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.

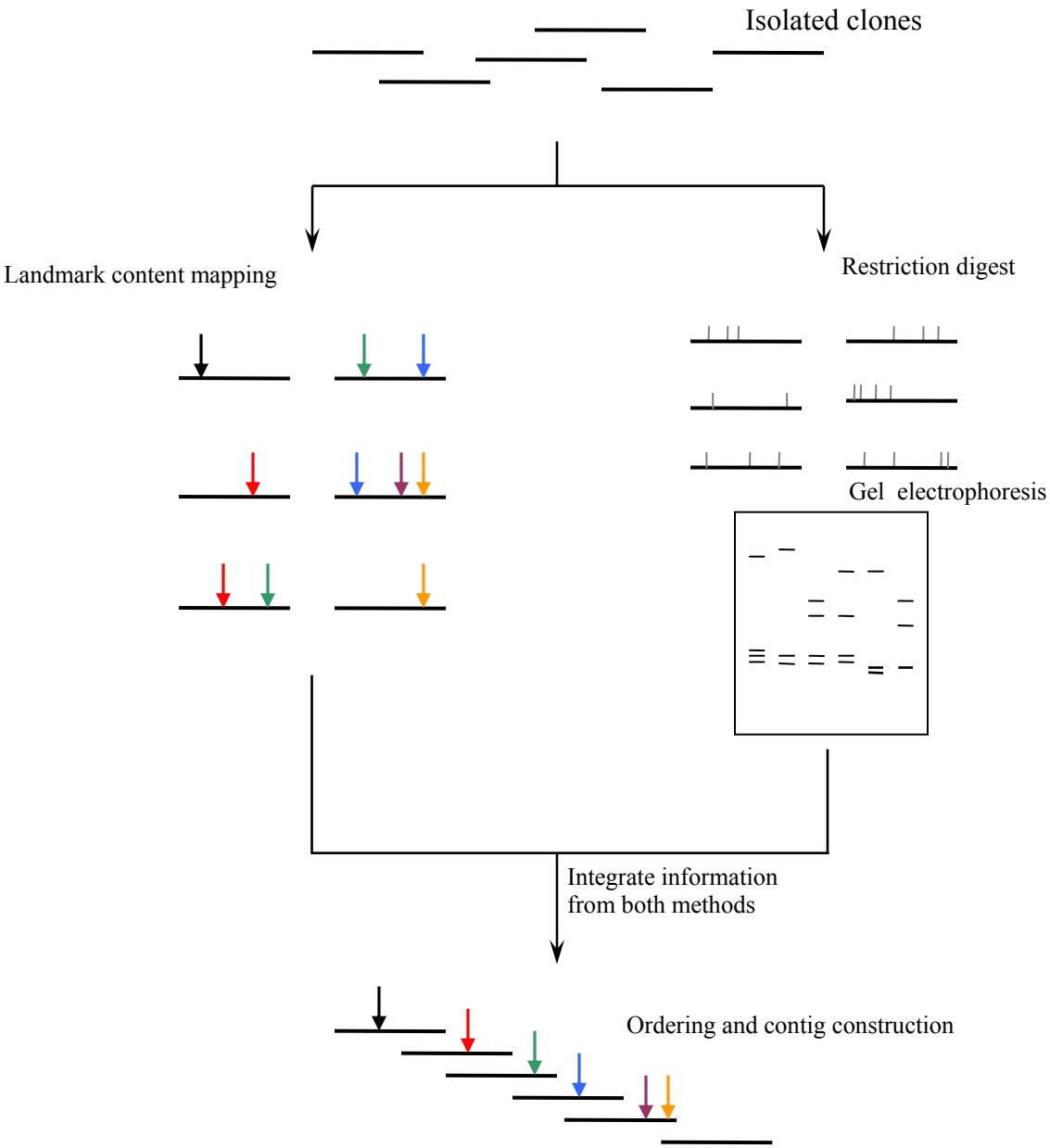


Figure 4.1: Strategy for contig construction, involving landmark content mapping and restriction enzyme fingerprinting.

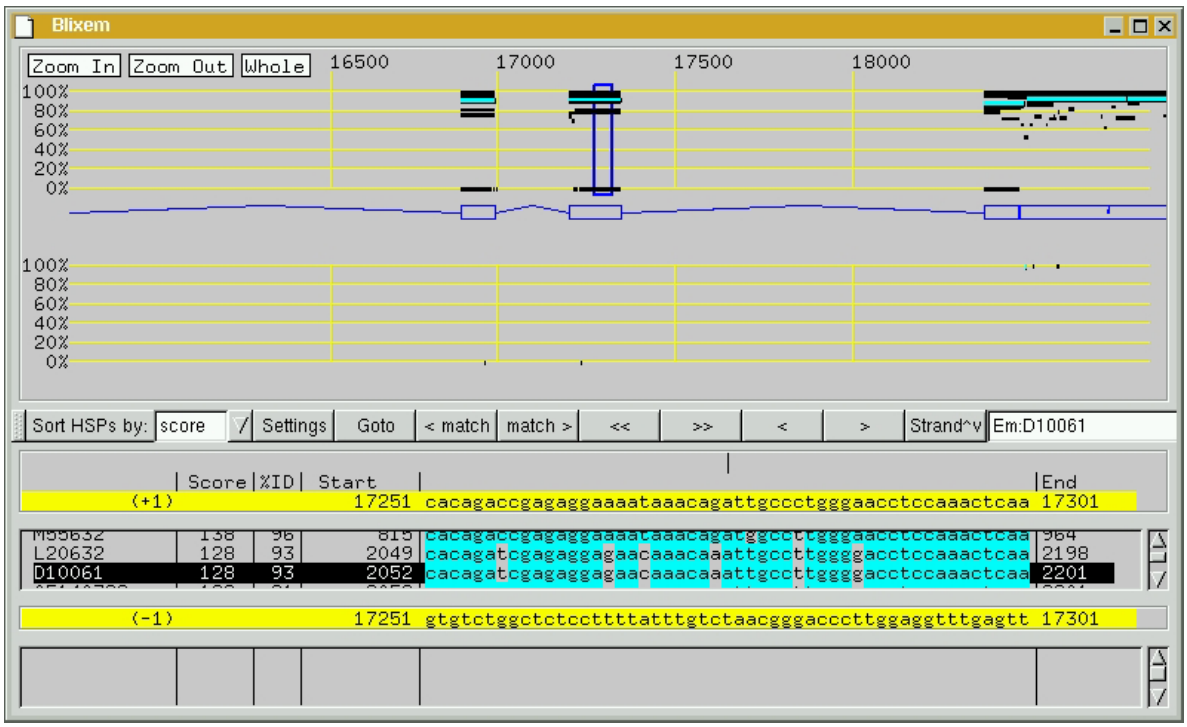
4.2.1 Marker selection and development

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.

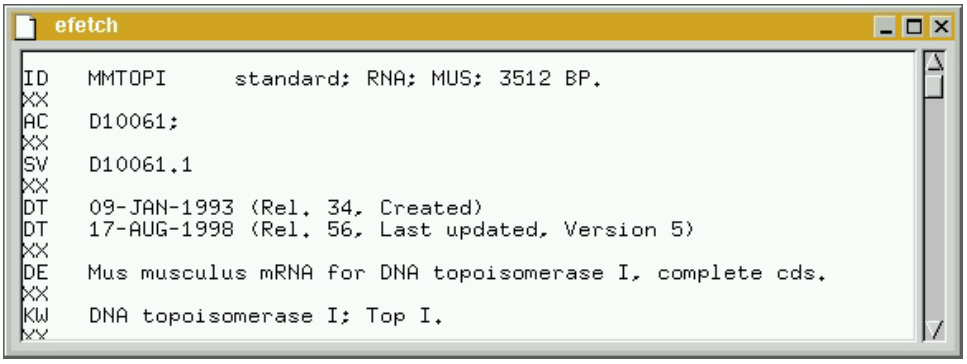


Figure 4.2: Design of mouse STSs. (A) Blixem view of homology clusters. Sequences homologous to the genomic region coding for TOP1 (human) are shown at the top part of the window as black lines whereas their sequence can be viewed at the bottom part of the window. The high quality genomic sequence is 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

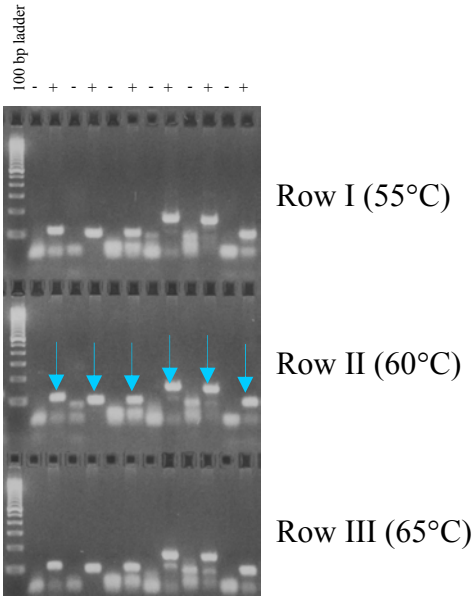


Figure 4.3: Primer testing. PCR products generated using the stSG77025, stSG77026, stSG77027, stSG77028, stSG77029 and stSG77030 primer sets were resolved on an agarose gel. The PCR products generated at 55°C, 60°C and 65°C annealing temperatures are shown on Rows I, II and III respectively. The expected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

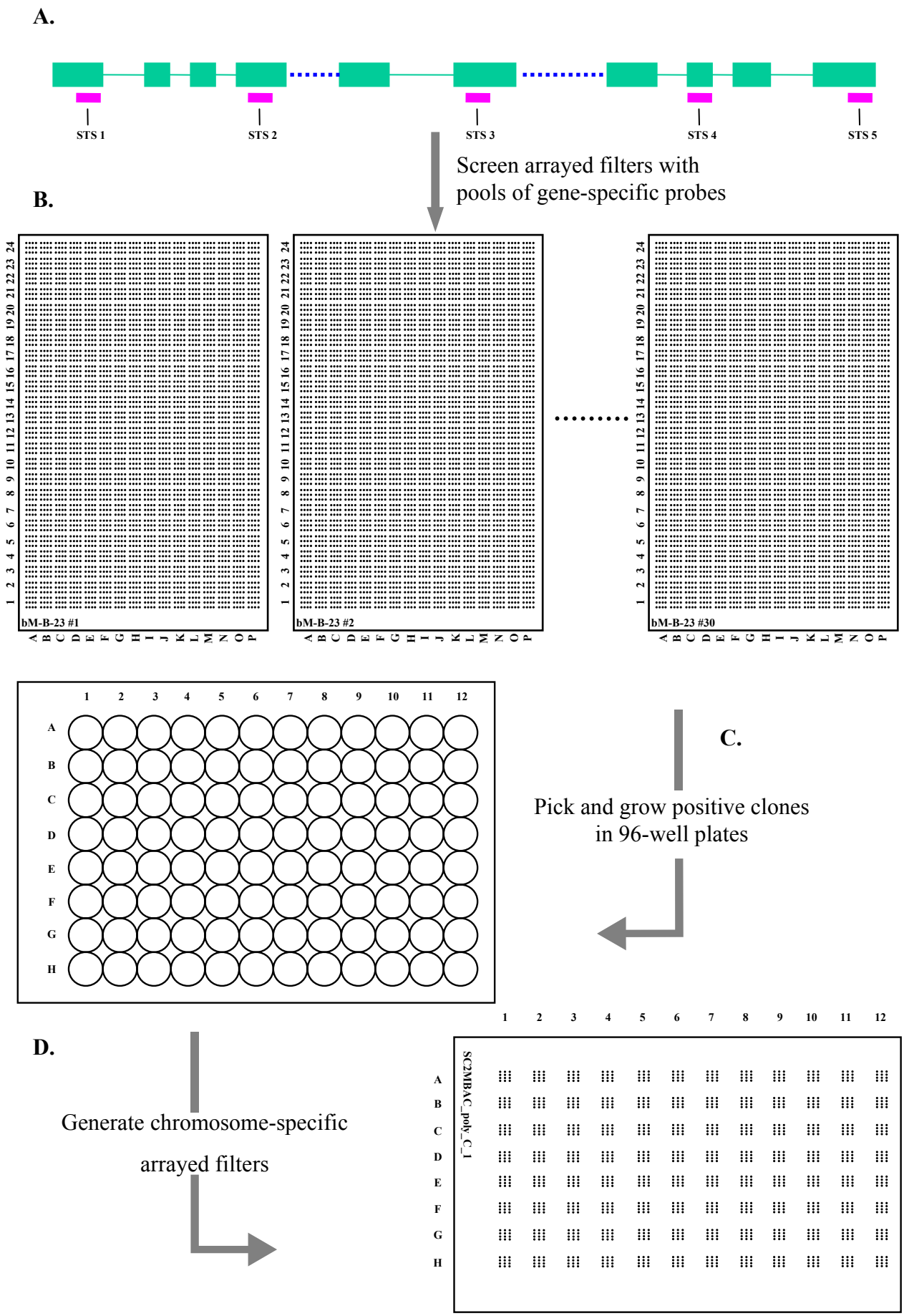
4.2.2 Bacterial clone identification

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

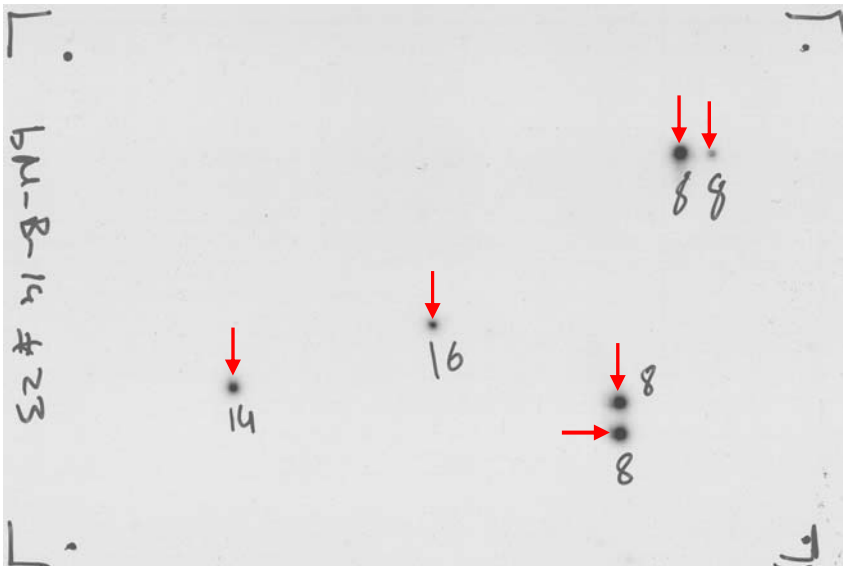
The generated probes (section 4.2.1) were labelled radioactively, pooled together (up to a maximum of 23) and used to hybridise clone filters from the RPCI-23 BAC library. All gene based probes were used in four pooled hybridisation experiments to identify 749 positive BAC clones. Data is stored in the 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.

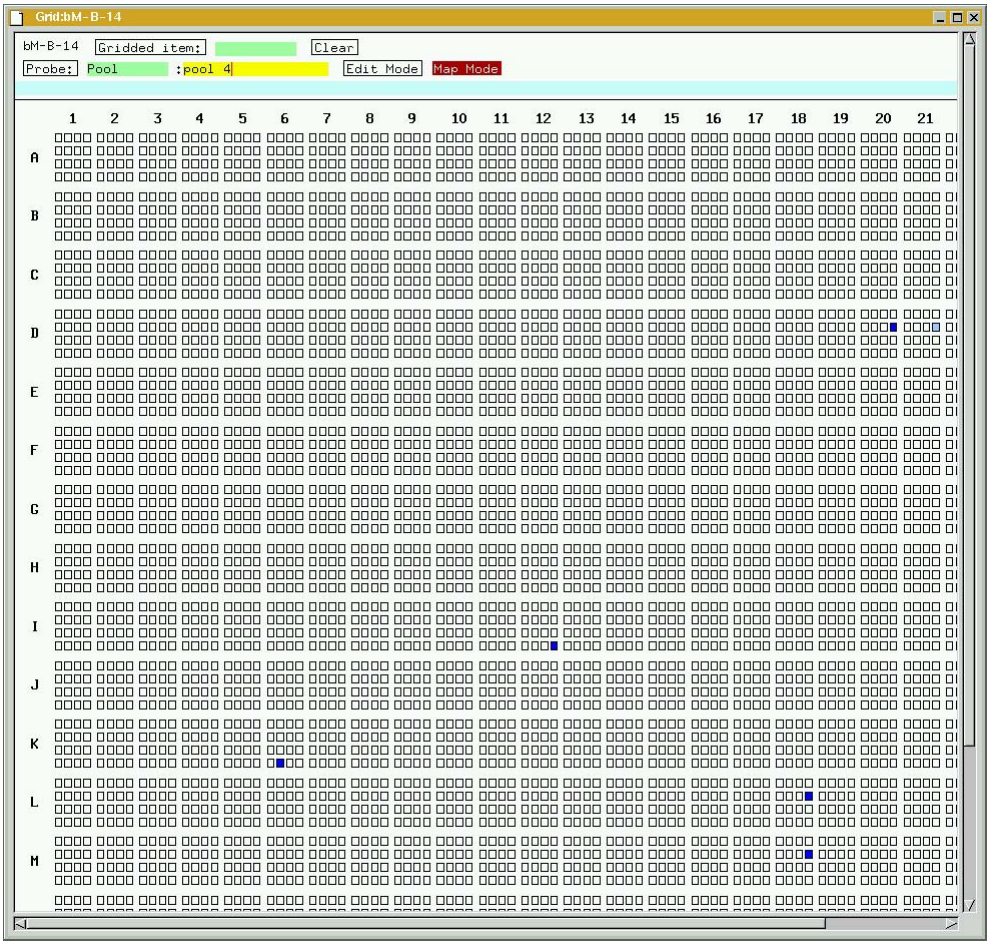


Figure 4.5: Positive clone identification and scoring. (A) Autoradiograph of filter bM-B-14 (from filter set 23) hybridised with a pool of radioactively labelled probes (stSG85199, stSG85204, stSG85205, stSG85301, stSG85302 and stSG85303; probe pool 4). Positive clones are indicated by arrows. (B) Part 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.

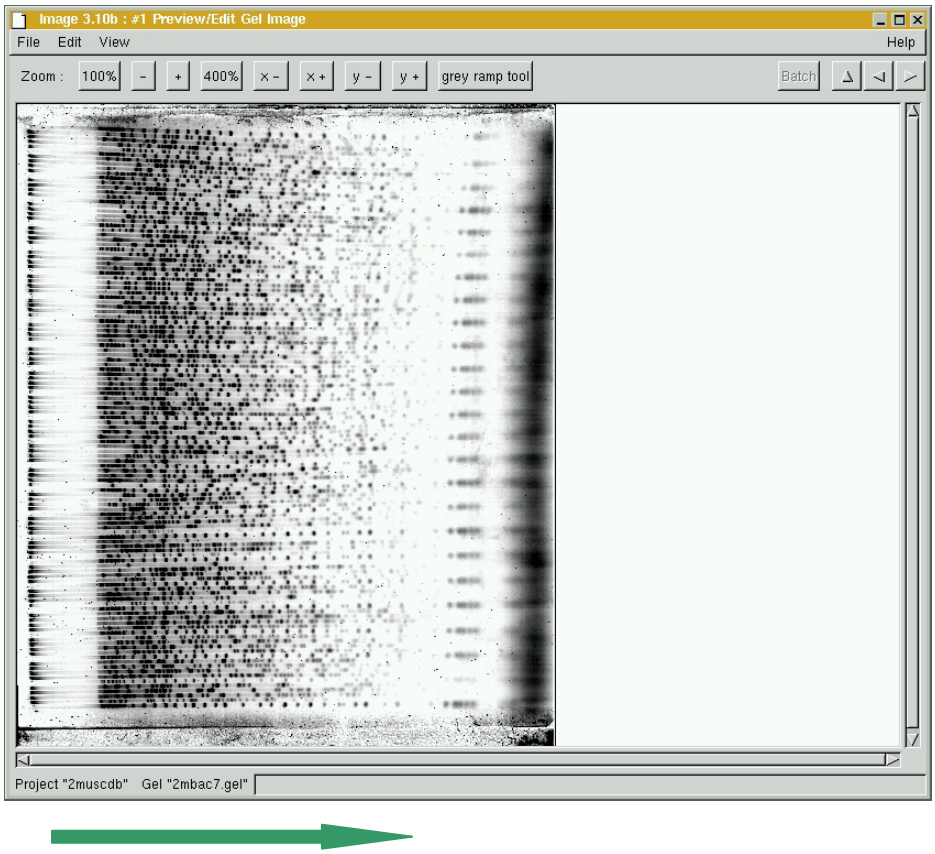


Image 3.10b : #1 Preview/Edit Gel Image
File Edit View Help
Zoom : 100% - + 400% x - x + y - y + grey ramp tool Batch
Project "2muscdb" Gel "2mbac7.gel"

B.

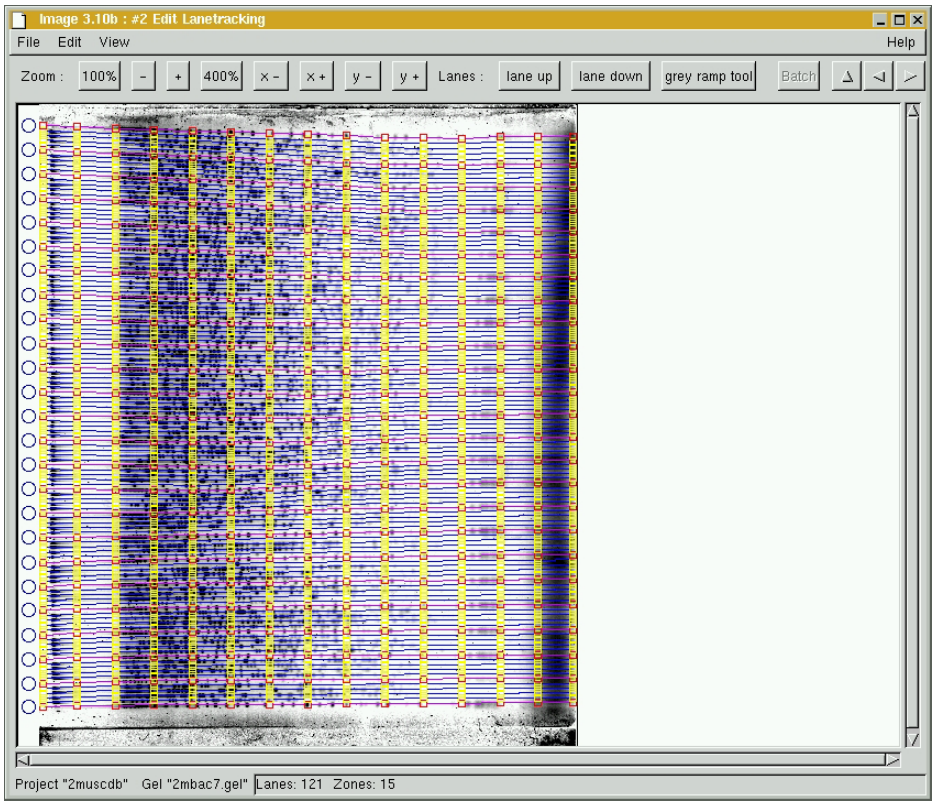


Image 3.10b : #2 Edit Lanetracking
File Edit View Help
Zoom : 100% - + 400% x - x + y - y + Lanes : lane up lane down grey ramp tool Batch
Project "2muscdb" Gel "2mbac7.gel" Lanes: 121 Zones: 15

C.

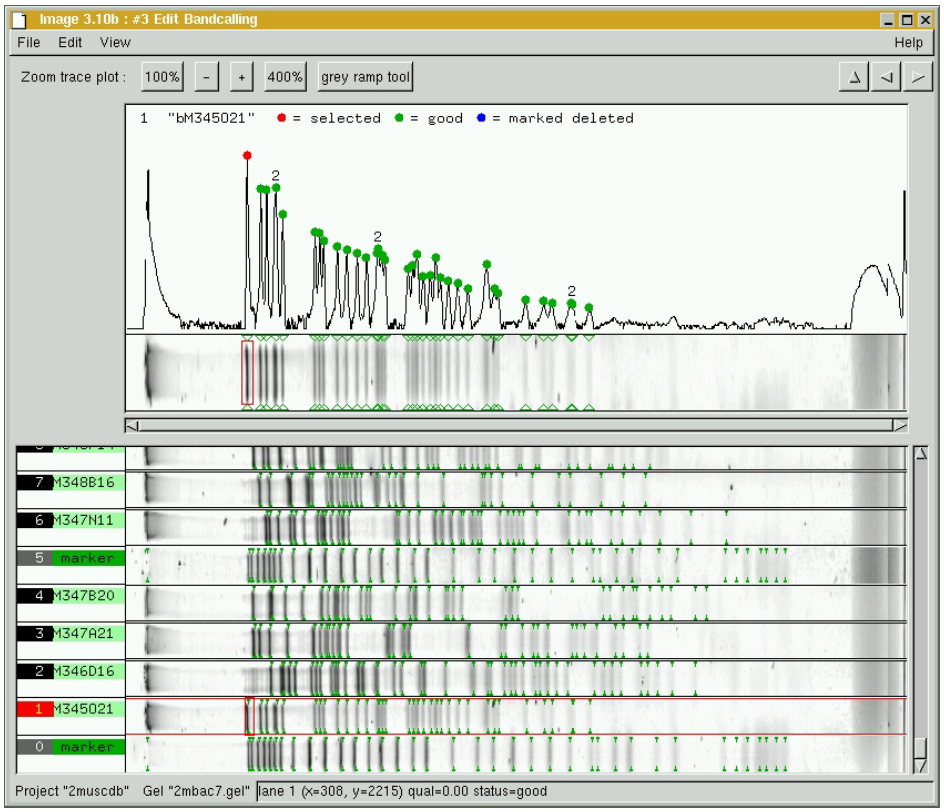


Image 3.10b : #3 Edit Bandcalling
File Edit View
Help
Zoom trace plot :
100%
-
+
400%
grey ramp tool
1 "bM345021" = selected = good = marked deleted
7 M348B16
6 M347N11
5 marker
4 M347B20
3 M347A21
2 M346D16
1 M345021
0 marker
Project "2muscdb" Gel "2mbac7.gel" lane 1 (x=308, y=2215) qual=0.00 status=good

D.

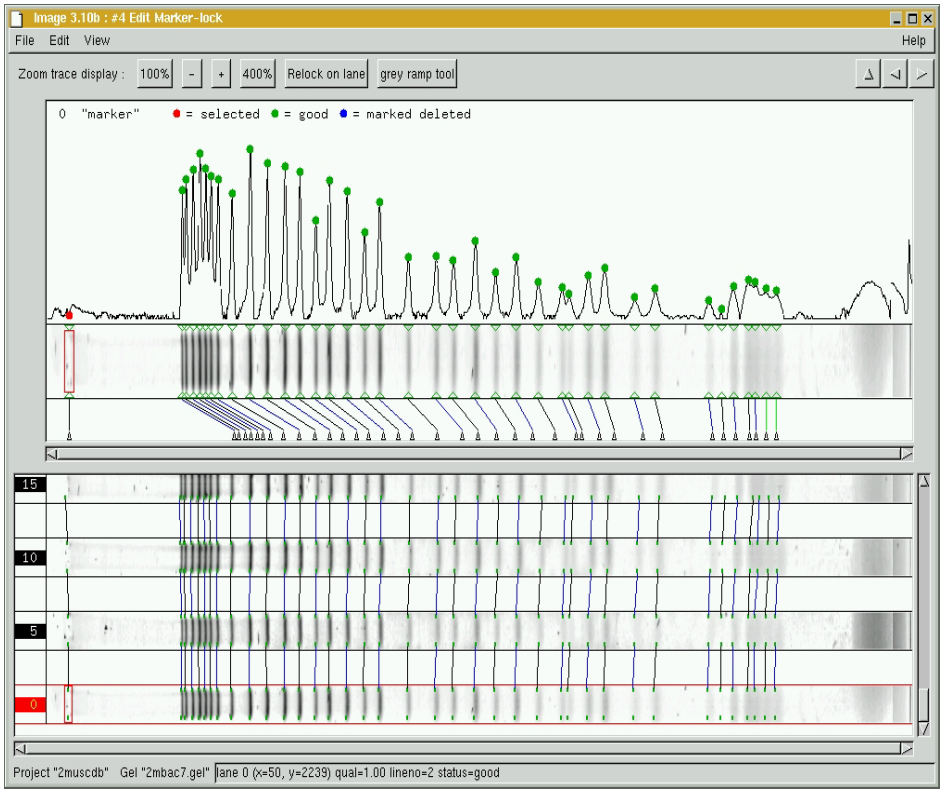


Image 3.10b : #4 Edit Marker-lock
File Edit View
Help
Zoom trace display :
100%
-
+
400%
Relock on lane
grey ramp tool
0 "marker" = selected = good = marked deleted
15
10
5
0
Project "2muscdb" Gel "2mbac7.gel" lane 0 (x=50, y=2239) qual=1.00 lineno=2 status=good

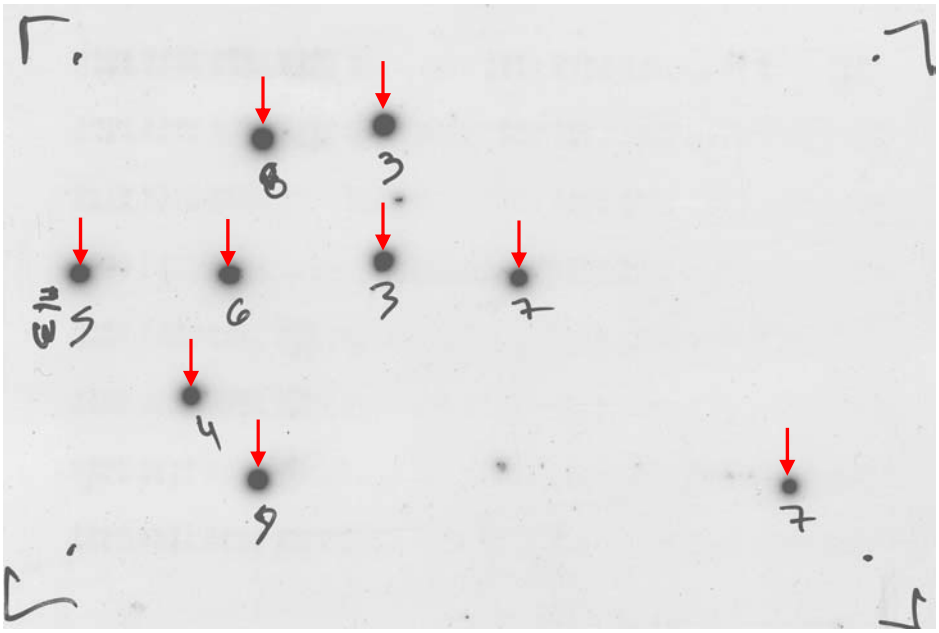
4.2.4 Landmark content mapping

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.

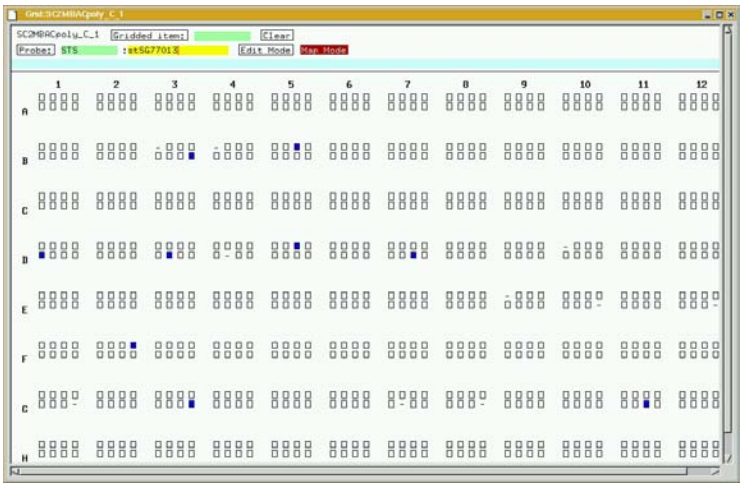
4.2.5 BAC contig assembly in FPC

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.

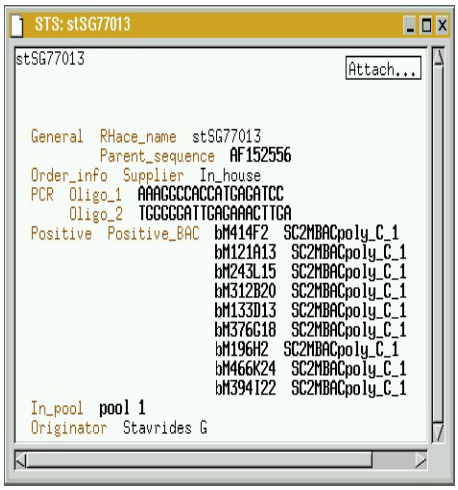


Figure 4.7: Example of landmark content mapping. (A) Hybridisation of polygrid filter 1 with STS stSG77013 (primer set designed using the mouse Ptprt mRNA sequence with EMBL accession number AF152556). Arrows indicate positive BAC clones. (B) 2musace view of polygrid filter 1. Each square represents 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.

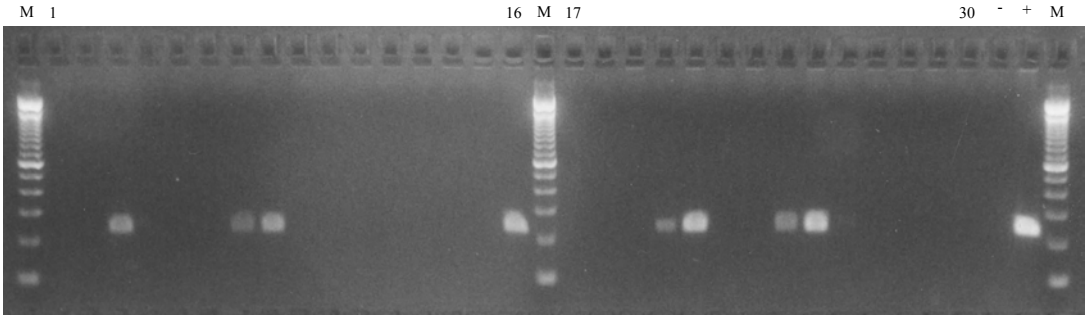
4.2.6 Gap closure

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.

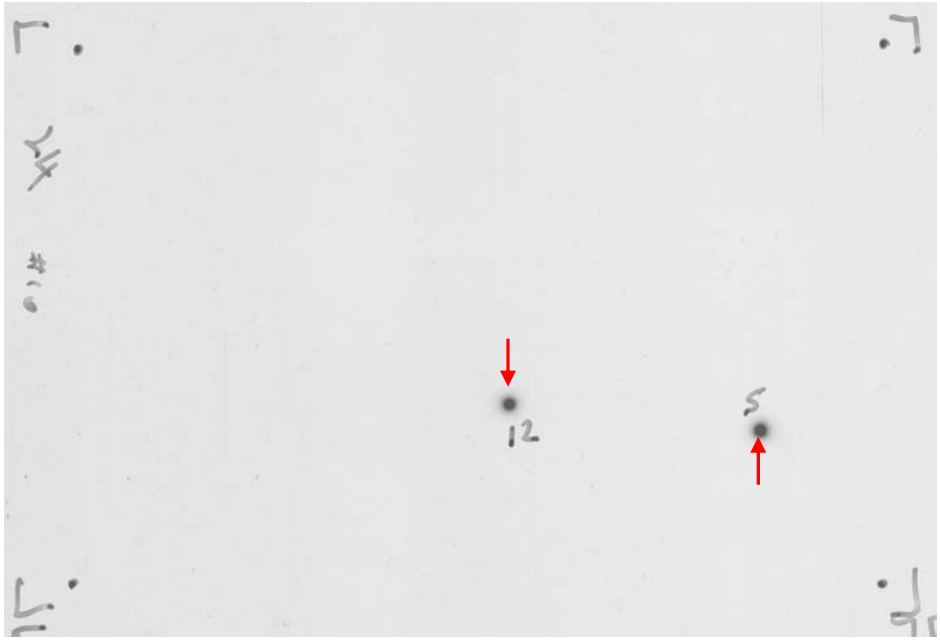


Figure 4.8: Example of PCR-based library screen. (A) End-STS stSG102484 was used to PCR screen the 30 DNA pools representing all RPCI-23 BAC clones. DNA pools representing BACs on filters 3, 7, 8, 16, 20, 21, 24 and 25 were positive. These filters were hybridised with the stSG102484 probe to identify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

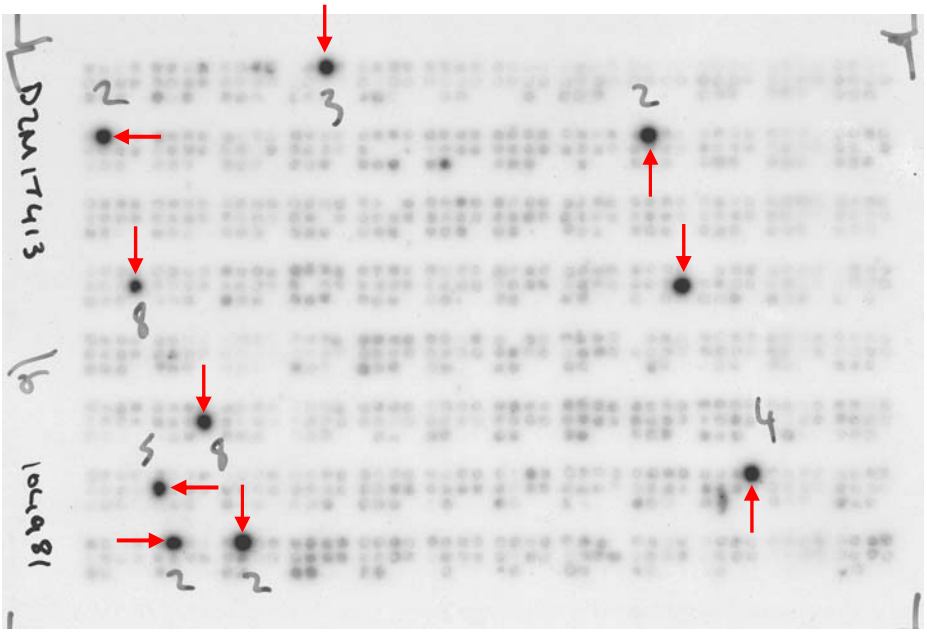


Figure 4.9: Hybridising marker D2MIT413 (stSG104981) to BAC polygrid 2. Arrows indicate the ten positive clones.

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/).

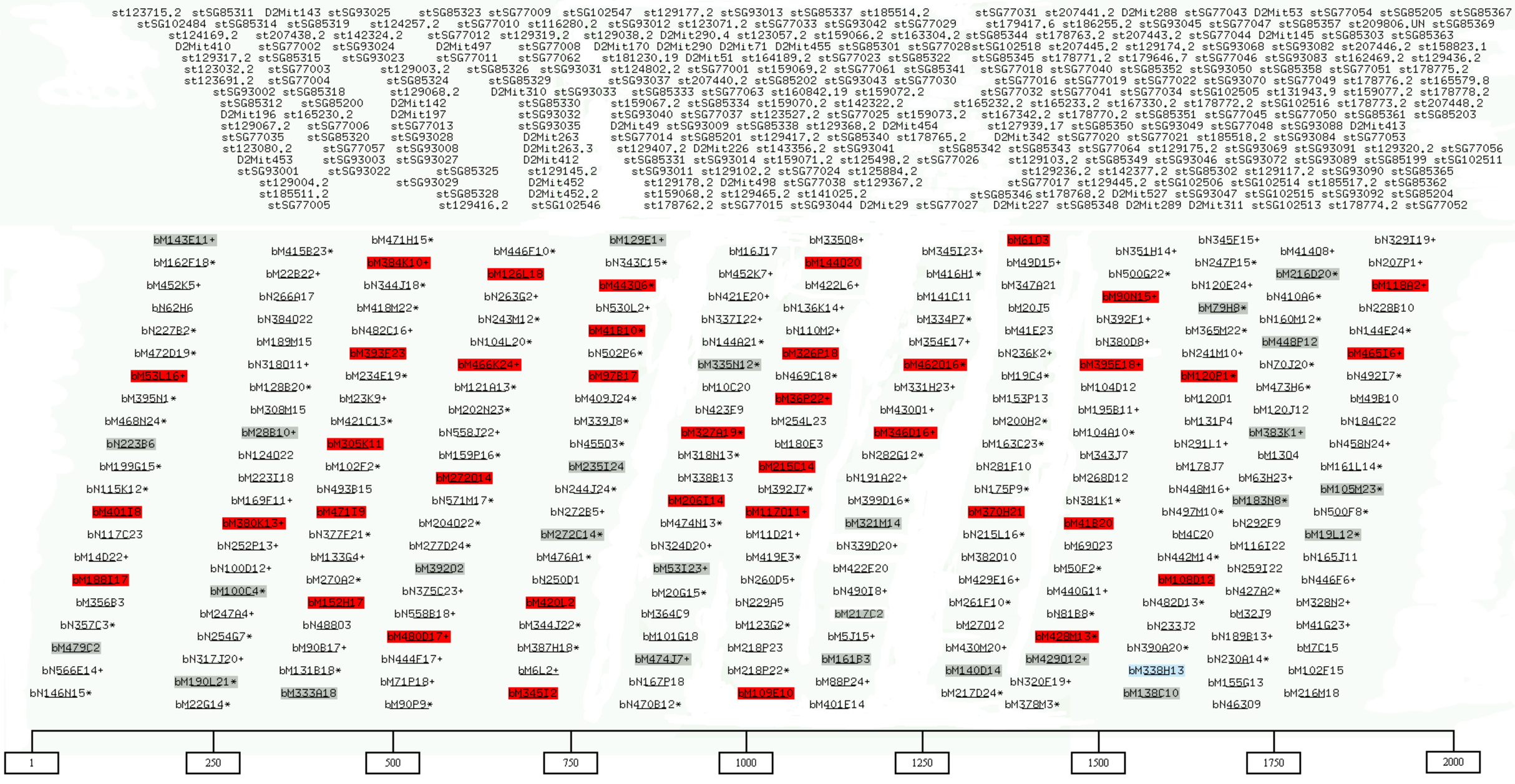
The order of the gene based STS markers on the clone map was compared to the order of the orthologous genes in the human sequence. Gene order was found to be conserved between the two species.

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.

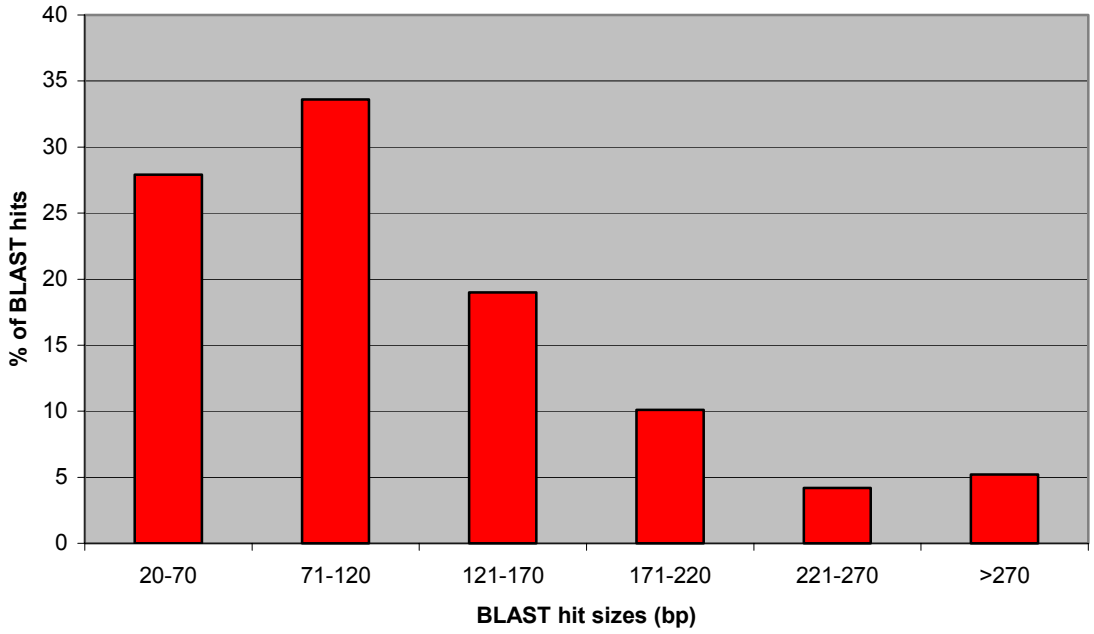


Figure 4.11: Size distribution of mouse BLAST hits

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.

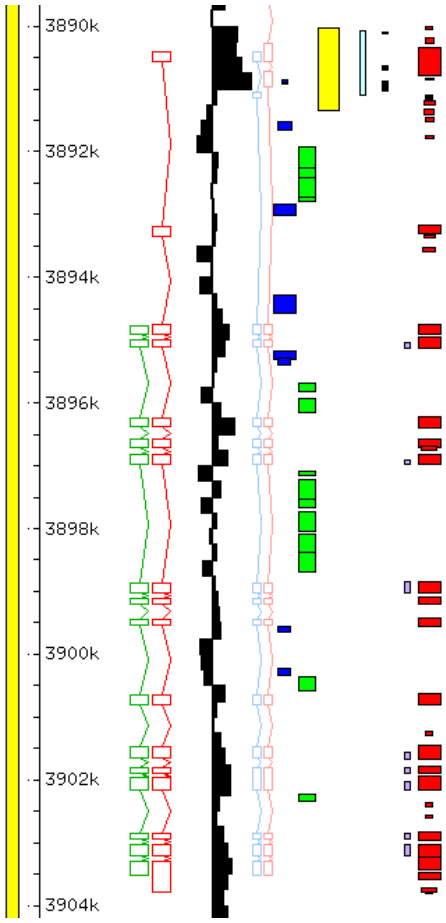


Figure 4.12: ACeDB view of human:mouse BLAST search results. The graphic overview shows the annotated features present in the *human* sequence between 3,890-3,904 Kb. This region encodes for a novel gene (C20orf67) that is similar to the *Drosophila melanogaster* gene CG11399. Sequence analysis identified 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).

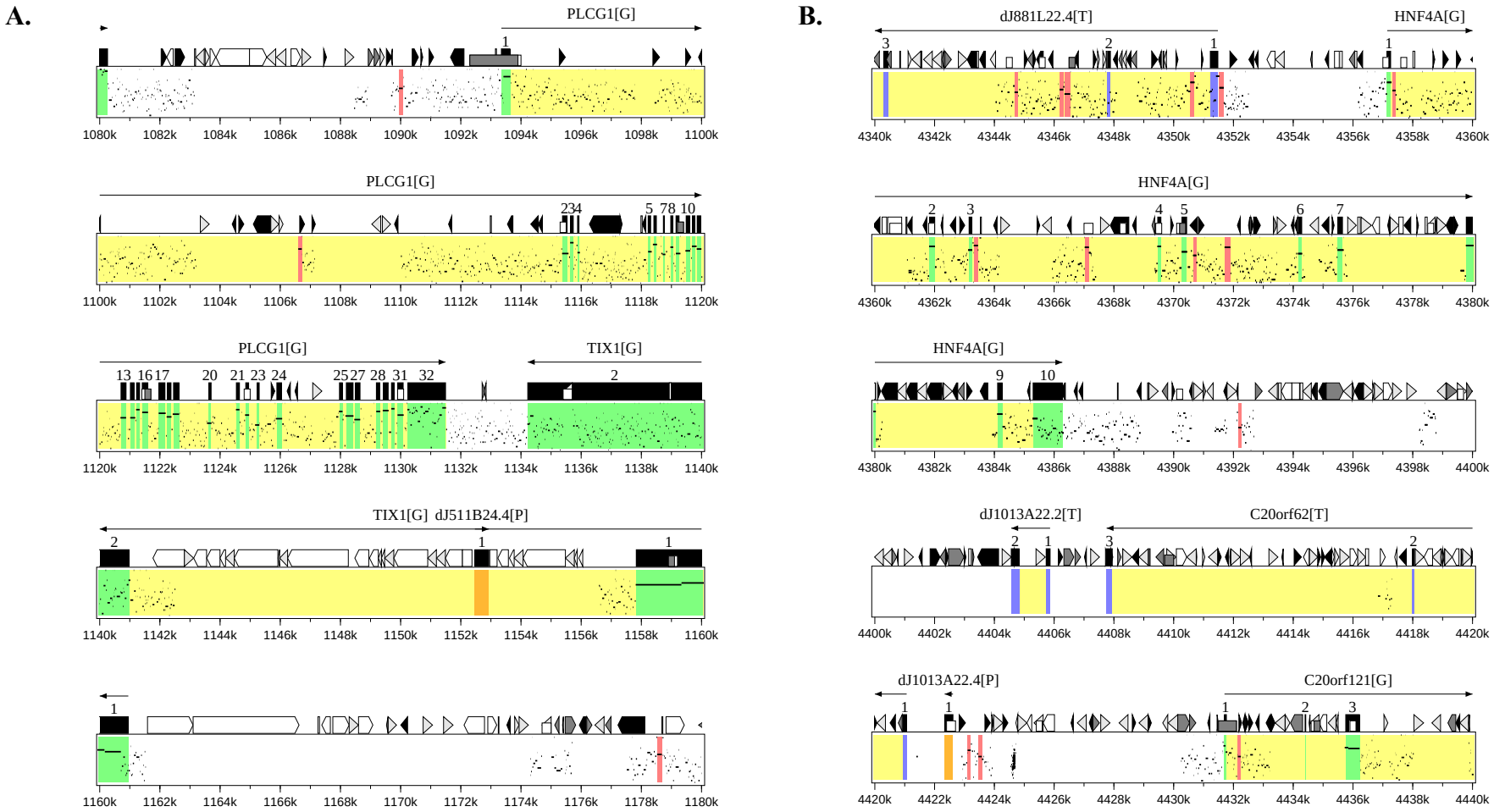


Figure 4.13. Pip plots for two genomic regions. (A) The region containing the PLCG1 and TIX genes and the pseudogene dJ511B24.4 (within the intron of TIX1). (B) The region containing the HNF4A gene, part of the C20orf121 gene, the putative genes (or parts of) dJ881L22.4, dJ1013A22.2 and C20orf62 and 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.

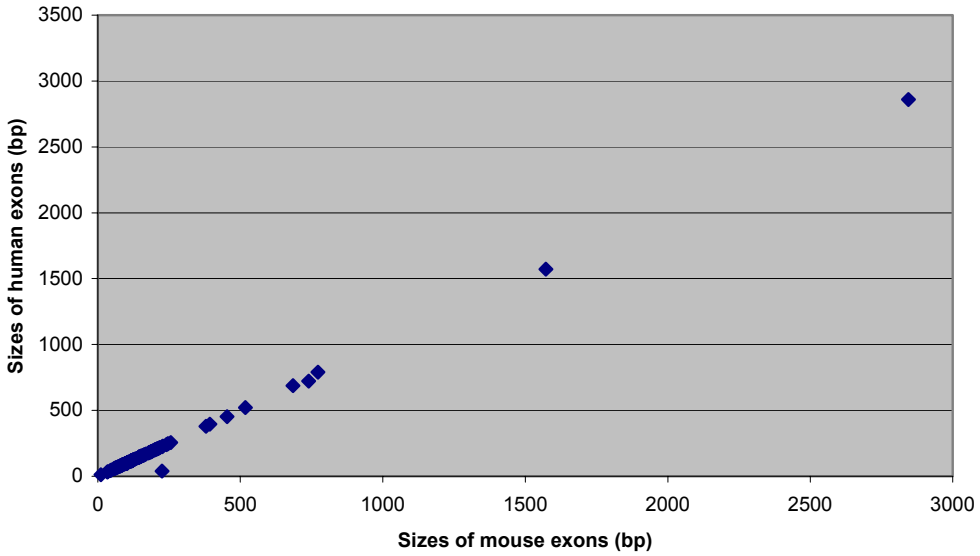
4.6 Finished mouse sequence analysis

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.

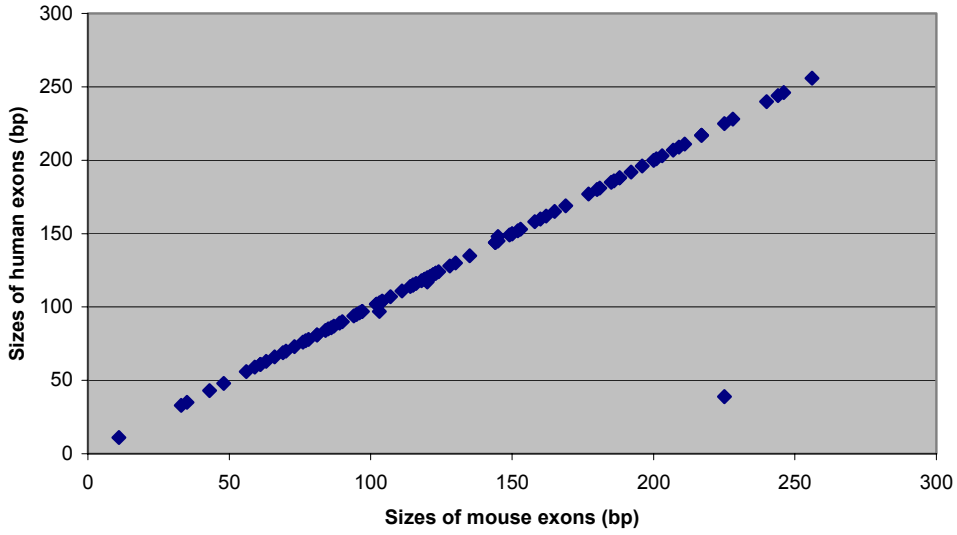
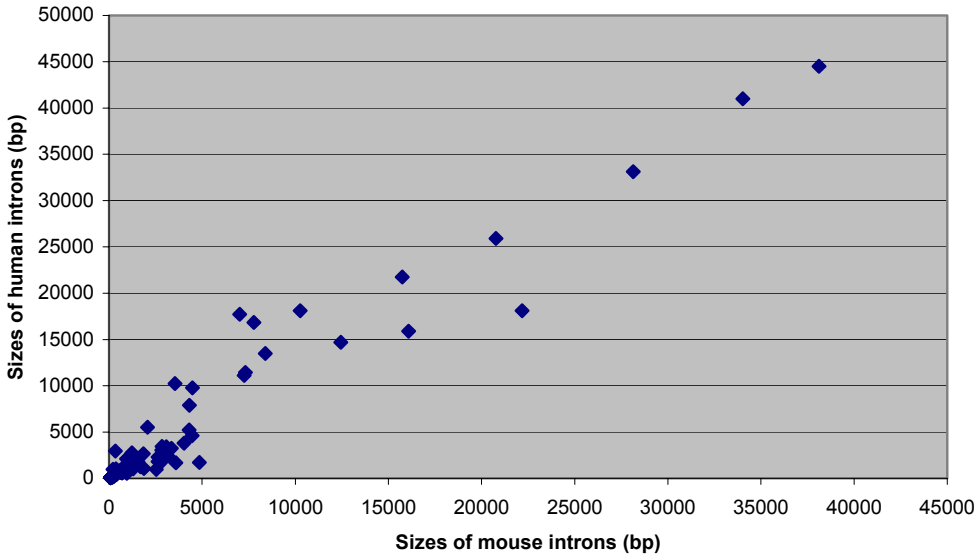


Figure 4.14: (A) Scatter plots for the exon sizes between human and mouse. (B) Detailed view of the 0-300 bp window. The only significant size deviation is between the first exon of human C20orf100 (39 bp) and the orthologous bM117O11.1 (225 bp). Note that the C20orf100 exon is incomplete (no starting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

A.



B.

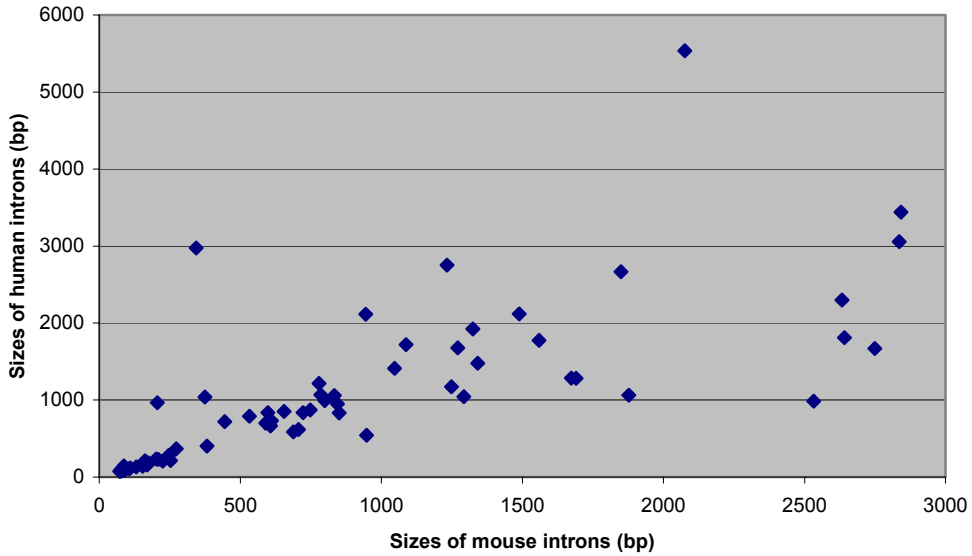


Figure 4.15: (A) Scatter plots for the intron sizes between human and mouse. (B) Detailed view of the 0-3,000 bp window.

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)

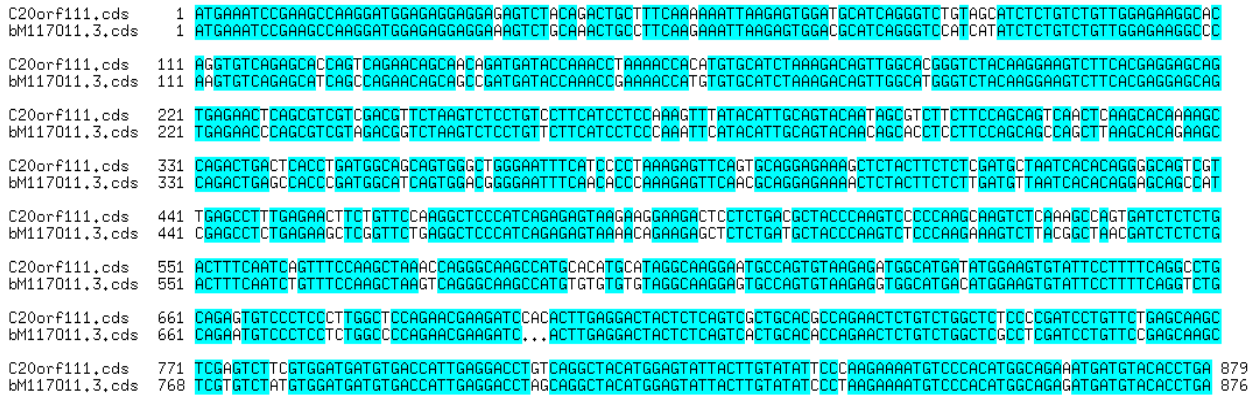


Figure 4.16: Coding sequence alignments (A-C). DNA sequences were aligned using CLUSTAL W (Thompson *et al.*, 1994) and the output was formatted using Belvu (Sonnhammer, unpublished). Identical base pairs are highlighted blue.

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
```

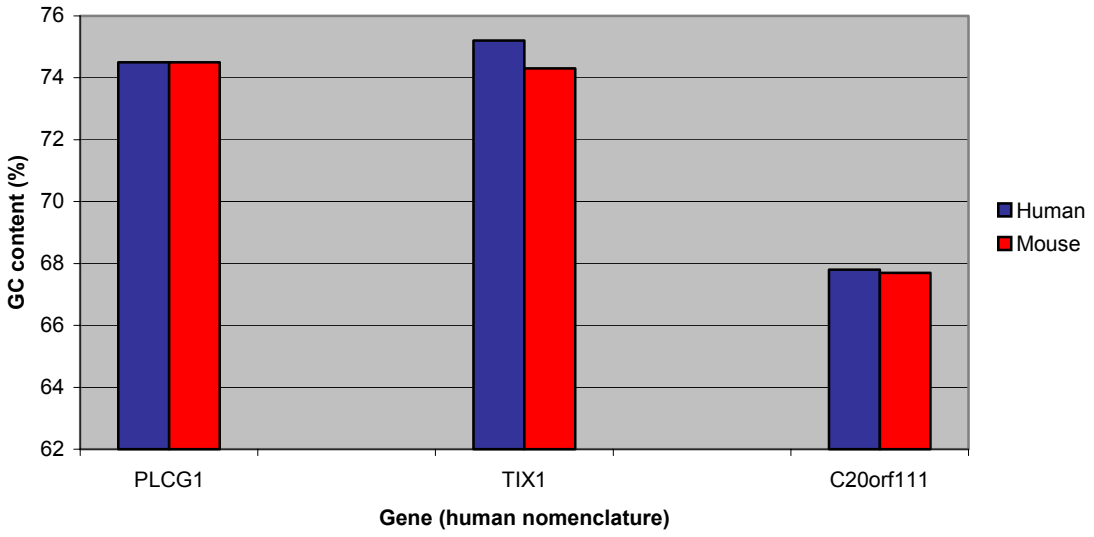
Figure 4.17: Protein sequence alignments (A-C). Sequences were aligned using CLUSTAL W (Thompson *et al.*, 1994) and the output was formatted using Belvu (Sonnhammer, unpublished). Identical aa are highlighted blue and similar, grey.

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.

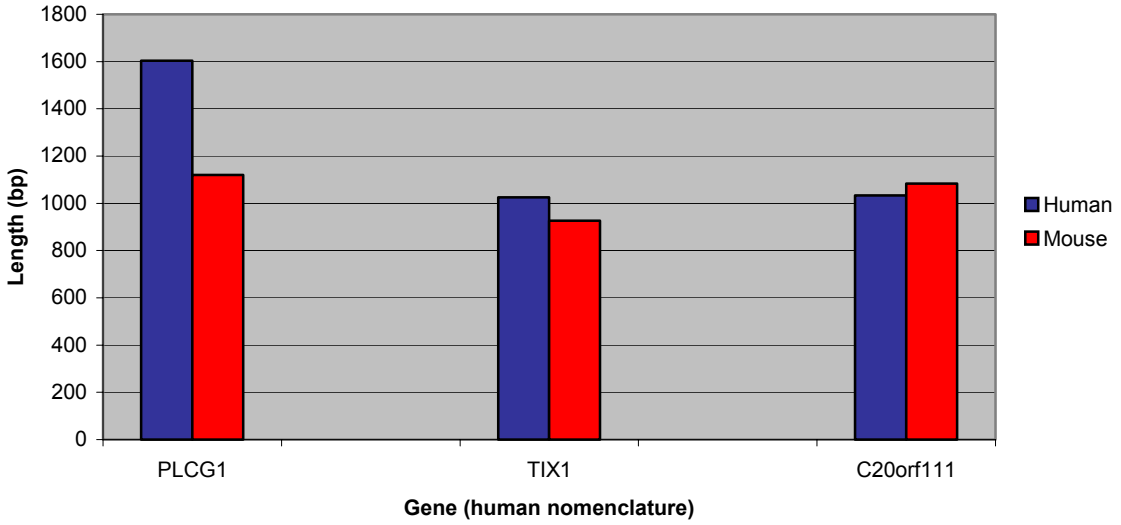
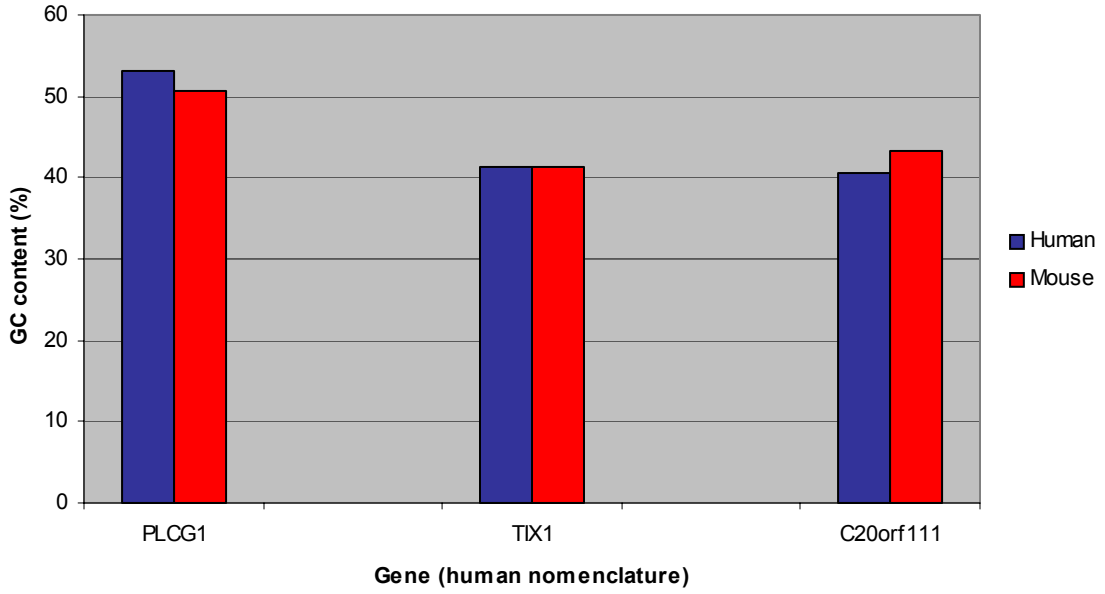


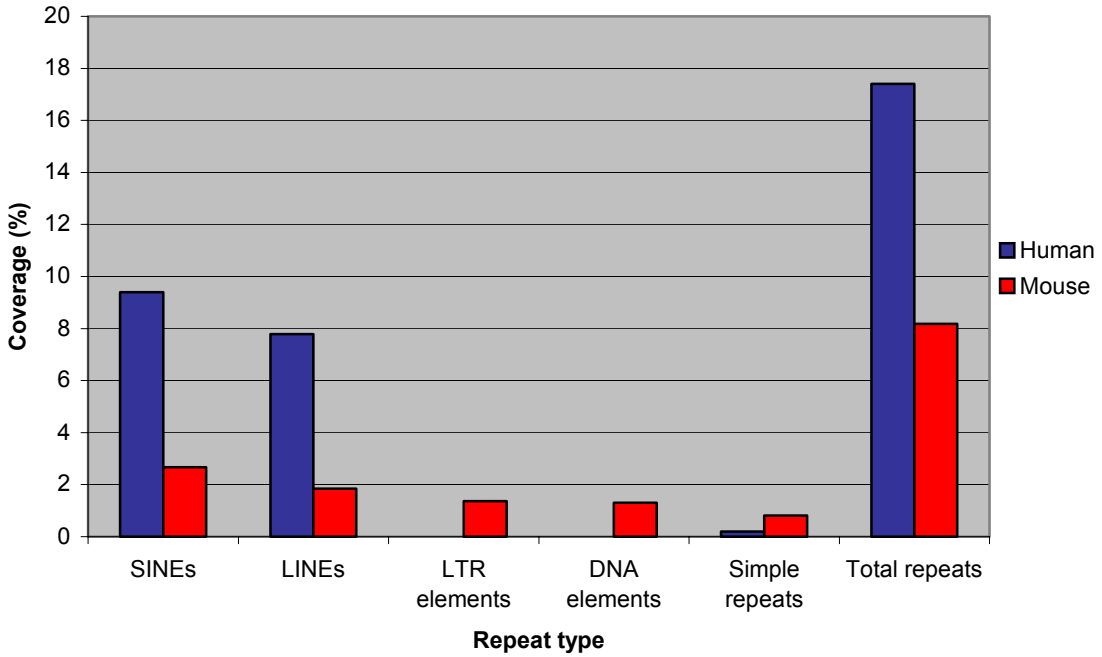
Figure 4.18: CpG island comparison. (A) GC content and (B) Length.

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).

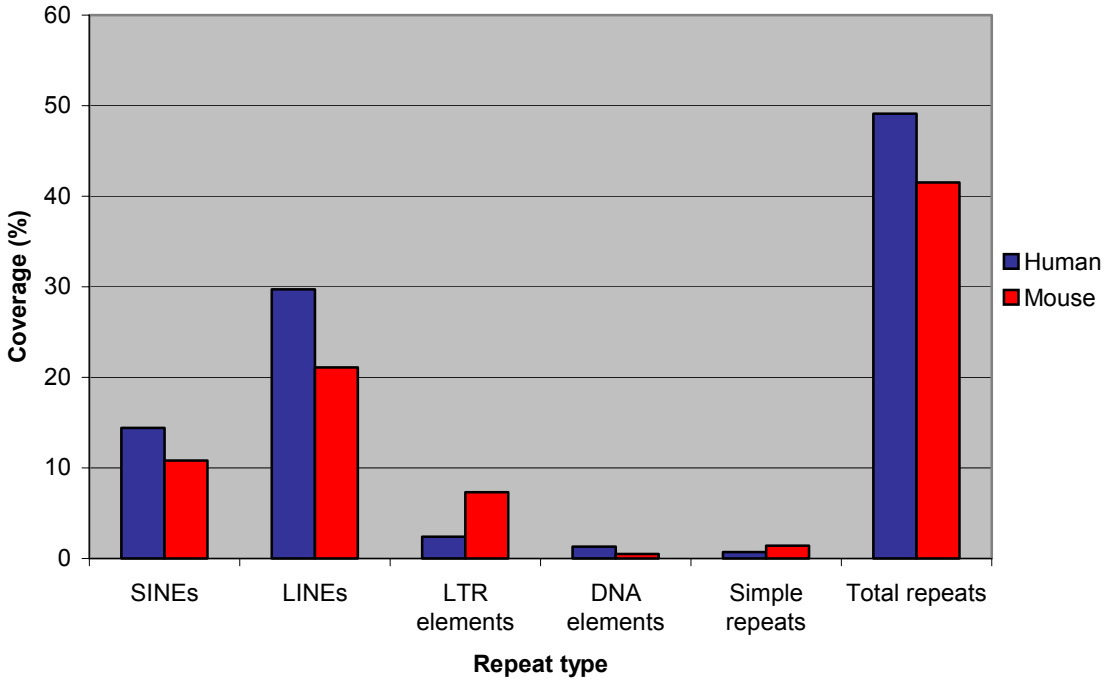
A.



B. PLCG1:Plcg1



C. TIX1:bM393F23.2



D. C20orf111:bM117O11.3

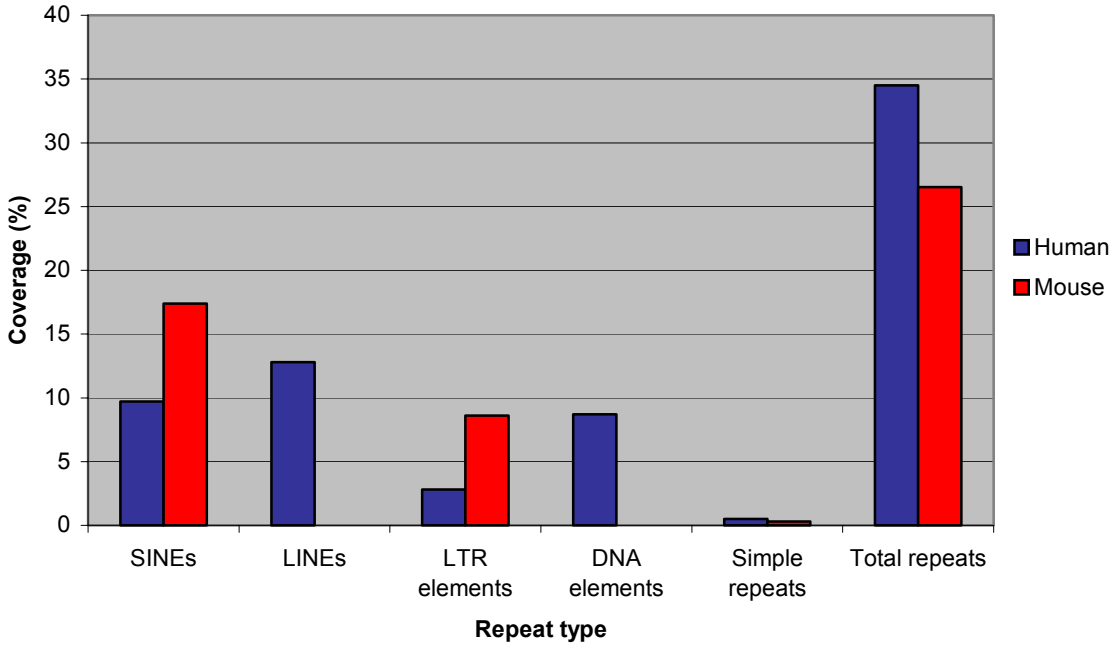


Figure 4.19 (A-D): (A) GC content. (B-D) Repeat content analysis.

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.