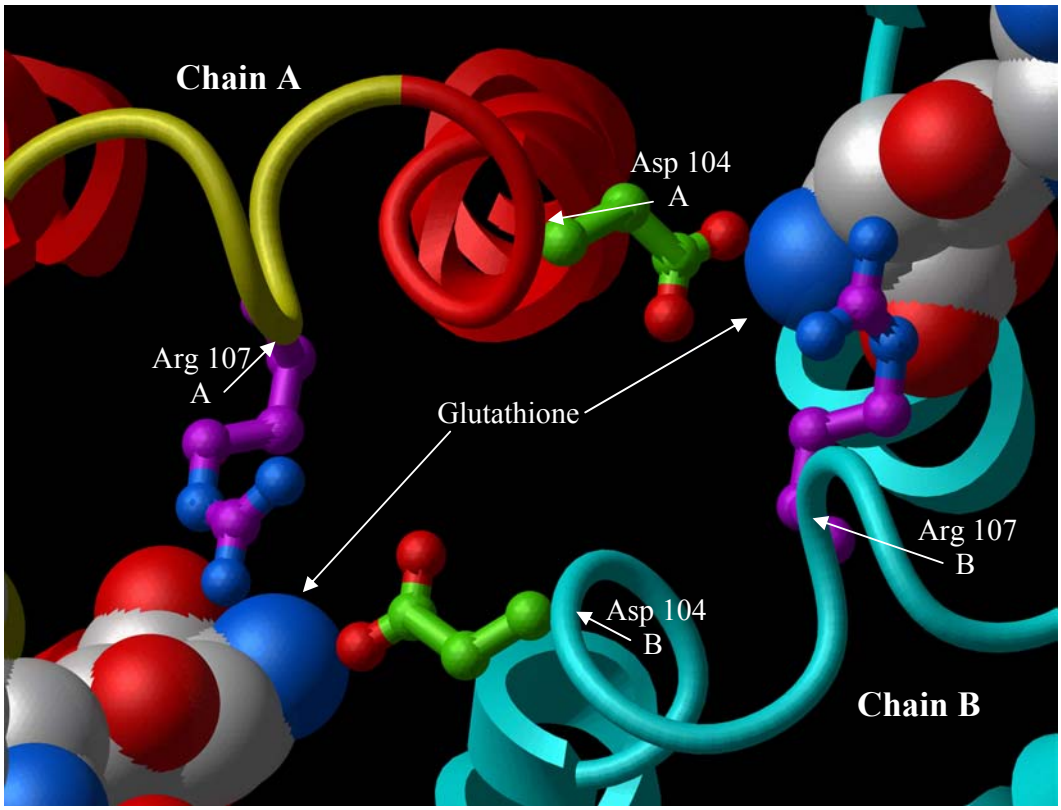


Figure 6.13: A vertical cross-sectional view of glutathione conjugating amino acid residues of 3LJR. Polymorphic amino acid residue, Asp 104, conjugates with the glutathione molecule on the opposite chain in the homodimer, chain B. Arg107 on chain A conjugates with glutathione residue within the same chain. The conformation of the dimmer means that hydrogen bonding conjugation is reciprocated between dimers.

6.9 Discussion

Novel sequence coverage has been generated for 77 (85%) exons from 12 genes that has identified 17 known and 9 novel SNPs. Exon specific primers were designed within highly homologous glutathione-S-transferase Mu 1 – 5 paralogous genes and seven other genes of medical interest to facilitate the unique assignment of known and identify novel SNPs.

The unambiguous localisation of SNPs within genomic sequence is central to determining the biological effects of coding SNPs or to their use in the construction of haplotype maps. Two known SNPs, with multiple loci according to the Ensembl genome database, were uniquely localised with sequence contigs assembled from the GSTM exon specific PCR products. A further 2 known SNPs with multiple loci, whose placement was not resolved within *de novo* sequence contigs, were uniquely assigned by sequence alignment of extended flanking sequence within an ACeDB database.

Fifty five of the 291 known SNPs localising to the genes in this study were not detected within exon, and partial flanking intron, specific sequence contigs. The absence of these SNPs does not necessarily mean that they are false but may instead be attributed to the ethnicity of the population from which the SNP was derived being different to the population tested here. Alternatively, the number of chromosomes contained within the reads of the sequence assembly may not be sufficient to detect the SNP if it has a low minor allele frequency (in this study, it would be unlikely for a minor allele of around 1% or less to be detected).

Eight known SNPs covered by *de novo* sequence failed to align to genomic sequence within genes structures in respective ACeDB databases. This may be caused by nucleotide differences within draft sequence from which the SNPs were identified (for example, dbSNP: 2545753 within exon 6 of CYP2A6) which were not subsequently present in the finished sequence. The alignment of mRNA sequence to genomic sequence to aid SNP identification may also cause inaccuracies in SNP assignment if there are errors in original mRNA sequence (for example, dbSNP: 1061604 within exon 8 of CYP2A6). In addition, if the mRNA was derived from a closely related paralogous gene, it is possible the mRNA sequence may be misaligned to the genomic sequence of a highly homologous locus.

There are approximately 60,000 coding SNPs present within the human genome (ISMWG, 2001), corresponding to 1-2 per gene per individual. However, the effect of these SNPs upon the function of genes is largely unknown. Many of these cSNPs will result in synonymous amino acid changes due to codon redundancy or conservative non-synonymous changes by the substitution of an amino acid of similar properties. However, a proportion of cSNPs will result in non-synonymous non-conservative changes that, depending on the amino acid substitution, may cause changes in the structural integrity and/or biological function of the protein. As discussed in the previous chapter, the structure, function and interaction of the majority of human gene products is largely unknown. The determination and characterisation of protein structures and the networks in which they interact, together with the elucidation of genuine SNPs within coding features will contribute to our understanding of the metabolic differences between individuals.

This chapter describes the possible effect that a non-synonymous, non-conservative coding SNP, Thr104 – Pro, may have upon the structure and function of the GSTT1 protein. Since this work, discovery of the same SNP in a Swedish population has been published (Alexandrie *et al*., 2002). Experimental investigation by Alexandrie *et al.,* (2002) has confirmed that the variant does indeed have a functional consequence. The so-called GSTT1*B allele was reported as having a frequency of 0.05 in Swedish Saamis, the same as observed within the CEPH DNAs used here. An ELISA assay showed that a stable protein product was not produced by individuals who had previously been genotyped as non-conjugators, lacking a GSTT1 protein that could functionally conjugate methyl chloride. Western blot analysis was also used to determine that the functional protein was absent within erythrocyte lysates from non-conjugating individuals. The association of the GSTT1*B, allele in conjunction with the functional GSTT1*A or the null GSTT1*0 allele, may explain why individuals that were previously thought to produce functional copies of the GSTT1 protein are associated with the non-conjugating phenotype.


Exon 1
Exon 2
Exon 3
Exon 4*
Exon 5
Exon 6
Exon 7
Exon 8*
Exon 5 - M3
Exon 6 - M3
Exon 8* - M3

Chapter 7

Discussion

7.1 Genome Mapping and Sequencing

The generation of genome-wide physical maps of biologically complex organisms with large (i.e. multi-gigabase) genomes has relied upon the application of strategies developed for the characterisation of smaller (~1-100 megabase) genomes. At the time of their construction, the cosmid and lambda clone physical maps of *Caenorhabditis elegans* (*C. elegans*) (Coulson *et al*., 1986) and *Saccharomyces cerevisiae* (*S. cerevisiae*) (Olson *et al*., 1986), respectively, were intended to facilitate the cloning of known genes and to serve as a genomic archive. It was during the construction of these physical maps that concomitant improvements in sequencing technologies enabled a whole genome sequencing strategy to be developed. Clones from the physical map that represented the genomic archive were used as the substrate for subcloning and shotgun sequencing. As size of the clones used to construct these maps was equivalent to large viral genomes, which had already been successfully sequenced using the random shotgun approach (e.g. 48 kb bacteriophage λ (Sanger *et al.,* 1982)), it was clear that the equivalent strategy could be applied to sequencing whole genomes. The limiting factor would be the availability of a complete clone map and sufficient resources to complete the sequencing. For example, results from the *C. elegans* pilot sequencing project were published (Sulston, *et al*., 1992) some 6 years after the physical map was reported to be 60% complete (Coulson *et al*., 1986). The publication reported the assembly of 120 kb of completed sequence by separate shotgun sequencing of 3 selected cosmid clones (two of which overlapped) from the physical map. This report, therefore, confirmed the feasibility of this approach and the value of the map in characterising large genomes.

At this point in time, clone coverage of the human genome was limited to small YAC or bacterial clone contigs which had been constructed during positional cloning projects. The use of YACs in attempts to rapidly construct chromosome specific (Chumakov *et al*., 1992, Foote *et al*., 1992) and genome wide clone maps (Chumakov *et al*., 1995, Hudson *et al*., 1995) demonstrated that larger insert clones could be used to generate a map of the whole genome. Indeed, the adoption of YACs, which have an average insert size of 1 Mb (Chumakov *et al*., 1995) theoretically reduced the complexity of mapping the human genome (1x coverage of 3 GB requiring approximately 3,000 clones) This approximated the complexity of the earlier work where the nematode genome was mapped using 40 kb cosmids (1x coverage of 100 Mb requiring approximately 2,500 clones). However, the instability and chimerism of YACs, in addition to the difficulties of isolating the cloned insert compared to bacterial cloning systems, meant that YACs could not practically provide a sufficiently reliable or convenient resource for the generation of genomic sequence. The development of large insert PACs (Iaonnou *et al.*, 1994) and BACs (Shizuya *et al*., 1992) provided clones with an insert large enough to permit the feasible generation of genomic coverage (1x coverage of the human genome requiring ~25,000 clones) (as described in chapter 4) and a cloning system that was amenable to large-scale sequencing (as described in chapter 5). The development of marker-based maps using genetic, RH or YAC mapping techniques (genetic: Weissenbach *et al*., 1992, Dib *et al*., 1996; RH: Schuler *et al*., 1996; Deloukas *et al*., 1998; YAC: Hudson *et al*., 1995; genome; Foote *et al*., 1992; Collins *et al*., 1995) provided the necessary long-range and independent map information to act as a framework on which bacterial clone coverage could be generated (Olson *et al*., 1993, Bentley *et al*., 2001). In addition, the evolution of fingerprinting techniques (Coulson *et al*., 1986, Olson *et al*., 1986), as described in chapter 3 (Gregory *et al*., 1997, Marra *et*

al., 1997), provided the tools by which sequence ready bacterial clone maps of large genomes could be constructed.

The chromosome specific, clone by clone approach that the Human Genome Project utilised to generate sequence of the human genome provided a means by which the work of many separate groups could be coordinated and a method by which highly accurate genomic sequence (>99.99%) could be produced. The principles demonstrated in the nematode project (Coulson *et al*., 1986) proved effective, and essential for the successful production of the human genome sequence. Problem solving (in particular resolution of the sequence of repeats, or technically difficult regions) was facilitated by the modular clone-based nature of the project; while long-range map information added weight to the clone order.

Celera, a private company that published a draft version of the human genome (Venter *et al*., 2001) in parallel to the publicly funded project (IHGSC 2001), purported to have assembled the human genome exclusively by a whole genome shotgun (WGS) strategy. However, they too ultimately relied upon a clone based map by incorporating a "perfect tiling path" of the HGP data into their sequence assembly. Specifically, 10 million 'faux' reads in a regular ordered pattern were generated by dividing up the public assembly into overlapping segments of 550 bp each, thus capturing the overall synthesis of all the information used to assemble the public domain sequence. The public sequence data used by Celera was thus derived from an accurately assembled clone based physical map and effectively representing complete genomic coverage (Waterston *et al*., 2002). Projects to sequence large genomes using a WGS strategy have also proved impossible to finish with high accuracy. This was illustrated by the reliance

upon clone based finishing of the WGS assembly for the *Drosophila* genome (Celniker *et al*., 2002). Similarly, both the human and the mouse genome sequences are being finished using a large-insert bacterial clone based approach.

The next phase in the evolution of physical map construction was driven by the availability of ordered genomic sequence. Conservation of sequence and long range order, between organisms that are sufficiently closely related, means that the genome of one species can act as the template upon which a physical map of another can built and, in doing so, elucidate the homologous relationship between them (Thomas *et al*., 2002). The success of the comparative physical mapping approach was demonstrated by construction of a clone map of the mouse genome using the assembled human genome sequence as a template. In this study, human genomic sequence was used to align stringently assembled BAC fingerprint contigs by matching mouse BAC end sequences (BESs) to their corresponding locations in the human genome. Ordered and orientated contigs (previously assembled by fingerprinting) were subsequently joined following further fingerprint analysis and addition of available genetic and radiation hybrid markers. The availability of BESs from a highly redundant fingerprint assembly of BAC clones, and using the strategy outlined above, greatly simplified the process of contig assembly, as the majority of the 7,500 contigs generated in the first fingerprinting phase were juxtaposed correctly relative to each other on the basis of homology between the two genomes. As a result, 7,500 x 7,500 possible joins (more than 56 million) was reduced to analysis of <10,000 putative joins. This permitted the construction of a physical map covering 98% of the 2500 Mb mouse genome, contained within 296 contigs, in approximately 12 months (see accompanying paper). The same approach could be adopted for any genome where there is sufficient sequence homology to allow

alignment of BESs (or equivalent sequence tags), plus sufficient homology between the template genome and the genome under study. The approach has important applications both for genomes where the full genome sequence is anticipated, and also (perhaps even more importantly) is a cost-effective way to provide access to regions of a genome for which there are no plans to generate genomic sequence on any scale.

Whilst a clone by clone approach proved successful for the generation of human sequence, the possible contribution of WGS data to large projects has continued to be evaluated. The main advantages of WGS are that the production of data is very rapid, can be highly automated, avoids cloning biases of BAC systems, and is very cost-effective. The assembly inherent from the sequence alignment also provides important mapping information which is unbiased by additional experimental mapping systems or procedures. While it remains true that WGS in isolation has disadvantages which prevent completion of either the map or finished sequence of a large genome, the possibility of combining the advantages of both approaches has been explored. A hybrid strategy emerged from the *Drosophila* project, and has since been adopted for the mouse genome. Seven fold WGS coverage was generated from sub-cloned plasmids of varying sizes which, when assembled with BESs, generated 96% coverage of the euchromatic portion of the mouse genome. This estimate was derived by assessing the amount of WGS coverage provided which matched 187 Mb of finished mouse sequence. For a second, independent estimate, a genomic alignment of a curated collection of cDNAs to the WGS assembly was also used. This alignment included 96.4% of cDNA bases. Paired-end reads from large insert plasmids and BACs provided the scaffold upon which the assembled whole genome shotgun sequence was ordered and orientated, and simultaneously integrated BAC clones into the sequence. A tiling

path of BAC clones from the physical map is currently being used for directed finishing of the draft genomic sequence. The physical map helped to assemble the sequence scaffold, whilst the WGS data increased the rate of clone based finishing (MGSC 2002).

If WGS sequence data can accurately place BACs via their BESs within the sequence assembly, is a restriction fingerprint database actually required? The answer is probably yes. Whilst BES localisation within a whole genome shotgun assembly facilitates a more optimal minimum tiling path selection, overlaps within fingerprinting contigs can link sequence assemblies (as reported by the assembly of the mouse WGS sequence (MGSC 2002)). Three hundred and seventy-seven anchored 'supercontigs' were generated by assembling plasmid and BAC end sequences in the WGS assembly. This number was reduced to 88 when two or more sequence supercontigs were localised within a single restriction fingerprint contig. The overlaps generated by fingerprint analysis may also be able to resolve errors in the genomic assembly where, for example, low copy repeats may have resulted in a compression of the sequence assembly. The proven success of assembling genome wide physical maps, the cost of constructing a >15 fold genomic BAC library and the ease with which genome-wide fingerprint databases can be assembled, has lead to the construction of several genomic fingerprint databases, table 7.1. Whilst genome-wide fingerprint maps will facilitate the large-scale characterisation of many varied species, the construction of small region specific sequence-ready maps will continue to be important for detailed inter-species sequence comparisons (Thomas *et al*., 2002).

Table 7.1: Organisms for which genome-wide fingerprint databases have or are being constructed.

Organism	Reference
A. thaliana	Marra et al 1999
Rice	Tao et al. 2000
H. sapiens	McPherson et al 2001
M. musculus	Gregory et al 2002
R. rattus	http://www.bcgsc.ca/lab/mapping
C. neoformans	http://www.bcgsc.ca/lab/mapping
Bovine	http://www.bcgsc.ca/lab/mapping
Porcine	http://www.nps.ars.usda.gov/
D. rario	http://www.sanger.ac.uk/Projects/D_rerio/
Soybean	http://hbz.tamu.edu/soybean.html

7.2 The identification of coding features

The availability of human genomic sequence provides the framework upon which the structure of genes and their regulatory elements can be placed. The existence of genes as protein-coding genes, pseudogenes (Harrison *et al*., 2002), non-protein-coding RNA transcripts (Mattick *et al*., 2001) and genes arising from genomic duplications (Bailey *et al*., 2002) requires a multifaceted approach to gene discovery. The three main methods used for large-scale genome analysis, *in silico* prediction of gene structures, sequence alignment of expressed and genomic sequences and identification of conserved sequences between different species, are described in chapter 5. These approaches are preferentially used in combination to assist correction for the shortcomings of each method. For example, *in silico* gene identification can lead to over prediction and false negatives (Guigo *et al*., 2000). EST and cDNA sequence alignment can be confounded by artefactual or unprocessed clones in cDNA libraries. Cross species sequence comparison, for example between human and mouse, can also identify

homologies outside genes and regulatory elements (Deloukas *et al*., 2001, Kondrashov and Shabalina *et al*., 2002).

An important facet of fully identifying all coding features within any organism is the availability of high quality finished genomic sequence. Annotation of genes using draft sequence, which may contain regions of low quality, incomplete or unordered data, may lead to inaccurate annotation of genes and possibly errors in inference of the protein products derived from them. Another important consideration is reanalysis of existing annotation. The continual emergence of new sequence data from independent studies, and reanalysis of genomic sequence on a regular basis (or on demand), is required to ensure incorporation of all available evidence for genes and other features. The genomic alignment of larger sets of non-redundant ESTs and cDNAs, derived from a wider range of tissues or of sequence from related organisms, may assist to fully define gene structures, identify novel coding features or regulatory regions. Recent publications detailing the re-annotation of human chromosome 22 (Collins *et al*., 2003) and *Drosophila* genomes (Misra *et al*., 2002), which list structural changes to genes previously identified, indicate that the characterisation of all coding features will require several iterations of automatic analysis, manual annotation and directed laboratory work. Other examples of comparative analysis have also revealed new regulatory elements or genes (Pennacchio *et al*., 2001, Gottgens *et al*., 2002). The identification of genic features such as the 5' and 3' ends of genes, splice variants and even the functional determination of non-coding genes such as anti-sense RNAs (Green *et al*., 1986), will take many years of painstaking study.

7.3 Assigning gene function

Characterisation of the functional product of a gene is not achieved directly by the identification of translated amino acid sequence contained within the open reading frame of the coding sequence. Within a genomic context, the final sequence and structure of an mRNA and the encoded protein may be influenced by priming from multiple promoters, splice variation within the coding exons or the existence of alternative polyadenylation sites, as discussed in chapter 5. Post-translational processing can also result in modification of a protein product. Whilst *in vivo* and *in vitro* studies within model organisms, by chemical mutagenesis and gene targeting can identify gene function by generating an observable phenotype it may not always be clear how a disruption of the target gene has given rise to a particular effect within a complex network of gene interactions. Alternatively, protein function may be predicted by *in silico* structural analysis. The assignment of new function to a novel protein at a nucleotide sequence level comparison, however, may fail as BLAST analysis within the Protein Data Bank (PDB) was shown to only find 10% of the known relationships (Brenner *et al*., 1998). Whilst iterative PSI-BLAST (Altschul *et al*., 1997) is more sensitive, relationships are still missed.

An alternative approach is the sequence-to-function method which uses pair-wise sequence or motif alignment to derive significant homologies between proteins and hence suggest similarity of function. Whilst these methods are powerful, they are not ideally suited to identify loss or gain of function during protein evolution and encounter difficulties when assigning function as protein databases become more diverse (Skolnick and Fetrow 2000). Alternatively, the possible function of a protein may be

suggested by comparison of three-dimensional structure to proteins of known function. Since the tertiary structure of proteins of common function is likely to be more conserved than their primary structures (amino acid sequences), attempts have been made to classify groups of proteins based on structural and phylogenetic relationships, e.g. SCOP (Murzin *et al*., 1995). A second approach describes proteins according to their structural characteristics, such as class of architecture and fold type. In practice both approaches are used, initially grouping proteins according to their sequence homology and then by their structural descriptors (Thornton *et al*., 1999).

The application of the sequence to structure to function approach aims to determine the structure of a protein and then to identify the functionally important residues, as described in chapter 6. *Ab initio* folding can be used to predict a native structure based on domains contained within the protein. A process known as threading utilises a known structure as a template upon which proteins of up to 500 residues can be moulded. These three dimensional structures can then be used to infer function by analysis of internal or external residues, the shape and molecular composition of the protein or the juxtaposition of individual groups. The prediction of protein folds, their 3D structure and function is, however, primarily reliant upon experimental evidence, either as a basis for modelling or as support for a prediction. X-ray crystallography and nuclear magnetic resonance spectroscopy are methods by which these proteins structures have been experimentally determined.

The identification of well conserved, functionally important residues within primary and tertiary structures of some protein families can assist to predict the function of novel proteins. Based upon the conservation of an Asp-Thr(Ser)-Gly amino acid sequence at

their active sites Pearl and Taylor (1987a, 1987b) concluded that retroviral proteases could belong to aspartic protease family. Modelling and subsequent structural comparisons between aspartic proteases, which contain 300 residues and an active site in each of its two domains, and retroviral proteases, that do not exceed 130 residues and have only one active site, led Pearl and Taylor to hypothesise that the retroviral enzymes may be dimeric aspartic proteases. This prediction was subsequently proven by the expression of the retroviral proteases in *E. coli* (Meek *et al*., 1989) and by the determination of it crystal structure (Navia *et al*., 1989)

In many cases, however, even in these well characterised families, the catalytic component may be recognisable, but the specific substrate binding properties may be difficult to determine. Additional protein domains (encoded by separate exons) which are required for function, but localise to other regions of the protein, are less readily elucidated by homology alone. For this, direct experimental approaches are required to determine the substrate and products in the appropriate biochemical pathway. For example protein binding assays, using yeast two hybrid systems, can identify interacting binding proteins. Knockouts, or natural mutants, may be investigated to determine the biochemistry of the altered phenotype in some detail. For example, a defective enzyme may result in accumulation of abnormally high levels of substrate, and comparison of normal vs. mutant systems will reveal candidates as possible substrates.

Additional information can be gained by determining the cellular and tissue localisation of the protein. The co-localisation of proteins in a highly tissue-specific pattern may provide evidence for some level of protein interaction. The fusion of the sequence encoding a novel protein to the sequence of a reporter molecule in a shuttle vector can

be used to determine cellular localisation if the construct can be introduced into a physiologically relevant cell line. This work may be followed up, for example. by manipulation of the construct and introduction into embryonic stem cells in order to create a transgenic animal model where the gene is under control of the endogenous promoter. This would enable investigation of the expression of the gene presumably in response to physiologically natural intracellular and extracellular signals. The cellular distribution of the signal molecule should, therefore, reflect the distribution of the natural gene product. Data from co-localisation experiments may be correlated with protein-protein interaction studies, and possibly analysis (e.g. by mass spectrometry) of the components of co-purified complexes, to build a picture of the interactions between specific proteins

7.4 Sequence variation

The key to understanding the effects of sequence variation within protein-coding genes is the identification of genuine sequence variants from accurately annotated gene structures and determination of the functional effect that a variant has upon the encoded protein, or on expression of the gene. Chapter 6 describes the generation of exon specific sequences from a collection of twelve medically important genes, including five closely related members of a gene family. The design of primers which are uniquely localised within the genome, in conjunction with detailed analysis of the sequence flanking the SNP, permitted identification of novel locus specific polymorphisms within highly homologous genes and the unique placement of SNPs which previously had been given multiple localisations within Ensembl and dbSNP. The functional effect a minor

coding SNP may have upon protein structure was predicted using protein modelling, as outlined in chapter 5. Whilst *in silico* prediction would always require experimental support (see also discussion above), it can be used to suggest the effects a SNP may have upon a protein structure and function. A novel non-synonymous, non-conservative coding SNP was identified within GSTT1, as part of this study.

The majority of genetic variation which contributes to cancer are somatic mutations caused by exposure to environmental carcinogens rather than inherited variants in susceptibility genes. However, genes that encode enzymes involved in the metabolism of carcinogens can be polymorphic and these polymorphisms may be related to elevated risk of cancer susceptibility. This mechanism illustrates the importance of genetic background and the effect on interaction between the individual and the environment. GSTT1 is a phase II enzyme that mediates the detoxification of xenobiotic chemicals by conjugation of glutathione which changes the polarity of the chemical and makes them more readily excreted. GSTT1 is a five exon gene which localises to 22q11.2 and is known to be absent from 38% of the population (Pemble *et al*., 1994). The null genotype of GSTT1 has been implicated with increased risk of myelodysplastic syndromes (MDS) (Chen *et al*., 1996), aplastic anemia (Lee *et al*., 2001) and, in conjunction with a CYP1A1 mutation, has effects on live birth weight (Wang *et al*., 2002). These studies indicate that the gene, by its absence, may influence the aetiology of disease. The identification of a novel non-synonymous non-conservative SNP, which may induce a change to conformation and function of the protein, as described in chapter 6, may also influence an individual's susceptibility to cancer.

7.5 Conclusions and future work

This thesis describes the application of a novel restriction fingerprinting technique to the generation of a sequence ready map of 1pcen – 1p13, the elucidation of the genic features within the interval and the characterisation of sequence variation within a selected number of these genes. In the short term, all these areas of research could be extended. The annotation of genes within the transcript map of the interval will require further work, as shown by the gain in information by the iterative annotation of *Drosophila* (Misra *et al*., 2002) and human chromosome 22 (Collins *et al*., 2003). Alignment of novel cDNAs and ESTs to genomic sequence may assist to further define gene structures, specifically the 5' and 3' ends, as well as identifying new genes and splice variants. Comparative sequence analyses, using species from a variety of evolutionary distances, will also help to better characterise the genes in the interval and also to identify regions that control their regulation.

Comparative sequence analysis could also be used to characterise the structural evolution of chromosome 1. Regions either side of the chromosome 1 centromere are contained within a contiguous homologous region on mouse chromosome 3. The investigation of gene order, number and orientation may assist to elucidate the number of chromosomal rearrangements that have taken place during the evolution of human chromosome 1 from the organisation of the ancestral karyotype.

The functional consequence of the novel non-synonymous, non-conservative SNP within exon 3 requires further investigation. The frequency of the non-synonymous, non-conservative GSTT1*B allele within other ethic groups (besides the Northern

European population used here) could be investigated. In addition, other phenotype to genotype correlations of the Thr104 – Pro SNP within large prospective DNA collections, such as the Avon Longitudinal Study of Parents and Children, could be determined.

In the longer term, it is anticipated that substantial efforts will result in all of the genes in the human genome being investigated in detail. These studies will result in a fuller understanding of the specificity and range of biochemical structures and functions that are encoded in the human genome sequence. In general there is likely to remain a distinction between the study of functions encoded at the DNA level, which affect gene expression via transcriptional control, and the study of functions reflected at the protein level following translation, taking into account post-translational modifications (processes which are largely genetically determined). Without a genic catalogue, functional studies are necessarily limited to the investigation of a specific target – a gene, a protein, or a disease. These approaches are an essential part of fully interpreting the genome as they provide a means by which hypotheses can be experimentally tested and which produce valid and supplementary results. However, the production of a complete gene catalogue (if completion can indeed be measured or achieved) will provide the raw material for modelling whole systems. The extensive use of computational biology to suggest how such complex systems are made up of their interacting components will, in itself, enable predictions to be made of the system model. These predictions can be tested, both to determine the validity of the modelled system, and also to test the success of the methods used to derive the system.

A more complete knowledge of biochemical processes will yield a better understanding of complex disease and how it should be treated. At present, our knowledge is primarily

based on monogenic diseases. As the problem is reduced to a single gene, hypotheses for function can be tested by biochemical assays, protein structural studies, experimental knock-outs, or the study of naturally occurring mutants. The approach to complex disease centres on a similar approach, i.e. trying to identify the one or few dominant genetic factors which contribute the most significant effect to the overall phenotype. However, there is a realisation that these genetic factors may not fully explain the observed phenotype and that a proportion of the remaining factors may not be identified. In these instances, a comprehensive knowledge of the systems involved will be more informative than the approach of complex disease genetics, both in how the phenotype arises, and how it might be possible to intervene more effectively. This is potentially a true long-term value of the genome sequence, and its interpretation in a biochemical, biological and genetic context, for the advancement of medicine in the future.

Chapter 8

References

Abagyan, R. A., Totrov, M.M., and Kuznetsov, D.N. ICM - a new method for protein modeling and design. Applications to docking and structure prediction from the distorted native conformation. (1994) *J Comp Chem,* **15,** 488 - 506.

Adachi, N. and Lieber, M. R. Bidirectional gene organization: a common architectural feature of the human genome. (2002) *Cell,* **109,** 807-9.

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., et al. The genome sequence of Drosophila melanogaster. (2000) *Science,* **287,** 2185-95.

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., et al. Complementary DNA sequencing: expressed sequence tags and human genome project. (1991) *Science,* **252,** 1651-6.

Adams, M. D., Soares, M. B., Kerlavage, A. R., Fields, C. and Venter, J. C. Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. (1993) *Nat Genet,* **4,** 373-80.

Alexandrie, A. K., Rannug, A., Juronen, E., Tasa, G. and Warholm, M. Detection and characterization of a novel functional polymorphism in the GSTT1 gene. (2002) *Pharmacogenetics,* **12,** 613-9.

Allan, J. M., Wild, C. P., Rollinson, S., Willett, E. V., Moorman, A. V., Dovey, G. J., et al. Polymorphism in glutathione S-transferase P1 is associated with susceptibility to chemotherapy-induced leukemia. (2001) *Proc Natl Acad Sci U S A,* **98,** 11592-7.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. Basic local alignment search tool. (1990) *J Mol Biol,* **215,** 403-10.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. (1997) *Nucleic Acids Res,* **25,** 3389-402.

Altshuler, D., Pollara, V. J., Cowles, C. R., Van Etten, W. J., Baldwin, J., Linton, L., et al. An SNP map of the human genome generated by reduced representation shotgun sequencing. (2000) *Nature,* **407,** 513-6.

Anderson, S. Shotgun DNA sequencing using cloned DNase I-generated fragments. (1981) *Nucleic Acids Res,* **9,** 3015-27.

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J., et al. Sequence and organization of the human mitochondrial genome. (1981) *Nature,* **290,** 457-65.

Angelisova, P., Vlcek, C., Stefanova, I., Lipoldova, M. and Horejsi, V. The human leucocyte surface antigen CD53 is a protein structurally similar to the CD37 and MRC OX-44 antigens. (1990) *Immunogenetics,* **32,** 281-5.

Ansari-Lari, M. A., Oeltjen, J. C., Schwartz, S., Zhang, Z., Muzny, D. M., Lu, J., et al. Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6. (1998) *Genome Res,* **8,** 29-40.

Antequera, F. and Bird, A. Number of CpG islands and genes in human and mouse. (1993) *Proc Natl Acad Sci U S A,* **90,** 11995-9.

Arabidopsis Genome Initiative, T. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. (2000) *Nature,* **408,** 796-815.

Attali, B., Romey, G., Honore, E., Schmid-Alliana, A., Mattei, M. G., Lesage, F., et al. Cloning, functional expression, and regulation of two K+ channels in human T lymphocytes. (1992) *J Biol Chem,* **267,** 8650-7.

Bailey, J. A., Yavor, A. M., Viggiano, L., Misceo, D., Horvath, J. E., Archidiacono, N., et al. Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. (2002) *Am J Hum Genet,* **70,** 83-100.

Bankier, A. T., Weston, K. M. and Barrell, B. G. Random cloning and sequencing by the M13/dideoxynucleotide chain termination method. (1987) *Methods Enzymol,* **155,** 51-93.

Bates, G. P., Valdes, J., Hummerich, H., Baxendale, S., Le Paslier, D. L., Monaco, A. P., et al. Characterization of a yeast artificial chromosome contig spanning the Huntington's disease gene candidate region. (1992) *Nat Genet,* **1,** 180-7.

Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J. M. and Gautheret, D. Patterns of variant polyadenylation signal usage in human genes. (2000) *Genome Res,* **10,** 1001-10.

Bentley, D. R. The Human Genome Project--an overview. (2000) *Med Res Rev,* **20,** 189-96.

Bentley, D. R., Deloukas, P., Dunham, A., French, L., Gregory, S. G., Humphray, S. J., et al. The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. (2001) *Nature,* **409,** 942-3.

Bernaola-Galvan, P., Roman-Roldan, R. and Oliver, J. L. Compositional segmentation and long-range fractal correlations in DNA sequences. (1996) *Physical Review. E. Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics,* **53,** 5181-5189.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., et al. The mosaic genome of warm-blooded vertebrates. (1985) *Science,* **228,** 953-8.

Bickmore, W. A. and Sumner, A. T. Mammalian chromosome banding--an expression of genome organization. (1989) *Trends Genet,* **5,** 144-8.

Bird, A. P. CpG-rich islands and the function of DNA methylation. (1986) *Nature,* **321,** 209-13.

Birnboim, H. C. and Doly, J. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. (1979) *Nucleic Acids Res,* **7,** 1513-23.

Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., et al. The complete genome sequence of Escherichia coli K-12. (1997) *Science,* **277,** 1453-74.

Board, P., Coggan, M., Johnston, P., Ross, V., Suzuki, T. and Webb, G. Genetic heterogeneity of the human glutathione transferases: a complex of gene families. (1990) *Pharmacol Ther,* **48,** 357-69.

Boguski, M. S. and Schuler, G. D. ESTablishing a human transcript map. (1995) *Nat Genet,* **10,** 369-71.

Bonfield, J. K., Smith, K. and Staden, R. A new DNA sequence assembly program. (1995) *Nucleic Acids Res,* **23,** 4992-9.

Booth, J., Boyland, E., and Sims, P. An enzyme from rat liver catalyzing conjugation with glutathione. (1961) *Biochem J,* **79,** 516 - 524.

Botstein, D., White, R. L., Skolnick, M. and Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. (1980) *Am J Hum Genet,* **32,** 314-31.

Bouffard, G. G., Idol, J. R., Braden, V. V., Iyer, L. M., Cunningham, A. F., Weintraub, L. A., et al. A physical map of human chromosome 7: an integrated YAC contig map with average STS spacing of 79 kb. (1997) *Genome Res,* **7,** 673-92.

Brenner, M., Lampel, K., Nakatani, Y., Mill, J., Banner, C., Mearow, K., et al. Characterization of human cDNA and genomic clones for glial fibrillary acidic protein. (1990) *Brain Res Mol Brain Res,* **7,** 277-86.

Brenner, S. E., Chothia, C. and Hubbard, T. J. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. (1998) *Proc Natl Acad Sci U S A,* **95,** 6073-8.

Brintnell, B., Hey, Y., Jones, D., Hoggard, N., James, L. and Varley, J. M. Generation of a contig comprising YACs and BACs within chromosome region 1p13.1. (1997) *Somat Cell Mol Genet,* **23,** 153-7.

Brown, L., Espinosa, R., 3rd, Le Beau, M. M., Siciliano, M. J. and Baer, R. HEN1 and HEN2: a subgroup of basic helix-loop-helix genes that are coexpressed in a human neuroblastoma. (1992) *Proc Natl Acad Sci U S A,* **89,** 8492-6.

Burge, C. and Karlin, S. Prediction of complete gene structures in human genomic DNA. (1997) *J Mol Biol,* **268,** 78-94.

Burke, D. T., Carle, G. F. and Olson, M. V. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. (1987) *Science,* **236,** 806-12.

C. elegans Sequencing Consortium, T. Genome sequence of the nematode C. elegans: a platform for investigating biology. The C. elegans Sequencing Consortium. (1998) *Science,* **282,** 2012-8.

Carpten, J., Nupponen, N., Isaacs, S., Sood, R., Robbins, C., Xu, J., et al. Germline mutations in the ribonuclease L gene in families showing linkage with HPC1. (2002) *Nat Genet,* **30,** 181-4.

Carrano, A. V., Lamerdin, J., Ashworth, L. K., Watkins, B., Branscomb, E., Slezak, T., et al. A high-resolution, fluorescence-based, semiautomated method for DNA fingerprinting. (1989) *Genomics,* **4,** 129-36.

Carrier, A., Rosier, M. F., Guillemot, F., Goguel, A. F., Pulcini, F., Bernheim, A., et al. Integrated physical, genetic, and genic map covering 3 Mb around the human NGF gene (NGFB) at 1p13. (1996) *Genomics,* **31,** 80-9.

Charroux, B., Pellizzoni, L., Perkinson, R. A., Shevchenko, A., Mann, M. and Dreyfuss, G. Gemin3: A novel DEAD box protein that interacts with SMN, the spinal muscular atrophy gene product, and is a component of gems. (1999) *J Cell Biol,* **147,** 1181-94.

Chasman, D. and Adams, R. M. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. (2001) *J Mol Biol,* **307,** 683-706.

Chelly, J., Tumer, Z., Tonnesen, T., Petterson, A., Ishikawa-Brush, Y., Tommerup, N., et al. Isolation of a candidate gene for Menkes disease that encodes a potential heavy metal binding protein. (1993) *Nat Genet,* **3,** 14-9.

Chen, C. L., Liu, Q. and Relling, M. V. Simultaneous characterization of glutathione S-transferase M1 and T1 polymorphisms by polymerase chain reaction in American whites and blacks. (1996) *Pharmacogenetics,* **6,** 187-91.

Choo, K. H., Gould, K. G., Rees, D. J. and Brownlee, G. G. Molecular cloning of the gene for human anti-haemophilic factor IX. (1982) *Nature,* **299,** 178-80.

Chumakov, I. M., Le Gall, I., Billault, A., Ougen, P., Soularue, P., Guillou, S., et al. Isolation of chromosome 21-specific yeast artificial chromosomes from a total human genome library. (1992) *Nat Genet,* **1,** 222-5.

Chumakov, I. M., Rigault, P., Le Gall, I., Bellanne-Chantelot, C., Billault, A., Guillou, S., et al. A YAC contig map of the human genome. (1995) *Nature,* **377,** 175-297.

Claudio, J. O., Liew, C. C., Ma, J., Heng, H. H., Stewart, A. K. and Hawley, R. G. Cloning and expression analysis of a novel WD repeat gene, WDR3, mapping to 1p12-p13. (1999) *Genomics,* **59,** 85-9.

Coffey, A. J., Roberts, R. G., Green, E. D., Cole, C. G., Butler, R., Anand, R., et al. Construction of a 2.6-Mb contig in yeast artificial chromosomes spanning the human dystrophin gene using an STS-based approach. (1992) *Genomics,* **12,** 474-84.

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

Collins, F. S., Brooks, L. D. and Chakravarti, A. A DNA polymorphism discovery resource for research on human genetic variation. (1998) *Genome Res,* **8,** 1229-31.

Collins, J. E., Cole, C. G., Smink, L. J., Garrett, C. L., Leversha, M. A., Soderlund, C. A., et al. A high-density YAC contig map of human chromosome 22. (1995) *Nature,* **377,** 367-79.

Collins, J. E., Goward, M. E., Cole, C. G., Smink, L. J., Huckle, E. J., Knowles, S., et al. Reevaluating human gene annotation: a second-generation analysis of chromosome 22. (2003) *Genome Res,* **13,** 27-36.

Connelly, M. A., Zhang, H., Kieleczawa, J. and Anderson, C. W. The promoters for human DNA-PKcs (PRKDC) and MCM4: divergently transcribed genes located at chromosome 8 band q11. (1998) *Genomics,* **47,** 71-83.

Coulondre, C., Miller, J. H., Farabaugh, P. J. and Gilbert, W. Molecular basis of base substitution hotspots in Escherichia coli. (1978) *Nature,* **274,** 775-80.

Coulson, A., Huynh, C., Kozono, Y. and Shownkeen, R. The physical map of the Caenorhabditis elegans genome. (1995) *Methods Cell Biol,* **48,** 533-50.

Coulson, A., Sulston, J., Brenner, S., Karn, J. Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. (1986) *Proc Natl Acad Sci U S A,* **83,** 7821-7825.

Couzin, J. Human genome. HapMap launched with pledges of $100 million. (2002) *Science,* **298,** 941-2.

Cox, D. R., Burmeister, M., Price, E. R., Kim, S. and Myers, R. M. Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. (1990) *Science,* **250,** 245-50.

Dauwerse, J. G., Kievits, T., Beverstock, G. C., van der Keur, D., Smit, E., Wessels, H. W., et al. Rapid detection of chromosome 16 inversion in acute nonlymphocytic leukemia, subtype M4: regional localization of the breakpoint in 16p. (1990) *Cytogenet Cell Genet,* **53,** 126-8.

Dawson, E., Chen, Y., Hunt, S., Smink, L. J., Hunt, A., Rice, K., et al. A SNP resource for human chromosome 22: extracting dense clusters of SNPs from the genomic sequence. (2001) *Genome Res,* **11,** 170-8.

Deininger, P. L. a. D., G.R. The recent evolution of DNA repetitive elements. (1986) *Trends Genet,* **2,** 76-80.

Dekker, J. W., Budhia, S., Angel, N. Z., Cooper, B. J., Clark, G. J., Hart, D. N., et al. Identification of an S-adenosylhomocysteine hydrolase-like transcript induced during dendritic cell differentiation. (2002) *Immunogenetics,* **53,** 993-1001.

Dell'Angelica, E. C., Mullins, C. and Bonifacino, J. S. AP-4, a novel protein complex related to clathrin adaptors. (1999) *J Biol Chem,* **274,** 7278-85.

Deloukas, P., Matthews, L. H., Ashurst, J., Burton, J., Gilbert, J. G., Jones, M., et al. The DNA sequence and comparative analysis of human chromosome 20. (2001) *Nature,* **414,** 865-71.

Deloukas, P., Schuler, G. D., Gyapay, G., Beasley, E. M., Soderlund, C., Rodriguez-Tome, P., et al. A physical map of 30,000 human genes. (1998) *Science,* **282,** 744-6.

DeMartino, G. N., Orth, K., McCullough, M. L., Lee, L. W., Munn, T. Z., Moomaw, C. R., et al. The primary structures of four subunits of the human, high-molecular-weight proteinase, macropain (proteasome), are distinct but homologous. (1991) *Biochim Biophys Acta,* **1079,** 29-38.

Deutsch, S., Iseli, C., Bucher, P., Antonarakis, S. E. and Scott, H. S. A cSNP map and database for human chromosome 21. (2001) *Genome Res,* **11,** 300-7.

Dib, C., Faure, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., et al. A comprehensive genetic map of the human genome based on 5,264 microsatellites. (1996) *Nature,* **380,** 152-4.

Ding, Y., Johnson, M. D., Chen, W. Q., Wong, D., Chen, Y. J., Benson, S. C., et al. Five-color-based high-information-content fingerprinting of bacterial artificial chromosome clones using type IIS restriction endonucleases. (2001) *Genomics,* **74,** 142-54.

Doggett, N. A., Goodwin, L. A., Tesmer, J. G., Meincke, L. J., Bruce, D. C., Clark, L. M., et al. An integrated physical map of human chromosome 16. (1995) *Nature,* **377,** 335-65.

Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K., et al. A genetic linkage map of the human genome. (1987) *Cell,* **51,** 319-37.

Down, T. A. and Hubbard, T. J. Computational detection and location of transcription start sites in mammalian genomic DNA. (2002) *Genome Res,* **12,** 458-61.

Dracopoli, N. C., Rettig, W. J., Whitfield, G. K., Darlington, G. J., Spengler, B. A., Biedler, J. L., et al. Assignment of the gene for the beta subunit of thyroid-stimulating hormone to the short arm of human chromosome 1. (1986) *Proc Natl Acad Sci U S A,* **83,** 1822-6.

Dumont, M., Luu-The, V., Dupont, E., Pelletier, G. and Labrie, F. Characterization, expression, and immunohistochemical localization of 3 beta-hydroxysteroid dehydrogenase/delta 5-delta 4 isomerase in human skin. (1992) *J Invest Dermatol,* **99,** 415-21.

Dunham, I., Shimizu, N., Roe, B. A., Chissoe, S., Hunt, A. R., Collins, J. E., et al. The DNA sequence of human chromosome 22. (1999) *Nature,* **402,** 489-95.

El-Rifai, W., Sarlomo-Rikala, M., Andersson, L. C., Knuutila, S. and Miettinen, M. DNA sequence copy number changes in gastrointestinal stromal tumors: tumor progression and prognostic significance. (2000) *Cancer Res,* **60,** 3899-903.

Evans, G. A. and Lewis, K. A. Physical mapping of complex genomes by cosmid multiplex analysis. (1989) *Proc Natl Acad Sci U S A,* **86,** 5030-4.

Ewing, B. and Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. (2000) *Nat Genet,* **25,** 232-4.

Ewing, B., Hillier, L., Wendl, M. C. and Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. (1998) *Genome Res,* **8,** 175-85.

Fields, C., Adams, M. D., White, O. and Venter, J. C. How many genes in the human genome? (1994) *Nat Genet,* **7,** 345-6.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. (1995) *Science,* **269,** 496-512.

Foote, S., Vollrath, D., Hilton, A. and Page, D. C. The human Y chromosome: overlapping DNA clones spanning the euchromatic region. (1992) *Science,* **258,** 60-6.

Forus, A., Weghuis, D. O., Smeets, D., Fodstad, O., Myklebost, O. and van Kessel, A. G. Comparative genomic hybridization analysis of human sarcomas: I. Occurrence of genomic imbalances and identification of a novel major amplicon at 1q21-q22 in soft tissue sarcomas. (1995) *Genes Chromosomes Cancer,* **14,** 8-14.

Francke, U. Digitized and differentially shaded human chromosome ideograms for genomic applications. (1994) *Cytogenet Cell Genet,* **65,** 206-18.

Galgoczy, P., Rosenthal, A. and Platzer, M. Human-mouse comparative sequence analysis of the NEMO gene reveals an alternative promoter within the neighboring G6PD gene. (2001) *Gene,* **271,** 93-8.

Gardiner, K. Base composition and gene distribution: critical patterns in mammalian genome organization. (1996) *Trends Genet,* **12,** 519-24.

Gardiner-Garden, M. and Frommer, M. CpG islands in vertebrate genomes. (1987) *J Mol Biol,* **196,** 261-82.

Gengyo-Ando, K. and Mitani, S. Characterization of mutations induced by ethyl methanesulfonate, UV, and trimethylpsoralen in the nematode Caenorhabditis elegans. (2000) *Biochem Biophys Res Commun,* **269,** 64-9.

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

Gilbert, W. Why genes in pieces? (1978) *Nature,* **271,** 501.

Gilbert, W. (1992) (Ed, Kelves, D. J. a. H., L.) Harvard University Press.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., et al. Life with 6000 genes. (1996) *Science,* **274,** 546, 563-7.

Goss, S. J. and Harris, H. New method for mapping genes in human chromosomes. (1975) *Nature,* **255,** 680-4.

Gottgens, B., Barton, L. M., Gilbert, J. G., Bench, A. J., Sanchez, M. J., Bahn, S., et al. Analysis of vertebrate SCL loci identifies conserved enhancers. (2000) *Nat Biotechnol,* **18,** 181-6.

Green, E. D. and Chakravarti, A. The human genome sequence expedition: views from the "base camp". (2001) *Genome Res,* **11,** 645-51.

Green, E. D. and Olson, M. V. Chromosomal region of the cystic fibrosis gene in yeast artificial chromosomes: a model for human genome mapping. (1990) *Science,* **250,** 94-8.

Green, E. D., Riethman, H. C., Dutchik, J. E. and Olson, M. V. Detection and characterization of chimeric yeast artificial-chromosome clones. (1991) *Genomics,* **11,** 658-69.

Green, P. Whole-genome disassembly. (2002) *Proc Natl Acad Sci U S A,* **99,** 4143-4.

Green, P. J., Pines O, Inouye M. The role of antisense RNA in gene regulation. (1986) *Annu Rev Biochem,* **55,** 569-97.

Gregory, S. G., Howell, G. R. and Bentley, D. R. Genome mapping by fluorescent fingerprinting. (1997) *Genome Res,* **7,** 1162-8.

Gregory, S. G., Sekhon, M., Schein, J., Zhao, S., Osoegawa, K., Scott, C. E., et al. A physical map of the mouse genome. (2002) *Nature,* **418,** 743-50.

Gregory, S. G., Vaudin, M., Wooster, R., Coleman, M., Mischke, D., Porter, C., et al. Report of the fourth international workshop on human chromosome 1 mapping 1998. (1998) *Cytogenet Cell Genet,* **83,** 147-75.

Guex, N. and Peitsch, M. C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. (1997) *Electrophoresis,* **18,** 2714-23.

Guigo, R., Agarwal, P., Abril, J. F., Burset, M. and Fickett, J. W. An assessment of gene prediction accuracy in large DNA sequences. (2000) *Genome Res,* **10,** 1631-42.

Guru, S. C., Olufemi, S. E., Manickam, P., Cummings, C., Gieser, L. M., Pike, B. L., et al. A 2.8-Mb clone contig of the multiple endocrine neoplasia type 1 (MEN1) region at 11q13. (1997) *Genomics,* **42,** 436-45.

Gyapay, G., Morissette, J., Vignal, A., Dib, C., Fizames, C., Millasseau, P., et al. The 1993-94 Genethon human genetic linkage map. (1994) *Nat Genet,* **7,** 246-339.

Hall, A. and Brown, R. Human N-ras: cDNA cloning and gene structure. (1985) *Nucleic Acids Res,* **13,** 5255-68.

Halushka, M. K., Fan, J. B., Bentley, K., Hsie, L., Shen, N., Weder, A., et al. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. (1999) *Nat Genet,* **22,** 239-47.

Hardison, R., Slightom, J. L., Gumucio, D. L., Goodman, M., Stojanovic, N. and Miller, W. Locus control regions of mammalian beta-globin gene clusters: combining phylogenetic analyses and experimental results to gain functional insights. (1997) *Gene,* **205,** 73-94.

Hardison, R., Xu, J., Jackson, J., Mansberger, J., Selifonova, O., Grotch, B., et al. Comparative analysis of the locus control region of the rabbit beta-like gene cluster: HS3 increases transient expression of an embryonic epsilon-globin gene. (1993) *Nucleic Acids Res,* **21,** 1265-72.

Harris, R. F. Hapmap flap. (2002) *Curr Biol,* **12,** R827.

Harrison, P. M., Hegyi, H., Balasubramanian, S., Luscombe, N. M., Bertone, P., Echols, N., et al. Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. (2002) *Genome Res,* **12,** 272-80.

Hattori, M., Fujiyama, A., Taylor, T. D., Watanabe, H., Yada, T., Park, H. S., et al. The DNA sequence of human chromosome 21. (2000) *Nature,* **405,** 311-9.

Hayes, J. D. and Pulford, D. J. The glutathione S-transferase supergene family: regulation of GST and the contribution of the isoenzymes to cancer chemoprotection and drug resistance. (1995) *Crit Rev Biochem Mol Biol,* **30,** 445-600.

Heding, I. J., Ivens, A. C., Wilson, J., Strivens, M., Gregory, S., Hoovers, J. M., et al. The generation of ordered sets of cosmid DNA clones from human chromosome region 11p. (1992) *Genomics,* **13,** 89-94.

Heiskanen, M., Karhu, R., Hellsten, E., Peltonen, L., Kallioniemi, O. P. and Palotie, A. High resolution mapping using fluorescence in situ hybridization to extended DNA fibers prepared from agarose-embedded cells. (1994) *Biotechniques,* **17,** 928-9, 932-3.

Henneberry, A. L. and McMaster, C. R. Cloning and expression of a human choline/ethanolaminephosphotransferase: synthesis of phosphatidylcholine and phosphatidylethanolamine. (1999) *Biochem J,* **339 (Pt 2),** 291-8.

Higgins, D. G. CLUSTAL V: multiple alignment of DNA and protein sequences. (1994) *Methods Mol Biol,* **25,** 307-18.

Hijikata, M., Ohta, Y. and Mishiro, S. Identification of a single nucleotide polymorphism in the MxA gene promoter (G/T at nt -88) correlated with the response of hepatitis C patients to interferon. (2000) *Intervirology,* **43,** 124-7.

Hillier, L. D., Lennon, G., Becker, M., Bonaldo, M. F., Chiapelli, B., Chissoe, S., et al. Generation and analysis of 280,000 human expressed sequence tags. (1996) *Genome Res,* **6,** 807-28.

Hoffman, H. M., Mueller, J. L., Broide, D. H., Wanderer, A. A. and Kolodner, R. D. Mutation of a new gene encoding a putative pyrin-like protein causes familial cold autoinflammatory syndrome and Muckle-Wells syndrome. (2001) *Nat Genet,* **29,** 301-5.

Horton, R., Niblett, D., Milne, S., Palmer, S., Tubby, B., Trowsdale, J., et al. Large-scale sequence comparisons reveal unusually high levels of variation in the HLA-DQB1 locus in the class II region of the human MHC. (1998) *Journal Of Molecular Biology,* **282,** 71-97.

Hoskins, R. A., Nelson, C. R., Berman, B. P., Laverty, T. R., George, R. A., Ciesiolka, L., et al. A BAC-based physical map of the major autosomes of Drosophila melanogaster. (2000) *Science,* **287,** 2271-4.

Hsia, D. Y.-Y., Naylor, J., Bigler, J. A. Gaucher's disease: report of two cases in father and son and review of the literature. (1959) *New Eng. J. Med.,* **261,** 164-169.

Hu, B., Trinh, K., Figueira, W. F. and Price, P. A. Isolation and sequence of a novel human chondrocyte protein related to mammalian members of the chitinase protein family. (1996) *J Biol Chem,* **271,** 19415-20.

Hubert, R. S., Mitchell, S., Chen, X. N., Ekmekji, K., Gadomski, C., Sun, Z., et al. BAC and PAC contigs covering 3.5 Mb of the Down syndrome congenital heart disease region between D21S55 and MX1 on chromosome 21. (1997) *Genomics,* **41,** 218-26.

Hudson, T. J., Engelstein, M., Lee, M. K., Ho, E. C., Rubenfield, M. J., Adams, C. P., et al. Isolation and chromosomal assignment of 100 highly informative human simple sequence repeat polymorphisms. (1992) *Genomics,* **13,** 622-9.

Hudson, T. J., Stein, L. D., Gerety, S. S., Ma, J., Castle, A. B., Silva, J., et al. An STS-based map of the human genome. (1995) *Science,* **270,** 1945-54.

Huntington's Disease Collaborative Research Group, T. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. (1993) *Cell,* **72,** 971-83.

Hurst, L. D. and Eyre-Walker, A. Evolutionary genomics: reading the bands. (2000) *Bioessays,* **22,** 105-7.

International Human Genome Sequencing Consortium, T. Initial sequencing and analysis of the human genome. (2001) *Nature,* **409,** 860-921.

International SNP Map Working Group, T. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. (2001) *Nature,* **409,** 928-33.

Ioannou, P. A., Amemiya, C. T., Garnes, J., Kroisel, P. M., Shizuya, H., Chen, C., et al. A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. (1994) *Nat Genet,* **6,** 84-9.

Irizarry, K., Kustanovich, V., Li, C., Brown, N., Nelson, S., Wong, W., et al. Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. (2000) *Nat Genet,* **26,** 233-6.

Isbrandt, D., Leicher, T., Waldschutz, R., Zhu, X., Luhmann, U., Michel, U., et al. Gene structures and expression profiles of three human KCND (Kv4) potassium channels mediating A-type currents I(TO) and I(SA). (2000) *Genomics,* **64,** 144-54.

Jang, W., Hua, A., Spilson, S. V., Miller, W., Roe, B. A. and Meisler, M. H. Comparative sequence of human and mouse BAC clones from the mnd2 region of chromosome 2p13. (1999) *Genome Res,* **9,** 53-61.

Jeffers, M., Paciucci, R. and Pellicer, A. Characterization of unr; a gene closely linked to N-ras. (1990) *Nucleic Acids Res,* **18,** 4891-9.

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

Jones, J. M., Morrell, J. C. and Gould, S. J. Identification and characterization of HAOX1, HAOX2, and HAOX3, three human peroxisomal 2-hydroxy acid oxidases. (2000) *J Biol Chem,* **275,** 12590-7.

Jurka, J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. (1997) *Proc Natl Acad Sci U S A,* **94,** 1872-7.

Kan, Y. W. and Dozy, A. M. Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation. (1978) *Proc Natl Acad Sci U S A,* **75,** 5631-5.

Karvonen, M. K., Pesonen, U., Heinonen, P., Laakso, M., Rissanen, A., Naukkarinen, H., et al. Identification of new sequence variants in the leptin gene. (1998) *J Clin Endocrinol Metab,* **83,** 3239-42.

Kent, W. J. BLAT--the BLAST-like alignment tool. (2002) *Genome Res,* **12,** 656-64.

Kessler, M. M., Beckendorf, R. C., Westhafer, M. A. and Nordstrom, J. L. Requirement of A-A-U-A-A-A and adjacent downstream sequences for SV40 early polyadenylation. (1986) *Nucleic Acids Res,* **14,** 4939-52.

Khan, A. S., Wilcox, A. S., Hopkins, J. A. and Sikela, J. M. Efficient double stranded sequencing of cDNA clones containing long poly(A) tails using anchored poly(dT) primers. (1991) *Nucleic Acids Res,* **19,** 1715.

Khan, S. G., Muniz-Medina, V., Shahlavi, T., Baker, C. C., Inui, H., Ueda, T., et al. The human XPC DNA repair gene: arrangement, splice site information content and influence of a single

nucleotide polymorphism in a splice acceptor site on alternative splicing and function. (2002) *Nucleic Acids Res,* **30,** 3624-31.

Kiechle, M., Hinrichs, M., Jacobsen, A., Luttges, J., Pfisterer, J., Kommoss, F., et al. Genetic imbalances in precursor lesions of endometrial cancer detected by comparative genomic hybridization. (2000) *Am J Pathol,* **156,** 1827-33.

Kim, U. J., Shizuya, H., de Jong, P. J., Birren, B. and Simon, M. I. Stable propagation of cosmid sized human DNA inserts in an F factor based vector. (1992) *Nucleic Acids Res,* **20,** 1083-5.

Kimberling, W. J., Weston, M. D., Moller, C., Davenport, S. L., Shugart, Y. Y., Priluck, I. A., et al. Localization of Usher syndrome type II to chromosome 1q. (1990) *Genomics,* **7,** 245-9.

Kitayama, H., Sugimoto, Y., Matsuzaki, T., Ikawa, Y. and Noda, M. A ras-related gene with transformation suppressor activity. (1989) *Cell,* **56,** 77-84.

Knight, J. C., Udalova, I., Hill, A. V., Greenwood, B. M., Peshu, N., Marsh, K., et al. A polymorphism that affects OCT-1 binding to the TNF promoter region is associated with severe malaria. (1999) *Nat Genet,* **22,** 145-50.

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

Kohl, S., Baumann, B., Rosenberg, T., Kellner, U., Lorenz, B., Vadala, M., et al. Mutations in the cone photoreceptor G-protein alpha-subunit gene GNAT2 in patients with achromatopsia. (2002) *Am J Hum Genet,* **71,** 422-5.

Kondo, S., Schutte, B. C., Richardson, R. J., Bjork, B. C., Knight, A. S., Watanabe, Y., et al. Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. (2002) *Nat Genet,* **32,** 285-9.

Kondoh, N., Nishina, Y., Tsuchida, J., Koga, M., Tanaka, H., Uchida, K., et al. Assignment of synaptonemal complex protein 1 (SCP1) to human chromosome 1p13 by fluorescence in situ hybridization and its expression in the testis. (1997) *Cytogenet Cell Genet,* **78,** 103-4.

Kondrashov, A. S. and Shabalina, S. A. Classification of common conserved sequences in mammalian intergenic regions. (2002) *Hum Mol Genet,* **11,** 669-74.

Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., et al. A high-resolution recombination map of the human genome. (2002) *Nat Genet,* **31,** 241-7.

Koop, B. F. and Hood, L. Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. (1994) *Nat Genet,* **7,** 48-53.

Kozak, M. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. (1987) *Nucleic Acids Res,* **15,** 8125-48.

Kritzler, R. A., Terner, J. Y., Lindenbaum, J., Magidson, J., Williams, R., Preisig, R., Phillips, G. B. Chediak-Higashi syndrome: cytologic and serum lipid observations in a case and family. (1964) *Am. J. Med,* **36,** 583-594.

Kruglyak, L. and Nickerson, D. A. Variation is the spice of life. (2001) *Nat Genet,* **27,** 234-6.

Kudoh, J., Nagamine, K., Asakawa, S., Abe, I., Kawasaki, K., Maeda, H., et al. Localization of 16 exons to a 450-kb region involved in the autoimmune polyglandular disease type I (APECED) on human chromosome 21q22.3. (1997) *DNA Res,* **4,** 45-52.

Kupfer, K., Smith, M. W., Quackenbush, J. and Evans, G. A. Physical mapping of complex genomes by sampled sequencing: a theoretical analysis. (1995) *Genomics,* **27,** 90-100.

Labay, V., Raz, T., Baron, D., Mandel, H., Williams, H., Barrett, T., et al. Mutations in SLC19A2 cause thiamine-responsive megaloblastic anaemia associated with diabetes mellitus and deafness. (1999) *Nat Genet,* **22,** 300-4.

Lachance, Y., Luu-The, V., Verreault, H., Dumont, M., Rheaume, E., Leblanc, G., et al. Structure of the human type II 3 beta-hydroxysteroid dehydrogenase/delta 5-delta 4 isomerase (3 beta-HSD) gene: adrenal and gonadal specificity. (1991) *DNA Cell Biol,* **10,** 701-11.

Lahat, H., Pras, E., Olender, T., Avidan, N., Ben-Asher, E., Man, O., et al. A missense mutation in a highly conserved region of CASQ2 is associated with autosomal recessive catecholamine-induced polymorphic ventricular tachycardia in Bedouin families from Israel. (2001) *Am J Hum Genet,* **69,** 1378-84.

Lamerdin, J. E. and Carrano, A. V. Automated fluorescence-based restriction fragment analysis. (1993) *Biotechniques,* **15,** 294-303.

Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. CpG islands as gene markers in the human genome. (1992) *Genomics,* **13,** 1095-107.

Lee, K. A., Kim, S. H., Woo, H. Y., Hong, Y. J. and Cho, H. C. Increased frequencies of glutathione S-transferase (GSTM1 and GSTT1) gene deletions in Korean patients with acquired aplastic anemia. (2001) *Blood,* **98,** 3483-5.

Levine, A. and Durbin, R. A computational scan for U12-dependent introns in the human genome sequence. (2001) *Nucleic Acids Res,* **29,** 4006-13.

Levy-Lahad, E., Wijsman, E. M., Nemens, E., Anderson, L., Goddard, K. A., Weber, J. L., et al. A familial Alzheimer's disease locus on chromosome 1. (1995) *Science,* **269,** 970-3.

Li, W. H. and Sadler, L. A. Low nucleotide diversity in man. (1991) *Genetics,* **129,** 513-23.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S. L. and Quackenbush, J. Gene index analysis of the human genome estimates approximately 120,000 genes. (2000) *Nat Genet,* **25,** 239-40.

Litt, M. and Luty, J. A. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. (1989) *Am J Hum Genet,* **44,** 397-401.

Little, P., Curtis, P., Coutelle, C., Van den Berg, J., Dalgleish, R., Malcolm, S., et al. Isolation and partial sequence of recombinant plasmids containing human alpha-, beta- and gamma-globin cDNA fragments. (1978) *Nature,* **273,** 640-3.

Liu, H., Nakagawa, T., Kanematsu, T., Uchida, T. and Tsuji, S. Isolation of 10 differentially expressed cDNAs in differentiated Neuro2a cells induced through controlled expression of the GD3 synthase gene. (1999) *J Neurochem,* **72,** 1781-90.

Loetscher, P., Alvarez-Gonzalez, R. and Althaus, F. R. Poly(ADP-ribose) may signal changing metabolic conditions to the chromatin of mammalian cells. (1987) *Proc Natl Acad Sci U S A,* **84,** 1286-9.

Makalowski, W., Zhang, J. and Boguski, M. S. Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. (1996) *Genome Res,* **6,** 846-57.

Mallon, A. M., Platzer, M., Bate, R., Gloeckner, G., Botcherby, M. R., Nordsiek, G., et al. Comparative genome sequence analysis of the Bpa/Str region in mouse and Man. (2000) *Genome Res,* **10,** 758-75.

Mannervik, B. and Danielson, U. H. Glutathione transferases--structure and catalytic activity. (1988) *CRC Crit Rev Biochem,* **23,** 283-337.

Marcelino, J., Carpten, J. D., Suwairi, W. M., Gutierrez, O. M., Schwartz, S., Robbins, C., et al. CACP, encoding a secreted proteoglycan, is mutated in camptodactyly-arthropathy-coxa vara-pericarditis syndrome. (1999) *Nat Genet,* **23,** 319-22.

Marra, M., Kucaba, T., Sekhon, M., Hillier, L., Martienssen, R., Chinwalla, A., et al. A map for sequence analysis of the Arabidopsis thaliana genome. (1999) *Nat Genet,* **22,** 265-70.

Marra, M. A., Kucaba, T. A., Dietrich, N. L., Green, E. D., Brownstein, B., Wilson, R. K., et al. High throughput fingerprint analysis of large-insert clones. (1997) *Genome Res,* **7,** 1072-84.

Marshall, E. Drug firms to create public database of genetic mutations. (1999) *Science,* **284,** 406-7.

Marth, G., Yeh, R., Minton, M., Donaldson, R., Li, Q., Duan, S., et al. Single-nucleotide polymorphisms in the public domain: how useful are they? (2001) *Nat Genet,* **27,** 371-2.

Mattick, J. S. Non-coding RNAs: the architects of eukaryotic complexity. (2001) *EMBO Rep,* **2,** 986-91.

Maxam, A. M. and Gilbert, W. A new method for sequencing DNA. (1977) *Proc Natl Acad Sci U S A,* **74,** 560-4.

McPherson, J. D., Marra, M., Hillier, L., Waterston, R. H., Chinwalla, A., Wallis, J., et al. A physical map of the human genome. (2001) *Nature,* **409,** 934-41.

Meek, T. D., Dayton, B. D., Metcalf, B. W., Dreyer, G. B., Strickler, J. E., Gorniak, J. G., et al. Human immunodeficiency virus 1 protease expressed in Escherichia coli behaves as a dimeric aspartic protease. (1989) *Proc Natl Acad Sci U S A,* **86,** 1841-5.

Mercher, T., Coniat, M. B., Monni, R., Mauchauffe, M., Khac, F. N., Gressin, L., et al. Involvement of a human gene related to the Drosophila spen gene in the recurrent t(1;22) translocation of acute megakaryocytic leukemia. (2001) *Proc Natl Acad Sci U S A,* **98,** 5776-9.

Milpetz, F., Argos, P. and Persson, B. TMAP: a new email and WWW service for membrane-protein structural predictions. (1995) *Trends Biochem Sci,* **20,** 204-5.

Mirzoeva, S., Weigand, S., Lukas, T. J., Shuvalova, L., Anderson, W. F. and Watterson, D. M. Analysis of the functional coupling between calmodulin's calcium binding and peptide recognition properties. (1999) *Biochemistry,* **38,** 3936-47.

Misra, S., 1, 2 , Madeline A Crosby3, Christopher J Mungall2, 4, Beverley B Matthews3, Kathryn S Campbell3, Pavel Hradecky3, Yanmei Huang3, Joshua S Kaminker1, 2, Gillian H Millburn5, Simon E Prochnik1, 2, Christopher D Smith1, 2, Jonathan L Tupy1, 2, Eleanor J Whitfied6, Leyla Bayraktaroglu3, Benjamin P Berman1, Brian R Bettencourt3, Susan E Celniker7, Aubrey DNJ de Grey5, Rachel A Drysdale5, Nomi L Harris2, 7, John Richter4, Susan Russo3, Andrew J Schroeder3, Sheng Qiang Shu1, 2, Mark Stapleton7, Chihiro Yamada5, Michael Ashburner5, William M Gelbart3, Gerald M Rubin1, 2, 4, 7 and Suzanna E Lewis1, 2 Annotation of the Drosophila melanogaster euchromatic genome: a systematic review. (2002) *Genome Biology,* **3**.

Mochizuki, N., Cho, G., Wen, B. and Insel, P. A. Identification and cDNA cloning of a novel human mosaic protein, LGN, based on interaction with G alpha i2. (1996) *Gene,* **181,** 39-43.

Modrek, B. and Lee, C. A genomic view of alternative splicing. (2002) *Nat Genet,* **30,** 13-9.

Moore, M. J. and Sharp, P. A. Evidence for two active sites in the spliceosome provided by stereochemistry of pre-mRNA splicing. (1993) *Nature,* **365,** 364-8.

Morton, N. E. Parameters of the human genome. (1991) *Proc Natl Acad Sci U S A,* **88,** 7474-6.

Mouse Genome Sequencing Consortium, T. Initial sequencing and comparative analysis of the mouse genome. (2002) *Nature,* **420,** 520-62.

Mulligan, L. M., Kwok, J. B., Healey, C. S., Elsdon, M. J., Eng, C., Gardner, E., et al. Germ-line mutations of the RET proto-oncogene in multiple endocrine neoplasia type 2A. (1993) *Nature,* **363,** 458-60.

Mullikin, J. C., Hunt, S. E., Cole, C. G., Mortimore, B. J., Rice, C. M., Burton, J., et al. An SNP map of human chromosome 22. (2000) *Nature,* **407,** 516-20.

Mungall, A. J., Edwards, C. A., Ranby, S. A., Humphray, S. J., Heathcott, R. W., Clee, C. M., et al. Physical mapping of chromosome 6: a strategy for the rapid generation of sequence-ready contigs. (1996) *DNA Seq,* **7,** 47-9.

Murray, J. C., Buetow, K. H., Weber, J. L., Ludwigsen, S., Scherpbier-Heddema, T., Manion, F., et al. A comprehensive human linkage map with centimorgan density. Cooperative Human Linkage Center (CHLC). (1994) *Science,* **265,** 2049-54.

Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. (1995) *J Mol Biol,* **247,** 536-40.

Myers, E. W., Sutton, G. G., Smith, H. O., Adams, M. D. and Venter, J. C. On the sequencing and assembly of the human genome. (2002) *Proc Natl Acad Sci U S A,* **99,** 4145-6.

Nagasaki, K., Maass, N., Manabe, T., Hanzawa, H., Tsukada, T., Kikuchi, K., et al. Identification of a novel gene, DAM1, amplified at chromosome 1p13.3-21 region in human breast cancer cell lines. (1999) *Cancer Lett,* **140,** 219-26.

Nagase, T., Ishikawa,K., Suyama,M., Kikuno,R., Hirosawa,M., and Miyajima, N., Tanaka,A., Kotani,H., Nomura,N. and Ohara,O. Prediction of the coding sequences of unidentified human genes.

XII. The complete sequences of 100 new cDNA clones from brain which

code for large proteins in vitro. (1998) *Dna Research,* **5,** 355-364.

Nagata, T., Mugishima, H., Shichino, H., Suzuki, T., Chin, M., Koshinaga, S., et al. Karyotypic analyses of hepatoblastoma. Report of two cases and review of the literature suggesting

chromosomal loci responsible for the pathogenesis of this disease. (1999) *Cancer Genet Cytogenet,* **114,** 42-50.

Nakachi, K., Imai, K., Hayashi, S., Watanabe, J. and Kawajiri, K. Genetic susceptibility to squamous cell carcinoma of the lung in relation to cigarette smoking dose. (1991) *Cancer Res,* **51,** 5177-80.

Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., et al. Variable number of tandem repeat (VNTR) markers for human gene mapping. (1987) *Science,* **235,** 1616-22.

Navia, M. A., Fitzgerald, P. M., McKeever, B. M., Leu, C. T., Heimbach, J. C., Herber, W. K., et al. Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1. (1989) *Nature,* **337,** 615-20.

Nicholas, K. B., Nicholas H.B. , and Deerfield, D.W. GeneDoc: Analysis and Visualization of Genetic Variation. (1997) *EMBNEW.NEWS,* **4**.

Ning, Z., Cox, A. J. and Mullikin, J. C. SSAHA: a fast search method for large DNA databases. (2001) *Genome Res,* **11,** 1725-9.

Nitta, N., Ochiai, M., Nagao, M. and Sugimura, T. Amino-acid substitution at codon 13 of the N-ras oncogene in rectal cancer in a Japanese patient. (1987) *Jpn J Cancer Res,* **78,** 21-6.

Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y., et al. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. (1992) *Nat Genet,* **2,** 173-9.

Oliver, J. L. and Marín, A. A relationship between GC content and coding-sequence length. (1996) *Journal Of Molecular Evolution,* **43,** 216-23.

Olson, M., Hood, L., Cantor, C. and Botstein, D. A common language for physical mapping of the human genome. (1989) *Science,* **245,** 1434-5.

Olson, M. V., Dutchik, J. E., Graham, M. Y., Brodeur, G. M., Helms, C., Frank, M., et al. Random-clone strategy for genomic restriction mapping in yeast. (1986) *Proc Natl Acad Sci U S A,* **83,** 7826-30.

Olson, M. V. and Green, P. Criterion for the completeness of large-scale physical maps of DNA. (1993) *Cold Spring Harb Symp Quant Biol,* **58,** 349-55.

Orias, M., Bray-Ward, P., Curran, M. E., Keating, M. T. and Desir, G. V. Genomic localization of the human gene for KCNA10, a cGMP-activated K channel. (1997) *Genomics,* **42,** 33-7.

Pan, S. S., Han, Y., Farabaugh, P. and Xia, H. Implication of alternative splicing for expression of a variant NAD(P)H:quinone oxidoreductase-1 with a single nucleotide polymorphism at 465C>T. (2002) *Pharmacogenetics,* **12,** 479-88.

Parvari, R., Hershkovitz, E., Grossman, N., Gorodischer, R., Loeys, B., Zecic, A., et al. Mutation of TBCE causes hypoparathyroidism-retardation-dysmorphism and autosomal recessive Kenny-Caffey syndrome. (2002) *Nat Genet,* **32,** 448-52.

Pearl, L. H. and Taylor, W. R. Sequence specificity of retroviral proteases. (1987a) *Nature,* **328,** 482.

Pearl, L. H. and Taylor, W. R. A structural model for the retroviral proteases. (1987b) *Nature,* **329,** 351-4.

Peitsch, M. C. and Jongeneel, C. V. A 3-D model for the CD40 ligand predicts that it is a compact trimer similar to the tumor necrosis factors. (1993) *Int Immunol,* **5,** 233-8.

Pemble, S., Schroeder, K. R., Spencer, S. R., Meyer, D. J., Hallier, E., Bolt, H. M., et al. Human glutathione S-transferase theta (GSTT1): cDNA cloning and the characterization of a genetic polymorphism. (1994) *Biochem J,* **300 (Pt 1),** 271-6.

Pennacchio, L. A., Olivier, M., Hubacek, J. A., Cohen, J. C., Cox, D. R., Fruchart, J. C., et al. An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. (2001) *Science,* **294,** 169-73.

Pennacchio, L. A. and Rubin, E. M. Genomic strategies to identify mammalian regulatory sequences. (2001) *Nat Rev Genet,* **2,** 100-9.

Pericak-Vance, M. A., Bebout, J. L., Gaskell, P. C., Jr., Yamaoka, L. H., Hung, W. Y., Alberts, M. J., et al. Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. (1991) *Am J Hum Genet,* **48,** 1034-50.

Persson, B. and Argos, P. Prediction of transmembrane segments in proteins utilising multiple sequence alignments. (1994) *J Mol Biol,* **237,** 182-92.

Petersen, C. M., Nielsen, M. S., Nykjaer, A., Jacobsen, L., Tommerup, N., Rasmussen, H. H., et al. Molecular identification of a novel candidate sorting receptor purified from human brain by receptor-associated protein affinity chromatography. (1997) *J Biol Chem,* **272,** 3599-605.

Pfost, D. R., Boyce-Jacino, M. T. and Grant, D. M. A SNPshot: pharmacogenetics and the future of drug therapy. (2000) *Trends Biotechnol,* **18,** 334-8.

Picoult-Newberg, L., Ideker, T. E., Pohl, M. G., Taylor, S. L., Donaldson, M. A., Nickerson, D. A., et al. Mining SNPs from EST databases. (1999) *Genome Res,* **9,** 167-74.

Pinkel, D., Landegent, J., Collins, C., Fuscoe, J., Segraves, R., Lucas, J., et al. Fluorescence in situ hybridization with human chromosome-specific libraries: detection of trisomy 21 and translocations of chromosome 4. (1988) *Proc Natl Acad Sci U S A,* **85,** 9138-42.

Pinkel, D., Straume, T. and Gray, J. W. Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization. (1986) *Proc Natl Acad Sci U S A,* **83,** 2934-8.

Platzer, M., Rotman, G., Bauer, D., Uziel, T., Savitsky, K., Bar-Shira, A., et al. Ataxia-telangiectasia locus: sequence analysis of 184 kb of human genomic DNA containing the entire ATM gene. (1997) *Genome Res,* **7,** 592-605.

Rabbitts, T. H. Bacterial cloning of plasmids carrying copies of rabbit globin messenger RNA. (1976) *Nature,* **260,** 221-5.

Ragoussis, J., and Olavesen, M.G. (1997) *Chromosome walking,* O.U.P.

Raich, N., Mattei, M. G., Romeo, P. H. and Beaupain, D. PHTF, a novel atypical homeobox gene on chromosome 1p13, is evolutionarily conserved. (1999) *Genomics,* **59,** 108-9.

Ramashwami, M., Gautam,M., Kamb,A.A., Rudy,B., Tanouye,M.A. and Mathew,M.K. Human potassium channel genes: molecular cloning and functional expression. (1990) *Mol Cell Biol Neurosci,* **1,** 214 - 223.

Rao, P. H., Houldsworth, J., Dyomina, K., Parsa, N. Z., Cigudosa, J. C., Louie, D. C., et al. Chromosomal and gene amplification in diffuse large B-cell lymphoma. (1998) *Blood,* **92,** 234-40.

Reymond, A., Meroni, G., Fantozzi, A., Merla, G., Cairo, S., Luzi, L., et al. The tripartite motif family identifies cell compartments. (2001) *Embo J,* **20,** 2140-51.

Riles, L., Dutchik, J. E., Baktha, A., McCauley, B. K., Thayer, E. C., Leckie, M. P., et al. Physical maps of the six smallest chromosomes of Saccharomyces cerevisiae at a resolution of 2.6 kilobase pairs. (1993) *Genetics,* **134,** 81-150.

Roach, J. C., Boysen, C., Wang, K. and Hood, L. Pairwise end sequencing: a unified approach to genomic mapping and sequencing. (1995) *Genomics,* **26,** 345-53.

Roest Crollius, H., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., et al. Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. (2000) *Nat Genet,* **25,** 235-8.

Roses, A. D. Pharmacogenetics and the practice of medicine. (2000) *Nature,* **405,** 857-65.

Roses, A. D. Genome-based pharmacogenetics and the pharmaceutical industry. (2002) *Nat Rev Drug Discov,* **1,** 541-9.

Royle, N. J., Clarkson, R. E., Wong, Z. and Jeffreys, A. J. Clustering of hypervariable minisatellites in the proterminal regions of human autosomes. (1988) *Genomics,* **3,** 352-60.

Sabina, R. L., Fishbein, W. N., Pezeshkpour, G., Clarke, P. R. and Holmes, E. W. Molecular analysis of the myoadenylate deaminase deficiencies. (1992) *Neurology,* **42,** 170-9.

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. (1988) *Science,* **239,** 487-91.

Salvatore, C. A., Jacobson, M. A., Taylor, H. E., Linden, J. and Johnson, R. G. Molecular cloning and characterization of the human A3 adenosine receptor. (1993) *Proc Natl Acad Sci U S A,* **90,** 10365-9.

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., et al. Nucliotide sequence of bacteriophage phi X174 DNA. (1977) *Nature,* **265,** 687-95.

Sanger, F., Coulson, A. R., Friedmann, T., Air, G. M., Barrell, B. G., Brown, N. L., et al. The nucleotide sequence of bacteriophage phiX174. (1978) *J Mol Biol,* **125,** 225-46.

Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. and Petersen, G. B. Nucleotide sequence of bacteriophage lambda DNA. (1982) *J Mol Biol,* **162,** 729-73.

Saupe, S., Roizes, G., Peter, M., Boyle, S., Gardiner, K. and De Sario, A. Molecular cloning of a human cDNA IGSF3 encoding an immunoglobulin-like membrane protein: expression and mapping to chromosome band 1p13. (1998) *Genomics,* **52,** 305-11.

Scherf, M., Klingenhoff, A. and Werner, T. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. (2000) *J Mol Biol,* **297,** 599-606.

Schuler, G. D. Sequence mapping by electronic PCR. (1997) *Genome Res,* **7,** 541-50.

Schuler, G. D., Boguski, M. S., Stewart, E. A., Stein, L. D., Gyapay, G., Rice, K., et al. A gene map of the human genome. (1996) *Science,* **274,** 540-6.

Schutte, B. C., Carpten, J. D., Forus, A., Gregory, S. G., Horii, A. and White, P. S. Report and abstracts of the sixth international workshop on human chromosome 1 mapping 2000. Iowa City, Iowa, USA. 30 September-3 October 2000. (2001) *Cytogenet Cell Genet,* **92,** 23-41.

Sedlacek, Z., Korn, B., Konecki, D. S., Siebenhaar, R., Coy, J. F., Kioschis, P., et al. Construction of a transcription map of a 300 kb region around the human G6PD locus by direct cDNA selection. (1993) *Hum Mol Genet,* **2,** 1865-9.

Seed, B., Aruffo A Molecular cloning of the CD2 antigen, the T-cell erythrocyte receptor, by a rapid immunoselection procedure. (1987) *Proc Natl Acad Sci U S A,* **84**.

Seidegard, J., Vorachek, W. R., Pero, R. W. and Pearson, W. R. Hereditary differences in the expression of the human glutathione transferase active on trans-stilbene oxide are due to a gene deletion. (1988) *Proc Natl Acad Sci U S A,* **85,** 7293-7.

Shackleton, S., Lloyd, D. J., Jackson, S. N., Evans, R., Niermeijer, M. F., Singh, B. M., et al. LMNA, encoding lamin A/C, is mutated in partial lipodystrophy. (2000) *Nat Genet,* **24,** 153-6.

Shen, L. X., Basilion, J. P. and Stanton, V. P., Jr. Single-nucleotide polymorphisms can cause different structural folds of mRNA. (1999) *Proc Natl Acad Sci U S A,* **96,** 7871-6.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. dbSNP: the NCBI database of genetic variation. (2001) *Nucleic Acids Res,* **29,** 308-11.

Shimada, T., Fujii, H. and Lin, H. A 165-base pair sequence between the dihydrofolate reductase gene and the divergently transcribed upstream gene is sufficient for bidirectional transcriptional activity. (1989) *J Biol Chem,* **264,** 20171-4.

Shizuya, H., Birren, B., Kim, U. J., Mancino, V., Slepak, T., Tachiiri, Y., et al. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. (1992) *Proc Natl Acad Sci U S A,* **89,** 8794-7.

Skolnick, J. and Fetrow, J. S. From genes to protein structure and function: novel applications of computational approaches in the genomic era. (2000) *Trends Biotechnol,* **18,** 34-9.

Slim, R., Le Paslier, D., Compain, S., Levilliers, J., Ougen, P., Billault, A., et al. Construction of a yeast artificial chromosome contig spanning the pseudoautosomal region and isolation of 25 new sequence-tagged sites. (1993) *Genomics,* **16,** 691-7.

Smit, A. F. Interspersed repeats and other mementos of transposable elements in mammalian genomes. (1999) *Curr Opin Genet Dev,* **9,** 657-63.

Soderlund, C., Longden, I. and Mott, R. FPC: a system for building contigs from restriction fingerprinted clones. (1997) *Comput Appl Biosci,* **13,** 523-35.

Solovyev, V. V., Salamov, A. A. and Lawrence, C. B. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. (1994) *Nucleic Acids Res,* **22,** 5156-63.

Solovyev, V. V., Salamov, A. A. and Lawrence, C. B. Identification of human gene structure using linear discriminant functions and dynamic programming. (1995) *Proc Int Conf Intell Syst Mol Biol,* **3,** 367-75.

Sonnhammer, E. L. and Durbin, R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. (1995) *Gene,* **167,** GC1-10.

Sonnhammer, E. L., von Heijne, G. and Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. (1998) *Proc Int Conf Intell Syst Mol Biol,* **6,** 175-82.

Soriano, P., Meunier-Rotival, M. and Bernardi, G. The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes. (1983) *Proc Natl Acad Sci U S A,* **80,** 1816-20.

Sparkes, R. S. C., V. H.; Mohandas, T.; Zollman, S.; Cire-Eversole, P.; Amatruda, T. T.; Reed, R. R.; Lochrie, M. A.; Simon, M. I Mapping of genes encoding the subunits of guanine nucleotide-binding protein (G-proteins) in humans. (1987) *Cytogenet Cell Genet,* **46,** 696.

Staden, R. A new computer method for the storage and manipulation of DNA gel reading data. (1980) *Nucleic Acids Res,* **8,** 3673-94.

Strittmatter, W. J., Saunders, A. M., Schmechel, D., Pericak-Vance, M., Enghild, J., Salvesen, G. S., et al. Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. (1993) *Proc Natl Acad Sci U S A,* **90,** 1977-81.

Sudo, K., Chinen, K. and Nakamura, Y. 2058 expressed sequence tags (ESTs) from a human fetal lung cDNA library. (1994) *Genomics,* **24,** 276-9.

Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., 3rd, Kondrashov, A. S. and Bork, P. Prediction of deleterious human alleles. (2001) *Hum Mol Genet,* **10,** 591-7.

Taillon-Miller, P., Gu, Z., Li, Q., Hillier, L. and Kwok, P. Y. Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. (1998) *Genome Res,* **8,** 748-54.

Takahashi, Y., Campbell, E. A., Hirata, Y., Takayama, T. and Listowsky, I. A basis for differentiating among the multiple human Mu-glutathione S-transferases and molecular cloning of brain GSTM5. (1993) *J Biol Chem,* **268,** 8893-8.

Tang, X., Wang, Y., Li, H. O., Sakatsume, O., Sarai, A. and Yokoyama, K. DNA fingerprinting involving fluorescence-labeled termini of any enzymatically generated fragments of DNA. (1994) *Jpn J Hum Genet,* **39,** 379-91.

Taniguchi, T., Fujii-Kuriyama, Y. and Muramatsu, M. Molecular cloning of human interferon cDNA. (1980) *Proc Natl Acad Sci U S A,* **77,** 4003-6.

Tao, Q., Chang, Y-L., Wang, J., Huaming , C., Islam-Faridi, M.N., Scheuring, C. Wang, B., Stelly, D.M. Zhang, H-B. Bacterial Artificial Chromosome-Based Physical Map of the Rice Genome Constructed by Restriction Fingerprint Analysis. (2001) *Genetics,* **158,** 1711-1724.

Taylor, K., Hornigold, N., Conway, D., Williams, D., Ulinowski, Z., Agochiya, M., et al. Mapping the human Y chromosome by fingerprinting cosmid clones. (1996) *Genome Res,* **6,** 235-48.

Thierry-Mieg, D. a. A C. elegans DataBase. (1994) *unpublished*.

Thiery, J. P., Macaya, G. and Bernardi, G. An analysis of eukaryotic genomes by density gradient centrifugation. (1976) *J Mol Biol,* **108,** 219-35.

Thomas, J. W., Prasad, A. B., Summers, T. J., Lee-Lin, S. Q., Maduro, V. V., Idol, J. R., et al. Parallel construction of orthologous sequence-ready clone contig maps in multiple species. (2002) *Genome Res,* **12,** 1277-85.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. (1994) *Nucleic Acids Res,* **22,** 4673-80.

Thornton, J. M., Orengo, C. A., Todd, A. E. and Pearl, F. M. Protein folds, functions and evolution. (1999) *J Mol Biol,* **293,** 333-42.

Tilghman, S. M., Tiemeier, D. C., Seidman, J. G., Peterlin, B. M., Sullivan, M., Maizel, J. V., et al. Intervening sequence of DNA identified in the structural portion of a mouse beta-globin gene. (1978) *Proc Natl Acad Sci U S A,* **75,** 725-9.

Tkachuk, D. C., Westbrook, C. A., Andreeff, M., Donlon, T. A., Cleary, M. L., Suryanarayan, K., et al. Detection of bcr-abl fusion in chronic myelogeneous leukemia by in situ hybridization. (1990) *Science,* **250,** 559-62.

Tremblay, L. O., Campbell Dyke, N. and Herscovics, A. Molecular cloning, chromosomal mapping and tissue-specific expression of a novel human alpha1,2-mannosidase gene involved in N-glycan maturation. (1998) *Glycobiology,* **8,** 585-95.

Trenkle, T., McClelland, M., Adlkofer, K. and Welsh, J. Major transcript variants of VAV3, a new member of the VAV family of guanine nucleotide exchange factors. (2000) *Gene,* **245,** 139-49.

Trofatter, J. A., MacCollin, M. M., Rutter, J. L., Murrell, J. R., Duyao, M. P., Parry, D. M., et al. A novel moesin-, ezrin-, radixin-like gene is a candidate for the neurofibromatosis 2 tumor suppressor. (1993) *Cell,* **75,** 826.

Tsukino, H., Kuroda, Y., Qiu, D., Nakao, H., Imai, H. and Katoh, T. Effects of cytochrome P450 (CYP) 2A6 gene deletion and CYP2E1 genotypes on gastric adenocarcinoma. (2002) *Int J Cancer,* **100,** 425-8.

Uberbacher, E. C., Xu, Y. and Mural, R. J. Discovering and understanding genes in human DNA sequence using GRAIL. (1996) *Methods Enzymol,* **266,** 259-81.

Van den Bergh, F., Sabina RL. Characterization of human AMP deaminase 2 (AMPD2) gene expression reveals alternative transcripts encoding variable N-terminal extensions of isoform L. (1995) *Biochem J,* **312,** 401-410.

Vega-Saenz de Miera, E., Moreno H, Fruhling D, Kentros C, Rudy B. Cloning of ShIII (Shaw-like) cDNAs encoding a novel high-voltage-activating, TEA-sensitive, type-A K+ channel. (1992) *Proc R Soc Lond B Biol Sci,* **248,** 9-18.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. The sequence of the human genome. (2001) *Science,* **291,** 1304-51.

Verma, I. M., Temple, G. F., Fan, H. and Baltimore, D. In vitro synthesis of DNA complementary to rabbit reticulocyte 10S RNA. (1972) *Nat New Biol,* **235,** 163-7.

Vetrie, D. Isolation of the defective gene in X linked agammaglobulinaemia. (1993) *J Med Genet,* **30,** 452-3.

Vorachek, W. R., Pearson, W. R. and Rule, G. S. Cloning, expression, and characterization of a class-mu glutathione transferase from human muscle, the product of the GST4 locus. (1991) *Proc Natl Acad Sci U S A,* **88,** 4443-7.

Walker, A. P., Muscatelli, F. and Monaco, A. P. Isolation of the human Xp21 glycerol kinase gene by positional cloning. (1993) *Hum Mol Genet,* **2,** 107-14.

Wallace, M. R., Marchuk, D. A., Andersen, L. B., Letcher, R., Odeh, H. M., Saulino, A. M., et al. Type 1 neurofibromatosis gene: identification of a large transcript disrupted in three NF1 patients. (1990) *Science,* **249,** 181-6.

Wallner, B. P., Frey, A. Z., Tizard, R., Mattaliano, R. J., Hession, C., Sanders, M. E., et al. Primary structure of lymphocyte function-associated antigen 3 (LFA-3). The ligand of the T lymphocyte CD2 glycoprotein. (1987) *J Exp Med,* **166,** 923-32.

Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. (1998) *Science,* **280,** 1077-82.

Wang, M., Duell, T., Gray, J. W. and Weier, H.-U. G. High sensitivity, high resolution physical mapping by fluorescence *in situ* hybridisation on to individually straightened DNA molecules. (1996) *Bioimaging,* **4,** 73-83.

Wang, X., Zuckerman, B., Pearson, C., Kaufman, G., Chen, C., Wang, G., et al. Maternal cigarette smoking, metabolic gene polymorphism, and infant birth weight. (2002) *Jama,* **287,** 195-202.

Wang, Z. and Moult, J. SNPs, protein structure, and disease. (2001) *Hum Mutat,* **17,** 263-70.

Waterston, R. H., Lander, E. S. and Sulston, J. E. On the sequencing of the human genome. (2002) *Proc Natl Acad Sci U S A,* **99,** 3712-6.

Watson, J. D. a. C., F. A Structure for Deoxyribose Nucleic Acid. (1953) *Nature,* **171,** 171.

Weber, J. L. and May, P. E. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. (1989) *Am J Hum Genet,* **44,** 388-96.

Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P., et al. A second-generation linkage map of the human genome. (1992) *Nature,* **359,** 794-801.

White, P. S., Forus, A., Matise, T. C., Schutte, B. C., Spieker, N., Stanier, P., et al. Report of the fifth international workshop on human chromosome 1 mapping 1999. (1999) *Cytogenet Cell Genet,* **87,** 143-71.

Wickens, M. P., Buell, G. N. and Schimke, R. T. Synthesis of double-stranded DNA complementary to lysozyme, ovomucoid, and ovalbumin mRNAs. Optimization for full length second strand synthesis by Escherichia coli DNA polymerase I. (1978) *J Biol Chem,* **253,** 2483-95.

Wilke, C. M., Guo, S. W., Hall, B. K., Boldog, F., Gemmill, R. M., Chandrasekharappa, S. C., et al. Multicolor FISH mapping of YAC clones in 3p14 and identification of a YAC spanning both FRA3B and the t(3;8) associated with hereditary renal cell carcinoma. (1994) *Genomics,* **22,** 319-26.

Wyman, A. R. and White, R. A highly polymorphic locus in human DNA. (1980) *Proc Natl Acad Sci U S A,* **77,** 6754-8.

Xiang, Z., Morse, E., Hu, X. L., Flint, J., Chi, H. C., Grady, D. L., et al. A sequence-ready map of the human chromosome 1q telomere. (2001) *Genomics,* **72,** 105-7.

Xie, Y. G., Han, F. Y., Peyrard, M., Ruttledge, M. H., Fransson, I., DeJong, P., et al. Cloning of a novel, anonymous gene from a megabase-range YAC and cosmid contig in the neurofibromatosis type 2/meningioma region on human chromosome 22q12. (1993) *Hum Mol Genet,* **2,** 1361-8.

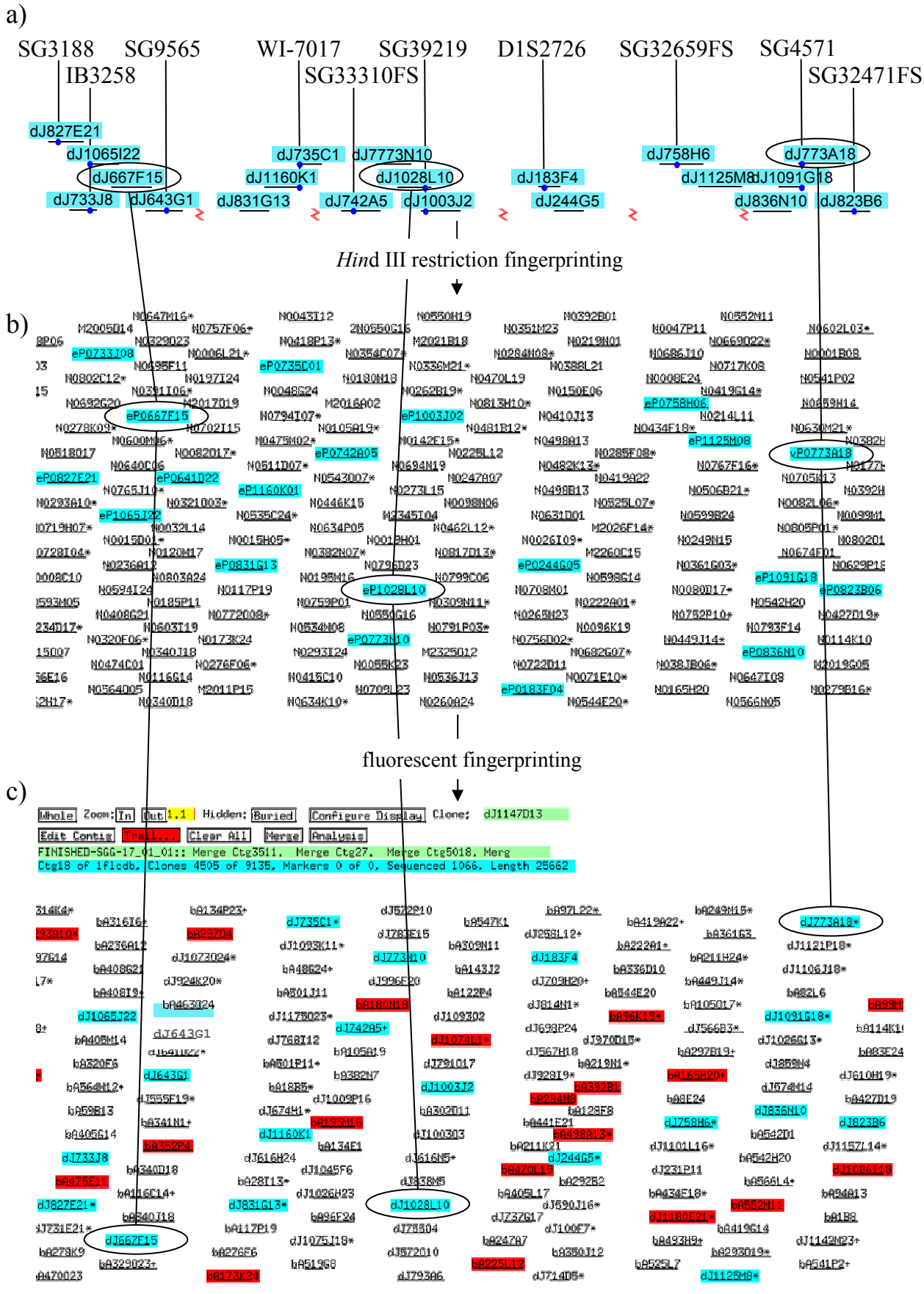
Xu, C. F., Chambers, J. A. and Solomon, E. Complex regulation of the BRCA1 gene. (1997) *J Biol Chem,* **272,** 20994-7.

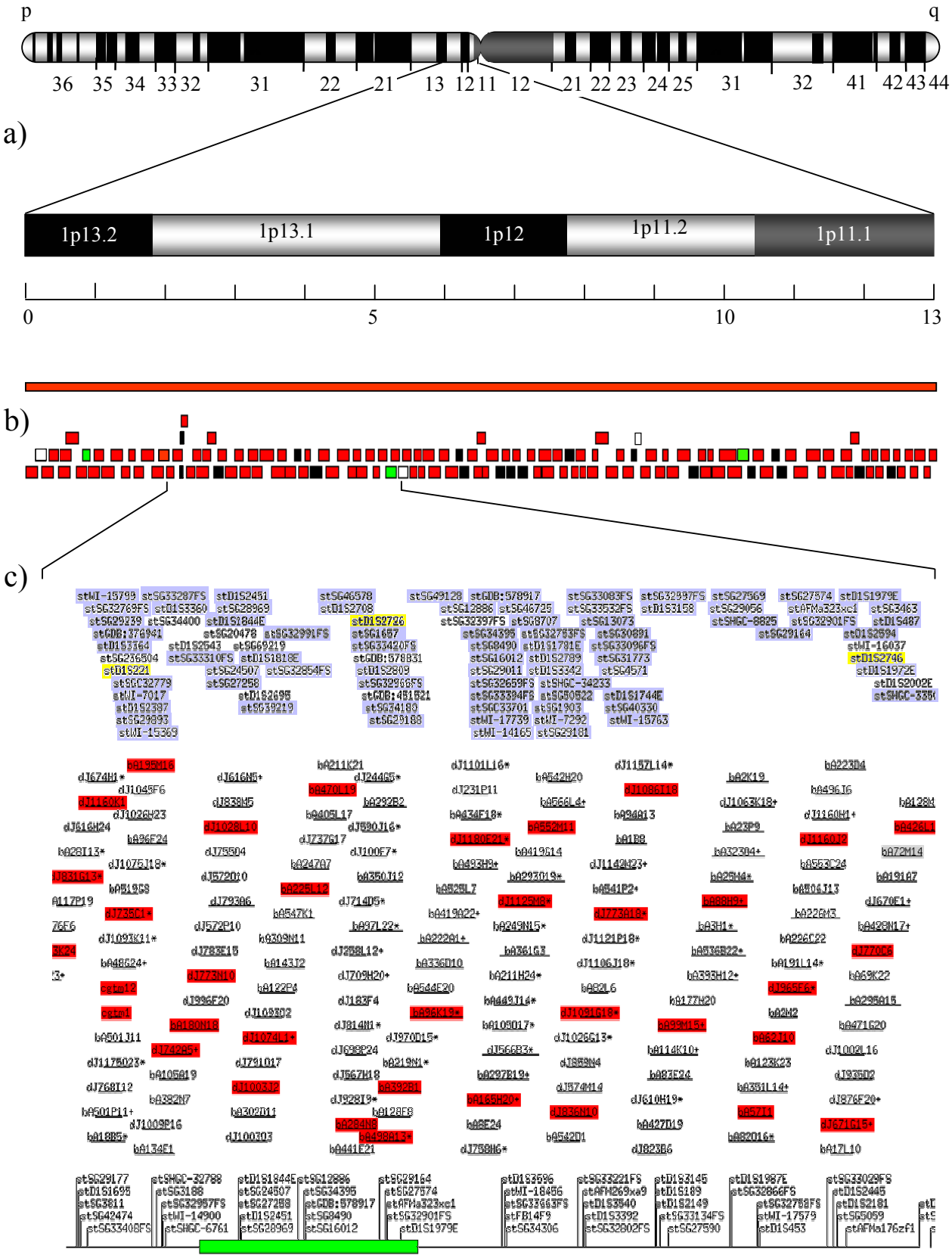
Xu, S., Wang, Y., Roe, B. and Pearson, W. R. Characterization of the human class Mu glutathione S-transferase gene cluster and the GSTM1 deletion. (1998) *J Biol Chem,* **273,** 3517-27.

Zachmann, M. and Prader, A. Unusual heterozygotes of congenital adrenal hyperplasia due to 21-hydroxylase deficiency confirmed by HLA tissue typing. (1979) *Acta Endocrinol (Copenh),* **92,** 542-6.

Zattara-Cannoni, H., Roll, P., Figarella-Branger, D., Lena, G., Dufour, H., Grisoli, F., et al. Cytogenetic study of six cases of radiation-induced meningiomas. (2001) *Cancer Genet Cytogenet,* **126,** 81-4.

Zoubak, S., Clay, O. and Bernardi, G. The gene distribution of the human genome. (1996) *Gene,* **174,** 95-102.


a)
SG3188
IB3258
SG9565
WI-7017
SG33310FS
SG39219
D1S2726
SG32659FS
SG4571
SG32471FS
dJ827E21
dJ1065I22
dJ667F15
dJ733J8
dJ643G1
dJ735C1
dJ1160K1
dJ831G13
dJ7773N10
dJ742A5
dJ1028L10
dJ1003J2
dJ183F4
dJ244G5
dJ758H6
dJ1125M8
dJ773A18
dJ1091G18
dJ836N10
dJ823B6
Hind III restriction fingerprinting
b)
fluorescent fingerprinting
c)
Whole
Zoom: In Out 1.1
Hidden: Buried
Configure Display
Clone: dJ1147D13
Edit Contig
Trail...
Clear All
Merge
Analysis
FINISHED-SGG-17_01_01:: Merge Ctg3511. Merge Ctg27. Merge Ctg5018. Merg
Ctg18 of 1flcdb, Clones 4505 of 9135, Markers 0 of 0, Sequenced 1066, Length 25662


p
q
36 35 34 33 32 31 22 21 13 12 11 12 21 22 23 24 25 31 32 41 42 43 44
a)
1p13.2
1p13.1
1p12
1p11.2
1p11.1
0
5
10
13
b)
c)