Large-scale sequencing in human chromosome 12p13: experimental and computational gene structure determination.

M A Ansari-Lari, Y Shen, D M Muzny, et al.

Genome Res. 1997 7: 268-280
Access the most recent version at doi:10.1101/gr.7.3.268

RESEARCH

Large-Scale Sequencing in Human Chromosome 12p13: Experimental and Computational Gene Structure Determination

M. Ali Ansari-Lari,[1] Ying Shen, Donna M. Muzny, William Lee, and Richard A. Gibbs

Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030

The detailed genomic organization of a gene-dense region at human chromosome 12p13, spanning 223 kb of contiguous sequence, was determined. This region is composed of 20 genes and several other expressed sequences. Experimental tools including RT–PCR and cDNA sequencing, combined with gene prediction programs, were utilized in the analysis of the sequence. Various computer software programs were employed for sequence similarity searches and functional predictions. The high number of genes with diverse functions and complex transcriptional patterns make this region ideal for addressing challenges of gene discovery and genomic characterization amenable to large-scale sequence analysis.

[The sequence data described in this paper have been submitted to the GenBank data library under accession nos. U72506–U72518.]

One of the challenges of the human genome project is to keep pace with the analysis of genomic sequence, which is being generated at an increasing rate (Gibbs 1995). Several computational software tools have been developed to speed up and improve the process, yet the complexity of the genome makes total reliance on these tools impractical. Although exon prediction/gene modeling programs can be useful for localizing genes from sequence, and in many cases can correctly predict individual exons, they cannot be used reliably as the only means for elucidation of gene structure (Burset and Guigo 1996). Sequences with more than one gene, genes with alternative splicing, and other complex transcriptional units can be particularly problematic (Burset and Guigo 1996).

Among other initiatives, generation of a large data base of expressed sequence tags (ESTs) has greatly improved the process of gene identification (Hillier et al. 1996). UniGene has also facilitated grouping of overlapping ESTs and unique transcriptional sequences into clusters that represent the transcriptional products of distinct genes (Boguski and Schuler 1995). ESTs are of sufficient sequence quality for generating transcription maps of the genome and for verifying some of the predictions of gene-finding programs. However, ESTs represent a single-pass sequence of 3′ and 5′ (or partial 5′) end of genes and hence are not sufficient for accurate gene structure determination.

Here, we describe the initial steps for the precise determination of gene structures from genomic sequence of a highly gene-packed region on human chromosome 12p13, using experimental tools such as RT–PCR and cDNA sequencing, as well as nonexperimental computational approaches. This region contains several important genes, including sequences involved in signal transduction (*CD4, GNB3, PTPN6*), glycolysis (*TPI* and *ENO2*), and a neurodegenerative disorder (*DRPLA*). This region serves as an excellent model for addressing the strengths and weaknesses of the current tools for large-scale genomic sequence analysis of regions with high gene density and complex transcriptional units.

RESULTS

In this study 106-kb of new genomic sequence was obtained with ≥99.99% accuracy. This extends the contiguous genomic sequence we generated previously on chromosome 12p13 (Ansari-Lari et al. 1996a) to ~223 kb (Fig. 1). Additionally, 11 kb of cDNA sequence was determined to help in characterizing several of the genes in this region (Table 1).

[1]Corresponding author.
E-MAIL ma029926@bcm.tmc.edu; FAX (713) 798-5741.

7:268–280 ©1997 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/97 $5.00

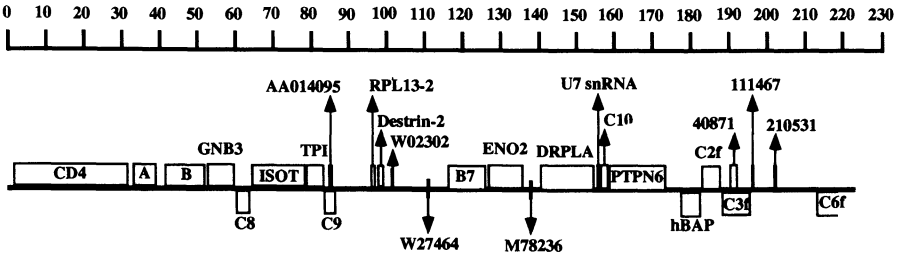

Figure 1 Schematic representation of a gene-rich cluster at human chromosome 12p13. The drawing is to scale (±100 bp). Genes are shown as open boxes, and ESTs as solid boxes. Orientations of the transcriptions are shown based on the position of the boxes above or below the solid line. Two of the ESTs, where the orientation of their transcription is not known, are drawn in the middle of the line.

The steps used for the analysis of the sequence are depicted in Figure 2.

First Pass to Screen for Genes

Through the use of PowerBLAST (Zhang 1996), several hundred EST matches were identified in this region, whereas GRAIL 2 (Xu et al. 1994) predicted several putative exons. This process greatly enhanced preliminary screening for identification of genes and transcribed regions. On the basis of these analyses, the region appeared to be highly gene rich. Although, in most cases the 3′ end of EST matches could be distinguished, the number of genes could not be determined accurately because of alternative polyadenylation signals and the presence of EST artifacts. Additionally, prediction of polyadenylation signals and promoter regions by GRAIL 2 were inaccurate. Therefore, except for the few genes corresponding to published complete cDNA sequences, defining the exact number and exon/intron organization of the genes required much more detailed analyses. When the studies detailed below were completed, we determined that EST matches had been identified for all of the genes, and GRAIL 2 predicted some of the putative exons for all of the genes except U7 small nuclear RNA (snRNA), which is the RNA component of a small ribonucleoprotein involved in post-transcriptional 3′-end processing of histone mRNA (Mowry and Steitz 1987). In addition, no tRNA gene could be identified using the tRNAscan-SE (v.1.0) program (Lowe and Eddy 1997).

Exon/Intron Organization of the Genes

On the basis of information from PowerBLAST and GRAIL 2, three general strategies were subsequently pursued for precise determination or accurate prediction of exon/intron organization of the genes. The first relied on characterization of cDNA clones corresponding to the EST matches that were accessible from the IMAGE consortium (Hillier et al. 1996). The sequencing of these clones provided more detailed transcriptional information (Table 1). Second, primers for RT–PCR were designed on the basis of exon prediction programs and EST matches to obtain the portion of transcripts not present in the cDNA clones. The RT–PCR products were then sequenced (Table 1). Compared with cDNA clone sequencing, RT–PCR provided the advantage of detecting alternative splicing patterns, as demonstrated for gene *B7* (Fig. 3A). Third, in the absence of RT–PCR or cDNA sequence information, predictions from GRAIL 2 and FGENEH (Solovyev et al. 1995) exon prediction/gene modeling programs were used, provided they were consistent with the available partial mouse and human EST sequences (Table 1).

This 223 kb of sequence contains 20 genes and potentially 4 more, making it one of the most gene-rich regions described to date, with an average of 1 every 10 kb. The intergenic distance between some of the genes in this region is as small as 100 bp. The genomic organization of *CD4, A, B, GNB3, TPI,* and *ISOT* genes was described previously (Ansari-Lari et al. 1996a). The intron/exon organizations of *ENO2, DRPLA, PTPN6,* and *U7* snRNAs were obtained on the basis of available published cDNA and/or genomic sequences (Table 1). For *B7, hBAP, C2f, C3f, C6f, RPL13-2,* and *Destrin-2* the exon/intron boundaries were determined by sequencing RT–PCR products and cDNA clones (Table 1). Using available partial mouse and human EST sequences in combination with GRAIL 2 and FGENEH predictions, the exon/intron organizations of *C8, C9,* and *C10* genes were predicted. Finally, six other EST matches were identified (Fig. 1). The sequences of two cDNA clones, 111467 and 210531, corresponding to two of the ESTs, were obtained. Both matched the genomic sequence without interruption by any intron. No identifiable long open reading frame (ORF) is present in either of the two sequences. Furthermore, both clones contain an *Alu* element and are flanked at one of their ends by another *Alu* element.

Table 1. Structural Features of the Genes at Human Chromosome 12p13

Gene	Number of exons	Length of cDNA (bp) (putative or determined)	Source of cDNA sequence	Accession nos.	References
C8	6	1247	available partial EST sequences; FGENEH	—	—
C9	2	877	available partial EST sequences; GRAIL-2 and FGENEH	—	—
RPL13-2	1	586	RT–PCR	U72513	Adams et al. (1992); Olvera and Wool (1994)
Destrin-2	1	1057	RT–PCR	U72518	Moriyama et al. (1990); Hawkins et al. (1993)
B7	7	1208	RT–PCR and IMAGE cDNA clone 302632	U72508 (note: U72509 and U72510 for alternatively spliced forms)	—
ENO2	12	2274	database	X51956; M22349	Oliva et al. (1991)
DRPLA	10	4339 (v)	database	D31840; U23851	Koide et al. (1994); Nagafuchi et al. (1994); Onodera et al. (1995)
U7 snRNA	1	64	database and comparison with the *M. musculus U7* snRNA sequence	M17910; X54165	Mowry and Steitz (1987); Soldati and Schumperli (1988); Philips and Turner (1991)
C10	3	537	available partial EST sequences; FGENEH	—	—
PTPN6	17	2158; 2156	database	U15528–U15537; M77273; X82818; X82817	Shen et al. (1991); Plutzky et al. (1992); Banville et al. (1995)
hBAP	9	1239	IMAGE cDNA clone 176864, and comparison with mouse *BAP37* sequence	U72511 (note; alternatively spliced form U72512)	Nuell et al. (1991); Sato et al. (1993); Terashima et al. (1994)
C2f	6	885	IMAGE cDNA clone 139446	U72514	—
C3f	12	1842	RT–PCR and IMAGE clone 188144	U72515 (note: U72507 is an alternatively spliced form and U72507 for the clone	—
C6f	I(>2)	I	IMAGE cDNA clone 113390	U72516	—

The GenBank accession no. for the genomic sequence is U72506. (I) incomplete; (v) variable. The structural features related to genes *CD4, A, B, GNB3, ISOT,* and *TPI* have been described previously (Ansari-Lari et al. 1996a)

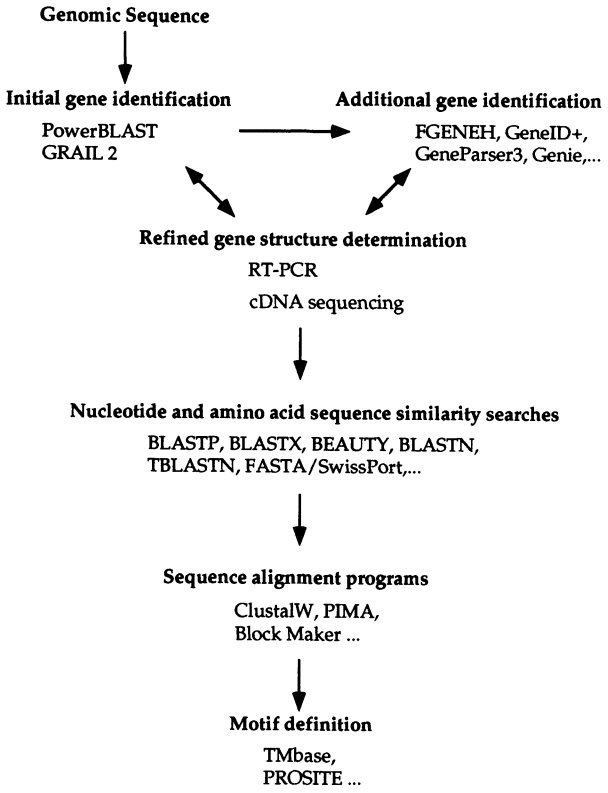

Figure 2 Stepwise analysis of genomic sequence.

The data suggest that these two cDNA clones do not represent bona fide expressed genes and are likely artifacts of cDNA library construction. Finally, of the other four EST matches (W02302, W27464, M78236, AA014095) that were identified by PowerBLAST, W02302 has a portion of an *Alu* element at its 5′ end and AA014095 appears to be in the opposite transcriptional direction of *C9*. Gene prediction programs in combination with available EST sequences do not predict a consistent and clear gene organization of these ESTs (as opposed to *C8, C9,* and *C10*). Additional analyses of these four ESTs are necessary to determine their relevance.

Exon Prediction by Computational Programs

After determining the exact intron/exon organization of the genes by obtaining cDNA and RT–PCR sequences, the performances of GRAIL 2, FGENEH, and GeneID (Guigo et al. 1992) on exon predictions at the nucleotide level were assessed, using sequence fragments each containing a single gene. This test was performed to compare the accuracy of these programs with the information obtained from RT–PCR and cDNA sequencing. With this test sequence set, all three programs missed at least one true exon and predicted only splice donor (SD) or splice acceptor (SA), but not both, for one or more exons (Table 2A,B). In addition, for some of the genes, other exons that were not part of the transcripts were predicted (apparent false positives). In general, more true exons, with fewer false positives were predicted by FGENEH and GRAIL 2 than by GeneID (Table 2A,B).

Sequence Similarity Searches

Following determination or prediction of the transcripts of the genes from this region, both the nucleotide sequence of the transcripts and their predicted conceptual translations were analyzed with various computer software programs, including BLASTP, BLASTX, BEAUTY, BLASTN, TBLASTX, and PROSITE (Smith et al. 1996), to identify related sequences by similarity. Some genes, such as *U7* snRNA are not translated, making nucleotide sequence searches necessary. Additionally, in the presence of insertions, deletions, or mismatches in the nucleotide sequence, an erroneous translation product can be predicted, reducing the power of protein similarity searches. Concordance between nucleotide and protein similarity searches therefore provide the most confident data.

The related sequences were aligned to predict conserved domains or functional motifs, using various sequence alignment programs, including ClustalW (Thompson et al. 1994), PIMA (Smith and Smith 1992), BLOCK MAKER (Henikoff et al. 1995), and BestFit from GCG (sequence analysis of software package, v. 8, Genetics Computer Group, Madison, WI). The alignments, combined with motif definition programs such as PROSITE (Bairoch 1993) and transmembrane prediction programs such as TMBASE (Hofmann and Stoffel 1993), enhanced the prediction of notable motifs.

These similarity searches provided insight into the possible functional roles of several of these genes. The human B-cell receptor-associated protein (hBAP), for example, was characterized by its similarity to other prohibitins. The hBAP protein shows 99.7% identity to the mouse BAP37 (mBAP37) protein (Terashima et al. 1994), 52% identity (73% similarity) to the human prohibitin gene (*PHB*) (Sato et al. 1992), and 38% identity (58% similarity) to the *Cc* gene product in *Drosophila melanogaster* (Eveleth and Marsh 1986) (Fig. 4A). The program TMBASE, predicts one putative transmembrane domain for hBAP and *PHB* at the amino termini of both proteins in support of membrane localization. Likewise, the predicted amino acid sequence for the *C2f* gene shows 72% similarity (55% identity) to a

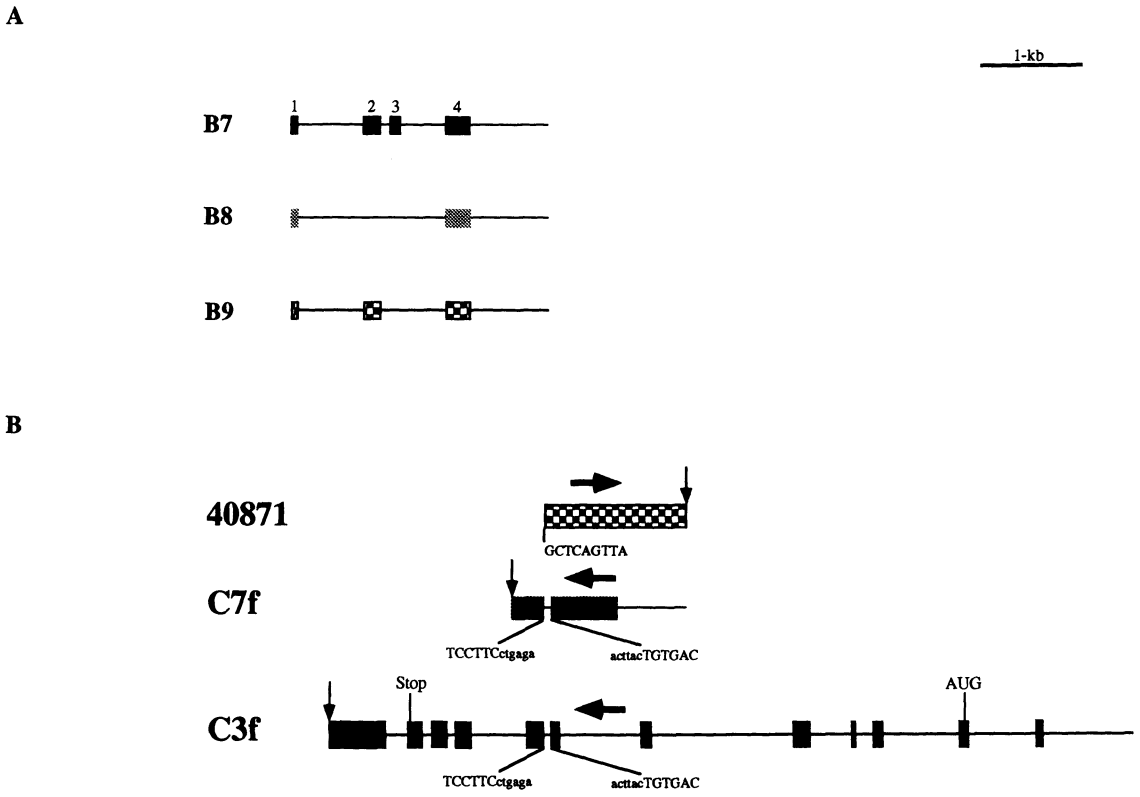

Figure 3 (*A*) Alternative splicing in gene *B7*. B8 and B9 are partial transcripts that were obtained by RT–PCR. The three alternatively spliced forms possess different putative amino termini, but the same reading frame is maintained in all three forms. (*B*) Alternative splicing and transcriptional pattern at gene *C3f*. The horizontal arrows show the direction of the transcripts. The vertical arrows show the polyadenylation sites. In *A* and *B*, boxes represents exons. The drawings are to scale. The relative positions of the exons are to scale.

Saccharomyces cerevisiae hypothetical protein L9470.5 and 71% similarity (54% identity) to the *Schizosaccharomyces pombe* hypothetical 34.9-kD protein, as shown in Figure 4B. On the basis of the absence of transmembrane prediction by the TMBASE program, the members of this gene family appear to encode intracellular proteins. Among several PROSITE pattern predictions, only two protein kinase C phosphorylation sites are present in all three members (Fig 4B). Also, C9 shows 57% similarity to amino acid sequence of a *Caenorhabditis elegans* protein (NCBI no. g532806). Among several PROSITE pattern predictions, a G-β repeat and an amidation site are in common between the two proteins. Although the B7 protein does not show amino acid sequence similarity to any known protein, it is highly leucine rich (16%), contains one putative leucine zipper domain, and has a high number of aspartic acid and glutamic acid residues (27 of 51) at its amino terminus. Finally, the C3f protein shows between 50% and 54% similarity and 25% and 29% identity to several proteins in *S. cerevisiae* and *C. elegans*. Comparison of these protein sequences is depicted in Figure 4C. Of the several annotated sites predicted by PROSITE, only one *N*-myristoylation site is in common among all five proteins. TMBASE program predicts between 8 and 10 transmembrane domains for these proteins (8 transmembrane domains for C3f).

Other Transcriptional Regions

Two transcriptionally active regions, *Destrin-2* and *RPL13-2*, with high sequence identity to transcripts from two known genes, were identified. By RT–PCR, the partial transcript for *Destrin-2* was isolated (absent in RT minus reactions) and sequenced. The *Destrin-2* sequence is highly similar to the transcript from the human actin depolymerizing factor gene (*ADF*, also known as *Destrin*) and cofilin (Moriyama et al. 1990; Hawkins et al. 1993). A putative nuclear localization-like motif (PEEIKKRTKAV) and a putative F-actin binding motif (DAIKKK) is present in the conceptual translation of *Destrin-2*. A similar

Table 2. Performance of GRAIL2, FGENEH, and GENEID on Exon Predictions

A. Comparison of exon predictions

Gene ID	GRAIL 2 Exons predicted and present	GRAIL2 Exons present but not predicted	GRAIL 2 Number of exons predicted but not present	FGENEH Exons predicted and present	FGENEH Exons present but not predicted	FGENEH Number of exons predicted but not present	GeneID Exons predicted and present	GeneID Exons present but not predicted	GeneID Number of exons predicted but not present
CD4	1, 2p, 3-9	10	3	2p, 3, 7, 8p, 9	1, 4-6, 10	7	5p, 6, 7, 8p	1-4, 9, 10	2
A	3, 4, 5	1, 2	2	1, 3p, 4p, 5	2	1	1	2-5	1
B	1, 2, 4, 5, 8, 10-14	3, 6, 7, 9	1	2p, 4, 5, 6p, 7p, 8-14	1, 3	-	3, 5, 8-13	1, 2, 4, 6, 7, 14	2
GNB3	3p, 4, 6-10	1, 2, 5, 11	2	3p, 4-9, 10p	1, 2, 11	-	7p, 8, 10, 11p	1-6, 9	2
ISOT-1	1,2, 3p, 5-8, 9p, 10, 11, 12p, 13, 14p, 15-17, 18p, 19, 20	4	2	1-8, 10-20	9	-	11p, 12, 13, 19, 20	1-10, 14-18	1
TPI	1, 2p, 4-7	3	-	1, 4-7	2, 3	-	2-6, 7p	1	2
B7	2p, 3p, 4p, 5p, 6	1, 7	1	3p, 4	1, 2, 5-7	1	2p, 3, 4, 5p	1, 6,7	6
ENO-2	2p, 3-6, 7p, 8p, 9, 10p, 11p, 12	1	2	2p, 3-12	1	-	4-7, 8p	1-3, 9-12	3
DRPLA	3p, 4, 5p, 6p, 9	1, 2, 7, 8, 10	3	2p, 3, 4, 5p, 6, 7p, 8-10	1	1	-	1-10	2
PTPN6	2p, 3, 4p, 5p, 6-11, 12p, 13, 14p	1, 15	-	1-7, 9-12, 13p	8, 14, 15	-	2-4, 5p, 7, 9, 11-13, 14p, 15	1, 6, 8, 10	1
C2f	2p, 3-6	1	-	2-6	1	1	3p, 4, 5	1, 2, 6	1
hBAP	1, 2p, 3, 5p, 6p, 8p	4, 7, 9	1	1-3, 5, 6p, 7p, 8p, 9	4	1	1, 3, 4p, 5, 6p, 9p	2, 7, 8	2
C3f	2p, 3-5, 6p, 7-9, 10p	1, 11, 12	-	2-5, 9, 10p, 12	1, 6-8, 11	-	4-9, 12	1-3, 10, 11	2

B. Summary of exon predictions

EXONS	Exon prediction / Gene modeling programs		
	GRAIL 2	FGENEH	GeneID
Predicted	54%	64%	35%
Partially predicted	24%	14%	11%
Not predicted	22%	22%	54%
Total number of apparent false positives	17	12	27

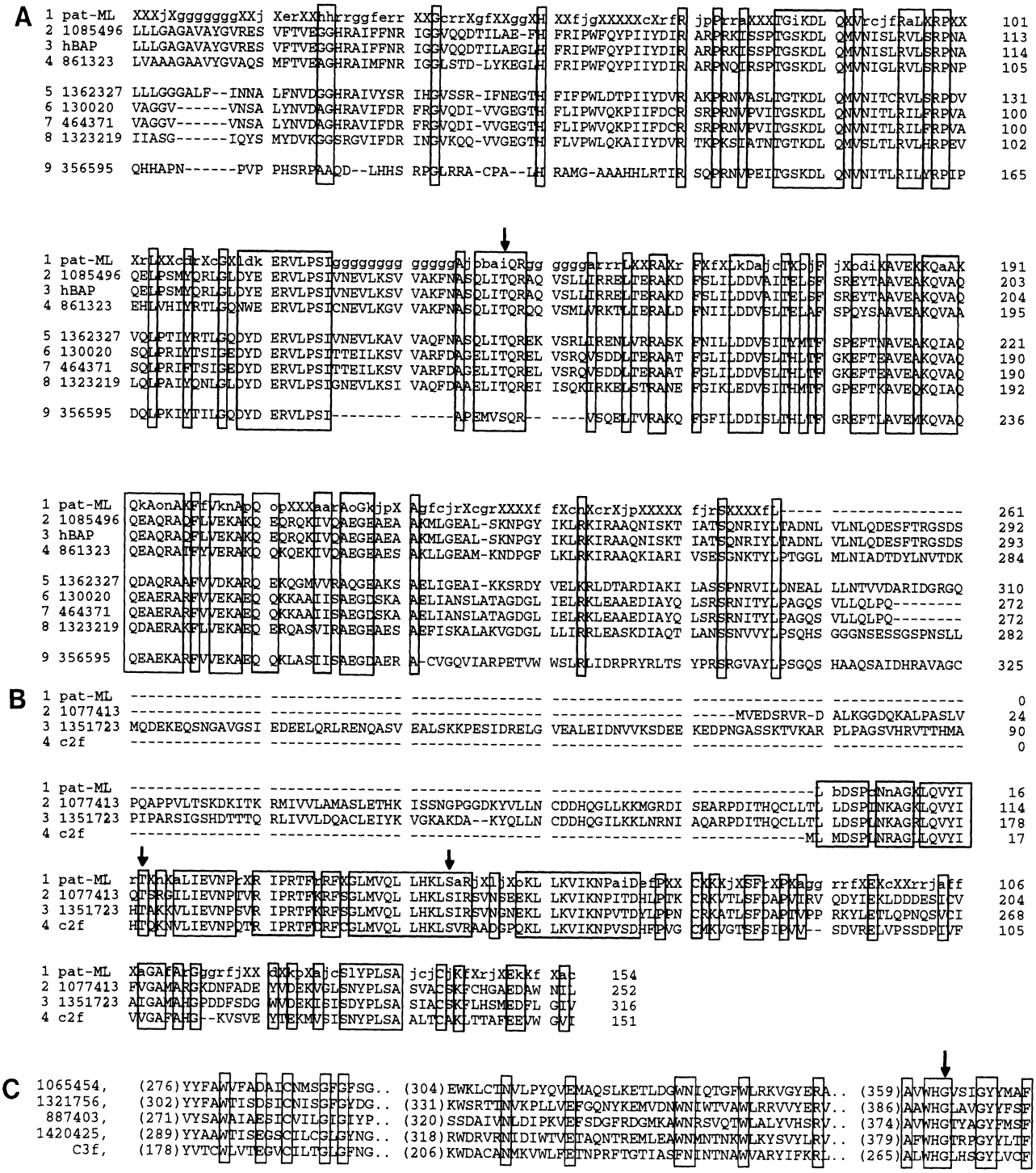

Figure 4 Multiple sequence alignments for identification of conserved domains. (*A*) The amino acid sequence alignment for *hBAP* and its related proteins was generated by PIMA. The NCBI protein sequence numbers are to the *left* (1085496, *Mus musculus;* 861323, *C. elegans;* 1362327, *S. cerevisiae;* 130020, *M. musculus;* 464371, 1323219, *S. cerevisiae;* 356595, *Drosophila;* 464371, *Homo sapiens*). The boxed sequences are either identical or highly similar amino acids. (pat-ML) The similarity annotation/score assigned by PIMA (http://dot.imgen.bcm.tmc.edu:9331/multi-align/Options/pima.html). (*B*) The sequence alignment of the *C2f* gene and two related proteins was generated by PIMA. The NCBI protein sequence numbers are to the *left* (1077413, *S. cerevisiae;* 1351723, *S. pombe*). (*C*) The conserved motifs predicted by BLOCK for gene *C3f* and related sequences. The NCBI protein sequence numbers are to the *left* (1065454, 1321756, and 887403, *C. elegans;* 1420425, *S. cerevisiae*). The arrow indicates the position of the conserved putative protein kinase C phosphorylation sites for *A* and *B,* and the N-myristylation site for *C,* all predicted by PROSITE.

nuclear localization motif (PEEIKKRKKAV) and the identical F-actin binding motif are present in cofilin and ADF (Moriyama et al. 1990). However, some of *Destrin-2* features are not present in the *Destrin* gene. Because of a nucleotide difference, the first methionine of Destrin-2 is 16 residues downstream of the first methionine of ADF. In addition, an inverted *Alu* element is present only at the 3′ end of *Destrin-2* transcript, and part of this *Alu* encodes the putative carboxyl terminus of the *Destrin-2* gene. No intron is present in the genomic sequence corresponding to *Destrin-2* cDNA. The genomic organization of the *Destrin* gene is not known.

The genomic sequence of *RPL13-2* shows a high degree of identity to the cDNA sequence of the gene encoding protein L13 of the 60S ribosomal subunit (*RPL13*) of human, rat, and several other species (Adams et al. 1992; Olvera and Wool 1994). In humans, *RPL13* (originally identified as the breast basic conserved gene or *bbc1*) is located on chromosome 16 (Adams et al. 1992). An RT–PCR product (absent in RT minus reaction) corresponding to the *RPL13-2* transcript was sequenced. There is 73% similarity and 67% identity between the amino acid sequence of the human *RPL13* and the putative translation product of *RPL13-2*. Some of the features of *RPL13-2* are not present in the *RPL13* gene. A stretch of adenosine is only present at the 3′ end of the *RPL13-2* genomic sequence with no identifiable poly(A) signal consensus. The cDNA sequence indicates absence of introns in the genomic sequence of *RPL13-2*. However, the genomic sequence of the human *RPL13* gene was found to have at least one intron (Adams et al. 1992), whereas the *Drosophila melanogaster* homolog of the *RPL13* gene (*Dmbbc1*) was found to have two introns, with one being the same as the intron in the human *RPL13* gene (Helps et al. 1995).

As more and more EST sequences are generated, ESTs overlapping the same genomic region, yet in opposite transcriptional orientation, can be identified. These ESTs can occasionally complicate the determination of the correct genomic organization of genes. Additionally, partially spliced transcripts and cDNA library artifacts can impede the analysis of the sequence. The following examples illustrate these points.

Multiple transcripts were identified in the genomic region corresponding to *C3f,* as shown in Figure 3B. *C7f* likely represents the 3′ end of an alteratively spliced form of the *C3f* gene. The sequence of a cDNA clone (40871), which exactly matches its corresponding genomic sequence, is transcriptionally in the opposite orientation as the *C3f* transcript. The 40871 clone does not have any intron or identifiable ORF and contains a poly(A) signal and a poly(A) tail. The 40871 cDNA clone could represent genomic contamination in the cDNA library; however, there are several EST matches to this region from more than one cDNA library. Clone 40871 could represent a transcriptionally active processed pseudogene, yet the features associated with processed pseudogenes is not fully recognizable. There is no poly(A) tract at the 3′ end of the genomic sequence. A heptanucleotide repeat (CTCCTTC) is present 11 bases 5′ and 4 bases 3′ of the 40871 genomic DNA, which might be the remainder of direct repeats associated with pseudogenes. Finally, 40871 could represent the 3′-untranslated region of an upstream gene. RT–PCR reactions failed to show any connection between 40871 and its closest upstream gene, *C2f,* both of which are in the same transcriptional orientation.

DISCUSSION

In this study 223 kb of contiguous genomic sequence from human chromosome 12p13 was analyzed using several computer software programs, and the sequences of cDNA clones and RT–PCR products. This is one of the most gene-dense regions described to date, containing 20 genes and at least 4 more potential genes based on EST sequence matches. The genes in this region exhibit diverse functions, including interconversion of products in the glycolytic pathway (*ENO2* and *TPI*), participation in signal transduction pathways (*CD4, GNB3, PTPN6*), regulation of cell proliferation (*hBAP*), and ubiquitin-dependent proteolysis (*ISOT*). In addition, the expansion of a CAG repeat in the coding region of the *DRPLA* gene is associated with the dentatorubral and pallidoluysian atrophy neurodegenerative disorder (Koide et al. 1994; Nagafuchi et al. 1994).

Identification of genes with one or several EST matches dispersed among multiple matches of EST from high-level expressing genes was problematic in this study. PowerBLAST improved the sequence analysis because of its graphic interface, automatic masking of the repeat elements, various search options, and direct links to GenBank entries. However, precise determination of exon/intron organizations, resolving complex transcriptional units, distinguishing "real transcripts" from artifacts of cDNA libraries, and differentiating genes from pseudogenes each required considerable attention from less systematic approaches.

The accurate determination of the exon/intron organization of the genes described here provides a model for analysis of other sequences. Previous studies frequently have relied on gene modeling and exon prediction programs, as their speed and ease of use make them amenable to large-scale genomic sequence analysis. However, the false-positive and false-negative exon determinations can provide erroneous transcriptional and translational predictions. Some of the false predictions can be resolved when gene modeling/prediction programs are combined with the available partial EST sequences. Here, cDNA and RT–PCR isolation and sequencing were shown to provide the most accurate and informative route for defining gene organization and justified the effort required to generate the data.

The performance of several gene modeling/exon prediction programs were examined in this study. Although the performance of FGENEH and GRAIL 2 was superior to GeneID, this might be limited to the sequence set that was used in this study. Others have shown a closer performance among these programs with a different sequence set (Burset and Guigo 1996). GRAIL 2 can predict exons regardless of alternative splicing patterns or presence of multiple genes in a sequence fragment. The performance of gene modeling programs such as GeneID and FGENEH, however, can be greatly affected if a sequence fragment has multiple genes. The accuracy of exon prediction/gene modeling programs improves when they are combined with protein database searches, as this has been demonstrated for GeneID+ (Burset and Guigo 1996), GeneParser3 (Snyder and Stormo 1995), Genie (Kulp et al. 1996), and FGENEH (Solovyev et al. 1995).

We now adhere to the stepwise approach to genomic sequence analysis depicted in the top portion of Figure 2. First the sequence is analyzed with programs that only predict exons, such as GRAIL 2, and HEXON (Solovyev et al. 1994). GRAIL 2, as a part of the XGRAIL package, has the advantage of showing the exon predictions in a graphic format. This information is then combined with the output from PowerBLAST to define the approximate gene boundaries. Other gene modeling programs, such as FGENEH, and GeneID+ (Burset and Guigo 1996), can subsequently be employed for analysis of segments of sequence that likely contain a single gene. Finally, the cumulative information from PowerBLAST and the exon prediction/gene modeling program is used to design appropriate primers for RT–PCR and to obtain the cDNA clones corresponding to some of the ESTs for sequencing.

Because of the current shortcomings of gene modeling programs for precise gene structure determination, and because RT–PCR and sequencing the individual cDNA clones are less amenable to automation, large-scale mouse genomic sequencing and systematic large-scale full cDNA sequencing should be combined with the human genomic sequencing effort to improve the ease, speed, and accuracy of gene characterization. Several pilot studies have demonstrated the advantages of concurrent large-scale mouse genomic sequencing for the analysis of the human sequences (Koop and Hood 1994; Oeltjen et al. 1997). In addition to the prediction of exon/intron organization, the human and mouse sequence comparison can effectively reveal transcriptional control elements (Koop and Hood 1994; Oeltjen et al. 1997). Furthermore, the information gained from mouse genomic sequencing can be used in structural and functional studies amenable to this animal model. Additionally, the pilot studies for large-scale full cDNA sequencing schemes have demonstrated their feasibility, usefulness, and speed (Andersson et al. 1997). Availability of tissue-specific cDNA from characterized EST libraries makes it possible to accomplish this task in a reasonable time scale with great impact on the scientific community.

Distinguishing pseudogenes from functional genes, in genomic sequence can sometimes pose special problems. As more and more genomic and EST data are accumulated, the task should become easier. Pseudogenes exhibit sequence similarity to functional genes, but they cannot be translated into functional proteins because of mutations that may prevent transcription, interfere with splicing, terminate translation prematurely, or inactivate conserved functional domains. Some pseudogenes retain similar exon/intron organization as the corresponding functional genes. Other pseudogenes, also called processed pseudogenes, lack introns and are believed to be generated via the reverse transcription of the mRNA of functional genes. Processed pseudogenes usually have a poly(A) tail and are flanked by a variable number of direct repeats. Few transcriptionally active pseudogenes have been reported previously (Goeddel et al. 1981; Sorge et al. 1990). The glucocerebrosidase pseudogene appears to be transcriptionally active but translationally nonfunctional (Sorge et al. 1990). The human testis-specific phosphoglycerate kinase (*PGK-2*) has all the characteristics of a processed pseudogene, yet it is transcriptionally and translationally active. Therefore *PGK-2* is considered to be an active gene (McCarrey and Thomas 1987).

Compared with RPL13 and Destrin genes, *RPL13-2* and *Destrin-2* exhibit some of the structural features of pseudogenes. However, *RPL13-2* and *Destrin-2* appear to be transcriptionally active and their putative translation products could possess function. Therefore, more experimental evidence is needed to determine whether *RPL13-2* and *Destrin-2* are active genes or merely pseudogenes without biological function that have not lost all transcriptional activity.

One of the outcomes of large sequencing efforts is the better understanding of the evolutionary relationship of related genes. One model example from this region is the identification of *hBAP* as a member of the prohibitin gene family. Several functions have been proposed for *PHB* and its mouse homolog *BAP32,* including inhibition of cell proliferation, tumor suppression, regulation of cell cycle, and programmed cell death (Sato et al. 1992; Terashima et al. 1994). *PHB,* which is located on chromosome 17q21, is composed of seven exons (Sato et al. 1992). However, *hBAP* is composed of nine exons, and it does not show exon/intron organization similar to *PHB*. *hBAP* and *PHB* might therefore have evolved from two different ancestral genes. However, the data also support the introns-late theory of gene evolution (Cavalier-Smith 1991). According to this theory, introns did not exist in the precursor genes and they inserted randomly during eukaryotic evolution. Recent analysis of exon/intron organization of triose-phosphate isomerase gene among diverse eukaryotes support the introns-late theory (Kwiatowski et al. 1995; Logsdon et al. 1995). Another model example in this region is represented by *ENO2.* Enolases are dimeric enzymes of the glycolytic pathway. There are three distinct subunits, α, β, and γ encoded by three separate loci, *ENO1, ENO3,* and *ENO2,* respectively. *ENO1* and *ENO3* are mapped to the short arm of 1pter–p36.13 and to the short arm of chromosome 17, respectively (Oliva et al. 1991). All three enolases show similar exon/intron organization, suggesting their generation by duplication from a single ancestral gene (Oliva et al. 1991).

Determination of the genomic organization of this region is the first step toward understanding the transcriptional regulation of this highly gene-dense sequence. Cross-species sequence comparison should reveal the conserved transcriptional regulatory elements and help to resolve overlapping bidirectional transcriptional segments. Additionally, the availability of detailed genomic organization in combination with known polymorphic markers provides the opportunity for haplotype analysis with relevance to population genetic studies and association of disease phenotypes to the loci.

METHODS

Isolation and Sequencing of the PAC

An arrayed human genomic P1-derived artificial chromosome (PAC) library was purchased from Research Genetics, Inc. (Huntsville, AL). A PCR fragment corresponding to the 3′ end of cosmid J0 (Ansari-Lari et al. 1996a) was generated using the oligonucleotides R1460 (5′-GAGCCTCAGTATCCTCTTC-3′) and R1461 (5′-GAACTGGGTTGGAGTCTTG-3′). Hybridization was performed according to standard protocols (Sambrook et al. 1989). Two positive clones were obtained showing similar restriction maps and a minimally overlapping PAC was isolated by PCR screening (overlapping J0 by 6.8 kb; L0 PAC with the filter coordinates 1.1.19.P11).

The generation of a shotgun M13 sequencing library was performed as described (Andersson et al. 1996a). M13 template preparations and DNA sequencing were performed as described (Civitello et al. 1993; Muzny et al. 1994; Andersson et al. 1996b). Sequence assembly was performed by XGAP software (Bonfield et al. 1995).

Sequencing of cDNA Clones and RT–PCR Products

Total RNA and mRNA were isolated from the HT-1080 cell line and white blood cells using standard protocols (Sambrook et al. 1989). First-strand cDNA synthesis was performed as described (Ansari-Lari et al. 1996b), and PCR was performed according to standard protocols (Chelly and Kahn 1994). RT–PCR products were sequenced using dye terminator sequencing kits (Perkin Elmer). When possible, the cDNA clones and large RT–PCR fragments were concatenated and sequenced as described (Andersson et al. 1997). The sequence editing and assembly were performed using Sequencher for Macintosh, version 3.0 (Gibbs and Cockerill 1995). The following primers were used for RT–PCR reactions: gene *B7,* R2404 (5′-GAAAGTGGCGTGGTAACCAG-3′) and R2327 (5′-AGAGATGCCTTCAGTGTCAG-3′); gene *C3f,* R2376 (5′-GAATCAGAGGGGAGATGTG-3′), R2343 (5′-CTCACTACCTTTTGCTTCCAG-3′), R2407 (5′-TCTACAAGGAGACCTACCTC-3′), and R2344 (5′-GAGGAATTCCATCTGGAAGC-3′); *RPL13-2,* R2590 (5′-GCGTCCTCTTCTACAGCTTC-3′), and R2879 (5′-GCATGATCCTGAAGCCCAAC-3′); *Destrin-2,* R2570 (5′-ACGCTGCCCAGACTCACTAC-3′), R2571 (5′-CTTCTGGTCCGTTTGTTTGC-3′), R2626 (5′-AACATGAATGGCAAACAAAC-3′), and R2880 (5′-GGATAACATTAAACGTGGAAGC-3′).

Computer Analysis Programs

Initially, the repeat elements in the genomic sequence were masked by CENSOR (Jurka et al. 1995), and 1 to 2-kb segments of sequence were analyzed by BLAST (Altschul et al. 1990) against nonredundant GenBank, GenBank EST division, EMBL, DDBJ, and PDB sequences via mailFASTA, a relatively time consuming process. The availability of PowerBLAST (Zhang 1996) speeded the process, as PowerBLAST can automatically mask the repeat elements and graphically depict

various sequence matches from sequence databases including nonredundant sequences, STSs, and ESTs. The matched sequences and sequence alignments from desired segments can then be viewed.

GRAIL 2 (Xu et al. 1994), FGENEH (Solovyev et al. 1995), and GeneID (Guigo et al. 1992) gene prediction programs were employed for gene modeling. For the overall view of the sequence, GRAIL 2 was accessed as a part of XGRAIL package that, in a graphic format, depicts exons and allows prediction of some of the standard promoters, polyadenylation signals, and CpG islands.

An array of programs that have been assembled into a World Wide Web page by the Baylor College of Medicine (BCM) informatics group (Smith et al. 1996; http://gc.bcm.tmc.edu:8088/search-launcher.html) for DNA and protein sequence analyses were utilized. For detection of tRNA genes in the genomic sequence, tRNAscan-SE (v. 1.0) was employed (Lowe and Eddy 1997; http://genome.wustl.edu/eddy/tRNAscan-SE/). For sequence alignment, ClustalW (Thompson et al. 1994), PIMA (Smith and Smith 1992), BLOCK MAKER (Henikoff et al. 1995), and BestFit from GCG were used. Among all of the protein sequence analysis programs listed in the BCM web page (Smith et al. 1996), BLASTP + BEAUTY (Altschul et al. 1990; Worley et al. 1995), TBLASTN/nr dna, BEAUTY/CRSeqAnnot, FASTA/SwissPort (Pearson 1990), and PPSEARCH/PROSITE Pattern DB (Bairoch 1993) were found to be the most informative. The default search parameters were used for all of the programs.

Exon Predictions by Computational Programs

The performances of GRAIL 2, FGENEH, and GeneID programs were assessed on the basis of their ability to predict exons at the nucleotide level. Sequence was fragmented such that each fragment contained only one gene. If the coding domains were in the complementary strand, the reverse complement of the sequence was analyzed. GRAIL 2 (GRAIL@ornl.gov), FGENEH (mail -s fgeneh services@bchs.uh.edu), and GeneID (geneid@darwin.bu.edu) were accessed through the e-mail server. For GeneID, sequences were sent with the options "-small_output" and "-noexonblast" to switch off exon BLAST. For each exon, if both SD and SA sites were predicted correctly, the prediction of the exon was considered complete. If only SD or SA was predicted correctly, the prediction was labeled "p" (for partial prediction). If neither SD nor SA was predicted correctly, regardless of whether any portion of the coding domain was predicted, the exon was considered as not being predicted. For the first exon, if the SD was predicted correctly, the prediction was considered complete. For the last exon, if the SA was predicted correctly, the prediction was considered complete. For GeneID, exons in the top-ranking complete gene model were considered. For GRAIL 2, only exons labeled "good" or "excellent" were considered. For FGENEH, exons in the predicted gene model were considered. The sequence position of the fragments that were analyzed (see GenBank accession no. U72506) are as follow: *CD4* (1–33000), *A* (33001–40000), *B* (40001–52000), *GNB3* (52001–60000), *ISOT* (64001–79000), *TPI* (79001–83150), *B7* (116001–126400), *ENO2* (126401–135750), *DRPLA* (140301–154401), *PTPN6* (158501–174001), *hBAP* (reverse-complement of 174001–182800), *C2f* (182801–188200), and *C3f* (reverse-complement of 188201-196001). For gene *A, GNB3, ISOT,* and *PTPN6, A-2, GNB3-1, ISOT-1,* and *PTPN6,* nonhematopoietic splicing forms were considered, respectively. *C8, C9, C10, RPL13-2,* and *Destrin-2* were excluded from the analysis.

ACKNOWLEDGMENTS

We thank Wen Liu for assistance in construction of the PAC subclone library, Harley Gorrell for informatic support, and Michail Haywood, Kecia Rowland, and Lisa Perez for technical assistance. We especially thank Dr. Kim Worley and Dr. Kirsten Timms for helpful discussions during the course of this study. This work was supported in part by grant RO1 HG01459 from the National Center for Human Genome Research.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adams, S.M., N.R. Helps, M.G.F. Sharp, W.J. Brammar, R.A. Walker, and J.M. Varley. 1992. Isolation and characterization of a novel gene with differential expression in benign and malignant human breast tumors. *Hum. Mol. Genet.* **1:** 91–96.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Andersson, B., M.A. Wentland, J.Y. Ricafrente, W. Liu, and R.A. Gibbs. 1996a. A "double adaptor" method for improved shotgun library construction. *Anal. Biochem.* **236:** 107–113.

Andersson, B., J. Lu, K.E. Edwards, D.M. Muzny, and R.A. Gibbs. 1996b. Method for 96-well M13 DNA template preparations for large-scale sequencing. *BioTechniques* **20:** 1022–1027.

Andersson, B., J. Lu, Y. Shen, M. Wentland, and R.A. Gibbs. 1997. Simultaneous shotgun sequencing of multiple cDNA clones. *DNA Sequence* (in press).

Ansari-Lari, M.A., D.M. Muzny, J. Lu, F. Lu, C.E. Lilley, S. Spanos, T. Malley, and R.A. Gibbs. 1996a. A gene-rich cluster between the CD4 and triosephosphate isomerase genes at human chromosome 12p13. *Genome Res.* **6:** 314–326.

Ansari-Lari, M.A., S.N. Jones, K.M. Timms, and R.A. Gibbs. 1996b. Improved ligation-anchored PCR strategy for identification of 5′ ends of transcripts. *BioTechniques* **21:** 34–38.

Bairoch, A. 1993. The PROSITE dictionary of sites and patterns in proteins, its current status. *Nucleic Acids Res.* **21:** 3097–3103.

Banville, D., R. Stocco, and S.-H. Shen. 1995. Human

protein tyrosine phosphatase 1C (PTPN6) gene structure: alternate promoter usage and exon skipping generate multiple transcripts. *Genomics* **27:** 165–173.

Boguski, M.S. and G.D. Schuler. 1995. ESTablishing a human transcript map. *Nature Genet.* **10:** 369–371.

Bonfield, J.K., K.F. Smith, and R. Staden. 1995. A new DNA sequence assembly program. *Nucleic Acids Res.* **25:** 4992–4999

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

Cavalier-Smith, T. 1991. Intron phylogeny: A new hypothesis. *Trends Genet.* **7:** 145–148.

Chelly, J. and A. Kahn. 1994. RT-PCR and mRNA quantitation. In *The polymerase chain reaction* (ed. K.B. Mullis, F. Ferre, and R.A. Gibbs), pp. 97–109. Birkhauser, Boston, MA.

Civitello, A.B., S. Richards, and R.A. Gibbs. 1993. A simple protocol for the automation of DNA cycle sequencing reactions and polymerase chain reactions. *DNA Sequence* **3:** 17–23.

Eveleth, D.D.J. and J.L. Marsh. 1986. Sequence and expression of *Cc* gene, a member of the dopa decarboxylase gene cluster of *Drosophila*: Possible translational regulation. *Nucleic Acids Res.* **14:** 6169–6183.

Gibbs, R.A. 1995. Pressing ahead with human genome sequencing. *Nature Genet.* **11:** 121–125.

Gibbs, R.A. and M. Cockerill. 1995. Working on the assembly line. *Trends Biochem. Sci.* **20:** 162–163.

Goeddel, D.V., D.W. Leung, T.J. Dull, M. Gross, R.M. Lawn, R. McCandliss, P.H. Seeburg, A. Ullrich, E. Yelverton, and P.W. Gray. 1981. The structure of eight distinct cloned human leukocyte interferon cDNAs. *Nature* **290:** 20–26.

Guigo, R., S. Knudsen, N. Drake, and T.F. Smith. 1992. Prediction of gene structure. *J. Mol. Biol.* **226:** 141–157.

Hawkins, M., B. Pope, S.K. Maciver, and A.G. Weeds. 1993. Human actin depolymerizing factor mediates a pH-sensitive destruction of actin filaments. *Biochemistry* **32:** 9985–9993.

Helps, N.R., S.M. Adams, W.J. Brammar, and J.M. Varley. 1995. The Drosophila melanogaster homologue of the human BBC1 gene is highly expressed during embryogenesis. *Gene* **162:** 245–248.

Henikoff, S., J.G. Henikoff, W.J. Alford, and S. Pietrokovski. 1995. Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene* **163:** GC17–26.

Hillier, L., G. Lennon, M. Becker, M.F. Bonaldo, B. Chiapelli, S. Chissoe, N. Dietrich, T. DuBuque, A. Favello, W. Gish, M. Hawkins, M. Hultman, T. Kucaba, M. Lacy, M. Le, N. Le, E. Mardis, B. Moore, M. Morris, J. Parsons, C. Prange, L. Rifkin, T. Rohlfing, K. Schellenberg, M.B. Soares, F. Tan, J. Thierry-Meg, E. Trevaski, K. Underwood, P. Wohldman, R. Waterston, R. Wilson, and M. Marra. 1996. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* **6:** 807–828.

Hofmann, K. and W. Stoffel. 1993. TMBASE—A database of membrane spanning protein segments. *Biol. Chem. Hoppe-Seyler* **374:** 166.

Jurka, J., P. Klonowski, V. Dagman, and P. Pelton. 1995. CENSOR—A program for identification and elimination of repetitive elements from DNA sequences. *Comput. & Chem.* **20:** 119–122.

Kwiatowski, J., M. Krawczyk, M. Kornacki, K. Bailey, and F.J. Ayala. 1995. Evidence against the exon theory of genes derived from the triose-phosphate isomerase gene. *Proc. Natl. Acad. Sci.* **92:** 8503–8506.

Koide, R., T. Ikeuchi, O. Onodera, H. Tanaka, S. Igarashi, K. Endo, H. Takahashi, R. Kondo, A. Ishikawa, T. Hayashi, M. Saito, A. Tomoda, T. Miike, H. Natio, F. Ikuta, and S. Tsuji. 1994. Unstabel expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA). *Nature Genet.* **6:** 9–13.

Koop, B.F. and L. Hood. 1994. Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. *Nature Genet.* **7:** 48–53.

Kulp, D., D. Haussler, M.G. Reese, and F.H. Eeckman. 1996. *A generalized hidden Markov model for the recognition of human genes in DNA* (ed. D.J. States, P. Agarwal, T. Gaasterland, L. Hunter, and R. Smith), pp. 134–142. AAAI Press, Menlo Park, CA.

Logsdon, J.M.J., M.G. Tyshenko, C. Dixon, J.D.-Jafari, V.K. Walker, and J.D. Palmer. 1995. Seven newly discovered intron positions in the triose-phosphate isomerase gene: Evidence for the introns-late theory. *Proc. Natl. Acad. Sci.* **92:** 8507–8511.

Lowe, T.M. and S.R. Eddy. 1997. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequences. *Nucleic Acids Res.* **25** (in press).

McCarrey, J.R. and K. Thomas. 1987. Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene. *Nature* **326:** 501–505.

Moriyama, K., E. Nishida, N. Yonezawa, H. Sakai, S. Matsumoto, K. Iida, and I. Yahara. 1990. Destrin, a mammalian actin-depolymerizing protein is closely related to cofilin. *J. Biol. Chem.* **265:** 5768–5773.

Mowry, K.L. and J.A. Steitz. 1987. Identification of the human U7 snRNP as one of several factors involved in the 3′ end maturation of histone premessenger RNA's. *Science* **238:** 1682–1687.

Muzny, D.M., S. Richards, Y. Shen, and R.A. Gibbs.1994. PCR based strategies for gap closure in large-scale sequencing projects. In *Automated DNA sequencing and analysis* (ed. M.D. Adams, C. Fields, and J.C. Venter), pp. 182–190. Academic Press, San Diego, CA.

Nagafuchi, S., H. Yanagisawa, E. Ohsaki, T. Shirayama, K.

Tadokoro, T. Inoue, and M. Yamada. 1994. Structure and expression of the gene responsible for the triplet repeat disorder, dentatorubral and pallidoluysian atrophy (DRPLA). *Nature Genet.* **8:** 177–182.

Nuell, M.J., D.A. Stewart, L. Walker, V. Friedman, C.M. Wood, G.A. Owens, J.R. Smith, E.L. Schneider, R. Dell'Orco, C.K. Lumpkin, D.B. Danner, and J.K. McClung. 1991. Prohibitin, an evolutionary conserved intracellular protein that blocks DNA synthesis in normal fibroblasts and Hela Cells. *Mol. Cell. Biol.* **11:** 1372–1381.

Oeltjen, J.C., T.M. Malley, D.M. Muzny, W. Miller, R.A. Gibbs, and J.W. Belmont. 1997. Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res.* **11:** (in press).

Oliva, D., L. Cali, S. Feo, and A. Giallongo. 1991. Complete structure of the human gene encoding neuron-specific enolase. *Genomics* **10:** 157–165.

Olvera, J. and I.G. Wool. 1994. The primary structure of rat ribosomal protein L13. *Biochem. Biophys. Res. Common* **201:** 102–107.

Onodera, O., M. Oyake, H. Takano, T. Ikeuchi, S. Igarashi, and S. Tsuji. 1995. Molecular cloning of a full-length cDNA for dentatorubral-pallidoluysian atrophy and regional expressions of the expanded alleles in the CNS. *Am. J. Hum. Genet.* **57:** 1050–1060.

Pearson, W.R. 1990. Rapid and sensitive sequence comparisons with FASTP and FASTA. *Methods Enzymol.* **183:** 63–98.

Phillips, S.C. and P.C. Turner. 1991. Nucleotide sequence of the mouse U7 snRNA gene. *Nucleic Acids Res.* **19:** 1344.

Plutzky, J., B.G. Neel, and R.D. Rosenberg. 1992. Isolation of a src homology 2-containing tyrosine phosphatase. *Proc. Natl. Acad. Sci.* **89:** 1123–1127.

Sambrook, J., E.F. Fritsch, and T. Maniatis. 1989. *Molecular cloning: A laboratory manual,* 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Sato, T., H. Saito, J. Swensen, A. Olifant, C. Wood, D. Danner, T. Sakamoto, K. Takita, F. Kasumi, Y. Miki, M. Skolnick, and Y. Nakamura. 1992. The human prohibitin gene located on chromosome 17q21 is mutated in sporadic breast cancer. *Cancer Res.* **52:** 1643–1646.

Sato, T., T. Sakamoto, K.-I. Takita, H. Saito, K. Okui, and Y. Nakamura. 1993. The human prohibitin (*PHB*) gene family and its somatic mutations in human tumors. *Genomics* **17:** 762–764.

Shen, S.-H., L. Bastien, B.I. Posner, and P. Chretien. 1991. A protein-tyrosine phosphatase with sequence similarity to the SH2 domain of the protein-tyrosine kinases. *Nature* **352:** 736–739.

Smith, R.F. and T.F. Smith. 1992. Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for comparative protein modelling. *Protein Eng.* **5:** 35–41.

Smith, R.F., B.A. Wiese, M.K. Wojzynski, D.B. Davison, and K.C. Worley. 1996. BCM search launcher-an integrated interface to molecular biology data base search and analysis services available on the World Wide Web. *Genome Res.* **6:** 454–462.

Snyder, E.E. and G.D. Stormo. 1995. Identification of protein coding regions in genomic DNA. *J. Mol. Biol.* **248:** 1–18.

Soldati, D. and D. Schumperli. 1988. Structural and functional characterization of mouse U7 small nuclear RNA active in 3′ processing of histone pre-mRNA. *Mol. Cell Biol.* **8:** 1518–1524.

Solovyev, V.V., A.A. Salamov, and C.B. Lawrence.1994. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.* **22:** 5156–5163.

———. 1995. Identification of human gene structure using linear discriminant functions and dynamic programming. In *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology* (ed. C. Rawling, D. Clark, R. Altman, L. Hunter, T. Lengauer, and S. Wodak), pp. 367–375. AAAI Press, Cambridge, UK.

Sorge, J., E. Gross, C. West, and E. Beutler. 1990. High level transcription of the glucocerebrosidase pseudogene in normal subjects and patients with Gaucher disease. *J. Clin. Invest.* **86:** 1137–1141.

Terashima, M., K.-M. Kim, T. Adachi, P.J. Nielsen, M. Reth, G. Kohler, and M.C. Lamers. 1994. The IgM antigen receptor of B lymphocytes is associated with prohibitin and a prohibitin-related protein. *EMBO J.* **13:** 3782–3792.

Thompson, J.D., D.G. Higgins, and T.J. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighing, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.

Worley, K.C., B.A. Weise, and R.F. Smith. 1995. BEAUTY: An enhanced BLAST-based search tool that integrates multiple biological information resources into sequence similarity search results. *Genome Res.* **5:** 173–184.

Xu, Y., R.J. Mural, M.B. Shah, and E.C. Uberbacher. 1994. Recognizing exons in genomic sequence using GRAIL II. In *Genetic engineering: Principles and methods* (ed. J. Setlow), Vol. 16, pp. 241–253. Plenum Press, New York, NY.

Zhang, J. 1996. PowerBLAST: A new network BLAST application for genomic sequence analysis. In *Genome mapping & sequencing,* p. 19. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Received November 8, 1996; accepted in revised form January 24, 1997.