Chapter 5

Sequence analysis of 1pcen – 1p13.2

5.1 Introduction

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

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

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

5.2 Sequence Composition Analysis

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

5.2.1 G-Banding

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

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

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

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

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

5.2.2 Isochores

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

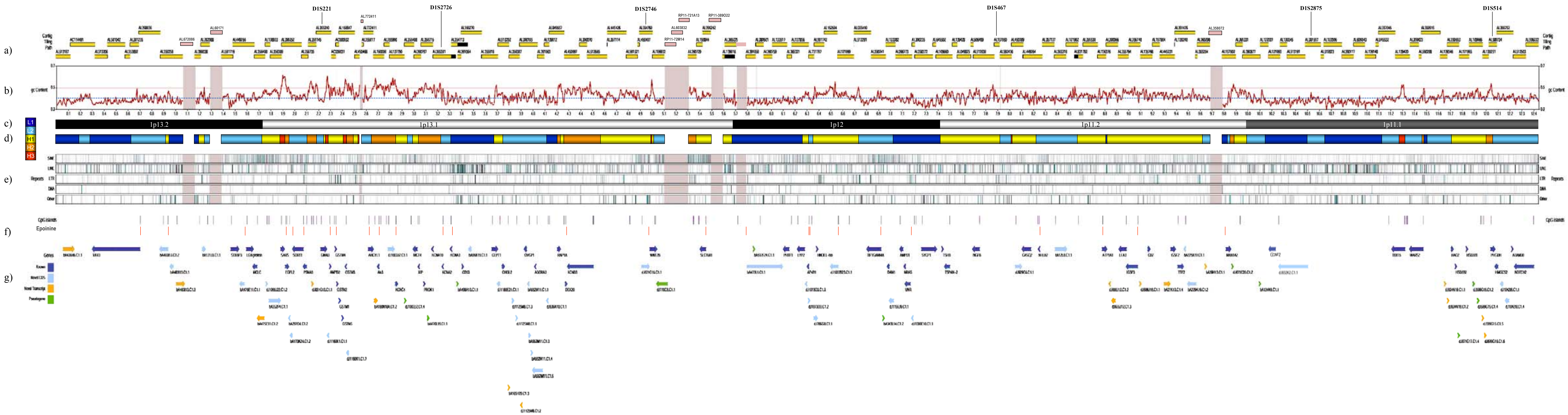

D1S221
D1S2726
D1S2746
D1S467
D1S2875
D1S514
a)
b)
c)
d)
e)
f)
g)
1p13.2
1p13.1
1p12
1p11.2
1p11.1
L1
L2
H1
H2
H3
Epoinine

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

5.2.3. Repeats

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

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

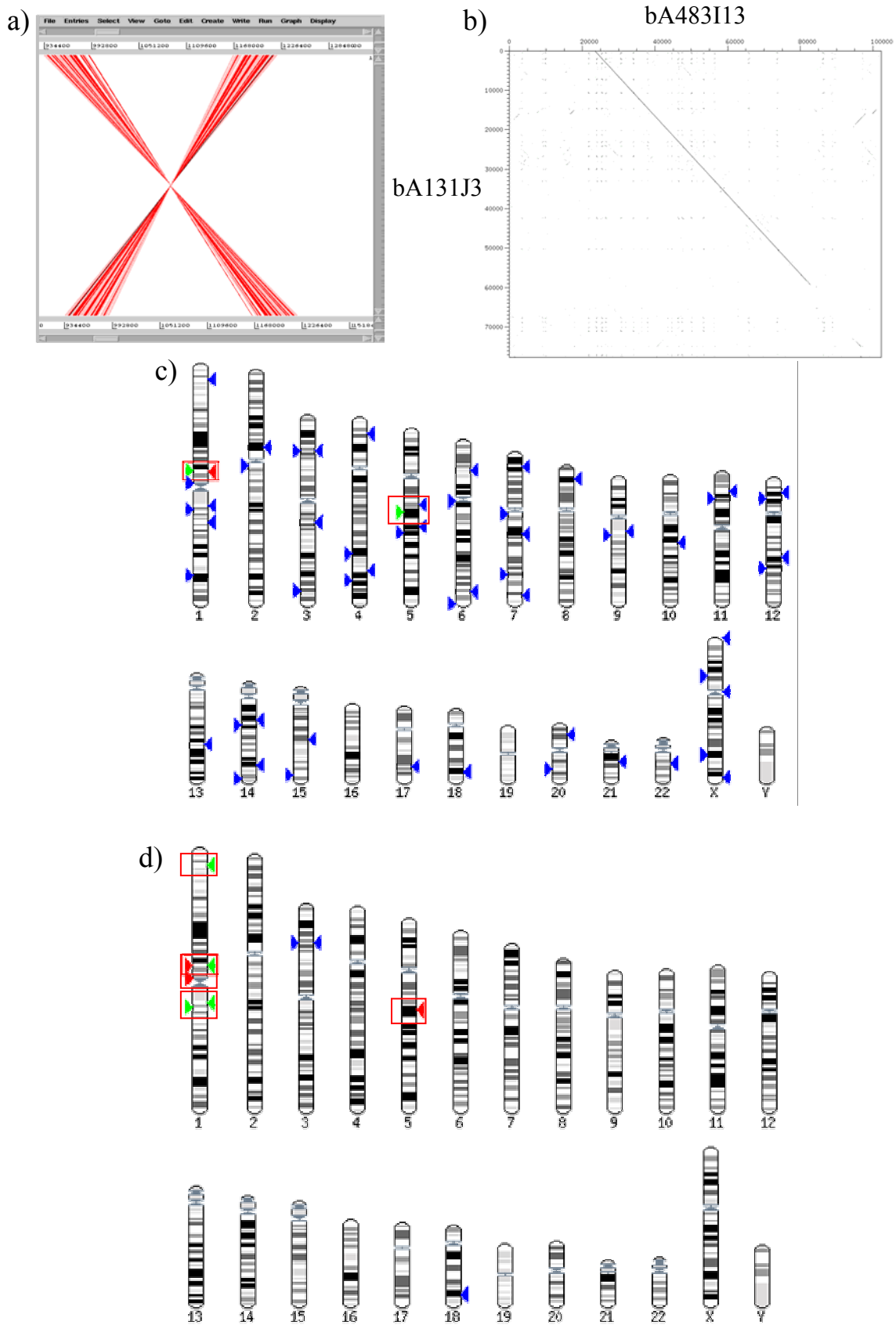
Table 5.2: The breakdown of repeat content within 1pcen – 1p13.2.

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

5.2.4. Low copy repeats

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

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


a)
b)
bA483I13
bA131J3
c)
d)

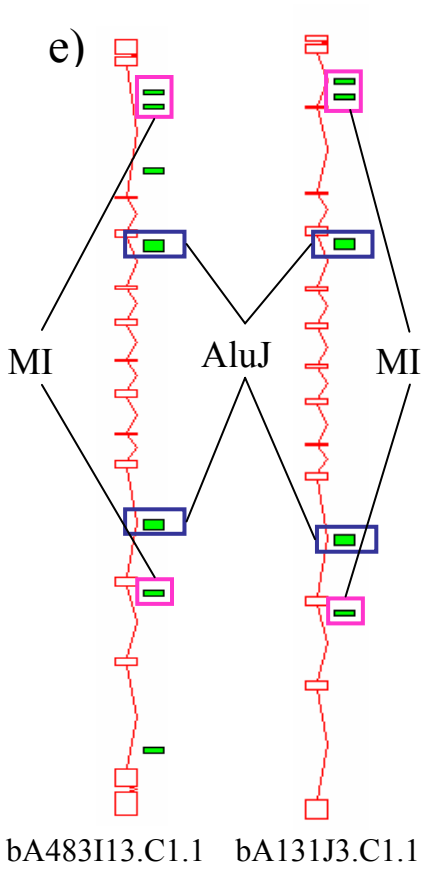


Figure 5.2: Low copy repeat detected within 1pc – 1p13. a) The relationship between two 59Kb inverted repeats, containing two novel genes, show an inverted relationship within ACT. b) Dotter displays the level of sequence homology between the low copy repeat regions. c) BLAST analysis of one of the low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

5.2.5. CpG islands

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

5.2.6 Eponine

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

5.3 Gene Identification

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

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

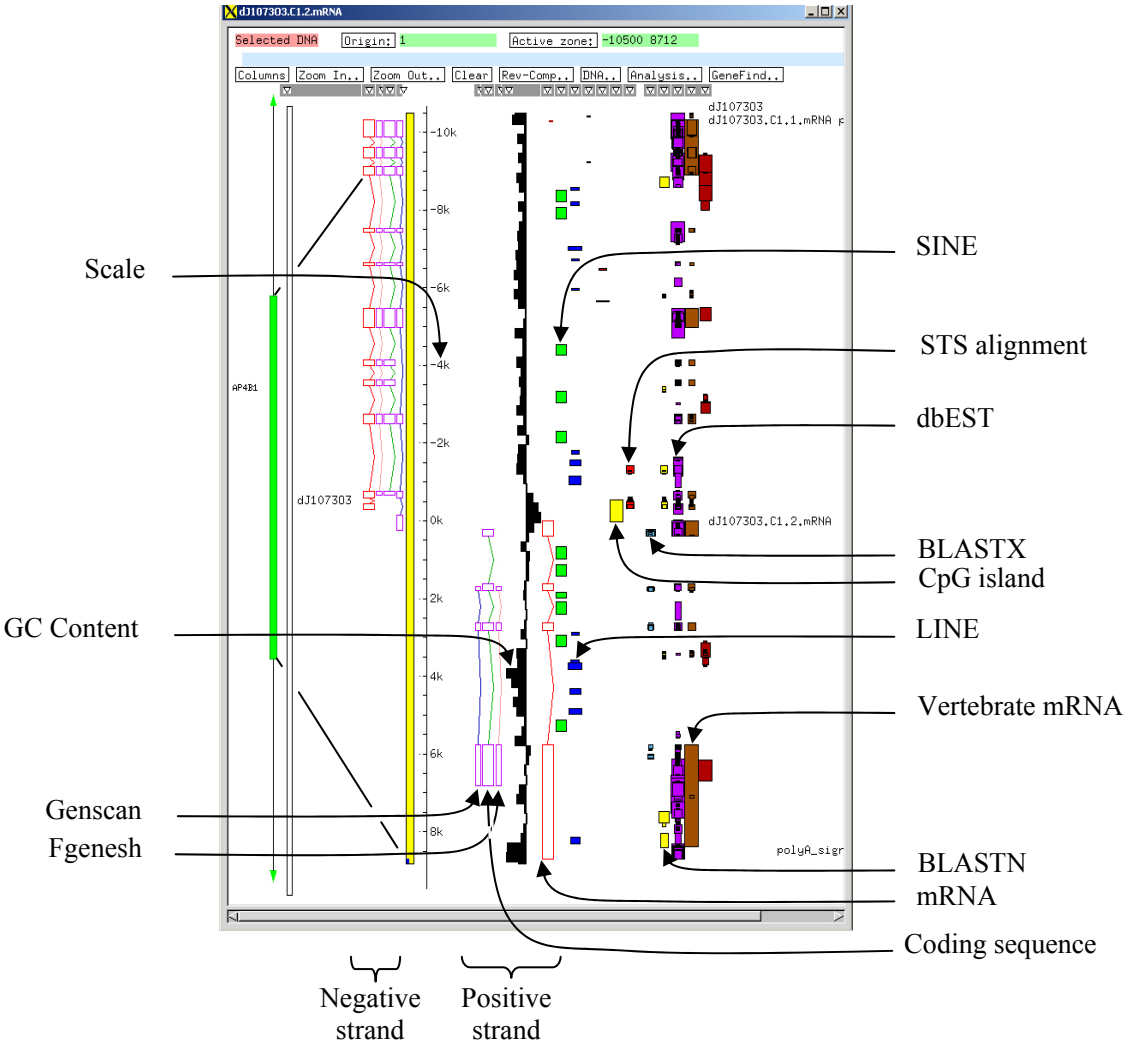


Figure 5.3: An ACeDB display of two annotated genes, including coding sequences, on opposite strands of DNA. Represented are LINE and SINE repeats as well as GC content and CpG island predictions. Vertebrate mRNAs, ESTs and STSs are positioned on genomic sequence by BLAST alignment.

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

5.3.1. Known genes

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

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

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

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

5.3.2. Novel genes

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

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

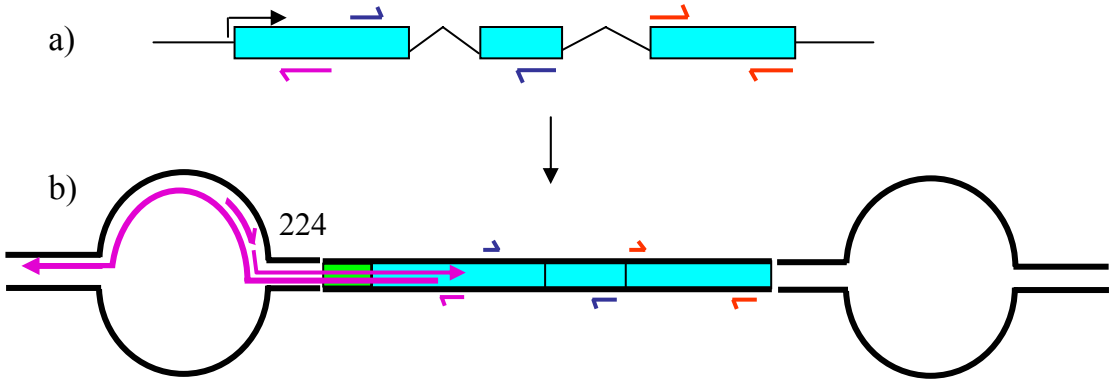


Figure 5.4: Primer combinations used to validate putative gene structures. a) Primers pairs, of the same colour, are designed to the annotated gene structure. Blue primers are designed between exons, red primers within an exon and the pink primer is designed to be used in conjunction with vectorette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

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

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

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

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

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

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

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

5.3.2.1 Splicing ESTs support the structure of a gene

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

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

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

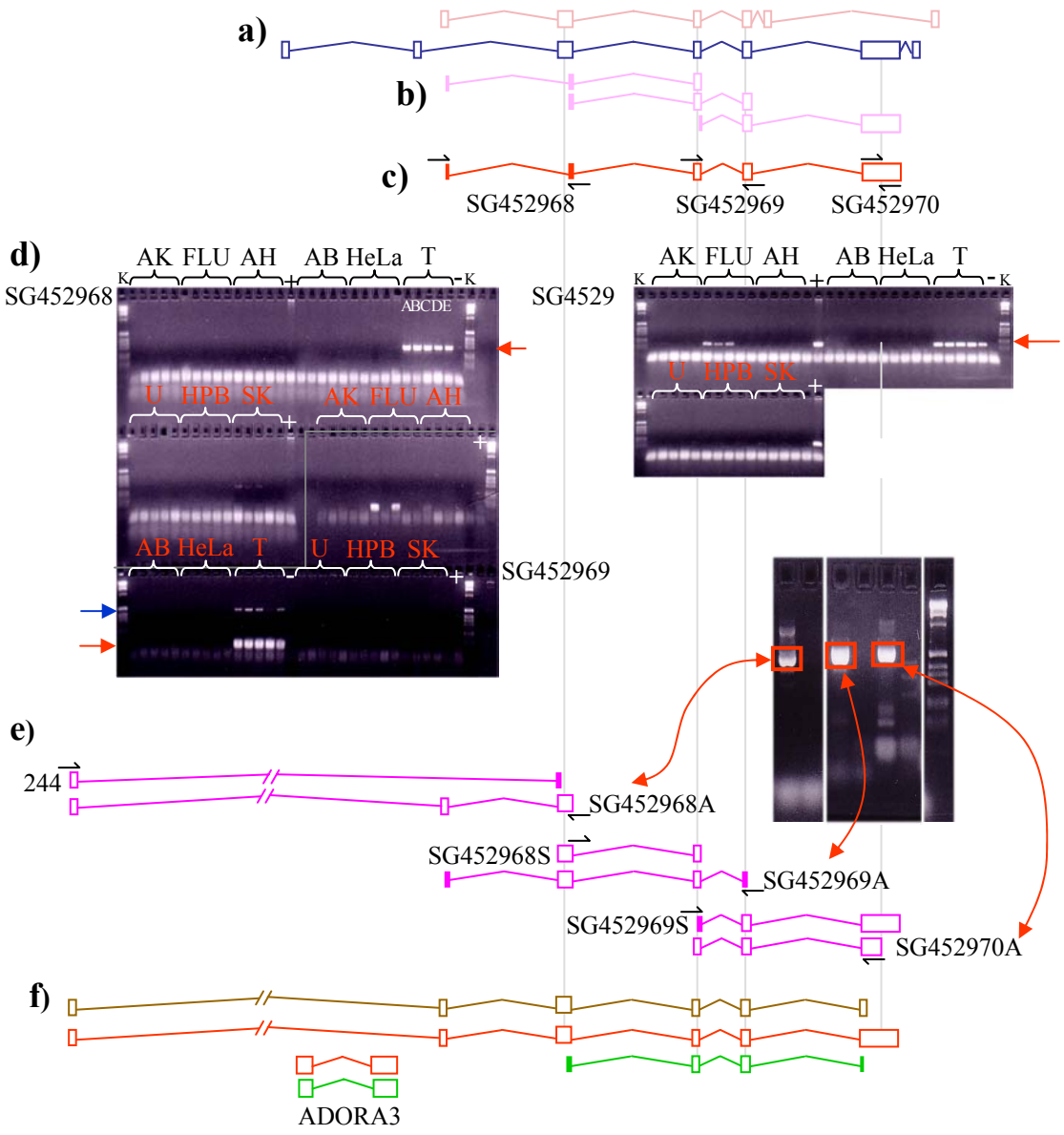


Figure 5.5: The annotation of a novel gene from *de novo* prediction and splicing EST alignment. a) *In silico* prediction of a novel gene is represented by Fgenesh (light pink) and GENSCAN (dark blue) structures. b) Alignment of splicing ESTs (pink) supports the presence of a gene. c) Annotation of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

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

5.3.2.2. mRNA support of novel coding features

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

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

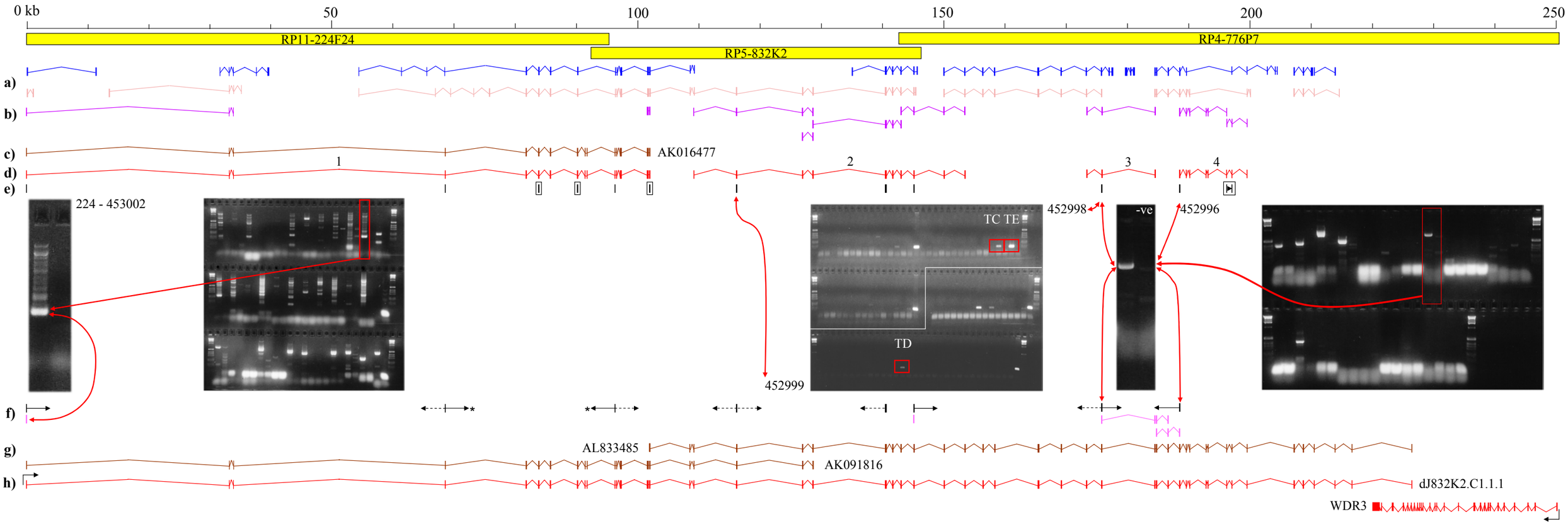

0 kb
50
100
150
200
250
RP11-224F24
RP5-832K2
RP4-776P7
a)
b)
c)
d)
e)
AK016477
1
2
3
4
224 - 453002
452998
-ve
452996
TC TE
TD
452999
f)
g)
AL833485
AK091816
h)
dJ832K2.C1.1.1
WDR3

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

5.3.3 Novel transcripts

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

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

5.3.4 Pseudogenes

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

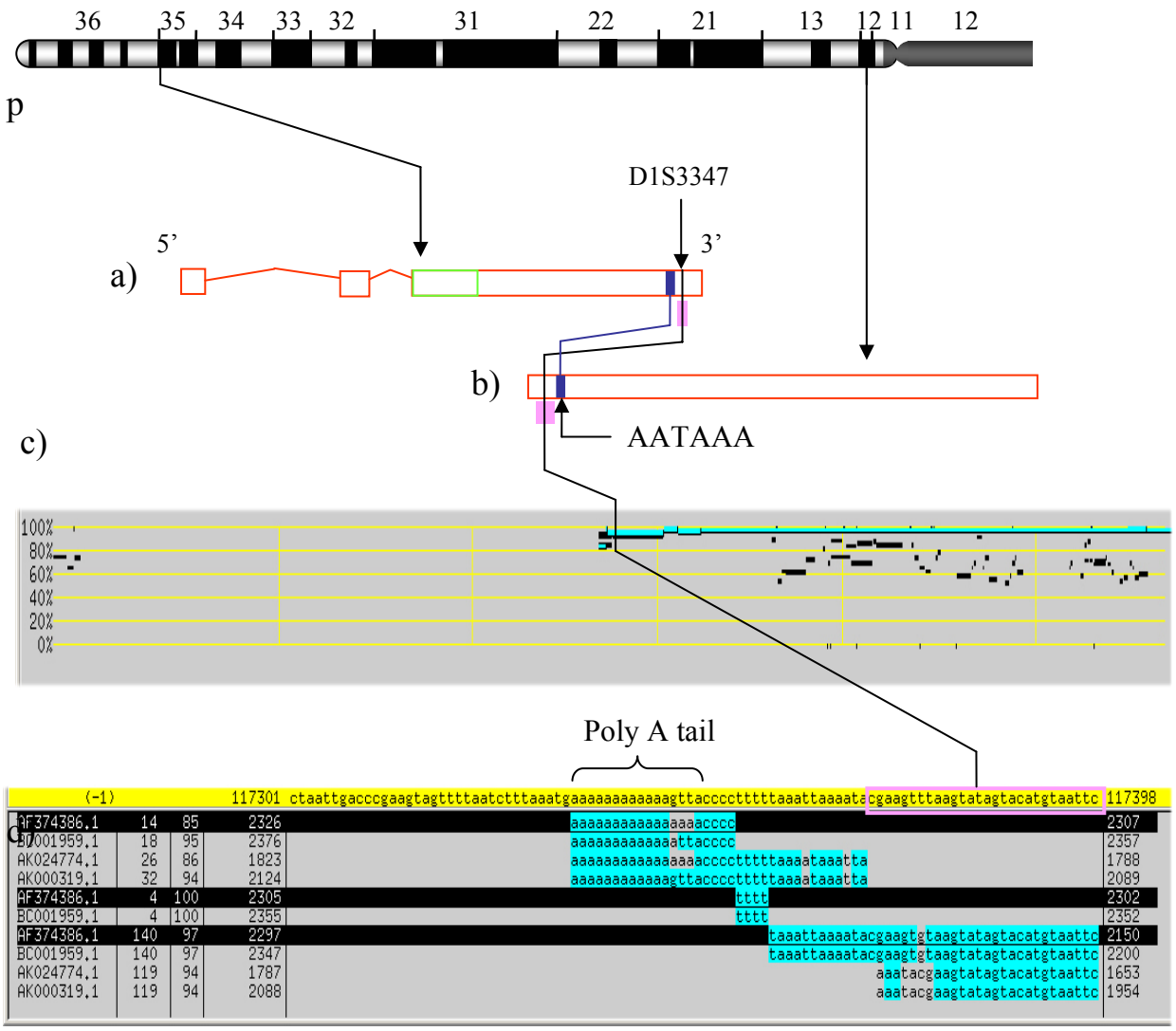


Figure 5.7: The characterisation of a processed pseudogene to 1p12, the original of which localised to 1p35. Arrows indicate where the functional gene, a), and pseudogene, b), are located on chromosome 1. D1S3347 was derived from the 3' UTR of the functional gene. Evidence that the pseudogene is processed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

5.4. Gene assessment

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

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

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

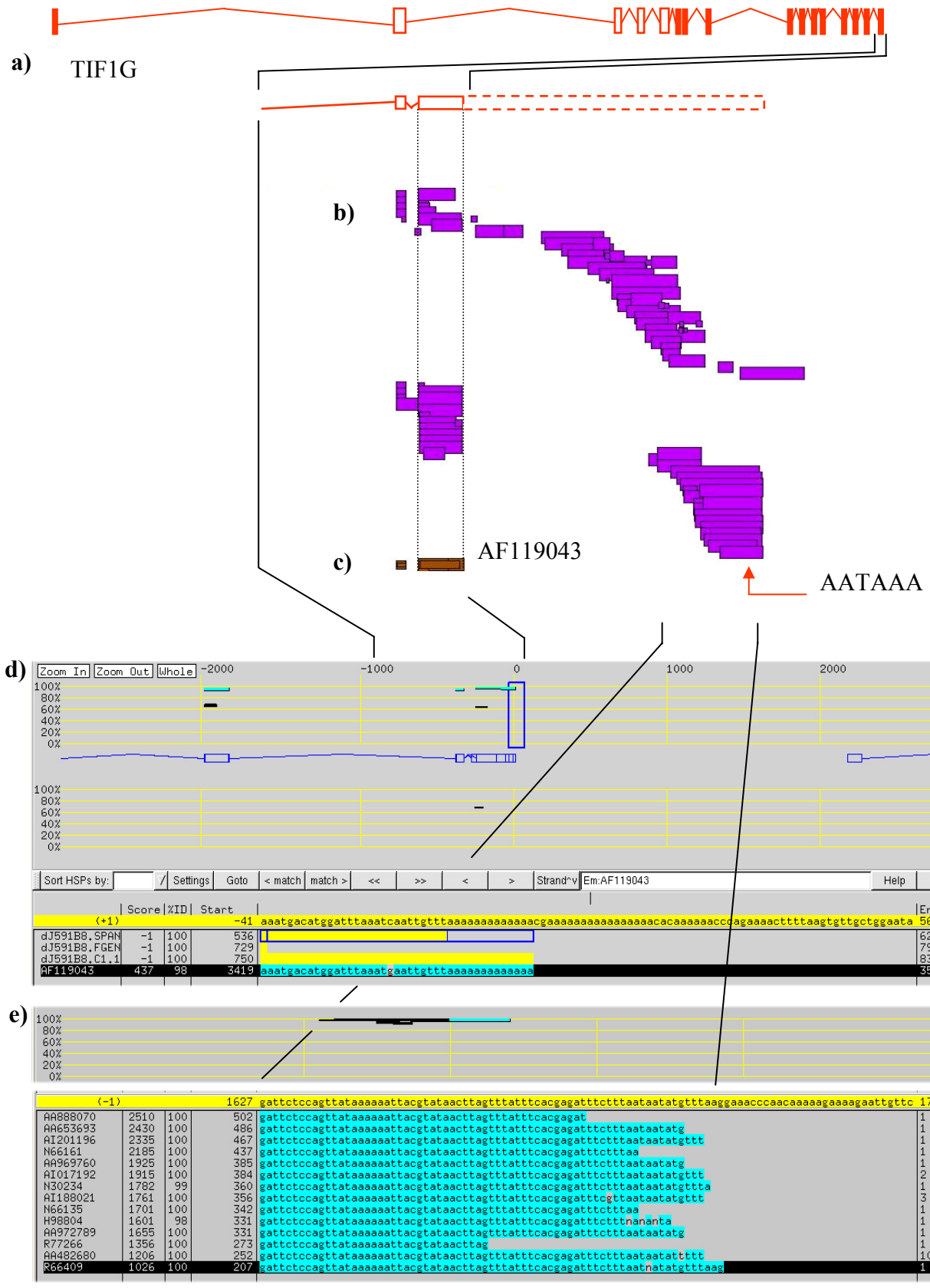

a)
TIF1G
b)
c)
AF119043
AATAAA
d)
Zoom In
Zoom Out
Whole
-2000
-1000
0
1000
2000
Sort HSPs by:
Settings
Goto
< match
match >
<<
>>
<
>
Strand^v
Em:AF119043
Help
Score
%ID
Start
End
dJ591B8.SPAN
dJ591B8.FGEN
dJ591B8.C1.1
AF119043
e)

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

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

5.4.1 Alternative splicing

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

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

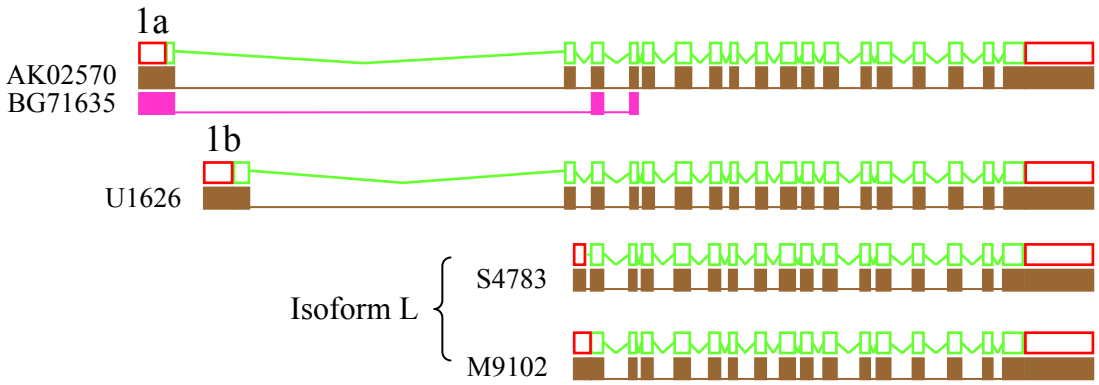


Figure 5.9: Splice variants of adenosine monophosphate deaminase 2 (AMPD2). Four splice variants were annotated (coding green boxes, UTR red boxes) by alignment of known mRNAs (brown boxes) to genomic sequence. Alignment of novel EST, BG716359 (pink box), identified a fifth potential splice variant.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

5.4.2. Genic features

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

5.4.2.1. Putative bidirectional promoters

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

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

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

5.4.2.2. Overlapping genes

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

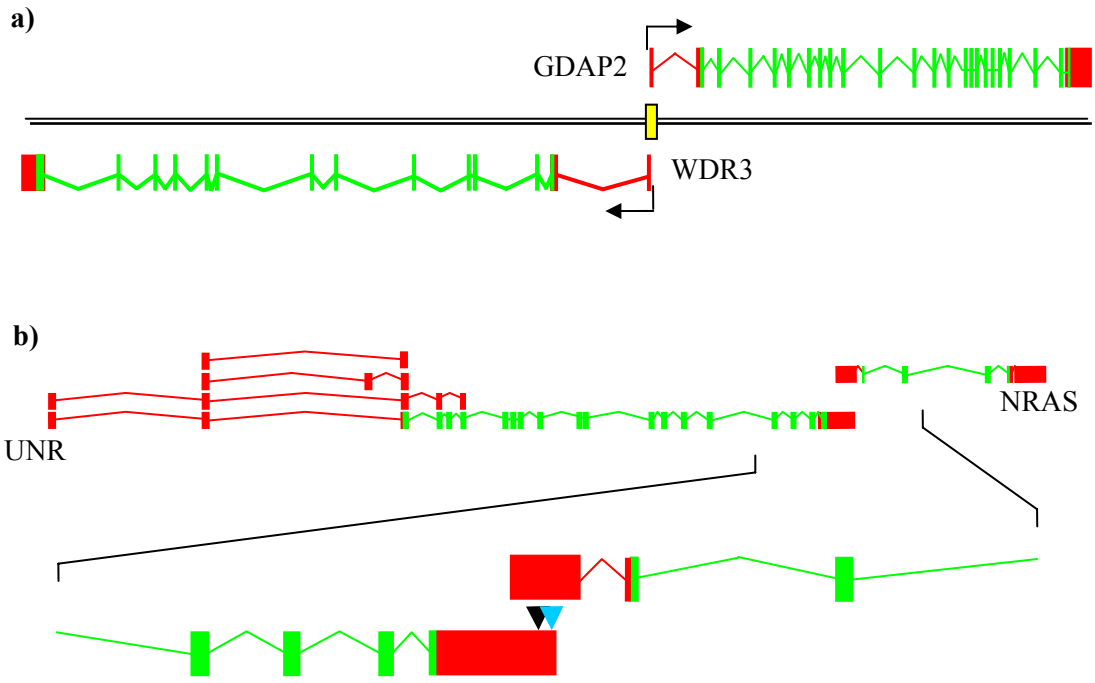


Figure 5.10: Genes in genomic context. Head to head orientation of a pair of genes (a) that share a CpG (yellow box) between the first non-coding exon (red box, coding green box) of each gene. Figure 5.10b depicts two genes, UNR and NRAS, whose 3' and 5' non-coding sequences partly overlap, respectively.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

5.5 Inferring function by protein homology

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

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

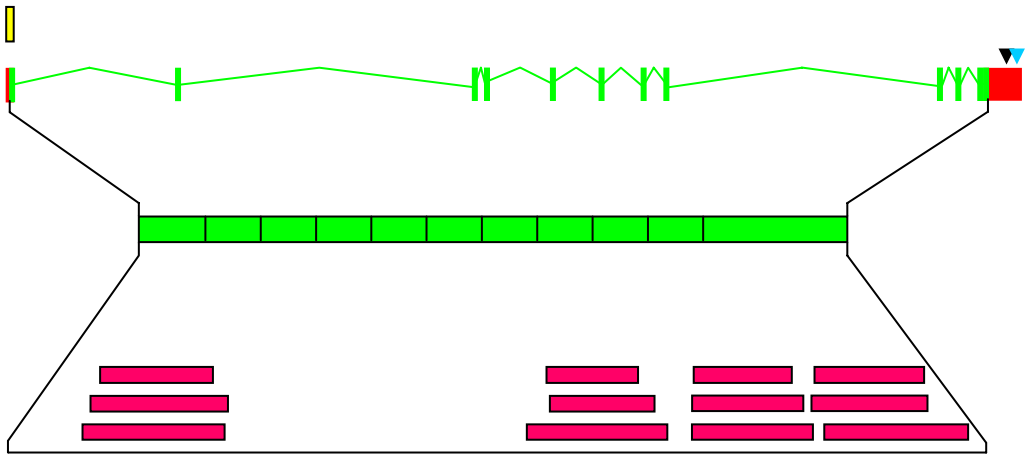
5.5.1. Identifying function through sequence homology

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

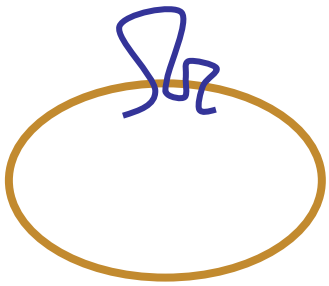
mRNA (but not aligned by annotation here) which is purported to show homology to the Drosophila Orct gene and mammalian carnitine transporters, a family of genes which are involved in transmembrane organic ion transportation.

a)

b)



c)



d)

e)

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

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

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

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

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

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

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

5.5.2. Identifying function by structural homology

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

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

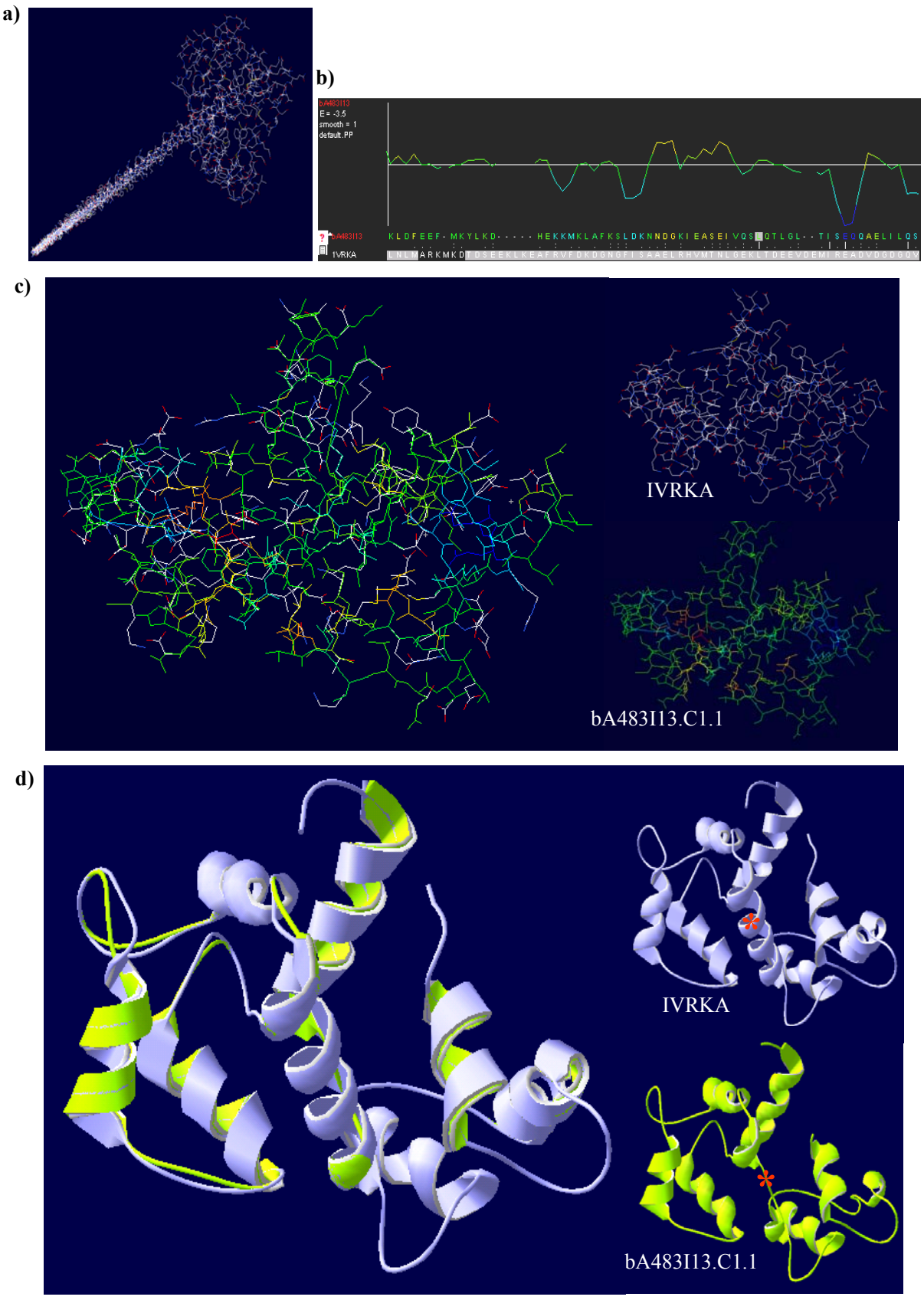

a)
b)
c)
IVRKA
bA483I13.C1.1
d)
IVRKA
bA483I13.C1.1

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

5.6. Discussion

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

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

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

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

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

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

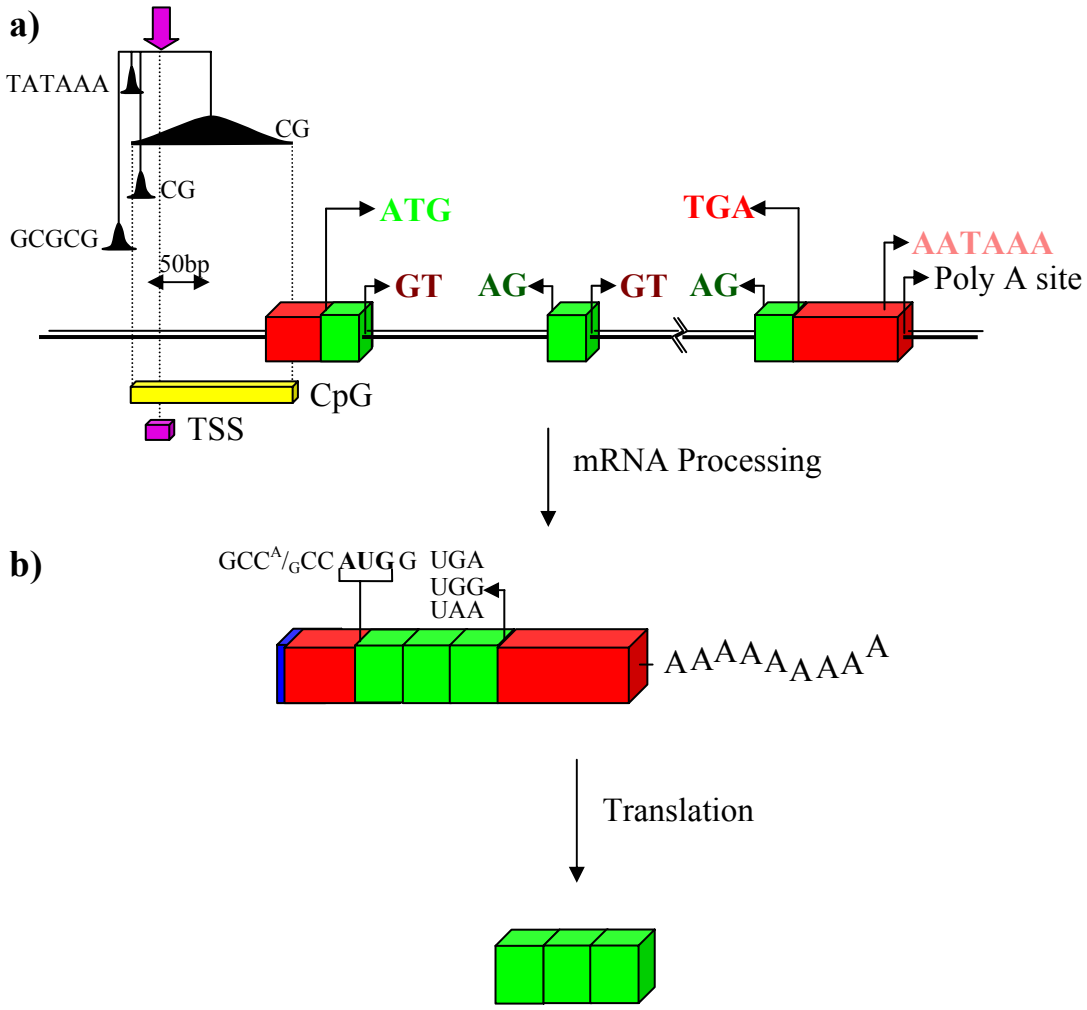


Figure 5.13: Generic structure of a gene. Figure 5.13a represents the regulatory elements (predicted CpG island and transcription start site consensus motifs (purple box)) at the 5' end of the gene, translation start (green lettering) and stop sites (red lettering), polyA signal (pink lettering) and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

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

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

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

5.7. Appendix

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

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

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

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

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

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

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