ELSEVIER

Gene 247 (2000) 215–232

GENE
AN INTERNATIONAL JOURNAL ON GENES, GENOMES AND EVOLUTION

www.elsevier.com/locate/gene

Criteria for gene identification and features of genome organization: analysis of 6.5 Mb of DNA sequence from human chromosome 21

Dobromir Slavov [a], , Masahira Hattori [b], Yoshiyuki Sakaki [b], André Rosenthal [c], Nobuyoshi Shimizu [d], Shinsei Minoshima [d], Jun Kudoh [d], Marie-Laure Yaspo [e], Juliane Ramser [e], Richard Reinhardt [e], Candy Reimer [a], Kevin Clancy [a], Alla Rynditch [a,1], Katheleen Gardiner [a,*]

[a] *Eleanor Roosevelt Institute, 1899 Gaylord Street, Denver, CO 80206, USA*
[b] *Riken/Kitasato University and the Human Genome Research Group, Genome Sciences Center, Riken, Wako, Saitama, Japan*
[c] *Institute of Molecular Biotechnology Jena and the Jena Sequencing Center, D-07745 Jena, Germany*
[d] *Department of Molecular Biology, Keio University School of Medicine, Tokyo, Japan 160-8582*
[e] *Keio University Sequencing Center; Max-Planck-Institute for Molecular genetics, Berlin, Germany*

Received 18 November 1999; received in revised form 26 January 2000; accepted 9 February 2000

Abstract

To establish criteria for and the limitations of novel gene identification, to identify novel genes of potential relevance to Down Syndrome and to investigate features of genome organization, 6.5 Mb of DNA sequence, dispersed throughout the long arm of human chromosome 21, have been annotated computationally and experimentally. Exon prediction with four programs, protein and EST database searches, two-sequence BLAST searches and CpG island characterization identified 41 genes with known or new protein homologies. Features of these genes suggested criteria for prediction of novel genes (those lacking any protein homology) with the following characteristics: (1) exon+EST genes: genes with excellent patterns of predicted exons and one or more matches in dbEST; (2) exon-EST genes: genes with good patterns of predicted exons and no matches in dbEST; (3) EST-exon genes: genes without any patterns of reliable exon prediction but with matches in dbEST; and (4) isolated CpG island genes: genes consisting of strong CpG islands that are apparently unique sequences and found in regions lacking any consistent exon predictions within >50 kb. In total, 41 novel gene models were predicted, and for a subset of these, RT-PCR experiments helped to verify and refine the models, and were used to assess expression in early development and in adult brain regions of potential relevance to Down syndrome. Results suggest generally low and/or restricted patterns of expression, and also reveal examples of complex alternative processing, especially in brain, that may have important implications for regulation of protein function. Analysis of complete gene structures of the known genes identified a number of very large introns, a number of very short intergenic distances, and at least one potentially bi-directional promoter. At least 3/4 of known genes and 1/2 of predicted genes are associated with CpG islands. For novel genes, three cases of overlapping genes are predicted. Results of these analyses illustrate some of the complexities inherent in mammalian genome organization and some of the limitations of current sequence analysis technologies. They also doubled the number of potential genes within the region.

Keywords: Down syndrome; Gene identification; Genome organization; Human chromosome 21; Sequence analysis

1. Introduction

The rapid accumulation of large-scale human genomic sequence provides new opportunities for gene identification that can complement the established experimental approaches. Computationally based analysis of genomic DNA sequence currently includes, among other features, use of exon prediction programs, BLASTX searches of protein databases, BLASTN and TBLASTX searches of the expressed sequence tag data-

Abbreviations: dbEST, database of expresed sequence tags; kb, kilobase pairs; Mb, megabase pairs; RT-PCR, reverse transcription polymerase chain reaction; UTR, untranslated region.

* Corresponding author. Tel.: +1-303-336-5652; fax: +1-303-333-8423.

E-mail address: gardiner@eri.uchsc.edu (K. Gardiner)

[1] Present address: Institute of Molecular Biology and Genetics, National Academy of Science of Ukraine, Kiev, Ukraine.

0378-1119/00/$ - see front matter
PII: S0378-1119(00)00089-5

bases (dbEST), and CpG island predictions. While each of these approaches has a high rate of success, each also has its limitations. Exon prediction programs have non-zero false positive and false negative rates, and have also been shown to perform less well on novel sequences than on test sequences (Burset and Guigo, 1996; Claverie, 1997). Truly novel protein sequences, of course, will have no matches in the protein databases, and while dbEST continues to expand in complexity, it remains incomplete. In particular for human genes, those that are developmentally specific or restricted in expression are likely to be absent from dbEST. Although rat and mouse dbESTs are more likely to contain these genes, the short sequence lengths and their confinement largely to 3′ untranslated regions will make matches to human genes often undetectable. Lastly, CpG islands, while most often located at the 5′ ends of genes, have also been documented upstream of transcription, within the coding regions and at the 3′ ends of genes (Larsen et al., 1992), and thus cannot be assumed to mark the 5′ end of a gene. Together, these limitations have several important implications for the analysis of large genomic sequence. First, gene identification is not completely unambiguous; in particular, gene predictions must be coupled to experimental verification. Second, because gene identification is the critical issue in beginning to determine the causes and cures of human disease, being able to find 'all genes' within a chromosomal segment is of interest. It is, therefore, useful to attempt an evaluation of the overall success rate of gene finding. Third, understanding organizational and regulatory features is further required for functional assessment.

Here, we report the analysis of 6.5 Mb of genomic sequence from human chromosome 21. The aims in this analysis were threefold: to use results from a set of sequence analysis tools to establish criteria for and to estimate the limitations of novel gene identification and experimental verification, to identify genes within 21q with potential relevance to Down syndrome or other chromosome 21 diseases or biological features, and to investigate features of human genome sequence organization relevant to function and/or regulation. This analysis revealed 38 genes already known to map to human chromosome 21 and an additional three genes representing new homologues of known mouse or human genes. Of these 41 genes, 38 would have been recognized by exon prediction in the absence of any protein homologies. Within the same 6.5 Mb of DNA, 41 novel genes are also predicted, based on patterns of predicted exons, one or more entries in dbEST, and/or CpG islands. Lastly, features of genome organization were documented, including intron/exon structures of the known genes, measurement of intergenic distances, examples of apparently overlapping genes, and gene density relative to base composition. Results of these analyses have general applicability to any segment of the human genome, as well as particular relevance to chromosome 21.

2. Methods

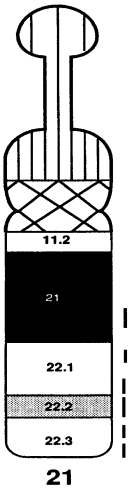
2.1. Sequence analysis

Human chromosome 21 sequences were generated at centers at the Riken/Kitasato University and Keio University in Japan, and at the Institute for Molecular Biology, Jena and the Max Planck Institute for Molecular Genetics, Berlin in Germany. Sequence segments are deposited in GenBank and at the International Chromosome 21 Sequencing Consortium Database maintained at the Eleanor Roosevelt Institute (ERI) (www-eri.uchsc.edu). Sequence segment numbers used here refer to the numbers assigned at the consortium site. The approximate map location of each segment is shown in Fig. 1, along with the size of each segment and examples of known genes contained within each. The Genotator tool (Harris, 1997) was used to produce graphical representations of exon prediction with programs Grail, Genscan, Genefinder and Xpound. For Genotator analysis, each sequence file is divided into 90 kb segments (designated as segment #-1, -2, etc). In this analysis, a consistent exon is defined as one that is predicted by at least Genscan and Grail programs.

BLASTX searches of the non-redundant protein database and BLASTN and TBLASTX searches of dbEST were run with individual exons from Genotator analysis and with genomic segments of interest (Altschul et al., 1997). The CpG islands in each Genotator segment were identified using the tool at Oak Ridge National Lab (http://avalon.epm.ornl.gov/Grail-bin/EmptyGrailForm); BLAST searches of each predicted island were conducted to eliminate the approximately 40% that were Alu sequences. Base composition was determined for 50–100 kb intervals using information in GenBank entries or the Composition program in the GCG sequence analysis package. Complete intron/exon structures of each known gene were determined using the Two Sequence BLAST program (Tatusova and Madden, 1999) at NCBI, comparing a complete cDNA sequence retrieved from GenBank and the appropriate genomic sequence segment. Graphical annotation of all Genotator segments, including locations and sequences of primers and complete sequences of ESTs and RT-PCR products of gene models, are available at http://www-eri.uchsc.edu.

2.2. Expressed sequence tag (EST) clones

IMAGE consortium clones identified in BLASTN searches of dbEST were purchased from Research Genetics. Inserts from single colonies were completely



Size	HC21000#	Sample Gene	# of Genes Known	# of Genes Predicted	Isochore
300 kb	1	APP	1	0	L
249,545 kb	95,94	SOD1,KIAA0539	4	8	L,H1
1050,215,708 kb	93,96,97	IFNAR,ATP50,AML1	16	18	L,H1,H2
959,280,1450 kb	86,100,102	SIM2,TRPD,ERG	9	13	L,H1
137 kb	101	ETS2	1	0	H1
87,157,330 kb	103,104,171	PFKL,PWP2,SMT3A	10	2	H3
			41	41	

Fig. 1. Sequence segments analyzed. A schematic of chromosome 21 at the 450 band stage is shown. Short vertical lines adjacent to the chromosome indicate the approximate locations of the sequences analyzed. Accession Nos in the ERI web site (www-eri.uchsc.edu) for each sequence are given as HC21000 numbers. Size in kb and examples of genes within each sequence are given for reference.

sequenced using vector primers and primer walking, as necessary, and the ABI prism dideoxy-dye terminator protocol; sequences were resolved on an ABI373a.

2.3. RT-PCR, Northern analysis and hybridization

Poly(A)$^+$ RNA from Hela cells, from 8 and 10 week fetus, and from adult brain regions obtained at autopsy was prepared by standard means (Chomzynski and Sacchi, 1987). Total RNA from adult human brain, cerebellum and placenta, and fetal brain was purchased from Clontech. Thermo-Script reverse transcriptase (BRL) was used at 60°C with random hexamer primers to produce cDNA. For PCR experiments, for each gene model, primers were designed for a subset of exons that were located within EST matches or that were predicted by at least Genscan and Grail programs. If exon sequences from the different programs did not coincide completely, primers were designed within the common region. Primers were also generally not designed within 20 nucleotides of the predicted exon boundaries because these are sometimes inaccurate. Inter-exon RT-PCR with 30–40 cycles was attempted first with RNA from Hela and adult brain and fetal brain; if these were negative, 8 and 10 week whole fetus were examined. If no product was observed in gels stained with ethidium bromide, gels were transferred to HybondN+ membranes and hybridized with corresponding single exons amplified from genomic DNA or with EST clone inserts, as appropriate. If these failed, 60-cycle RT-PCR was examined similarly, and then intra-exon RT-PCR, both followed by hybridization. Additional tissues and adult brain regions were examined with a subset of genes.

Northerns containing human fetal and adult tissue poly(A)$^+$ RNAs were purchased from Clontech. Probes (>600 bp in size) were either obtained by appropriate RT-PCR as described above or were generated from EST clones. Probes were labeled by random hexamer priming (Feinberg and Vogelstein, 1984); hybridizations were carried out either in Express Hyb (Clontech) for 1–18 h at 68°C or using a standard 50% formamide SSPE buffer at 42°C overnight (Sambrook et al., 1989). After washing, filters were exposed overnight and analyzed on a phosphoimager (Molecular Dynamics).

3. Results

3.1. Exon prediction of known genes

Fig. 1 shows the approximate locations of the sequence segments analyzed and examples of known genes contained within each. Fig. 2a shows the Genotator exon prediction graphical output for the 59% GC, 87 kb segment 103. On the forward (upper) strand, there is a strikingly consistent exon prediction with all four programs in the region 25–70 kb. On the reverse strand, all but Genefinder predict a set of exons between 70 and 80 kb. BLASTX searches, as indicated in the figure, reveal identities with the APECED autoimmune disease gene (Accession No. Z97990), the Phospho-FructoKinase Liver type (PFKL) gene (X15573), and the gene of unknown function, c21orf2 (Y11392). It is of interest to note that the distances between the coding regions of APECED and PFKL and between the coding regions of PFKL and c21orf2 are each less than 5 kb, while there are no genes or consistent exons predicted in the 25 kb segment 5′ to the APECED gene.

Genotator analysis of a contrasting region, segments 96-1 and 96-2 of 39% and 46% GC, is shown in Fig. 2b and c. Throughout 96-1 and in 96-2 up to 55 kb, consistent exon predictions are seen largely with

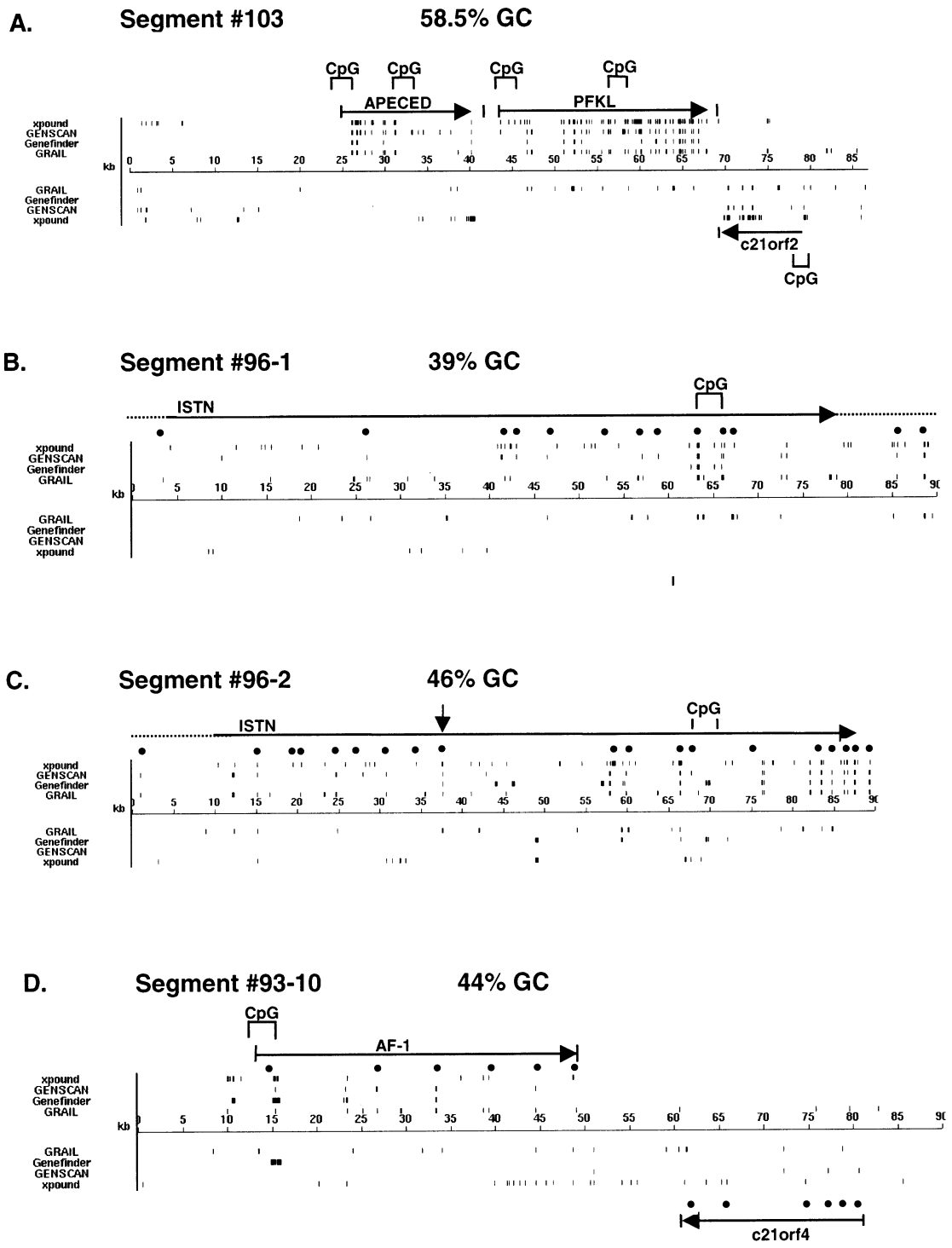


Fig. 2. Genotator graphical output from four exon prediction programs, Xpound, Genscan, Genefinder and Grail, as indicated at the left of each segment, is shown. For simplicity of presentation, other annotations possible with Genotator (Harris, 1997) (GenPept and EST matches, repeat sequences, promoters and ORFs) are not shown. In each frame, a segment up to 90 kb in size has been analyzed. Arrows indicate the 5′ and 3′ ends of the known genes. (**a**) sequence segment 103, containing the autoimmune disease gene, APECED, the phosphofructokinase gene, PFKL, and the anonymous gene, c21orf2. (**b**) and (**c**) segments 96-1 and 96-2 with a similar analysis of the Intersectin gene. (**d**) segment 93-10 with the AF-1 and c21orf4 gene exons predictions. '·' indicates the locations of exons identified by Two Sequence BLAST analysis between the complete cDNA and the genomic segment.

Genscan and Grail and sometimes with Xpound, while in 96-2 from 55-90 kb, all four exon prediction programs are highly consistent. In the course of this work, BLASTX searches revealed homologies throughout both segments with two transcripts of the Intersectin gene (Guipponi et al., 1998). The most 3′ predicted exons (>55 kb in 96-2) are specific to the longer, brain specific, Intersectin transcript.

Fig. 2d shows the Genotator output of segment 93-10, a 45% GC region, but where exon prediction is remarkably less convincing. BLASTX searches revealed the AF-1 gene (U05875) on the forward strand and the anonymous cDNA c21orf4 (AF045606) on the reverse strand. In both cases, several exons were predicted by only one program or were missed completely.

A similar analysis of all 13 segments comprising the complete 6.5 Mb revealed a total of 38 genes already known to map to human chromosome 21. Names and locations within the Genotator files of each of these are included in Table 1.

3.2. New homologies

Three additional gene models showed new homologies. One example is located in sequence segment 94-6-85R. The B9 gene model is represented by a pattern of seven consistent exons each of which shows homology to a mouse putative serine-threonine kinase, Mak-V (AF055919). The murine Mak-V gene was isolated from highly metastatic tumors, where it showed increased expression levels relative to normal tissues (Korobko et al., 1997). Further analysis of 94-6 and the adjacent sequence segment in comparison with the Mak-V cDNA provides convincing evidence that B9 is a human homologue and not a pseudogene. Ten coding exons in B9, each with consensus splice sites, are identified in a Two Sequence BLAST search between the mouse Mak-V cDNA and the 94 genomic sequence. The open reading frame of 715 amino acids averages >90% identity and >95% similarity to the mouse gene.

In segment 97-6-68F, the B35 gene model consists of five Genscan+Grail exons with significant similarity to several mammalian chloride channel proteins, including the human channels, CLIC2 (AF097330) and CLIC4 (Y12696), both of 253 amino acids. The five exons of the B35 model, each with consensus splice sites, show similarities ranging from 75 to 90% to separate segments of CLIC2 and CLIC4, spanning amino acids 20 through 253. The locations of the putative introns of B35 correspond to those reported for CLIC2 (Heiss and Poustka, 1997), although the intron sizes are different. CLIC2 maps to Xq28 and CLIC4 to chromosome 6. Thus, it seems likely that B35 represents a new member of the intracellular chloride channel gene family. 5′ RACE with B35 sequences should identify the first 20 amino acids and the 5′UTR to complete the gene structure.

3.3. Efficiency of exon prediction

To evaluate the efficiency of exon prediction for the known genes, Two Sequence BLAST searches (Tatusova and Madden, 1999) were conducted for each, using a cDNA sequence retrieved from GenBank and the appropriate genomic sequence retrieved either from GenBank or from the ERI web site. This analysis provided the locations and sizes of all exons, coding and non-coding, for each gene. These were then compared to the number of coding exons (prediction programs are not designed to recognize non-coding exons) successfully predicted by any of the four programs in Genotator and to the number of exons predicted consistently by both Grail and Genscan programs (Table 1).

Similar to the examples shown in Fig. 2, the remainder of the 38 known genes differed in the characteristics of their patterns of consistent exons. A total of 21 showed consistent exon predictions with 3–4 programs, i.e. similar characteristics to those seen for APECED, PFKL and c21orf2. For these genes, typically >80% of exons were recognized by at least two programs, and few exons would be missed by relying on Genscan plus Grail predictions. In genes where fewer than 3–4 programs were consistent, it was uniformly Genscan and Grail that correctly predicted the greatest number of exons, and that together provided a reliable visual model of the gene. Fourteen genes followed this pattern. For these genes, it was typical that a proportion of exons, often as many as 50%, were not reliably predicted, although the genes were still visibly recognizable.

Exons of three genes, DCRB, c21orf4 (Fig. 2d) and c21orf3, were not well predicted. DCRB is composed of three coding exons, spanning >60 kb. Of these, only the third is predicted and that only partially and only by Grail. c21orf3 is composed of six coding exons dispersed over 22 kb. Two of the exons, separated by 8 kb, are predicted by Genscan and Grail, a third is predicted by Grail alone, and the remaining three are not recognized by any of the programs. To compound the problem, there are two consistent exons predicted that are apparently not part of the mRNA. In the absence of the BLASTX results, the DCRB and c21orf3 genes would mostly likely not be recognized; the absence of a pattern of consistent exon predictions would make the design of successful RT-PCR experiments problematic.

In summary, the robust predictions of recognizable patterns of exons, ones that would prompt RT-PCR experiments, from 35 of 38 known genes and three of three new homologies, suggest that this analysis alone should reveal more than 90% of genes.

3.4. Novel gene prediction and classification

A major focus of sequence analysis is the identification of novel genes, defined here as those models with

Table 1
Summary of gene identification[a]

Gene	Accession No.	Location	Class	Size (kb)	Number of exons			Number of programs	Comments
					Total	Predicted	Gs+Gr		
SEG #1, 40% GC, q21, 300 kb; D87675									
APP	M34862-79	1-1-9F	V	286	19	17	14	4	CpG:945/.90/70.1% within 5′ UTR
SEG #95, 43.6% GC, q22.1, 249 kb; AP000030–31									
BC1		95-1-55	CpG						2300/.92/69.7% 100 kb 5′ to SOD1 CpG
β-trans		95-2-35	ψ						No introns; 5 stop codons
SOD1	K00065	95-2-65F	V	10	5	4	3	2	CpG:920/.91/71% within 5′ coding
rA4	AF023142	95-3-48R	V	>61	20	19	8	2	CpG:1500/.92/74%; within 5′ coding
SEG #94, 44.6% GC, q22.1, 545 kb; AP000032-36									
B18		94-1-15R	EST	>10		3	0	1	Intronic exon with *C. elegans* partial homology; CpG 15 kb 5′: 350/1.2/69%
B19		94-2-6R	Exon+EST	45		4	3	4	5′ CpG; 1600/.85/70.6%
B1		94-2-25F	Exon	50		23	23	4	CpG:600/.77/70% within first exon
KIAA0539	AB011111	94-2-77F	V	27	14	14	12	4	
B38		94-3-20F	EST	>3		2	0	0	
B27		94-3-30R	Exon	8		2	2	4	Internal CpG 600/1.1/55%; overlaps opposite strand B38
B28		94-3-50F	Exon+EST	10		5	2	0–2	5'CpG 640/.83/68%
B20		94-4-60R	Exon	35			4	2	
B9		94-6-90R	H	>30		4	7	4	Mouse MAK V homologue
SEG #93, 43.3% GC, q22.1, 1050 kb; AP000037-47									
B36		93-1-5R	H	>7	(3)	2	1	4	Homology to T complex; 3′ end outside region analyzed
B2		93-1-45R	Exon+EST	27			8	3–4	internal CpG:1100/.82/63%
SYNJ1	AF009039	93-2-60R	V	96	31	24	12	2	CpG:1000/.8/75% within 5′ coding
B3		93-3-14R	Exon+EST	40			14	2	5′ CpG:750/1.1/72%; GC rich protein homology
B4		93-3-35F	Exon	10			4	2	
B37		93-3-55R	EST	20		3	0	1	Overlaps opposite strand B4
PKCBP		93-5-80F	V	1	1	1	1	4	Intronless; CpG:1200/.92/71%
BC15		93-6-5	CpG						770/.79/67%; 5 kb 3′ of PKCBP; 37 kb 5′ of BC16
BC16		93-6-42	CpG						1900/.85/71%; 37 kb 3′ of BC15; >100 kb from any exon gene model
B29		93-7-37F	EST	5		2	1	2	
IFNARB	L41942	93-8-34F	V	21	8	5	4	2	CpG:800/.93/73% within 5′ UTR
CRFB4	Z17227	93-8-60F	V	30	7	7	6	2	CpG:600/.72/68.5% within 5′ coding
IFNARA	J03171	93-9-35F	V	30	11	10	5	2	CpG:930/.97/64%; within 5′ coding
B6		93-9-65F	Exon	10			3	2	
AF1	U05875	93-10-25F	V	34	7	4	4	2	CpG:640/.88/79% within 5′ coding
c21 orf4	AF045606	93-10-85R	V	17	6	4	1	1	CpG:1100/.9/72% 10 kb 5′
5SRP		93-11-4R	ψ						Intronless
B30		93-11-12R	EST	5		1			CpG:650/.8/64%; ~ 3kb 5′ to EST
GART	X54199	93-11-62R	V	35	21	20	17	2	CpG:350/.92/63% ~3 kb 5′
SON	X63753	93-11-70F	V	25	10	9	9	2	CpG:850/.86/60% ~5 kb 5′
B17		93-12-20R	Exon+EST	10		7	4	2	5′ CpG:1000/.87/71%
B31		93-12-60R	EST	40		9	2	2	CpG
SEG #92, 41.8% GC, q22.1, 57 kb; AP000048									
B31		92-5R							CpG: 490/.79/60% within 5′ UTR
ISTN		92-5F							CpG: 1074/.96/72% within 5′ UTR

Table 1 (*continued*)

Gene	Accession No.	Location	Class	Size (kb)	Number of exons			Number of programs	Comments
					Total	Predicted	Gs+Gr		
SEG #96, 43% GC, q22.1, 215 kb; AP000049-50									
ISTN	AF064244	96-1-25F	V	>220	38	34	23	2	Internal CpG: 1200/.5/55%; CpG, within 5′ UTR
ATP50	X83218	96-3-25R	V	13	7	3	2	2	
SEG #97, 44.3% GC, q22.1, 708 kb; AP000051-57									
B10		97-1-15R	Exon	5			4	2–4	3′ CpG 290/.84/63%
B11		97-2-3F	Exon	58			7	2	
MINK		97-3-1F	ψ?				1	4	Only AA 45-95/125
B32		97-3-5F	EST	27		4	0	0	5′ CpG:800/.77/67%
B12		97-3-56R	Exon	50			8	2	Overlaps opposite strand B32
ISK	M26685	97-3-80R	V	1	1	1	1	4	Intronless; 550/.7/69% 10 kb 5′
DSCR1	U85266	97-4-67R	V	100	7	5	5	4	4 alternative first exons; 5′ CpG:1200/.9/75%
B13		97-5-10F	Exon	15			4	2	Within 90 kb first intron DSCR1
BC8		97-6-30	CpG						1800/.85/72%; >50 kb 5′ of DSCR1; >35kb 5′ of B35
B35		97-6-67F	H	9	5	5	5	2	Homology to human chloride channel
B14		97-7-15F	Exon	22			4	2	
AML1	D43969	97-8-70R	V	>100	7	6	6	2	5′ CpG:2500/.87/70%; 3′ CpG:800/.9/74%
SEG #86, 40% GC, q22.2, 959 kb; AJ229041									
B21		86-1-60R	Exon	32			5	2	
B22		86-4-40R	Exon	220			12	2	
B23		86-4-47F	Exon	85			8	2	Internal EZH2 truncated 3′ homology; internal CpG: 300/.7/57%
B24		86-5-85R	Exon	35			3	2	5′ CpG:400/.72/61%
B8		86-9-78R	Exon	163			11	2	
B25		86-10-1F	Exon	100			7	2	3′ CpG:320/.91/58%
SEG #100, 45.8% GC, q22.2, 280 kb									
BC9		100-1-28	CpG						230/.60/55%; >40 kb 5′ of SIM2
SIM2	U80456	100-1-73F	V	49	11	11	9	4	CpG: >5000/>.7/>60%; within 5′ coding
HCS	D23672	100-4-1R	V	>164	(5)	4	3	4	5′ UTR within 102-1-5; 5′ CpG:985/.84/78%; internal exons within gap in sequence
SEG #102, 42.2% GC, q22.2, 1530 kb; AP000008-14; 17-22									
TRPD	D84294	102-2-45F	V	115	45	35	22	2	CpG:1700/.87/69.7% 15 kb 5′
BC10		102-4-1	CpG						650/.74/65%; 20 kb 3′ of TRPD; 5 kb 3′ of DCRA
DCRA	D87343	102-4-46R	V	41	8	8	6	3–4	CpG:1270/.78/67% within 5′ coding
BC11		102-4-63	CpG						200/.65/71%; >15 kb 5′ DCRA
MNB	U58496	102-5-55F	V	146	11	10	10	2	CpG:2400/196/75% within 5′ UTR
B15		102-7-73R	Exon	35			7	2–4	5′ CpG:565/.91/68%
BC14		102-9-5	CpG						350/.77/67%; within KCNJ6 third intron
KCNJ6	D87327	102-11-65R	V	280	3	3	2	4	CpG:930/.86/63% within 5′ UTR
B26		102-12-10F	Exon	140			12	2	
DCRB	AB000099	102-14-2R	V	70	3	1	0	0	
B33		102-15-28R	EST ψ?			1		4	tBlastX mouse cDNA AI036903; $E<10^{-33}$
KCNJ15	NM_002243	102-15-90F	V	2	1	1	1	4	
B16		102-16-20F	Exon	12			4	2	
ERG	M17254	>102-17-60	V	>64	(8)	7	5	4	5′ end outside region analyzed

(*continued overleaf*)

Table 1 (*continued*)

Gene	Accession No.	Location	Class	Size (kb)	Number of exons			Number of programs	Comments
					Total	Predicted	Gs+Gr		
SEG #101, 43.5% GC, q22.3, 137 kb									
ETS2	J04102	101-1-30F	V	12	9	9	9	4	CpG:1700/.82/67.5% within 5′ UTR
SEG #103, 58.9% GC, q22.3, 87 kb									
APECED	Z97990	103-25F	V	12	14	13	7	4	CpG:380/.89/76% within 5′ coding
PFKL	X15573	103-40F	V	27	22	22	22	4	CpG:1030/.86/79% within 5′ coding
C21 orf2	Y11392	103-80R	V	7	7	6	4	3	CpG:960/.94/75.6% within 5′ coding
SEG #104, 49.6% GC, q22.3, 157 kb; (AB001523; 122 kb)									
TMEM[b]	U61500	104-1-25F	V	91	23	22	15	3–4	CpG:1500/.9/72% within 5′ coding
PWP2[b]	X95263	104-2-32F	V	24	21	20	16	4	CpG:800/.95/71% within 5′ coding
ES1[b]	U53003	104-2-55F	V	>33	(6)	6	4	4	CpG:1040/.81/68% within 5′ coding
SEG #171, 50.6% GC, q22.3, 333 kb; AJ011930									
B34		171-2-11F	Exon	47			5	2	3′ CpG:>1000/>.7/>64%
Motor protein		171-2-25R	ψ						
BC17		171-2-70	CpG						300/.75/68%+215/.58/58%
UBEG2	AF032456	171-3-53R	V	27	6	4	3	4	CPG:1040/.86/71% within 5′ coding
SMT3A	X99584	171-3-73R	V	11	4	4	3	4	CpG:2100/.89/74% within 5′ coding
C21 orf3	Z50022	171-4-38R	V	22	6	3	2	1	
CD18	M81233	>171-4-90	V	>12	(8)	8	7	4	5′ end outside region analyzed

[a] For each sequenced segment, the GenBank Accession Nos (where available) and the ERI Accession No. is given; the base composition in %GC is calculated as an average over the entire sequence; the approximate band location and sequence size are also listed. Gene: the standard gene name or GenBank designation is used for genes with known protein identity; accession numbers are for cDNAs used in two sequence BLAST searches; 'B' numbers indicate novel genes predicted from EST matches or patterns in exon prediction as described in the text; 'BC' numbers indicate isolated CpG islands. Location refers to the strand and the kb of the 5′ end of the gene or gene model in the Genotator file designation. Genotator analyzes sequences in 90 kb segments; therefore, location is given in three numbers: ERI sequence segment number, the number of the 90 kb interval within the segment, and the kb location within that 90 kb interval; e.g. the 5′ end of the SOD1 gene is found in ERI segment 95 in the second 90 kb segment at 65 kb on the forward strand (95-2-65F). Class refers to the type of gene or gene model: V is a verified or known gene (i.e. with 100% protein identity); H is a new member of a gene family or a new human homologue of a known gene based on a strong protein homology and no evidence of a pseudogene; EST is a gene model based solely on matches to a dbEST entry with criteria as described in the text; Exon is a gene model based solely on consistent patterns of exons predicted by two (Genscan and Grail) or more programs; Exon+EST models have consistent exon predictions, one or more of which show a match in dbEST; CpG is a gene model based upon an apparently isolated CpG island; Ψ is a likely pseudogene based on a lack of introns and/or open reading frame in spite of a strong homology to a database protein. Size, for known genes and new homologues, is based on identification of all exons in a Two Sequence BLAST search between the genomic sequence and the complete cDNA for the gene; for Exon, EST and Exon+EST models, the size is the minimum based on exon predictions, EST sequences and 5′ RACE data, where available. The number of exons has been counted in three ways. 'Total', for known genes, is the number of coding exons found in Two Sequence BLAST matches; for gene models, it is based on exon prediction, EST sequence and available RT-PCR data, and must be considered a minimum. 'Predicted' refers to the number of correct exons predicted by one or more programs or by dbEST sequence (this number may differ from information obtained by RT-PCR). 'Gs+Gr' is the number of exons predicted correctly by both Genscan and Grail programs. 'Number of programs' is the average number of programs predicting each exon within a known gene or gene model. 'Comments' includes information on the size, strength and location of CpG islands, given as #bp/CpG frequency obs:exp/GC%. For a description of the known genes, see references in GenBank entries and in Antonarakis (1998).
[b] Gene structures from genomic sequence also reported in Nagamine et al. (1996, 1997a,b).

no discernible significant similarity to any protein or complete cDNA in the databases. To classify putative novel gene models, we used criteria based on exon prediction, EST matches, and CpG islands.

Exon gene models were based on patterns of consistent exon prediction. The term 'consistent exons' means that more than one program predicts the same sequence to be an exon; the overlap need not be perfect, and predictions can vary as to the precise locations of putative splice sites. Requirements for an exon gene model are: (1) a minimum of two consistent exons predicted by 3–4 programs, specifically Genscan and Grail, plus Xpound and/or Genefinder, or (2) a minimum of three consistent exons predicted by Genscan and Grail. Obviously, 'intron' size is relevant but it is not possible to specify limits. Visually, the greater the number of consistent exons in a pattern, the larger the intron sizes can be and still provide a convincing gene model that can be tested experimentally.

An EST gene model is defined as a match to a human EST clone (from Soares libraries, TIGR or NCI-CGAP libraries) fulfilling the following criteria: the match must be over the complete sequence of the EST, the identity must be >88%, and the match must be either non-contiguous, i.e. it must identify in the genomic sequence two or more exons showing consensus splice sites, or, if contiguous, it must be in the vicinity of additional gene features such as consistent exons or a CpG island. These criteria were chosen to eliminate spurious matches that arise from the nature of EST clones, i.e. similarities among repetitive elements or sequence motifs found in 3′ untranslated regions (UTRs) and ESTs resulting from priming from intronic sites. While the stringency of these criteria (in particular the non-contiguous match requirement) will eliminate some bone fide matches (e.g. intronless genes, and those with long 3′UTRs located at a distance from convincing patterns of exon predictions), it will also avoid numerous false positives.

Using Exon and EST gene characteristics, the following classes of novel gene models were defined: Exon+EST genes, those based on patterns of consistent exon prediction, one or more of which are also associated with EST matches; Exon-EST genes, those based solely on patterns of consistent exons; and EST-Exon genes, those that are based on EST matches but lack an association with a consistent pattern of exons. A fourth class of novel gene was defined as the isolated CpG island gene (see below).

Using these criteria, five Exon+EST gene models were identified, B19, B28, B2, B3 and B17. All were associated with nearby CpG islands. As an example, B17, is shown in Fig. 3a. Genotator output shows four consensus exons within 10 kb, with a strong CpG island located at the most 5′ exon. EST analysis confirmed two of these exons and added six more.

Twenty Exon-EST models were defined. For example, the B27 gene model (94-3-30R) is composed of two exons, both predicted by all four programs. In contrast, the B1 gene (Fig. 3b) is predicted to be composed of 23 exons, most of which are predicted by all four programs. An intriguing set of Exon-EST gene models are predicted in the most AT-rich segment, 86, located within 21q22.2. Many of these included >8–10 Genscan+Grail exons within a distance of 80–100 kb (data not shown). There was little background of inconsistent exon prediction in these regions, and adjacent gene models could be defined by abrupt switches in the strand in which consistent exons were located. Each of these Exon-EST models must be regarded as hypothetical, unless and until transcription and splicing are experimentally demonstrated.

Seven EST-Exon gene models were identified; examples are described here:

1. B31 (92 and 93-12) (Fig. 3c). Exon prediction in segment 92 and between 93-12-60R to 93-12-20R is sporadic, inconsistent among programs and visually uninspiring. However, partially overlapping EST matches together predict 13 exons within a 52 kb region. Only two of these exons are predicted by both Genscan and Grail; three others are predicted by Genscan alone and three by Grail +/− Xpound. Additional ESTs overlap the 5′ end of this model, which terminates in the adjacent sequence segment (the size of the gap is unknown at the time of this analysis), 92, at 5 kb, where a CpG island is found. There are more than 80 matches in human dbEST. It is noteworthy that there are also significant matches (E<10–30) for sequences in mouse and rat dbEST.
2. B37 (93-3). The 5′ sequence of a single EST clone predicts three exons on the reverse strand, spanning 93-3-55R to 93-3-35R. The first exon is predicted by Grail, the second is not predicted, and the third is roughly predicted on both strands with Genscan and/or Grail. The match with the 3′ sequence from the same EST clone suggests that the third exon is long and includes the 3′ UTR. This picture is complicated by the prediction by both Grail and Genscan of three exons on the forward strand in what would be the 17 kb second intron of the gene model on the reverse strand. This latter Exon-EST model has been designated as B4, to distinguish it from the EST-Exon gene, B37.
3. B32 (97-3) (Fig. 3d). Five ESTs identify four exons within 97-3-5F to 97-3-32F, none of which is predicted by Genscan and only one of which is predicted by Grail. The first exon of the EST is located within a strong CpG island. As with the B37 gene, the picture is complicated by Grail+Genscan exon predictions on the opposite strand. In this case, however, the EST exons on the forward strand are located in three different introns of the Exon model predicted on the reverse strand.

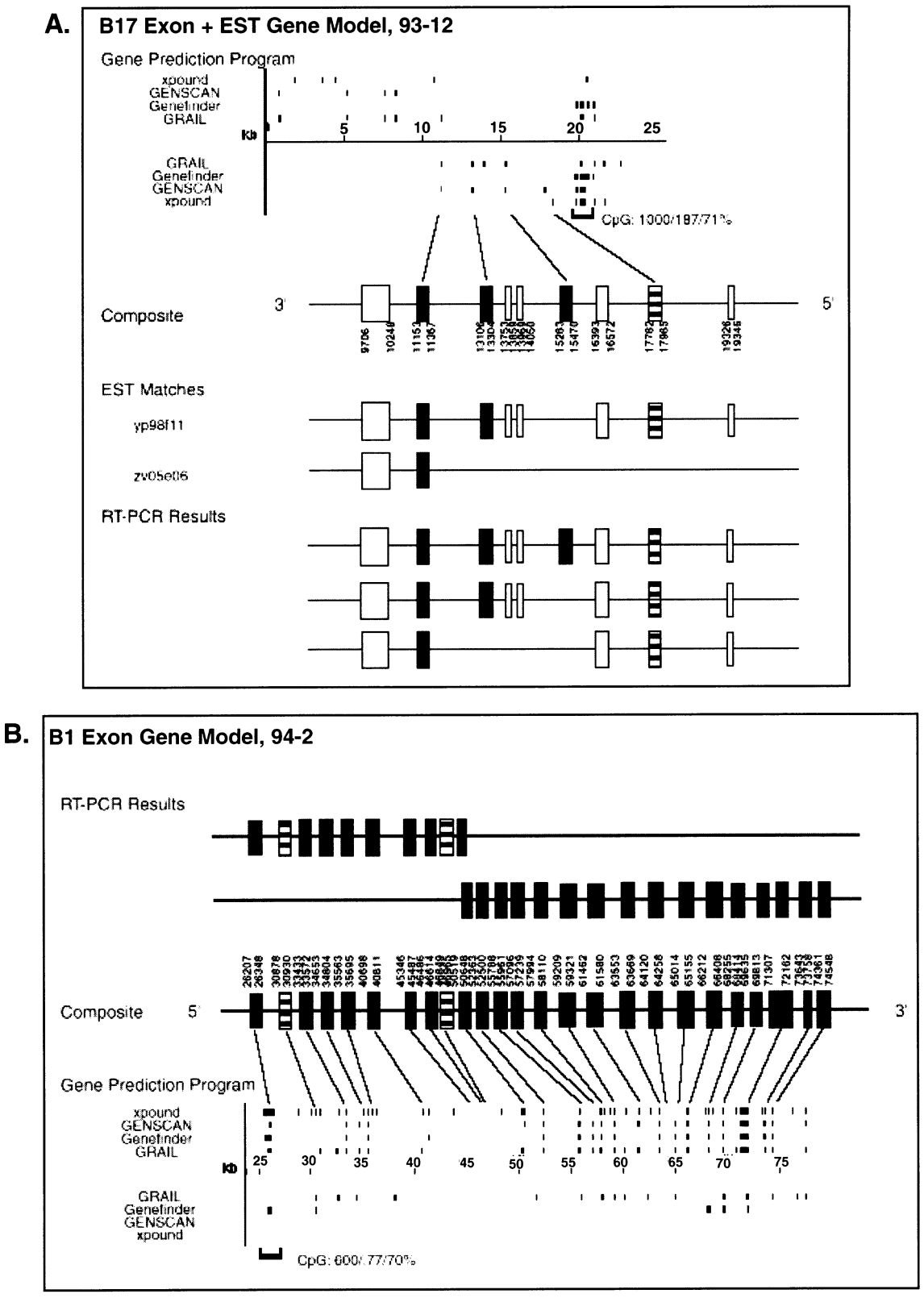
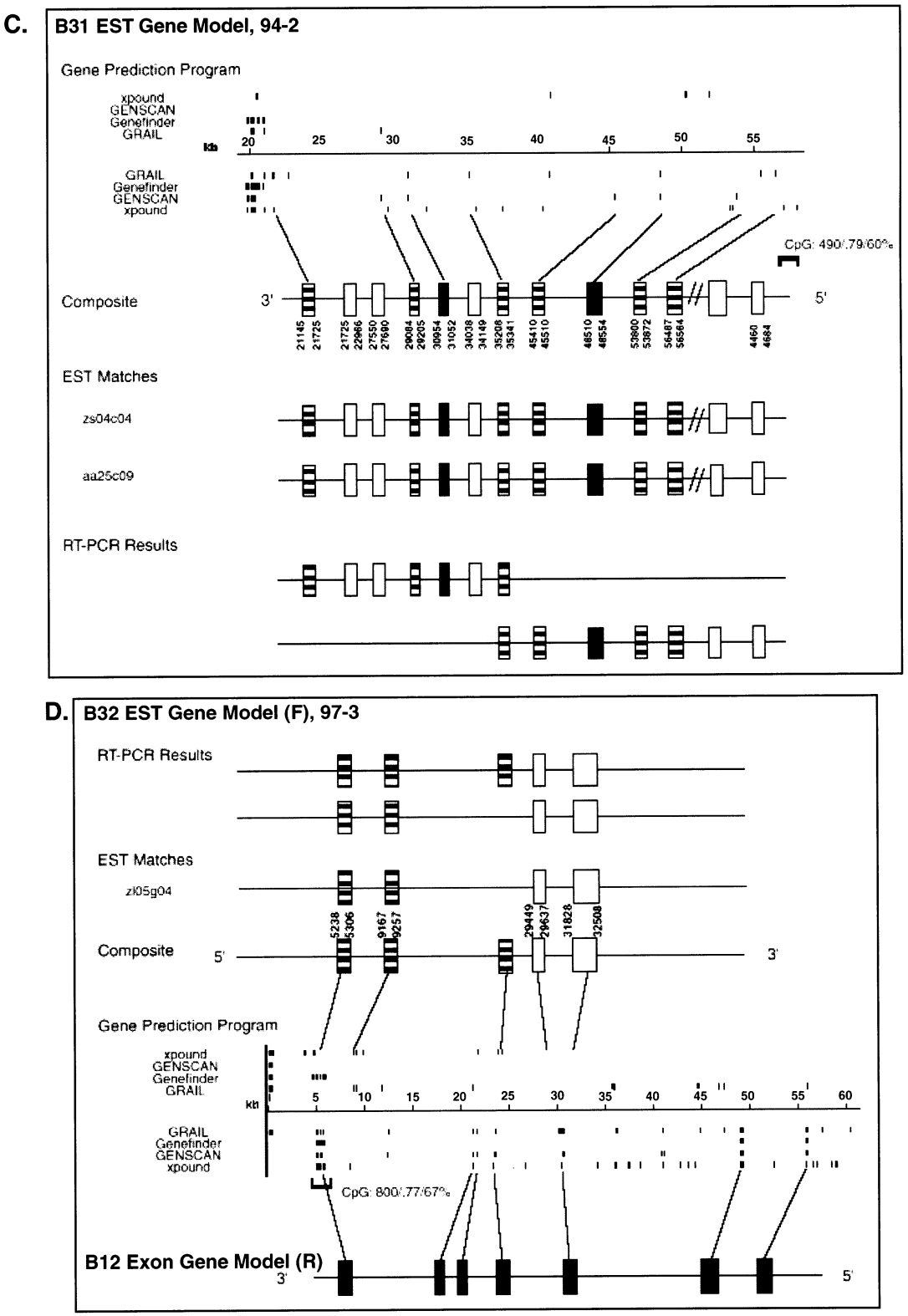


Fig. 3. Partial gene structures of several gene models, as determined from exon prediction programs, EST sequences and RT-PCR experiments. The exon prediction output of Genotator analysis is shown in each panel. EST data indicate exons detected by sequencing of ESTs of the given clone number; RT-PCR data indicate exons detected by sequencing of RT-PCR products. More than one EST or RT-PCR pattern indicates the exon content of alternatively processed transcripts. Directly above or below the Genotator pattern is shown the composite exon pattern derived from RT-PCR and EST sequence data, with nucleotide positions indicated. Solid exons are predicted by both Genscan and Grail; striped, by



Genscan OR Grail; open exons, by neither Genscan nor Grail. (**a**) B17: Exon prediction and CpG island identification strongly suggest the presence of a gene. Both EST sequencing and RT-PCR were required to define the complete gene structure and define three transcripts; (**b**) B1: exon prediction is excellent and verified by RT-PCR, yet there are no ESTs from Soares, TIGR or CGAP libraries; (**c**) B31: only two of 13 exons verified by ESTs and RT-PCR were predicted by both Grail and Genscan; (**d**) B32 and B12 gene models interdigitate: four exons of the B32 EST gene are predicted on the forward strand; these exons interdigitate with those of the B12 Exon model on the reverse strand; B12 is negative by RT-PCR in all tissues tested.

It is worth reiterating that software prediction of exons for none of the seven EST-Exon genes produced pictures as convincing as those seen in even the least inspiring Exon-EST genes described above. While Grail and/or Genscan did predict in each case one or more exons found by the EST matches, it is unlikely that their patterns would have prompted attempts at expression verification in any real-world RT-PCR program. Proposing RT-PCR experiments for every exon predicted by each program would be ineffective and highly expensive, given that currently, each program has a non-zero false positive rate that has been evaluated as high as 30% (Fickett and Tung, 1992; Burset and Guigo, 1996; Claverie, 1997). dbEST information has therefore been extremely valuable.

3.5. Genomic features

3.5.1. CpG islands and CpG island genes

CpG islands were predicted using the criteria of Gardiner-Garden and Frommer (1987), i.e. a stretch of >200 bp, with a GC content of $>50\%$ and a CpG dinucleotide frequency ratio of observed to expected of >0.6. Segments of such characteristics stand out against the mammalian genome average of 38% GC and an observed frequency of CpG only 0.2–0.25 of that expected from base composition. Larsen et al. (1992), in an analysis of GenBank data, found that $\sim 50\%$ genes were associated with CpG islands. In the 6.5 Mb analyzed here, 125 such CpG islands were found; BLAST searches revealed that 44 of these were Alu sequences, and these were eliminated from further consideration. The remaining CpG islands were classified, based on sequence length, as strong (>750 bp), moderate (400–750 bp), or weak (200–400). As shown in Table 2, strong islands averaged 1260 bp in length, moderate, 590 bp and weak islands, 280 bp. The average value of %GC and CpG ratio did not vary significantly among island classes.

Thirty-one of the 38 known genes are associated with CpG islands, most of them strong. For 19 of these genes, the CpG island spanned 5′ coding regions; for seven genes, the islands spanned 5′ untranslated regions only, and for the remaining five genes, the islands appear to be located 5′ to the untranslated regions, although this could be due to incomplete data on the 5′UTRs of the associated cDNAs. Three genes, KIAA00539, KCNJ15 and c21orf3 lack CpG islands. ERG and CD18 could not be evaluated because their 5′ ends lie outside the regions analyzed.

Of the Exon+EST, Exon-EST and EST-Exon gene models, a total of five of five, eight of twenty and three of seven, respectively, have likely CpG island associations. For the remaining novel gene models, no islands were in the vicinity, but this is obviously inconclusive because the sizes of these genes are not reliably estimated. Linking a CpG island to the appropriate exon/EST predictions is particularly ambiguous given the possibility of very large first introns (see below). The existence of additional CpG islands should, however, be annotated for future reference; this led to the designation of a fourth class of gene, the isolated CpG island.

CpG island genes are defined as strong and moderate CpG islands that are not associated with convincing patterns of exons or spliced ESTs for tens of kb both 5′ and 3′, or that are found 3′ to a gene that has a 5′ CpG island. Nine such genes are included in Table 1, where they are described by their CG content and size, and by their distance from known genes or robust models. It is clear that this is not a robust classification of gene. Given the observation that many genes are poorly predicted by exon programs, and given that dbEST is as yet incomplete, many or all of these islands may be associated with unidentified exon patterns.

3.5.2. Large introns

Fourteen genes have one or more introns $>\sim 20$ kb, most often located within the 5′ UTRs or 5′ coding regions. The locations and sizes of all 22 large introns are listed in Table 3. Together, these total >1 Mb, more than 15% of the DNA analyzed.

Table 2
CpG islands: characteristics and gene associations[a]

Class	Size	CpG o/e	Percentage GC	Total number	Gene association		
					Known	Exon/Est	CpG
Strong	>750 bp (1260 bp)[b]	0.86[b]	70%[b]	37	26	5	6
Moderate	400–750 bp (590)[b]	0.82[b]	67%[b]	12	3	6	1
Weak	200–400 bp (280)[b]	0.85[b]	62%[b]	32	2	5	4

[a] Strong, moderate and weak CpG islands are described by size (average for the class is given in brackets), ratio of CG frequency observed/expected, and GC level. Total number: number of islands of each class observed in the 6.5 Mb. Gene associations for each class with known genes; Exon/EST: novel Exon +/− EST models, and CpG: isolated CpG islands. The Intersectin gene is associated with two strong islands, one 5′ and one internal.
[b] Average values.

Table 3
Introns > ~20 kb (AML1, D. Levanon, personal communication)

Gene	Intron location	Size	Isochore location
APP	First coding	55.9 kb	L
	Five other coding	21.7–31.6 kb	
rA4	First coding	25.3 kb	L
SYNJ1	First coding	24.7 kb	L
	(w/i 5′ UTR)	(1.0) kb	
ISTN	5′ UTR	>62.2 kb	L-H1
DSCR1	First coding (alternative)	~90 kb	H1
HCS	Internal coding	129.6 kb	H1
DCRA	First coding	26.9 kb	L
MNB	5′ UTR	51 kb	L
	First coding	55 kb	
KCNJ6	5′ UTR	75.2 kb	L
	First coding	125.5 kb	
	Second coding	88.2 kb	
DCRB	Second coding	65.3 kb	L
KCNJ15	5′ UTR	61.6 kb	L
AML1	5′ UTR (alternative)	150 kb	H1
ERG	~First coding	>40 kb	L
TMEM	First coding	19.5 kb	H3

3.5.3. Intergenic distances

For a number of genes, the extents of 5′ and/or 3′ UTRs are known or could be inferred. This information was used to determine the intergenic distances listed in Table 4. Thirteen of the nineteen defined intergenic distances are less than 5 kb, some considerably less. When this occurs at the 5′ ends of genes, it leaves little room for promoters and other regulatory sequences. Small intergenic distances are anticipated in the most GC-rich regions, where isochore analysis predicts a high gene density. Thus, 1.8 kb between APECED and PFKL, 1.1 kb between TMEM and PWP2, and 2.5 kb between PWP2 and ES1, all from regions that are >50% GC, are not surprising. Limited intergenic distances are also seen, however, in moderate and low GC segments. For example, the 5′ end of KIAA00539 is only 629 bp downstream from the first possible polyadenylation site of the B1 Exon gene model. The 5′ ends of B31 and Intersectin are located in CpG islands separated by <400 bp, and thus may share a bi-directional promoter. The 3′ end of SR-A4 is less than 2.5 kb from SOD1. These genes derive from regions of 43, 42 and 40% GC, respectively. Interpretation of larger intergenic distances remains problematic because the failure to predict a gene in a 20–30 kb segment is clearly inconclusive.

Table 4
Intergenic distances[a]

			bp
SOD1 3′ UTR	–	A4 3′ UTR	2327
B1 ‘3′ UTR’	–	KIAA00539 5′ UTR	≤629
IFNARB 3′ UTR	–	CRFB4 5′ UTR	1898
CRFB4 3′ UTR	–	IFNARA 5′ UTR	27763
AF1 3′ UTR	–	c21orf 4 3′ UTR	13190
GART 5′ UTR	–	SON 5′ UTR	8791
SON 3′ UTR	–	B17 3′ UTR	681
B17 5′ UTR	–	B31 3′ UTR	<2000
B31 5′ UTR	–	ISTN 5′ UTR	<900
ISTN 3′ UTR	–	ATP50 3′ UTR	13208
SIM2 3′ UTR	–	HCS 3′ UTR	4014
TRPD 3′ UTR	–	DCRA 3′ UTR	13439
APECED 3′ UTR	–	PFKL 5′ UTR	1832
PFKL 3′ UTR	–	c21orf 2 3′ UTR	1571
TMEM 3′ UTR	–	PWP2 5′ UTR	1133
PWP2 3′ UTR	–	5′ UTR ES1	2502
UBEG2 5′ UTR	–	SMT3A 3′ UTR	3833
SMT3A 5′ UTR	–	c21orf 2 3′	31325
c21orf 2 5′ UTR	–	CD18 3′ UTR	12277

[a] Locations and extents of 5′ and 3′ UTRs of known genes were determined by two-sequence BLAST; they were inferred for novel gene models from locations of CpG islands (5′) and/or polyadenylation site (3′) where possible.

3.6. Experimental analysis of novel gene models

One or more EST clones were obtained for each of the Exon+EST and EST-Exon genes. Clones were completely sequenced and compared to the genomic sequence to determine all exons present. Primers were designed to various predicted exons and used in RT-PCR from Hela and various tissue RNAs. Similarly, a subset of the Exon-EST gene models were examined by primers designed to consistent exons. Each RT-PCR product was sequenced to verify included exons and correct splicing. The results of some examples are shown in the lower portions of Fig. 3a–d. For B17, RT-PCR confirmed all predicted and EST exons, and added information on alternative splicing (Fig. 3a). Three RT-PCR products are observed; the largest is 1.7 kb and contains an open reading frame (orf) (not necessarily complete

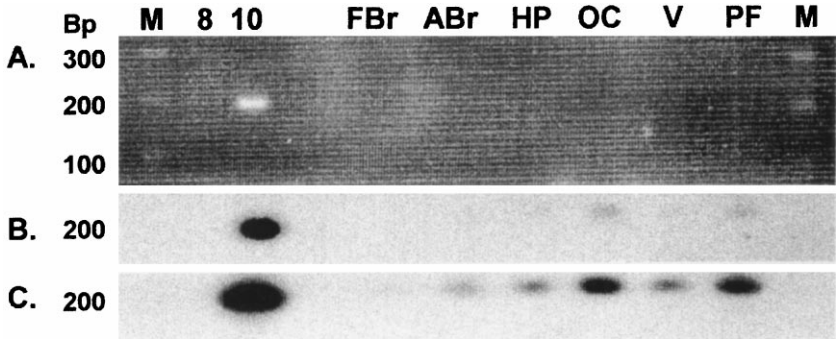


Fig. 4. RT-PCR analysis of Exon model B27. Primers designed to the exons predicted in 94-3R at 22 and 30 kb were used in RT-PCR analysis of eight tissues. Products were resolved in an agarose gel, transferred to a nylon membrane and hybridized. (**a**) Ethidium bromide staining of the gel, showing a visible product only in 10 week fetus; (**b**) after hybridization, 10 min exposure; (**c**) 1 h exposure. M, markers; 8, cDNA from 8 week fetus; 10, 10 week fetus; FBr, ~18 week fetal brain; ABr, adult brain; HP, hippocampus; OC, occipital bulb; PF, prefrontal cortex; V, vermis.

Table 5
RT-PCR analysis of novel gene models[a]

Gene	Type	ESTs		Tissues											Alternative splicing
		Number	Clone Name	H	Pl	8	10	FB	AB	CR	HP	PF	OC	V	
B1	Exon	0		+	+	+	+	+	+	+					
B2	Exon + EST	>36	aa24g05	+	+	+	+	+	+	+	+	+	+	+	Yes
B3	Exon + EST	29	zj32b10	+	+	+	+	+	+	+	+	+	+*	+*	Yes
B4	Exon	0					−	−							
B8[a]	Exon	0		−	−	−	+*	+*	−	−	−	−	−	−	
B9	H	2	zw01h08	+	+	+	+	+	+	+	+	+	+	+	
B15	Exon	0		−			−	−	−						
B17	Exon + EST	45	yp98f11	+		+	+	+	+	+	+	−	+	+	Yes
B18	EST	3	zl03h07	+	+	+	+	+	+	+	+	+	+	+	
B19	Exon + EST	3	yy54b04	+	+	+	+	+	+	+	+*	+*	+	+	Yes
B21	Exon	0		−	−				−	−					
B23	Exon	0		−				−	−						
B27	Exon	0		−	−	−	+	−	−	+*	+*	+*	+*	+*	
B28	Exon + Est	73	qy92e02	+		+	+	+	+	+	+	+	+	+	
B31	EST	88	zr21h04	+	+	+	+	+	+	+	+	+	+	+	Yes
B32	EST	22	zl05g04	+	+*	+	+	+*	+	+	+	+	+	+	Yes
B37	EST	3	ym12a05	−	+*	−	+	+	+*	+	+*	−	+*	−	Yes

[a] H, Hela; Pl, placenta; 8, 8 week fetus; 10, 10 week fetus; FB, ~18 week fetal brain; AB, adult brain; CR, cerebellum; HP, hippocampus; PF, prefrontal cortex; OC, occipital lobe; V, vermis; +, product visible in ethidium bromide stained gel; +*, product visible only after hybridization to Southern transfer of the gel; −, negative in hybridization; blanks indicated sample was not tested; a, positive only in intra exon RT-PCR

at the 5′ end) of 430 amino acids. The alternative splicing in forms 2 and 3 (1.5 and 1.2 kb products) truncates the orf to 232 and 230 amino acids, respectively. Interestingly, the longest transcript is not detected in several brain regions, although it is seen in whole brain (data not shown). For the B27 gene, the two consistent exons were verified by RT-PCR to be expressed at significant (visibly detectable in gels stained with ethidium bromide) levels only in a 10 week fetus among the tissues tested, although hybridization revealed lower levels of expression in brain regions (Fig. 4). Fig. 3d illustrates splice forms detected in RT-PCR for the B32 gene.

As shown in Table 5, patterns of expression varied, as did levels of expression, as indicated by the requirement for hybridization to verify the presence of RT-PCR products in some cases, in some tissues. Of the 17 models tested, 13 were verified to be expressed in one or more tissues examined. Because RT-PCR was across one or more introns and because RT-PCR products were verified to be correct by sequencing, these results illustrate the reliability of these models. Interestingly, the low representation of many (seven of 13 models tested had three or fewer matches) in dbEST is consistent with the failure to identify these genes in previous efforts at chromosome 21 gene discovery.

4. Discussion

The 40 Mb of 21q is estimated to contain between 500 and 1000 genes, assuming between 50 000 and

100 000 genes total in the human genome and a uniform chromosomal distribution of genes. While there is evidence to suggest that such estimates for chromosome 21 are likely high (Gardiner, 1997), the current catalogue of genes is still far from complete. Fewer than 100 genes, as defined by sequencing of complete coding regions, have been reported in the literature (for a summary, see Antonarakis, 1998, and references therein; also GenBank entries in Table 1). In addition to these characterized genes, a large number of gene fragments have been identified by extensive efforts at cDNA selection and exon trapping (Tassone et al., 1995; Yaspo et al., 1995; Cheng et al., 1994; Chen et al., 1996; Ohira et al., 1997; Dahmane et al., 1998). More recently, over 300 ESTs were reported with map locations on 21q (Deloukas et al., 1998). The nature and number of genes that these cDNA fragments, exons and EST sequences together represent are unknown.

The analysis of the 6.5 Mb presented here was designed to establish criteria for identification of novel genes within genomic sequence, to identify novel genes of potential relevance to Down syndrome and other chromosome 21 diseases, and to examine features of mammalian genome organization. As such, results have a general applicability to any segment of the human genome, as well as particular relevance to chromosome 21.

4.1. Criteria for novel gene identification

Novel genes are defined here as those Exon +/− EST models with no significant homology to known proteins as determined by BLASTX searches of the non-redundant protein database. Estimations of the success of novel gene identification must depend in part on the successes in computational prediction of known genes. In the 6.5 Mb, 38 genes already known to map to chromosome 21 were identified by protein homologies. Only three of these did not also present a reliable exon model with at least Grail and Genscan. An additional three genes representing new homologies to mouse or human gene families were also found by both protein matches and patterns of predicted exons.

For novel gene identification, both computational and expression analyses are required, i.e. exon prediction, EST matches (because there are no protein matches), CpG island location, coupled to RT-PCR/Northern analysis. By these means, 41 additional genes were predicted. Twenty-five novel gene models are proposed, based on patterns of consistent exon prediction. Five of these are also associated with one or more ESTs. Of these 25 models, 12 were tested in RT-PCR experiments designed to verify splicing between two or more exons. For eight of the 12 genes tested, expression was verified in one or more tissues. One of these, B8, could only be detected via intra-exon RT-PCR, suggesting that while the exons were correct, the gene model may not have been.

For four Exon-EST models, even 60 cycles of amplification plus hybridization failed to verify expression. Such negative results could be due to restricted time and/or place of expression, which could be addressed by assaying additional tissues and developmental time points. The latter is not always practical for human tissues, but the growing rat and mouse EST databases may be helpful in this regard. However, for transcripts with longer (>1–2 kb) 3′ UTRs, the probability of the EST sequence including any coding sequence is small, and therefore, the probability of detecting similarity with a human sequence is small. It is also possible that the failure of RT-PCR is a result of incorrect gene modeling. Intra-exon RT-PCR could be used, followed by more exhaustive inter-exon RT-PCR and RACE to define the correct gene model. Lastly, the failure could be a result of false positive exon predictions. There are no data to indicate how often Genscan and Grail both predict a false positive, and given the ambiguities involved in proving that a sequence is not a gene, such data will be hard to generate. Exon-EST models must therefore be regarded as hypothetical.

Seven EST-Exon gene models were found. In requiring consistent predictions with at least Genscan and Grail, these genes were missed by the Exon prediction criteria used here. It remains impractical to propose RT-PCR experiments for all exons predicted by all programs, until false positive rates are reduced. While, in these cases, the complexity of dbEST solved the problem of gene detection, how frequent and in what proportion of cases is this probable? For the 41 known and homologous genes recognized in this 6.5 Mb, only three (7%) would have been missed by exon prediction criteria. Given the finding of an additional 25 Exon+/−EST putative genes, in comparison, a 7% failure rate suggests that only one or two genes would be expected to be missed. The need to rely on dbEST for identifying seven additional genes was surprising and suggests that there is a significant class of genes that current exon prediction programs do not see. This becomes even more important when considering that these observations imply yet another class of genes, the -Exon-EST genes, those with poor exon prediction and no entries in dbEST. Such genes will be refractory to identification by current methods.

Dahmane et al. (1998) recently reported results of exon trapping and cDNA selection within a 2.5 Mb region from 21q22.2. Because this region overlaps with segments 100 and 102 analyzed here, it is relevant to compare the experimentally derived gene models with those obtained from genomic sequence analysis. Four novel Transcriptional Units (TUs) are of particular interest. TU17 is contiguous with the genomic sequence, located at 102-8-79. It is possible that this is the 3′UTR

of a gene formed by BC14 16 kb upstream at 102-9-5 plus a consistent exon a 102-8-88R. It would be of interest to test this model. TU24 is of interest because its 550 nucleotides are split among three exons, spanning 102-14-2682 to 36872. All three exons have consensus splice sites, but none is predicted well by software. TU25, at 102-14- 58264, could represent part of the 3′UTR of TU24. Again, this is a model worth testing. Lastly, TU27 is a 692 nucleotide cDNA contiguous with genomic sequence at 102-16-56052. Conceivably, this forms a 3′UTR of the B16 exon gene model.

Based on all of these data, success in gene identification from human genomic sequence would be enhanced by several additions/improvements in available information and resources. The availability of mouse genomic sequence would allow comparative exon prediction. Consistent exon prediction between human and mouse would add considerable confidence to the existence of spliced transcripts when experimental verification is lacking due to restricted expression. Efforts to generate sequence of complete coding regions of mouse cDNAs, especially for transcripts of developmental or restricted tissue specificity, would greatly increase the rate of positive BLASTX searches with human Exon models and overcome the 3′UTR limitation of mouse and rat dbESTs. Such data would also allow appropriate tissue selection, again especially for developmental genes, and increase the chances of verification of gene models by expression analysis. In addition to mouse resources, increasingly reliable computational tools of promoter prediction and SAR/MAR identification, as well as retraining of exon prediction programs using data from experimentally verified EST-Exon genes may improve subsequent predictions. Of these improvements, the generation of mouse sequence is arguably the most critical, both because it would also provide information on conserved regulatory sequences, and also because it may be less difficult than generation of complete coding sequences of rare mouse transcripts. Last, but not at all least, as evidenced by the data of Dahmane et al. (1998), concerted efforts at the direct experimental approaches of exon trapping and cDNA selection still have valuable and unique contributions to make.

4.2. Expression analysis

Expression analysis is essential for verification of computationally derived gene models, which are, after all, only predictions. Expression data, however, also add functionally important information on tissue specificity and alternative processing. While experiments done here were not quantitative, variation among genes in expression levels in a single tissue and among tissues with a single gene can be inferred from the relative variation in quantity of RT-PCR product obtained. Table 5 shows, for example, that B3 and B19 are expressed in all 11 tissues tested; however, B3 required hybridization for detection in occipital bulb and vermis, while B19 required hybridization in hippocampus and prefrontal cortex.

Expression experiments also revealed a high frequency of alternative processing, seen in seven of 13 models, as defined by the presence of more than one band in RT-PCR verified by hybridization and/or sequencing. Because few of the gene models described here are complete, the general effect of alternative processing on protein sequence remains to be addressed.

4.3. Genome organization

The picture of the human genome presented here is one comprising both large introns and short intergenic distances. This suggests the occurrence both of large regions of no obvious functional relevance and of compressed regulatory regions. Twenty-two introns of $> \sim 20$ kb were identified and total > 1 Mb, or $> 15\%$ of the DNA analyzed. Given that observation, it must be considered that the isolated CpG islands may be marking the 5′ ends of additional genes with very large first introns. With one exception, there is no evidence of gene models within the large introns. Here, comparative sequence analysis with mouse will not help; if exon prediction fails in the human sequence, it is unlikely to be effective on the homologous mouse sequence. In contrast to large first introns, CpG islands containing the 5′ UTRs of B31 and ISTN are only 350 bp apart. In eight other cases, either the 3′ end or the 5′ end of one gene is within < 2500 bp of the 5′ end of an adjacent gene. Such short distances offer opportunities for examining overlapping or intermingled regulatory sequences; perhaps large first introns are not useless, but rather are repositories for regulatory sequences of adjacent genes. The sequence of regions of the corresponding mouse genomic DNA may be particularly valuable in these cases for identification of conserved and therefore functionally important sequences.

Three instances were noted where a verified EST gene model was found to overlap or interdigitate with an exon gene model on the opposite strand. The EST gene models B37 and B32 are composed of at least three and five exons, respectively, and include at least one alternatively spliced exon each. Consensus splice sites are found in each case only on the appropriate strand, so these are not ESTs with reversed orientations due to cloning artefacts. On the opposite strands to B37 and B32 are found Genscan + Grail exon models, B4 and B12. Exons of these models also display appropriate consensus splice sites, but RT-PCR has so far been negative in the tissues tested. These data suggest that, if these are indeed overlapping/interdigitated genes, in each case, the pair of genes are restricted in expression, in time and/or place. Such limited or restricted expression may be a

Table 6
Gene distribution and isochores

Isochore	Percentage genome[a]	Mb 21q	Number of genes known and predicted	Total number of genes in 21q	
				[b]	[a]
L	~67%	27	30 in 3.1 Mb (20)[a]	260	180
H1/H2	~25	10	40 in 2.9 Mb (54)[a]	138	185
H3	~5%	2	12 in 0.58 Mb (64)[a]	40	220
Totals		39		438	585

[a] Based on whole genome isochore data (Bernardi, 1995; Zoubak et al., 1996).
[b] Extrapolated from 21q data.

requirement for appropriate regulation of genes that are in such close physical proximity. Alternatively, the B4 and B12 Exon models may be examples of Genscan and Grail both predicting false positive patterns.

Lastly, the estimation of approximately eighty genes within this 6.5 Mb permits revised calculations of gene numbers. Information from the isochore analysis of the mammalian genome correlates the base composition of a region with gene density. Specifically, Bernardi (1995) and Zoubak et al. (1996) have shown that the most GC-rich regions of the genome (>48% GC), the H3 isochores, comprise approximately 5% of the genome, the moderately GC-rich regions (43–48% GC), the H1 and H2 isochores, comprise approximately 25%, and the most AT-rich regions, the L isochores (<43% GC), comprise the bulk of the genome, 67%. Gene densities in the isochore classes H3, H1/H2 and L have been determined at one per 9 kb, one per 54 kb and one per 150 kb, respectively, based on whole genome analysis (Zoubak et al., 1996). Given these proportions and densities, the number of megabases of each isochore expected within 21q, the number of megabases of each isochore analyzed here, and the gene density, expected and observed, within each, can be calculated (Table 6). From these data, a total of 400–500 genes within 21q can be estimated. This is at the low end of estimates of 500–1000 based on the size of chromosome 21 and assuming 50 000–100 000 genes in the whole genome. One other organizational feature is of interest. The high GC region, H3 isochore, gene density found on chromosome 21 is significantly different from that found in the whole genome. Here, the 21q gene density is very low, only one gene per 50 kb (Table 1, segment numbers 103, 104 and 171) versus the expected. This could be due to some feature of chromosome 21 genes, e.g. an average larger size although still GC-rich, or a feature that results in consistent failure in exon prediction. If the former is true, 21q is indeed relatively gene-poor, consistent with its association with viable trisomy. If the latter is true, many genes remain to be discovered, even in the 0.6 Mb of H3 isochore analyzed here.

Acknowledgements

This is a contribution (#1754) of the Thomas G. and Mary W. Vessels Laboratory for Molecular Biology and the John C. Mitchell Laboratory for the Study of Genetic Diseases and Human Development of the Eleanor Roosevelt Institute. The authors thank Andrew Fortna and Roger Lucas for excellent technical assisance and Richard Mural for helpful discussions. This work was supported by NIH grant HD17449 and by the Boettcher Foundation.

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Antonarakis, S.E., 1998. 10 years of genomics, chromosome 21 and Down syndrome. Genomics 51, 1–16.

Bernardi, G., 1995. The human genome: organisation and evolutionary history. Annu. Rev. Genet. 29, 445–476.

Burset, M., Guigo, R., 1996. Evaluation of gene structure prediction programs. Genomics 34, 353–367.

Chen, H.M., Chrast, R., Rossier, C., Morris, M.A., Lalioti, M.D., Antonarakis, S.E., 1996. Cloning of 559 potential exons of genes of human chromosome 21 by exon trapping. Genome Res. 6, 747–760.

Cheng, J.F., Boyartchuk, V., Zhu, Y.W., 1994. Isolation and mapping of human chromosome 21 cDNA: Progress in constructing a chromosome 21 expression map. Genomics 23, 75–84.

Chomzynski, B., Sacchi, N., 1987. Rapid RNA isolation. Anal. Biochem. 162, 156–159.

Claverie, J.-M., 1997. Computational methods for the identification of genes in vertebrate genomic sequences. Hum. Mol. Genet. 6, 1735–1744.

Dahmane, A., Ait Ghezala, G., Gosset, P., Chamoun, Z., Dufresne-Zacharia, M.C., Lopes, C., Rabatel, N., Gassanova-Maugenre, S., Chettouh, Z., Abramowski, V., Fayet, E., Yaspo, M.L., Korn, B., Blouin, J.L., Lehrach, H., Poustka, A., Antonarakis, S.E., Sinet, P.M., Creau, N., Delabar, L.M., 1998. Transcriptional map of the 2.5-Mb Down syndrome chromosomal region (DCR1). Genomics 48, 12–23.

Deloukas, P., et al., 1998. A physical map of 30,000 human genes. Science 282, 744–746.

Feinberg, A., Vogelstein, B., 1984. A technique for radiolabelling DNA to high specific activity. Anal. Biochem. 137, 266–268.

Fickett, J.W., Tung, C.S., 1992. Assessment of protein coding measures. Nucleic Acids Res. 20, 6441–6450.

Gardiner-Garden, M., Frommer, M., 1987. CpG islands in vertebrate genomes. J. Mol. Biol. 196, 261–282.

Gardiner, K., 1997. Clonability and gene distribution on human chromosome 21: reflections of junk DNA content? Gene 205, 39–46.

Guipponi, M., Scott, H.S., Chen, H., Schebesta, A., Rossier, C., Antonarakis, S.E., 1998. Two isoforms of a human Intersectin (ISTN) protein are produced by brain-specific alternative splicing in a stop codon. Genomics 53, 369–376.

Harris, N.L., 1997. Genotator: A workbench for sequence annotation. Genome Res. 7, 754–761.

Heiss, N.S., Poustka, A., 1997. Genomic structure of a novel chloride channel gene, CLIC2, in Xq28. Genomics 45, 224–228.

Korobko, I.V., Kabishev, A.A., Kiselev, S.L., 1997. Identification of the new protein kinase specifically transcribed in mouse tumors with high metastatic potential. Dokl. Adad. Nauk. 354, 554–556.

Larsen, F., Gunderson, G., Lopez, R., Prydz, H., 1992. CpG islands as gene markers in the human genome. Genomics 13, 1095–1107.

Nagamine, K., Kudoh, J., Minoshima, S., Kawasaki, K., Asakawa, S., Ito, F., Shimizu, N., 1996. Isolation of cDNA for a novel human protein KNP-I that is homologous to the *E. coli* SCRP-27A protein from the autoimmune polyglandular disease type I (APECED) region of chromosome 21q22.3. Biochem. Biophys. Res. Commun. 225, 608–616.

Nagamine, K., Kudoh, J., Kawasaki, K., Minoshima, S., Asakawa, S., Ito, F., Shimizu, N., 1997a. Genomic organization and complete nucleotide sequence of the TMEM1 gene on human chromosome 21q22.3. Biochem. Biophys. Res. Commun. 235, 185–190.

Nagamine, K., Kudoh, J., Minoshima, S., Kawasaki, K., Asakawa, S., Ito, F., Shimizu, N., 1997b. Genomic organization and complete nucleotide sequence of the human PWP2 gene on chromosome 21. Genomics 42, 528–531.

Ohira, M., Seki, N., Nagase, T., Suzuki, E., Nomura, N., Ohara, O., Hattori, M., Sakaki, Y., Eki, T., Murakami, Y., Saito, T., Ichikawa, H., Ohki, M., 1997. Gene identification in 1.6-Mb region of the Down syndrome region on chromosome 21. Genome Res. 7, 47–58.

Sambrook, J., Fritsch, E.F., Maniatis, E., 1989. Molecular Cloning: A Laboratory Manual. second ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Tassone, F., Xu, H., Burkin, H., Weissman, S., Gardiner, K., 1995. cDNA selection from 10 Mb of chromosome 21 DNA: Efficiency in transcriptional mapping and reflections of genome organization. Hum. Mol. Genet. 4, 1509–1518.

Tatusova, T., Madden, T., 1999. Blast2 sequences — a new tool for comparing protein and nucleotide sequences. FEMS Microbiol. Lett. 174, 247–250.

Yaspo, M.L., Gellen, L., Mott, R., Korn, B., Nizetic, D., Poustka, A.M., Lehrach, H., 1995. Model for a transcript map of human chromosome 21: Isolation of new coding sequences from exon and enriched cDNA libraries. Hum. Mol. Genet. 4, 1291–1304.

Zoubak, S., Clay, O., Bernardi, G., 1996. The gene distribution of the human genome. Gene 174, 95–102.