well as their overall quantity as a subfraction of DOC, is only summarily known[2,3,8,28]. Likewise, we lack precise knowledge about the real straining (and aggregation or agglutination) capacity of oikopleurid food-concentrating filters[12]. But even a very conservative estimate of 10% of DOC as grazable by oikopleurids, means that this source of food is as important for the oikopleurids as is total POC. This agrees well with previous findings that POC accounts for a maximum of 30% of the energy needs of *O. dioica*[19,20].

Oikopleurid tunicates often occur in high densities in discrete strata at various depths[21,22], and may under such conditions clear 30–60% of the water mass in 24 h[1,4]. Obviously, they may remove and repack colloidal DOC (>0.2 μm particle size) rapidly under such conditions. On the basis of filter parameters, this ability to graze on colloidal DOC is probably shared by caddisfly larvae[23], pedal worms[24], ascidians[25], salps[26] and amphioxus[9]. □

Received 10 September; accepted 8 November 1991.

1. Alldredge, A. L. *Limnol. Oceanogr.* **26,** 247–257 (1981).
2. Sugimura, Y. & Suzuki, Y. *Mar. Chem.* **24,** 105–131 (1988).
3. Koike, I., Hara, S., Terauchi, K. & Kogure, K. *Nature* **345,** 242–244 (1990).
4. Knoechel, R. & Steel-Flynn, D. *Mar. Ecol. Prog. Ser.* **53,** 257–266 (1989).
5. Shelbourne, J.E. *J. mar. biol. Ass. U.K.* **42,** 243–252 (1962).
6. Gadomski, D. M. & Boelert, G.W. *Mar. Ecol. Prog. Ser.* **20,** 1–12 (1984).
7. Keats, D. W., Steele, D. H. & South, G. R. *Canad. J. Zool.* **65,** 49–53 (1987).
8. Flood, P. R. *Experientia* **34,** 173–175 (1978).
9. Flood, P. R. *Biomed. Res. Suppl. 2* 49–53 (1981).
10. Deibel, D., Dickson, M.-L. & Powell, C. V. L. *Mar. Ecol. Prog. Ser.* **27,** 79–86 (1987).
11. Deibel, D. & Powell, C. V. I. *Mar. Ecol. Prog. Ser.* **39,** 81–85 (1987).
12. Flood, P. R. *Mar. Biol.* **111,** 95–111 (1991).
13. Deibel, D. & Powel, C. V. L. *Mar. Ecol. Prog. Ser.* **35,** 243–250 (1987).
14. Flood, P. R., Deibel, D. & Morris, C. C. *Biol. Bull. Mar. biol. Lab., Woods Hole* **178,** 118–125 (1990).
15. Johnson, B. D. & Wangersky, P. J. *Limnol. Oceanogr.* **30,** 966–971 (1985).
16. Deibel, D. *Mar. Biol.* **99,** 177–186 (1988).
17. Bagnara, J. T. & Hadley, M. E. *Chromatophores and Color Change, the Comparative Physiology of Animal Pigmentation* 46–50 (Prentice-Hall, Englelwood Cliffs, NJ, 1973).
18. Cauwet, G. *Oceanologica Acta* **1,** 99–105 (1978).
19. Gorsky, G. thesis, Univ. de P. et. M. Curie, Paris VI (1980).
20. King, K. R. thesis, Univ. Washington (1981).
21. Youngbluth, M. J., Bailey, T. G. & Jacoby, C. A. in *Man in the Sea* (eds Lin, Y. C. & Shida, K. K.) Vol. 2 191–208, (Best, San Pedro, California, 1990).
22. Magnesen, T., Aksnes, D. L. & Skjoldal, H. R. *Sarsia* **74,** 115–126 (1989).
23. Wallace, J. B. & Malas, D. *Arch. Hydrobiol.* **77,** 205–212 (1976).
24. Flood, P. R. & Fiala-Medioni, A. *Mar. Biol.* **72,** 27–33 (1982).
25. Flood, P. R. & Fiala-Medioni, A. *Acta Zool. Stockh.* **62,** 53–65 (1981).
26. Bone, Q., Braconnot, J.-C. & Ryan, K. P. *Acta Zool.* Stockh. **72,** 55–60 (1991).
27. Harbison, G. R. and McAlister, V. L. *Limnol. Oceanogr.* **24,** 875–892 (1979).
28. Williams, P. M. & Druffel, E. R. M. *Oceanography* **1,** 14–17 (1988).

ACKNOWLEDGEMENTS. We thank E. Bru, B. Hansen, M. Riehl and the scuba divers of the Ocean Science Centre for technical assistance. This work was supported by the Norwegian Fisheries Research Council, the Norwegian Research Council for Science and the Humanities (to P.R.F.) and by the Natural Sciences and Engineering Research Council of Canada (to D.D.).

Sequence identification of 2,375 human brain genes

Mark D. Adams, Mark Dubnick, Anthony R. Kerlavage, Ruben Moreno, Jenny M. Kelley, Teresa R. Utterback, James W. Nagle, Chris Fields & J. Craig Venter*

Receptor Biochemistry and Molecular Biology Section, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, Maryland 20892, USA

WE recently described a new approach for the rapid characterization of expressed genes by partial DNA sequencing to generate 'expressed sequence tags'[1]. From a set of 600 human brain complementary DNA clones, 348 were informative nuclear-encoded messenger RNAs. We have now partially sequenced 2,672 new, independent cDNA clones isolated from four human brain cDNA libraries to generate 2,375 expressed sequence tags to nuclear-encoded genes. These sequences, together with 348 brain expressed sequence tags from our previous study, comprise more than 2,500 new human genes and 870,769 base pairs of DNA sequence. These data represent an approximate doubling of the number of human genes identified by DNA sequencing and may represent as many as 5% of the genes in the human genome.

Most (83%) of the 2,375 partial cDNA sequences reported here (Table 2) are not related to any previously described sequences. Based on database matches to known genes from humans and from such evolutionarily distant organisms as *Escherichia coli*, yeast, *Caenorhabditis elegans*, *Drosophila*, barley, *Arabidopsis*, rice and green algae, we have putatively identified 217 of the expressed sequence tags (ESTs; Table 1). These include a novel gene similar to *Notch/TAN-1* (refs 1, 2), a new neurotransmitter transporter gene, and a new member of the multidrug resistance gene family. Several genes involved in development or cell differentiation in *Drosophila* are represented by similar human ESTs, including *seven in absentia*[3], *big-brain*[4], the *discs-large* tumour suppressor[5] and the homeotic gene *orthodenticle*[6]. New members of previously known gene families in humans include a Ca^{2+}-transporting ATPase, an ADP ribosylation factor and a new neural-cell adhesion molecule gene.

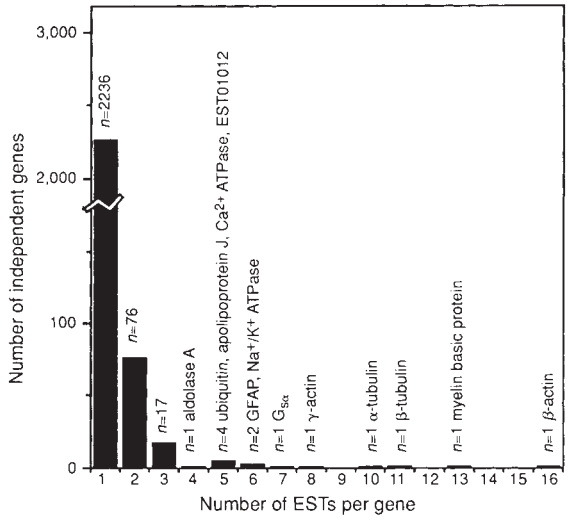

FIG. 1 Redundancy of sequencing of ESTs. The number of putatively identified EST clones plus the groups of ESTs that form contigs are plotted against the number of independent genes represented. The number of genes is given above each bar, with the names of the genes for redundancies of 4–16. We define redundancy as the number of times each gene is represented by an EST; for example β-actin has an EST redundancy of 16. One five-member EST contig was constructed that did not match any known sequences, indicating that at least one common transcript in brain, EST 01012, has not been reported previously. GFAP, glial fibrillary acidic protein.

The 1,971 ESTs without a putative identification were analysed using the coding-region prediction program CRM with the GRAIL server[7]. Some of the unknown ESTs (15%) scored a likely probability of containing protein-coding sequence. Half of the ESTs to known human genes contain protein-coding sequences, so at most half of the unknown ESTs are likely to contain coding sequences. We have found no evidence that genomic DNA or cDNA to unspliced precursor RNA is a major contaminant of either the hippocampus or fetal brain library.

The limited extent of redundancy of EST sequencing is shown in Fig. 1. Of the nuclear-encoded messenger RNAs, the most common ESTs were to the β-actin (0.6% of the EST clones)

* To whom correspondence should be addressed.

TABLE 1 Gene composition of human brain cDNA libraries

EST	Putative Identification	Species	DB	Accession	Length	%ID
EST02243	14.3.3 protein (PK† regulator)	H	(G)	HUM1433	272	98.9
EST01757	2',3'-cyclic nucleotide phoshodiesterase	H	(G)	HUMCNPDEA	272	98.5
EST01784	60K filarial antigen	A	(P)	A28209	88	50.6
EST01455	80K-H protein	H	(G)	HUMG19P1A	263	97.3
EST01982	ADP-ribosylation factor 1	H	(P)	B33283	84	41.2
EST02077	ADP/ATP translocase	H	(G)	HUMTLCA	200	96.5
EST01620	AMP deaminase, brain	R	(P)	A37056	57	100.0
EST01504	Actin, beta, cytoskeletal	H	(G)	HUMACCYBA	247	98.8
EST01543	Actin, gamma, cytoskeletal	H	(G)	HUMACTGNM	359	98.9
EST01891	Actinin, alpha	H	(G)	HUMACTAR	315	81.6
EST01801	Adaptin, beta	H	(G)	HUMBADPTA	380	99.5
EST01710	Adenosine deaminase	H	(G)	HUMADAG	316	100.0
EST01625	Agrin	R	(G)	RATAGR	103	84.6
EST02113	Ala	H	(G)	HUMALA	92	82.8
EST00675	Alcohol dehydrogenase	I	(H)	RICGOS2G_1	38	59.0
EST01660	Aldolase A	H	(G)	HUMALDA	346	98.0
EST01761	Aldolase C	H	(G)	HUMALDC	317	99.0
EST02688	Alpha-Enolase	H	(G)	HUMENOA	345	98.8
EST01635	Alpha-2-Macroglobulin	H	(G)	HUMA2M	319	99.4
EST01664	Amyloid A4	H	(P)	A29030	52	54.7
EST01700	Anion exchanger homolog AE3	M	(P)	A33638	95	97.9
EST01585	Apolipoprotein J	H	(G)	HUMAPOJ	317	99.0
EST01825	Aspartate aminotransferase, cytosolic	H	(G)	HUMASAM	309	98.7
EST02671	Aspartate aminotransferase, mitochondria	H	(G)	HUMASPAT	231	99.1
EST01634	Axonal glycoprotein TAG-1	R	(P)	A34695	69	87.1
EST02530	B cell-specific Mo-MLV integration site	M	(G)	MUSHMI1A	111	87.5
EST02306	Bib protein	D	(P)	S09699	57	53.4
EST01443	CDP-SO-p†	E	(P)	JH0368	33	41.2
EST01800	Ca2+-transporting ATPase	H	(G)	HUMBK1A	380	99.2
EST02146	Calbindin D28	R	(G)	RATCALBD28	81	87.8
EST01823	Calcineurin A2	H	(G)	HUMCNRAB	287	99.7
EST02055	Calcium channel	L	(P)	S05054	33	67.6
EST01849	Calmodulin	H	(G)	HUMCAMA	327	98.5
EST01466	Calmodulin-dependent PK†, βII	R	(P)	A26464	93	98.9
EST02378	cAMP-dependent PK† inhibitor	M	(G)	MUSPKI	234	91.5
EST01644	cAMP-dependent PK† regulatory subunit RIß	H	(G)	HUMRIB	198	97.5
EST01628	cAMP-dependent PK† regulatory subunit RIIß	H	(G)	HUMPKCAMD	294	98.3
EST01041	cAMP-regulated phosphoprotein	B	(P)	B35308	21	86.4
EST02447	cAMP-specific phosphodiesterase	H	(G)	HUMPDEAA	363	69.0
EST01536	Cannabinoid receptor	H	(S)	CANRSHUMAN	97	93.9
EST01795	Carboxypeptidase E	H	(G)	X51405	330	99.4
EST01606	Casein kinase II alpha subunit	H	(G)	HUMACKII	186	97.3
EST01733	Catalase	H	(G)	HUMCATR	270	96.3
EST01810	Cathepsin D	H	(G)	HUMCTHD	246	98.4
EST01799	Cell surface glycoprotein MUC18	H	(G)	HUMMUC18B	413	97.8
EST01487	C/D PG40†	H	(G)	HUMCSPG1A	261	100.0
EST01913	Clathrin coat assembly protein homologue	Y	(H)	YSCYAP54_1	62	63.5
EST01796	Coagulation factor VII	H	(G)	HUMCFVII	227	97.8
EST01676	Cofilin	P	(G)	PIGCOFIL	132	89.5
EST01774	Cysteine proteinase inhibitor	H	(G)	HUMCYSTCR	163	97.6
EST01824	Cysteine-rich intestinal protein	R	(P)	GYRTI	56	66.7
EST01502	Cytochrome P-450IIE1	H	(G)	HUMCYPIIE	175	97.2
EST01951	Cytochrome c	H	(G)	HUMCYCAA	172	96.0
EST01808	DNA binding protein YAVREB67	H	(G)	HUMYAVREB	429	98.8
EST01721	DNA repair helicase (ERCC3)	H	(G)	HUMERCC3A	334	98.5
EST01622	DNF1552 (lung) mRNA	H	(G)	HUMDNFA	137	96.4
EST01257	Diacylglycerol kinase, lymphocyte	P	(P)	S09156	44	42.2
EST02477	Diamine acetyltransferase	H	(S)	ATDASHUMAN	74	45.3
EST01508	Dihydrolipoamide acetyltransferase	H	(G)	HUMPDCE2	254	98.4
EST00642	Dilute (myosin heavy chain)	M	(H)	MUSDILUTE_1	27	100.0
EST01779	Discs-large tumor suppressor	D	(H)	DRODLGA_1	53	63.0
EST02627	Elongation Factor 1 alpha	H	(G)	HUMEF1AR	361	99.2
EST01165	Elongation Factor 1 beta	N	(P)	A24806	36	64.9
EST02596	Enolase, gamma-2, neuron specific	H	(G)	HUMENOG	345	98.5
EST01743	Epoxide hydrolase	H	(G)	HUMEH	363	99.7
EST01946	Ezrin	H	(G)	HUMEZRIN	153	99.3
EST01971	Familial adenomatous polyposis coli	H	(G)	HUMFAPAPC	316	98.1
EST01325	Fatty acid synthase	R	(G)	RATFAS	98	79.8
EST01476	Fibrillarin	H	(G)	HUMFIBAA	201	99.5
EST01790	Fibroblast growth factor receptor	H	(G)	HUMFGF1A	376	99.2
EST01967	Fibronectin	H	(G)	HUMFNMC	120	98.3
EST02186	Fodrin, alpha	H	(G)	HUMASPX	282	98.2
EST02428	Fumarate Hydratase, mitochondrial	H	(G)	HUMFUMA	96	96.9
EST01665	G(i) alpha	H	(G)	HUMGIAA	310	98.1
EST02389	G(s) alpha	H	(G)	HUMGSA1R	275	99.3
EST02362	GA binding protein, beta subunit	M	(H)	MUSGAC_1	86	90.8
EST01745	GTP-binding protein beta chain homologue	H	(G)	HUMHBA123	319	99.1
EST00825	Gamma-aminobutyric acid transporter	R	(P)	A35918	26	59.3
EST01738	Gelation factor ABP-280	H	(P)	A37098	74	80.0
EST01776	Gelsolin precursor, plasma	H	(G)	HUMPGSR	171	99.4
EST01649	Glial fibrillary acidic protein	H	(G)	HUMGFAP	399	99.2
EST01965	Globin, gamma	H	(G)	HUMGGL3	147	100.0
EST02192	Glutamate decarboxylase	H	(G)	HUMGAD	213	100.0
EST01702	Glutamate dehydrogenase	H	(G)	HUMGDHL	223	96.4
EST02446	Glutamate-aspartate carrier protein	E	(P)	JV0092	57	37.9
EST02034	Glutaminase	T	(S)	GLSSRAT	34	74.3
EST01576	Glyceraldehyde-3-phosphate dehydrogenase	H	(G)	HUMG3PD	428	99.8
EST01621	HLA-B-associated transcript 2 (BAT2)	H	(G)	HUMBAT2A	301	98.7
EST01462	Heat shock cognate protein 70	H	(G)	HUMHSC70	290	97.2
EST01637	Heat shock protein 90	H	(G)	HUMHSP90	265	99.6
EST01610	Heparin-binding growth factor 1	H	(G)	X51943	235	94.5
EST02315	Hexokinase I	H	(P)	A31869	107	96.3
EST01709	High density lipoprotein binding protein	H	(G)	HUMHBP	282	99.6
EST01639	Histocompatibility antigen modifier 1	M	(P)	A37779	63	75.0
EST02556	Histone H1(0)	H	(E)	HSHIS10G	172	97.7
EST02528	hnRNP core protein A1	H	(G)	HUMHNRNPA	151	84.9
EST02533	Hypothetical 43.5K protein	M	(P)	JU0319	43	52.3
EST01531	Hypothetical protein (similar to BCR)	H	(P)	S09291	104	95.2
EST01861	IMP dehydrogenase	H	(G)	HUMIMP	298	99.7
EST01791	Inositol-1,4,5-trisphosphate 3-kinase	H	(P)	JN0129	65	68.2
EST01605	Insulin-like growth factor II receptor	H	(G)	HUMIGFIIR	240	95.0
EST01813	Interferon regulatory factor 1	H	(G)	HUMIRF1	174	98.3
EST01529	Interferon-induced 54K protein	H	(S)	INI4SHUMAN	76	70.1
EST02051	J1 protein	M	(G)	MUSJ1PRO	362	85.7
EST02258	KUP protein	H	(H)	HUMKUPMR_1	54	36.4
EST02007	Kinase 5 protein	C	(H)	CHKCEK5_1	68	94.2
EST01748	Kinesin	T	(P)	A35075	98	52.5
EST01491	Ku autoimmune antigen	H	(G)	HUMKUANT	288	99.7
EST01651	Lactate dehydrogenase A	H	(G)	HUMLDHA	378	99.5
EST01470	Lactate dehydrogenase B	H	(G)	HUMLDHBR	401	99.5
EST01764	Lamin B receptor	C	(P)	A36427	76	71.4
EST01724	Lon protease	E	(P)	JQ0901	103	41.3
EST02463	Long-chain-fatty-acid-CoA ligase	R	(P)	A36275	36	62.2
EST02418	MARCKS homolog	M	(E)	MMF52	237	92.4
EST01952	MHC class I HLA-Cw2 heavy chain	H	(G)	HUMHCW2B	231	97.8
EST01640	MHC class III HSP70-1 gene (HLA)	H	(G)	HUMHHSP	181	100.0
EST02505	Matrin 3	R	(G)	RATMATRIN3	137	93.5
EST01918	Metalloproteinase inhibitor	H	(G)	HUMMET	236	99.2
EST01865	Microtubule-associated protein 1B	R	(G)	RATNEU	293	86.4
EST01473	Microtubule-associated protein 4	H	(G)	HUMMAP4	249	97.6
EST01678	Milk fat globule membrane protein	M	(P)	A36479	48	61.2
EST01704	Monoamine oxidase A (MAOA)	H	(G)	HUMMOAA	255	98.8
EST01629	Myelin basic protein	H	(G)	HUMMBPB	314	99.4
EST01580	Myeloid differentiation primary response	M	(H)	MUSMYD118_1	76	88.3
EST02585	Myoblast cell surface antigen 24.1D5	H	(G)	HUM241D5	177	98.3
EST01614	Myosin heavy chain, non-muscle	H	(G)	HUMMYONM	291	99.3
EST01672	N-formylpeptide receptor	H	(G)	HUMFMLP	237	99.2
EST01744	NAD(P)+ transhydrogenase (B-specific)	B	(P)	DEBOXM	86	93.1
EST01338	NF-kappa-B transcription factor	H	(G)	HUMP50NFKB	98	98.0
EST01805	Na+/K+ ATPase, alpha subunit	H	(G)	HUMATP1A2	277	99.6
EST02610	Neural cell adhesion molecule L1	M	(P)	S05479	82	43.4
EST01471	Neuraxin	R	(P)	S06017	120	84.3
EST01519	Neurofibromatosis type 1	H	(G)	HUMNF1MRNA	369	98.6
EST01749	Neurofilament heavy chain	H	(G)	HUMNFH4	356	99.7
EST02455	Neurofilament subunit M (NF-M)	H	(G)	HUMNFM	164	99.4
EST02609	Neutrophil oxidase factor	H	(P)	A34855	43	47.7
EST00632	Nodulation protein I	K	(P)	A24400	63	39.1
EST01643	Nodulation protein I	K	(P)	A24400	71	50.0
EST01961	Notch/Xotch homologue	H	(H)	HUMTAN1_1	85	57.0
EST02429	Nuclear factor 1-like protein (NF1)	H	(G)	HAMNF1A	111	92.0
EST01573	Nucleoside diphosphate kinase	H	(P)	A33386	71	52.8
EST01657	Osteonectin/SPARC	H	(G)	HUMSPARC	348	99.7
EST01822	Osteopontin	H	(G)	HUMOP	170	96.5
EST01828	Otd homeotic protein	D	(P)	A35912	35	52.8
EST01486	Pancreatic tumor-related protein	H	(G)	HUMPANCAN	422	99.3
EST01798	Peptidylprolyl isomerase (cyclophilin)	H	(G)	HUMCYCR	382	99.5
EST01751	P-4,5-BPI†	R	(P)	A28807	40	90.2
EST01656	Phosphoglycerate kinase	H	(G)	HUMPGKA11	374	98.7
EST00992	Polymyxin B resistance	Y	(P)	A32714	20	76.2
EST01806	Prohibitin	R	(H)	RATPROHIB_1	120	97.5
EST01775	Prohormone cleavage enzyme	M	(H)	MUSMPC1A_1	91	93.5
EST01461	Protein kinase C, beta subunit	H	(G)	HUMPKCBA	155	94.9
EST02087	Protein kinase C, zeta	H	(G)	HUMPKCL	382	58.7
EST01650	Protein phosphatase 2A beta subunit	H	(G)	HUMPROP2AB	288	76.8
EST01584	PR65 (alpha)†	H	(G)	HUMP2A	242	98.8
EST01786	Protein-tyrosine kinase (clone JTK14)	H	(P)	C38269	38	100.0
EST01572	Protochlorophyllide reductase	W	(P)	S04783	34	57.1
EST01538	Pyruvate dehydrogenase alpha subunit	H	(G)	HUMPDHA	322	99.7
EST01587	Pyruvate kinase isozyme M2	H	(G)	HUMPKM2L	395	98.5
EST02683	Rab11	H	(H)	HUMRASLP_1	78	97.5
EST02515	Rab5	H	(P)	F34323	91	82.6
EST01072	Rac protein kinase	H	(G)	HUMRACPC	91	100.0
EST01389	Radial spoke protein 3	F	(P)	S05962	58	52.5
EST01787	Retinaldehyde-binding protein	H	(G)	HUMRALBPC	129	100.0
EST01579	Retrovirus-related gag polyprotein	H	(P)	FOHUE2	95	77.1
EST02550	Retrovirus-related pol polyprotein	H	(P)	GNLJGL	50	54.9
EST01578	Ribophorin II	H	(G)	HUMRIBIIR	289	99.3
EST01928	Ribosomal phosphoprotein P0	H	(G)	HUMPPARP0	273	98.2
EST02158	Ribosomal phosphoprotein P1	H	(G)	HUMPPARP1	126	91.3
EST01583	Ribosomal protein l18a	R	(P)	R5RT18	68	95.7
EST01627	Ribosomal protein l1a	X	(P)	A24579	75	63.1
EST01667	Ribosomal protein L3	Q	(P)	JQ0771	74	80.0
EST01826	Ribosomal protein S10	S	(P)	R3YM10	36	51.4
EST01459	Ribosomal protein YL10	Y	(P)	S11581	40	68.3
EST01608	Ribosomal protein S3	H	(G)	HUMHUMS3	279	95.7
EST02442	Seven in absentia	D	(P)	A36195	46	80.8
EST01960	Spectrin, beta	H	(G)	HUMSPTB	268	67.7
EST01699	Sperm membrane protein	R	(P)	A35981	52	58.5
EST01760	Spermidine/spermine N1-acetyltransferase	H	(G)	HUMSPDMAT	102	97.1
EST01545	Sphingomyelin phosphodiesterase	H	(G)	HUMASM2MR	159	97.5
EST01984	Stathmin (p18)	H	(G)	HUMP18	180	100.0
EST01697	Succinate dehydrogenase flavoprotein	B	(H)	BOVSDHFP1_1	44	100.0
EST00472	Synaptotagmin (p65)	H	(S)	SY65SHUMAN	27	53.6
EST01586	T-cell surface glycoprotein E2	H	(G)	X16996	277	99.6
EST01809	TAPA-1, 26-kDa cell surface protein	H	(G)	HUMTAPA1	271	98.5
EST01955	TCTE1†	H	(G)	HUMTCTEB10	162	98.8
EST01575	TFEB transcription factor	H	(G)	HUMTFEB	369	98.1
EST02402	Talin	M	(H)	MUSTALINR_1	79	81.2
EST01601	Thiosulfate sulfurtransferase (rhodanese)	B	(P)	ROBO	65	81.8
EST01435	Threonyl-tRNA synthetase	H	(G)	HUMTHRSYNT	228	99.6
EST02420	Thrombospondin precursor	H	(G)	HUMTSPA	58	98.3
EST01783	Thy-1 glycoprotein	H	(G)	HUMTHY1A	353	98.6
EST01235	Thymosin beta-10	H	(G)	HUMTHYB10	88	97.8
EST02451	Transcription factor ETR101	H	(G)	HUMETR101	287	99.3
EST01705	Transducin beta-2 subunit	H	(G)	HUMTRNB	126	100.0
EST02047	Triose-phosphate isomerase	H	(G)	HUMTPI	258	98.8
EST01512	Tubulin, alpha	H	(G)	HUMTUBAG	223	75.0
EST01490	Tubulin, beta	H	(G)	HUMTBB5	298	93.6
EST01530	Tubulin, beta	H	(G)	HUMTBB5	212	99.1
EST02169	Tyrosine kinase	H	(G)	HUMECK	384	74.3
EST01557	Ubiquitin	H	(G)	HUMUBCP	329	97.3
EST01769	Ubiquitin activating enzyme E1	H	(G)	HUMUBIQAA	362	91.2
EST00853	Unc-104 (kinesin homologue)	J	(P)	JN0114	37	45.5
EST01687	Uroporphyrinogen decarboxylase	H	(G)	HUMUROD	297	99.3
EST02557	Vacuolar H+ ATPase proton channel	H	(G)	HUMPCHSUCA	152	93.5
EST01711	Valine-tRNA ligase	Y	(P)	A29871	56	57.9
EST01556	Wilm's tumor-related protein	H	(G)	HUMQM	363	97.8
EST01588	XPR2 alkaline extracellular protease	Y	(P)	B26955	88	46.1
EST01511	Y box binding protein-1 (YB-1)	H	(G)	HUMYB1A	135	97.0
EST01352	Zinc finger protein kox1	H	(P)	S10397	29	56.7
EST01460	Zinc finger protein 9	H	(P)	C32891	34	54.3
EST01446	Zinc finger protein HF.12	H	(P)	S00754	45	60.9
EST01560	Zinc finger protein HF.12	H	(P)	S00754	105	67.0
EST02259	Zinc finger protein HF.12	H	(P)	S00754	71	62.5
EST02534	Zinc finger protein 7	H	(P)	A34612	50	45.1

Gene names are listed with representative ESTs. Quality of match is given as per cent identity (%ID) and length in base pairs for nucleotide matches and in amino acid residues for peptide matches. In many cases ESTs match multiple domains on several related proteins; for example, EST00825 matches two transmembrane domains on both GABA (γ-aminobutyric acid) and noradrenaline transporters. Where more than one EST matched a gene, one representative match is given. Nucleotide databases (DB): GenBank (G) and EMBL (E); peptide databases: GenPept (H), Swiss-Prot (S) and PIR (P). Species names: A, *Brugia malayai*; B, bovine; C, chicken; D, *Drosophila*; E, *E. coli*; F, green alga; H, human; I, rice; J, *C. elegans*; K, *Rhizobium*; L, rabbit; M, mouse; N, Brine shrimp; P, pig; Q, *Arabidopsis*; R, rat; S, *Mycoplasma capricolum*; T, squid; W, barley; X, *Xenopus laevis*; Y, yeast.

† Abbreviations: CDP-SO-P, CDPdiacylglycerol-serine *O*-phosphatidyltransferase; PK, protein kinase; C/D PG40, chondroitin/dermatin sulphate proteoglycan core protein; P-4,5-BPI, phosphatidylinositol-4,5-biphosphate phosphodiesterase I; alpha-PR65, protein phosphatase alpha-2A regulatory subunit; TCTE 1, human t-complex associated testes-expressed-1 homologue.

TABLE 2 Gene composition of human brain cDNA libraries

	Total	Hippocampus unscreened	Hippocampus prescreened	Fetal brain*	Fetal brain†	Whole brain‡
No database match	1,942	474	394	1,025	32	17
Exact human match	255	76	88	87	0	4
Non-exact human match	51	8	10	33	0	0
Non-human match	99	34	26	35	2	2
Alu repeat	313	70	70	172	1	0
L1 repeat	58	13	5	39	0	1
THE-Itr sequence	17	1	1	14	1	0
Other repeat	9	3	4	2	0	0
Mitochondrial	181	115	33	27	0	6
rRNA	57	29	7	16	5	0
poly(A) insert only	339	171	161	5	0	2
Total	3,321	994	799	1,455	41	32

Four cDNA libraries were used as sources of clones for sequencing. Human hippocampus and fetal brain (*) libraries, plasmid template preparation, sequencing reactions, and automated sequencing were performed as described[1]. A pooled probe consisting of inserts from 10 different EST clones with sequences that matched either mitochondrial genes or the 18S or 28S rRNAs was used to prescreen a gridded filter array of the hippocampus library; nonhybridizing clones are referred to as the 'prescreened library'. Another fetal brain library (†) was constructed by and was a gift from B. Soares (Columbia University). A directionally-cloned library (‡) was prepared[14] using human adult brain mRNA from Clontech (Palo Alto). Of 482 clones analysed by restriction-enzyme digestion, 33% contained inserts at least 1,500 base pairs long. Stratagene hippocampus and fetal brain library totals include data from ref. 1. Sequences of nuclear-encoded cDNAs that did not include the interspersed repeats Alu, L1 or THE-Itr[15-17] were searched against GenBank and, in 6-frame translation, against a comprehensive, non-redundant peptide database using the network BLAST[18] server at the National Center for Biotechnology Information. For significant similarities, a putative gene name and protein identification resource (PIR) gene family identification[19] for the EST were assigned. ESTs without significant matches using BLAST were searched in translation against PIR using the program FASTA. Ten additional marginal matches were found. A total of 2,300 new EST sequences comprising 765,505 nucleotides from the current data set have been submitted to GenBank and assigned accession numbers M77851–M79278 and M85308–M86179. All ESTs except those multiply representing actin, tubulin, and myelin basic protein clones were submitted. cDNA clones from which ESTs were derived are available from the American Type Culture Collection (Rockville, Maryland) with accession numbers 77501-78999 and 81000–81756. The Genome Data Base[20] expressed D-segment numbers for these clones are DOS1E–DOS2300E. We have developed a database which includes the clone and sequence data, sequence analysis results, physical mapping data, tissue localization and cross-references to the public databases and distribution of the clones, mapping and sequence data using the Sybase relational database management system (Sybase Inc., Emeryville, California). Comprehensive reports on the sequences described in this paper are available in electronic form. A README file describing how to access the EST database reports is available via anonymous file transfer protocol (FTP) to briggs.ninds.nih.gov. Questions on data access and database structure can be addressed via electronic mail to arkerlav@briggs.ninds.nih.gov.

and myelin basic protein genes (0.5% of the clones). Myelin basic protein, a highly expressed structural component of nerve tissue[8], displays four alternate splicing forms, of which at least two are present among the ESTs reported here. Other common ESTs were G-protein subunit $G_{s\alpha}$, γ-actin and both α- and β-tubulin.

All of the genes for which four or more ESTs were found have been sequenced in humans, except for one which was matched by five unknown ESTs. Assuming that most brain mRNAs are rare transcripts[9], the chance of finding a new gene by EST sequencing is fairly high when ribosomal and mitochondrial transcripts are eliminated. Therefore, although normalization may be important as we near closure in sequencing every human gene, it is not necessary at this stage to reduce sequencing redundancy or to increase gene diversity. Furthermore, a certain amount of redundancy is desirable to the extent that it promotes assembly of EST contigs into full-length cDNA sequences.

By matching ESTs to known database sequences, a phenotypic characterization of the tissue begins to emerge. Protein superfamilies matched by ESTs were grouped into three broad functional categories to assess the biological spectrum represented by these randomly selected cDNA clones. Structural and metabolic classes comprised about 30% of the ESTs each, 25% were involved in regulatory pathways and the remainder were not classifiable. Eleven of the eighteen enzymes of glycolysis and the citric acid cycle are represented by at least one subunit or isozyme. In addition, several genes not previously known to be expressed in the brain were matched, including spermine/spermidine acetyltransferase[10] and osteopontin[11]. Isolation of 171 ESTs from mouse testes was recently reported[12], including four with database matches in common with human ESTs.

The genomic mapping of these new human expressed genes is among our highest priorities. Physical mapping of the 2,375 EST clones reported here would provide human chromosome markers spaced an average of 1.2 megabases apart and would roughly double the number of expressed sequences that have been localized to chromosomes[13]. Mapped ESTs are a new resource for identifying candidate genes for the estimated 5,000 single-locus diseases[13]. All the sequences and clones described here are publicly available (Table 2). We shall update EST clone identification and map information through the NIH cDNA database. □

Received 27 November 1991; accepted 3 January 1992.

1. Adams, M D. *et al. Science* **252,** 1651–1656 (1991).
2. Ellisen, L. *et al. Cell* **66,** 649–661 (1991).
3. Carthew, R. & Rubin, G. *Cell* **63,** 561–577 (1990).
4. Rao, Y., Jan, L. & Jan, Y. *Nature* **345,** 163–167 (1990).
5. Woods, D. & Bryant, P. *Cell* **66,** 451–464 (1991).
6. Finkelstein, R., Smouse, D. Capaci, T., Spradling, A. & Perrimon, N. *Genes Dev.* **4,** 1516–1527 (1990).
7. Uberbacher, E. & Mural, R. *Proc. natn. Acad. Sci. U.S.A.* **88,** 11261–11265 (1991).
8. Kamholtz, J., de Ferra, F., Puckett, C. & Lazzarini, R. *Proc. natn. Acad. Sci. U.S.A.* **83,** 4962–4966 (1986).
9. Galau, G., Klein, W., Britten, R. & Davidson, E. *Archs Biochem. Biophys.* **197,** 584–599 (1977).
10. Casero, R. *et al. J. biol. Chem.* **266,** 810–814 (1991).
11. Young, M. *et al. Genomics* **7,** 491–502 (1990).
12. Höög, C. *Nucleic Acids Res.* **19,** 6123–6127 (1991).
13. McKusick, V. *FASEB J.* **5,** 12–20 (1991).
14. Rubenstein, J. *et al. Nucleic Acids Res.* **18,** 4833–4842.
15. Schmid, C. W. & Jalinek, W. R. *Science* **216,** 1065–1070 (1982).
16. Paulson, K. E. *et al. Nature* **316,** 359–361 (1985).
17. Fanning, T. G. & Singer, M. F. *Biochim. biophys. Acta* **910,** 203–212 (1987).
18. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. *J. molec. Biol.* **215,** 403–410 (1990).
19. Barker, W., George, D., Hunt, L. & Garavelli, J. *Nucleic Acids Res.* **19** (Suppl), 2231–2236 (1991).
20. Pearson, P. *Nucleic Acids Res.* **19** (Suppl), 2237–2239 (1991).

ACKNOWLEDGEMENTS. We thank J. Powell and J. Kelley of the Division of Computer Research and Technology at NIH for computer systems support and D. Lipman of the National Center for Biotechnology Information at NIH for access to the network BLAST server and nonredundant peptide database.