

Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project Author(s): Mark D. Adams, Jenny M. Kelley, Jeannine D. Gocayne, Mark Dubnick, Mihael H. Polymeropoulos, Hong Xiao, Carl R. Merril, Andrew Wu, Bjorn Olde, Ruben F. Moreno, Anthony R. Kerlavage, W. Richard McCombie, J. Craig Venter Source: *Science*, New Series, Vol. 252, No. 5013 (Jun. 21, 1991), pp. 1651-1656 Published by: <u>American Association for the Advancement of Science</u> Stable URL: <u>http://www.jstor.org/stable/2876333</u> Accessed: 22/03/2011 12:49

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <a href="http://www.jstor.org/page/info/about/policies/terms.jsp">http://www.jstor.org/page/info/about/policies/terms.jsp</a>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/action/showPublisher?publisherCode=aaas.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.



American Association for the Advancement of Science is collaborating with JSTOR to digitize, preserve and extend access to Science.

- E. R. Fearon et al., Science 247, 49 (1990).
   K. W. Kinzler et al., ibid. 251, 1366 (1991).
- 89. S. Srivastava, Z. Zou, K. Pirollo, W. Blattner, E. H. Chang, Nature 348, 747 (1990).
- S. Kwok and J. J. Sninsky, PCR Technology: Principles and Applications for DNA Amplification, H. Erlich, Ed. (Stockton, New York, 1989), pp. 235–244.
   T. J. White, R. Madej, D. H. Persing, Adv. Clin. Chem., in press; S. D. Williams and S. Kwok, in Laboratory Diagnosis of Viral Infections, E. H. Lennette, Ed. (Debter, New York, 100), pp. 1472.

- and S. KWOK, III Laboratory Diagnosis of Viral Infections, E. H. Lennette, Ed. (Dekker, New York, 1991), pp. 147–173.
  92. P. S. Walsh, D. A. Matzger, R. Higuchi, Biotechniques 10, 506 (1991).
  93. E. Blake, J. Mihalovich, R. Higuchi, P. S. Walsh, H. Erlich, unpublished data.
  94. J. Bill et al., J. Exp. Med. 169, 115 (1989); B. Budowle et al., Am. J. Hum. Genet. (1997). . 48, 137 (1991).
- 95. R. C. Allen, G. Graves, B. Budowle, Biotechniques 7, 736 (1989); K. Kasai et al., J. Forensic Sci. 35, 1196 (1990).
- 96. S. Scharf, unpublished data.
- 97. R. Helmuth et al., Am. J. Hum. Genet. 47, 515 (1990).
- 98. M. Stoneking, D. Hedgecock, R. Higuchi, L. Vigilant, H. A. Erlich, *ibid.* 48, 370 (1991).
- We thank our colleagues who contributed to the development and application of PCR. The space constraints of this review and the many publications on PCR 99. prevent a comprehensive survey of advances and applications; we apologize to any of our colleagues whose studies have not been noted specifically. We are grateful to R. Saiki, S. Scharf, R. Higuchi, and R. Abramson for allowing us to cite their unpublished work; E. Rose for critical review; and K. Levenson for preparation of this manuscript.

# **Complementary DNA Sequencing: Expressed** Sequence Tags and Human Genome Project

MARK D. ADAMS, JENNY M. KELLEY, JEANNINE D. GOCAYNE, MARK DUBNICK, MIHAEL H. POLYMEROPOULOS, HONG XIAO, CARL R. MERRIL, ANDREW WU, BJORN OLDE, RUBEN F. MORENO, ANTHONY R. KERLAVAGE, W. RICHARD MCCOMBIE, J. CRAIG VENTER\*

Automated partial DNA sequencing was conducted on more than 600 randomly selected human brain complementary DNA (cDNA) clones to generate expressed sequence tags (ESTs). ESTs have applications in the discovery of new human genes, mapping of the human genome, and identification of coding regions in genomic sequences. Of the sequences generated, 337 represent new genes, including 48 with significant similarity to genes from other organisms, such as a yeast RNA polymerase II subunit; Drosophila kinesin, Notch, and Enhancer of split; and a murine tyrosine kinase receptor. Forty-six ESTs were mapped to chromosomes after amplification by the polymerase chain reaction. This fast approach to cDNA characterization will facilitate the tagging of most human genes in a few years at a fraction of the cost of complete genomic sequencing, provide new genetic markers, and serve as a resource in diverse biological research fields.

HE HUMAN GENOME IS ESTIMATED TO CONSIST OF 50,000to 100,000 genes, up to 30,000 of which may be expressed in the brain (1). However, GenBank lists the sequence of only a few thousand human genes and <200 human brain messenger RNAs (mRNAs) (2). Once dedicated human chromosome

be required to complete the sequence of the genome (3). It is therefore likely that the majority of human genes will remain unknown for at least the next decade. The merits of sequencing cDNA, reverse transcribed from mRNA, as a part of the human genome project have been vigorously debated since the idea of determining the complete nucleotide sequence of humans first surfaced. Proponents of cDNA sequencing have argued that because the coding sequences of genes represent the vast majority of the information content of the genome, but only 3% of the DNA, cDNA sequencing should take precedence over genomic sequencing (4). Proponents of genomic sequencing have argued the difficulty of finding every mRNA expressed in all tissues, cell types, and developmental stages and have pointed out that much valuable information from intronic and intergenic regions, including control and regulatory sequences, will be missed by cDNA sequencing (5). However, many genome enthusiasts have incorrectly stated that gene coding regions, and therefore mRNA sequences, are readily predictable from genomic sequences and have concluded that there is no need for large-scale cDNA sequencing. In fact, prediction of transcribed regions of human genomic sequence is currently feasible only for relatively large exons (6). On the basis of our high output with automated DNA sequence

sequencing begins in 5 years, it is expected that 12 to 15 years will

analysis of 96 templates per day and consideration of the above issues, we initiated a pilot project to test the use of partial cDNA sequences (ESTs) in a comprehensive survey of expressed genes.

Sequence-tagged sites (STSs) are becoming standard markers for the physical mapping of the human genome (7). These short sequences from physically mapped clones represent uniquely identified map positions. ESTs can serve the same purpose as the random genomic DNA STSs and provide the additional feature of pointing directly to an expressed gene. An EST is simply a segment of a sequence from a cDNA clone that corresponds to an mRNA. ESTs longer than 150 bp were found to be the most useful for similarity searches and mapping.

M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, A. Wu, B. Olde, R. F. Moreno, A. R. Kerlavage, W. R. McCombie, and J. C. Venter are in the Section of Receptor Biochemistry and Molecular Biology, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 20892. M. H. Polymeropoulos, H. Xiao, and C. R. Merril are in the Laboratory of Biochemical Genetics, National Institute of Mental Health, Neuroscience Center at St. Elizabeth's Hospital, Washington, DC 20032.

<sup>\*</sup>To whom correspondence should be addressed.

## Libraries of Complementary DNAs

Of the estimated 30,000 genes expressed in the human brain, as many as 20,000 may encode low-abundance, brain-specific transcripts (1). The fact that up to one-fourth of all genetic diseases affect neurological functions is an indication of the diversity and importance of genes expressed in the brain (8).

An assumption in our choice of cDNA libraries was that randomprimed and partial cDNA clones would be more informative in identifying genes and constructing a useful EST database than sequencing from the ends of full-length cDNAs (which contain 5' and 3' untranslated sequences) would be. By obtaining coding sequences, we hoped to take advantage of more sensitive peptide comparisons, in addition to nucleotide sequence comparisons. To discover the inherent limitations to be overcome in a large-scale cDNA sequencing project, we wanted to examine the diversity of representative cDNA libraries, identify desirable and undesirable characteristics of the libraries, and determine the information content and accuracy of single-run sequencing from both coding and flanking regions. Single-run sequencing involves performing a single sequence reaction, rather than relying on multiple, redundant reactions from each strand. We chose three commercial human brain cDNA libraries made from mRNA isolated from the hippocampus and temporal cortex of a 2-year-old female and from a fetal brain (9).

Single-run DNA sequence data were obtained from 609 randomly chosen cDNA clones (Table 1). Double-stranded cDNA clones in the pBluescript vector (Stratagene) were sequenced by a cycle sequencing protocol (10) with dye-labeled primers and 373A DNA Sequencers (Applied Biosystems). The average length of usable sequence was 397 bases with a standard deviation of 99 bases.

Subtractive hybridization has been used by researchers to reduce the population of highly represented sequences in a cDNA library (11, 12) by selectively removing sequences shared by another library. We tested subtractive hybridization as a way of enhancing the number of brain-specific clones in the hippocampus library by hybridizing the hippocampus library with a WI38 human lung fibroblast cell line cDNA library and removing the common sequences (Table 1) (12, 13).

**Table 1.** Composition of cDNA library determined by random clone sequencing. Each  $\lambda$  ZAP library (Stratagene) was converted en masse to pBluescript plasmids, transfected into *Escherichia coli* XL1-Blue (Stratagene) cells, and plated on plates with X-gal, isopropyl-1-thio- $\beta$ -D-galactoside, and ampicillin. A total of 1058 clones were picked at random from three human brain cDNA libraries: fetal brain, 2-year-old hippocampus, and 2-year-old temporal cortex (9). Clones selected from the hippocampus library after subtraction with the fibroblast library are listed in the "Subtracted" column. Templates for DNA sequencing were PCR products or plasmids prepared by the alkaline lysis method. About half of the templates prepared by PCR failed to yield an amplified fragment suitable for sequencing. This was primarily due to use of PCR conditions that minimized the need for further purification of the product but selected against amplification of long inserts (5  $\mu$ l of *E. coli* fresh or frozen overnight carrying the pBluescript plasmid, 7.5  $\mu$ M

### **EST** Characterization

Initially, EST sequences were examined for similarities in the GenBank nucleic acid database (14). ESTs without exact GenBank

**Table 2.** EST matches to human genes. Matches of at least 97% were considered to indicate that the EST corresponds directly to the human gene. Number, GenBank accession number of the matched sequence. Map positions are from (8), except where indicated. LDL, low-density lipoprotein; EST names in GenBank are the three-digit number given here preceded by "EST00."

EST	Identification	Number	Map location
001	B-Actin (cytoplasmic)	M10277	
002	B-Actin (cytoplasmic)	M10277	
003	<pre>B-Actin (cytoplasmic)</pre>	M10277	
264	γ-Actin (nonmuscle)	M24241	17p11-qter
265	γ-Actin (nonmuscle)	M24241	17p11-qter
005	CNPase	M19650	-
006	CNPase	M19650	
237	ADP/ATP translocase	J03591	Xq13-q26
238	Fructose-1,6-bisphosphatase	X07292	17cen-q12
239	α-2-Macroglobulin	M11313	12p13.3-p12.3
240	α-Fodrin	M18627	9q33-34
242	α-Tubulin	K00558	<b>*</b>
243	α-Tubulin	K00558	+
004	<b>B-Tubulin</b>	X02344	+
244	Amyloid A4	Y00264	21q21.3-22.05
245	Apolipoprotein J	J02908	8*
246	Breakpoint cluster region	X02596	22q11-q12
251	c-erbA-a-2	J03239	3p24.3
253	Calelectrin	J03578	-
254	Calmodulin	J04046	
261	Elongation factor-1 $\alpha$	X03558	<b>‡</b>
262	Filaggrin	M24355	1q21
263	G <sub>S</sub> protein α subunit	X04408	20q13.2-q13.3
266	Glial fibrillary acidic protein	J04569	
268	Gln synthetase	¥00387	
280	Hexokinase	t	10p11.2
269	High-mobility group 1 protein	X12597	
278	LDL receptor-related protein	X13916	
284	Na <sup>+</sup> , K <sup>+</sup> -ATPase $\alpha$ subunit	X04297	1p13-p11
285	Neurofilament light chain	X05608	8p21
288	Phosphoglycerate kinase	L00160	Xq13
362	Ret proto-oncogene	M16029	10g11.2
363	RhoB	X06820	
366	Osteonectin	J03040	5q31-q33
367	Synaptophysin (p38)	X06389	Xp11.23-p11.22
	-]F.ob]		

\*Indicates that the EST was mapped in this study by PCR. The human hexokinase nucleotide sequence has been published (29) but does not appear in GenBank or EMBL. This EST was initially identified by matches to the mouse and rat nucleotide sequences and the human peptide sequence. to not available.

each deoxynucleotide triphosphate, and 0.1  $\mu$ M each primer for 35 cycles: 94°C, 40 s; 55°C, 40 s; 72°C, 90 s). A further percentage of the PCRgenerated templates failed to sequence, largely because of primer-dimer or other amplification artifacts. Qiagen columns (Studio City, California) improved the percentage of plasmid templates that yielded usable sequences from about 60% with a standard alkaline lysis protocol to over 90%. Overall, 117 PCR-generated templates and 497 plasmid templates gave usable sequences. Dideoxy chain-termination sequencing reactions were performed with fluorescent dye-labeled M13 universal or reverse primers (Applied Biosystems). After a cycle sequencing reactions were run on a 373A automated DNA sequencer (Applied Biosystems). Some sequencing reactions were performed on an Applied Biosystems robotic workstation (28). For each column, numbers are indicated followed by percents in parentheses.

EST category	Hippocampus		Subtracted		Fetal brain		Temporal cortex	
Database match—human								
Mitochondrial genes	48	(12.8)	10	(8.6)	3	(7.9)	6	(7.5)
Repeated sequences	39	(10.4)	14	(12.2)	6	(15.8)	0	( <b>0</b> )
Ribosomal RNA	10	(2.7)	7	(6.0)	0	<b>`(0)</b> ´	11	(13.8)
Other nuclear genes	32	(8.6)	7	(6.0)	4	(10.5)	0	<b>`(0)</b> ´
Database match-other	32	(8.6)	7	(6.0)	5	(13.2)	4	(5.0)
No database match	160	( <b>42.8</b> )	44	(37.9)	20	(52.6)	6	(7.5)
Polyadenylate insert	53	(14.1)	24	(20.7)	0	<b>`(0)</b> ´	27	(33.7)
No insert	1	(0.3)	3	(2.6)	0	(0)	26	(32.5)

matches were translated in all six reading frames, and each translation was compared with the protein sequence database Protein Information Resource (PIR) and the ProSite protein motif database

Table 3. EST similarities in the GenBank and PIR databases. All significant similarities (P < 0.01) with GenBank or PIR entries are listed. Matches indicate percent identical bases for nucleotides and percent similarity (identical plus conservative substitutions) for peptides. Number indicates the accession number or locus name of the matched sequence. Abbreviations used are as follows: B, bovine; BM, Brugia malayi; BMDV, bovine mucosal disease virus; C, chicken; CE, Caenorhabditis elegans; D, Drosophila melanogaster; E, E. coli; H, human; L, lamprey; M, mouse; N, Neurospora crassa; P, pig; PP, Pseudomonas putida; PRV, Pseudorabies virus; R, rat; S, squid; T, Torpedo californica; TN, transposon Tn 4556; X, Xenopus laevis; Y, yeast; UT, untranslated; MARCKS, myristoylated alanine-rich C kinase substrate; HPRT, hypoxanthine-guanine phosphoribosyltransferase; GTP, guanosine triphosphate; LAMP, lysosomal-associated membrane protein; tRNA, transfer RNA; snRNP, small nuclear ribonucleoprotein; IGF, insulin growth factor; Mito, mitochondrial; DBP, albumin promoter D site-binding protein; and Pol, polymerase. EST names in GenBank are the three-digit number given here preceded by "EST00."

299       ras-like (R)       138       57       X06889         300       RP L30 (R)       189       69       K02932         301       RP S10 (R)       273       90.8       X13549         305       UT conserved sequence element (H)       85       81       M24686         365       UT conserved sequence element (T)       112       64       M30271         371       Maternal G10 mRNA (X)       234       80       X15243         372       Catalase T (Y)       65       72.3       X04625         374       RNA Pol II 6th subunit (Y)       216       64.7       M33924         Peptide similarities (PIR)         247       80-87tkD MARCKS (B)       62       82.3       S08341         377       Mito ATPase & subunit (B)       97       92.8       S00763         249       GTP-binding protein (BMDV)       27       74.1       GNWEWS         250       60K filarial antigen (BM)       109       78.0       A28209         252       Collagen 1 (CE)       57       57.9       A31219         255       Cadherin, neuronal (C)       42       64.3       A29964         250       Notch (D)       102	EST	Description	Length	Match	Number
377         With ATDESE Subunit (B)         421         85.1         X06088           248         p ADP-ribosyltransferase					
248         p ADP-ribosyltransferase					
substrate (B)         256         Bhancer of split (D)         264         71         M20571           257         Kinesin (D)         263         70.4         M24411           259         Xotch (X)         435         75.4         M3374           70         A-tubulin (H)         495         82.3         X00734           71         G-Actinin (H)         495         82.3         X00734           73         Apolipoprotein A-I 5'-UT (H)         110         69         M20655           74         HPRT 3'-UT (H)         85         75         M26434           75         KImpel-related Zn2+ fingers (H)         86         67         M20678           283         Ascnitase (P)         318         89         J05224           293         ras-like (K)         138         57         X06489           291         IGF-binding protein 5'-UT (R)         115         77.3         J04486           299         ras-like (K)         138         57         X06689           201         RP 130 (R)         138         57         X04895           201         RP 130 (R)         216         64.7         M3324           211         Catalase T (2)	- · ·		421	85.1	X06088
256         Enhancer of split (D)         264         71         M20571           257         Kinesin (D)         263         70.4         M2441           259         Actinin (D)         435         75.4         M33874           270         B-Tubulin (H)         495         82.3         X00734           271         G-Actinin (H)         272         85         X15804           273         Apolipoprotein A-I 5'-UT (H)         110         69         M20656           274         HPR 1's'-UT (H)         85         75         71.5         J04182           276         LAMP-1 (H)         257         71.5         J04186           276         LAMP-1 (H)         257         71.5         J04186           273         ras-like (R)         138         89         J05224           273         ras-like (R)         189         89         K05223           201         RP 510 (R)         189         89         K02933           201         RP 510 (R)         273         90.8         K13543           210         RR 710 (R)         216         64.7         M3924           211         Rohenaley protein (R)         216         6	248		250		107779
Zimesin (b)         Zimesin (c)         Zimesin (c) <thzimesin (c)<="" th=""> <thzimesin (c)<="" th=""></thzimesin></thzimesin>					
259         Xotch (X)         435         75.4         M3874           270         B-Tubulin (H)         495         82.3         X0073           271         G-Actinin (H)         272         85         X15804           273         Apolipoprotein A-I 5'-UT (H)         110         69         M20656           274         HRT 3'-UT (H)         85         75         M26434           275         Kruppel-related Zn <sup>2+</sup> fingers (H)         86         67         M20678           276         LAMP-1 (H)         257         71.5         J04146           289         Aconitase (P)         318         89         J05224           293         ras-like (K)         115         77.3         J04466           299         ras-like (R)         189         89         K02932           301         RP 510 (R)         189         80         K02932           301         RP 510 (R)         65         72.3         X04625           371         Maternal G10 mRNA (X)         234         80         X15243           372         Catalase T (Y)         65         72.3         X04625           274         RNA POI II 6th subunit (Y)         216         <					
270         B-TUDUIIN (H)         495         82.3         X00734           271         G-Actinin (H)         272         85         X15804           273         Apolipoprotein A-I 5'-UT (H)         110         69         M20656           274         HPRT 3'-UT (H)         85         75         M26434           275         Kruppel-related Zn <sup>2+</sup> fingers (H)         86         67         M20678           276         LAMP-1 (H)         257         71.5         J04182           289         Aconitase (P)         318         89         J05224           293         ras-like (*)         71         74         X01669           295         IGF-binding protein 5'-UT (R)         138         57         X06889           300         RP L30 (R)         273         90.6         X13549           301         RP S10 (R)         273         90.6         X13549           305         UT conserved sequence element (H)         85         81         M24666           317         Mito ATPase & subunit (Y)         216         64.7         M3324           204         Gatalase T (X)         216         64.7         M3324           217         MACKS (B)					
271         Q-Actinin (H)         272         85         X1804           273         Apoliporotein A-I 5'-UT (H)         110         69         M2065           274         HPRT 3'-UT (H)         85         75         M26434           275         Kruppel-related Zn <sup>2+</sup> fingers (H)         85         75         J04182           276         LAMP-1 (H)         257         71.5         J04182           289         Aconitase (P)         318         89         J05224           293         ras-like (*)         71         74         X01669           295         IGF-binding protein 5'-UT (R)         115         77.3         J04486           299         ras-like (R)         189         89         K02932           301         RP S10 (R)         189         89         K02932           301         RP S10 (R)         273         90.8         X13543           302         Maternal G10 mRNA (X)         234         80         X15243           301         RM Actal ACKS (B)         62         82.3         S08341           307         Genome polyprotein smg p25A (B)         98         89.4         A35652           307         Genome polyprotein (BMDV					
Apolipoprotein A-I 5'-UT (H)         110         69         M20656           274         HPRT 3'-UT (H)         85         75         M206434           275         Kruppel-related Zn <sup>2+</sup> fingers (H)         86         67         M20678           276         LAMP-1 (H)         257         71.5         J04182           289         Aconitase (P)         318         89         J05224           293         ras-like (*)         115         77.3         J04486           294         Aconitase (P)         138         57         X06889           296         IGF-binding protein 5'-UT (R)         115         77.3         J04486           299         ras-like (R)         138         57         X06889           200         RP L30 (R)         213         64         M30217           217         Maternal G10 mRNA (X)         234         60         X15243           217         Catalase T (Y)         216         62         72.3         S08341           217         B0-87XD MARCKS (B)         98         98         8.8         85652           217         GALalast I antigen (BM)         109         74.1         GRWPANS           216					
274         HPRT 3'-UT (H)         85         75         M26434           275         Kruppel-related Zn <sup>2+</sup> fingers (H)         86         67         M20678           276         LAMP-1 (H)         257         71.5         J04182           289         Aconitase (P)         318         89         6705224           293         ras-like (*)         71         74         X01669           295         IGF-binding protein 5'-UT (R)         115         77.3         J04486           296         ras-like (R)         138         57         X06899           300         RP 130 (R)         189         89         K02932           301         RP S10 (R)         126         64         M3271           317         Maternal G10 mRNA (X)         234         80         K12433           322         Catalase T (Y)         65         72.3         X04623           323         Mito ATPase & subunit (Y)         216         64.7         M3924           324         80-87kD MARCKS (B)         97         92.8         S00763           3247         80-87kD MARCKS (B)         97         92.8         S00763           3247         80-87kD MARCKS (B)         <					
Intro De Le (P)         B8         67         M20678           275         Kruppel-related Zn <sup>2+</sup> fingers (H)         88         67         M20678           276         LAMP-1 (H)         257         71.5         704182           289         Aconitase (P)         318         89         J05224           293         ras-like (*)         71         74         X01669           295         IGF-binding protein 5'-UT (R)         115         77.3         J04446           299         ras-like (R)         138         57         X06689           300         RP L30 (R)         119         89         K02932           301         RP S10 (R)         273         90.8         X13549           3168         Electromotor neuron protein (T)         112         64         M30271           317         Maternal G10 mRNA (X)         234         80         X15243           317         Maternal G10 mRNA (X)         234         80         X15243           317         Maternal G10 mRNA (X)         234         80         X36525           317         Grabarin, neuronal (C)         216         64.7         M3924           3121         GTP-binding protein MbeA (E)	-				
276       LAMP-1 (H)       257       71.5       J04182         289       Aconitase (P)       318       89       J05224         293       ras-like (*)       71       74       X01669         295       IGF-binding protein 5'-UT (R)       115       77.3       J04486         299       ras-like (R)       138       57       X06889         300       RP L30 (R)       189       89       K02932         301       RP S10 (R)       273       90.8       X13543         305       UT conserved sequence element (H)       85       11       M24686         306       Electromotor neuron protein (T)       112       64       M30271         317       Mitornal of MRNA (X)       234       80       X15243         200       Catalase T (2)       216       61.7       M3924         217       RAN Pol II 6th subunit (Y)       216       64.7       M3924         227       Catalase T (2)       62       82.3       S00763         2375       Genome polyprotein (BMDV)       27       74.1       GWWW         250       GOL filarial antigen (EM)       109       78.0       A28209         255       Caherin,					
289       Aconitase (P)       318       89       J05224         293       ras-like (*)       71       74       X01669         295       IGF-binding protein 5'-UT (R)       115       77.3       J04486         299       ras-like (R)       138       57       X06899         300       RP L30 (R)       189       89       K02932         301       RP S10 (R)       273       90.8       X13549         365       UT conserved sequence element (H)       85       81       M24686         668       Electromotor neuron protein (T)       112       64       M30271         372       Catalase T (Y)       65       72.3       X04625         374       RN Pol II 6th subunit (Y)       216       64.7       M33924         Peptide similarities (PIR)         VOID ROTEIN (BMDV)       27       74.1       GNWVEV         270       60K filarial antigen (EMD)       109       78.0       A28209         252       Collagen 1 (CE)       57       57.9       A31219         253       Gaderin, neuronal (C)       42       64.3       A2964         254       Ankyrin (H)       89       95.5       S0503					
293       ras-like (*)       71       74       X01662         295       IGF-binding protein 5'-UT (R)       115       77.3       J04486         299       ras-like (R)       138       57       X06889         300       RP L30 (R)       189       89       K02932         301       RP S10 (R)       273       90.8       X13549         305       UT conserved sequence element (H)       85       81       M24686         365       UT conserved sequence element (T)       112       64       M3021         372       Catalase T (Y)       65       72.3       X04625         374       RNA Pol II 6th subunit (Y)       216       64.7       M33924         277       80-87kD MARCKS (B)       62       82.3       S08341         377       Mico ATPase & subunit (B)       97       92.8       S00763         276       GRP-binding protein smg p25A (B)       98       89.8       A35652         375       Genome polyprotein (BMDV)       27       74.1       GNWWW         256       Calherin, neuronal (C)       42       64.3       A29964         255       Calherin, neuronal (C)       42       67.8       S04790 <tr< td=""><td></td><td></td><td></td><td></td><td></td></tr<>					
295         TGP-binding protein 5'-UT (R)         115         77.3         J04466           299         ras-like (R)         138         57         X06889           200         RP S10 (R)         189         89         K02932           301         RP S10 (R)         273         90.8         X13549           365         UT conserved sequence element (H)         85         81         M24666           371         Maternal G10 mRNA (X)         234         80         X15243           372         Catalase T (Y)         65         72.3         X04625           374         RNA Pol II 6th subunit (Y)         216         64.7         M33924           77         Mito ATPase 8 subunit (B)         97         92.8         S00763           375         Genome polyprotein (BMDV)         27         74.1         GNW2029           252         Collagen 1 (CE)         57         79         A31219           255         Cadherin, neuronal (C)         42         64.3         A23904           254         Enhancer of split (D)         102         72.5         A24768           257         Finger protein MbeA (E)         47         63.8         435104           257<					
299       ras-like (R)       138       57       X068932         300       RP L30 (R)       189       89       K02932         301       RP S10 (R)       189       89       K02932         305       UT conserved sequence element (H)       85       81       M24666         365       UT conserved sequence element (T)       112       64       M30271         374       RNA Pol II 6th subunit (X)       214       80       X15243         374       RNA Pol II 6th subunit (Y)       216       64.7       M33924         277       80-87kD MARCKS (B)       62       82.3       S08341         377       Mico ATPase & subunit (B)       97       92.8       S00763         289       GTP-binding protein smg p25A (B)       98       89.8       A35652         290       GK filarial antigen (EM)       109       78.0       A28209         252       Collagen 1 (CE)       57       57.9       A31219         255       Cadherin, neuronal (C)       42       64.3       A29564         266       Enhancer of split (D)       87       78.2       A30047         271       G-Actinin (H)       84       60.7       A35043					J04486
300         RP L30 (R)         189         89         K02323           301         RP S10 (R)         273         90.8         X13549           305         UT conserved sequence element (H)         85         81         M24686           365         UT conserved sequence element (T)         112         64         M30271           371         Maternal G10 mRNA (X)         234         80         X15243           372         catalase T (X)         65         72.3         X04625           374         RNA Pol II 6th subunit (Y)         216         64.7         M33924           247         80-87kD MARCKS (B)         62         82.3         S08341           377         Mito ATPase & subunit (B)         97         92.8         S00763           249         GTP-binding protein (BMDV)         27         74.1         GNWARW           250         Collagen 1 (CE)         57         57.9         A31219           252         Collagen 1 (CE)         102         72.5         A24768           254         Makrin (H)         89         95.5         S05503           257         Finger protein MbeA (E)         47         63.8         S04790           261			138		X06889
301         RP S10 (R)         273         90.8         X13549           365         UT conserved sequence element (H)         85         81         M24686           366         Electromotor neuron protein (T)         112         64         M30271           371         Maternal G10 mRNA (X)         234         80         X15243           372         Catalase T (Y)         216         64.7         M33924           Peptide similarities (PTR)           247         80-87kD MARCKS (B)         67         92.8         S00763           375         Genome polyprotein (BMDV)         27         74.1         GNWVBW           250         60K filarial antigen (BM)         109         78.0         A28209           252         Collagen 1 (CE)         57         57.9         A31219           256         Enhancer of split (D)         102         72.5         A24768           260         Mothilization protein MbeA (E)         47         63.8         S04790           271         Q-Actinin (H)         89         95.5         S05503           275         Finger protein XlcGF20-1 (X)         30         80.0         S0562           279         Elongation factor Tu (*)			189	89	K02932
365       UT conserved sequence element (H)       85       81       M24686         368       Electromotor neuron protein (T)       112       64       M30271         371       Maternal G10 mRNA (X)       234       80       X15243         372       Catalase T (Y)       65       72.3       X04625         374       RNA Pol II 6th subunit (Y)       216       64.7       M33924         Peptide similarities (PIR)         247       80-87kD MARCKS (B)       97       92.8       S00763         375       Genome polyprotein smg p25A (B)       98       98.8       A35652         375       Genome polyprotein (EMDV)       27       74.1       GNWEW         250       60K filarial antigen (EM)       109       78.0       A28209         252       Collagen 1 (CE)       57       57.9       A31219         255       Cadherin, neuronal (C)       42       64.3       A29964         256       Enhancer of split (D)       87       78.2       A30047         257       Finger protein XlcGF20-1 (X)       30       80.0       S06562         279       Elongation factor Tu (*)       24       79.2       S06703         281	301	RP S10 (R)	273	90.8	x13549
368         Electromotor neuron protein (T)         112         64         M30271           371         Maternal G10 mRNA (X)         234         80         X15243           372         Catalase T (X)         65         72.3         X04625           374         RNA Pol II 6th subunit (Y)         216         64.7         M33924           Peptide similarities (PIR)           247         80-87kD MARCKS (B)         62         82.3         S08341           377         Mito ATPase & subunit (B)         97         92.8         S00763           249         GTP-binding protein smg p25A (B)         98         89.8         A35652           375         Genome polyprotein (EMDV)         27         74.1         GNWVEW           250         60K filarial antigen (EM)         109         78.0         A282090           252         Colherin, neuronal (C)         42         64.3         A29964           256         Enhancer of split (D)         87         78.2         A30047           257         Ankyrin (H)         84         60.7         A35048           260         Mothilization protein MbeA (E)         47         9.5         S05503           257         Finger protein			85	81	M24686
371       Maternal G10 mRNA (X)       234       80       X15243         372       Catalase T (Y)       65       72.3       X04625         374       RNA Pol I 6th subunit (Y)       216       64.7       M33924         Peptide similarities (PIR)         247       80-87kD MARCKS (B)       62       82.3       S08341         375       Genome polyprotein smg p25A (B)       98       89.8       A35652         375       Genome polyprotein (BMDV)       27       74.1       GNWUW20         250       GOX filarial antigen (BM)       109       78.0       A28209         252       Collagen 1 (CE)       57       57.9       A31219         255       Cadherin, neuronal (C)       42       64.3       A29964         256       Enhancer of split (D)       87       78.2       A30047         259       Notch (D)       102       72.5       A24768         260       Mobilization protein MbeA (E)       47       63.8       S04792         271       α-Actinin (H)       84       60.7       A35049         271       α-Actinin (H)       89       95.5       S05503         275       Finger protein X1cGF20-1 (X) <t< td=""><td>368</td><td></td><td>112</td><td>64</td><td>M30271</td></t<>	368		112	64	M30271
374RNA Pol II 6th subunit (Y)21664.7M33924Peptide similarities (PIR)24780-87kD MARCKS (B)6282.3S08341377Mito ATPase B subunit (B)9792.8S00763249GTP-binding protein smg p25A (B)9889.8A35652375Genome polyprotein (BMDV)2774.1GNWVWD25060K filarial antigen (BM)10978.0A28209252Collagen 1 (CE)5757.9A31219255Cadherin, neuronal (C)4264.3A29964256Enhancer of split (D)8778.2A30047259Notch (D)10272.5A24768260Mobilization protein MbeA (E)4763.8S04790271 $\alpha$ -actinin (H)8460.7A35049272Ankyrin (H)8995.5S05503275Finger protein XlcGF20-1 (X)3080.0S0655279Elongation factor Tu (*)2479.2S04336281Monophenol monoxygenase (M)2969.0YRMSCS282Neurogenic receptor trkB (M)5683.4A35104283U1 snRNP 70K protein (M)5957.6S04336284Pro-rich protein (Clone cP7)5664.3E25372291NtrA (PP)3161.3JG0338292IE180 protein (PRV)2286.4A34845293ras-like (*)5358.558.2<	371		234		X15243
Peptide similarities (PIR)24780-87kD MARCKS (B)6282.3\$08341377Mito ATPase & subunit (B)9792.8\$00763249GTP-binding protein smg p25A (B)9889.8A35652375Genome polyprotein (BMDV)2774.1GNWYEW25060K filarial antigen (BM)10978.0A282096252Collagen 1 (CE)5757.9A31219255Cadherin, neuronal (C)4264.3A29964256Enhancer of split (D)8778.2A30047259Notch (D)10272.5A24766260Mobilization protein MbeA (E)4763.8\$04790272Ankyrin (H)8460.7A35049271 $\alpha$ -Actinin (H)8995.5\$05503275Finger protein XLcGF20-1 (X)3080.0\$06565279Elongation factor Tu (*)2479.2\$06703281Monophenol monooxygenase (M)2969.0YMMSC282Neurogenic receptor trkB (M)5683.4A35104283Ul snRNP 70K protein (M)5779.4\$03966284Aconitase (P)10698.1A35544290Pro-rich protein (clone cP7)5664.3\$25342291NtrA (FP)3161.3JG0336292IE180 protein (PRV)2286.4EDBETE293ras-like (*)5358.5847468	372	Catalase T (Y)			X04625
247 $80-87kD$ MARCKS (B)62 $82.3$ $$00341$ 377Mito ATPase & subunit (B)9792.8 $$00763$ 249GTP-binding protein smg p25A (B)98 $89.8$ $A35652$ 375Genome polyprotein (BMDV)2774.1GNWYBW25060K filarial antigen (EM)10978.0 $A28209$ 252Collagen 1 (CE)5757.9 $A31212$ 255Cadherin, neuronal (C)4264.3 $A29964$ 256Enhancer of split (D)8778.2 $A30047$ 259Notch (D)10272.5 $A24766$ 260Mobilization protein MbeA (E)4763.8 $S04790$ 272Ankyrin (H)8460.7 $A35043$ 271 $\alpha$ -Actinin (H)8995.5 $S05503$ 275Finger protein XlcGF20-1 (X)30 $80.0$ $S06565$ 279Elongation factor Tu (*)2479.2 $S06703$ 281Monophenol monoxygenase (M)2969.0YMMSC2282Neurogenic receptor trkB (M)56 $83.4$ $A35104$ 283U1 snRNP 70K protein (M)5957.6 $S04336$ 286Leu-tRNA ligase (N)4858.3 $A33475$ 289Aconitase (P)106 $86.1$ $A35544$ 290Pro-rich protein (clone cP7)56 $64.3$ $E25372$ 291NtrA (PP)31 $61.3$ JG0335294Alcohol sulfotransferase (R)3571.4 $A348$	374		216	64.7	M33924
377       Mito ATPase B subunit (B)       97       92.8       \$00763         249       GTP-binding protein smg p25A (B)       98       89.8       A35652         375       Genome polyprotein (BMDV)       27       74.1       GNWTBW         250       60K filarial antigen (BM)       109       78.0       A28209         252       Collagen 1 (CE)       57       57.9       A31219         255       Cadherin, neuronal (C)       42       64.3       A29964         256       Enhancer of split (D)       87       78.2       A30047         256       Enhancer of split (D)       102       72.5       A24768         260       Mobilization protein MbeA (E)       47       63.8       S0792         271       α-Actinin (H)       89       95.5       \$05503         275       Finger protein XlcGF20-1 (X)       30       80.0       \$06655         279       Elongation factor Tu (*)       24       79.2       \$06703         281       Monophenol monoxygenase (M)       29       69.0       YRMSCS         282       Neurogenic receptor trkB (M)       56       83.4       A35104         283       U1 snRNP 70K protein (M)       59       57.6<					
3/1AltonConstructor989.8 <td>247</td> <td>80-87kD MARCKS (B)</td> <td></td> <td></td> <td></td>	247	80-87kD MARCKS (B)			
375Genome polyprotein (BMDV)2774.1GNWVBU25060K filarial antigen (BM)10978.0A28203252Collagen 1 (CE)5757.9A31219255Cadherin, neuronal (C)4264.3A29964256Enhancer of split (D)8778.2A30047259Notch (D)10272.5A24766260Mobilization protein MbeA (E)4763.8S04790271 $\alpha$ -Actinin (H)8995.5S05503275Finger protein XlcGF20-1 (X)3080.0S06565279Elongation factor Tu (*)2479.2S06703281Monophenol monooxygenase (M)2969.0YMMSC2282Neurogenic receptor trkB (M)5683.4A35104283Ul snRNP 70K protein (M)5957.6S04336286Leu-tRNA ligase (N)4858.3A33475289Aconitase (P)10698.1A35544290Pro-rich protein (clone cP7)5664.3E25372291NtrA (PP)3161.3JG0336292IE180 protein (PRV)2286.4EDBETE293ras-like (*)5358.5B34786294Alcohol sulfotransferase (R)3571.4A33652295ras-like (R)5558.2TVHURF300RP L30 (R)6797.0S01848299ras-like (R)5558.2TVHURF </td <td></td> <td></td> <td></td> <td></td> <td></td>					
25060K filarial antigen (BM)10978.0A28209252Collagen 1 (CE)5757.9A31219255Cadherin, neuronal (C)4264.3A29964256Enhancer of split (D)8778.2A30047259Notch (D)10272.5A24768260Mobilization protein MbeA (E)4763.8S04790272Ankyrin (H)8460.7A35049271 $\alpha$ -actinin (H)8995.5S05503275Finger protein XlcGF20-1 (X)3080.0S06565279Elongation factor Tu (*)2479.2S06703281Monophenol monooxygenase (M)2969.0YRMSCS282Neurogenic receptor trkB (M)5683.4A35104283U1 snRNP 70K protein (M)5957.6S04336286Leu-tRNA ligase (N)4858.3A33475287Processing-enhancing protein (N)9779.4S03966289Aconitase (P)10698.1A35544290Pro-rich protein (clone cP7)5664.3E25372291NtrA (PP)3161.3JG0333292IE180 protein (PRV)2286.4EDBETE293ras-like (*)5358.558.5T1.4294Alcohol sulfotransferase (R)3571.4A33507295ras-like (R)5558.2TVHURE300RP J30 (R)6797.0					
252Collagen 1 (CE)5757.9A31219255Cadherin, neuronal (C)4264.3A29964256Enhancer of split (D)8778.2A30047259Notch (D)10272.5A24768260Mobilization protein MbeA (E)4763.8S04790272Ankyrin (H)8460.7A35049271 $\alpha$ -actinin (H)8995.5S05503275Finger protein XlcGF20-1 (X)3080.0S06565279Elongation factor Tu (*)2479.2S06703281Monophenol monoxygenase (M)2969.0YRMSC2282Neurogenic receptor trkB (M)5683.4A35104283UI snRNP 70K protein (M)5957.6S04336286Leu-tRNA ligase (N)4858.3A33475287Processing-enhancing protein (N)9779.4S03966298Aconitase (P)10698.1A35544290Pro-rich protein (Clone cP7)5664.3E25372291NtrA (PP)3161.3JG0338292IE180protein (PRV)2286.4EDBETE293ras-like (*)5358.5B34768294Alcohol sulfotransferase (R)3571.4A33565295ras-like (R)5558.2TVHURE296Transcriptional activator DBP (R)3974.4A34845299ras-like (R)58 <td< td=""><td></td><td></td><td></td><td></td><td></td></td<>					
255Cadherin, neuronal (C)4264.3A29964256Enhancer of split (D)8778.2A30047259Notch (D)10272.5A24766260Mobilization protein MbeA (E)4763.8S04790271 $\alpha$ -Actinin (H)8460.7A35043271 $\alpha$ -Actinin (H)8995.5S05503275Finger protein XlcGF20-1 (X)3080.0S06565279Elongation factor Tu (*)2479.2S06703281Monophenol monooxygenase (M)2969.0YRMSC5282Neurogenic receptor trkB (M)5683.4A35104283U1 snRNP 70K protein (M)5957.6S04336286Leu-tRNA ligase (N)4858.3A33470287Processing-enhancing protein (N)9779.4S03966289Aconitase (P)10698.1A35544290Pro-rich protein (clone cP7)5666.4ED5172291NtrA (PP)3161.3JG0336292IE180 protein (PRV)2286.4ED5176293ras-like (*)5358.5B34786294Alcohol sulfotransferase (R)3571.4A33507295ras-like (R)5558.2TVHURF296rascriptional activator DBP (R)3971.4A34844299ras-like (R)5558.2TVHURF300RP L30 (R)6797.0					
256Enhancer of split (D)8778.2 $A30047$ 259Notch (D)10272.5 $A24766$ 260Mobilization protein MbeA (E)4763.8 $S04790$ 272Ankyrin (H)8460.7 $A35049$ 271 $\alpha$ -Actinin (H)8995.5 $S05503$ 275Finger protein XlcGF20-1 (X)3080.0 $S06565$ 279Elongation factor Tu (*)2479.2 $S06703$ 281Monophenol monooxygenase (M)2969.0YRMSCS282Neurogenic receptor trkB (M)5683.4 $A35104$ 283Ul snRNF 70K protein (M)5957.6 $S04336$ 286Leu-tRNA ligase (N)4858.3 $A33475$ 287Processing-enhancing protein (N)9779.4 $S03966$ 289Aconitase (P)10698.1 $A35544$ 290Pro-rich protein (clone cP7)5664.3E25372291NtrA (PP)3161.3JG0336292IE180 protein (PRV)2286.4EDBEIF293ras-like (*)5358.5B34768294Alcohol sulfotransferase (R)3571.4A33565295ras-like (R)5558.2TVHURF300RP 1J0 (R)6797.0S01881304RP 510 (R)6797.0S01881304RP 510 (R)575898.3306RP 510 (R)6797.0S01881 <td></td> <td></td> <td></td> <td></td> <td></td>					
10010110272.5 $A24768$ 259Notch (D)10272.5 $A24768$ 260Mobilization protein MbeA (E)4763.8 $S04790$ 272Ankyrin (H)8460.7 $A35049$ 271 $\alpha$ -Actinin (H)8995.5 $S05503$ 275Finger protein XlcGF20-1 (X)3080.0 $S06565$ 279Elongation factor Tu (*)2479.2 $S06703$ 281Monophenol monoxygenase (M)2969.0YRMSCS282Neurogenic receptor trkB (M)5683.4 $A35104$ 283U1 snRNP 70K protein (M)5957.6 $S04336$ 286Leu-tRNA ligase (N)4858.3 $A33475$ 287Processing-enhancing protein (N)9779.4 $S03968$ 298Aconitase (P)10698.1 $A3544$ 290Pro-rich protein (Clone cP7)5664.3E25372291NtrA (PP)3161.3JG0336292IE180protein (PRV)2286.4EDBETE293ras-like (*)5358.5B34786294Alcohol sulfotransferase (R)3571.4A33565295ras-like (R)5558.2TVHURR296ras-like (R)5558.2TVHUR207Myosin heavy chain (L)5558.2TVHUR208ras-like (R)5693.331.6235075201RP 510 (R)6797.0 <td></td> <td></td> <td></td> <td></td> <td></td>					
260Mobilization protein MbeA (E)4763.8 $504790$ 272Ankyrin (H)8460.7A35043271 $\alpha$ -Actinin (H)8995.5 $505503$ 271 $\alpha$ -Actinin (H)8995.5 $505503$ 275Finger protein XlcGF20-1 (X)3080.0 $506565$ 279Elongation factor Tu (*)2479.2 $506703$ 281Monophenol monooxygenase (M)2969.0YRMSC52282Neurogenic receptor trkB (M)5683.4A35104283U1 snRNP 70K protein (M)5957.6 $504336$ 286Leu-tRNA ligase (N)4858.3 $A33475$ 287Processing-enhancing protein (N)9779.4 $503966$ 289Aconitase (P)10698.1 $A35544$ 290Pro-rich protein (clone cP7)5664.3E25372291NtrA (PP)3161.3JG0336292IE180 protein (PRV)2286.4EDBETT293ras-like (*)5358.5B34786294Alcohol sulfotransferase (R)3571.4A33507295ras-like (R)5558.2TVHURF296ras-like (R)569344.4299ras-like (R)5758.2TVHURF300RP L30 (R)6797.0S01881364Fibrinogen $\gamma$ chain (L)3577.1FGLMG5376Electromotor neuron protein (T)32					
272Ankyrin (H)8460.7A35049271 $\alpha$ -Actinin (H)8995.5\$05503275Finger protein XlcGP20-1 (X)3080.0\$06565279Elongation factor Tu (*)2479.2\$06703281Monophenol monooxygenase (M)2969.0YRMSCS282Neurogenic receptor trkB (M)5683.4A35104283UI snRNP 70K protein (M)5957.6\$04336286Leu-tRNA ligase (N)4858.3A33475287Processing-enhancing protein (N)9779.4\$03966289Aconitase (P)10698.1A35564290Pro-rich protein (clone cP7)5664.3E25372291NtrA (PP)3161.3JG0336292IE180 protein (PRV)2286.4EDBEIR293ras-like (*)5358.5B34786294Alcohol sulfotransferase (R)3571.4A33565295ras-riptional activator DBP (R)3974.4A34844297myosin heavy chain (R)6058.3MWRTS298Protein-tyrosine phosphatase (R)2286.4A34845299ras-like (R)5558.2TVHURR300RP 130 (R)6797.0S01881364Fibrinogen $\gamma$ chain (L)3577.1FGIM657376Hypothetical protein (TN)3764.9JQ0431370Various actins (*)<					
271 $\alpha$ -Actinin (H)8995.5\$05503275Finger protein XlcGF20-1 (X)3080.0\$06565279Elongation factor Tu (*)2479.2\$06703281Monophenol monooxygenase (M)2969.0YRMSCS282Neurogenic receptor trkB (M)5683.4A35104283UI snRNP 70K protein (M)5957.6\$04336286Leu-tRNA ligase (N)4858.3A33475287Processing-enhancing protein (N)9779.4\$03966290Pro-rich protein (clone cP7)5664.3E25372291NtrA (PP)3161.3JG0336292IE180 protein (PRV)2286.4EDBEIF293ras-like (*)5358.5B34786294Alcohol sulfotransferase (R)3571.4A33566295ras-criptional activator DBP (R)3974.4A34844296ras-like (R)5558.2TVHURF300RP L30 (R)5898.3S11622301RP S10 (R)6797.0S01881305TLike (S)5558.2TVHURF306Hypothetical protein (TN)3764.9JQ0431370Various actins (*)3775.7\$06662371Maternal G10 mRNA (X)3994.9\$059575373Hypothetical protein (Y)2475.0C27066					
275Finger protein XlcGF20-1 (X)3080.0 $$06565$ 279Elongation factor Tu (*)2479.2 $$06703$ 281Monophenol monocygenase (M)2969.0YRMSC2282Neurogenic receptor trkB (M)5683.4A35104283UI snRNP 70K protein (M)5957.6 $$04336$ 286Leu-tRNA ligase (N)4858.3 $$33475$ 287Processing-enhancing protein (N)9779.4 $$03966$ 289Aconitase (P)10698.1 $$A35544$ 290Pro-rich protein (clone cP7)5664.3E25372291NtrA (PP)3161.3JG0336292IE180 protein (PRV)2286.4EDBETH293ras-like (*)5358.5B34786294Alcohol sulfotransferase (R)3571.4A33562296Transcriptional activator DBP (R)3974.4A34894297Myosin heavy chain (R)6058.3MWRTS298ras-like (R)5558.2TVHURF300RP L30 (R)5898.3\$11622301RP S10 (R)6797.0\$01888364Fibrinogen $\gamma$ chain (L)3577.1FGLMG6374Hypothetical protein (TN)3764.9JQ0431370Various actins (*)3775.7\$06662371Maternal G10 mRNA (X)3994.9\$05955373Hypothetical protein		· ·			
279Elongation factor Tu (*)2479.2 $506703$ 281Monophenol monooxygenase (M)2969.0YRMSCS282Neurogenic receptor trkB (M)5683.4 $A35104$ 283Ul snRNP 70K protein (M)5957.6 $S04336$ 286Leu-tRNA ligase (N)4858.3 $A33475$ 287Processing-enhancing protein (N)9779.4 $S03968$ 289Aconitase (P)10698.1 $A35544$ 290Pro-rich protein (clone cP7)5664.3E25372291NtrA (PP)3161.3JG0336292IE180 protein (PRV)2286.4EDBETR293ras-like (*)5358.5B34786294Alcohol sulfotransferase (R)3571.4A33556295Transcriptional activator DBP (R)3974.4A34894296ras-like (R)5558.2TVHURF298Protein-tyrosine phosphatase (R)2286.4A34845299ras-like (R)5558.2TVHURF300RP L30 (R)6797.0S01881364Fibrinogen $\gamma$ chain (L)3577.1FGIMGS376Hypothetical protein (TN)3764.9JQ0431370Various actins (*)3775.7S06562371Maternal G10 mRNA (X)3994.9905955373Hypothetical protein (Y)2475.0C270667					
281Monophenol monoxygenase (M)2969.0YRMSCS282Neurogenic receptor $trkB$ (M)5683.4A35104283U1 snRNP 70K protein (M)5957.6S04336286Leu-tRNA ligase (N)4858.3A33475287Processing-enhancing protein (N)9779.4S03966289Aconitase (P)10698.1A35544290Pro-rich protein (clone cP7)5664.3E25372291NtrA (PP)3161.3JG0338292IE180 protein (PRV)2286.4EDBIF293ras-like (*)5358.5B34786294Alcohol sulfotransferase (R)3571.4A33569295Transcriptional activator DBP (R)3974.4A4844297Myosin heavy chain (R)6058.3MWRTS298Protein-tyrosine phosphatase (R)2286.4A34845299ras-like (R)5558.2TVHURR300RP 130 (R)5898.3S11622301RP 510 (R)6797.0S01881364Fibrinogen $\gamma$ chain (L)3577.1FGLMGS376Hypothetical protein (TN)3764.9JQ0431370Various actins (*)3775.7S0605373Hypothetical protein (Y)2475.0C27064					S06703
282Neurogenic receptor $trkB$ (M)5683.4A35104283UI snRNP 70K protein (M)5957.6S04336286Leu-tRNA ligase (N)4858.3A33475287Processing-enhancing protein (N)9779.4S03966289Aconitase (P)10698.1A35544290Pro-rich protein (clone cP7)5664.3E25372291NtrA (PP)3161.3JG0336292IE180 protein (PRV)2286.4EDBETF293ras-like (*)5358.5B34786294Alcohol sulfotransferase (R)3571.4A33562296Transcriptional activator DBP (R)3974.4A34894297Myosin heavy chain (R)6058.3MWRTS298protein-tyrosine phosphatase (R)2286.4A34842299ras-like (R)5558.2TVHURF300RP L30 (R)6797.0S01881622301RP S10 (R)6797.0S01881622304Fibrinogen $\gamma$ chain (L)3577.1FGLMGS376Electromotor neuron protein (T)3281.3B3337368Electromotor neuron protein (T)3764.9JQ0431370Various actins (*)3775.7S06062371Maternal G10 mRNA (X)3994.9S05955373Hypothetical protein (Y)2475.0C27064					
283U1snNP70Kprotein (M)5957.6S04336286Leu-tRNALigase (N)4858.3A33475287Processing-enhancing protein (N)9779.4S03966289Aconitase (P)10698.1A35544290Pro-rich protein (clone cP7)5664.3E25372291NtrA (PP)3161.3JG03336292IE180protein (PRV)2286.4EDBEIT293ras-like (*)5358.5B34786294Alcohol sulfotransferase (R)3571.4A33565296Transcriptional activator DBP (R)3974.4A34894297Myosin heavy chain (R)6058.3MWRTS298Protein-tyrosine phosphatase (R)2286.4A34845299ras-like (R)5558.2TVHURR300RP L30 (R)6797.0S01881301RP S10 (R)6797.0S01885364Fibrinogen $\gamma$ chain (L)3577.1FGLMGS376Electromotor neuron protein (T)3281.3B33317368Electromotor neuron protein (T)3764.9JQ0431370Various actins (*)3775.7S06662371Maternal G10 mRNA (X)3994.9S05955373Hypothetical protein (Y)2475.0C27064					A35104
286Leu-tRNA ligase (N)4858.3A33475287Processing-enhancing protein (N)9779.4S03966289Aconitase (P)10698.1A35544290Pro-rich protein (clone cP7)5664.3E25372291NtrA (PP)3161.3JG0336292IE180 protein (PRV)2286.4EDBEIF293ras-like (*)5358.5B34786294Alcohol sulfotransferase (R)3571.4A33565295Transcriptional activator DBP (R)3974.4A34894296Protein-tyrosine phosphatase (R)2286.4A34845297myosin heavy chain (R)6058.3MWRTS208Protein-tyrosine phosphatase (R)2286.4A34845299ras-like (R)5558.2TVHURR300RP L30 (R)6797.0S01881364Fibrinogen $\gamma$ chain (L)3577.1FGLMGS257Kinesin (S)3391.4A35075368Electromotor neuron protein (T)3281.3B33313369Hypothetical protein (TN)3764.9JQ0431371Maternal G10 mRNA (X)3994.9S05955373Hypothetical protein (Y)2475.0C27065			59	57.6	S04336
287Processing-enhancing protein (N)9779.4S03966289Aconitase (P)10698.1A35544290Pro-rich protein (clone cP7)5664.3E25372291NtrA (PP)3161.3JG0336292IE180 protein (PRV)2286.4EDBEIF293ras-like (*)5358.5B34786294Alcohol sulfotransferase (R)3571.4A33569297Myosin heavy chain (R)6058.3MWRTS298Protein-tyrosine phosphatase (R)2286.4A34845299ras-like (R)5558.2TVHURF301RP L30 (R)6797.0S01881622304Fibrinogen $\gamma$ chain (L)3577.1FGLMGS376Hypothetical protein (TN)3764.9JQ0431370Various actins (*)3775.7S06052373Hypothetical protein (Y)2475.0C27065			48		A33475
289         Aconitase (P)         106         98.1         A35544           290         Pro-rich protein (clone cP7)         56         64.3         E25372           291         NtrA (PP)         31         61.3         JG0336           292         IE180 protein (PRV)         22         86.4         EDBEIR           293         ras-like (*)         53         58.5         B34786           294         Alcohol sulfotransferase (R)         35         71.4         A33562           296         Transcriptional activator DBP (R)         39         74.4         A34894           297         Myosin heavy chain (R)         60         58.3         MWRTS           298         Protein-tyrosine phosphatase (R)         22         86.4         A34844           299         ras-like (R)         55         58.2         TVHURF           301         RP 130 (R)         67         97.0         S01881           364         Fibrinogen γ chain (L)         35         77.1         FGLMGE           376         Kinesin (S)         93         91.4         A35075           368         Electromotor neuron protein (T)         32         81.3         B33316           370<			97		S03968
290         Pro-rich protein (clone cP7)         56         64.3         E25372           291         NtrA (PP)         31         61.3         JG0336           292         IE180 protein (PRV)         22         86.4         EDBEIT           293         ras-like (*)         53         58.5         B34786           294         Alcohol sulfotransferase (R)         35         71.4         A33565           296         Transcriptional activator DBP (R)         39         74.4         A34894           297         Myosin heavy chain (R)         60         58.3         MWRTS           298         Protein-tyrosine phosphatase (R)         22         86.4         A34894           299         ras-like (R)         55         58.2         TVHURF           300         RP L30 (R)         58         98.3         S11622           301         RP S10 (R)         67         97.0         S01881           364         Fibrinogen Y chain (L)         35         77.1         FGIMGS           257         Kinesin (S)         93         91.4         A35075           368         Electromotor neuron protein (T)         32         81.3         B33315           369	289		106	98.1	A35544
291         NtrA (PP)         31         61.3         JG0338           292         IE180 protein (PRV)         22         86.4         EDBEIF           293         ras-like (*)         53         58.5         B34786           294         Alcohol sulfotransferase (R)         35         71.4         A33565           296         Transcriptional activator DBP (R)         39         74.4         A34894           297         Myosin heavy chain (R)         60         58.3         MWRTS           298         Protein-tyrosine phosphatase (R)         22         86.4         A34845           299         ras-like (R)         55         58.2         TVHURF           300         RP L30 (R)         58         98.3         S11622           201         RP S10 (R)         67         97.0         S01881           364         Fibrinogen γ chain (L)         35         77.1         FGLMGS           376         Hypothetical protein (TN)         37         64.9         JQ0431           370         Various actins (*)         37         75.7         S0662           371         Maternal G10 mRNA (X)         39         94.9         S059553           373			56	64.3	E25372
292         IE180 protein (PRV)         22         86.4         EDBEIR           293         ras-like (*)         53         58.5         B34786           294         Alcohol sulfotransferase (R)         35         71.4         A33565           296         Transcriptional activator DBP (R)         39         74.4         A34894           297         Myosin heavy chain (R)         60         58.3         MWRTS           298         Protein-tyrosine phosphatase (R)         22         86.4         A34844           299         ras-like (R)         55         58.2         TVHURF           301         RP L30 (R)         67         97.0         S0181622           301         RP S10 (R)         67         97.0         S0181622           304         Fibrinogen γ chain (L)         35         77.1         FGLMGS           364         Fibrinogen γ chain (L)         35         77.1         FGLMGS           376         Electromotor neuron protein (T)         32         81.3         B333163           369         Hypothetical protein (TN)         37         64.9         JQ0431           370         Various actins (*)         37         75.7         S06062 <tr< td=""><td>291</td><td></td><td>31</td><td>61.3</td><td>JG0338</td></tr<>	291		31	61.3	JG0338
293       ras-like (*)       53       58.5       B34786         294       Alcohol sulfotransferase (R)       35       71.4       A33569         296       Transcriptional activator DBP (R)       39       74.4       A34894         297       Myosin heavy chain (R)       60       58.3       MWRTS         298       Protein-tyrosine phosphatase (R)       22       86.4       A34845         299       ras-like (R)       55       58.2       TVHURE         300       RP L30 (R)       67       97.0       S01881         301       RP S10 (R)       67       97.0       S01881         364       Fibrinogen Y chain (L)       35       77.1       FGLMGS         257       Kinesin (S)       93       91.4       A3507         368       Electromotor neuron protein (T)       32       81.3       B33319         369       Hypothetical protein (TN)       37       64.9       JQ0431         370       Various actins (*)       37       75.7       S06662         371       Maternal G10 mRMA (X)       39       94.9       S05955         373       Hypothetical protein (Y)       24       75.0       C27066 <td>292</td> <td></td> <td>22</td> <td></td> <td>EDBEIF</td>	292		22		EDBEIF
296         Transcriptional activator DBP (R)         39         74.4         A34894           297         Myosin heavy chain (R)         60         58.3         MWRTS           298         Protein-tyrosine phosphatase (R)         22         86.4         A34894           299         ras-like (R)         55         58.2         TVHURE           300         RP L30 (R)         58         98.3         \$11622           301         RP S10 (R)         67         97.0         \$018181           364         Fibrinogen γ chain (L)         35         77.1         FGLMGS           368         Electromotor neuron protein (T)         32         81.3         B33315           369         Hypothetical protein (TN)         37         64.9         JQ0431           370         Various actins (*)         37         75.7         \$06062           371         Maternal G10 mRNA (X)         39         94.9         \$059557           373         Hypothetical protein (Y)         24         75.0         \$270661	293		53		в34788
297         Myosin heavy chain (R)         60         58.3         MWRTS           298         Protein-tyrosine phosphatase (R)         22         86.4         A34845           299         ras-like (R)         55         58.2         TVHURF           300         RP L30 (R)         58         98.3         S11622           301         RP S10 (R)         67         97.0         S01883           364         Fibrinogen Y chain (L)         35         77.1         FGLMGS           257         Kinesin (S)         93         91.4         A35075           368         Electromotor neuron protein (T)         32         81.3         B3319           369         Hypothetical protein (TN)         37         64.9         JQ0431           370         Various actins (*)         37         75.7         S06662           371         Maternal G10 mRMA (X)         39         94.9         S05955           373         Hypothetical protein (Y)         24         75.0         C27063	294				A33569
298         Protein-tyrosine phosphatase (R)         22         86.4         A34845           299         ras-like (R)         55         58.2         TVHURE           300         RP L30 (R)         58         98.3         S11622           301         RP S10 (R)         67         97.0         S01881           364         Fibrinogen γ chain (L)         35         77.1         FGLMGS           368         Electromotor neuron protein (T)         32         81.3         B33315           369         Hypothetical protein (TN)         37         64.9         JQ0431           370         Various actins (*)         37         75.7         S06062           371         Maternal G10 mRNA (X)         39         94.9         S05955           373         Hypothetical protein (Y)         24         75.0         C27065	296	Transcriptional activator DBP (R)			
299       ras-like (R)       55       58.2       TVHURE         300       RP L30 (R)       58       98.3       S11622         301       RP S10 (R)       67       97.0       S01881         364       Fibrinogen γ chain (L)       35       77.1       FGLMGS         257       Kinesin (S)       93       91.4       A35075         368       Electromotor neuron protein (T)       32       81.3       B33315         369       Hypothetical protein (TN)       37       64.9       JQ0431         370       Various actins (*)       37       75.7       S06662         373       Hypothetical protein (Y)       24       75.0       C27061					
300         RP L30 (R)         58         98.3         \$11622           301         RP S10 (R)         67         97.0         \$01881           364         Fibrinogen γ chain (L)         35         77.1         FGLMGS           257         Kinesin (S)         93         91.4         A35075           368         Electromotor neuron protein (T)         32         81.3         B33315           369         Hypothetical protein (TN)         37         64.9         JQ0431           370         Various actins (*)         37         37.7         \$06662           371         Maternal G10 mRNA (X)         39         94.9         \$059555           373         Hypothetical protein (Y)         24         75.0         C27061		Protein-tyrosine phosphatase (R)			
301         RP S10 (R)         67         97.0         S01881           364         Fibrinogen Y chain (L)         35         77.1         FGLMGS           257         Kinesin (S)         93         91.4         A35075           368         Electromotor neuron protein (T)         32         81.3         B33315           369         Hypothetical protein (TN)         37         64.9         JQ0431           370         Various actins (*)         37         75.7         S06662           371         Maternal G10 mRNA (X)         39         94.9         S05955           373         Hypothetical protein (Y)         24         75.0         C27066					
364         Fibrinogen γ chain (L)         35         77.1         FGLMGS           257         Kinesin (S)         93         91.4         A35075           368         Electromotor neuron protein (T)         32         81.3         B33315           369         Hypothetical protein (TN)         37         64.9         JQ0431           370         Various actins (*)         37         75.7         S06062           371         Maternal G10 mRNA (X)         39         94.9         S05955           373         Hypothetical protein (Y)         24         75.0         C27065					
257         Kinesin (S)         93         91.4         A35075           368         Electromotor neuron protein (T)         32         81.3         B33315           369         Hypothetical protein (TN)         37         64.9         JQ0431           370         Various actins (*)         37         75.7         S06062           371         Maternal G10 mRNA (X)         39         94.9         S05955           373         Hypothetical protein (Y)         24         75.0         C27061					
368         Electromotor neuron protein (T)         32         81.3         B33315           369         Hypothetical protein (TN)         37         64.9         JQ0433           370         Various actins (*)         37         75.7         S06062           371         Maternal G10 mRNA (X)         39         94.9         S05955           373         Hypothetical protein (Y)         24         75.0         C27063					
369         Hypothetical protein (TN)         37         64.9         JQ0431           370         Various actins (*)         37         75.7         S06062           371         Maternal G10 mRNA (X)         39         94.9         S05955           373         Hypothetical protein (Y)         24         75.0         C27061					
370         Various actins (*)         37         75.7         S06062           371         Maternal G10 mRNA (X)         39         94.9         S05955           373         Hypothetical protein (Y)         24         75.0         C27061					
371         Maternal G10 mRNA (X)         39         94.9         \$05955           373         Hypothetical protein (Y)         24         75.0         C27061					
373 Hypothetical protein (Y) 24 75.0 C27061					
3/4 KNA POL II 6th subunit (Y) /3 90.4 B34588					
	374	RNA POL II 6th subunit (Y)	13	90.4	54586

(14). Comparisons with the ProSite motif database were done by means of the program MacPattern from the EMBL Data Library (14a). GenBank and PIR searches were conducted with our modifications of the "basic local alignment search tool" programs for nucleotide (BLASTN) and peptide (BLASTX) comparisons (15). These modifications permit many query sequences to be automatically searched in a sequential fashion. PIR searches were run on the National Center for Biotechnology Information BLAST network service. The BLAST programs contain a rapid database-searching algorithm that searches for local areas of similarity between two sequences and then extends the alignments on the basis of defined match and mismatch criteria. The algorithm does not consider the potential of gaps to improve the alignment, thus sacrificing some sensitivity for 60- to 80-fold increase in speed over other database-searching programs such as FASTA (16).

Sequence similarities identified by the BLAST programs were considered statistically significant with a Poisson *P*-value <0.01. The Poisson *P*-value is the probability of as high a score occurring by chance, given the number of residues in the query sequence and the database. After the BLASTN search, 30 unmatched ESTs were compared against GenBank by FASTA to determine if significant matches were missed because of the use of BLASTN for the database search. No additional statistically significant matches were found. Statistical significance does not necessarily mean functional similarity; some of the matches reported here may indicate the presence of a conserved domain or motif or simply a common protein structure pattern. Statistically significant matches to GenBank and PIR are reported in Tables 2 and 3. The length and percent identity or similarity of each alignment is given in Table 3 to aid in evaluation of match quality.

On the basis of database searches, the 609 EST sequences were classified into eight groups as shown in Table 1. Four groups, with 197 of the sequences (32% of the total), consist of matches to human sequences: repetitive elements, mitochondrial genes, ribosomal RNA genes, and other nuclear genes. Forty-eight of the sequences (8%) matched nonhuman entries in GenBank or PIR, whereas 230 (38%) had no significant matches. The remaining 134 (22%) sequences contained no insert between the Eco RI cloning sites or consisted entirely of polyadenylate.

**Table 4.** Matches to the ProSite motif database. Pattern matches from the ProSite database (except posttranslational modification sites) are shown. Abbreviations used are as follows: AA, amino acyl; HIGH, motif consensus in single-letter amino acid code (30); ILGF, insulin-like growth factor; DHFR, dihydrofolate reductase; EGF, epidermal growth factor; Gal-P-UDP, galactose-1-phosphate-uridyl; C2H2, two Cys and two His residues. EST names in GenBank are the three-digit number given here preceded by "EST00."

Motif name	EST
AA-tRNA ligase "HIGH"	094
ATP-binding site A	052,068,158,177,207,091,008,261
Carboxypeptidase/Zn <sup>2+</sup>	112
COX1	249*
Cytochrome c	060,128,120,139,279*,218,063,106
DHFR	235
Elongation factor	261
EGF	187,203
2Fe/2S Ferredoxin	067
Gal-P-UDP-transferase	101
Glycoprotein hormone	112
ILGF-binding protein	193
Leu zipper	071,072,095,055,070,106,025,200,221,
	107,102,114,131,260*,290*,291*,294*,
	164,287*,061,369*
Nuclear localization	182,020,183,214,062
Rubredoxin	226
Snake toxin	085
Zinc finger (C2H2)	188,275*

\*See Table 3 for ESTs with similarity to GenBank or PIR sequences.

\*Matches with sequences from several organisms.

Table 5. Accuracy of single-run double-stranded automated sequencing.
ESTs listed in Table 2 and those matching mitochondrial and ribosomal
genes were aligned with sequences from GenBank with the GCG program
BESTFIT. The first 85 nucleotides were the polylinker sequence that was not

aligned with the pBluescript SK reference sequence. Tabulation of errors began 15 bases into the BESTFIT alignment and thus is reported beginning with bases 101 to 200.

Bases from	Mismatches- ambiguities*	Gaps*		Accuracy	Aligned
primer		Insertions	Deletions	%	bases
101-200	1.45	0.18	0.19	98.2	8800
201-300	1.72	0.25	0.11	97.9	8130
301-400	2.07	0.98	0.37	96.6	5404
>400	3.53	2.63	1.06	92.8	3197

\*Error rates are reported as number of mismatches, insertions, or deletions per hundred aligned bases. "Mismatches" includes ambiguous base calls.

Thirty-six ESTs matched previously sequenced human nuclear genes with more than 97% identity (Table 2). Four of these ESTs were from genes encoding enzymes involved in maintaining metabolic energy, including ADP/ATP (adenosine diphosphate/adenosine triphosphate) translocase, aldolase C, hexokinase, and phosphoglycerate kinase. Human homologs of genes for the bovine mitochondrial ATP synthase  $F_0\beta$  subunit and porcine aconitase were also found (Table 3). Brain-specific cDNAs included synaptophysin, glial fibrillary acidic protein (GFAP), and neurofilament light chain. At least six ESTs were from genes encoding proteins involved in signal transduction: 2',3'-cyclic nucleotide 3'-phosphodiesterase (CNPase) (two ESTs), calmodulin, c-erbA-a-2, G stimulating protein (G<sub>s</sub>)  $\alpha$  subunit, and Na<sup>+</sup>, K<sup>+</sup>-ATPase  $\alpha$  subunit. Other ESTs were matches to genes for ubiquitous structural proteins-actins, tubulins, and fodrin (nonerythroid spectrin). Eight ESTs were from genes known to be associated with genetic disorders (8). More than half of the human-matched ESTs have been mapped to chromosomes, indicating the bias of GenBank entries toward well-studied genes and proteins.

ESTs without significant GenBank matches were also compared to the ProSite database of recognized protein motifs. Not counting posttranslational modification signatures, 54 sequences contained motifs from the database (Table 4). Some patterns are found in scores or even, as in the case of the leucine zipper, hundreds of proteins that do not share the functional property implied by the presence of the motif.

Similarities to sequences from other organisms were also detected in the BLAST searches of GenBank and PIR (Table 3). Several ESTs were similar to "housekeeping" genes, including the ribosomal proteins (RP) S10 and L30 (in the rat) and the above glycolytic enzymes. EST00257 showed strong nucleotide sequence similarity to the squid (67.4%) and Drosophila (70.4%) kinesin heavy chain. Kinesin was first described as a microtubule-associated motor protein involved in organelle transport in the squid giant axon (17). Six oncogene-related sequences were also among the cDNA clones sequenced. EST00299 and EST00283 showed similarity to several ras-related genes, and EST00248 matched the 3' untranslated region of the bovine substrate of botulinum toxin ADP-ribosyltransferase. We also observed similarities with a Saccharomyces cerevisiae RNA polymerase subunit and Torpedo californica electromotor neuron-associated protein. Two ESTs may represent new members of known human gene families: EST00270 matched the three  $\beta$ -tubulin genes with 88 to 91% identity and EST00271 matched  $\alpha$ -actinin with 85% identity at the nucleotide level.

Among the most interesting of the primary sequence relationships was the similarity of ESTs to the Drosophila genes Notch and Enhancer of split. Nucleotide and peptide alignments of EST00256 and EST00259 with the Drosophila genes are shown in Fig. 1. Both genes are part of a signal cascade encoded by the "neurogenic" genes that are involved in the differentiation of neuronal and epidermal cell lineages in the neuroectoderm of the developing Drosophila embryo (18). It has been proposed that the Enhancer of split protein interacts with a membrane protein that is the product of the Notch gene to convert a developmental signal into an altered pattern of gene expression (18). EST00256 matched near the 5' end of the Enhancer of split coding sequence, away from the mammalian G protein  $\beta$ subunit and yeast cdc4-like elements (19). Part of the EST00259 match to Notch is in the cdc10/SW16 region that is similar to three

Fig. 1. Sequence alignments of ESTs with Drosophila neurogenic genes. ESTs and EST translations were aligned with nucleotide and peptide sequences of two Drosophila neurogenic genes with the GCG program BESTFIT. The peptide alignment (30) of EST00259 with the Xotch product, the Xenopus laevis homolog of Notch, is also shown. (A) EST00256 with Drosophila Enhancer of split (M20571); 69.202% identity; 1 gap. (B) EST00256 product with the product of Drosophila Enhancer of split (M20571); 72.826% similarity; 58.696% identity, 0 gaps. (C) EST00259 product with the Drosophila Notch (K03508); 60.294% product similarity; 43.382% identity; 5 gaps. (D) EST00259 product with the Xenopus Xotch product (M33874); 82.143% similarity; 75.714% identity, 4 gaps. Gaps have been introduced to increase identity and similarity (indicated by dots in lines). Numbers in parentheses indicate the GenBank accession numbers. Symbols between lines: dashes indicate identity; double dots indicate a similarity score of 0.5 to 1.4; single dots represent a

Α	C
5 CRACTCARATTCACCACCTCGGGACTCCTGCGACCGCATCAAAGACGAATT 	1
55 TCAGCTACTGCAAGCTCAGTACCACAGCCTCAAGCTCGAWTGTNACAAGT 	2
105 TGGCCAGTGAGAAGTCAGAGATGCAGCRTCACTATKTGATGTACTACGAG 	:
155 AKGTCCTACGGCTTGACCATCGAGATGCACAAACAGGCTGAGACCGTCAA 1:11111 11 1 1 1 1 1111111111111111111	:
205 AAGGCTGAACGGGATTTGTGCCCAGGTCCTGCCTACCTTTCCCAAGGAG 	
255 CACCAGCAGCAGGT              568 CACCAGCAGCAGT	D
В	18
1 QOLKFTTSDSCDRIKDEFQLLQAQYESLKLXCXKLASEKSEMQXHYXMYY :    .:  !! .:  !! !      .       18 GPIKFTIADTLERIKEEPNFLQAHYHSIKLECEKLSNEKTEMQRHYMYY	10

51 EXSYGLTIEMHKQAETVKRLNGICAQVLPYLSKEHQQQVFGG | ||||.:|||||...||||.:..|:||:| :|||||:.: EMSYGLNVEMHKOTETAKRINTLINOLLPFLOADHOOQVLQA

68

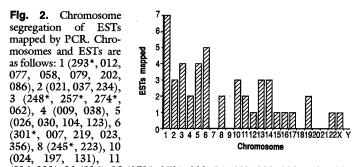
- 2 KKFRFEEPVVLPDLD......DOTXHROWTOOHLDAADLRM | . .:|| .:|: ::. :| |.|.||..|:| 192 PKR@RSDPVSGVGLGNNGGYASDHTMVSEYEEAD@RVMSQAHLDVVDVR.
- 37 XAMAPTPPQGEVDADCMDVNVRGPDGFTPIMIASCSGGGLETGNSEE..E | .|.:. |:: ||:.||| |:|||||...||||:||:. | | 241 ...AIMTPPAHQDGGKHDVDARGPCGLTPIMIAAVRGGGLDTGEDIENNE

85 EDAPAVISDFIYQGATCHNQTDRTGETALHLAA.VTYALMPQGLLRP.AK :..:||||:: ||]. :. |:|||.|||| .. | :. || : |. 288 DSTAQVISDLLAQGAELNATMDKTGETSLHLAARFARADAAKRLLDAGAD

- 133 MPTSGNMGF
- 338 ANCODNTGE

- 1 TKKFRFEEPVVLPDL.DDQTXHRQWTQQHLDAADLRWXAMAPTPPQGEVD .|:)||||.|:||:| ||.| .||||||||||||: .|||||||:: 1814 NKFFRFEEQVILPELVDDKTDFRQWTRQHLDAADLRISSMAPTPPQGEIE

similarity score of 0.1 to 0.4. Scores are from pairwise alignments based on the matrix of Schwartz and Dayhoff (31) as modified in the GCG package.



(016, 111), 12 (014), 13 (372\*, 273\*, 200) 14 (221, 201, 008), 15 (165), 16 (373\*), 17 (068), 19 (368\*, 080), and X (276\*). PCR conditions were as follows: 60 ng of genomic DNA was used as a template for PCR with 80 ng of each oligonucleotide primer, 0.6 unit of Taq polymerase, and 1  $\mu$ Ci of  $\alpha^{-32}$ P-labeled deoxycytidine triphosphate. The PCR was performed in a microplate thermocycler (Techne) under the following conditions: 30 cycles of 94°C, 1.4 min; 55°C, 2 min; and 72°C, 2 min; with a final extension at 72°C for 10 min. The amplified products were analyzed on a 6% polyacrylamide sequencing gel and visualized by autoradiography. Asterisks indicate those ESTs with similarity to GenBank or PIR sequences (Tables 2 and 3). EST names in GenBank are the three-digit number given here preceded by "EST00."

cell-cycle control genes in yeast and is tightly conserved in the *Xenopus laevis Notch* homolog, *Xotch*. In *Drosophila*, *Enhancer of split* is required for formation of epidermal tissue. *Notch* contains several epidermal growth factor–like repeats and appears to be involved in cell-cell communication during development (20).

Seven genes were represented by more than one EST. Comparisons of all the ESTs against one another revealed two overlaps of unknown ESTs: EST00233 and EST00234 matched in opposite orientations, and EST00235 and EST00236 matched in the same orientation beginning at the same nucleotide. Five human genes were represented by more than one EST:  $\beta$ -actin (three),  $\gamma$ -actin (two),  $\alpha$ -tubulin (two),  $\alpha$ -2-macroglobulin (two), and CNPase (two).

#### Mapping of ESTs to Human Chromosomes

We used the polymerase chain reaction (PCR) to screen a series of somatic cell hybrid cell lines containing defined sets of human chromosomes for the presence of a given EST (21). In this process, only the hybrids that contain the human gene corresponding to the EST will yield an amplified fragment. An EST is assigned to a chromosome by analysis of the segregation pattern of PCR products from hybrid DNA templates. The single human chromosome present in all hybrids that give rise to an amplified fragment is the location of the EST.

PCR mapping has been applied to 46 clones, as summarized in Fig. 2. The EST of the human gene for apolipoprotein J (also called SP-40,40 complement-associated protein and sulfated glycoprotein 2) was localized to chromosome 8. Eleven other ESTs with Gen-Bank or PIR similarities were mapped to chromosomes. Although PCR mapping of somatic cell hybrids is relatively rapid—up to three clones can be assigned per day with a single thermal cycler—it is relatively expensive, costing about ten times as much as EST sequencing. With the same oligonucleotide primers, sublocalization can be achieved with panels of fragments from specific chromosomes or pools of large genomic clones in an analogous manner. Other mapping strategies that have been proposed are multiplex in situ hybridization, prescreening with labeled flow-sorted chromosomes, and preselection by hybridization to construct chromosome specific cDNA libraries. However, these methods are limited by the purity of the chromosome-specific material or the specific activity necessary for detection.

# Automated DNA Sequencing Accuracy and GenBank Submission

ESTs that match human sequences in GenBank are excellent tools for the analysis of the accuracy of double-strand automated DNA sequencing. Ninety EST-GenBank matches were examined for the number of nucleotide mismatches and gaps required to achieve optimal alignment by the Genetics Computer Group (GCG) program BESTFIT (22). The number of mismatches, insertions, and deletions was counted for each hundred bases of the sequence (Table 5). As expected, the sequence quality was best closest to the primer and decreased rapidly after about 400 bases. The number of deletions and insertions relative to the GenBank reference sequence increased five- to tenfold beyond 400 bases, whereas the number of mismatches doubled. The average accuracy rate for individual double-stranded sequencing runs was 97.7% for up to 400 bases.

The minimum criteria for submission of ESTs to GenBank were that sequences be at least 150 bases in length and contain <3% ambiguous base calls. The overall accuracy of sequences submitted from each template group was at least 97%, based on matches to known human genes. Three hundred forty-eight ESTs met these criteria and were submitted to GenBank with accession numbers M61953 through M62300, inclusive. All ESTs except those matching mitochondrial or ribosomal RNA (rRNA) genes and simple repetitive elements were submitted to GenBank.

### **Conclusions and Prospects**

Single-run DNA sequencing has proven to be an efficient method of obtaining preliminary data on cDNA clones. Our results demonstrate that sufficient information is contained in 150 to 400 bases of a nucleotide sequence from one sequencing run for preliminary identification of the cDNA and localization to a chromosome. In addition to the 35 ESTs homologous to known human genes, 48 ESTs matched sequences in GenBank or PIR with moderate to striking similarity, including high-quality matches with genes from such evolutionarily distant organisms as yeast (EST00374) and *Neurospora* (EST00287) (Table 3).

Two hundred thirty ESTs did not match any current database entries and therefore represent new, previously uncharacterized genes. A multitude of approaches for classifying these genes exists, including complete sequencing and expression, chromosome mapping, tissue distribution, and immunological characterization. Currently unidentified cDNAs will also be classified by similarity to genes from other organisms as those sequences become available. Three ESTs reported here (EST00257, EST00259, and EST0374) were identified by similarity to sequences that have appeared since the last full release of GenBank.

The random selection approach used here revealed an unacceptably large number of highly represented clones in these cDNA libraries. Over 30% of the clones from the hippocampus cDNA library consisted of rRNA, mitochondrial cDNAs, or inserts consisting entirely of polyA. Sixty-eight ESTs matched 12 different mitochondrial genes, including 18 matches to cytochrome oxidase I. Although elimination of these uninformative clones is a priority for developing ideal cDNA libraries, techniques to reduce repeated sequencing of clones will become increasingly important as large numbers of cDNAs are sequenced. The use of library preprocessing techniques such as subtraction, which preferentially reduces the

population of certain sequences in the library (11, 12), and normalization, which results in all sequences being represented in approximately equal proportions in the library (23), should reduce repeated sequencing of high and intermediate abundance clones and maximize the chances of finding rare messages from specific cell populations. In our initial experiments with subtractive hybridization of the hippocampus library with a human fibroblast cDNA library, CNPase and GFAP clones were enriched greater than tenfold and twofold, respectively. Another characteristic of the ideal cDNA library would be directional cloning so that either a coding sequence or a 3' noncoding sequence could be selectively obtained.

The EST data, in conjunction with physical mapping, will provide a high resolution map of the location of genes along chromosomes, a map that would be more costly to construct by genomic sequencing and analysis. By performing a single DNA sequencing reaction on each cDNA clone, a key piece of information was obtained for the relatively low cost of about \$0.12 to \$0.15 per base. The EST approach will provide a new resource for the analysis of chromosome sequence and for human gene discovery.

The screening of cDNA clones to identify the protein complement of a tissue has been explored by others to a limited extent. In 1983, Putney and co-workers sequenced over 150 clones from a rabbit muscle cDNA library and identified clones for 13 of the 19 known muscle proteins, including one new isotype, but no unknown coding sequences (24). Over 400 adult head-specific cDNA clones from Drosophila have been identified by differential screening of cDNA libraries from different developmental stages (25). Improvements in DNA sequencing technologies have now made feasible essentially complete screening of the expressed gene complement of an organism.

In our own laboratory, the EST approach should result in the partial sequencing of most human brain cDNAs in a few years. Similar approaches begun elsewhere (26) could result in a database of most human expressed genes in less than 5 years. The presence of these minimally characterized sequences in GenBank will assist research efforts in several areas of biology. The EST database will provide identification and confirmation of coding regions in naive genomic sequences. Sublocalization of cDNAs that have been mapped to chromosomes will help define the genetic content of specific chromosomal regions and permit correlation with patterns of inheritance in genetic disease. In a related experiment, chromosome sublocalization was the key to establishing that the  $\gamma$ -aminobutyric acid-benzodiazepine receptor  $\beta_3$  subunit is deleted in individuals with Angelman-Prader-Willi syndrome (27). We anticipate that ESTs from human brain will further the identification of genes associated with other neurological diseases and will provide a more complete view of gene expression in the brain.

#### REFERENCES AND NOTES

- 1. J. G. Sutcliffe, Annu. Rev. Neurosci. 11, 157 (1988).
- 2. GenBank Release 66.0, December 1990
- 3. D. Baltimore, Ed., "Report of the Ad Hoc Program Advisory Committee on Complex Genomes," Reston, VA, February 1988 (National Institutes of Health, Bethesda, MD, 1988).
- S. Brenner, Ciba Found. Symp. 149, 6 (1990).
- "Report of the Committee on Mapping and Sequencing the Human Genome" (National Academy Press, Washington, DC, 1988).
- J. Fickett, Nucleic Acids Res. 10, 5303 (1982).
- M. Olson, L. Hood, C. Cantor, D. Botstein, Science 245, 1434 (1989) V. A. McKusick, Online Mendelian Inheritance in Man, from the Welch Medical 8. Library, Johns Hopkins University School of Medicine, Baltimore, MD.
- Stratagene catalog numbers 936206, 936205, and 935205, respectively
- Cycle sequencing was performed in a Perkin-Elmer Thermal Cycler for 15 cycles of 95°C, 30 s; 60°C, 1 s; 70°C, 60 s; and 15 cycles of 95°C, 30 s; 70°C, 60 s with the 10.
- Taq Dye Primer Cycle Sequencing Core Kit protocol (Applied Biosystems).
  D. W. Schmid and C. Girou, J. Neurochem. 48, 307 (1987); J. Fargnoli, N. J.
  Holbrook, A. J. Fornace, Jr., Anal. Biochem. 187, 364 (1990); J. R. Duguid and M.
  C. Dinauer, Nucleic Acids Res. 18, 2789 (1990); F. H. Travis and J. G. Sutcliffe, Proc. 11. Natl. Acad. Sci. U.S.A. 85, 1696 (1988); K. Kato, Eur. J. Neurosci. 2, 704 (1990). 12. C. W. Schweinfest et al., Genet. Anal. Tech. Appl. 7, 64 (1990).
- 13. H. L. Sive and T. St. John, Nucleic Acids Res. 16, 10937 (1988)
- 14. GenBank Release 65.0 and collective updates through 21 January 1990, PIR
- Release 26.0, and ProSite (EMBL protein motif database) Release 5.0 were used.
   14a. R. Fuchs, Comput. Appl. Biosci. 7, 105 (1991).
   15. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. Lipman, J. Mol. Biol. 215,
- 403 (1990).

- 403 (1990).
   16. W. R. Pearson and D. J. Lipman, Proc. Natl. Acad. Science. U.S.A. 85, 2444 (1988).
   17. R. D. Vale, T. S. Reese, M. P. Sheetz, Cell 42, 39 (1985).
   18. J. A. Campos-Ortega, Trends Neurosci. 11, 400 (1988).
   19. D. A. Hartley, A. Preiss, S. Artavanis-Tsakonas, Cell 55, 785 (1988); C. Klämbt, E. Knust, K. Tietze, J. A. Campos-Ortega, EMBO J. 8, 203 (1989).
   20. U. Banerjee and S. L. Zipursky, Neuron 4, 177 (1990).
   21. We designed olicingnucleatide primer pairs from EST sequences to minimize the
- We designed oligonucleotide primer pairs from EST sequences to minimize the chance of amplifying through an intron. Introns are less likely to be present in 3'untranslated sequences, so ESTs were examined for the presence of stop codons in each reading frame and for consensus splice junctions [J. D. Hawkins, Nucleic Acids Res. 16, 9893 (1988); M. B. Shapiro and P. Senapathy, *ibid*. 15, 7155 (1987)]. Also desirable for mapping are 3' untranslated sequences because they are less likely Also desirable for mapping are 3' untranslated sequences because tney are less inkey to be identical in other members of a gene family or in pseudogenes. We used the primers in a PCR to amplify from total human genomic DNA templates as described in the legend to Fig. 2. If the resulting product was of the expected size, then the PCR reaction was repeated with DNA templates from two panels of human-rodent somatic cell hybrids: Bios PCRable DNA (BIOS Corporation) and NIGMS Human-Rodent Somatic Cell Hybrid Mapping Panel Number 1 ONCOLE Conden NU (NIGMS, Camden, NJ)
- 22. J. Devereux, P. Haeberli, O. Smithies, Nucleic Acids Res. 12, 387 (1984).
- J. Deversux, T. Hactern, O. Smithnes, Nucleic Acids Res. 12, 387 (1984).
  S. R. Patanjali, S. Parimoo, S. M. Weissman, Proc. Natl. Acad. Sci. U.S.A. 88, 1943 (1991); M. S. H. Koch, Nucleic Acids Res. 18, 5705 (1990).
  S. D. Putney, W. C. Herlihy, P. Schimmel, Nature 302, 718 (1983).
  M. J. Palazzolo et al., Neuron 3, 527 (1989). 23.
- 24.
- 25.

- M. J. Falazzolo et al., Neuron 3, Ser (1997).
   Hum. Genome News 2 (no. 6), 1 (1991).
   J. Wagstaff et al., Am. J. Hum. Genet., in press.
   R. Cathcart, Nature 347, 310 (1990).
   S. Nishi, S. Seino, G. I. Bell, Biochem. Biophys. Res. Commun. 157, 937 (1988).
- Abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
   R. O. Schwartz and M. O. Dayhoff, in *Atlas of Protein Sequences and Structure*, M. O. Dayhoff, Ed. (National Biomedical Research Foundation, Washington, DC, 1077) 1272 259
- 1979), pp. 353-358.
- 32. Dedicated to the memory of J. E. Venter, 18 August 1923 to 10 June 1982.