Chapter III

Sequence and transcript map of 20q12-13.2

3.1 Introduction

3.1.1 Strategies for gene identification

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

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

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

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

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

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

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

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

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

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

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

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

3.1.2 Overview

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

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

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

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

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

3.2 Sanger annotation pipeline

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

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

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

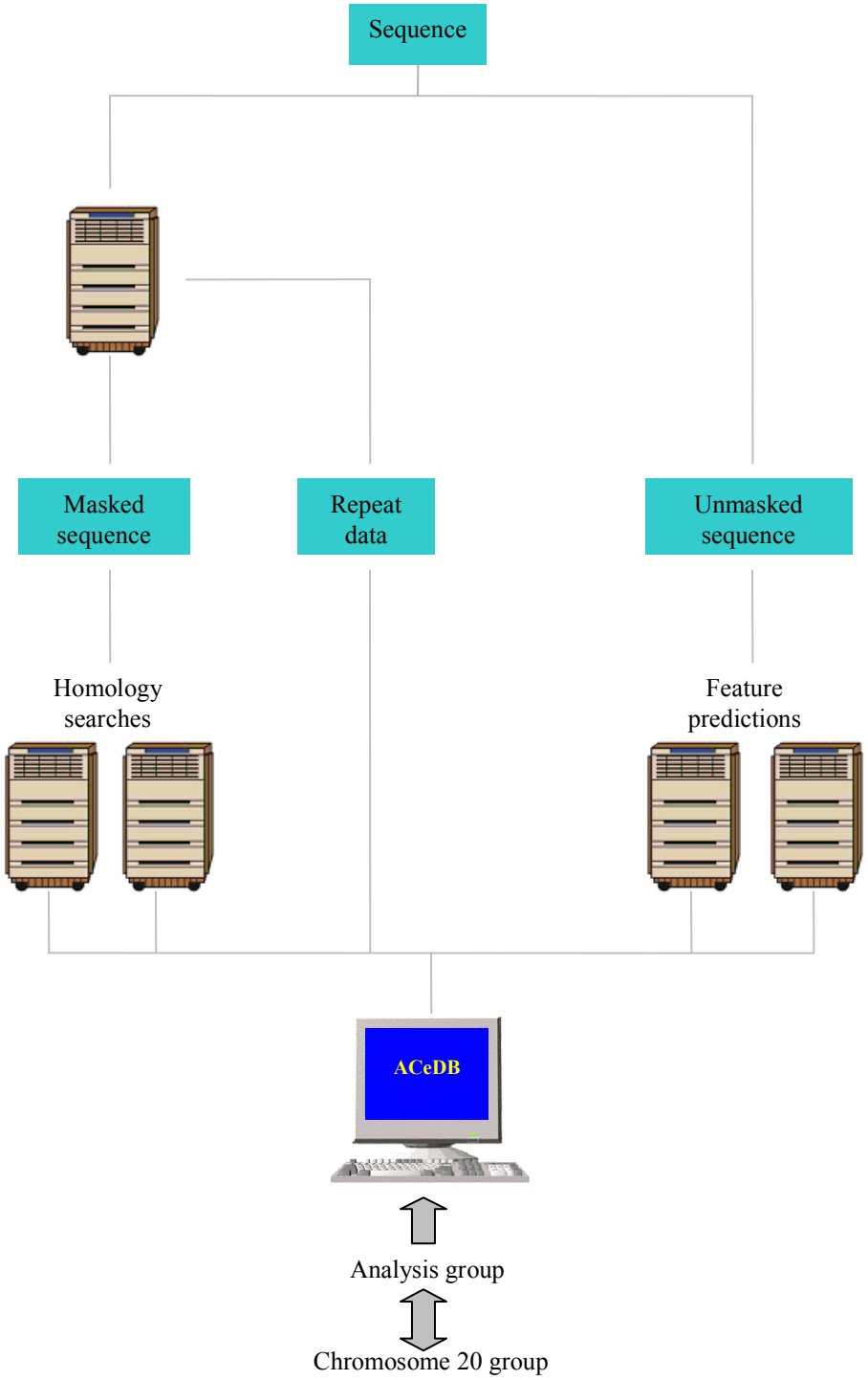


Figure 3.1: Sequence analysis pipeline. Software packages were used to predict gene structures and CpG islands in the unmasked finished sequence. RepeatMasker was used to identify repeats. The masked sequence was used to perform homology searches. All generated data was visualised in an ACeDB database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

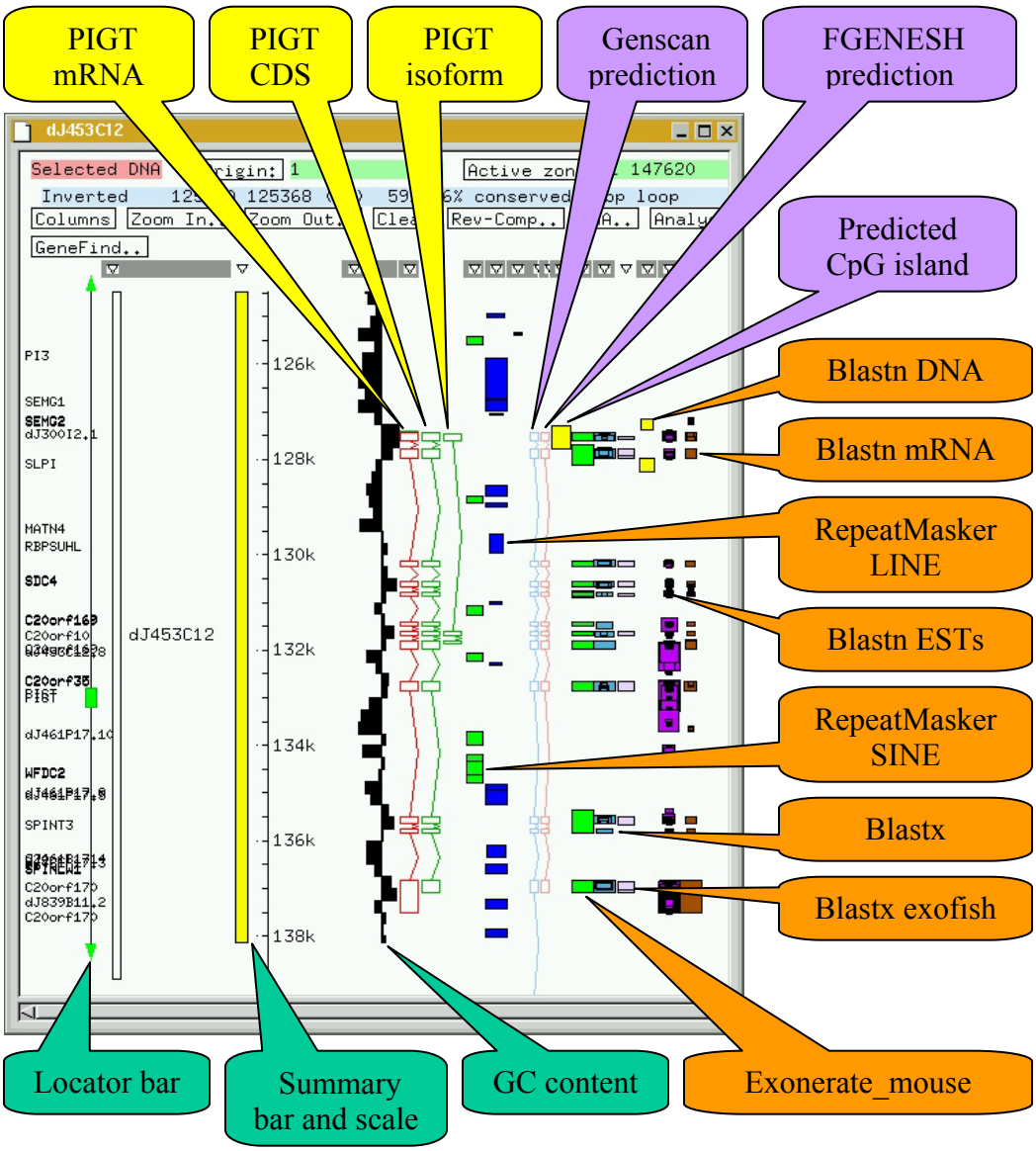


Figure 3.2: ACeDB view of sequence features. Orange description boxes indicate homologies, lavender boxes show prediction package data, green boxes show ACeDB general features and the yellow boxes indicate the manual annotation. The region shown here surrounds the PIGT gene locus.

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

3.3 Experimental confirmation of 20q12-13.2 genes

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

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

3.3.1 Vectorette

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

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

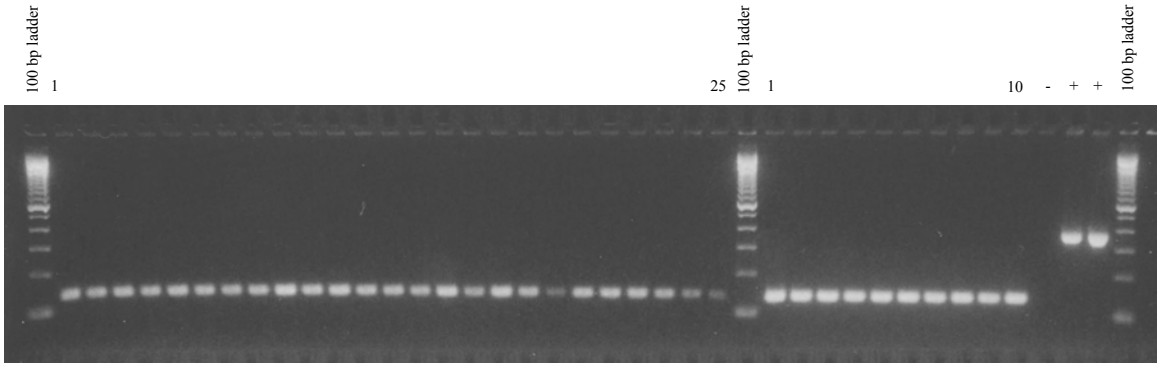


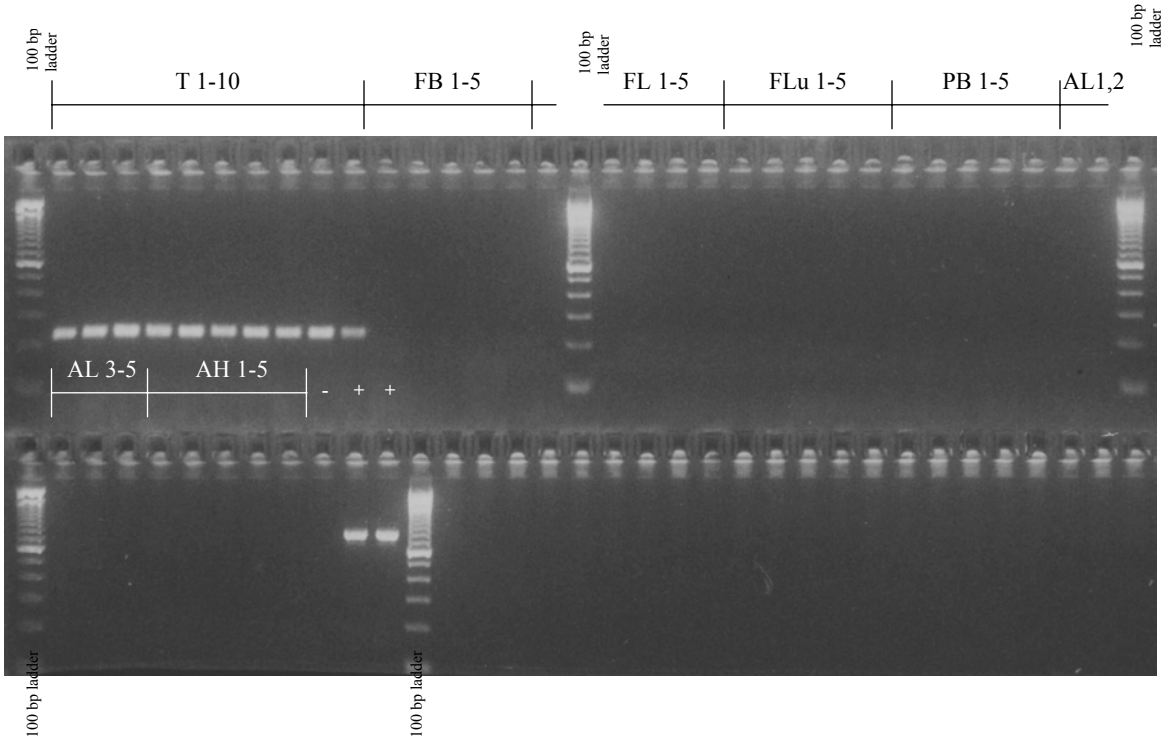
Figure 3.3: Vectorette library construction (final step PCR screening to check cDNA recovery and contamination). Newly constructed vectorette pools (25 from plated cultures and 10 from liquid cultures) generated from an Adult Heart cDNA library were PCR screened with the stSG71396 primer set. The primers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

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

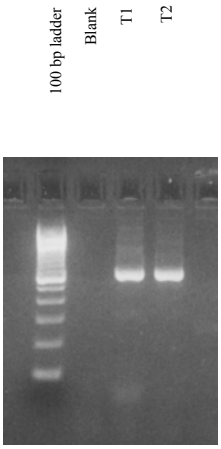
generated sequence reads provided novel expressed data and were used to confirm and extend annotation. They were also submitted to the EMBL database as ESTs.

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

A.



B.



C.

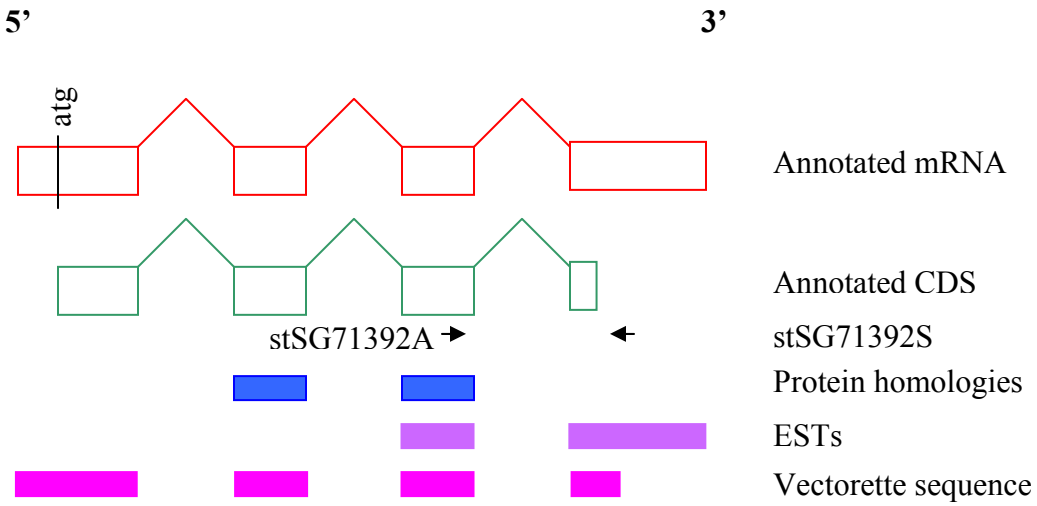


Figure 3.4: Example of cDNA-end isolation using the vectorette method. Initial annotation of SPINLW1 was based on homology with splicing ESTs (supporting two exons) and various protein homologies (supporting one EST-supported exon, and an exon further 5'). (A) the stSG71392 primer set was used to PCR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

3.3.2 Single specificity primer PCR

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

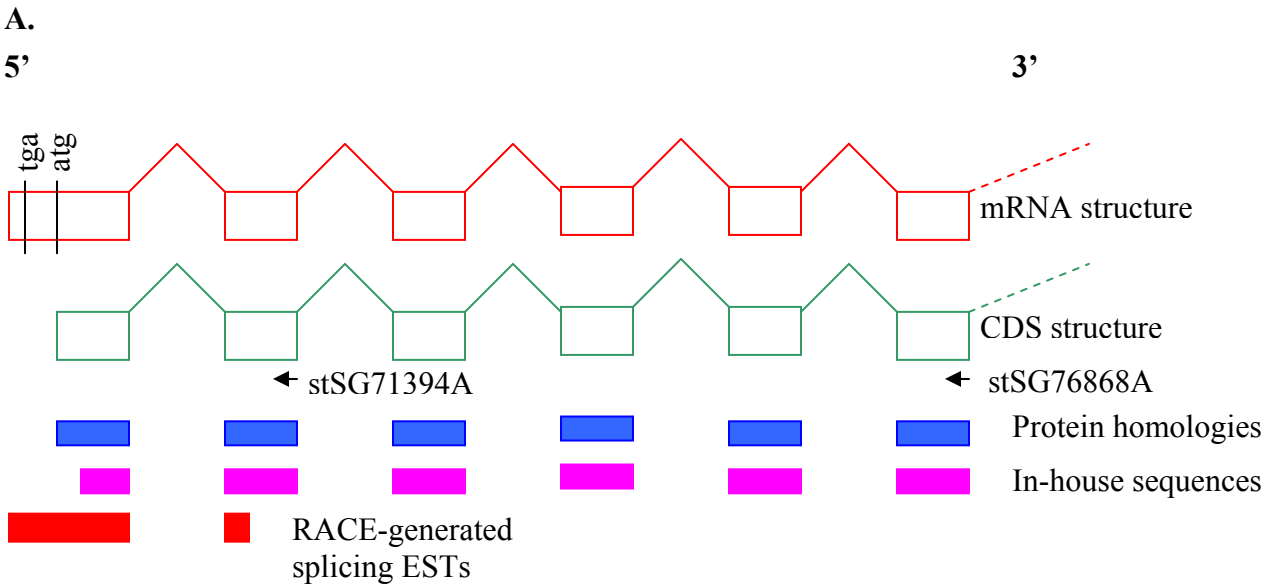
Unlike vectorette, SSP-PCR does not require specially adapted cDNA pools and can use boiled cDNA clones as templates for end-sequence isolation. To expand the available Sanger library resource I generated cDNA pools from an Adult Heart (Invitrogen) cDNA library. During this project, eighteen SSP-PCR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

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

3.3.3 RACE

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

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



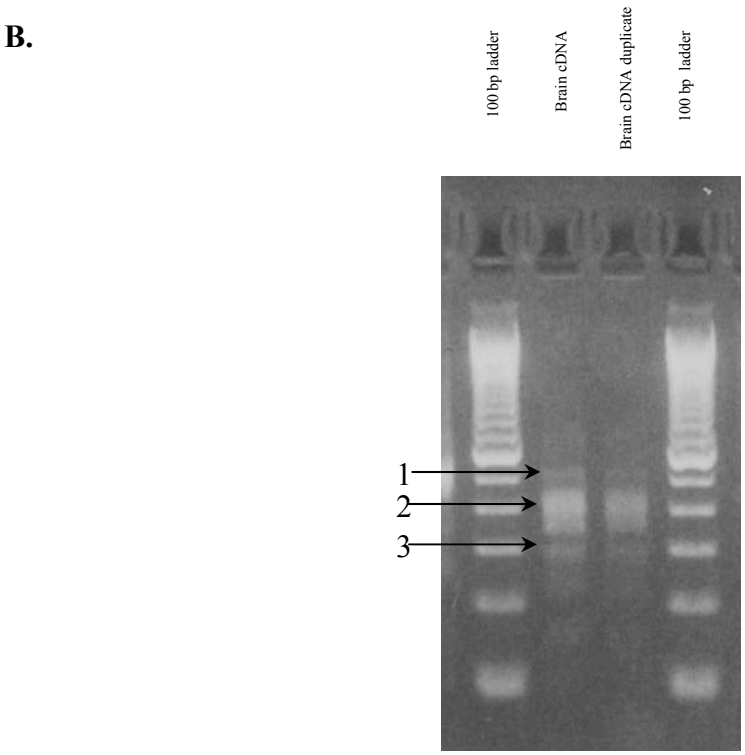


Figure 3.5 (previous page): Enhancing the annotation using RACE. (A) Figure shows the first six exons of C20orf119 (not to scale). Similarities with polyA-binding proteins from several organisms (blue boxes) were used to annotate a CDS (structure shown in green). cDNA fragments isolated from Fetal Brain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

3.3.4 Summary of experimental efforts

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

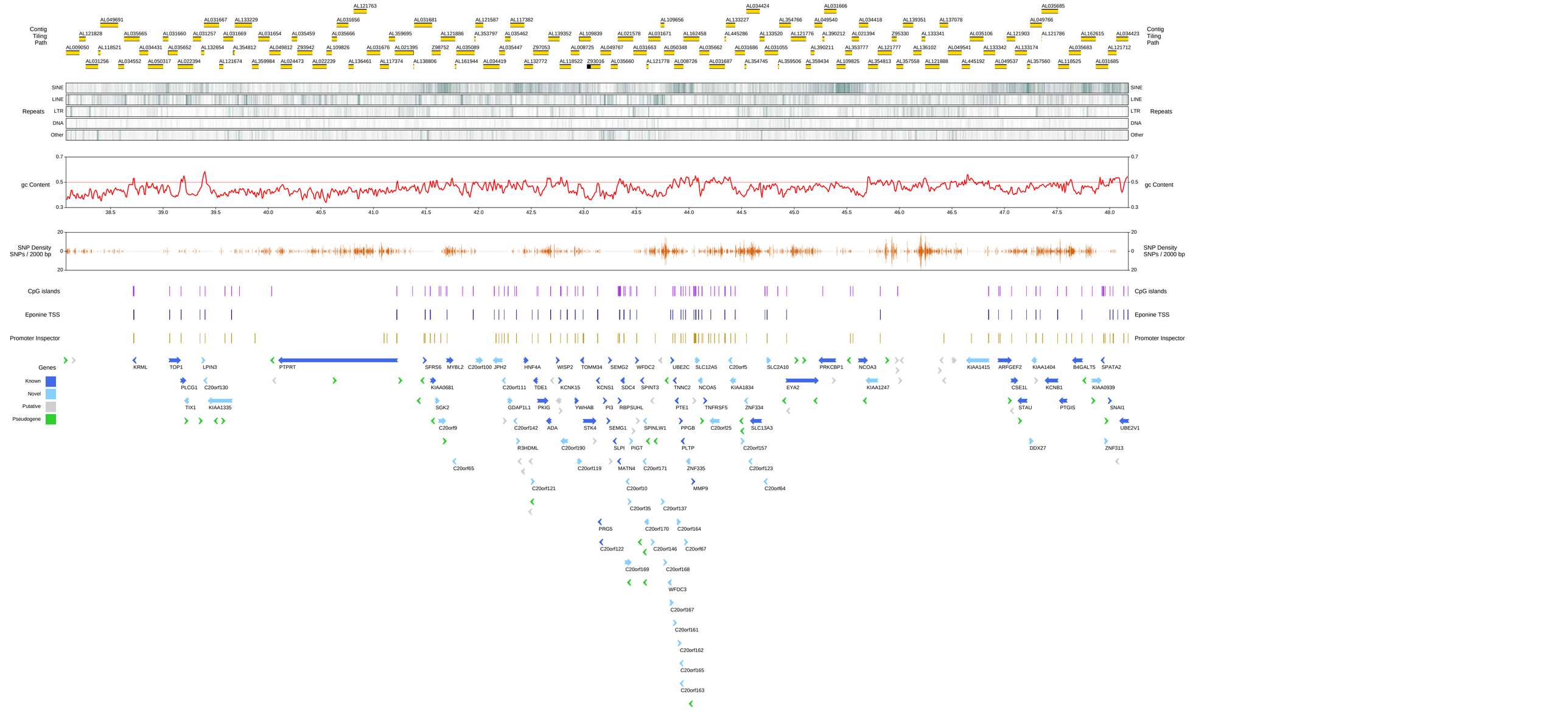
An additional 142 good sequence reads corresponding to novel genes mapping across the whole of chromosome 20 (but outside 20q12-13.2) were also isolated and subsequently submitted to EMBL (Appendix 4).

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

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

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

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


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

3.4.1 Broad genome landscape

3.4.1.1 Repeats

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

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

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

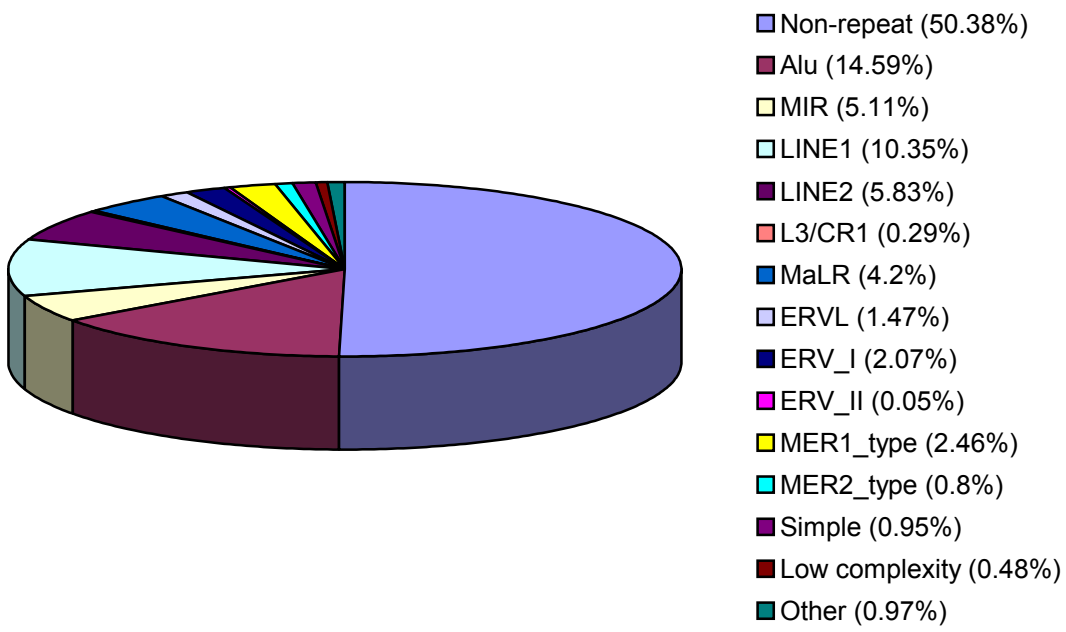


Figure 3.7: Repeat content distribution of 20q12-13.2.

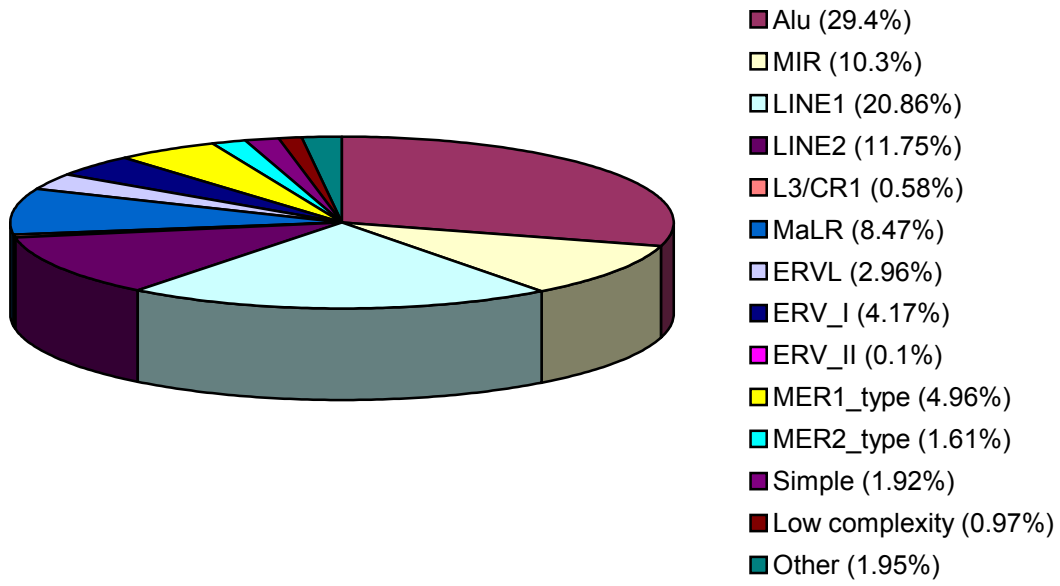


Figure 3.8: Repeat distribution for each family.

3.4.1.2 A segmental duplication

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

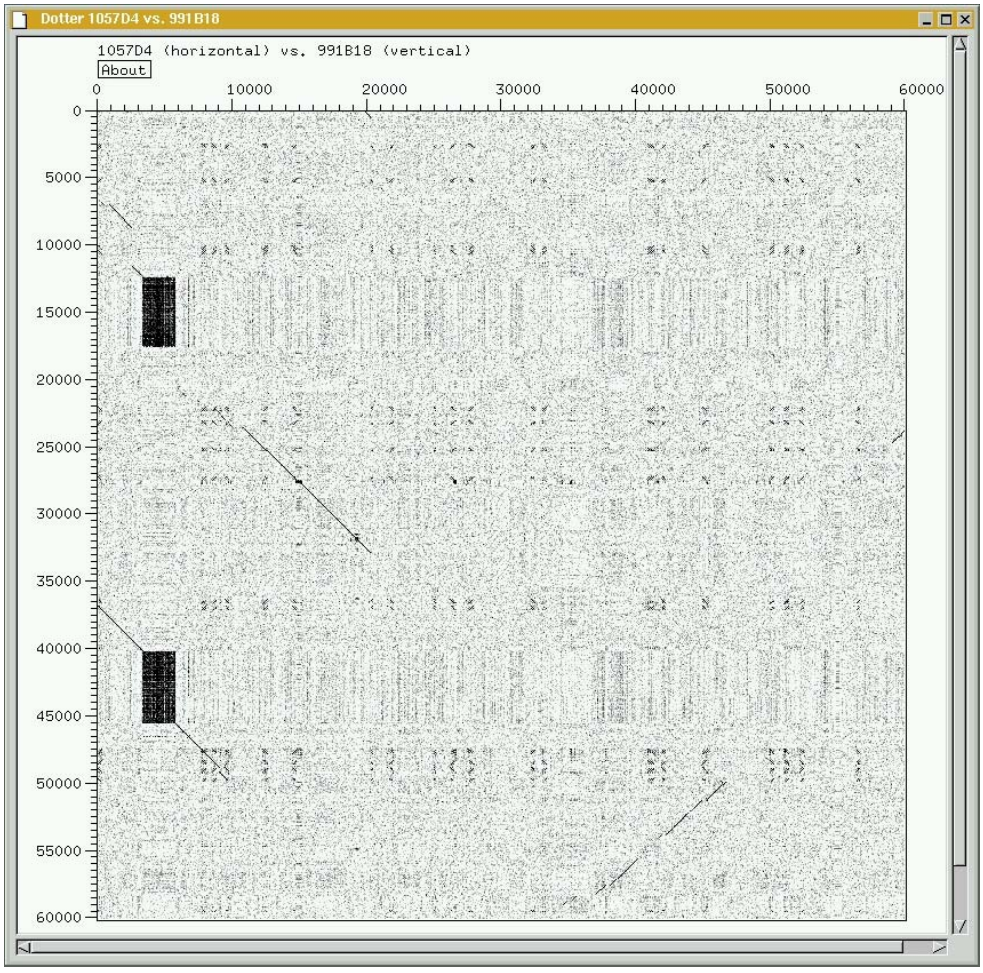


Figure 3.9: Dot plot (output from Dotter, Sonnhammer and Durbin, 1995) of the two regions present in the sequences of clones RP5-1057D4 and RP5-991B18 (sequence accession numbers AL121777 and AL049541). The non-linear output is probably due to rearrangements occurring after the duplication event.

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

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

3.4.2 Supporting evidence for annotated loci

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

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

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

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

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

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

Table 3.2: Supporting evidence for annotated features.

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

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

3.4.3 First-pass expression data for novel and putative genes

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

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

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

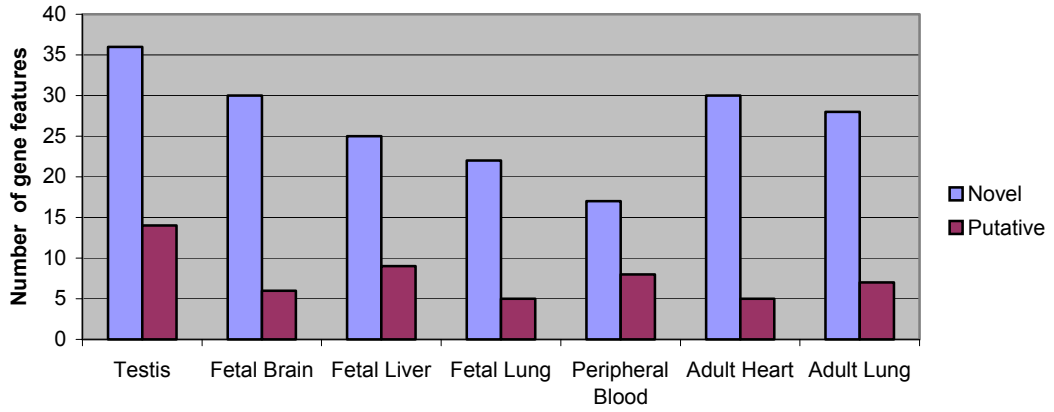


Figure 3.10: Positive genes per cDNA library.

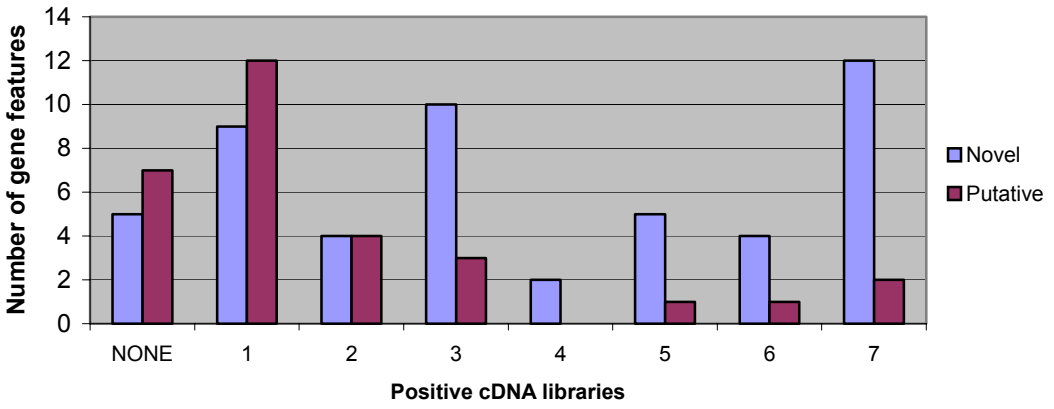


Figure 3.11: Positive cDNA libraries per gene.

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

3.4.4 Gene structures

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

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

Table 3.3: Size of gene loci.

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

Table 3.4: Structural features of annotated gene features.

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

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

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

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

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

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

3.4.5 Splice sites

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

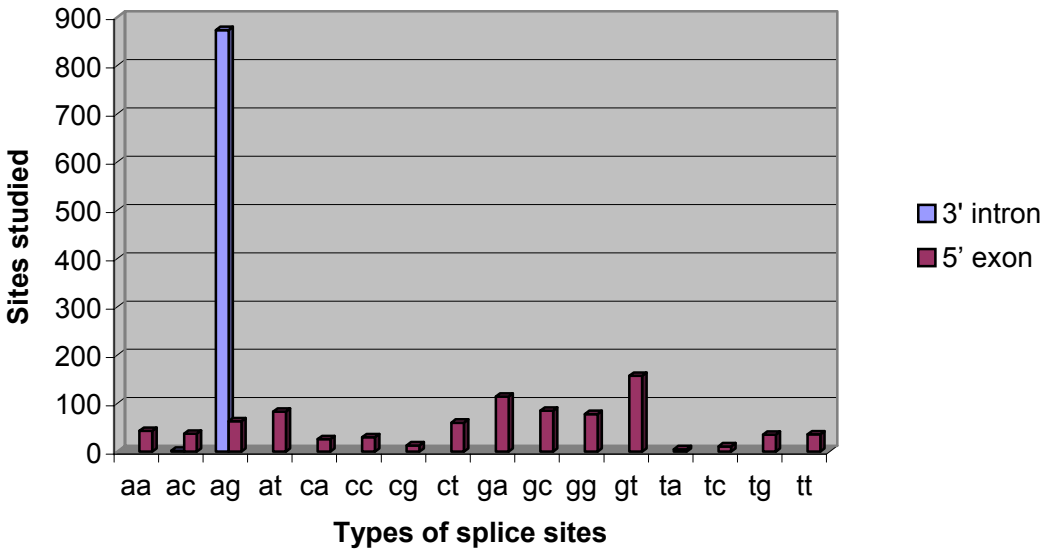


Figure 3.12: 3' intron-5' exon splice sites.

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

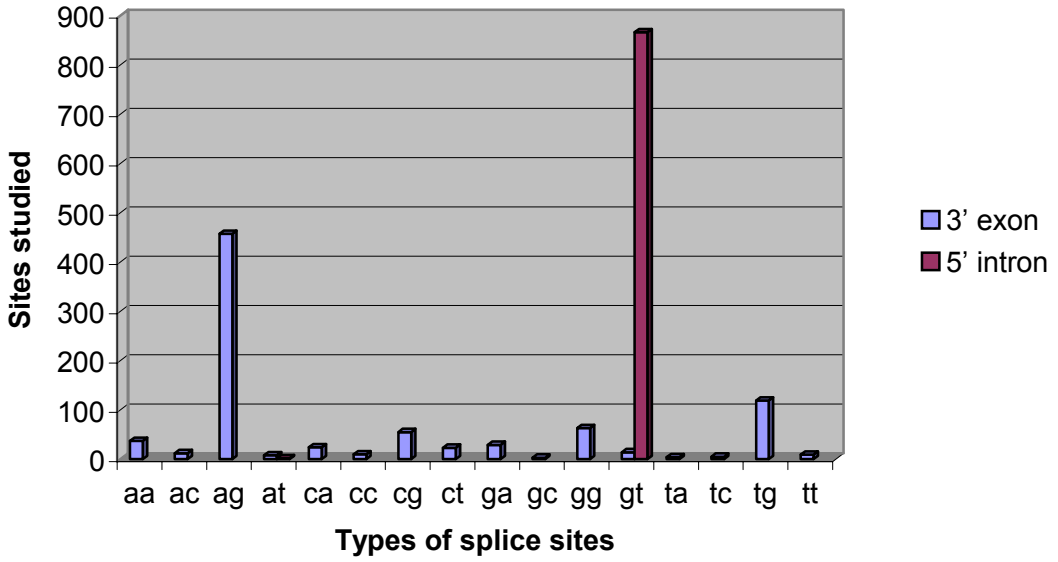


Figure 3.13: 3' exon-5' intron splice sites.

3.4.6 Splice isoforms

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

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

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

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

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

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

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

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

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

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

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

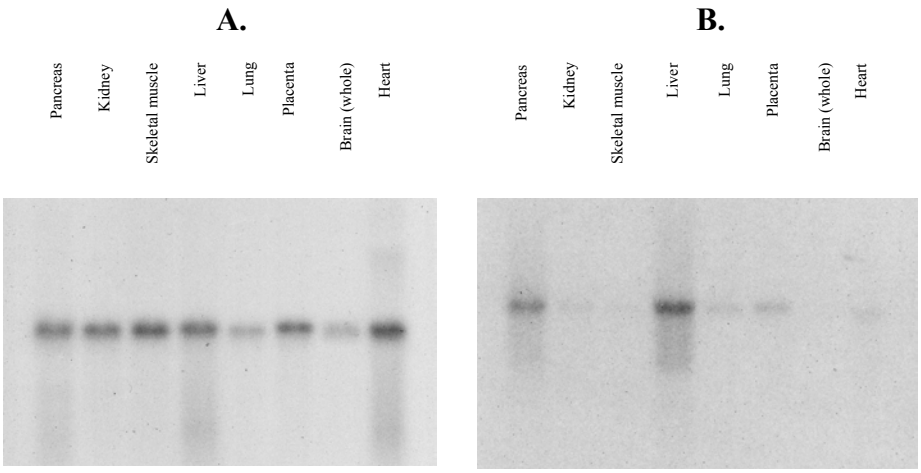


Figure 3.14: Northern blots. PCR-based probes were generated for the PIGT and SLC2A10 genes using the stSG85271 and stSG76895 primers respectively. The probes were radioactively labelled and used in hybridisation experiments with commercial Northern blots containing polyA RNA from eight tissues (A and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

3.5.1 Polyadenylation signals (3' end)

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

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

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

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

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

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

3.5.2 Promoters (5' end)

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

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

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

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

3.5.2.1 Correlation of predictions and all gene structures

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

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

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

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

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

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

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

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

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

3.5.2.2. Focusing on coding genes

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

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

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

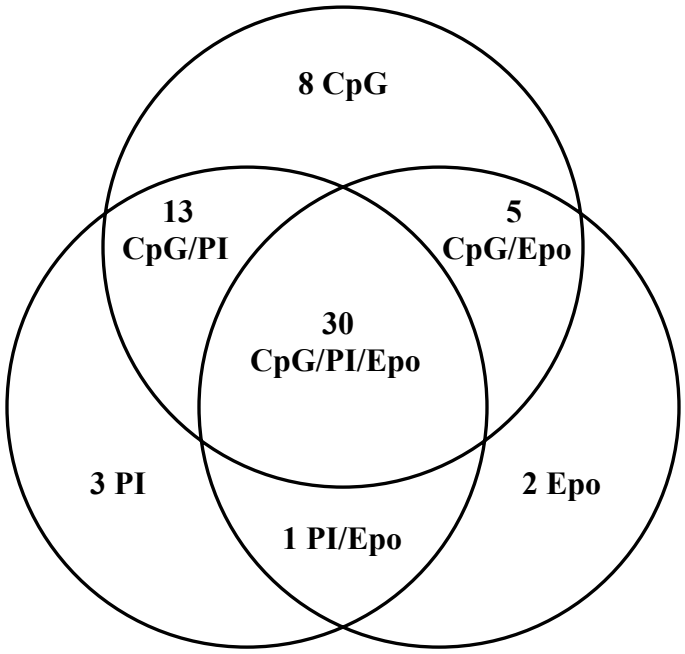


Figure 3.15: Correlation of genes with predicted promoters at their 5' end and predictions by the three methods.

3.5.3 Summary

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

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

3.6 Measuring completion of annotation

3.6.1 Homology searches

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

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

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

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

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

3.6.2 Genscan and FGENESH

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

Table 3.9: Analysis of predicted coding sequence.

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

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

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

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

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

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

3.7 Protein analysis

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

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

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

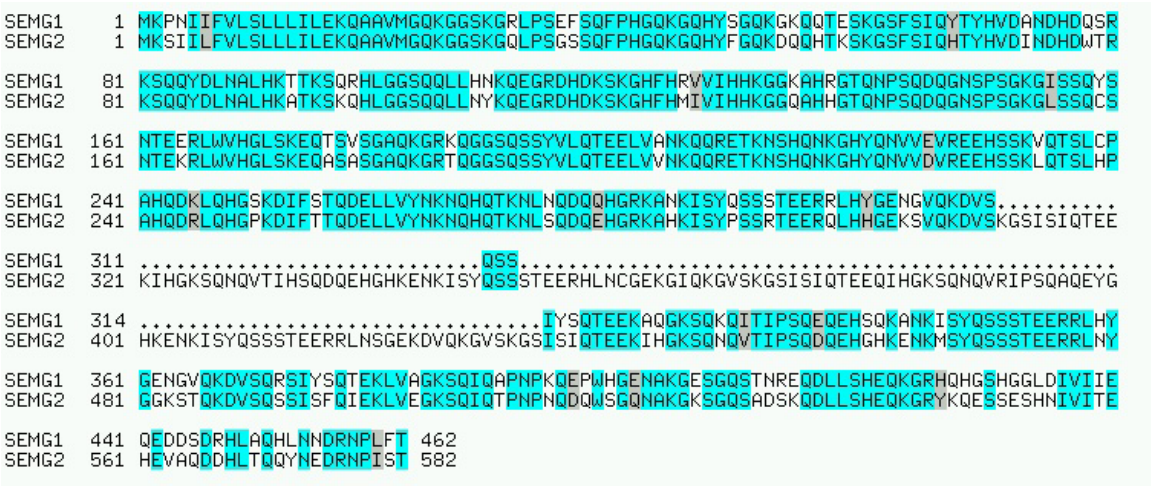


Figure 3.16: SEMG1 and SEMG2 protein alignment. The protein sequences were aligned using CLUSTAL W (Thompson *et al*., 1994) and the output was formatted using Belvu (Sonnhammer, unpublished). Identical aa are highlighted blue and similar, grey.

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

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

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

3.8 Discussion

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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