Human chromosome 20q12-13.2: Structural, comparative and sequence variation studies

George Stavros Stavrides

A thesis submitted in partial fulfilment of the requirements of Cambridge University for the degree of Doctor of Philosophy

Wolfson College, University of Cambridge

September 2002

This dissertation is the result of my own work and includes nothing that is the outcome of work done in collaboration. The dissertation does not exceed the length limit set by the Biology Degree Committee.

To my parents

Abstract

As the human genome sequencing effort nears completion, there is a great need to identify and characterise the structural pieces of genetic information embedded in the generated sequence. The aim of this project was to explore how this goal can be achieved, in order to maximise the impact of the Human Genome Project on genome biology. A 10 Mb region on human chromosome 20q12-13.2 (representing 1/6th of the whole chromosome) provided the basis for a number of studies.

The assembly of a detailed transcript map across this region is described. Candidate gene features were identified from publicly available expressed sequences and *ab initio* gene predictions, then experimentally verified and extended. The final transcript map contains 99 coding genes, 30 putative genes and 36 pseudogenes. The expression of all novel genes was investigated by PCR screening of seven cDNA libraries. All annotated structures were studied in terms of total sequence coverage and their sequence environment was investigated. Splice sites, polyadenylation signals, isoforms and predicted transcription start sites/promoters/CpG islands are also discussed. The predicted encoded proteins were compared to various proteomes (including human), whereas data from three-species genomic sequence comparisons was used to confirm that virtually all exons and probably all genes in this region have been identified.

A gene/homology-based approach was used to construct a contiguous, 10 Mb long bacterial clone map on mouse chromosome 2 spanning the syntenic region on human 20q12-13.2. A tile path of 66 BAC clones was used to generate approximately 10.3 Mb of sequence. The mouse and human sequences were compared and the distribution of regions showing sequence conservation is discussed in the context of the annotated human sequence. The two syntenic

regions showed strong conservation of gene order and content, but no conservation of human putative genes and pseudogenes was observed within the mouse sequence. Non-exonic conserved sequences and *ab initio* predictions were used to estimate the completion of human annotation.

Human expressed sequences aligned to the annotated exons in 20q12-13.2 were used to identify over 100 exonic SNPs. A set of 2,208 SNPs mapping across the region was used to obtain allele frequencies in three populations (95 Caucasians, 12 Asians and 12 African Americans). A first generation linkage disequilibrium (LD) map of the region was constructed in Caucasians. Over half of the region is covered in "LD blocks", segments with three or more SNPs and for which all possible SNP pairs have D'>0.9.

Acknowledgements

Many people have kindly provided me with their help and advice throughout the course of this project. The most important of these is Panos Deloukas, who has given me guidance and encouragement since the beginning of my time at the Sanger Institute. Thank you for all the time you have spent in helping me to complete this project.

A project of this type inevitably involves many people. I would like to thank Susan Rhodes, Jackie Bye and John Collins, for help with the gene discovery experiments. Thanks also to Ian Mullenger and Lisa French for their advice and guidance during the mouse map construction. A big thanks to Carol Scott, Sarah Hunt and Jilur Ghori, for being always able to help with data analysis and database management. Many thanks to Webb Miller and Francesca Chiaromonte for the help with the comparative studies and Robert Lawrence for the analysis of the genotype data. Many other people from the Sanger Institute have provided me with invaluable assistance and I would particularly like to thank Pam Whittaker, Kate Downes, Tom Dibling, Melanie Goward, Rhian Gwilliam, Elisabeth Huckle and Carol Carder. Thanks should also be extended to people in informatics, especially Jennifer Ashurst, James Gilbert, Thomas Down and Michele Clamp, as well as to all the sequencing team members for generating the human and mouse sequence.

I would like to also acknowledge the helpful discussions and critical reading of this manuscript by Panos Deloukas and Rhian Gwilliam. The transcript map fold out diagram was created by James Gilbert and printed by Richard Summers.

I would also like to thank the residents (past and present) of the G205 office, Ele Holloway, Luk Smink, Andy Mungall and Rhian Gwilliam for sharing an office with me, and putting up with me.

A warm thank you to my dearest friends, Vassilis Koudounas, Marios Pirishis, Thanasis Athanasiou, Stelios Koursaris, Chris Markides, Simoni Ioannou, George Petropoulos, Vangelis Andreakos, Alex Papadatos, Yiannis Iliopoulos and (of course) cousin George for all the good times, home and abroad, and their support.

Finally, I would like especially to thank my parents and sister, for their support and encouragement. Your love and belief in me kept me going. Naki, good luck with your final experiments and the forthcoming write-up!

Table of contents

List of tables

List of figures

List of abbreviations

20ace	chromosome 22 implementation of ACeDB
aa	amino acid
ACeDB	A C. elegans DataBase
AITDs	AutoImmune Thyroid Diseases
ALL	Acute Lymphoblastic Leukaemia
AML	Acute Myeloid Leukaemia
BAC	Bacterial Artificial Chromosome
BLAST	Basic Local Alignment Search Tool
bp	base pair(s)
BSA	Bovine Serum Albumin
cDNA	complementary DNA
CDS	CoDing Sequence
CEPH	Centre d'Etude du Polymorphisme Humain
cM	centiMorgan
CpG	5'CG3' dinucleotide
cR	centiRay
cRSC	coding Region of Sequence Conservation
cSNP	complementary SNP
dATP	2'-deoxyAdenosine 5'-TriPhosphate
dbEST	database of ESTs
dbSNP	database of SNPs
dCTP	2'-deoxyCytidine 5'-TriPhosphate
DDBJ	Dna DataBase of Japan
ddCTP	2', 3'-dideoxyCytidine 5'-TriPhosphate
dGTP	2'-deoxyGuanosine 5'-TriPhosphate
DNA	DeoxyriboNucleic Acid
dsDNA	double strand DNA
DTT	DiThioThreitol
dTTP	2'-deoxyThymidine 5'-TriPhosphate
EDTA	EthyleneDiamineTetraAcetic acid
EMBL	European Molecular Biology Laboratory
ePCR	Electronic PCR
EST	Expressed Sequence Tag
FBS	Fetal Bovine Serum
FEN	Flap EndoNucleases
FISH	Fluorescent In Situ Hybridisation
FMF	Familial Mediterranean Fever
FPC	FingerPrinting Contigs
G-band	Giemsa band
GD	Graves disease
GSS	Genome Survey Sequence
HERV	Human Endogenous RetroVirus-like elements
HGP	Human Genome Project
HGSP	Human Genome Sequencing Project

HSA20	Homo SApiens chromosome 20
HT	Hashimoto's Thyroiditis
iATG	Translation Initiation site
IHGMC	International Human Genome Mapping Consortium
IHGSC	International Human Genome Sequencing Consortium
INSD	International Nucleotide Sequence Databases
ISNPMWG	International SNP Map Working Group
Kb	Kilo base pairs
LINE	Long INterspersed repeat Element
LTR	Long Terminal Repeat
MaLR	Mammalian LTR
Mb	Mega base pairs
MDS	Myelodysplastic Syndromes
MER	Medium Reiterative Repeat
MGC	Mouse Genome Consortium
MGD	Mouse Genome Database
MGSC	Mouse Genome Sequencing Consortium
MIR	Mammalian-wide Interspersed Repeat
MIR	Mammalian-wide Interspersed Repeat
MMU2	Mus MUsculus chromosome 2
MPD	MyeloProliferative Disorders
mRNA	messenger RNA
MS	Mass Spectroscopy
NCBI	National Center for Biotechnology Information
ncRNA	non-coding RNA
NIH	National Institute of Health
nt	nucleotide
OMIM	Online Mendelian Inheritance In Man
ORF	Open Reading Frame
PAC	P1 Artificial Chromosome
PCR	Polymerase Chain Reaction
PIP	Percentage Identity Plot
Q-banding	Quinacrine banding
R	Purine
R-banding	Reverse banding
RCS	Region of Sequence Conservation
RefSNP	Reference SNP
RFLP	Restriction Fragment Length Polymorphism
RH	Radiation Hybrid
RNA	RiboNucleic Acid
RNAi	RNA interference
rRNA	ribosomal RNA
RT-PCR	Reverse Transcription PCR
SAGE	Serial Analysis of Gene Expression
SDS	Sodium Dodecyl Sulphate
SINE	Short INterspersed repeat Element
SNP	Single Nucleotide Polymorphism

snRNA	small nuclear RNA
SRS	Sequence Retrieval System
SSR	Simple Sequence Repeat
STS	Sequence Tagged Site
TIR	Terminal Inverted Repeat
TrEMBL	Translated EMBL
tRNA	transfer RNA
TS site	Transcription Start site
TSC	The Snp Consortium
UTR	UnTranslated Region
VNTR	Variable Number Tandem Repeat
WGS	Whole Genome Shotgun
WWW	World Wide Web
Y	Pyrimidine
YAC	Yeast Artificial Chromosome

Publications arising from this work

Bench A. J., Nacheva E. P., Hood T. L., Holden J. L., French L., Swanton S., Champion K. M., Li J., Whittaker P., ***Stavrides G***., Hunt A. R., Huntly B. J., Campbell L. J., Bentley D. R., Deloukas P., and Green A. R. (2000). Chromosome 20 deletions in myeloid malignancies: reduction of the common deleted region, generation of a PAC/BAC contig and identification of candidate genes. UK Cancer Cytogenetics Group (UKCCG). *Oncogene* **19:** 3902-13.

Deloukas P., Matthews L. H., Ashurst J., Burton J., Gilbert J. G., Jones M., ***Stavrides G***., Almeida J. P., Babbage A. K., Bagguley C. L., Bailey J., Barlow K. F., Bates K. N., Beard L. M., Beare D. M., Beasley O. P., Bird C. P., Blakey S. E., Bridgeman A. M., Brown A. J., Buck D., Burrill W., Butler A. P., Carder C., Carter N. P., Chapman J. C., Clamp M., Clark G., Clark L. N., Clark S. Y., Clee C. M., Clegg S., Cobley V. E., Collier R. E., Connor R., Corby N. R., Coulson A., Coville G. J., Deadman R., Dhami P., Dunn M., Ellington A. G., Frankland J. A., Fraser A., French L., Garner P., Grafham D. V., Griffiths C., Griffiths M. N., Gwilliam R., Hall R. E., Hammond S., Harley J. L., Heath P. D., Ho S., Holden J. L., Howden P. J., Huckle E., Hunt A. R., Hunt S. E., Jekosch K., Johnson C. M., Johnson D., Kay M. P., Kimberley A. M., King A., Knights A., Laird G. K., Lawlor S., Lehvaslaiho M. H., Leversha M., Lloyd C., Lloyd D. M., Lovell J. D., Marsh V. L., Martin S. L., McConnachie L. J., McLay K., McMurray A. A., Milne S., Mistry D., Moore M. J., Mullikin J. C., Nickerson T., Oliver K., Parker A., Patel R., Pearce T. A., Peck A. I., Phillimore B. J., Prathalingam S. R., Plumb R. W., Ramsay H., Rice C. M., Ross M. T., Scott C. E., Sehra H. K., Shownkeen R., Sims S., Skuce C. D., et al. (2001). The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414:** 865-71.

Chapter I

Introduction

1.1 Introduction

The greatest achievement in biology over the past millennium has been the elucidation of the mechanism of heredity. Organisms encapsulate instructions for creating a member of their species in their gametes; these instructions are passed on to a fertilised egg and are then used to give rise to offspring. Although heredity intrigued philosophers since the time of Hippocrates and Aristotle, the absence of a way to probe the physical nature of these phenomena meant that they could do no more than speculate for more than two millennia.

Milestones in understanding the central dogmas of biology include Mendel's observations of genetic dominance and segregation of traits (1865), the discovery of chromosomes (end of the 19th century), the explanation of the biochemical basis of DNA (1953) and the deciphering of the genetic code (1964). Advances in molecular biology and sequencing sparked the genomics revolution and allowed for the first time the systematic study of individual genes and their environment. The almost "unthinkable" was then proposed in 1985: sequence the whole genome and generate a complete catalogue of every human gene. The Human Genome Project (HGP) was en route.

In the nineties, the HGP became a truly international collaboration involving research centres and funding agencies around the world (Bentley, 2000). Framework maps were generated and used to construct detailed physical and gene maps whereas the introduction of improved sequencing technologies dramatically increased sequence output (Collins,

2001). Model organisms were selected for systematic studies of their genetic make-up and new computational methods have been constantly devised to analyse the huge amount of generated data (Collins and Galas, 1993; Collins *et al*., 1998a). The list of completed genomes is constantly growing and includes *Saccharomyces cerevisiae* (Goffeau *et al*., 1996), *Caenorhabditis elegans* (The C. elegans Sequencing Consortium, 1998) and *Drosophila melanogaster* (Adams *et al*., 2000). In addition, the intermediate goal of the HGP has been achieved with the announcement of the completion and initial analysis of the draft human genome sequence (International Human Genome Sequencing Consortium (IHGSC), 2001).

Besides the tremendous progress that has been made in assembling a human reference sequence, more work is required to extract the full information contained in the genome. A high priority is to establish a complete catalogue of every gene, including all alternative splice forms, and accurately determine the structure of each one of them. Establishing and sequencing a comprehensive collection of full-length cDNAs (IHGSC, 2001) is one way to reach this goal but other approaches also have very strong potential, for example genome sequence comparisons to identify conserved sequences. At the computational level, new algorithms are required to efficiently predict genes and gene-related features. Easy data access should also be ensured through improved, user-friendly genome browsers.

Similar to obtaining the sequence, finding all the genes is an additional layer of genetic information. Many more layers are gradually targeted in a systematic way, from gene expression and regulation to protein localisation and function. For example, the

identification of regulatory regions and their study, in combination with expression array technologies, can identify sets of genes that are co-ordinately regulated. An integral part of all these efforts is the ongoing project that aims to complete the catalogue of human sequence variation and identify the common ancestral haplotypes present in the current population. The generated data from the HGP will serve as a toolbox to probe function in a more systematic manner. Genome data will be combined with improved techniques and databases for the global analysis of RNA and protein expression, protein localisation, protein-protein interactions, and chemical inhibition of pathways.

The generated resources from the HGP had an immediate impact on human health by accelerating the identification of human disease genes (roughly 200 disease-associated genes have been discovered using HGP information, Waterston and Sulston, 1998). In the future, understanding the human genome should illuminate the molecular pathogenesis of disorders that are currently poorly understood, and for which treatments are largely empirical and frequently sub-optimal. Deciphering the pathogenic pathways involved in illnesses could provide the greatest opportunity for the development of targeted therapies since the development of antibiotics (Collins and Guttmacher, 2001).

This chapter provides the background for the work described in the rest of this thesis. Genome organisation and mapping are discussed in section 1.2. Section 1.3 describes the features of the human genome. The computational approaches used for collecting, analysing and representing genomic data are discussed in section 1.4. Mapping and sequencing of model organisms (comparative studies) are discussed in 1.5. Section 1.6 describes the most popular methods employed by functional genomics to study

transcriptomes and proteomes. Human variation is discussed in section 1.7. Finally, chromosome 20 and the aims of this thesis are discussed in sections 1.8 and 1.9 respectively.

1.2 Mapping the human genome

1.2.1 The genome

The size of the haploid nuclear genome is estimated to be circa 3,200 Mb of DNA (Morton, 1991) and two copies are present in human somatic cells (note that the emerging sequence data suggests a smaller size. For example, the reported sizes of all finished chromosomes (20, 21, and 22) are smaller than the Morton estimates). Somatic nuclear DNA is organised into 23 pairs of chromosomes. There are 22 autosomes (numbered 1-22) and two sex chromosomes (X and Y, males are XY and females XX). The shape, relative size and distinctive banding pattern produced by various staining techniques, can identify each chromosome cytogenetically.

The basic shape of a chromosome is defined by the position of the centromere. Metacentric chromosomes have their centromere midway between the ends, submetacentric somewhat closer to one end, and acrocentric close to one end. The banding patterns generated by staining reflect the longitudinal structural heterogeneity of each chromosome (Bickmore and Sumner, 1989).

Giemsa (G) and reverse (R) banding are two of the most frequently used techniques for staining chromosomal regions (reviewed in Craig and Bickmore, 1993). The banding pattern reflects the base composition of a genomic region. Differences in base composition have been correlated with variations in gene density, time of replication, density of repeat sequences and chromatin packaging (Table 1.1) (Holmquist, 1992; Craig and Bickmore, 1993).

Table 1.1: Properties of Giemsa (G) and Reverse (R) bands (reproduced from Gardiner, 1995).

G bands	R bands
Stain strongly with Giemsa and quinacrine	Stain weakly with Giemsa and quinacrine
AT rich	GC rich
Chromomeres	Interchromomeres
DNase insensitive	DNase sensitive
Few breakpoints or rearrangements	Most breakpoints or rearrangements
Gene poor	Gene rich
Alu poor	Alu rich
LINE rich	LINE poor
Replicate late	Replicate early

1.2.2 Genome mapping

Various types of maps can be constructed to provide a more detailed view of a genome. Genetic and physical maps can be used to order landmarks along the length of a chromosome. The landmark-based maps serve as a scaffold for anchoring and orientating overlapping genomic clones (contigs) during clone map construction.

1.2.2.1 Genetic maps

Genetic mapping depends on following the segregation of alleles at two or more loci during meiosis. The unit of the genetic map, the Morgan (M), is defined as the length of chromosomal segment that on average undergoes one exchange per individual chromatid strand during meiosis. Over short chromosomal regions, the recombination fraction is directly proportional to the genetic map distance, so that a recombination fraction of 0.01 corresponds to a genetic map distance of 1 cM.

The construction of genetic maps had been limited by the availability of markers until 1980 when Botstein *et al*. observed that DNA sequence variation could provide a large source of polymorphic markers. These markers, known as Restriction Fragment Length Polymorphisms (RFLPs), could detect DNA polymorphisms when hybridised to restriction digests of an individual's DNA. The collection of RFLP markers was soon supplemented by the more informative minisatellites (Jeffreys *et al*., 1985), or variable number of tandem repeats (VNTR, Nakamura *et al*., 1987), which were used to construct low-resolution maps of the human genome (Donis-Keller *et al*., 1987).

Simple sequence repeats (SSRs), or microsatellites, provide another source of DNA polymorphisms. They are widely dispersed throughout eukaryotic genomes, are highly polymorphic (Weber and May, 1989; Weber, 1990) and when converted to Sequence Tagged Sites (STSs) (Olson *et al*., 1989) they can easily be typed using the polymerase chain reaction (PCR) (Saiki *et al*., 1985, 1988; Litt and Luty, 1989; Tautz, 1989). These technical advances aided the construction of human genetic maps of increasingly higher resolution (Hudson *et al*., 1992; Weissenbach *et al*., 1992; Gyapay *et al*., 1994; Murray *et al*., 1994; Dib *et al*., 1996). Genetic maps have also been constructed for a number of model organisms (section 1.5).

The available genetic map data was extensively used during the construction of physical maps for the human chromosomes. In addition, genetic maps continue to have an instrumental role in identifying disease genes. A new generation of genetic map, able to extract most of the inheritance information from human pedigrees, is under construction. The new, Single Nucleotide Polymorphism (SNP)-based map will contain several-fold more markers and genetic analysis will be performed using high-throughput, automated platforms (section 1.7).

1.2.2.2 Radiation hybrid maps

Based on the original approach by Goss and Harris (1975) and initially modified to study single chromosomes (Cox *et al*., 1990), Radiation Hybrid (RH) mapping was applied to study whole genomes (Walter *et al*., 1994; Gyapay *et al*., 1996).

RH cell lines are constructed by fusing lethally irradiated donor cells to recipient rodent cells deficient in a selectable marker. In RH mapping, the presence or absence of an STS-based marker is tested across a panel of radiation hybrids. The further apart two markers are on the chromosome, the more likely they are to be separated by an irradiation induced break, placing the markers on two separate chromosomal fragments. By estimating the frequency of breakage, and thus the distance between markers, it is possible to determine their order (Cox *et al*., 1990). The method combines aspects of both genetic and physical mapping. In contrast to recombination events, the frequency of breaks induced by the irradiation appears be linearly related to physical distance without cold or hot spots of breakage along the chromosome. The unit of map distance is the centiRay (cR) and represents 1% probability of breakage between two markers for a given radiation dosage (that is, the one used to construct the RH panel). Therefore the correlation between cR units and physical distance in bp will differ from one panel to the other and can only be determined by extrapolation.

In humans, RH mapping has been used to produce high-resolution gene maps by integrating genetic and EST-based markers (Schuler *et al*., 1996; Deloukas *et al*., 1998; http://www.sanger.ac.uk/HGP/Rhmap/). Dense RH framework maps were also constructed to assist the assembly of bacterial clone maps (Mungall *et al*., 1996).

1.2.2.3 Yeast artificial chromosome maps

The construction of clone maps is based on ordering and orientating cloned DNA fragments. As a cloning system, Yeast Artificial Chromosomes (YACs) (Burke *et al*.,

1987) can accommodate large DNA fragments (>1 Mb) and were initially used for the construction of chromosome-specific maps (Chumakov *et al*., 1992; Foote *et al*., 1992). Later, YAC contig maps covering most of the human genome were published (Chumakov *et al*., 1995; Hudson *et al*., 1995). Although YACs allow long-range continuity, they are not optimal substrates for a genome project because they suffer from instability and chimerism (Green *et al*., 1991; Nagaraja *et al*., 1994).

1.2.2.4 Bacterial clone maps

Bacterial (BAC) and P1 (PAC) artificial chromosomes (Shizuya *et al*., 1992; Ioannou *et al*., 1994) can accommodate DNA inserts of approximately 200 Kb (~5 times more than cosmids and fosmids (30-45 Kb), Collins and Hohn, 1978; Kim *et al*., 1992). Both PACs and BACs are stable and very few clones have been found so far to contain rearrangements (Shizuya *et al*., 1992; Ioannou *et al*., 1994). Because of the lower (1-2) copy number per cell and the smaller vector size, BACs are favoured over PACs. The available human PAC and BAC libraries (http://www.chori.org/bacpac/) represent more than 60 genome equivalents and are the preferred resource for sequence-ready map construction.

The initial strategy for constructing bacterial clone maps for whole chromosomes consisted of screening bacterial clones using a high density of STSs (15/Mb on average) obtained from framework maps (The Sanger Institute and Washington University Genome Sequencing Center, 1998). The bacterial clones were then assembled into contigs by comparative restriction fingerprint analysis (Coulson *et al*., 1986; Olson *et al*.,

1986; Gregory *et al*., 1997; Marra *et al*., 1997) and landmark content mapping (Green and Olson, 1990). The generated contigs were then extended and joined by chromosome walking. The strategy was slightly modified once a whole genome fingerprint map was assembled (International Human Genome Mapping Consortium (IHGMC), 2001). Clone maps for all human chromosomes are being constructed (Dunham *et al*., 1999; Hattori *et al*., 2000; Deloukas *et al*., 2001; Bentley *et al*., 2001; Tilford *et al*., 2001; Montgomery *et al*., 2001; Brüls *et al*., 2001; IHGMC, 2001).

1.3 Sequencing and the landscape of the human genome

1.3.1 Construction of sequence maps

Genome sequencing is used to construct maps of base pair resolution. The high quality DNA sequence allows the accurate annotation of all genomic features such as repeats, genes and their control elements. Sequence maps can be assembled either by a hierarchical clone-by-clone sequencing approach or a whole genome shotgun (WGS) approach, or a combination of the two.

The hierarchical approach employed by the IHGSC involves three steps: selecting a set of minimally overlapping clones from the map (tiling path), sequencing each clone and assembling the finished clone sequences into an overall genome sequence map (IHGSC, 2001).

Selected clones (predominantly BACs and PACs) are subjected to shotgun sequencing where the cloned DNA is fragmented and 1.4-2.2 Kb-long fragments are sub-cloned into M13 or plasmid vectors (Bankier *et al*., 1987). The subclones are then sequenced by a modified chain termination method (Sanger *et al*., 1977; IHGSC, 2001). A directed manual editing approach follows the automatic assembly of the generated sequence reads into sequence contigs. Any remaining sequence gaps and problems are resolved during "finishing". Finished sequence contains no gaps and is at least 99.99% accurate. Using this approach, the IHGSC aims to generate finished sequence for the euchromatic regions of all human chromosomes by 2003. Three chromosomes, 20, 21 and 22 (Deloukas *et al*., 2001; Hattori *et al*., 2000; Dunham *et al*., 1999) have already reached this gold standard.

The Whole Genome Shotgun (WGS) approach was first used to determine the sequence of *Haemophilus influenzae* (Fleischmann *et al*., 1995) and is routinely used to sequence small genomes. A hybrid approach was used to sequence the significantly larger and more complex genome of *Drosophila melanogaster* (Adams *et al*., 2000). The authors reported that they determined nearly all of the ~120 Mb euchromatic portion of the Drosophila genome using a whole-genome sequencing strategy supported by extensive clone-based sequence and a high-quality BAC clone map. Venter *et al*. (2001) reported the construction of a human genome sequence map using a WGS approach. Plasmid libraries of various sizes were constructed and sequenced, followed by a whole-genome sequence assembly without any prior mapping information. Sequence data generated by the public effort (using the hierarchical approach) was included in the final assembly; this sparked a debate as to how successful was the WGS approach (Waterston *et al*., 2002; Green, 2002; Myers *et al*., 2002).

A hybrid approach has been adopted for sequencing the mouse genome (Mouse Genome Sequencing Consortium, 2001; http://www.sanger.ac.uk/Info/Press/010508.shtml). The initial whole-genome sequence assembly already provides the scientific community with rapid access to the features of the mouse genome. This will be followed by the time-consuming process of clone-by-clone finishing to generate contiguous sequence and remove errors (such as mis-assemblies) that are common in draft sequence assemblies (Deloukas *et al*., 2001).

1.3.2 The genomic landscape-sequence features

1.3.2.1 GC content and isochores

Equilibrium centrifugation in analytical CsCl-density-gradient shows that DNA preparations from warm-blooded vertebrates are characterised by strong compositional heterogeneity, whereas those from cold-blooded vertebrates exhibit a weak heterogeneity (Thiery *et al*., 1976).

Fractionation of human DNA by equilibrium centrifugation in Cs_2SO_4 density gradients in the presence of sequence-specific DNA ligands (for example Ag^+) showed that human DNA is characterised by a very broad GC range. This analysis also indicated that the human genome is a mosaic of long (>200 Kb) DNA segments that have a homogeneous base composition, the isochores (equal regions) (Bernardi *et al*., 1985, 1989). The L1 and L2 isochore families are GC poor and represent about two-thirds of the genome whereas the H1, H2 and H3 isochore families are GC rich and represent the remaining third

(reviewed in Bernardi, 1989). H3 was later subdivided into three increasingly GC rich sub-fractions, $H3^{-}$, H3* and $H3^{+}$ (Saccone *et al*., 1996).

Saccone *et al*., (1993) used chromosome in situ hybridisation to correlate isochores and chromosomal bands. They showed that G-bands essentially consist of L1 and L2 isochores whereas the GC-richest regions of R-bands consist of H1, H2 and H3 isochores.

Analysis of the draft sequence shows an average GC content of 41% and confirms the variation of GC level across the human genome (for example, regions with GC contents of 33.1% and 59.3% have been identified), but rules out a strict notion of isochores as compositionally homogeneous (IHGSC, 2001). Although the genome does contain large regions of distinct GC content, the substantial variation present at many different scales indicates that a more moderate name such as "GC content domains" would be more appropriate (IHGSC, 2001).

1.3.2.2 CpG islands

The CpG dinucleotides are notable because they are usually methylated on the cytosine base and are greatly under-represented in human DNA, occurring at only about one-fifth of the expected frequency (IHGSC, 2001). A process that can lead to CpG suppression was described by Coulondre *et al*. (1978) who demonstrated that in the *Escherichia coli* lacI gene spontaneous base substitution hotspots occur at 5-methylcytosine residues.

Deamination of these residues to uracil and failure of DNA repair mechanisms can result in G:C→A:T transitions.

CpG islands constitute a distinctive fraction of the genome because they contain the dinucleotide CpG at its expected frequency and in the non-methylated form and their GC content is significantly higher than that of non-island DNA. CpG islands are rich in sites for methyl-sensitive restriction enzymes such as *Hpa*II that recognise unmethylated CpG dinucleotides (Bird, 1986). It is estimated that there are approximately 30,000 CpG islands in the haploid human genome (reviewed in Bird, 1987), but higher numbers have also been reported (Antequera and Bird, 1993; Cross and Bird, 1995).

Restriction enzymes were used to experimentally associate CpG islands and human genes (Lindsay and Bird, 1987). Computational sequence analysis of 375 human genes identified predicted CpG islands at the 5' end of all housekeeping genes, whereas 40% of the genes with tissue-specific or limited expression are also associated with islands. Overall, more than half of the genes analysed were associated with islands (Larsen *et al*., 1992). Analysis of a larger gene set suggests that approximately 58% of human coding genes have a CpG island at their start site (Ponger *et al*., 2001). Because of this association, CpG islands can be used as potential gene markers (Cross *et al*., 1994, 1999, 2000).

Analysis of the draft genome identified 50,267 putative CpG islands, of which at least 21,377 reside in repeats. The density of CpG islands varies substantially across

chromosomes but correlates reasonably well with estimates of relative chromosomal gene densities (IHGSC, 2001).

1.3.2.3 Repeats

An early observation was that genome size does not correlate well with organismal complexity (Lewin, 1994). For example *Homo sapiens* has a genome that is 200 times as large as that of *Saccharomyces cerevisiae*, but 200 times as small as that of *Amoeba dubia*. This phenomenon (known as the C-value paradox) was largely resolved with the recognition that genomes can contain a large quantity of repetitive sequence, far in excess of that devoted to protein-coding genes. Analysis of the draft sequence of the human genome revealed that coding sequences comprise less than 5% of the genome whereas repeat sequences account for at least 50% (IHGSC, 2001).

Generally, repeats fall into five classes:

A. Transposon-derived repeats, often referred to as interspersed repeats.

B. Inactive (partially) retroposed copies of cellular genes (including protein-coding genes and small structural RNAs) usually referred to as processed pseudogenes.

C. Simple Sequence Repeats (SSRs), consisting of direct repetitions of relatively short k-mers such as $(A)_n$, $(CA)_n$ or $(CGG)_n$.

D. Segmental duplications, consisting of blocks of around 10-300 Kb that have been copied from one region of the genome to another region.

E. Blocks of tandemly repeated sequences, such as centromeres, telomeres, the short arms of acrocentric chromosomes and ribosomal gene clusters (such regions are not present in the sequence of 20q12-13.2 and will not be discussed).

A. Transposon-derived repeats

Most of human repeat sequence is derived from transposable elements. 45% of the genome sequence is currently identified as such, and it is believed that much of the remaining "unique" sequence also belongs to this class but has diverged too widely to be recognised (IHGSC, 2001).

In humans there are four main types of transposable elements. These four types fall in two classes: DNA transposons (one type of transposable element) and retrotransposons (three types of transposable elements; Prak and Kazazian, 2000).

A.1 DNA transposons

DNA transposons move by excision and reintegration into the genome ***without*** an RNA intermediate. The main characteristics of this type include the Terminal Inverted Repeats (TIRs) that are 10-500 bp long. DNA transposons encode a transposase gene. When the gene is expressed, the transposase binds specifically to the TIRs and catalyses the cutting and pasting of the element. Integration results in a short, constant length duplication of the target site visible as directed repeats flanking the element (Smit and Riggs, 1996). Although DNA transposition is not replicative, it can result in duplication if (i) it moves from a replicated to a still non-replicated part of the genome (Chen *et al*., 1992) or (ii) the gap resulting from the excision of the transposon is repaired by using the sister chromatid as template (Engels *et al*., 1990).

The human genome contains at least seven major classes of DNA transposons, which can be subdivided into many families with independent origins. DNA transposons cannot exercise a cis-preference. The transposase is produced in the cytoplasm and when it returns to the nucleus it cannot distinguish active from inactive elements. As inactive elements accumulate in the genome, transposition becomes less efficient. This controls the expansion of any DNA transposon family and in due course causes it to die out (IHGSC, 2001).

A.2-4 Retrotransposons

Retrotransposons are duplicated through an RNA intermediate; the original transposon is maintained *in situ*, where it is transcribed. Its RNA transcript is then reverse transcribed into DNA and integrated into a new genomic location.

Long terminal repeat retrotransposons contain partly overlapping regions for group-specific antigen (gag), protease (prt), polymerase (pol) and envelope (env) genes. They are flanked on both ends by Long Terminal Repeats (LTRs) with promoter activity. The transcript is reverse transcribed in a cytoplasmic virus-like particle, primed by a tRNA.

Although a variety of LTR retrotransposons exist, only the vertebrate-specific endogenous retroviruses (ERVs) appear to have been active in the mammalian genome. Mammalian retroviruses fall into three classes (I-III), each comprising many families with independent origins. Most (85%) of the LTR retrotransposon-derived fossils consist only of an isolated LTR, with the internal sequence being lost by homologous recombination between the flanking LTRs (IHGSC, 2001).

Long interspersed elements (LINEs). Three distantly related LINE families are found in the human genome: LINE1, LINE2 and LINE3. Of these only LINE1 is active (IHGSC, 2001). 3,000-4,000 human LINE1s are full length, and of those only 40-60 are estimated to be active (Sassaman *et al*., 1997). In contrast the mouse genome has an estimated 3,000 active LINE1s (DeBerardinis *et al*., 1998).

A full-length (6.1 Kb) L1 element consists of a 5' untranslated region (5' UTR) that has a promoter activity; 2 open reading frames (ORF1 and ORF2) are separated by an intergenic spacer, followed by a 3'UTR and a polyA tail. ORF1 encodes a 40 KDa protein that binds nucleic acids, whereas ORF2 contains reverse transcriptase (RT), an endonuclease domain (EN) and a cysteine-rich region (C). Genomic L1 insertions are often flanked by 7-20 nucleotide target-site duplications (TSDs) (Prak and Kazazian, 2000). Note that at the stage where the reverse transcriptase uses the nicked DNA to prime reverse transcription from the 3' end of the LINE RNA, the enzyme frequently fails to reach the 5' end, resulting in many truncated, non-functional insertions (Smit, 1996; IHGSC, 2001).

Three distinct families of **short interspersed elements (SINEs)** are found in the human genome: the active Alu family, and the inactive MIR and Ther/MIR3 families. SINEs are characterised by an internal polymerase III promoter that ensures a fair chance for transcriptional activity of new copies (Smit, 1996). SINEs do not code for any proteins.

MIRs (mammalian-wide interspersed repeats) are approximately 260 bp long. Both families are tRNA-derived interspersed repeats that are believed to have spread through

the genome before the mammalian radiation (Jurka *et al*., 1995). MIR elements are the most mammalian-wide interspersed SINEs and are thought to be the most ancient mammalian SINE family (Rogozin *et al*., 2000; Smit and Riggs, 1995).

The most abundant and thoroughly studied SINEs are those belonging to the Alu family. Alus are named after the *Alu*I restriction site they carry (Houck *et al*., 1979; Gu *et al*., 2000). Alu repeats were derived from the signal recognition particle component 7SL (the 300 bp 7SL RNA is an essential component of the Signal Recognition Particle (SRP), which mediates the translocation of secretory proteins across the endoplasmic reticulum (Mighell *et al*., 1997)). A typical human Alu is an approximately 300 bp head-to-tail dimer of 7SL RNA-derived elements. The left monomer has significant similarity with a RNA pol III promoter; an A-rich linker connects right and left monomers. (Rogozin *et al*., 2000).

Based on the presence of diagnostic nucleotide substitutions, Alus are divided into three branches, which are further classified into sub-branches reflecting the age of individual elements from oldest (J), to intermediate (S), to youngest (Y) (Gu *et al*., 2000). The intermediate and older subfamilies have a significant amount of heterogeneity and there are many examples of intermediates between these various subfamilies. Thus, this categorisation is not exhaustive and is simply used to provide a reasonable working nomenclature for these older subfamilies (Batzer *et al*., 1996).

The AluJ repeats are divided into the Jo and Jb sub-branches and it is estimated that they were introduced to the genome 50 to 80 million years ago. The AluS repeats are divided

into the Sq, Sp, Sx, Sc, Sg, and Sg1 sub-branches. It is estimated that they were introduced to the genome 35 million years ago (Jurka and Milosavljevic, 1991; Gu *et al*., 2000).

The AluY repeats (Y, Ya5, Ya8, and Yb8) date back only to 20 million years ago. (Mighell *et al*., 1997; Gu *et al*., 2000). All Alu repeats presently known to retropose differ from the Y subfamily consensus sequence by only a few additional diagnostic mutations. This suggests that the youngest subfamilies of Alu repeats were ancestrally derived from the Y subfamily. Therefore, young subfamily sub-branches are defined as lineages that descended from this gold standard (Batzer *et al*., 1996).

LINE elements have been proposed to be the main generators of Alus. LINEs are thought to mobilise Alus because of the similarity of their target site duplications and the similarity of their insertion sites (the DNA nick for Alu insertions is probably made by L1 endonuclease). This parasitism of LINEs by SINEs remains difficult to reconcile with the observation that LINEs seem to insert preferentially into AT rich regions, whereas SINEs such as Alus accumulate in GC regions. One theory suggests that Alu elements integrate randomly but those that are actively transcribed (and are therefore more likely to reside in GC rich regions of the genome) are more likely to become fixed in the population. This explanation predicts that Alu RNA may have some advantageous function (Smit *et al*., 1999; Prak and Kazazian, 2000).

Possible roles for the transposon-derived repeats include the use of their regulatory elements by genes and regulation of protein translation. LINE-mediated transduction

mobilises DNA sequences around the genome, whereas at least 43 genes probably derived form DNA transposons (IHGSC, 2001). The ability of LINE retrotransposons to cause reverse transcription of genic mRNA can give rise to processed pseudogenes (Esnault *et al*., 2000). The possible role of these elements in the evolution of species suggests that it is important to further understand and evaluate their functional impact (Kass, 2001).

B. Processed pseudogenes

Pseudogenes were first identified in the *Xenopus* laevis genome (Jacq *et al*., 1977). The main characteristics of pseudogenes are the close similarity they have to one or more paralogous genes and the fact that most are non-functional due to failure of either transcription or translation (Mighell *et al*., 2000). Pseudogenes arise either by retrotransposition or duplication of genomic DNA. Pseudogenes that arise by retrotransposition are called processed pseudogenes and their main characteristics include the lack of introns and 5' promoter sequences and the presence of a 3' polyA tail. They are also flanked by short target site repeats (<15 bp) (Maestre *et al*., 1995).

Although pseudogenes are generally considered as defective entities, examples have been described where they retain their open reading frames (Vanin, 1985). Transcribing pseudogenes have been identified and it is postulated that some may also be functional (examples reviewed in Mighell *et al*., 2000; Makalowski, 2000). In addition, pseudogenes can potentially be copied to generate further pseudogenes.

C. Simple sequence repeats

SSRs are a common feature in the human genome. They are perfect or slightly imperfect tandem repeats of a particular k-mer. SSRs with a short repeat unit (n=1-13 bp) are called microsatellites, whereas those with longer repeat units (n=14-500 bp) are called minisatellites. SSRs comprise about 3% of the genome. The biggest contribution is by dinucleotide repeats (0.5% of the genome; IHGSC, 2001).

Of the twelve equivalence classes of triplet repeats (64 classes, twelve taking into account reverse complement and shift), three (CAG, CGG, GAA) have been associated with triplet disease disorders (Baldi and Baisnee, 2000). Expansions of unstable trinucleotide repeats have been associated with at least fifteen inherited neurologic diseases (Lieberman and Fischbeck, 2000).

D. Segmental duplications

Segmental duplications involve the transfer of blocks of sequence (1-200 Kb) to one or more locations of the genome. Interchromosomal duplications involve blocks of sequence duplicated among non-homologous chromosomes whereas intrachromosomal duplications involve blocks of sequence duplicated within a particular chromosome or chromosomal arm. Recombination between duplicons leads to chromosomal rearrangements that could lead to genomic disorders (Ji *et al*., 2000).

1.3.2.4 Genes

The task of nuclear gene transcription is shared by three RNA polymerases (pol I, pol II and pol III). Pol I synthesises rRNA and pol III makes 5S rRNA, tRNA, 7SL RNA, U6 snRNA and a few other small stable RNAs. In contrast, the huge variety of protein-coding genes is transcribed by pol II. So, although pol I and pol III account for 80% of total RNA synthesis (reviewed in Paule and White, 2000), only pol II-transcribed genes will be discussed.

Human protein-coding DNA sequences have complex structures (Figure 1.1), typically segmented by intervening sequences (Tilghman *et al*., 1978a, 1978b), called introns (Gilbert, 1978). A large ribonucleoprotein complex, the spliceosome, recognises sites at the 5' and 3' ends of introns (the donor and acceptor sites respectively), as well as an internal site, the branch point, and removes introns from gene transcripts (Moore and Sharp, 1993). With a few exceptions (Sharp and Burge, 1997; Burset *et al*., 2000), nearly all spliceosomal introns begin with GT and end with AG (donor and acceptor sites, respectively). The retained segments, called exons, form the messenger RNA (mRNA). Differential removal of RNA sequences from gene transcripts of a particular gene (alternative splicing) generates isoforms that can encode for protein variants. EST-based studies estimate that 35-38% of human genes undergo alternative splicing (Mironov *et al*., 1999; Brett *et al*., 2000). Higher numbers of alternatively spliced genes have also been reported (Kan *et al*., 2001; IHGSC, 2001).

The discovery of introns sparked a debate regarding their origin and evolution (reviewed in Long *et al*., 1995; Logsdon *et al*., 1995). The fragmentation of protein-coding genes by

introns may have conferred an advantage, by facilitating the modular shuffling of eukaryotic protein domains in evolutionary time and in real time via alternative splicing (Mattick, 2001).

Other features of protein-coding genes include a translation start site (usually ATG), often contained in an optimal consensus sequence (Kozak, 1987) and in most cases, a highly conserved hexamer (the polyA signal; Kessler *et al*., 1986), which is involved in the polyadenylation of the RNA transcript. A significant fraction of genes display multiple polyadenylation sites (Gautheret *et al*., 1998) and the patterns of variant polyA signals used in this process are currently under study (Beaudoing *et al*., 2000).

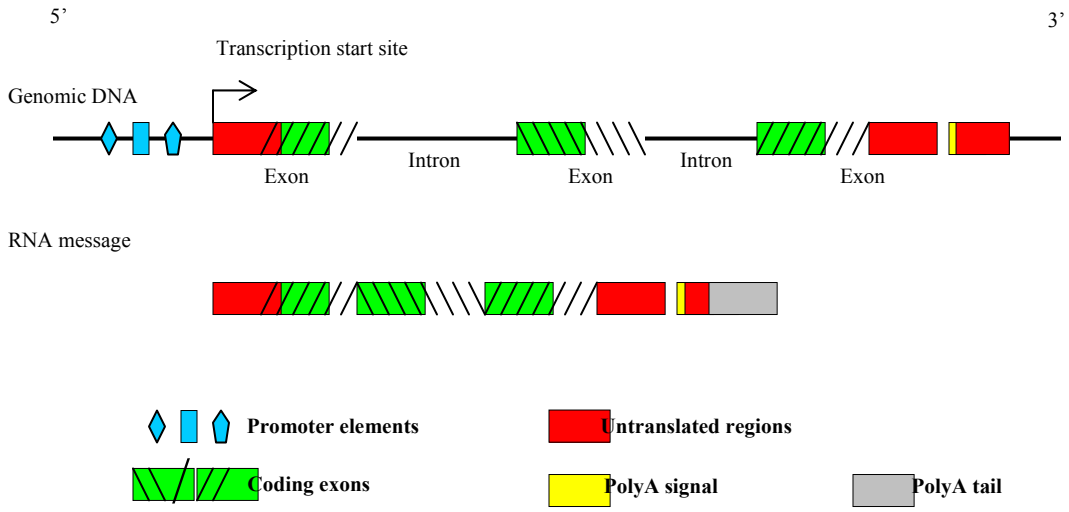


Figure 1.1: Basic gene structure. The organisation of a typical gene locus is shown on top whereas the resulting messenger RNA is shown below. The coloured boxes represent various gene features.

Promoter sequences upstream of the site of transcriptional initiation of protein-coding genes bind the basal transcriptional machinery (and additional potentiating factors) and specify the transcription start (TS) site (reviewed in Fickett and Hatzigeorgiou, 1997; Dillon and Sabbattini, 2000). The complex biochemical processes that govern promoter recognition, binding and control of transcription are currently under intense investigation. In addition to promoters, other cis-acting regulatory sequences could also be present upstream, downstream, or within a gene locus (reviewed in Fraser and Grosveld, 1998; Li *et al*., 1999; Lee and Young, 2000).

Identification of transcribing sequences in genomic DNA is a recurrent problem in transcript mapping and positional cloning studies. Traditional techniques employed to address this problem include using CpG islands as positional signposts for the starts of some transcription units (Lindsay and Bird, 1987), 'zoo' blots to detect cross-species conserved genomic sequences that could represent genes (Monaco *et al*., 1986) and hybridisation of entire genomic clones directly to cDNA libraries (Elvin *et al*., 1990; methods reviewed in Gardiner and Mural, 1995).

A variation of the latter is cDNA selection where the genomic sequence of interest is immobilised on a nylon filter and hybridised to an amplified cDNA library. The hybridised cDNA sequences are then eluted and re-amplified using PCR. The process can be repeated to achieve further enrichment (Parimoo *et al*., 1991; Lovett *et al*., 1991). The method was improved by using biotin-labelled genomic DNA and streptavidin-coated magnetic beads to capture the DNA-cDNA hybrids (Korn *et al*., 1992; Morgan *et al*.,

1992). The cDNA selection method was successfully used to identify expressed sequences for specific chromosomes (Touchman *et al*., 1997).

The exon trapping/amplification technique (Duyk *et al*., 1990; Buckler *et al*., 1991) is based on the selection of exonic sequences that are flanked by functional 5' and 3' splice sites. Fragments of genomic DNA are cloned into a vector flanked by 5' and 3' splice sites. The constructs are used to transfect COS-7 cells and the resulting RNA transcripts are processed *in vivo*. Splice sites of exons contained within the inserted genomic fragment are paired with those of the flanking intron. The resulting mRNA contains the previously unidentified exons that can then be PCR-amplified and cloned. Exon trapping has been used to identify candidate disease genes (Vulpe *et al*., 1993; Trofatter *et al*., 1993) and exons from entire chromosomes (Church *et al*., 1993; Trofatter *et al*., 1995). Unfortunately, the methods described above are technically challenging and time-consuming which makes them unsuitable for analysing large genomic regions (Gardiner and Mural, 1995).

Currently, large-scale sequence analysis relies on DNA and protein similarity searches. Extensive collections of Expressed Sequence Tags (ESTs) from human (Adams *et al*., 1991; Wilcox *et al*., 1991), mouse (Marra *et al*., 1999), rat and other model organisms have become available (Boguski *et al*., 1993). Additional resources include the mRNAs of known genes and a large number of anonymous cDNA clone sequences produced by different centres (KIAA collection, http://www.kazusa.or.jp, Nomura *et al*., 1994; RIKEN collection, http://www.riken.go.jp/, The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium, 2001; Genoscope collection,

http://www.Genoscope.cns.fr; DKFZ collection, http://mbi.dkfz-heidelberg.de/, Wiemann *et al*., 2001). In addition, sequence comparisons between species at the genomic level are utilised to identify conserved regions, which may represent coding exons.

Ab initio methods are also used to complement the homology-based analysis results. These algorithms can predict coding sequences, promoters, transcription start sites and CpG islands (section 1.4). A disadvantage of this approach is the need for experimental confirmation of the various predictions (see chapter III).

Published estimates regarding the total number of human coding genes vary substantially. For example, Crollius *et al*. (2000a) used a human:*Tetraodon nigroviridis* comparative sequence approach and estimated that the human genome contains 28,000-34,000 genes. Human EST analyses by other groups suggest that the total number is approximately 120,000 (Liang *et al*., 2000). The Chromosome 22 Sequencing Consortium estimated a minimum of 45,000 genes based on their annotation of the complete chromosome although their data suggests that there may be additional genes (Dunham *et al*., 1999). Other whole-chromosome studies suggest that the gene number may be closer to 40,000 (Hattori *et al*., 2000), 35,000 (Ewing and Green, 2000) or 31,500 (Deloukas *et al*., 2001). Two independent analyses of the draft genome sequence suggest a total of 39,114 (Venter *et al*., 2001) and 31,778 (IHGSC, 2001) but the minimal overlap between the novel genes of the two sets casts doubts on these gene estimates (Hogenesch *et al*., 2001).

The “gene guessing game” (Fields *et al*., 1994; Dunham, 2000; Aparicio, 2000; Simpson *et al*., 2001) and the recent consensus of there being less than 40,000 genes in the human

genome raises questions regarding the origins of human species complexity (Claverie, 2001). The structure and control architecture of genes are probably at the heart of eukaryotic complexity and phenotypic variation (Mattick, 2001).

1.4 Computational genomics (Bioinformatics)

The "tidal wave of data" (Reichhardt, 1999) caused by the Human Genome Project gave birth to bioinformatics, the computer-assisted data management discipline that helps the gathering, analysis and representation of genomic information (Persidis, 1999, 2000).

1.4.1 Sequence databases

DNA sequences are made publicly available through the International Nucleotide Sequence Databases (INSD) that consist of GenBank (http://www.ncbi.nlm.nih.gov/Genbank/index.html; Benson *et al*., 2002), EMBL (http://www.ebi.ac.uk/embl/; Stoesser *et al*., 2002) and DDBJ (http://www.ddbj.nig.ac.jp; Tateno *et al*., 2002). Data is exchanged between the three sites on a daily basis to ensure that each maintains a comprehensive collection of sequence information. The amount of sequence data stored continues to grow at an exponential rate and more than 105,000 different species are represented in the databases (Benson, 2002; 121,736 as of May 2002, http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=howmany).

A large number of other DNA databases are also maintained, including the dbEST database (http://www.ncbi.nlm.nih.gov/dbEST/, Boguski *et al*., 1993), the Unigene database (http://www.ncbi.nlm.nih.gov/UniGene/index.html), the Reference Sequence Database (RefSeq; http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html, Pruitt and Maglott, 2001) and the database of Single Nucleotide Polymorphisms (dbSNP; http://www.ncbi.nlm.nih.gov/SNP/, Sherry *et al*., 2001). LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/, Pruitt and Maglott, 2001) provides summary pages for each gene and links to other relevant databases.

Annotated (curated) protein sequences are stored in SWISS-PROT (http://www.ebi.ac.uk/swissprot/) whereas TrEMBL contains EMBL nucleotide translated sequences (http://www.ebi.ac.uk/trembl/index.html; Bairoch *et al*., 2000). InterPro (http://www.ebi.ac.uk/interpro/index.html; Apweiler *et al*., 2000) provides an integrated documentation resource for protein families, domains and functional sites.

Data is freely accessible to the scientific community via web-based, integrated database retrieval systems such as the Sequence Retrieval System (SRS; http://srs.ebi.ac.uk/; Zdobnov *et al*., 2002) and Entrez (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi; Wheeler *et al*., 2002).

1.4.2 Sequence analysis

In addition to text-based searches, tools such as BLAST (Altschul *et al*., 1990, 1997; Table 1.2) enable sequence-based data mining. Sequence comparisons at the nucleotide

and/or amino acid level are performed to identify homologous sequences. Alignment of such sequences using software tools such as CLUSTAL W (Thompson *et al*., 1994) can be used to predict structural, functional and evolutionary relationships.

Table 1.2: Blast types (reproduced from Brenner, 1998).

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

New homology search tools such as SSAHA (Ning *et al*., 2001), Exonerate (Slater, unpublished) and BLAT (Kent, 2002) can be used to perform fast searches of databases containing multiple gigabases of DNA, whereas software such as PipMaker (Schwartz *et al*., 2000), Vista (Mayor *et al*., 2000), GLASS (Batzoglou *et al*., 2000) and SynPlot (Göttgens *et al*., 2001) are available for performing comparative sequence analyses (reviewed in Miller, 2001).

A program widely used for genomic analysis is RepeatMasker (Smit and Green, unpublished). This program scans sequences to identify full-length and partial members of all known repeat families represented in Repbase (Jurka, 2000). RepeatMasker analysis of the draft genome sequence has shown that at least 50% corresponds to repeat

elements (IHGSC, 2001). So, masking repeats before performing homology searches is an important step that filters out spurious matches.

Software programs that have been developed to predict sequence features are summarised in Table 1.3. Various evaluations of exon/gene prediction programs indicate that each has its individual strengths and weaknesses that are reflected by its sensitivity and specificity scores (Burset and Guigo, 1996; Claverie, 1997; Guigo *et al*., 2000; Rogic *et al*., 2001). Currently, none of the available software can correctly predict all protein-coding genes in a given sequence, although more reliable predictions can be obtained by the combined use of software (Murakami and Takagi, 1998).

Computational analysis of polymerase II promoters can contribute to gene identification. In 1997, Fickett and Hatzigeorgiou reviewed the available prediction programs and concluded that on average they report one false positive per Kb. Although this may not be a problem when analysing short genomic sequences, analysis of a whole chromosome or genome would produce too many false positives. Since then, new software tools have been developed. PromoterInspector (Scherf *et al*., 2000) and Eponine (Down and Hubbard, 2002) have a sensitivity of 43% and 40% respectively and nearly 40% of their predictions correspond to true promoters (Scherf *et al*., 2001; Deloukas *et al*., 2001). Other available software include CPGFIND (Micklem, unpublished) for CpG island prediction (which are associated with the 5' end of genes) and a newly developed algorithm, FirstEF, which according to Davuluri *et al*. (2001), predicts 86% of first exons with only 17% false positives.

Table 1.3: Overview of the main sequence-feature prediction programs.

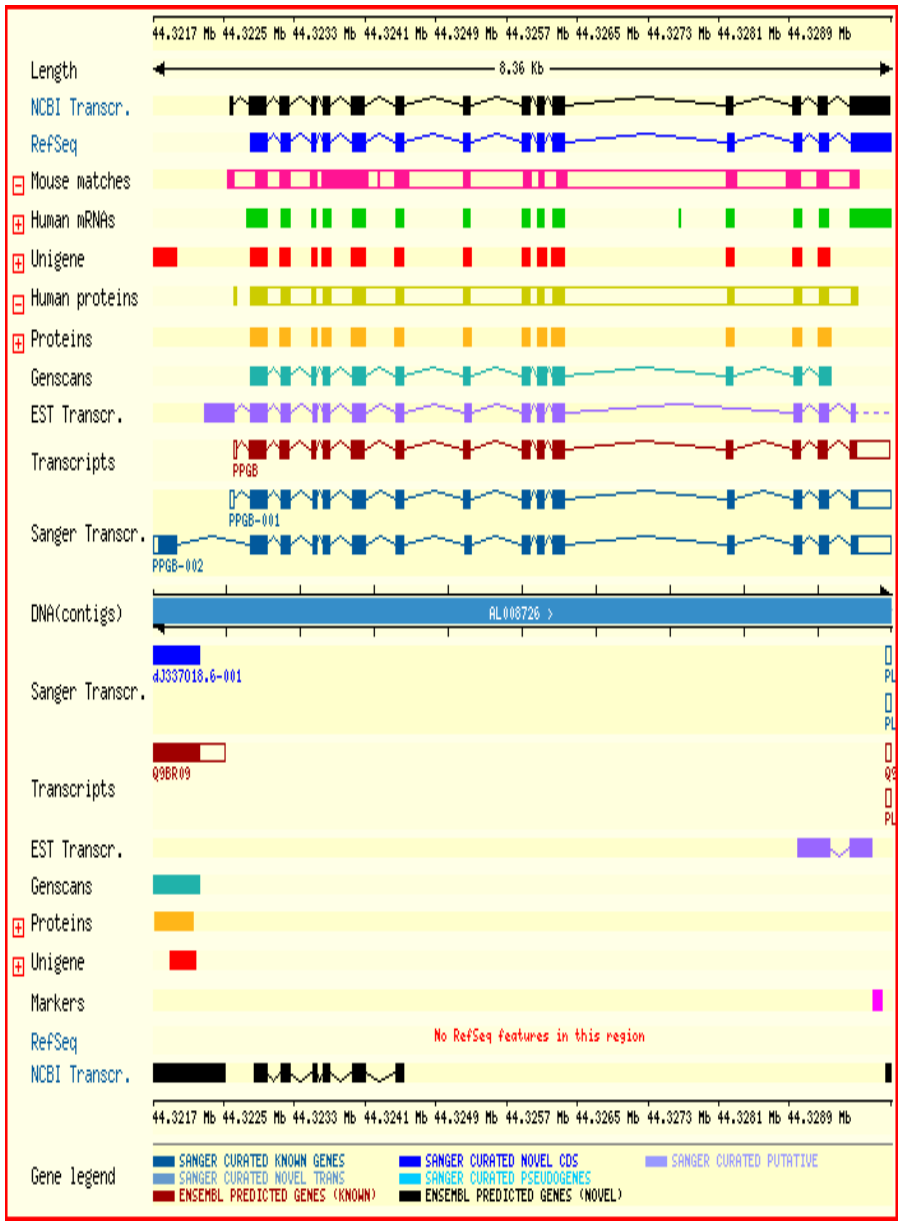
Program	Description	Reference
Genscan	Gene prediction	Burge and Karlin, 1997
FGENESH	Gene prediction	Salamov and Solovyev, 2000
Hexon	Exon prediction	Solovyev *et al*., 1994
GRAIL	Exon prediction	Uberbacher *et al*., 1996
Xpound	Exon prediction	Thomas and Skolnick, 1994
PromoterInspector	Promoter prediction	Scherf *et al*., 2000
Eponine	TS site prediction	Down and Hubbard, 2002
FirstEF	First exon prediction	Davuluri *et al*., 2001
CPGFIND	CpG island prediction	Micklem, unpublished
RepeatMasker	Repeat sequences prediction	Smit and Green, unpublished

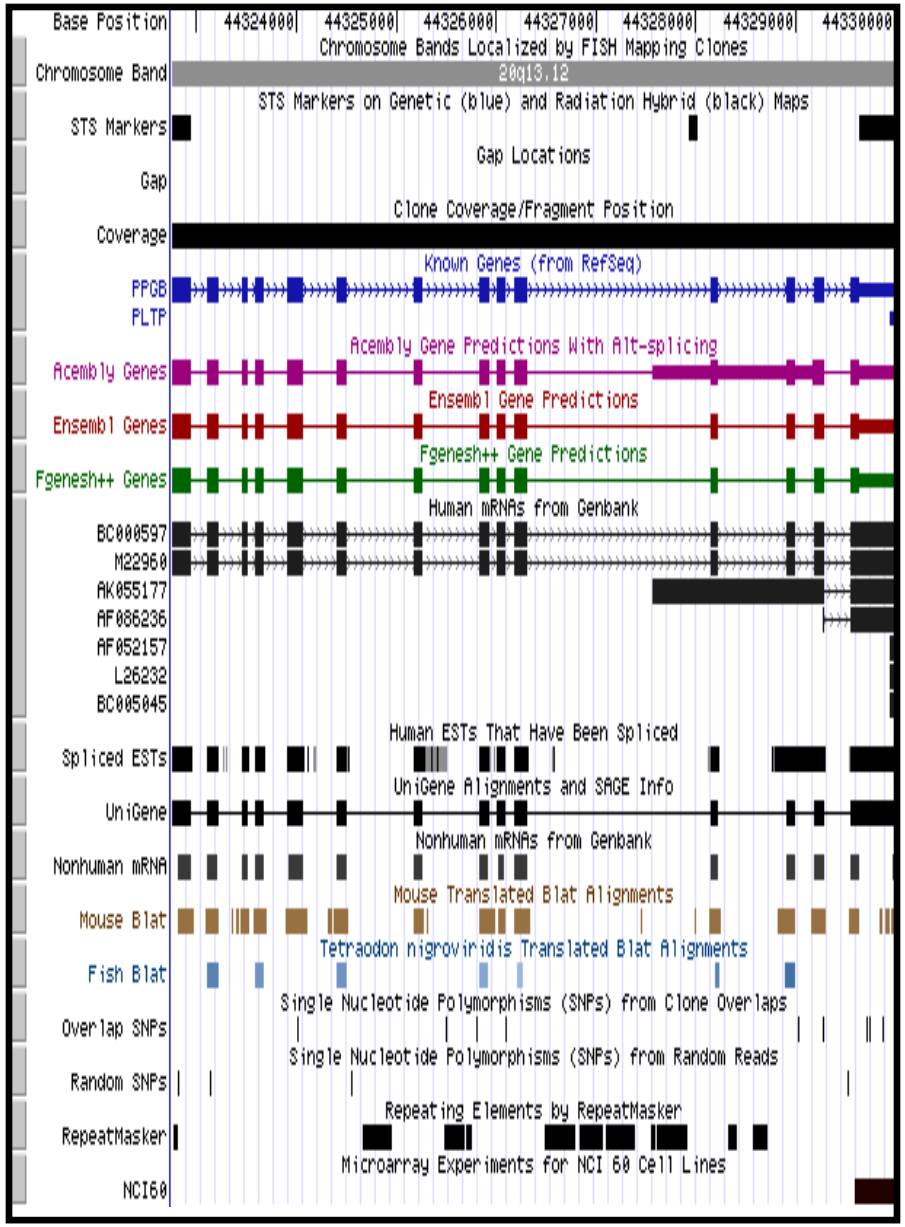
1.4.3 Viewing genomic information

Individual laboratories use different software to represent genomic information. For example, ACeDB, which was originally developed for the *Caenorhabditis elegans* community (A C. elegans DataBase; Durbin and Thierry-Mieg, 1994), is extensively used at the Sanger Institute. ACeDB is used for graphical display and browsing of biological data such as DNA/peptide sequence and annotation, map data and hybridisation data (Kelley, 2000). Several of the human chromosome projects rely on ACeDB for data management (http://www.sanger.ac.uk/HGP/Humana/).

Ensembl (http://www.ensembl.org/) is one of the leading sources of automated genome sequence annotation. Ensembl provides a bioinformatics framework to organise biology around the sequences of large genomes. Annotation of gene features is performed automatically using confirmed gene predictions that have been integrated with external data resources (Hubbard *et al*., 2002). The other main genome browser is based at the University of California Santa Cruz (UCSC) (Kent *et al*., 2002; http://genome.ucsc.edu/). Like Ensembl, the UCSC genome browser displays a variety of information, including assembly contigs and gaps, mRNA and expressed sequence tag alignments, gene predictions, cross-species homologies, single nucleotide polymorphisms, sequence tagged sites and repeat elements. An example of how the sequence features are displayed by the Ensembl and the UCSC genome browser is shown in Figure 1.2.

Figure 1.2 (next page): The Ensembl (red box) and UCSC (black box) genome browsers. In this example both browsers display the sequence features of the PPGB locus.


44.3217 Mb 44.3225 Mb 44.3233 Mb 44.3241 Mb 44.3249 Mb 44.3257 Mb 44.3265 Mb 44.3273 Mb 44.3281 Mb 44.3289 Mb
Length
8.36 Kb
NCBI Transcr.
RefSeq
Mouse matches
Human mRNAs
Unigene
Human proteins
Proteins
Genscans
EST Transcr.
Transcripts
PPGB
PPGB-001
Sanger Transcr.
PPGB-002
DNA(contigs)
AL008726 >
Sanger Transcr.
dJ337018.6-001
Transcripts
Q9BR09
EST Transcr.
Genscans
Proteins
Unigene
Markers
RefSeq
No RefSeq features in this region
NCBI Transcr.
Gene legend
SANGER CURATED KNOWN GENES
SANGER CURATED NOVEL TRANS
ENSEMBL PREDICTED GENES (KNOWN)
SANGER CURATED NOVEL CDS
SANGER CURATED PSEUDOGENES
ENSEMBL PREDICTED GENES (NOVEL)
SANGER CURATED PUTATIVE


Base Position
44324000 44325000 44326000 44327000 44328000 44329000 44330000
Chromosome Bands Localized by FISH Mapping Clones
Chromosome Band
20q13.12
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
STS Markers
Gap Locations
Gap
Clone Coverage/Fragment Position
Coverage
Known Genes (from RefSeq)
PPGB
PLTP
Acembly Gene Predictions With Alt-splicing
Acembly Genes
Ensembl Gene Predictions
Ensembl Genes
Fgenesh++ Gene Predictions
Fgenesh++ Genes
Human mRNAs from Genbank
BC000597
M22960
AK055177
AF086236
AF052157
L26232
BC005045
Human ESTs That Have Been Spliced
Spliced ESTs
UniGene Alignments and SAGE Info
UniGene
Nonhuman mRNAs from Genbank
Nonhuman mRNA
Mouse Translated Blat Alignments
Mouse Blat
Tetraodon nigroviridis Translated Blat Alignments
Fish Blat
Single Nucleotide Polymorphisms (SNPs) from Clone Overlaps
Overlap SNPs
Single Nucleotide Polymorphisms (SNPs) from Random Reads
Random SNPs
Repeating Elements by RepeatMasker
RepeatMasker
Microarray Experiments for NCI 60 Cell Lines
NCI60

1.5 Comparative genomics

Much of the power of molecular genetics arises from the ability to isolate and study genes from one species based on knowledge about related genes in other species. Comparisons between genomes that are distantly related provide insight into the universality of biological mechanisms and identify experimental models for studying complex processes. Furthermore, comparisons between genomes that are closely related provide unique insights into the details of gene structure and function (Collins *et al*., 1998a).

Five model organisms, *Escherichia coli*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Mus musculus* were targeted by the first five-year plan for the Human Genome Project (1993-1998; Collins and Galas, 1993). The genomes of these organisms can provide vital insights regarding the function and organisation of the human genome. For example, the genome of a unicellular organism could reveal a minimal set of proteins required for life.

Genomic analysis of multicellular organisms can help unravel the complex developmental pathways, whereas the mouse genomic sequence can be used to highlight conserved functional features such as genes and regulatory elements (Watson, 1990; Collins *et al*., 1998a). Of the five model organisms initially proposed, four have been fully sequenced so far (table 1.4), whereas the mouse genome sequencing project is still in progress.

Table 1.4: Genome sizes of the model organisms initially proposed. Current protein numbers and a summary of InterPro analysis are also shown (as of January 2002).

Organism	Genome size (Mb)	Number of proteins	Proteins with InterPro matches
E. coli[1]	4.6	4,363	71.2 %
S. cerevisiae[2]	12	6,123	65.6 %
C. elegans[3]	97	18,940	67.9 %
D. melanogaster[4]	120	13,958	71.6 %
M. musculus	3,000 (est)	15,856	72.6 %

[1]Blattner *et al*., 1997
[2]Goffeau *et al*., 1996
[3]The C. elegans sequencing consortium, 1998
[4]Adams *et al*., 2000

Initial sequence analysis of the *Escherichia coli* genome revealed that protein-coding genes account for approximately 88% of the genome. Insertion sequence (transposable) elements, phage remnants and many other patches of unusual composition were also identified, indicating genome plasticity through horizontal transfer (Blattner *et al*., 1997).

In *Saccharomyces cerevisiae* only 4% of protein-coding genes contain introns (Goffeau *et al*., 1996) and so Open Reading Frames (ORFs) can be easily identified. The functions of individual ORFs can be evaluated readily because it is relatively easy to disrupt them *in vivo* (Rothstein, 1991; Burns *et al*., 1994). A variety of approaches are used to analyse the yeast sequence on a genome-wide scale (Bassett *et al*., 1996; Ross-MacDonald *et al*., 1999). LacZ/transposon insertion-based approaches were used to perform large-scale analysis of gene expression, protein localisation and gene disruption (Burns *et al*., 1994; Ross-MacDonald *et al*., 1999). A PCR-based approach was also developed to perform targeted ORF deletions (Baudin *et al*., 1993). This was later refined by adding two unique

20 bp sequences in each deletion construct that can serve as molecular barcodes for the strain carrying them (Shoemaker *et al*., 1996). Other methods used to study the yeast genome and proteome include genetic footprinting to assess the phenotypic effects of induced Ty1 transposon insertions (Smith *et al*., 1996), serial analysis of gene expression (SAGE; Velculescu *et al*., 1997), microarray analysis (DeRisi *et al*., 1997), two-dimensional (2D) gel/protein identification analysis (Maillet *et al*., 1996), two-hybrid system analysis (Fromont-Racine *et al*., 1997) and genome-wide protein tagging (Martzen *et al*., 1999).

The worm *Caenorhabditis elegans* is an ideal animal to study basic gene functions. It is fully transparent at all stages of its life, allowing cell divisions, migrations, and differentiation to be monitored in live animals. Its anatomy is simple, yet the 959 somatic cells of the adult represent most major differentiated tissue types. In addition it is the only animal for which the complete neuronal wiring pattern is known (reviewed in Ahringer, 1997). On average, the amino acid (aa) similarity and identity between aligned human and *Caenorhabditis elegans* orthologous gene products are 69.3% and 49.1% respectively, and the nucleotide identity is 49.8% (Wheelan *et al*., 1999). *Caenorhabditis elegans* genes can be studied by generating knockouts using a PCR/Tc1-transposon-based approach (Collins *et al*., 1987; Mori *et al*., 1988; Rushforth *et al*., 1993; Zwaal *et al*., 1993). Gene expression patterns can be determined using high resolution FISH to visualise mRNA distributions at the cellular and sub-cellular level (Birchall *et al*., 1995), and manipulated using double-stranded RNA interference (Fire *et al*., 1998).

RNA interference (RNAi) describes the use of double-stranded RNA (dsRNA) to target mRNAs for degradation. When dsRNA is injected into worms the RNAi machinery uses the sequence information in the dsRNA to generate a protein-RNA complex that destroys the corresponding mRNA. This new approach led to the development of new strategies for blocking gene function, which have been successfully applied to silence worm and fly genes (reviewed in Schmid *et al*., 2002; Bargmann, 2001; Zamore, 2001).

In terms of evolutionary sequence conservation, the fruit fly *Drosophila melanogaster* is the closest of the invertebrate model organisms to humans (Sidow and Thomas, 1994). It has been studied for more than 80 years and numerous publications are dedicated to describing its genetics and hundreds of individual genes (reviewed in Rubin, 1996). The fruit fly provides a powerful system to study the function of conserved genes because any of its ORFs can be mutated and subjected to detailed functional analysis within the context of an intact organism. An ongoing gene-disruption project establishes mutant strains that each contains a single, genetically engineered P transposable element in a defined genomic region (Spradling *et al*., 1995; Deák *et al*., 1997).

The mouse has served over the past century as an excellent experimental system for studying mammalian genetics and physiology (Dietrich *et al*., 1995). A number of detailed mouse genetic and physical maps have been constructed (discussed in section 4.1.1), whilst in May 2001, the Mouse Sequencing Consortium reported a first sequence draft of the mouse genome (http://www.sanger.ac.uk/Info/Press/010508.shtml). Detailed human:mouse comparative studies at distinct genomic regions are used to identify coding regions and regulatory elements (discussed in section 4.1.2). Methods such as Mb-long

chromosomal rearrangements (Mills and Bradley, 2001) and tagged random mutagenesis (Zambrowicz *et al*., 1998) in embryonic stem cells are used to study gene function and generate mammalian models for developmental processes and cancer.

In 1998(a), Collins *et al*. suggested the selection of additional model organisms. For example, the vertebrate fish *Danio rerio* (zebrafish) is an excellent genetic system for the identification and functional analysis of genes that control pattern formation and organogenesis. Zebrafish complements other experimental systems, since information gained from analyses of its functionally important genes can readily be extended to homologous systems in mice and other organisms (Driever *et al*., 1994; Talbot and Hopkins, 2000). Genetic screens have identified mutations in hundreds of genes with essential functions in the zebrafish embryo (Dreiever *et al*., 1996; Haffter *et al*., 1996). Genetic linkage (Shimoda *et al*., 1999; Gates *et al*., 1999; Kelly *et al*., 2000) and radiation hybrid (Geisler *et al*., 1999) maps have been constructed, followed by a whole-genome human:zebrafish comparative map (Woods *et al*., 2000). A hybrid approach to sequence the zebrafish genome is underway at the Sanger Institute.

The genome of the pufferfish *Fugu rubripes* (Fugu) is only four times larger than that of *Caenorhabditis elegans* and it was shown that a random genomic sequence is 7.5 times more likely to be coding than a random human sequence (Brenner *et al*., 1993). This compact genome provides a simple and economic approach to compare sequence data from mammals and fish and could enable the identification of essential conserved elements because of the large evolutionary divergence (Elgar *et al*., 1996). The presence (or not) of synteny between human and Fugu sparked a debate on whether Fugu is a good

model organism (Gilley *et al*., 1997; Aparicio and Brenner, 1997; Elgar *et al*., 1997). Miles *et al*. (1998) reported extensive conservation of synteny between a 1.5 Mb region of human chromosome 11 and <100 Kb of the Fugu genome, but a comparative study of seven genes on human chromosome 9 revealed extensive gene order differences within regions of conserved synteny (Gilley and Fried, 1999). A recent study based on chromosome 20 genes identified considerable conservation of synteny, but not good conservation of gene order (Smith *et al*., 2002).

Tetraodon nigroviridis is a freshwater pufferfish 20-30 million years distant from *Fugu rubripes* (Roest Crollius *et al*., 2000b). It has been suggested that studying a species related to Fugu but distant by 20-30 million years would enable the identification of functionally important sequences that appeared after the human/teleostan divergence (Crnogorac-Jurcevic *et al*., 1997). Human:*Tetraodon nigroviridis* comparative analysis was used to provide an estimate for the total human gene number (Crollius *et al*., 2000a), whilst similar analysis was used to estimate the completion of annotation of human chromosome 20 (Deloukas *et al*., 2001). No genetics are available in *Tetraodon nigroviridis*.

The laboratory rat is one of the most important animal models for the genetic mapping of complex phenotypes (James and Lindpaintner, 1997). Over the past ten years approximately 150 genetic loci controlling multifactorial traits have been identified in rats (http://ratmap.gen.gu.se). Rat genomic resources include genetic, RH and EST maps (Bihoreau *et al*., 1997; Watanabe *et al*., 1999; McCarthy *et al*., 2000; Scheetz *et al*., 2001; Bihoreau *et al*., 2001), as well as rat:human:mouse comparative maps (Summers *et*

al., 2001; Kwitek *et al*., 2001). Shotgun sequencing of the rat genome is undertaken at the Baylor College of Medicine (http://www.hgsc.bcm.tmc.edu/projects/rat/) in collaboration with other institutes. DNA sequence alignments of transcribing human:rat orthologous sequences indicate that untranslated exons share on average >68% identity. The mean aligned identity of human/rat coding sequences is 85.9% and the mean aligned identity of human/rat proteins is 88% (Makalowski and Boguski, 1998).

1.6 Functional genomics

Although genome sequence analyses provide a wealth of information on predicted gene products, the majority of these have no known function (Blackstock and Weir, 1999). The generated resources and data can be used to develop and apply global (genome-wide) experimental approaches to assess gene function. This approach (referred to as functional genomics; Hieter and Boguski, 1997) promises to rapidly narrow the gap between sequence and function and provide a molecular understanding of biological processes (Thornton, 2001).

A number of powerful approaches are available to monitor the RNA (transcriptome) and protein (proteome) molecules in a cell. For example, differential-display-reverse transcription PCR (DDRT-PCR) can be used to identify and compare mRNAs during different cell processes. In DDRT-PCR, RNAs isolated from different cell populations are reverse transcribed with a set of degenerate, anchored oligo(dT) primers to generate cDNA pools. This is followed by PCR amplification and labelling using the original

primer and a degenerate one. The products can be visualised on a sequencing gel and differentially expressed molecules can be isolated for further studies (Liang and Pardee, 1992; refinements and modifications reviewed in Liang and Pardee, 1995; Matz and Lukyanov, 1998; Sturtevant, 2000).

Serial analysis of gene expression (SAGE; Velculescu *et al*., 1995) allows the quantitative and simultaneous analysis of a large number of transcripts. Double-stranded cDNA is digested, ligated to linkers and amplified. The linkers contain a restriction site for a type II restriction enzyme that cuts DNA twenty nucleotides away from the recognition site, which is used to digest the cDNA pool. 13-20 bp-long tagged cDNAs are then ligated to generate a library of clones, each representing twenty or more tagged genes. The expression profile can be obtained by sequencing each clone.

Microarray technology provides a format for the simultaneous measurement of the expression level of thousands of genes in a single hybridisation assay. Each array consists of a reproducible pattern of thousands of different DNAs (primarily PCR products or oligonucleotides) attached to a solid support, usually glass. Fluorescently labelled DNA or RNA is hybridised to complementary DNA on the array and signals are detected by laser scanning. Hybridisation intensities for each arrayed DNA sequence are determined using an automated process and converted to a quantitative read-out of relative gene expression levels. The data can then be further analysed to identify expression patterns and variation and to correlate with cellular development, physiology and function (reviewed in Harrington *et al*., 2000; Noordwier and Warren, 2001; Lee and Lee, 2000).

Work describing microarray studies of gene expression across the entire yeast genome were first reported in 1997 (De Risi *et al*., 1997). Since then, numerous transcription profiles in various conditions have been generated (Devaux *et al*., 2001). A similar approach has been used to create a reference database of 300 full-genome expression profiles in yeast corresponding to mutations in ORFs, as well as treatments with compounds with known molecular targets (Hughes *et al*., 2000). The generated profiles identified many co-regulated sets of genes, allowing dissection of transcriptional responses and isolation of candidate genes for many cellular processes.

cDNA microarrays were used to characterise gene expression patterns in 49 mouse tissues. Clustering genes coding for known enzymes into metabolic pathways revealed coordination of expression within each pathway among different tissues (Miki *et al*., 2001). Other applications of the DNA microarray technology include the identification of genome-wide locations and functions of DNA binding proteins (Ren *et al*., 2000; White, 2001) and the experimental annotation of the human genome (Shoemaker *et al*., 2001).

Currently, proteomic analysis relies on a limited number of techniques. These include 2D-gel electrophoresis and mass spectrometry (MS) to separate and analyse thousands of proteins, and ICAT (isotope-coded affinity tag)/MS technology for qualitative and quantitative comparisons of complex protein mixtures. Applications such as the yeast two-hybrid system and phage display are also used to study protein-protein interactions (proteomic analysis methods reviewed in Pandey and Mann, 2000; Dutt and Lee, 2000; Legrain and Selig, 2000; Martin and Nelson, 2001; Yaspo, 2001; Lee, 2001).

1.7 Human variation

The central aim of genetics is to correlate specific molecular variation with phenotypic changes by exploiting the polymorphic nature of the genome. The human genome sequence is not one sequence but rather many variations on a common theme, each of which alters the inherent molecular circuitry, and thus, consequent phenotypes, in a specific manner (Chakravarti, 1999).

Since the mid-1980s sequence variations such as RFLPs (Botstein *et* al., 1980), minisatellites (Jeffreys *et* al., 1985) and microsatellites (Weber and May, 1989) (also see section 1.2.2.1) have been successfully used in genome-wide linkage and positional cloning analyses to identify hundreds of genes for human diseases (Collins, 1995; http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM). Unfortunately, most successes in finding genes that contribute to disease risk have been for highly penetrant diseases, caused by single genes. The identification of genes involved in complex diseases requires the use of dense genetic maps that could be obtained through the systematic discovery, analysis and characterisation of SNPs (Collins *et al*., 1998b; International SNP Map Working Group (ISNPMWG), 2001).

SNPs are single base-pair substitutions in genomic DNA at which different alleles exist with a frequency of at least 1% in one or more populations (Brookes, 1999; Table 1.5). About 90% of sequence variants in humans are SNPs (Collins *et al*., 1998b) and comparison of two haploid genomes yields one SNP per 1,331 bp (ISNPMWG, 2001).

Analysis of all chromosomes from 40 individuals is expected to identify 17 million SNPs, of which 500,000 will map within coding regions (Collins *et al*., 1998b).

Table 1.5: Essential SNP facts (Risch, 2000; Taylor *et al*., 2001; Collins *et al*., 1998b).

- Defined by a frequency of >1% in at least one population
- Stable inheritance
- Building block of haplotypes
- Estimated density of 1 in ~2 Kb (when 2 chromosomes are compared)
- Bi-allelic, suitable for high-throughput analysis
- Topological classification
  Coding amino acid change (non-synonymous, non-conservative aa change)
  Coding amino acid change (non-synonymous, conservative aa change)
  No change in amino acid (synonymous coding)
  Non-coding (5' or 3' UTR)
  Other non-coding (intra- and inter-genic regions)
- More stable, more numerous and easier to score than microsatellite repeat variants

1.7.1 SNP identification

SNP discovery and characterisation is a multi-step process (Kwok and Gu, 1999). Sequences such as random STS and/or EST sequences can be used for identifying candidate SNPs (Wang *et al*., 1998a; Gu *et al*., 1998; Deutsch *et al*., 2001; Picoult-Newberg *et al*., 1999; Irizarry *et al*., 2000). More systematic approaches were centred on the emerging sequence of the human genome. For example, candidate SNPs were identified by analysing the sequence overlaps from the clone tile paths (Dawson *et al*., 2001; Taillon-Miller *et al*., 1998), or by reduced representation shotgun (RRS) sequencing (Altshuler *et al*., 2000; Mullikin *et al*., 2000). In the RRS approach, pooled

DNA from a group of individuals is digested with restriction enzyme(s) and size fractionated on an agarose gel. Fragments approximately 1.5 Kb in size are isolated and used to construct a small insert library that is shotgun sequenced to 2-5-fold redundant coverage. The sequence traces are then aligned and compared for mismatches (SNPs). The sequence reads can also be aligned to the reference sequence of the human genome to identify additional SNPs.

In February 2001 the ISNPMWG published a map of the human genome containing 1.42 million SNPs (one SNP every ~1.9 Kb of available sequence). The SNP Consortium (TSC, http://snp.cshl.org/; Marshall, 1999) and the HGP identified over 95% of the SNPs by using the RRS sequencing approach and the sequence from overlapping clones, respectively (ISNPMWG, 2001).

The most comprehensive public repository of SNPs is the dbSNP database (http://www.ncbi.nlm.nih.gov/SNP/index.html; Sherry *et al*., 2001; Barnes, 2002). Established in 1998, dbSNP currently contains 4.2 million SNPs (Build 104 - May 2002) that can be grouped into a non-redundant set of 2.6 million SNPs (RefSNPs; http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi). dbSNP is currently integrated with other large public variation databases such as the NCI CGAP-GAI database of EST-derived SNPs (http://lpgws.nci.nih.gov/; Masood, 1999), the TSC (http://snp.cshl.org/; Masood, 1999) and HGBASE (http://hgbase.cgr.ki.se; Brookes *et al*., 2000).

Validation studies on a small subset of TSC and HGP candidate SNPs showed that more than 80% are polymorphic and that 50% have a minor allele frequency of ≥20% (Marth

et al., 2001; ISNPMWG, 2001). In a recent publication Kruglyak and Nickerson (2001) suggest that the 2001 SNP map comprises 11-12% of all human sequence variations. In addition, they estimate that obtaining a nearly complete catalogue (95%) of human polymorphic sites (allele frequency >1%) would require a comparison between 96 haploid genomes and highlight the fact that the biggest challenge lying ahead is not SNP discovery, but SNP analysis (genotyping) (Kruglyak and Nickerson, 2001).

1.7.2 SNP analysis

Most current SNP analysis methods rely on PCR amplification of the sequence of interest, which is then tested for the presence, or absence of the polymorphism using an assay system. The multitude of assay systems in use are described in Landegren *et al*. (1998), Gut (2001), Jenkins and Gibson (2002), Syvänen (2001), and summarised in Figure 1.3. Some examples are also discussed below.

The most straightforward gel-based assay exploits the introduction/removal of a restriction enzyme site at the polymorphic site. Primers designed on either site of the SNP are used to PCR amplify the sequence of interest. Restriction digest and differential migration on an agarose gel by electrophoresis of the PCR product indicates the presence or absence of the restriction site in the DNA sample tested (Dawson *et al*., 2001). Alternatively, the presence of a SNP in a PCR product can be determined by the Sanger method of sequencing (Wang *et al*., 1998a; Taillon-Miller *et al*., 1998; Mullikin *et al*., 2000; Deutsch *et al*., 2001; ISNPMWG, 2001).

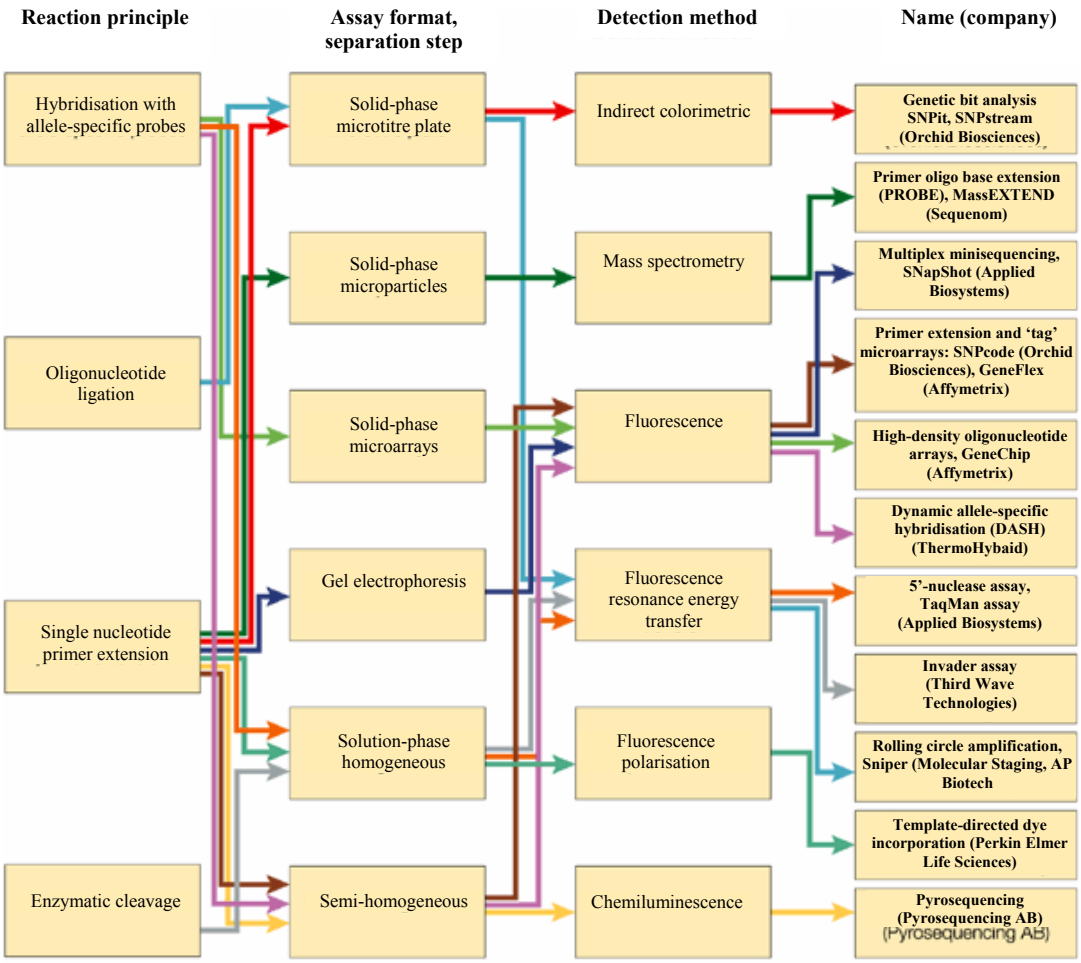


Figure 1.3 (reproduced from Syvänen, 2001): "Modular" design of some of the assays for SNP genotyping. Coloured arrows are used to show the reaction principles, assay format and detection methods that make up a particular genotyping method. For example, the TaqMan™ assay involves hybridisation with allele-specific oligonucleotides, a solution-phase assay and detection by fluorescence resonance energy transfer. The figure illustrates principles for assay design, and the list of assays is not intended to be comprehensive.

The Invader assay was used to systematically analyse SNPs during a whole-chromosome-22 study (Dawson *et al*., 2002). This method utilises the invasive cleavage of oligonucleotide probes (Lyamichev *et al*., 1999) in a two-stage reaction set-up. The first step involves the hybridisation of two target-specific hybridisation oligonucleotides, an allele-specific signalling probe with a 5'-region that is non-complementary to the target sequence and an upstream Invader oligonucleotide. When the allele-specific probe is

perfectly matched at the SNP, the three-dimensional structure formed by the oligonucleotides and the target sequence at the SNP is recognised and is cleaved by a 5'-endonuclease, called FLAP endonuclease, which is specific for this structure. The cleavage releases the 5'-sequence of the signalling probe, which can be detected using fluorescence- or mass spectrometry-based detection.

The TaqMan™ and Molecular Beacon approaches employ Allele-Specific Oligonucleotide (ASO) hybridisation coupled to fluorescence detection. They are both based on an energy transfer principle in which fluorescence is detected as a result of a change in physical distance between a reported fluorophore and a quencher molecule.

In the TaqMan™ (or 5' nuclease allelic discrimination) method, the region flanking the polymorphism is amplified in the presence of allele-specific probes (labelled with a different fluorophore). Probes have a fluorophore, called reporter, at the 5' end and a quencher at the 3' end that absorbs fluorescence from the reporter (Livak *et al*., 1995). During the extension phase of PCR, the DNA polymerase encounters probes specifically base-paired with its target and unwinds them. The 5'→3' exonuclease activity of the DNA polymerase degrades the partially unwound probes and liberates the reporter fluorophore from the quencher, thereby increasing net fluorescence. Mismatched probes are displaced from the target without degradation (Livak, 1999; Holloway *et al*., 1999; Ranade *et al*., 2001).

Molecular Beacon probes consist of sequence that is complementary to the target sequence and a short stretch of self-complementary 5' and 3' nucleotides with a

fluorophore at the 5' end and a quencher at the 3' end. When not hybridised to a target sequence the probes adopt a stem-loop conformation, bringing the fluorophore and quencher pair close together, thereby extinguishing the donor fluorescence. When the probes hybridise to a perfectly matched target during the primer-annealing phase of PCR, the stem-loop structure opens, and the distance between the quencher and fluorophore increases, resulting in a 900-fold increase in fluorescence. Fluorescence from mismatched probes is quenched because they readily adopt the stem-loop structure, enabling allele discrimination (Tyagi and Kramer, 1996; Tyagi *et al*., 1998; Marras *et al*., 1999).

Arrays of oligonucleotides (DNA chips) can also be used to detect human variation in a number of different ways (Hacia and Collins 1999). For example SNP detection can be achieved by carrying out multiplex ASO reactions on microarrays that carry several probes for each SNP to be analysed (Hacia *et al*., 1998). Alternatively, hybridised targets can be used as templates for the extension of the immobilised probes (mini-sequencing; Pastinen *et al*., 1997; Pastinen *et al*., 2000).

The mass spectrometry-based approaches are discussed in section 5.1.

1.7.3 Utilising SNP data

There are great hopes that SNP data can be used to improve biological understanding and to advance medicine (Isaksson *et al*., 2000). Shen *et al*. (1999) used structural mapping and structure-based targeting strategies to show that SNPs can have marked effects on the structural folds of mRNAs. These results suggest that phenotypic consequences of SNPs

could arise from mechanisms that involve allele-specific structural motifs in mRNA, which could influence a diverse range of events such as mRNA splicing, processing, translational control and regulation.

SNPs that map in coding exons can also affect protein function, if they result in amino acid substitutions (non-synonymous cSNPs). The impact of such an amino acid replacement on the three-dimensional structure and function of a protein can be predicted computationally with methods that rely on protein structure information from previous research, and/or the structural context of known disease-causing mis-sense mutations and/or the amino acid similarity with homologous proteins (Chasman and Adams, 2001; Sunyaev *et al*., 2001; Wang and Moult, 2001; Ng and Henikoff, 2002). Such analyses predict that (i) the SNP discovery efforts have discovered several hundred of nsSNPs (ii) in the genome of the average human there are hundreds of nsSNPs that have a direct effect on protein function (a subset of which could impact on human health).

Pharmacogenetics (the study of how genetic differences influence the variability in patients' responses to drugs) holds great promise for the optimisation of new drug development and the individualisation of clinical therapeutics. It is predicted that the use of pharmacogenetics in pre-marketing clinical trials will enable a greater percentage of those trials to produce significant results, because patients whose genetic profile suggests that the drug will be harmful or ineffective to them will be intentionally excluded (Pfost *et al*., 2000). In addition, physicians will be able to use genetic testing to predict the patient's response to a drug, which can aid in individual dosing of medications or avoidance of side effects (Roses, 2000; Pfost *et al*., 2000; Chakravarti, 2001).

The most challenging role envisaged for SNPs is their use for the identification of genetic factors in common disease. For example, SNPs can be used as genetic markers in linkage equilibrium studies (Kruglyak, 1997; Carlson *et al*., 2001). Simulations by Kruglyak (1997) indicate that a map of 700-900 moderately polymorphic SNPs is equivalent to, and a map of 1,500-3,000 superior to, a 300-400 microsatellite marker set. More importantly, they can also be used in association studies of complex diseases to test whether a SNP is enriched in patients compared to suitable controls (Risch and Merikangas, 1996; Risch, 2000; Gray *et al*., 2000; Nowotny *et al*., 2001).

Marker (SNP) selection is critical to the success of such studies. The densities of SNPs to be used depend on the haplotype features of the region they map to. Haplotype information is obtained by determining the combinations of SNPs that are inherited together on the same DNA strand using linkage disequilibrium (LD) analysis.

LD analysis measures the degree of association between two genetic markers. The extent of LD in the human genome is still under debate, mainly because of lack of experimental evidence. The emerging experimental evidence suggests that the human genome can be parsed objectively into haplotype blocks: sizeable regions over which there is little evidence of historical recombination, and within which only a few common haplotypes are observed. It is also suggested that the boundaries of blocks and specific haplotypes they contain are highly correlated across populations (Patil *et al*., 2001; Olivier *et al*., 2001; Gabriel *et al*., 2002). LD will be discussed in more detail in section 5.1.

1.8 Chromosome 20

Chromosome 20 is a metacentric chromosome and represents ~2.2% of the human genome (Morton, 1991). The clone map, which was assembled by fingerprinting and STS content analysis (Bentley *et al*., 2001), is in six contigs of which one spans the entire p arm. The four gaps in the q arm have all been sized by fibre FISH and together account for less than 320 Kb of DNA. 59,421,637 bp of non-redundant sequence generated from 629 overlapping clones represents 99.4% of euchromatic DNA (Deloukas *et al*., 2001).

The finished sequence was analysed on a clone-by-clone basis using a combination of similarity searches and *ab initio* gene predictions, followed by manual annotation of 895 gene features. Excluding pseudogenes, chromosome 20 has a gene density of 12.18 genes/Mb, which is intermediate to 6.71 (low) and 16.31 (high) reported for chromosome 21 and 22, respectively. 32,763 unique SNPs have been identified on chromosome 20. Comparative analysis of chromosome 20 to mouse and *Tetraodon nigroviridis* genomic sequences indicates that the current analysis may account for over 95% of all coding exons and almost all genes (Deloukas *et al*., 2001). The above analysis included data generated by this study (discussed in chapter III).

The best-known disorders linked to chromosome 20 (in total 46 are reported by OMIM (May 2002), http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM) are Creutzfeldt-Jacob disease (Collinge *et al*., 1996) and severe combined immunodeficiency (Wiginton *et al*., 1986). The resources generated by the Chromosome 20 Project accelerated the cloning of disease genes such as those involved in the Alagille (JAG1; Li *et al*., 1997),

the McKusick-Kaufman (MKKS; Stone *et al*., 2000), the ICF (DNMT3B; Hansen *et al*., 1999) and the Hallervorden-Spatz (PANK2; Zhou *et al*., 2001) syndromes. In addition, the candidate disease loci for myeloproliferative disorders, myelodyplastic syndromes and congenital hereditary endothelial dystrophy were refined (Bench *et al*., 2000; Ebenezer *et al*., unpublished). Other disorders linked to chromosome 20 for which the underlying genetic defect is still unknown include type 2 diabetes, cataract, obesity and Graves disease.

The studies described in this thesis (chapters III, IV, and V) are centred on the region of 20q12-13.2, which is linked to many characterised and uncharacterised candidate disease loci. Deletion of the long arm of chromosome 20 represents the most common chromosomal abnormality associated with the myeloproliferative disorders (MPD) and is also found in other malignancies including myelodysplastic syndromes (MDS) and acute myeloid leukaemia (AML) (Bench *et al*., 2000). Earlier studies identified a common deleted region in 20q11.2-13.1 spanning circa 11 Mb (Wang *et al*., 1998b; Bench *et al*., 1998; Bench *et al*., 2000; Wang *et al*., 2000). MDS/AML and MPD associated with a del(20q) are distinct diseases, raising the possibility that multiple target genes on 20q are involved in the pathogenesis of these neoplastic disorders (Wang *et al*., 2000). Acute lymphoblastic leukaemia (ALL) is the most common malignancy in childhood and has also been mapped to 20q (Chambon-Pautas *et al*., 1998; Couque *et al*., 1999).

Autoimmune disease occurs when the immune system mounts a response directed against self (Gough, 2000). The autoimmune thyroid diseases (AITDs) including Graves disease (GD) and Hashimoto's thyroiditis (HT) are the commonest human autoimmune diseases

and are responsible for significant morbidity in pre-menopausal women. The pathogenesis of both diseases appears to develop a result of a complex interaction between predisposing genes and environmental triggers (Tomer and Davies, 1997; Tomer *et al*, 1999). A whole genome linkage study of families with AITD was used to identify three loci that are linked with GD, one of which (GD2) was identified on 20q11.2 (Tomer *et al*., 1998; Tomer *et al*., 1999; Barbesino *et al*., 1998; Pearce *et al*., 1999). GD2 showed the strongest evidence for linkage to GD and it was fine-mapped to a 1 cM interval between markers D20S107 and D20S108 (Tomer *et al*., 1999).

Diabetes is a heterogeneous disorder characterised by a chronic elevation of plasma glucose and its mode of inheritance remains unclear (Rich, 1990). Several studies reported evidence for linkage on chromosome 20q for type 2 diabetes (Ji *et al*., 1997; Bowden *et al*., 1997; Zouali *et al*., 1997; Ghosh *et al*., 1999; Ghosh *et al*., 2000).

The interest in identifying and characterising disease-causing genes prompted the construction of several physical maps of the region. Physical YAC/PAC/BAC clone maps (such as those described in Bench *et al*., 1998; Wang *et al*., 1998b; Price *et al*., 1999; Wang *et al*., 2000) were constructed and used to position genes utilising the available information from resources such as the Gene Map (Deloukas *et al*., 1998). The chromosome 20 mapping effort (Bentley *et al*., 2001) produced a PAC/BAC contiguous map across the region of 20q12-13.2. A set of 111 minimally overlapping clones was selected and sequenced as part of the chromosome 20 sequencing project (Deloukas *et al*., 2001).

1.9 This thesis

The raw DNA sequence is the output of large-scale genome sequencing. This thesis aims to show ways of increasing our knowledge about the genetic information content of this product in order to maximise the impact of the Human Genome Project on genome biology. A 10 Mb region on human chromosome 20q12-13.2 (representing 1/6th of the whole chromosome) provided the basis for the following studies:

- Computational and experimental annotation of sequence features (chapter III). Three experimental approaches were used to confirm and/or extend gene structures and the advantages and disadvantages of each approach are discussed. All annotated features were studied in terms of total sequence coverage (e.g. exon sizes). Splice sites and isoforms, as well as first-pass expression data obtained for the studied transcripts are also discussed. Estimates for gene annotation completion were obtained using different computational approaches and the translated annotated coding features were compared to the proteome of other species. The sequence environment of gene features was also investigated.

- Human:mouse comparative studies (chapter IV). This chapter describes the construction of a bacterial clone map for a mouse chromosome 2 region syntenic to human 20q12-13.2, using a homology-based, gene-orientated approach. The mouse map was used to select a sequence tile path and the annotation and analysis of the generated mouse sequence is described. The human:mouse comparative sequence analysis findings are also discussed.

- Human variation across 20q12-13.2 (chapter V). In this final results chapter, I discuss the analysis of human variation across the region. The use of a data mining approach to identify more than a hundred novel cSNPs is described. Circa 2,000 candidate SNPs were selected for genotyping across three populations (Caucasians, Asians and African Americans). The genotypes obtained from 95 Caucasian individuals (from twelve CEPH families) were then used to generate a first generation linkage disequilibrium map of the region.

Chapter II

Materials and Methods

2.1 Gene identification

2.1.1 DNA manipulation methods

2.1.1.1 Polymerase chain reaction

PCR reactions were performed in 96-well microtitre plates (Costar Thermowell^{TM} C- or M-type) in an Omnigene (Hybaid) (C-type) or a PTC-225 (MJ Research) (M-type). For most applications, 15 µl reactions were prepared.

1. A premix sufficient for all reactions was prepared, allowing for a 1 X reaction mix once the DNA template was added (10 µl of mix and 5 µl of template).
2. The standard reaction mix contained 2 µl of Buffer 1, 2 µl of 5 mM dNTPs, 0.28 µl of 1/20th βME, 0.07 µl of 5 mg/ml BSA, 5.165 µl of 40% sucrose, 0.325 µl of primer mix (each primer at 100 ng/µl) and 0.16 µl (0.8 U) of Taq DNA polymerase (AmpliTaq). All reaction mixture variants are listed in Table 2.1.
3. Amplifications were performed under the following cycling profile (unless specified otherwise): 94^{o}C for 5 minutes, 35 cycles at 94^{o}C for 30 seconds, annealing temperature (specific to each primer) for 30 seconds and 72^{o}C for 3 minutes, and finally 1 cycle at 72^{o}C for 10 minutes. All cycling reaction mixtures used are listed in Table 2.1.
4. Reaction products were visualised by agarose gel electrophoresis and stained with ethidium bromide (section 2.1.1.2).

DNA templates

The templates used were:

1. cDNA pools.
2. DNA excised from agarose gels in 100 µl $T_{0.1}E$ and left overnight; 5 µl used directly.
3. Human genomic DNA at 12.5 ng/µl.

2.1.1.2 Gel electrophoresis

1. An agarose gel was prepared (2.5% for most PCR amplified products and 1% for fragments over 1 Kb) in 1 X TBE and ethidium bromide (250 ng/µl).
2. PCR reaction products were loaded directly. For purified DNA samples, the appropriate amount of 6 X loading buffer was added prior to loading (e.g. 5 µl of purified DNA and 1 µl of 6 X loading buffer).
3. Size markers (100 bp or 1 Kb ladder) were also loaded.
4. Minigels were run at 80 volts for 10-15 minutes and larger gels were run at 200 volts for approximately one hour.
5. DNA was visualised under UV on a transilluminator and photographed with a Polaroid camera.

2.1.1.3 DNA purification

2.1.1.3.1 Gel purification

The DNA fragment was excised from the agarose gel with a clean scalpel.

1. The gel slice was weighed in a 1.5 ml Eppendorf tube.
2. The gel slice was then purified using a Qiaquick Gel Extraction Kit™ (Qiagen) according to the manufacturer's instructions.
3. Recovery was tested by gel electrophoresis (section 2.1.1.2).

2.1.1.3.2 Ethanol precipitation

1. In a 1.5 ml microcentrifuge tube, 0.1 volumes of 3 M sodium acetate and either one volume of isopropanol or two and a half volumes of ethanol were added to the DNA.
2. The samples were mixed well by vortexing and incubated for 20 minutes at –20°C.
3. DNA was pelleted in a microcentrifuge at 13,000 rpm and washed with 70% ethanol.
4. The pellet was left to dry and then resuspended in the appropriate amount of $T_{0.1}E$.
5. Recovery was tested by gel electrophoresis (section 2.1.1.2).

2.1.1.4 Restriction enzyme digests of DNA

1. Up to 10 µg of DNA was digested in a reaction containing the appropriate 1 X buffer, 1 mM spermidine, 100 µg/ml BSA and 20-50 units of the appropriate enzyme.
2. The DNA was digested for 2 hours or overnight at the appropriate temperature for the enzyme.
3. The DNA was subjected to agarose gel electrophoresis and visualised (section 2.1.1.2).

Table 2.1: PCR mixes (A) and cycling programs (B, next page). All primer concentrations are at 100 ng/μl unless specified otherwise.

A. PCR mixes

Default		SSP-PCR1		SSP-PCR2		Vectorette	
Buffer 1	2μl	Buffer 1	2μl	Buffer 1	2μl	Buffer 1	1.5μl
5mM dNTPs	2μl	5mM dNTPs	1.1μl	5mM dNTPs	1.1μl	5mM dNTPs	1.5μl
1/20 βME	0.28μl	1/20 βME	0.28μl	1/20 βME	0.28μl	1/20 βME	0.21μl
5mg/ml BSA	0.07μl	5mg/ml BSA	0.66μl	5mg/ml BSA	0.66μl	0.5mg/ml BSA	0.495μl
40% sucrose	5.165μl	40% sucrose	5.11μl	40% sucrose	5.11μl	40% sucrose	4.545μl
Primer mix	0.325μl	$T_{0.1}E$	1.85μl	$T_{0.1}E$	2.85μl	Primer 224	0.375μl
Amplitaq	0.16μl					Primer specific	0.375μl
Total	10μl	Total	11μl	Total	12μl	Total	9μl

RACE1		RACE2		Vectorette/RACE enzyme mix		SSP-PCR enzyme mix	
Buffer 1	1.5μl	Buffer 1	1.5μl	AmpliTaq	0.12μl	Buffer 1	0.1μl
5mM dNTPs	1.5μl	5mM dNTPs	1.5μl	Taq Extender	0.12μl	AmpliTaq	0.2μl
1/20 βME	0.21μl	1/20 βME	0.21μl	Perfect Match	0.12μl	$T_{0.1}E$	0.7μl
5mg/ml BSA	0.495μl	5mg/ml BSA	0.495μl	40% sucrose	0.64μl		
40% sucrose	4.745μl	40% sucrose	4.32μl				
Primer specific	0.45μl	Primer specific	0.65μl				
Primer AP1	0.1μl	Primer AP2	0.325μl				
Total	9μl	Total	9μl	Total	1μl	Total	1μl

B. Cycling programs

Default		SSP-PCR		Vectorette	
95°C	5min	95°C	3min	95°C	3min
		Add 1μl SSP-PCR enzyme mix	Pause PCR machine	Add 1μl vectorette enzyme mix	Pause PCR machine
		95°C	2min	95°C	2min
94°C	30sec	94°C	30sec	94°C	5sec
60°C	30sec 30cycles	60°C	30sec 25cycles	68°C	30sec 17cycles
72°C	3min	72°C	3min	72°C	3min
				94°C	5sec
				60°C	30sec 18cycles
				72°C	3min
72°C	10min	72°C	10min	72°C	10min

RACE		PCR product re-amplification	
95°C	3min	95°C	5min
Add 1μl RACE enzyme mix	Pause PCR machine		
95°C	2min		
94°C	5sec	94°C	30sec
68°C	30sec 25cycles	60°C	30sec 20cycles
72°C	3min	72°C	3min
72°C	10min	72°C	10min

.

2.1.1.5 Primer design, synthesis and storage

1. Primers were designed using the Primer3 program (Rozen and Skaletsky, 2000; http://www.genome.wi.mit.edu/genome_software/other/primer3.html) from http://www.sanger.ac.uk/cgi-bin/primer3.cgi.

2. Primers were synthesised at the Sanger Institute by David Fraser and Diane Gibson. A subset of the primers were synthesised by GenSet (http://www.genxy.com/index.html).

3. Primers were stored at –20°C and working dilutions were prepared at 100 ng/µl for each primer.

2.1.2 Clone resources

Different types of clone resources have been used throughout this project. The Sanger Institute clone resource and the gene identification groups maintain the clone resources.

2.1.2.1 cDNA libraries used

The cDNA libraries used during the course of this project are described in Table 2.2.

Table 2.2 (next page): cDNA resources. Pools available for each technique are listed at the far right. Column A reports the number of SSP-PCR Super Pools available for each library (each Super Pool representing cDNA inserts from 100,000 clones). Column B reports the number of SSP-PCR individual Pools (each representing cDNA inserts from 20,000 clones). Column C reports the number of vectorette Super Pools (each representing cDNA inserts from 100,000 clones). Column D reports the number of individual plated pools (each representing cDNA inserts from 20,000 clones). Column E reports the number of liquid pools (each representing cDNA inserts from 250,000 clones grown in liquid media). Column F reports the number of liquid pools (each representing cDNA inserts from 100,000 clones grown in liquid media).

cDNA library code	cDNA library description	Supplier	Vector	Pools available					
				SSP-PCR		Vectorette			
				A	**B**	**C**	**D**	**E**	**F**
T	Adult testis	Clontech	pCDM8	5	25	-	-	-	10
FB	Fetal Brain	Invitrogen	pcDNAI	5	25	5	25	10	-
FL	Fetal Liver	Invitrogen	pcDNAI	5	25	5	25	10	-
FLu	Fetal Lung	Invitrogen	pcDNAI	5	25	5	25	10	-
HL60	Peripheral blood	Invitrogen	pcDNAI	5	25	5	25	10	-
AH	Adult Heart	Invitrogen	pcDNA3	5	25	5	25	10	-
ALu	Adult Lung	Clontech	pcDNAI	5	25	5	25	10	-
SK-N-MC	Neuroblastoma cells	Invitrogen	pcDNAI	5	25	-	-	-	-
PF	Adult brain	Pfizer	pcDNAI	3	15	-	-	-	-
U937+	(Monocyte, NOT activated, from a patient with promonocytic leukaemia)	DS*	pCDM8	5	25	-	-	-	-
U937ACT	(Monocyte, PMA activated, from a patient with promonocytic leukaemia)	DS*	pCDM8	5	25	-	-	-	-
H9	Placental, full term normal pregnancy	DS*	pH3M	5	25	-	-	-	-
YT	HTLV-1 +ve adult leukaemia T cell	DS*	pH3M	5	25	-	-	-	-
NK	Natural killer cell	DS*	pH3M	5	25	-	-	-	-
Daudi	B lymphoma	DS*	pH3M	5	25	-	-	-	-
HPBall	T cell from a patient with acute lymphocytic leukaemia	DS*	pH3M	5	25	-	-	-	-
BM	Bone marrow	DS*	pH3M	5	25	-	-	-	-
DX3	Melanoma	DS*	pH3M	5	25	-		-	-

*Libraries marked with an asterisk were generously provided by David Simmons, Oxford

2.1.2.2 Construction of cDNA pools

2.1.2.2.1 Library titration

1. 2 litres of LB agar were used to prepare 35 agar plates. Molten LB agar was left to cool down to 55°C and the appropriate amounts of antibiotics were added (ampicillin at 50 μg/ml and tetracycline at 10 μg/ml final volume). Approximately 50 ml of the mixture was used for each colony picker plate. The plates were left to set and stored at 4°C.

2. The cDNA library, consisting of bacteria stored in glycerol, was defrosted on ice and 2 μl were diluted in 198 μl LB. Six ten-fold serial dilutions were prepared and 100 μl of each was plated on LB agar plates with Hybond N+ filters, using an ethanol-flamed bent Pasteur pipette.

3. The plates were left inverted, at 37°C for 4 hours. They were then transferred to 30°C for 16 hours, and then back to 37°C for an additional 4 hours. The colonies of each plate were counted and the library titre estimated.

2.1.2.2.2 Low-density plated pools

1. The cDNA library was diluted in 20% glycerol/LB/antibiotics. 20,000 clones were plated out on each of 25 LB/antibiotics plates with Hybond N+ filters (500,000 clones in total).

2. The plates were left to dry for 3-5 minutes, and the clones were then left to grow (4 hours at 37°C, 16 hours at 30°C and 4 hours at 37°C).

2.1.2.2.3 High-density liquid pools

1. Ten 50 ml Falcon tubes containing 20 ml of LB/antibiotics were set up.
2. The cDNA library was diluted in LB/antibiotics and 250,000 clones were added to each tube (2,500,000 clones in total). The clones were left to grow at 37°C/240 rpm for 20 hours in a shaking incubator.
3. Dilutions of 1000, 100, 10 and 1 were plated out on colony picker plates/Hybond+ filters to check titration. The plates were inverted and left to grow overnight at 37°C.

2.1.2.2.4 Preparation of SSP-PCR pools

1. All SSP-PCR pools were prepared from low-density plated clones (section 2.1.2.2.2).
2. The contents of each filter were scraped into 3 ml of 20% glycerol in LB/antibiotics using a glass Pasteur pipette, and transferred to a 15 ml Falcon tube. The cells were shaken off and the filters were removed.
3. Super Pools were prepared by pooling 1 ml from each of a group of five tubes (Pools) in a new Falcon tube thus generating 5 Super Pools. The tubes were frozen in dry ice and stored at –70°C.
4. Pool templates for use in PCR screens and SSP-PCR reactions were prepared by transferring 0.3 ml of each Pool (or Super Pool) in screw-top Eppendorf tubes. The aliquots were boiled for 5 minutes and then quenched in ice. The contents were briefly spun, and stored at –20°C. 1/100 dilutions of the boiled template/$T_{0.1}E$ were prepared for the first and second steps of SSP-PCR (Super

Pool and individual pool screens), respectively. 1/10 dilutions of the boiled Pools template/$T_{0.1}E$ were prepared for the third step of SSP-PCR (cDNA-end recovery). All dilutions were stored at –20°C.

2.1.2.2.5 Preparation of vectorette pools

1. For plated vectorette pools, each Hybond N+ filter was removed from the colony picker plates, rolled-up and placed in a 50 ml Falcon tube containing 20 ml of SET. The cells were shaken off and the filters were removed.
2. The cells in the thirty-five, 50 ml Falcon tubes (25 plated pools/SET and 10 liquid pools/LB), were pelleted at 4,000 rpm for 10 minutes at room temperature, using a Beckman J6-MC centrifuge.
3. The media (or SET) was removed and the pellets were re-suspended in 200 µl of GTE on ice, transferred to a 1.5 ml tube and left to stand for 5 minutes.
4. 400 µl of freshly made 0.2 M NaOH/1% SDS (briefly cooled in ice) was added and the tubes were left to stand on ice for 5 minutes. 300 µl of 5 M acetate/3 M K^+ were added to each tube. The tubes were gently inverted once and left on ice for 10 minutes.
5. Cell debris was pelleted in a microcentrifuge at 13,000 rpm for 10 minutes and the clear supernatants were removed and put into clean tubes. These were spun for 5 minutes to remove any remaining debris.
6. 600 µl of isopropanol stored at –20°C was added in each tube. The tubes were then well shaken and left on ice for at least 10 minutes.

7. DNA was pelleted by spinning the tubes in a microcentrifuge at 4°C /13,000 rpm, for 15 minutes. The pellets were re-suspended in 200 µl of $T_{0.1}E$.
8. 200 µl of phenol:chloroform:isoamyl alcohol (25:24:1) was added in each tube. The tubes were shaken and spun for 5 minutes. The top (aqueous) layers were removed and placed in fresh tubes.
9. The DNA was ethanol precipitated (section 2.1.1.3.2).
10. The DNA was pelleted in a microcentrifuge and washed with 70% ethanol.
11. The DNA was resuspended in 30 µl $T_{0.1}E$ and stored at –20°C.
12. 1 µl of 10 mg/ml RNase was added to each tube and incubated at 37°C for 1 hour.
13. 1 µl of the DNA was run on a 0.8% agarose gel to check the extraction outcome.
14. 1 µg of each extracted DNA was digested with the appropriate enzyme in a total volume of 30 µl (section 2.1.1.4).
15. 70 µl of water was added and the DNA was extracted with 200 µl of phenol:chloroform:isoamyl alcohol (25:24:1).
16. The DNA was ethanol precipitated (section 2.1.1.3.2).
17. The DNA was pelleted in a microcentrifuge, washed with 70% ethanol and left to dry for 5 minutes. The pellets were then re-suspended in 100 µl of ligation buffer.
18. 10 µl of 1 pmol/µl annealed vectorette bubbles, 1.1 µl adenosine 5'-triphosphate and 2.5 units of T4 DNA ligase were added to each tube and left at 16°C overnight.

19. The contents of each tube were diluted to 500 µl with $T_{0.1}E$ to generate Stock Pools.

20. Equal volumes of sets of five plated Stock Pools were mixed to generate Stock Super Pools.

21. 1/100 dilutions of Stock Super Pools were prepared using $T_{0.1}E$, for Super Pool PCR screens.

22. 1/100 dilutions of the plated Stock Pools were prepared using $T_{0.1}E$, for the second step of the vectorette method (individual Pool screens).

23. 1/10 dilutions of the plated Stock Pools were prepared using $T_{0.1}E$, for cDNA-end recovery (vectorette PCR). 1/10 dilutions of the liquid Stock Pools were also prepared using $T_{0.1}E$, for PCR pool screening and cDNA-end recovery.

2.1.3 Isolation of cDNA fragments

Three methods (SSP-PCR, vectorette and RACE) were used to isolate expressed sequences from cDNA pools. The PCR mixes and cycling programs used for each technique are listed in Table 2.1. The various universal primers used are listed in Table 2.3.

Table 2.3: Universal primer sequences.

SSP-PCR	Primer Sequences	Comments
pH3M-1FP	CTT CTA GAG ATC CCT CGA	Amplifies inserts from pH3M and pCDM8 vectors
pH3M-2FP	GAT CCC TCG ACC TCG AGA T	Amplifies inserts from pH3M and pCDM8 vectors
pH3M-1RP	CGC AGA ACT GGT AGG TAT	Amplifies inserts from pH3M and pCDM8 vectors
pH3M-2RP	CGA CCT GCA GGC GCA GAA	Amplifies inserts from pH3M and pCDM8 vectors
T7-2FP	TAA TAC GAC TCA CTA TAG G	Amplifies inserts from pCDM8, pcDNA3 and pcDNAI vectors
pCDM8-RP	TAA GGT TCC TTC ACA AAG	Amplifies inserts from pCDM8 and pcDNAI vectors
SP6	ATT TAG GTG ACA CTA TAG	Amplifies inserts from pcDNA3 vectors
VECTORETTE	**Primer Sequences**	**Comments**
224	CGA ATC GTA ACC GTT CGT ACG AGA ATC GCT	Universal primer for all vectorette pools. Used for cDNA-end isolation
Xho-I	TCG AGC AAG GAG AGG ACC AAG GAG AGG ACG CTG TCT GTC GAA GGT AAG GAA CGG ACG AGA GAA GGG AGA G	Used for the construction of the Testis vectorette library
Xho-II	CTC TCC CTT CTC GAA TCG TAA CCG TTC GTA CGA GAA TCG CTG TCC TCT CCT TGG TCC TCT CCT TGC	Used for the construction of all vectorette libraries
RACE	**Primer Sequences**	**Comments**
AP1	CCA TCC TAA TAC GAC TCA CTA TAG GGC	Adaptor Primer 1
AP2	ACT CAC TAT AGG GCT CGA GCG GC	Nested Adaptor Primer 2

2.1.3.1 SSP-PCR

The technique of end-fragment isolation from cDNA libraries is an adaptation (Bye and Rhodes, unpublished) of the original SSP-PCR (Shyamala and Ames, 1989; Shyamala and Ames, 1993). cDNA pools for SSP-PCR were constructed by Jackie Bye, Suzan Rhodes and George Stavrides.

2.1.3.1.1 Identification of positive pools

1. 5 μl from each of the 88 available Super Pools (five Super Pools from seventeen cDNA libraries and 3 Super Pools from 1 cDNA library) were PCR screened.
2. For each positive Super Pool, the five individual constituent Pools (1/100 dilutions) were PCR screened.

2.1.3.1.2 Amplification of cDNA ends

1. The first part of the SSP-PCR cDNA-end isolation (SSP-PCR1) was performed using the 1/10 dilutions of the positive individual Pools. Because the orientation of the cDNA insert in the clone was unknown, two reactions were set-up for each gene-specific primer.
2. 1 μl from each positive pool and 4 μl of $T_{0.1}E$ were added to 11 μl of SSP-PCR1 buffer mix. 2 μl of the sense or antisense gene-specific primer (100 ng/μl), 1 μl (10 ng/μl) of either the forward or reverse vector primer and a drop of mineral oil were also added.
3. The PCR reaction was performed under the SSP-PCR cycling profile, which was briefly paused after the first step to add 1 μl of SSP-PCR mix in each well.

4. The second part of the SSP-PCR method of cDNA-end recovery (SSP-PCR2) was performed using 1/50 and 1/10 dilutions (in $T_{0.1}E$) of the SSP-PCR1 products as templates.

5. 5 µl from each dilution was added to 12 µl of SSP-PCR2 buffer mix. In agreement with the SSP-PCR1 step, 1 µl of either a sense or an anti-sense gene-specific nested primer was also added. 1 µl of either a forward or reverse nested vector primer and a drop of oil were also added.

6. The PCR reaction was performed under the SSP-PCR cycling profile, paused after the first step to add 1 µl of SSP-PCR mix in each well.

2.1.3.1.3 Isolation and re-amplification of cDNA ends

1. Reaction products were visualised by agarose gel electrophoresis (section 2.1.1.2).

2. The amplified DNA fragments were excised and the gel slices were placed in 1.5 ml tubes containing 100 µl $T_{0.1}E$ and left overnight at 4°C.

3. To obtain sufficient DNA for sequencing, liquid from around the gel slice was re-amplified by PCR (re-amplification PCR mix and cycling program). Four reactions were set-up to obtain sufficient DNA for sequencing. The amplified products were separated by gel electrophoresis (section 2.1.1.2) and gel purified (section 2.1.1.3.1). Recovery was checked by gel electrophoresis (section 2.1.1.2).

4. Elizabeth Huckle performed the DNA sequencing reactions.

2.1.3.2 Vectorette

The technique of vectorette cDNA end isolation is an adapted version (Collins, unpublished) of the original vectorette PCR (Riley et al., 1990; Arnold and Hodgson, 1991). cDNA pools for vectorette were constructed by John Collins, Melanie Goward and George Stavrides.

2.1.3.2.1 Identification of positive pools

1. 5 μl from each of the 30 available Super Pools (five Super Pools from six cDNA libraries) or the 60 liquid Pools (ten Pools from six cDNA libraries) were PCR screened.
2. For each positive Super Pool, the five individual constituent Pools (1/100 dilutions) were PCR screened.

2.1.3.2.2 Amplification of cDNA ends

1. Vectorette cDNA-end recovery was performed on 1/10 dilutions of positive Pools. 5 μl of each positive Pool were added to 9 μl of the vectorette reaction mixture that contained one gene-specific primer and a universal primer (224). One drop of oil was also added to each well.
2. The PCR reaction was performed under the vectorette cycling profile that was briefly paused after the first step in order to add 1 μl of vectorette enzyme mix.

2.1.3.2.3 Isolation and re-amplification of cDNA ends

1. The PCR products were isolated, re-amplified and sequenced, as above (section 2.1.3.1.3).

2.1.3.3 RACE

RACE was performed on either Human Brain or Testis Marathon-Ready™ cDNA.

2.1.3.3.1 Identification of positive Marathon pools

1. 5 µl of RACE template (2 µl of Marathon-Ready™ cDNA diluted in 3 µl $T_{0.1}E$) was PCR screened.

2.1.3.3.2 Amplification of cDNA ends

1. The first step of RACE (RACE1) was performed with 5 µl of RACE template (2 µl of Marathon-Ready™ cDNA diluted in 3 µl $T_{0.1}E$) added to 9 µl of RACE1 reaction mixture (which included one gene-specific primer and the adaptor primer AP1). PCR was performed under the RACE cycling profile, which was briefly paused after the first step to add 1 µl of RACE enzyme mix.
2. 1/50 and 1/10 dilutions of the RACE1 products were prepared using $T_{0.1}E$. 5 µl of each dilution was used as templates in a second round of RACE.
3. 9 µl of RACE2 buffer mix (which includes one gene-specific primer and the nested adaptor primer AP2) was added in each template and PCR was performed under the RACE cycling profile, which was briefly paused after the first step to add 1 µl of RACE enzyme mix.

2.1.3.3.3 Isolation and re-amplification of cDNA ends

The PCR products were isolated, re-amplified and sequenced, as above (section 2.1.3.1.3).

2.1.4 Northern Blots

2.1.4.1 Probe generation and labelling

1. Probes were generated by PCR (section 2.1.1.1) from cDNA templates (Table 2.2).
2. PCR products were separated by gel electrophoresis (section 2.1.1.2).
3. The expected-size band was excised and stored in 100 µl $T_{0.1}E$, at 4°C.
4. Labelling was performed (to be described in section 2.2.1.2).

2.1.4.2 Hybridisation

Labelled probes were hybridised to Multiple Tissue Northern (MTN®) Blots (Clontech). Each blot contains 2 µg of polyA mRNAs from different adult and fetal human tissues (Table 2.4).

Table 2.4: Northern Blots.

Human MTN Blot #7760-1	Human MTN Blot II #7759-1	Human Fetal MTN Blot II #7756-1
Heart	Spleen	Brain
Brain	Thymus	Lung
Placenta	Prostate	Liver
Lung	Testis	Kidney
Liver	Ovary	
Skeletal muscle	Small Intestine	
Kidney	Colon	
Pancreas	Peripheral Blood Leukocyte	

1. The blots were pre-hybridised for 1 hour and then hybridised for 18 hours at 65°C in ExpressHyb Hybridisation solution (Clontech).
2. The blots were washed twice in Northern Wash Solution I for 5 minutes at room temperature, then twice in Northern Wash Solution II for 3 minutes at 55°C.
3. The blots were subjected to autoradiography for an average of 4 days, at room temperature.

2.2 Mouse studies

The RPCI-23 female (C57Bl/6J) mouse BAC library (Osoegawa *et al*., 2000) was screened in this study. Library details are shown in Table 2.5.

Table 2.5: Details of the mouse genomic library.

Library	Library type	Library code	Antibiotic	Vector	Cloning site	Genomic digest
RPCI-23	BAC	bM	Chloramphenicol 12.5 μg/ml	pBACe3.6	*Eco*RI	*Eco*RI

2.2.1 Probe preparation

2.2.1.1 Primer testing and probe generation with PCR

1. 20 μl reaction mixtures were set-up in a 96-well plate containing 10 μl of dilution-buffer/primer-mix solution, 7.063 μl sucrose solution, 2 μl of PCR Buffer 2, 0.187 μl of 1/10 β-ME, 0.25 μl of 5mM dNTP solution, 0.4 μl of mouse genomic DNA (or $T_{0.1}E$ as negative control) and 0.1 μl (0.5 U) of AmpliTaq.
2. PCR was performed using a touch-down PCR program (Table 2.6).
3. PCR products were separated by gel electrophoresis (section 2.1.1.2). The expected-size band was excised and stored in 100 μl of $T_{0.1}E$ at 4°C.

Table 2.6: Touch-down PCR programs.

Steps	65T		60T		55T	
1	94°C	5min	94°C	5min	94°C	5min
2	93°C	30s	93°C	30s	93°C	30s
3	65°C	50s	60°C	50s	55°C	50s
4	-0.5 per cycle		-0.5 per cycle		-0.5 per cycle	
5	72°C	50s	72°C	50s	72°C	50s
6	repeat steps 2-5, 9 times		repeat steps 2-5, 9 times		repeat steps 2-5, 9 times	
7	93°C	30s	93°C	30s	93°C	30s
8	60°C	50s	60°C	50s	60°C	50s
9	72°C	50s	72°C	50s	72°C	50s
10	repeat steps 7-9, 29 times		repeat steps 7-9, 29 times		repeat steps 7-9, 29 times	
11	72°C	5min	72°C	5min	72°C	5min

2.2.1.2 PCR labelling

1. A 9.5 µl reaction mixture was set-up in a 0.5 ml tube, for each probe. The reaction mixture contained 1 µl of PCR Buffer 3, 0.4 µl of primer mix, 2.5 µl of the liquid surrounding the gel slice, 0.4 µl of d(ATG), 4.6 µl of H_2O and 0.1 µl of AmpliTaq.
2. A single drop of mineral oil was added on top of the reaction mixture, followed by 0.5 µl of α-^{32}P.
3. PCR was performed as follows: 94°C for 5 minutes, 20 cycles of 93°C for 30 seconds, 55°C for 30 seconds and 72°C for 30 seconds, followed by 1 cycle at 72°C for 5 minutes.
4. The PCR products were then denatured at 99°C for 5 minutes in the thermal cycler, snap chilled on ice and added to the hybridisation mix (unless competitive re-association was required, section 2.2.1.3).

2.2.1.3 Competitive re-association of radiolabelled probes

Where appropriate, the radiolabelled probes were competed using polyCA•GT to suppress the non-specific binding of PCR-labelled probes containing polyCA•GT.

1. The labelling reaction products (section 2.2.1.2) were transferred in a screw-top microcentrifuge tube containing 5 μl of 1 mg/ml polyCA•GT, 125 μl of 20 X SSC and 360 μl of H_2O.
2. The tube was left to boil for 5 minutes in a water bath.
3. The tube was snap chilled on ice and the contents were then added to the hybridisation mix.

2.2.2 Screening

2.2.2.1. Screening library filters

1. The probes were prepared as described above (section 2.2.1).
2. **A.** 2-30 library filters were placed in a 15 X 10 X 5 cm sandwich box containing approximately 150 ml of hybridisation buffer (enough to cover all filters). A plastic sheet was placed on top of the filters. The filters were then left to pre-hybridise at 65°C for at least 3 hours, in an orbital shaker. The filters and the plastic sheet were removed from the sandwich box and the denatured, labelled probe(s) added to the hybridisation solution and mixed well. The filters were added back to the box one by one and the plastic sheet was again placed on top.

B. When only one filter was to be screened, a 15 ml Falcon tube was used instead of sandwich boxes. The filter was placed in the 15 ml Falcon tube containing 13 ml of hybridisation buffer. The filter was left to pre-hybridise at 65°C for at least 3 hours, in an orbital shaker. The labelled probe was then added and mixed well.

3. The filters were left to hybridise overnight at 65°C, in an orbital shaker.
4. The filters were then rinsed twice in 2 X SSC at room temperature, followed by two 30 minute washes with 0.5 X SSC/1% N-Lauroyl Sarcosine at 65°C in an orbital shaker.
5. The filters were rinsed twice in 0.2 X SSC and wrapped in Saran™ wrap.
6. The wrapped filters were placed overnight in a cassette with an X-ray film and two intensifying screens.
7. The X-ray films were developed and data was entered in the 2musace database.

2.2.2.2 PCR screening of BAC DNA pools

1. BAC DNA pools were PCR screened with the STS of interest (section 2.1.1.1).
2. Positive pools were identified by gel electrophoresis (section 2.1.1.2).

2.2.2.3 Colony PCR

1. The colony of interest was picked in 70 µl of H_2O and boiled at 95°C.
2. 5 µl aliquots were used as templates in PCR screens with the STS(s) of interest (section 2.1.1.1) and PCR products were visualised by gel electrophoresis (section 2.1.1.2).

2.2.3 Fingerprinting

2.2.3.1 Clone picking and microprepping

1. BAC clones were picked in 96-deep-well plates containing 1.5 ml LB broth/antibiotics, using a wooden cocktail stick.
2. The picked clones were grown at 37°C/300 rpm for 16 hours.
3. Eight plate copies were prepared by aliquoting 170 µl from each well into 96-well plates. Each plate copy was frozen on dry ice and stored at –70°C.
4. Bacterial clones were microprepped (Marra *et al*., 1997) by Carol Carder (Sanger Institute).

2.2.3.2 Restriction enzyme digestion

1. Simultaneous digestion of 96 clones was achieved with the use of 96-well plates.
2. The restriction digest mixture for each BAC DNA consisted of 2.6 µl of ddH_2O, 0.9 µl of 10 X enzyme buffer B, and 0.5 µl of *Hind*III. The mixture was delivered in each well using a Hamilton combitip dispenser.
3. The 96-well plate was sealed, gently agitated using a whirlimixer, briefly centrifuged at 1,000 rpm and incubated at 37°C for 2 hours.
4. The reaction was terminated by the addition of 2 µl of 6 X loading dye. The plate was resealed, and briefly centrifuged at 1,000 rpm.

2.2.3.3 Agarose gel electrophoresis and data acquisition

1. 450 ml of molten agarose were used to prepare 1% agarose gels in 1 X TAE.
2. Each solidified gel was placed in an electrophoresis unit containing 2-3 litres of 1 X TAE.
3. 0.8 µl of marker mix was added in the first well and every fifth well. 1 µl from each restriction enzyme digestion/loading dye mix was then loaded.
4. Samples were electrophoresed at room temperature for 30 minutes at 90 volts. The electrophoresis apparatus was then moved into the cold room where the gels were left to run for 15 hours at 90 volts.

5. Following electrophoresis, the gels were trimmed to ~19 cm, placed in plastic trays containing Vistra Green stain mix and agitated in the dark, on an orbital flatform shaker, for 45 minutes.

6. The gels were then briefly rinsed with de-ionised H_2O and imaged using a FluorImager SI.

2.2.3.4 Fingerprint analysis and contig construction

1. Fingerprint analysis was performed interactively using the Image 3.10 software (Sulston *et al*., 1980; Platt and Wobus, unpublished; also see section 4.2.3).

2. Band data was collected and used to perform an automatic contig assembly using FPC V4 (Soderlund *et al*., 2000). The parameters used were an overlap statistic of $3x10^{-12}$ (about 75% clone overlap) and 0.7 mm tolerance.

3. To identify potential joins, fingerprints of clones at the extreme ends of contigs were used to query the FPC database at a lower fingerprint overlap stringency (overlap statistic of $1x10^{-8}$ or about 50% clone overlap).

4. Joins were incorporated into the map if the fingerprint data was logically consistent with the proposed map order.

2.3 Human variation

2.3.1 DNA samples

The Caucasian, Asian and African American samples were obtained from the Coriell Cell Repository (http://locus.umdnj.edu/ccr/). The Caucasian panel consists of 95 DNA samples and is drawn from the UTAH CEPH pedigree collection. The Asian panel consists of twelve Japanese DNA samples from unrelated individuals. The African American panel consists of twelve African American DNA samples from unrelated individuals. Specific sample identifiers are listed in Table 2.7.

All DNA samples were diluted to 3.5 ng/µl using $T_{0.1}E$ (working DNA solutions).

2.3.2 SNP selection and primer design

Publicly available SNPs from various discovery efforts were utilised. SNPs were selected in a hierarchical way so as to generate a polymorphic SNP map of increasing density. Genotyping was performed after each round of SNP selection followed by additional SNP selection and genotyping. Where possible, SNPs that mapped outside repeats were selected.

Sarah Hunt designed the primers and probes in a quadruplex format, using the SpectroDesigner software (Sequenom, San Diego, CA). GenSet (http://www.genset.fr/) synthesised all SNP primers and probes.

Table 2.7: (A) The Caucasian family samples. (B) and the Caucasian, Asian and African American samples (twelve unrelated individuals from each ethnic group). PGF, paternal grandfather; PGM, paternal grandmother; MGF, maternal grandfather; MGM, maternal grandmother; F, father; M, mother; S, son; D, daughter.

A.

Family ID	Relation	DNA name	Family ID	Relation	DNA name	Family ID	Relation	DNA name
1331	PGF	NA07007	1341	PGF	NA07034	1408	PGF	NA12154
1331	PGM	NA07340	1341	PGM	NA07055	1408	PGM	NA12236
1331	MGM	NA07016	1341	MGM	NA06993	1408	MGM	NA12155
1331	MGF	NA07050	1341	MGF	NA06985	1408	MGF	NA12156
1331	F	NA07057	1341	F	NA07048	1408	F	NA10830
1331	M	NA06990	1341	M	NA06991	1408	M	NA10831
1331	S	NA06983	1341	S	NA07020	1408	S	NA12148
1331	D	NA06988	1341	D	NA07006	1408	D	NA12149
1333	PGF	NA07049	1346	PGF	NA12043	1416	PGF	NA12248
1333	PGM	NA07002	1346	PGM	NA12044	1416	PGM	NA12249
1333	MGM	NA07017	1346	MGM	NA12045	1416	MGM	NA12250
1333	MGF	NA07341	1346	F	NA10857	1416	MGF	NA12251
1333	F	NA07038	1346	M	NA10852	1416	F	NA10835
1333	M	NA06987	1346	S	NA12039	1416	M	NA10834
1333	S	NA07009	1346	D	NA12040	1416	S	NA12243
1333	D	NA07011	1347	PGF	NA11879	1416	D	NA12244
1334	PGF	NA12144	1347	PGM	NA11880	1420	PGF	NA12003
1334	PGM	NA12145	1347	MGM	NA11881	1420	PGM	NA12004
1334	MGM	NA12146	1347	MGF	NA11882	1420	MGM	NA12005
1334	MGF	NA12239	1347	F	NA10858	1420	MGF	NA12006
1334	F	NA10846	1347	M	NA10859	1420	F	NA10838
1334	M	NA10847	1347	S	NA11871	1420	M	NA10839
1334	S	NA12138	1347	D	NA11870	1420	S	NA12007
1334	D	NA12139	1362	PGF	NA11992	1420	D	NA11997
1340	PGF	NA06994	1362	PGM	NA11993	1423	PGF	NA11917
1340	PGM	NA07000	1362	MGM	NA11994	1423	PGM	NA11918
1340	MGM	NA07022	1362	MGF	NA11995	1423	MGM	NA11919
1340	MGF	NA07056	1362	F	NA10860	1423	MGF	NA11920
1340	F	NA07029	1362	M	NA10861	1423	F	NA10842
1340	M	NA07019	1362	S	NA11984	1423	M	NA10843
1340	S	NA07040	1362	D	NA11985	1423	S	NA11909
1340	D	NA07053				1423	D	NA11910

B.

No.	Asian	African American	Caucasian[1]
1	NA17051	NA17109	NA11879
2	NA17053	NA17111	NA11880
3	NA17056	NA17114	NA11881
4	NA17057	NA17115	NA11882
5	NA17058	NA17117	NA12248
6	NA17060	NA17119	NA12249
7	NE00251	NA17122	NA12250
8	NE00374	NA17124	NA12251
9	NE00904	NA17125	NA07340
10	NE00810	NA17132	NA07016
11	NE00299	NA17134	NA07050
12	NE00744	NA17136	NA07007

[1]Also part of the family panel

2.3.3 Working PCR primer mix and probe dilutions

1. For each set of 384 SNPs to be genotyped, 375 nM (for each primer) quadruplex primer mix dilutions were prepared in either a 96-well V bottom plate or a 0.5 ml Costar Assay Block on a Genesis RSP Tecan, using ddH_2O.

2. Similarly, 10 μM (for each probe) quadruplex probe dilutions were also prepared.

2.3.4 PCR amplification

Reactions were performed in 384-well microtitre plates. Each microtitre plate was used to genotype 384 SNPs assays across four DNA samples (1,536 assays).

1. The quadruplex PCR reaction mixtures (5 μl final volume) consisted of 2 μl of the appropriate primer mix, 0.75 μl of 10 X PE buffer, 0.2 μl of 5mM dNTP mix,

1.01 µl of ddH_2O, 0.04 µl (2 X) of Titanium Taq and 1 µl of the DNA to be tested.

2. Each 384-well plate was sealed with Microseal 'A' film and PCR was performed under the following cycling profile: 95°C for 1 minute, 45 cycles of 95°C for 20 seconds, 56°C for 30 seconds and 72°C for 1 minute, followed by 1 cycle at 72°C for 5 minutes.
3. The plates were spun at 1,000 rpm for one minute, their seals replaced with ScotchPad™ Tape Pads, and stored at –20°C until the next step.

2.3.5 SAP

After the PCR reaction, the un-incorporated dNTPs were inactivated using Shrimp Alkaline Phosphatase (SAP).

1. SAP deactivation of dNTPs was performed in a total volume of 7 µl consisting of the PCR reaction products (5 µl), 1.5 µl of ddH_2O, 0.2 µl of 10 X TS buffer and 0.3 µl (0.3 U) of SAP enzyme.
2. Each 384-well plate was sealed with Microseal 'A' film. The SAP reaction was performed at 37°C for 20 minutes, followed by inactivation of the SAP enzyme at 80°C for 5 minutes.
3. The plates were spun at 1,000 rpm for one minute, their seals replaced with ScotchPad™ Tape Pads, and stored at –20°C until the next step.

2.3.6 Extension of primer probe

1. Primer probe extension was performed in a total volume of 9 µl consisting of the SAP reaction products (7 µl), 0.5 µl of the appropriate extend primer probes, 0.382 µl of ddH_2O, 0.2 µl of 10 X TS buffer, 0.9 µl of the appropriate dNTP/ddNTP combination and 0.018 µl (0.58 U) of Thermosequenase.
2. Each 384-well plate was sealed with Microseal 'A' film and reactions were cycled as follows: 94°C for 2 minutes and 55 cycles of 94°C for 5 seconds, 52°C for 5 seconds and 72 °C for 5 seconds, followed by 1 cycle at 72°C for 5 minutes.
3. The plates were spun at 1,000 rpm for one minute, their seals replaced with ScotchPad™ Tape Pads, and stored at –20°C until the next step.

2.3.7 Water and resin addition

1. After the extension of primer probes, the 384-well plates were spun at 1,000 rpm for 1 minute.
2. 16 µl of Milli-Q water was delivered in each well using a BECKMAN Multimek-96 automated 96-Channel Pipettor.
3. To remove residual salt from the reactions, a cation exchange resin (SpectroCLEAN™) was added. The resin was initially applied on specially adapted trays and then delivered in each well. Approximately 1 g of resin was used for each 384-well plate.

4. The plates were rotated at medium speed on a Roto-Shake Genie™ rotator for 4 minutes.
5. The plates were then spun for 4 minutes at 4,000 rpm.

2.3.8 Mass spectroscopy

1. 7 nl of the purified primer extension reaction was loaded on to a matrix pad (3-hydroxypicoloinic acid) of a SpectroCHIP.
2. SpectroCHIPs were analysed using a Bruker Biflex III Maldi-TOF mass spectrometer (SpectroReader, Sequenom).
3. Spectra were processed using SpectroTYPER (Sequenom).

2.4 Bioinformatics and computational support

The software used in these studies is listed in Table 2.8. The names of people involved in database management and sequence analysis are reported in Table 2.9. Table 2.10 gives the URLs of web sites used.

Table 2.8: Software used in this study.

Software	Description	Reference
ACeDB	Data storage/graphical display	Durbin and Thierry-Mieg, 1994
Ensembl	Genome browser	Hubbard *et al*., 2002
UCSC GB	Genome browser	Kent *et al*., 2002
Genscan	Gene prediction	Burge and Karlin, 1997
FGENESH	Gene prediction	Salamov and Solovyev, 2000
RepeatMasker	Repeat sequences prediction	Smit and Green, unpublished
CPGFIND	CpG island prediction	Micklem, unpublished
PromoterInspector	Promoter prediction	Scherf *et al*., 2000
Eponine	TS site prediction	Down and Hubbard, 2002
BLAST	Similarity searches	Altschul *et al*., 1990, 1997
InterProScan	Protein motif analysis	Zdobnov and Apweiler, 2001
Dotter	Dot plot DNA comparisons	Sonnhammer and Durbin, 1995
CLUSTAL W	Sequence alignments	Thompson *et al*., 1994
Belvu	Formatting of aligned sequences	Sonnhammer, unpublished
FPC 4V	Contig building	Soderlund *et al*., 2000
Image 3.10	Processing of raw fingerprint data	Sulston *et al., 1980*
PipMaker	Comparative sequence alignments	Schwartz *et al*., 2000
Spectro Designer	SNP assay design	Sequenom™, unpublished
SpectroTYPER RT	Genotype analysis	Sequenom™, unpublished
SpectroTYPER DB	SNP data storage	Sequenom™, unpublished
SpectroCHECK	Genotype Quality Control check	Sequenom™, unpublished

Table 2.9: People involved in sequence analysis and data storage and management.

James Gilbert	Automated sequence analysis and chromosome 20 Ensembl database curator
Michele Clamp Guy Slater	Exonerate analysis of chromosome 20 sequence and WGS homologous mouse reads
Sarah Hunt	SNP analysis and data management
Carol Scott	Management of chromosome-specific fingerprint and sequence databases
Jilur Ghori	Oligo ordering and management of primace database
Jennifer Ashurst Laurens Wilming Andrew King Kerstin Jekosch	Sequence analysis and annotation
Panos Deloukas George Stavrides	Chromosome 20 gene annotation group
Lisa French Ian Mullenger	Manual curation of mouse FPC database

Table 2.10: URLs used in this study.

Description	URL
ACeDB	http://www.acedb.org/
BACPAC resources	http://www.chori.org/bacpac/home.htm
BLAST at NCBI	http://www.ncbi.nlm.nih.gov/BLAST/
BLAT	http://genome.ucsc.edu/cgi-bin/hgBlat?db=hg7
DKFZ	http://mbi.dkfz-heidelberg.de/
Dotter	http://www.cgr.ki.se/cgr/groups/sonnhammer/Dotter.html
Ensembl	http://www.ensembl.org/Docs/
Ensembl Trace server	http://trace.ensembl.org/
European Bioinformatics Institute	http://www.ebi.ac.uk/
European Molecular Biology Laboratory	http://www.embl.org/
FPC	http://www.sanger.ac.uk/Software/fpc
GeneMap '99	http://www.ncbi.nlm.nih.gov/genemap/
Genoscope	http://www.genoscope.cns.fr
Human BLAST server at the Sanger Institute	http://www.sanger.ac.uk/HGP/blast_server.shtml
Human Chromosome 20	http://www.sanger.ac.uk/HGP/Chr20/
Image	http://www.sanger.ac.uk/Software/Image
Mouse BAC end sequences	http://www.tigr.org/tdb/bac_ends/mouse/bac_end_intro.html
Mouse BAC fingerprints	http://www.bcgsc.bc.ca/projects/mouse_mapping/
Mouse genome sequence FTP at the Sanger Institute	ftp://ftp.sanger.ac.uk/pub/mouse/
PipMaker	http://bio.cse.psu.edu/pipmaker/
PromoterInspector	http://www.genomatix.de/cgi-bin/promoterinspector/promoterinspector.pl
PubMed	http://www.ncbi.nlm.nih.gov/PubMed/
RIKEN Genomic Sciences Centre	http://www.gsc.riken.go.jp/
SMART	http://smart.embl-heidelberg.de/help/smart_about.shtml
SSAHA	http://www.sanger.ac.uk/Software/analysis/SSAHA/
The Baylor College of Medicine search launcher	http://searchlauncher.bcm.tmc.edu/
The Coriell Cell Repository	http://locus.umdnj.edu/ccr/
The dbSNP database	http://www.ncbi.nlm.nih.gov/SNP/

The EMBL nucleotide sequence database	http://www.ebi.ac.uk/embl/
The Ensembl Human genome server	http://www.ensembl.org/Homo_sapiens/
The Ensembl Mouse genome server	http://www.ensembl.org/Mus_musculus/
The EST database dbEST	http://www.ncbi.nlm.nih.gov/dbEST/
The GenBank DNA sequence database	http://www.ncbi.nlm.nih.gov/Genbank/index.html
The Genome Sequence Centre (BCGSC)	http://www.bcgsc.bc.ca/
The Institute for Genomic Research	http://www.tigr.org/
The InterPro database	http://www.ebi.ac.uk/interpro/
The InterProScan package	http://www.ebi.ac.uk/interpro/scan.html
The Jackson Laboratory mice web site	http://jaxmice.jax.org/index.shtml
The LocusLink query interface	http://www.ncbi.nlm.nih.gov/LocusLink/
The Mouse Sequencing Consortium	http://www.sanger.ac.uk/Info/Press/001006.shtml
The MRC Mouse Genome Centre	http://www.mgc.har.mrc.ac.uk/
The National Center for Biotechnology Information	http://www.ncbi.nlm.nih.gov/
The Online Mendelian Inheritance in Man database	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
The Pfam collection of protein sequence alignments	http://www.sanger.ac.uk/Software/Pfam/
The PRINTS compendium of conserved protein motifs	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/
The ProDom (protein domain database)	http://prodes.toulouse.inra.fr/prodom/doc/prodom.html
The PROSITE database of protein families and domains	http://www.expasy.ch/prosite/
The Reference Sequence project	http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html
The Sanger Institute	http://www.sanger.ac.uk/
The Sequence Retrieval System (SRS) at the EBI	http://srs.ebi.ac.uk/
The SNP Consortium	http://snp.cshl.org/
The SWISS-PROT protein sequence database	http://www.expasy.org/sprot/
The TIGRFAMs collection of protein families	http://www.tigr.org/TIGRFAMs/Explanations.shtml
The TrEMBL computer-annotated protein database	http://www.ebi.ac.uk/swissprot/
The UniGene human sequences collection	http://www.ncbi.nlm.nih.gov/UniGene/Hs.Home.html
The Whitehead Institute Center for genome research	http://www-genome.wi.mit.edu/
The Whole Mouse Catalog	http://www.rodentia.com/wmc/toc.html

2.5 Materials

Culture plates
(Genomic Solutions #CONS1001)

Ampicillin sodium salt (Sigma #A9518)

Hybond N+ filters (Amersham #NK9655)

1.2 ml screw-top propylene tubes
(Costar #2027)

Beckman J2-MC centrifuge

RNase solution, 500 μg/ml
(Boehringer Mannheim #119 915)

100 X NEB BSA, 10 mg/ml
(New England Biolabs #B9001S)

10 X Ligation buffer
(Boehringer Mannheim #1243 292)

Phenol:chloroform:isoamyl alcohol,
25:24:1 v/v
(GIBCO BRL #15593-031)

Taq Extender PCR additive 5 U/μl
(Stratagene #600-148-81)

BSA (Albumin, Bovine), 5% solution
(Sigma A-4628)

Tetracycline hydrochloride
(Sigma #T3383)

Glucose (BDH #10117)

EDTA (IBI #IB70182)

Innova 4000 Incubator Shaker
(New Brunswick Scientific)

Isopropanol (BDH #102246L)

*Eag*I, 10,000 U/ml
(New England Biolabs #505L)

5 U/μl T4 DNA ligase
(Boehringer Mannheim #799 009)

Glacial acetic acid
(Mallinckrodt Baker #9507-02)

100 mM adenosine 5'-triphosphate solution
(Amersham Pharmacia
Biotech #27-2056-01)

Perfect Match 1 U/μl
(Stratagene #600-129-81)

Ammonium sulphate, enzyme grade
(GIBCO BRL #5501UA)

Ultrapure dNTP set 100 mM solution (Amersham Pharmacia Biotech #27-20-35-01)	Cresol red sodium salt (Sigma #C-9877)
Sucrose (BDH #102744B)	Tris base (Anachem #0826)
Human Brain Marathon-Ready™ cDNA (Clontech #7400-1)	Human Testis Marathon-Ready™ cDNA (Clontech #7414-1)
Light white oil (Sigma #M3516)	MJ thermal cycler
BAC genomic filters	Hybaid OmniGene cycler
Costar 6511 96-well plates (M-type)	Omni seals (Hybaid #HB-TD-MT-SRS-5)
AmpliTaq DNA polymerase, 5 U/µl (Roche #N808-0145)	Mouse genomic DNA, 0.1 µg/µl (Clontech #6650-1)
2-mercaptoethanol (Bio-Rad #161-0710)	BSA (Sigma A-2153)
Perkin Elmer DNA thermal cycler	HAAKE SW20 waterbath
Kodak M35I film processor	Decon 90 (Decon Laboratories)
Beta cabinets, shields, bins, racks and boxes (Anachem-Scotlab)	Innova 4080 Incubator shaker (New Brunswick Scientific)
N- Lauroyl Sarcosine (Sigma #L-5125)	Saran™ wrap (Dow Chemical Co.)
Poly (dA-dC)•Poly (dG-dT) (Amersham Pharmacia Biotech #27-7940-01)	Redivue [α-^{32}P]dCTP 10 mCi/ml, 3000 Ci/mmol (Amersham Pharmacia Biotech #AA0005)
Sodium chloride (BDH #301237S)	Super RX Fuji X-Ray film (#03G010)
Tri-sodium citrate (BDH #301287F)	Whatman filter paper (#1001 240)

Sodium dodecyl sulphate (SDS, BDH #442444H)	4N Sodium hydroxide solution (BDH 191373M)
Whatman 3MM chromatography paper (#3030 931)	Dextran sulphate (Amersham #17-0340-02)
Polyvinylpyrrolidone (Sigma #PVP-40)	Ficoll (Sigma #F-9378)
Innova 4000 Incubator Shaker (New Brunswick Scientific)	2ml deep 96-square-well titre plates (Beckman #140504)
96-well Microtest, flat bottom plates (Falcon #353072)	Glacial acetic acid (Mallinckrodt Baker #9507-02)
HindIII, 40 U/µl (Roche Molecular Biochemicals #798983)	10 X SuRE/Cut Buffer B (Roche Molecular Biochemicals)
DNA molecular weight marker V (Roche Molecular Biochemicals #821705)	Gel tanks (Owl Scientific Gator Wide Forma System model A3-1 #008100191)
Analytical marker DNA, wide range (Promega #DG1931)	Seakem LE Agarose (FM Bioproducts #50004)
Mylar plate sealers (Dynex Technologies #5701)	Hamilton repeat dispenser (Hamilton Company)
Benchtop centrifuge (Eppendorf #5415C)	Wooden cocktail sticks
Ficoll Type 400-DL (Sigma #F-9378)	Cold room regulated to 4°C
Vistra Green (Amersham Life Sciences #RPN5786)	FluorImager SI Vistra Fluorescence (Molecular Dynamics)
Tabletop centrifuge (Sorvall #RT6000D)	Eppendorf Combitip Repeat Dispenser
Xylene cyanol (BDH #44306)	Bromophenol blue (BDH #20015)

ExpressHyb Hybridisation solution (Clontech #8015)

RoboDesign SpectroPOINT (Sequenom)

Titanium Taq DNA Polymerase, 50 X (Clontech #8434)

Clontech MTN® Blots (#7756-1, #7759-1, #7760-1)

Bruker Biflex™ III MALDI-TOF mass spectrometer

Thermo-Fast® 384-well plates (Abgene #TF-0384)

96-well V bottom plates with lids (SARSTEDT #82.1583.001)

0.5 ml Assay block (Costar #3956)

Genesis RSP (Robotic Sample Processor) 100/8 Tecan with MβP® 50µl tips (BioRobotic Molecular BioProducts #902-262), linked to an AcerPower4100

MATRIX technologies 100 ml disposable reagent reservoir (MATRIX #8086)

TOMTEC THINLID ™ plate sealers (Costar #3095)

ScotchPad™ Tape Pads (3M #0212-61618)

BECKMAN Multimek-96 automated 96-Channel Pipettor

Microseal 'A' film (MJ Research #MSA-5001)

Impact multichannel pipettes by MATRIX

Benchtop centrifuge (Eppendorf #5403)

Shrimp alkaline phosphatase 1 U/µl (Amersham #70092)

SpectroCLEAN™ (Sequenom™ #10053)

Roto-Shake Genie™ (Scientific Industries Inc.)

Thermosequenase DNA polymerase, 32 U/µl (Amersham #79000)

3-Pt Calibrant (Sequenom #335)

SpectroCHIP (Sequenom #000153)

2.6 DNA ladders

2.6.1 1 Kb ladder (Gibco-BRL #15615-024). This contains 1 to 12 repeats of a 1,018 bp concatenated fragment and vector fragments from 75 to 1,636 bp, producing the following sized fragments (bp):

12,216	6,108	1,018	201
11,198	5,090	517/506	154
10,180	4,072	396	134
9,162	3,054	344	75
8,144	2,036	298	
7,126	1,636	220	

2.6.2 100 bp DNA ladder (Gibco-BRL #15628-019). This consists of 15 blunt-ended fragments between 100 and 1500 bp in multiples of 100 bp and an additional fragment at 2072 bp, producing the following sized fragments (bp):

2,072	1,000	400
1,500	900	300
1,400	800	200
1,300	700	100
1,200	600	
1,100	500	

2.6.3 Wide Range Analytical Marker DNA (Promega)

The Analytical Marker DNA, Wide Range, provides an evenly spaced distribution of 37 DNA fragments ranging from 702 bp to 29,950 bp in size and was used for band sizing in fingerprint experiments. This marker is composed of a mixture of restriction enzyme digests of Lambda DNA and ϕX174 DNA.

2.7 Solutions

$T_{0.1}E$

10 mM Tris-HCl (pH8.0)

0.1 mM EDTA

1 X GTE

50 mM glucose

25 mM Tris-HCl (pH8.0)

1 mM EDTA

Vectorette bubbles (1 ml)

1 nmole of XhoI or EagI primer

1 nmole of XhoII primer

25 µl of 1 M NaCl

H_2O up to 1 ml

Heated to 65°C for 5 minutes and left to cool at room temperature

5 M acetate-3M K^+ (100 ml)

60 ml 5 M potassium acetate

11.5 ml glacial acetic acid

28.5 ml H_2O

SET

10 mM Tris-HCl (pH8.0)

0.1 mM EDTA

100 mM NaCl

PCR buffer 1 (1 litre)

80 g Tris Base

22 g $(NH_4)_2SO_4$

65 ml 1 M $MgCl_2$

Made up to 1000 ml with ddH_2O

40% sucrose/cresol red (1litre)

400 g sucrose

0.1 g cresol red

Made up to 1000 ml with ddH_2O

34.6% sucrose/cresol red (1 litre)

346 g sucrose

0.1 g cresol red

Made up to 1000 ml with ddH_2O

Northern wash solution I

2 X SSC

0.05% SDS

Northern wash solution II

0.1 X SSC

0.1% SDS

Nucleotide mix (PCR labelling)

dATG, dGTP and dTTP at a concentration of 5 mM each.

Nucleotide mix (PCR amplification)

dATG, dCTP, dTTP and dGTP at a concentration of 5 mM each

Primer mix

Sense and antisense primers at a concentration of 100 ng/µl each

PCR buffer 2 (10 ml)

4.5 ml 1 M Tris-HCl (pH8.8)

0.15 ml 1 M $MgCl_2$

0.1453 g $(NH_4)_2SO_4$

0.35 ml H_20

5 ml cresol red solution

Cresol red solution

0.1 g/l cresol red in T0.1E

Dilution buffer (pH8.5)

8 ml $T_{0.1}E$

0.13 ml cresol red solution

14 µl NaOH

16 ml H_2O

PCR buffer 3

500 mM KCl

100 mM Tris-HCl (pH 8.3)

15 mM $MgCl_2$

Northern strip solution

H_2O/0.5% SDS

Dilution buffer/primer mix solution

2.5 ng/µl of primer mix in dilution buffer

20 X SSC

3 M NaCl

0.3 M Tri-sodium citrate

0.2 X SSC

0.03 M NaCl

0.003 M Tri-sodium citrate

2 X SSC

0.3 M NaCl

0.03 M Tri-sodium citrate

Wash solution

0.5 X SSC

1% N-Lauroyl Sarcosine

50 X TAE (1litre)

242 g Tris

0.1 M EDTA

57.1 ml glacial acetic acid

Made up to 1,000 ml with milli-Q water

6 X loading dye

0.25% bromophenol blue

0.25% xylene cyanol

15% Ficoll

Hybridisation buffer

6 X SSC

2 mg/ml polyvinylpyrrolidone

2 mg/ml Ficoll

2 mg/ml BSA

50 mM Tris-HCl (pH7.4)

1% N-Lauroyl Sarcosine

10% w/v dextran sulphate

LB broth

10 mg/ml bacto-tryptone

5 mg/ml yeast extract

10 mg/ml NaCl

pH 7.4

LB culture medium (for BAC clones)

92.4 ml LB broth

7.5 ml 100% glycerol

0.1 ml 25 mg/ml chloramphenicol

Strip solution I

0.4 N NaOH

Strip solution II

0.1 X SSC

0.2 M Tris-HCl (pH 7.4)

1% N-Lauroyl Sarcosine

Marker mix

1.5 µl Analytical marker DNA wide range

0.1 µl DNA molecular weight marker V

4.2 µl 6 X loading dye

19.2 µl $T_{0.1}E$

10 X TS buffer

260 mM Tris-HCl (pH 9.5)

65 mM $MgCl_2$

PE 10 X PCR buffer

500 mM KCl

100 mM Tris-HCl (pH 8.3)

15 mM $MgCl_2$

0.01% (w/v) gelatine

CGT stop mix

ddCTP, ddGTP, ddTTP and dATG
at a concentration of 5 mM each

ACT stop mix

ddATG, ddCTP, ddTTP and dGTP
at a concentration of 5 mM each

ACG stop mix

ddATG, ddCTP, ddGTP and dTTP
at a concentration of 5 mM each

Chapter III

Sequence and transcript map of 20q12-13.2

3.1 Introduction

3.1.1 Strategies for gene identification

Genes are the basic units of genetic information and, as such, they have been at the centre of genome research. Initial efforts to construct human transcript maps were regional, often part of a positional cloning project. Experimental approaches such as cDNA selection and exon trapping were successfully used to identify disease genes for several monogenic disorders including Duchenne muscular dystrophy (Monaco *et al*., 1986) and cystic fibrosis (Rommens *et al*., 1989). These methods, which typically yield fragments rather than entire transcripts, are expensive, time-consuming and do not guarantee the identification of all genes. The latter was demonstrated during the construction of transcript maps of the Familial Mediterranean Locus. Two independent groups (Centola *et al*., 1998; Bernot *et al*., 1998) constructed transcript maps of this region, in parallel, using the same gene identification approaches (cDNA selection, exon amplification/trapping, EST mapping, limited sequencing and computational gene prediction). Within the ~225 Kb of overlap between the two maps, each group identified genes that were not identified by the other. In addition, obtaining the overall structure of the identified gene fragments (exact exon/intron boundaries and sizes, splice sites and regulatory elements) requires further work.

The emerging finished sequence of the human genome provides a solid foundation for the systematic identification of genes. In general, large-scale gene identification projects use

two main sequence analysis approaches to identify genes: sequence similarity searches and *ab initio* predictions.

The success of gene identification using similarity searches is heavily dependent on the size and quality of the available data sets. The availability of large numbers of partial 5' and 3' cDNA sequences in the form of ESTs (Adams *et al*., 1991) greatly enhances gene identification. As of May 2002, the dbEST database (http://www.ncbi.nlm.nih.gov/dbEST/; release 052402) contained 11,779,868 entries from 403 organisms. Using such extensive collections of both human and non-human ESTs offers an increased chance of identifying genes that are expressed at low levels, or have a restricted pattern of expression. Note that although ESTs are a valuable gene identification tool, they are not 'full length' mRNA sequences.

In addition to the known genes, systematic efforts to sequence entire cDNA clones include the KIAA collection (Nomura *et al*., 1994), the RIKEN collection (The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium, 2001), the Genoscope collection (http://www.Genoscope.cns.fr), the DKFZ collection (Wiemann *et al*., 2001) and the Mammalian Gene Collection (Strausberg *et al*., 1999). Comparison of the human genome sequence with genome sequences of other organisms can also be used to identify conserved regions, which could represent exonic sequences (also see chapters I and IV).

Similarity searches at the protein level compare the translated genomic sequence (all six frames) to known and predicted proteins from a variety of organisms, including the protein indices of fully sequenced model organisms such as *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Arabidopsis thaliana*. With

species evolutionarily distant to humans, homology is typically maintained only at the protein level.

Ab initio prediction software such as Genscan (Burge and Karlin, 1997) and FGENESH (Salamov and Solovyev, 2000) can be used to predict genes independently from expression data. The use of multiple prediction software can help identify the sequences that are more likely to contain exonic regions and filter out some of the over-predictions (Burset and Guigo, 1996; Claverie, 1997; Reese *et al*., 2000; also see section 1.4.2). Genic regions can also be identified using software algorithms that scan the genome sequence for gene-related features such as CpG islands (Micklem, unpublished), promoters (Scherf *et al*., 2000) and transcription start sites (Down and Hubbard, 2002).

Recently, automated approaches that combine *ab initio* predictions and similarity search results to annotate human genes received a lot of attention. A fully automated approach enables the fast analysis of large genomes such as the human genome and provides gene sets that are free from human error.

Several reasons prompted the pursuit of this approach. During the process of manual gene identification, computational analysis (*ab initio* predictions and homology searches) is followed by manual annotation (evaluation of evidence and determination of the exon/intron structure of genes by a trained annotator). The rate-limiting step in this process is manual annotation, an arduous task that is prone to human error. Erroneous annotations contaminate the sequence databases and avoiding this requires a high level of expertise from each annotator. This is very difficult to achieve, since large-scale gene identification and annotation projects require a large number of annotators, who have to be well trained and co-ordinated to ensure the accuracy and consistency of annotation.

Two catalogues of human genes were generated independently by automated annotation of the draft genome sequence (IHGSC, 2001; Venter *et al*., 2001). In both cases the automated annotation was performed using similarity searches and *ab initio* predictions and in each case gene catalogues of approximately 30,000 genes were constructed. Although this figure appears to be very close to the current estimate of the total number of human genes (~35,000), both groups stressed that their gene lists were far from complete. Issues raised included fragmented genes and the annotation of pseudogenes as genes. Comparison with well-defined sets of genes showed that the IHGSC gene list contained 60% of novel genes and that on average, 79% of each gene was detected (IHGSC, 2001). Venter *et al*. (2001) concluded that "extensive manual curation to establish precise characterisation of gene structure will be necessary to improve the results from this initial computational approach". Comparisons between the two gene sets confirmed their incomplete nature (Hogenesch *et al*., 2001).

In the absence of algorithms capable of correctly identifying all genes with high confidence, groups involved in large sequencing projects rely on the manual approach. The sequence of all finished chromosomes to date (20, 21 and 22) has undergone intense manual annotation to generate detailed lists of gene features (Deloukas *et al*., 2001; Hattori *et al*., 2000; Dunham *et al*., 1999). In these studies, gene structures were identified with high confidence, but not all exons/genes were found. One reason for this was that the 5' ends of genes are under-represented in most EST collections. In other cases, the only available evidence was non-human expressed sequences, or homology with paralogous proteins. As a result, in order to generate an accurate and complete list of all genes, an experimental approach is required to confirm and extend annotated genes and to discover those missed by the annotation.

3.1.2 Overview

This chapter discusses the sequence analysis a 10 Mb segment of chromosome 20q12-13.2. The contiguous genomic sequence was used to study the genomic landscape of this region through analysis of the GC and repeat content. A combination of *ab initio* predictions and homology searches were used to identify coding regions and generate a first generation transcript map. This map was then refined by experimental confirmation and extension of the annotated gene structures. Three different experimental approaches were used, each with specific strengths and weaknesses.

Various features of the annotated gene structures were examined, for example exon/intron structure and splice sites, alternative transcripts and polyadenylation signals. Software algorithms were used to scan the sequence for gene-related features such as CpG islands, promoters and transcription start sites, and the generated data were correlated with the annotated genes.

Analysis of the 20q12-13.2 proteome was also performed. The translated ORFs of annotated genes were analysed using InterProScan (Zdobnov and Apweiler, 2001) to look at the distribution of known protein domains. The data from 20q12-13.2 was also compared to the proteomes of six organisms (including humans) to investigate whether 20q12-13.2 is enriched in genes encoding proteins with particular domains.

Two approaches were taken to estimate how complete is the current annotation. One was based on a comparative analysis with draft genomic sequences from the mouse *Mus musculus* and the puffer fish *Tetraodon nigroviridis* (the two sets were not used for the annotation of the region). The conserved sequences were studied and correlated to estimate the number of exons that remain un-annotated. The other type of analysis was

based on *ab initio* predictions. Genscan (Burge and Karlin, 1997) and FGENESH (Salamov and Solovyev, 2000) predictions were compared to the annotation and the number of exons exactly predicted by both algorithms was measured. This data was then used to extrapolate the number of missed exons that remain un-annotated (also discussed in chapter IV).

3.2 Sanger annotation pipeline

Sequence analysis of the whole of chromosome 20, including my region of interest (20q12-13.2), has been an ongoing process and as such, time points are often difficult to define. This section gives a short summary of my contribution in the standard Sanger annotation of the whole chromosome. Section 3.3 describes the experimental approaches I used to confirm and extend genes in 20q12-13.2, whereas the 20q12-13.2 transcript map generated by combining the results from all analyses is presented in section 3.4.

The euchromatic sequence of chromosome 20 was determined by sequencing a set of 629 minimally overlapping clones. The finished non-redundant sequence comprises 59,187,298 bp and is assembled in 6 contigs (Deloukas *et al*., 2001).

Automated computational analysis proceeded on a clone-by-clone basis (Figure 3.1; references for software used and names of people involved are summarised in chapter II). The analysis files were imported in an implementation of ACeDB (humace) and displayed graphically. Based on the available supportive evidence (e.g. EST, mRNA and protein homologies), the annotators defined gene structures (Figure 3.2). To ensure a uniform and high quality annotation we re-checked the analysis of all 629 clones. The annotators implemented the proposed alterations and the final version of annotation was submitted to EMBL.

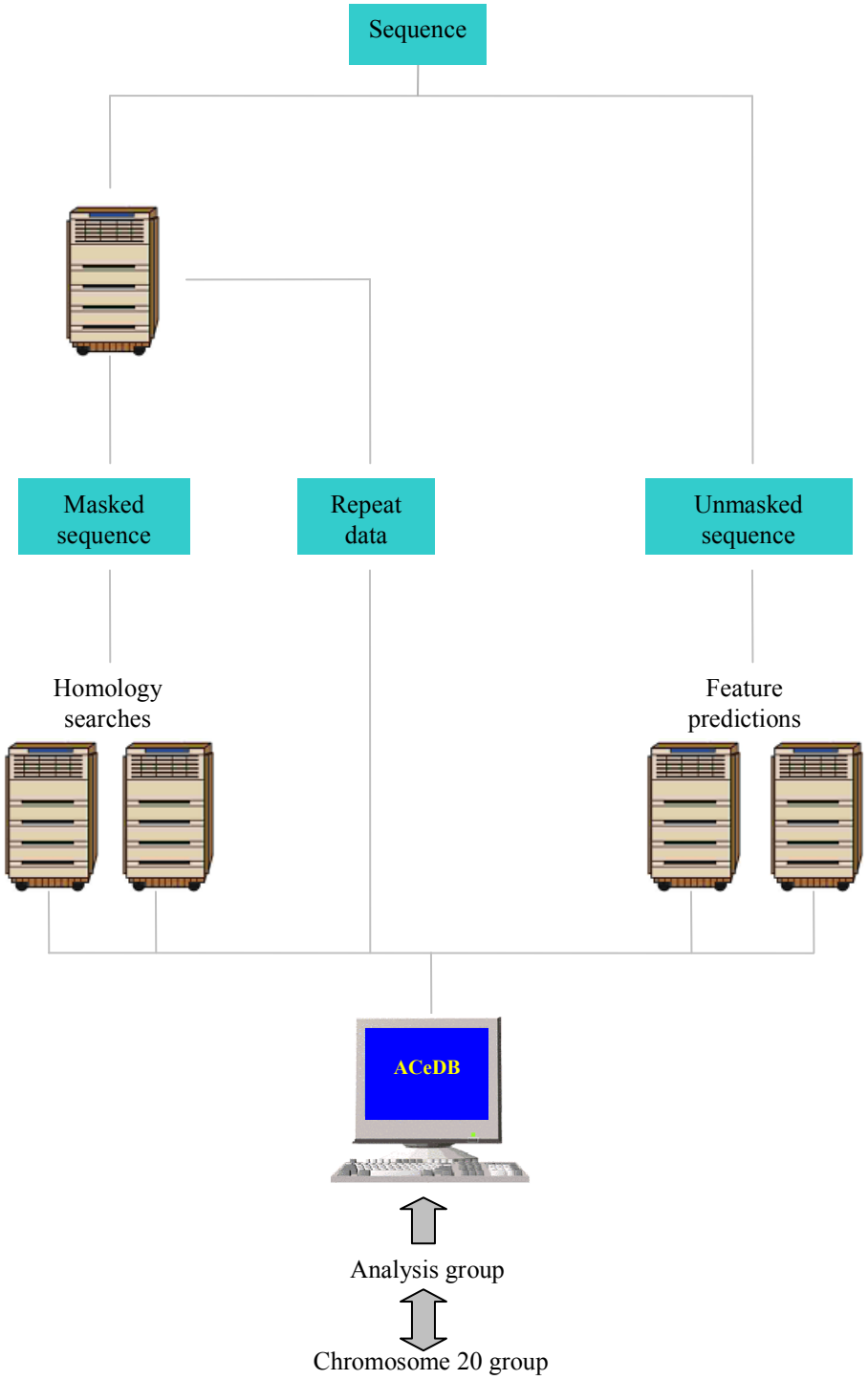


Figure 3.1: Sequence analysis pipeline. Software packages were used to predict gene structures and CpG islands in the unmasked finished sequence. RepeatMasker was used to identify repeats. The masked sequence was used to perform homology searches. All generated data was visualised in an ACeDB database and used for gene annotation.

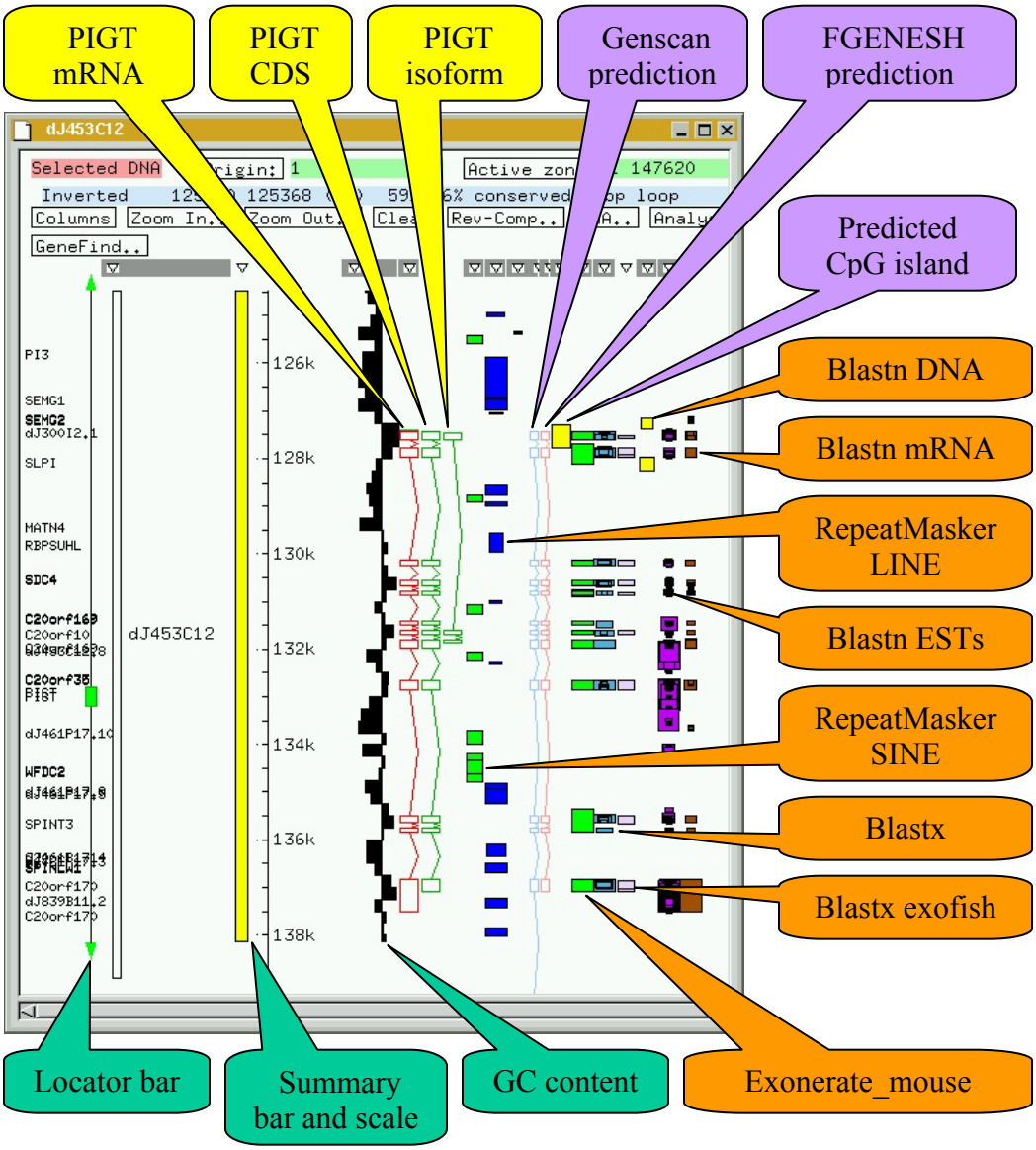


Figure 3.2: ACeDB view of sequence features. Orange description boxes indicate homologies, lavender boxes show prediction package data, green boxes show ACeDB general features and the yellow boxes indicate the manual annotation. The region shown here surrounds the PIGT gene locus.

A total of 895 structures were annotated in the finished sequence of chromosome 20 (Deloukas *et al*., 2001). The annotated structures were divided to five groups: (1) "known" genes: those that are identical to known human complementary DNA or protein sequences (all known genes were in the LocusLink database, http://www.ncbi.nlm.nih.gov/LocusLink); (2) "novel" genes: those that have an open reading frame (ORF), are identical to human ESTs that splice into two or more exons, and/or have homology to known genes or proteins (all species); (3) "novel" transcripts: genes as in (2) but for which a unique ORF cannot be determined; (4) "putative" genes: sequences identical to human ESTs that splice into two or more exons but without an ORF; and (5) "pseudogenes": sequences homologous to known genes and proteins but with a disrupted ORF.

3.3 Experimental confirmation of 20q12-13.2 genes

A lab-based approach was used to confirm and extend the first-pass annotation of the region. Since the annotation of all (except one) known genes was supported by very strong evidence (identical cDNA sequences), my efforts focused on the novel genes. In the absence of a "complete" gene structure (ORF with a predicted starting methionine, a 5' end and a 3' end), or when the annotation lacked human expressed evidence, I attempted to isolate and characterise cDNA fragments by sequencing. The generated sequences were aligned to the genomic sequence and used to improve annotation.

Three different experimental methods (vectorette, SSP-PCR and RACE) were used to isolate cDNAs of interest. The three methods permit amplification of sequences that are only partially known and allow uni-directional walking from known to unknown gene regions. This is achieved through amplification of the DNA of interest using a single sequence-specific primer and a generic primer (see chapter II). The amplified DNA can be separated on an agarose gel and the band of interest excised and sequenced.

3.3.1 Vectorette

The technique of vectorette cDNA-end isolation is an adapted version (Collins, unpublished) of the original vectorette PCR (Riley *et al*., 1990). The method is applied to pools of modified cDNAs that have DNA "bubbles" ligated on both ends. Because of the DNA "bubble", the vectorette method has the advantage of screening highly complex cDNA pools whilst retaining high specificity. In addition, the relatively simple experimental protocol allows the parallel screening for several genes in a large number of cDNA pools (high-throughput method).

I prepared two vectorette cDNA pools (as described in chapter II) from Adult Heart (Invitrogen) and Adult Lung (Clontech) cDNA libraries. The final step of the Adult Heart cDNA pool construction is shown in Figure 3.3. The generated pools became part of the Sanger vectorette library resource and have since been extensively used for cDNA isolation by other research groups. In total, seven vectorette libraries were used in this study.

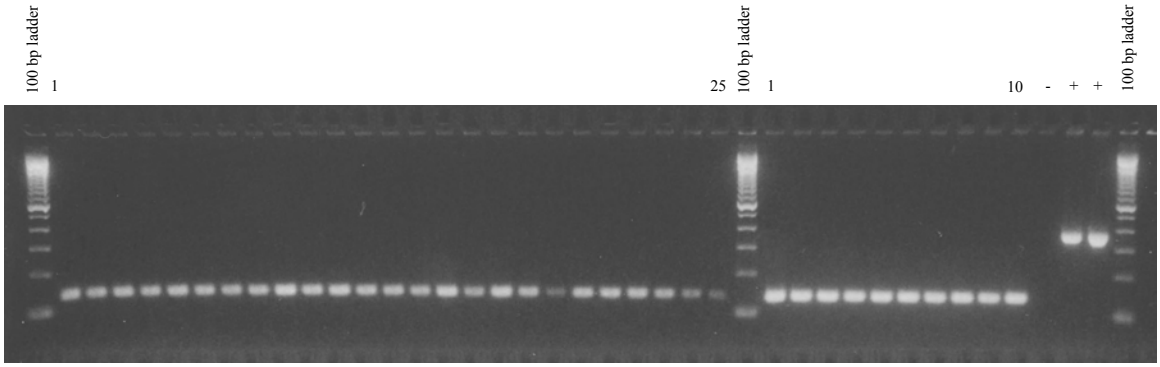


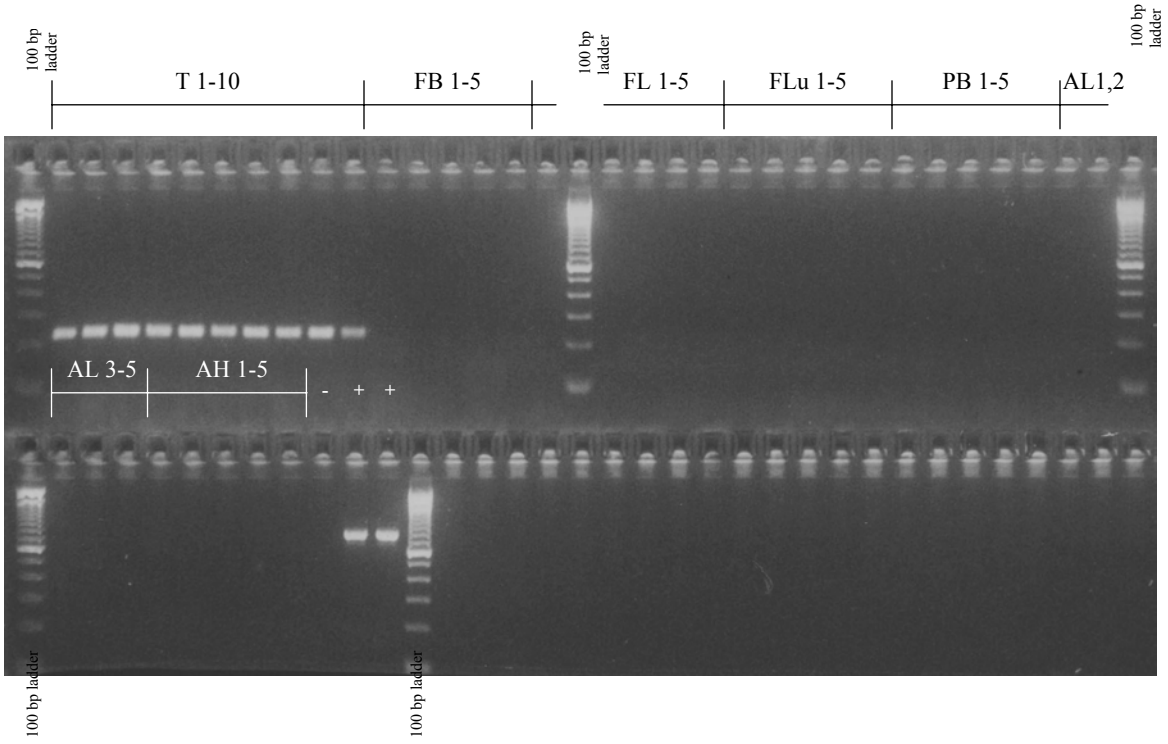
Figure 3.3: Vectorette library construction (final step PCR screening to check cDNA recovery and contamination). Newly constructed vectorette pools (25 from plated cultures and 10 from liquid cultures) generated from an Adult Heart cDNA library were PCR screened with the stSG71396 primer set. The primers are designed in a coding region of KIAA1247 gene, across a 240 bp intron. The amplified fragments from genomic templates are 375 bp long whereas fragments from cDNA templates are 135 bp long.

The vectorette method was used to isolate and sequence cDNA-end fragments (probes) for twenty novel genes across the region of 20q12-13.2 (Appendix 6). An example is shown in Figure 3.4. The generated probes were gel cleaned and sent for bi-directional sequencing. Aliquots were also kept as a reference repository. In total, 296 sequence reads were obtained from 258 probes (sequence success rate 296/516 = 57%). 226 of the

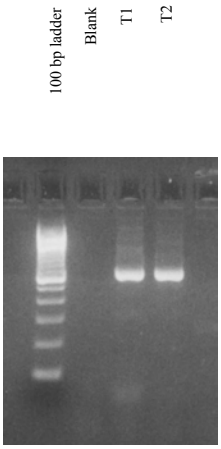
generated sequence reads provided novel expressed data and were used to confirm and extend annotation. They were also submitted to the EMBL database as ESTs.

Attempts to confirm the annotation of six genes that are not supported by human splicing expressed data (also see section 3.4.2) were unsuccessful because either a positive signal from the PCR screens was not obtained, or due to the inability to amplify and isolate gene specific cDNA-end fragments from positive pools. In addition, the structures of four other novel genes (LPIN3, C20orf100, R3HDML and C20orf137) and one known gene (SPINT3) remained incomplete for the same reasons.

A.



B.



C.

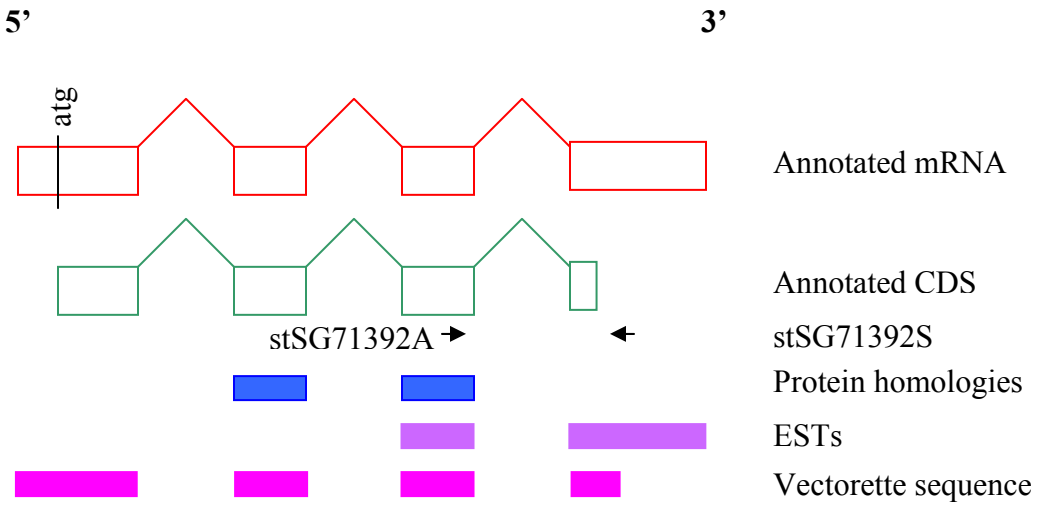


Figure 3.4: Example of cDNA-end isolation using the vectorette method. Initial annotation of SPINLW1 was based on homology with splicing ESTs (supporting two exons) and various protein homologies (supporting one EST-supported exon, and an exon further 5'). (A) the stSG71392 primer set was used to PCR screen the vectorette cDNA pools. The PCR products were loaded on the gel in the following order: Testis, Fetal Brain, Fetal Liver, Fetal Lung, Peripheral Blood, Adult Lung, Adult Heart. (B) Positive testis pools 1 and 2 were then used as templates with the stSG71392 sense primer in a vectorette reaction. The PCR products were separated using gel electrophoresis. The fragment from Testis 1 (at 600 bp) was gel purified, sequenced and the generated sequence (sccd1284.224) was aligned to the genomic sequence. (C) Schematic representation of the revised gene annotation (not to scale). The alignment confirmed the three EST and/or protein-supported exons and identified an additional exon further upstream.

3.3.2 Single specificity primer PCR

The technique of SSP-PCR end-sequence isolation from cDNA libraries is an adaptation (Bye and Rhodes, unpublished) of the original SSP-PCR (Shyamala and Ames, 1989; Shyamala and Ames, 1993).

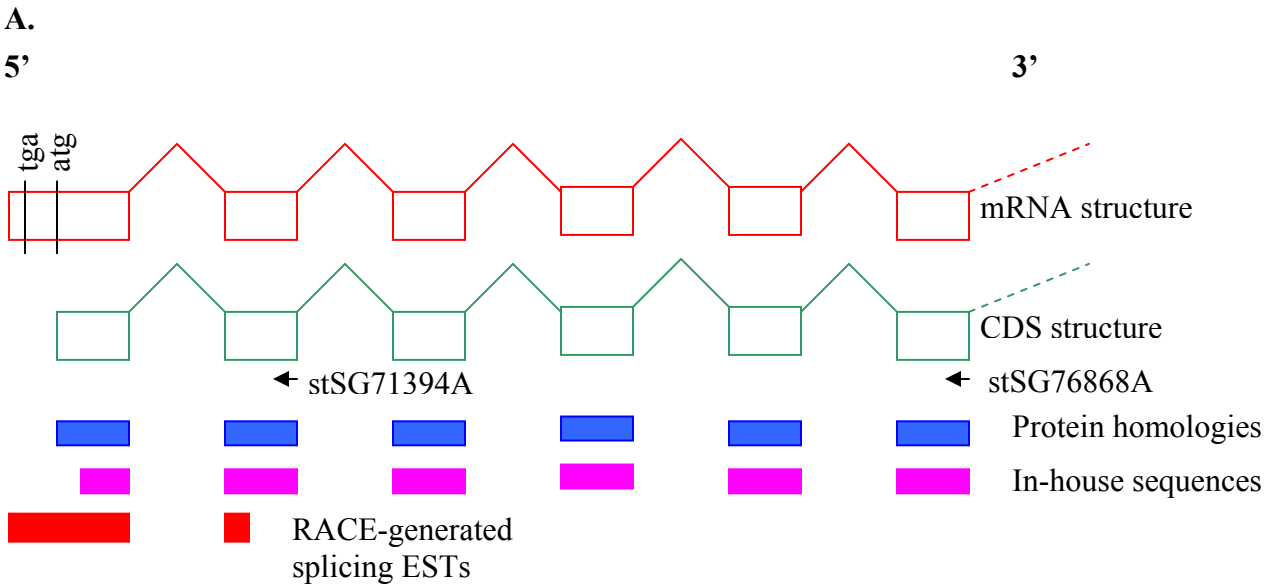
Unlike vectorette, SSP-PCR does not require specially adapted cDNA pools and can use boiled cDNA clones as templates for end-sequence isolation. To expand the available Sanger library resource I generated cDNA pools from an Adult Heart (Invitrogen) cDNA library. During this project, eighteen SSP-PCR libraries were available for screening and cDNA isolation (see chapter II).

Compared to vectorette, end-sequence isolation using SSP-PCR requires an additional PCR amplification step. This additional step makes the application of SSP-PCR less amenable to parallel analysis of several genes and more expensive due to the requirement for extra (nested) primer sets. Thus, SSP-PCR was used only at the beginning of this project to isolate nine cDNA-end sequences corresponding to three genes (C20of169, C20orf35 and PIGT) (Appendix 6). The generated sequences were used to confirm and expand the annotations and identify novel isoforms (section 3.4.6).

3.3.3 RACE

Due to the 3' end bias of the available EST collections, determining the most 5' end of genes is probably the most challenging step in gene annotation. In addition, experimental confirmation of the 5' end requires full-length cDNAs. Since our vectorette and SSP-PCR libraries were not enriched in full-length cDNAs, I applied RACE on two commercially available "full-length" cDNA pools (Clontech Marathon Adult Brain and

Testis) to extend the 5'-end annotation of incomplete genes. Gene specific primers were designed and used to isolate 5' cDNA-ends for fourteen novel genes (Appendix 6) and generate 31 and 28 quality sequence reads from Adult Brain and Testis cDNA pools, respectively. An example is shown in Figure 3.5.



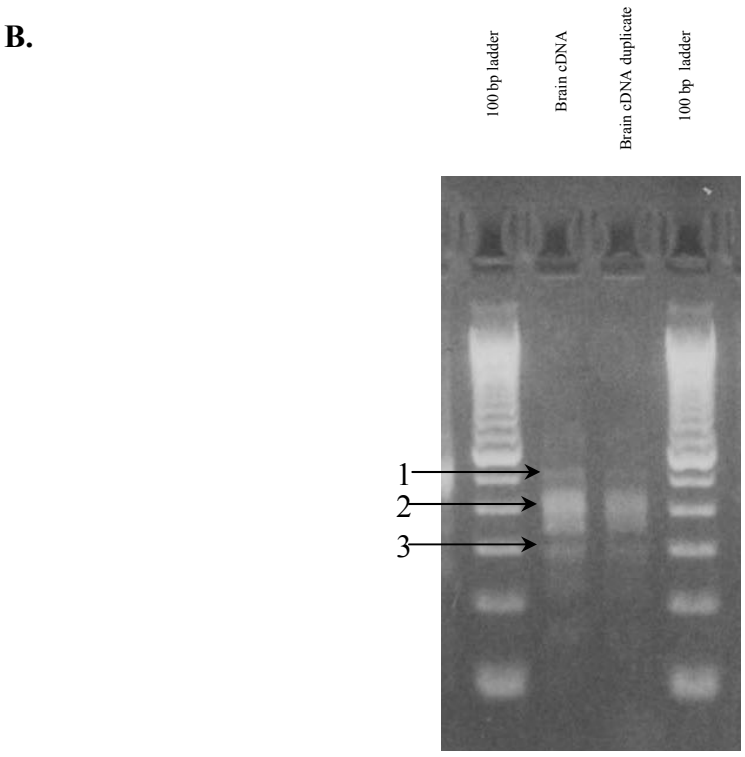


Figure 3.5 (previous page): Enhancing the annotation using RACE. (A) Figure shows the first six exons of C20orf119 (not to scale). Similarities with polyA-binding proteins from several organisms (blue boxes) were used to annotate a CDS (structure shown in green). cDNA fragments isolated from Fetal Brain vectorette pools were used to confirm exons 2 to 6 as well as part of exon 1 (pink boxes). Adult Brain Marathon cDNA was used to apply RACE with the stSG76868 and stSG71394 (nested) antisense primers, in duplicate. (B) The RACE products of the nested (second) amplification step were separated on an agarose gel. Three bands (1), (2) and (3) were excised and used in PCR re-amplification reactions. Sequence (red boxes) generated from the re-amplified products (sccd4368, sccd4370 and sccd4372, respectively; Appendix 4) confirmed the whole of exon 1 and part of exon 2. sccd4368 and sccd4370 extended beyond the annotated CDS and an mRNA with a 50 bp 5'UTR was annotated. The presence of a stop (tga) codon (within the annotated 5' UTR) in frame with the predicted ORF indicates that the annotation includes the start of the gene's coding sequence.

3.3.4 Summary of experimental efforts

Vectorette, SSP-PCR and RACE were used to confirm and/or extend the annotation of 24 novel genes mapping in 20q12-13.2 (Appendix 6). All isolated probes were used to generate a repository of cDNA fragments (Appendix 3) and generate 364 informative sequence reads. 294 reads containing novel expressed sequences were submitted to EMBL as novel ESTs (Appendix 4).

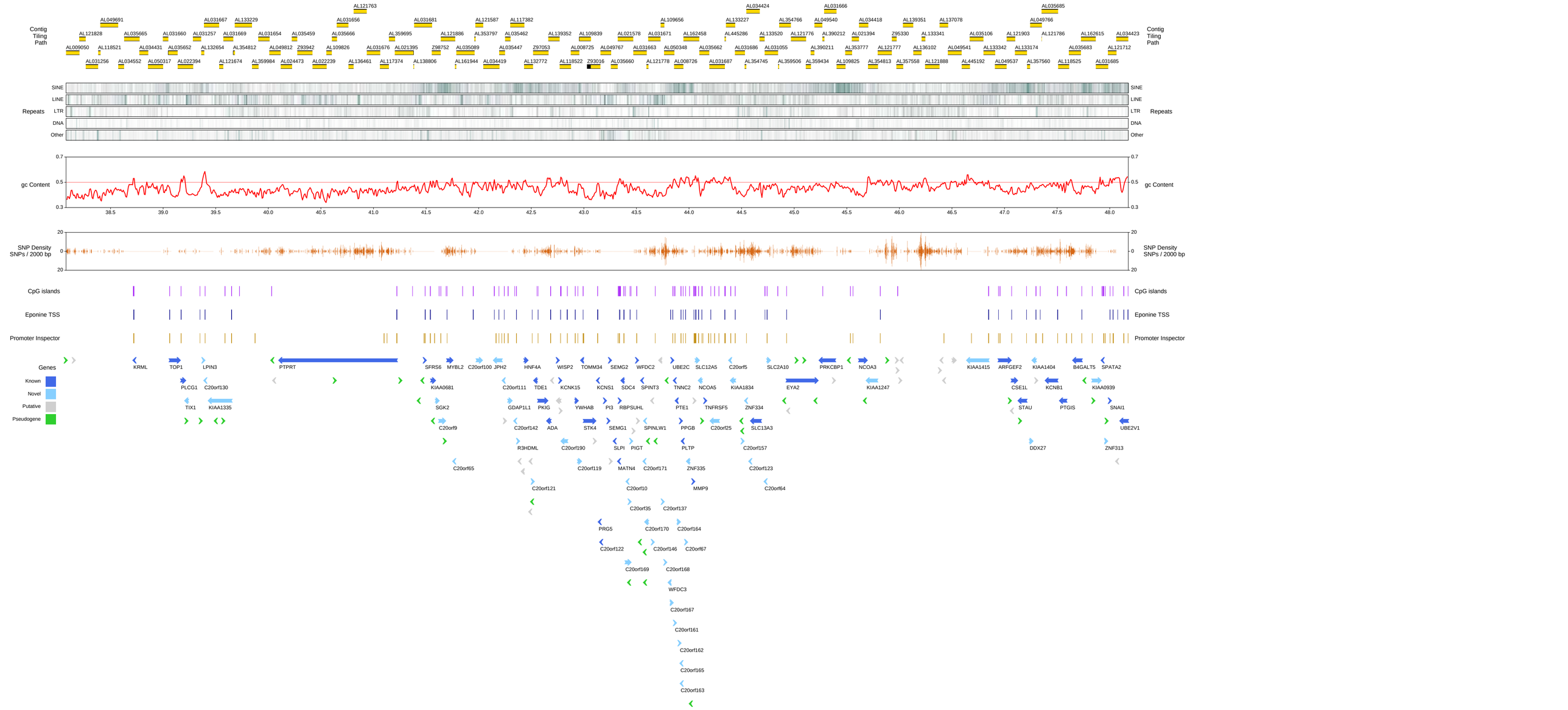
An additional 142 good sequence reads corresponding to novel genes mapping across the whole of chromosome 20 (but outside 20q12-13.2) were also isolated and subsequently submitted to EMBL (Appendix 4).

3.4 Combining all computational and experimental data – re-annotation of 20q12-13.2

The finished sequence of the 111 overlapping sequence clones from 20q12-13.2 provided the basis for systematic re-analysis of the region. Combining all available data to provide a comprehensive annotation of the region required a single metric system that can easily be accessed and manipulated. I manually edited the individual clone sequences to construct a virtual contiguous sequence contig that spans 10,099,164 bp and starts at base pair position 1 of the sequence AL009050 (PAC clone 191L6). The end of the contig is at base pair position 113,589 of the sequence AL034423 (PAC clone 1185N5). The sequence is stored in the chromosome 20 implementation of the ACeDB database, 20ace (available at http://webace.sanger.ac.uk/cgi-bin/webace?db=acedb20 under the name "CDR_region").

A new version (v_6_2001) of RepeatMasker was used to identify repeats. Genscan, FGENESH and CPGFINDER were used to predict genes and CpG islands. In addition, I used PromoterInspector (Scherf *et al*., 2000) to predict promoters and Eponine (Down and Hubbard, 2002) to predict Transcription Start (TS) sites. Incorporating and displaying all available computational (similarity searches and *ab initio* predictions) and experimental data on a contiguous virtual contig allowed the joining of annotated gene structures spanning two or more clone sequences. Storing the data in ACeDB provided an easy means to access and complement the data with newly emerging information. Data was exported from the ACeDB database and manipulated in Microsoft Excel. Figure 3.6 was generated by importing the generated data into an Ensembl database (James Gilbert).

Figure 3.6 (fold-out): The sequence map of human chromosome 20q12-13.2. The various features are shown from top to bottom as follows: (1) The finished sequence of each clone in the tiling path as a yellow line. Sequence positions relative to the whole-chromosome sequence are indicated in megabases along the x-axis of GC content (see 3, below). (2) The distribution of the main types of repeats in the sequence. (3) Plot of the GC content of the sequence. (4) Plot of the SNP density long the sequence (as of January 2002). (5) The location of predicted CpG islands. (6) The location of predicted TS sites. (7) The location of PromoterInspector predictions. (8) The location of annotated gene structures. Right and left coloured arrows indicate gene structures on the + and – strand, respectively. Only known (dark blue) and novel genes (blue) are named.


Contig Tiling Path
AL121763
AL049691
AL031667
AL133229
AL031656
AL031681
AL121587
AL117382
AL109656
AL034424
AL031666
AL133227
AL354766
AL049540
AL034418
AL139351
AL137078
AL035685
AL049766
AL121828
AL035665
AL031660
AL031257
AL031669
AL031654
AL035459
AL035666
AL359695
AL121886
AL353797
AL035462
AL139352
AL109839
AL021578
AL031671
AL162458
AL445286
AL133520
AL121776
AL390212
AL021394
Z95330
AL133341
AL035106
AL121903
AL121786
AL162615
AL034423
AL009050
AL118521
AL034431
AL035652
AL132654
AL354812
AL049812
Z93942
AL109826
AL031676
AL021395
Z98752
AL035089
AL035447
Z97053
AL008725
AL049767
AL031663
AL050348
AL035662
AL031686
AL031055
AL390211
AL353777
AL121777
AL136102
AL049541
AL133342
AL133174
AL035683
AL121712
AL031256
AL034552
AL050317
AL022394
AL121674
AL359984
AL024473
AL022239
AL136461
AL117374
AL138806
AL161944
AL034419
AL132772
AL118522
Z93016
AL035660
AL121778
AL008726
AL031687
AL354745
AL359506
AL359434
AL109825
AL354813
AL357558
AL121888
AL445192
AL049537
AL357560
AL118525
AL031685
Repeats
SINE
LINE
LTR
DNA
Other
gc Content
0.7
0.5
0.3
38.5
39.0
39.5
40.0
40.5
41.0
41.5
42.0
42.5
43.0
43.5
44.0
44.5
45.0
45.5
46.0
46.5
47.0
47.5
48.0
SNP Density
SNPs / 2000 bp
20
0
CpG islands
Eponine TSS
Promoter Inspector
Genes
Known
Novel
Putative
Pseudogene
KRML
TOP1
LPIN3
PTPRT
SFRS6
MYBL2
C20orf100
JPH2
HNF4A
WISP2
TOMM34
SEMG2
WFDC2
UBE2C
SLC12A5
C20orf5
SLC2A10
PRKCBP1
NCOA3
KIAA1415
ARFGEF2
KIAA1404
B4GALT5
SPATA2
PLCG1
C20orf130
KIAA0681
C20orf111
TDE1
KCNK15
KCNS1
SDC4
SPINT3
TNNC2
NCOA5
KIAA1834
EYA2
KIAA1247
CSE1L
KCNB1
KIAA0939
TIX1
KIAA1335
SGK2
GDAP1L1
PKIG
YWHAB
PI3
RBPSUHL
PTE1
TNFRSF5
ZNF334
STAU
PTGIS
SNAI1
C20orf9
C20orf142
ADA
STK4
SEMG1
SPINLW1
PPGB
C20orf25
SLC13A3
UBE2V1
R3HDML
C20orf190
SLPI
PIGT
PLTP
C20orf157
DDX27
ZNF313
C20orf65
C20orf119
MATN4
C20orf171
ZNF335
C20orf123
C20orf121
C20orf10
MMP9
C20orf64
C20orf35
C20orf137
PRG5
C20orf170
C20orf164
C20orf122
C20orf146
C20orf67
C20orf169
C20orf168
WFDC3
C20orf167
C20orf161
C20orf162
C20orf165
C20orf163

3.4.1 Broad genome landscape

3.4.1.1 Repeats

In total, 49.62% of the sequence is occupied by repeats. Table 3.1 shows that interspersed repeats account for 96.8% of the repeat sequence. SINEs are the most abundant elements both in terms of number and coverage. Figure 3.7 shows the percentage of total sequence covered by the different repeat families whereas Figure 3.8 shows the percentage of repeat sequence covered by the different families (also see chapter IV).

Table 3.1: Number, coverage and density of different classes of repeats.

Repeat type	Number of elements	Coverage (%)	Density (Kb/element number)
SINEs	8,998	19.7	1.12
LINEs	3,852	16.47	2.62
LTR elements	1,956	7.81	5.16
DNA elements	1,727	3.93	5.85
Unclassified	13	0.14	777
Total interspersed repeats	**16,546**	**48.04**	**0.61**
Other	2,435	1.58	4.15
Total repeats	**18,981**	**49.62**	**0.53**

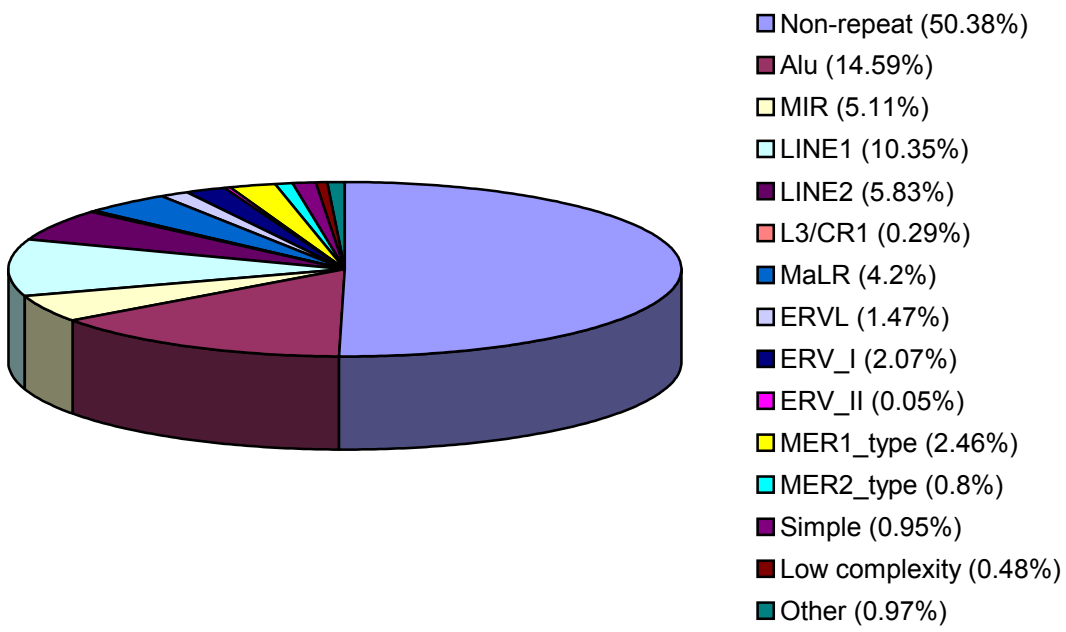


Figure 3.7: Repeat content distribution of 20q12-13.2.

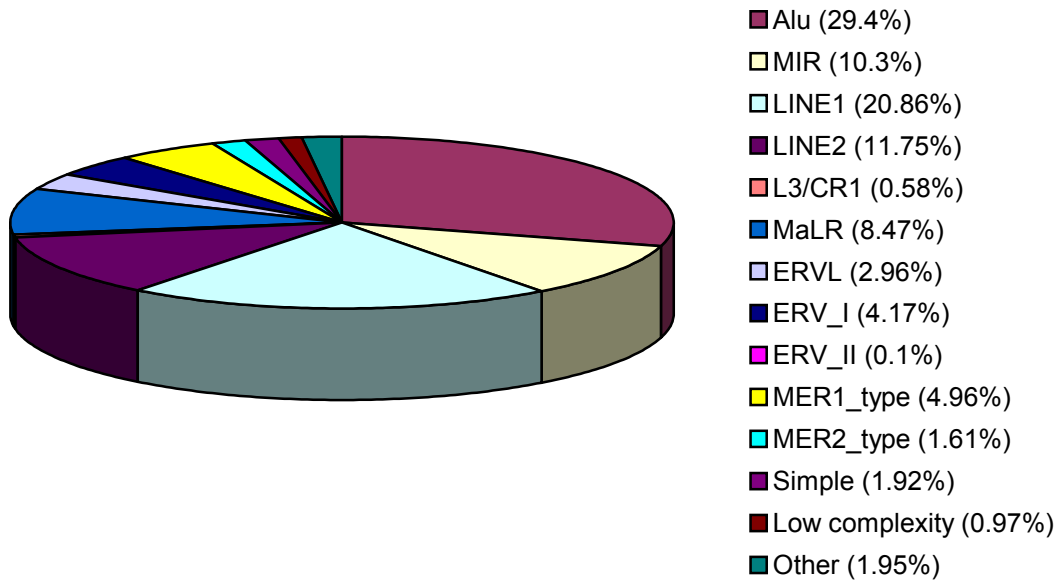


Figure 3.8: Repeat distribution for each family.

3.4.1.2 A segmental duplication

Two copies of a 60 Kb intrachromosomal duplication were found in the region between 7,780 Kb and 8,480 Kb. The two sequences were compared using Dotter (Figure 3.9).

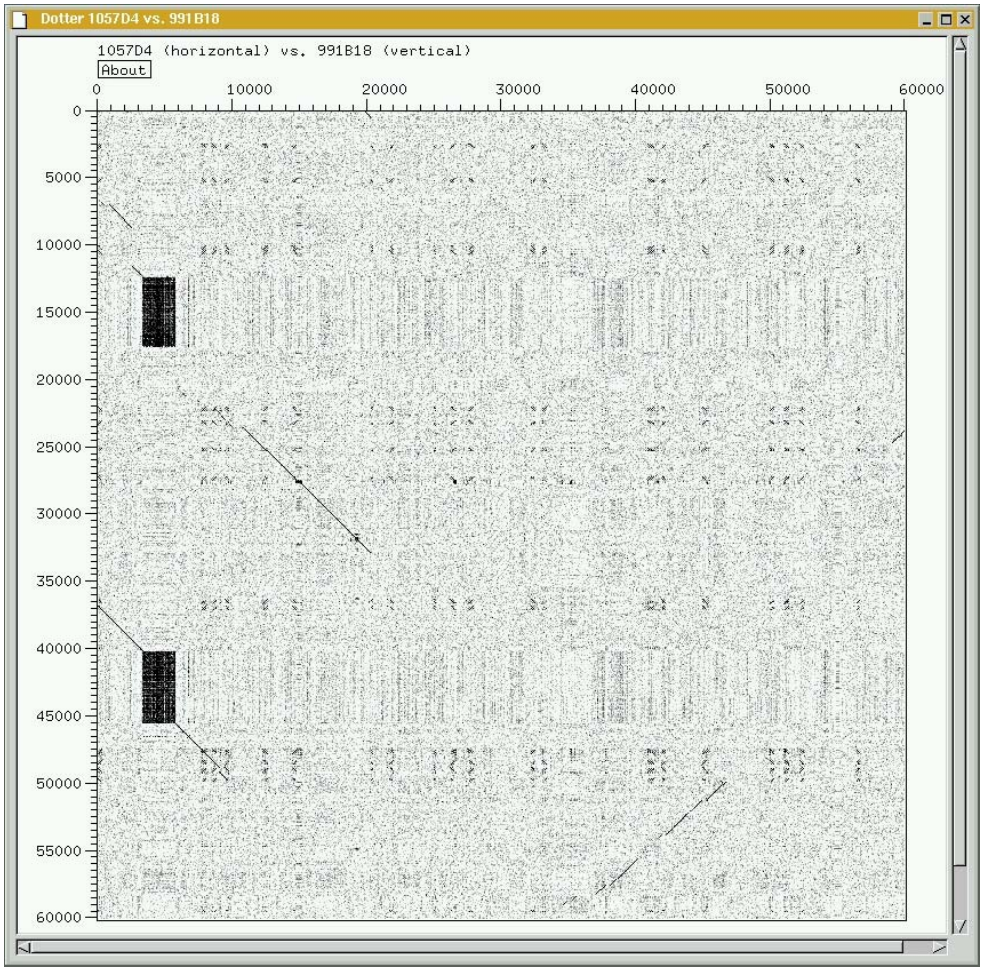


Figure 3.9: Dot plot (output from Dotter, Sonnhammer and Durbin, 1995) of the two regions present in the sequences of clones RP5-1057D4 and RP5-991B18 (sequence accession numbers AL121777 and AL049541). The non-linear output is probably due to rearrangements occurring after the duplication event.

The region between the duplicated segments contains seven putative genes but no coding genes. The more telomeric copy of the duplication contains an additional putative gene whereas a pseudogene is present in the more centromeric copy.

The steepest increase in recombination frequency across chromosome 20 is observed between genetic markers D20S178 and D20S176 (Deloukas *et al*, 2001). The region bordered by the duplications and the area defined by these genetic markers share significant overlap.

3.4.2 Supporting evidence for annotated loci

The evidence used for the annotation of gene structures is summarised in Table 3.2. I annotated 99 coding genes (48 known and 51 novel), 36 pseudogenes and 30 putative genes.

Less than 3% of the total annotated exons remain without human expressed-sequence evidence. Six novel genes (C20orf171, C20orf168, C20orf157, C20orf164, C20orf165 and C20orf123) are not supported by human splicing ESTs. C20orf171 is annotated as a three exon gene, whereas C20orf168 is annotated as a single exon gene. The annotation of both genes is based on similarities with proteins containing a WAP (Whey Acidic Protein)-type 'four-disulphide core' domain (IPR002221) and/or a Kunitz/Bovine pancreatic trypsin inhibitor domain (IPR002223). C20orf157 is annotated as a two exon gene, based on similarities with proteins containing zinc-finger domains.

C20orf164 is annotated as a two exon gene based on homology with a mouse RIKEN cDNA clone (clone 4921517A06; EMBL accession number AK014904) and several non-

splicing human ESTs. C20orf165 is annotated as a two exon gene, based on homology with a mouse RIKEN cDNA clone (clone 1700020C07; EMBL accession number AK006145) and a non-splicing human EST (EMBL accession number AA972728). C20orf123 is annotated as a two exon gene based on homology with a RIKEN cDNA clone (clone 4833422F24; EMBL accession number AK014751) and two non-splicing human ESTs (EMBL accession numbers D20888 and BG058578).

Annotation of the 36 pseudogenes was based on BLASTX homologies with a variety of different proteins. The annotation of twelve pseudogenes was based on homology to ribosomal proteins, whereas three pseudogenes (dJ450M14.1, dJ138B7.4, dJ1041C10.3) were based on BLASTX homologies with human predicted proteins of unknown function (translations of predicted ORFs of anonymous cDNA sequences).

The 30 putative gene structures were annotated using human splicing ESTs. Putative genes do not have a clearly detectable ORF and do not have any BLASTX homologies. Attempts to expand all these structures by the vectorette method failed.

Table 3.2: Supporting evidence for annotated features.

Gene features	Total	Forward strand	Reverse strand	Human EST evidence	Human cDNA evidence	Protein evidence
Known	48	25	23	47	47[1]	48
Novel	51	25	26	48[2]	37[2]	49
All coding genes	99	50	49	95	84	97
Putative	30	15	15	30	3	-
All transcripts	129	65	64	125	87	97
Pseudogenes	36	15	21	-	-	36
All features	165	80	85	125	87	133

[1]SPINT3 is supported by DNA submission X77166 and ESTs (AW118166 and AA812696)
[2]Three novel genes were annotated without any human expression data (C20orf171, C20orf168 and C20orf157).

3.4.3 First-pass expression data for novel and putative genes

The number of genes testing positive per vectorette cDNA library (section 3.3.1) is shown in Figure 3.10 and the correlation between the number of positive cDNA libraries and genes is shown in Figure 3.11.

The generated data suggests that 69 of the 81 (85%) attempted genes (novel coding and putative) are expressed in at least one of the seven cDNA libraries tested. The majority of novel genes appear to be expressed in three or more libraries, whereas the opposite is true for putative genes. The expression profile of the putative genes should be treated with

caution because all attempts to isolate cDNA fragments for putative genes from positive pools failed.

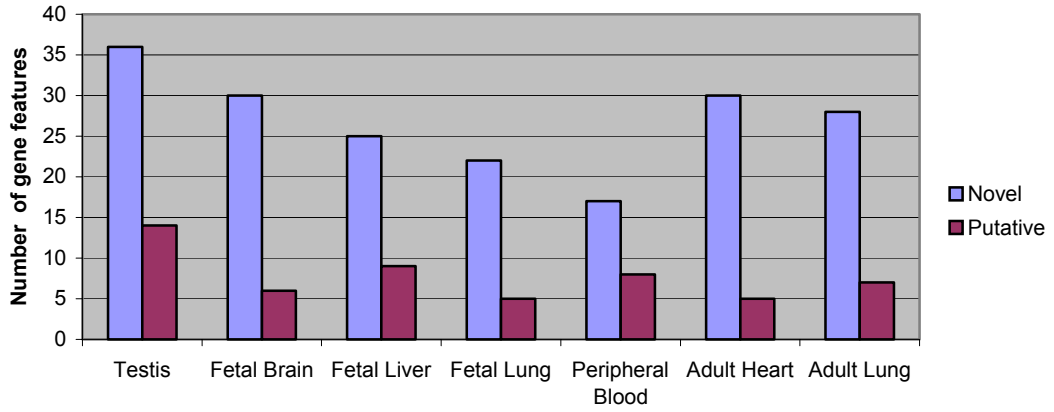


Figure 3.10: Positive genes per cDNA library.

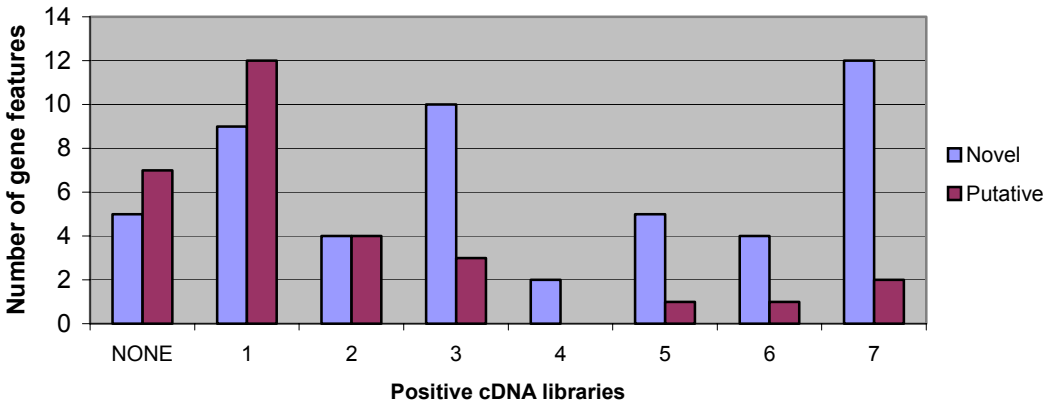


Figure 3.11: Positive cDNA libraries per gene.

PCR screens for five novel (C20orf171, C20orf137, C20orf165, C20orf157 and C20orf123) and seven putative (dJ1121H13.3, dJ781B1.4, dJ461P17.9, bA347D21.3, bA347D21.4, dJ66N13.1 and dJ66N13.3) genes did not produce any positive results (Figure 3.11). PCR screening of all available (eighteen; chapter II) SSP-PCR libraries did not yield any positive results either (data not shown).

3.4.4 Gene structures

Size characteristics of each category of annotated genes are given in Table 3.3 whilst their structural features are summarised in Table 3.4. The corresponding data for the three finished chromosomes is shown in Table 3.5.

44.5% of the region is occupied by coding genes but of that, only 5.7% represents mRNA sequences. The mean mRNA length is 2.8 Kb for coding genes and 502 bp for putative genes. Coding gene sizes were also found to vary substantially. For example SPINT3 is 384 bp long whereas PTPRT is 1,117,219 bp long.

Table 3.3: Size of gene loci.

Gene Features	Total length (Kb)	Mean length (bp)	Median length (bp)	Percentage of locus occupied by mRNA	Percentage of region occupied by mRNA
Known genes	3,072	66,797	20,554	4.3%	1.3%
Novel genes	1,423	31,639	17,481	8.7%	1.2%
All coding genes	4,495	49,411	18,302	5.7%	2.5%
Putative genes	207	6,690	2,596	7.2%	0.15%
Pseudogenes	30	830	796	-	-

Table 3.4: Structural features of annotated gene features.

Gene type	Exon number (mean)	Exon number (median)	Exon size (mean)	Exon size (median)	Coding exon number (mean)	Coding exon number (median)	Coding exon size (mean)	Coding exon size (median)
Known	10.78	9	267	136	10.57	9	154	126
Novel	9.71	6	285	132	9	5.5	179	128
All coding	10.2	7	276	135	9.7	6	166	127
Putative	2.6	2	188	149	-	-	-	-

	5' UTR (mean)	5' UTR (median)	3' UTR (mean)	3' UTR (median)	Intron size (mean)	Intron size (median)
Known	110	73	1,194	767	6,533	1,392
Novel	94	68	850	328	3,313	1,330
All coding	94	72	1,015	493	5,034	1,354
Putative	-	-	-	-	3,836	2,112

The average number of exons encoded by known and novel genes is approximately the same whilst their exon sizes are also very similar. Compared to coding genes, putative genes are smaller in size and have significantly less exons. On average, 3' UTR sequences are six times longer than coding exons and more than ten times longer than 5' UTR sequences. The longest 3' UTR annotated in the region is part of the PTPRT gene and spans 8,181 bp (average 3' UTR size is 1,194 bp).

Table 3.5: Structural features of genes annotated in chromosomes 20, 21 and 22 (reproduced from Deloukas *et al*., 2001).

Chromosome, gene type	Mean size (Kb)	Mean exon size (bp)	Mean exon number
Chr20, known genes	51.3	294	10.3
Chr21, known genes	57.0	-	-
Chr20, novel genes	25.1	278	5.7
Chr20, putative genes	9.1	217	2.5
Chr21, novel+putative genes	27.0	-	-
Chr20, all coding genes	34.7	283	7.1
Chr21, all coding genes	39.0	-	-
Chr20, pseudogenes	1.9	499	1.4
Chr20, all	27.6	292	6.0
Chr22, all	19.2	266	5.4

The mean size of all the annotated gene structures (including pseudogenes) in the region is 28.7 Kb and compares favourably with the 27.6 Kb and 19.2 Kb mean sizes reported for chromosomes 20 and 22, respectively (the average transcript size reported for chromosome 21 is 27 Kb compared to 36.7 Kb for 20q12-13.2).

3.4.5 Splice sites

Splice sites from all transcribed multi-exon structures (known, novel and putative genes) were used. All splice sites included in this study were annotated with high confidence and are supported by identical, human expressed sequences. The classification of the 875 3' intron-5'exon junctions and 868 3' exon-5' intron junctions is examined in Figure 3.12 and Figure 3.13 respectively. Note that 3' intron-5' exon sites are not available for the 5' exons and 3' exon-5' intron sites are not available for the 3' exons of each gene.

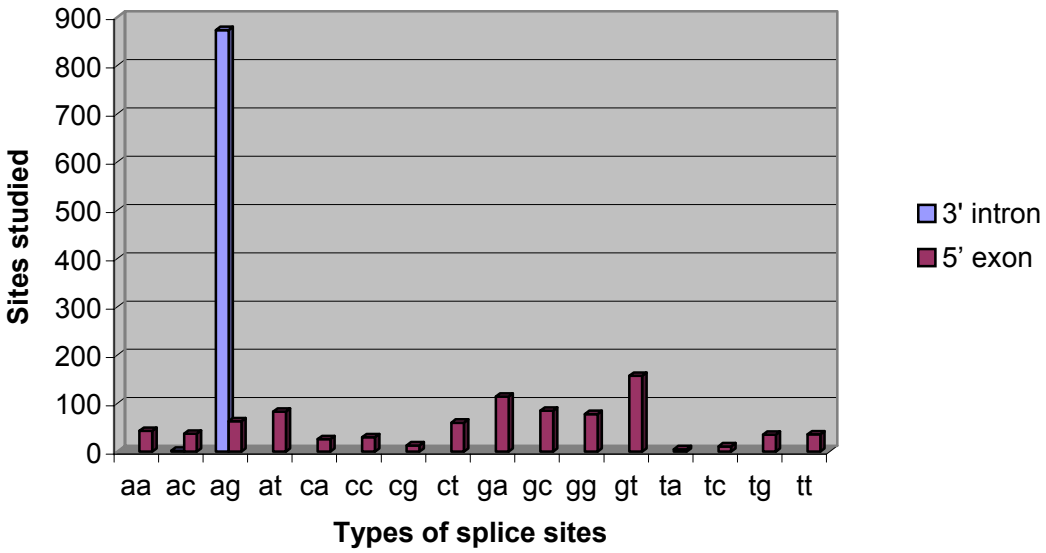


Figure 3.12: 3' intron-5' exon splice sites.

Two AC 3' intron sites were identified. The first one belongs to MATN4. Like the matrilin-1 gene, the human matrilin-4 gene contains an AT-AC intron between the two exons encoding the coiled-coil domain (Wagener *et al*., 1998). The second belongs to an AT-AC intron of the KIAA0939 gene and is supported by a cDNA sequence (EMBL accession number AB023156).

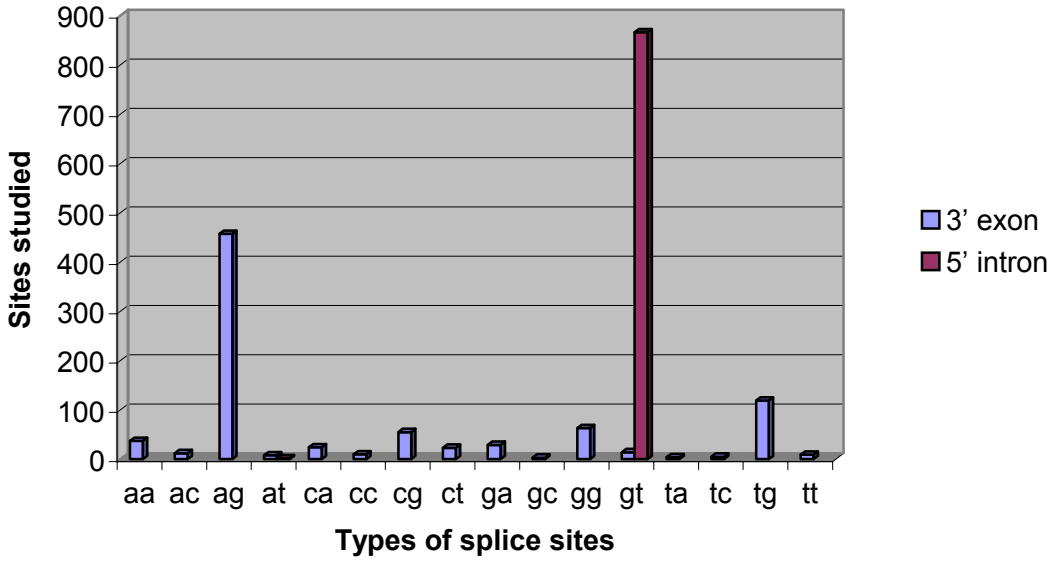


Figure 3.13: 3' exon-5' intron splice sites.

3.4.6 Splice isoforms

The wealth of available expression sequence data provides an ideal tool to examine the alternative splicing of genes. As part of the annotation process we identified splice isoforms for 43/99 (43%) coding (known and novel) genes and 2/30 (6.7%) putative genes. Because of the limited amount of supporting evidence, putative genes were excluded from the following analysis.

A total of 122 transcripts were annotated for the 43 coding genes showing alternative splicing. These loci were first annotated for the longest possible structure. Additional structures (variants) were subsequently annotated based on ESTs or cDNAs that differ

from the longest annotated structure. I did not attempt to use overlapping evidence to extend the variants, so their structures are mostly incomplete and span a relatively small number of exons. They can be categorised as follows:

i. Nineteen variants (corresponding to twelve genes) have an alternative 5' start site.

ii. Seventeen variants (corresponding to thirteen genes) either lack, or have an extra exon.

iii. Fifteen variants (corresponding to twelve genes) have an exon that differs in size.

iv. Twenty variants (corresponding to seventeen genes) have an alternative 3' exon.

v. Eight variants (corresponding to eight genes) were based on several ESTs terminating within the 3' UTR and may represent alternative polyadenylation sites.

A 47.5 Kb region of clone RP3-453C12 is of particular interest. Using vectorette, SSP-PCR, RACE and publicly available sequences, three genes were annotated: C20orf169 and C20orf35 on the forward strand and C20orf10 in their intergenic region, on the reverse strand. In total, ten different structures were annotated for the three genes. All variants identified in the region have different ORFs. Interestingly, two variants annotated for locus C20orf169 skip the annotated 3' end of C20orf169 and contain parts of exon 2 (variant 5) or exons 2, 3 and 4 (variant 4) of C20orf35 (based on a few EST and SSP-PCR sequences). Therefore, it is possible that C20orf10 may in fact be located in the intron of a bigger gene that is currently annotated as two different genes (C20orf169 and C20orf35).

3.5 Investigating the annotated 5' and 3' ends of coding genes

Experimental approaches can be used to investigate the completeness of gene annotation. For example, the annotated transcript size can be verified using Northern blot analysis (Figure 3.14). In addition the 5' and 3' annotated ends can be investigated by reporter assays. These experimental approaches are laborious, time-consuming and cannot be easily applied in a high-throughput manner. Therefore, I used a computational approach to investigate the completion of gene structures. Note that only the longest transcript of each gene was investigated (most of the annotated variants remain incomplete because their annotation is based on a limited amount of evidence).

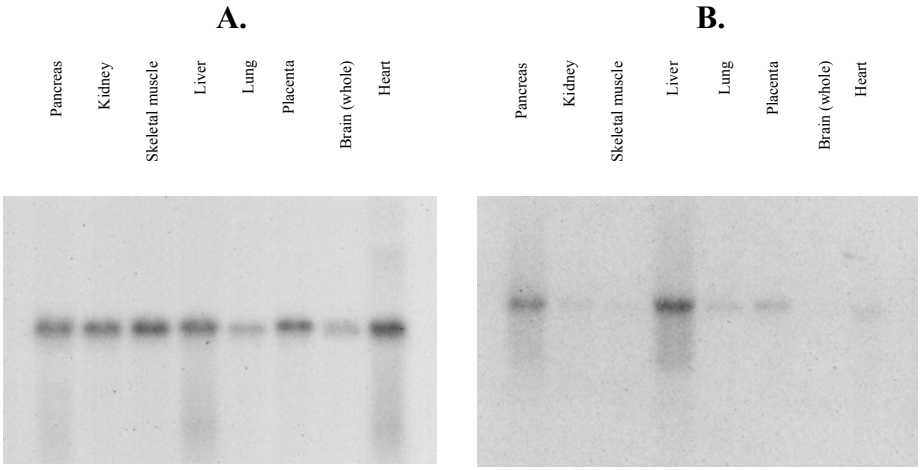


Figure 3.14: Northern blots. PCR-based probes were generated for the PIGT and SLC2A10 genes using the stSG85271 and stSG76895 primers respectively. The probes were radioactively labelled and used in hybridisation experiments with commercial Northern blots containing polyA RNA from eight tissues (A and B). The annotated PIGT mRNA is 2,148 bp long and the SLC2A10 mRNA is 4,126 bp long. Northern blot analysis suggests that the size of PIGT mRNA is ~2.4 Kb and the SLC2A10 mRNA ~4.4 Kb.

3.5.1 Polyadenylation signals (3' end)

The formation of nearly all vertebrate, mature mRNAs involves the cleavage and polyadenylation of the pre-mRNA 15-30 nucleotides downstream of a conserved hexanucleotide polyadenylation signal. The mechanism and regulation of mRNA polyadenylation is reviewed by Colgan and Manley (1997).

The 3' UTRs of the 99 annotated coding genes were examined for the presence of potential polyadenylation signals. Putative cleavage sites were recognised by alignment of 3' EST sequences to the mRNA through the graphical BLAST viewer Blixem (Sonnhammer and Durbin, 1994)

The highly conserved AATAAA hexanucleotide was identified near the 3' end of 63 UTRs, whereas its most common variant (ATTAAA) was present in 22 UTRs (only the most 3' hexanucleotide reported). To determine whether a different putative polyadenylation signal was present I scanned the sequence for the presence of ten other reported hexanucleotide variants (Beaudoing *et al*., 2000). The results are given in Table 3.6.

Table 3.6: Polyadenylation signals found in the 3' UTR of annotated coding genes (only the most 3' signal reported).

Signal	Number of genes
AATAAA	63
ATTAAA	22
TATAAA	2
GATAAA	1
AGTAAA	2
NONE	9 (the annotated UTRs of 6 genes are incomplete)

The absence of a polyadenylation signal from three genes with 'complete' UTRs could suggest that rare polyadenylation variants are present. This is in agreement with a recent study that involved thousands of 3' ends and where the authors reported that only 88% of the mRNA 3' ends contained a characteristic polyadenylation signal (Beaudoing *et al*., 2000). An alternative explanation could be that the annotation of these UTRs is incomplete and that the mRNA transcripts extend further 3'. Currently, there is no evidence to support this.

3.5.2 Promoters (5' end)

Since CpG islands (section 1.3.2.2) are associated with the 5' end of genes they can be used to estimate the completion of the annotated genes. I used the prediction program CPGFIND (Micklem, unpublished) and found 99 CpG islands. Predicted CpG islands were at least 400 bp long, have a GC content greater than 50% and an expected/observed CpG count of greater than 0.6.

PromoterInspector (PI) was also used to identify putative polymerase II promoters (Scherf *et al*., 2000). PI locates genomic regions of 0.2 Kb to 2 Kb that contain or overlap with polymerase II promoters. In a recent study on chromosome 22, PI predicted correctly 43% of known promoters (Scherf *et al*., 2001). PI predicted 93 promoters in the 10 Mb region of 20q12-13.2.

The sequence was also scanned for putative TS sites using the probabilistic TS site detector program Eponine (Down and Hubbard, 2002). Eponine is optimised for mammalian sequences and detects likely TS sites on the basis of the surrounding sequence. Eponine has a sensitivity of 40%, on the basis of an analysis of human

chromosome 22. Multiple predictions are often clustered, suggesting alternative TS sites for a gene. Eponine predicted 266 transcription start sites. Multiple TS sites predicted in DNA sequences less than 500 bp long were grouped in Eponine clusters. In total, Eponine predicted 67 such clusters.

3.5.2.1 Correlation of predictions and all gene structures

Predictions near the start of annotated structures may indicate the presence of a promoter, whereas predictions downstream of the 5' annotated start may indicate promoters for isoforms, or false positives. Predictions upstream of annotated loci may indicate promoters of genes that extend beyond the current annotation. Predictions not associated with annotated genes either correspond to promoters of un-annotated genes or false positives. Table 3.7 reports the associations between predictions and all annotated gene structures (coding genes, putative genes and pseudogenes).

Table 3.7: Correlation of predicted regions and annotation. Total predictions by each method are reported in column two. Predictions that map within annotations, or up to 1 Kb upstream are reported in column three. Predictions that map 1-20 Kb upstream of the annotations are reported in column four, whereas predictions that map elsewhere in the sequence are reported in column five.

	Total	In loci	Upstream (1-20 Kb)	Not associated
CpG islands	99	80/99 (80.8%)	8/99 (8.1%)	11/99 (11.1%)
PI	93	76/93 (81.7%)	5/93 (5.4%)	12/93 (12.9%)
Eponine	67	57/67 (85.1%)	3/67 (4.5%)	7/67 (10.4%)

Predictions were also investigated in terms of the type of gene structure they are associated with. Predictions were considered to be associated with a particular structure if mapped within, or up to 20 Kb upstream. Table 3.8 gives a summary of the data obtained.

Table 3.8: Correlation of types of annotated structure and predictions.

	Annotated gene features associated with predictions		
	Putative	Pseudogenes	Coding (known and novel)
CpG	5/30 (16.7%)	4/36 (11.1%)	63/99 (63.6%)
PI	4/30 (13.3%)	3/36 (8.3%)	56/99 (56.6%)
Eponine	2/30 (6.7%)	0 (0%)	48/99 (48.5%)

Predictions associated with putative genes provide further support to the notion that these loci are transcribed and that the ESTs used to annotate them are not artefacts of cDNA libraries. Note that only 7-17% of the putative genes show such associations.

Although very few were observed, predictions associated with pseudogenes indicate that some of these structures may be transcribed (examples of transcribed pseudogenes are reviewed in Mighell *et al*., 2000).

In six cases, two genes were found sharing the same promoter prediction. In all cases one gene is annotated on the forward strand and the other on the reverse strand.

3.5.2.2. Focusing on coding genes

Predictions at the 5' end of coding genes (within 500 bp of the first annotated exon) were studied in more detail. CpG islands were predicted at the 5' end of 56.6% of genes, which is in agreement with previous studies (Ponger *et al*., 2001). Similarly, PI and Eponine predicted promoters for 47.4% and 38.3% of the genes, respectively.

Based on the sensitivity of PI and Eponine, 43% and 40% respectively, the number of correctly predicted promoters suggests that the 5' end of most coding genes has been annotated. The evidence for predicted promoter-containing regions at the 5' end of coding genes provided by the three prediction programs are shown in Figure 3.15.

The data suggests that for a given sequence, two or more of the three prediction programs predict approximately half the promoters; note that only 6.2% of the promoters predicted by PI and/or Eponine are not supported by a CPGFIND prediction. The use of all three methods identifies promoters for approximately 62% of coding genes. At least one more promoter prediction software will be required to predict promoters for the remaining genes.

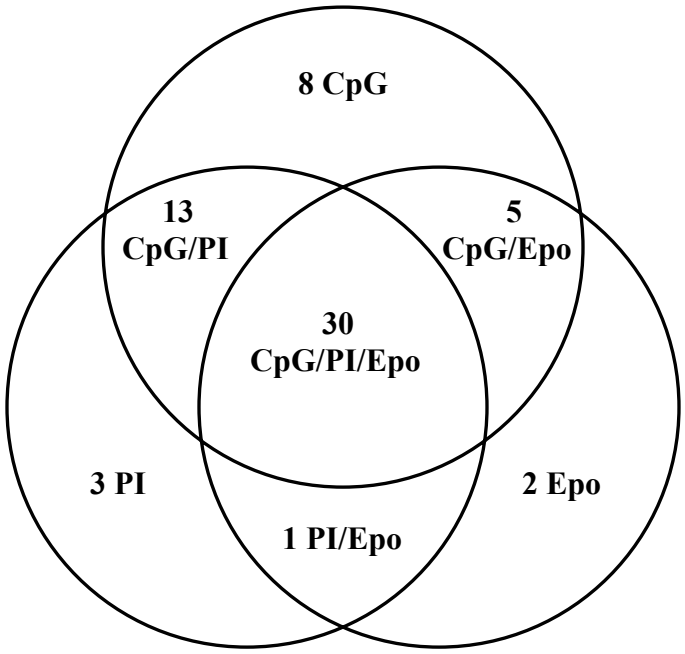


Figure 3.15: Correlation of genes with predicted promoters at their 5' end and predictions by the three methods.

3.5.3 Summary

Of the 99 coding genes, 90 appear to have 'complete' structures, i.e. they have an open reading frame (beginning with a predicted starting methionine and ending with a stop codon) and both a 5' and a 3' UTR. The computational analyses described above confirm this observation and suggest that the current annotation includes the 5' and 3' ends of nearly all coding genes.

The nine incomplete genes (missing the 5' and/or the 3' end) lack promoter predictions and/or polyadenylation signals. In addition, it is worth noting that the annotation of three incomplete genes (C20orf171, C20orf168 and C20orf157) is based on protein homologies only, suggesting that they may have a restricted expression pattern, or that they are not real genes, but probably pseudogenes.

3.6 Measuring completion of annotation

3.6.1 Homology searches

Two data sets were used for this analysis: thirteen million sequence reads of a mouse whole-genome shotgun giving an estimated genome coverage of 2.3-fold (http://trace.ensembl.org/) and 816,262 single sequence reads from BAC and plasmid ends of the *Tetraodon nigroviridis* genome, totalling 664 Mb and corresponding to 1.72 genome equivalents (generated at Genoscope; http://www.genoscope.cns.fr/). Mouse sequences were aligned to the sequence of chromosome 20 using Exonerate version 0.3d (Slater, unpublished), whereas Tetraodon sequences were aligned using Exofish (Crollius *et al*., 2000a).

2,165 mouse Regions of Sequence Conservation (RSC) and 580 *Tetraodon nigroviridis* evolutionary conserved regions (ecores) were identified in the sequence of 20q12-13.2. Genscan predicted that 700 of the 2,165 RSC (32.3%) to be coding (cRSC).

92.9% of cRSC were found to map within annotated exons (including pseudogenes), whereas 61.3% of annotated exons are supported by cRSC. Eighteen cRSC were found to map in introns, whereas 32 map in intergenic regions. Of the remaining 1,465 non-coding RSC, 15% were found to map in exons (including pseudogenes), whereas 16.1% of annotated exons are supported by a non-cRSC. 72.7% of all annotated exons are supported by at least one cRCS and/or a non-RSC.

95% of ecores were found to map within annotated exons (including pseudogenes), whereas 50% of the annotated exons are supported by ecores. Eight ecores were found to map in introns and a further twenty in intergenic regions.

Five conserved regions, two in introns and three in intergenic regions, are matched by both an ecore and a mouse hit. These may represent exons that remain un-annotated. Since 45.8% of annotated exons have both ecore and RSC hits we can postulate that eleven exons still remain un-annotated.

3.6.2 Genscan and FGENESH

Genscan and FGENESH were used to predict gene structures. The data from the prediction program analysis are shown in Table 3.9.

Table 3.9: Analysis of predicted coding sequence.

	Genscan	FGENESH
Predicted genes	327	255
Predicted exons	1,989	1,629
Total coding length	340,018 bp	273,270 bp
GC level	52.69%	55.61%
Sequence in repeats	72,751 bp (21.4%)	37,156 bp (13.6%)

The predicted structures were compared to the annotated coding exons that are supported by expressed sequences. Approximately one in ten annotated coding exons were completely missed by either Genscan or FGENESH, whereas Genscan predicted exact matches for 79.7% of coding exons compared to 77.6% by FGENESH (Table 3.10).

Table 3.10: Comparison of Genscan and FGENESH predictions and annotated, supported, coding exons.

	Exact matches	5' end of exon missed	3' end of exon missed	Both exon ends missed	Exon missed
Genscan	79.7%	4.9%	4.5%	0.3%	10.6%
FGENESH	77.6%	6.4%	4.8%	0.3%	10.9%

6.7% of coding exons were completely missed by both methods. 6.3% of annotated coding exons had their 5' end and/or their 3' end incorrectly predicted by both programs, whereas 87% of annotated exons were correctly predicted by at least one prediction program. Both methods produced exact matches for 72.2% of exons.

In addition, both programs produced exact predictions for 226 sequences outside annotated exons. If we assume that all exact predictions are real and that both algorithms in 72.2% of the cases can correctly predict exons then approximately 314 coding exons remain to be annotated. This remains to be investigated (section 4.5.3).

3.7 Protein analysis

I analysed the proteome of 20q12-13.2 using InterProScan (http://www.ebi.ac.uk/interpro/scan.html) to look at the distribution of known protein domains. The InterPro database combines information on protein families, domains and functional sites from the databases Pfam, PRINTS, PROSITE, SMART and SWISS-PROT (see http://www.ebi.ac.uk/interpro/ for links). Of all proteins encoded in the region, 85.8% have an InterPro match and 45.4% are multi-domain with an average of 3.4 different InterPro domains. Table 3.11 lists the most widespread domains in the region and the number of proteins with these domains in various organisms. At the time of analysis (January 2002) 71.4% of Homo sapiens, 70.8% of *Mus musculus*, 70.9% of *Drosophila melanogaster*, 66.7% of *Caenorhabditis elegans*, 68.9% of *Arabidopsis thaliana* and 65.1% of *Saccharomyces cerevisiae* proteins had at least one InterPro domain.

The region is enriched in proteins containing IPR002221 and IPR002223 domains. The thirteen genes that encode proteins with a WAP-type four disulphide core domain and/or a pancreatic trypsin inhibitor (Kunitz) domain are clustered in the sequence between Z93016 and AL050348. A smaller cluster is located within this ~700 Kb region that encodes for SEMG1 and SEMG2 (semen proteins involved in reproduction). SEMG1 and SEMG2 are located in tandem on the same sequence strand. Both encode for three exon genes and share high homology at the nucleotide and protein level (Figure 3.16). The secreted forms of SEMG1 and SEMG2 proteins are composed of 434 and 554 amino acids respectively, mainly consisting of sixty-residue tandem repeats. Comparison of the two loci suggests that they evolved by the duplication of an approximately 8 Kb DNA

segment, probably by a mechanism involving recombination between L1 elements (Lundwall, 1996).

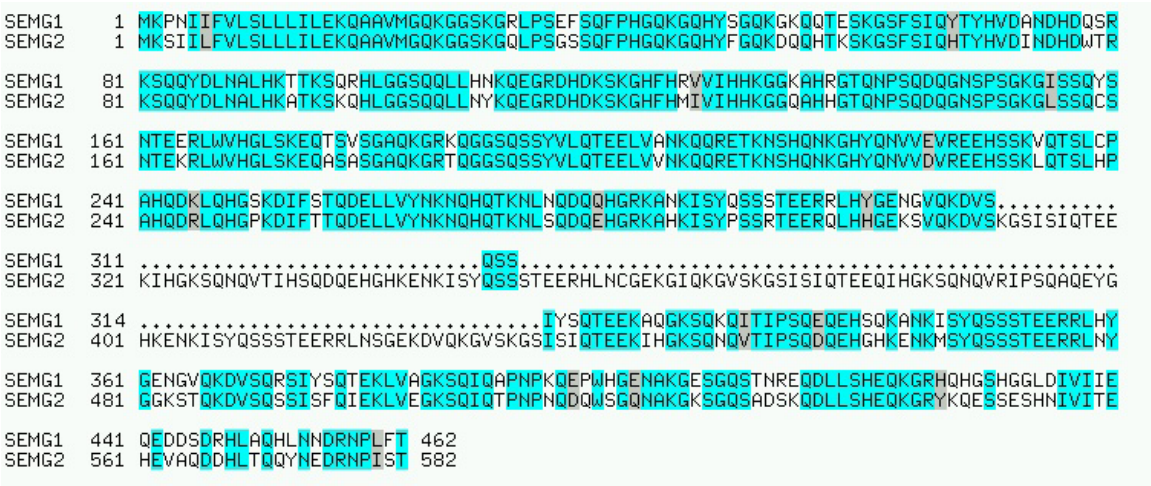


Figure 3.16: SEMG1 and SEMG2 protein alignment. The protein sequences were aligned using CLUSTAL W (Thompson *et al*., 1994) and the output was formatted using Belvu (Sonnhammer, unpublished). Identical aa are highlighted blue and similar, grey.

The coding potential of putative genes was also investigated. Assuming that the entire mRNA structure was annotated I looked for ORFs with predicted translation start sites (atg base pairs). Predicted peptide sequences were identified for all putative genes with an average length of 37 amino acids (median 27 amino acids). Like BLASTX searches, InterProScan did not identify any similarities with known protein motifs. Alignment of just the predicted peptides also failed to indicate the presence of any similarities. In addition, although coding exons of coding genes rarely overlap with repeat sequences, approximately 60% of putative genes have exons overlapping with repeats.

Table 3.11: Most common InterPro domains in 20q12-13.2 and their abundance in other species. At the time of analysis the InterPro database contained 24,680 *Homo sapiens* (Hs), 15,884 *Mus musculus* (Mm), 13,844 *Drosophila melanogaster* (Dm), 18,935 *Caenorhabditis elegans* (Ce), 25,773 *Arabidopsis thaliana* (At) and 6,140 *Saccharomyces cerevisiae* (Sc) protein entries. InterPro domains IPR001472 and IPR000694 are excluded from proteome analysis, owing to low specificity. Whole proteome data reproduced from http://www.ebi.ac.uk/interpro/.

Rank	InterPro code	Abundance (number of proteins with InterPpro domain)							Name
		20q12-13.2	Hs	Mm	Dm	Ce	At	Sc	
1	IPR001472	11	ND	ND	ND	ND	ND	ND	Bipartite nuclear localisation signal
2	IPR002221	9	11	9	4	6	0	0	Whey acidic protein, core region
3	IPR002223	6	17	13	22	37	0	0	Pancreatic trypsin inhibitor (Kunitz)
4	IPR000822	5	791	341	340	209	169	53	Zinc finger, C2H2 type
5	IPR000694	4	ND	ND	ND	ND	ND	ND	Proline rich
6	IPR000636	3	148	88	51	80	29	3	Cation channel
6	IPR000719	3	614	387	239	439	1041	115	Eukaryotic protein kinase
6	IPR001245	3	271	180	90	154	475	3	Tyrosine protein kinase
6	IPR001622	3	90	59	43	83	21	1	Potassium channel, pore region
6	IPR001687	3	73	46	63	58	158	35	ATP/GTP binding motif

3.8 Discussion

In this chapter I presented the sequence analysis of a 10 Mb region of human chromosome 20q12-13.2. Computational and experimental analyses ensured the assembly of the most detailed gene map of the region to date. Placement of the genes on the sequence map enabled the investigation of their structure and environment.

The gene map of this region contains 165 gene structures that are divided to four groups: "novel", "known", "putative" and "pseudogenes". Genes that belong to the "known" and "novel" groups are supported by strong evidence that they are expressed and have an easily detectable open reading frame. The "known" or "novel" classification was used to differentiate between genes that during the first round of chromosome 20 sequence analysis were known and genes that, at the time, were only supported by ESTs and/or protein homologies and/or anonymous partial cDNA sequences. The group of the mainly incomplete "novel" genes was the focus of my experimental work. Following this study and the publication of the whole chromosome-20 sequence analysis (Deloukas *et al*., 2001), all genes in both groups are "known" genes.

The choice of experimental method to confirm gene structures depends on various constraints, such as the number of genes to be investigated, available resources and time constraints. If only a small number of genes are under investigation, cDNA sequences can be isolated relatively easily and at a low cost by screening arrayed cDNA libraries. For larger numbers of genes a different approach is required. For example, the SSP-PCR strategy achieves rapid identification of positive cDNA pools by the parallel PCR

screening of several cDNA libraries, in a single step. cDNA fragments can then be systematically isolated from the positive pools. Like SSP-PCR, the vectorette method ensures the rapid identification of positive cDNA pools. In addition, vectorette-based cDNA isolation is less time-consuming than SSP-PCR. A disadvantage of this method is that it requires the use of several cDNA libraries from different tissues (expensive) and a relatively complex process of generating modified cDNA pools (time-consuming).

The ability to isolate the cDNA fragment of interest depends on the characteristics of the cDNA libraries used. For example, none of the cDNA vectorette pools are enriched in full-length cDNAs. This is mainly due to the fact that during the construction of these cDNA libraries (as well as of those used to generate a significant portion of the available public EST resources), the reverse transcriptase dissociates at any random point from the template, resulting in incomplete cDNAs.

The advantage of using such libraries is that it enables the isolation of different size cDNA fragments confirming different parts of the gene. Use of full-length cDNA libraries would result in the isolation of large fragments. Vectorette isolation of cDNA fragments longer than 2 Kb is inefficient and thus confirmation of genes encoding for long transcripts using full-length cDNA libraries would be quite challenging.

I found the vectorette libraries used in this study sub-optimal for isolation of the 5' end of genes. RACE was the preferred approach and although more time-consuming than the vectorette method it does allow the study of several genes simultaneously.

Only three of the annotated genes remain without human evidence of expression. In addition, less than 3% of all annotated exons (excluding pseudogenes) are not supported

by human expressed sequence across their whole length. This is mainly due to the fact that I was unable to identify positive cDNA pools, or isolate cDNA-end fragments corresponding to these exons.

Of the 99 protein coding genes (known and novel), 90/99 (91%) have "complete" structures (i.e. both 5' and 3' UTRs and an open reading frame with a predicted starting methionine). Compared to all novel genes annotated on chromosome 20, those in 20q12-13.2 have on average 1.7-fold more exons. The only difference between the two datasets is the addition of the experimental data generated by this study. This approach can be easily scaled up and applied to the rest of the genome.

By definition, all putative genes (30) have incomplete structures. The structure of putative genes is different from that of coding genes in several ways:

i. They are significantly smaller both in terms of locus size and exon number. Their mRNA transcripts are also significantly smaller (502 bp compared to 2.8 Kb of coding genes).

ii. They are frequently found in sequences that overlap either with coding genes or other putative genes.

iii. Their predicted putative ORF does not encode for peptides similar to any known protein sequences.

iv. Approximately 60% have exons that partially overlap with repeat elements.

v. PCR-based analysis indicates that their expression is less abundant than that of coding genes (note that primary signals in vectorette cDNA pool screenings could not be followed to isolate distinct fragments).

Experimental approaches to extend their structure were not successful, even though positive cDNA pools were identified for most of them. The inherent complexity of the vectorette reaction makes it difficult to pinpoint the reason for this failure. It is worth noting that, failure to isolate cDNA-ends for coding genes from a particular vectorette cDNA pool is not uncommon. That is the reason why vectorette is usually applied to several positive cDNA pools instead of only one. This was not always possible for the putative genes since they were usually found in fewer pools than the coding genes. Another possible explanation may be that putative gene transcripts are less abundant compared to transcripts of coding genes, and thus isolation of the transcript of interest is more difficult.

On average, 1.8 transcripts were annotated for each coding gene. This is in agreement with the whole chromosome 20 study (1.65 transcripts per gene, excluding different polyadenylation sites; Deloukas *et al*., 2001). Significantly higher numbers of transcripts were reported for the gene-rich chromosomes 19 and 22 (3.2 and 2.6 transcripts per gene, respectively; IHGSC, 2001). Of the annotated variants, 44/71 (62%) are predicted to have a different ORF. 30 genes are predicted to have two or more protein isoforms. Whether these annotated variants represent real isoforms or are artefacts of the EST libraries remains to be investigated.

InterProScan analysis of the proteome of the region identified a gene cluster that encodes for thirteen proteins enriched in IPR002221 and/or IPR002223 domains. A second gene cluster encoding for the SEMG1 and SEMG2 proteins was also identified, mapping within the first cluster. Further analysis will be required to decipher the evolutionary history of these two clusters.

A three-species comparative analysis was used to estimate the level of completion of annotation. Sequence comparison of the human annotation to the mouse whole genome shotgun and the ecores generated from the *Tetraodon nigroviridis* genomic sequence suggests that the vast majority of exons in the region have been identified. Exon identification cannot be performed solely using the mouse data set because of the high degree of sequence conservation between human and mouse across the whole region. Despite this, the combination of the mouse and Tetraodon data provides an excellent tool for assisting the identification of new, and the completion of existing, gene structures (Deloukas *et al*., 2001).

The combined use of Genscan and FGENESH identified 93.3% of annotated coding exons in the region, whereas 72.2% of exons have exact matches by both programs. Exact double matches were obtained for 155 intergenic regions and 71 intronic regions. Whether these represent real exons is investigated in chapter IV.

The sequence of chromosome 20q12-13.2 has an average GC content of 45.2%, which is higher than the chromosome 20 and the genome average (44.1% and 41% respectively). The distribution of GC content fluctuates along the chromosome and regions with higher GC have higher gene density (Figure 3.6).

20q12-13.2 has a gene density of 12.6 genes per Mb, which is similar to the gene density across the whole chromosome 20 (12.18 per Mb). This is intermediate to 6.71 (low) and 16.31 (high) per Mb reported for chromosome 21 and 22, respectively.

The coding-gene density of this region overall corresponds to one gene/104 Kb but it varies significantly across different sub-regions. For example, a single intron-less coding gene was identified in the first 980 Kb compared to twelve coding genes in a 140 Kb segment between 5,730 Kb and 5,870 Kb. Compared to an average coding gene density of 9.6 genes per Mb across the whole region, the two sub-regions have an average gene density of 1 and 85.7 genes per Mb (gene poor and gene rich, respectively). The gene poor region has a GC content of 42.85% compared to 49.44% of the gene rich region. 41.15% of the gene poor sequence is covered by interspersed repeats compared to 48.45% of the gene rich sequence (20q12-13.2 average is 48.04%). 11.92% of the gene poor sequence is covered by SINEs and 17.98% by LINEs. 38.27% of the gene rich sequence is covered by SINEs and 6.49% is covered by LINEs. Alu repeats occupy 4.6-fold more sequence in the gene rich region whereas LINE1 repeats occupy 8.4-fold more sequence in the gene poor region.

Overall, this study has shown that the combination of genomic sequencing coupled to computational analysis and laboratory-based efforts is a very powerful gene-finding approach. This approach was successfully used to generate the most detailed transcript map of this region to date, and the reported analyses and annotation will be a valuable tool in tackling the various diseases linked to this region. For example, we have reported the refinement of a Commonly Deleted Region (CDR) of 20q12-13.1 found in patients with myeloproliferative disorders and myelodysplastic syndromes (Bench *et al*., 2000).

The transcript map can also be used to identify possible candidate genes for the various diseases, for which mutation analyses can be performed. Finally, the generated data provided the basis for the experimental work described in the following chapters (IV and V).

Chapter IV

Comparative mapping, sequencing and analysis

4.1 Introduction

The challenges that arise after sequencing the human genome include finding and verifying all genes, obtaining their expression patterns and functional characteristics and studying how they interact with each other and the environment. The mouse has served over the past century as an excellent experimental system for studying mammalian genetics and physiology and is expected to greatly enhance these efforts (Dietrich *et al*., 1995).

Blocks of synteny between the human and mouse genomes can provide an insight into genome organisation and evolution. Comparative analysis at the DNA level can be used to identify coding exons or regulatory elements, which are often highly conserved. Once established, the resources can be used to further characterise genomic regions (for example gene knockouts to assess function) and if applicable, provide an animal model for a human disease (Hardison *et al*., 1997; O'Brien *et al*., 1999; Murphy *et al*., 2001).

4.1.1 The mouse genome

The mouse genome is roughly 3,000 Mb in size and a number of genetic maps have been constructed. Dietrich *et al*. (1996) published a high-density, intermediate-resolution genetic map of the mouse genome. The map contained 7,377 genetic markers consisting of 6,580 microsatellite markers integrated with 797 RFLPs in mouse genes (Dietrich *et al*., 1996; Jordan and Collins, 1996). The construction of a high-resolution genetic map incorporating 3,368 microsatellites was reported two years later (Rhodes *et al*., 1998).

The available genetic maps provided the scaffold for the construction of a YAC-based physical map of the mouse genome (Nusbaum *et al*., 1999). STSs were screened against 21,120 YAC clones with an average insert length of 820 Kb and the STS-content information was integrated with the genetic map. The resulting map showed the location of 9,787 loci with an average spacing of approximately 300 Kb and affording YAC coverage of approximately 92% of the mouse genome (Nusbaum *et al*., 1999).

Van Etten *et al*. (1999) described the construction of an RH map of the mouse genome. The map contained 2,486 loci screened against an RH panel of 93 cell lines. Most (93%) were microsatellite loci taken from the genetic map, thereby providing direct integration between these two key maps.

ESTs are key in providing rapid access to the gene repertoire of an organism. To provide a broad overview of genes expressed throughout the mouse development, ESTs were sequenced from fifteen normalised libraries and 26 early-stage libraries (Marra *et al*., 1999). In a more systematic effort, RIKEN sequenced and annotated 21,076 cDNA clones from 160 "full-length", normalised and subtracted cDNA libraries from various tissues and developmental stages (The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium, 2001).

Genomic clone resources are vital for genome studies. Clones from the RPCI-23 and RPCI-24 mouse BAC libraries (Osoegawa *et al*., 2000, http://www.chori.org/bacpac/) were fingerprinted at the Genome Sequence Centre (GSC) of British Columbia Cancer Research Center at Vancouver, Canada (http://www.bcgsc.bc.ca/projects/mouse_

mapping/) and end-sequenced at The Institute for Genomic Research (TIGR) (Zhao *et al*., 2001, http://www.tigr.org/).

The available resources were used to assemble a microsatellite-marker anchored BAC framework map of the mouse genome (Cai *et al*., 2001). The map was subsequently improved by aligning mouse BAC-end sequences to the human genome (Gregory *et al*., unpublished; http://mouse.ensembl.org/assembly.html). The mouse map database currently contains a total of 554 contigs providing an estimated coverage of 3,028 Mb and is available at http://mouse.ensembl.org/.

In May 2002, the Mouse Sequencing Consortium (http://www.sanger.ac.uk/Info/Press/001006.shtml) reported a draft sequence of the mouse genome. The mouse strain C57BL/6J was used and the coverage is estimated at ~seven-fold (http://www.ensembl.org/Mus_musculus/). Completion of the mouse genome is scheduled for 2005. Raw sequence data (traces) is publicly available from the Ensembl trace server (http://trace.ensembl.org/). The large volumes of mouse sequence data can be searched using new homology search methods such as SSAHA (Ning *et al*., 2001), Exonerate (Slater, unpublished) or BLAT (Kent, 2002).

4.1.2 Comparative studies

Human:mouse comparative sequence analyses have been performed for a number of gene loci (Collins and Weissman, 1984; Shehee *et al*., 1989; Lamerdin *et al*., 1995; Koop and Hood, 1994; Blechschmidt *et al*., 1999; Brickner *et al*., 1999). The findings suggest that

coding regions are generally well conserved, whilst conservation at the intronic and intergenic regions varies extensively.

Comparative analysis of 1,196 orthologous mouse and human full-length mRNA and protein sequences showed that protein sequence conservation varies between 36% and 100% identity, with an average value of 85%. The average degree of nucleotide sequence identity for the corresponding coding sequences was also approximately 85% whilst 5' and 3' UTRs were found to be less conserved (Makalowski *et al*., 1996). A comprehensive study of 77 orthologous mouse and human gene pairs revealed that the proportion of the non-coding regions covered by blocks of over 60% identity was 36% for upstream regions, 50% for 5' UTRs, 23% for introns and 56% for 3' UTRs (Jareborg *et al*., 1999).

Comparative analyses of sequence from the human and mouse α/δ T-cell receptor loci (Koop and Hood, 1994) revealed a high degree of conservation across both coding and non-coding regions. In contrast, studies at the XRCC1 locus (Lamerdin *et al*., 1995), the β-globin gene cluster (Collins and Weissman, 1984; Shehee *et al*., 1989), the ERCC2 gene region (Lamerdin *et al*., 1996), the AIRE (Blechschmidt *et al*, 1999) and the ADA genes (Brickner *et al*., 1999) found less conservation across non-coding sequences. Sequence analysis of regions encoding for several genes, such as the human and murine Bruton's tyrosine kinase loci and a gene rich cluster at human chromosome 12 syntenic to mouse chromosome 6, revealed that the extent of conservation between the non-coding regions of neighbouring genes can vary (Oeltjen *et al*., 1997; Ansari-Lari *et al*., 1998).

Comparative mapping and sequencing aims to identify new genes and detect regulatory elements (Koop and Hood, 1994; Lamerdin *et al*., 1995; Oeltjen *et al*., 1997; Hardison *et al*., 1997; Jackson, 2001) on the basis that these sequences are highly conserved during evolution. Wasserman *et al.* (2000) reported that 98% of experimentally determined binding sites of skeletal-muscle-specific transcription factors were found in the highest fraction (19%) of conserved sequence in the orthologous genomic segments (human and mouse). Novel regulatory elements of the SCL (Göttgens *et al*., 2000, 2001), the interleukins 4, 13 and 5 loci (Loots *et al*., 2000), the α-synuclein genes (Touchman *et al*., 2001) and the ABCA1 genes (Qiu *et al*., 2001) were identified through comparative analysis and further validated by experimental approaches.

The promise of comparative studies to contribute to the in-depth analysis and characterisation of genomic regions prompted the construction of several syntenic maps. This was followed by sequencing and comparative sequence analysis (Wenderfer *et al*., 2000; Martindale *et al*., 2000; Mallon *et al*., 2000; Footz *et al*., 2001; Pletcher *et al*., 2001; Wilson *et al*., 2001). This homology-based approach for map construction was successfully used to generate megabase-long mouse contigs for regions encoding genes homologous to those found on human chromosome 4 (Crabtree *et al*., 2001), 7 (Thomas *et al*., 2000) and the whole euchromatic portion of human chromosome 19 (Kim *et al*., 2001; Dehal *et al*., 2001). The generated data was used to determine gene order, identify novel genes, compare GC and repeat content and characterise breakpoints of evolutionary rearrangements. The need to compare genomic sequences (reviewed in Miller, 2001) resulted in the development of new software such as PipMaker (Schwartz *et al*., 2000),

Vista (Mayor *et al*., 2000), GLASS (Batzoglou *et al*., 2000) and SynPlot (Göttgens *et al*., 2001).

Finished sequence is the ideal tool for the exhaustive search and accurate annotation of gene features. The finished and annotated sequence can then be further analysed by comparison to the genome sequence of other species. For example, it was shown by mouse sequence comparison to finished, annotated human sequence that some human genomic regions tend to accumulate changes due to both point mutation and retrotransposition at a higher rate than others which appear to be protected from these two types of sequence alteration (Chiaromonte *et al*., 2001).

4.1.3 Overview

Comparative studies promise to be an essential tool in furthering our understanding of the emerging human genome sequence. The aim of this chapter is to test this approach through the systematic analysis of the mouse genomic region that is syntenic to human 20q12-13.2. The use of a gene based, homology-driven approach to construct a 10 Mb-long mouse clone contig spanning this region is described. The contig is located on mouse chromosome 2 and the comparative mapping data suggests gene order conservation between human and mouse.

The clone map was used to select a tiling path (66 clones) for sequencing the entire region. At the time of analysis, 38 and 27 mouse clones had finished and unfinished sequence, respectively. The available mouse clone sequences were used in comparative analyses with the orthologous human sequence. Similarity searches were used to

investigate the extent of synteny between annotated human gene features and mouse genomic sequences. Non-exonic conserved sequences were also examined to determine the presence of un-annotated exons and provide an estimate of the completeness of the human annotation. A comparison of the GC and repeat content is also reported.

As described for human (section 3.2), finished mouse sequence undergoes systematic computational analysis and gene annotation. Data from these analyses was used to perform size comparisons between orthologous exon (and intron) pairs. In addition, three orthologous gene pairs were selected for in-depth analyses of their DNA and predicted protein sequences.

4.2 Mouse clone map construction

Since YACs are generally considered-sub optimal substrates for genomic sequencing due to chimerism and deletions (Green *et al*., 1991; Nagaraja *et al*., 1994), I used the RPCI-23 mouse BAC library (Osoegawa *et al*., 2000) for map construction. Bacterial contig construction was performed by a parallel approach (Figure 4.1) of landmark-content mapping (Green and Olson, 1990) and restriction enzyme fingerprinting (Marra *et al*., 1997; Humphray *et al*., 2001).

Landmark content mapping is based on the detection of the presence or absence of a DNA marker in a set of clones. The major advantages of this method are that it allows the ordering of clones based on their landmark content and the detection of overlaps of any length (typically >100 bp) between clones. Fingerprinting assesses clones over their entire length and provides a size estimate of their DNA inserts. As a result, the extent of overlap between the two clones can be estimated (unlike landmark content mapping, fingerprinting does not detect small overlaps). The parallel use of the two methods provides an accurate means to confidently construct contig maps of specific genomic regions.

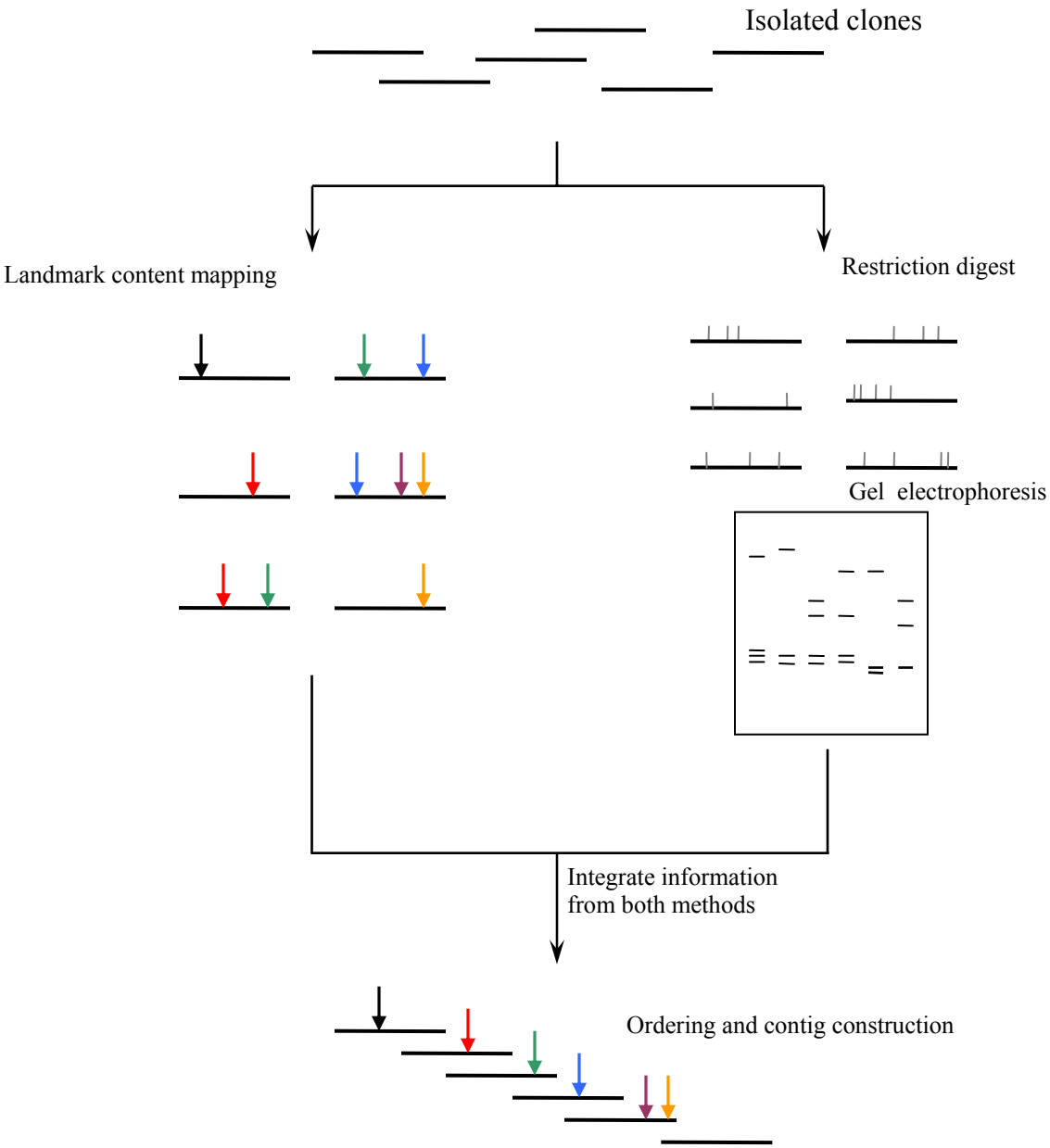


Figure 4.1: Strategy for contig construction, involving landmark content mapping and restriction enzyme fingerprinting.

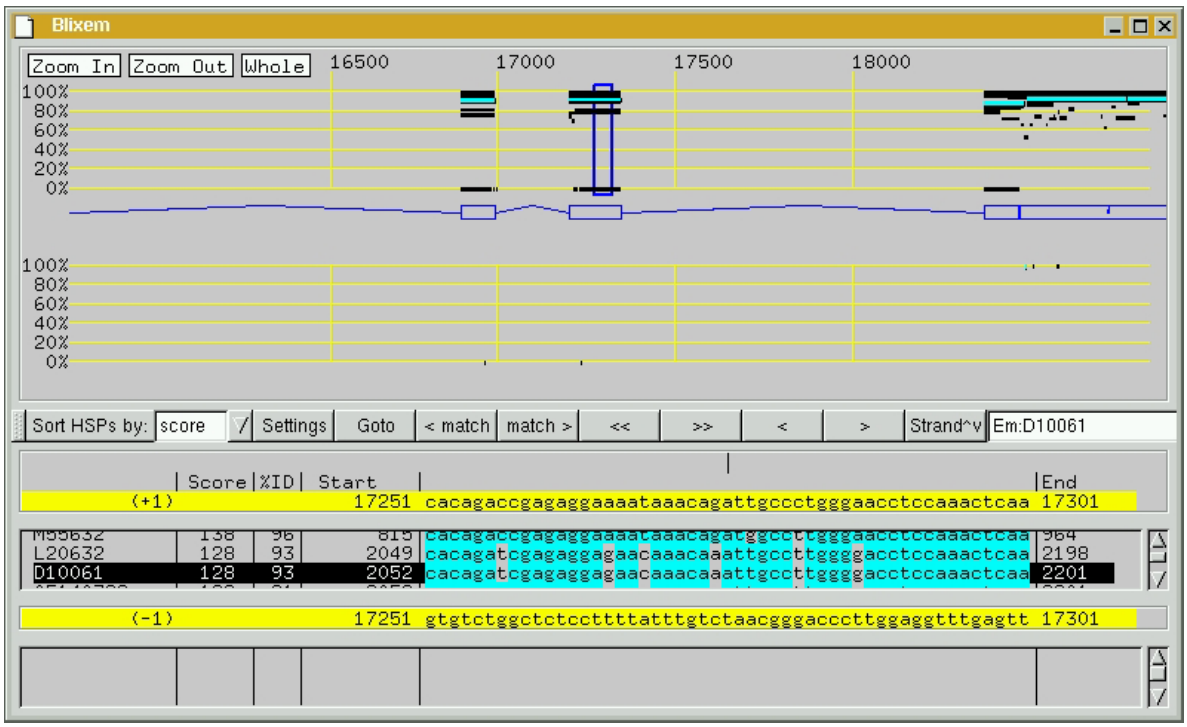
4.2.1 Marker selection and development

To target the mouse regions of interest (i.e. syntenic to the 10 Mb region of human chromosome 20q12-13.2), I identified mouse-expressed sequences sharing extensive homology with the annotated genes in the human region. Selected mouse sequences were then used to develop STS-based markers. Where possible, mouse sequences were selected at 70-100 Kb intervals on the basis of the human sequence.

Mouse-expressed sequences (53 in total, showing homology to 47 genes) were selected and used to design 71 STS primer pairs (Figure 4.2); primer design was performed as described in chapter II. Primer pairs were tested at three annealing temperatures (55°C, 60°C and 65°C) for PCR amplification of mouse genomic DNA. Gel electrophoresis was used to separate the PCR products on an agarose gel (Figure 4.3). The expected size bands were excised and stored in water (probes).

The 66 working STSs show homology to exons of 47 human genes and are listed in Table 4.1 (also see Appendix 9). The average size of the generated probes is 143 bp.

A.



B.

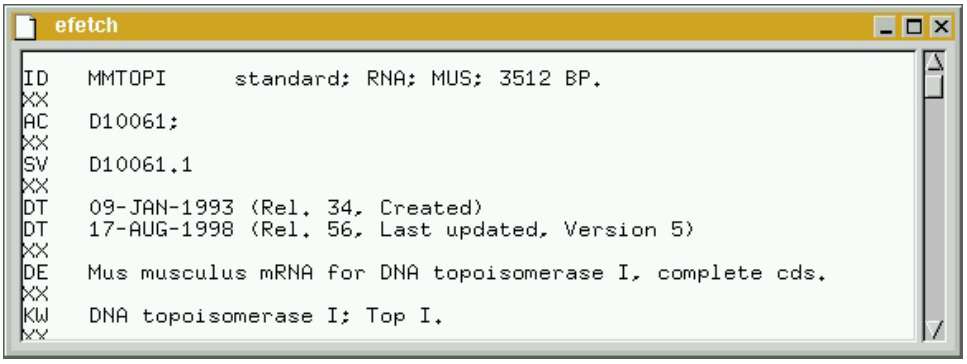


Figure 4.2: Design of mouse STSs. (A) Blixem view of homology clusters. Sequences homologous to the genomic region coding for TOP1 (human) are shown at the top part of the window as black lines whereas their sequence can be viewed at the bottom part of the window. The high quality genomic sequence is highlighted yellow (+1 forward strand; -1 reverse strand). The percentage identity of each sequence is also shown on both the top and bottom part of the window (as %ID). The gene structure is coloured blue and is shown between the top and bottom stands, at the top part of the window (exons are shown as boxes, introns as lines). The (highlighted) sequence with the accession number D10061 is an mRNA submission for the mouse topoisomerase I gene. (B) Part of the efetch window under which the EMBL submission for this sequence is stored. Regions of this mouse mRNA sequence homologous to exon 12 and the 3' UTR of the annotated human TOP1 gene were used to design the stSG77003 and stSG77004 STSs respectively.

Table 4.1: Gene based (working) STS markers. Human genes are listed according to the order with which they map on human chromosome 20. The names of mouse-specific STS markers are listed in the next column. Where available, the orthologous mouse gene names were obtained from LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/).

	Human gene name	Primer name	Mouse gene name
1	KRML	stSG77035	Mafb
2	TOP1	stSG77002	Top1
3	TOP1	stSG77003	Top1
4	TOP1	stSG77004	Top1
5	PLCG1	stSG77005	Plcg1
6	LPIN3	stSG85200	N/A
7	KIAA1335	stSG77006	5430439G14Rik
8	KIAA1335	stSG77057	5430439G14Rik
9	PTPRT	stSG77008	Ptprt
10	PTPRT	stSG77009	Ptprt
11	PTPRT	stSG77010	Ptprt
12	PTPRT	stSG77011	Ptprt
13	PTPRT	stSG77012	Ptprt
14	PTPRT	stSG77013	Ptprt
15	PTPRT	stSG77062	Ptprt
16	SFRS6	stSG77014	1210001E11Rik
17	C20orf9	stSG77001	N/A
18	MYBL2	stSG85201	Mybl2
19	MYBL2	stSG77037	Mybl2
20	C20orf100	stSG77063	N/A
21	C20orf111	stSG77015	N/A
22	GDAP1L1	stSG77033	N/A
23	HNF4A	stSG77038	Hnf4A
24	HNF4A	stSG85202	Hnf4A
25	TDE1	stSG77024	Tde1
26	ADA	stSG77023	Ada
27	YWHAB	stSG77025	Ywhab
28	TOM34	stSG77061	N/A
29	STK4	stSG85301	Stk4
30	SLPI	stSG77030	Slpi
31	MATN4	stSG77029	Matn4
32	SDC4	stSG77028	Sdc4
33	C20orf169	stSG77027	N/A
34	PIGT	stSG77026	N/A
35	C20orf167	stSG77032	N/A
36	TNNC2	stSG77031	Tncs

37	C20orf161	stSG77018	N/A
38	PPGB	stSG77017	Ppgb
39	PLTP	stSG77016	Pltp
40	TNFRSF5	stSG77040	Tnfrsf5
41	TNFRSF5	stSG77041	Tnfrsf5
42	C20orf25	stSG77019	N/A
43	C20orf25	stSG77020	N/A
44	KIAA1834	stSG77064	N/A
45	SLC13A3	stSG77034	N/A
46	C20orf64	stSG77021	N/A
47	SLC2A10	stSG77022	N/A
48	EYA2	stSG77043	Eya2
49	EYA2	stSG85302	Eya2
50	PRKCBP1	stSG77044	3632413B07Rik
51	PRKCBP1	stSG77045	3632413B07Rik
52	PRKCBP1	stSG77046	3632413B07Rik
53	NCOA3	stSG77047	Ncoa3
54	KIAA1247	stSG77048	2010004N24Rik
55	KIAA1415	stSG77049	N/A
56	KIAA1415	stSG77050	N/A
57	ARFGEF2	stSG85303	N/A
58	CSE1L	stSG77053	Cse1l
59	CSE1L	stSG77054	Cse1l
60	DDX27	stSG77051	N/A
61	KCNB1	stSG85204	Kcnb1
62	KCNB1	stSG85205	Kcnb1
63	PTGIS	stSG85199	Ptgis
64	B4GALT5	stSG77052	B4galt5
65	ZNF313	stSG85203	Zfp313
66	UBE2V1	stSG77056	Ube2v1

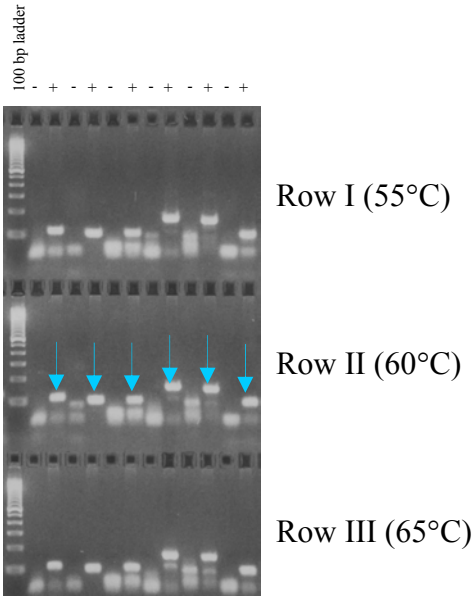


Figure 4.3: Primer testing. PCR products generated using the stSG77025, stSG77026, stSG77027, stSG77028, stSG77029 and stSG77030 primer sets were resolved on an agarose gel. The PCR products generated at 55°C, 60°C and 65°C annealing temperatures are shown on Rows I, II and III respectively. The expected size bands (indicated by arrows) were excised and stored in water.

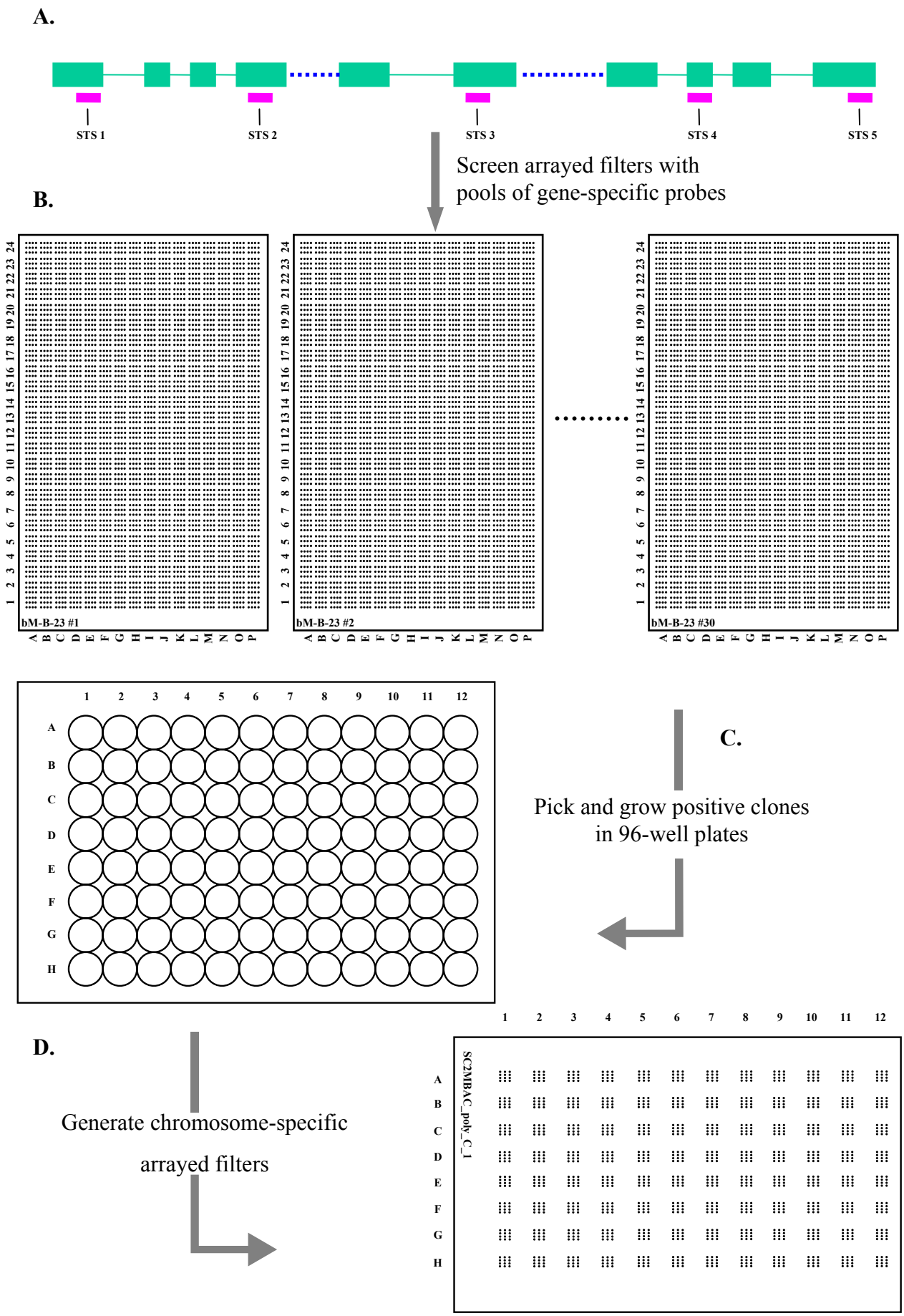
4.2.2 Bacterial clone identification

An overview of the strategy followed for BAC clone identification and isolation is shown in Figure 4.4.

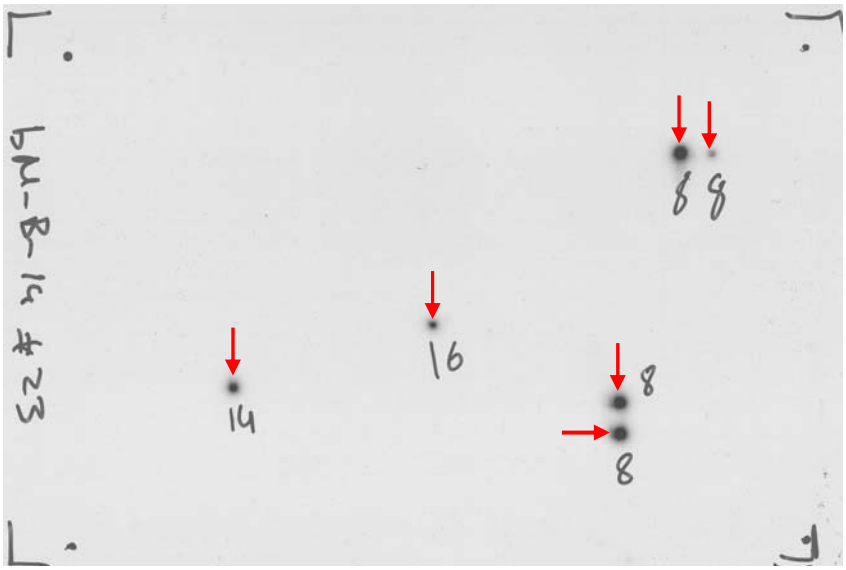
The generated probes (section 4.2.1) were labelled radioactively, pooled together (up to a maximum of 23) and used to hybridise clone filters from the RPCI-23 BAC library. All gene based probes were used in four pooled hybridisation experiments to identify 749 positive BAC clones. Data is stored in the mouse chromosome 2 ACeDB database (2musace). An example of clone identification and scoring is shown in Figure 4.5.

Positive clones were picked and grown as liquid cultures in 96-deep-well plates. Aliquots of the liquid cultures were used for fingerprinting and to generate mouse chromosome 2-specific grids (polygrids) for landmark content mapping.

Figure 4.4 (next page): Overview of BAC identification strategy. (A) Virtual region encoding three human genes. Exons are shown as green boxes, introns as green lines, intergenic regions as dotted blue lines. Mouse homologous sequences (pink boxes) were used to design mouse STS primers. (B) Pools of probes from gene-based STS markers were used to screen 30 filters representing the RPCI-23 library to identify positive clones. (C) Positive clones were picked and grown in 96-well plates. (D) The cultures were used to generate polygrids.


A.
STS 1
STS 2
STS 3
STS 4
STS 5
Screen arrayed filters with pools of gene-specific probes
B.
bM-B-23 #1
bM-B-23 #2
bM-B-23 #30
C.
Pick and grow positive clones in 96-well plates
D.
Generate chromosome-specific arrayed filters
SC2MBAC_poly_C_1

A.



B.

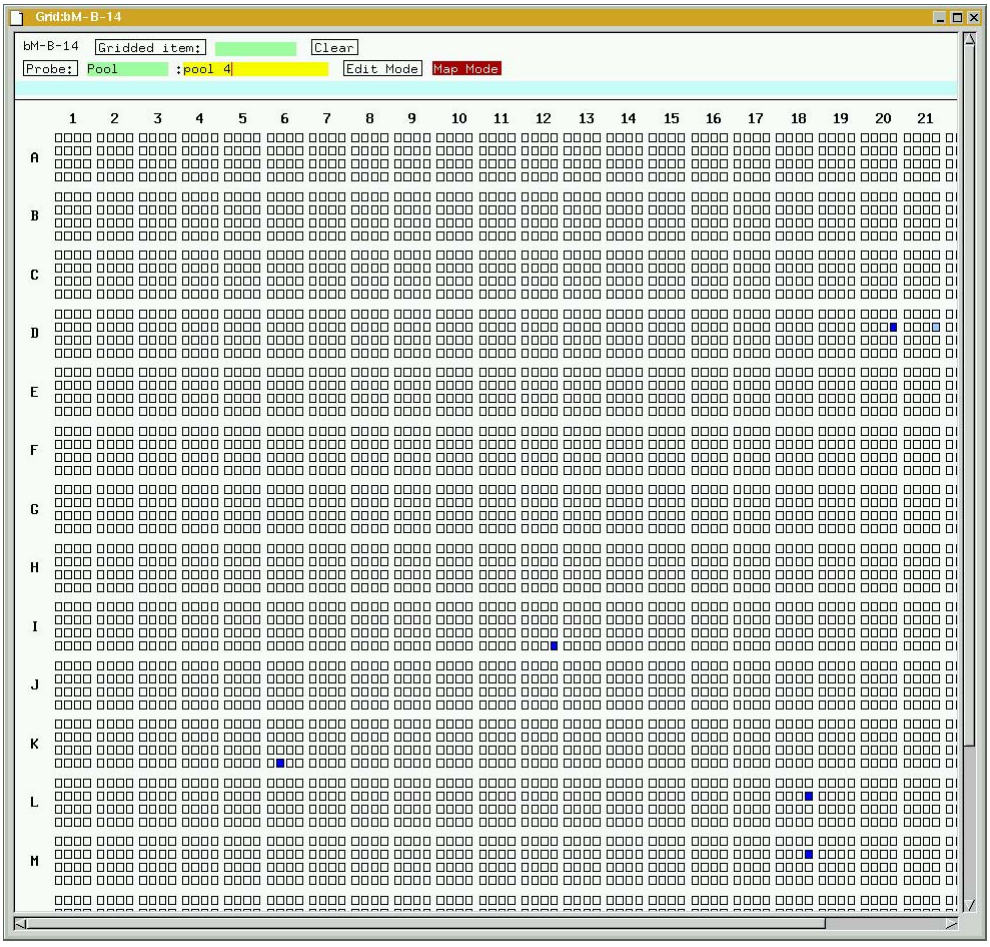


Figure 4.5: Positive clone identification and scoring. (A) Autoradiograph of filter bM-B-14 (from filter set 23) hybridised with a pool of radioactively labelled probes (stSG85199, stSG85204, stSG85205, stSG85301, stSG85302 and stSG85303; probe pool 4). Positive clones are indicated by arrows. (B) Part of grid display from 2musace. The positive clones are scored on the virtual grid (virtual grid 14). Each square represents a clone on the grid. Dark blue indicate positives and light blue indicate weak positives.

4.2.3 Fingerprint analysis

BAC DNA templates were digested in 96-well plates using *Hind*III (Marra *et al*., 1997; Humphray *et al*., 2001). A tandem 121-lane agarose gel format was used, allowing the simultaneous electrophoresis of 25 'marker' DNA samples and 96 BAC restriction digests. DNA fragments were visualised using Vistra-green staining.

Importing and editing the data in Image (Humphray *et al*., 2001) is an interactive and multi-step process (Figure 4.6). Editing of the digitised raw gel image (Figure 4.6 A) starts with lane tracking (Figure 4.6 B). Lines are manually traced along each lane across the length of the gel. The next step is band calling (Figure 4.6 C); the position of true bands is registered and spurious band calls are removed. Marker locking (Figure 4.6 D) is the final step in Image; marker lane data is normalised across the whole gel. This is used for the automatic normalisation of BAC fingerprint band values.

Figure 4.6 (next two pages): Viewing and editing fingerprint data in Image. (A) Interface for viewing the raw gel image. The grey arrow indicates the loading order, the green arrow migration. (B) Lanetracking. Blue circles show marker lanes whose corresponding lines are also traced with red coloured open boxes. (C) Bandcalling. The lane number of the selected BAC is highlighted red. (D) Standard marker locking. The number of the selected marker lane is highlighted red.

A.

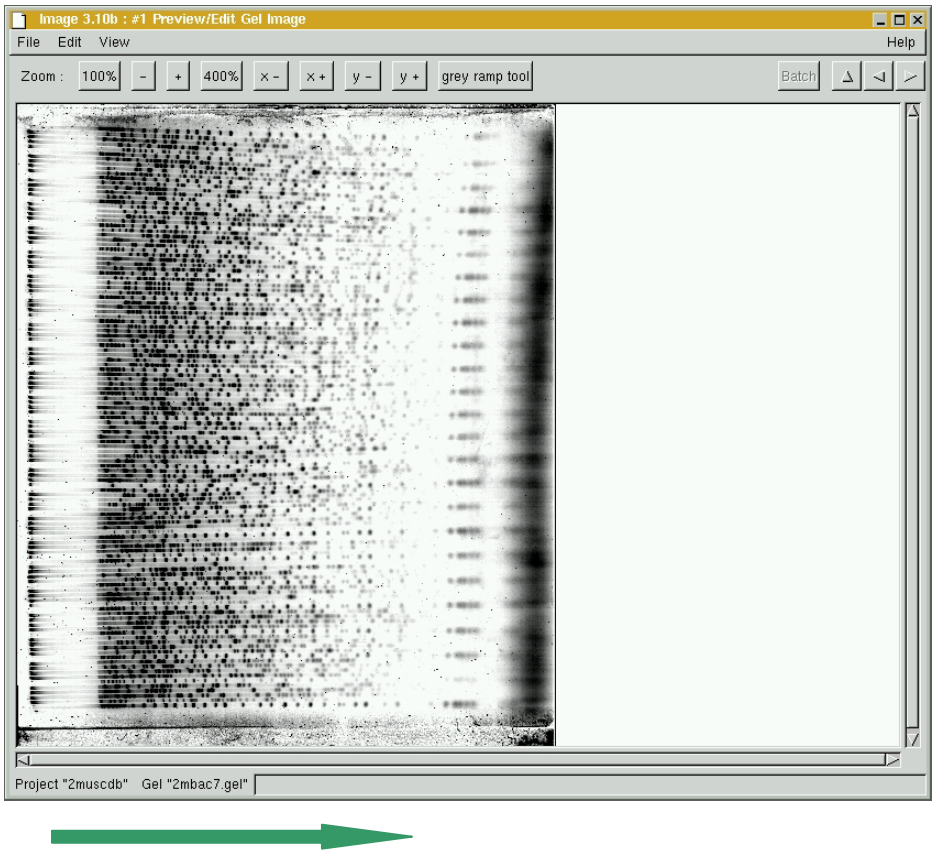


Image 3.10b : #1 Preview/Edit Gel Image
File Edit View Help
Zoom : 100% - + 400% x - x + y - y + grey ramp tool Batch
Project "2muscdb" Gel "2mbac7.gel"

B.

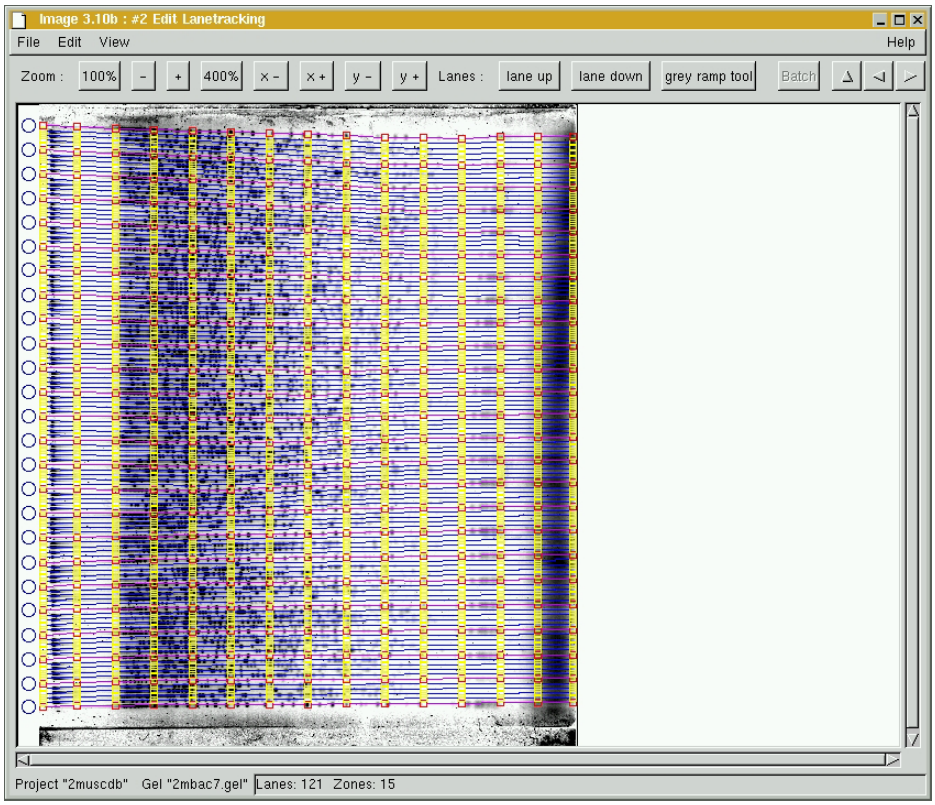


Image 3.10b : #2 Edit Lanetracking
File Edit View Help
Zoom : 100% - + 400% x - x + y - y + Lanes : lane up lane down grey ramp tool Batch
Project "2muscdb" Gel "2mbac7.gel" Lanes: 121 Zones: 15

C.

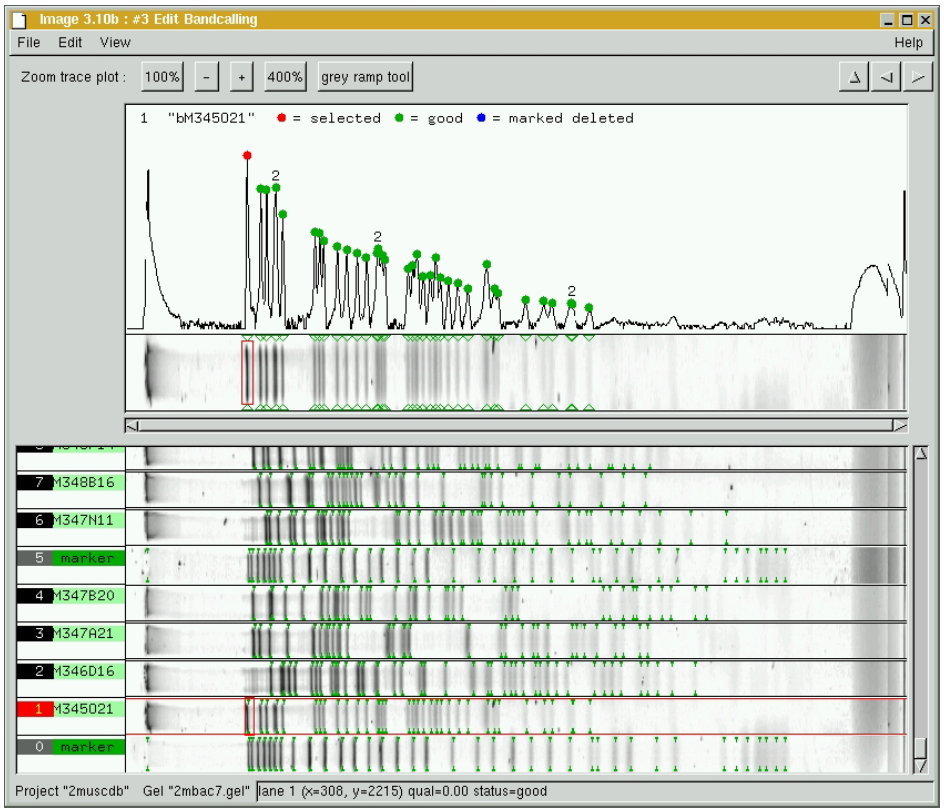


Image 3.10b : #3 Edit Bandcalling
File Edit View
Help
Zoom trace plot :
100%
-
+
400%
grey ramp tool
1 "bM345021" = selected = good = marked deleted
7 M348B16
6 M347N11
5 marker
4 M347B20
3 M347A21
2 M346D16
1 M345021
0 marker
Project "2muscdb" Gel "2mbac7.gel" lane 1 (x=308, y=2215) qual=0.00 status=good

D.

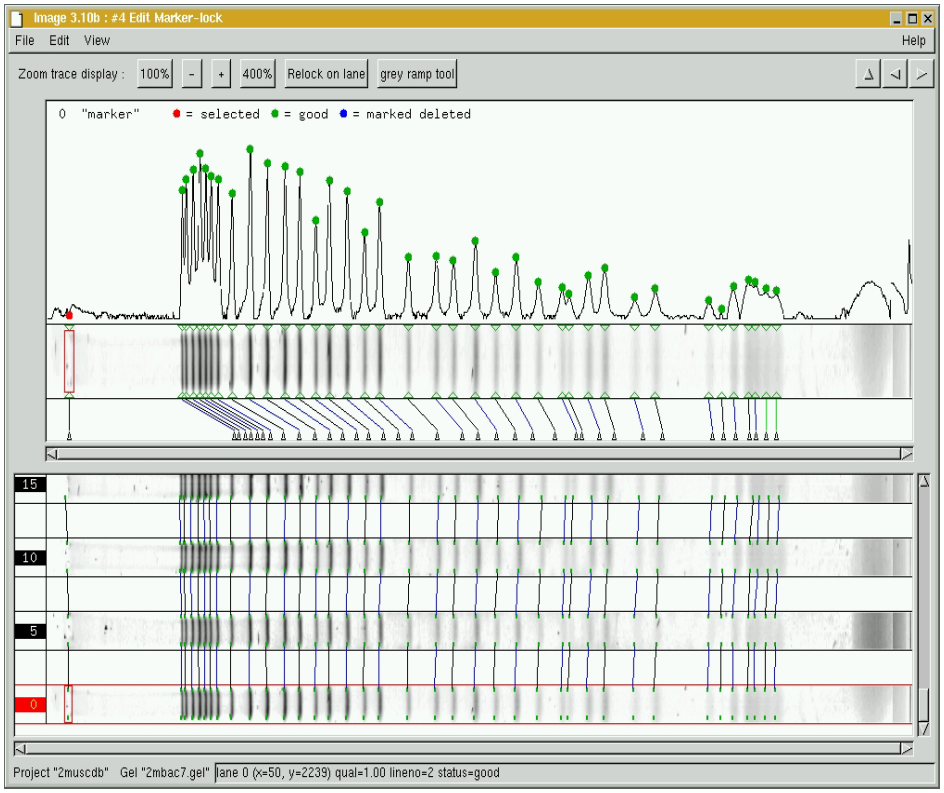


Image 3.10b : #4 Edit Marker-lock
File Edit View
Help
Zoom trace display :
100%
-
+
400%
Relock on lane
grey ramp tool
0 "marker" = selected = good = marked deleted
15
10
5
0
Project "2muscdb" Gel "2mbac7.gel" lane 0 (x=50, y=2239) qual=1.00 lineno=2 status=good

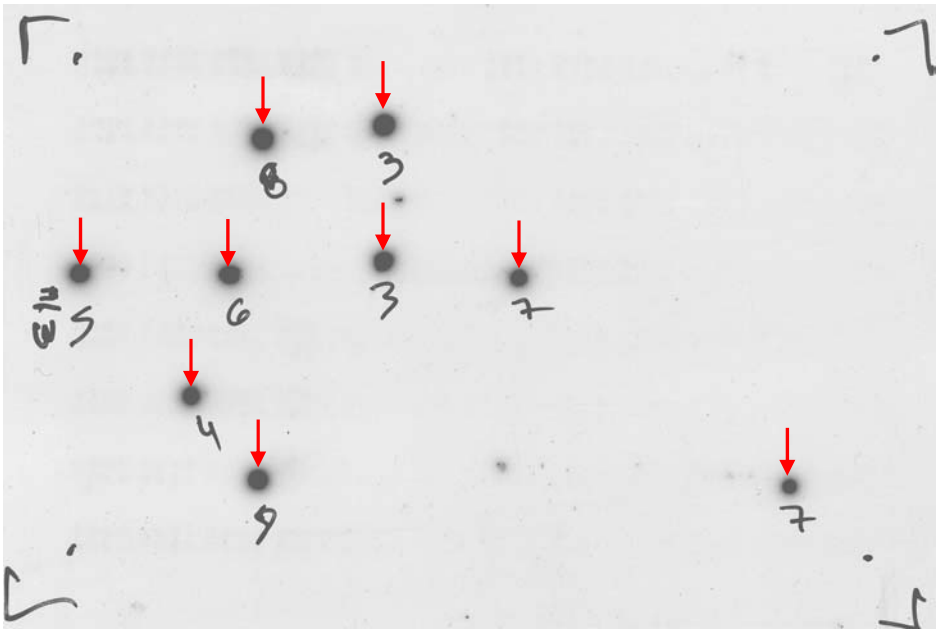
4.2.4 Landmark content mapping

Landmark content mapping was used in parallel to fingerprint analysis to obtain additional mapping information. The BAC clones identified by the pooled hybridisation approach were gridded on chromosome 2 specific filters, polygrids (section 4.2.2). The polygrids were hybridised using one gene-specific probe at a time (Figure 4.7) and hybridisation results were scored in 2musace.

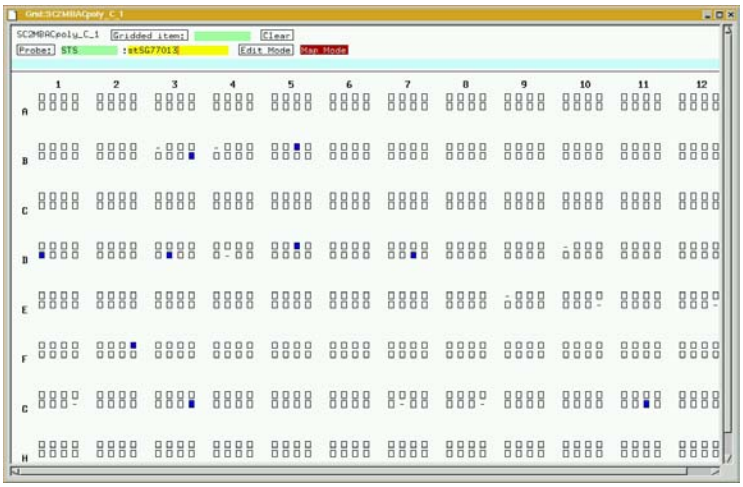
4.2.5 BAC contig assembly in FPC

Fingerprint and landmark content mapping data was imported in an FPC database. Following automated assembly, manual editing (Ian Mullenger and Lisa French) resulted in the construction of eleven seed contigs 0.4-1.4 Mb long.

A.



B.



C.

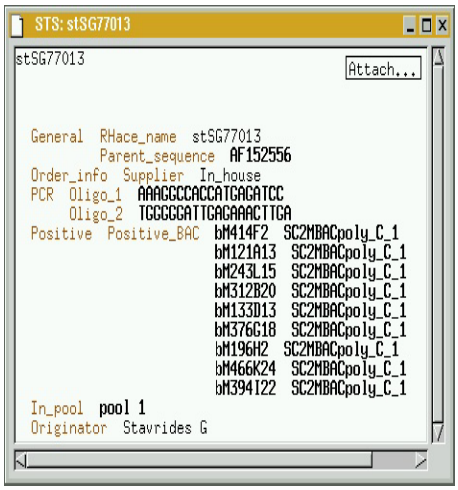


Figure 4.7: Example of landmark content mapping. (A) Hybridisation of polygrid filter 1 with STS stSG77013 (primer set designed using the mouse Ptprt mRNA sequence with EMBL accession number AF152556). Arrows indicate positive BAC clones. (B) 2musace view of polygrid filter 1. Each square represents a clone on the grid whereas dashes represent empty spaces. The dark blue filled squares indicate positives. (C) The positive clone names can be viewed in 2musace from the window of STS stSG77013.

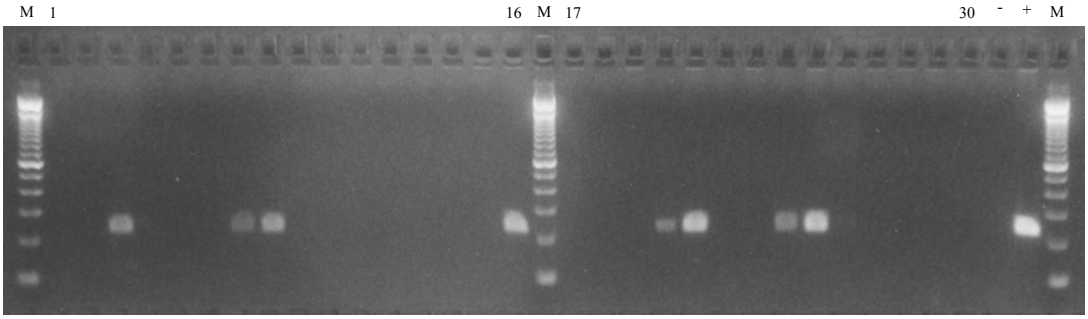
4.2.6 Gap closure

New STS markers were designed at the ends of each contig, using the publicly available BAC end sequences (Zhao *et al*., 2001; see Appendix 10). Clones having end-sequences for both ends were usually preferred.

Marker development, library screening, clone identification and analysis were performed as described for the gene based markers. The process was repeated until all gaps were closed, resulting in a single contig.

When only a few STS markers were available for library screening, BACs were identified by PCR screening of BAC DNA pools and selective hybridisation of the corresponding positive library filters (Figure 4.8).

A.



B.

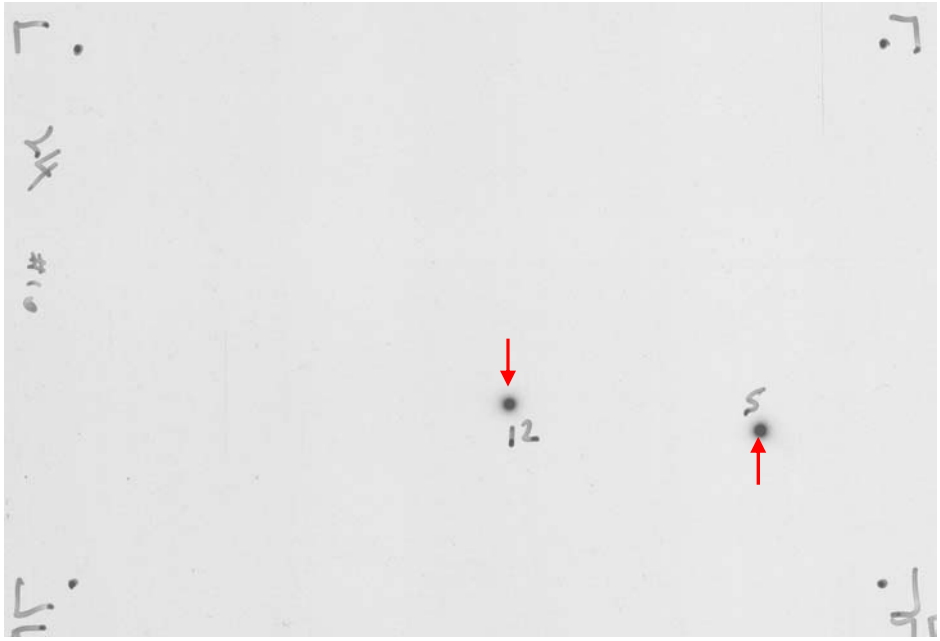


Figure 4.8: Example of PCR-based library screen. (A) End-STS stSG102484 was used to PCR screen the 30 DNA pools representing all RPCI-23 BAC clones. DNA pools representing BACs on filters 3, 7, 8, 16, 20, 21, 24 and 25 were positive. These filters were hybridised with the stSG102484 probe to identify the individual positive clones. (B) Autoradiograph of filter 24. Arrows indicate positive clones. In total, eleven positive BACs were identified on the eight filters.

4.2.7 Genetic markers

The synteny between human chromosome 20 and mouse chromosome 2 is well established (Peters *et al*., 1999; Carver and Stubbs, 1997; DeBry and Seldin, 1996). Markers from the mouse chromosome 2 genetic map (Dietrich *et al*., 1996, http://www-genome.wi.mit.edu/) were used to position the generated BAC contig on the chromosome. Initially, markers mapping at various positions on the genetic map were tested by hybridisation to the polygrids. Testing the genetic markers surrounding the positive ones followed that preliminary step. In total, 33/84 genetic markers tested were incorporated into the clone map. An example is shown in Figure 4.9 (for more details regarding these positive markers see Appendix 11). Besides the 44 markers that map outside the region of the clone map, seven markers were not placed either due to PCR failure, or non-specific hybridisation. The 33 mapped markers place the contig between 77.6-84.2 cM on the mouse chromosome 2 genetic map.

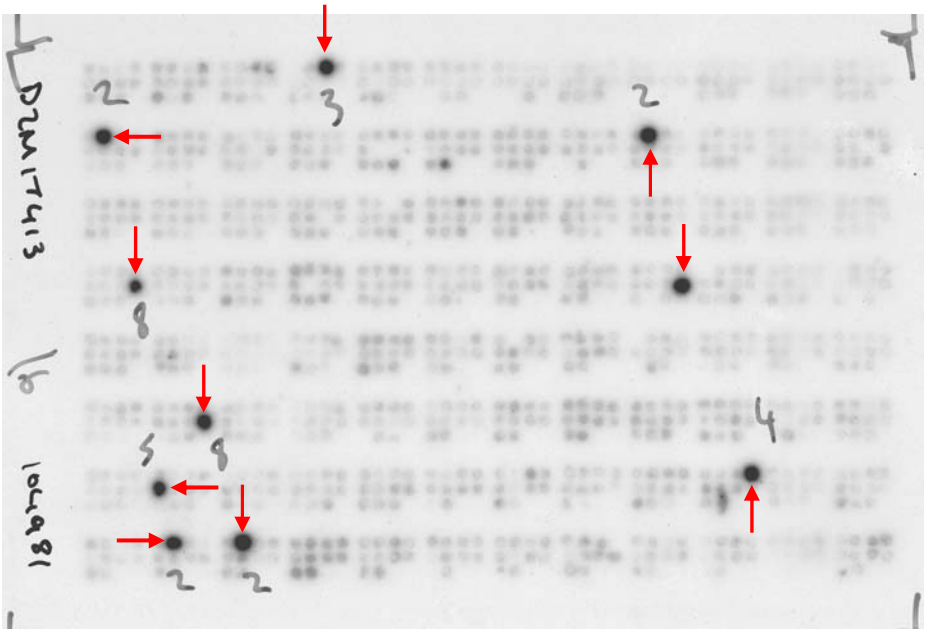


Figure 4.9: Hybridising marker D2MIT413 (stSG104981) to BAC polygrid 2. Arrows indicate the ten positive clones.

4.3 The sequence-ready bacterial clone map

The final sequence-ready map spans 9.8 Mb, based on an empirical, average-size estimate of 5 Kb per fingerprinting band (Figure 4.10). The map contains 66 gene based, 91 end-STS and 33 genetic markers. The genetic markers confirm the position and orientation of the contig on mouse chromosome 2 and allow integration to other maps, such as the genetic (Dietrich *et al*., 1996) and YAC physical (Nusbaum *et al*., 1999) maps.

The clone map contains 996 BACs. 524 are RPCI-23 clones and were placed on the map as described. A set of 472 RPCI-24 BACs was incorporated into the map (Ian Mullenger) using fingerprint data obtained from the publicly available database at GSC (http://www.bcgsc.bc.ca/projects/mouse_mapping/).

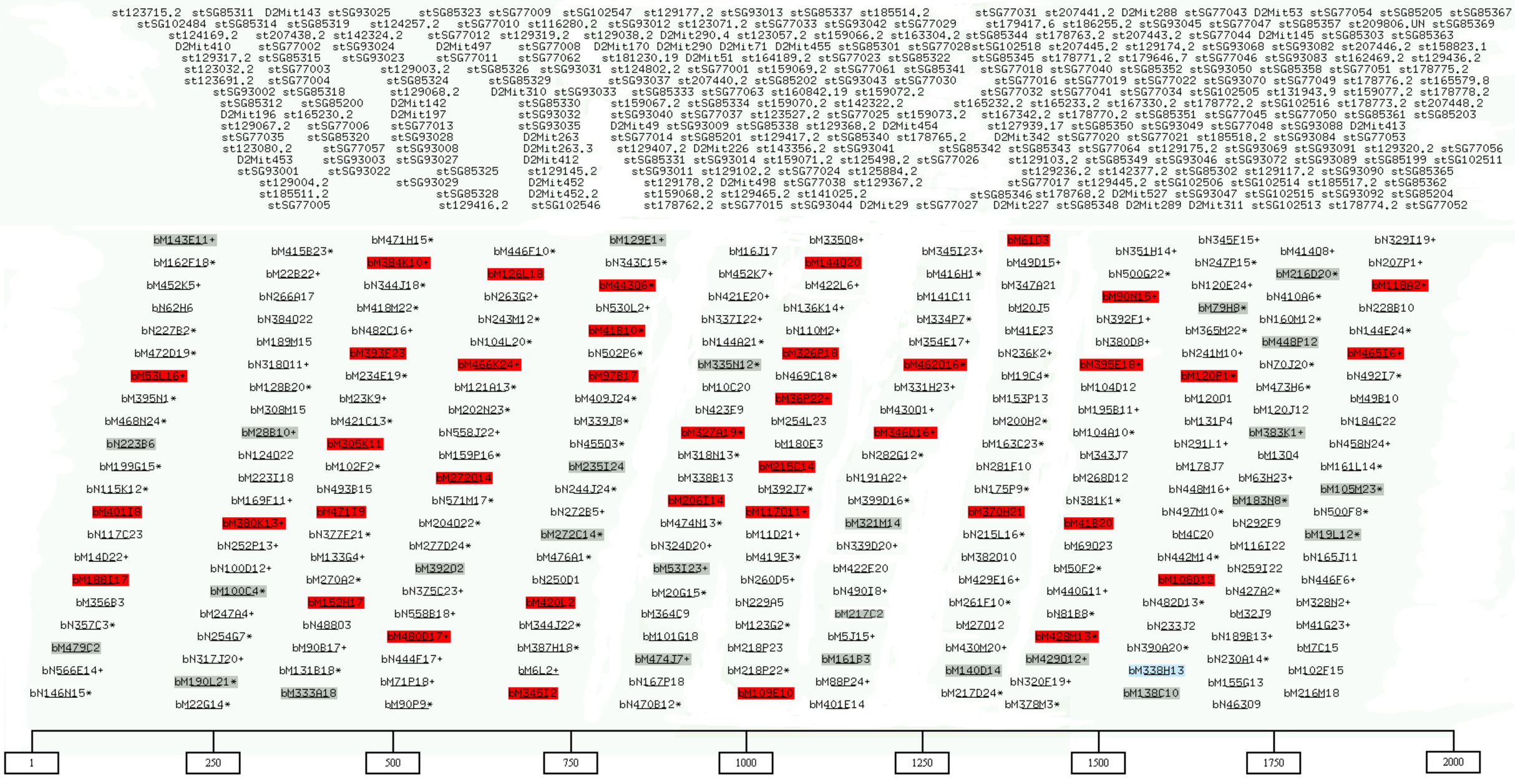
The order of the gene based STS markers on the clone map was compared to the order of the orthologous genes in the human sequence. Gene order was found to be conserved between the two species.

4.4 Tile path selection and sequencing

A set of 66 minimally overlapping BAC clones, the tile path, was selected and is currently being sequenced. To date, 5,541,112 bp of finished and 4,778,718 bp of unfinished (redundant) sequence have been generated from 38 and 27 clones, respectively. The unfinished sequence is in 105 contigs (>1 Kb long) with a minimum of six-fold sequence coverage per clone. The combined 10.3 Mb of mouse sequence is available at ftp://ftp.sanger.ac.uk/pub/mouse/.

All mapping and sequencing data reported here have been incorporated into the Ensembl mouse genome browser (v7.3b.2; http://www.ensembl.org/Mus_musculus/). In this release, the co-ordinates of the sequence contig on mouse chromosome 2 are between 159.8 Mb and 168.56 Mb. The size estimate of 8.76 Mb of non-redundant sequence is likely to be an underestimate because of the sequence gaps that remain in the unfinished clone sequence. At the time of analysis, one clone (bM338H13) was still in pre-sequencing.

Figure 4.10 (foldout): The mouse clone map. The STS markers are shown at the top, genetic markers are prefixed with D2Mit, whereas gene and clone end-sequence-based markers with stSG. Markers from the whole mouse-genome mapping effort, prefixed with st, are also shown but were not used during the construction of the clone map. BAC clones are represented as lines and prefixed with bM (RPCI-23 library) and bN (RPCI-24 library). BACs with identical fingerprints are not shown. Highlighted BAC clones are part of the tile path (blue, pre-sequencing stage; grey, being sequenced (includes those in finishing); and red, finished). The scale is in fingerprint bands (5 Kb per band).


st123715.2 stSG85311 D2Mit143 stSG93025 stSG85323 stSG77009 stSG102547 st129177.2 stSG93013 stSG85337 st185514.2 stSG77031 st207441.2 D2Mit288 stSG77043 D2Mit53 stSG77054 stSG85205 stSG85367
stSG102484 stSG85314 stSG85319 st124257.2 stSG77010 st116280.2 stSG93012 st123071.2 stSG77033 stSG93042 stSG77029 st179417.6 st186255.2 stSG93045 stSG77047 stSG85357 st209806.UN stSG85369
st124169.2 st207438.2 st142324.2 stSG77012 st129319.2 st129038.2 D2Mit290.4 st123057.2 st159066.2 st163304.2 stSG85344 st178763.2 st207443.2 stSG77044 D2Mit145 stSG85303 stSG85363
D2Mit410 stSG77002 stSG93024 D2Mit497 stSG77008 D2Mit170 D2Mit290 D2Mit71 D2Mit455 stSG85301 stSG77028 stSG102518 st207445.2 st129174.2 stSG93068 stSG93082 st207446.2 st158823.1
st129317.2 stSG85315 stSG93023 stSG77011 stSG77062 st181230.19 D2Mit51 st164189.2 stSG77023 stSG85322 stSG85345 st178771.2 st179646.7 stSG77046 stSG93083 st162469.2 st129436.2
st123032.2 stSG77003 st129003.2 stSG85326 stSG93031 st124802.2 stSG77001 st159069.2 stSG77061 stSG85341 stSG77018 stSG77040 stSG85352 stSG93050 stSG85358 stSG77051 st178775.2
st123691.2 stSG77004 stSG85324 stSG85329 stSG93037 st207440.2 stSG85202 stSG93043 stSG77030 stSG77016 stSG77019 stSG77022 stSG93070 stSG77049 st178776.2 st165579.8
stSG93002 stSG85318 st129068.2 D2Mit310 stSG93033 stSG85333 stSG77063 st160842.19 st159072.2 stSG77032 stSG77041 stSG77034 stSG102505 st131943.9 st159077.2 st178778.2
stSG85312 stSG85200 D2Mit142 stSG85330 st159067.2 stSG85334 st159070.2 st142322.2 st165232.2 st165233.2 st167330.2 st178772.2 stSG102516 st178773.2 st207448.2
D2Mit196 st165230.2 D2Mit197 stSG93032 stSG93040 stSG77037 st123527.2 stSG77025 st159073.2 st167342.2 st178770.2 stSG85351 stSG77045 stSG77050 stSG85361 stSG85203
st129067.2 stSG77006 stSG77013 stSG93035 D2Mit49 stSG93009 stSG85338 st129368.2 D2Mit454 st127939.17 stSG85350 stSG93049 stSG77048 stSG93088 D2Mit413
stSG77035 stSG85320 stSG93028 D2Mit263 stSG77014 stSG85201 st129417.2 stSG85340 st178765.2 D2Mit342 stSG77020 stSG77021 st185518.2 stSG93084 stSG77053
st123080.2 stSG77057 stSG93008 D2Mit263.3 st129407.2 D2Mit226 st143356.2 stSG93041 stSG85342 stSG85343 stSG77064 st129175.2 stSG93069 stSG93091 st129320.2 stSG77056
D2Mit453 stSG93003 stSG93027 D2Mit412 stSG85331 stSG93014 st159071.2 st125498.2 stSG77026 st129103.2 stSG85349 stSG93046 stSG93072 stSG93089 stSG85199 stSG102511
stSG93001 stSG93022 stSG85325 st129145.2 stSG93011 st129102.2 stSG77024 st125884.2 st129236.2 st142377.2 stSG85302 st129117.2 stSG93090 stSG85365
st129004.2 stSG93029 D2Mit452 st129178.2 D2Mit498 stSG77038 st129367.2 stSG77017 st129445.2 stSG102506 stSG102514 st185517.2 stSG85362
st185511.2 stSG85328 D2Mit452.2 st159068.2 st129465.2 st141025.2 stSG85346 st178768.2 D2Mit527 stSG93047 stSG102515 stSG93092 stSG85204
stSG77005 st129416.2 stSG102546 st178762.2 stSG77015 stSG93044 D2Mit29 stSG77027 D2Mit227 stSG85348 D2Mit289 D2Mit311 stSG102513 st178774.2 stSG77052
bM143E11+
bM162F18*
bM452K5+
bN62H6
bN227B2*
bM472D19*
bM53L16+
bM395N1*
bM468N24*
bN223B6
bM199G15*
bN115K12*
bM401I8
bN117C23
bM14D22+
bM188I17
bM356B3
bN357C3*
bM479C2
bN566E14+
bN146N15*
bM415B23*
bM22B22+
bN266A17
bN384D22
bM189M15
bN318D11+
bM128B20*
bM308M15
bM28B10+
bN124D22
bM223I18
bM169F11+
bM380K13+
bN252P13+
bN100D12+
bM100C4*
bM247A4+
bN254G7*
bN317J20+
bM190L21*
bM22G14*
bM471H15*
bM384K10+
bN344J18*
bM418M22*
bN482C16+
bM393F23
bM234E19*
bM23K9+
bM421C13*
bM305K11
bM102F2*
bN493B15
bM471I9
bN377F21*
bM133G4+
bM270A2*
bM152H17
bN488D3
bM90B17+
bM131B18*
bM333A18
bM446F10*
bM126L18
bN263G2+
bN243M12*
bN104L20*
bM466K24+
bM121A13*
bM202N23*
bN558J22+
bM159P16*
bM272D14
bN571M17*
bM204D22*
bM277D24*
bM392D2
bN375C23+
bN558B18+
bM480D17+
bN444F17+
bM71P18+
bM90P9*
bM129E1+
bN343C15*
bM443D6*
bN530L2+
bM41B10*
bN502P6*
bM97B17
bM409J24*
bM339J8*
bN455D3*
bM235I24
bN244J24*
bN272B5+
bM272C14*
bM476A1*
bN250D1
bM420L2
bM344J22*
bM387H18*
bM6L2+
bM345I2
bM16J17
bM452K7+
bN421E20+
bN337I22+
bN144A21*
bM335N12*
bM10C20
bN423E9
bM327A19*
bM318N13*
bM338B13
bM206I14
bM474N13*
bN324D20+
bM53I23+
bM20G15*
bM364C9
bM101G18
bM474J7+
bN167P18
bN470B12*
bM335D8+
bM144D20
bM422L6+
bN136K14+
bM110M2+
bM326P18
bN469C18*
bM36P22+
bM254L23
bM180E3
bM215C14
bM392J7*
bM117D11+
bM11D21+
bM419E3*
bN260D5+
bN229A5
bM123G2*
bM218P23
bM218P22*
bM109E10
bM345I23+
bM416H1*
bM141C11
bM334P7*
bM354E17+
bM462D16*
bM331H23+
bM430D1+
bM346D16+
bN282G12*
bN191A22+
bM399D16*
bM321M14
bN339D20+
bM422E20
bN490I8+
bM217C2
bM5J15+
bM161B3
bM88P24+
bM401E14
bM61D3
bM49D15+
bM347A21
bM20J5
bM41E23
bN236K2+
bM19C4*
bM153P13
bM200H2*
bM163C23*
bN281E10
bN175P9*
bM370H21
bN215L16*
bM382D10
bM429E16+
bM261F10*
bM27D12
bM430M20+
bM140D14
bM217D24*
bN320F19+
bM378M3*
bN351H14+
bN500G22*
bM90N15+
bN392F1+
bN380D8+
bM395E18+
bM104D12
bM195B11+
bM104A10*
bM343J7
bM268D12
bN381K1*
bM41B20
bM69D23
bM50F2*
bM440G11+
bN81B8*
bM428M13*
bM429D12+
bN345F15+
bN247P15*
bN120E24+
bM79H8*
bM365M22*
bN241M10+
bM120P1*
bM120D1
bM131P4
bN291L1+
bM178J7
bN448M16+
bN497M10*
bM4C20
bN442M14*
bM108D12
bN482D13*
bN233J2
bN390A20*
bM338H13
bM138C10
bM414D8+
bM216D20*
bN410A6*
bN160M12*
bM448P12
bN70J20*
bM473H6*
bM120J12
bM383K1+
bM13D4
bM63H23+
bM183N8*
bN292E9
bM116I22
bN259I22
bN427A2*
bM32J9
bN189B13+
bN230A14*
bM155G13
bN46309
bN329I19+
bN207P1+
bM118A2+
bN228B10
bN144E24*
bM465I6+
bN492I7*
bM49B10
bN184C22
bN458N24+
bM161L14*
bM105M23*
bN500F8*
bM19L12*
bN165J11
bN446F6+
bM328N2+
bM41G23+
bM7C15
bM102F15
bM216M18
1
250
500
750
1000
1250
1500
1750
2000

4.5 Long range comparative sequence analysis

4.5.1 Repeat content analysis

The 10,319,830 bp of redundant (finished and unfinished) mouse sequence has an average GC content of 46.2% compared to 45.2% of the human sequence. The results of RepeatMasker (Smit and Green, unpublished) analysis of both human and mouse sequences (Table 4.2) suggest that in both organisms approximately 38% of repeat sequence is due to SINE elements. The more abundant LINE element in both organisms is L1, whereas the sequence coverage of LTRs is approximately the same.

The lower repeat content (32.1%) detected in the mouse sequence compared to the human (49.6%) does not necessarily imply a higher percentage of non-repetitive sequence in the mouse. For example, it is known that the faster rate of substitution per million years in rodent lineages compared to hominid lineages (Li *et al*., 1996) makes the detection of ancient elements more difficult (IHGSC, 2001). In addition, the list of known repeats in the mouse may be less complete than for the human (IHGSC, 2001).

Table 4.2: Repeat content analysis. 10,099,164 bp of non-redundant human sequence and 10,319,830 bp of redundant mouse sequence were analysed using RepeatMasker v_6_2001.

		HUMAN		MOUSE	
		Total length (base pairs)	*Percentage of sequence*	*Total length (base pairs)*	*Percentage of sequence*
SINE		**1,989,403**	**19.70**	**1,227,560**	**11.9**
	Alu	1,473,655	14.59	-	-
	MIR	515,748	5.11	111,621	1.1
	B1	-	-	377,268	3.6
	B2-B4	-	-	716,333	6.9
	ID	-	-	22,338	0.2
LINE		**1,663,145**	**16.47**	**776,355**	**7.5**
	L1	1,045,075	10.35	687,572	6.6
	L2	588,625	5.83	83,821	0.8
	L3/CR1	29,445	0.29	4,962	0.05
LTR		**788,760**	**7.81**	**776,223**	**7.5**
	MaLRs	424,261	4.20	467,951	4.5
	ERVL	147,960	1.47	59,827	0.6
	ERVL classI	209,244	2.07	12,527	0.12
	ERVL classII	5,218	0.05	130,359	1.2
DNA elements		**396,682**	**3.93**	**101,820**	**1**
	MER1 type	248,047	2.46	79,110	0.76
	MER2 type	80,503	0.80	10,733	0.1
Unclassified		13,667	0.14	18,377	0.17
Total IR		**4,851,657**	**48.04**	**2,900,335**	**28.1**
Small RNA		5,180	0.05	7,986	0.08
Satellites		10,786	0.11	305	0.00
Simple repeats		96,377	0.95	332,003	3.2
Low complexity		48,731	0.48	76,552	0.74
Total bases masked		**5,010,905**	**49.62**	**3,316,142**	**32.1**

4.5.2 BLAST searches

The finished human sequence was used to perform BLAST searches against the available mouse sequence of each clone. At a 60% identity cut-off, a redundant set of 6,213 mouse BLAST hits was obtained (893 BLAST hits were duplicates because of the sequence redundancy, thus the non-redundant BLAST hit set was 5,320). The size distribution of BLAST hits is shown on Figure 4.11.

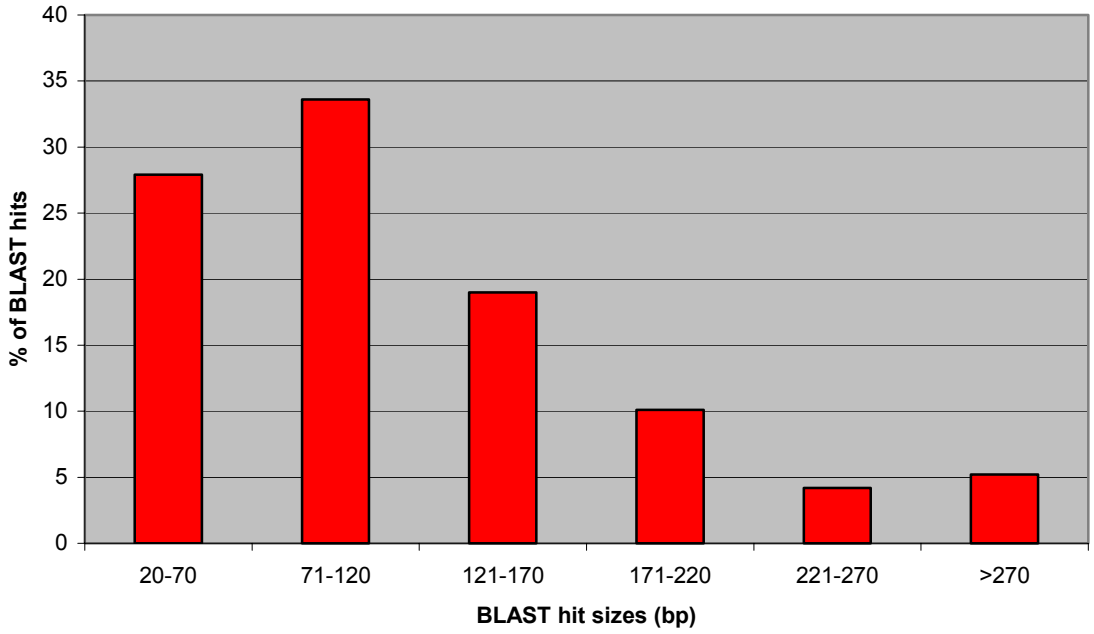


Figure 4.11: Size distribution of mouse BLAST hits

Results were imported in 20ace and inspected manually (Figure 4.12). At the 60% identity cut-off, mouse matches were obtained for all (at least one exon) of the 99 annotated human coding genes (Table 4.3). Overall mouse hits were obtained for >96% of annotated coding gene exons.

The high number of mouse BLAST hits across exonic regions of coding genes contrasts sharply with the very low number of hits across the exonic regions of annotated putative genes and pseudogenes. Putative genes have been annotated on the basis of spliced ESTs although no open reading frame was determined (Deloukas *et al*., 2001). Matches were obtained for only five of the 30 putative genes and only two of the 36 annotated pseudogenes.

On the basis of the mouse clone map and human annotation, (coding) gene order is fully conserved between human and mouse, suggesting absence of major (>1 Mb) rearrangements in this region. However, until we obtain the full finished mouse sequence we cannot exclude the presence of any small local rearrangements.

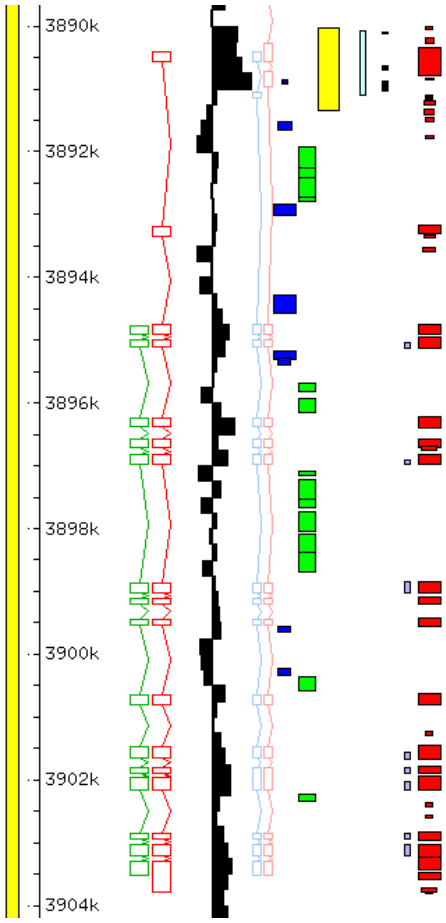


Figure 4.12: ACeDB view of human:mouse BLAST search results. The graphic overview shows the annotated features present in the *human* sequence between 3,890-3,904 Kb. This region encodes for a novel gene (C20orf67) that is similar to the *Drosophila melanogaster* gene CG11399. Sequence analysis identified a large number of splicing ESTs and an IMAGE clone sequence (clone 3640928 (BC013365)) that shows high homology to this sequence. The evidence was used to determine the exon/intron boundaries of the transcribed mRNA (structure shown in red). A 2,112 bp long putative ORF was also annotated (shown in green). Other features shown (from left to right) include a GC-content plot, Genscan (light blue) and FGENESH (light red) predictions, LINE repeats (blue), SINE repeats (green), predicted CpG islands (yellow), PromoterInspector results (light blue), and Eponine predicted TS sites (black). Homologies with *Tetraodon nigroviridis* identified using Exofish are shown in violet. BLAST search results against the mouse genomic sequence are shown at the far right (red). The size of the boxes indicates the extent and percentage identity of sequence homology (box length and width respectively). All red boxes (mouse hits) correspond to sequences from the mouse clone bM61O3. Other sequence features are not shown for clarity.

Table 4.3: Human:mouse BLAST searches. Column one reports all annotated human coding genes in the order they map in the sequence of 20q12-13.2. Column two reports the type of coding gene (known or novel). Column three reports the names of mouse clones, the sequence of which was found to share homology with the exons of genes in column one.

Human gene name	Type of human gene	Mouse BAC clone
KRML	Known	333A18
TOP1	Known	471I9
PLCG1	Known	393F23
TIX1	Novel	393F23
LIPN3L	Novel	393F23
C20orf130	Novel	393F23
KIAA1335	Novel	384K10
PTPRT	Known	466K24
SFRS6	Known	206I14
KIAA0681	Known	335N12
SGK2	Novel	335N12
C20orf9	Novel	335N12
MYBL2	Known	335N12
C20orf65	Novel	335N12
C20orf100	Novel	117O11
JPH2	Novel	117O11
C20orf111	Novel	215C14
GDAP1L1	Novel	36P22
C20orf142	Novel	36P22
R3HDML	Novel	36P22
HNF4A	Known	36P22
C20orf121	Novel	36P22
TDE1	Known	144O20
PKIG	Known	144O20
ADA	Known	144O20
WISP2	Known	217C2
KCNK15	Known	217C2
C20orf190	Novel	321M14
YWHAB	Known	321M14
C20orf119	Novel	321M14
TOMM34	Known	321M14
STK4	Known	346D16
KCNS1	Known	346D16
PRG5	Known	346D16
C20orf122	Known	346D16
PI3	Known	462O16
SEMG1	Known	462O16
SEMG2	Known	462O16
SLPI	Known	462O16
MATN4	Known	462O16
RBPSUHL	Known	462O16
SDC4	Known	462O16
C20orf169	Novel	140D14

Human gene name	Type of human gene	Mouse BAC clone
C20orf10	Novel	140D14
C20orf35	Novel	140D14
PIGT	Novel	140D14
WFDC2	Known	140D14
SPINT3	Known	140D14
C20orf171	Novel	140D14
SPINLW1	Novel	140D14
C20orf170	Novel	140D14
C20orf146	Novel	140D14
C20orf137	Novel	370H21
C20orf168	Novel	370H21
WFDC3	Novel	370H21
C20orf167	Novel	370H21
UBE2C	Known	370H21
TNNC2	Known	370H21
C20orf161	Novel	370H21
PTE1	Known	370H21
C20orf164	Novel	370H21
C20orf162	Novel	61O3
C20orf165	Novel	61O3
C20orf163	Novel	61O3
PPGB	Known	61O3
PLTP	Known	61O3
C20orf67	Novel	61O3
ZNF335	Novel	61O3
MMP9	Known	61O3
SLC12A5	Novel	61O3
NCOA5	Novel	61O3
TNFRSF5	Known	428M13
C20orf25	Novel	428M13
C20orf5	Novel	41B20
KIAA1834	Novel	41B20
C20orf157	Novel	41B20
ZNF334	Novel	41B20
C20orf123	Novel	395E 18
SLC13A3	Known	395E 18
C20orf64	Novel	395E 18
SLC2A10	Novel	90N15
EYA2	Known	138C10
PRKCBP1	Known	138C10
NCOA3	Known	120P1
KIAA1247	Novel	120P1
KIAA1415	Novel	183N8
ARFGEF2	Known	216D20
CSE1L	Known	216D20
STAU	Known	19L12
DDX27	Novel	19L12

Human gene name	Type of human gene	Mouse BAC clone
KIAA1404	Novel	19L12
KCNB1	Known	105M23
PTGIS	Known	105M23
B4GALT5	Known	105M23
KIA0939	Novel	328K5
SPATA2	Known	465I6
ZNF313	Novel	465I6
SNAI1	Known	118A2
UBE2V1	Known	118A2

4.5.3 An evaluation of the current human sequence annotation

Of all the annotated coding exons in the region, 72.2% are identical to exons predicted by both FGENESH (Salamov and Solovyev, 2000; optimised for human gene prediction, Solovyev, unpublished) and Genscan (Burge and Karlin, 1997). Identical predictions by both programs were also obtained in 226 loci outside annotated exons (155 in intergenic regions and 71 in intragenic regions). These loci may represent un-annotated coding exons. When assessed, only 28 of these double predictions are supported by mouse-conserved sequences (eleven map in introns and seventeen in intergenic regions). Since more than 96% of the annotated coding exons are supported by mouse hits it is likely that most of the 198/226 FGENESH-Genscan predictions, which are not supported by mouse hits, do not represent real exons.

Furthermore, STSs were designed for five of the seventeen loci that map outside annotated genes and used to PCR screen seven cDNA libraries. No positives were obtained for the five loci tested. In contrast, screening of the 51 novel coding genes identified at least one positive cDNA library in 46 cases (90% detection; chapter III). These findings imply that the 28 loci either do not correspond to exonic sequences or do

correspond to parts of transcripts not represented in the screened cDNA libraries. In fact, the recently released mouse mRNA BC002161 supports two of the seventeen loci as exons of C20orf130 suggesting that this set of 28 loci may be enriched in un-annotated exons. Even if we assume that all 28 loci represent coding exons, then they would only represent 3% of annotated coding exons in the region. This is in agreement with our published estimate (Deloukas *et al*., 2001).

4.5.4 PipMaker analysis

PipMaker (Schwartz *et al*., 2000) was used to align the complete human sequence with 9.4 Mb of finished and unfinished mouse sequence (Webb Miller, Pennsylvania State University; this earlier sequence version consisted of 410 sequence contigs). Part of the generated Pip plot is shown in Figure 4.13. At the time of analysis, three regions at 120-145 Kb, 2,845-3,235 Kb and 4,670-4,710 Kb from the human reference sequence had no mouse sequence and were excluded from the statistical analysis.

Table 4.4 reports the percentage of nucleotides covered by PipMaker alignments for the following types of regions: (1) exons of the annotated genes (including UTRs but excluding pseudogenes and transcripts), (2) introns, (3) the 200 bp upstream of the annotated TS site, (4) the 1,000 bp upstream of the annotated TS site, and (5) intergenic regions. For the same regions, we also determined the percentage of nucleotides contained in "strongly aligning regions" (at least 100 bp that align without a gap and at least 70% nucleotide identity).

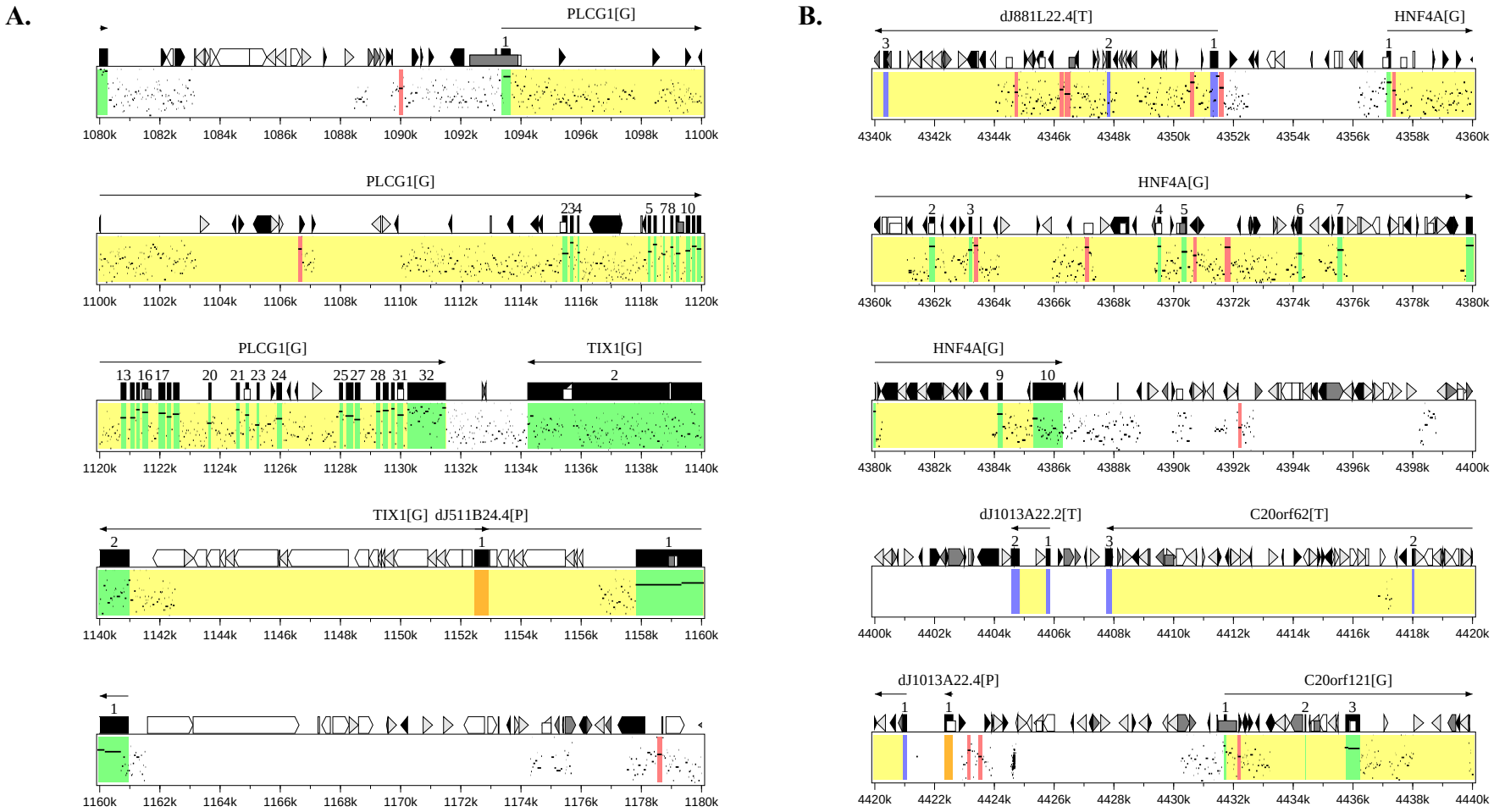


Figure 4.13. Pip plots for two genomic regions. (A) The region containing the PLCG1 and TIX genes and the pseudogene dJ511B24.4 (within the intron of TIX1). (B) The region containing the HNF4A gene, part of the C20orf121 gene, the putative genes (or parts of) dJ881L22.4, dJ1013A22.2 and C20orf62 and the dJ1013A22.4 pseudogene. The positions of gap-free segments of alignments are plotted along the horizontal axis by using co-ordinates in the human sequence, and the percent identity is plotted along the vertical axis (from 50% to 100%). Features of the human sequence are annotated along the top of each graph. Annotated features are labelled above arrows showing the direction of transcription, and exons are shown as numbered black rectangles. Low rectangles denote CpG islands, shown as white if 0.6 ≤ CpG/GpC < 0.75 and as grey if CpG/GpC ≥ 0.75. Interspersed repeats are shown by the following icons: light grey triangles are SINEs other than MIRs, black triangles are MIRs, black pointed boxes are LINE2s, and dark grey triangles and pointed boxes are other kinds of interspersed repeats, such as long terminal repeat elements and DNA transposons. Areas within the pip are coloured yellow for introns, green for exons of coding genes, blue for exons of putative genes, orange for pseudogenes and light red for matches longer than 100 bp in non-coding, non-repetitive regions with percent identities of at least 70%.

Table 4.4: PipMaker analysis. Column two reports the percentage of non-repetitive nucleotides that align in various classes of genomic segments. Column three reports the percentage of non-repetitive nucleotides contained in regions of at least 100 bp that align without a gap and at least 70% nucleotide identity.

Regions studied	Aligns	Strong
Exon	93.7	53.7
Intron	51.5	4.5
Upstream 200	83.2	13.5
Upstream 1000	68.4	7.1
Intergenic	42.4	4.1

Excluding exons, the regions 200 bp upstream of the annotated gene-starts show the highest alignability. This is not unexpected, since these regions probably correspond to un-annotated 5' UTRs.

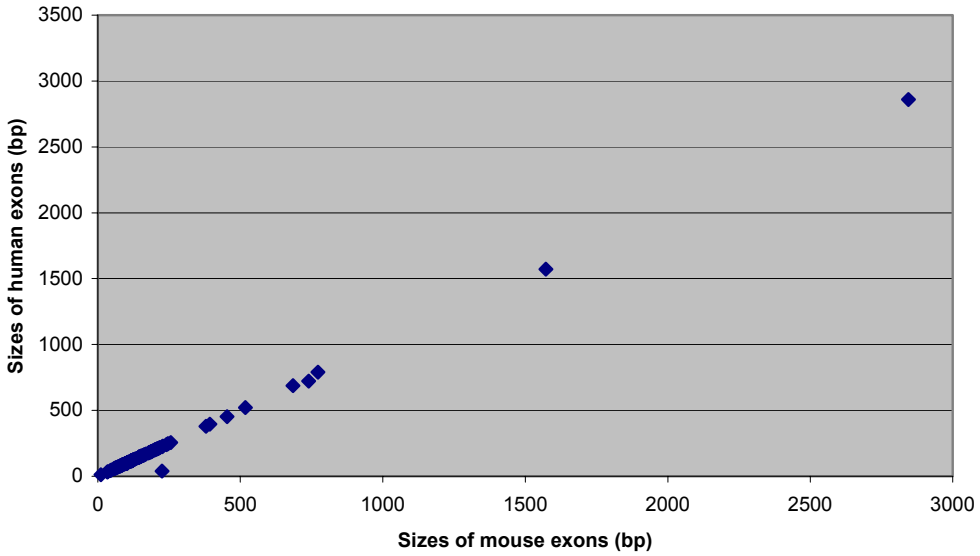
4.6 Finished mouse sequence analysis

The sequence of all finished clones is subjected to the standard Sanger Institute analysis (section 3.2). Manual annotation of gene structures is performed by Dr. Laurens Wilming (Sanger Institute). As for the human chromosome 20 sequence, I conduct the interactive checking process. As of June 2002, 26 mouse clones were analysed and an additional five were also annotated.

The computational analysis and annotation of approximately 700 Kb of mouse sequence (five clones) resulted in the identification of >100 mouse exons (approximately 10% of the total number expected in the whole region). A size correlation of 103 pairs of orthologous human and mouse coding exons (fully supported by human expressed data) is shown in Figure 4.14, whereas the equivalent comparison of 96 introns is shown in Figure 4.15. The average exon and intron sizes were 198.5 bp and 4,478.1 bp for the human, and 200.6 bp and 3,531.8 bp for the mouse respectively.

In total, size differences were found in ten of the 103 orthologous pairs examined and in all cases they are either three, or a multiple of three nucleotides, indicating conservation of the open reading frame. As shown in Figure 4.15, orthologous introns lack size conservation. Absence of size conservation was also observed across non-coding exons such as 5' and 3' UTRs. In addition, differences were also observed in the number of 5' untranslated exons. On average, introns were 1.27-fold longer in human, suggesting that genic regions are more compact in the mouse genome. This could be due to differences in the type of repetitive elements in the sequence of each species as well as the relative abundance of repeats.

A.



B.

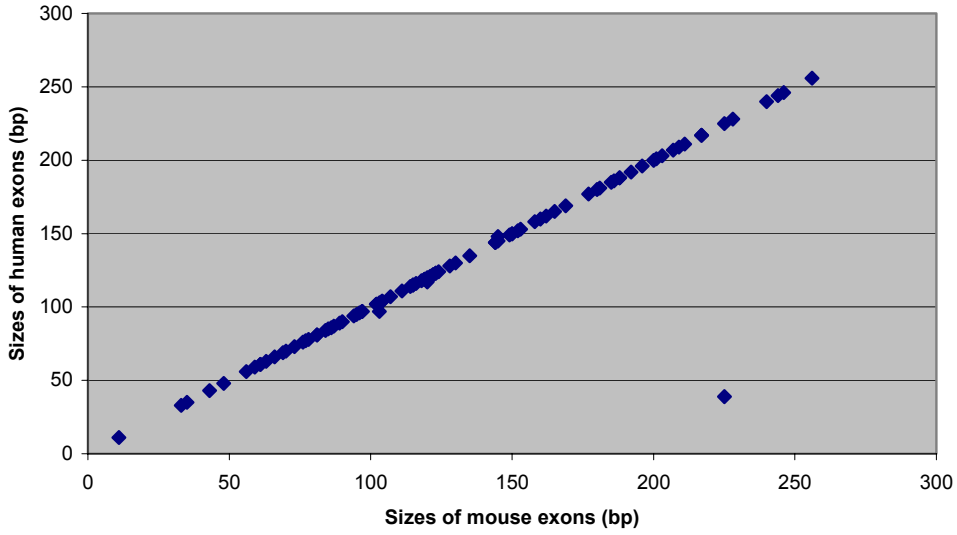
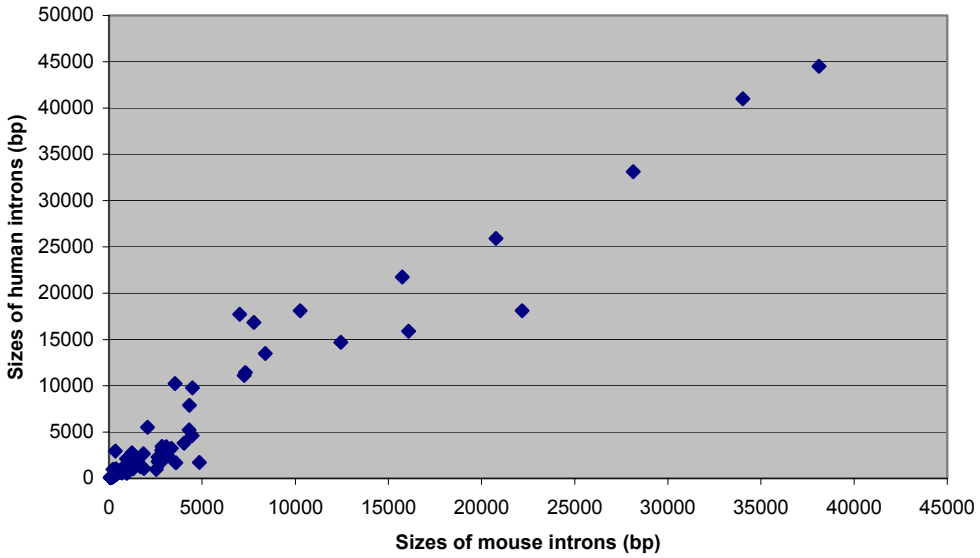


Figure 4.14: (A) Scatter plots for the exon sizes between human and mouse. (B) Detailed view of the 0-300 bp window. The only significant size deviation is between the first exon of human C20orf100 (39 bp) and the orthologous bM117O11.1 (225 bp). Note that the C20orf100 exon is incomplete (no starting methionine has been found).

A.



B.

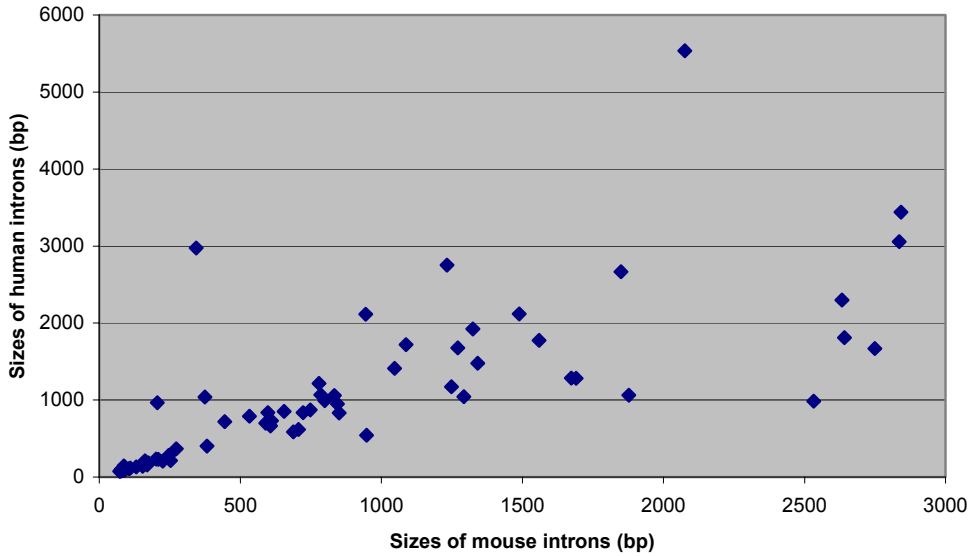


Figure 4.15: (A) Scatter plots for the intron sizes between human and mouse. (B) Detailed view of the 0-3,000 bp window.

As an example, the sequences of three orthologous human:mouse gene pairs (PLCG1:Plcg1, TIX1:bM393F23.2 and C20orf111:bM117O11.3) were studied in more detail. The various features of these genes are reported in Table 4.5.

PLCG1 and Plcg1 have similar genomic sizes and encode for the same number of exons. The same is also true for C20orf111 and bM117O11.3. In contrast, TIX1 and bM393F23.2 differ both in size and exon number. These differences are due to mouse ESTs that splice further upstream and identify two additional 5' UTR exons (exons 1a and 1b). Human ESTs and vectorette sequences do not support these exons, which also lack human:mouse sequence homology. In the mouse sequence, a rat splicing EST (AI071486) supports an alternative transcript that has an alternative 5' exon for bM393F23.2 (exon 1a'), upstream of exon 1a. BLAST searches with exon 1a' against the human sequence obtain a hit ~113 Kb upstream of the annotated TIX1 start. Whether this sequence is part of the TIX1 gene remains to be experimentally verified. This extended TIX1 genomic sequence was used for the analysis described below.

It is worth mentioning that two pseudogenes have been annotated within the orthologous sequences of TIX1 and bM393F23.2. The human pseudogene (dJ511B24.4) resides within the intron of TIX1 and is similar to the 60S ribosomal protein L23A. The mouse pseudogene (bM393F23.3) resides between exons 1a and 1b of bM393F23.2 and is similar to the glyceraldehyde-3-phosphate dehydrogenase protein (Gapd). bM393F23.3 and dJ511B24.4 are not present in the corresponding human and mouse sequences, respectively.

Nucleotide alignments of the orthologous CDSs are shown in Figure 4.16 whereas predicted-protein alignments are shown in Figure 4.17. The 15 bp difference in CDS size between TIX1 and bM393F23.2 is due to two insertions of 12 and 3 bp in human

exon 1. Compared to bM117O11.3, the CDS of C20orf111 also has a 3 bp insertion in exon 4. DNA and protein sequence identities per gene are given in Table 4.6. In all three cases, coding regions (CDSs) share higher homology than untranslated regions.

Table 4.5: Sequence features of three gene pairs

	Locus size (bp)	Exons (total)	mRNA size (bp)	CDS size (bp, including stop codon)	No of 5' UTR exons	No of 3' UTR exons	Stop codon	Poly (A) signal (bp)	Size differences of splicing exons	5' CpGs	Protein size (aa)
PLCG1	38,146	32	5,151	3,873	1 (68bp)	1(1,210bp)	tag	AATAAA(-23)	-	yes	1,290
Plcg1	31,554	32	5,107	3,873	1 (84bp)	1(1,150bp)	tag	AATAAA(-22)	-	yes	1,290
TIX1	26,726 (extended, 139,530)	2	9,871	2,871	1(260bp)	1(6,740bp)	tga	AATAAA(-22)	ex1[1]	no	956
bM393F23.2	106,671	4	9,109	2,856	3(420bp)	1(5,833bp)	tga	AATAAA(-25)	ex3[1]	yes yes	951
C20orf111	14,296	4	1,572	879	2(137bp)	1(556bp)	tga	AATAAA(-25)	ex2(118bp)[2]	yes	292
bM117O11.3	13,707	4	1,620	876	2(161bp)	1(583bp)	tga	AATAAA(-31)	ex2(121bp)[2]	yes	291

Splice site differences	3' intron/5' exon		Splice site differences	3' exon/5' intron	
	Human	Mouse		Human	Mouse
PLCG1- Plcg1 (ex8)	ag/gg	ag/ga	PLCG1- Plcg1 (ex14)	ac/gt	at/gt
PLCG1- Plcg1 (ex10)	ag/tt	ag/ct	PLCG1- Plcg1 (ex23)	tg/gt	cg/gt
PLCG1- Plcg1 (ex16)	ag/gt	ag/gc	PLCG1- Plcg1 (ex26)	tg/gt	ag/gt
C20orf111- bM117O11.3(ex2)	ag/tg	ag/ta	PLCG1- Plcg1 (ex28)	gg/gt	ag/gt

[1]The annotated TIX1 is in two exons whereas the orthologous bM393F23.2 is in four (1a, 1b, 3 and 4. Exon 1a', which is based on rat homologies is not included because it is part of an annotated isoform). Exon 1 of TIX1 corresponds to exon 3 of bM393F23.2. The coding part of the TIX1 exon 1 is 2,860 bp whereas the corresponding size for bM393F23.2 exon 3 is 2,845 bp.

[2] The size difference is in the 5' UTR.

A. PLCG1:Plcg1 (CDS)

```
PLCG1.cds     1 ATGGCGGGCGCCGCGTCCCCTTGCGCCAACGGCTGCGGGCCCGGCGCGCCCTCGGACGCCGAGGTGCTGCACCTCTGCCGCAGCCTCGAGGTGGGCACCGTCATGACTTT
plcg1.cds     1 ATGGCGGGCGTCGCGACCCCCTGCGCCAACGGCTGCGGGCCTGGCGCACCCTCGAAGCCGAGGTGCTGCACCTCTGCCGCAGCCTCGAGGTGGGCACCGTCATGACTTT

PLCG1.cds   111 GTTCTACTCCAAGAAGTCGCAGCGACCCGAGCGGAAGACCTTCCAGGTCAAGCTGGAGACGCGCCCAGATCACGTGGAGCCGCGGCGCCGACAAGATCGAGGGGGCCATTG
plcg1.cds   111 GTTCTACTCCAAGAAGTCGCAGCGGCCAGAACGGAAGACCTTCCAGGTCAAGTTGGAGACGCGCCCAGATCACATGGAGCCGCGGCGCGGACAAAATCGAGGGGTCCATTG

PLCG1.cds   221 ACATTCGTGAAATTAAGGAGATCCGCCCAGGGAAGACCTCACGGGACTTTGATCGCTATCAAGAGGACCCAGCTTTCCGGCCGGACCAGTCACATTGCTTTGTCATTCTC
plcg1.cds   221 ATATCCGAGAAATCAAGGAGATCCGTCCAGGGAAGACTTCTCGGGACTTTGACCGCTACCAAGAAGACCCTGCCTTCCGGCCAGATCAATCACACTGCTTTGTCATCCTC

PLCG1.cds   331 TATGGAATGGAATTTCGCCTGAAAACGCTGAGCCTGCAAGCCACATCTGAGGATGAAGTGAACATGTGGATCAAGGGCTTAACTTGGCTGATGGAGGATACATTGCAGGC
plcg1.cds   331 TATGGAATGGAATTCCGCTTGAAGACCCTGAGCCTGCAAGCCACATCTGAGGATGAAGTGAACATGTGGATCAAGGGCTTAACTTGGCTAATGGAGGATACACTGCAGGC

PLCG1.cds   441 ACCCACACCCCTGCAGATTGAGAGGTGGCTCCGGAAGCAGTTTTACTCAGTGGATCGGAATCGTGAGGATCGTATATATCAGCCAAGGACCTGAAGAACATGCTGTCCCAGG
plcg1.cds   441 AGCCACACCCCTGCAAATTGAGAGGTGGCTCCGGAAGCAGTTCTACTCAGTGGATCGTAACGGAGGGATCGTATATCAGCCAAGGACTTGAAGAACATGCTGTCTCAGG

PLCG1.cds   551 TCAACTACCGGGTCCCCAACATGCGCTTCCTCCGAGAGCGGCTGACGGACCTGGAGCAGCGCAGCGGGGACATCACCTACGGGCAGTTTGCTCAGCTGTACCGCAGCCTC
plcg1.cds   551 TCAACTACCGAGTCCCCAATATGCGCTTCCTCCGAGAGCGGCTGACGGACCTTGAACAGCGCAGCGGGGACATCACCTACGGGCAGTTTGCTCAGCTGTACCGCAGCCTC

PLCG1.cds   661 ATGTACAGCGCCCAGAAGACGATGGACCTCCCCTTCTTGGAAGCCAGTACTCTGAGGGCTGGGAGCGGCCGGAGCTTTGCCGAGTGTCCCTTCCTGAGTTCCAGCAGTT
plcg1.cds   661 ATGTACAGCGCCCAGAAGACGATGGACCTTCCCTTCTTGGAAACCAACGCTTTGAGGACTGGAGAGCGTCCGGAGCATTGCCAGGTGTCCCTTTCTGAGTTCCAGCAGTT

PLCG1.cds   771 CCTTCTTGACTACCAGGGGGAGCTGTGGGCTGTTGATCGCCTCCAGGTGCAGGAGTTCATGCTCAGCTTCCTCCGAGACCCCTTACGAGAGATCGAGGAGCCATACTTCT
plcg1.cds   771 CCTCCTTGAATACCAGGGGGAGCTGTGGGCTGTCGACCGGCTCCAGGTGCAGGAATTCATGCTCAGCTTCCTTCGAGACCCCTTGCGAGAGATTGAGGAGCCATACTTCT

PLCG1.cds   881 TCCTGGATGAGTTTGTCACCTTCCTGTTCTCCAAAGAGAACAGTGTGTGGAACTCGCAGCTGGATGCAGTATGCCCGGACACCATGAACAACCCTCTTTCCCACTACTGG
plcg1.cds   881 TCTTGGATGAGCTTGTCACCTTCCTGTTCTCCAAAGAGAACAGCGTGTGGAACTCACAGCTGGATGCTGTGTGCCCGGACACCATGAACAACCCACTCTCTCACTATTGG

PLCG1.cds   991 ATCTCCTCCTCGCACAACACGTACCTGACCGGGGACCAGTTCTCCAGTGAGTCCTCCTTGGAAGCCTATGCTCGCTGCCTGCGGATGGGCTGTCGCTGCATTGAGTTGGA
plcg1.cds   991 ATCTCCTCCTCACATAATACGTACCTGACTGGGGACCAGTTCTCCAGTGAGTCCTCCCTGGAAGCCTATGCTCGCTGCCTGAGGATGGGCTGTCGCTGCATTGAGTTGGA

PLCG1.cds  1101 CTGCTGGGACGGCCCGGATGGGATGCCAGTTATTTACCATGGGCACACCCTTACCACCAAGATCAAGTTCTCAGATGTCCTGCACACCATCAAGGAGCATGCCTTTGTGG
plcg1.cds  1101 CTGCTGGGATGGCCCAGATGGGATGCCAGTCATTTACCATGGGCACACTCTCACCACCAAGATCAAGTTCTCTGACGTCCTGCACACCATCAAGGAGCATGCTTTCGTAG

PLCG1.cds  1211 CCTCAGAGTACCCAGTCATCCTGTCCATTGAGGACCACTGCAGCATTGCCCAGCAGAGAAACATGGCCCAATACTTCAAGAAGGTGCTGGGGCACACACTCCTCACCAAG
plcg1.cds  1211 CCTCAGAGTACCCAGTCATCCTGTCCATCGAGGACCACTGCAGCATTGCCCAGCAGAGAAACATGGCTCAGCACTTCAGAAAGGTGCTTGGCGACACACTCCTTACCAAG

PLCG1.cds  1321 CCCGTGGAGATCTCTGCCGACGGGCTCCCCTCACCCAACCAGCTTAAGAGGAAGATCCTCATCAAGCACAAGAAGCTGGCTGAGGGCAGTGCCTACGAGGAGGTGCCTAC
plcg1.cds  1321 CCTGTGGACATTGCTGCTGATGGCCTTCCTTCTCCCAACCAGCTCAGGAGGAAGATCCTTATTAAGCACAAGAAGCTGGCTGAGGGCAGTGCCTATGAGGAGGTGCCTAC

PLCG1.cds  1431 ATCCATGATGTACTCTGAGAACGACATCAGCAACTCTATCAAGAATGGCATCCTCTACCTGGAGGACCCTGTGAACCACGAATGGTATCCCCACTACTTTGTTCTGACCA
plcg1.cds  1431 CTCTGTGATGTACTCCGAGAATGACATCAGTAACTCCATCAAGAATGGTATCCTCTACTTGGAGGACCCCGTGAATCATGAGTGGTATCCCCACTACTTCGTTCTGACTA

PLCG1.cds  1541 GCAGCAAGATCTACTACTCTGAGGAGACCAGCAGTGACCAGGGCAACGAGGATGAGGAGGAGCCCAAGGAGGTCAGCAGCAGCACAGAGCTGCACTCCAATGAGAAGTGG
plcg1.cds  1541 GCAGCAAGATCTACTACTCTGAGGAGACCAGCAGTGACCAGGGCAATGAGGATGAAGAGGAGCCGAAGGAGGCCAGCAGCAGCACAGAGCTGCACTCGAGTGAGAAGTGG

PLCG1.cds  1651 TTCCATGGGAAGCTAGGGGCAGGCGTGACGGGCGTCACATCGCTGAGCGCCTGCTTACTGAGTACTGCATCGAGACCGGAGCCCTGACGGCTCCTTCCTCGTGCGAGA
plcg1.cds  1651 TTCCACGGGAAGCTCGGGGCTGGACGTGACGGGCGGCACATTGCTGAGCGCCTGCTCACTGAGTACTGCATAGAGACTGGGGCTCCTGATGGCTCCTTCCTAGTGCGAGA

PLCG1.cds  1761 GAGTGAGACCTTCGTGGGCGACTACACGCTCTCTTTCTGGCGGAACGGGAAAGTCCAGCACTGCCGTATCCACTCCCGGCAAGATGCTGGGACCCCCAAGTTCTTCTTGA
plcg1.cds  1761 AAGTGAGACCTTCGTGGGTGACTACACGCTGTCTTTCTGGCGGAATGGGAAAGTCCAGCACTGCCGCATCCACTCCCGGCAGGATGCTGGGACTCCTAAGTTCTTCTTGA

PLCG1.cds  1871 CAGACAACCTCGTCTTTGACTCCTCTATGACCTCATCACGCACTACCAGCAGGTGCCCCTGCGCTGTAATGAGTTTGAGATGCGACTTTCAGAGCCTGTCCCACAGACC
plcg1.cds  1871 CAGATAACCTTGTCTTTGACTCTCTCTATGACCTCATCACACATTATCAGCAAGTACCCCTGCGCTGCAATGAGTTTGAGATGCGCCTTTCAGAGCCTGTTCCACAGACG

PLCG1.cds  1981 AACGCCCACGAGAGCAAAGAGTGGTACCACGCGAGCCTGACCAGAGCACAGGCTGAGCACATGCTAATGCGCGTCCCTCGTGATGGGGCCTTCCTGGTGCGGAAGCGGAA
plcg1.cds  1981 AATGCCCATGAGAGCAAAGAGTGGTACCACGCAAGCCTGACTAGAGCTCAGGCTGAACATATGCTGATGCGAGTGCCCCGGGATGGGGCCTTCCTGGTGCGGAAACGCAA

PLCG1.cds  2091 TGAACCCAACTCATATGCCATCTCTTTCCGGGCTGAGGGCAAGATCAAGCATTGCCGTGTCCAGCAAGAGGGCCAGACAGTGATGCTAGGGAACTCGGAGTTCGACAGCC
plcg1.cds  2091 TGAGCCTAACTCATATGCCATCTCTTTCCGGGCTGAGGGAAAGATCAAGCACTGCCGAGTACAGCAGGAAGGCCAGACAGTGATGCTGGGGAACTCTGAGTTTGACAGCC

PLCG1.cds  2201 TTGTTGACCTCATCAGCTACTATGAGAAACACCCGCTATACCGCAAGATGAAGCTGCGCTATCCCATCAACGAGGAGGCACTGGAGAAGATTGGCACAGCTGAGCCTGAC
plcg1.cds  2201 TGGTTGACCTCATCAGCTACTATGAGAAGCACCCCCTGTACCGCAAAATGAAGCTACGCTACCCCATCAACGAGGAGGCACTGGAGAAGATCGGGACAGCTGAACCCGAT

PLCG1.cds  2311 TACGGGGCCCTGTATGAGGGACGCAACCCTGGCTTCTATGTAGAGGCAAACCCTATGCCAACTTTCAAGTGTGCAGTCAAAGCCCTCTTTGACTACAAGGCCCAGAGGGA
plcg1.cds  2311 TATGGGGCACTATACGAGGGCCGCAACCCTGGTTTCTATGTGGAGGCAAACCCTATGCCAACTTTCAAGTGTGCAGTAAAAGCCCTCTTCGACTACAAGGCCCAGAGAGA

PLCG1.cds  2421 GGACGAGCTGACCTTCATCAAGAGCGCCATCATCCAGAATGTGGAGAAGCAAGAGGGAGGCTGGTGGCGAGGGGACTACGGAGGGAAGAAGCAGCTGTGGTTCCCATCAA
plcg1.cds  2421 GGATGAGCTGACCTTCACCAAGAGTGCCATCATCCAGAATGTGGAAAAGCAAGATGGTGGCTGGTGGCGAGGGGACTATGGTGGGAAGAAGCAGCTGTGGTTCCCCTCAA

PLCG1.cds  2531 ACTACGTGGAAGAGATGGTCAACCCCGTGGCCCTGGAGCCGGAGAGGGAGCACTTGGACGAGAACAGCCCCCTAGGGGACTTGCTGCGCGGGGTCTTGGATGTGCCGCT
plcg1.cds  2531 ACTATGTGGAAGAGATGATCAATCCAGCAGTCCTAGAGCCTGAGAGGGAGCACCTGGATGAGAACAGCCCACTGGGGGACTTGCTGCGAGGGGTCTTAGATGTGCCAGCT

PLCG1.cds  2641 TGTCAGATTGCCATCCGTCCTGAGGGCAAGAACAACCGGCTCTTCGTCTTCTCCATCAGCATGGCGTCGTGGTGGCCCACTGGTCCCTGGATGTTGCTGCCGACTCACAGGA
plcg1.cds  2641 TGTCAGATCGCCATCCGTCCTGAGGGCAAAAACAACCGGCTCTTCGTCTTCTCCATCAGCATGCCATCAGTGGCTCAGTGGTCCCTGGATGTTGCAGCTGACTCACAGGA

PLCG1.cds  2751 GGAGCTGCAGGACTGGGTGAAAAAGATCCGTGAAGTGGCCCAGACAGCAGACGCCAGGCTCACTGAAGGGAAGATAATGGAACGGAGGAAGAAGATTGCCCTGGAGCTCT
plcg1.cds  2751 GGAGTTACAGGACTGGGTGAAAAAGATCCGTGAAGTTGCCCAGACTGCAGATGCCAGGCTCACTGAGGGAAAGATGATGGAGAGGAGGAAGAAGATCGCCTTGGAGCTCT

PLCG1.cds  2861 CTGAACTTGTCGTCTACTGCCGGCCTGTTCCCTTTGATGAAGAGAAGATTGGCACAGAACGTGCTTGCTACCGGGACATGTCATCCTTCCCGGAAACCAAGGCTGAGAAA
plcg1.cds  2861 CCGAGCTTGTGGTCTACTGCCGGCCCGTTCCCTTTGATGAAGAGAAGATTGGCACAGAACGTGCTTGTTACCGGGACATGTCCTCCTTTCCGGAAACCAAGGCTGAGAAG

PLCG1.cds  2971 TACGTGAACAAGGCCAAAGGCAAGAAGTTCCTTCAGTACAATCGACTGCAGCTCTCCCGCATCTACCCCAAGGGCCAGCGACTGGATTCCTCCAACTACGATCCTTTGCC
plcg1.cds  2971 TATGTGAACAAGGCCAAAGGCAAGAAGTTCCTCCAGTACAACCGGCTGCAGCTCTCGCGCATCTACCCTAAGGGCCAGAGGCTAGACTCCTCCAATTATGACCCTCTGCC

PLCG1.cds  3081 CATGTGGATCTGTGGCAGTCAGCTTGTGGCCCTCAACTTCCAGACCCCTGACAAGCCTATGCAGATGAACCAGGCCCTCTTCATGACGGGCAGGCACTGTGGCTACGTGC
plcg1.cds  3081 CATGTGGATCTGCGGTAGCCAGCTTGTAGCACTCAATTTCCAGACCCCAGACAAGCCTATGCAGATGAACCAGGCCCTCTTCATGGCTGGTGGGCATTGTGGCTATGTGC

PLCG1.cds  3191 TGCAGCCAAGCACCATGCGGGATGAGGCCTTCGACCCCTTTGACAAGAGCAGCCTCCGCGGGCTGGAGCCATGTGCCATCTCTATTGAGGTGCTGGGGGCCCGACATCTG
plcg1.cds  3191 TGCAGCCAAGCACCATGAGAGACGAAGCCTTTGACCCCTTTGATAAGAGCAGTCTCCGAGGTCTGGAACCCTGTGTCATTTGCATTGAGGTGCTGGGGGCCAGGCATCTG

PLCG1.cds  3301 CCAAAGAATGGCCGAGGCATTGTGTGTCCTTTTGTGGAGATTGAGGTGGCTGGAGCTGAGTATGACAGCACCAAGCAGAAGACAGAGTTTGTGGTGGACAATGGACTCAA
plcg1.cds  3301 CCGAAGAATGGCCGGGGTATTGTGTGTCCTTTTGTGGAGATTGAGGTGGCTGGGCTGAGTACGACAGCACCAAGCAAAAGACGGAGTTTGTAGTGGACAACGGACTGAA

PLCG1.cds  3411 CCCTGTATGGCCAGCCAAGCCCTTCCACTTCCAGATCAGTAACCCTGAATTTGCCTTTCTGCGCTTCGTGGTGTATGAGGAAGACATGTTTAGTGACCAGAATTTCCTGG
plcg1.cds  3411 CCCTGTGTGGCCTGCTAAGCCCTTCCACTTCCAGATCAGTAACCCAGAGTTTGCCTTTCTGCGCTTTGTGGTGTATGAGGAAGACATGTTTAGTGACCAGAACTTCTTGG

PLCG1.cds  3521 CTCAGGCTACTTTCCCAGTAAAAGGCCTGAAGACAGGATACAGAGCAGTGCCTTTGAAGAACAACTACAGTGAGGACCTGGAGTTGGCCTCCCTGCTGATCAAGATTGAC
plcg1.cds  3521 CTCAGGCTACTTTCCCAGTAAAAGGCCTGAAGACAGGATATAGAGCAGTGCCTTTGAAGAACAACTACAGTGAAGACCTGGAGTTGGCCTCCCTGCTCATCAAGATTGAC

PLCG1.cds  3631 ATTTTCCCTGCCAAGGAGAATGGTGACCTCAGTCCCTTCAGTGGTACGTCCCTGCGGGAGCGGGGCTCAGATGCCTCAGGCCAGCTGTTTCATGGCCGAGCCCGGGAAGG
plcg1.cds  3631 ATTTTCCCTGCTAAGGAGAACGGTGACCTCAGTCCTTTCAGTGGCATATCCCTAAGGGAACGGGCCTCAGATGCCTCCAGCCAGCTGTTCCATGTCCGGCCCGGGAAGG

PLCG1.cds  3741 CTCCTTTGAATCCCGCTACCAGCAGCCGTTTGAGGACTTCCGCATCTCCCAGGAGCATCTCGCAGACCATTTTGACAGTCGAGAACGAAGGGCCCCAAGAAGGACTCGGG
plcg1.cds  3741 GTCCTTTGAAGCCAGATACCAGCAGCCATTTGAAGATTTCCGCATCTCGCAGGAGCATCTAGCAGACCATTTTGACAGTCGGGAACGAAGGGCCCCAAGAAGGACTCGGG

PLCG1.cds  3851 TCAATGGAGACAACCGCCTCTAG 3873
plcg1.cds  3851 TCAATGGAGACAACCGCCTCTAG 3873
```

B. TIX1:bM393F23.2 (CDS)

```
TIX1.cds           1 ATGGCCAGCAAGAGGAAATCCACCACACCATGCATGATCCCAGTGAAGACTGTGGTGTTGCAAGATGCCAGCATGGAGGCCCAGCCCGCTGAGACCTTGCCTGAAGGACC
bM393F23.2.cds     1 ATGGCCAGCAAAAGGAAGTCCACCACCCCATGCATGATCCCCGTTAAGACCGTGGTGCTGCCGGGTGCCAGCACAGAGCCTCAGCCTGTGGAGTCTCTGCCTGAAGGTCC

TIX1.cds         111 CCAGCAGGATCTGCCCCCAGAAGCATCTGCTGCCAGCAGTGAGGCAGCACAGAACCCCAGCAGTACTGATGGCTCTACACTGGCCAATGGGCATCGGAGCACTTTAGATG
bM393F23.2.cds   111 CCAGCAGGATCTTCCCTCAGAGGCACCTGATGCGAGCAGCGAGGCAGCCCCAAACCCCAGCAGCACTGATGGCTCTGCCCTGGCAAATGGGCACCGGAGCACTTTGGATG

TIX1.cds         221 GCTATTTATATTCCTGTAAATACTGCGATTTCAGATCCCATGACATGACCCAATTTGTGGGACATATGAACTCAGAGCACACAGACTTTAATAAAGACCCAACCTTTGTA
bM393F23.2.cds   221 GCTATGTATATTGCTGTAAAGAGTGTGAGTTCAGGTCCCAGGATGTGACCCACTTTATAGGACATATGAACTCAGAGCACACAGACTTTAATAAAGACCCAACTTTTGTA

TIX1.cds         331 TGCAGTGGGTGCAGTTTTCTGGCAAAAACCCCTGAGGGCTTTCCTTGCACAATGCCACATGTCACTCCGGGGAAGCCAGCTTTGTGTGGAACGTGGCCAAGCCAGACAA
bM393F23.2.cds   331 TGTACAGGGTGCAGTTTCCTGGCAAAAACCCCTGAAGGACTTTCCCTGCATAATGCCAAGTGTCACTCAGGGGAAGCCAGCTTTTTGTGGAATGTGACCAAGCCAGACAA

TIX1.cds         441 TCATGTGGTTGTGGAGCAGAGCATCCCTGAGAGCACCAGCACTCCTGACCTAGCGGGTGAGCCCAGTGCTGAAGGGGCTGATGGACAGGCAGAAATCATCATTACCAAAA
bM393F23.2.cds   441 TCATGTAGTAGTGGAGCAGAGTGTCCCTGACAGTGCTAGCTCATCTGTCCTAGCAGGTGAGTCTACTACGGAAGGGACTGA.............AATCATCATTACCAAGA

TIX1.cds         551 CTCCAATCATGAAGATAATGAAAGGCAAAGCTGAAGCCAAAAAAAATTCATACACTCAAGGAGAATGTCCCTAGCCAGCCTGTGGGTGAGGCCTTACCAAAGCTGTCGACT
bM393F23.2.cds   539 CTCCAATCATGAAGATAATGAAAGGCAAAGCTGAAGCCAAAAAAAATCCATATGCTTAAAGAGAACGCTCCTAATCAGCCTGGGAGTGAGGCCTTGCCTAAGCCATTGGCA

TIX1.cds         661 GGAGAAATGGAGGTGAGAGAGGGGGACCATTCCTTCATCAATGGGGCAGTTCCAGTCAGCCAGGCATCTGCCAGCTCTGCAAAAAACCCCCATGCCGCCAACGGGCCCCT
bM393F23.2.cds   649 GGGGAGAGAGAGGTTAAAGAGGGGGACCACACTTTCATCAATGGGGCAGCTCCAGGCAGTCAGGCATCTGCTAAGTCCACAAAACCCCCTCCTGCTGCCAATGGGCCCCT

TIX1.cds         771 GATAGGAACAGTGCCAGTTTTGCCAGCTGGCATAGCACAGTTCCTCTCCCTCCAGCAGCAGCCCCCAGTGCATGCCCAACACCATGTCCACCAGCCACTGCCCACGGCCA
bM393F23.2.cds   759 GATAGGAACAGTGCCTGTCTTGCCAGCTGGTATAGCACAGTTCCTCTCCCTCCAGCAGCAGCCCCCAGTACATGCCCAACACCACACCCACCAACCCCTGCCTACATCCA

TIX1.cds         881 AGGCCCTTCCCAAAGTGATGATCCCCCTGAGCAGCATTCCAACGTACAATGCAGCCATGGACTCTAACAGCTTCCTGAAGAACTCCTTCCACAAGTTCCCCTACCCCACC
bM393F23.2.cds   869 AGACCCTTCCGAAAGTCATGATCCCCTTGAGCAGCATCCCAACATATAATGCAGCTATGGACTCCAACAGCTTTCTGAAGAACTCCTTCCACAAATTCCCCTACCCAACC

TIX1.cds         991 AAAGCCGAGCTCTGCTATTTGACTGTGGTGACCAAGTATCCAGAAGAACAGCTCAAGATCTGGTTCACAGCCCAAAGGCTGAAGCAGGGATCAGCTGGTCCCCTGAGGA
bM393F23.2.cds   979 AAAGCTGAGCTCTGCTATTTGACTGTGGTGACCAAGTATCCAGAAGAACAGCTTAAAATCTGGTTCACAGCCCAGAGGCTGAAGCAAGGAATCAGCTGGTCTCCTGAGGA

TIX1.cds        1101 GATTGAGGATGCCCGGAAAAAGATGTTCAATACAGTCATCCAGTCTGTGCCTCAGCCCACAATTACGGTTCTAAATACCCCACTCGTCGCCAGTGCTGGCAATGTCCAGC
bM393F23.2.cds  1089 GATTGAAGATGCCAGGAAAAAGATGTTCAATACAGTCATCCAGTCTGTGCCTCAGCCCACAATCACAGTTCTAAACACCCCGCTTGTTGCCAGTGCTGGCAATGTACAGC

TIX1.cds        1211 ATCTCATCCAGGCCGCTCTTCCAGGTCACGTTGTGGGGCAGCCAGAGGGTACAGGAGGGGGACTTCTGGTCACTCAGCCATTGATGGCCAATGGGTTGCAAGCAACAAGT
bM393F23.2.cds  1199 ACCTCATCCAGGCTACTCTCCCAGGTCATGCTGTGGGACAGCCTGAGGGCACAGCAGGGGGGCTGTTGGTCACTCAACCACTGATGGCCAATGGGTTACAAGCATCAAGC

TIX1.cds        1321 TCCCCTCTCCCCTCACGGTGACATCCGTCCCCAAGCAGCCAGGTGTGGCACCCATTAACACTGTGTGTTCAAATACAACGTCAGCTGTGAAGGTGGTCAATGCGGCCCA
bM393F23.2.cds  1309 TCATCTCTCCCTCTGACAACTGCATCTGTTCCCAAGC...CAACTGTGGCACCCATTAACACTGTGTGTTCAAATTCGGCATCAGCTGTGAAAGTTGTCAATGCAGCCCA

TIX1.cds        1431 GTCGCTCCTCACGGCCTGCCCCAGCATAACCTCCCAAGCCTTCCTTGATGCTAGCATCTACAAAAATAAGAAATCTCATGAACAGCTGTCAGCTCTGAAAGGGAGCTTCT
bM393F23.2.cds  1416 GTCACTCCTCACAGCATGCCCCAGCATAACTTCCCAAGCTTTCCTTGATGCAAACATCTACAAAAATAAGAAGTCTCATGAACAGCTCTCGGCTCTGAAAGGAAGCTTCT

TIX1.cds        1541 GTCGGAACCAGTTCCCAGGGCAGAGCGAAGTTGAACATCTCACAAAGTGACGGGCCTCAGTACCAGAGAGGTGCGGAAATGGTTCAGTGATCGTAGATACCACTGCCGG
bM393F23.2.cds  1526 GTCGGAACCAGTTTCCTGGGCAGAGTGAAGTAGAGCATCTGACCAAAGTGACAGGCCTCAGTACCAGAGAGGTGCGGAAGTGGTTCAGTGACCGGAGATATCACTGCCGG

TIX1.cds        1651 AACTTGAAGGGCTCCAGAGCGATGATACCTGGAGATCACAGTTCCATCATCATTGACTCTGTGCCAGAGGTGTCCTTCTCCCATCGTCCAAGGTCCCTGAGGTAACCTG
bM393F23.2.cds  1636 AATCTGAAGGGCTCCAGGGCCATGATGCCCGGAGAGCATGGCTCTGTTCTCATTGACTCTGTGCCAGAGGTGCCCTTTCCCCTGGCATCTAAAGTTCCAGAGGTGACCTG

TIX1.cds        1761 CATTCCGACAACAGCCACACTAGCAACCCACCCTTCTGCCAAACGACAATCTTGGCACCAGACTCCTGACTTCACACCAACCAAATACAAGGAGAGAGCCCCTGAGCAGC
bM393F23.2.cds  1746 TATCCCAACAGCAACCTCACTAGTAAGCCACCCTGCCACCAAACGACAATCTTGGCACCAGACCCCAGACTTCACACCAACCAAATACAAAGAAGAGCCCCAGAGCAGC

TIX1.cds        1871 TCAGAGCCCTGGAGAGCAGTTTTGCACAAAACCCTCTTCCTGTTGATGAGGAACTGGACCGCCTGAGAAGTGAAACCAAAATGACCGACGAGAAATTGATAGCTGGTTT
bM393F23.2.cds  1856 TGCGAGTCCTAGAGAACAGCTTTGCACAAAACCCACTTCCCCCCGAGGAGGAGCTGGACCGCCTAAGAAGTGAAACGAAAATGACTCGAAGGGAAATTGATGGCTGGTTC

TIX1.cds        1981 TCAGAGAGACGGAAAAAAGTGAATGCTGAGGAGACCAAGAAGGCTGAGGAGAATGCCTCTCAGGAGGAAGAGGAGGCTGCTGAGGATGAGGGTGGAGAAGAGGATTTGGC
bM393F23.2.cds  1966 TCAGAGAGGCGGAAAAAAGGTGAATACTGAGGAGACAAAGAAAGCTGATGGGCACATGCCTAAGGAGGAGGAGGAGGTGCTGAGCAGGAGGGTAGAGATGAAGAGTTGGC

TIX1.cds        2091 CAGTGAGCTAAGGGTCTCTGGTGAAAATGGCTCTCTGGAAATGCCCAGCAGCCATATCTTGGCAGAGCGCAAAGTCAGCCCCATTAAAATCAACCTGAAGAACCTGAGGG
bM393F23.2.cds  2076 CAATGAGCTGAGGGTCCCTGGTGAAAATGGCTCTCCTGAAATGTTTCTTAGCCATGCCTTGGCAGAACGGAAGGTCAGCCCCATCAAAATCAACCTCAAGAACCTGCGGG

TIX1.cds        2201 TCACTGAAGCCAATGGCAGGAACGAGATTCCAGGGCTGGGTGCCTGTGACCCTGAGGATGATGAGTCAAACAAACTGGCAGAGCAGCTCCCAGGCAAAGTGAGCTGCAAA
bM393F23.2.cds  2186 TCACTGAAGCCAGTGGCAAGAGTGAGTTTCCAGGGATGGGTGTCTGTGAGCCTGAGGAAGATGGTTTGAACAAGTTGGTAGAACAGCCACCAAGCAAAGTAAGCTACAAA

TIX1.cds        2311 AAGACTGCCCAGCAGCGGCACTTGCTGCGGCAGCTCTTTGTCCAGACACAGTGGCCAAGCAACCAGGACTATGACTCCATCATGGCCCAGACGGGTCTGCCACGGCCAGA
bM393F23.2.cds  2296 AAGACCGCACAGCAGCGTCATTTGCTTCGGCAGCTCTTTGTCCAAACACAGTGGCCAAGCAACCAGGACTATGACTCCATTATGGCCCAAACAGGGCTGCCCCGGCCAGA

TIX1.cds        2421 GGTGGTGCGCTGGTTTGGAGATAGCAGGTACGCACTGAAGAACGGCCAACTCAAATGGTACGAAGACTATAAGCGAGGCAACTTCCCACCAGGGCTACTGGTCATTGCCC
bM393F23.2.cds  2406 GGTGGTACGCTGGTTTGGAGATAGCAGGTATGCCCTTAAGAATGGCCAGCTCAAGTGGTACGAAGACTATAAGAGGGGAAACTTCCCACCAGGTTTGCTGGTCATAGCCC

TIX1.cds        2531 CTGGCAACCGGGAGCTCCTGCAGGACTATTACATGACACACAAGATGCTGTATGAAGAGGACCTGCAGAACCTCTGTGACAAGACCCAGATGAGCTCCCAGCAGGTCAAG
bM393F23.2.cds  2516 CTGGCAACCGAGAGCTGCTGCAAGATTATTACATGACACACAAGATGTTGTGTGAGGAAGATCTGCAGACCCTCTGTGATAAGACCCAGATGAGTGCCCAGCAGGTCAAG

TIX1.cds        2641 CAGTGGTTTGCTGAGAAAATGGGGGAGGAGACCAGAGCCGTGGCAGACACAGGCAGTGAGGACCAGGGCCCTGGTACTGGTGAGCTCACAGCAGTTCACAAAGGATGGG
bM393F23.2.cds  2626 CAGTGGTTTGCTGAGAAGATGGGTGAAGAGACCCGGGCTGTGGCAGATATAAGCAGTGAGGATCAGGGCCCTAGAAATGGAGAGCCTGTGGCAGTTCACAAGGTGCTGGG

TIX1.cds        2751 TGACACCTATTCAGAGGTGTCTGAGAACAGTGAGTCGTGGGAGCCCCGTGTCCCTGAGGCCAGCTCAGAGCCCTTTGACACATCGAGTCCCCAGGCTGGACGTCAGCTCG
bM393F23.2.cds  2736 TGATGCCTATTCGGAGCTGTCTGAGAACAGCGAATCATGGGAGCCAAGTGCTCCTGAGGCCAGCTCAGAGCCTTTTGACACTTCAAGTCCCCAGTCTGGACGTCAACTCG

TIX1.cds        2861 AAACAGACTGA 2871
bM393F23.2.cds  2846 AAGCAGACTGA 2856
```

C. C20orf111:bM117O11.3 (CDS)

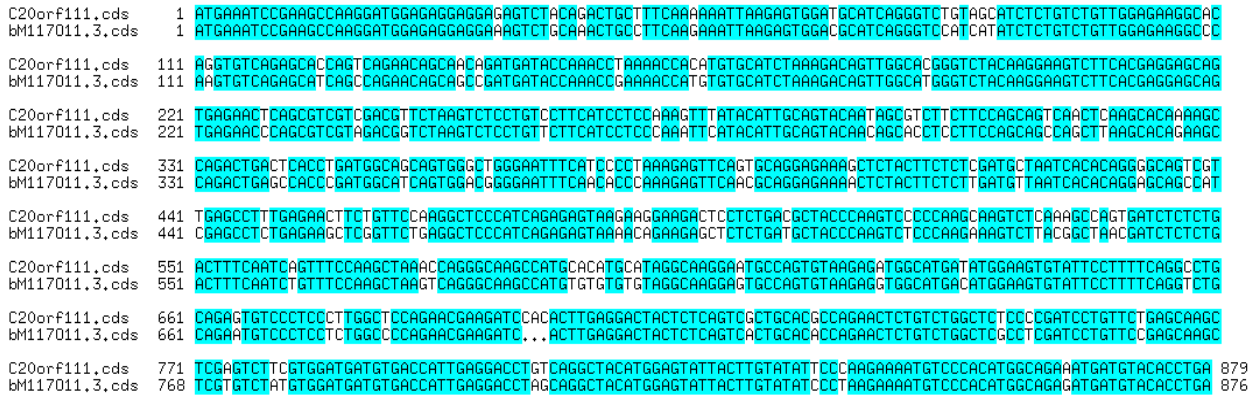


Figure 4.16: Coding sequence alignments (A-C). DNA sequences were aligned using CLUSTAL W (Thompson *et al.*, 1994) and the output was formatted using Belvu (Sonnhammer, unpublished). Identical base pairs are highlighted blue.

A. PLCG1:Plcg1 (protein)

B. TIX1:bM393F23.2 (protein)

```
TIX1.pep          1 MASKRKSTTPCMIPVKTVVLQDASMEAQPAETLPEGPQQDLPPEASAASSEAAQNPSSTDGSTLANGHRSTLDGYLYSCK
bM393F23.2.pep    1 MASKRKSTTPCMIPVKTVVLPGASTEPQPVESLPEGPQQDLPSEAPDASSEAAPNPSSTDGSALANGHRSTLDGYVYCCK

TIX1.pep         81 YCDFRSHDMTQFVGHMNSEHTDFNKDPTFVCSGCSFLAKTPEGLSLHNATCHSGEASFVWNVAKPDNHVVVEQSIPESTS
bM393F23.2.pep   81 ECEFRSQDVTHFIGHMNSEHTDFNKDPTFVCTGCSFLAKNPEGLSLHNAKCHSGEASFLWNVTKPDNHVVVEQSVPDSAS

TIX1.pep        161 TPDLAGEPSAEGADGQAEIIITKTPIMKIMKGKAEAKKIHTLKENVPSQPVGEALPKLSTGEMEVREGDHSFINGAVPVS
bM393F23.2.pep  161 SSVLAGESTTEG....TEIIITKTPIMKIMKGKAEAKKIHMLKENAPNQPGSEALPKPLAGEREVKEGDHTFINGAAPGS

TIX1.pep        241 QASASSAKNPHAANGPLIGTVPVLPAGIAQFLSLQQQPPVHAQHHVHQPLPTAKALPKVMIPLSSIPTYNAAMDSNSFLK
bM393F23.2.pep  237 QASAKSTKPPPAANGPLIGTVPVLPAGIAQFLSLQQQPPVHAQHHTHQPLPTSKTLPKVMIPLSSIPTYNAAMDSNSFLK

TIX1.pep        321 NSFHKFPYPTKAELCYLTVVTKYPEEQLKIWFTAQRLKQGISWSPEEIEDARKKMFNTVIQSVPQPTITVLNTPLVASAG
bM393F23.2.pep  317 NSFHKFPYPTKAELCYLTVVTKYPEEQLKIWFTAQRLKQGISWSPEEIEDARKKMFNTVIQSVPQPTITVLNTPLVASAG

TIX1.pep        401 NVQHLIQAALPGHVVGQPEGTGGGLLVTQPLMANGLQATSSPLPLTVTSVPKQPGVAPINTVCSNTTSAVKVVNAAQSLL
bM393F23.2.pep  397 NVQHLIQATLPGHAVGQPEGTAGGLLVTQPLMANGLQASSSSLPLTTASVPK.PTVAPINTVCSNSASAVKVVNAAQSLL

TIX1.pep        481 TACPSITSQAFLDASIYKNKKSHEQLSALKGSFCRNQFPGQSEVEHLTKVTGLSTREVRKWFSDRRYHCRNLKGSRAMIP
bM393F23.2.pep  476 TACPSITSQAFLDANIYKNKKSHEQLSALKGSFCRNQFPGQSEVEHLTKVTGLSTREVRKWFSDRRYHCRNLKGSRAMMP

TIX1.pep        561 GDHSSIIIDSVPEVSFSPSSKVPEVTCIPTTATLATHPSAKRQSWHQTPDFTPTKYKERAPEQLRALESSFAQNPLPLDE
bM393F23.2.pep  556 GEHGSVLIDSVPEVPFPLASKVPEVTCIPTATSLVSHPATKRQSWHQTPDFTPTKYKERAPEQLRVLENSFAQNPLPPEE

TIX1.pep        641 ELDRLRSETKMTRREIDSWFSERRKKVNAEETKKAEENASQEEEEAAEDEGGEEDLASELRVSGENGSLEMPSSHILAER
bM393F23.2.pep  636 ELDRLRSETKMTRREIDGWFSERRKKVNTEETKKADGHMPKEEEEGAEQEGRDEELANELRVPGENGSPEMFLSHALAER

TIX1.pep        721 KVSPIKINLKNLRVTEANGRNEIPGLGACDPEDDESNKLAEQLPGKVSCKKTAQQRHLLRQLFVQTQWPSNQDYDSIMAQ
bM393F23.2.pep  716 KVSPIKINLKNLRVTEASGKSEFPGMGVCEPEEDGLNKLVEQPPSKVSYKKTAQQRHLLRQLFVQTQWPSNQDYDSIMAQ

TIX1.pep        801 TGLPRPEVVRWFGDSRYALKNGQLKWYEDYKRGNFPPGLLVIAPGNRELLQDYYMTHKMLYEEDLQNLCDKTQMSSQQVK
bM393F23.2.pep  796 TGLPRPEVVRWFGDSRYALKNGQLKWYEDYKRGNFPPGLLVIAPGNRELLQDYYMTHKMLCEEDLQTLCDKTQMSAQQVK

TIX1.pep        881 QWFAEKMGEETRAVADTGSEDQGPGTGELTAVHKGMGDTYSEVSENSESWEPRVPEASSEPFDTSSPQAGRQLETD 956
bM393F23.2.pep  876 QWFAEKMGEETRAVADISSEDQGPRNGEPVAVHKVLGDAYSELSENSESWEPSAPEASSEPFDTSSPQSGRQLEAD 951
```

C. C20orf111:bM117O11.3 (protein)

```
C20orf111.pep     1 MKSEAKDGEEESLQTAFKKLRVDASGSVASLSVGEGTGVRAPVRTATDDTKPKTTCASKDSWHGSTRKSSRGAVRTQRRR
bM117O11.3.pep    1 MKSEAKDGEEESLQTAFKKLRVDASGSIISLSVGEGPSVRASARTAADDTKPKTMCASKDSWHGSTRKSSRGAVRTQRRR

C20orf111.pep    81 RSKSPVLHPPKFIHCSTIASSSSSQLKHKSQTDSPDGSSGLGISSPKEFSAGESSTSLDANHTGAVVEPLRTSVPRLPSE
bM117O11.3.pep   81 RSKSPVLHPPKFIHCSTTAPPSSSQLKHRSQTEPPDGISGRGISTPKEFNAGENSTSLDVNHTGAAIEPLRSSVLRLPSE

C20orf111.pep   161 SKKEDSSDATQVPQASLKASDLSDFQSVSKLNQGKPCTCIGKECQCKRWHDMEVYSFSGLQSVPPLAPERRSTLEDYSQS
bM117O11.3.pep  161 SKTEELSDATQVSQESLTANDLSDFQSVSKLSQGKPCVCVGKECQCKRWHDMEVYSFSGLQNVPPLAPERR.SLEDYSQS

C20orf111.pep   241 LHARTLSGSPRSCSEQARVFVDDVTIEDLSGYMEYYLYIPKKMSHMAEMMYT 292
bM117O11.3.pep  240 LHTRTLSGSPRSCSEQARVYVDDVTIEDLAGYMEYYLYIPKKMSHMAEMMYT 291
```

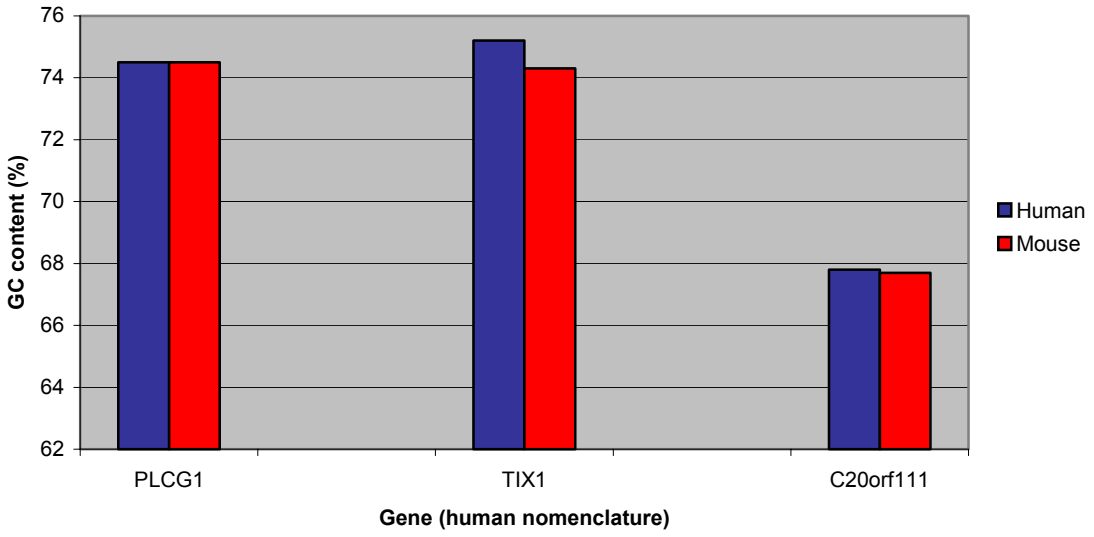
Figure 4.17: Protein sequence alignments (A-C). Sequences were aligned using CLUSTAL W (Thompson *et al.*, 1994) and the output was formatted using Belvu (Sonnhammer, unpublished). Identical aa are highlighted blue and similar, grey.

Table 4.6: Percentage identities of human and mouse sequences.

Orthologous gene pair	CDS sequence identity (%)	Amino acid sequence identity (%)
PLCG1:Plcg1	90.5	97
TIX1:bM393F23.2	84.9	85.8
C20orf111:bM117O11.3	88	86.6

The 5' UTRs of all three gene loci overlap with predicted CpG islands (CPGFIND; Micklem, unpublished). The orthologous CpG-island pairs have similar GC content; for PLCG1:plcg1, CpG islands differ considerably in size (Figure 4.18).

A.



B.

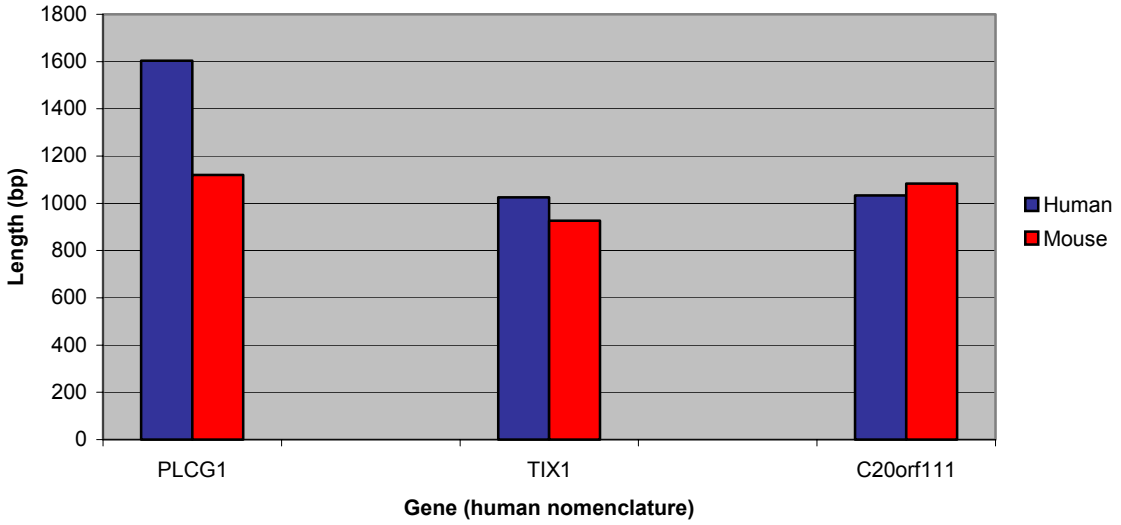
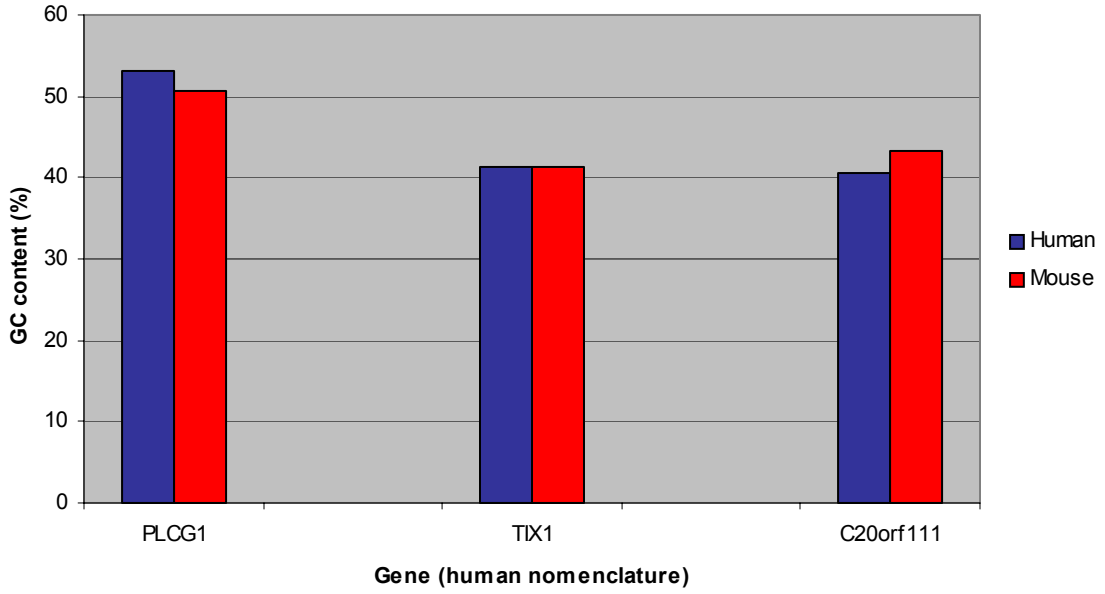


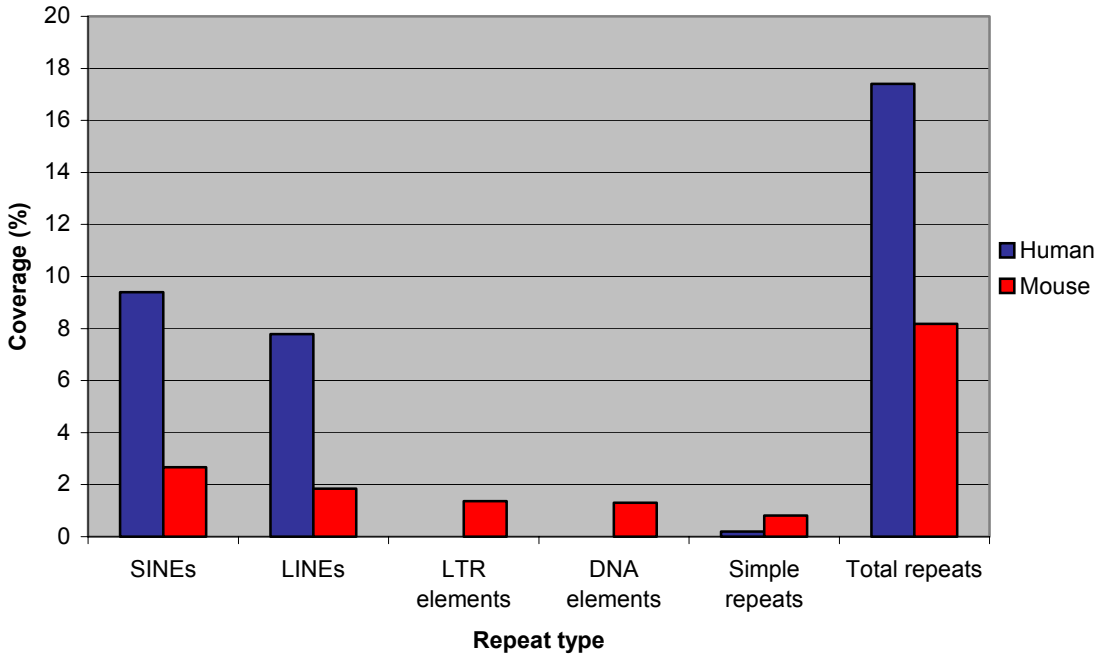
Figure 4.18: CpG island comparison. (A) GC content and (B) Length.

All orthologous gene pairs have similar GC content (exons and introns; Figure 4.19 A) and differ in repeat content, which is higher in the human genes (Figure 4.19 B-D). Interestingly, the LTR content is higher in the mouse for each of the orthologous gene pairs considered. Whole genome human:mouse repeat content analyses have shown that the age distribution of human and mouse transposons is strikingly different (IHGSC, 2001). Transposon activity in the mouse genome has not undergone the decline seen in humans and proceeds at a much higher rate. This phenomenon may be responsible for the LTR coverage differences in Figure 4.19 B-D, but the sequence sets studied are far too small to suggest a similar trend across the whole mouse genome. In fact, the human and mouse LTR coverage across the whole region is quite similar (Table 4.2).

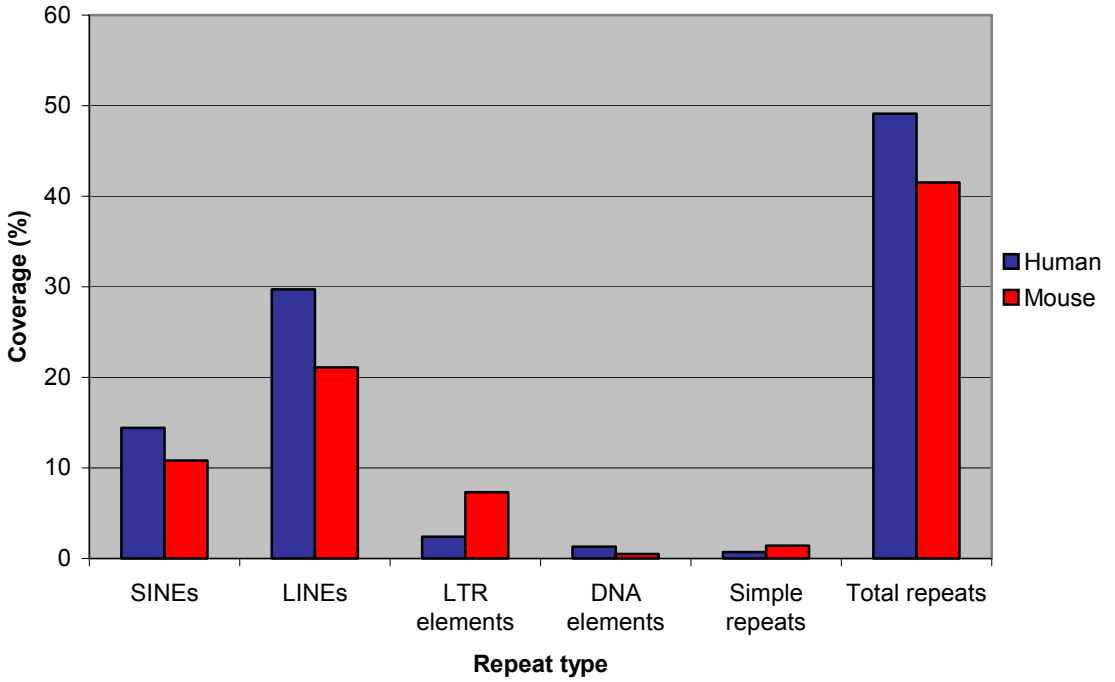
A.



B. PLCG1:Plcg1



C. TIX1:bM393F23.2



D. C20orf111:bM117O11.3

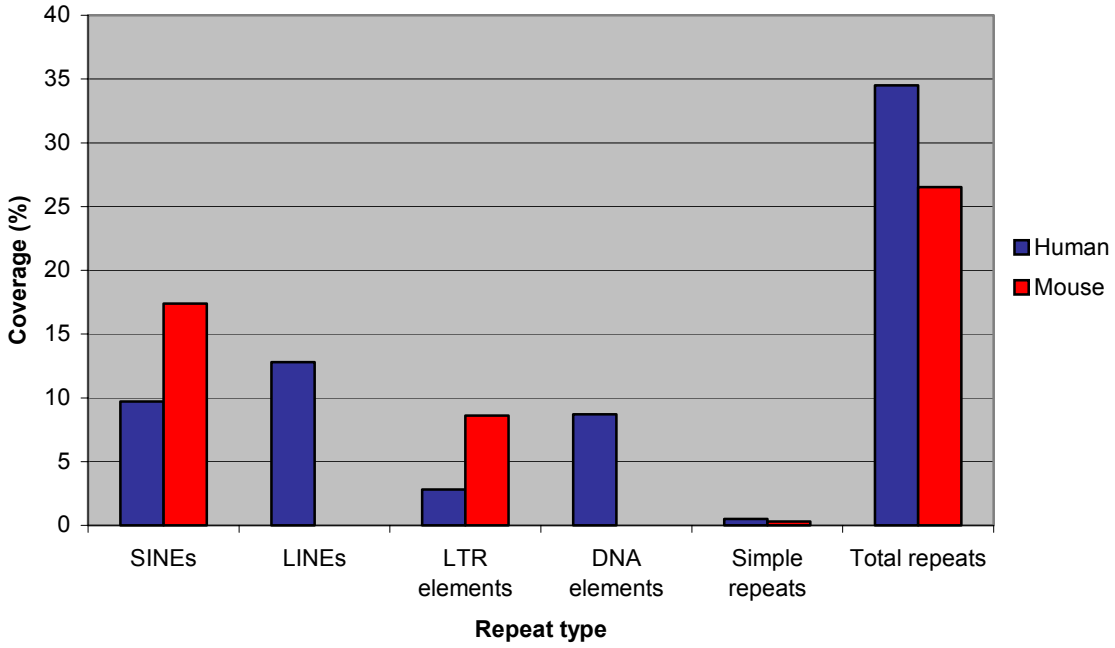


Figure 4.19 (A-D): (A) GC content. (B-D) Repeat content analysis.

4.7 Discussion

This chapter has described the construction, sequencing and comparative sequence analysis of approximately 10 Mb of the mouse genome, spanning a region of synteny with human chromosome 20q12-13.2. The final map consists of a single clone contig, with no gaps, that according to fingerprint estimates spans 9.8 Mb. It consists of 996 BAC clones, 66 gene based markers, 91 end based markers and 33 genetic markers. All data has been incorporated into the mouse chromosome 2 physical map produced by the MGSC (http://mouse.ensembl.org).

The annotation of human chromosome 20q12-13.2 was utilised during the initial selection of mouse landmark STSs. The use of mouse-expressed sequences (mRNA and ESTs) that shared extensive homology with annotated human genes provided an easy means to target the mapping efforts to the region of interest.

Landmark content mapping and fingerprinting data were used for the rapid assembly of eleven seed clone contigs. Unlike landmark content mapping, restriction enzyme fingerprinting allows analysis over the length of the clone and the construction of contigs relies on the number of bands shared between overlapping clones. Unfortunately fingerprinting does not allow the orientation of the contigs relative to each other, but incorporating landmark content data can alleviate this problem. This combined approach offered the best strategy for contig construction, accurately determining the overlap between clones.

Chromosome walks were performed to close the remaining gaps. With the availability of public data, this process was significantly accelerated. It is worth noting that the presence of gene deserts longer than 250 Kb long did not pose any problems, since all gaps were closed with chromosome walks. The choice of genomic library greatly aided this effort because of the large average insert size of the RPCI-23 BACs (197 Kb). Nevertheless, chromosome walks are time-consuming and extended gene deserts (>1 Mb) would have delayed the mapping process.

Comparisons between the generated mouse map and human genes confirmed the high degree of synteny between mouse chromosome 2 and human chromosome 20 (as proposed by earlier studies, Peters *et al*., 1999; Carver and Stubbs, 1997; DeBry and Seldin, 1996). Markers corresponding to 48% of annotated human coding genes were used in this mapping effort. The landmark content data from these markers suggests that there are no megabase-long rearrangements between these human and mouse regions.

A set of 66 overlapping BACs is being sequenced and as of May 2002, 10.3 Mb of mouse sequence has been generated (5,541,112 bp of finished and 4,778,718 bp of unfinished (redundant) sequence). According to the mouse genome assembly at Ensembl, the region is 8.76 Mb. This suggests that the mouse region is 10-15% smaller than the syntenic human region, which is in agreement with previous studies (Mural *et al*., 2002).

The GC content of the human and mouse sequences is similar but the repeat content differs (49.6% in humans and 32.1% in mice). It has been suggested that the observed difference in repeat content is probably due to the failure of the current algorithms to detect all repeats, rather than the presence of more additional unique sequence in mouse.

Mouse homologous sequences were identified for all annotated human coding genes and in some cases, supported new exons (e.g. C20orf130, TIX1). Gene order is completely conserved in human and mouse, but the presence of small local rearrangements cannot be excluded until finished mouse sequence is obtained across the whole region. Complete conservation of gene order has recently been reported for other large human and mouse syntenic regions. For example, Mural *et al*. (2002) identified two such regions, both residing on HSA3 and MMU16. Both regions are >10 Mb long and contain >100 genes each.

The high degree of conservation observed across the exons of coding genes is absent from pseudogenes. It may be that the non-functional sequences of pseudogenes have diverged more quickly in the mouse genome, possibly because of the much shorter generation time of the mouse. Alternatively, most of the pseudogenes may have arisen in the human lineage after divergence from the common human:mouse ancestor.

Human putative genes were not conserved in the mouse; against our expectations the mouse sequence did not highlight additional un-annotated exons for these structures in the human sequence. The lack of sequence conservation across exonic regions of human putative genes may mean that:

i. The orthologous structures are altogether absent from the mouse genome.

ii. Orthologous structures exist but because of the absence of coding constraints they have diverged to such an extent that searches do not identify them.

iii. They represent mistakes in the transcription machinery.

This study has shown that putative gene structures differ significantly from coding genes and further computational and experimental analysis will be required to address their significance as functional elements. In fact, in a preliminary investigation of approximately 4 Mb of mouse finished and computationally analysed sequence from this region, I did not identify any splicing ESTs to indicate the presence of mouse putative genes.

The mouse sequence was used to provide an estimate for the degree of completeness for the 20q12-13.2 annotation. Analysis and correlation of gene predictions and mouse hits suggest that at least 97% of exons in this region have been annotated, which is in agreement with our published estimate (Deloukas *et al*., 2001). The putative coding, un-annotated exons identified by this analysis require experimental verification (isolation of cDNA sequences or identification of homologous ESTs). The benefit of using comparative sequence analysis and prediction programs to identify coding regions is that it is not limited by spatial or temporal restrictions on transcription. However, this also means that expression of these regions is difficult to confirm.

PipMaker was used to perform "global" comparative analysis. If only strong alignments are considered (>70% identity), the percentage of exonic regions that align is ~twelve-fold higher compared to introns and intergenic regions, respectively. Excluding exons, the regions upstream of annotated gene-starts show the highest alignability. Promoter analysis (chapter III) indicates that most annotated genes have virtually complete gene structures, which suggests that these regions correspond to either 5' UTRs or promoter sequences, enriched in various protein-binding signals.

The emerging finished mouse sequence is subject to the established high standards of sequence analysis and annotation. Comparison of gene features found in the finished sequence of human and mouse showed that the lengths of coding exons are conserved, whereas those of introns are not. All size differences identified across orthologous exons involve 3 bp (or multiples of) insertions/deletions suggesting conservation of the ORF. Overall, the total lengths of human introns were found to be 1.27-fold longer, compared to that of mouse. Finished sequence across the region will be required to investigate features such as untranslated exons and intergenic regions.

The sequence features of three orthologous gene pairs were studied in more detail. In all cases, the orthologous ORFs shared extensive homology (>84%) both at the nucleotide and protein level. Gene pairs had the same number of coding exons and used the same translation stop codon. The UTR sizes were not conserved, but the same polyA signal was found at approximately the same position, within their 3' UTRs.

For the three gene pairs studied, most (62/70) of the orthologous splice site junctions were identical. Exceptions include four 3' exon/5' intron and four 3' intron/5' exon sites. The GC content of the orthologous regions was similar. CpG islands were identified at the gene-starts of all genes. CpG islands of orthologous genes have similar GC contents but on average, in the human sequence, CpG islands are 17% longer than in the mouse sequence. Whether this is true for all CpG island pairs in the whole region remains to be investigated.

This study shows that comparative sequence analysis can be used to systematically identify coding exons; the identification of non-coding functional features is less efficient

because of the large number of conserved regions identified by homology searches. Even in a region previously subjected to extensive computational and experimental gene annotation (chapter III) this approach contributes new, although very few, exons. In my opinion, the combination of the methods described in this and the previous chapter (computational annotation, experimental verification/extension and comparative analysis) provide the most robust approach for large-scale identification of sequence features. In addition, I would favour the parallel analysis with additional genomes to further investigate the non-coding conserved regions.

Chapter V

Sequence variation

5.1 Introduction

5.1.1 Human variation

Inherited differences in DNA sequence contribute to phenotypic variation, influencing an individual's anthropometric characteristics, risk of disease and response to the environment (ISNPMWG, 2001). SNPs are the most common type of DNA variation in the human genome and large-scale discovery projects, aiming to catalogue them, are under way (section 1.7.1).

Although common human genetic variation is limited compared to other species, it remains impractical to discover and test all SNPs for a role in each disease. One attractive proposal suggests the testing of only functional SNPs (Collins *et al*., 1997). Functional variants are likely to either introduce an amino acid change or alter the base composition of regulatory sequences. With a partial gene list in hand and poor knowledge of the location of regulatory elements, this approach is currently not fully applicable.

The availability of an alternative strategy was revealed by empirical studies, which demonstrated that nearby SNPs often display strong correlation in the population: Inheritance of one SNP allele is tightly linked to the state of other, closely linked sites. These correlations exist because the unit of human inheritance is not the individual SNP, but rather the ancestral segment that has undergone minimal historical recombination, and thus has been handed down from generation to generation with little modification. A biological basis for defining these ancestral segments (haplotypes) is to examine the

genomic patterns of recombination. This can be achieved using linkage disequilibrium analysis.

5.1.2 Theoretical aspects of linkage disequilibrium

Linkage Disequilibrium (LD) refers to the non-random association of alleles at linked loci. Such associations underlie all forms of genetic mapping. However, linkage analysis is based upon associations in well-characterised pedigrees, whereas LD refers to the associations within populations of "unrelated" individuals. Nonetheless, there is a close relationship between the two approaches, because the "unrelated" individuals in a population are unrelated only in a relative and approximate sense (Nordborg and Tavare, 2002). In other words, the "unrelated" individuals will have a common ancestor at some point in the distant past. This makes LD particularly suitable to fine-scale mapping because it allows a lot more opportunities for recombination to take place.

LD is quantified using statistics of association between the allelic states at pairs of loci. D is one of the earliest measures of LD proposed (Lewontin, 1964), and quantifies disequilibrium as the difference between the observed frequency of a two-locus haplotype and the expected frequency if the alleles were segregating at random (i.e. if they were in linkage equilibrium). Consider two loci A and B with alleles A_1/A_2 and B_1/B_2. The proportion of chromosomes on which alleles A_1 and B_1 co-occur in the population is the observed frequency, denoted by P_{11}. The expected frequency under linkage equilibrium is the product of the allele frequencies in the population. Thus,

$$D = P_{11} - p_1 q_1 \qquad (1)$$

where the allele frequencies are symbolised as follows: p_1 = f(A_1); $p2$ = 1 - p_1 = f(A_2); q_1 = f(B_1); q_2= 1- q_1 = f(B_2).

If D differs significantly from zero, LD is said to exist. The degree of LD between two loci is dependent on both the recombination fraction, θ, and time in generations, *t*. Thus, D will tend to be smaller when two loci are located further apart, and D will decrease through time as a result of recombination (Jorde, 2000).

Although D captures the intuitive concept of LD, its numerical value is of little use for measuring the strength of and comparing levels of LD. This is due to the dependence of D on allele frequencies in the population: its maximum value is given by D_{max} = min(p_1q_2, p_2q_1), whereas its minimum value is given by D_{min} = max(-p_1q_1, -p_2q_2) As a result, several alternative measures, based on D, have been devised (reviewed in Devlin and Risch, 1995; note that although they are all based on Lewontin's D, they have different properties and measure different things (Ardlie *et al*., 2002)). The most common measures are the absolute value of D' and r^2.

The absolute value of D' is determined by dividing D by D_{max} (Lewontin, 1964).

$$D' = \frac{D}{D_{max}} \qquad (2)$$

D' = 1 (complete LD) if, and only if, two SNPs have not been separated by recombination (or recurrent mutation or gene conversion) during the history of the sample. In this case, at most three of the four possible two-locus haplotypes are observed

in the sample (Figure 5.1). Values of D' < 1 indicate that the complete ancestral LD has been disrupted, but the relative magnitude of values of D' < 1 has no clear interpretation. Estimates of D' are inflated in small samples, especially for SNPs with rare alleles. In addition, samples are difficult to compare because the magnitude of D' depends strongly on sample size. Therefore, statistically significant values of D' that are near 1 provide a useful indication of minimal historical recombination, but intermediate values should not be used for comparisons of the strength of LD between studies, or to measure the extent of LD (Ardlie *et al*., 2002).

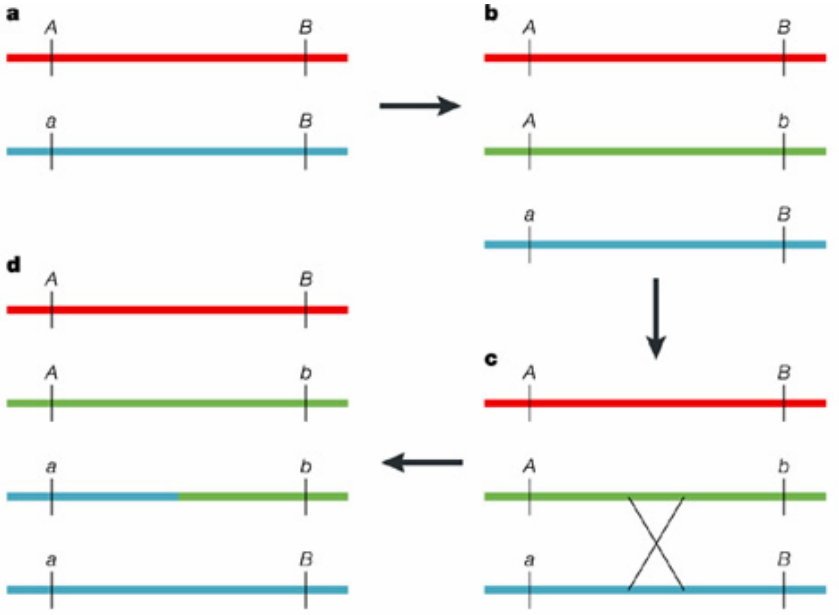


Figure 5.1 (reproduced from Ardlie *et al*., 2002): The erosion of linkage disequilibrium by recombination. (a) At the outset, there is a polymorphic locus with alleles A and a. (b) When a mutation occurs at a nearby locus, changing an allele B to b, this occurs on a single chromosome bearing either allele A or a at the first locus (A in this example). So, early in the lifetime of the mutation, only three out of the four possible haplotypes will be observed in the population. The b allele will always be found on a chromosome with the A allele at the adjacent locus. (c) The association between alleles at the two loci will gradually be disrupted by recombination between the loci. (d) This will result in the creation of the fourth possible haplotype and an eventual decline in LD among the markers in the population as the recombinant chromosome (a, b) increases in frequency.

The measure r^2 (Hill and Robertson, 1968; also labelled R^2 or Δ^2) is in some ways complementary to D', and has recently emerged as the measure of choice for quantifying and comparing LD in the context of mapping (Pritchard and Przeworski, 2001; Weiss and Clark, 2002).

$$r^2 = \frac{D^2}{(p_1 p_2 q_1 q_2)} \quad (3)$$

$r^2 = 1$ (perfect LD) if, and only if, the markers have not been separated by recombination and have the same allele frequency. In this case, exactly two out of the four possible two-locus haplotypes are observed in the sample and observations at one marker (locus) provide complete information about the other marker (locus). Note that the value of r^2 is related to the amount of information provided by one locus about the other and that this property correctly takes into account differences in allele frequencies at the two loci. However, it also means that two markers that are immediately adjacent might show different r^2 values with a third marker, and that a low pairwise r^2 is not necessarily indicative of high ancestral recombination in the region. The sample size required to detect statistically significant LD is inversely proportional to r^2. r^2 also shows much less inflation in small samples than does D'. Typically, r^2 is lower than D' for any chromosomal distance (Ardlie *et al*., 2002; Weiss and Clark, 2002).

Mutation and recombination might have the most evident impact on LD, but there are additional contributors to the extent and distribution of disequilibrium. Most of these involve demographic aspects of a population, and tend to sever the relationship between LD strength and the physical distance between loci. Examples include population growth,

admixture/migration, population structure, natural selection, variable recombination rates, variable mutation rates and gene conversion.

5.1.3 Allelic associations and common disease

Many studies have examined LD across small regions, like genes. These studies have generally concluded that the extent of disequilibrium is highly variable both across and between regions, and also differs between populations (reviewed in Pritchard and Przeworski, 2001; Boehnke, 2000; Jorde, 2000).

Studies across large contiguous genomic regions show a variable pattern of LD with regions of nearly complete LD interspersed with regions that show little, or no LD. Furthermore, they increasingly suggest that the human genome can be parsed objectively into haplotype blocks: sizeable regions over which there is little evidence for historical recombination, and within which only a few common haplotypes are observed (Olivier *et al*., 2001; Reich *et al*., 2001; Patil *et al*., 2001; Dawson *et al*., 2002; Gabriel *et al*., 2002).

With most human genetic variation being attributable to a limited set of common haplotypes, scanning the genome for regions of association to disease becomes feasible by testing the subset of variants able to report the common variants. However, for the study of complex, common diseases, a key question is whether the causative variants that confer disease susceptibility are likely to be common or rare. The answer cannot be known with certainty, of course, until these variants are identified and characterised; but there is a growing list of examples of common variants that predispose to common disease (examples include the SNPs in the apolipoprotein E gene (Davignon *et al*., 1988) and the factor V Leiden mutation (Bertina *et al*., 1994)).

To systematically test this hypothesis in a given population, we need a map of haplotype blocks that captures most of the genome. Although a dense SNP map is already available, empirical studies suggest that many more SNPs may be needed to achieve coverage of >90% of the genome in haplotype blocks (Jeffreys *et al*., 2001); such a project will require testing well above one million SNPs in multiple populations. This can only be achieved through the development of high throughput and cost effective genotyping platforms (reviewed in section 1.7.2). Although no one technique can be designated as the method of choice, the application of mass spectrometry in SNP genotyping emerges as a serious contender.

5.1.4 Mass spectrometry

5.1.4.1 Background

Recent technological innovations have made nucleic acids accessible to mass spectrometric analysis. Due to its inherently high specificity, accuracy and throughput, mass spectrometry is an attractive detection method for SNP genotyping (Jackson *et al*., 2000; Kwok, 1998; Leushner, 2001).

Mass spectrometry (MS) is used to measure the mass-to-charge ratio (m/z) of ions, which can be used to infer their molecular weight. The recent advent of electrospray ionisation (ESI, Fenn *et al.*, 1989) and matrix-assisted laser desorption-ionisation (MALDI, Karas and Hillenkamp, 1988) techniques enable the routine mass spectroscopic analysis of various biomolecules, including peptides/proteins, lipids and carbohydrates (reviewed in Griffiths *et al*., 2001; Harvey, 2001), as well as nucleic acids.

Before the advent of ESI and MALDI, it was not possible to acquire the mass spectra of non-volatile, thermally labile, intact molecules with molecular weights greater than ~1-2 KDa. ESI and MALDI allow the production of gas-phase ions from solution and solid phases respectively. Different ionisation methods can be used with different mass analysers but most commonly ESI is coupled to either a quadrupole or ion-trap analyser whereas MALDI is coupled to a time-of-flight (TOF) analyser (Jackson *et al*., 2000). Of these, only MALDI-TOF will be discussed in more detail.

In MALDI-TOF the compound to be analysed (the analyte) is co-crystallised with excess light-absorbing matrix. Under high vacuum, the sample crystal is irradiated with an ultraviolet laser pulse that vaporises and ionises both analyte and matrix at the same time. Since the matrix absorbs most of the laser energy, sample fragmentation does not usually occur for smaller molecules. In the TOF analyser, the molecular ions are accelerated and passed over a flight tube, during which the ions are separated according to their m/z ratios. The smaller the ion, the faster it reaches the end of the tube. By measuring the ions at the end of the tube over a short time, a mass spectrum is generated. Under optimal conditions, MALDI produces singly charged ions. This simplifies the calculation of molecular masses that can be determined with high accuracy (0.01% to 0.1%). The entire process, including data acquisition, can be completed in approximately 3 seconds (Leushner and Chiu, 2000). A schematic diagram of MALDI-TOF is shown in Figure 5.2.

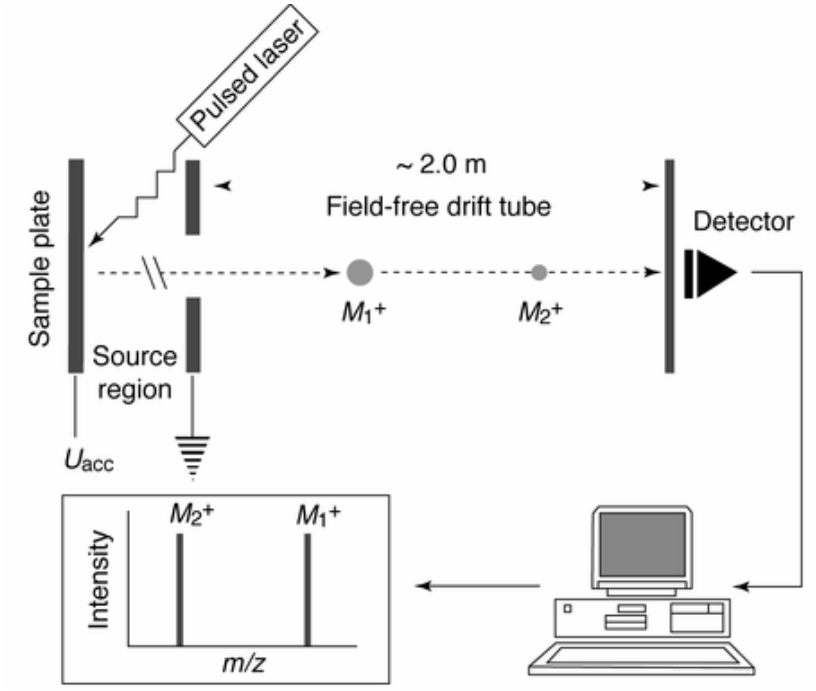


Figure 5.2: MALDI-TOF MS (reproduced from Griffin and Smith, 2000, 2002). Matrix and analyte ions are desorbed and ionised upon irradiance with a laser pulse in the source region; a potential (U_{acc}) applied to the sample accelerates the ions into the field-free drift tube. The time-of-flight of each ion is measured and converted into m/z. The diagram shows the separation and detection of two positive, single-charged ions with different masses, M_1 and M_2.

5.1.4.2 Genotyping methods using mass spectrometry

The short oligonucleotide mass analysis (SOMA) is one of the few techniques reported that employs ESI rather than MALDI for SNP analysis (Laken *et al*., 1998). The genomic region to be analysed is PCR amplified with primers containing a sequence for a type IIS restriction enzyme. Enzymatic digestion of the PCR products yields fragments as small as 7 bp. HPLC is used for improved purification of the samples, which are then analysed by ESI MS.

One of the MALDI methods developed for genotyping uses allele specific, mass-labelled, peptide nucleic acid (PNA) hybridisation probes (Griffin *et al*., 1997). PNA probes are structural analogs that have increased hybridisation stability and specificity over conventional DNA probes. The biotinylated target DNA (e.g. a PCR amplicon) is immobilised by binding to streptavidin-coated magnetic beads. The non-biotinylated strand is removed, followed by PNA probe hybridisation. Stringent washing conditions are used to remove the unbound probe and achieve proper discrimination. MALDI is then used to determine the mass of the bound PNA probes.

MALDI-TOF MS can also determine genotypes by characterising the products of primer extension (mini-sequencing) reactions. Application of a mini-sequencing based approach involves PCR amplification of the sequence of interest, followed by the annealing of a primer extension probe hybridising immediately upstream of the variable site. The probe is extended and the mass of the extended products is then used to determine the composition of the variable site. In the PROBE (primer oligonucleotide base extension) assay (Braun *et al*., 1997a, 1997b), the annealed probe is extended through the SNP site in the presence of three dNTPs and one ddNTP (incorporation of a ddNTP terminates the extension reaction). In the PinPoint assay (Haff and Smirnov, 1997; Ross *et al*., 1998; Fei *et al*., 1998) the primer extension probe is extended by only one base, in the presence of four ddNTPs. Finally, in the very short extension assays (Sun *et al*., 2000), the primer extension probe is extended in the presence of one dNTP and three ddNTPs, which tends to produce very short extension products. Platforms that employ this approach include the MassEXTEND assay (Sequenom) and the GOOD assay (Sauer *et al*., 2000).

MALDI was also been used in a miniaturised, chip-based probe annealing, extension and termination approach (Tang *et al*., 1999). It was also employed to analyse the products of Invader assays (Griffin *et al*., 1999). Finally, genotyping could potentially be achieved through sequencing by mass spectrometry (Köster *et al*., 1996; Kirpekar *et al*., 1998; Fu *et al*., 1998)

5.1.5 This chapter

This chapter describes:

i. The identification of exonic SNPs in 20q12-13.2 by comparing the finished reference sequence to EST and mRNA sequences. The annotation of the region was used to classify each SNP as either non-synonymous (amino acid change), or synonymous (no amino acid change), or in UTR, whereas a subset was experimentally verified using the homogeneous MassEXTEND assay (Sequenom).

ii. Selection, genotyping and analysis of a set of circa 2,200 SNPs distributed across 119 individuals from three populations: twelve unrelated individuals of Asian origin, twelve unrelated individuals of African American origin and 95 Caucasian individuals from twelve multigenerational pedigrees. The three panels (each of twelve individuals) were used to verify the SNPs and estimate their allele frequencies in each population. The genotype data from the twelve pedigrees was used to obtain haplotype data and investigate the extent of LD across the region.

5.2 Exonic SNP discovery in 20q12-13.2

5.2.1 Identification of exonic SNPs in silico

BLAST searches (performed as part of the automated analysis, section 3.2) were used to identify human expressed sequences with high homology to the finished sequence. The graphical BLAST viewer Blixem (Sonnhammer and Durbin, 1994) was used to manually inspect the alignments of the homologous human EST, cDNA and in-house generated sequences (Figure 5.3).

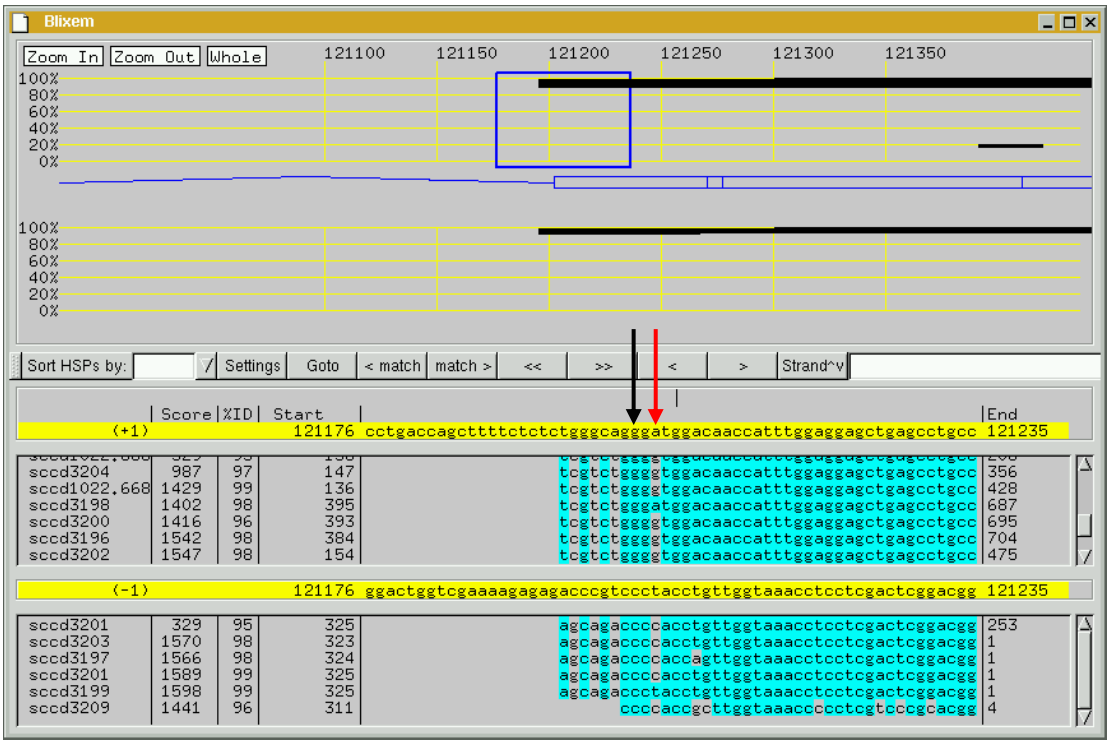


Figure 5.3: Blixem view of homologous sequences (also see Figure 4.2). The region shown corresponds to the sequence from clone dJ453C12 that encodes part of the last exon of C20orf35. The exon starts at position 121,202 of the clone sequence (black arrow). Homologous expressed sequences corresponding to either the forward (+1) or reverse (-1) strand are shown below (sequence names at far left; not all sequences are shown). For each expressed sequence, nucleotides identical to the reference sequence are highlighted blue. Various vectorette sequences and other ESTs (e.g. BE908675 and AW37112; not shown) identify a candidate A□G variation at position 121,204 (red arrow).

Each position, where two or more expressed sequences differed from the reference genomic sequence, was tagged as harbouring a putative exonic SNP. Expressed sequences that differ at multiple positions from the reference sequence (< 95% ID) were not used for SNP identification.

This approach yielded 124 putative SNPs. As shown in Figure 5.4 more than 60% of these SNPs are supported by at least four expressed sequences. Fifteen of the SNPs were known (previously identified by other SNP discovery projects). The 109 new SNPs were incorporated into 20ace (http://webace.sanger.ac.uk/cgi-bin/webace?db=acedb20) and submitted to dbSNP (http://www.ncbi.nlm.nih.gov/SNP/snp_search.cgi?searchType=byBatch&batch_id=4640).

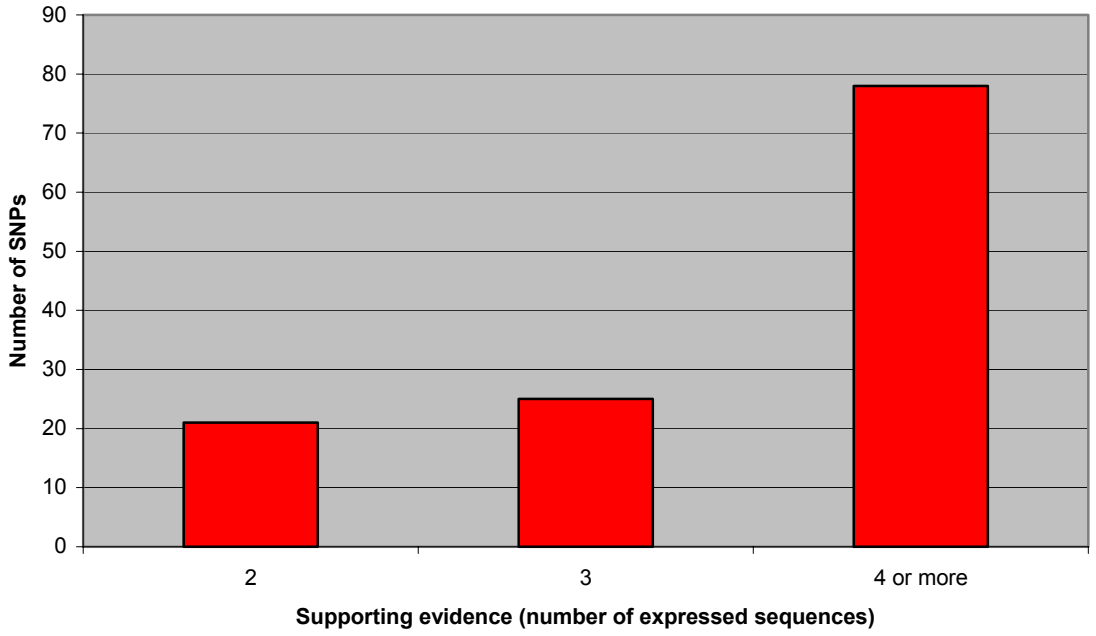


Figure 5.4: Supporting evidence for exonic SNPs.

5.2.2 Features of exonic SNPs

Under random mutation, half as many transitions as transversions are expected to occur (Dawson *et al*., 2001; Table 5.1). Since the strand on which the changes occurred is not known, no distinction can be made between A↔G and C↔T and between C↔A and G↔T. In this data set (124 SNPs) transitions (66.9%) occur twice as often as transversions (33.1%) (Table 5.2), a pattern already reported in other SNP identification studies (Horton *et al*., 1998; Dawson *et al*., 2001; Deutsch *et al*., 2001). The most common change was C↔T (G↔A) (83/124), which probably reflects the deamination of 5-methylcytosine that occurs frequently at CpG dinucleotides. The other thing to note from Table 5.2 is that the occurrence of A↔T (T↔A) is less than half compared to any other variation (Dawson *et al*., 2001; Smink, 2000).

Table 5.1: Expected distribution of transitions and transversions.

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

Table 5.2: Distribution of transitions and transversions.

Variation	Number	Number
C↔T (G↔A)	83	66.9%
Transitions	**83**	**66.9%**
C↔A (T↔G)	28	22.6%
C↔G	9	7.3%
A↔T	4	3.2%
Transversions	**41**	**33.1%**

On the basis of the 20q12-13.2 annotation, at least one exonic SNP was identified for 47 of the 99 coding genes in the region (Appendix 14). Most (92/124) SNPs are in the UTRs, whereas the remaining 32 (25.8%) map within the ORF of twenty genes (Table 5.3). Of the 21 variations that correspond to a 3' codon position change (Figure 5.5), only one results in an amino acid change. In contrast, five of the six first base changes and four of the five second base changes result in amino acid changes.

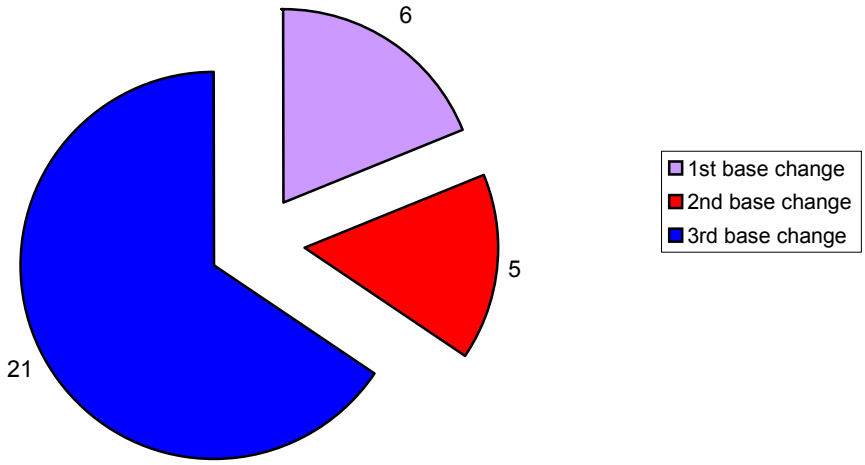


Figure 5.5: Codon position changes for coding exonic SNPs.

Table 5.3: Coding changes for exonic SNPs.

Clone name	SNP position (clone sequence)	Gene	Allele change	Codon change	Encoded amino acid	Amino acid change
dJ511B24	64,001	PLCG1	T ↔ C	ATC ↔ ACC	ile ↔ thr	yes
dJ862K6	107,817	SFRS6	T ↔ C	CCT ↔ CCC	pro ↔ pro	
dJ138B7	73,871	C20orf9	G ↔ A	GTG ↔ GTA	val ↔ val	
dJ1028D15	93,173	MYBL2	T ↔ C	CCT ↔ CCC	pro ↔ pro	
dJ1028D15	96,061	MYBL2	C ↔ G	GCC ↔ GCG	ala ↔ ala	
dJ148E22	56,722	YWHAB	A ↔ C	CGA ↔ CGC	arg ↔ arg	
dJ148E22	60,103	YWHAB	T ↔ A	GTG ↔ GAG	val ↔ glu	yes
dJ1069P2	89,939	TOMM34	A ↔ G	GCA ↔ GCG	ala ↔ ala	
dJ1069P2	94,339	C20orf119	G ↔ A	CTC ↔ CTT	leu ↔ leu	
dJ1069P2	94,345	C20orf119	G ↔ A	CTG ↔ TTG	leu ↔ leu	
dJ1069P2	95,369	C20orf119	T ↔ C	TCA ↔ TCG	ser ↔ ser	
dJ453C12	121,204	C20orf35	A ↔ G	ATG ↔ GTG	met ↔ val	yes
dJ453C12	135,622	PIGT	G ↔ A	ACG ↔ ACA	thr ↔ thr	
dJ453C12	135,781	PIGT	C ↔ T	AGC ↔ AGT	ser ↔ ser	
dJ453C12	136,979	PIGT	T ↔ C	TAT ↔ TAC	tyr ↔ tyr	
dJ461P17	119,487	C20orf170	T ↔ C	AAT ↔ AGT	asn ↔ ser	yes
dJ447F3	93,739	TNNC2	C ↔ A	ACG ↔ ACT	thr ↔ thr	
dJ337O18	14,252	PTE1	G ↔ A	GTC ↔ GTT	val ↔ val	
bA465L10	100,183	MMP9	G ↔ C	CGG ↔ CCG	arg ↔ pro	yes
bA394O2	49,352	KIAA1834	A ↔ G	GAT ↔ GAC	asp ↔ asp	
dJ28H20	2,734	C20orf64	G ↔ A	GCC ↔ GCT	ala ↔ ala	
dJ1049G16	55,543	NCOA3	G ↔ A	GGG ↔ GGA	gly ↔ gly	
dJ1049G16	67,054	NCOA3	A ↔ G	CAA ↔ CAG	gln ↔ gln	
dJ1049G16	67,057	NCOA3	G ↔ A	CAG ↔ CAA	gln ↔ gln	
dJ1049G16	67,084	NCOA3	A ↔ G	CAA ↔ CAG	gln ↔ gln	
bA269H4	64,588	KIAA1415	T ↔ C	AAG ↔ GAG	lys ↔ glu	yes
dJ998C11	2,266	KIAA1415	G ↔ C	CAC ↔ CAG	his ↔ gln	yes
dJ155G6	63,791	CSE1L	T ↔ C	TCT ↔ CCT	ser ↔ pro	yes
dJ155G6	63,887	CSE1L	A ↔ C	AAA ↔ CAA	lys ↔ gln	yes
dJ686N3	10,895	DDX27	C ↔ T	TTC ↔ TTT	phe ↔ phe	
dJ686N3	34,718	KIAA1404	A ↔ G	CAT ↔ CAC	his ↔ his	
dJ686N3	34,741	KIAA1404	G ↔ C	CTT ↔ GTT	leu ↔ val	yes

5.3 Studying sequence variation across 20q12-13.2

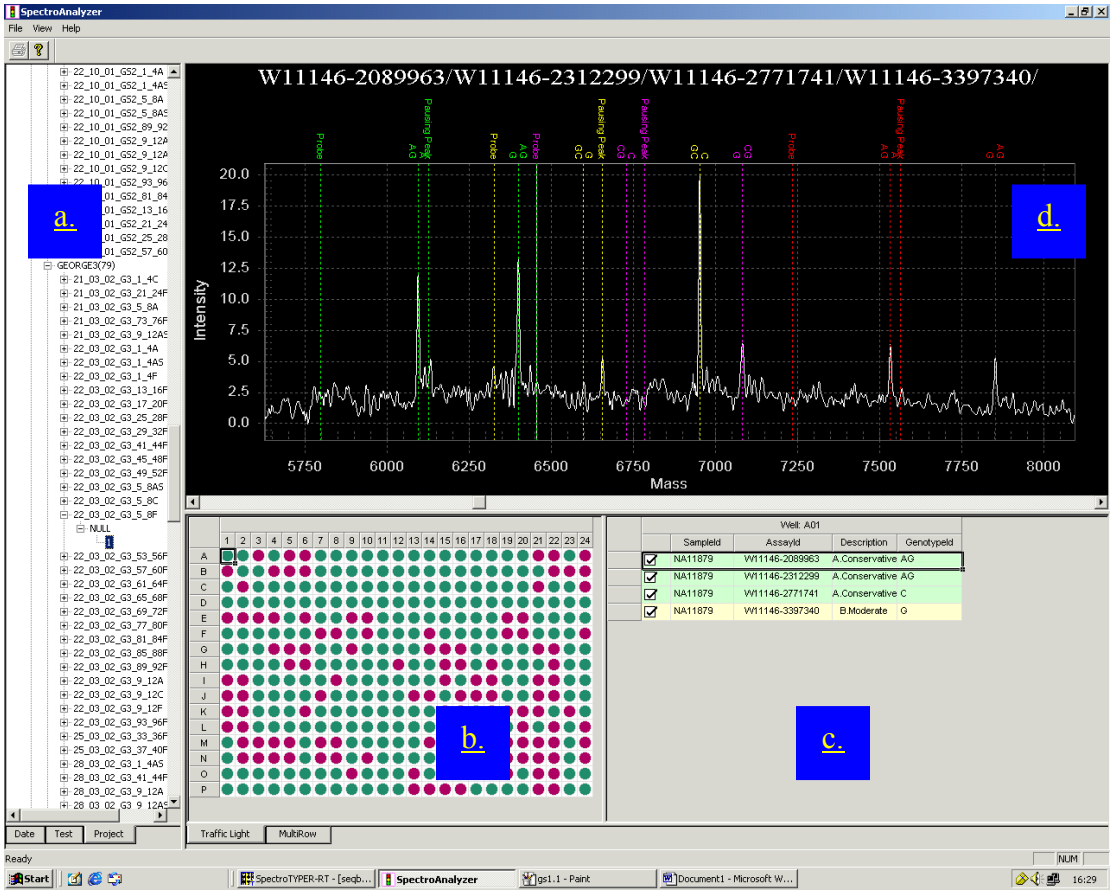
5.3.1 SNP selection and high-throughput genotyping

Homogeneous MassEXTEND assays were designed for 2,208 SNPs mapping in 20q12-13.2 using the SpectroDesigner software. During the first round of SNP selection, SNPs mapping more than >15 Kb apart were selected. During subsequent SNP selection rounds, neighbouring SNPs were selected to replace failed or non-polymorphic ones; additional SNPs mapping between polymorphic SNPs were also selected to decrease the average SNP distance to ~5 Kb. In general, SNPs mapping <2 Kb apart were avoided. 106 of the SNPs selected were identified by this study (section 5.2), whilst the remaining 2,102 were imported from dbSNP or generated in-house by a parallel SNP discovery project on chromosome 20. In brief, chromosome 20 was flow-sorted from a Caucasian, an African American, an African pygmy and a Chinese cell line. Individual 2 Kb insert libraries were prepared and shotgun sequenced to a depth of 2x coverage. The SSAHA (Ning *et al*., 2001) software was used to align reads to the genome assembly and over 110,000 new SNPs were discovered (Deloukas, pers. communication). Note that this rich resource became available at a very late stage of this study and was only used to smooth the initially uneven distribution of the selected SNPs.

Assays were designed as quadruplex reactions and genotyping was performed as described in section 2.3. Genotyping was attempted in 119 individuals from three ethnic groups (twelve African Americans, twelve Asians and 95 Caucasians). DNA sample information is listed in Table 2.7 (Chapter II). Automated call analysis was performed

using the SpectroTyper RT package and data was stored in an Oracle database. An example of genotype calls is shown in Figure 5.6.

A.



B.

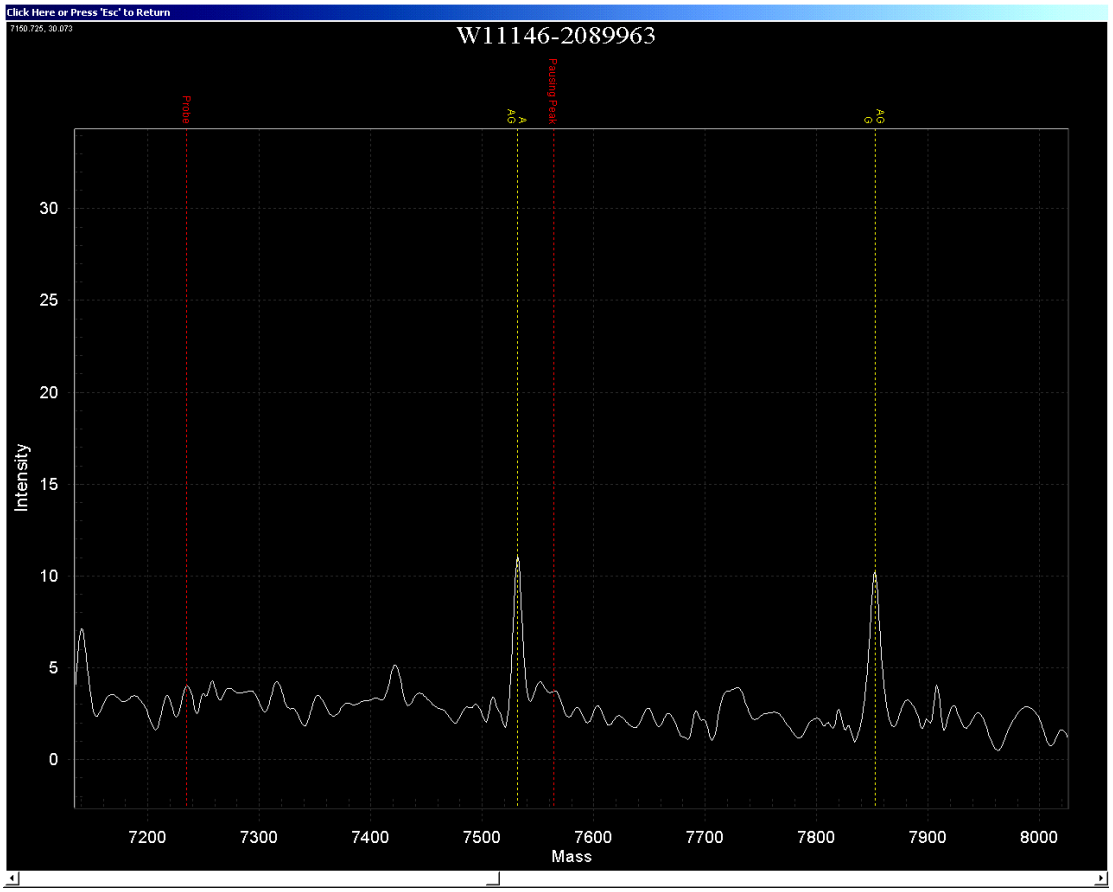


Figure 5.6: Viewing genotyping results in SpectroAnalyser (SpectroTyper RT software package). (A) Sub window a lists genotyped plates; plate 22_03_02_G3_5_8F was selected for viewing. Sub window b shows a virtual 384-well plate, highlighting the type of genotyping results obtained for each well. Wells with four conservative/moderate calls are highlighted green, the rest are highlighted red. For a selected virtual well (A1 in this example) the individual assay information are listed in sub window c. Data listed include the DNA tested, well/SNP assay numbers, type of calls and the genotypes obtained. Sub window d shows the spectra obtained for these assays. Each colour corresponds to a particular assay. (B) Detailed view of the W11146-2089963 spectra. The "Probe" line (red) shows the expected peak position for un-incorporated (not extended) probe. Both "Probe" and "Pausing Peak" lines help monitor the completeness of the extension reaction. The two yellow lines show the peak positions for the two alleles obtained (A and G).

5.3.2 Error checks and quality assessment of data

All checks described in this section were performed by Sarah Hunt, using commercial software and customised in-house perl scripts.

The genotypes obtained from the twelve Caucasian families were used to test the SNP assays for Hardy-Weinberg equilibrium (Pedigree Statistics (c) 1999-2001, Gonçalo Abecasis). 48 assays violated H-W equilibrium (χ^2>10) and they were excluded from further analyses.

Mendelian checks (PedCheck 1.1 (c) 1997-1999, Jeff O'Connell; MERLIN 0.8.8 (c) 2000-2001, Gonçalo Abecasis) were first performed in parents/offspring trios and then in whole families. 1,459 genotypes (from 202 SNPs) were involved in Mendelian errors and were excluded from further analyses.

Two independent tests were used to estimate error rates. Comparison of duplicate calls in the raw Sequenom data (all genotype calls) from this study suggests an error rate of 1.4% (983/69,618). Note that this percentage corresponds to the raw data (which includes all assays).

The results were also compared to data obtained by an independent study, which used the Illumina SNP assay platform. In total, 566 SNPs had conclusive data from both methods and 295 differences were detected in the combined set of 27,901 genotypes (1.1%). Of the total 295 discrepant calls, 215 (72.9%) are associated with only 21 SNP assays (3.7%). This and previous studies, using the Sequenom platform in our laboratory have shown that 2-3% of SNP assays (randomly selected from public resources) are sub-optimal. Although the exact reason is not fully understood, imbalanced allele

amplification appears to be the most likely cause. 490 SNPs (86.6%) showed no discrepant calls for any DNA. Less than half of the total differences (119/27,901; 0.43%) corresponded to high quality genotype calls (i.e. conservative calls for Sequenom and confidence of "5" for Illumina calls).

5.3.3 Estimation of allele frequencies in three populations

Of the 2,208 SNPs genotyped in the three ethnic groups, circa 15% (331) failed, i.e. did not produce any genotype data, or failed during data checks (section 5.3.2). For a further 438 (20%) SNP assays, genotype data was incomplete for at least one ethnic group (genotype data was obtained for less than seven of the twelve individuals from each group). In total, 1,439/2,208 (65%) SNP assays gave genotype data for at least seven individuals in each group (Figure 5.7). This set of 1,439 "complete" SNPs was used for the analyses described below.

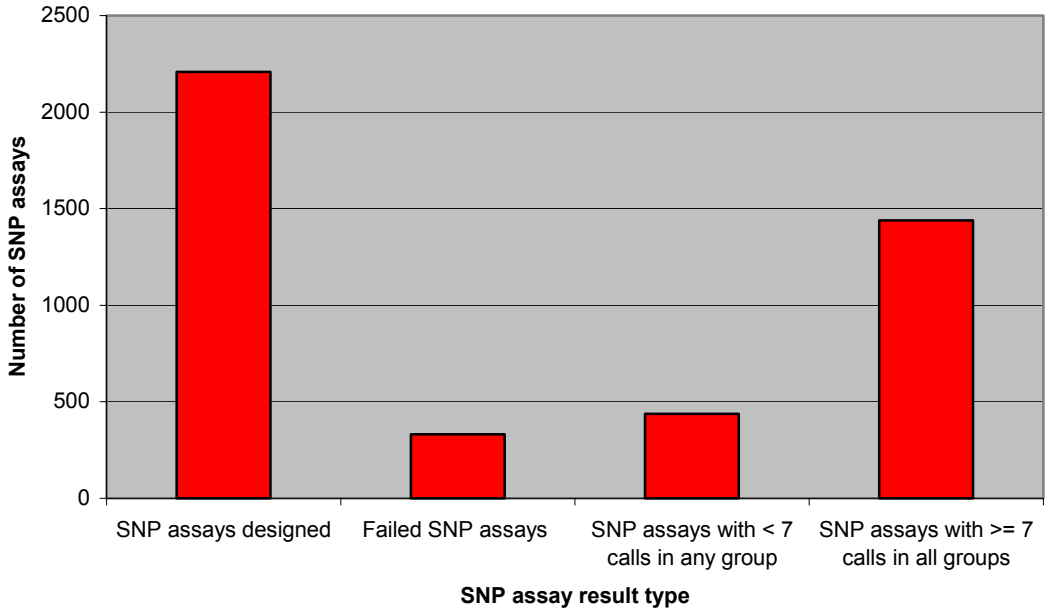


Figure 5.7: Overall breakdown of SNP assay results.

A total of 46,627 non-redundant genotypes were obtained for the 1,439 SNPs with "complete" assays (out of 51,804 possible genotypes). On average, 32.4 individuals were successfully genotyped for each SNP (32.4/36, 90% complete). 340 (23.5%) SNPs were found to be non-polymorphic in all ethnic groups (for the individuals tested), whilst the remaining 1,099 SNPs (76.5%) were found to be polymorphic in at least one population. The results are shown in Figure 5.8. The corresponding figures for the exonic SNPs identified *in silico* (section 5.2) were 48% non-polymorphic and 52% polymorphic in at least one population. This suggests that the *in silico* approach either yields more rare SNPs, or has a high false positive rate due to sequencing errors.

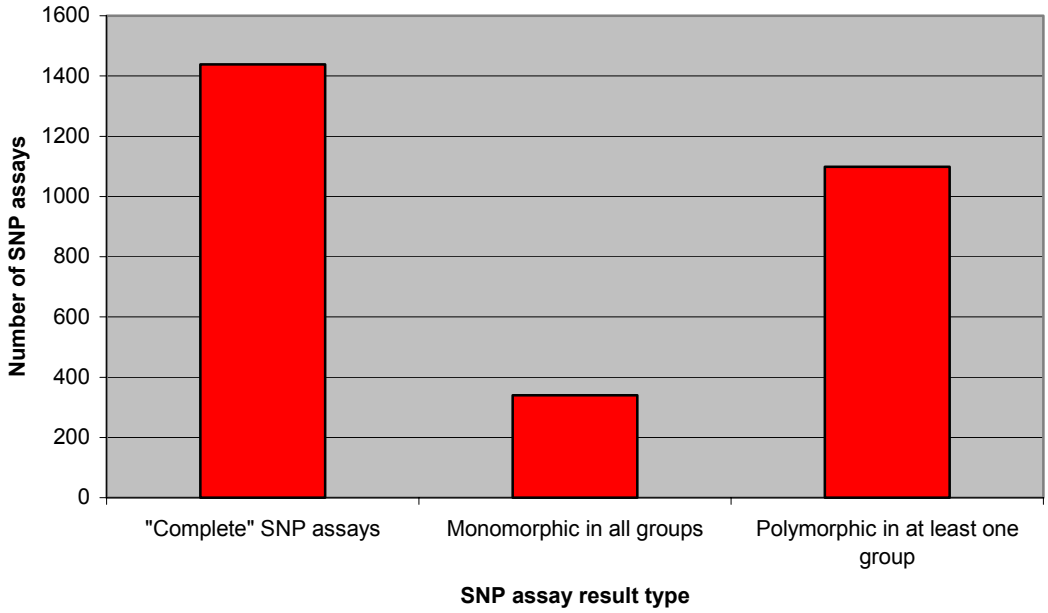


Figure 5.8: Breakdown of "complete" SNP assay results.

Figure 5.9 shows the distribution of polymorphic and non-polymorphic SNPs across the three populations. Compared to the other two populations, African Americans had the highest number of polymorphic SNPs (985/1,099). The corresponding numbers for Asians and Caucasians were 850 and 904, respectively. Approximately 66% (725/1,099) of SNPs were polymorphic in all groups.

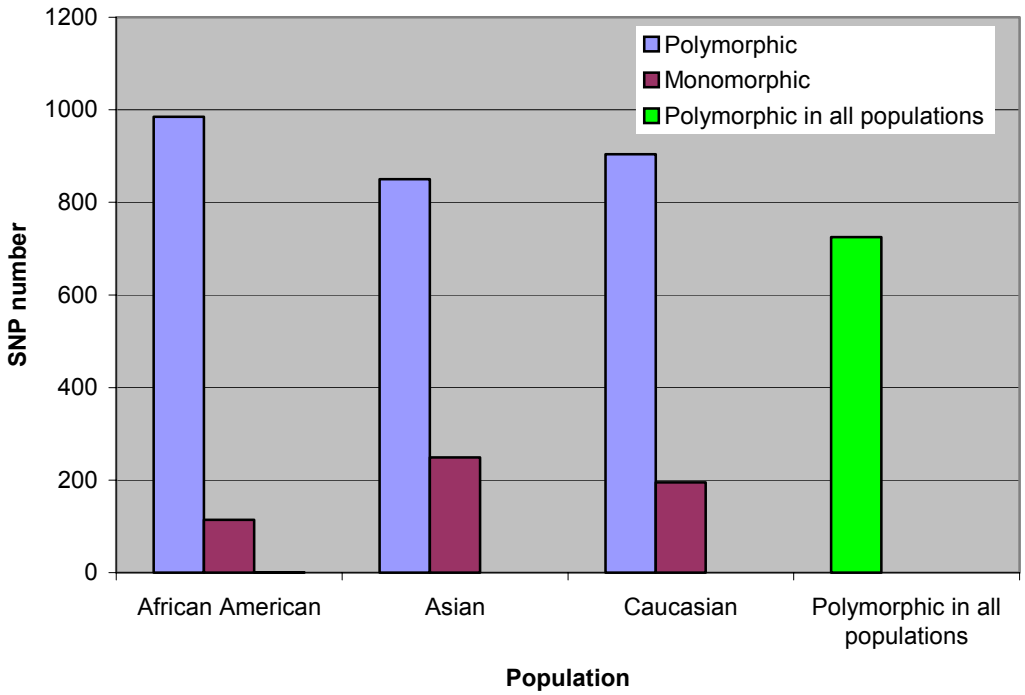


Figure 5.9: Distribution of polymorphic and monomorphic SNPs across the three ethnic groups. Only polymorphic SNPs, in at least one group, were considered (1,099 SNPs).

As illustrated in Figure 5.10, the African Americans showed more than two-fold more unique polymorphisms than either the Asians, or Caucasians. Of the SNPs found to be polymorphic in only two groups, the African American and Caucasian groups shared the most. Compared to Asians and Caucasians, the African Americans and Asians also shared more SNPs (polymorphic in only two groups).

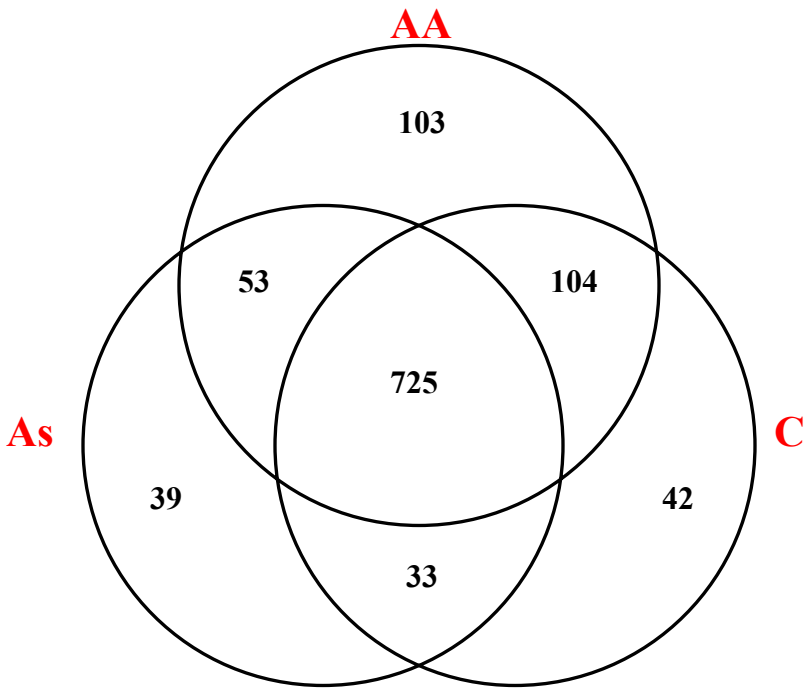


Figure 5.10: Distribution of polymorphic SNPs in the three groups (African Americans, AA; Asians, As; Caucasians, C).

Overall, the African Americans had the largest proportion of SNPs at the lower end of minor allele frequencies (40% of SNPs had a Minor Allele Frequency (MAF) of 4-20%). The corresponding numbers for Asians and Caucasians were 31.5% and 27.4%, respectively. The reverse was observed in the Caucasian population, for which 55% of SNPs had a MAF of 21-50%. The corresponding numbers for the Asian and African American populations were 45.8% and 49.5%, respectively (Figure 5.11).

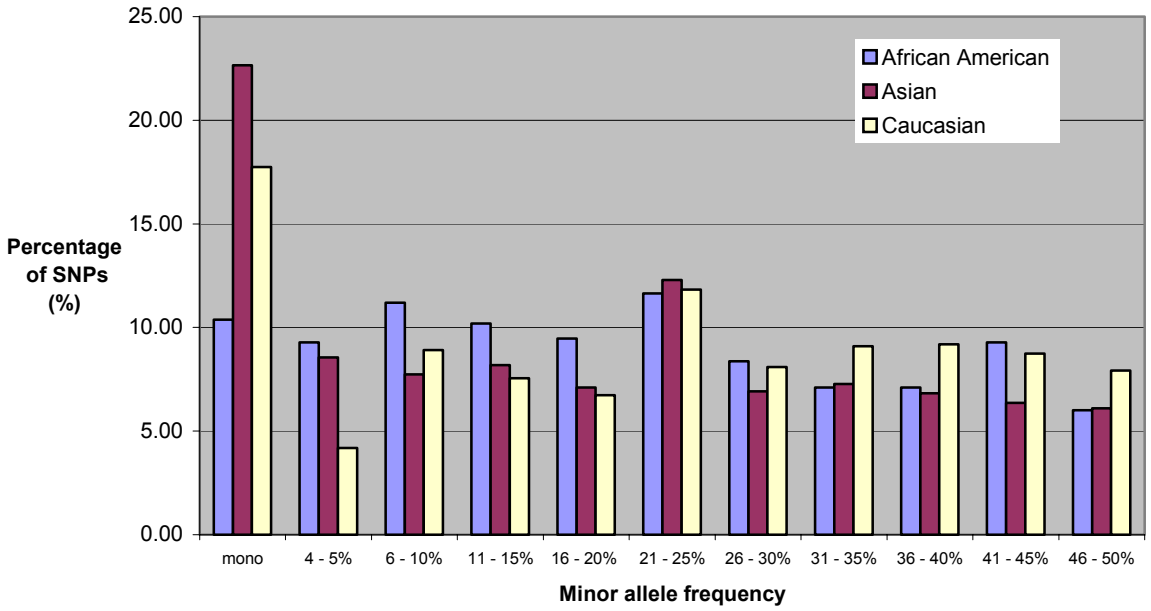
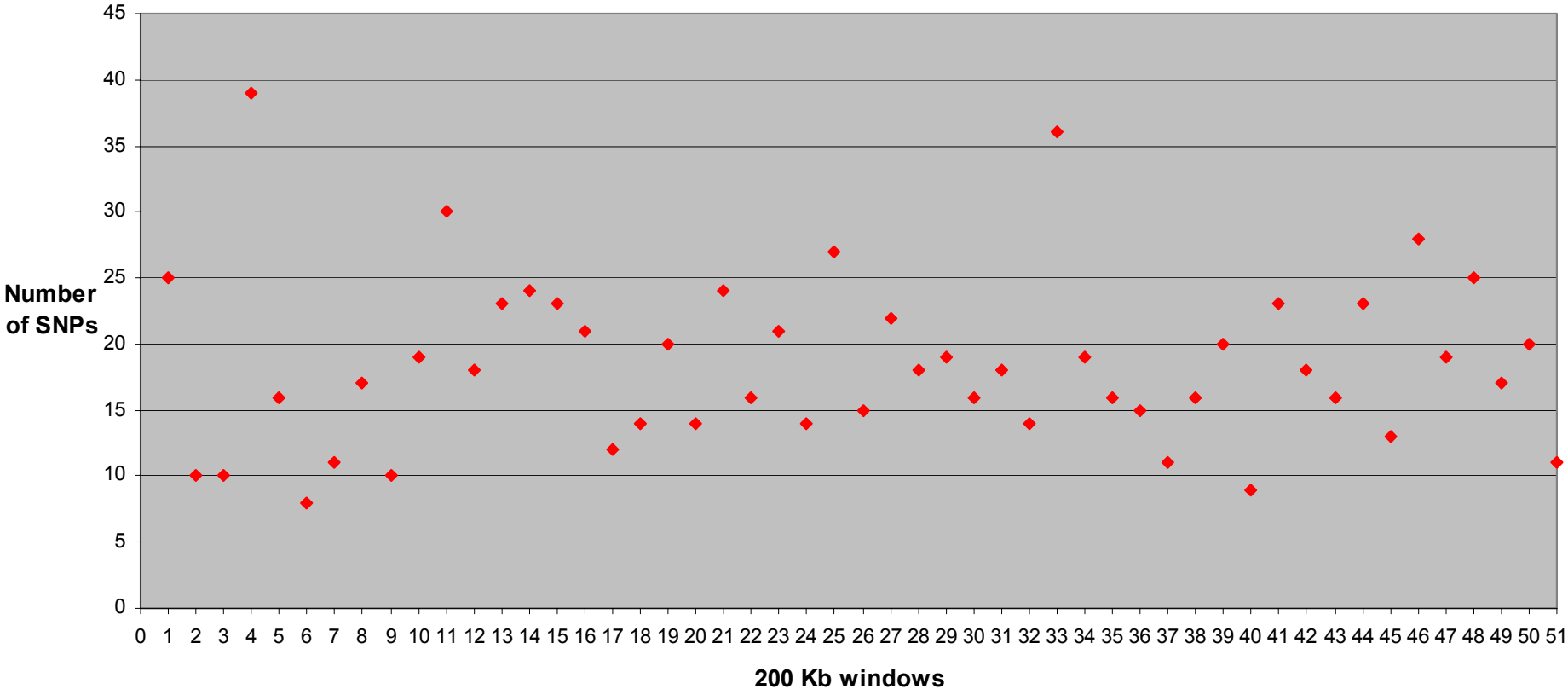


Figure 5.11: Distribution of minor allele frequencies in ethnic populations.

5.4 A first generation LD map of 20q12-13.2 in Caucasians

The 2,208 SNP assays were used to genotype twelve three-generation CEPH families with a total of 95 individuals (section 5.3.1). Allele frequencies were calculated using only the founder chromosomes (grandparents; 47 individuals). Following quality checks, a set of 943 SNPs, with MAF of ≥5% and with at least thirteen calls from founder chromosomes, was selected (Appendix 15). The distribution of these SNPs across 20q12-13.2 is shown in Figure 5.12.

Figure 5.12: SNP distribution across the region of interest. The y-axis reports the total number of SNPs in non-overlapping windows of 200 Kb (x-axis).



The average distance between SNPs is 10,709 bp (median 7,564 bp). As shown in Figure 5.13, 62.6% of neighbouring SNPs (590 of 942 pairs) are separated by less than 10 Kb, whilst 2.8% of SNP pairs (27) are separated by more than 40 Kb. The longest interval is 85,498 bp.

The sequence coverage per interval size (given in 5 Kb windows) is shown in Figure 5.14.

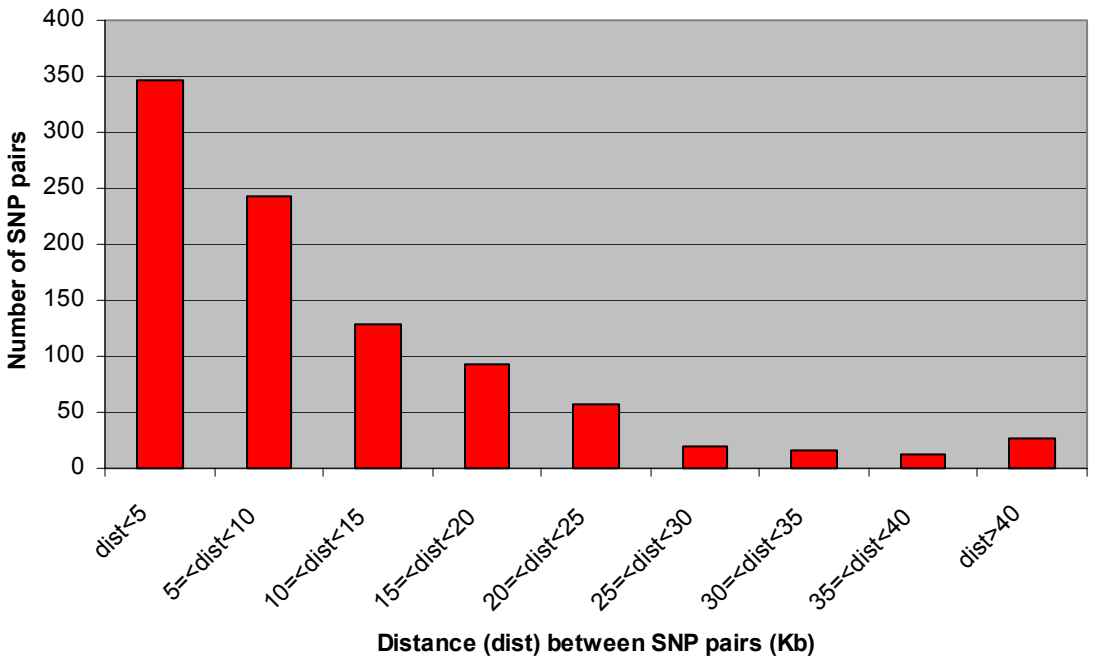


Figure 5.13: Distance between neighbouring SNPs (SNP pairs).

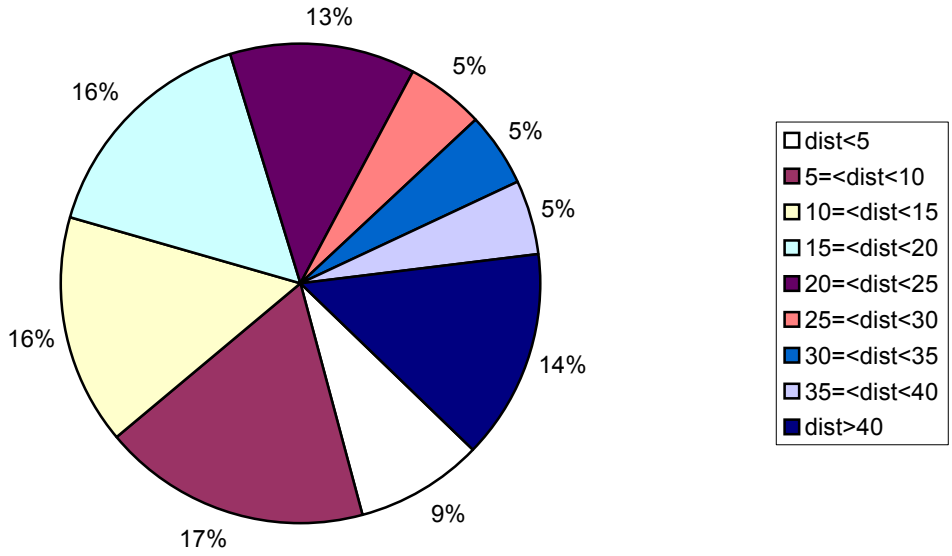
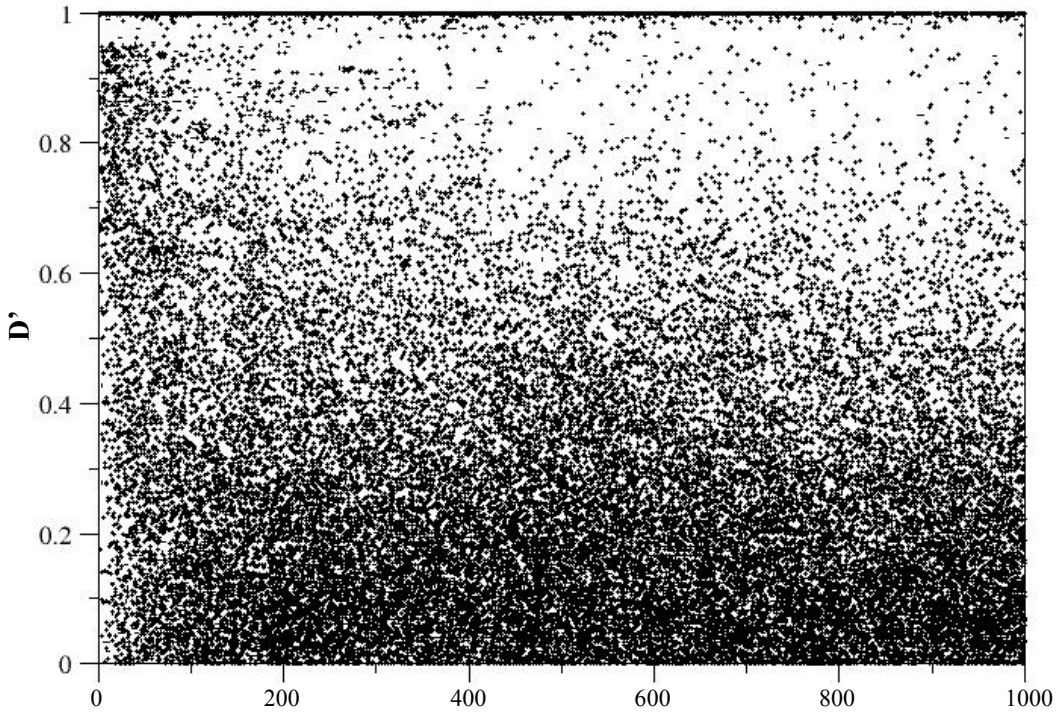


Figure 5.14: Proportion of sequence occupied by the various types of SNP pairs.

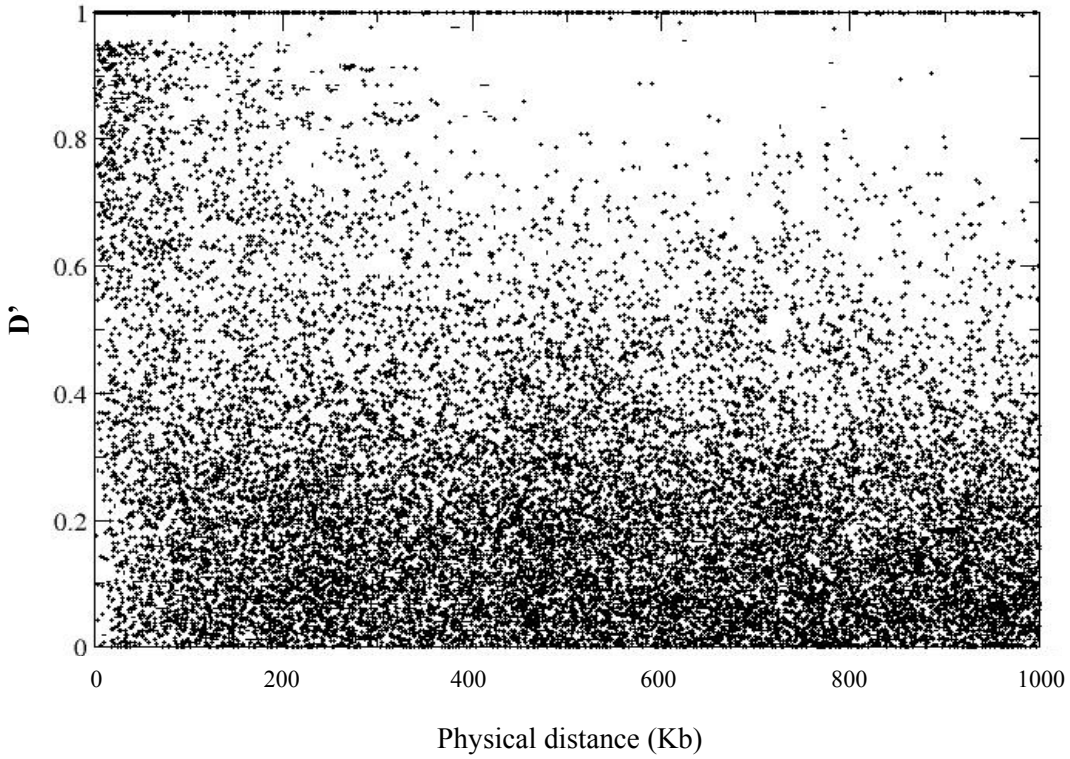
Of all polymorphic SNPs in Caucasians, those with >80% of all possible genotype calls were selected for further analysis. These 879 SNPs give an average spacing of 11.5 Kb. On average >93% of all genotyping calls were obtained for these SNPs. Statistical analyses were performed by Robert Lawrence at Oxford University.

LD between pairs of markers was calculated using D' and r^2. Calculations were performed for SNPs with MAF>10% (685), SNPs with MAF>20% (485), and all polymorphic SNPs (879). As shown in Figure 5.15, LD decays with increasing distance, but also shows extensive variability. Very high D' values (>0.9) extend up to distances over 250 Kb, contrasting with occurrences of very low D' values (<0.2) for SNPs less than 5 Kb apart (Figure 5.15 A, B). The corresponding r^2 values also show extensive variability (Figure 5.15 C, D). LD decay as a function of increasing physical distance was calculated using the average D' and r^2 in successive 10 Kb windows and the plots are shown in Figure 5.16. For common SNPs (MAF>20%), average D' declines from 0.89 for markers less than 10 Kb apart to ~0.22 for markers > 250 Kb apart (Figure 5.16 B), whereas the corresponding values for r^2 are 0.5 and 0.03 respectively (Figure 5.16 D). Although the two measures differ in scale, their decay profiles are similar. Note that when the average D' is calculated using the 685 SNPs with MAF>10% the average D' values decline from 0.89 to 0.31. The observed differences between the two D' values are expected, since the estimates of D' are known to be inflated by SNPs with rare alleles. (section 5.1, also see Appendix 16). Also note that when the less common SNPs are included, the maximal average r^2 decreases (Figure 5.16 C, D and Appendix 16).

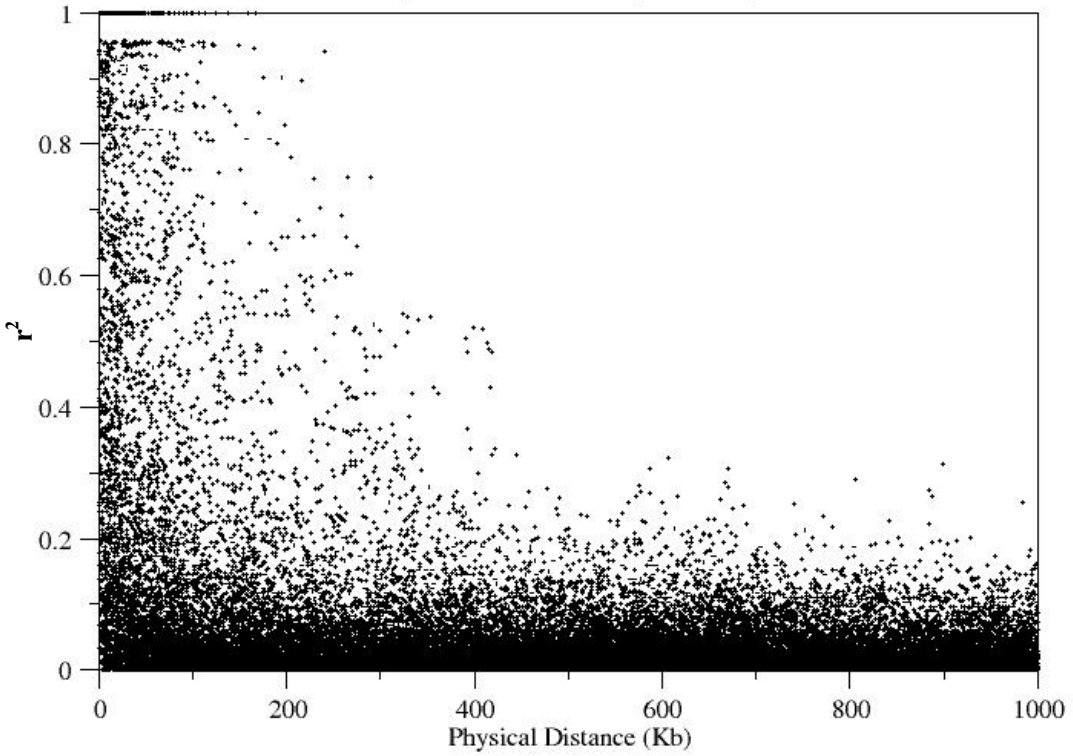
A. D' (using SNPs with minor allele frequency >10%)



B. D' (using SNPs with minor allele frequency >20%)



C. r^2 (using SNPs with minor allele frequency >10%)



D. r^2 (using SNPs with minor allele frequency >20%)

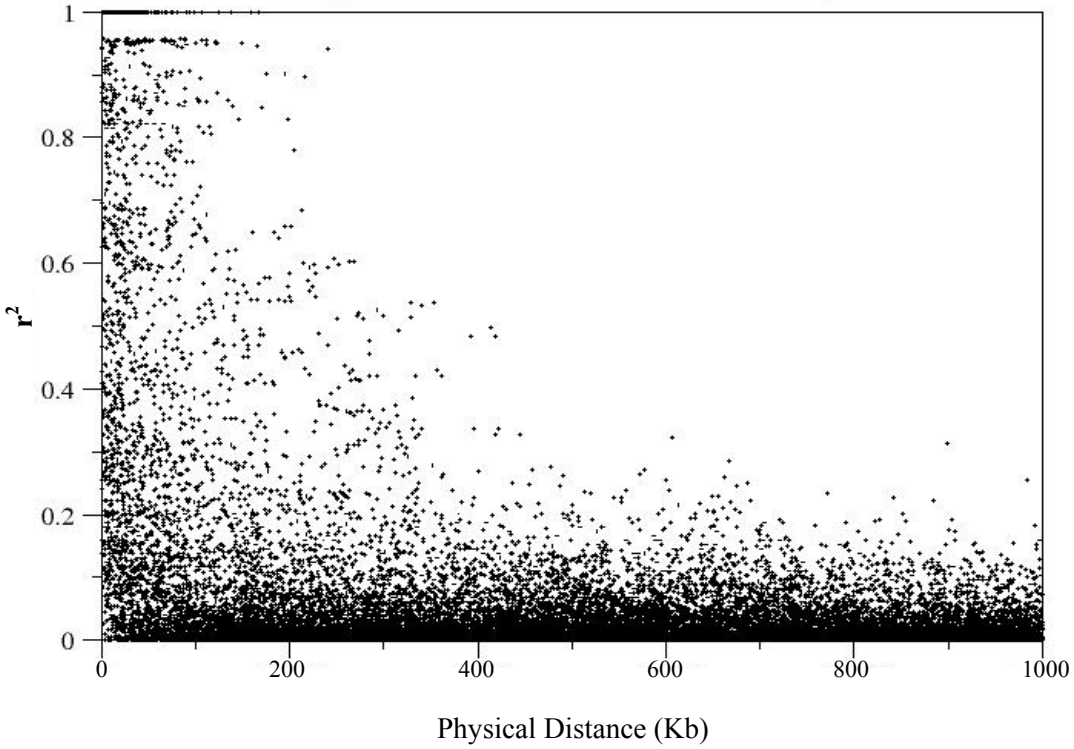
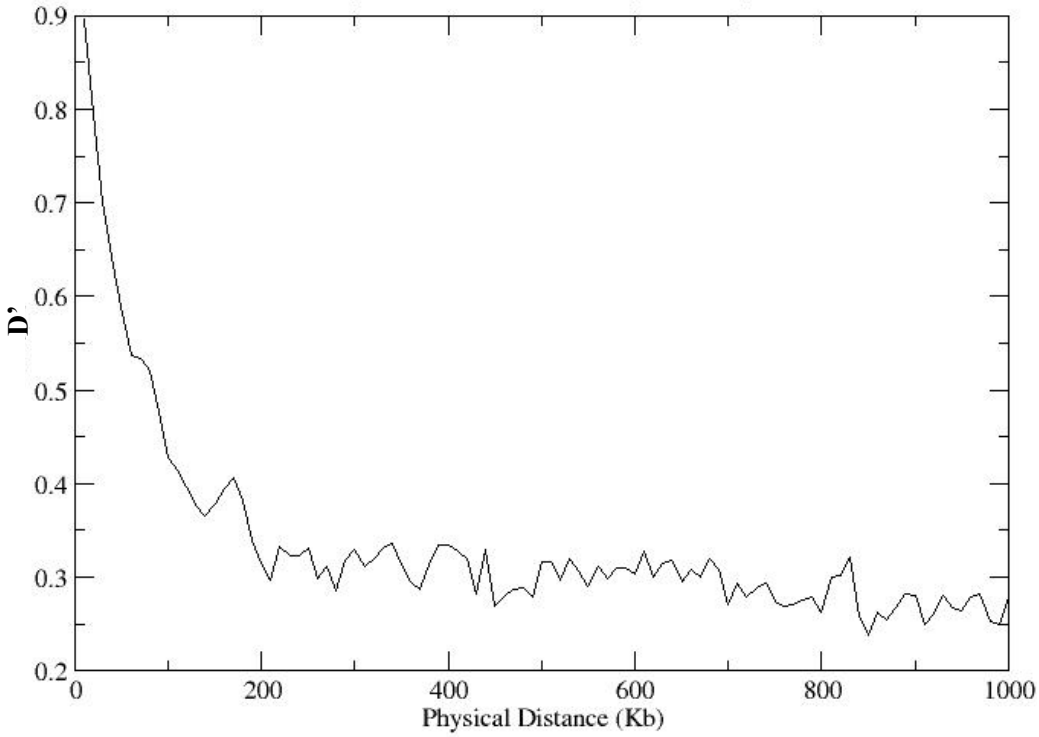
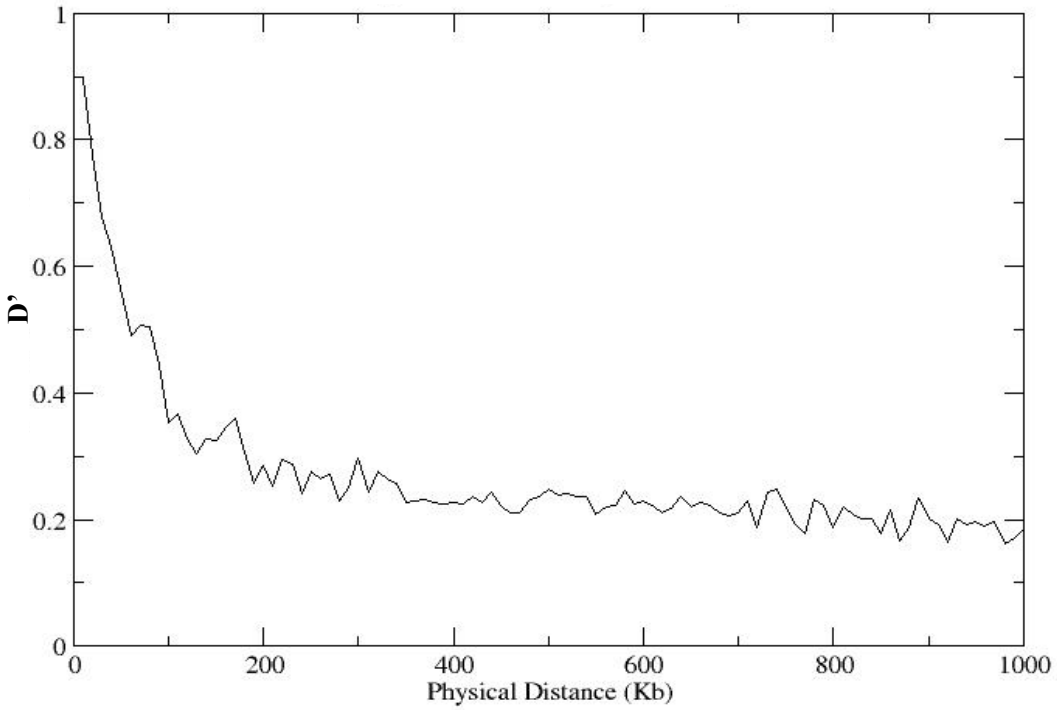


Figure 5.15: Variability of D' (A, B) and r^2 (C, D) using the pairwise values for SNPs separated by □1 Mb. The corresponding D' and r^2 scatter plots using all polymorphic SNPs are shown in Appendix 16.

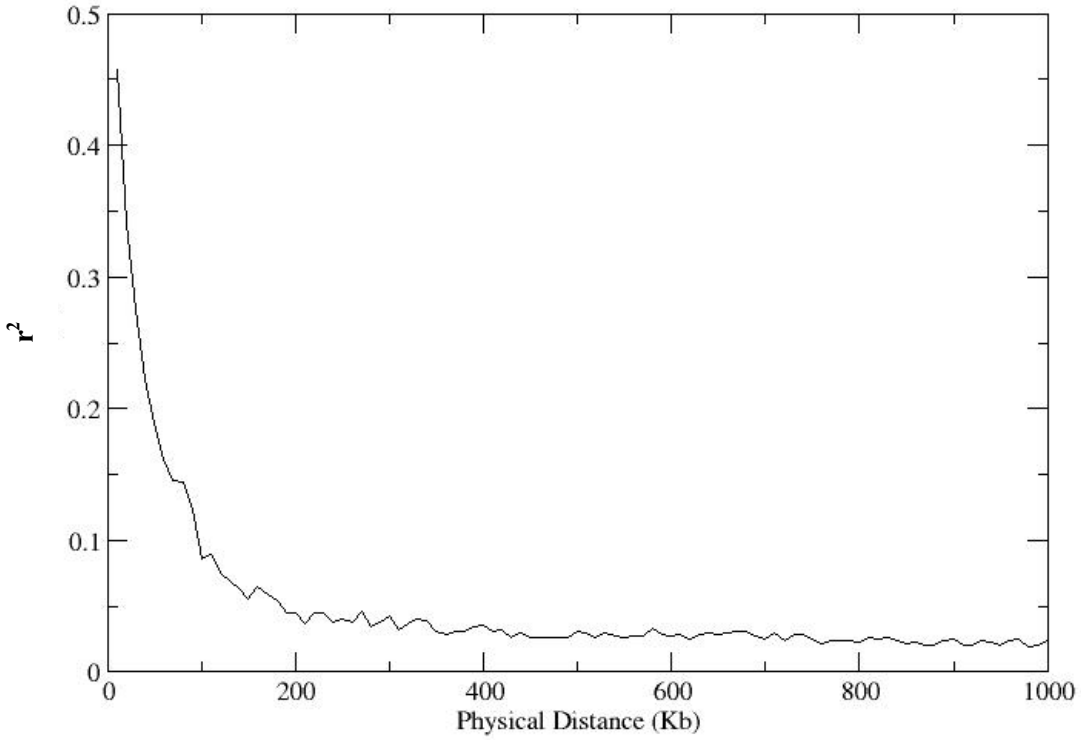
A. D' (using SNPs with minor allele frequency >10%)



B. D' (using SNPs with minor allele frequency >20%)



C. r^2 (using SNPs with minor allele frequency >10%)



D. r^2 (using SNPs with minor allele frequency >20%)

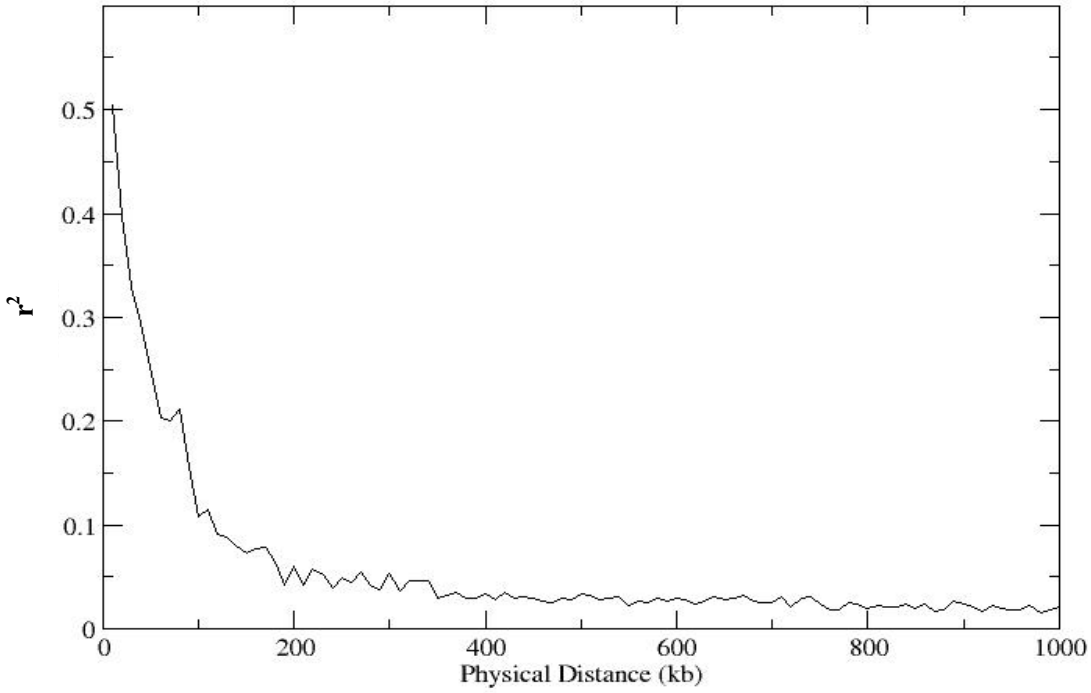


Figure 5.16: Average D' (A, B) and r^2 (C, D) in 10 Kb, non-overlapping, physical distance bins (up to 1 Mb). The average D' and r^2 plots using all polymorphic SNPs are shown in Appendix 16.

The pattern of LD along the 20q12-13.2 sequence was assessed by calculating the average D' and r^2 for polymorphic SNPs within continuous stretches of DNA (500 Kb windows, sliding by 250 Kb). As shown in Figure 5.17, the results highlight areas with high levels of LD, with peaks at 1-1.5 Mb, 3.5-4 Mb (detected only by D'), 5.5-6 Mb, 7.5-8 Mb (detected only by D') and 8.7-9.2 Mb.

There is considerable evidence that sites of recombination in humans are not randomly distributed, but are often localised into specific hotspots (Jeffreys et al., 2001). Current, low-resolution genetic maps can be used to model local recombination rates and provide additional predictors of LD beyond physical distance (Dawson *et al*., 2002). The region of 20q12-13.2 has an elevated degree of recombination, averaging 2.1 cM Mb^{-1} in the Marshfield sex-averaged genetic map (Figure 5.18), compared to the genome average of approximately 1.3 cM Mb^{-1}. With the exception of the LD peak at ~5.5-6 Mb all other peaks are situated within regions of very low recombination (<1 cM Mb^{-1}). Note that in the 5.5-6 Mb area two markers are wrongly placed on the genetic map, which complicates the correlation of the two maps (Figure 5.18). In addition, areas of higher recombination frequency are associated to steep decreases in LD. For example, this is observed at 4-5 Mb and 8-8.6 Mb (recombination rate >3 cM Mb^{-1}). These data suggest a strong correlation between the rate of recombination and the extent of LD which is in agreement with the findings of the recent chromosome 22 study (Dawson *et al*., 2002).

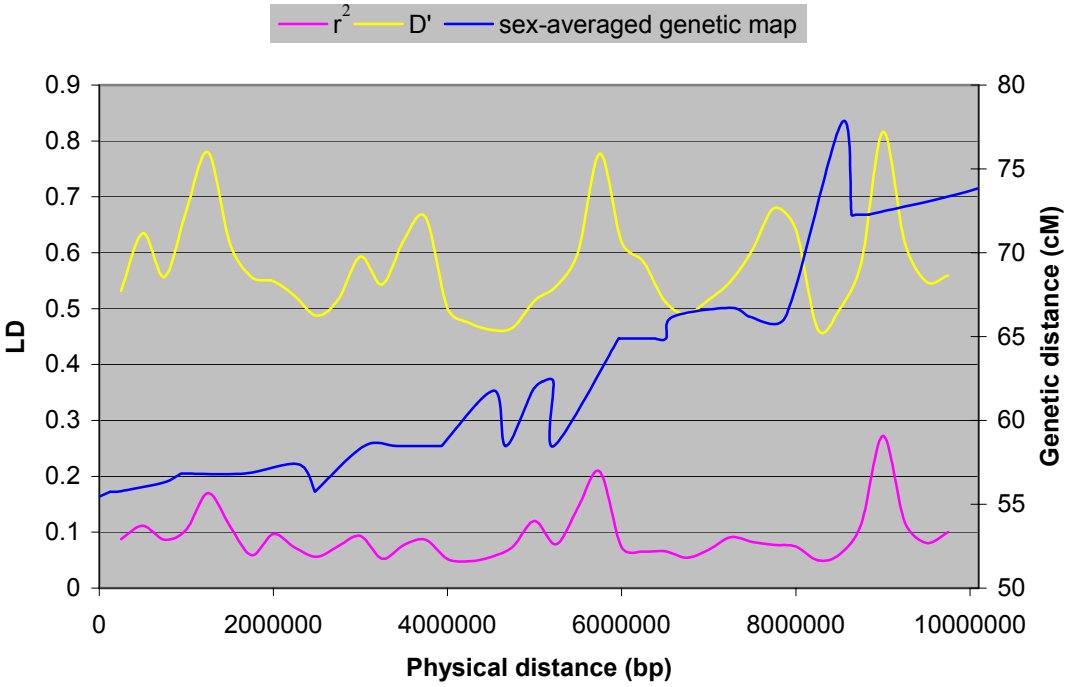


Figure 5.17: Linkage disequilibrium across 20q12-13.2. The average D' and r^2 are plotted in 500 Kb windows (sliding by 250 Kb) containing all polymorphic SNPs. The sex averaged genetic map for the region is also shown (also see Figure 5.18).

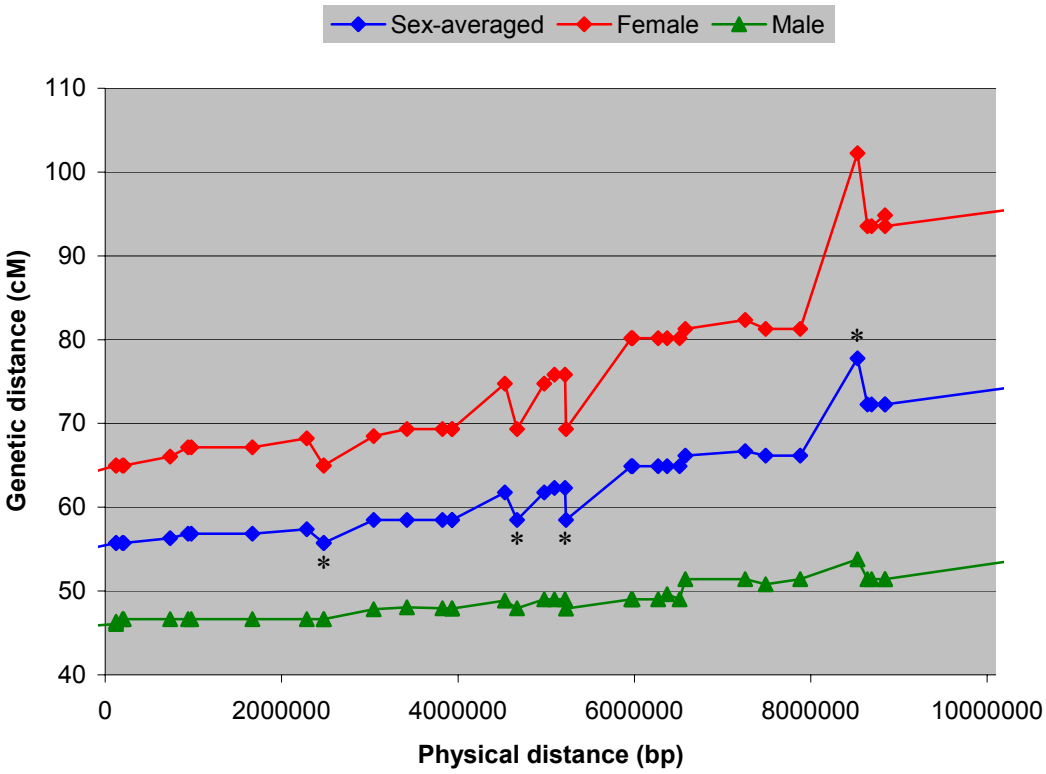
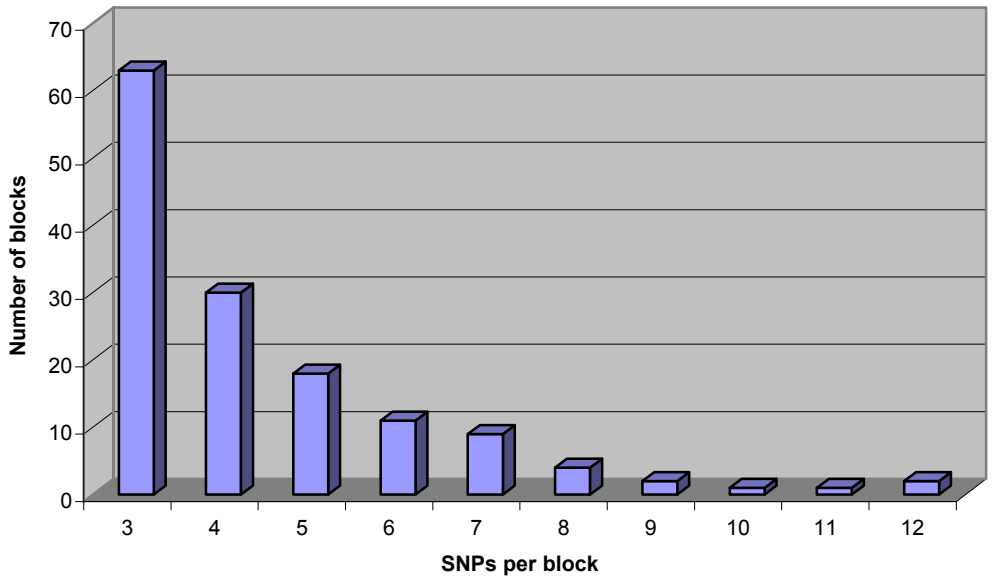


Figure 5.18 (reproduced from Deloukas *et al*., 2001 (only the region of interest shown)): The integrated Marshfield male, female and sex-averaged genetic maps (Yu *et al*., 2001), aligned to the sequence map of the region. The four discrepant points, marked with an asterisk, correspond to markers D20S466, D20S454, D20S424 and D20S427 (Utah Center of Genome Research markers UT1688, UT275, UT654 and UT1521 respectively).

Regions with three or more SNPs, for which all SNP pairs have D'>0.9, were identified as "LD blocks". A total of 141 such blocks were identified, which span 5.02 Mb (~50% of the region) and harbour 597 of the 879 analysed SNPs (68%). The mean block size is 35.6 Kb and the mean number of SNPs per block is 4.2 (corresponding median values 27.1 Kb and 4 respectively). The smallest and largest block identified span 2,332 and 142,291 bp respectively. As shown in Figure 5.19 A, most blocks (55%) contain four or more SNPs with two blocks having twelve SNPs each. Only 25% of the total sequence in blocks belongs to blocks with three markers (Figure 5.19 B).

A.



B.

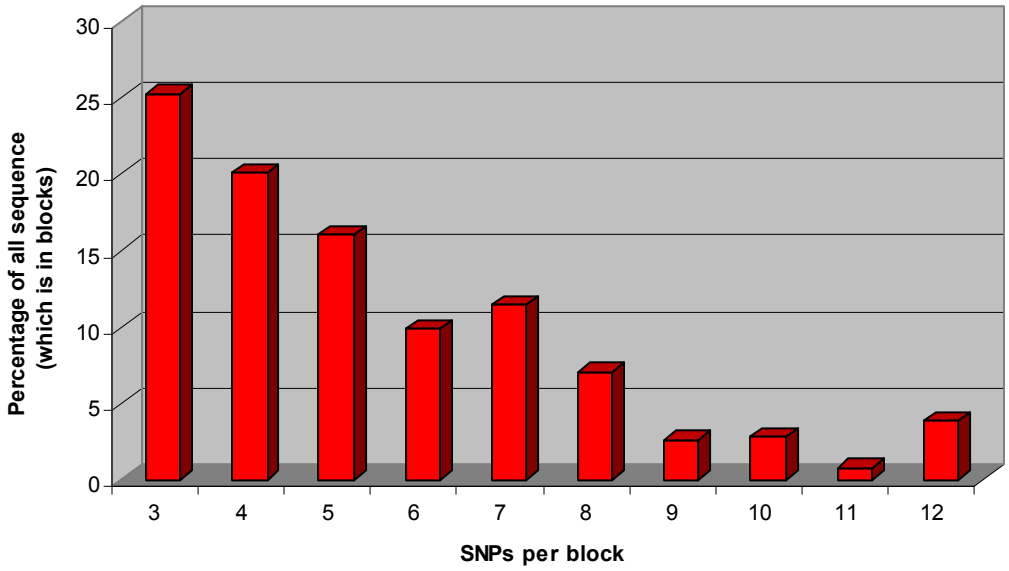
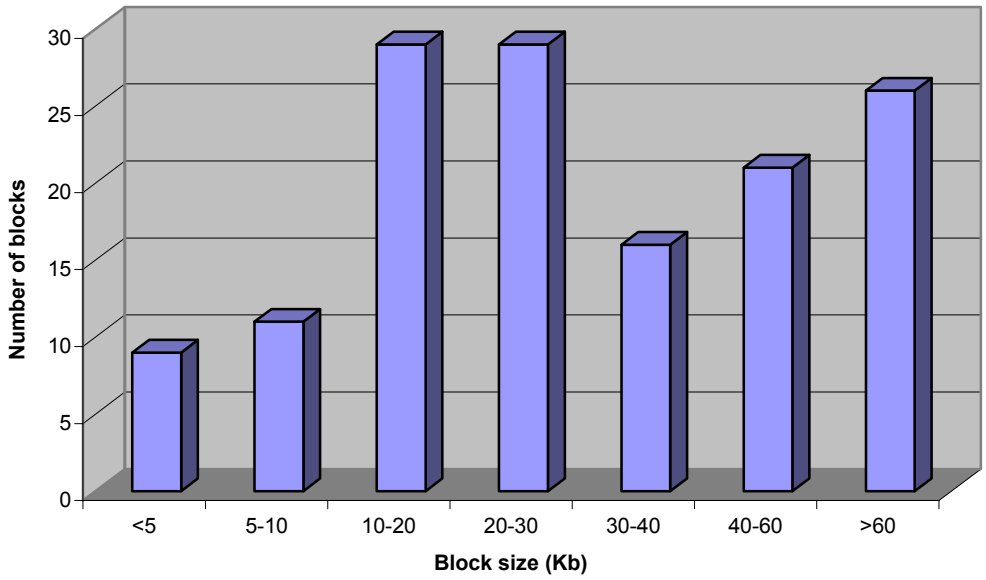


Figure 5.19: Correlation of "LD blocks" and SNPs. (A) "LD blocks" binned according to the number of SNPs in each block. (B) Proportion of all genome sequence spanned by blocks, binned according to the number of SNPs in each block.

Most of the LD blocks (65%) are more than 20 Kb long and 18% are more than 60 Kb long (Figure 5.20 A). Also, blocks over 40 Kb in size account for most (63%) sequence in blocks (Figure 5.20 B). The distribution of the various types of blocks is shown in

Figure 5.21 A, whereas the percentage of the total sequence in blocks across the region is shown in Figure 5.21 B.

A.



B.

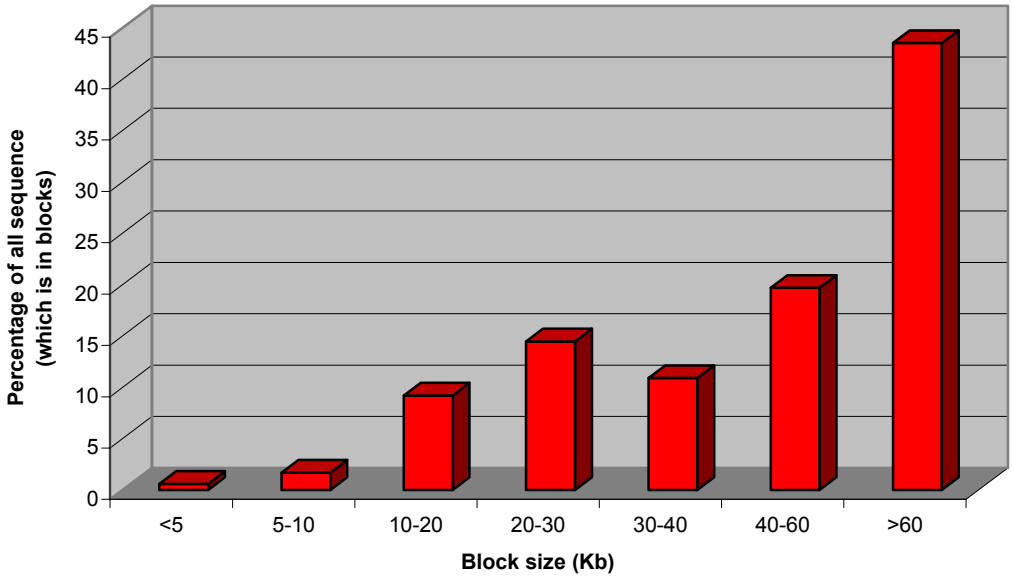
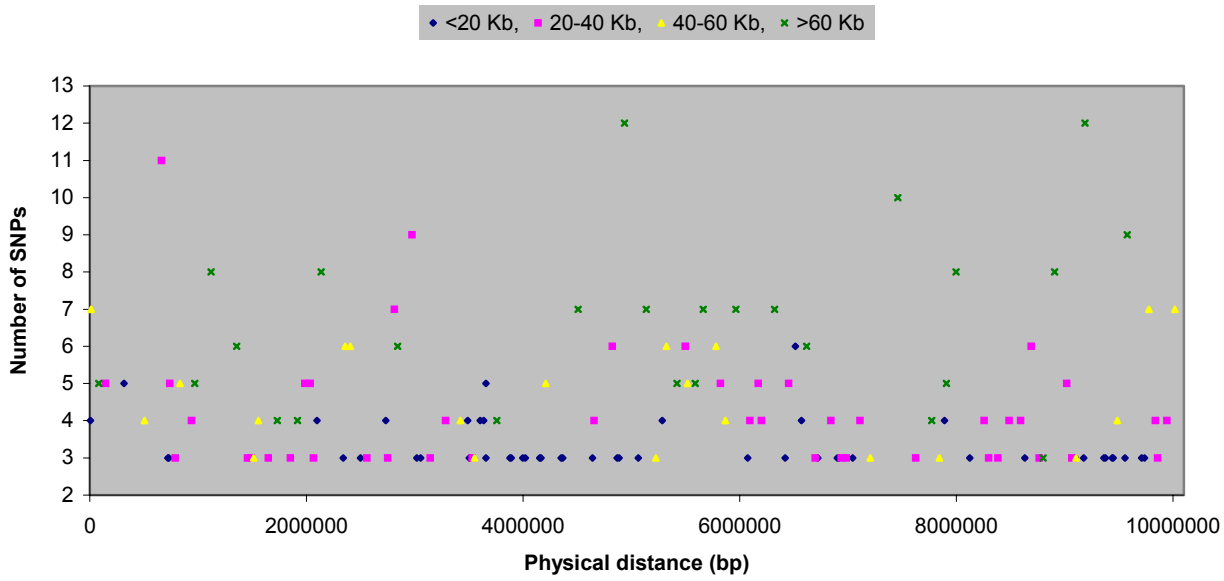


Figure 5.20: Size correlation of "LD blocks". (A) Size distribution of "LD blocks" (B) Proportion of all sequence in blocks binned according to the size of each block.

A.



B.

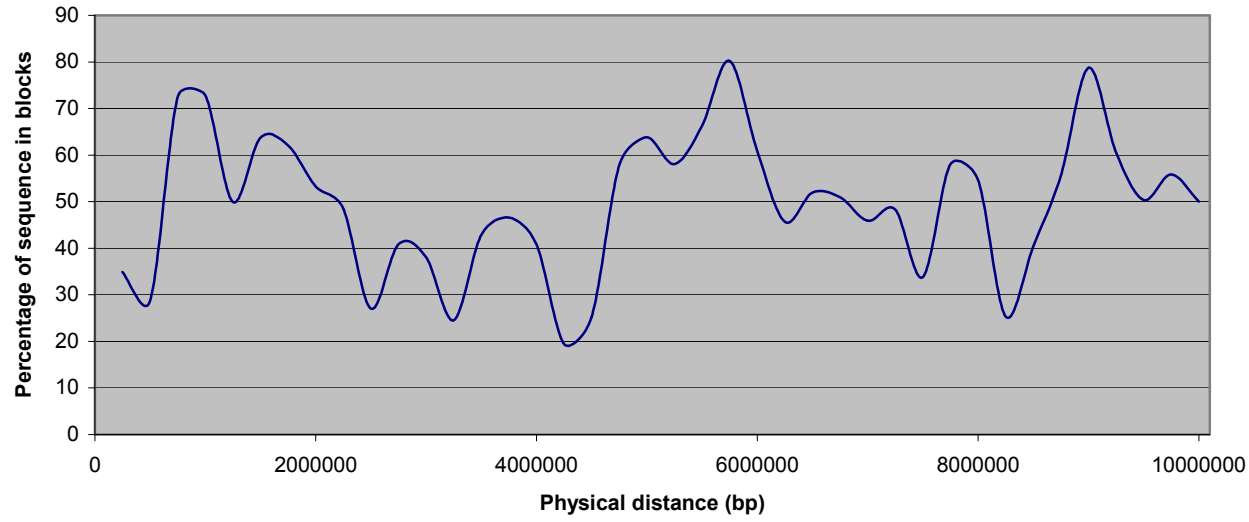


Figure 5.21: The distribution of "LD blocks" across 20q12-13.2. (A) The position of blocks across the region. Blocks were binned in four size ranges, each represented by a different type of data point. The number of SNPs in each block is also shown. (B) The percentage of all genomic sequence in blocks across the region (calculated as 500 Kb sliding windows, overlapping by 250 Kb).

5.5 Discussion

The available reference sequence of human chromosome 20 and expressed sequences were used to identify over 100 putative SNPs *in silico*, corresponding to 47 genes. This approach of SNP discovery is simple and inexpensive. In addition, it tackles exonic regions, which are more likely to harbour SNPs that are functionally important. The major disadvantage of this approach is that SNP discovery depends on the expression levels of genes: genes with lower expression levels will be represented by fewer ESTs, decreasing the chance of identifying SNPs. In addition, due to the 3' end bias of the EST databases, most identified SNPs reside within the 3' UTR of genes. Functionally important SNPs residing in non-transcribing genic regions (like promoters and introns) are also missed. Validation of the identified SNPs suggested that either there is a higher rate of false positives or many of these SNPs are rare (52% polymorphic in at least one population). Finally, the manual inspection of sequence reads and scoring of SNPs is a tedious process; thus for large-scale applications, an automated approach is required (as described for example in Deutsch *et al.*, 2001).

A set of 2,208 SNPs mapping across 20q1-13.2 were genotyped across 119 individuals from three populations, using the Sequenom MassEXTEND platform. In total, a non-redundant set of 188,307 genotype calls was obtained. Error checking using an independent platform (Illumina) suggests an error rate of 0.4%. However, 2-3% of assays are not robust and tend to inflate error rates in the raw data. Imbalanced allelic amplification is the most likely cause.

Approximately 50% of the SNPs with "complete" results were polymorphic in all three populations, whilst 12% were polymorphic in only one population. Overall, 76.5% of SNPs with "complete" results were polymorphic in at least one population.

The African American ethnic group represents a population that has very recently (in demographic terms) migrated from Africa. Whilst expected to be subject to the forces of admixture with the population in the new geographic region, studies found almost complete correlation between African and African American samples (Gabriel *et al*., 2002). For the purposes of this discussion, African Americans will be treated as Africans.

In agreement with previous studies (Frisse *et al*., 2001; Przeworski *et al*., 2000; Wall and Przeworski, 2000), this study has found higher levels of variation in African/African American populations relative to non-African populations and a skew in the frequency spectrum towards more common variants in non-African populations relative to African/African American populations. One possibility is that non-African populations experienced a phase of population size reduction, during which the rare variants were lost more quickly than the common ones. The deficit of less common variants outside Africa, combined with the fact that the non-African variation is a fraction of that found in Africa, is consistent with the "Out of Africa" model of modern human origins (Wall, 2001). This suggests that modern humans evolved in a small region in Africa 120,000-150,000 years ago, and from there they expanded and replaced existing hominid populations around the world (Stringer and Andrews, 1988).

The study shows that Caucasians and African Americans have more SNPs in common than the Asians with the African Americans. Assuming an "Out of Africa" model, this difference could imply that Asians and Caucasians may have arisen from separately

migrating populations. This is supported by the fact that the Asian and Caucasian groups share fewer polymorphisms with each other than either share with the African Americans.

Of course, even the most complex population history models that are currently available are likely to oversimplify the real history of human populations. Scenarios affecting the patterns of sequence variation could include population subdivision with a change in migration rates over time and admixture with archaic humans. Recent developments such as admixture, population growth and founder effects that have occurred in historical times are also likely to affect patterns of variation (Przeworski *et al*., 2000).

This study identified a set of 943 SNPs with minor allele frequencies of ≥5% in Caucasians. This set is already a useful resource for ongoing and future association studies for the common diseases linked to 20q12-13.2 (e.g. Graves, diabetes and obesity). Although the average distance between neighbouring SNPs is less than 11 Kb some larger gaps still remain despite repeated attempts to identify and verify polymorphic SNPs. This is due to the non-uniform distribution of public SNPs across the region. The new set of ~110,000 chromosome 20 SNPs identified recently at the Sanger Institute will allow the selection of SNPs in the remaining gaps. It is, however, clear that additional SNP discovery efforts are needed for tackling the whole genome.

The extent of LD across 20q12-13.2 was assessed in Caucasians using 879 polymorphic SNPs, for which at least 80% of the 95 possible genotypes were available (average 93%). Both the D' and r^2 measures were applied. The decay of average D' with distance is sensitive to the minor allele frequency cut off used. The inclusion of rare SNPs tends to

inflate D' estimates as opposed to the use of only common SNPs (MAF>20%). Little difference was seen between the MAF>10% only and MAF>20% only "half length of decay" (approximately 80 Kb in both cases). This is slightly elevated (by ~20 Kb), when compared to similar studies (Reich *et al*. (2001) and Dawson *et al*. (2002)).

The pattern of LD across the region (average D' and r^2 in 500 Kb sliding windows) shows clear fluctuation with areas of high and low LD. There is strong correlation between recombination rate and extent of LD; the same observation was made in the chromosome 22 study Dawson *et al*. (2002). It would be interesting to use a higher resolution genetic map like the one recently reported by deCODE (Kong *et al*., 2002) for a more detailed analysis.

DNA regions were defined as "LD blocks" if they harboured three or more polymorphic SNPs with D' values of more than 0.9 for all possible SNP pairs (stringent cut off). In total, 141 such blocks were identified, covering approximately 50% of the sequence. Sequence coverage is a minimum as blocks are likely to extend further with the addition of more SNPs. Block size varied enormously between 142.3 Kb (largest) and 2.3 Kb (smallest). As in the present study we did not include the number of common haplotypes per block (ongoing effort), we used the term "LD" instead of "haplotype" block. The long-range haplotype of each founder chromosome is available. The next steps beside the calculation of the number of common haplotypes per block are to increase the coverage of the region in blocks and extend the study to other populations.

This study contributes to the overall effort to understand the long-range organisation of LD across the human genome and provides a first generation LD map as a tool for association studies in 20q12-13.2.

Chapter VI

Discussion

6.1 Summary

This thesis has described structural, comparative and human variation studies for a 10 Mb region on human chromosome 20q12-13.2. The sequence features of the region were investigated and a detailed transcript map was produced. The syntenic mouse region was mapped and sequenced and the generated data were used for systematic human:mouse comparative analyses. Expressed sequences were used to identify human exonic SNPs *in silico*. A set of 2,208 SNPs mapping across the region was used to obtain allele frequencies in three populations and generate a first generation linkage disequilibrium map of 20q12-13.2 in Caucasians.

6.2 Analysis of genomic sequence

The sequence data generated by the HGP is paving the way for the identification of the entire complement of human genes. As described in Chapter I, the vast amount of data produced has prompted the development of fully automated annotation systems, which at the moment have severe limitations. As a result, the various genome sequencing centres prefer to utilise the semi-automatic approach of computational analysis and manual annotation for their clone sequence output.

This thesis described the assembly of a high quality transcript map. Central to this was the availability of a contiguous, finished genomic sequence spanning the entire region.

Expressed sequences and *ab initio* predictions were manually inspected and gene structures were annotated only when supported by experimental evidence. Where necessary, cDNA isolation and sequencing were undertaken to verify exon-intron boundaries and extend the annotation of incomplete gene structures. Although arduous, this approach is necessary to ensure high levels of accuracy. Un-supported gene predictions were nevertheless stored in the 20ace database for future investigations.

In total, 99 coding genes, 30 putative genes and 36 pseudogenes were annotated, and their structural features including splice sites, alternative isoforms and polyadenylation signals were studied in detail. Predicted CpG islands, promoters and transcription start sites were then correlated with the above data suggesting that the structural annotation of most genes is complete. Furthermore, three species sequence comparisons were used to show that very few exons (<2%), and probably no genes remain un-annotated in this region.

The expression pattern of novel genes was experimentally investigated by screening cDNA libraries. Also, protein analysis of the translated ORF of all coding genes revealed that this region is enriched in genes encoding for proteins with particular domains. All generated data has been integrated with other sequence features such as repeats, segmental duplications and recombination in a single map of 20q12-13.2 to provide an advanced tool for future research in this region. In addition, this study can serve as guide of how to proceed with the annotation of the rest of the genome.

6.3 Mouse genomics

Comparative mapping and sequencing of human chromosome 20q12-13.2 in mouse has shown that the syntenic regions share conservation of gene order and content. The benefit of the early data release policy implemented by the Sanger Institute (and other public domain sequencing centres) was also illustrated, as a large amount of information was derived from unfinished mouse genomic sequence.

Even in a region previously subjected to extensive computational and experimental gene annotation, this approach contributes new, although very few, exons. Although no additional human genes were annotated in this study, conserved regions correlated well with gene annotation implying that mouse genomic sequence could provide a powerful tool for gene annotation. In particular, the mouse sequence can be used to identify conserved regions, corresponding to genes with spatially or temporally limited expression patterns. This possibility was examined in this region by identifying 28 human loci that show conservation and are supported by identical predictions from two prediction software. Testing a subset of these by PCR screening of cDNA libraries did not confirm any as being expressed. Confirming any of these regions as being expressed will be challenging and will probably require a high-throughput method. For example, the mouse sequence could be used to construct DNA chips to be used in hybridisation experiments with several mouse cDNA libraries from a large number of tissues.

It is worth noting that although the data from this study strongly suggest that all human genes of 20q12-13.2 are present in the mouse sequence, the total number of mouse genes has not been assessed.

Unlike coding genes, the putative genes were not conserved in the mouse sequence (discussed in Chapters III and IV). This study has demonstrated that these structures differ significantly from coding genes, and that their possible role remains elusive. Preliminary analysis of finished mouse sequence failed to identify any similar gene structures.

The relatively high extent of synteny between human and mouse across intra- and inter-genic regions does not allow the systematic identification of non-transcribed gene features. This could be resolved by performing sequence comparisons with the genome of another organism; the region of 20q12-13.2 could be an ideal candidate for a study of this type. As for the mouse, the available resources and annotation could provide an easy means for the rapid construction of comparative maps. In addition, the virtually complete gene annotation would enable the easy discrimination between transcribed gene elements (exons) and non-transcribed gene elements (e.g. promoters, enhancers).

6.4 Human variation and linkage disequilibrium

Unlike most previous studies, the LD map constructed by this study (chapter V) spans a large (~10 Mb), contiguous segment of the genome. The pairwise values for markers obtained using D' and r^2 confirmed the extensive variability of LD across the region, and

that average half-length LD (D'>0.5) extends to ~80 Kb. Inspection of the pattern of LD along the region revealed that tracts of high LD are situated in regions of low recombination, whereas tracts of low LD are situated in regions of high recombination. Stringent criteria were used to identify 141 "LD blocks", which cover approximately 50% of the region. Increasing the SNP coverage of the regions that are currently outside the defined blocks will extend the current blocks, as well as define other, smaller blocks that have been missed by this study. Overall, the generated data will provide the basis for determining the features of haplotype blocks across 20q12-13.2 and accelerate future association studies for common diseases linked to the region.

6.5 Conclusions and future work

This thesis focused on the structural and comparative analysis of 20q12-13.2. More work will be required to elucidate gene expression and function.

The systematic study of the expression profile of the annotated genes using microarray technology is an attractive option. Given that the sequence of the orthologous mouse genes is available, mouse DNA chips could also be used for detailed expression studies of the various developmental stages, as well as responses to all kinds of stimuli (heat/cold shock, starvation, irradiation, addition of compounds with known molecular targets etc) to provide data on cellular and molecular pathways. The DNA chip approach could also be used to test regions conserved between human and mouse, or regions supported by multiple exon predictions for their coding potential.

Our understanding of the region could also be expanded through comparative analysis with other mammals, for example, the dog. The generated data from such studies could shed light on the evolutionary history of the region. In addition, three-species sequence comparisons could streamline the identification of non-coding regulatory regions. Such regions could then be experimentally tested in a systematic fashion.

The analysis and annotation of the generated finished mouse sequence should also be pursued. The features of a detailed mouse map could be compared to the human orthologs and identify similarities and differences between the two species. Such a map would also be a priceless tool for functional studies, such as mouse knockouts.

Overall, the studies described above will provide the basis for further protein analyses. Combined with a refined haplotype map, these resources could lead to the identification of the genes associated with disorders and provide a strong foundation for understanding the molecular basis of the diseases linked to 20q12-13.2.

Chapter VII

References

7.1 References

Abecasis G. R., Noguchi E., Heinzmann A., Traherne J. A., Bhattacharyya S., Leaves N. I., Anderson G. G., Zhang Y., Lench N. J., Carey A., Cardon L. R., Moffatt M. F., and Cookson W. O. (2001). Extent and distribution of linkage disequilibrium in three genomic regions. *Am J Hum Genet* **68:** 191-197.

Adams M. D., Celniker S. E., Holt R. A., Evans C. A., Gocayne J. D., Amanatides P. G., Scherer S. E., Li P. W., Hoskins R. A., Galle R. F., George R. A., Lewis S. E., Richards S., Ashburner M., Henderson S. N., Sutton G. G., Wortman J. R., Yandell M. D., Zhang Q., Chen L. X., Brandon R. C., Rogers Y. H., Blazej R. G., Champe M., Pfeiffer B. D., Wan K. H., Doyle C., Baxter E. G., Helt G., Nelson C. R., Gabor G. L., Abril J. F., Agbayani A., An H. J., Andrews-Pfannkoch C., Baldwin D., Ballew R. M., Basu A., Baxendale J., Bayraktaroglu L., Beasley E. M., Beeson K. Y., Benos P. V., Berman B. P., Bhandari D., Bolshakov S., Borkova D., Botchan M. R., Bouck J., Brokstein P., Brottier P., Burtis K. C., Busam D. A., Butler H., Cadieu E., Center A., Chandra I., Cherry J. M., Cawley S., Dahlke C., Davenport L. B., Davies P., de Pablos B., Delcher A., Deng Z., Mays A. D., Dew I., Dietz S. M., Dodson K., Doup L. E., Downes M., Dugan-Rocha S., Dunkov B. C., Dunn P., Durbin K. J., Evangelista C. C., Ferraz C., Ferriera S., Fleischmann W., Fosler C., Gabrielian A. E., Garg N. S., Gelbart W. M., Glasser K., Glodek A., Gong F., Gorrell J. H., Gu Z., Guan P., Harris M., Harris N. L., Harvey D., Heiman T. J., Hernandez J. R., Houck J., Hostin D., Houston K. A., Howland T. J., Wei M. H., Ibegwam C., *et al*. (2000). The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185-95.

Adams M. D., Kelley J. M., Gocayne J. D., Dubnick M., Polymeropoulos M. H., Xiao H., Merril C. R., Wu A., Olde B., Moreno R. F., and *et al*. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252:** 1651-6.

Ahringer J. (1997). Turn to the worm! *Curr Opin Genet Dev* **7:** 410-5.

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

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

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

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

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

Aparicio S., and Brenner S. (1997). How good a model is the Fugu genome? *Nature* **387:** 140.

Aparicio S. A. J. R. (2000). How to count...human genes. *Nat. Genet.* **25:** 129-130.

Apweiler R., Attwood T. K., Bairoch A., Bateman A., Birney E., Biswas M., Bucher P., Cerutti L., Corpet F., Croning M. D., Durbin R., Falquet L., Fleischmann W., Gouzy J., Hermjakob H., Hulo N., Jonassen I., Kahn D., Kanapin A., Karavidopoulou Y., Lopez R., Marx B., Mulder N. J., Oinn T. M., Pagni M., Servant F., Sigrist C. J., and Zdobnov E. M. (2000). InterPro-an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* **16:** 1145-50.

Ardlie K. G., Kruglyak L., and Seielstad M. (2002). Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* **3:** 299-309.

Arnold C., and Hodgson I. J. (1991). Vectorette PCR: a novel approach to genomic walking. *PCR Methods Appl* **1:** 39-42.

Bairoch A., and Apweiler R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* **28:** 45-8.

Baldi P., and Baisnee P. F. (2000). Sequence analysis by additive scales: DNA structure for sequences and repeats of all lengths. *Bioinformatics* **16:** 865-89.

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

Barbesino G., Tomer Y., Concepcion E., Davies T. F., and Greenberg D. A. (1998). Linkage analysis of candidate genes in autoimmune thyroid disease: 1. Selected immunoregulatory genes. International Consortium for the Genetics of Autoimmune Thyroid Disease. *J Clin Endocrinol Metab* **83:** 1580-4.

Barbesino G., Tomer Y., Concepcion E. S., Davies T. F., and Greenberg D. A. (1998). Linkage analysis of candidate genes in autoimmune thyroid disease. II. Selected gender-related genes and the X-chromosome. International Consortium for the Genetics of Autoimmune Thyroid Disease. *J Clin Endocrinol Metab* **83:** 3290-5.

Bargmann C. I. (2001). High-throughput reverse genetics: RNAi screens in *Caenorhabditis elegans*. *Genome Biol* **2:** REVIEWS1005.

Barnes M. R. (2002). SNP and mutation data on the Web - hidden treasures for uncovering. *Comparative and Functional Genomics* **3:** 67-74.

Bassett D. E., Jr., Basrai M. A., Connelly C., Hyland K. M., Kitagawa K., Mayer M. L., Morrow D. M., Page A. M., Resto V. A., Skibbens R. V., and Hieter P. (1996). Exploiting the complete yeast genome sequence. *Curr Opin Genet Dev* **6:** 763-6.

Batzer M. A., Deininger P. L., Hellmann-Blumberg U., Jurka J., Labuda D., Rubin C. M., Schmid C. W., Zietkiewicz E., and Zuckerkandl E. (1996). Standardized nomenclature for Alu repeats. *J Mol Evol* **42:** 3-6.

Batzoglou S., Pachter L., Mesirov J. P., Berger B., and Lander E. S. (2000). Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res* **10:** 950-8.

Baudin A., Ozier-Kalogeropoulos O., Denouel A., Lacroute F., and Cullin C. (1993). A simple and efficient method for direct gene deletion in *Saccharomyces cerevisiae*. *Nucleic Acids Res* **21:** 3329-30.

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

Bench A. J., Aldred M. A., Humphray S. J., Champion K. M., Gilbert J. G., Asimakopoulos F. A., Deloukas P., Gwilliam R., Bentley D. R., and Green A. R. (1998). A detailed physical and transcriptional map of the region of chromosome 20 that is deleted in myeloproliferative disorders and refinement of the common deleted region. *Genomics* **49:** 351-62.

Bench A. J., Nacheva E. P., Hood T. L., Holden J. L., French L., Swanton S., Champion K. M., Li J., Whittaker P., Stavrides G., Hunt A. R., Huntly B. J., Campbell L. J., Bentley D. R., Deloukas P., and Green A. R. (2000). Chromosome 20 deletions in myeloid malignancies: reduction of the common deleted region, generation of a PAC/BAC contig and identification of candidate genes. UK Cancer Cytogenetics Group (UKCCG). *Oncogene* **19:** 3902-13.

Benson D. A., Karsch-Mizrachi I., Lipman D. J., Ostell J., Rapp B. A., and Wheeler D. L. (2002). GenBank. *Nucleic Acids Res* **30:** 17-20.

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

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., Scott C. E., Ashcroft K. J., Atkinson A. L., Aubin K., Beare D. M., Bethel G., Brady N., Brook J. C., Burford D. C., Burrill W. D., Burrows C., Butler A. P., Carder C., Catanese J. J., Clee C. M., Clegg S. M., Cobley V., Coffey A. J., Cole C. G., Collins J. E., Conquer J. S., Cooper R. A., Culley K. M., Dawson E., Dearden F. L., Durbin R. M., de Jong P. J., Dhami P. D., Earthrowl M. E., Edwards C. A., Evans R. S., Gillson C. J., Ghori J., Green L., Gwilliam R., Halls K. S., Hammond S., Harper G. L., Heathcott R. W., Holden J. L., Holloway E., Hopkins B. L., Howard P. J., Howell G. R., Huckle E. J., Hughes J., Hunt P. J., Hunt S. E., Izmajlowicz M., Jones C. A., Joseph S. S., Laird G., Langford C. F., Lehvaslaiho M. H., Leversha M. A., McCann O. T., McDonald L. M., McDowall J., Maslen G. L., Mistry D., Moschonas N. K., Neocleous V., Pearson D. M., Phillips K. J., Porter K. M., Prathalingam S. R., Ramsey Y. H., Ranby S. A., Rice C. M., Rogers J., Rogers L. J., Sarafidou T., Scott D. J., Sharp G. J., Shaw-Smith C. J., Smink L. J., Soderlund C., Sotheran E. C., Steingruber H. E., Sulston J. E., Taylor A., Taylor R. G., Thorpe A. A., Tinsley E., Warry G. L., Whittaker A., Whittaker P., Williams S. H., Wilmer T. E., Wooster R., *et al*. (2001). The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. *Nature* **409:** 942-3.

Bernardi G. (1989). The isochore organization of the human genome. *Annu Rev Genet* **23:** 637-61.

Bernardi G., Olofsson B., Filipski J., Zerial M., Salinas J., Cuny G., Meunier-Rotival M., and Rodier F. (1985). The mosaic genome of warm-blooded vertebrates. *Science* **228:** 953-8.

Bernot A., Heilig R., Clepet C., Smaoui N., Da Silva C., Petit J. L., Devaud C., Chiannilkulchai N., Fizames C., Samson D., Cruaud C., Caloustian C., Gyapay G., Delpech M., and Weissenbach J. (1998). A transcriptional Map of the FMF region. *Genomics* **50:** 147-60.

Bertina R. M., Koeleman B. P., Koster T., Rosendaal F. R., Dirven R. J., de Ronde H., van der Velden P. A., and Reitsma P. H. (1994). Mutation in blood coagulation factor V associated with resistance to activated protein C. *Nature* **369:** 64-7.

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

Bihoreau M. T., Gauguier D., Kato N., Hyne G., Lindpaintner K., Rapp J. P., James M. R., and Lathrop G. M. (1997). A linkage map of the rat genome derived from three F2 crosses. *Genome Res* **7:** 434-40.

Bihoreau M. T., Sebag-Montefiore L., Godfrey R. F., Wallis R. H., Brown J. H., Danoy P. A., Collins S. C., Rouard M., Kaisaki P. J., Lathrop M., and Gauguier D. (2001). A high-resolution consensus linkage map of the rat, integrating radiation hybrid and genetic maps. *Genomics* **75:** 57-69.

Birchall P. S., Fishpool R. M., and Albertson D. G. (1995). Expression patterns of predicted genes from the *C. elegans* genome sequence visualized by FISH in whole organisms. *Nat Genet* **11:** 314-20.

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

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

Blackstock W. P., and Weir M. P. (1999). Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol* **17:** 121-7.

Blattner F. R., Plunkett G., 3rd, Bloch C. A., Perna N. T., Burland V., Riley M., Collado-Vides J., Glasner J. D., Rode C. K., Mayhew G. F., Gregor J., Davis N. W., Kirkpatrick H. A., Goeden M. A., Rose D. J., Mau B., and Shao Y. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* **277:** 1453-74.

Blechschmidt K., Schweiger M., Wertz K., Poulson R., Christensen H. M., Rosenthal A., Lehrach H., and Yaspo M. L. (1999). The mouse Aire gene: comparative genomic sequencing, gene organization, and expression. *Genome Res* **9:** 158-66.

Boehnke M. (2000). A look at linkage disequilibrium. *Nat Genet* **25:** 246-7.

Boguski M. S., Lowe T. M., and Tolstoshev C. M. (1993). dbEST-database for expressed sequence tags. *Nat Genet* **4:** 332-3.

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

Bowden D. W., Sale M., Howard T. D., Qadri A., Spray B. J., Rothschild C. B., Akots G., Rich S. S., and Freedman B. I. (1997). Linkage of genetic markers on human chromosomes 20 and 12 to NIDDM in Caucasian sib pairs with a history of diabetic nephropathy. *Diabetes* **46:** 882-6.

Braun A., Little D. P., Reuter D., Muller-Mysok B., and Koster H. (1997a). Improved analysis of microsatellites using mass spectrometry. *Genomics* **46:** 18-23.

Braun A., Little D. P., and Koster H. (1997b). Detecting CFTR gene mutations by using primer oligo base extension and mass spectrometry. *Clin Chem* **43:** 1151-8.

Brenner S., Elgar G., Sandford R., Macrae A., Venkatesh B., and Aparicio S. (1993). Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome. *Nature* **366:** 265-8.

Brett D., Hanke J., Lehmann G., Haase S., Delbruck S., Krueger S., Reich J., and Bork P. (2000). EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett* **474:** 83-6.

Brickner A. G., Koop B. F., Aronow B. J., and Wiginton D. A. (1999). Genomic sequence comparison of the human and mouse adenosine deaminase gene regions. *Mamm Genome* **10:** 95-101.

Brookes A. J. (1999). The essence of SNPs. *Gene* **234:** 177-86.

Brookes A. J., Lehvaslaiho H., Siegfried M., Boehm J. G., Yuan Y. P., Sarkar C. M., Bork P., and Ortigao F. (2000). HGBASE: a database of SNPs and other variations in and around human genes. *Nucleic Acids Res* **28:** 356-60.

Brüls T., Gyapay G., Petit J. L., Artiguenave F., Vico V., Qin S., Tin-Wollam A. M., Da Silva C., Muselet D., Mavel D., Pelletier E., Levy M., Fujiyama A., Matsuda F., Wilson R., Rowen L., Hood L., Weissenbach J., Saurin W., and Heilig R. (2001). A physical map of human chromosome 14. *Nature* **409:** 947-8.

Buckler A. J., Chang D. D., Graw S. L., Brook J. D., Haber D. A., Sharp P. A., and Housman D. E. (1991). Exon amplification: a strategy to isolate mammalian genes based on RNA splicing. *Proc Natl Acad Sci U S A* **88:** 4005-9.

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

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

Burns N., Grimwade B., Ross-Macdonald P. B., Choi E. Y., Finberg K., Roeder G. S., and Snyder M. (1994). Large-scale analysis of gene expression, protein localization, and gene disruption in *Saccharomyces cerevisiae*. *Genes Dev* **8:** 1087-105.

Burset M., and Guigo R. (1996). Evaluation of gene structure prediction programs. *Genomics* **34:** 353-67.

Burset M., Seledtsov I. A., and Solovyev V. V. (2000). Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res* **28:** 4364-75.

Cai W. W., Chow C. W., Damani S., Gregory S. G., Marra M., and Bradley A. (2001). An SSLP marker-anchored BAC framework map of the mouse genome. *Nat Genet* **29:** 133-4.

Carlson C. S., Newman T. L., and Nickerson D. A. (2001). SNPing in the human genome. *Curr Opin Chem Biol* **5:** 78-85.

Carver E. A., and Stubbs L. (1997). Zooming in on the human-mouse comparative map: genome conservation re-examined on a high-resolution scale. *Genome Res* **7:** 1123-37.

Centola M., Chen X., Sood R., Deng Z., Aksentijevich I., Blake T., Ricke D. O., Wood G., Zaks N., Richards N., Krizman D., Mansfield E., Apostolou S., Liu J., Shafran N., Vedula A., Hamon M., Cercek A., Kahan T., Gumucio D., Callen D. F., Richards R. I., Moyzis R. K., Kastner D. L., and *et al*. (1998). Construction of an approximately 700-kb transcript map around the familial Mediterranean fever locus on human chromosome 16p13.3. *Genome Res* **8:** 1172-91.

Chakravarti A. (1999). Population genetics-making sense out of sequence. *Nat Genet* **21:** 56-60.

Chakravarti A. (2001). To a future of genetic medicine. *Nature* **409:** 822-3.

Chambon-Pautas C., Cave H., Gerard B., Guidal-Giroux C., Duval M., Vilmer E., and Grandchamp B. (1998). High-resolution allelotype analysis of childhood B-lineage acute lymphoblastic leukemia. *Leukemia* **12:** 1107-13.

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

Chen J., Greenblatt I. M., and Dellaporta S. L. (1992). Molecular analysis of Ac transposition and DNA replication. *Genetics* **130:** 665-76.

Chiaromonte F., Yang S., Elnitski L., Yap V. B., Miller W., and Hardison R. C. (2001). Association between divergence and interspersed repeats in mammalian noncoding genomic DNA. *Proc Natl Acad Sci U S A* **98:** 14503-8.

Chiaromonte F., Yap V. B., and Miller W. (2002). Scoring pairwise genomic sequence alignments. *Pac Symp Biocomput***:** 115-26.

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

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

Church D. M., Banks L. T., Rogers A. C., Graw S. L., Housman D. E., Gusella J. F., and Buckler A. J. (1993). Identification of human chromosome 9 specific genes using exon amplification. *Hum Mol Genet* **2:** 1915-20.

Claverie J. M. (1997). Computational methods for the identification of genes in vertebrate genomic sequences. *Hum Mol Genet* **6:** 1735-44.

Claverie J. M. (2001). Gene number. What if there are only 30,000 human genes? *Science* **291:** 1255-7.

Colgan D. F., and Manley J. L. (1997). Mechanism and regulation of mRNA polyadenylation. *Genes Dev* **11:** 2755-66.

Collinge J., Sidle K. C., Meads J., Ironside J., and Hill A. F. (1996). Molecular analysis of prion strain variation and the aetiology of 'new variant' CJD. *Nature* **383:** 685-90.

Collins F., and Galas D. (1993). A new five-year plan for the U.S. Human Genome Project. *Science* **262:** 43-6.

Collins F. S. (1995). Positional cloning moves from perditional to traditional. *Nat Genet* **9:** 347-50.

Collins F. S. (2001). Contemplating the end of the beginning. *Genome Res* **11:** 641-3.

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

Collins F. S., and Guttmacher A. E. (2001). Genetics moves into the medical mainstream. *Jama* **286:** 2322-4.

Collins F. S., Guyer M. S., and Charkravarti A. (1997). Variations on a theme: cataloging human DNA sequence variation. *Science* **278:** 1580-1.

Collins F. S., Patrinos A., Jordan E., Chakravarti A., Gesteland R., and Walters L. (1998a). New goals for the U.S. Human Genome Project: 1998-2003. *Science* **282:** 682-9.

Collins F. S., and Weissman S. M. (1984). The molecular genetics of human hemoglobin. *Prog Nucleic Acid Res Mol Biol* **31:** 315-462.

Collins J., and Hohn B. (1978). Cosmids: a type of plasmid gene-cloning vector that is packageable in vitro in bacteriophage lambda heads. *Proc Natl Acad Sci U S A* **75:** 4242-6.

Collins J., Saari B., and Anderson P. (1987). Activation of a transposable element in the germ line but not the soma of *Caenorhabditis elegans*. *Nature* **328:** 726-8.

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

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

Couque N., Chambon-Pautas C., Cave H., Bardet V., Duval M., Vilmer E., and Grandchamp B. (1999). Mapping of chromosome 20 for loss of heterozygosity in childhood ALL reveals a 1,000-kb deletion in one patient. *Leukemia* **13:** 1972-4.

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

Crabtree J., Wiltshire T., Brunk B., Zhao S., Schug J., Stoeckert C. J., Jr., and Bucan M. (2001). High-resolution BAC-based map of the central portion of mouse chromosome 5. *Genome Res* **11:** 1746-57.

Craig J. M., and Bickmore W. A. (1993). Chromosome bands--flavours to savour. *Bioessays* **15:** 349-54.

Crnogorac-Jurcevic T., Brown J. R., Lehrach H., and Schalkwyk L. C. (1997). Tetraodon fluviatilis, a new puffer fish model for genome studies. *Genomics* **41:** 177-84.

Crollius H. R., Jaillon O., Bernot A., Dasilva C., Bouneau L., Fischer C., Fizames C., Wincker P., Brottier P., Quetier F., Saurin W., and Weissenbach J. (2000a). Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nat Genet* **25:** 235-8.

Crollius H. R., Jaillon O., Dasilva C., Ozouf-Costaz C., Fizames C., Fischer C., Bouneau L., Billault A., Quetier F., Saurin W., Bernot A., and Weissenbach J. (2000b). Characterization and repeat analysis of the compact genome of the freshwater pufferfish *Tetraodon nigroviridis*. *Genome Res* **10:** 939-49.

Cross S. H., and Bird A. P. (1995). CpG islands and genes. *Curr Opin Genet Dev* **5:** 309-14.

Cross S. H., Charlton J. A., Nan X., and Bird A. P. (1994). Purification of CpG islands using a methylated DNA binding column. *Nat Genet* **6:** 236-44.

Cross S. H., Clark V. H., and Bird A. P. (1999). Isolation of CpG islands from large genomic clones. *Nucleic Acids Res* **27:** 2099-107.

Cross S. H., Clark V. H., Simmen M. W., Bickmore W. A., Maroon H., Langford C. F., Carter N. P., and Bird A. P. (2000). CpG island libraries from human chromosomes 18 and 22: landmarks for novel genes. *Mamm Genome* **11:** 373-83.

Davignon J., Gregg R. E., and Sing C. F. (1988). Apolipoprotein E polymorphism and atherosclerosis. *Arteriosclerosis* **8:** 1-21.

Davuluri R. V., Grosse I., and Zhang M. Q. (2001). Computational identification of promoters and first exons in the human genome. *Nat Genet* **29:** 412-7.

Dawson E., Abecasis G. R., Bumpstead S., Chen Y., Hunt S., Beare D. M., Pabial J., Dibling T., Tinsley E., Kirby S., Carter D., Papaspyridonos M., Livingstone S., Ganske R., Lohmussaar E., Zernant J., Tonisson N., Remm M., Magi R., Puurand T., Vilo J., Kurg A., Rice K., Deloukas P., Mott R., Metspalu A., Bentley D. R., Cardon L. R., and Dunham I. (2002). A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **10:** 544-8.

Dawson E., Chen Y., Hunt S., Smink L. J., Hunt A., Rice K., Livingston S., Bumpstead S., Bruskiewich R., Sham P., Ganske R., Adams M., Kawasaki K., Shimizu N., Minoshima S., Roe B., Bentley D., and Dunham I. (2001). A SNP resource for human chromosome 22: extracting dense clusters of SNPs from the genomic sequence. *Genome Res* **11:** 170-8.

Deak P., Omar M. M., Saunders R. D., Pal M., Komonyi O., Szidonya J., Maroy P., Zhang Y., Ashburner M., Benos P., Savakis C., Siden-Kiamos I., Louis C., Bolshakov V. N., Kafatos F. C., Madueno E., Modolell J., and Glover D. M.

(1997). P-element insertion alleles of essential genes on the third chromosome of Drosophila melanogaster: correlation of physical and cytogenetic maps in chromosomal region 86E-87F. *Genetics* **147:** 1697-722.

DeBerardinis R. J., Goodier J. L., Ostertag E. M., and Kazazian H. H., Jr. (1998). Rapid amplification of a retrotransposon subfamily is evolving the mouse genome. *Nat Genet* **20:** 288-90.

DeBry R. W., and Seldin M. F. (1996). Human/mouse homology relationships. *Genomics* **33:** 337-51.

Dehal P., Predki P., Olsen A. S., Kobayashi A., Folta P., Lucas S., Land M., Terry A., Ecale Zhou C. L., Rash S., Zhang Q., Gordon L., Kim J., Elkin C., Pollard M. J., Richardson P., Rokhsar D., Uberbacher E., Hawkins T., Branscomb E., and Stubbs L. (2001). Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science* **293:** 104-11.

Deloukas P., Matthews L. H., Ashurst J., Burton J., Gilbert J. G., Jones M., Stavrides G., Almeida J. P., Babbage A. K., Bagguley C. L., Bailey J., Barlow K. F., Bates K. N., Beard L. M., Beare D. M., Beasley O. P., Bird C. P., Blakey S. E., Bridgeman A. M., Brown A. J., Buck D., Burrill W., Butler A. P., Carder C., Carter N. P., Chapman J. C., Clamp M., Clark G., Clark L. N., Clark S. Y., Clee C. M., Clegg S., Cobley V. E., Collier R. E., Connor R., Corby N. R., Coulson A., Coville G. J., Deadman R., Dhami P., Dunn M., Ellington A. G., Frankland J. A., Fraser A., French L., Garner P., Grafham D. V., Griffiths C., Griffiths M. N., Gwilliam R., Hall R. E., Hammond S., Harley J. L., Heath P. D., Ho S., Holden J. L., Howden P. J., Huckle E., Hunt A. R., Hunt S. E., Jekosch K., Johnson C. M., Johnson D., Kay M. P., Kimberley A. M., King A., Knights A., Laird G. K., Lawlor S., Lehvaslaiho M. H., Leversha M., Lloyd C., Lloyd D. M., Lovell J. D., Marsh V. L., Martin S. L., McConnachie L. J., McLay K., McMurray A. A., Milne S., Mistry D., Moore M. J., Mullikin J. C., Nickerson T., Oliver K., Parker A., Patel R., Pearce T. A., Peck A. I., Phillimore B. J., Prathalingam S. R., Plumb R. W., Ramsay H., Rice C. M., Ross M. T., Scott C. E., Sehra H. K., Shownkeen R., Sims S., Skuce C. D., *et al*. (2001). The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414:** 865-71.

Deloukas P., Schuler G. D., Gyapay G., Beasley E. M., Soderlund C., Rodriguez-Tome P., Hui L., Matise T. C., McKusick K. B., Beckmann J. S., Bentolila S., Bihoreau M., Birren B. B., Browne J., Butler A., Castle A. B., Chiannilkulchai N., Clee C., Day P. J., Dehejia A., Dibling T., Drouot N., Duprat S., Fizames C., Bentley D. R., and *et al*. (1998). A physical map of 30,000 human genes. *Science* **282:** 744-6.

DeRisi J. L., Iyer V. R., and Brown P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278:** 680-6.

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

Devaux F., Marc P., and Jacq C. (2001). Transcriptomes, transcription activators and microarrays. *FEBS Lett* **498:** 140-4.

Devlin B., and Risch N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29:** 311-22.

Dib C., Faure S., Fizames C., Samson D., Drouot N., Vignal A., Millasseau P., Marc S., Hazan J., Seboun E., Lathrop M., Gyapay G., Morissette J., and Weissenbach J. (1996). A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380:** 152-4.

Dietrich W. F., Copeland N. G., Gilbert D. J., Miller J. C., Jenkins N. A., and Lander E. S. (1995). Mapping the mouse genome: current status and future prospects. *Proc Natl Acad Sci U S A* **92:** 10849-53.

Dietrich W. F., Miller J., Steen R., Merchant M. A., Damron-Boles D., Husain Z., Dredge R., Daly M. J., Ingalls K. A., O'Connor T. J., and *et al*. (1996). A comprehensive genetic map of the mouse genome. *Nature* **380:** 149-52.

Dillon N., and Sabbattini P. (2000). Functional gene expression domains: defining the functional unit of eukaryotic gene regulation. *Bioessays* **22:** 657-65.

Donis-Keller H., Green P., Helms C., Cartinhour S., Weiffenbach B., Stephens K., Keith T. P., Bowden D. W., Smith D. R., Lander E. S., and *et al*. (1987). A genetic linkage map of the human genome. *Cell* **51:** 319-37.

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

Driever W., Solnica-Krezel L., Schier A. F., Neuhauss S. C., Malicki J., Stemple D. L., Stainier D. Y., Zwartkruis F., Abdelilah S., Rangini Z., Belak J., and Boggs C. (1996). A genetic screen for mutations affecting embryogenesis in zebrafish. *Development* **123:** 37-46.

Driever W., Stemple D., Schier A., and Solnica-Krezel L. (1994). Zebrafish: genetic tools for studying vertebrate development. *Trends Genet* **10:** 152-9.

Dunham I. (2000). The gene guessing game. *Yeast* **17:** 218-24.

Dunham I., Shimizu N., Roe B. A., Chissoe S., Hunt A. R., Collins J. E., Bruskiewich R., Beare D. M., Clamp M., Smink L. J., Ainscough R., Almeida J. P., Babbage A., Bagguley C., Bailey J., Barlow K., Bates K. N., Beasley O., Bird C. P., Blakey S., Bridgeman A. M., Buck D., Burgess J., Burrill W. D., O'Brien K. P., and *et al*. (1999). The DNA sequence of human chromosome 22. *Nature* **402:** 489-95.

Dunning A. M., Durocher F., Healey C. S., Teare M. D., McBride S. E., Carlomagno F., Xu C. F., Dawson E., Rhodes S., Ueda S., Lai E., Luben R. N., Van Rensburg E. J., Mannermaa A., Kataja V., Rennart G., Dunham I., Purvis I., Easton D., and Ponder B. A. (2000). The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am J Hum Genet* **67:** 1544-54.

Durbin R., and Thierry-Mieg J. (1994). The ACEDB genome database. *Proceedings of the International Symposium on Computational Methods in Genome Research (1992; Heidelberg, Germany)*.

Dutt M. J., and Lee K. H. (2000). Proteomic analysis. *Curr Opin Biotechnol* **11:** 176-9.

Duyk G. M., Kim S. W., Myers R. M., and Cox D. R. (1990). Exon trapping: a genetic screen to identify candidate transcribed sequences in cloned mammalian genomic DNA. *Proc Natl Acad Sci U S A* **87:** 8995-9.

Elgar G., Clark M., Green A., and Sandford R. (1997). How good a model is the Fugu genome? *Nature* **387:** 140.

Elgar G., Sandford R., Aparicio S., Macrae A., Venkatesh B., and Brenner S. (1996). Small is beautiful: comparative genomics with the pufferfish (Fugu rubripes). *Trends Genet* **12:** 145-50.

Elvin P., Slynn G., Black D., Graham A., Butler R., Riley J., Anand R., and Markham A. F. (1990). Isolation of cDNA clones using yeast artificial chromosome probes. *Nucleic Acids Res* **18:** 3913-7.

Engels W. R., Johnson-Schlitz D. M., Eggleston W. B., and Sved J. (1990). High-frequency P element loss in Drosophila is homolog dependent. *Cell* **62:** 515-25.

Eppig J. T. (1996). Comparative maps: adding pieces to the mammalian jigsaw puzzle. *Curr Opin Genet Dev* **6:** 723-30.

Eppig J. T., and Nadeau J. H. (1995). Comparative maps: the mammalian jigsaw puzzle. *Curr Opin Genet Dev* **5:** 709-16.

Esnault C., Maestre J., and Heidmann T. (2000). Human LINE retrotransposons generate processed pseudogenes. *Nat Genet* **24:** 363-7.

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

Fei Z., Ono T., and Smith L. M. (1998). MALDI-TOF mass spectrometric typing of single nucleotide polymorphisms with mass-tagged ddNTPs. *Nucleic Acids Res* **26:** 2827-8.

Fenn J. B., Mann M., Meng C. K., Wong S. F., and Whitehouse C. M. (1989). Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246:** 64-71.

Fickett J. W., and Hatzigeorgiou A. G. (1997). Eukaryotic promoter recognition. *Genome Res* **7:** 861-78.

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

Fire A., Xu S., Montgomery M. K., Kostas S. A., Driver S. E., and Mello C. C. (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391:** 806-11.

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

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

Footz T. K., Brinkman-Mills P., Banting G. S., Maier S. A., Riazi M. A., Bridgland L., Hu S., Birren B., Minoshima S., Shimizu N., Pan H., Nguyen T., Fang F., Fu Y., Ray L., Wu H., Shaull S., Phan S., Yao Z., Chen F., Huan A., Hu P., Wang Q., Loh P., Qi S., Roe B. A., and McDermid H. E. (2001). Analysis of the cat eye syndrome critical region in humans and the region of conserved

synteny in mice: a search for candidate genes at or near the human chromosome 22 pericentromere. *Genome Res* **11:** 1053-70.

Fraser P., and Grosveld F. (1998). Locus control regions, chromatin activation and transcription. *Curr Opin Cell Biol* **10:** 361-5.

Frisse L., Hudson R. R., Bartoszewicz A., Wall J. D., Donfack J., and Di Rienzo A. (2001). Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am J Hum Genet* **69:** 831-43.

Fromont-Racine M., Rain J. C., and Legrain P. (1997). Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat Genet* **16:** 277-82.

Fu D. J., Tang K., Braun A., Reuter D., Darnhofer-Demar B., Little D. P., O'Donnell M. J., Cantor C. R., and Koster H. (1998). Sequencing exons 5 to 8 of the p53 gene by MALDI-TOF mass spectrometry. *Nat Biotechnol* **16:** 381-4.

Gabriel S. B., Schaffner S. F., Nguyen H., Moore J. M., Roy J., Blumenstiel B., Higgins J., DeFelice M., Lochner A., Faggart M., Liu-Cordero S. N., Rotimi C., Adeyemo A., Cooper R., Ward R., Lander E. S., Daly M. J., and Altshuler D. (2002). The structure of haplotype blocks in the human genome. *Science* **296:** 2225-9.

Gardiner K. (1995). Human genome organization. *Curr Opin Genet Dev* **5:** 315-22.

Gardiner K., and Mural R. J. (1995). Getting the message: identifying transcribed sequences. *Trends Genet* **11:** 77-9.

Gates M. A., Kim L., Egan E. S., Cardozo T., Sirotkin H. I., Dougan S. T., Lashkari D., Abagyan R., Schier A. F., and Talbot W. S. (1999). A genetic linkage map for zebrafish: comparative analysis and localization of genes and expressed sequences. *Genome Res* **9:** 334-47.

Gautheret D., Poirot O., Lopez F., Audic S., and Claverie J. M. (1998). Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. *Genome Res* **8:** 524-30.

Geisler R., Rauch G. J., Baier H., van Bebber F., Brobeta L., Dekens M. P., Finger K., Fricke C., Gates M. A., Geiger H., Geiger-Rudolph S., Gilmour D., Glaser S., Gnugge L., Habeck H., Hingst K., Holley S., Keenan J., Kirn A., Knaut H., Lashkari D., Maderspacher F., Martyn U., Neuhauss S., Haffter P., and *et al*. (1999). A radiation hybrid map of the zebrafish genome. *Nat Genet* **23:** 86-9.

Ghosh S., Watanabe R. M., Hauser E. R., Valle T., Magnuson V. L., Erdos M. R., Langefeld C. D., Balow J., Jr., Ally D. S., Kohtamaki K., Chines P., Birznieks G., Kaleta H. S., Musick A., Te C., Tannenbaum J., Eldridge W., Shapiro S., Martin C., Witt A., So A., Chang J., Shurtleff B., Porter R., Boehnke M., and *et al*. (1999). Type 2 diabetes: evidence for linkage on chromosome 20 in 716 Finnish affected sib pairs. *Proc Natl Acad Sci U S A* **96:** 2198-203.

Ghosh S., Watanabe R. M., Valle T. T., Hauser E. R., Magnuson V. L., Langefeld C. D., Ally D. S., Mohlke K. L., Silander K., Kohtamaki K., Chines P., Balow Jr J., Birznieks G., Chang J., Eldridge W., Erdos M. R., Karanjawala Z. E., Knapp J. I., Kudelko K., Martin C., Morales-Mena A., Musick A., Musick T., Pfahl C., Porter R., and Rayman J. B. (2000). The Finland-United States

investigation of non-insulin-dependent diabetes mellitus genetics (FUSION) study. I. An autosomal genome scan for genes that predispose to type 2 diabetes. *Am J Hum Genet* **67:** 1174-85.

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

Gilley J., Armes N., and Fried M. (1997). Fugu genome is not a good mammalian model. *Nature* **385:** 305-6.

Gilley J., and Fried M. (1999). Extensive gene order differences within regions of conserved synteny between the Fugu and human genomes: implications for chromosomal evolution and the cloning of disease genes. *Hum Mol Genet* **8:** 1313-20.

Goffeau A., Barrell B. G., Bussey H., Davis R. W., Dujon B., Feldmann H., Galibert F., Hoheisel J. D., Jacq C., Johnston M., Louis E. J., Mewes H. W., Murakami Y., Philippsen P., Tettelin H., and Oliver S. G. (1996). Life with 6000 genes. *Science* **274:** 546, 563-7.

Goldstein D. B. (2001). Islands of linkage disequilibrium. *Nat Genet* **29:** 109-11.

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

Göttgens B., Barton L. M., Gilbert J. G., Bench A. J., Sanchez M. J., Bahn S., Mistry S., Grafham D., McMurray A., Vaudin M., Amaya E., Bentley D. R., Green A. R., and Sinclair A. M. (2000). Analysis of vertebrate SCL loci identifies conserved enhancers. *Nat Biotechnol* **18:** 181-6.

Göttgens B., Gilbert J. G., Barton L. M., Grafham D., Rogers J., Bentley D. R., and Green A. R. (2001). Long-range comparison of human and mouse SCL loci: localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences. *Genome Res* **11:** 87-97.

Gough S. C. (2000). The genetics of Graves' disease. *Endocrinol Metab Clin North Am* **29:** 255-66.

Gray I. C., Campbell D. A., and Spurr N. K. (2000). Single nucleotide polymorphisms as tools in human genetics. *Hum Mol Genet* **9:** 2403-8.

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

Green E. D., and Olson M. V. (1990). Systematic screening of yeast artificial-chromosome libraries by use of the polymerase chain reaction. *Proc Natl Acad Sci U S A* **87:** 1213-7.

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

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

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

Griffin T. J., Hall J. G., Prudent J. R., and Smith L. M. (1999). Direct genetic analysis by matrix-assisted laser desorption/ionization mass spectrometry. *Proc Natl Acad Sci U S A* **96:** 6301-6.

Griffin T. J., and Smith L. M. (2000). Single-nucleotide polymorphism analysis by MALDI-TOF mass spectrometry. *Trends Biotechnol* **18:** 77-84.

Griffin T. J., and Smith L. M. (2002). Single-nucleotide polymorphism analysis by MALDI-TOF mass spectrometry. *A TRENDS Guide to Genetic Variation and Genomic medicine***:** S10-18.

Griffin T. J., Tang W., and Smith L. M. (1997). Genetic analysis by peptide nucleic acid affinity MALDI-TOF mass spectrometry. *Nat Biotechnol* **15:** 1368-72.

Griffiths W. J., Jonsson A. P., Liu S., Rai D. K., and Wang Y. (2001). Electrospray and tandem mass spectrometry in biochemistry. *Biochem J* **355:** 545-61.

Gu Z., Hillier L., and Kwok P. Y. (1998). Single nucleotide polymorphism hunting in cyberspace. *Hum Mutat* **12:** 221-5.

Gu Z., Wang H., Nekrutenko A., and Li W. H. (2000). Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequence. *Gene* **259:** 81-8.

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

Gut I. G. (2001). Automation in genotyping of single nucleotide polymorphisms. *Hum Mutat* **17:** 475-92.

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

Gyapay G., Schmitt K., Fizames C., Jones H., Vega-Czarny N., Spillett D., Muselet D., Prud'Homme J. F., Dib C., Auffray C., Morissette J., Weissenbach J., and Goodfellow P. N. (1996). A radiation hybrid map of the human genome. *Hum Mol Genet* **5:** 339-46.

Hacia J. G., and Collins F. S. (1999). Mutational analysis using oligonucleotide microarrays. *J Med Genet* **36:** 730-6.

Hacia J. G., Sun B., Hunt N., Edgemon K., Mosbrook D., Robbins C., Fodor S. P., Tagle D. A., and Collins F. S. (1998). Strategies for mutational analysis of the large multiexon ATM gene using high-density oligonucleotide arrays. *Genome Res* **8:** 1245-58.

Haff L. A., and Smirnov I. P. (1997). Single-nucleotide polymorphism identification assays using a thermostable DNA polymerase and delayed extraction MALDI-TOF mass spectrometry. *Genome Res* **7:** 378-88.

Haffter P., Granato M., Brand M., Mullins M. C., Hammerschmidt M., Kane D. A., Odenthal J., van Eeden F. J., Jiang Y. J., Heisenberg C. P., Kelsh R. N., Furutani-Seiki M., Vogelsang E., Beuchle D., Schach U., Fabian C., and Nusslein-Volhard C. (1996). The identification of genes with unique and essential functions in the development of the zebrafish, *Danio rerio*. *Development* **123:** 1-36.

Hansen R. S., Wijmenga C., Luo P., Stanek A. M., Canfield T. K., Weemaes C. M., and Gartler S. M. (1999). The DNMT3B DNA methyltransferase gene is mutated in the ICF immunodeficiency syndrome. *Proc Natl Acad Sci U S A* **96:** 14412-7.

Hardison R. C., Oeltjen J., and Miller W. (1997). Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res* **7:** 959-66.

Harrington C. A., Rosenow C., and Retief J. (2000). Monitoring gene expression using DNA microarrays. *Curr Opin Microbiol* **3:** 285-91.

Harvey D. J. (2001). Identification of protein-bound carbohydrates by mass spectrometry. *Proteomics* **1:** 311-28.

Hattori M., Fujiyama A., Taylor T. D., Watanabe H., Yada T., Park H. S., Toyoda A., Ishii K., Totoki Y., Choi D. K., Soeda E., Ohki M., Takagi T., Sakaki Y., Taudien S., Blechschmidt K., Polley A., Menzel U., Delabar J., Kumpf K., Lehmann R., Patterson D., Reichwald K., Rump A., Schillhabel M., Schudy A., Zimmermann W., Rosenthal A., Kudoh J., Schibuya K., Kawasaki K., Asakawa S., Shintani A., Sasaki T., Nagamine K., Mitsuyama S., Antonarakis S. E., Minoshima S., Shimizu N., Nordsiek G., Hornischer K., Brant P., Scharfe M., Schon O., Desario A., Reichelt J., Kauer G., Blocker H., Ramser J., Beck A., Klages S., Hennig S., Riesselmann L., Dagand E., Haaf T., Wehrmeyer S., Borzym K., Gardiner K., Nizetic D., Francis F., Lehrach H., Reinhardt R., and Yaspo M. L. (2000). The DNA sequence of human chromosome 21. *Nature* **405:** 311-9.

Hieter P., and Boguski M. (1997). Functional genomics: it's all how you read it. *Science* **278:** 601-2.

Hill W. G., and Robertson A. (1968). Linkage-disequilibrium in finite populations. *Theor. Appl. Genet.* **38:** 226-31.

Hillier L. D., Lennon G., Becker M., Bonaldo M. F., Chiapelli B., Chissoe S., Dietrich N., DuBuque T., Favello A., Gish W., Hawkins M., Hultman M., Kucaba T., Lacy M., Le M., Le N., Mardis E., Moore B., Morris M., Parsons J., Prange C., Rifkin L., Rohlfing T., Schellenberg K., Marra M., and *et al*. (1996). Generation and analysis of 280,000 human expressed sequence tags. *Genome Res* **6:** 807-28.

Hogenesch J. B., Ching K. A., Batalov S., Su A. I., Walker J. R., Zhou Y., Kay S. A., Schultz P. G., and Cooke M. P. (2001). A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* **106:** 413-5.

Holloway J. W., Beghe B., Turner S., Hinks L. J., Day I. N., and Howell W. M. (1999). Comparison of three methods for single nucleotide polymorphism typing for DNA bank studies: sequence-specific oligonucleotide probe hybridisation, TaqMan liquid phase hybridisation, and microplate array diagonal gel electrophoresis (MADGE). *Hum Mutat* **14:** 340-7.

Holmquist G. P. (1992). Chromosome bands, their chromatin flavors, and their functional features. *Am J Hum Genet* **51:** 17-37.

Horton R., Niblett D., Milne S., Palmer S., Tubby B., Trowsdale J., and Beck S. (1998). Large-scale sequence comparisons reveal unusually high levels of variation in the HLA-DQB1 locus in the class II region of the human MHC. *J Mol Biol* **282:** 71-97.

Houck C. M., Rinehart F. P., and Schmid C. W. (1979). A ubiquitous family of repeated DNA sequences in the human genome. *J Mol Biol* **132:** 289-306.

Hubbard T., Barker D., Birney E., Cameron G., Chen Y., Clark L., Cox T., Cuff J., Curwen V., Down T., Durbin R., Eyras E., Gilbert J., Hammond M., Huminiecki L., Kasprzyk A., Lehvaslaiho H., Lijnzaad P., Melsopp C., Mongin E., Pettett R., Pocock M., Potter S., Rust A., Schmidt E., Searle S., Slater G., Smith J., Spooner W., Stabenau A., Stalker J., Stupka E., Ureta-Vidal A., Vastrik I., and Clamp M. (2002). The Ensembl genome database project. *Nucleic Acids Res* **30:** 38-41.

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

Hudson T. J., Stein L. D., Gerety S. S., Ma J., Castle A. B., Silva J., Slonim D. K., Baptista R., Kruglyak L., Xu S. H., and *et al*. (1995). An STS-based map of the human genome. *Science* **270:** 1945-54.

Hughes T. R., Marton M. J., Jones A. R., Roberts C. J., Stoughton R., Armour C. D., Bennett H. A., Coffey E., Dai H., He Y. D., Kidd M. J., King A. M., Meyer M. R., Slade D., Lum P. Y., Stepaniants S. B., Shoemaker D. D., Gachotte D., Chakraburtty K., Simon J., Bard M., and Friend S. H. (2000). Functional discovery via a compendium of expression profiles. *Cell* **102:** 109-26.

Hugot J. P., Chamaillard M., Zouali H., Lesage S., Cezard J. P., Belaiche J., Almer S., Tysk C., O'Morain C. A., Gassull M., Binder V., Finkel Y., Cortot A., Modigliani R., Laurent-Puig P., Gower-Rousseau C., Macry J., Colombel J. F., Sahbatou M., and Thomas G. (2001). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411:** 599-603.

Humphray S. J., Knaggs S. J., and Ragoussis I. (2001). Contiguation of bacterial clones. *Methods Mol Biol* **175:** 69-108.

International Human Genome Mapping Consortium (IHGMC): McPherson J. D., Marra M., Hillier L., Waterston R. H., Chinwalla A., Wallis J., Sekhon M., Wylie K., Mardis E. R., Wilson R. K., Fulton R., Kucaba T. A., Wagner-McPherson C., Barbazuk W. B., Gregory S. G., Humphray S. J., French L., Evans R. S., Bethel G., Whittaker A., Holden J. L., McCann O. T., Dunham A., Soderlund C., Scott C. E., Bentley D. R., Schuler G., Chen H. C., Jang W., Green E. D., Idol J. R., Maduro V. V., Montgomery K. T., Lee E., Miller A., Emerling S., Kucherlapati, Gibbs R., Scherer S., Gorrell J. H., Sodergren E., Clerc-Blankenburg K., Tabor P., Naylor S., Garcia D., de Jong P. J., Catanese J. J., Nowak N., Osoegawa K., Qin S., Rowen L., Madan A., Dors M., Hood L., Trask B., Friedman C., Massa H., Cheung V. G., Kirsch I. R., Reid T., Yonescu R., Weissenbach J., Bruls T., Heilig R., Branscomb E., Olsen A., Doggett N., Cheng J. F., Hawkins T., Myers R. M., Shang J., Ramirez L., Schmutz J., Velasquez O., Dixon K., Stone N. E., Cox D. R., Haussler D., Kent W. J., Furey T., Rogic S., Kennedy S., Jones S., Rosenthal A., Wen G.,

Schilhabel M., Gloeckner G., Nyakatura G., Siebert R., Schlegelberger B., Korenberg J., Chen X. N., Fujiyama A., Hattori M., Toyoda A., Yada T., Park H. S., Sakaki Y., Shimizu N., *et al*. (2001). A physical map of the human genome. *Nature* **409:** 934-41.

International Human Genome Sequencing Consortium (IHGSC): Lander E. S., Linton L. M., Birren B., Nusbaum C., Zody M. C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W., Funke R., Gage D., Harris K., Heaford A., Howland J., Kann L., Lehoczky J., LeVine R., McEwan P., McKernan K., Meldrim J., Mesirov J. P., Miranda C., Morris W., Naylor J., Raymond C., Rosetti M., Santos R., Sheridan A., Sougnez C., Stange-Thomann N., Stojanovic N., Subramanian A., Wyman D., Rogers J., Sulston J., Ainscough R., Beck S., Bentley D., Burton J., Clee C., Carter N., Coulson A., Deadman R., Deloukas P., Dunham A., Dunham I., Durbin R., French L., Grafham D., Gregory S., Hubbard T., Humphray S., Hunt A., Jones M., Lloyd C., McMurray A., Matthews L., Mercer S., Milne S., Mullikin J. C., Mungall A., Plumb R., Ross M., Shownkeen R., Sims S., Waterston R. H., Wilson R. K., Hillier L. W., McPherson J. D., Marra M. A., Mardis E. R., Fulton L. A., Chinwalla A. T., Pepin K. H., Gish W. R., Chissoe S. L., Wendl M. C., Delehaunty K. D., Miner T. L., Delehaunty A., Kramer J. B., Cook L. L., Fulton R. S., Johnson D. L., Minx P. J., Clifton S. W., Hawkins T., Branscomb E., Predki P., Richardson P., Wenning S., Slezak T., Doggett N., Cheng J. F., Olsen A., Lucas S., Elkin C., Uberbacher E., Frazier M., *et al*. (2001). Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

International SNP Map Working Group (ISNPMWG): Sachidanandam R., Weissman D., Schmidt S. C., Kakol J. M., Stein L. D., Marth G., Sherry S., Mullikin J. C., Mortimore B. J., Willey D. L., Hunt S. E., Cole C. G., Coggill P. C., Rice C. M., Ning Z., Rogers J., Bentley D. R., Kwok P. Y., Mardis E. R., Yeh R. T., Schultz B., Cook L., Davenport R., Dante M., Fulton L., Hillier L., Waterston R. H., McPherson J. D., Gilman B., Schaffner S., Van Etten W. J., Reich D., Higgins J., Daly M. J., Blumenstiel B., Baldwin J., Stange-Thomann N., Zody M. C., Linton L., Lander E. S., and Altshuler D. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409:** 928-33.

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

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

Isaksson A., Landegren U., Syvanen A. C., Bork P., Stein C., Ortigao F., and Brookes A. J. (2000). Discovery, scoring and utilization of human single nucleotide polymorphisms: a multidisciplinary problem. *Eur J Hum Genet* **8:** 154-6.

Jackson I. J. (2001). Mouse genomics: making sense of the sequence. *Curr Biol* **11:** R311-4.

Jackson P. E., Scholl P. F., and Groopman J. D. (2000). Mass spectrometry for genotyping: an emerging tool for molecular medicine. *Molecular Medicine Today* **6:** 271-76.

Jacq C., Miller J. R., and Brownlee G. G. (1977). A pseudogene structure in 5S DNA of *Xenopus laevis*. *Cell* **12:** 109-20.

James M. R., and Lindpaintner K. (1997). Why map the rat? *Trends Genet* **13:** 171-3.

Jareborg N., Birney E., and Durbin R. (1999). Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res* **9:** 815-24.

Jeffreys A. J., Kauppi L., and Neumann R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* **29:** 217-22.

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

Jenkins S., and Gibson N. (2002). High-throughput SNP genotyping. *Comparative and Functional Genomics* **3:** 57-66.

Ji L., Malecki M., Warram J. H., Yang Y., Rich S. S., and Krolewski A. S. (1997). New susceptibility locus for NIDDM is localized to human chromosome 20q. *Diabetes* **46:** 876-81.

Ji Y., Eichler E. E., Schwartz S., and Nicholls R. D. (2000). Structure of chromosomal duplicons and their role in mediating human genomic disorders. *Genome Res* **10:** 597-610.

Jordan E., and Collins F. S. (1996). A march of genetic maps. *Nature* **380:** 111-2.

Jorde L. B. (2000). Linkage disequilibrium and the search for complex disease genes. *Genome Res* **10:** 1435-44.

Jurka J. (2000). Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* **16:** 418-20.

Jurka J., and Milosavljevic A. (1991). Reconstruction and analysis of human Alu genes. *J Mol Evol* **32:** 105-21.

Jurka J., Zietkiewicz E., and Labuda D. (1995). Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the mesozoic era. *Nucleic Acids Res* **23:** 170-5.

Kan Z., Rouchka E. C., Gish W. R., and States D. J. (2001). Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res* **11:** 889-900.

Karas M., and Hillenkamp F. (1988). Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem* **60:** 2299-301.

Kass D. H. (2001). Impact of SINEs and LINEs on the Mammalian Genome. *Current Genomics* **2:** 199-219.

Kelley S. (2000). Getting started with Acedb. *Brief Bioinform* **1:** 131-7.

Kelly P. D., Chu F., Woods I. G., Ngo-Hazelett P., Cardozo T., Huang H., Kimm F., Liao L., Yan Y. L., Zhou Y., Johnson S. L., Abagyan R., Schier A. F., Postlethwait J. H., and Talbot W. S. (2000). Genetic linkage mapping of zebrafish genes and ESTs. *Genome Res* **10:** 558-67.

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

Kent W. J., Sugnet C. W., Furey T. S., Roskin K. M., Pringle T. H., Zahler A. M., and Haussler A. D. (2002). The Human Genome Browser at UCSC. *Genome Res* **12:** 996-1006.

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

Kim J., Gordon L., Dehal P., Badri H., Christensen M., Groza M., Ha C., Hammond S., Vargas M., Wehri E., Wagner M., Olsen A., and Stubbs L. (2001). Homology-driven assembly of a sequence-ready mouse BAC contig map spanning regions related to the 46-Mb gene-rich euchromatic segments of human chromosome 19. *Genomics* **74:** 129-41.

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

Kirpekar F., Nordhoff E., Larsen L. K., Kristiansen K., Roepstorff P., and Hillenkamp F. (1998). DNA sequence analysis by MALDI mass spectrometry. *Nucleic Acids Res* **26:** 2554-9.

Kong A., Gudbjartsson D. F., Sainz J., Jonsdottir G. M., Gudjonsson S. A., Richardsson B., Sigurdardottir S., Barnard J., Hallbeck B., Masson G., Shlien A., Palsson S. T., Frigge M. L., Thorgeirsson T. E., Gulcher J. R., and Stefansson K. (2002). A high-resolution recombination map of the human genome. *Nat Genet* **31:** 241-7.

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

Korn B., Sedlacek Z., Manca A., Kioschis P., Konecki D., Lehrach H., and Poustka A. (1992). A strategy for the selection of transcribed sequences in the Xq28 region. *Hum Mol Genet* **1:** 235-42.

Köster H., Tang K., Fu D. J., Braun A., van den Boom D., Smith C. L., Cotter R. J., and Cantor C. R. (1996). A strategy for rapid and efficient DNA sequencing by mass spectrometry. *Nat Biotechnol* **14:** 1123-8.

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

Kristensen V. N., Kelefiotis D., Kristensen T., and Borresen-Dale A. L. (2001). High-throughput methods for detection of genetic variation. *Biotechniques* **30:** 318-22, 324, 326 passim.

Kruglyak L. (1997). The use of a genetic map of biallelic markers in linkage studies. *Nat Genet* **17:** 21-4.

Kruglyak L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* **22:** 139-44.

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

Kwitek A. E., Tonellato P. J., Chen D., Gullings-Handley J., Cheng Y. S., Twigger S., Scheetz T. E., Casavant T. L., Stoll M., Nobrega M. A., Shiozawa M., Soares M. B., Sheffield V. C., and Jacob H. J. (2001). Automated construction of

high-density comparative maps between rat, human, and mouse. *Genome Res* **11:** 1935-43.

Kwok P. Y. (1998). Genotyping by mass spectroscopy takes flight. *Nat. Biotech.* **16:** 1314-15.

Kwok P. Y., and Gu Z. (1999). Single nucleotide polymorphism libraries: why and how are we building them? *Mol Med Today* **5:** 538-43.

Laken S. J., Jackson P. E., Kinzler K. W., Vogelstein B., Strickland P. T., Groopman J. D., and Friesen M. D. (1998). Genotyping by mass spectrometric analysis of short DNA fragments. *Nat Biotechnol* **16:** 1352-6.

Lamerdin J. E., Montgomery M. A., Stilwagen S. A., Scheidecker L. K., Tebbs R. S., Brookman K. W., Thompson L. H., and Carrano A. V. (1995). Genomic sequence comparison of the human and mouse XRCC1 DNA repair gene regions. *Genomics* **25:** 547-54.

Lamerdin J. E., Stilwagen S. A., Ramirez M. H., Stubbs L., and Carrano A. V. (1996). Sequence analysis of the ERCC2 gene regions in human, mouse, and hamster reveals three linked genes. *Genomics* **34:** 399-409.

Landegren U., Nilsson M., and Kwok P. Y. (1998). Reading bits of genetic information: methods for single-nucleotide polymorphism analysis. *Genome Res* **8:** 769-76.

Lander E. S. (1996). The new genomics: global views of biology. *Science* **274:** 536-9.

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

Lee K. H. (2001). Proteomics: a technology-driven and technology-limited discovery science. *Trends Biotechnol* **19:** 217-22.

Lee P. S., and Lee K. H. (2000). Genomic analysis. *Curr Opin Biotechnol* **11:** 171-5.

Lee T. I., and Young R. A. (2000). Transcription of eukaryotic protein-coding genes. *Annu Rev Genet* **34:** 77-137.

Legrain P., and Selig L. (2000). Genome-wide protein interaction maps using two-hybrid systems. *FEBS Lett* **480:** 32-6.

Leushner J. (2001). MALDI TOF mass spectrometry: an emerging platform for genomics and diagnostics. *Expert Rev Mol Diagn* **1:** 11-8.

Leushner J., and Chiu N. H. (2000). Automated mass spectrometry: a revolutionary technology for clinical diagnostics. *Mol Diagn* **5:** 341-8.

Lewin B. (1994). *Genes V***:** 657-660.

Lewontin R. C. (1964). The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49:** 49-67.

Li L., Krantz I. D., Deng Y., Genin A., Banta A. B., Collins C. C., Qi M., Trask B. J., Kuo W. L., Cochran J., Costa T., Pierpont M. E., Rand E. B., Piccoli D. A., Hood L., and Spinner N. B. (1997). Alagille syndrome is caused by mutations in human Jagged1, which encodes a ligand for Notch1. *Nat Genet* **16:** 243-51.

Li Q., Harju S., and Peterson K. R. (1999). Locus control regions: coming of age at a decade plus. *Trends Genet* **15:** 403-8.

Li W. H., Ellsworth, D.L., Krushkal, J., Chang, B.H., and Hewett-Emmett, D (1996). Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol. Phylogenet. Evol.* **5:** 182-187.

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

Liang P., and Pardee A. B. (1992). Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* **257:** 967-71.

Liang P., and Pardee A. B. (1995). Recent advances in differential display. *Curr Opin Immunol* **7:** 274-80.

Lieberman A. P., and Fischbeck K. H. (2000). Triplet repeat expansion in neuromuscular disease. *Muscle Nerve* **23:** 843-50.

Lindsay S., and Bird A. P. (1987). Use of restriction enzymes to detect potential gene sequences in mammalian DNA. *Nature* **327:** 336-8.

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

Livak K. J. (1999). Allelic discrimination using fluorogenic probes and the 5' nuclease assay. *Genet Anal* **14:** 143-9.

Livak K. J., Flood S. J., Marmaro J., Giusti W., and Deetz K. (1995). Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridisation. *PCR Methods Appl.* **4:** 357-62.

Logsdon J. M., Jr., Tyshenko M. G., Dixon C., J D. J., Walker V. K., and Palmer J. D. (1995). Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory. *Proc Natl Acad Sci U S A* **92:** 8507-11.

Long M., Rosenberg C., and Gilbert W. (1995). Intron phase correlations and the evolution of the intron/exon structure of genes. *Proc Natl Acad Sci U S A* **92:** 12495-9.

Loots G. G., Locksley R. M., Blankespoor C. M., Wang Z. E., Miller W., Rubin E. M., and Frazer K. A. (2000). Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288:** 136-40.

Lovett M., Kere J., and Hinton L. M. (1991). Direct selection: a method for the isolation of cDNAs encoded by large genomic regions. *Proc Natl Acad Sci U S A* **88:** 9628-32.

Lundwall A. (1996). The structure of the semenogelin gene locus--nucleotide sequence of the intergenic and the flanking DNA. *Eur J Biochem* **235:** 466-70.

Lyamichev V., Mast A. L., Hall J. G., Prudent J. R., Kaiser M. W., Takova T., Kwiatkowski R. W., Sander T. J., de Arruda M., Arco D. A., Neri B. P., and Brow M. A. (1999). Polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probes. *Nat Biotechnol* **17:** 292-6.

Maestre J., Tchenio T., Dhellin O., and Heidmann T. (1995). mRNA retroposition in human cells: processed pseudogene formation. *Embo J* **14:** 6333-8.

Maillet I., Lagniel G., Perrot M., Boucherie H., and Labarre J. (1996). Rapid identification of yeast proteins on two-dimensional gels. *J Biol Chem* **271:** 10263-70.

Makalowski W. (2000). Genomic scrap yard: how genomes utilize all that junk. *Gene* **259:** 61-7.

Makalowski W., and Boguski M. S. (1998). Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc Natl Acad Sci U S A* **95:** 9407-12.

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

Mallon A. M., Platzer M., Bate R., Gloeckner G., Botcherby M. R., Nordsiek G., Strivens M. A., Kioschis P., Dangel A., Cunningham D., Straw R. N., Weston P., Gilbert M., Fernando S., Goodall K., Hunter G., Greystrong J. S., Clarke D., Kimberley C., Goerdes M., Blechschmidt K., Rump A., Hinzmann B., Mundy C. R., Miller W., Poustka A., Herman G. E., Rhodes M., Denny P., Rosenthal A., and Brown S. D. (2000). Comparative genome sequence analysis of the Bpa/Str region in mouse and Man. *Genome Res* **10:** 758-75.

Marra M., Hillier L., Kucaba T., Allen M., Barstead R., Beck C., Blistain A., Bonaldo M., Bowers Y., Bowles L., Cardenas M., Chamberlain A., Chappell J., Clifton S., Favello A., Geisel S., Gibbons M., Harvey N., Hill F., Jackson Y., Kohn S., Lennon G., Mardis E., Martin J., Waterston R., and *et al*. (1999). An encyclopedia of mouse genes. *Nat Genet* **21:** 191-4.

Marra M. A., Kucaba T. A., Dietrich N. L., Green E. D., Brownstein B., Wilson R. K., McDonald K. M., Hillier L. W., McPherson J. D., and Waterston R. H. (1997). High throughput fingerprint analysis of large-insert clones. *Genome Res* **7:** 1072-84.

Marras S. A., Kramer F. R., and Tyagi S. (1999). Multiplex detection of single-nucleotide variations using molecular beacons. *Genet Anal* **14:** 151-6.

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

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

Martin D. B., and Nelson P. S. (2001). From genomics to proteomics: techniques and applications in cancer research. *Trends Cell Biol* **11:** S60-5.

Martindale D. W., Wilson M. D., Wang D., Burke R. D., Chen X., Duronio V., and Koop B. F. (2000). Comparative genomic sequence analysis of the Williams syndrome region (LIMK1-RFC2) of human chromosome 7q11.23. *Mamm Genome* **11:** 890-8.

Martzen M. R., McCraith S. M., Spinelli S. L., Torres F. M., Fields S., Grayhack E. J., and Phizicky E. M. (1999). A biochemical genomics approach for identifying genes by the activity of their products. *Science* **286:** 1153-5.

Masood E. (1999). As consortium plans free SNP map of human genome. *Nature* **398:** 545-6.

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

Matz M. V., and Lukyanov S. A. (1998). Different strategies of differential display: areas of application. *Nucleic Acids Res* **26:** 5537-43.

Mayor C., Brudno M., Schwartz J. R., Poliakov A., Rubin E. M., Frazer K. A., Pachter L. S., and Dubchak I. (2000). VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16:** 1046-7.

McCarthy L. C., Bihoreau M. T., Kiguwa S. L., Browne J., Watanabe T. K., Hishigaki H., Tsuji A., Kiel S., Webber C., Davis M. E., Knights C., Smith A., Critcher R., Huxtall P., Hudson J. R., Jr., Ono T., Hayashi H., Takagi T., Nakamura Y., Tanigami A., Goodfellow P. N., Lathrop G. M., and James M. R. (2000). A whole-genome radiation hybrid panel and framework map of the rat genome. *Mamm Genome* **11:** 791-5.

Meisler M. H. (2001). Evolutionarily conserved noncoding DNA in the human genome: how much and what for? *Genome Res* **11:** 1617-8.

Mighell A. J., Markham A. F., and Robinson P. A. (1997). Alu sequences. *FEBS Lett* **417:** 1-5.

Mighell A. J., Smith N. R., Robinson P. A., and Markham A. F. (2000). Vertebrate pseudogenes. *FEBS Lett* **468:** 109-14.

Miki R., Kadota K., Bono H., Mizuno Y., Tomaru Y., Carninci P., Itoh M., Shibata K., Kawai J., Konno H., Watanabe S., Sato K., Tokusumi Y., Kikuchi N., Ishii Y., Hamaguchi Y., Nishizuka I., Goto H., Nitanda H., Satomi S., Yoshiki A., Kusakabe M., DeRisi J. L., Eisen M. B., Iyer V. R., Brown P. O., Muramatsu M., Shimada H., Okazaki Y., and Hayashizaki Y. (2001). Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays. *Proc Natl Acad Sci U S A* **98:** 2199-204.

Miles C., Elgar G., Coles E., Kleinjan D. J., van Heyningen V., and Hastie N. (1998). Complete sequencing of the Fugu WAGR region from WT1 to PAX6: dramatic compaction and conservation of synteny with human chromosome 11p13. *Proc Natl Acad Sci U S A* **95:** 13068-72.

Miller R. D., and Kwok P. Y. (2001). The birth and death of human single-nucleotide polymorphisms: new experimental evidence and implications for human history and medicine. *Hum Mol Genet* **10:** 2195-8.

Miller W. (2001). Comparison of genomic DNA sequences: solved and unsolved problems. *Bioinformatics* **17:** 391-7.

Mills A. A., and Bradley A. (2001). From mouse to man: generating megabase chromosome rearrangements. *Trends Genet* **17:** 331-9.

Mironov A. A., Fickett J. W., and Gelfand M. S. (1999). Frequent alternative splicing of human genes. *Genome Res* **9:** 1288-93.

Moffatt M. F., Traherne J. A., Abecasis G. R., and Cookson W. O. (2000). Single nucleotide polymorphism and linkage disequilibrium within the TCR alpha/delta locus. *Hum Mol Genet* **9:** 1011-9.

Monaco A. P., Neve R. L., Colletti-Feener C., Bertelson C. J., Kurnit D. M., and Kunkel L. M. (1986). Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene. *Nature* **323:** 646-50.

Montgomery K. T., Lee E., Miller A., Lau S., Shim C., Decker J., Chiu D., Emerling S., Sekhon M., Kim R., Lenz J., Han J., Ioshikhes I., Renault B., Marondel I., Yoon S. J., Song K., Murty V. V., Scherer S., Yonescu R., Kirsch I. R., Ried T., McPherson J., Gibbs R., and Kucherlapati R. (2001). A high-resolution map of human chromosome 12. *Nature* **409:** 945-6.

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

Morgan J. G., Dolganov G. M., Robbins S. E., Hinton L. M., and Lovett M. (1992). The selective isolation of novel cDNAs encoded by the regions surrounding the human interleukin 4 and 5 genes. *Nucleic Acids Res* **20:** 5173-9.

Mori I., Benian G. M., Moerman D. G., and Waterston R. H. (1988). Transposable element Tc1 of Caenorhabditis elegans recognizes specific target sequences for integration. *Proc Natl Acad Sci U S A* **85:** 861-4.

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

Mullikin J. C., Hunt S. E., Cole C. G., Mortimore B. J., Rice C. M., Burton J., Matthews L. H., Pavitt R., Plumb R. W., Sims S. K., Ainscough R. M., Attwood J., Bailey J. M., Barlow K., Bruskiewich R. M., Butcher P. N., Carter N. P., Chen Y., Clee C. M., Coggill P. C., Davies J., Davies R. M., Dawson E., Francis M. D., Joy A. A., Lamble R. G., Langford C. F., Macarthy J., Mall V., Moreland A., Overton-Larty E. K., Ross M. T., Smith L. C., Steward C. A., Sulston J. E., Tinsley E. J., Turney K. J., Willey D. L., Wilson G. D., McMurray A. A., Dunham I., Rogers J., and Bentley D. R. (2000). An SNP map of human chromosome 22. *Nature* **407:** 516-20.

Mungall A. J., Edwards C. A., Ranby S. A., Humphray S. J., Heathcott R. W., Clee C. M., East C. L., Holloway E., Butler A. P., Langford C. F., Gwilliam R., Rice K. M., Maslen G. L., Carter N. P., Ross M. T., Deloukas P., Bentley D. R., and Dunham I. (1996). Physical mapping of chromosome 6: a strategy for the rapid generation of sequence-ready contigs. *DNA Seq* **7:** 47-9.

Murakami K., and Takagi T. (1998). Gene recognition by combination of several gene-finding programs. *Bioinformatics* **14:** 665-75.

Mural R. J., Adams M. D., Myers E. W., Smith H. O., Miklos G. L., Wides R., Halpern A., Li P. W., Sutton G. G., Nadeau J., Salzberg S. L., Holt R. A., Kodira C. D., Lu F., Chen L., Deng Z., Evangelista C. C., Gan W., Heiman T. J., Li J., Li Z., Merkulov G. V., Milshina N. V., Naik A. K., Qi R., Shue B. C., Wang A., Wang J., Wang X., Yan X., Ye J., Yooseph S., Zhao Q., Zheng L., Zhu S. C., Biddick K., Bolanos R., Delcher A. L., Dew I. M., Fasulo D.,

Flanigan M. J., Huson D. H., Kravitz S. A., Miller J. R., Mobarry C. M., Reinert K., Remington K. A., Zhang Q., Zheng X. H., Nusskern D. R., Lai Z., Lei Y., Zhong W., Yao A., Guan P., Ji R. R., Gu Z., Wang Z. Y., Zhong F., Xiao C., Chiang C. C., Yandell M., Wortman J. R., Amanatides P. G., Hladun S. L., Pratts E. C., Johnson J. E., Dodson K. L., Woodford K. J., Evans C. A., Gropman B., Rusch D. B., Venter E., Wang M., Smith T. J., Houck J. T., Tompkins D. E., Haynes C., Jacob D., Chin S. H., Allen D. R., Dahlke C. E., Sanders R., Li K., Liu X., Levitsky A. A., Majoros W. H., Chen Q., Xia A. C., Lopez J. R., Donnelly M. T., Newman M. H., Glodek A., Kraft C. L., Nodell M., Ali F., An H. J., Baldwin-Pitts D., Beeson K. Y., Cai S., *et al*. (2002). A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296:** 1661-71.

Murphy W. J., Stanyon R., and O'Brien S. J. (2001). Evolution of mammalian genome organization inferred from comparative gene mapping. *Genome Biol* **2:** REVIEWS0005.

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

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

Nagaraja R., Kere J., MacMillan S., Masisi M. J., Johnson D., Molini B. J., Halley G. R., Wein K., Trusgnich M., Eble B., and *et al*. (1994). Characterization of four human YAC libraries for clone size, chimerism and X chromosome sequence representation. *Nucleic Acids Res* **22:** 3406-11.

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

Ng P. C., and Henikoff S. (2002). Accounting for human polymorphisms predicted to affect protein function. *Genome Res* **12:** 436-46.

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

Nomura N., Miyajima N., Sazuka T., Tanaka A., Kawarabayasi Y., Sato S., Nagase T., Seki N., Ishikawa K., and Tabata S. (1994). Prediction of the coding sequences of unidentified human genes. I. The coding sequences of 40 new genes (KIAA0001-KIAA0040) deduced by analysis of randomly sampled cDNA clones from human immature myeloid cell line KG-1. *DNA Res* **1:** 27-35.

Noordewier M. O., and Warren P. V. (2001). Gene expression microarrays and the integration of biological knowledge. *Trends Biotechnol* **19:** 412-5.

Nordborg M., and Tavare S. (2002). Linkage disequilibrium: what history has to tell us. *Trends Genet* **18:** 83-90.

Nowotny P., Kwon J. M., and Goate A. M. (2001). SNP analysis to dissect human traits. *Curr Opin Neurobiol* **11:** 637-41.

Nusbaum C., Slonim D. K., Harris K. L., Birren B. W., Steen R. G., Stein L. D., Miller J., Dietrich W. F., Nahf R., Wang V., Merport O., Castle A. B., Husain Z., Farino G., Gray D., Anderson M. O., Devine R., Horton L. T., Jr., Ye W., Wu X., Kouyoumjian V., Zemsteva I. S., Wu Y., Collymore A. J., Courtney D. F., and *et al*. (1999). A YAC-based physical map of the mouse genome. *Nat Genet* **22:** 388-93.

O'Brien S. J., Menotti-Raymond M., Murphy W. J., Nash W. G., Wienberg J., Stanyon R., Copeland N. G., Jenkins N. A., Womack J. E., and Marshall Graves J. A. (1999). The promise of comparative genomics in mammals. *Science* **286:** 458-62, 479-81.

Oeltjen J. C., Malley T. M., Muzny D. M., Miller W., Gibbs R. A., and Belmont J. W. (1997). Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res* **7:** 315-29.

Ogura Y., Bonen D. K., Inohara N., Nicolae D. L., Chen F. F., Ramos R., Britton H., Moran T., Karaliuskas R., Duerr R. H., Achkar J. P., Brant S. R., Bayless T. M., Kirschner B. S., Hanauer S. B., Nunez G., and Cho J. H. (2001). A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411:** 603-6.

Olivier M., Bustos V. I., Levy M. R., Smick G. A., Moreno I., Bushard J. M., Almendras A. A., Sheppard K., Zierten D. L., Aggarwal A., Carlson C. S., Foster B. D., Vo N., Kelly L., Liu X., and Cox D. R. (2001). Complex high-resolution linkage disequilibrium and haplotype patterns of single-nucleotide polymorphisms in 2.5 Mb of sequence on human chromosome 21. *Genomics* **78:** 64-72.

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

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

Osoegawa K., Tateno M., Woon P. Y., Frengen E., Mammoser A. G., Catanese J. J., Hayashizaki Y., and de Jong P. J. (2000). Bacterial artificial chromosome libraries for mouse sequencing and functional analysis. *Genome Res* **10:** 116-28.

Pandey A., and Mann M. (2000). Proteomics to study genes and genomes. *Nature* **405:** 837-46.

Paracchini S., Arredi B., Chalk R., and Tyler-Smith C. (2002). Hierarchical high-throughput SNP genotyping of the human Y chromosome using MALDI-TOF mass spectrometry. *Nucleic Acids Res* **30:** e27.

Parimoo S., Patanjali S. R., Shukla H., Chaplin D. D., and Weissman S. M. (1991). cDNA selection: efficient PCR approach for the selection of cDNAs encoded in large chromosomal DNA fragments. *Proc Natl Acad Sci U S A* **88:** 9623-7.

Pastinen T., Kurg A., Metspalu A., Peltonen L., and Syvanen A. C. (1997). Minisequencing: a specific tool for DNA analysis and diagnostics on oligonucleotide arrays. *Genome Res* **7:** 606-14.

Pastinen T., Raitio M., Lindroos K., Tainola P., Peltonen L., and Syvanen A. C. (2000). A system for specific, high-throughput genotyping by allele-specific primer extension on microarrays. *Genome Res* **10:** 1031-42.

Patil N., Berno A. J., Hinds D. A., Barrett W. A., Doshi J. M., Hacker C. R., Kautzer C. R., Lee D. H., Marjoribanks C., McDonough D. P., Nguyen B. T., Norris M. C., Sheehan J. B., Shen N., Stern D., Stokowski R. P., Thomas D. J., Trulson M. O., Vyas K. R., Frazer K. A., Fodor S. P., and Cox D. R. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294:** 1719-23.

Paule M. R., and White R. J. (2000). Survey and summary: transcription by RNA polymerases I and III. *Nucleic Acids Res* **28:** 1283-98.

Pearce S. H., Vaidya B., Imrie H., Perros P., Kelly W. F., Toft A. D., McCarthy M. I., Young E. T., and Kendall-Taylor P. (1999). Further evidence for a susceptibility locus on chromosome 20q13.11 in families with dominant transmission of Graves disease. *Am J Hum Genet* **65:** 1462-5.

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

Persidis A. (1999). Bioinformatics. *Nat Biotechnol* **17:** 828-30.

Persidis A. (2000). Data mining in biotechnology. *Nat Biotechnol* **18:** 237-8.

Peters J., Church D., Zuberi A. R., and Selley R. (1999). Mouse chromosome 2. *Mamm Genome* **10:** 941.

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

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

Pletcher M. T., Wiltshire T., Cabin D. E., Villanueva M., and Reeves R. H. (2001). Use of comparative physical and sequence mapping to annotate mouse chromosome 16 and human chromosome 21. *Genomics* **74:** 45-54.

Ponger L., Duret L., and Mouchiroud D. (2001). Determinants of CpG islands: expression in early embryo and isochore structure. *Genome Res* **11:** 1854-60.

Prak E. T., and Kazazian H. H., Jr. (2000). Mobile elements and the human genome. *Nat Rev Genet* **1:** 134-44.

Price J. A., Brewer C. S., Howard T. D., Fossey S. C., Sale M. M., Ji L., Krolewski A. S., and Bowden D. W. (1999). A physical map of the 20q12-q13.1 region associated with type 2 diabetes. *Genomics* **62:** 208-15.

Pritchard J. K., and Przeworski M. (2001). Linkage disequilibrium in humans: models and data. *Am J Hum Genet* **69:** 1-14.

Pruitt K. D., and Maglott D. R. (2001). RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res* **29:** 137-40.

Przeworski M., Hudson R. R., and Di Rienzo A. (2000). Adjusting the focus on human variation. *Trends Genet* **16:** 296-302.

Qiu Y., Cavelier L., Chiu S., Yang X., Rubin E., and Cheng J. F. (2001). Human and mouse ABCA1 comparative sequencing and transgenesis studies revealing novel regulatory sequences. *Genomics* **73:** 66-76.

Ranade K., Chang M. S., Ting C. T., Pei D., Hsiao C. F., Olivier M., Pesich R., Hebert J., Chen Y. D., Dzau V. J., Curb D., Olshen R., Risch N., Cox D. R., and Botstein D. (2001). High-throughput genotyping with single nucleotide polymorphisms. *Genome Res* **11:** 1262-8.

Rapp J. P. (2000). Genetic analysis of inherited hypertension in the rat. *Physiol Rev* **80:** 135-72.

Reese M. G., Hartzell G., Harris N. L., Ohler U., Abril J. F., and Lewis S. E. (2000). Genome annotation assessment in *Drosophila melanogaster*. *Genome Res* **10:** 483-501.

Reich D. E., Cargill M., Bolk S., Ireland J., Sabeti P. C., Richter D. J., Lavery T., Kouyoumjian R., Farhadian S. F., Ward R., and Lander E. S. (2001). Linkage disequilibrium in the human genome. *Nature* **411:** 199-204.

Reichhardt T. (1999). It's sink or swim as a tidal wave of data approaches. *Nature* **399:** 517-20.

Remm M., and Metspalu A. (2002). High-density genotyping and linkage disequilibrium in the human genome using chromosome 22 as a model. *Curr Opin Chem Biol* **6:** 24-30.

Ren B., Robert F., Wyrick J. J., Aparicio O., Jennings E. G., Simon I., Zeitlinger J., Schreiber J., Hannett N., Kanin E., Volkert T. L., Wilson C. J., Bell S. P., and Young R. A. (2000). Genome-wide location and function of DNA binding proteins. *Science* **290:** 2306-9.

Rhodes M., Straw R., Fernando S., Evans A., Lacey T., Dearlove A., Greystrong J., Walker J., Watson P., Weston P., Kelly M., Taylor D., Gibson K., Mundy C., Bourgade F., Poirier C., Simon D., Brunialti A. L., Montagutelli X., Gu'enet J. L., Haynes A., and Brown S. D. (1998). A high-resolution microsatellite map of the mouse genome. *Genome Res* **8:** 531-42.

Rich S. S. (1990). Mapping genes in diabetes. Genetic epidemiological perspective. *Diabetes* **39:** 1315-9.

Riley J., Butler R., Ogilvie D., Finniear R., Jenner D., Powell S., Anand R., Smith J. C., and Markham A. F. (1990). A novel, rapid method for the isolation of terminal sequences from yeast artificial chromosome (YAC) clones. *Nucleic Acids Res* **18:** 2887-90.

Rioux J. D., Daly M. J., Silverberg M. S., Lindblad K., Steinhart H., Cohen Z., Delmonte T., Kocher K., Miller K., Guschwan S., Kulbokas E. J., O'Leary S., Winchester E., Dewar K., Green T., Stone V., Chow C., Cohen A., Langelier D., Lapointe G., Gaudet D., Faith J., Branco N., Bull S. B., McLeod R. S., Griffiths A. M., Bitton A., Greenberg G. R., Lander E. S., Siminovitch K. A.,

and Hudson T. J. (2001). Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat Genet* **29:** 223-8.

Risch N., and Merikangas K. (1996). The future of genetic studies of complex human diseases. *Science* **273:** 1516-7.

Risch N. J. (2000). Searching for genetic determinants in the new millennium. *Nature* **405:** 847-56.

Rogic S., Mackworth A. K., and Ouellette F. B. (2001). Evaluation of gene-finding programs on mammalian sequences. *Genome Res* **11:** 817-32.

Rogozin I. B., Mayorov V. I., Lavrentieva M. V., Milanesi L., and Adkison L. R. (2000). Prediction and phylogenetic analysis of mammalian short interspersed elements (SINEs). *Brief Bioinform* **1:** 260-74.

Rommens J. M., Iannuzzi M. C., Kerem B., Drumm M. L., Melmer G., Dean M., Rozmahel R., Cole J. L., Kennedy D., Hidaka N., and *et al*. (1989). Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* **245:** 1059-65.

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

Ross P., Hall L., Smirnov I., and Haff L. (1998). High level multiplex genotyping by MALDI-TOF mass spectrometry. *Nat Biotechnol* **16:** 1347-51.

Ross-MacDonald P., Coelho P. S., Roemer T., Agarwal S., Kumar A., Jansen R., Cheung K. H., Sheehan A., Symoniatis D., Umansky L., Heidtman M., Nelson F. K., Iwasaki H., Hager K., Gerstein M., Miller P., Roeder G. S., and Snyder M. (1999). Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402:** 413-8.

Rothstein R. (1991). Targeting, disruption, replacement, and allele rescue: integrative DNA transformation in yeast. *Methods Enzymol* **194:** 281-301.

Rozen S., and Skaletsky H. (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132:** 365-86.

Rubin G. M. (1996). Around the genomes: the Drosophila genome project. *Genome Res* **6:** 71-9.

Rushforth A. M., Saari B., and Anderson P. (1993). Site-selected insertion of the transposon Tc1 into a Caenorhabditis elegans myosin light chain gene. *Mol Cell Biol* **13:** 902-10.

Saccone S., Caccio S., Kusuda J., Andreozzi L., and Bernardi G. (1996). Identification of the gene-richest bands in human chromosomes. *Gene* **174:** 85-94.

Saccone S., De Sario A., Wiegant J., Raap A. K., Della Valle G., and Bernardi G. (1993). Correlations between isochores and chromosomal bands in the human genome. *Proc Natl Acad Sci U S A* **90:** 11929-33.

Saiki R. K., Gelfand D. H., Stoffel S., Scharf S. J., Higuchi R., Horn G. T., Mullis K. B., and Erlich H. A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239:** 487-91.

Saiki R. K., Scharf S., Faloona F., Mullis K. B., Horn G. T., Erlich H. A., and Arnheim N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230:** 1350-4.

Salamov A. A., and Solovyev V. V. (2000). Ab initio gene finding in Drosophila genomic DNA. *Genome Res* **10:** 516-22.

Sanger F., Nicklen S., and Coulson A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74:** 5463-7.

Sassaman D. M., Dombroski B. A., Moran J. V., Kimberland M. L., Naas T. P., DeBerardinis R. J., Gabriel A., Swergold G. D., and Kazazian H. H., Jr. (1997). Many human L1 elements are capable of retrotransposition. *Nat Genet* **16:** 37-43.

Sauer S., Lechner D., Berlin K., Lehrach H., Escary J. L., Fox N., and Gut I. G. (2000). A novel procedure for efficient genotyping of single nucleotide polymorphisms. *Nucleic Acids Res* **28:** E13.

Schalkwyk L. C., Cusack B., Dunkel I., Hopp M., Kramer M., Palczewski S., Piefke J., Scheel S., Weiher M., Wenske G., Lehrach H., and Himmelbauer H. (2001). Advanced integrated mouse YAC map including BAC framework. *Genome Res* **11:** 2142-50.

Scheetz T. E., Raymond M. R., Nishimura D. Y., McClain A., Roberts C., Birkett C., Gardiner J., Zhang J., Butters N., Sun C., Kwitek-Black A., Jacob H., Casavant T. L., Soares M. B., and Sheffield V. C. (2001). Generation of a high-density rat EST map. *Genome Res* **11:** 497-502.

Scherf M., Klingenhoff A., Frech K., Quandt K., Schneider R., Grote K., Frisch M., Gailus-Durner V., Seidel A., Brack-Werner R., and Werner T. (2001). First pass annotation of promoters on human chromosome 22. *Genome Res* **11:** 333-40.

Scherf M., Klingenhoff A., and Werner T. (2000). Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J Mol Biol* **297:** 599-606.

Schmid A., Schindelholz B., and Zinn K. (2002). Combinatorial RNAi: a method for evaluating the functions of gene families in Drosophila. *Trends Neurosci* **25:** 71-4.

Schuler G. D., Boguski M. S., Stewart E. A., Stein L. D., Gyapay G., Rice K., White R. E., Rodriguez-Tome P., Aggarwal A., Bajorek E., Bentolila S., Birren B. B., Butler A., Castle A. B., Chiannilkulchai N., Chu A., Clee C., Cowles S., Day P. J., Dibling T., Drouot N., Dunham I., Duprat S., East C., Hudson T. J., and *et al*. (1996). A gene map of the human genome. *Science* **274:** 540-6.

Schwartz S., Zhang Z., Frazer K. A., Smit A., Riemer C., Bouck J., Gibbs R., Hardison R., and Miller W. (2000). PipMaker-a web server for aligning two genomic DNA sequences. *Genome Res* **10:** 577-86.

Sharp P. A., and Burge C. B. (1997). Classification of introns: U2-type or U12-type. *Cell* **91:** 875-9.

Shehee W. R., Loeb D. D., Adey N. B., Burton F. H., Casavant N. C., Cole P., Davies C. J., McGraw R. A., Schichman S. A., Severynse D. M., and *et al*. (1989). Nucleotide sequence of the BALB/c mouse beta-globin complex. *J Mol Biol* **205:** 41-62.

Shen L. X., Basilion J. P., and Stanton V. P., Jr. (1999). Single-nucleotide polymorphisms can cause different structural folds of mRNA. *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., and Sirotkin K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29:** 308-11.

Shimoda N., Knapik E. W., Ziniti J., Sim C., Yamada E., Kaplan S., Jackson D., de Sauvage F., Jacob H., and Fishman M. C. (1999). Zebrafish genetic map with 2000 microsatellite markers. *Genomics* **58:** 219-32.

Shizuya H., Birren B., Kim U. J., Mancino V., Slepak T., Tachiiri Y., and Simon M. (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc Natl Acad Sci U S A* **89:** 8794-7.

Shoemaker D. D., Lashkari D. A., Morris D., Mittmann M., and Davis R. W. (1996). Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat Genet* **14:** 450-6.

Shoemaker D. D., Schadt E. E., Armour C. D., He Y. D., Garrett-Engele P., McDonagh P. D., Loerch P. M., Leonardson A., Lum P. Y., Cavet G., Wu L. F., Altschuler S. J., Edwards S., King J., Tsang J. S., Schimmack G., Schelter J. M., Koch J., Ziman M., Marton M. J., Li B., Cundiff P., Ward T., Castle J., Krolewski M., Meyer M. R., Mao M., Burchard J., Kidd M. J., Dai H., Phillips J. W., Linsley P. S., Stoughton R., Scherer S., and Boguski M. S. (2001). Experimental annotation of the human genome using microarray technology. *Nature* **409:** 922-7.

Shyamala V., and Ames G. F. (1989). Genome walking by single-specific-primer polymerase chain reaction: SSP-PCR. *Gene* **84:** 1-8.

Shyamala V., and Ames G. F. (1993). Genome walking by single specific primer-polymerase chain reaction. *Methods Enzymol* **217:** 436-46.

Sidow A., and Thomas W. K. (1994). A molecular evolutionary framework for eukaryotic model organisms. *Curr Biol* **4:** 596-603.

Simpson A. J. G., de Souza S. J., Camargo A. A., and Brentani R. R. (2001). Definition of the gene content of the human genome: the need for deep experimental verification. *Comparative and Functional Genomics* **2:** 169-175.

Smink L. (2000). Genome studies of human chromosome 22q13.3. PhD thesis.

Smit A. F. (1996). The origin of interspersed repeats in the human genome. *Curr Opin Genet Dev* **6:** 743-8.

Smit A. F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* **9:** 657-63.

Smit A. F., and Riggs A. D. (1995). MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. *Nucleic Acids Res* **23:** 98-102.

Smit A. F., and Riggs A. D. (1996). Tiggers and DNA transposon fossils in the human genome. *Proc Natl Acad Sci U S A* **93:** 1443-8.

Smith S. F., Snell P., Gruetzner F., Bench A. J., Haaf T., Metcalfe J. A., Green A. R., and Elgar G. (2002). Analyses of the extent of shared synteny and conserved gene orders between the genome of *Fugu rubripes* and human 20q. *Genome Res* **12:** 776-84.

Smith V., Chou K. N., Lashkari D., Botstein D., and Brown P. O. (1996). Functional analysis of the genes of yeast chromosome V by genetic footprinting. *Science* **274:** 2069-74.

Soderlund C., Humphray S., Dunham A., and French L. (2000). Contigs built with fingerprints, markers, and FPC V4.7. *Genome Res* **10:** 1772-87.

Solovyev V. V., Salamov A. A., and Lawrence C. B. (1994). Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res* **22:** 5156-63.

Sonnhammer E. L., and Durbin R. (1994). A workbench for large-scale sequence homology analysis. *Comput Appl Biosci* **10:** 301-7.

Sonnhammer E. L., and Durbin R. (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167:** GC1-10.

Spradling A. C., Stern D. M., Kiss I., Roote J., Laverty T., and Rubin G. M. (1995). Gene disruptions using P transposable elements: an integral component of the Drosophila genome project. *Proc Natl Acad Sci U S A* **92:** 10824-30.

Steemers F. J., Ferguson J. A., and Walt D. R. (2000). Screening unlabeled DNA targets with randomly ordered fiber-optic gene arrays. *Nat Biotechnol* **18:** 91-4.

Stoesser G., Baker W., van den Broek A., Camon E., Garcia-Pastor M., Kanz C., Kulikova T., Leinonen R., Lin Q., Lombard V., Lopez R., Redaschi N., Stoehr P., Tuli M. A., Tzouvara K., and Vaughan R. (2002). The EMBL Nucleotide Sequence Database. *Nucleic Acids Res* **30:** 21-6.

Stone D. L., Slavotinek A., Bouffard G. G., Banerjee-Basu S., Baxevanis A. D., Barr M., and Biesecker L. G. (2000). Mutation of a gene encoding a putative chaperonin causes McKusick-Kaufman syndrome. *Nat Genet* **25:** 79-82.

Strausberg R. L., Feingold E. A., Klausner R. D., and Collins F. S. (1999). The mammalian gene collection. *Science* **286:** 455-7.

Stringer C. B., and Andrews P. (1988). Genetic and fossil evidence for the origin of modern humans. *Science* **239:** 1263-8.

Stumpf M. P. (2002). Haplotype diversity and the block structure of linkage disequilibrium. *Trends Genet* **18:** 226-8.

Sturtevant J. (2000). Applications of differential-display reverse transcription-PCR to molecular pathogenesis and medical mycology. *Clin Microbiol Rev* **13:** 408-27.

Sulston J. E., Mallet F., Staden R., Durbin R., Horsnell T., and Coulson A. (1980). Software for genome mapping by fingerprinting techniques. *CABIOS* **4:** 125-132.

Summers T. J., Thomas J. W., Lee-Lin S. Q., Maduro V. V., Idol J. R., and Green E. D. (2001). Comparative physical mapping of targeted regions of the rat genome. *Mamm Genome* **12:** 508-12.

Sun X., Ding H., Hung K., and Guo B. (2000). A new MALDI-TOF based mini-sequencing assay for genotyping of SNPS. *Nucleic Acids Res* **28:** E68.

Sunyaev S., Ramensky V., Koch I., Lathe W., 3rd, Kondrashov A. S., and Bork P. (2001). Prediction of deleterious human alleles. *Hum Mol Genet* **10:** 591-7.

Syvänen A. C. (2001). Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat Rev Genet* **2:** 930-42.

Taillon-Miller P., Gu Z., Li Q., Hillier L., and Kwok P. Y. (1998). Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. *Genome Res* **8:** 748-54.

Talbot W. S., and Hopkins N. (2000). Zebrafish mutations and functional analysis of the vertebrate genome. *Genes Dev* **14:** 755-62.

Tang K., Fu D. J., Julien D., Braun A., Cantor C. R., and Koster H. (1999). Chip-based genotyping by mass spectrometry. *Proc Natl Acad Sci U S A* **96:** 10016-20.

Tateno Y., Imanishi T., Miyazaki S., Fukami-Kobayashi K., Saitou N., Sugawara H., and Gojobori T. (2002). DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic Acids Res* **30:** 27-30.

Tautz D. (1989). Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Res* **17:** 6463-71.

Taylor J. G., Choi E. H., Foster C. B., and Chanock S. J. (2001). Using genetic variation to study human disease. *Trends Mol Med* **7:** 507-12.

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

The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium: Kawai J., Shinagawa A., Shibata K., Yoshino M., Itoh M., Ishii Y., Arakawa T., Hara A., Fukunishi Y., Konno H., Adachi J., Fukuda S., Aizawa K., Izawa M., Nishi K., Kiyosawa H., Kondo S., Yamanaka I., Saito T., Okazaki Y., Gojobori T., Bono H., Kasukawa T., Saito R., Kadota K., Matsuda H., Ashburner M., Batalov S., Casavant T., Fleischmann W., Gaasterland T., Gissi C., King B., Kochiwa H., Kuehl P., Lewis S., Matsuo Y., Nikaido I., Pesole G., Quackenbush J., Schriml L. M., Staubli F., Suzuki R., Tomita M., Wagner L., Washio T., Sakai K., Okido T., Furuno M., Aono H., Baldarelli R., Barsh G., Blake J., Boffelli D., Bojunga N., Carninci P., de Bonaldo M. F., Brownstein M. J., Bult C., Fletcher C., Fujita M., Gariboldi M., Gustincich S., Hill D., Hofmann M., Hume D. A., Kamiya M., Lee N. H., Lyons P., Marchionni L., Mashima J., Mazzarelli J., Mombaerts P., Nordone P., Ring B., Ringwald M., Rodriguez I., Sakamoto N., Sasaki H., Sato K., Schonbach C., Seya T., Shibata Y., Storch K. F., Suzuki H., Toyo-oka K., Wang K. H., Weitz C., Whittaker C., Wilming L., Wynshaw-Boris A., Yoshida K., Hasegawa Y., Kawaji H., Kohtsuki S., and Hayashizaki Y. (2001). Functional annotation of a full-length mouse cDNA collection. *Nature* **409:** 685-90.

The Sanger Institute and Washington University Genome Sequencing Center (1998). Toward a complete human genome sequence. *Genome Res* **8:** 1097-108.

Thiery J. P., Macaya G., and Bernardi G. (1976). An analysis of eukaryotic genomes by density gradient centrifugation. *J Mol Biol* **108:** 219-35.

Thomas A., and Skolnick M. H. (1994). A probabilistic model for detecting coding regions in DNA sequences. *IMA J Math Appl Med Biol* **11:** 149-60.

Thomas J. W., Summers T. J., Lee-Lin S. Q., Maduro V. V., Idol J. R., Mastrian S. D., Ryan J. F., Jamison D. C., and Green E. D. (2000). Comparative genome mapping in the sequence-based era: early experience with human chromosome 7. *Genome Res* **10:** 624-33.

Thompson J. D., Higgins D. G., and Gibson T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:** 4673-80.

Thornton J. M. (2001). From genome to function. *Science* **292:** 2095-7.

Tilford C. A., Kuroda-Kawaguchi T., Skaletsky H., Rozen S., Brown L. G., Rosenberg M., McPherson J. D., Wylie K., Sekhon M., Kucaba T. A., Waterston R. H., and Page D. C. (2001). A physical map of the human Y chromosome. *Nature* **409:** 943-5.

Tilghman S. M., Curtis P. J., Tiemeier D. C., Leder P., and Weissmann C. (1978a). The intervening sequence of a mouse beta-globin gene is transcribed within the 15S beta-globin mRNA precursor. *Proc Natl Acad Sci U S A* **75:** 1309-13.

Tilghman S. M., Tiemeier D. C., Seidman J. G., Peterlin B. M., Sullivan M., Maizel J. V., and Leder P. (1978b). Intervening sequence of DNA identified in the structural portion of a mouse beta-globin gene. *Proc Natl Acad Sci U S A* **75:** 725-9.

Tomer Y., Barbesino G., Greenberg D. A., Concepcion E., and Davies T. F. (1998). A new Graves disease-susceptibility locus maps to chromosome 20q11.2. International Consortium for the Genetics of Autoimmune Thyroid Disease. *Am J Hum Genet* **63:** 1749-56.

Tomer Y., Barbesino G., Greenberg D. A., Concepcion E., and Davies T. F. (1999). Mapping the major susceptibility loci for familial Graves' and Hashimoto's diseases: evidence for genetic heterogeneity and gene interactions. *J Clin Endocrinol Metab* **84:** 4656-64.

Tomer Y., and Davies T. F. (1997). The genetic susceptibility to Graves' disease. *Baillieres Clin Endocrinol Metab* **11:** 431-50.

Touchman J. W., Bouffard G. G., Weintraub L. A., Idol J. R., Wang L., Robbins C. M., Nussbaum J. C., Lovett M., and Green E. D. (1997). 2006 expressed-sequence tags derived from human chromosome 7-enriched cDNA libraries. *Genome Res* **7:** 281-92.

Touchman J. W., Dehejia A., Chiba-Falek O., Cabin D. E., Schwartz J. R., Orrison B. M., Polymeropoulos M. H., and Nussbaum R. L. (2001). Human and mouse alpha-synuclein genes: comparative genomic sequence analysis and identification of a novel gene regulatory element. *Genome Res* **11:** 78-86.

Trofatter J. A., Long K. R., Murrell J. R., Stotler C. J., Gusella J. F., and Buckler A. J. (1995). An expression-independent catalog of genes from human chromosome 22. *Genome Res* **5:** 214-24.

Trofatter J. A., MacCollin M. M., Rutter J. L., Murrell J. R., Duyao M. P., Parry D. M., Eldridge R., Kley N., Menon A. G., Pulaski K., and *et al*. (1993). A novel moesin-, ezrin-, radixin-like gene is a candidate for the neurofibromatosis 2 tumor suppressor. *Cell* **72:** 791-800.

Tyagi S., Bratu D. P., and Kramer F. R. (1998). Multicolor molecular beacons for allele discrimination. *Nat Biotechnol* **16:** 49-53.

Tyagi S., and Kramer F. R. (1996). Molecular beacons: probes that fluoresce upon hybridization. *Nat Biotechnol* **14:** 303-8.

Uberbacher E. C., Xu Y., and Mural R. J. (1996). Discovering and understanding genes in human DNA sequence using GRAIL. *Methods Enzymol* **266:** 259-81.

Van Etten W. J., Steen R. G., Nguyen H., Castle A. B., Slonim D. K., Ge B., Nusbaum C., Schuler G. D., Lander E. S., and Hudson T. J. (1999). Radiation hybrid map of the mouse genome. *Nat Genet* **22:** 384-7.

Vanin E. F. (1985). Processed pseudogenes: characteristics and evolution. *Annu Rev Genet* **19:** 253-72.

Velculescu V. E., Zhang L., Vogelstein B., and Kinzler K. W. (1995). Serial analysis of gene expression. *Science* **270:** 484-7.

Velculescu V. E., Zhang L., Zhou W., Vogelstein J., Basrai M. A., Bassett D. E., Jr., Hieter P., Vogelstein B., and Kinzler K. W. (1997). Characterization of the yeast transcriptome. *Cell* **88:** 243-51.

Venter J. C., Adams M. D., Myers E. W., Li P. W., Mural R. J., Sutton G. G., Smith H. O., Yandell M., Evans C. A., Holt R. A., Gocayne J. D., Amanatides P., Ballew R. M., Huson D. H., Wortman J. R., Zhang Q., Kodira C. D., Zheng X. H., Chen L., Skupski M., Subramanian G., Thomas P. D., Zhang J., Gabor Miklos G. L., Nelson C., Broder S., Clark A. G., Nadeau J., McKusick V. A., Zinder N., Levine A. J., Roberts R. J., Simon M., Slayman C., Hunkapiller M., Bolanos R., Delcher A., Dew I., Fasulo D., Flanigan M., Florea L., Halpern A., Hannenhalli S., Kravitz S., Levy S., Mobarry C., Reinert K., Remington K., Abu-Threideh J., Beasley E., Biddick K., Bonazzi V., Brandon R., Cargill M., Chandramouliswaran I., Charlab R., Chaturvedi K., Deng Z., Di Francesco V., Dunn P., Eilbeck K., Evangelista C., Gabrielian A. E., Gan W., Ge W., Gong F., Gu Z., Guan P., Heiman T. J., Higgins M. E., Ji R. R., Ke Z., Ketchum K. A., Lai Z., Lei Y., Li Z., Li J., Liang Y., Lin X., Lu F., Merkulov G. V., Milshina N., Moore H. M., Naik A. K., Narayan V. A., Neelam B., Nusskern D., Rusch D. B., Salzberg S., Shao W., Shue B., Sun J., Wang Z., Wang A., Wang X., Wang J., Wei M., Wides R., Xiao C., Yan C., *et al*. (2001). The sequence of the human genome. *Science* **291:** 1304-51.

Wagener R., Kobbe B., and Paulsson M. (1998). Genomic organisation, alternative splicing and primary structure of human matrilin-4. *FEBS Lett* **438:** 165-70.

Wall J. D. (2001). Insights from linked single nucleotide polymorphisms: what we can learn from linkage disequilibrium. *Curr Opin Genet Dev* **11:** 647-51.

Wall J. D., and Przeworski M. (2000). When did the human population size start increasing? *Genetics* **155:** 1865-74.

Walt D. R. (2000). Techview: molecular biology. Bead-based fiber-optic arrays. *Science* **287:** 451-2.

Walter M. A., Spillett D. J., Thomas P., Weissenbach J., and Goodfellow P. N. (1994). A method for constructing radiation hybrid maps of whole genomes. *Nat Genet* **7:** 22-8.

Wang D. G., Fan J. B., Siao C. J., Berno A., Young P., Sapolsky R., Ghandour G., Perkins N., Winchester E., Spencer J., Kruglyak L., Stein L., Hsie L., Topaloglou T., Hubbell E., Robinson E., Mittmann M., Morris M. S., Shen N., Kilburn D., Rioux J., Nusbaum C., Rozen S., Hudson T. J., Lander E. S., and *et al*. (1998a). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280:** 1077-82.

Wang P. W., Eisenbart J. D., Espinosa R., 3rd, Davis E. M., Larson R. A., and Le Beau M. M. (2000). Refinement of the smallest commonly deleted segment of chromosome 20 in malignant myeloid diseases and development of a PAC-based physical and transcription map. *Genomics* **67:** 28-39.

Wang P. W., Iannantuoni K., Davis E. M., Espinosa R., 3rd, Stoffel M., and Le Beau M. M. (1998b). Refinement of the commonly deleted segment in myeloid leukemias with a del(20q). *Genes Chromosomes Cancer* **21:** 75-81.

Wang Z., and Moult J. (2001). SNPs, protein structure, and disease. *Hum Mutat* **17:** 263-70.

Wasserman W. W., Palumbo M., Thompson W., Fickett J. W., and Lawrence C. E. (2000). Human-mouse genome comparisons to locate regulatory sites. *Nat Genet* **26:** 225-8.

Watanabe T. K., Bihoreau M. T., McCarthy L. C., Kiguwa S. L., Hishigaki H., Tsuji A., Browne J., Yamasaki Y., Mizoguchi-Miyakita A., Oga K., Ono T., Okuno S., Kanemoto N., Takahashi E., Tomita K., Hayashi H., Adachi M., Webber C., Davis M., Kiel S., Knights C., Smith A., Critcher R., Miller J., James M. R., and *et al*. (1999). A radiation hybrid map of the rat genome containing 5,255 markers. *Nat Genet* **22:** 27-36.

Waterston R., and Sulston J. E. (1998). The Human Genome Project: reaching the finish line. *Science* **282:** 53-4.

Waterston R. H., Lander E. S., and Sulston J. E. (2002). On the sequencing of the human genome. *Proc Natl Acad Sci U S A* **99:** 3712-6.

Watson J. D. (1990). The human genome project: past, present, and future. *Science* **248:** 44-9.

Weber J. L. (1990). Human DNA polymorphisms and methods of analysis. *Curr Opin Biotechnol* **1:** 166-71.

Weber J. L., and May P. E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am J Hum Genet* **44:** 388-96.

Weiss K. M., and Clark A. G. (2002). Linkage disequilibrium and the mapping of complex human traits. *Trends Genet* **18:** 19-24.

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

Wenderfer S. E., Slack J. P., McCluskey T. S., and Monaco J. J. (2000). Identification of 40 genes on a 1-Mb contig around the IL-4 cytokine family gene cluster on mouse chromosome 11. *Genomics* **63:** 354-73.

Wheelan S. J., Boguski M. S., Duret L., and Makalowski W. (1999). Human and nematode orthologs-lessons from the analysis of 1800 human genes and the proteome of *Caenorhabditis elegans*. *Gene* **238:** 163-70.

Wheeler D. L., Church D. M., Lash A. E., Leipe D. D., Madden T. L., Pontius J. U., Schuler G. D., Schriml L. M., Tatusova T. A., Wagner L., and Rapp B. A. (2002). Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res* **30:** 13-6.

White K. P. (2001). Functional genomics and the study of development, variation and evolution. *Nat Rev Genet* **2:** 528-37.

Wiemann S., Weil B., Wellenreuther R., Gassenhuber J., Glassl S., Ansorge W., Bocher M., Blocker H., Bauersachs S., Blum H., Lauber J., Dusterhoft A., Beyer A., Kohrer K., Strack N., Mewes H. W., Ottenwalder B., Obermaier B., Tampe J., Heubner D., Wambutt R., Korn B., Klein M., and Poustka A. (2001). Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res* **11:** 422-35.

Wiginton D. A., Kaplan D. J., States J. C., Akeson A. L., Perme C. M., Bilyk I. J., Vaughn A. J., Lattier D. L., and Hutton J. J. (1986). Complete sequence and structure of the gene for human adenosine deaminase. *Biochemistry* **25:** 8234-44.

Wilcox A. S., Khan A. S., Hopkins J. A., and Sikela J. M. (1991). Use of 3' untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to STSs: implications for an expression map of the genome. *Nucleic Acids Res* **19:** 1837-43.

Wilson M. D., Riemer C., Martindale D. W., Schnupf P., Boright A. P., Cheung T. L., Hardy D. M., Schwartz S., Scherer S. W., Tsui L. C., Miller W., and Koop B. F. (2001). Comparative analysis of the gene-dense ACHE/TFR2 region on human chromosome 7q22 with the orthologous region on mouse chromosome 5. *Nucleic Acids Res* **29:** 1352-65.

Woods I. G., Kelly P. D., Chu F., Ngo-Hazelett P., Yan Y. L., Huang H., Postlethwait J. H., and Talbot W. S. (2000). A comparative map of the zebrafish genome. *Genome Res* **10:** 1903-14.

Yaspo M. L. (2001). Taking a functional genomics approach in molecular medicine. *Trends Mol Med* **7:** 494-501.

Yu A., Zhao C., Fan Y., Jang W., Mungall A. J., Deloukas P., Olsen A., Doggett N. A., Ghebranious N., Broman K. W., and Weber J. L. (2001). Comparison of human genetic and sequence-based physical maps. *Nature* **409:** 951-3.

Zambrowicz B. P., Friedrich G. A., Buxton E. C., Lilleberg S. L., Person C., and Sands A. T. (1998). Disruption and sequence identification of 2,000 genes in mouse embryonic stem cells. *Nature* **392:** 608-11.

Zamore P. D. (2001). RNA interference: listening to the sound of silence. *Nat Struct Biol* **8:** 746-50.

Zdobnov E. M., and Apweiler R. (2001). InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17:** 847-8.

Zdobnov E. M., Lopez R., Apweiler R., and Etzold T. (2002). The EBI SRS server-recent developments. *Bioinformatics* **18:** 368-73.

Zhao S., Shatsman S., Ayodeji B., Geer K., Tsegaye G., Krol M., Gebregeorgis E., Shvartsbeyn A., Russell D., Overton L., Jiang L., Dimitrov G., Tran K., Shetty J., Malek J. A., Feldblyum T., Nierman W. C., and Fraser C. M. (2001). Mouse BAC ends quality assessment and sequence analyses. *Genome Res* **11:** 1736-45.

Zhou B., Westaway S. K., Levinson B., Johnson M. A., Gitschier J., and Hayflick S. J. (2001). A novel pantothenate kinase gene (PANK2) is defective in Hallervorden-Spatz syndrome. *Nat Genet* **28:** 345-9.

Zouali H., Hani E. H., Philippi A., Vionnet N., Beckmann J. S., Demenais F., and Froguel P. (1997). A susceptibility locus for early-onset non-insulin dependent (type 2) diabetes mellitus maps to chromosome 20q, proximal to the phosphoenolpyruvate carboxykinase gene. *Hum Mol Genet* **6:** 1401-8.

Zwaal R. R., Broeks A., van Meurs J., Groenen J. T., and Plasterk R. H. (1993). Target-selected gene inactivation in *Caenorhabditis elegans* by using a frozen transposon insertion mutant bank. *Proc Natl Acad Sci U S A* **90:** 7431-5.

Appendices

Appendix 1: A genomics timeline (reproduced from Lander *et al*., 2000).

1800s

1865	Gregor Mendel reports the results of his pea plant experiments, from which he discerned several fundamental laws of heredity. His results, which appeared in an obscure journal article in 1866, were ignored for 34 years.
1882	Walther Flemming publishes his observations of tiny threads--later known as chromosomes--inside salamander larvae cells that appear to be dividing.

1900s

1900	Hugo de Vries in the Netherlands, Erich Tschermak von Seysenegg in Austria, and Karl Correns in Germany simultaneously rediscover and verify Mendel's principles of heredity.
1902	Walter Sutton points out connection between chromosomes and Mendel's "factors," thereby expanding the science of genetics from the organismal level to the subcellular level.
1910	Thomas Hunt Morgan and co-workers in the "fly lab" show that some genetically determined traits are sex linked. They also confirm that some trait-determining genes are located on specific chromosomes.
1927	Working with fruit flies, Hermann Muller determines that x-rays can cause genetic mutations.
1928	Fred Griffith discovers the phenomenon of transformation, in which some unknown "principle" transforms a harmless strain of bacteria into a virulent one.
1944	Oswald Avery, Colin MacLeod, and Maclyn McCarty prove that DNA, not protein, embodies the heredity material in most living organisms.
Late 1940s	Erwin Chargaff discovers one-to-one correspondence between adenine and thymine and between cytosine and guanine--a key piece of information for determining the structure of DNA.
1952	Rosalind Franklin obtains x-ray diffraction data of DNA, which become central to the elucidation of DNA's molecular structure. Martha Chase and Alfred Hershey report experiments with bacteriophages that help prove DNA is the molecule of heredity.
1953	James Watson and Francis Crick announce their discovery of the double-helix structure of DNA. They write in a 958-word Nature article: "It has not escaped our notice that the specific pairings we have postulated immediately suggest a possible copying mechanism for the genetic material."
Mid-1960s	Marshall Nirenberg, H. Gobind Khorana, and others crack the triplet code that maps messenger RNA codons to specific amino acids.
1969	A team at Harvard Medical School led by Jonathan Beckwith isolates the first gene, specifically, a bacterial gene whose protein product is involved in sugar metabolism.
1970	A team at the University of Wisconsin, led by H. Gobind Khorana, synthesizes a gene from scratch, beginning what might be called chemical genetics.
1972	Using restriction enzymes from Herbert Boyer's research group, Paul Berg and colleagues produce the first recombinant DNA molecules.

1973	The era of genetic engineering begins when Stanley Cohen, Herbert Boyer, and co-workers insert a gene from an African clawed toad into bacterial DNA.
1976	Genentech, the first genetic engineering company, is founded in South San Francisco.
1983	James Gusella and co-workers locate a genetic marker for Huntington's disease on chromosome 4. This leads to scientists having the ability to screen people for a disease without being able to cure it. Meanwhile, Kary Mullis conceives of the polymerase chain reaction, a chemical DNA replication process that will greatly quicken the pace of genetic science and technology development.
1984	Alec Jeffreys develops "genetic fingerprinting," a molecular biological analog of traditional fingerprinting for identifying individuals by analyzing polymorphic (variable) sequences in their DNA.
1986	The Human Genome Initiative, later called the Human Genome Project, is announced. The goal is to sequence the entire human genome and provide a complete catalog of every human gene.
1987	A large, collaborative effort yields the first comprehensive human genetic map with 400 signposts.
1988	The National Center for Human Genome Research is created, with the goal of mapping and sequencing all human DNA by 2005.
1990	The Human Genome Project is formally launched, with a completion date set for 2005. W. French Anderson performs the first gene therapy procedure on a 4-year-old girl with an immune disorder known as ADA deficiency. (It didn't work.) Mary-Claire King finds evidence that a gene on chromosome 17 causes an inherited form of breast cancer and increases the risk of ovarian cancer.
1992	An international collaboration produces a rough map of genetic polymorphism: the variable genetic regions along all 23 pairs of human chromosomes that govern person-to-person biological variation.
1995	The Institute for Genomic Research reports the first complete DNA sequence of the genome of a free-living organism--the bacterium Haemophilus influenzae.
1996	The first complete sequence of the genome of a eukaryote (the yeast Saccharomyces cerevisiae) is reported by an international effort involving some 600 scientists in Europe, North America, and Japan.
1998	The first genome of a multicellular organism--the 97-megabase DNA sequence of the roundworm Caenorhabditis elegans--is published by the C. elegans Sequencing Consortium.
	2000
2000	Completion of the first draft of the entire human genome sequence.

Appendix 2: List of primers designed and used for gene identification. Additional information for each STS can be obtained at http://webace.sanger.ac.uk/cgi-bin/webace?db=acedb20&class=STS.

STS	Sense primer	Antisense primer	Parent sequence
stSG64569	ATGTCCTTCATCCTCAACGC	TGGATTTAGGGGCTGAGTTG	dJ453C12
stSG64570	GCGCCTCTCTCAACTTCATC	CAAGAGCTTCTTCCAGGGG	dJ453C12
stSG64571	CCGTGGAGTACATGGAACG	TCTGAGGTTCTCCAGCAGGT	dJ453C12
stSG64572	TCTTGATGTCTCTCAGCTGCA	TGGTTCACTTTTCCAGGGTC	dJ48A11
stSG64573	CTGGGTCTGGCCAATGAC	CACAGCATAGCGCAGAAAGT	dJ453C12
stSG64574	TACGGATGATCTGCAATGGA	ATATCCTCCGGCTCAAAGGT	dJ453C12
stSG64575	GGAGCGGCAAAAATTCTTC	CAGCTCCTCCAAATGGTTGT	dJ453C12
stSG64576	CAGTCTGAAGAAACCCAGGTG	GTTCAAGTCAGACCCATGGA	dJ179M20
stSG64577	CATCTCAGATGCAGACAAGAGG	TCTCCTTGGGCAGAGTGC	dJ453C12
stSG66868	AATGAGCCTCCCCTCCAG	AGAATGGATGTCCGAGGAGT	dJ1119D9
stSG66869	TGGGCTGCTGGTTCTACAG	GGGGTATCTCCTTGAGCTCC	dJ453C12
stSG66870	CTTTCCTCTGTCCCATCTGC	GGGTCCCCAAGGTCAATAAG	dJ453C12
stSG66871	TCTTCTCATACCCTGGCACC	CTGTGGTTGTGACCAGTTGG	dJ715N11
stSG66873	GGACAACATGGTGAAATCCC	AGAGGCTGGTGCATCTGTTT	dJ715N11
stSG66874	TGTCTCATGGCAGAGTTTGC	GGTCAATGCGATATCGGC	dJ1119D9
stSG66875	ATCTTCACCAGGTTGATGTGG	GTGAAGCTGCTCCTCTCCC	dJ28F12
stSG66876	AACAGCAACAACGACCGAG	CGTGTGGCTATCCCCAAG	dJ816K17
stSG66877	AACAGCTTCAACTGGGCG	GGCCTTTTGTTTCAGGAACA	dJ28H20
stSG71391	CTGAAACTCATTTGGCTGCA	CAACAGATTCTAGGGCACTGC	dJ931K24
stSG71392	GGGAAGGGCTAAGATCTTGG	ATGTTTGTCTATGGTGGCTGC	dJ461P17
stSG71394	TCTATGAGAAGTTCTCTCCCGC	TTCTTGATGAAGATGTTGCCC	dJ148E22
stSG71395	GCTGGGTACCTTCCACAGC	AGCCCACTGTCGTGCACT	dJ453C12
stSG71396	TGGAGGCCTCAGTAGGCTTA	AGGCCTCACGTGCTTCAC	dJ931K24
stSG71396	TGGAGGCCTCAGTAGGCTTA	TCTCCAAGACCTGGCTCAGT	dJ931K24
stSG71396	TTGTGAGTCTCATTGATGGTCC	AGGCCTCACGTGCTTCAC	dJ931K24
stSG71396	TTGTGAGTCTCATTGATGGTCC	TCTCCAAGACCTGGCTCAGT	dJ931K24
stSG71398	CTCTGGCTCCTGGAACTTTG	TCAGCATAGGTCTCGGTGC	dJ1183I21
stSG71399	TGATGTACATGCTCCTGTAATGC	CTGCCATCTTCCTAGAATCCC	dJ738P15
stSG71400	TTCACAGTGACTCTTGGAATTCC	TACCACCAAGACCTGGAAGG	dJ971N18
stSG71402	ATGCGGAGGAGGAAAAAGAT	TTTACGCTGCCTCAAGGAGT	dJ971N18
stSG71403	CAGGAACCCACGTTGATTTT	TTGCATAGTCATTGTCCCCA	dJ569M23
stSG71404	GGCTGGGGAACTTTCTTTTC	TCAACACAGGAATGAGGCAG	dJ569M23
stSG71405	GAGAACCAGTGGGGAGACTG	TGACAAATGTCTGCTGCCTC	dJ569M23
stSG71406	GATGGCATCTTGCTCTGTCA	CCAGCTACTCCTACTGGGGA	dJ569M23
stSG71407	CTTTGCTTCTTCCTCCTCCC	GAGGGCCTAGCCTCTCTTGT	dJ569M23
stSG71408	CTCCCCCGGCTAATTTGTAT	CACATATTGGCATGCTGACC	dJ209B9
stSG71409	TGGATGCCATTGCCTTTATT	AAATGTATGGACCCCAGACG	dJ461P17
stSG71414	TGTCTCATGGCAGAGTTTGC	GGTCAATGCGATATCGGC	dJ1119D9
stSG74158	GAGATGGAGTCTCTGTCGCC	TTAGCTGGGCATGGTGGT	dJ994L9
stSG74159	ATCCTCCCCAAGGAGTATGG	GGAGTCGTCACACTGTGCAT	dJ179M20
stSG74160	AATTTCCCATTTATTCCCCG	TTCTCCCCACTCCAGAAAAA	dJ738P15
stSG74161	CGTTTGGAAGCTAAGAAGCC	CACGTAACCTCCAAAGCCAT	dJ931K24
stSG74162	ATCCTCATTACCCAGGTCCC	GCATCTGGGAGCTTCTTGTC	dJ28H20

STS	Sense primer	Antisense primer	Parent sequence
stSG74163	ATCCTGCTCTCCTATGCCCT	GAAGAGGAGGCTGAGGGATT	dJ28H20
stSG74164	TTTCCCATGGGGTCGTAGTA	GATGACAATGGTGTGCTTGC	dJ816K17
stSG74165	CAAAGGCACAATGATCTCCA	TCTCTCGTGATCTGTGGACG	dJ1100H13
stSG76865	ACCAGAGGACACACTCAGGG	TGGCATTCATAAGGCCTCTC	dJ694B14
stSG76866	GAAGCCTTTTCTGTGCAAGG	TTAAACCTTCGCCCACACTC	dJ694B14
stSG76867	AAAGGTGGTGCAGCCAATAG	GTCTGTCTGTCCCCATTCGT	dJ694B14
stSG76868	GTCCATATGAACGGGAAGGA	TCCTGCTTCATCTGCTCAAA	dJ148E22
stSG76869	ATCCCTGGACTCCATTGTGA	AACCTTCTGGATGGACTGGA	dJ343K2
stSG76870	CAACTCAAGCACAAAAGCCA	CTTGGGTAGCGTCAGAGGAG	dJ1183I21
stSG76871	TGGGATGGGTACTGAGGAAG	AGACAATTGGGCAGGTTTTG	dJ28H20
stSG76872	GATATCCAACCTGACGTGGC	TGAGGAACATTTGCCAAAAA	dJ715N11
stSG76873	TGCTTTTGGTGTTACATCTAAGAAAT	GCCTAGGCAACGTGGATATACT	dJ715N11
stSG76874	AAGTGACTCCTACAGTCCCCC	CCAACCTCCTGGAAAAGTGA	dJ1100H13
stSG76875	AGCCTGTCAGAGGAGAGCAG	CCAAAGAACTGAAAGCCTCG	dJ1054C24
stSG76876	GCGAGGAGAGATCATCGACA	GATCTCGGCCATGATGTAGC	dJ633020
stSG76877	AGAAGATCTCCCGGAGCAGT	AGAGGGGAAAGATGGTTCGT	dJ633020
stSG76878	TTCACCTACACCCCGGAATA	ACTCCTGATGGATGCTCTCC	dJ453C12
stSG76879	AAACAGAAAACTAACATTGGAACCT	CAAAATACCAATGTCTCTAGGAAGG	dJ1178H5
stSG76880	TACTTCAGTCACCACCAGCG	AACGCAGGATCTCCTTCAGA	dJ1178H5
stSG76881	GAGGTGAACCAGAAGCCAAG	TTGCTTGGTGATCATTTTGC	dJ1178H5
stSG76882	TCAGGAGTAGCAGCCAAGGT	GATGCAGCTTCTCTGGCAAT	dJ1178H5
stSG76883	TCAGGAGTGATACTTCACAGATCC	TCTAGCCGGTCCACTTTG	dJ850H21
stSG76884	GATCCCCAATGTTGACGTTT	ATCCTAAGCAGTGGATTGGC	dJ850H21
stSG76885	CCCCCGAGTCATCTTCATTA	GAGGATCTGCACATCACCCT	dJ453C12
stSG76886	CACCAAAGCTGTGTTCCAGA	GTGAGAGGAAAGCCTGATGC	dJ816K17
stSG76887	GGACATCCTGAACATCTGCC	CTGGTGACGTAGATGGGGTT	dJ816K17
stSG76888	CTTCGTGTTTGCGGAGGT	GCATCTCCCAATCTTCTTCG	dJ816K17
stSG76889	GCAGCTTCCTGTGTAAAGGC	GTATCCCTGCTCAGCCTCTG	dJ730D4
stSG76890	GGCAAGACGTTTGGTTTCAT	GGACTGCCATTTTCTGGATT	dJ1119D9
stSG76891	AAGGTGCTAGAGCCTCCCAT	GGGTATAGAGGATGGTGGCA	dJ734P14
stSG76892	GGTCACTTCTGACAGGGGAA	AACCAATACGAACTCGACCG	dJ963K23
stSG76893	AACTGCTACTGGAAAGGAGAAA	TCTGGAATGAGGAACAAGCC	dJ963K23
stSG76894	CCAAACAGGGAAGGAACGTA	TGTACGATCAAGCTGGCATT	dJ576H24
stSG76895	GTCTGCCTGGAAATTCTGGA	CATCCTTTTTACCCCAGCAA	dJ28H20
stSG76896	AGTCCTTGGTCTGCTGGTGT	TTGACAAGCAGGTTTGTTGG	dJ28H20
stSG76897	TCAACCCTCATCAGACACCA	GGTTTCTCACCGGAGTGTGT	dJ734P14
stSG76898	GGCCTGACCTTTGGTTATGA	TGCCATAGCAGTCAATGAGG	dJ28H20
stSG76899	CACCTATGAGCGAATTGATGGGCG	CACGTGGTCTCGCCCTCCGGACCCC	dJ620E11
stSG76900	TGCCATCCTCGAGAAGAACT	CCTGCAATCAGCTTAGGGAG	dJ620E11
stSG76901	AATGGTGTACAGAGACGCCC	TATGTTCCAGGGCCATAAGC	dJ620E11
stSG76902	ACTAACTGGGAGCGGGAGTT	GGGCGTCTCTGTACACCATT	dJ620E11
stSG76905	AGTGCAGCTGTTCCTGGTCT	TTCGTCTGTTTTGTTCGCAG	dJ28H20
stSG76906	AGATGTGGTTGTAGGGCAGC	GACGGCGTCTGTGACAAGTA	dJ1065O2
stSG76907	GTGAATATCCCCAACATGGC	TGGAGGTGAGAGATGTGCAG	dJ576H24
stSG76909	CAAATTGACGCAAATAGCGA	CCAGACCAAAGGATGTGGAT	dJ569M23
stSG76910	GGCCTGACCTTTGGTTATGA	GTCCTCATTGGTCCTTGGAA	dJ28H20

STS	Sense primer	Antisense primer	Parent sequence
stSG76911	AAATCGCCCTCAGAATATGG	AAGTTACACGGGAGGGAGGT	dJ1183I21
stSG76912	CATGGCACTGACTAGCCGAC	GTCCACTCGCCCTTGTAGAG	dJ1183I21
stSG76913	GCTGGAGCTTTGCCTCTCTA	GATCGAAGTGTCTGAGCGTG	dJ453C12
stSG76914	GGAGCTCAAGGAGATACCCC	CATTTCCAGGGTTCCTCTCA	dJ453C12
stSG76915	CTGCAGGAGCTGACATGGAC	GACTCGAGATGCAGATGGGA	dJ453C12
stSG76916	AGACTGACCACGCTGGCTAT	GGTACACAATCCCACCATCC	dJ453C12
stSG76917	CTAGCAGCAGGTGATGGTGA	TTGGCTATGAATACCTGGGC	dJ257E24
stSG76918	CAGTCCCCCGTACAGGAAG	GTGTGACGTGGTGAATTTCG	dJ257E24
stSG76919	TGGACTCCAACAAGCATCAG	ATCGTCCTCTTCCCCTTCAT	dJ257E24
stSG76922	TTTGACTGCTGTTGTGACCC	GGGTGAAGAATGGTCAAGGA	dJ620E11
stSG76923	GGGATGGACCTGGTAGGACT	CCACTTTTCGTTGTGCTTGA	dJ620E11
stSG76924	TCATTTCCCGTTTGGACAAG	TTCCATGTGGGTAGACGACA	dJ620E11
stSG76926	AGGTTGTGGGAAGTGAGGC	TCTTCACGTTTCAGCACCAC	dJ967N21
stSG76927	AAGAGGGACAGTGGTTCGTG	TCCATCTGGAAGAGGAAAGC	dJ967N21
stSG76928	AGGCCCAAAAGTTTACCAGC	CGTGCACGCATATCTTCACT	dJ967N21
stSG76929	GTGTCCCAGGCTGCTGAG	AACAGTGGTTCCGCTGGC	dJ967N21
stSG76930	AGCTTTCACCACAGCTGCAT	GCCCTGTTCCCATTGATGTA	dJ967N21
stSG76931	TCTCTGTGCTCTCGATCTGC	GGAATGTCTGAAGCCGAAGA	dJ963K23
stSG76932	CCAGGGTTATGTCCCAAAGA	TTGAAAGGGATCCATGCACT	dJ461P17
stSG76933	CTAGCAGCAGGTGATGGTGA	TTGGCTATGAATACCTGGGC	dJ257E24
stSG76934	TGGACTCCAACAAGCATCAG	ATCGTCCTCTTCCCCTTCAT	dJ257E24
stSG76935	GGACACCCTCTCTAGGGTCC	GCTTGTGCATTCAGACCAGA	dJ718J7
stSG76936	CAGGGTGTACCACGTAGGC	CCCACCATCCTTGTCATCTT	dJ1049G16
stSG76938	AAGAGGTCACCAAGGGCAG	CGGGTAGAGGAGCAAAACAA	dJ453C12
stSG76939	GTTGCCATGGAGACAGGC	AGCCGTAATACACGGTCTGC	dJ453C12
stSG76940	AACATCATGATAGGGCCTGG	TTTCATTCAGTTTCACCCCC	dJ620E11
stSG76943	CTCACCCCAGATGAAGGTGT	AAAGCATAGGCCACCATGAC	X59747
stSG76944	TGGACTTCGAATCCCAGC	CTCGTCCACGTCGGTCAC	dJ998H6
stSG76945	AGCTGCTGACTGCAAGGTCT	AGTGGGTGAGAAACAGGAGG	dJ816K17
stSG76946	AGCTGCTGACTGCAAGG	AGTGGGTGAGAAACAG	dJ816K17
stSG76946	AGCTGCTGACTGCAAGG	GGGTCTCCTGATTCCCTCTC	dJ816K17
stSG76946	GGAAAAACCTCTGCATTGGA	AGTGGGTGAGAAACAG	dJ816K17
stSG76946	GGAAAAACCTCTGCATTGGA	GGGTCTCCTGATTCCCTCTC	dJ816K17
stSG76947	CCTTGTCCCTCCTGAGAAGA	CTGCTGCTGATCCCTGATG	dJ816K17
stSG76948	GGGGGATCCTTGAGGAAGTA	CTAAGCACAGCCTCTCTGGG	dJ816K17
stSG76949	TTAGATAGAGGAAGCTGGGGG	ATGCGGTTCTTGAGGTAGGA	dJ816K17
stSG76950	TATGGAAGATGAGGCCTTGC	CACAGACAAAGATGGGGTGG	dJ816K17
stSG76951	GGTAAGGAAAAGCAGGGGTC	ACTCTAGGCAGCTTCACCCA	dJ816K17
stSG76952	GTAATAAAGGCCGCCATGTG	TTTCCTTACCCATGCTGCTC	dJ816K17
stSG76953	TGAAAAAGGAGGTGGTTTGG	ATTAAAGGAGCAAGGGGGTG	dJ734P14
stSG76954	AGTTAAAGGCACAAACCCCC	TTCCCTTCCTTCTTCCCAGT	dJ734P14
stSG76955	ACAAGGAAGAGAGCCCCATT	TGGGGGTGGACATGACTAAT	dJ734P14
stSG76956	GAGACAAAGAATGGGCAGGA	ACAGTCTGCCTTTCTCCCCT	dJ734P14
stSG76957	GCTGTTAGTCACTGGCCCTC	TCCTGCCCATTCTTTGTCTC	dJ734P14
stSG76958	ATGTAGATGGCCACTCGGTC	AGGGGAGACCACTCAGATCA	dJ18C9
stSG76959	GGCCATATCAATGACTCCCA	AACCATGGCCACAATGAAGT	dJ18C9

STS	Sense primer	Antisense primer	Parent sequence
stSG76960	TTCTGCCCCAAGATCAAATC	TTCTTCAGCTTTGCCACCTT	dJ18C9
stSG76961	CGACCTCACAGACGACACAG	ATCTCCACTGTCTGCCTCCA	dJ1085F17
stSG76962	CTGACTCTTGCAGGGGTAGC	GGTAGACCCAGGAAGGGAAA	dJ1085F17
stSG76963	AGCTCCTGTTGGCCATTCTA	CTCGGAAGATCCTGAAGTGG	dJ322G13
stSG76964	CGAATTAGCCTCTGGACTCG	CTGGGTCACTTCCAAAGGTG	dJ322G13
stSG76965	TGTTGCTTCGCATCAAAAAG	GCCATTAAATAGGCATCCCA	dJ322G13
stSG76966	TGCTGGTACAGTGAATTCCG	TGTGTGAAACATGTGGCAAG	dJ322G13
stSG76967	GAGTGGCCTTTCATCAAACC	AGGCGTTCTGCTTTCTGTGT	dJ322G13
stSG76968	GCCATGTGAAACCGTCAGTA	ATGACGTTCCCCAGAGTCAC	dJ777L9
stSG76969	ACCAATGTGGCAGAAACCTC	TAGCTGATGCCTGGGAGACT	dJ616B8
stSG76970	CGTCTTTCAGCTGGACATCA	GCTAAGCAGAACCTTGCCAC	dJ1068E13
stSG76971	CCTCAGCTGGAGAAGGACAC	GGTGGTTCTGTACGTGGCTT	dJ717M23
stSG76972	TCCTGGAAACCACATCCTTC	ATGTGCCAGGAACTTCAACC	dJ1112F19
stSG76973	TCCGGTCCTGGAAGTAAATG	TGTGGAAGAGCCACTACCCT	dJ1103G7
stSG76974	ATGTGATACCTTTGAGGGCG	ACGCTACAGTTCTCCGCAGT	dJ1103G7
stSG76975	TCTTCCCCTCCTCTTCCAAT	GACCAGCATCTCAACTGCAA	dJ1103G7
stSG76976	ACATCAAGAGGCCGATGAAC	GGTGGGCTCAGTAGGTGAAA	dJ1103G7
stSG76977	GTTCCTGGTGTTGGATCAGC	CTATCATGGCCCAGAAGAGG	dJ1103G7
stSG76978	GAGATACTCAGCTCACGGGC	AGTCCTGGAAGGGGTAGTGG	dJ1103G7
stSG76979	GAAAGAAAAGGTCCCAAGGC	CTCCAATTCTGTCCAAGGGA	dJ1068E13
stSG76980	CTGGGGCATGTTCTTCTCAT	GGAGGCCTGTGCTTATTTGA	dJ1068E13
stSG76981	TTAGATAGAGGAAGCTGGGGG	ATGCGGTTCTTGAGGTAGGA	dJ816K17
stSG76982	GGTCCCCCAAACTCTCATTT	TCTGAACCTGGCACAGTGAG	dJ453C12
stSG76983	TCCAGGGTATCAAGGTCCAC	GAGCTCCTCTTCCCACCTGT	dJ620E11
stSG76984	TGAAGATGGGTTGCAGCATA	GGACAGGATTTGACCACGAT	dJ816K17
stSG76985	CTTCTGGCAAGGACCTGATG	CCTCGCCCACTTCATTAGAA	dJ816K17
stSG76986	TCCTTGAATGAACAAGGGCT	AACTTACACCCTTCCCCCAC	dJ816K17
stSG76987	TTTTCTAGCTTAGGCCAGGG	CTCCACAGTCTGACCAGTGC	dJ816K17
stSG76988	AGATGTGGGACTTCCCATTG	TCCGAGGATTCTAGTCGCTG	dJ816K17
stSG76989	GTGTGCGTGCCTAGCTCATA	GTTCCAGAGGCCACAGAGTC	dJ816K17
stSG76990	TGGAGATGGTACCTTGGCAT	ATCTCCCCACCTTGTCCTCT	dJ816K17
stSG76991	GTCCCAGAAGGATGGTGAGA	AGGTACCTTGGGTGTGGTGT	dJ816K17
stSG76992	GGGCTTTGGAGTTACTTCCC	GGGCAAAGATGTGCTGTCTT	dJ816K17
stSG76993	CCCACAATAGAACCTCTCGC	ACTGGGTAGGTTGTGGCCTT	dJ816K17
stSG76994	GATCTCCTCCAACTGCGG	CCACCACACAGGCTCTGAC	dJ816K17
stSG76995	GTGGCTCACAGGGACAGTG	GGCATCTGTTCTGAGGAAGG	dJ816K17
stSG76996	AGCTCTGTCTGTCCCACCC	CACCCCCTTCACTCAGTCAC	dJ816K17
stSG76997	GCTCCGTGTTGTCCAAGTTT	ACAGGACAGGAGGTCACAGG	dJ816K17
stSG76998	AGATCAAGCCCACTCCCTTT	CGTTTTGATATCCTGCCCTC	dJ816K17
stSG77027	GGGCTTGCTGTACTTCATCC	GGGAAACGGGAGCTGTAGA	AI226097
stSG77046	TATGAAGCAGCCACTCACCA	TTCAGGACAAAACCAGTCCC	AA178103
stSG85180	CTGATTTATAAACCCCGCCA	CCAGCTGTCTGGTGACAGAG	dJ686N3
stSG85181	TAGCCAACTCCTGGTCTGCT	AAGGGGATGAGGAGGAAGAA	dJ686N3
stSG85182	ATGCTGCCAACCAACCTAAC	CACTCTCATCGACGTCCTCA	dJ1002M8
stSG85183	ACTCTTCTCAATGGGGCAAA	GACCTCCAAAACTGGGTGAA	dJ828K20
stSG85185	CGAACATCGAAGACATCCTG	AACATTCCCAAGTTCATGCTG	dJ998C11

STS	Sense primer	Antisense primer	Parent sequence
stSG85186	GCCTGTGCCATACACCTCTAA	AGATGTCATCAACATGGGCTC	dJ718P11
stSG85187	GCAGAGGAACAACAGAAGCA	GGCTGCAGAGAGTACAGCAA	dJ718P11
stSG85188	CCAGGGCTCCATACACTGAT	GGGATATTGCCCAGAGTTTG	dJ300I2
stSG85189	TTCATCTGCACTGCCAAGAC	TCTCCTGAGCTACGGAAGGA	dJ269M15
stSG85190	TCAAACTCCGAAGACAGCCT	CCTTTTTGCACGGAGAAGAC	dJ1018E9
stSG85191	GGATGGGACATTTCCTTGAC	CCTGCACACACAATGGAAAG	dJ824F16
stSG85192	TTATCGGTGCAATGTTTGGA	CCAGGTCATTGACAGCAGTG	dJ842G6
stSG85193	CCTGTTGCATACATCTTGGC	TGGCTCTGGAATATGGAAGG	dJ842G6
stSG85194	TTTGGAATCATGAAGGGAGG	CTGCTAAGCTTTGTTTCCCG	dJ447F3
stSG85195	GCGGTGATTGTCCAAAAGTT	AGGACTGGTGGGACACAGAA	dJ447F3
stSG85196	CTCCTGGATGGTGGAGAAGA	CATCTAGGAGCAGTCCCAGC	dJ447F3
stSG85197	AAAACCAGGTCTGGGAAGGT	TGCTTGTCCTTGCACTTCTG	dJ568C11
stSG85198	CCCTGCTTGGAATGTTTTGT	CAGGGCCAAGAGAAATGTGT	dJ686C3
stSG85206	TGCTGCTGGAGATTGACAAC	TTAGGGCTGAGAGCCAGAAA	dJ1069P2
stSG85207	GCTGTGCAGTCCAGCATTT	AATGCTTCCTTCTGCTGCTC	dJ644L1
stSG85208	AGATGTGAGGAGGACCATGC	GTTCTGAAAGGCAGAGTGGC	dJ970A17
stSG85209	AACCGTAATTGTGGGCTGAG	CTGAGCAGCATTCCAACGTA	dJ511B24
stSG85210	GGTGACAGAGGTGAGCAACA	ACACTCACTCTGGACGCCTT	dJ620E11
stSG85211	TATTCAAACCATCGCAGCAG	GCCTTTGGTTTGAGTTTTGG	dJ753D4
stSG85212	TAGGCATGGAAGGGAACAAG	CTTCGGCAAGGTGAAAGAAG	dJ269M15
stSG85213	CCTCTGGTCCTTGTGGAGAA	GCCCTTCTACAGCCAAGATG	dJ138B7
stSG85214	CACGATTTTGGACACGTCAC	CGGGAGAAAAGAGACCTTCA	dJ1108D11
stSG85215	AAGGTGCGCTCCACATAGTC	ACAGAGCGAGCACAAGGAG	dJ995J12
stSG85216	TGAGCAGCTTCTTAGGCACA	AGGTGTTTGTTGCTATCGGG	dJ688G8
stSG85217	TGTTGCGCTTTTGAAATGTT	CCTGCATATCAGCACCTGAA	dJ601O1
stSG85218	CATCTAGGAGCAGTCCCAGC	CTCCTGGATGGTGGAGAAGA	dJ447F3
stSG85219	GCGGTGATTGTCCAAAAGTT	CTCACAGTCCAGTTGGGGTT	dJ447F3
stSG85220	GGGAATGACAATTTTGGTCG	CCTCCCTTCATGATTCCAAA	dJ447F3
stSG85221	ACTCTGTGCCGCCTAGTGAT	TGCACTGTGGGTCTTCACTC	dJ337O18
stSG85222	CAGGATGGGACAAGGAAAAA	CAGGTGGTGAGGTTGAGGTT	dJ337O18
stSG85223	GGTGCTTATCCAGGGTCTCA	ATTGGCTCTGGCTCAGAAAA	dJ337O18
stSG85224	CAGAAGGTAGAACTCGCCAGA	GAGCCTATGAATGAGCTGCC	dJ981L23
stSG85225	CTGTCGTGGCCTGAAGAAAT	CGTACCATACCACCACACCA	dJ686N3
stSG85226	AAGCTGGTTGCTTCTTTCCA	ATCCTCGTGGTCACTGGTTC	dJ686N3
stSG85227	GGCGATGGAATATGAGAGGA	TGACATTCTTGAAGTGGGCA	dJ686N3
stSG85228	TCTCCACCCTCAGAGGCTTA	ATTCTGGCATCCCAGTGAAG	dJ257E24
stSG85229	TGCTGCTTATGTCCTGATGC	TGACCAAATGTGAGACTGGG	dJ179M20
stSG85230	CCTTAACCTGGCCTCACAAA	TGTGTCACCAATTAGCCCTG	dJ179M20
stSG85231	TTGATGTTTGGGCTAAGGCT	CAGCCCCACTTGAGTTTCTC	dJ179M20
stSG85232	CAGGTCCCTTCTGGTCTTTG	CACTCAGCCCTGATGTTTCA	dJ179M20
stSG85233	TCCTTAAGAGCGACCTCAGC	GCGTTTGTCGTTCTTCCACT	dJ1108D11
stSG85234	CTTCACACTTGCCGTTTCAA	GGACAAGAATCTCCGAGACG	dJ1108D11
stSG85235	TGCAGGAGATCCTGGAGAAC	CTCGGACGGAGTGACTGAC	dJ1108D11
stSG85236	GAGACCAAGAGCACTCAGGC	GGGACATGCTAACAGGTGGA	dJ1108D11
stSG85237	GGGTTTGGCTTTGTGTGTTT	AGTGCCACGTAGAGTGGCTT	dJ1069P2
stSG85238	TCAAACTGCAACAGGAGCAC	TCTTGAGGGTTCTGGAGCTG	dJ453O12

STS	Sense primer	Antisense primer	Parent sequence
stSG85239	GTTACTCGGACTTTGAGCGG	CTGCAGGTGACCCAAAAACT	dJ337O18
stSG85241	CACCTCTGGCACAGAGTCAA	ATCTCACAAAAGTGACGGGC	dJ511B24
stSG85242	AGCCATCTAAAGTGCTCCGA	ACTGTGGTGTTGCAAGATGC	dJ511B24
stSG85243	CTAGGGCTTGCACCTCAAAG	CAGTGCACCAAACTCTGCTC	dJ686N3
stSG85244	AAAACCCACCTTGCCTTCTT	TACCTGCCTTCCCAGATCAC	dJ686N3
stSG85245	GCAGAGCTGGACAATTCCTC	AGGGCCTGAGGAAGAGAAAG	dJ686N3
stSG85246	GCTTTTAACACCCCTGCAAA	ACCAAATGTCCAGGAACCAA	dJ620E11
stSG85247	AAGGCGAGATCACTCTGCTC	ACACCAGCCTTGAGTTGCTC	dJ620E11
stSG85248	CTCCCAAATGTATGCGCTTT	AAGTTGTTGCTGTTGCCTCC	dJ461P17
stSG85249	TCCTCTTTCTCCTCCTCTGG	GTCTGGAGGAGCAATTGGG	dJ309K20
stSG85250	ATGGTGAAGAACCTGAAGCG	GGCTTCATGATCCAGGTGAT	dJ310O13
stSG85251	CACAGCACCTCACAGCATTC	GTGATTGGCTGTCTCCTGGT	dJ310O13
stSG85252	GGGCAGATCCTGTCCTTACA	TAACTTGAACCCCCTCAACC	dJ310O13
stSG85253	CTCCTACCAGAAGGTGCTGC	TGACGTTGAACCAGCAGAAG	dJ310O13
stSG85254	GAGCAGGACAGCATCCTGAG	CGGTCTTGAGCACTAGCAGG	dJ310O13
stSG85255	AGGACATCGGCTACTTCGAG	GATGTCGTGCGTGATGTCAG	dJ310O13
stSG85256	CCACTCAACCTCGAGTAGCC	TCAGGGGACTCACAGGATTC	dJ460J8
stSG85257	TGAAAGAGCCCAGGTACAGG	CGGTACAGCCTCGTCCTAGA	dJ568C11
stSG85258	TGCTTGTCCTTGCACTTCTG	AAAACCAGGTCTGGGAAGGT	dJ568C11
stSG85259	CTCATTTTCGGTCATGCAGT	ATCGGATAGCGTCCATCTGT	dJ687F11
stSG85260	GACGCTGCCTGTGGAACT	GTGGCCCGAGAATAAAGAGC	dJ726C3
stSG85261	TGGAGGAGGACTCATCGACT	GCCATTGGTGATATCCAGGT	dJ726C3
stSG85262	ATGTCCATTCCAGATTAGCAA	GAGGCCCTGGGGTCTTATTA	dJ727I10
stSG85263	GAAGTACCCAGGGGGACAGT	GCAACTGCACAGGCTGTATC	dJ824F16
stSG85264	AGTCACACTTGAAGGGACGG	GCATCAAGCACACCTTCAAA	dJ831D17
stSG85265	TAGTCAGAAGCCCCAGCACT	CACTTCTCCCTGAGCTGTCC	dJ860F19
stSG85266	GCATAGACAGTGCTGAGCCA	CTCGTGGTGTCCTACCCATT	dJ860F19
stSG85267	GCTAGGTGAAAACTGGTGGC	TCAAACTCCGAAGACAGCCT	dJ1018E9
stSG85268	AAGGACTCTGCACTGAAGCC	ACTACTGGAGACCACGTGCC	dJ1025A1
stSG85269	GAGTTGAGCACGGAGTCTGG	CTGGTTTCAAAACCGCAGG	dJ1025A1
stSG85270	CAAACATTATGTCCACCGGG	ATCTCCTTCAGCATCTTCCG	dJ686N3
stSG85271	CTGGAAAGGGCAAGAATCAG	AATCTCCTCACCCGAACCTT	dJ453C12
stSG92802	CGCTGAGAGTGAGCTATCCC	GGGTAGATGTCTTTGGGCTG	dJ450M14
stSG92803	GTGAACGCCTTGATAATGCC	TCTCTCACAGAGATGTGGCG	dJ881L22
stSG92804	ACGTGTGAAGGCTGAGGACT	GATGATGGCCTCCTGAGTGT	dJ998H6
stSG92807	AAGGCCAAGATTCCCTGAGT	CCTACCCTTCTTTCCCGAAC	dJ191L6
stSG92808	TGGGAATTTCATCCCCTAAA	ACTCTCTGATGGGAGCCTTG	dJ1183I21
stSG92809	AGCTGGCAAAAATATGACGG	GGCTAGGTGTCCTGGGGTAT	dJ620E11
stSG92810	GAGCCCGCAAGTTTGATTAC	CGGAAGCAAGGACATCTTTT	dJ1013A22
stSG92811	CTGGCCTCTCTTCAACATGG	AATCAGTCAGACCAGGTCCC	dJ461P17
stSG92812	ATTCAGGGAGACCCAGTGTG	AGCCCTGCTATGCTGTCTGT	dJ450M14
stSG92813	TGGTTCAGTAGCCCTGTGTT	TTGTTACAAGGCAGGCACAA	dJ450M14
stSG92814	ATTTTGGTGAGCTCAATGGC	ACCCAGACTCCTCCTGGAAT	dJbA394O2
stSG92815	TGGAGGTGTGAGATGAGCTG	TGGACGTGTCAATCCTTCAG	bA394O2
stSG92816	GTAACCACTTACAGGCCGGA	CTCTGAAAACTGTCGCGATG	dJ28H20
stSG92817	CGCCGCCGAGTCCCGCTCGC	CGCCATGGGCCACTCCCCAC	dJ28H20

STS	Sense primer	Antisense primer	Parent sequence
stSG92819	CAATGATGTGGTCTTCAGCG	CCCAGAGCATCTGACTTGGT	dJ450M14
stSG92820	GATGGAGTGAACCCAGCAGT	GTCCACCACAGAGGGACAGT	dJ450M14
stSG92821	CTAAGATCGTTCTCCGCAGG	CTGGAGGAGCTGGACTTGAC	dJ781B1
stSG92822	GGTACCTCCAATCAGAGCCA	GTTGGGGTTCCTGGGATACT	dJ781B1
stSG92823	AACCGATTCACCAACCAGTC	GTGGGTGTCACTGGGATTTC	dJ781B1
stSG92824	CCCCCTGCCTAATAGCTACC	GGGTCTGGGAGTGCACTAGA	dJ781B1
stSG92825	ACACGGCACAAGAAGAACCT	GCTGCTTCAGCTCCTCAATC	dJ781B1
stSG92826	GTAACGGGCACCTCAAGTTC	TCAGGATGATGGTCTGGGTT	dJ781B1
stSG92827	TCATAAGGATTTCTTGGCCG	CACGCAGAACTTGTCCTTGA	dJ998C11
stSG92828	CTCTGAACCAAAGGAGCAGG	CTGCTTTGGTGAGGATGTTC	dJ686N3
stSG92829	TTGGAAGAGGGAGTCACCAC	CTACTTCCAACACCCGCATT	dJ686N3
stSG92830	GTGTCCCGTGTGCTTAGAGG	GCTTCTTCGGCTTCAGACAT	dJ963K23
stSG92832	AGAGCACTTCCTGGCATGTT	CAGCAAAGACACAGAGGCAC	dJ28H20
stSG92833	GTGCTATGCCTCCACCATCT	GTCCTCATTGGTCCTTGGAA	dJ28H20
stSG92851	CCATGCCTCTTGAAGCTTTT	CCAGAACTGGGGTTGAGAAA	dJ1028D15
stSG92852	AATTAGCGCCATCGACATTC	AGATGGCCACCTGCAAATAC	dJ1028D15
stSG92853	GCAGGAAGGAGTGACCTGAG	ATTTTGTCAAAGTGGCCTGG	dJ881L22
stSG92854	CTGGTCTGAGGAGAAGTGGC	CTGGGTATAATGGGAGCAGG	dJ881L22
stSG92855	CCAGGCCTTGTGCTAGACTC	TGACGATCTGATTAAGGTCCG	dJ881L22
stSG92856	TGTGAGCAGAAACAACCTGTG	GCGTTGTGGAACGATTCAT	dJ881L22
stSG92857	CCATATGTGTAGGATGCCCA	TGAGTACACACCGTCTCCCA	dJ1013A22
stSG92858	TCCAATATTAGCGGCAGAGC	GGGCATAGTGAGCACCAGAT	dJ1013A22
stSG92859	GTCCTGTGCACCCTGTGTTA	GGTGATGGCGAAGTAGAAGG	dJ781B1
stSG92860	GGTGGCAGAAATGTGCTTCT	CTTGGCTGAAGGAAAGATGG	dJ781B1
stSG92861	ACTACAAGGCCGTCCACAAC	GTAGGCAGCACTCAGCTCCT	bA465L10
stSG92862	AGACAGTGCCCGAACACATT	CTAACCCCAACAGCACAGGT	bA465L10
stSG92863	CCGCTGTCACCGTTTACATA	GCACTGAGGAGCTCTGGTCT	bA465L10
stSG92864	CAAGTGCTCCCCTCTGCTAC	CTGTTTGCTGCTGTGGAAAA	bA465L10
stSG92865	GTTCACGGTCACCTCGCT	CTTTGACGTCTTCCCCCG	bA465L10
stSG92866	GCAAATGGCTTAGAAGCAGG	CTTGTGTGACCAGGCTCTGA	bA394O2
stSG92867	GAGAACATCCAGATTCCCGA	TGGCCACCAAAATCACTACA	bA394O2
stSG92868	CCCTCAAGAAAAATGCCCTA	TTCTCATGCTTCCCAAATCC	bA179N14
stSG92869	TTCCCAACCAAGAAGGACAC	GTCCACCACGTCCTGAGTTT	bA323C15
stSG92870	CAACCTCTCGCTCTCTGGAC	CCCACCTCAGTCGTAAGCTC	dJ66N13
stSG92871	AGCAGCCCCTTCAAGACATA	GGGAGATGACAGTTTCCCAA	dJ963K23
stSG92872	CGCAGACTCACACATGTCAA	CTTGCACCTAAAGCACACCA	bA347D21
stSG92873	GCAGAGTTGGCTGCTTCTCT	CAGCGAGGAGGCTTTCATAC	bA347D21
stSG92874	GTCTTGCCTGCTTGTTCTCC	TTCATCACATCTGAGCCAGG	bA347D21
stSG92875	CCAGGGAGGAAGAACCACTT	GACCTTTGAGCATCTCCTGC	bA347D21
stSG92876	ATGCCGAGGAGCCGGG	CCGGGAGGCAGCAGAG	dJ614O4
stSG92877	CCTATGACTCACGGGCATCT	TCGATCGTTGTCTGATCCAA	dJ614O4
stSG92878	CTTGAAATGGACCCCAACTG	AGGAGCAGCAGCTCTTCTTG	dJ614O4
stSG92880	ATCTGCTGCTCAATGCCAG	GGATGACGTACATGACAAGGC	dJ1022E24
stSG92881	CTGAGGCACCTAGGGAGTTG	AAGATGCAGGCGAGAAAGAC	dJ353C17
stSG92882	TGTCAAGCCCTGCTCTTTCT	GTTCCTTCTGCTGCAACTCC	dJ995J12
stSG92883	ATTCTGAAAGAGTGGCCCAG	CACATGGGCTAATGGAAAGG	dJ461P17

STS	Sense primer	Antisense primer	Parent sequence
stSG92884	TGCTGCTTGCTGTGTAGACC	CTCACTCCTGCTTCTCCAGG	dJ461P17
stSG92885	TACCAGCTAGAGCCAGTGGG	CCCCTAGATTCTATTGGCCG	dJ688G8
stSG92886	CTCACCGAGATTGTCAAGCA	TGTGGGCACAGTCTTGGATA	dJ688G8
stSG92887	TGCACCATCCGAAATAAAGA	GGGTCAGGAAAATAGCAAACA	dJ601O1
stSG92888	AACAGCAACACAGGCTCCTC	CGGTTGTTCAAATCTGAGGG	dJ991B18
stSG92889	AAGGATTGTTGGGAAGGTCC	GCCTGGGAAAAGCTACTTCA	dJ710H13
stSG92890	AATCAGTTCCATCTGGGTCC	GAAAGACCCTGCAGAATGGA	dJ1123D4
stSG92891	GAAATGCTTCCTTCTGCTGC	CAGTCCAGCATTTGCTCCAT	dJ644L1
stSG92892	GTTCTGAAAGGCAGAGTGGC	AGATGTGAGGAGGACCATGC	dJ970A17
stSG92893	GCCCTCAGCAAATCTGAGAA	GAAGCCTCAGACTCCTGGCT	dJ970A17
stSG92894	CTGGCATTGGCTCTGTGAG	TTTGCAGTATTGCTGAGTGCC	dJ128O17
stSG92895	ATGTGATTGGATGCCATGTG	GTTCCATGGCTAGTCCCAGA	dJ128O17
stSG92896	GGTATAGATGCCGGGGAAGA	CGCAGCAGAGAACCAAAAAT	dJ1030M6
stSG92897	CACGCTGATCAGAGACAGGA	GAGACTTGTGGGAAACGGAG	dJ981L23
stSG92898	CATCTTCCCCATCCTTGAAA	TCCCATCTAACCTCACTCGG	dJ1057D4
stSG92899	GCCAGGAGTGCACAGATGTA	ATTCCACGTCATCAGCCTCT	dJ73E16
stSG92900	AAGCAACCTCAAGTTCCACG	CTCCACCCTCCTTTTTCCTT	dJ906C1
stSG92903	TTCCATCCAGGAACCTTCAC	TGAGTCACTTCACACAGCCC	dJ453O12
stSG102600	CCCATCTACACTGCCGGTAT	CCCTCTGTGCCTATGTGGTG	dJ620E11
stSG102601	GTCAGTGTGGCACTGTCCTG	TTACAGCAAGGTCCCAGGTT	dJ620E11
stSG102602	AAGTTCCTGTACTGCTCGGC	CTGTGGAGGGCAGAGAAGAG	dJ620E11
stSG102603	CCTGAGCAGTGATGACGATG	GCGGAGGGACTTCTTGTAGG	dJ450M14
stSG102604	GATGGAGTGAACCCAGCAGT	GTCCACCACAGAGGGACAGT	dJ450M14
stSG102605	GTCTGTCTGGCTTCCCCTC	GGGATAGCTCACTCTCAGCG	dJ450M14
stSG102606	GACGGTGAAGGAGCTGTACC	CCAAAACGCACGTGGAAG	dJ450M14
stSG102607	TCACTCAGGCGTAAGATCACC	TTTGGAAATGACCAACTCCC	dJ1121H13
stSG102608	CTCCAAGTCTCTGCTGCCA	GGGACATGCTAACAGGTGGA	dJ1108D11
stSG102609	TACCTGAGATGTGCTGGCTG	GGCCTCAGTTTCTCGATCTG	dJ1108D11
stSG102610	ACTGTGCTCCACTTTCTGGC	CTCAGCCTCCTGGCCAAT	dJ1108D11
stSG102611	TCTTGTGGGGTGAATGTGAA	AGCCAGAAGTTCCTGCTGAG	dJ1108D11
stSG102612	CTGGGGACTCTCAGGAAACA	CATCTACTTCCCGTGCATGA	dJ781B1
stSG102613	GACCATTCTTCCCTGGTCC	GGCCAGCCTGGAAGACTTA	dJ781B1
stSG102614	AGAAAATTGTTCCTGCCCAA	TTTTTCCAAGTTGGTCCGAG	dJ211D12
stSG102615	TTCTTCCAGTTCCTTCACAGG	CTGCTTCAAGTCCGATCCTC	dJ211D12
stSG102616	TTTTGCAGGTGACCAGACAG	CTGGCCTCCATACCACACTG	dJ461P17
stSG102617	AACATTTCTTTCCACGGCTG	TCAAAGTGGAATGCGAAGTG	dJ461P17
stSG102618	ATCCTCTGAAGAAGGGCACC	CCTCCACAGCCACTGAAGAC	dJ447F3
stSG102619	ACAATGGTACCACCTGCCAT	AGCCCCACCTAGGAAGACAT	dJ337O18
stSG102620	CTCTTTCGGGAACAGCAGTC	GTTCCGTCCTTGCTTCTCAG	dJ337O18
stSG102621	TTCCTTCAGTGTGGGTGACA	GCCTTCTTGGACTTCAGTGG	dJ337O18
stSG102622	ATTCGGACCCAATCAAACAG	TGTCACCCACACTGAAGGAA	dJ337O18
stSG102623	GGTAGAACCCGTGGTGACAG	CGGCAGATCACTATTCCCTC	dJ337O18
stSG102624	GCAGCACGTTCAGCTCATAG	GATTCGGGTGCACTCTGG	dJ337O18
stSG102625	ATTTAGAGATGTGGCCGTGG	CCAAAGACGCAAGGTATTTACA	dJ981L23
stSG102626	ACCGGTTATCCAGGCCTATC	ACACTGTCCTGGGCTTCATC	dJ257E24
stSG102627	ATTGGTAGGAGCTGTGGACG	TCCTCAATACCACTCACCAGC	dJ257E24

STS	Sense primer	Antisense primer	Parent sequence
stSG102628	GCAGCTGGTGAGTGGTATTG	CACTTGGGGTTCTGGAAGG	dJ257E24
stSG102629	TGACAAATGTCTGCTGCCTC	CATTTCCCCTTTTCTTTCCC	dJ569M23
stSG102630	AAATTGCCAGCCTCATAAGC	CACACAATGTCAAGATCCCG	dJ191L6
stSG102631	GACCATTGTGGAAAAATGCC	GCAGCCCTCTGCTACTGTTT	dJ892M9
stSG102632	TTTCCAAAGCCTGTTTGCTT	GCTGGATTGGCAGAGCTAAG	dJ892M9
stSG102633	TGTAGCAAATTGGGAAGGCT	CCTTCCCCTTTAACCCTCAC	dJ892M9
stSG102634	TCATGCCCACAGAATACAGC	TTTTCCAAAAACAGCAAGCA	dJ892M9
stSG102635	TGGGGAAACCATTGATCTGT	GGAGGCTTGGAGGAAAGATT	dJ409O10
stSG102636	CAGCCCCTCCACATTGTTTA	CCAGCCAGGTAGACGTGTTT	dJ644L1
stSG102637	TTCAAAAGTGCTTGCTGGTG	GCAAAAGGCCATTTTTCACT	dJ644L1
stSG102638	CCCCATTTAGGGTCTGGTTT	AAATGCCTGCAATTCCAATA	dJ970A17
stSG102639	CAAAGACTGGGAAAAATGGC	TTGTTTTCAGCAGTTGTCGC	dJ970A17
stSG102640	TTCGTTTATTCTGGGGATCG	TGCTCTGCAAGGCTATTTTG	dJ94E24
stSG102641	GCAGCAACTAAATGAGGCGT	TAATTATTCTGCGGCTGCCT	dJ94E24
stSG102642	TGAGCTATCATAGAGGAGCCG	GCCTTTTCTCTTCCATTCCC	dJ1167E19
stSG102643	TAGTCTGGGCTTCCTCATGC	TACCAAGTTTCTGCGGGTGT	dJ1030M6
stSG102644	CGCCAAATGAAAATGTTTGT	CATTTGGAGCTCTCTCCCAT	dJ495O3
stSG102645	TTGCAGCAACAACTTAACGG	CCTCTTTCTCCCTTGTCCCT	dJ394O2
stSG102646	GCAGTTAAAGTCCCCCTGGT	ATCAGCCAAAGGTGCTGACT	dJ394O2
stSG102647	GTGGTCTTTTCAGAGCAGCC	CTTCATCTTCCTGGCGGTAG	dJ394O2
stSG102648	CTTGCTCTGCTCGTCCTGTT	CTGAAGCTCCGAATCCCAG	dJ453C12
stSG102649	CTTCCTGGTTGCTTCTTTGC	GCAGGAGATGCTCAAAGGTC	dJ66N13
stSG102650	AAGCCTTGAGGAGTTCAGCA	ACCCTGTCTCCTTTCGGTCT	dJ66N13
stSG102651	TTCTTCCACTGCAACCTTGA	ATCTACCGCTGTCCATCTGC	dJ155G6
stSG102652	AAGAAAAATATGCCCCCTGG	AGCTGATGCAGCAGGAGTTT	dJ1123D4
stSG102653	CCAATGCTGGAGCTGTTAGG	GACTCCATCAGGGAGCTTTG	dJ1123D4
stSG102654	GCTTTTGGTTGCTGTGTTGA	TGCTAACTGTGCATCCTTGC	dJ644L1
stSG102655	CCAGTGGCACAGTGGGTAG	AAATCCAGTGCAAGTCACCA	dJ644L1
stSG102656	ATCTCGTACACAAGGTGGGC	GTCACTGATTCTGCTGGGGT	dJ644L1
stSG102657	CATTAACAGGGTGATGCCAA	CATCCCTGTCTGGAAAGCAT	dJ644L1
stSG102658	ATGATCAATCTTGCCGGATG	TCTTTCTTGCCAAAAGCGAC	dJ970A17
stSG102659	CTCTTTTCCTGCTCTGCCTG	TCAATCATGGTCAGGACTCG	dJ94E24
stSG102660	TGCTTTTTCAGTGGAGGTTG	TTCATCTTGGGTGTCCAATG	dJ179J15
stSG102661	ATGCTGGGCTGAAAAGAGAA	TTTCACAGGATGCAGAACCT	dJ179J15
stSG102662	AGCCCCATTAGCAGAGGAGT	TCCAGACTAGGAAGGCTGTCA	dJ1030M6
stSG102663	TTGAAGCAAAGCAAACATGC	ATGAATGGGTTTCAGTTGGC	dJ1005L2
stSG102664	ATATGGGCAGCTACACCAGC	ATTGGGGAAGGCATCTCTCT	bA445H22
stSG102665	CTGGGCTGTGAAGAAGGAAG	CGTCCACAGCTCCTACCAAT	dJ257E24
stSG102666	GAGACAAGAGACCACGAGGC	AACATTGGGTAGTGCCTTGC	dJ511B24
stSG102689	GTGGTGCAGAGTACTGGGGT	GGACCAACGGAAAGAGTTCA	dJ211D12
ststSG76940	AACATCATGATAGGGCCTGG	TTTCATTCAGTTTCACCCCC	dJ620E11

Appendix 3: cDNA probe repository (also available from http://webace.sanger.ac.uk/cgi-bin/webace?db=acedb20&class=Probe).

Probe	cDNA pool	Probe	cDNA pool
pr8000.A.FBI	Fetal_brain_vecI	pr71398.A.T5	Testis_vec5
pr8000.A.FBJ	Fetal_brain_vecJ	pr71398.S.T4	Testis_vec4
pr8000.A.FluR	Fetal_lung_vecR	pr71402.A.T2	Testis_vec2
pr8000.A.T1	Testis_vec1	pr71404.S.T1	Testis_vec1
pr8000.S.FBI	Fetal_brain_vecI	pr71404.T1.1	Testis_vec1
pr8000.S.FBJ	Fetal_brain_vecJ	pr71404.T1.2	Testis_vec1
pr8000.S.T1	Testis_vec1	pr71405.A.Flo	Fetal_liver_vecO
pr8000.S.T3	Testis_vec3	pr71405.S.FlO	Fetal_liver_vecO
pr34084.A.AH8	Heart_vec8	pr71408.A.FlA	Fetal_liver_vecA
pr34084.S.AH8.600	Heart_vec8	pr71408.A.FlJ	Fetal_liver_vecJ
pr34084.S.AH8.1500	Heart_vec8	pr71408.A.FlM.1	Fetal_liver_vecM
pr64002.A.FbL	Fetal_brain_vecL	pr71408.A.FlM.2	Fetal_liver_vecM
pr64002.A.FbN	Fetal_brain_vecN	pr71408.S.FlA	Fetal_liver_vecA
pr64002.S.FbN	Fetal_brain_vecN	pr71409.A.T2.1	Testis-vec2
pr64002.S.T10	Testis_vec10	pr71409.A.T2.2	Testis-vec2
pr64573.A.R	Marathon_Brain	pr71409.A.T2.400	Testis_vec2
pr66874.A.FBP.1	Fetal_brain_vecP	pr71409.A.T2.500	Testis_vec2
pr66874.A.FBP.2	Fetal_brain_vecP	pr71409.A.T3	Testis_vec3
pr66874.S.FBP.1	Fetal_brain_vecP	pr71409.A.T5	Testis_vec5
pr66874.S.FBP.2	Fetal_brain_vecP	pr71409.A.T7	Testis_vec7
pr66877.A.FluV	Fetal_lung_vecV	pr71409.A.T9	Testis_vec9
pr66877.S.74162.A	Fetal_lung_vecV	pr71409.A.T9.400	Testis_vec9
pr66877.S.FluV.1	Fetal_lung_vecV	pr71409.A.T9.600	Testis_vec9
pr66877.S.FluV.2	Fetal_lung_vecV	pr71409.A.T10	Testis_vec10
pr66877.S.FluV.3	Fetal_lung_vecV	pr71409.A.T10.400	Testis_vec10
pr66877.S.FluV.4	Fetal_lung_vecV	pr71409.A.T10.600	Testis_vec10
pr66877S76895A.FLuV	Fetal_lung_vecV	pr71409.S.T2	Testis-vec2
pr71374S76868A.FB2	Fetal_brain_SP2	pr71409.S.T5	Testis_vec5
pr71374S76868A.FB2.B	Fetal_brain_SP2	pr71409.S.T7	Testis_vec7
pr71392.A.T3	Testis_vec3	pr71409.S.T9	Testis_vec9
pr71392.A.T6.500	Testis_vec6	pr71409.S.T10	Testis_vec10
pr71392.A.T6.700	Testis_vec6	pr74159.A.FlE	Fetal_liver_vecE
pr71392.A.T7.500	Testis_vec7	pr74159.A.FlH	Fetal_liver_vecH
pr71392.A.T7.700	Testis_vec7	pr74159.A.T2	Testis-vec2
pr71392.A.T8.500	Testis_vec8	pr74159.A.T2	Testis_vec2
pr71392.A.T8.700	Testis_vec8	pr74159.A.T6	Testis_vec6
pr71392.A.T9.500	Testis_vec9	pr74159.S.FlE	Fetal_liver_vecE
pr71392.A.T9.700	Testis_vec9	pr74159.S.FLH	Fetal_liver_vecH
pr71392.S.T3	Testis_vec3	pr74159.S.T6	Testis_vec6
pr71394.A.R	Marathon_Brain	pr74161.A.FBV	Fetal_brain_vecV
pr71394.A.R2	Marathon_Brain	pr74161.A.FBY	Fetal_brain_vecY
pr71394.A.R4	Marathon_Brain	pr74161.S.FBV	Fetal_brain_vecV
pr71394.A.R5	Marathon_Brain	pr74161.S.FBY	Fetal_brain_vecY
pr71396.A.T1	Testis-vec1	pr74162.S.FBD	Fetal_brain_D
pr71396.A.T3	Testis-vec1	pr74162.S.FLJ	Fetal_liver_vecJ
pr71396.S.T1	Testis-vec1	pr74162.S.FluV	Fetal_lung_vecV
pr71398.A.T4	Testis_vec4	pr74162.S.T1.1000	Testis_vec1

Probe	cDNA pool	Probe	cDNA pool
pr74162.S.T1.1100	Testis_vec1	pr76874.A.FLT	Fetal_liver_vecT
pr74162.S.T5.350	Testis_vec5	pr76874.S.AH1.300	Heart_vec1
pr74162.S.T5.550	Testis_vec5	pr76874.S.AH1.900	Heart_vec1
pr74163.A.ALu3	Adult_lung_vec3	pr76874.S.AH1.900.B	Heart_vec1
pr74163.A.FluV	Fetal_lung_vecV	pr76874.S.FB3.350	Fetal_brain_vec3
pr74163.S.66877.A	Fetal_lung_vecV	pr76874.S.FB3.1000	Fetal_brain_vec3
pr76866.A.FLB	Fetal_liver_vecB	pr76874.S.FBG	Fetal_brain_vecG
pr76866.A.FLD	Fetal_liver_vecD	pr76874.S.FBJ	Fetal_brain_vecJ
pr76866.A.FLuJ.500	Fetal_lung_vecJ	pr76874.S.FLT	Fetal_liver_vecT
pr76866.A.FLuJ.1000	Fetal_lung_vecJ	pr76875.A.FBO	Fetal_brain_vecO
pr76866.A.T1	Testis_vec1	pr76875.A.FLuS	Fetal_lung_vecS
pr76866.S.T1	Testis_vec1	pr76875.S.FBO	Fetal_brain_vecO
pr76868.A.FBI.1	Fetal_brain_vecI	pr76875.S.FLuS	Fetal_lung_vecS
pr76868.A.FBI.2	Fetal_brain_vecI	pr76876.A.FBK	Fetal_brain_vecK
pr76868.A.FLO	Fetal_liver_vecO	pr76876.A.FBL	Fetal_brain_vecL
pr76868.S.FBI	Fetal_brain_vecI	pr76876.A.T1.1000	Testis_vec1
pr76869.A.T1	Testis_vec1	pr76876.A.T1.1200	Testis_vec1
pr76869.A.T3.1200	Testis_vec3	pr76876.S.FBK	Fetal_brain_vecK
pr76869.A.T3.1400	Testis_vec3	pr76876.S.FBL	Fetal_brain_vecL
pr76870.A.FBI	Fetal_brain_vecI	pr76876.S.T1	Testis_vec1
pr76870.A.FLL	Fetal_liver_vecL	pr76877.A.FBK	Fetal_brain_vecK
pr76870.A.FLO.500	Fetal_liver_vecO	pr76877.A.FLQ	Fetal_liver_vecQ
pr76870.A.FLO.1100	Fetal_liver_vecO	pr76877.A.FLR	Fetal_liver_vecR
pr76870.A.T1.800	Testis_vec1	pr76877.A.T1.1100	Testis_vec1
pr76870.A.T1.1000	Testis_vec1	pr76877.A.T1.1200	Testis_vec1
pr76870.S.FBI	Fetal_brain_vecI	pr76877.S.FBK	Fetal_brain_vecK
pr76870.S.FLL	Fetal_liver_vecL	pr76877.S.FLQ	Fetal_liver_vecQ
pr76870.S.FLO.800	Fetal_liver_vecO	pr76877.S.FLR	Fetal_liver_vecR
pr76870.S.FLO.1000	Fetal_liver_vecO	pr76877.S.T1	Testis_vec1
pr76870.S.T1	Testis_vec1	pr76880.A.AH1	Heart_vec1
pr76871.A.AHA	Heart_vecA	pr76880.A.AH10	Heart_vec10
pr76871.A.AHA.350	Heart_vecA	pr76880.S.AH1.850	Heart_vec1
pr76871.A.AHA.700	Heart_vecA	pr76880.S.AH1.1100	Heart_vec1
pr76871.A.AHB	Heart_vecB	pr76880.S.AH10	Heart_vec10
pr76871.A.AHB.350	Heart_vecB	pr76883.A.FBP	Fetal_brain_vecP
pr76871.A.AHB.1000	Heart_vecB	pr76883.A.FLuV	Fetal_lung_vecV
pr76871.A.AHD	Heart_vecD	pr76883.S.FBP	Fetal_brain_vecP
pr76871.A.AHD.350	Heart_vecD	pr76883.S.FLuV	Fetal_lung_vecV
pr76871.A.AHD.1000	Heart_vecD	pr76885.A.AH1	Heart_vec1
pr76871.A.AHK	Heart_vecK	pr76885.A.T3.700	Testis_vec3
pr76871.A.AHK.350	Heart_vecK	pr76885.A.T3.900	Testis_vec3
pr76871.A.AHK.700	Heart_vecK	pr76885.A.T4.400	Testis_vec4
pr76871.S.AHA	Heart_vecA	pr76885.A.T4.900	Testis_vec4
pr76871.S.AHB	Heart_vecB	pr76885.A.T5	Testis_vec5
pr76871.S.AHD	Heart_vecD	pr76885.A.T5.400	Testis_vec5
pr76871.S.AHK	Heart_vecK	pr76885.A.T5.900	Testis_vec5
pr76874.A.AH1	Heart_vec1	pr76885.A.T9	Testis_vec9
pr76874.A.FB3	Fetal_brain_vec3	pr76885.S.AH1	Heart_vec1
pr76874.A.FBG	Fetal_brain_vecG	pr76885.S.T5	Testis_vec5
pr76874.A.FBJ	Fetal_brain_vecJ	pr76885.S.T9	Testis_vec9

Probe	cDNA pool	Probe	cDNA pool
pr76895.S.A.FLuV	Fetal_lung_vecV	pr76917.A.FBF	Fetal_brain_vecF
pr76895S76896A.FLuV	Fetal_lung_vecV	pr76917.A.FBL	Fetal_brain_vecL
pr76895S76896A.FLuV.2	Fetal_lung_vecV	pr76917.A.FBO	Fetal_brain_vecO
pr76896.S.A.FLuV	Fetal_lung_vecV	pr76917.S.76918.A.T9	Testis_vec9
pr76896S74162A.FLuV	Fetal_lung_vecV	pr76917.S.FBF	Fetal_brain_vecF
pr76898.A.ALu3	Adult_lung_vec3	pr76917.S.FBL	Fetal_brain_vecL
pr76898.A.ALu8	Adult_lung_vec8	pr76917.S.FBO	Fetal_brain_vecO
pr76898.A.FLu7	Fetal_lung_vec7	pr76917.S.T9	Testis_vec9
pr76898.A.R1	Marathon_Brain	pr76918.A.T3	Testis_vec3
pr76898.A.R2	Marathon_Brain	pr76918.A.T5.250	Testis_vec5
pr76898.A.R3	Marathon_Brain	pr76918.A.T5.700	Testis_vec5
pr76898.A.R4	Marathon_Brain	pr76918.A.T6.250	Testis_vec6
pr76898.A.R5	Marathon_Brain	pr76918.A.T6.700	Testis_vec6
pr76898.S.74163.A.FLu7	Fetal_lung_vec7	pr76918.A.T8.250	Testis_vec8
pr76899.A.T4	Testis_vec4	pr76918.A.T8.700	Testis_vec8
pr76899.A.T5	Testis_vec5	pr76918.S.T9	Testis_vec9
pr76899.S.T4	Testis_vec4	pr76923.A.AH5	Heart_vec5
pr76899.S.T5	Testis_vec5	pr76923.A.AH9	Heart_vec9
pr76900.A.FB7	Fetal_brain_vec7	pr76923.A.FB9	Fetal_brain_vec9
pr76900.S.FB1	Fetal_brain_vec1	pr76923.S.AH5	Heart_vec5
pr76900.S.T5	Testis_vec5	pr76923.S.AH9	Heart_vec9
pr76901.A.FB7	Fetal_brain_vec7	pr76923.S.FB6	Fetal_brain_vec6
pr76901.A.T5	Testis_vec5	pr76923.S.FB7	Fetal_brain_vec7
pr76901.S.FB1	Fetal_brain_vec1	pr76923.S.T9	Testis_vec9
pr76901.S.FB7	Fetal_brain_vec7	pr76924.A.T7	Testis_vec7
pr76901.S.T5	Testis_vec5	pr76931.S.AH1	Heart_vec1
pr76902.A.FB1	Fetal_brain_vec1	pr76931.S.ALu1	Adult_lung_vec1
pr76902.A.FB7	Fetal_brain_vec7	pr76931.S.T1	Testis_vec1
pr76902.S.FB1	Fetal_brain_vec1	pr76931.S.T3	Testis_vec3
pr76902.S.FB7	Fetal_brain_vec7	pr76931.S.T4	Testis_vec4
pr76902.S.T5	Testis_vec5	pr76932.A.R	Marathon_Testis
pr76904.76898.A.	Neils_eye_cDNA	pr76932.A.R2	Marathon_Testis
pr76911.SA.T5	Testis_vec5	pr76932.A.T3	Testis_vec3
pr76912.SA.T4	Testis_vec4	pr76932.A.T5	Testis_vec5
pr76912.SA.T5	Testis_vec5	pr76932.A.T9	Testis_vec9
pr76913.A.R	Marathon_Brain	pr76932.A.T10	Testis_vec10
pr76913.A.R2	Marathon_Brain	pr76932.S.T3	Testis_vec3
pr76913.A.R3	Marathon_Brain	pr76932.S.T5	Testis_vec5
pr76913S76914A.AH10	Heart_vec10	pr76932.S.T7	Testis_vec7
pr76913S76914A.FLu9	Fetal_lung_vec9	pr76932.S.T9	Testis_vec9
pr76914A.AH5	Heart_vec5	pr76932.S.T10	Testis_vec10
pr76914A.AH7	Heart_vec7	pr76936.A.AHA	Heart_vecA
pr76916A.AH1	Heart_vec1	pr76936.A.AHC.600	Heart_vecC
pr76916A.AH10.800	Heart_vec10	pr76936.A.AHC.2000	Heart_vecC
pr76916A.AH10.1000	Heart_vec10	pr76936.A.FBB	Fetal_brain_vecB
pr76916A.FB6.800	Fetal_brain_vec6	pr76936.S.AHA	Heart_vecA
pr76916A.FB6.1000	Fetal_brain_vec6	pr76936.S.AHB	Heart_vecB
pr76916A.FL1	Fetal_liver_vec1	pr76936.S.AHC	Heart_vecC
pr76916A.T1.600	Testis_vec1	pr76936.S.FBB	Fetal_brain_vecB
pr76916A.T1.800	Testis_vec1	pr76938S76916A.AH8	Heart_vec8

Probe	cDNA pool	Probe	cDNA pool
pr76938S76916A.AH10	Heart_vec10	pr85181.A.ALuG.150	Adult_lung_vecG
pr76938S76916A.T2	Testis_vec2	pr85181.A.ALuG.350	Adult_lung_vecG
pr76940.S.FB1	Fetal_brain_vec1	pr85181.A.T4	Testis_vec4
pr76940.S.FB7	Fetal_brain_vec7	pr85181.S.AHK	Heart_vecK
pr76940.S.T5	Testis_vec5	pr85181.S.T4	Testis_vec4
pr76941.A.AH1	Heart_vec1	pr85185.A.ALuP	Adult_lung_vecP
pr76941.A.AH2	Heart_vec2	pr85185.A.ALuQ	Adult_lung_vecQ
pr76941.A.AH5	Heart_vec5	pr85185.A.R	Marathon_Testis
pr76941.A.AH8	Heart_vec8	pr85185.A.R2	Marathon_Testis
pr76941.S.AH5	Heart_vec5	pr85185.A.R3	Marathon_Testis
pr76941.S.AH8	Heart_vec8	pr85185.A.R4	Marathon_Testis
pr76944.A.T6	Testis_vec6	pr85185.A.R5	Marathon_Testis
pr76944.A.T7	Testis_vec7	pr85185.S.ALuQ	Adult_lung_vecQ
pr76944.A.T8	Testis_vec8	pr85188.A.AHP	Heart_vecP
pr76944.A.T10	Testis_vec10	pr85189.A.AHR	Heart_vecR
pr76958.A.T1	Testis_vec1	pr85189.A.T4	Testis_vec4
pr76959.A.AH6	Heart_vec6	pr85189.S.AHR	Heart_vecR
pr76959.A.FB1	Fetal_brain_vec1	pr85189.S.ALuY	Adult_lung_vecY
pr76959.A.T3	Testis_vec3	pr85189.S.T4.500	Testis_vec4
pr76959.S.AH6	Heart_vec6	pr85189.S.T4.1700	Testis_vec4
pr76959.S.FB1	Fetal_brain_vec1	pr85195.A.T7	Testis_vec7
pr76960.A.FLJ.500	Fetal_liver_vecJ	pr85195.A.T8	Testis_vec8
pr76960.A.FLJ.700	Fetal_liver_vecJ	pr85195.S.T6	Testis_vec6
pr76960.A.FLuK	Fetal_lung_vecK	pr85195.S.T7	Testis_vec7
pr76960.A.T7	Testis_vec7	pr85195.S.T10	Testis_vec10
pr76960.S.FLJ	Fetal_liver_vecJ	pr85196.A.AHA	Heart_vecA
pr76960.S.FLuK	Fetal_lung_vecK	pr85196.A.T4	Testis_vec4
pr76963.A.AH1	Heart_vec1	pr85196.S.AHA	Heart_vecA
pr76963.A.T1	Testis_vec1	pr85196.S.ALuV	Adult_lung_vecV
pr76963.S.T3	Testis_vec3	pr85196.S.T4	Testis_vec4
pr76966.S.FB3	Fetal_brain_vec3	pr85196.S.T6	Testis_vec6
pr76967.A.T4	Testis_vec4	pr85196.S.T7	Testis_vec7
pr76967.S.T4	Testis_vec4	pr85196.S.T8	Testis_vec8
pr76968.A.AH7	Heart_vec7	pr85196.S.T9	Testis_vec9
pr76969.A.AH4	Heart_vec4	pr85196.S.T10	Testis_vec10
pr76970.A.T6.500	Testis_vec6	pr85206.A.T5.800	Testis_vec5
pr76970.A.T6.600	Testis_vec6	pr85206.A.T5.1300	Testis_vec5
pr76970.S.T6	Testis_vec6	pr85206.A.T6	Testis_vec6
pr76979.A.T2	Testis_vec2	pr85206.S.Alu.Q	Adult_lung_vecQ
pr76979.S.FLB	Fetal_liver_vecB	pr85206.S.T3	Testis_vec3
pr76979.S.T5	Testis_vec5	pr85206.S.T5	Testis_vec5
pr76979.S.T6	Testis_vec6	pr85206.S.T6.300	Testis_vec6
pr76983.A.ALuT	Adult_lung_vecT	pr85206.S.T6.600	Testis_vec6
pr76983.A.FLB	Fetal_liver_vecB	pr85209.A.T5	Testis_vec5
pr76983.S.ALuT	Adult_lung_vecT	pr85209.S.T5	Testis_vec5
pr76983.S.FLB	Fetal_liver_vecB	pr85209.S.T10	Testis_vec10
pr85180.A.AHY	Heart_vecY	pr85214.S.R1	Marathon_Brain
pr85180.A.T2	Testis_vec2	pr85214.S.R2	Marathon_Brain
pr85180.S.AHY	Heart_vecY	pr85214.S.R3	Marathon_Brain
pr85181.A.AHK	Heart_vecK	pr85214.S.R4	Marathon_Brain

Probe	cDNA pool	Probe	cDNA pool
pr85214.S.R5	Marathon_Brain	pr85247.A.AH1.1500	Heart_vec1
pr85225.A.85180.S.T7	Testis_vec7	pr85247.A.AH2.700	Heart_vec2
pr85225.S.AHY.300	Heart_vecY	pr85247.A.AH2.1500	Heart_vec2
pr85225.S.AHY.400	Heart_vecY	pr85247.S.AH1	Heart_vec1
pr85225.S.T2	Testis_vec2	pr85247.S.AH2	Heart_vec2
pr85225.S.T4	Testis_vec4	pr92808.A.R	Marathon_Brain
pr85225.S.T8.150	Testis_vec8	pr92810.A.R	Marathon_Brain
pr85225.S.T8.200	Testis_vec8	pr92811.A.R	Marathon_Testis
pr85225.S.T10	Testis_vec10	pr92811.A.R2	Marathon_Testis
pr85228.A.FBF	Fetal_brain_vecF	pr92816.A.R	Marathon_Brain
pr85228.A.FBO	Fetal_brain_vecO	pr92816.A.R2	Marathon_Brain
pr85228.A.T10	Testis_vec10	pr92824.A.R	Marathon_Testis
pr85241.S.T5.400	Testis_vec5	pr92824.A.R2	Marathon_Testis
pr85241.S.T5.600	Testis_vec5	pr92824.A.R3	Marathon_Testis
pr85241.S.T5.2000	Testis_vec5	pr92824.A.R4	Marathon_Testis
pr85241.S.T10	Testis_vec10	pr92828.A.R	Marathon_Testis
pr85242.A.T5	Testis_vec5	pr92828.A.R2	Marathon_Testis
pr85242.A.T10	Testis_vec10	pr92828.A.R3	Marathon_Testis
pr85242.S.R	Marathon_Brain	pr92829.A.R	Marathon_Testis
pr85242.S.T5	Testis_vec5	pr92830.A.R	Marathon_Testis
pr85247.A.AH1.1000	Heart_vec1	pr92830.A.R2	Marathon_Testis

Appendix 4: The 508 cDNA sequences generated during the gene identification effort. Column one lists the sequence identification numbers whereas the EMBL accession numbers (where available) are listed in column two. An example of EST submission is also shown.

Sequence ID	EMBL
sccd1013.64573S	
sccd1014.1rp	
sccd1014.64573A	
sccd1015.1rp	
sccd1015.64573A	
sccd1016.1rp	
sccd1016.64573A	
sccd1017.66868A	AL449719
sccd1018.66869S	AL449720
sccd1019.66869A	AL449721
sccd1019.t7.2fp	AL449722
sccd1020.66872	AL449723
sccd1020.t7.2fp	AL449724
sccd1021.1rp	
sccd1021.66870S	
sccd1022.66870S	
sccd1023.2fp	
sccd1023.66870A	
sccd1024.66870A	
sccd1025.66870A	
sccd1025.t7.2fp	
sccd1026.66870A	
sccd1026.t7.2fp	
sccd1282.66877S	
sccd1283.224	
sccd1283.66877A	
sccd1284.224	
sccd1285.224	
sccd1286.224	AL449504
sccd1286.66875S	AL449505
sccd1287.66875A	AL449506
sccd1288.224	AL449725
sccd1288.66874A	AL449726
sccd1289.224	
sccd1289.66875S	
sccd1290.224	
sccd1290.66875A	
sccd1291.224	
sccd1291.71391A	
sccd1292.224	
sccd1294.71394S	
sccd1295.224	
sccd1295.71394A	
sccd1296.224	
sccd1296.71396S	

Sequence ID	EMBL
sccd1297.224	
sccd1297.71397S	
sccd1298.224	
sccd1298.71397A	
sccd1299.224	
sccd1299.66877S	
sccd1300.224	
sccd1300.66877A	
sccd1302.224	
sccd1302.63375S	
sccd1304.63423A	
sccd1305.224	
sccd1305.63611S	
sccd1306.224	
sccd1306.63646S	
sccd1307.224	
sccd1307.63646A	
sccd1309.224	
sccd1309.65059A	
sccd1314	AL449507
sccd1315	AL449508
sccd1318	AL449509
sccd1319	AL449510
sccd1320	AL449511
sccd1321	AL449512
sccd1322	AL449513
sccd1323	AL449514
sccd1324	AL449515
sccd1325	AL449516
sccd1328	
sccd1329	
sccd1336	AL449886
sccd1338	AL449887
sccd1339	AL449888
sccd1340	AL449889
sccd1341	AL449890
sccd1342	AL449891
sccd1343	AL449892
sccd1350	AL449517
sccd1352	AL449518
sccd1353	AL449519
sccd1354	AL449520
sccd1355	AL449521
sccd1356	AL449667
sccd1363	AL449668

Sequence ID	EMBL
sccd1364	AL449669
sccd1365	AL449670
sccd1366	AL449671
sccd1368	AL449522
sccd1369	AL449523
sccd1370	AL449524
sccd1373	AL449525
sccd1378	AL449727
sccd1379	AL449728
sccd1380	AL449729
sccd1381	AL449730
sccd1382	AL449526
sccd1383	AL449527
sccd1384	AL449528
sccd1386	AL449529
sccd1387	AL449530
sccd1390	AL449731
sccd1391	AL449732
sccd1392	AL449733
sccd1393	AL449734
sccd1394	AL449735
sccd1395	AL449736
sccd1396	AL449737
sccd1397	AL449738
sccd1399	AL449893
sccd1404	AL449739
sccd1406	AL449740
sccd1407	AL449741
sccd1408	AL449531
sccd1409	AL449532
sccd1413	AL449672
sccd1414	AL449894
sccd1415	AL449895
sccd1416	AL449673
sccd1417	AL449674
sccd1420	AL449533
sccd1421	AL449534
sccd1427	AL449675
sccd1429	AL449676
sccd1433	AL449742
sccd1438	AL449743
sccd1439	AL449744
sccd1440	AL449896
sccd1441	AL449897
sccd1443	AL449535

Sequence ID	EMBL	Sequence ID	EMBL	Sequence ID	EMBL
sccd1444	AL449745	sccd3121	AL449752	sccd3194	AL449825
sccd1445	AL449746	sccd3124	AL449753	sccd3195	AL449826
sccd1446	AL449747	sccd3125	AL449754	sccd3196	AL449566
sccd1447	AL449748	sccd3126	AL449550	sccd3197	AL449567
sccd1455	AL449677	sccd3127	AL449551	sccd3198	AL449827
sccd1457	AL449678	sccd3128	AL449755	sccd3199	AL449828
sccd1467	AL449679	sccd3129	AL449756	sccd3200	AL449829
sccd1468	AL449749	sccd3130	AL449757	sccd3201	AL449830
sccd1469	AL449750	sccd3131	AL449758	sccd3202	AL449568
sccd1470	AL449898	sccd3132	AL449759	sccd3203	AL449569
sccd1471	AL449899	sccd3133	AL449760	sccd3204	AL449682
sccd1473	AL449536	sccd3138	AL449761	sccd3205	AL449683
sccd3070	AL449537	sccd3139	AL449762	sccd3208	AL449831
sccd3071	AL449538	sccd3140	AL449763	sccd3209	AL449832
sccd3074	AL449539	sccd3141	AL449764	sccd3212	AL449570
sccd3075	AL449540	sccd3142	AL449552	sccd3213	AL449571
sccd3077	AL449541	sccd3143	AL449553	sccd3215	AL449770
sccd3079	AL449542	sccd3144	AL449554	sccd3217	AL449771
sccd3080	AL449543	sccd3145	AL449555	sccd3219	AL449833
sccd3081	AL449544	sccd3146	AL449556	sccd3221	AL449843
sccd3082	AL449545	sccd3147	AL449557	sccd3222	AL449772
sccd3084	AL449680	sccd3148	AL449765	sccd3223	AL449773
sccd3085	AL449681	sccd3149	AL449766	sccd3224	AL449844
sccd3086	AL449900	sccd3150	AL449767	sccd3225	AL449845
sccd3087	AL449901	sccd3151	AL449768	sccd3226	AL449846
sccd3088	AL449902	sccd3155	AL449769	sccd3227	AL449847
sccd3089	AL449903	sccd3161	AL449818	sccd3228	AL449572
sccd3090	AL449546	sccd3162	AL449819	sccd3229	AL449573
sccd3091	AL449547	sccd3163	AL449820	sccd3230	AL449574
sccd3094	AL449904	sccd3166	AL449821	sccd3231	AL449575
sccd3095	AL449905	sccd3167	AL449822	sccd3232	AL449576
sccd3096	AL449906	sccd3168	AL449823	sccd3233	AL449577
sccd3097	AL449907	sccd3169	AL449824	sccd3235	AL449578
sccd3098	AL449908	sccd3171	AL449558	sccd3237	AL449579
sccd3100	AL449909	sccd3172	AL449559	sccd3238	AL449580
sccd3101	AL449910	sccd3173	AL449560	sccd3239	AL449581
sccd3104	AL449807	sccd3174	AL449561	sccd3241	AL449582
sccd3105	AL449808	sccd3175	AL449562	sccd3242	AL449583
sccd3107	AL449809	sccd3177	AL449563	sccd3243	AL449584
sccd3108	AL449810	sccd3179	AL449564	sccd3245	AL449585
sccd3109	AL449811	sccd3181	AL449565	sccd3247	AL449586
sccd3110	AL449812	sccd3182		sccd3248	AL449587
sccd3111	AL449813	sccd3183		sccd3249	AL449588
sccd3112	AL449814	sccd3184		sccd3250	AL449589
sccd3113	AL449815	sccd3186		sccd3251	AL449590
sccd3116	AL449816	sccd3187		sccd3252	AL449591
sccd3117	AL449817	sccd3190		sccd3255	AL449592
sccd3118	AL449548	sccd3191		sccd3257	AL449593
sccd3119	AL449549	sccd3192	AL449911	sccd3258	AL449848
sccd3120	AL449751	sccd3193	AL449912	sccd3259	AL449849

Sequence ID	EMBL
sccd3261	AL449594
sccd3263	AL449595
sccd3265	AL449596
sccd3267	AL449597
sccd3268	AL449774
sccd3269	AL449775
sccd3275	AL449776
sccd3276	AL449850
sccd3277	AL449851
sccd3278	AL449852
sccd3279	AL449853
sccd3280	AL449854
sccd3281	AL449855
sccd3282	AL449856
sccd3283	AL449857
sccd3287	AL449858
sccd3290	AL449777
sccd3291	AL449778
sccd3292	AL449779
sccd3293	AL449780
sccd3296	AL449781
sccd3297	AL449782
sccd3298	AL449783
sccd3299	AL449784
sccd3302	AL449785
sccd3304	AL449684
sccd3305	AL449685
sccd3306	AL449598
sccd3307	AL449599
sccd3308	AL449686
sccd3313	
sccd3314	
sccd3315	
sccd3316	
sccd3317	
sccd3318	AL449794
sccd3319	AL449795
sccd3320	AL449600
sccd3324	
sccd3325	
sccd3326	
sccd3329	
sccd3334	AL449913
sccd3335	AL449914
sccd3336	AL449601
sccd3337	AL449602
sccd3340	AL449859
sccd3341	AL449860
sccd3342	AL449603
sccd3343	AL449604

Sequence ID	EMBL
sccd3347	AL449861
sccd3353	AL449862
sccd3364	
sccd4088	AL449786
sccd4089	AL449787
sccd4090	AL449788
sccd4091	AL449789
sccd4092	AL449790
sccd4093	AL449791
sccd4094	AL449605
sccd4095	AL449606
sccd4098	AL449607
sccd4099	AL449608
sccd4105	AL449609
sccd4106	AL449610
sccd4107	AL449611
sccd4109	AL449612
sccd4111	AL449863
sccd4113	AL449796
sccd4115	AL449797
sccd4116	AL449798
sccd4117	AL449799
sccd4118	AL449800
sccd4119	AL449801
sccd4123	AL449613
sccd4125	AL449614
sccd4127	AL449864
sccd4129	AL449615
sccd4131	AL449865
sccd4133	AL449802
sccd4135	AL449616
sccd4139	AL449617
sccd4143	AL449618
sccd4144	AL449619
sccd4145	AL449620
sccd4147	AL449621
sccd4148	AL449622
sccd4149	AL449623
sccd4150	AL449624
sccd4151	AL449625
sccd4152	AL449687
sccd4153	AL449688
sccd4154	AL449626
sccd4155	AL449627
sccd4156	AL449689
sccd4162	AL449628
sccd4165	AL449866
sccd4168	AL449629
sccd4169	AL449630
sccd4173	AL449631

Sequence ID	EMBL
sccd4175	AL449632
sccd4176	AL449633
sccd4177	AL449634
sccd4179	AL449635
sccd4180	AL449636
sccd4185	AL449637
sccd4186	AL449638
sccd4187	AL449639
sccd4190	AL449640
sccd4191	AL449641
sccd4193	AL449642
sccd4194	AL449643
sccd4195	AL449644
sccd4197	AL449645
sccd4199	AL449867
sccd4201	AL449646
sccd4203	AL449647
sccd4216	AL449868
sccd4217	AL449869
sccd4221	AL449915
sccd4233	AL449648
sccd4235	AL449649
sccd4240	AL449870
sccd4241	AL449871
sccd4242	AL449803
sccd4243	AL449804
sccd4244	AL449792
sccd4245	AL449793
sccd4248	AL449872
sccd4250	AL449873
sccd4251	AL449874
sccd4252	AL449875
sccd4253	AL449876
sccd4254	AL449690
sccd4261	AL449877
sccd4266	AL449878
sccd4267	AL449879
sccd4270	AL449880
sccd4271	AL449881
sccd4273	AL449882
sccd4274	AL449883
sccd4275	AL449884
sccd4277	AL449885
sccd4278	AL449650
sccd4279	AL449651
sccd4280	AL449652
sccd4281	AL449653
sccd4282	AL449654
sccd4283	AL449655
sccd4284	AL449656

Sequence ID	EMBL
sccd4285	AL449657
sccd4286	AL449658
sccd4288	AL449659
sccd4289	AL449660
sccd4292	AL449661
sccd4293	AL449662
sccd4294	AL449663
sccd4295	AL449664
sccd4296	AL449665
sccd4297	AL449666
sccd4298	AL449916
sccd4299	AL449917
sccd4300	AL449805
sccd4301	AL449806
sccd4334	AL449467
sccd4335	AL449468
sccd4336	AL449469
sccd4337	AL449470
sccd4341	AL449471
sccd4342	AL449472
sccd4360	AL449473
sccd4362	AL449474
sccd4363	AL449475
sccd4364	AL449476
sccd4366	AL449477
sccd4367	AL449478
sccd4368	AL449479

Sequence ID	EMBL
sccd4370	AL449480
sccd4372	AL449481
sccd4376	AL449482
sccd4378	AL449483
sccd4380	AL449484
sccd4382	AL449485
sccd4384	AL449486
sccd4385	AL449487
sccd4386	AL449488
sccd4387	AL449489
sccd4388	AL449490
sccd4389	AL449491
sccd4390	AL449492
sccd4392	AL449503
sccd4393	AL449493
sccd4394	AL449494
sccd4395	AL449495
sccd4396	AL449496
sccd4398	AL449691
sccd4400	AL449692
sccd4401	AL449693
sccd4402	AL449694
sccd4404	AL449695
sccd4406	AL449696
sccd4407	AL449697
sccd4408	AL449698
sccd4412	AL449699

Sequence ID	EMBL
sccd4413	AL449700
sccd4414	AL449701
sccd4415	AL449702
sccd4417	AL449703
sccd4419	AL449704
sccd4420	AL449705
sccd4421	AL449706
sccd4422	AL449707
sccd4423	AL449708
sccd4424	AL449709
sccd4425	AL449710
sccd4426	AL449711
sccd4427	AL449712
sccd4428	AL449713
sccd4430	AL449714
sccd4432	AL449715
sccd4434	AL449716
sccd4436	AL449717
sccd4437	AL449718

An example of an EMBL submitted cDNA sequence

```
ID   HSCCD1447  standard; RNA; EST; 232 BP.
XX
AC   AL449748;
XX
SV   AL449748.1
XX
DT   09-NOV-2000 (Rel. 65, Created)
DT   09-NOV-2000 (Rel. 65, Last updated, Version 1)
XX
DE   A Homo sapiens single pass sequence read. The sequence corresponds to a
DE   cDNA fragment isolated from a fetal brain cDNA library using a cDNA end
DE   rescue technique.
XX
KW   chromosome 20; EST.
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Primates; Catarrhini; Hominidae; Homo.
XX
RN   [1]
RP   1-232
RA   Stavrides G.S., Huckle E.J., Deloukas P.;
RT   ;
RL   Submitted (08-NOV-2000) to the EMBL/GenBank/DDBJ databases.
RL   The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire,
RL   CB10 1SA, UK. E-mail contact: humquery@sanger.ac.uk Unpublished
XX
CC   Sanger Centre name :  sccd1447
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..232
FT                   /chromosome="20"
FT                   /db_xref="taxon:9606"
FT                   /db_xref="ESTLIB:7114"
FT                   /organism="Homo sapiens"
FT                   /tissue_type="Brain"
XX
SQ   Sequence 232 BP; 60 A; 64 C; 61 G; 47 T; 0 other;
     ctacctccgg cgcctggatg cggggctctt tgttctccag cacatctgct acatcatggc        60
     cgagatctgc aatgccaatg tcccccagat tcgccagagg gttcaccaga tcctaaacat       120
     gcgaggaagc tccatcaaaa ttgtcaggca tatcatcaag gagtatgcag agaacatcgg       180
     ggacggccgc aaggagagga cagcgattct cgtacgaacg gttacgattc ga               232
//
```

17 mRNA sequences constructed in silico using overlapping In_house generated cDNA sequences and publicly available ESTs were also submitted to EMBL (Accession numbers: AL133001, AL132998, AL591562, AL591713, AL121742, AL591714, AL591565, AL591715, AL118493, AL096778, AL118494, AL121740, AL133002, AL133000, AL132999, AL121739, AL137188).

Appendix 5: Gene data.

Locus	Type	Genomic size (bp)	Largest transcript (bp)	Transcript size (bp)	Exon number	Comments
dJ191L6.1	Pseudogene	238			1	Similar to src-like proto-oncogene
dJ191L6.2	Putative	5,317	191L6.2.mRNA	426	3	
KRML	Known	3,049	644L1.1.mRNA	3,049	1	Kreisler maf-related leucine zipper homolog
TOP1	Known	95,622	1J6.1.mRNA	3,690	21	Topoisomerase I
PLCG1	Known	38,146	511B24.2.1.mRNA	5,151	32	Phospholipase C, gamma 1
TIX1	Novel	26,726	511B24.3.mRNA	9,871	2	Probable homeobox protein
dJ511B24.4	Pseudogene	463			1	60S Ribosomal Protein L23A (RPL23A) pseudogene
dJ450M14.1	Pseudogene	2,079			3	
LPIN3	Novel	13,063	620E11.2.mRNA	2,190	17	Protein similar to KIAA0188, KIAA0249 and yeast SMP2
C20orf130	Novel	6,761	620E11.4.mRNA	3,679	4	
KIAA1335	Novel	216,269	620E11.1.mRNA	10,964	37	Helicase C-terminaland SNF2 N-terminal domains containing protein
dJ620E11.3	Pseudogene	2,590			2	Similar to LY6
RPL12L2	Pseudogene	612			1	Similar to RPL12 (ribosomal protein L12)
dJ1121H13.1	Pseudogene	2,846			1	Similar to intermediate filament proteins
dJ1121H13.3	Putative	1,445	1121H13.3.mRNA	417	3	
PTPRT	Known	1,117,219	269M15.2.mRNA	12,708	31	Protein tyrosine phosphatase, receptor type T
dJ730D4.1	Pseudogene	866			2	
PPIAL	Pseudogene	494			1	Similar to peptidylprolyl isomerase (cyclophilin)
dJ862K6.4	Pseudogene	319			1	Similar to part of NBP (Nucleotide Binding Protein)

Locus	Type	Genomic size (bp)	Largest transcript (bp)	Transcript size (bp)	Exon number	Comments
dJ862K6.3	Pseudogene	346			1	Similar to 4E-BP2 (4E-Binding Protein 2)
SFRS6	Known	5,678	862K6.2.2.mRNA	3,951	7	Splicing factor, arginine/serine-rich 6
KIAA0681	Known	27,552	862K6.1.1.mRNA	3,286	19	Lethal (3) malignant brain tumor (l(3)mbt)
dJ138B7.4	Pseudogene	312			1	HSPC194 pseudogene
SGK2	Novel	26,582	138B7.2.1.mRNA	1,852	13	Serum/glucocorticoid regulated kinase 2
C20orf9	Novel	56,233	138B7.1.1.mRNA	1,624	14	CGI-53 protein
RPL27AP	Pseudogene	833			2	Similar to RPL27A
MYBL2	Known	49,340	1028D15.3.mRNA	2,639	14	V-myb avian myeloblastosis viral oncogene homolog-like 2
C20orf65	Novel	577	1028D15.4.mRNA	577	1	
C20orf100	Novel	110,874	1108D11.2.mRNA	1,296	7	Novel HMG (high mobility group) box protein
JPH2	Novel	73,752	1183I21.2.1.mRNA	2,655	6	Novel protein similar to C. elegans T22C1.7
C20orf111	Novel	14,296	1183I21.1.1.mRNA	1,572	4	
dJ995J12.2	Putative	13,697	995J12.2.mRNA	665	4	
GDAP1L1	Novel	33,182	995J12.1.1.mRNA	2,320	6	Similar to mouse ganglioside-induced differentiation associated protein
C20orf142	Novel	4,894	881L22.2.mRNA	1,159	2	
R3HDML	Novel	13,635	881L22.3.mRNA	762	5	Similar to trypsin inhibitors
dJ881L22.5	Putative	7,634	881L22.5.mRNA	754	4	
dJ881L22.4	Putative	17,964	881L22.4.1.mRNA	648	4	
HNF4A	Known	29,128	1013A22.1.mRNA	2,260	10	Hepatocyte nuclear factor 4, alpha
dJ1013A22.2	Putative	1,301	1013A22.2.mRNA	396	2	
C20orf62	Putative	377	1013A22.3.mRNA	377	3	
dJ1013A22.4	Pseudogene	279			1	RPL37A (ribosomal protein L37a) pseudogene
C20orf121	Novel	14,164	179M20.3.mRNA	1,681	6	Similar to cellular retinaldehyde-binding protein
TDE1	Known	24,589	179M20.2.1.mRNA	4,355	10	Tumour differentially expressed 1(DIFF33)
PKIG	Known	87,107	179M20.4.1.mRNA	1,045	4	Protein kinase (cAMP-dependent, catalytic)

Locus	Type	Genomic size (bp)	Largest transcript (bp)	Transcript size (bp)	Exon number	Comments
						Inhibitor gamma
ADA	Known	32,180	179M20.1.mRNA	1,498	12	Adenosine deaminase
bA445H22.3	Putative	6,837	445H22.3.mRNA	790	4	
dJ781B1.4	Putative	34,161	445H22.4.mRNA	529	5	
WISP2	Known	12,626	445H22.2.mRNA	1,462	4	WNT1 inducible signaling pathway protein 2
KCNK15	Known	5,246	781B1.1.mRNA	1,311	2	Two pore potassium channel KT3.3
dJ781B1.2	Putative	3,268	781B1.2.mRNA	465	2	
C20orf190	Novel	54,606	781B1.3.mRNA	1,281	6	Similar to myeloblast KIAA0237 protein and rat protein NIM2
YWHAB	Known	22,808	148E22.1.1.mRNA	3,095	7	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, beta polypeptide
C20orf119	Novel	29,223	148E22.2.mRNA	2,017	14	Similar to poly(A)-binding protein
TOMM34	Known	18,303	1069P2.2.mRNA	2,008	7	Translocase of outer mitochondrial membrane 34 (TOM34)
STK4	Known	109,027	211D12.2.mRNA	1,883	11	Serine/threonine kinase 4
dJ211D12.3	Putative	2,142	211D12.3.mRNA	1,508	2	
KCNS1	Known	8,802	211D12.1.mRNA	4,537	5	Potassium voltage-gated channel, delayed-rectifier, subfamily S, member 1
PRG5	Known	5,711	211D12.5.mRNA	1,034	4	p53-responsive gene 5
C20orf122	Known	1,104	211D12.4.mRNA	838	3	Similar to Elafin-like protein
PI3	Known	1,570	172H20.3.mRNA	478	3	Protease inhibitor 3, skin-derived (SKALP)
SEMG1	Known	2,734	172H20.2.mRNA	1,620	3	Semenogelin I
SEMG2	Known	3,068	172H20.1.mRNA	1,920	3	Semenogelin II
dJ300I2.1	Putative	476	300I2.1.mRNA	227	2	
SLPI	Known	2,323	300I2.2.mRNA	593	4	Secretory leukocyte protease inhibitor
MATN4	Known	12,138	453C12.1.2.mRNA	2,182	10	Matrilin 4

Locus	Type	Genomic size (bp)	Largest transcript (bp)	Transcript size (bp)	Exon number	Comments
RBPSUHL	Known	10,117	453C12.2.mRNA	1,739	11	Recombining binding protein suppressor of hairless-like
SDC4	Known	23,137	453C12.3.mRNA	2,613	5	Syndecan 4 (amphiglycan, ryudocan)
C20orf169	Novel	6,058	453C12.4.1.mRNA	2,684	4	Similar to Drosophila CG11753
C20orf10	Novel	4,508	453C12.5.mRNA	1,051	5	clg01 (p53 response element)
dJ453C12.8	Pseudogene	359			1	NADH-ubiquinone oxidoreductase B15 subunit pseudogene
C20orf35	Novel	4,388	453C12.6.1.mRNA	1,479	4	Uncharacterized hypothalamus protein (HSMNP1)
PIGT	Novel	10,081	453C12.7.1.mRNA	2,148	12	Phosphatidyl inositol glycan class T
dJ461P17.10	Putative	945	461P17.10.mRNA	518	2	
WFDC2	Known	11,777	461P17.6.mRNA	564	4	Major epididymis-specific protein E4 precursor
dJ461P17.9	Putative	2,597	461P17.9.mRNA	316	2	
dJ461P17.5	Pseudogene	854			1	RPL5 (60S Ribosomal Protein L5) pseudogene
SPINT3	Known	384	461P17.8.mRNA	384	1	Serine protease inhibitor, Kunitz type, 3
C20orf171	Novel	4,251	461P17.11.mRNA	395	3	Contains a WAP-type (Whey Acidic Protein) "four-disulfide core" domain and a Kunitz/Bovine pancreatic trypsin inhibitor domain
dJ461P17.4	Pseudogene	219			1	COX6C pseudogene
SPINLW1	Novel	6,756	461P17.2.1.mRNA	1,944	4	Contains a WAP-type (Whey Acidic Protein) "four-disulfide core" domain and a Kunitz/Bovine pancreatic trypsin inhibitor domain
dJ461P17.3	Pseudogene	245			1	Similar to part of the HSPD1
C20orf170	Novel	27,538	461P17.1.mRNA	1,042	6	Contains a WAP-type (Whey Acidic Protein) "four-disulfide core" domain and Kunitz/Bovine

Locus	Type	Genomic size (bp)	Largest transcript (bp)	Transcript size (bp)	Exon number	Comments
						pancreatic trypsin inhibitor domains
dJ839B11.2	Pseudogene	818			1	Enhancer of invasion 10 (HEI10) pseudogene
dJ688G8.2	Putative	6,740	688G8.2.mRNA	550	4	
C20orf146	Novel	1,370	688G8.3.mRNA	405	2	Weakly similar to basic protease inhibitor chelonianin
RPS2L1	Pseudogene	882			1	Similar to ribosomal protein S2 (RPS2)
dJ688G8.4	Putative	20,083	688G8.4.mRNA	371	2	
C20orf137	Novel	3,480	601O1.1.mRNA	512	3	Contains a WAP-type (Whey Acidic Protein) "four-disulfide core" domain and a Kunitz/Bovine pancreatic trypsin inhibitor domain
dJ447F3.1	Pseudogene	465			1	Pseudogene
C20orf168	Novel	154	447F3.6.mRNA	154	1	Contains a Kunitz/Bovine pancreatic trypsin inhibitor domain
WFDC3	Novel	16,733	447F3.3.1.mRNA	897	7	Contains WAP-type (Whey Acidic Protein) "four-disulfide core" domains
C20orf167	Novel	19,483	447F3.4.1.mRNA	1,282	13	Similar to synaptotagmin 1
UBE2C	Known	4,294	447F3.2.1.mRNA	762	6	Ubiquitin-conjugating enzyme E2 H10
TNNC2	Known	10,531	447F3.5.2.mRNA	779	7	Fast troponin C2
C20orf161	Novel	7,872	337O18.4.mRNA	1,597	4	
PTE1	Known	15,663	337O18.3.1.mRNA	1,145	6	Peroxisomal acyl-CoA thioesterase
C20orf164	Novel	21,302	337O18.7.mRNA	2,569	2	
C20orf162	Novel	3,071	337O18.5.1.mRNA	1,492	3	
C20orf165	Novel	1,147	337O18.8.mRNA	757	2	
C20orf163	Novel	2,367	337O18.6.mRNA	991	2	
PPGB	Known	8,355	337O18.2.2.mRNA	2,082	15	Lysosomal protective protein precursor
PLTP	Known	13,388	337O18.1.1.mRNA	1,743	16	Lipid transfer protein II

Locus	Type	Genomic size (bp)	Largest transcript (bp)	Transcript size (bp)	Exon number	Comments
C20orf67	Novel	13,396	465L10.1.mRNA	2,747	17	
ZNF335	Novel	23,520	465L10.2.mRNA	4,430	28	C2H2 type zinc finger protein similar to chicken FZF-1
bA465L10.3	Pseudogene	827			2	Ferritin light polypeptide (FTL) pseudogene
MMP9	Known	7,654	465L10.4.mRNA	2,336	13	Matrix metalloproteinase 9
bA465L10.7	Putative	6,024	465L10.7.mRNA	248	2	
SLC12A5	Novel	30,904	465L10.5.mRNA	5,994	26	Solute carrier family 12, member 5
NCOA5	Novel	28,909	599F21.1.mRNA	3,161	8	Nuclear receptor coactivator 5
dJ599F21.2	Pseudogene	659			1	RPL13 (60S ribosomal protein L13) pseudogene
TNFRSF5	Known	11,592	599F21.3.mRNA	1,729	9	Tumor necrosis factor receptor superfamily, member 5
C20orf25	Novel	77,957	998H6.1.mRNA	3,655	11	Similar to rat PB-Cadherin
C20orf5	Novel	14,858	394O2.1.mRNA	2,236	10	
KIAA1834	Novel	40,561	394O2.2.2.mRNA	3,569	21	Similar to KIAA0281
dJ981L23.1	Pseudogene	1,319			1	Zinc-finger protein pseudogene
dJ981L23.2	Pseudogene	1,262			1	Makorin (MKRN1 or MKRN3) ring finger protein pseudogene
C20orf157	Novel	635	981L23.3.mRNA	223	2	Novel KRAB box type zinc-finger protein
ZNF334	Novel	12,463	179N14.1.mRNA	2,430	5	Zinc finger protein
C20orf123	Novel	5,384	257E24.3.mRNA	1,985	2	
SLC13A3	Known	93,608	257E24.2.1.mRNA	4,017	13	Sodium-dependent high-affinity dicarboxylate transporter
C20orf64	Novel	3,132	28H20.2.mRNA	1,231	2	
SLC2A10	Novel	26,598	28H20.1.mRNA	4,126	5	Solute carrier family 2 (facilitated glucose transporter), member 10
bA323C15.1	Pseudogene	549			1	RPL13 (60S ribosomal protein L13) pseudogene
EYA2	Known	294,206	1050K3.1.mRNA	2,687	16	Eyes absent (Drosophila) homolog 2

Locus	Type	Genomic size (bp)	Largest transcript (bp)	Transcript size (bp)	Exon number	Comments
bA323C15.2	Putative	584	323C15.2.mRNA	423	2	
dJ1050K3.2	Pseudogene	1,008			1	GAPDH pseudogene
dJ1050K3.3	Pseudogene	278			1	RPL27A (60S ribosomal protein L27A) pseudogene
dJ890O15.3	Pseudogene	863			1	RPS2 (40S ribosomal protein S2) pseudogene
PRKCBP1	Known	147,134	890O15.2.mRNA	4,536	22	Protein kinase C binding protein 1
dJ569M23.2	Putative	677	569M23.2.mRNA	413	2	
bA456N23.1	Pseudogene	797			2	RPL35A (60S ribosomal protein L35a) pseudogene
NCOA3	Known	153,863	1049G16.2.1.mRNA	6,838	23	Nuclear receptor coactivator 3
dJ1049G16.3	Pseudogene	1,173			2	RPS3A pseudogene
KIAA1247	Novel	100,060	1049G16.1.1.mRNA	3,671	20	Similar to glucosamine-6-sulfatases
dJ1057D4.1	Pseudogene	885			1	Spermidine synthase pseudogene
bA347D21.1	Putative	1,257	347D21.1.mRNA	425	2	
bA347D21.2	Putative	582	347D21.2.mRNA	456	2	
bA347D21.4	Putative	910	347D21.4.mRNA	450	2	
bA347D21.3	Putative	8,257	347D21.3.mRNA	429	2	
dJ66N13.2	Putative	1,012	66N13.2.mRNA	551	2	
dJ66N13.3	Putative	893	66N13.3.mRNA	438	2	
dJ66N13.1	Putative	1,573	66N13.1.mRNA	404	2	
dJ991B18.1	Putative	30,225	991B18.1.mRNA	277	2	
KIAA1415	Novel	203,594	269H4.1.mRNA	6,499	40	
ARFGEF2	Known	114,804	155G6.1.mRNA	8,852	39	ADP-ribosylation factor guanine nucleotide-exchange factor 2
dJ155G6.3	Pseudogene	246			1	SNAP-25 pseudogene
dJ155G6.4	Putative	5,582	155G6.4.1.mRNA	556	3	
CSE1L	Known	50,641	470L14.1.1.mRNA	3,553	25	Chromosome segregation 1 (yeast homolog)-

Locus	Type	Genomic size (bp)	Largest transcript (bp)	Transcript size (bp)	Exon number	Comments
						like
STAU	Known	74,845	470L14.2.1.mRNA	3,500	14	Staufen (RNA binding protein)
ARPC3B	Pseudogene	800			1	Arp2/3 protein complex subunit p21-Arc (ARC21) pseudogene
DDX27	Novel	24,703	686N3.1.mRNA	2,615	21	Similar to ATP dependent RNA helicases (contains conserved C-terminal helicase domains and DEAD/DEAH boxes)
KIAA1404	Novel	32,166	686N3.2.1.mRNA	7,209	14	
dJ686N3.3	Putative	10,621	686N3.3.1.mRNA	546	5	
KCNB1	Known	110,681	791K14.1.mRNA	3,760	2	Potassium voltage-gated channel, Shab-related subfamily, member 1
PTGIS	Known	60,676	298O6.1.mRNA	1,982	10	Prostaglandin I2 (prostacyclin) synthase
B4GALT5	Known	80,857	1063B2.1.mRNA	4,645	9	beta-1,4-galactosyltransferase, polypeptide 5
dJ1041C10.2	Pseudogene	255			1	small nuclear ribonucleoprotein polypeptide F (SNRPF) pseudogene
KIAA0939	Novel	79,342	1041C10.4.1.mRNA	6,095	16	Novel member of sodium/hydrogen exchanger family
dJ1041C10.3	Pseudogene	2,017			1	
SPATA2	Known	12,118	963K23.3.mRNA	4,022	3	Spermatogenesis associated protein 2, PD1
ZNF313	Novel	17,482	963K23.2.mRNA	2,440	6	Zinc finger protein 313
dJ963K23.1	Pseudogene	1,288			1	KRT18 (Keratin type I, Cytoskeletal 18 (Cytokeratin 18, CK18,CYK18)) pseudogene
SNAI1	Known	5,888	710H13.1.mRNA	1,686	3	Snail 1 (drosophila homolog), zinc finger protein
dJ710H13.2	Putative	789	710H13.2.mRNA	462	2	
UBE2V1	Known	48,277	1185N5.1.4.mRNA	2,134	5	Ubiquitin-conjugating enzyme E2 variant 1

Appendix 6: Novel gene expression results and isolation of cDNA sequences.

Number of positive libraries	Testis	Fetal Brain	Fetal Liver	Fetal Lung	Peripheral Blood	Adult Heart	Adult Lung	Novel gene	Primer name (stSG)	Vectorette sequences	SSP-PCR Sequences	RACE sequences
5	+	+	+			+	+	TIX1	85209	+		+
2			+				+	LPIN3	76983	+		
3	+			+		+		C20orf130	85247	+		
6	+	+	+	+		+	+	KIAA1335	76922	+		
5	+		+	+	+	+		SGK2	85213			
2		+				+		C20orf9	92851			
3	+					+	+	C20orf65	92852			
7	+	+	+	+	+	+	+	C20orf100	85236			+
3				+		+	+	JPH2	85233	+		
7	+	+	+	+	+	+	+	C20orf111	76870	+		+
3	+	+					+	GDAP1L1	85215			
3		+			+	+		C20orf142	92853			
1						+		R3HDML	92854			
3	+	+	+					C20orf121	74159	+		+
2	+					+		C20orf190	92821			
3		+	+				+	C20orf119	76868	+		+
6	+	+	+	+		+	+	C20orf169	76914		+	+
1	+							C20orf10	76885	+		
6	+	+	+	+		+	+	C20orf35	76916		+	
7	+	+	+	+	+	+	+	PIGT	85271		+	+
NONE								C20orf171	102616			
1	+							SPINLW1	71392	+		+

Number of positive libraries	Testis	Fetal Brain	Fetal Liver	Fetal Lung	Peripheral Blood	Adult Heart	Adult Lung	Novel gene	Primer name (stSG)	Vectorette sequences	SSP-PCR Sequences	RACE sequences
1	+							C20orf170	71409	+		
1	+							C20orf146	92886			
NONE								C20orf137	92887			
1					+			C20orf168	102618			
2	+	+						WFDC3	85219	+		
1	+							C20orf167	85218			
3	+					+	+	C20orf161	85239			
4	+	+		+		+		C20orf164	102619			
6	+	+	+	+		+	+	C20orf162	85223	+		
NONE								C20orf165	102623			
5	+		+	+	+		+	C20orf163	85221			
7	+	+	+	+	+	+	+	C20orf67	92861			
7	+	+	+	+	+	+	+	ZNF335	92862	+		+
1		+						SLC12A5	92863			
5	+	+		+		+	+	NCOA5	92864			
1		+						C20orf25	76944			
7	+	+	+	+	+	+	+	C20orf5	92866			
7	+	+	+	+	+	+	+	KIAA1834	92867			
NONE								C20orf157	102625			
4	+	+	+			+		ZNF334	92868			
NONE								C20orf123	102628			
7	+	+	+	+	+	+	+	C20orf64	76871	+		+
3		+	+				+	SLC2A10	76895	+		+
7	+	+	+	+	+	+	+	KIAA1247	71396	+		+

Number of positive libraries	Testis	Fetal Brain	Fetal Liver	Fetal Lung	Peripheral Blood	Adult Heart	Adult Lung	Novel gene	Primer name (stSG)	Vectorette sequences	SSP-PCR Sequences	RACE sequences
5	+	+	+		+		+	KIAA1415	85185			
7	+	+	+	+	+	+	+	DDX27	85225	+		+
7	+	+	+	+	+	+	+	KIAA1404	85226	+		
3	+					+	+	KIAA0939	92871			
7	+	+	+	+	+	+	+	ZNF313	76931	+		+

Appendix 7: Putative gene expression results.

Number of positive libraries	Testis	Fetal Brain	Fetal Liver	Fetal Lung	Peripheral Blood	Adult Heart	Adult Lung	Putative gene	Primer name (stSG)
2	+				+			dJ191L6.2	92807
NONE								dJ1121H13.3	102607
6	+	+	+	+		+	+	dJ995J12.2	92882
1	+							dJ881L22.5	92855
2	+		+					dJ881L22.4	92856
1	+							dJ1013A22.2	92857
1	+							C20orf62	92858
1	+							bA445H22.3	102664
NONE								dJ781B1.4	102613
1	+							dJ781B1.2	92860
7	+	+	+	+	+	+	+	dJ211D12.3	102614
3			+			+	+	dJ300I2.1	85188
1				+				dJ461P17.10	92884
NONE								dJ461P17.9	92883
3	+		+				+	dJ688G8.2	92885
5	+	+	+		+	+		dJ688G8.4	85217
1		+						bA465L10.7	92865
1							+	bA323C15.2	92869
1			+					dJ569M23.2	102629
2		+					+	bA347D21.1	92872
1					+			bA347D21.2	92873
NONE								bA347D21.3	92874
NONE								bA347D21.4	92875
NONE								dJ66N13.1	92870
3			+	+	+			dJ66N13.2	102649
NONE								dJ66N13.3	102650
2	+				+			dJ991B18.1	92888
1	+							dJ155G6.4	102651
7	+	+	+	+	+	+	+	dJ686N3.3	85227
1					+			dJ710H13.2	92889

Appendix 8: Supporting evidence for annotation.

Gene name	Type	Identical cDNA(s)	Identical EST(s)	Similar to Protein(s)/Predicted protein(s)
dJ191L6.1	Pseudogene			+
dJ191L6.2	Putative		+	
KRML	Known	+	+	+
TOP1	Known	+	+	+
PLCG1	Known	+	+	+
TIX1	Novel	+	+	+
dJ511B24.4	Pseudogene			+
dJ450M14.1	Pseudogene			+
LPIN3	Novel	+	+	+
C20orf130	Novel	+	+	+
dJ620E11.3	Pseudogene			+
KIAA1335	Novel	+	+	+
RPL12L2	Pseudogene			+
dJ1121H13.1	Pseudogene			+
dJ1121H13.3	Putative		+	
PTPRT	Known	+	+	+
dJ730D4.1	Pseudogene			+
PPIAL	Pseudogene			+
dJ862K6.4	Pseudogene			+
dJ862K6.3	Pseudogene			+
SFRS6	Known	+	+	+
KIAA0681	Known	+	+	+
dJ138B7.4	Pseudogene			+
SGK2	Novel	+	+	+
C20orf9	Novel	+	+	+
RPL27AP	Pseudogene			+
MYBL2	Known	+	+	+
C20orf65	Novel		+	+
C20orf100	Novel	+	+	+
JPH2	Novel	+	+	+
C20orf111	Novel	+	+	+
dJ995J12.2	Putative		+	
GDAP1L1	Novel	+	+	+
C20orf142	Novel		+	+
R3HDML	Novel		+	+
dJ881L22.5	Putative		+	
dJ881L22.4	Putative	+	+	
HNF4A	Known	+	+	+
dJ1013A22.2	Putative		+	
C20orf62	Putative		+	
dJ1013A22.4	Pseudogene			+
C20orf121	Novel		+	+
TDE1	Known	+	+	+
PKIG	Known	+	+	+

Gene name	Type	Identical cDNA(s)	Identical EST(s)	Similar to Protein(s)/Predicted protein(s)
ADA	Known	+	+	+
WISP2	Known	+	+	+
bA445H22.3	Putative	+	+	
dJ781B1.4	Putative		+	
KCNK15	Known	+	+	+
dJ781B1.2	Putative		+	
C20orf190	Novel	+	+	+
YWHAB	Known	+	+	+
C20orf119	Novel	+	+	+
TOMM34	Known	+	+	+
STK4	Known	+	+	+
KCNS1	Known	+	+	+
dJ211D12.3	Putative	+	+	
PRG5	Known	+	+	+
C20orf122	Known	+	+	+
PI3	Known	+	+	+
SEMG1	Known	+	+	+
SEMG2	Known	+	+	+
dJ300I2.1	Putative		+	
SLPI	Known	+	+	+
MATN4	Known	+	+	+
RBPSUHL	Known	+		+
SDC4	Known	+	+	+
C20orf169	Novel	+	+	+
C20orf10	Novel	+	+	+
dJ453C12.8	Pseudogene			+
C20orf35	Novel	+	+	+
PIGT	Novel	+	+	+
dJ461P17.10	Putative		+	
WFDC2	Known	+	+	+
dJ461P17.9	Putative		+	
dJ461P17.5	Pseudogene			+
SPINT3	Known		+	+
C20orf171	Novel			+
dJ461P17.4	Pseudogene			+
dJ461P17.3	Pseudogene			+
SPINLW1	Novel	+	+	+
C20orf170	Novel		+	+
dJ839B11.2	Pseudogene			+
dJ688G8.2	Putative		+	
C20orf146	Novel		+	+
RPS2L1	Pseudogene			+
dJ688G8.4	Putative		+	
C20orf137	Novel		+	+
dJ447F3.1	Pseudogene			+
C20orf168	Novel			+

Gene name	Type	Identical cDNA(s)	Identical EST(s)	Similar to Protein(s)/Predicted protein(s)
WFDC3	Novel	+	+	+
C20orf167	Novel	+	+	+
UBE2C	Known	+	+	+
TNNC2	Known	+	+	+
C20orf161	Novel	+	+	+
PTE1	Known	+	+	+
C20orf164	Novel		+	
C20orf162	Novel	+	+	+
C20orf165	Novel		+	
C20orf163	Novel		+	+
PPGB	Known	+	+	+
PLTP	Known	+	+	+
C20orf67	Novel	+	+	+
ZNF335	Novel	+	+	+
bA465L10.3	Pseudogene			+
MMP9	Known	+	+	+
bA465L10.7	Putative		+	
SLC12A5	Novel	+	+	+
NCOA5	Novel	+	+	+
dJ599F21.2	Pseudogene			+
TNFRSF5	Known	+	+	+
C20orf25	Novel	+	+	+
C20orf5	Novel	+	+	+
KIAA1834	Novel	+	+	+
dJ981L23.1	Pseudogene			+
dJ981L23.2	Pseudogene			+
C20orf157	Novel			+
ZNF334	Novel	+	+	+
C20orf123	Novel		+	+
SLC13A3	Known	+	+	+
C20orf64	Novel	+	+	+
SLC2A10	Novel	+	+	+
bA323C15.1	Pseudogene			+
bA323C15.2	Putative		+	
dJ1050K3.3	Pseudogene			+
dJ1050K3.2	Pseudogene			+
EYA2	Known	+	+	+
dJ890O15.3	Pseudogene			+
PRKCBP1	Known	+	+	+
dJ569M23.2	Putative		+	
bA456N23.1	Pseudogene			+
NCOA3	Known	+	+	+
KIAA1247	Novel	+	+	+
dJ1049G16.3	Pseudogene			+
dJ1057D4.1	Pseudogene			+
bA347D21.1	Putative		+	

Gene name	Type	Identical cDNA(s)	Identical EST(s)	Similar to Protein(s)/Predicted protein(s)
bA347D21.2	Putative		+	
bA347D21.3	Putative		+	
bA347D21.4	Putative		+	
dJ66N13.1	Putative		+	
dJ66N13.2	Putative		+	
dJ66N13.3	Putative		+	
dJ991B18.1	Putative		+	
KIAA1415	Novel	+	+	+
ARFGEF2	Known	+	+	+
dJ155G6.3	Pseudogene			+
dJ155G6.4	Putative		+	
CSE1L	Known	+	+	+
STAU	Known	+	+	+
ARPC3B	Pseudogene			+
DDX27	Novel	+	+	+
KIAA1404	Novel	+	+	+
dJ686N3.3	Putative		+	
KCNB1	Known	+	+	+
PTGIS	Known	+	+	+
B4GALT5	Known	+	+	+
dJ1041C10.2	Pseudogene			+
dJ1041C10.3	Pseudogene			+
KIAA0939	Novel	+	+	+
SPATA2	Known	+	+	+
ZNF313	Novel	+	+	+
dJ963K23.1	Pseudogene			+
SNAI1	Known	+	+	+
dJ710H13.2	Putative		+	
UBE2V1	Known	+	+	+

Appendix 9: The sequences of gene-based, working STSs.

stSG number	Mouse sequence accession number	Status	Sense primer	Antisense primer	Product length (bp)
stSG77001	AA986794	EST	CAGTGGTGGGGATATCCTTG	ACACGTTTCTACCCACAGCA	400
stSG77002	X70958	EST	AAGGTACGTGCGGATCAAAC	CTCTCTGCTTCTGCCTTGCT	222
stSG77003	D10061	mRNA	TAGAGAGCGGATTGCCAACT	TCCTCAGGCATGATCCTTCT	105
stSG77004	D10061	mRNA	CAGGGGTTCTGTGAAAAGGA	CAAATTCCCATCCACTTGCT	268
stSG77005	AA215096	EST	CCACCTAGTTGGAGAGCAGC	AGGGAGGCCCAAAACAATAC	174
stSG77006	AI427671	EST	ATTGGAGTCTTTGAGCCCCT	AGCCAGGCAGTGATCAGAAC	101
stSG77008	AF152556	mRNA	TGCCAACCTTGAAGGAGAAT	CTCAGTGACGACGCCAGATA	158
stSG77009	AF152556	mRNA	GTGCCACCTACCTGTGGATT	CAGCACCCGGATCTCATACT	182
stSG77010	AF152556	mRNA	CTGCCACAGCTACAACCTCA	CCACCAACTCCTCACTCTCC	185
stSG77011	AF152556	mRNA	CTCCAACCTCGATTCGCTAC	GCTGTAGCTCTTCAATGGGG	142
stSG77012	AF152556	mRNA	CTAACACTGTGGAGCCGGAG	ATCACCCCCAGGAGAATGA	100
stSG77013	AF152556	mRNA	AAAGGCCACCATGAGATCC	TGGGGGATTGAGAAACTTGA	126
stSG77014	AI060744	EST	GGCAGGAAATCCAGATCAAA	TGTCTCCTTTGCCGTTCTCT	148
stSG77015	W89632	EST	GAGGAGCAGTGAGAACCCAG	GCTTCTGTGCTTAAGCTGGC	118
stSG77016	U58941	mRNA	CTAGGCTTGTCACAGGCTCC	TTTATAAGTCTGGGCCTCCG	134
stSG77017	J05261	mRNA	TGGCGGAGAACAATTATGAA	TCCTGCATGACCAGTACAGC	136
stSG77018	AA530750	EST	CCATGTCTGCCATCTCCTTT	GTGGCCCAAAAACTGTTCAA	101
stSG77019	AB019618	mRNA	GGGACCATCTTCCTGATTGA	CGTGGTTCACTGTCGTTGAT	182
stSG77020	AB019618	mRNA	ATTGATCGGGACTCGGATTT	CAGTAATGTTGTGCCAGCCA	112
stSG77021	AA756018	EST	CGGTGACTGTTCGGGATTAT	ATGTTGGAGGTGGTGAGGTC	148
stSG77022	AA461907	EST	TGGACCTTCCTGCTCTATGG	TGTCTGAAACTGCTGCTCTATTTC	114
stSG77023	M10319	mRNA	GAGGATCGCCTACGAGTTTG	AGTCTGGTTCCAGGGCATT	121
stSG77024	L29441	mRNA	CCTCTGGACTCTTGTGGCTC	TTGATGAAACGCAAAATGGA	132
stSG77025	AF058797	mRNA	GAGCAGGGACACGAACTCTC	TCTGTTCGATGCTGGAGATG	112
stSG77026	AI322317	EST	GTGCACGTAAGGCAGATCAC	TGGCTCCAGCTACTTTGTCC	102

stSG number	Mouse sequence accession number	Status	Sense primer	Antisense primer	Product length (bp)
stSG77027	AI226097	EST	GGGCTTGCTGTACTTCATCC	GGGAAACGGGAGCTGTAGA	105
stSG77028	D89571	mRNA	CCACTGGATAACCACATCCC	CTCAAAGATGTTGCTGCCCT	174
stSG77029	AJ006140	mRNA	GTTTTGGCCGTGGAGTACAT	AAATGTCATCCTGAGAGCGG	163
stSG77030	AF002719	mRNA	GCCTTAAGCTTGAGAAGCCA	GGAACAGGATTCACGCACTT	103
stSG77031	M57590	mRNA	CCCTGTGTCCTTCTCTGTCC	TCCTTCATCAGGGATTCGAT	132
stSG77032	W15900	EST	ATGCTTGGACCTGAAGTTGG	GCTCATTTTCTGTCCGCAGT	104
stSG77033	W97825	EST	CAGAGCGAGCATAAGGAACC	AAGGTGCGTTCCACGTAGTC	125
stSG77034	BF687342	EST	CTAGAGCATGTTCCCCCACT	AGGCAGGAAGATGATGATGG	105
stSG77035	AF180338	mRNA	GGGGCCAAATTGACATACAC	CTTAGGACGCAAAGCCTGTC	133
stSG77037	X70472	mRNA	ACGTGGTAATACCCCTTCCC	GAGGCCAGACTGGATGAGAG	146
stSG77038	D29015	mRNA	GACCAGTCCCAGAGCAGCTA	TTGAGAGGAGTCCAGGCAGT	157
stSG77040	M83312	mRNA	CTTCGGGTTAAGAAGGAGGG	CATAACTCCAAAGCCAGGGA	132
stSG77041	M83312	mRNA	AAAACACATTCCAAGGCAGG	GCACACATGGAGGTCAAATG	147
stSG77043	U81603	mRNA	CTTCTCCACTGATGGCTTCC	TCCTTCACACGACAGTAGCG	120
stSG77044	AA222144	EST	ACCAGCACGAAGAACCTCAT	CTGTGGTTTGGGCTGTTTCT	124
stSG77045	AA004036	EST	TGGCTGAAGAGGAAATAAAAACCGAGTAGAGATATCCATTCCCTCCA		60
stSG77046	AA178103	EST	TATGAAGCAGCCACTCACCA	TTCAGGACAAAACCAGTCCC	184
stSG77047	AF000581	mRNA	TACAGTGGTGAGAAGTGGCG	TGGCACATTTATCTGGCTTG	121
stSG77048	AA120567	EST	AGGCTTGGATACTGTGCCTC	GTCACTTTGATGGGATTGGG	146
stSG77049	AA796530	EST	AAACATTCCCAAGTTCGTGC	CAACATTGAGGACATCCTGG	101
stSG77050	AA796530	EST	TTCTGATGTTGGAGCAGACAG	AGGAGTTGGCCAAGAGGAC	95
stSG77051	W45951	EST	TGAAAGATCTGGCTTCGGTC	ATCCGTATGAACTCTTGCCG	100
stSG77052	AA023567	EST	ACTCAGTCACCTGAGCCAGC	ATGGCCTCAACAACCTGAAC	102
stSG77053	AA881599	EST	TCTCTCATTGGCCTGTTTGA	CAGATGGATTTTGGGGTTGT	174
stSG77054	AI549625	EST	CCCGTCTTCCTATATGGCCT	CTAAGAATGCCTGGAGCAGC	104
stSG77056	AA611074	EST	ACCATCTTTGGCACCTTCAC	ATGGTGTTTCCGTGGTAGGA	117
stSG77057	AI640027	EST	TTGTCTGCCCCATAAGAACC	TGCCTCCACTATCCTCCATC	101

stSG number	Mouse sequence accession number	Status	Sense primer	Antisense primer	Product length (bp)
stSG77061	AU079453	EST	CAGAGCTCTCATGGACTCCC	GGTAGCTGTGGCTTCTTTGG	154
stSG77062	AF152556	mRNA	TGCCAACCTTGAAGGAGAAT	CTCAGTGACGACGCCAGATA	158
stSG77063	AI851523	EST	CGACAAGAACCAAACCCAAA	CGGCAATGATTTTGGTTATCTA	101
stSG77064	AI553523	EST	CCTGGCTGCTCTCTGGTTAC	CACAGCCTCTTTCCTTGGAC	251
stSG85199	AB001607	mRNA	GGGGCAGATACGTCACTGTT	ATCCTGGCCTTCTCCTCACT	163
stSG85200	W08433	EST	CCTGAGTCCCGAATCTTCAC	GAGCAGCTCAACCACCTCTC	195
stSG85201	X70472	mRNA	AACAGTGGACGCTGATAGCC	ATCCTGTCTTCCTCCTCGGT	124
stSG85202	D29015	mRNA	CTGAAGGTGCCAACCTCAAT	ATGTGGTTCTTCCTCACGCT	151
stSG85203	W83123	EST	AGAAACCTGTCTGTGGGGTG	CAGCCATGGCAAGAAGTCTC	103
stSG85204	M64228	mRNA	CGTGTCCATGAACATGAAGG	ACTTGTTGGGAGACAGGTGG	128
stSG85205	M64228	mRNA	AGCCTCGAGGACAACGAGTA	CCTCGTTCATCTGCTCCTTC	211
stSG85301	AA177721	EST	AAGAGCTGGACTGTGGAGGA	TTCACAAAGGTGCTGTGGAG	252
stSG85302	U71208	mRNA	GTCCCTCCTGGTCAGTCCTC	CGGAGTTGGGTACGCTGTAT	102
stSG85303	AI835478	EST	TTTTCCACCTGATCAAAGCA	CCTTGCCCAATGGATGTAAA	137

Appendix 10: Mouse BAC-end sequence-based STSs.

STS	Sense primer	Antisense primer	End-sequence	Parent clone
stSG85311	GAGCATCAGGACCGCTTTAG	TGGCATCTCTTAACACAGCG	Em:AZ105369	bM22B22
stSG85312	AGTGAGAGCTGAGCAGGGAG	TTCATCATCAAAGACGCTCG	Em:AZ105371	bM22B22
stSG85313	TGAGAAGGCTTGGCTTTTGT	ACCAGCTTGCCTGACAGTCT	Em:AZ088114	bM23N22
stSG85314	AGTGAGTGAAGCCAGGCAGT	ATGGTTGTATCTGCAAGCCC	Em:AZ103297	bM23N22
stSG85315	GCTTTGTAGATCAGGCTGGC	TCTGTGGATTTCCCTCTTGG	Em:AZ011393	bM378K6
stSG85316	CAACCCTGAGATTCCACCTC	TTACCAGTTGGGGCATCTTC	Em:AZ011399	bM378K6
stSG85317	TTTAACCCCAGCTCTCAGGA	GCCCCATCTTTCTTTCTTCC	Em:AZ087306	bM23K9
stSG85318	GTGTGATGAAGCCATTCCCT	GAAGAGAATGACACTGGGGC	Em:AZ087307	bM23K9
stSG85319	ATAAAGCCAGCCATCACAGG	CTTGGCCAGATGGGTCTTAG	Em:AQ971386	bM328D8
stSG85320	CAACCCCAGTGGAGAGAAAG	TATAACGTCCAGCAGGTCCC	Em:AZ038580	bM328D8
stSG85321	TGGCTCCCAACACTTTTACC	CCACACTCACCCTTTTTGCT	Em:AZ108171	bM471I14
stSG85322	GTATTTGCAAGGCAAGCACA	GATCCTTTCCACCGAGTTCA	Em:AZ108174	bM471I14
stSG85323	GTGTTTAGGGCTGGCCTCTT	GCCCACTCATGATATCCCAC	Em:AZ060029	bM414F2
stSG85324	CATAGCAAACAGGGAAAGCC	AGTTCCACAGTGAGGCTGCT	Em:AZ060034	bM414F2
stSG85325	CAAGGCCAGCAGGTGTTTTA	CCCCCACATCATACTTCTTTTT	Em:AZ052261	bM424E13
stSG85326	GCTGGTCATCTATGCTGCAA	GCAAACCCTGCAAGTTTCAT	Em:AZ052265	bM424E13
stSG85327	GTTCTAGGACAGCAGCCAGG	TGGCTTTTTCTCAAGGTTTTG	Em:AZ071410	bM435K23
stSG85328	AGTGGAGGCTCCTTTCCATT	CATTTGAGCCCCAACAAACT	Em:AZ071412	bM435K23
stSG85329	TTTCTCATGCTGTGGTCTCG	AAGCCCTTGGTTCCTCAGAT	Em:AZ075271	bM397C16
stSG85330	CATGTCCCCATGAAATCTCC	TCCTCTACTCAAGGGGCAGA	Em:AZ075276	bM397C16
stSG85331	AACACAGTTCCTGTCCCTGG	ACAGCCTCATGACCATCTCC	Em:AZ034106	bM327A19
stSG85332	CCAGCTCTGTCACCATCAGA	TCCCTGGTTTGAATTTGCTC	Em:AZ034108	bM327A19
stSG85333	TCCTTTGCCTCTCTAAGCCA	CTCTGAGAGCCCAGCTCAGT	Em:AZ021160	bM335N12
stSG85334	ATGGAGAAGCTGGGGAAGAT	GGCTTCTGTTCAGTTCCTGC	Em:AZ021162	bM335N12
stSG85335	CATTAAGGACTCGGGGGAAT	GATGGTGACAATGATGCTGG	Em:AZ039371	bM285O5
stSG85336	TTATTTTGCCCCACTCTTGC	GGTGGCACCAAATCCTTAGA	Em:AZ039372	bM285O5
stSG85337	CCACTTCAGGGACTGCATTT	TTAGCTGGCACCATTAACCC	Em:AZ087726	bM36P22
stSG85338	GGACACCCAGTGAGACCCTA	TGGCAAGTGCTGTTACAAGG	Em:AZ087731	bM36P22
stSG85339	CAACACAAGGCAGAACAGGA	ACTGAAAGGAAACCAAGGGG	Em:AZ080682	bM399D16
stSG85340	ACCAGCACAAGGTCCAAAAC	TGAGCGATGTCAAGAAGCAC	Em:AZ080684	bM399D16
stSG85341	TAATCTTTTCCGTGGCCTTG	CTGGCCTTGGACTCAGAGAG	Em:AZ007440	bM345I23
stSG85342	TGGAATCTTTGGCTTTGTCC	GTCAGGCATTGCCCTGTACT	Em:AZ007446	bM345I23
stSG85343	CTAGGCTTGTCACAGGCTCC	CCTCCCCACCTGGATTTAGT	Em:AQ993950	bM355H6
stSG85344	AAAAAGATGCTTCCCCCATC	TTTTTCCTACCTTCCCACCC	Em:AQ993952	bM355H6
stSG85345	GTTCTCTGGGTACAGCAGCC	CTTAGTAGGCTCTGCCGCTC	Em:AZ042052	bM382N13
stSG85346	GGCAAACAGTGAGAGCAACA	TTTCACAACACTGGCCACAT	Em:AZ042054	bM382N13
stSG85347	CCACCATGTGATTGCTGGTA	CCATGTTAAGGAACCCGAAA	Em:AZ057660	bM428M13
stSG85348	CAGAGCCATGGGGTGTACTT	AATTTCTGACGACAGGGCAC	Em:AZ088690	bM428M13
stSG85349	TTACTACCATCCCTGCACCC	AGGAGCTGGAGAAGACACGA	Em:AZ068663	bM394J1
stSG85350	ACGTGCACACACTTCCACAT	CGGTCCATCTGTATGAGGCT	Em:AZ068667	bM394J1
stSG85351	AGGAAAACCATTTTCATGCG	GTAGGAAGCAGTGCAGACCC	Em:AQ923673	bM294F12
stSG85352	CAGGAAGGAAGATTTCAGCG	CCACCGACATTGTGTGCTTA	Em:AZ019160	bM294F12
stSG85353	AAACAGAATGCAGAATGGGG	GGTTTCCACAAGCCTTCCTA	Em:AQ932728	bM284F12
stSG85354	TGGTCCCATTTATTCGGTGT	GGTCATGATGTTTGTGCAGG	Em:AZ035262	bM284F12
stSG85355	TTCCACTGACTGCAACTAGAGC	GGTGAATATCCAATTTTCAAATCA	Em:AZ025852	bM345N18
stSG85356	ACCCCCAAAATTCTACCAGC	ACGCATGAGCATCATTTAGC	Em:AZ025857	bM345N18

STS	Sense primer	Antisense primer	End-sequence	Parent clone
stSG85357	GCCTCCTGAGAGCTAGGGTT	ATGACTCCAAGGTTTGGCAC	Em:AQ995176	bM383K1
stSG85358	GCAGATAGCACCCTGGAGAG	AATGTGAAGGGAACAGACCG	Em:AQ995178	bM383K1
stSG85359	GAACAAAACCCCAGGAACG	TTTCCCTTTTGTGTTTTGCC	Em:AZ105691	bM455P19
stSG85360	GACACCAGCGGCACAGAA	CTGCTGTGGGGCCTCTTC	Em:AZ105695	bM455P19
stSG85361	GCTTCAAAGCACCTCAGAGC	CGGGTAACAGGTGCTTCTTC	Em:AZ094209	bM448B9
stSG85362	CATTACTGCCTGGCTCCCTA	GGTAGGTAGGGTGCGTGTGT	Em:AZ094214	bM448B9
stSG85363	GAAGAAAAGTCCCCCACCTC	TTTTAAATGTGGAGCCAGCC	Em:AQ918418	bM271D20
stSG85364	TCACGCATCACGAGTGTGTA	CAGCCTTCCAGCTCAATCTC	Em:AQ918421	bM271D20
stSG85365	TTCTTGAGAGCTCCCCATGT	ATTGCTGAGGTGATGCTGTG	Em:AQ972328	bM328K5
stSG85366	GTCCTCCTGACCAAAGTCCA	TCTTCTGGCTTCTTTGGGAA	Em:AZ039808	bM328K5
stSG85367	CTCAGTCCAGAGCCTTGTCC	ACAATCAGCTCACAAACCCC	Em:AQ998366	bM356E13
stSG85368	AGGGCTAGCTCCCTCTATGC	TGGTCATCTGGAGAAGGGAC	Em:AQ998368	bM356E13
stSG85369	CCTGGAGATTGTTCCCAGAA	CCTCAGTTTCCCCAACCATA	Em:AZ031930	bM363O18
stSG85370	GGTGGGCAGGTTAACTTCAA	TGGGATTTGAACTCAGGACC	Em:AZ031934	bM363O18
stSG85371	CTAAATGTCGACGGCCAGA	TCCAGTTTTCTGAGGAACCG	Em:AZ030549	bM245F10
stSG85372	CTAGGGAGTGGGCTAGGCTT	CTTCTACGGCTCCAGACCAG	Em:AZ030550	bM245F10
stSG92901	AGCTAGTGAAGGCCAGGTCA	CTTGCCTTTGAGTGGGAGAG	Em:AZ280759	bM122E21
stSG92902	TTTGCAGTTCTCACAGCAGG	ATGTGCTGCTAGCCTTGGAT	Em:AZ280756	bM122E21
stSG93001	TAAACCGGGGTGTCTGGTTA	AGTTATGGGCCCTGCCTAGT	Em:AZ090774	bM7G21
stSG93002	TCAACACAGAAAATGCCCAA	TCTCCACGCACAAACAAGAG	Em:AZ090784	bM7G21
stSG93003	CACTCGGAGGTGAGAACACA	GTCCCCAAATGCTGGAGTTA	Em:AZ099280	bM480D17
stSG93004	TTGGATCAGAAGGCCAAAAG	AGCAACCTCCAACTCTCCCT	Em:AZ099281	bM480D17
stSG93005	GAGGGGCCCTAAAATAGCTG	CCGCACTGGACTGACACTAA	Em:AZ092826	bM7O16
stSG93006	CTAGCAAGAATGGCCGAAAG	ATCCCCCAACACACTACAGC	Em:AZ099281	bM480D17
stSG93007	TATTGCTTTGCTGGCTCAAA	AATCTCATGCTGGGGTTCAG	Em:AQ928451	bM264D2
stSG93008	CGATCTCACTTTGAGTGGCA	TCTGTGACTGGCTTCTCACG	Em:AQ993701	bM264D2
stSG93009	GCCCATGCGCTGTAGATTAT	AAAAGTCCTCCCATTGTCCC	Em:AZ116954	bM16J17
stSG93010	GGCTCAGTGGAGGTGAAGAG	GGCTAGGGATGGTTGTGAAA	Em:AZ116955	bM16J17
stSG93011	GCACACTCTGATTCTTGCCA	CTTGGAATGTGAGGAGCCAT	Em:AQ975929	bM338B13
stSG93012	CTTCCCACATACCCAAGCAT	AAGCACATTATGCCTGACCC	Em:AQ995655	bM338B13
stSG93013	ATCGGCCTGGCTATCCTATT	GTGGAGGGACAGGAAACTGA	Em:AQ997651	bM371C18
stSG93014	ATTCTCCATGCCAGACCATC	TTGGCCTCTGCCTTTTCTTA	Em:AQ997654	bM371C18
stSG93021	AGAGGAAGCCGCCTATCATT	CTACCCTGATGGGCAGAAAA	Em:AZ036759	bM392O2
stSG93022	TGGTAACATGGAGGCAGACA	GGAGTAGGTGTGGCATTGGT	Em:AZ036755	bM392O2
stSG93023	CACTGTGGCATGTAGATGGG	GGAGGGAGGGACTTGGTTAG	Em:AZ122159	bM30C22
stSG93024	ATACATGGCCCACCATTGTT	CGCCACATTTTCAACATCAC	Em:AZ122161	bM30C22
stSG93025	AGCCCAAGGCAGTAGGAAAT	CTGTGTCAGGCTATGGCAGA	Em:AZ079157	bM400C18
stSG93026	CAGAACATTGCAAGGCAGAA	CCAAATGAGGTGTCACATGC	Em:AZ079153	bM400C18
stSG93027	CCAACACAAAGAGAGGGGAA	GCATATCTGCAGCGAGTTCA	Em:AZ029651	bM376G18
stSG93028	GTTTGGGAAAGATCCAGCAA	GCCTTTCTCTGGTGTTCAGC	Em:AZ029656	bM376G18
stSG93029	GAACCGCAGGAGAGACAGAC	TTCCCCTCTCTTTGTGTTGG	Em:AZ124020	bM466K24
stSG93030	TTTGTGGTCCTCAGCCTTCT	CATTTAGCCAGTCCTCAGGG	Em:AZ124022	bM466K24
stSG93031	TTGGCTGGTTGAGGTAAAGG	CATTGTCACACCTGACGACC	Em:AQ928402	bM264B16
stSG93032	AAGGAATGGACCCAAACTCC	TATCATGCCAACGGATGTGT	Em:AQ993569	bM264B16
stSG93033	TCCTCATGCATTCTGCAAAG	AGGGATGTGTGTGACAGCAG	Em:AZ077944	bM452H13
stSG93034	ATCTGTTGGTGGCCTTTTTG	GAAGATATGGGCACAAGGGA	Em:AZ101722	bM452H13
stSG93035	AGCCCTTGTGTTTGCATAGC	AGTTTGGGGGAGGATGAACT	Em:AZ245759	bM41B10
stSG93036	TCTGACTCCTGGGACAGCTT	TCCACACACAGTGGGGATAA	Em:AZ245736	bM41B10

STS	Sense primer	Antisense primer	End-sequence	Parent clone
stSG93037	TTAGTCACCGGAAGAGGTGG	TAGGCTGGGCTAGAAATCCA	Em:AZ083890	bM20G15
stSG93038	TGGACCTAGAGGGCATCATC	TGGGTGTTTTGTTCCCAAAT	Em:AZ083889	bM20G15
stSG93039	CCCCTACTGCAGCATTTCAT	TGAGGATGCACCAGAGACAG	Em:AZ233207	bM53I23
stSG93040	TTAGCATGTGGCTATGCTGC	CTTGTAAGGGTGGGACTGGA	Em:AZ233206	bM53I23
stSG93041	CTGAGTGCAATAGGGGGAAA	CCTTGGGTGGTTTTCTTTGA	Em:AZ070451	bM422E20
stSG93042	GAGGGGACAATGACCCTGTA	TGTCTCTGCCTTCCAGGTCT	Em:AZ070453	bM422E20
stSG93043	ACACCAGTGGCCAATCTAGG	ATAACCATCTCCCTCCCTGG	Em:AZ055536	bM401E14
stSG93044	AGTGCAGCCAAGCAACAGTA	GAAGCAGATAATTGTGGCCC	Em:AZ055537	bM401E14
stSG93045	AGGCTTATGAGTGGCTGGAA	AATCGACGCACCTTTTATGG	Em:AZ242223	bM90N15
stSG93046	CTTAGTACAGGGCACACGCA	CGGCAGTACAGGGCATAGAT	Em:AZ242221	bM90N15
stSG93047	AACGCCTCTCCACTCAGAAA	GGACTGCCCTGTGACTTCTC	Em:AZ241536	bM75D18
stSG93048	CAGAAGTATGGCCAGCAACA	AAGGATACAGGCATGAACCG	Em:AZ241529	bM75D18
stSG93049	GAGAGAGAGAGCCTGCTGGA	TAGGTTATTAGGGGCCACCC	Em:AZ234410	bM67L5
stSG93050	AAACAGACCAAGCAAAGCGT	CCGTGGATCTTTCTGCTCTC	Em:AZ234406	bM67L5
stSG93051	CCACAGTCATCTGGAGGGAT	AGCCATGTTGTCAATAGCCC	Em:AZ068118	bM392L18
stSG93052	GCCTTGCAGAAACTTTGGAG	GAGAAAGTGGGGAAAAGCCT	Em:AZ068124	bM392L18
stSG93053	ATATGGATCATGGCATGGGT	GGCTCTAGGGTCATTTTGAGG	Em:AZ067326	bM392H4
stSG93054	GAAAGCACTCTGCACCCTTC	GCCAGCATGTGGGTTAAGAT	Em:AZ067328	bM392H4
stSG93055	ATTGGATGGAGCACAAGGTC	CGGGGGTACTGGTTAGTTCA	Em:AZ106933	bM34L21
stSG93056	TCCTGGTTGTAGATGGCACA	CTCACGGTGTGCATGTGTTT	Em:AZ106937	bM34L21
stSG93057	CTGGTTTGTCTTGCTCCCAT	GAGAGCAGCACTCCATAGGC	Em:AZ093224	bM14A22
stSG93058	CCAGAGCAATTCGACAACAA	TGAGAAGGGTGTTGTTTCCC	Em:AZ093226	bM14A22
stSG93059	ATCTGTTGGTGGCCTTTTTG	GAGCAATGGCTTGTGCTGTA	Em:AZ040471	bM329C16
stSG93060	GAAGATGTCCCAACCGGTAA	ACCTCACTCAGGATGATGCC	Em:AQ973463	bM329C16
stSG93061	GGCAGAGGAGCACGTAAAAA	AGCATCTGCTGTCACCTGAG	Em:AZ094011	bM14D12
stSG93063	CCTCCATTGTTGGTGGGATA	TTACCGGTTGGGACATCTTC	Em:AZ237927	bM87H22
stSG93064	CCTAGCATTTGGGGCATAGA	CCAAGAAAATGGTCCCAAGA	Em:AZ237922	bM87H22
stSG93067	TGGCACTAGTACCCACACCA	ACTTGCAGGTTGCCCAATAC	Em:AQ981886	bM303D1
stSG93068	AACCTTCACATGTTGGCACA	TCTGGAATGCTCTTGTGACG	Em:AQ995814	bM303D1
stSG93069	GCTGAAAGGAACGGAGTCTG	TCCTGTGATGCGCTTAGATG	Em:AZ102205	bM4C20
stSG93070	TGCCTGGAAACAAATGAACA	ATGGTTGGATTTGGGAAACA	Em:AZ102206	bM4C20
stSG93071	GTCCTCACGTGGATGAACCT	GTTACTTGCAGACCCCCAAA	Em:AQ931798	bM282E8
stSG93072	CCCACTCCTACACAGCCCTA	GTGTGTGTGCATGCATGTGT	Em:AQ931801	bM282E8
stSG93073	TTTGTTTTGTTGCCAGACCA	CCACAAAAGTGCCTCATCCT	Em:AZ026394	bM345O21
stSG93074	TCCCTTCTCCTCCTCCCTTA	CTGTGTTCCCTCATCCTCGT	Em:AZ026399	bM345O21
stSG93075	GCTGTTGCTCTTGCTAATGC	AGTTCTTTGGGGGCAATTTT	Em:AZ228499	bM54O6
stSG93076	GATGGCTTCGTGGTTAGGAA	CACAATTCAGCACATCAGGG	Em:AZ228496	bM54O6
stSG93077	CCCTTGTTGGCTAGCTTTTG	GAGCCCAGTCATGTGGTTTT	Em:AZ232817	bM68L12
stSG93078	AGAGGCCCAAAATGTCTCCT	TGGTTCTTCCCATTCTCAGC	Em:AZ232811	bM68L12
stSG93079	ACAGGTCCTGCCAAGAGAGA	TCCCATCCCAGCTGTTCTAC	Em:AQ921960	bM275P22
stSG93080	TTTGGCAACTGGTTTTGTCA	TGCACTCTCAGTATCACGGC	Em:AQ921963	bM275P22
stSG93081	TAGTGGATTGCGTTGATGGA	CCGAGAACCAACTCAAAACAA	Em:AZ102380	bM32J9
stSG93082	AACAGTTTGCAAAGGGGCTA	GAAAGCTAAGGGGAGCAAGAA	Em:AZ102377	bM32J9
stSG93083	GGGGGTCCTAGCTAAACAGG	CCCTTCTGTCATGGTGGACT	Em:AZ008703	bM362M10
stSG93084	ATTATTCCCCATTTCCAGCC	GAGGAAGACAGCAGTGGGAG	Em:AZ008707	bM362M10
stSG93085	TACATTTTGCCATTGCTCCA	GCTCTCCCATTTCAGAGCAC	Em:AQ986388	bM326F18
stSG93086	ACCAATCCAACATAGGCTGC	GCTGGCTAGCAAGTCCATTC	Em:AQ986390	bM326F18
stSG93087	CTCCCCATAACACACAGCCT	CTGGGACTGGGCAGTAAAAA	Em:AZ008128	bM348B16

STS	Sense primer	Antisense primer	End-sequence	Parent clone
stSG93088	GCGTGATATGTTCAGGGGTC	GCCCAACAGTAGAAACCCAA	Em:AZ008133	bM348B16
stSG93089	CTCAGCTGTAGCTTCCTGGG	ATGGGTCAGCAGGAGCATAA	Em:AZ056789	bM414O8
stSG93090	TGTTCCTCGCACTAGACCCT	GTGCAGGAAGAACGTAAGGC	Em:AZ056791	bM414O8
stSG93091	GAGGCTCTGTGGGAGAACTG	AGAGCCCATGAATTTGCATC	Em:AQ982691	bM307D13
stSG93092	ACTTTGTACCCAGCCCTCCT	GCATAGGACAGCAGCACTCA	Em:AQ982692	bM307D13
stSG93093	CCTAGGAAGTGCAAGCAAGG	TGTGCCTGCCTGTGAGATAG	Em:AZ114921	bM13J16
stSG93094	GGAGTGATCCCGTGTTAGGA	TAGCCAGCAAGGAAGACTGG	Em:AZ114920	bM13J16
stSG93095	CAGTGCTGCTTCTGCTTCAG	GGTCAGAAAGGAAGATGCCA	Em:AZ279592	bM113I22
stSG93096	CTTGTCCTCTTGCCCACAAT	GAAAGGAGTGAATTGGCCTG	Em:AZ279591	bM113I22
stSG102483	ATGTATGGTGTCCACTCGCA	ATCAGAATTCACCTGGGTCG	Em:AZ041304	bM380K13
stSG102484	CGGCTTAAAGGCAGAACAAG	GCACAATTTCCAGGGAGAAA	Em:AZ041307	bM380K13
stSG102485	AGCTGGCTGTTTCCTCATTG	TGCGAGTGGACACCATACAT	Em:AQ927197	bM257I13
stSG102486	TTTTGCCAATCCTCGATCTT	AGAGGAGAAAGTGGGGAAAAG	Em:AQ971030	bM257I13
stSG102487	ACCTGCTACCTTGCAGATGG	ACAGCTGCAACACAAACTGG	Em:AZ296378	bM160F6
stSG102488	TGTCTGGCGTGAAAGAGATG	AGTCTTCCCAGAGCTGACCA	Em:AZ296399	bM160F6
stSG102489	TCTGACTCCTGGGACAGCTT	TCCACACACAGTGGGGATAA	Em:AZ245736	bM41B10
stSG102490	AGCCCTTGTGTTTGCATAGC	AGTTTGGGGGAGGATGAACT	Em:AZ245759	bM41B10
stSG102491	GTTAGGGTCATGGGATGCAG	TCTGTTCAAGGCAGTTTCCC	Em:AZ230080	bM97B17
stSG102492	CTCAGAAATCTGCCTGCCTC	GGCTGGATTGGTGTCTCAGT	Em:AZ230088	bM97B17
stSG102493	AAAACAGGGGAGATCCAGGT	GGTGAGCCTACCTGGACAGA	Em:AZ227384	bM52H2
stSG102494	GCACCCTGACAACCATTTTT	CACACACACACACACACTATGG	Em:AZ227387	bM52H2
stSG102495	CCAGGCCTGAGAAGACAAAG	AGCCAGCTGAATACCAGCAT	Em:AZ288877	bM129E1
stSG102496	TGGCTTGTGCTGTAAGATCG	CCCAGTCTGTTGGTGGTCTT	Em:AZ288880	bM129E1
stSG102497	GAGGCTCTGTGAACTCCTGG	TACTGAAGAGCCTCCCGCTA	Em:AZ245837	bM90C9
stSG102498	AATCCTCGTGGACTGTCGTC	TTCCGGTACTACACCAAGCC	Em:AZ245843	bM90C9
stSG102499	AGGATGGAGCACAATTCACTC	CATGGTTCCACAGATGTATGC	Em:AZ112090	bM446P19
stSG102500	CAACGGTCATTGCTTTCCTT	CTGAAAGCCTCTCATGGCTC	Em:AZ112093	bM446P19
stSG102501	CCCTTTCTAAAGCTGGACCC	CCTTGCCAGCCAGTTATCAT	Em:AZ289501	bM156F17
stSG102502	CCAGCTTCAGAAAAGTTGGC	CTCCCAACCAGAAAGAGCAG	Em:AZ111263	bM472C2
stSG102503	GTGAGAGCAAACCAAGGAGC	ACACTTGACTGCTGTGTGCC	Em:AZ281170	bM138C10
stSG102504	CAGTTGGAGCATCTTCTGGG	CAAGCGGGTACAACCACTCT	Em:AZ247477	bM58B6
stSG102505	AAACAGACCAAGCAAAGCGT	CCGTGGATCTTTCTGCTCTC	Em:AZ003494	bM372G22
stSG102506	ATTGGACACAGAAGGGGATG	TATCCGGAGGCGATAATCAG	Em:AZ003498	bM372G22
stSG102507	CAGTGCGCAGACTCATTCAT	TAGGGAGAGCGGCTTTTACA	Em:AZ266354	bM127G5
stSG102508	AGGACGAACTCCTTTCACCA	GTCCAGGGGCTTTACCTTCT	Em:AZ266357	bM127G5
stSG102509	TCAGGAATTTTTCCCCTGTG	ACCCAGAAATGAACCCACAC	Em:AQ980679	bM350M23
stSG102510	CATCCCATGAAAATGGAACC	GCTTCCTCTCCTCTGCATTG	Em:AZ034955	bM350M23
stSG102511	TACCTCTGTACCCTGCCCAC	CGGAACTTGTAGGCCATGTT	Em:AZ272974	bM118A2
stSG102512	AACCCAAGGCTTCATGTACG	TGATGGTAGTCGACCCTTCC	Em:AZ273083	bM118A2
stSG102513	TCTGTGGTTTTCACTGTCGC	TAGAAAGGCCCAGAGAAGCA	Em:AZ263712	bM140D8
stSG102514	TGCCCTGTCATGTGATGAAT	GGCTACCCTAAACGAGGGAC	Em:AZ263715	bM140D8
stSG102515	CCGCTGCTTTTCTCTTCATC	AGACCCCACACCAAGACAAG	Em:AZ285540	bM155G13
stSG102516	CCACAAAGACTCCACCCTGT	CTACCAGCATGCACCTCTGA	Em:AZ285544	bM155G13
stSG102517	GACAAGCATGTGTTGGTTGG	AGTTCAGTCCCATGGCAAAC	Em:AZ043050	bM244C17
stSG102518	TATGGCTACAGCCACCATCA	TGGGTGCTACTCACTACCCC	Em:AZ020223	bM299J24
stSG102519	TTTCAAAACCGGGTAGGTGA	TGTTTTGTTCAGGACCTCCC	Em:AZ071388	bM435K19
stSG102520	CATGGAGACAGCAAAGGACA	TCAAATGCTATCCCCAAAGC	Em:AZ086915	bM36K24
stSG102521	GATGGCTTCGTGGTTAGGAA	CACAATTCAGCACATCAGGG	Em:AZ228496	bM54O6

STS	Sense primer	Antisense primer	End-sequence	Parent clone
stSG102522	GCTGTTGCTCTTGCTAATGC	AGTTCTTTGGGGGCAATTTT	Em:AZ228499	bM54O6
stSG102523	CAGAAGTATGGCCAGCAACA	AAGGATACAGGCATGAACCG	Em:AZ241529	bM75D18
stSG102524	AACGCCTCTCCACTCAGAAA	GGACTGCCCTGTGACTTCTC	Em:AZ241536	bM75D18
stSG102525	GAGGGTGGTGCTTCTGAGAG	GGAAGAAGTCGCTGGTCTTG	Em:AZ023266	bM336L6
stSG102526	CAAGGCCACAGCAACTCTTA	TGTCTGCCCATCCTGATGTA	Em:AZ052658	bM336L6
stSG102527	AAAGCAATTCATGGGCAAAC	GGGGAGAGTGTGCATTCTTG	Em:AZ025839	bM316C16
stSG102528	GGGTTGTTCTCAGCTTCTGG	CTGGGGGTTCCAAAGAAAAT	Em:AZ025844	bM316C16
stSG102529	ACCCATTCTGTTCTCTGGAA	CCCACAGATTGGGAAAAGAA	Em:AZ061951	bM414K20
stSG102530	CAGTGCCATTTCAACCAAGA	TGTCCAAGTGAGGCACAAAG	Em:AZ061954	bM414K20
stSG102531	TCTCAAAGCTTTAACCAAAGGC	AAGAACCAAGCAGCATTTGAA	Em:AZ252288	bM57L24
stSG102532	TCCCAATCTGTATGTTGCCA	ACTCACATGGCTCAGGCTCT	Em:AZ252289	bM57L24
stSG102533	AATAAGTCACCGGATGGTGC	AGGAGACATTTTGGGCCTCT	Em:AZ232811	bM68L12
stSG102534	AAAACCACATGACTGGGCTC	GAACCCATGAAGGCAACCTA	Em:AZ232817	bM68L12
stSG102535	AGTCTGCTTGGGGGAATTTT	GTGATATGCCAACGGCTCTT	Em:AZ294646	bM105O8
stSG102536	TTTCTGAAGTCCACCTTCTTGA	CTCCAAAGTTTCTGCAAGGC	Em:AZ294647	bM105O8
stSG102544	TGCCCCTAGCTTCTACCTCA	GACTGAAGCACTCAGGAGCC	Em:AZ267556	bM443O6
stSG102545	TGGGTCTTGGGTTGTCTAGG	TTGTGTCAAAACCCCACAAA	Em:AZ267562	bM443O6
stSG102546	GATGCGTACCACTCCTCCAT	CTGGCCTCAGGGTGTGTATT	Em:AQ995696	bM387A13
stSG102547	CTCACAGTCAAGCAGGGACA	GCAGATGGTCTGGTTCCATT	Em:AQ995699	bM387A13

Appendix 11: Mouse genetic markers mapped on mouse contig.

Name	STS (stSG) number	Sense primer	Antisense primer	Product length (bp)	Position (cM)
D2Mit310	114695	TTTAAATGAAGAATAAGGTCAGAAACA	GCATTAATTCTCATTCTCAATAATGG	138	77.6
D2Mit142	104950	AGCAAGGCATAAGGAGACCA	GGGTGGCTCTATTAGGCACA	118	78.7
D2Mit263.3	104951	TTTCTGCTGTGGGTTCTAAGAA	ATGCGACTCAGGTCTCATGA	111	78.7
D2Mit497	104952	ACTCTGTCTCTGTTTCTATGTCTCTCT	CAACTAAAAGTTCACTTGGTAAAGATT	249	78.7
D2Mit143	104953	ACATTAGAAGGAAATGAAAACATGC	CCCTTTTACTTCCCCATGCT	98	78.7
D2Mit263	104954	ACTGAATCATCTCTCTCCTCAGC	AGTTCAGTTCTTAGAACCCACAGC	138	78.7
D2Mit196	104956	AAACACATGCATGTGCACG	TTCAAAACCCCTCCCTCC	144	78.7
D2Mit452	104957	TCCCACATTTCTGGCATACA	CACATGTGCATTTAAGCATGC	123	78.7
D2Mit453	104958	CCTGAAATTTCCCTTCATAGTAGG	GAAGACACCCACAAGACTAATGC	114	78.7
D2Mit197	104959	CTGGCAGTGACCATGGTG	CATGATCAAAGTACACTCTATTTCCC	147	78.7
D2Mit452.2	104960	GCATGCATGCTTAAATGCAC	GTTAGAGCTCCAGATTCAGTGAGAGA	143	78.7
D2Mit410	104961	TGGAATGTATCCTTTGGGGA	TTGTTTGTTTATTTGTTTGTTCAGG	119	78.7
D2Mit412	104962	ACAGGGCTGCAGAGACCTAA	GACTATCAAAAAGATGGTATTGATGG	125	78.7
D2Mit49	104963	CTGTAACTCCAGGGGATCCA	TGGTGCTCTCAAGGCTAACA	149	79.8
D2Mit51	104964	GTGAGGGGTCAATGCCAC	GGCTCAGTTGTAAGCACAAGG	126	79.8
D2Mit455	104966	CTCCAAGGTCTGATATACATACATACA	TTATGTGCACTGGTGTCTAGCC	121	80.9
D2Mit226	104967	TTTTTGCAACTTTGTTAAGAATTCC	AAAACACCCTCCCACCCTT	101	80.9
D2Mit454	104968	ATTGAGTTGCAGAGGGTAACTAGG	TGAATTGTTCACCACTGGGA	111	80.9
D2Mit498	104969	GCAGCCTTTCCTTCCTTTCT	CAGATAGAGCACTCAGACATACATACA	122	80.9
D2Mit29	104970	CGGTGACGAAGCTTCTGAG	CTTTGAATATGAACTCTCACCTTCC	115	80.9
D2Mit71	103373	CATCTGTGTGACCCACAAGG	ACATCAAAATGCAAGAGGGC	163	80.9
D2Mit170	103374	GGGCTTCCATCAACTCTCTG	CATTGGTGCAGCACTTGC	137	80.9
D2Mit527	103376	AAAGATGGGTGGGCTCTTCT	TTTTTGTAACTCAATCCCCCC	90	82.0
D2Mit289	103377	CTGCCCTTCTCCCTCCTC	CACGTCTTTGTGTGAGAAAACG	129	83.1

D2Mit288	104971	TGCTGTCACTGGGGTTGTTA	TCCTTCCTCTGACCACCAAC	199	83.1
D2Mit311	104972	ACAGGCAGCCTTCCCTTC	TCTGTCCCGCTTCTGTTTCT	126	83.1
D2Mit290.4	104974	AATGCTTGCTCATGCGAATA	GCAGACATGAGAAACCCTGT	150	83.1
D2Mit290	104975	TTATTTTTGGATGAGAGAGAACTGG	AATGGGAGAGAACTGACCCC	191	83.1
D2Mit342	104976	CTCAAGAAACAGCAACAAGGG	CCTGCCTATGTGGCCCTG	92	83.1
D2Mit227	104977	TCTGTCTGCCTGCCTATGTG	TGAGATCTTGTCTCAAGAAACAGC	111	83.1
D2Mit53	104980	GTGGACATTCCCTGAGAAACA	GGGGTTTGATCAGCTCATGT	148	84.2
D2Mit413	104981	GATAATGTCCTCAGAAGGTGGC	AATTTAGCAGGCACTCGTGG	117	84.2
D2Mit145	114697	TGGGGAGGAGACCAGACTC	AAAGGCTTCCGGAAGAGGTA	144	84.2

Appendix 12: Human clone sequences.

Clone name	Library	Sequence accession number	Chromosome	Sequenced by
bA120I11	RPCI-11.1	AL357560	20	Sanger Institute
bA151A11	RPCI-11.1	AL359695	20	Sanger Institute
bA179J15	RPCI-11.1	AL359984	20	Sanger Institute
bA179N14	RPCI-11.1	AL354745	20	Sanger Institute
bA269H4	RPCI-11.1	AL445192	20	Sanger Institute
bA298O6	RPCI-11.2	AL118525	20	Sanger Institute
bA314A4	RPCI-11.2	AL359434	20	Sanger Institute
bA321P16	RPCI-11.2	AL139351	20	Sanger Institute
bA323C15	RPCI-11.2	AL354766	20	Sanger Institute
bA32G22	RPCI-11.1	AL136461	20	Sanger Institute
bA347D21	RPCI-11.2	AL357558	20	Sanger Institute
bA394O2	RPCI-11.2	AL133227	20	Sanger Institute
bA445H22	RPCI-11.2	AL139352	20	Sanger Institute
bA456N23	RPCI-11.2	AL353777	20	Sanger Institute
bA465L10	RPCI-11.2	AL162458	20	Sanger Institute
bK2007A7	CIT-HSP-D1	AL359506	20	Sanger Institute
bK2653D5	CIT-HSP-D2	AL354813	20	Sanger Institute
dJ1005L2	RPCI-5	AL445286	20	Sanger Institute
dJ1013A22	RPCI-5	AL132772	20	Sanger Institute
dJ101A2	RPCI-1	AL133520	20	Sanger Institute
dJ1028D15	RPCI-5	AL121886	20	Sanger Institute
dJ1030M6	RPCI-5	AL035089	20	Sanger Institute
dJ1041C10	RPCI-5	AL162615	20	Sanger Institute
dJ1049G16	RPCI-5	AL034418	20	Sanger Institute
dJ1050C22	RPCI-5	AL121888	20	Sanger Institute
dJ1050K3	RPCI-5	AL121776	20	Sanger Institute
dJ1057D4	RPCI-5	AL121777	20	Sanger Institute
dJ1063B2	RPCI-5	AL035683	20	Sanger Institute
dJ1069P2	RPCI-5	AL109839	20	Sanger Institute
dJ1079N22	RPCI-5	AL161944	20	Sanger Institute
dJ1108D11	RPCI-5	AL034419	20	Sanger Institute
dJ1121H13	RPCI-5	AL049812	20	Sanger Institute
dJ1121P14	RPCI-5	AL138806	20	Sanger Institute
dJ1123D4	RPCI-5	AL049691	20	Sanger Institute
dJ1164I10	RPCI-5	AL049537	20	Sanger Institute
dJ1167E19	RPCI-5	AL133229	20	Sanger Institute
dJ1183I21	RPCI-5	AL035447	20	Sanger Institute
dJ1185N5	RPCI-5	AL034423	20	Sanger Institute
dJ128O17	RPCI-1	AL031654	20	Sanger Institute
dJ138B7	RPCI-1	Z98752	20	Sanger Institute
dJ148E22	RPCI-1	AL008725	20	Sanger Institute
dJ148H17	RPCI-1	AL109825	20	Sanger Institute

Clone name	Library	Sequence accession number	Chromosome	Sequenced by
dJ155G6	RPCI-1	AL121903	20	Sanger Institute
dJ172H20	RPCI-1	AL049767	20	Sanger Institute
dJ179M20	RPCI-1	Z97053	20	Sanger Institute
dJ191L6	RPCI-1	AL009050	20	Sanger Institute
dJ1J6	RPCI-1	AL035652	20	Sanger Institute
dJ211D12	RPCI-1	Z93016	20	Sanger Institute
dJ230I19	RPCI-1	Z93942	20	Sanger Institute
dJ232N11	RPCI-1	AL031656	20	Sanger Institute
dJ237J2	RPCI-1	AL021394	20	Sanger Institute
dJ257E24	RPCI-1	AL034424	20	Sanger Institute
dJ269M15	RPCI-1	AL021395	20	Sanger Institute
dJ272H18	RPCI-1	AL109826	20	Sanger Institute
dJ28H20	RPCI-1	AL031055	20	Sanger Institute
dJ29M10A	RPCI-1	AL390211	20	Sanger Institute
dJ29M10B	RPCI-1	AL390212	20	Sanger Institute
dJ300I2	RPCI-1	AL035660	20	Sanger Institute
dJ337O18	RPCI-3	AL008726	20	Sanger Institute
dJ342O24	RPCI-3	AL133341	20	Sanger Institute
dJ387E22	RPCI-3	AL031660	20	Sanger Institute
dJ3E5	RPCI-1	AL022239	20	Sanger Institute
dJ409O10	RPCI-3	AL031256	20	Sanger Institute
dJ447F3	RPCI-3	AL050348	20	Sanger Institute
dJ450M14	RPCI-3	AL132654	20	Sanger Institute
dJ453C12	RPCI-3	AL021578	20	Sanger Institute
dJ453O12	RPCI-3	AL136102	20	Sanger Institute
dJ461P17	RPCI-3	AL031663	20	Sanger Institute
dJ470L14	RPCI-3	AL133174	20	Sanger Institute
dJ47A22	RPCI-1	AL117374	20	Sanger Institute
dJ485M8	RPCI-3	AL353797	20	Sanger Institute
dJ495O3	RPCI-3	AL121587	20	Sanger Institute
dJ508O2	RPCI-3	AL354812	20	Sanger Institute
dJ511B24	RPCI-3	AL022394	20	Sanger Institute
dJ540H1	RPCI-4	AL121674	20	Sanger Institute
dJ569M23	RPCI-4	AL031666	20	Sanger Institute
dJ599F21	RPCI-4	AL035662	20	Sanger Institute
dJ601O1	RPCI-4	AL109656	20	Sanger Institute
dJ620E11	RPCI-4	AL031667	20	Sanger Institute
dJ644L1	RPCI-4	AL035665	20	Sanger Institute
dJ661I20	RPCI-4	AL031669	20	Sanger Institute
dJ66N13	RPCI-1	AL137078	20	Sanger Institute
dJ686N3	RPCI-4	AL049766	20	Sanger Institute
dJ688G8	RPCI-4	AL031671	20	Sanger Institute
dJ690O1	RPCI-4	AL118521	20	Sanger Institute
dJ707K17	RPCI-4	AL024473	20	Sanger Institute

Clone name	Library	Sequence accession number	Chromosome	Sequenced by
dJ710H13	RPCI-4	AL121712	20	Sanger Institute
dJ730D4	RPCI-4	AL035666	20	Sanger Institute
dJ73E16	RPCI-1	Z95330	20	Sanger Institute
dJ753D4	RPCI-4	AL031676	20	Sanger Institute
dJ781B1	RPCI-4	AL118522	20	Sanger Institute
dJ791K14	RPCI-4	AL035685	20	Sanger Institute
dJ796I11	RPCI-4	AL031257	20	Sanger Institute
dJ81G23	RPCI-1	AL035459	20	Sanger Institute
dJ824J5	RPCI-5	AL034552	20	Sanger Institute
dJ839B11	RPCI-5	AL121778	20	Sanger Institute
dJ862K6	RPCI-5	AL031681	20	Sanger Institute
dJ881L22	RPCI-5	AL117382	20	Sanger Institute
dJ890O15	RPCI-5	AL049540	20	Sanger Institute
dJ892M9	RPCI-5	AL121828	20	Sanger Institute
dJ906C1	RPCI-5	AL133342	20	Sanger Institute
dJ914M10	RPCI-5	AL121763	20	Sanger Institute
dJ94E24	RPCI-1	AL050317	20	Sanger Institute
dJ963K23	RPCI-5	AL031685	20	Sanger Institute
dJ970A17	RPCI-5	AL034431	20	Sanger Institute
dJ981L23	RPCI-5	AL031686	20	Sanger Institute
dJ991B18	RPCI-5	AL049541	20	Sanger Institute
dJ993C19	RPCI-5	AL121786	20	Sanger Institute
dJ995J12	RPCI-5	AL035462	20	Sanger Institute
dJ998C11	RPCI-5	AL035106	20	Sanger Institute
dJ998H6	RPCI-5	AL031687	20	Sanger Institute

Appendix 13: Mouse clone sequences.

Clone name	Library	Sequence accession number	Chromosome	Sequenced by
bM100C4	RPCI-23	AL672162	2	Sanger Institute
bM105M23	RPCI-23	AL591854	2	Sanger Institute
bM108D12	RPCI-23	AL589902	2	Sanger Institute
bM109E10	RPCI-23	AL589874	2	Sanger Institute
bM117O11	RPCI-23	AL589876	2	Sanger Institute
bM118A2	RPCI-23	AL589870	2	Sanger Institute
bM120P1	RPCI-23	AL589873	2	Sanger Institute
bM126L18	RPCI-23	AL591586	2	Sanger Institute
bM129E1	RPCI-23	AL645794	2	Sanger Institute
bM138C10	RPCI-23	AL591712	2	Sanger Institute
bM140D14	RPCI-23	AL591478	2	Sanger Institute
bM143E11	RPCI-23	AL591967	2	Sanger Institute
bM144O20	RPCI-23	AL591490	2	Sanger Institute
bM152H17	RPCI-23	AL591905	2	Sanger Institute
bM161B3	RPCI-23	AL731698	2	Sanger Institute
bM183N8	RPCI-23	AL591884	2	Sanger Institute
bM188I17	RPCI-23	AL669917	2	Sanger Institute
bM190L21	RPCI-23	AL591970	2	Sanger Institute
bM19L12	RPCI-23	AL591711	2	Sanger Institute
bM206I14	RPCI-23	AL590418	2	Sanger Institute
bM215C14	RPCI-23	AL591607	2	Sanger Institute
bM216D20	RPCI-23	AL591911	2	Sanger Institute
bM217C2	RPCI-23	AL669906	2	Sanger Institute
bM235I24	RPCI-23	AL731671	2	Sanger Institute
bM272C14	RPCI-23	AL591598	2	Sanger Institute
bM272O14	RPCI-23	AL591606	2	Sanger Institute
bM28B10	RPCI-23	AL591936	2	Sanger Institute
bM305K11	RPCI-23	AL590414	2	Sanger Institute
bM321M14	RPCI-23	AL591542	2	Sanger Institute
bM326P18	RPCI-23	AL672196	2	Sanger Institute
bM327A19	RPCI-23	AL606473	2	Sanger Institute
bM333A18	RPCI-23	AL591665	2	Sanger Institute
bM335N12	RPCI-23	AL591584	2	Sanger Institute
bM338H13	RPCI-23	N/A	2	Sanger Institute
bM345I2	RPCI-23	AL626766	2	Sanger Institute
bM346D16	RPCI-23	AL591512	2	Sanger Institute
bM36P22	RPCI-23	AL591488	2	Sanger Institute
bM370H21	RPCI-23	AL591127	2	Sanger Institute
bM380K13	RPCI-23	AL592042	2	Sanger Institute
bM383K1	RPCI-23	AL732357	2	Sanger Institute
bM384K10	RPCI-23	AL590430	2	Sanger Institute
bM392O2	RPCI-23	AL596263	2	Sanger Institute

Clone name	Library	Sequence accession number	Chromosome	Sequenced by
bM393F23	RPCI-23	AL590389	2	Sanger Institute
bM395E18	RPCI-23	AL591064	2	Sanger Institute
bM401I8	RPCI-23	AL663062	2	Sanger Institute
bM41B10	RPCI-23	AL645766	2	Sanger Institute
bM41B20	RPCI-23	AL591430	2	Sanger Institute
bM420L2	RPCI-23	AL591675	2	Sanger Institute
bM428M13	RPCI-23	AL591411	2	Sanger Institute
bM429O12	RPCI-23	AL732312	2	Sanger Institute
bM443O6	RPCI-23	AL645827	2	Sanger Institute
bM448P12	RPCI-23	AL591703	2	Sanger Institute
bM462O16	RPCI-23	AL590429	2	Sanger Institute
bM465I6	RPCI-23	AL591762	2	Sanger Institute
bM466K24	RPCI-23	AL606841	2	Sanger Institute
bM471I9	RPCI-23	AL591673	2	Sanger Institute
bM474J7	RPCI-23	AL645736	2	Sanger Institute
bM479C2	RPCI-23	AL669913	2	Sanger Institute
bM480D17	RPCI-23	AL591763	2	Sanger Institute
bM53I23	RPCI-23	AL732362	2	Sanger Institute
bM53L16	RPCI-23	AL591882	2	Sanger Institute
bM61O3	RPCI-23	AL591495	2	Sanger Institute
bM79H8	RPCI-23	AL669836	2	Sanger Institute
bM90N15	RPCI-23	AL591805	2	Sanger Institute
bM97B17	RPCI-23	AL590415	2	Sanger Institute
bN223B6	RPCI-24	AL732310	2	Sanger Institute

Appendix 14: Exonic SNPs. Novel SNPs were submitted to dbSNP (http://www.ncbi.nlm.nih.gov/SNP/).

Clone	cSNP coordinates	Gene	Allele 1 (reference sequence)	Allele 2	Allele 2 supporting evidence (A) [b]	Allele 2 supporting evidence (B) [b]
dJ511B24	64,001	PLCG1	T	C	BF345779	AL045204
dJ511B24	73,832	TIX1	G	A	AW340006	AI885778
dJ511B24	73,854	TIX1	C	T	AI885778	BE856363
dJ511B24	80,265	TIX1	G	A	AI867932	AA969866
dJ620E11	2,329	LPIN3	A	T	BF055162	AI127104
dJ620E11	44,899	KIAA1335	G	A	AW235493	BF063641
dJ620E11	45,017	KIAA1335	T	G	AW235493	AI951142
dJ1121H13	93,709	PTPRT	G	A	R50970	F02638
dJ1121H13	95,065	PTPRT	G	A	AL138251	AW577442
dJ862K6	107,817	SFRS6	T	C	BE937954	AV758335
dJ138B7	73,871	C20orf9	G	A	BE788978	AV703478
dJ1028D15	93,173 [a]	MYBL2	T	C	BE265253	AA356175
dJ1028D15	96,061	MYBL2	C	G	BG114635	AA496491
dJ179M20	135,657	PKIG	C	T	AI422674	AI679348
dJ179M20	135,660	PKIG	C	T	AW340090	AW167468
dJ179M20	135,869	PKIG	C	T	AI814404	BF109931
dJ179M20	135,879	PKIG	C	T	AI422674	BE328253
dJ179M20	135,890	PKIG	A	C	T77999	H51415
dJ179M20	14,353	TDE1	G	A	AW044595	AI188582
dJ179M20	14,482	TDE1	C	T	AI188582	BF111462
dJ179M20	14,497	TDE1	C	T	AI188582	AI299818
bA445H22	97,067	WISP2	C	T	AI807970	AI973060
dJ148E22	56,722	YWHAB	A	C	AU132709	BF306710
dJ148E22	60,103	YWHAB	T	A	BE709497	BE709825
dJ148E22	62,399	YWHAB	A	G	R19741	H12905
dJ148E22	62,943	YWHAB	C	T	AA167429	BE895870
dJ148E22	63,061	YWHAB	C	T	AV721832	AW955480
dJ1069P2	89,939	TOMM34	A	G	H75519	BE933215
dJ1069P2	94,339	C20orf119	G	A	AW967588	AI689512
dJ1069P2	94,345	C20orf119	G	A	BG152831	AW243900
dJ1069P2	95,369	C20orf119	T	C	AW954161	BE177459
dJ1069P2	100,623	C20orf119	C	T	AA251071	BF818192
dJ172H20	50,495	PI3	A	C	BF094327	BE772171
dJ172H20	51,122	PI3	C	T	BF002099	AI924155
dJ172H20	51,151	PI3	C	T	AW082097	AI807596
dJ300I2	27,359	SLPI	G	A	BE044073	AI222907
dJ453C12	36,688 [a]	SDC4	G	C	AA912436	W80465
dJ453C12	78,510	C20orf169	T	G	AV702492	BE781485
dJ453C12	78,642	C20orf169	G	T	AI343524	BF002486
dJ453C12	78,669	C20orf169	A	C	AI808184	AI681435
dJ453C12	78,762	C20orf169	C	T	BG116204	R55837
dJ453C12	79,096 [a]	C20orf169	C	T	AA258683	AI096624

Clone	cSNP coordinates	Gene	Allele 1 (reference sequence)	Allele 2	Allele 2 supporting evidence (A) [b]	Allele 2 supporting evidence (B) [b]
dJ453C12	79,870	C20orf169	A	G	AW994246	BE166488
dJ453C12	80,132 [a]	C20orf169	C	T	AI126562	T17223
dJ453C12	80,276 [a]	C20orf169	C	A	AI573128	AA258588
dJ453C12	121,204	C20orf35	A	G	BE908675	AW371128
dJ453C12	135,622	PIGT	G	A	AW370635	AW361708
dJ453C12	135,781	PIGT	C	T	AI907602	AI907608
dJ453C12	136,979 [a]	PIGT	T	C	BE869193	AA309604
dJ453C12	137,317	PIGT	A	C	AI201054	AW167891
dJ453C12	137,395	PIGT	A	C	AI670082	AI201054
dJ461P17	105,824	SLPINLW1	C	A	AA860417	AW182111
dJ461P17	119,487	C20orf170	T	C	AL449578	AL449581
dJ447F3	43,911	WFDC3	C	T	AI985094	AI243277
dJ447F3	61,679	C20orf167	G	A	AW250651	BF337137
dJ447F3	61,724	C20orf167	T	C	BE267410	AW250529
dJ447F3	92,943 [a]	TNNC2	T	C	AI313386	F30367
dJ447F3	93,739 [a]	TNNC2	C	A	W94954	F00841
dJ337O18	14,252	PTE1	G	A	AW025493	AI435343
dJ337O18	51,519	C20orf162	C	A	AL135265	R06780
dJ337O18	70,945	PPGB	C	T	AI924882	AW148956
dJ337O18	70,916	PPGB	C	A	AI818205	AW337488
bA465L10	34,279	C20orf67	T	C	AA865977	AI351098
bA465L10	34,427	C20orf67	A	C	AW516562	AI024388
bA465L10	98,002	MMP9	A	G	BE831794	T64837
bA465L10	100,183	MMP9	G	C	AA078733	AW838034
bA465L10	102,787	MMP9	C	T	AI075104	AI268673
bA465L10	102,880	MMP9	C	T	AI241706	AI819628
bA465L10	146,507	SLC12A5	G	A	R38486	H09108
bA465L10	147,505	NCOA5	G	A	AI148312	BE675918
bA394O2	47,862	KIAA1834	G	T	AI982596	AW449931
bA394O2	49,352	KIAA1834	A	G	AI933679	AI653338
dJ257E24	46,793 [a]	SLC13A3	A	G	AA808117	AI634218
dJ257E24	47,057	SLC13A3	C	A	AA558111	R50872
dJ28H20	2,734	C20orf64	G	A	AV648314	AV648389
dJ28H20	2,761	C20orf64	C	G	AV648314	AV648431
dJ28H20	48,458 [a]	SLC2A15	T	C	N44116	N40351
dJ890O2	28,401	EYA2	G	A	AW392633	AI368727
dJ890O15	28,475	EYA2	G	A	AW392633	AW392620
dJ890O2	28,507	EYA2	A	C	AW392629	BF060660
dJ890O2	64,289	PRKCBP1	A	G	AA102321	BE884364
dJ29M10B	2,058	PRKCBP1	C	T	T06694	AW498886
dJ29M10B	3,281	PRKCBP1	C	A	AA504355	AI671052
dJ1049G16	55,543	NCOA3	G	A	BE733896	BE902710
dJ1049G16	67,054	NCOA3	A	G	AF010227	AW503332
dJ1049G16	67,057	NCOA3	G	A	AA065271	AU138524
dJ1049G16	67,084	NCOA3	A	G	AF010227	AW502847

Clone	cSNP coordinates	Gene	Allele 1 (reference sequence)	Allele 2	Allele 2 supporting evidence (A) [b]	Allele 2 supporting evidence (B) [b]
dJ1049G16	71,448	NCOA3	C	G	AI808698	AW015672
dJ1049G16	71,612 [a]	NCOA3	A	T	AI808698	AI809327
dJ1049G16	71,620 [a]	NCOA3	G	A	AA229716	BE168160
bA347D21	12,833	bA347D21.1	C	T	AI954602	AA974244
bA347D21	45,157	bA347D21.4	T	C	AA972413	AI219998
bA269H4	52,232	KIAA1415	T	G	AW663740	AI204500
bA269H4	52,338	KIAA1415	C	A	AW039564	AA029710
bA269H4	52,481	KIAA1415	G	A	AI809960	AW663740
bA269H4	52,555	KIAA1415	G	A	AI810901	AI394548
bA269H4	52,577	KIAA1415	C	T	AI765687	AI204500
bA269H4	52,796	KIAA1415	C	T	AI459219	BF111136
bA269H4	53,056 [a]	KIAA1415	A	G	AW073339	AI204980
bA269H4	53,141	KIAA1415	C	G	T90531	T81327
bA269H4	53,287	KIAA1415	G	A	AW073339	AI638231
bA269H4	53,463	KIAA1415	C	A	BE729461	BF027564
bA269H4	64,588	KIAA1415	T	C	AB037836	AX018071
dJ998C11	2,266	KIAA1415	G	C	BF309614	BF204719
dJ155G6	36,373 [a]	ARFGEF2	A	G	AI873207	N23053
dJ155G6	37,064	ARFGEF2	A	C	AI741843	AI671441
dJ155G6	63,791	CSE1L	T	C	AL048759	L44546
dJ155G6	63,887	CSE1L	A	C	BF361281	AW386335
dJ470L14	16,598	CSE1L	C	T	AV723252	BE004378
dJ686N3	10,895 [a]	DDX27	C	T	BF305904	BE018458
dJ686N3	23,191	KIAA1404	T	G	AI139877	AI090233
dJ686N3	23,242	KIAA1404	G	A	BE219265	AI202444
dJ686N3	23,352	KIAA1404	G	A	AI085706	AI859676
dJ686N3	34,718	KIAA1404	A	G	T92574	W72798
dJ686N3	34,741	KIAA1404	G	C	BE837809	BF359333
dJ686N3	58,171	dJ686N3.3	C	G	N47913	H95224
dJ791K14	39,149	KCNB1	T	G	H84228	AA069746
dJ791K14	39,329	KCNB1	A	G	AI885070	AI360273
dJ1063B2	41,068	B4GALT5	T	C	AI039893	AA780169
dJ1063B2	41,840	B4GALT5	C	A	AA699676	AI248228
dJ1185N5	39,338	UBE2V1	G	A	AI373095	AI042109
dJ1185N5	39,348	UBE2V1	T	A	AW664872	AI499008
dJ1185N5	39,418	UBE2V1	T	G	AA603807	AW058376
dJ1185N5	82,102	UBE2V1	G	A	BF589921	AW235963

[a] These candidate cSNPs were previously identified by other SNP discovery projects and were not submitted to dbSNP.

[b] EMBL accession numbers for expressed sequences harbouring Allele 2 (only two shown -A, B- for clarity).

Appendix 15: Verified polymorphic SNPs across the region (MAF□5%). All SNPs reported have been genotyped in at least 13 Caucasian individuals. SNPs without a dbSNP identifier (rs#) are listed with their provisional Sanger Institute SNP identifier (G#).

Position	SNP	DNAs tested	A1	AF1	A2	AF2
4554	rs2425387	47	A	0.904	G	0.095
7185	rs2425388	38	C	0.052	T	0.947
8938	rs926258	47	A	0.797	G	0.202
9094	rs926259	46	A	0.815	C	0.184
11833	rs1033379	36	C	0.777	T	0.222
12992	rs2425392	42	A	0.095	G	0.904
14104	rs2425393	43	C	0.116	T	0.883
20038	rs2425398	34	C	0.882	T	0.117
29387	rs2179084	38	A	0.144	G	0.855
43257	rs1883146	47	C	0.585	T	0.414
60887	rs2010380	44	A	0.568	G	0.431
71032	rs1474568	42	G	0.88	T	0.119
81752	rs1883147	47	A	0.691	G	0.308
88376	rs1406965	32	C	0.562	T	0.437
89152	rs983835	46	G	0.554	T	0.445
145216	rs1570019	42	C	0.107	T	0.892
159817	rs911071	45	C	0.122	T	0.877
160893	rs1321337	35	A	0.142	G	0.857
162526	rs1321338	38	C	0.855	T	0.144
167256	rs1321339	41	A	0.841	G	0.158
172154	rs1358787	47	C	0.851	T	0.148
176776	rs1321333	44	A	0.511	G	0.488
197178	rs1980592	37	A	0.202	T	0.797
197774	rs2870432	44	C	0.625	G	0.375
198768	rs1406964	35	C	0.7	T	0.3
202151	rs1004483	45	A	0.944	G	0.055
217514	G3943295	38	C	0.684	G	0.315
230663	G3956444	41	A	0.378	G	0.621
316161	rs2224282	47	C	0.861	G	0.138
318642	rs1535209	47	G	0.851	T	0.148
319722	rs980355	45	A	0.855	G	0.144
327061	G4052842	42	C	0.845	G	0.154
334900	G4060681	42	C	0.892	T	0.107
357971	G4083752	36	C	0.694	T	0.305
394049	G4119830	46	A	0.054	G	0.945
418621	rs2902940	33	A	0.651	G	0.348
442762	rs926728	14	A	0.178	G	0.821
449916	rs2207135	46	A	0.358	C	0.641
470432	rs1022581	45	G	0.644	T	0.355
500961	rs1883713	38	A	0.276	G	0.723
503981	rs2903023	39	A	0.628	T	0.371
546194	G4271975	45	G	0.3	T	0.7
572909	rs926663	25	A	0.72	G	0.28
582682	G4308463	36	G	0.625	T	0.375
597902	G4323683	35	C	0.6	T	0.4
613920	rs910135	19	A	0.736	G	0.263
625118	rs882192	47	C	0.148	G	0.851
625324	rs882191	43	C	0.848	T	0.151
634797	rs742745	46	G	0.902	T	0.097
641804	rs3577	33	A	0.075	G	0.924
653649	rs932440	46	A	0.086	G	0.913
661618	rs2425419	38	A	0.407	G	0.592
664907	rs742742	42	A	0.785	G	0.214
666908	rs2064406	41	A	0.17	C	0.829
669926	rs1883839	43	A	0.906	G	0.093
670212	rs1883841	37	A	0.918	T	0.081
672400	rs2425421	38	A	0.684	G	0.315
674491	rs2425425	38	G	0.315	T	0.684
679318	rs2425429	39	C	0.32	T	0.679
684599	rs723080	44	A	0.454	G	0.545
684831	rs723081	46	G	0.413	T	0.586
694631	rs2425447	24	C	0.395	T	0.604
696403	rs2425451	39	C	0.692	T	0.307
700961	rs2050204	29	A	0.275	G	0.724
701132	rs2064346	46	C	0.597	T	0.402
702826	rs732961	40	A	0.787	G	0.212
704415	rs1883702	43	A	0.627	G	0.372
706329	rs2903114	40	A	0.325	T	0.675
707799	rs974672	44	C	0.625	G	0.375
708884	rs1007563	37	A	0.445	T	0.554
710513	rs2206670	38	A	0.157	G	0.842
714991	rs1160307	40	A	0.825	G	0.175
715348	rs1014748	45	C	0.855	T	0.144
721866	rs2143513	47	A	0.861	G	0.138
725306	rs2206665	46	C	0.858	T	0.141
729663	rs1535025	47	A	0.393	G	0.606
737670	rs742912	47	A	0.872	C	0.127
738283	rs761566	46	G	0.076	T	0.923
746107	rs1546905	40	A	0.075	G	0.925
751407	rs2206663	43	A	0.593	C	0.406
755956	rs2223657	47	C	0.382	T	0.617
764791	rs2206662	46	C	0.663	T	0.336
789607	rs932489	46	C	0.597	T	0.402
792672	rs2865899	46	C	0.565	T	0.434
802273	rs1000349	28	A	0.232	G	0.767
814352	rs1005534	40	A	0.1	G	0.9
821901	G4547682	43	A	0.476	G	0.523
831903	rs1884101	37	C	0.459	T	0.54
847921	G4573702	41	C	0.5	T	0.5
853184	rs2425452	37	A	0.121	G	0.878

Position	SNP	DNAs tested	A1	AF1	A2	AF2
894908	G4620689	43	C	0.825	T	0.174
904876	rs2207309	47	C	0.244	T	0.755
906767	rs1548244	39	A	0.846	G	0.153
918789	rs2425458	36	G	0.75	T	0.25
926135	rs1980479	47	A	0.223	G	0.776
938936	rs734532	46	C	0.336	T	0.663
942079	rs2011809	42	A	0.297	G	0.702
961500	rs1000410	47	C	0.585	G	0.414
968651	rs1883511	47	A	0.914	C	0.085
994200	G4719981	30	G	0.183	T	0.816
1040486	rs2076576	47	G	0.191	T	0.808
1047964	rs2076575	38	C	0.184	T	0.815
1062430	G4788211	38	C	0.276	T	0.723
1099081	rs926345	47	C	0.553	T	0.446
1115071	rs2072881	45	C	0.355	T	0.644
1124599	rs753381	36	C	0.555	T	0.444
1140863	rs2664537	39	A	0.192	G	0.807
1157256	rs2235367	43	A	0.43	G	0.569
1204879	rs1543324	46	C	0.554	T	0.445
1224157	rs1540908	39	C	0.474	T	0.525
1237238	rs1018389	46	C	0.445	G	0.554
1245391	G4971172	42	C	0.535	G	0.464
1253467	G4979248	42	C	0.892	G	0.107
1335364	G5061145	44	A	0.125	G	0.875
1350180	G5075961	46	C	0.782	T	0.217
1357254	G5083035	44	C	0.636	T	0.363
1368612	rs742434	44	A	0.284	G	0.715
1375774	G5101555	40	C	0.325	T	0.675
1383770	rs761024	47	A	0.382	C	0.617
1403642	rs909882	47	C	0.095	T	0.904
1426098	rs2143227	44	C	0.681	T	0.318
1439033	rs2143228	45	A	0.277	G	0.722
1439621	rs2072970	23	C	0.673	T	0.326
1445038	rs2143229	20	A	0.05	G	0.95
1452592	rs742431	43	A	0.302	C	0.697
1456363	rs1010901	46	C	0.228	G	0.771
1471412	rs967083	22	C	0.818	T	0.181
1479616	G5205397	42	A	0.214	G	0.785
1489747	rs987343	39	C	0.102	G	0.897
1490588	rs2866742	40	A	0.287	G	0.712
1498079	G5223860	43	C	0.127	T	0.872
1512397	rs2272959	46	A	0.913	G	0.086
1518785	rs2294579	45	C	0.077	G	0.922
1533029	rs2143232	47	C	0.287	T	0.712
1557352	rs2903379	45	A	0.244	G	0.755
1574654	G5300435	32	A	0.125	C	0.875
1604327	G5330108	40	C	0.95	T	0.05
1617939	rs941796	15	A	0.3	G	0.7
1648420	rs1885174	39	A	0.743	C	0.256
1657495	rs2866705	47	C	0.436	T	0.563
1678813	rs2866707	41	C	0.817	T	0.182
1708376	rs2038525	26	A	0.942	T	0.057
1709373	rs2057343	39	A	0.205	C	0.794
1726391	rs1465122	44	C	0.772	T	0.227
1750115	rs2208523	43	C	0.767	T	0.232
1796399	G5522180	42	C	0.833	T	0.166
1806636	rs2866823	43	A	0.406	G	0.593
1809842	rs2866824	45	C	0.566	T	0.433
1814245	rs2211353	39	G	0.666	T	0.333
1829405	rs1543344	44	C	0.795	G	0.204
1851605	rs2076394	38	A	0.447	C	0.552
1853697	rs926477	47	A	0.17	G	0.829
1870460	rs1569602	23	C	0.152	T	0.847
1874449	rs2206413	27	C	0.851	T	0.148
1879025	rs2206414	35	A	0.542	G	0.457
1894256	rs2050083	25	A	0.58	G	0.42
1898650	rs926478	45	A	0.088	G	0.911
1918558	rs2425465	37	A	0.905	G	0.094
1928154	rs1000337	31	C	0.79	T	0.209
1931677	rs2866986	41	A	0.378	C	0.621
1935741	rs875791	35	C	0.471	T	0.528
1975294	rs2223912	44	C	0.056	T	0.943
1985051	rs2223913	39	A	0.897	G	0.102
1993634	rs2207219	45	C	0.388	T	0.611
1996863	rs2207220	45	C	0.766	T	0.233
2001918	rs748526	43	A	0.639	G	0.36
2006871	rs2144006	45	A	0.6	T	0.4
2015649	G5741430	41	A	0.634	G	0.365
2023360	rs1010377	45	C	0.633	T	0.366
2024905	rs733673	41	A	0.353	C	0.646
2028589	rs2596	40	A	0.937	G	0.062
2029945	rs2664587	31	A	0.435	G	0.564
2033370	rs1884039	33	C	0.712	T	0.287
2033428	rs877431	44	C	0.261	T	0.738
2035012	rs2866943	46	C	0.76	T	0.239
2038778	rs750782	46	A	0.38	G	0.619
2041130	rs2072913	36	A	0.638	G	0.361
2045292	rs1126101	41	A	0.585	G	0.414
2047007	rs2866947	21	A	0.738	G	0.261
2048086	rs2425470	37	G	0.283	T	0.716
2057648	rs2425473	40	C	0.362	T	0.637
2063797	rs2425478	39	C	0.32	T	0.679
2068970	rs742293	42	C	0.702	T	0.297
2086583	G5812364	44	C	0.068	T	0.931
2097057	rs2076241	26	A	0.5	G	0.5
2117570	rs2867061	46	G	0.576	T	0.423

Position	SNP	DNAs tested	A1	AF1	A2	AF2
2117766	rs2867062	47	A	0.414	C	0.585
2129393	rs2223424	47	G	0.787	T	0.212
2132140	rs2143008	33	A	0.409	G	0.59
2134051	rs873538	47	C	0.212	T	0.787
2135798	rs2903465	42	A	0.619	T	0.38
2135865	rs2867063	47	A	0.361	G	0.638
2144981	rs932358	47	C	0.574	T	0.425
2149546	rs979634	30	C	0.45	T	0.55
2181755	rs926647	45	G	0.422	T	0.577
2206264	rs1984399	47	A	0.425	G	0.574
2211762	rs727337	38	A	0.421	T	0.578
2220892	G5946673	42	A	0.559	G	0.44
2251124	G5976905	40	C	0.762	T	0.237
2283026	G6008807	26	A	0.557	G	0.442
2291285	rs763474	41	A	0.451	C	0.548
2306563	rs929071	47	C	0.776	T	0.223
2323710	rs2867437	41	A	0.5	C	0.5
2333691	rs1156500	37	C	0.864	G	0.135
2336119	rs1155382	47	C	0.095	G	0.904
2339182	rs916325	40	C	0.637	G	0.362
2349397	rs2425484	41	C	0.134	T	0.865
2362723	rs932320	39	A	0.115	T	0.884
2371615	rs1883269	39	A	0.82	G	0.179
2372943	rs2179217	46	A	0.163	G	0.836
2380014	rs978457	47	C	0.308	T	0.691
2388003	G6113784	40	C	0.225	G	0.775
2397665	rs760703	46	A	0.206	G	0.793
2403672	rs2076081	47	C	0.372	G	0.627
2435278	rs126180	45	C	0.944	T	0.055
2437054	rs230155	45	A	0.133	G	0.866
2441473	rs230158	47	A	0.053	G	0.946
2460772	rs230165	40	A	0.3	G	0.7
2463694	rs2205939	47	A	0.063	G	0.936
2466037	rs230170	47	C	0.946	T	0.053
2477389	G6203170	42	C	0.059	T	0.94
2488646	rs986830	47	C	0.234	G	0.765
2497071	rs2867484	47	C	0.351	T	0.648
2500464	rs2224166	41	G	0.426	T	0.573
2515451	G6241232	46	A	0.51	G	0.489
2537879	rs764444	41	A	0.414	G	0.585
2545205	rs1015387	47	A	0.819	G	0.18
2550195	rs1883842	44	G	0.329	T	0.67
2551695	rs722556	18	A	0.555	G	0.444
2558099	rs2206905	30	A	0.9	T	0.1
2559658	rs2057074	47	A	0.212	C	0.787
2566051	rs877440	29	G	0.086	T	0.913
2579381	rs874921	47	A	0.829	C	0.17
2580132	rs909862	45	A	0.166	G	0.833

Position	SNP	DNAs tested	A1	AF1	A2	AF2
2583901	rs997893	45	A	0.355	G	0.644
2591533	rs2867490	47	C	0.372	T	0.627
2603355	rs1157668	40	A	0.25	T	0.75
2605363	rs2425500	45	C	0.711	T	0.288
2610109	rs2425503	43	A	0.488	G	0.511
2617371	G6343152	28	C	0.339	T	0.66
2635865	rs2206428	37	C	0.743	T	0.256
2647357	rs2425528	47	A	0.829	G	0.17
2658009	rs926481	42	C	0.107	T	0.892
2661715	rs742420	42	C	0.75	T	0.25
2664443	rs1040481	47	A	0.712	G	0.287
2670739	rs976249	46	C	0.097	G	0.902
2686025	G6411806	36	A	0.902	G	0.097
2696722	rs1022819	22	A	0.84	G	0.159
2699553	rs2867554	33	C	0.06	T	0.939
2704999	rs2038668	38	C	0.75	T	0.25
2729339	rs2867555	39	A	0.756	G	0.243
2733627	rs2146615	36	C	0.666	T	0.333
2734195	rs980929	47	A	0.521	G	0.478
2739710	rs1973949	42	A	0.416	G	0.583
2741210	rs208254	43	A	0.523	C	0.476
2744591	rs208248	29	C	0.344	G	0.655
2748169	rs208243	18	A	0.888	C	0.111
2750589	rs68049	38	A	0.236	G	0.763
2770522	rs208219	41	A	0.329	G	0.67
2772116	rs208220	41	A	0.304	G	0.695
2806215	rs208262	34	A	0.5	G	0.5
2811818	rs208268	45	C	0.833	T	0.166
2816214	rs208269	47	C	0.234	T	0.765
2822030	rs208274	40	A	0.787	T	0.212
2827812	rs208186	43	G	0.267	T	0.732
2828293	rs208187	46	A	0.771	G	0.228
2829817	rs172303	35	C	0.785	G	0.214
2833075	rs208195	43	A	0.72	C	0.279
2886731	rs761026	42	A	0.726	C	0.273
2926128	G6651909	44	A	0.829	C	0.17
2936321	rs2867591	47	A	0.191	G	0.808
2950271	rs2223558	38	A	0.184	C	0.815
2953774	rs2206456	29	A	0.344	C	0.655
2960216	rs2294591	29	A	0.741	T	0.258
2973694	rs2223556	39	A	0.705	G	0.294
2976180	rs206150	47	C	0.287	T	0.712
2978981	rs206156	47	C	0.18	T	0.819
2984149	rs206162	47	C	0.819	G	0.18
2990143	rs2208048	41	A	0.207	T	0.792
2990594	rs2224259	47	C	0.819	T	0.18
2996596	rs2425552	44	C	0.818	T	0.181
2998906	rs2425559	45	A	0.822	G	0.177

Position	SNP	DNAs tested	A1	AF1	A2	AF2
2999004	rs2425561	44	C	0.829	T	0.17
3001641	rs910829	26	C	0.134	G	0.865
3003275	rs2425567	47	C	0.191	G	0.808
3009655	rs984962	30	C	0.15	T	0.85
3019388	rs755149	47	A	0.127	G	0.872
3021342	rs985693	47	C	0.755	T	0.244
3021882	rs2425578	42	C	0.809	T	0.19
3031763	rs990773	34	C	0.676	T	0.323
3041698	rs1006746	47	C	0.67	T	0.329
3053856	rs2425589	44	C	0.625	G	0.375
3055675	rs2425591	43	G	0.627	T	0.372
3057442	rs927058	47	A	0.478	G	0.521
3061721	rs2425599	47	A	0.446	C	0.553
3066559	rs2425608	42	A	0.345	G	0.654
3075264	G6801045	46	A	0.684	G	0.315
3085102	rs1572925	39	C	0.705	G	0.294
3102651	G6828432	44	A	0.625	G	0.375
3133994	rs2425614	41	C	0.829	T	0.17
3141297	rs926288	27	C	0.87	T	0.129
3143846	rs2235206	46	A	0.793	G	0.206
3157402	rs2205773	46	A	0.152	G	0.847
3166458	G6892239	42	A	0.88	G	0.119
3216500	rs760630	39	A	0.064	T	0.935
3241997	G6967778	40	A	0.85	G	0.15
3258734	rs926294	45	C	0.655	T	0.344
3274561	rs926291	32	A	0.546	G	0.453
3284053	G7009834	41	A	0.353	G	0.646
3293069	G7018850	41	C	0.097	T	0.902
3300323	G7026104	41	A	0.719	G	0.28
3309831	rs909892	47	C	0.904	T	0.095
3338938	rs1883544	47	C	0.606	T	0.393
3345192	rs1569623	45	A	0.6	G	0.4
3363024	rs1997749	46	C	0.51	T	0.489
3386746	rs941441	39	C	0.448	G	0.551
3449593	rs765148	47	A	0.585	G	0.414
3486339	rs2285186	47	A	0.074	G	0.925
3488095	rs2071969	46	C	0.902	T	0.097
3490901	rs2071968	47	C	0.648	G	0.351
3494369	rs2269625	39	A	0.769	G	0.23
3505689	rs2269622	47	C	0.531	G	0.468
3506578	rs3205	47	C	0.095	T	0.904
3513356	rs763228	47	C	0.744	G	0.255
3523683	rs2067061	42	A	0.666	C	0.333
3532194	rs763227	38	A	0.789	G	0.21
3543780	rs1055334	38	A	0.184	G	0.815
3552247	rs2664519	19	A	0.684	G	0.315
3586769	rs714998	44	C	0.818	T	0.181
3591859	rs2273523	32	C	0.828	T	0.171

Position	SNP	DNAs tested	A1	AF1	A2	AF2
3602632	rs393115	27	A	0.907	C	0.092
3606275	rs285191	34	A	0.102	G	0.897
3608967	rs285194	44	A	0.159	G	0.84
3618700	rs396373	47	A	0.627	G	0.372
3619094	rs385345	47	C	0.202	T	0.797
3623777	rs285205	38	A	0.118	C	0.881
3637642	rs442143	46	C	0.902	T	0.097
3638988	rs387769	46	A	0.097	G	0.902
3655772	rs285162	47	C	0.946	T	0.053
3657097	rs285167	47	A	0.063	G	0.936
3658590	rs2070235	37	A	0.905	G	0.094
3665008	rs285171	47	C	0.127	G	0.872
3667817	rs285172	47	A	0.872	G	0.127
3686616	rs371484	47	A	0.925	G	0.074
3690239	rs285197	47	C	0.808	T	0.191
3714831	rs244071	47	G	0.17	T	0.829
3739421	G7465202	46	C	0.869	T	0.13
3758171	rs244066	38	A	0.184	G	0.815
3780143	rs2235765	39	A	0.833	G	0.166
3798617	G7524398	39	A	0.128	C	0.871
3848704	rs1569697	40	C	0.437	T	0.562
3878621	rs1569698	41	C	0.268	T	0.731
3882692	rs1076723	47	C	0.159	T	0.84
3890704	rs916474	44	A	0.136	G	0.863
3893142	rs1888982	40	A	0.15	G	0.85
3904495	rs761919	47	A	0.585	G	0.414
3925293	rs736992	47	C	0.84	G	0.159
3933353	G7659134	44	G	0.829	T	0.17
3954237	rs1980578	21	C	0.119	T	0.88
3956870	rs1547002	45	A	0.666	G	0.333
3973965	G7699746	45	C	0.311	G	0.688
3987419	rs2179593	46	A	0.728	C	0.271
3995666	rs1040545	41	C	0.243	T	0.756
3996896	rs1555118	47	C	0.265	T	0.734
4006324	G7732105	43	A	0.255	G	0.744
4020783	rs2143495	46	A	0.25	G	0.75
4025984	rs1555124	46	C	0.673	G	0.326
4037628	rs1883682	46	C	0.326	T	0.673
4047232	rs932415	45	A	0.8	C	0.2
4054623	rs1883684	47	A	0.468	G	0.531
4070587	G7796368	41	A	0.658	G	0.341
4078168	rs2038168	46	C	0.75	T	0.25
4090850	rs969570	17	A	0.235	G	0.764
4102809	rs761205	46	A	0.282	G	0.717
4110645	rs2057029	47	A	0.691	G	0.308
4110678	rs2143494	31	A	0.516	G	0.483
4119262	rs2425630	44	A	0.261	G	0.738
4135042	rs738498	44	A	0.465	G	0.534

Position	SNP	DNAs tested	A1	AF1	A2	AF2
4142323	rs1883790	28	A	0.892	G	0.107
4152620	rs9875	46	C	0.369	T	0.63
4157331	rs1474753	43	C	0.127	T	0.872
4162890	rs2232286	46	C	0.097	T	0.902
4164259	rs956610	46	A	0.119	G	0.88
4164490	rs1007125	47	A	0.127	G	0.872
4166537	rs731498	45	A	0.822	G	0.177
4166594	rs731499	40	C	0.737	T	0.262
4184909	G7910690	42	C	0.119	T	0.88
4195279	rs2867799	40	C	0.875	T	0.125
4209342	G7935123	43	C	0.058	T	0.941
4218942	rs2293	47	C	0.787	G	0.212
4227781	G7953562	43	G	0.802	T	0.197
4244900	rs2868090	38	A	0.815	T	0.184
4266559	G7992340	44	G	0.806	T	0.193
4280248	G8006029	36	C	0.875	T	0.125
4293158	rs1884612	38	C	0.289	T	0.71
4312850	rs2144908	40	A	0.187	G	0.812
4350974	rs2425635	33	A	0.5	G	0.5
4352917	rs717247	42	A	0.678	G	0.321
4353762	rs2425638	41	C	0.109	T	0.89
4356418	rs1800963	47	A	0.319	C	0.68
4361826	rs745975	44	A	0.284	G	0.715
4366173	rs1885088	33	A	0.227	G	0.772
4366396	rs1885089	47	C	0.765	T	0.234
4378114	rs1028584	28	A	0.303	C	0.696
4450161	rs1983	44	A	0.59	G	0.409
4455136	rs6606	37	C	0.918	T	0.081
4489604	G8215385	46	A	0.195	G	0.804
4496763	rs244128	47	C	0.574	G	0.425
4499257	rs244127	46	G	0.119	T	0.88
4507748	rs2425648	40	G	0.787	T	0.212
4511112	rs244125	38	A	0.802	C	0.197
4516411	rs244123	15	C	0.166	T	0.833
4522838	rs244120	29	A	0.051	G	0.948
4543469	rs244107	40	A	0.812	C	0.187
4543614	rs244106	47	C	0.18	G	0.819
4552599	rs244099	44	A	0.84	T	0.159
4570678	rs1061662	45	C	0.844	T	0.155
4572324	rs244082	28	C	0.16	T	0.839
4574528	rs2664557	32	C	0.906	T	0.093
4577724	rs244079	45	C	0.055	T	0.944
4578900	rs929089	42	C	0.892	T	0.107
4587792	rs395209	47	A	0.936	T	0.063
4590169	rs371927	47	C	0.17	T	0.829
4594398	rs446125	45	A	0.222	C	0.777
4598525	rs406383	43	C	0.639	G	0.36
4602134	rs2299686	46	A	0.521	G	0.478

Position	SNP	DNAs tested	A1	AF1	A2	AF2
4618353	rs2903744	47	C	0.574	G	0.425
4639530	rs912914	36	C	0.875	T	0.125
4640727	rs912921	47	A	0.67	C	0.329
4656140	rs2148041	44	C	0.568	T	0.431
4673118	G8398899	30	A	0.116	G	0.883
4697239	rs1555299	47	C	0.351	T	0.648
4723347	rs1078342	36	A	0.625	G	0.375
4740331	rs1080026	15	A	0.8	C	0.2
4755456	rs1117080	39	C	0.666	G	0.333
4757623	rs1555300	42	A	0.404	G	0.595
4763024	rs1123402	21	C	0.523	G	0.476
4772599	G8498380	43	C	0.558	G	0.441
4782878	rs2903760	47	C	0.574	T	0.425
4822621	rs2903761	47	C	0.68	G	0.319
4843289	rs2239535	45	A	0.188	G	0.811
4844334	rs2284266	45	A	0.788	G	0.211
4850897	rs2253712	47	C	0.308	T	0.691
4853270	rs2425672	47	A	0.478	G	0.521
4857367	rs4931	45	A	0.788	C	0.211
4863588	rs8356	26	C	0.865	T	0.134
4863706	rs2664577	25	C	0.84	T	0.16
4868316	rs2235172	39	C	0.217	G	0.782
4874810	rs2075960	47	G	0.234	T	0.765
4886042	rs1998033	47	A	0.436	G	0.563
4888666	rs2664567	45	C	0.866	T	0.133
4892168	rs2273358	42	A	0.107	G	0.892
4893920	rs11780	33	C	0.409	T	0.59
4899350	rs2234209	26	A	0.73	G	0.269
4900686	rs2284267	31	A	0.709	C	0.29
4907251	rs1079900	36	C	0.402	T	0.597
4916174	rs2234197	42	C	0.785	T	0.214
4925838	rs927000	30	A	0.15	G	0.85
4929054	rs910671	25	A	0.42	C	0.58
4936111	rs2267857	45	C	0.477	G	0.522
4949236	rs2903772	33	G	0.454	T	0.545
4958360	rs1015520	33	C	0.09	T	0.909
4958569	rs1015519	47	C	0.521	T	0.478
4971164	rs2284272	39	C	0.153	T	0.846
4973303	rs2299975	45	C	0.522	T	0.477
4976059	rs2010195	34	C	0.147	T	0.852
5004917	rs2284274	41	A	0.512	T	0.487
5009084	rs2284275	44	C	0.522	T	0.477
5012643	rs1894572	44	C	0.465	T	0.534
5025305	rs2267862	47	G	0.468	T	0.531
5042321	rs1003855	46	C	0.456	T	0.543
5047118	rs1540310	47	C	0.531	G	0.468
5050760	rs734784	46	A	0.554	G	0.445
5063666	rs916311	47	G	0.617	T	0.382

Position	SNP	DNAs tested	A1	AF1	A2	AF2
5067623	rs2157360	47	C	0.414	G	0.585
5081551	G8807332	42	C	0.19	T	0.809
5108041	rs2207199	37	C	0.256	T	0.743
5131655	rs2664581	27	A	0.888	C	0.111
5138094	rs1997892	47	A	0.18	T	0.819
5143726	rs1007137	43	C	0.232	T	0.767
5160265	G8886046	39	C	0.82	T	0.179
5209139	rs2206888	32	A	0.812	G	0.187
5225272	rs991049	40	A	0.775	G	0.225
5225360	rs991048	47	A	0.808	G	0.191
5269809	rs2076026	47	A	0.946	G	0.053
5273901	rs736389	47	C	0.17	T	0.829
5284643	rs985586	40	C	0.725	T	0.275
5286005	rs2070638	47	C	0.51	T	0.489
5291421	rs2072792	37	A	0.5	G	0.5
5293036	rs2267867	44	A	0.784	G	0.215
5302262	rs2741450	37	A	0.5	T	0.5
5302584	rs1981431	44	A	0.477	C	0.522
5303940	rs2078423	41	A	0.402	G	0.597
5323013	rs2245717	31	G	0.854	T	0.145
5323145	rs2664543	42	G	0.297	T	0.702
5323265	rs2664583	45	C	0.288	T	0.711
5323599	rs707576	43	C	0.267	T	0.732
5324373	rs2248637	46	A	0.152	G	0.847
5365707	rs2243553	46	A	0.13	G	0.869
5372703	rs1028307	39	A	0.192	C	0.807
5380125	rs13217	47	A	0.489	G	0.51
5381482	rs707577	34	C	0.764	T	0.235
5392942	rs2741553	46	A	0.467	G	0.532
5401683	rs2745065	37	C	0.067	T	0.932
5412593	rs1016496	36	C	0.152	T	0.847
5421508	rs973446	47	C	0.925	G	0.074
5452896	rs2868301	45	C	0.633	T	0.366
5497845	rs11594	43	A	0.79	C	0.209
5505718	rs2050095	31	C	0.306	T	0.693
5508856	rs2294559	47	A	0.67	G	0.329
5511508	rs2250860	46	C	0.336	T	0.663
5515056	rs1569612	43	C	0.313	T	0.686
5518157	rs909879	40	A	0.537	G	0.462
5518901	rs1977096	47	A	0.329	C	0.67
5552111	rs2425707	47	A	0.404	G	0.595
5554454	rs2425709	46	A	0.347	G	0.652
5564120	rs1487320	46	C	0.543	G	0.456
5570975	rs2072973	46	A	0.543	G	0.456
5585876	rs1157672	20	C	0.25	T	0.75
5587812	rs1825776	44	C	0.613	T	0.386
5592051	rs1005456	42	A	0.63	G	0.369
5622944	rs1487318	43	A	0.616	G	0.383

Position	SNP	DNAs tested	A1	AF1	A2	AF2
5640124	rs1825775	44	C	0.67	T	0.329
5641742	rs232729	34	A	0.323	G	0.676
5657724	rs714595	46	A	0.206	G	0.793
5659431	rs432448	15	A	0.7	C	0.3
5661527	rs232291	37	G	0.472	T	0.527
5679873	rs1487317	40	C	0.6	T	0.4
5681520	rs232259	40	G	0.4	T	0.6
5682313	rs761810	45	A	0.611	G	0.388
5705703	rs380421	39	A	0.423	G	0.576
5707597	rs454874	40	A	0.325	G	0.675
5729995	rs2664529	45	C	0.366	T	0.633
5745597	rs411945	35	C	0.271	G	0.728
5747763	rs2664539	30	A	0.716	G	0.283
5765425	rs399672	46	C	0.315	T	0.684
5766100	rs390386	38	A	0.605	C	0.394
5773195	rs445503	25	A	0.6	C	0.4
5779823	rs369138	44	A	0.522	C	0.477
5784475	rs383112	41	A	0.487	G	0.512
5802539	rs174745	46	C	0.543	T	0.456
5814890	rs722669	44	C	0.5	G	0.5
5833099	rs2903808	46	C	0.532	T	0.467
5844486	rs1516580	42	A	0.523	G	0.476
5849131	rs742034	45	A	0.477	G	0.522
5857964	rs553359	26	A	0.442	C	0.557
5866618	rs441346	41	C	0.414	G	0.585
5867304	rs394643	28	C	0.392	T	0.607
5890133	rs1057208	38	C	0.828	T	0.171
5904440	rs3363	47	C	0.851	T	0.148
5919490	G9645271	44	C	0.863	T	0.136
5951093	rs1888235	43	C	0.825	T	0.174
5959493	G9685274	43	A	0.174	G	0.825
5967351	rs2664538	33	A	0.651	G	0.348
5972136	rs20544	32	C	0.546	T	0.453
5985422	rs2868364	18	A	0.388	G	0.611
6015125	rs1128536	34	A	0.294	G	0.705
6018476	rs1537028	42	G	0.297	T	0.702
6034490	rs1406826	36	A	0.486	G	0.513
6041530	rs1950174	47	A	0.148	C	0.851
6044872	rs2206892	43	C	0.883	T	0.116
6053363	rs1569722	43	C	0.674	T	0.325
6059872	rs1358719	35	A	0.671	G	0.328
6069190	rs1569723	28	A	0.946	C	0.053
6073529	rs1800686	47	A	0.276	G	0.723
6073864	rs752118	37	C	0.716	T	0.283
6075225	rs1535045	43	C	0.686	T	0.313
6086292	rs2143699	47	A	0.085	G	0.914
6094850	rs1535043	38	A	0.407	T	0.592
6124047	rs2425760	38	A	0.118	T	0.881

Position	SNP	DNAs tested	A1	AF1	A2	AF2
6132564	G9858345	43	G	0.662	T	0.337
6151750	rs2868767	38	A	0.828	G	0.171
6166448	rs2425785	44	C	0.386	G	0.613
6181719	rs2092384	36	C	0.944	G	0.055
6201464	G9927245	45	A	0.2	G	0.8
6214726	rs2425809	47	C	0.478	T	0.521
6219351	rs2425811	45	G	0.477	T	0.522
6225016	rs2425825	47	A	0.51	G	0.489
6239080	rs2425856	47	A	0.446	G	0.553
6248983	rs2425859	19	C	0.684	T	0.315
6255855	rs2425861	44	G	0.363	T	0.636
6274892	rs2425863	47	C	0.627	T	0.372
6276873	rs2180911	41	C	0.365	T	0.634
6277977	rs2425866	44	C	0.363	T	0.636
6280775	rs1998253	36	A	0.319	T	0.68
6293784	rs2869219	39	C	0.294	T	0.705
6323431	rs2297057	18	C	0.833	T	0.166
6335855	rs2260959	44	A	0.943	G	0.056
6403741	rs391117	44	C	0.261	T	0.738
6411300	rs593048	42	C	0.202	T	0.797
6423036	rs678086	38	A	0.21	C	0.789
6438755	rs457918	46	C	0.13	T	0.869
6446155	rs438687	44	A	0.897	T	0.102
6450547	rs464406	46	C	0.195	T	0.804
6452419	rs460067	40	C	0.837	G	0.162
6461292	rs1880899	39	C	0.205	T	0.794
6463570	rs2780231	37	C	0.459	T	0.54
6478399	rs847104	44	A	0.795	G	0.204
6480329	rs1527139	38	A	0.776	G	0.223
6485576	rs202370	33	C	0.393	G	0.606
6492984	rs202380	47	A	0.18	G	0.819
6496931	rs2411	43	A	0.825	G	0.174
6514631	rs10218	45	A	0.311	C	0.688
6517500	rs847062	42	C	0.821	T	0.178
6518827	rs847063	41	C	0.402	T	0.597
6520115	rs847065	44	A	0.625	G	0.375
6522030	rs202391	42	C	0.809	G	0.19
6524341	rs847069	45	C	0.411	T	0.588
6525657	rs436978	42	A	0.214	G	0.785
6531392	rs1880898	40	C	0.425	T	0.575
6532608	rs413635	21	A	0.761	G	0.238
6537213	rs202388	43	C	0.43	T	0.569
6539581	rs389905	47	G	0.414	T	0.585
6539746	rs863675	46	C	0.054	T	0.945
6549477	rs85024	38	C	0.828	T	0.171
6550764	rs431072	46	C	0.521	T	0.478
6565338	rs383551	30	G	0.716	T	0.283
6566882	rs2141113	20	C	0.25	G	0.75

Position	SNP	DNAs tested	A1	AF1	A2	AF2
6569062	rs847084	37	A	0.351	T	0.648
6569395	rs2273024	47	C	0.297	G	0.702
6571758	rs847100	13	A	0.538	G	0.461
6583050	rs847071	39	C	0.358	T	0.641
6584310	rs761215	44	C	0.534	T	0.465
6588167	rs941206	46	G	0.108	T	0.891
6605295	rs1004571	37	A	0.31	G	0.689
6617100	rs1204641	47	A	0.063	G	0.936
6627222	rs1011074	46	A	0.108	G	0.891
6645195	rs2664574	25	C	0.92	G	0.08
6677640	rs2425901	35	A	0.285	C	0.714
6677706	rs2425902	45	A	0.733	G	0.266
6684002	rs2425904	26	C	0.692	T	0.307
6687345	rs2425911	41	C	0.743	G	0.256
6690892	rs707507	41	C	0.365	T	0.634
6698210	rs2179357	37	C	0.351	T	0.648
6705287	G10431068	43	C	0.453	T	0.546
6720442	rs760874	46	A	0.13	G	0.869
6721346	rs976938	45	A	0.633	G	0.366
6724642	rs1997711	43	A	0.418	G	0.581
6729945	rs760877	36	A	0.166	C	0.833
6747631	G10473412	41	C	0.5	T	0.5
6755198	G10480979	45	G	0.488	T	0.511
6763119	G10488900	38	C	0.407	T	0.592
6783303	rs928486	34	C	0.72	T	0.279
6806280	rs2026509	42	C	0.202	T	0.797
6812873	rs947078	46	C	0.358	G	0.641
6832032	rs2903948	36	A	0.5	G	0.5
6840561	rs1984076	46	A	0.423	G	0.576
6853247	rs1007330	47	C	0.627	T	0.372
6856697	rs2236519	43	A	0.29	G	0.709
6860474	rs963978	37	G	0.581	T	0.418
6864515	rs914829	42	A	0.642	T	0.357
6899982	rs1206745	46	C	0.521	G	0.478
6923504	G10649285	43	A	0.686	G	0.313
6948585	rs1889143	45	A	0.355	G	0.644
6971209	rs1572867	22	A	0.5	G	0.5
6973010	rs2903942	43	A	0.604	G	0.395
6973748	rs1340774	46	C	0.597	G	0.402
6982972	rs1340775	43	A	0.174	G	0.825
6985572	rs947011	20	C	0.1	T	0.9
7001373	rs947019	47	G	0.734	T	0.265
7015566	rs1206808	47	A	0.212	G	0.787
7016301	rs878636	42	C	0.321	T	0.678
7023003	rs1890990	41	C	0.817	G	0.182
7058335	rs1890987	47	C	0.436	G	0.563
7076565	rs2066226	47	C	0.606	G	0.393
7109050	rs1105402	47	C	0.563	T	0.436

Position	SNP	DNAs tested	A1	AF1	A2	AF2
7120657	rs1008031	31	A	0.887	G	0.112
7136302	rs2073172	45	A	0.488	G	0.511
7147606	rs2057085	46	C	0.489	T	0.51
7176890	rs2235908	46	A	0.434	G	0.565
7180163	rs2664544	20	A	0.05	G	0.95
7184703	G10910484	39	C	0.166	T	0.833
7193602	rs2664579	46	A	0.086	C	0.913
7194978	rs2275801	39	A	0.91	G	0.089
7205500	rs2298195	47	A	0.585	G	0.414
7214820	rs2281208	30	A	0.233	G	0.766
7218315	rs2294560	16	C	0.437	T	0.562
7221085	rs761021	41	A	0.073	G	0.926
7238247	rs2235588	46	C	0.282	T	0.717
7250509	rs2076402	47	A	0.563	C	0.436
7273313	rs1013715	34	C	0.073	T	0.926
7284192	G11009973	44	C	0.909	T	0.09
7328804	rs2038376	46	C	0.804	T	0.195
7340303	rs2868824	43	C	0.825	T	0.174
7388225	rs2903925	38	A	0.394	G	0.605
7407650	rs1210832	46	C	0.88	G	0.119
7435267	G11161048	44	A	0.227	G	0.772
7449410	rs2026401	35	C	0.1	T	0.9
7458924	rs1537306	45	C	0.122	G	0.877
7472403	rs1206883	34	A	0.882	G	0.117
7484072	rs2425940	19	A	0.157	G	0.842
7486745	rs1212595	33	A	0.878	T	0.121
7496045	rs2425954	47	A	0.106	C	0.893
7511236	G11237017	43	C	0.883	G	0.116
7534591	rs2425974	39	C	0.153	T	0.846
7555793	rs2143491	39	A	0.41	G	0.589
7572726	G11298507	42	G	0.119	T	0.88
7583326	rs2273022	27	C	0.888	G	0.111
7593810	rs720500	16	A	0.125	G	0.875
7597027	rs2780389	31	A	0.903	G	0.096
7598724	rs627601	34	A	0.132	G	0.867
7601215	rs396221	41	A	0.451	C	0.548
7601311	rs427967	35	A	0.185	G	0.814
7603626	rs642417	14	A	0.785	G	0.214
7605530	rs434258	39	C	0.576	T	0.423
7618608	rs450110	47	C	0.936	T	0.063
7623406	rs425433	45	A	0.177	C	0.822
7632757	rs1577063	31	C	0.08	G	0.919
7638295	rs1591171	47	C	0.787	T	0.212
7650265	rs1889175	35	A	0.757	C	0.242
7660693	rs1610375	47	C	0.5	G	0.5
7668100	G11393881	44	A	0.875	G	0.125
7680319	rs2150809	47	A	0.382	G	0.617
7691843	rs1889170	33	C	0.681	T	0.318

Position	SNP	DNAs tested	A1	AF1	A2	AF2
7698794	rs1970137	32	A	0.234	T	0.765
7707362	rs914870	43	C	0.453	T	0.546
7730054	rs2840278	47	A	0.744	G	0.255
7742097	rs928495	47	C	0.882	T	0.117
7750421	G11476202	42	A	0.857	G	0.142
7761606	rs1889178	47	C	0.893	T	0.106
7772734	rs2869224	37	C	0.945	T	0.054
7815713	rs2145134	41	A	0.463	C	0.536
7831818	G11557599	33	A	0.909	G	0.09
7840964	rs981303	47	A	0.914	G	0.085
7843842	rs2426003	39	A	0.769	C	0.23
7868233	G11594014	44	A	0.079	G	0.92
7907760	rs2664588	37	C	0.527	T	0.472
7908242	rs2869315	47	C	0.542	G	0.457
7918857	rs2869318	46	A	0.706	G	0.293
7984412	G11710193	44	A	0.511	G	0.488
8020712	rs2869343	44	C	0.863	G	0.136
8067196	rs1325760	34	C	0.279	T	0.72
8086429	rs645580	30	A	0.516	G	0.483
8093304	rs162084	37	A	0.918	G	0.081
8102275	rs911953	35	C	0.757	T	0.242
8107481	rs162083	40	A	0.262	G	0.737
8109045	rs227879	26	C	0.865	G	0.134
8112096	rs162077	38	C	0.815	T	0.184
8114342	rs827949	45	C	0.433	T	0.566
8116481	rs827942	37	C	0.216	T	0.783
8121186	rs382820	46	A	0.445	G	0.554
8123079	rs730944	43	C	0.802	T	0.197
8131959	rs676815	32	A	0.078	C	0.921
8133019	rs2034811	47	C	0.585	T	0.414
8136606	rs1047605	47	C	0.17	T	0.829
8140295	rs852328	38	A	0.25	G	0.75
8155286	rs852350	38	A	0.868	C	0.131
8160780	rs852353	32	A	0.453	C	0.546
8165169	rs852363	41	C	0.67	G	0.329
8170016	rs852290	46	C	0.619	G	0.38
8180850	rs2426060	44	A	0.34	G	0.659
8184275	rs226831	39	A	0.307	G	0.692
8189258	rs610494	43	A	0.674	T	0.325
8205848	rs170536	39	C	0.346	T	0.653
8212848	rs2021812	35	A	0.9	G	0.1
8216385	rs226811	46	G	0.239	T	0.76
8219174	rs226808	34	C	0.705	T	0.294
8221545	rs226820	35	C	0.371	T	0.628
8228572	G11954353	46	C	0.76	T	0.239
8257798	rs430071	39	A	0.448	G	0.551
8279379	rs442855	43	G	0.546	T	0.453
8282424	rs414644	36	A	0.138	G	0.861

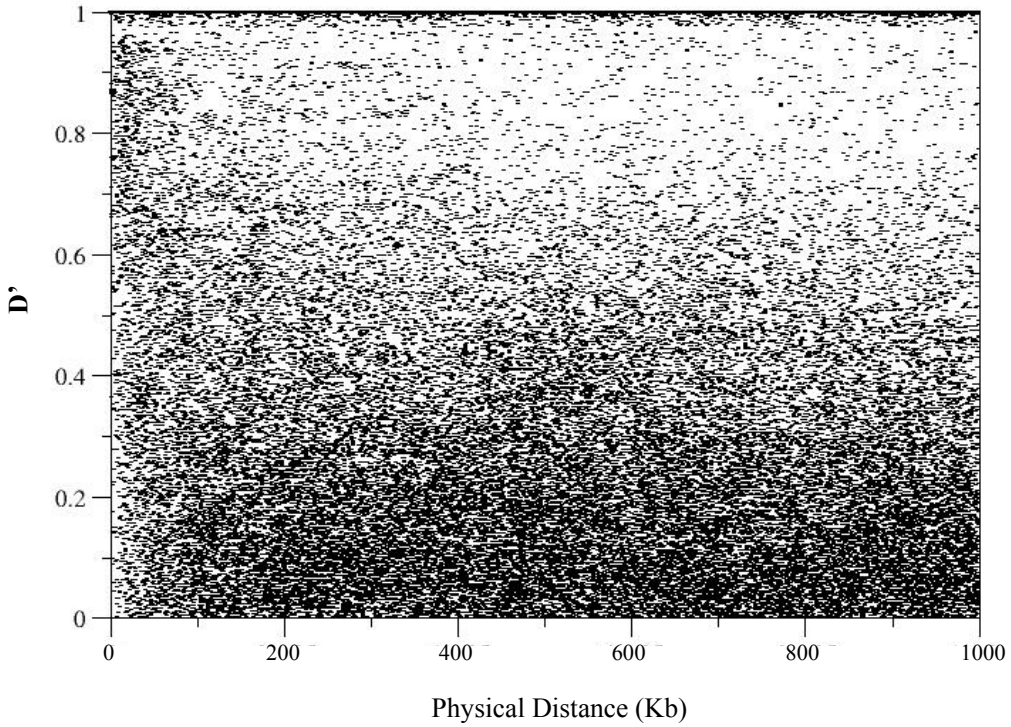
Position	SNP	DNAs tested	A1	AF1	A2	AF2
8286584	rs2426064	17	A	0.852	G	0.147
8294793	rs435954	40	A	0.1	C	0.9
8297380	rs911106	40	G	0.637	T	0.362
8300078	rs410426	40	A	0.587	G	0.412
8359190	rs2869368	41	C	0.731	T	0.268
8366488	rs737329	43	G	0.848	T	0.151
8382395	rs999151	38	C	0.736	T	0.263
8385803	rs993425	46	C	0.347	G	0.652
8393387	rs1321006	39	A	0.333	G	0.666
8406083	rs926693	39	C	0.435	T	0.564
8409400	rs926692	47	G	0.468	T	0.531
8415858	rs910191	45	C	0.311	T	0.688
8439073	rs2869385	40	C	0.225	T	0.775
8444302	rs2010276	17	C	0.852	T	0.147
8468407	rs1546923	39	A	0.448	G	0.551
8487158	rs2904081	47	C	0.755	G	0.244
8495249	G12221030	45	A	0.222	G	0.777
8502800	rs1358721	47	A	0.308	G	0.691
8519582	G12245363	39	A	0.333	G	0.666
8527917	G12253698	43	C	0.558	T	0.441
8568484	rs2664570	38	C	0.855	T	0.144
8576581	G12302362	43	G	0.104	T	0.895
8580276	rs2664521	28	C	0.946	T	0.053
8583605	G12309386	43	C	0.244	T	0.755
8593044	rs2281287	37	C	0.945	T	0.054
8603507	rs2294910	35	A	0.885	G	0.114
8623140	rs2904167	43	C	0.093	T	0.906
8631366	rs729664	44	A	0.147	G	0.852
8641571	rs1040559	40	A	0.575	G	0.425
8647337	rs2143563	37	A	0.513	G	0.486
8650015	rs735084	42	A	0.738	G	0.261
8650836	rs736659	41	C	0.817	T	0.182
8659966	rs2206742	40	C	0.837	T	0.162
8662862	rs926629	21	C	0.214	T	0.785
8666868	rs968478	42	A	0.654	G	0.345
8687541	rs742644	43	C	0.825	T	0.174
8691206	rs2073071	47	G	0.5	T	0.5
8701921	rs1883745	46	C	0.532	T	0.467
8706937	rs2426087	47	A	0.053	G	0.946
8711246	rs2426091	39	A	0.474	G	0.525
8714825	rs998915	40	G	0.362	T	0.637
8720892	rs749339	46	A	0.402	G	0.597
8727430	rs2224504	47	C	0.904	T	0.095
8741429	rs747786	36	C	0.277	T	0.722
8743024	rs750085	38	A	0.802	G	0.197
8762182	rs752420	33	A	0.303	G	0.696
8763650	rs1885289	35	A	0.114	G	0.885
8787153	G12512934	42	A	0.238	G	0.761

Position	SNP	DNAs tested	A1	AF1	A2	AF2
8845178	rs2024596	42	C	0.476	T	0.523
8865459	rs2273101	15	C	0.4	T	0.6
8881518	G12607299	42	A	0.345	G	0.654
8892905	rs1883881	43	A	0.313	C	0.686
8907473	rs2295031	47	A	0.648	G	0.351
8928635	rs1569749	46	A	0.641	G	0.358
8938022	rs2295033	41	C	0.646	T	0.353
8953862	rs2295579	39	A	0.628	T	0.371
8957575	rs2281582	47	C	0.84	T	0.159
8979380	rs707533	45	A	0.644	G	0.355
8979464	rs707534	46	A	0.652	G	0.347
8981690	rs927160	47	C	0.638	G	0.361
8989753	rs2426109	38	C	0.671	G	0.328
9010019	rs2246266	41	C	0.731	T	0.268
9018337	rs2075676	35	A	0.357	C	0.642
9028108	rs2426125	44	A	0.659	G	0.34
9032362	rs2426127	27	C	0.814	T	0.185
9040189	rs17632	36	C	0.652	T	0.347
9043577	rs1556876	35	A	0.657	C	0.342
9048083	rs755587	47	C	0.361	T	0.638
9050253	rs2426132	37	C	0.513	G	0.486
9061435	rs873689	18	C	0.944	T	0.055
9065470	rs168345	42	C	0.19	T	0.809
9075836	rs348298	30	A	0.066	G	0.933
9090860	rs348279	43	A	0.267	G	0.732
9097882	rs2273653	37	A	0.675	C	0.324
9103047	rs348267	44	C	0.647	T	0.352
9128661	rs348284	31	C	0.161	T	0.838
9133460	rs348293	44	C	0.715	T	0.284
9147215	rs1318950	45	A	0.2	G	0.8
9166717	rs2273145	28	C	0.767	T	0.232
9175163	rs238171	43	C	0.197	T	0.802
9175841	rs761499	27	G	0.796	T	0.203
9177308	rs238148	40	C	0.212	T	0.787
9179045	rs238150	46	A	0.684	G	0.315
9186048	rs238203	47	A	0.223	C	0.776
9186141	rs238204	46	A	0.206	G	0.793
9189576	rs7689	28	C	0.892	T	0.107
9192635	rs238209	40	A	0.75	G	0.25
9198565	rs238217	33	A	0.712	T	0.287
9199503	rs238221	31	C	0.096	G	0.903
9201131	rs3021	47	A	0.68	G	0.319
9206235	rs238174	46	C	0.684	T	0.315
9210241	rs238180	24	A	0.145	T	0.854
9216730	rs238186	40	G	0.187	T	0.812
9221384	rs238192	46	A	0.228	G	0.771
9222323	rs479474	47	A	0.797	G	0.202
9232634	rs8197	39	A	0.743	G	0.256

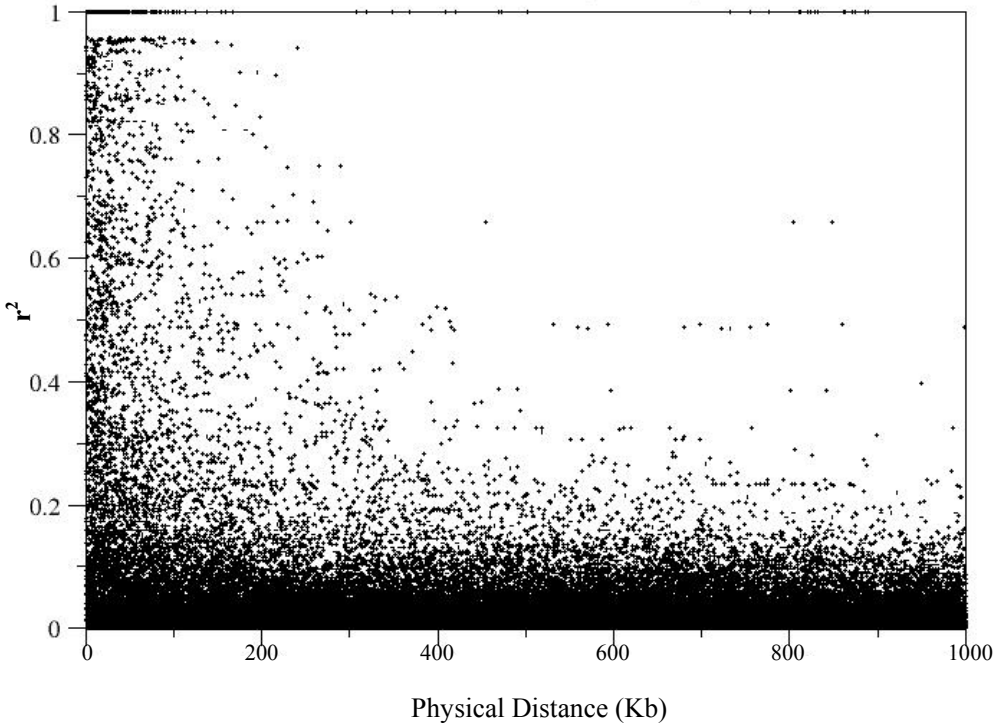
Position	SNP	DNAs tested	A1	AF1	A2	AF2
9246798	rs237720	47	C	0.797	G	0.202
9297507	rs1393948	44	A	0.147	G	0.852
9308102	rs237449	47	A	0.914	G	0.085
9316031	rs1051295	38	A	0.657	G	0.342
9338134	rs756529	27	A	0.481	G	0.518
9352518	rs237452	44	C	0.431	T	0.568
9360297	rs237461	46	C	0.26	T	0.739
9362723	rs119416	45	A	0.255	G	0.744
9365238	rs237468	36	C	0.166	T	0.833
9370936	rs237473	43	A	0.755	C	0.244
9376632	rs237475	47	C	0.425	T	0.574
9378508	rs237476	47	G	0.585	T	0.414
9403740	rs572845	46	C	0.673	T	0.326
9405328	rs610412	46	A	0.63	C	0.369
9419580	rs477135	47	A	0.351	T	0.648
9424690	rs553213	43	C	0.534	T	0.465
9430480	rs554985	41	A	0.926	T	0.073
9436803	rs496511	44	A	0.59	G	0.409
9446556	rs491025	38	A	0.539	G	0.46
9454302	rs729824	31	C	0.112	T	0.887
9456832	rs5629	39	A	0.153	C	0.846
9466097	rs508757	42	C	0.214	T	0.785
9467808	rs5628	47	C	0.946	T	0.053
9504964	rs498646	36	C	0.875	G	0.125
9510256	rs493694	34	A	0.352	G	0.647
9515205	rs693649	47	A	0.18	G	0.819
9524088	rs538748	45	A	0.288	G	0.711
9535505	rs507120	47	C	0.819	G	0.18
9556062	rs235013	41	C	0.597	G	0.402
9559489	rs408618	47	C	0.436	T	0.563
9566770	rs235039	44	A	0.636	G	0.363
9571200	rs235025	45	A	0.622	T	0.377
9575378	rs235030	27	C	0.462	T	0.537
9580040	rs235035	40	C	0.937	T	0.062
9584275	rs421801	39	C	0.166	T	0.833
9586160	rs2235855	40	C	0.662	T	0.337
9593917	rs803194	42	A	0.833	G	0.166
9600585	rs2273088	36	A	0.43	G	0.569
9621222	rs2057075	33	A	0.621	T	0.378
9629778	rs1016234	47	C	0.18	T	0.819
9636933	G13362714	42	C	0.369	T	0.63
9645140	rs913476	45	C	0.388	T	0.611
9658646	rs951497	46	A	0.184	G	0.815
9694319	rs590397	47	A	0.138	G	0.861
9710810	rs651195	45	A	0.611	G	0.388
9714843	rs679376	47	C	0.755	T	0.244
9729705	rs588375	47	A	0.223	G	0.776
9737529	rs678634	43	A	0.162	G	0.837
9747259	rs605138	43	C	0.872	T	0.127
9754474	G13480255	45	C	0.744	T	0.255
9762534	rs719219	19	A	0.184	G	0.815
9767149	rs1328454	43	A	0.558	G	0.441
9777066	rs2182963	46	A	0.195	G	0.804
9794842	rs530461	41	C	0.585	T	0.414
9801945	rs595777	46	C	0.576	G	0.423
9805780	rs645182	30	A	0.116	G	0.883
9807380	rs627898	47	G	0.425	T	0.574
9809295	rs517501	43	C	0.883	G	0.116
9814983	rs1033474	38	C	0.71	T	0.289
9820417	rs633829	40	A	0.437	G	0.562
9828477	rs536092	27	A	0.351	G	0.648
9832592	rs1062840	46	C	0.934	T	0.065
9836642	rs479007	33	C	0.863	G	0.136
9847736	rs632376	45	A	0.6	G	0.4
9859218	rs1008687	45	G	0.788	T	0.211
9881796	rs2038127	47	A	0.691	G	0.308
9889019	G13614800	44	A	0.136	G	0.863
9901242	rs1883553	30	C	0.866	T	0.133
9922778	G13648559	40	A	0.875	G	0.125
9942829	G13668610	43	C	0.906	T	0.093
9964783	rs2869940	39	C	0.256	T	0.743
9970776	rs1555317	44	C	0.352	T	0.647
9976053	rs1973945	38	C	0.381	T	0.618
9996796	rs926602	44	C	0.125	T	0.875
10017355	rs1007580	47	A	0.595	G	0.404
10024917	rs1049871	38	A	0.118	T	0.881
10026922	rs2073053	14	G	0.892	T	0.107
10033521	rs2269217	47	C	0.265	G	0.734
10050075	rs2869956	41	A	0.573	T	0.426
10052424	rs2269214	40	A	0.325	G	0.675
10057157	rs383495	45	C	0.688	T	0.311
10060293	rs367033	41	A	0.926	C	0.073
10060929	rs1928545	46	C	0.728	T	0.271
10077265	rs761213	47	C	0.446	G	0.553
10091159	rs232737	45	C	0.544	G	0.455

Appendix 16: (A,B) Variability of D' (A) and r^2 (B) using the pairwise values for all polymorphic SNPs separated by □ 1 Mb. (C,D) The corresponding plots for LD decay. (E) Detailed view of the 0-300 Kb window for SNPs with MAF>20% .

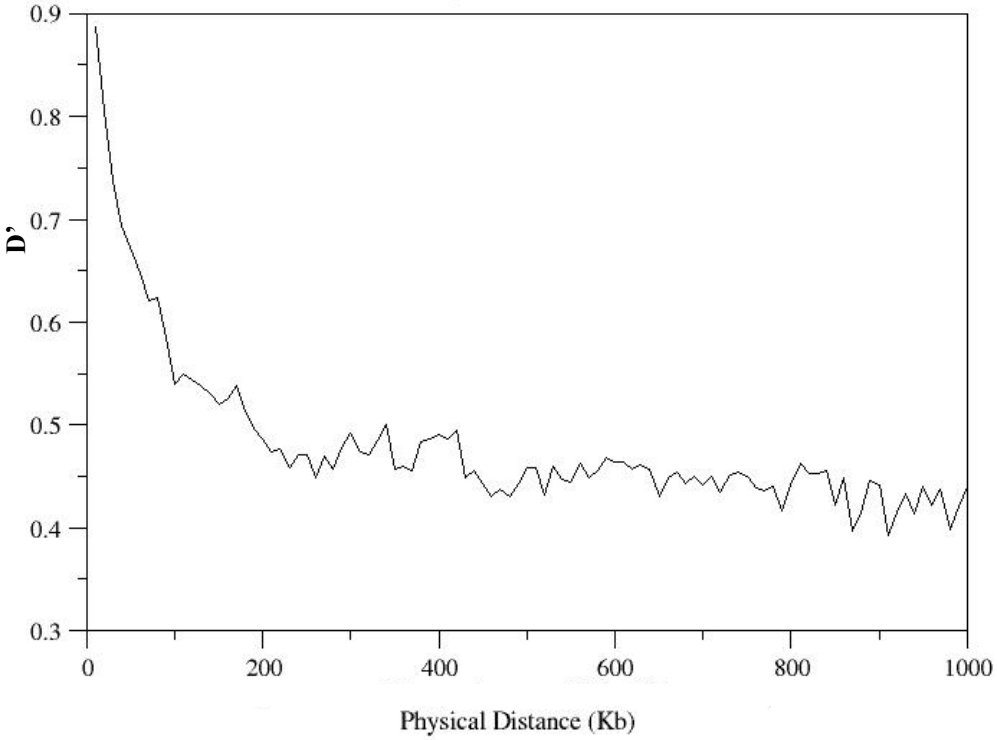
A. D' (using all polymorphic SNPs)



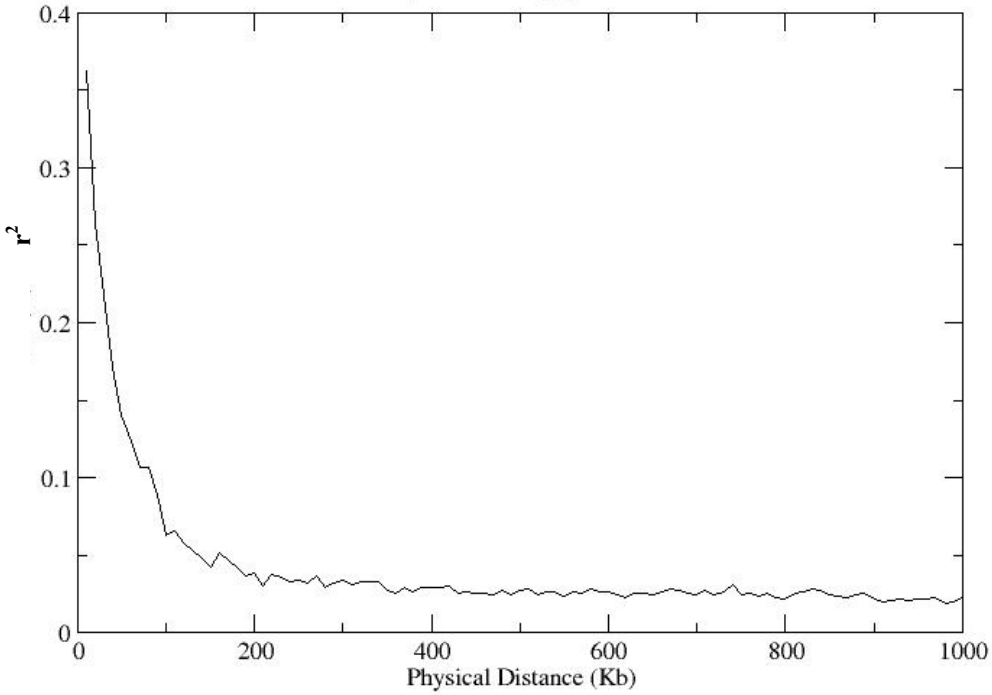
C. r^2 (using all polymorphic SNPs)



A. D' (using all polymorphic SNPs)



C. r^2 (using all polymorphic SNPs)



E. (using SNPs with MAF>20%)