J. Biochem. 122, 129-147 (1997)

A Survey of Genes Expressed in Mouse Embryonal Carcinoma F9 Cells: Characterization of Expressed Sequence Tags Matching No Known Genes[1]

Midori Nomura,* Seiji Nishiguchi,* Md. Abdul Motaleb,* Yoshihiro Takihara,* Tatsuya Takagi,† Teruo Yasunaga,† and Kazunori Shimada*,[2]

*Department of Medical Genetics, Division of Molecular Biomedicine, Research Institute for Microbial Diseases, Osaka University, 3-1 Yamadaoka, Suita, Osaka 565; and †Genome Information Research Center, Osaka University, 3-1 Yamadaoka, Suita, Osaka 565

Received for publication, February 17, 1997

We prepared 2,132 expressed sequence tags (ESTs) from undifferentiated mouse embryonal carcinoma F9 cells and found that 1,416 match known gene and/or protein sequences [Nishiguchi *et al.* (1996) *J. Biochem.* 119, 749-767]. To obtain information on the functions of the remaining 716 unidentified ESTs and to develop a system for characterizing ESTs matching no known genes, we analyzed their sequences by (i) repeated database searches, using the BLASTN, BLASTX, TBLASTX, and FASTA programs, (ii) using computer programs developed or modified for this work, such as the WFASTA, ORFTRNS, and MFASTA programs, together with the DBPROSITE and GRAIL programs, and (iii) examining the expression patterns of the corresponding mRNAs in F9 cells and several organs of adult mice, using the digoxigenin-labeled dot-blot method. We found that 216 of the 716 ESTs match known gene and/or protein sequences, and 307 show significant similarities to these sequences, with a Poisson p-value <0.01. The strategy and usefulness of such analysis for characterizing unidentified ESTs are discussed.

Key words: digoxigenin, embryonal carcinoma cells, expressed sequence tags, hierarchical cluster analysis, unidentified EST.

Mouse embryonal carcinoma F9 cells differentiate into endoderm-like cells resembling those of the mouse blastocyst in response to retinoic acid (RA) (*1*). F9 cells constitute an attractive model system for studying early events in mammalian development, and retinoid signaling (*1*). Since F9 cells and embryonic stem (ES) cells show a close resemblance (*2*), a catalogue of genes expressed in F9 cells should be of significant interest not only to those using this system to study early mammalian development (*3*), but also to those trying to characterize various gene functions using ES cells and gene targeting techniques (*4*).

We initiated a series of studies to classify mRNAs present in F9 cells by examining cDNA libraries prepared on poly(A)$^+$RNAs from F9 cells (*5-9*). We found that 2,132 expressed sequence tags (ESTs) prepared from mRNAs in undifferentiated F9 cells can be classified into the following three groups: the 1st group (=identified ESTs), matching previously documented gene and/or protein sequences, and including 1,416 ESTs; the 2nd group (=novel ESTs), matching no known gene and/or protein sequences in the databases, and including 562 ESTs; and the 3rd group (=dbEST ESTs), matching previously registered, but not characterized EST sequences in the dbEST (=EST database), and including 154 ESTs (*6*). All these results indicated that we were not able to obtain information on the functions of at least 34% (562+154 out of 2,132 ESTs) of the prepared ESTs (*6*). In the present paper, we call the 2nd and 3rd groups collectively "unidentified ESTs," and before increasing the numbers of ESTs, we tried to develop a system for further characterization and classification of these unidentified ESTs. To obtain clues as to their functions and also to screen ESTs possibly involved in the regulation of early mammalian development, we analyzed the nucleotide sequences of the unidentified ESTs, using several computer programs. We also examined the expression patterns of the corresponding mRNAs in undifferentiated F9 cells, RA-treated F9 cells, and several organs of adult mice.

MATERIALS AND METHODS

Computer-Assisted Analyses—i) Database searches: The ESTs showing sequence similarities to the non-redundant

[1] This work was supported by a Grant-in-Aid for Creative Basic Research (06NP0401) and a Grant-in-Aid for Scientific Research (04454170) from the Ministry of Education, Science, Sports and Culture of Japan. A part of this work was also supported by a Grant for Aging from the Ministry of Health and Welfare of Japan. The nucleotide sequence data reported in this paper can be found in the GSDB, DDBJ, EMBL, and NCBI nucleotide sequence databases under the following accession numbers, D21355 to D21795, D28603 to D28744, and D76447 to D77996.

[2] To whom all correspondence should be addressed. Tel: +81-6-879-8324, Fax: +81-6-879-8326, E-mail: shimada@biken.osaka-u.ac.jp

Abbreviations: AP, alkaline phosphatase; dbESTs, EST databases; DIG, digoxigenin; EC, embryonal carcinoma; EGF, epidermal growth factor; ES, embryonic stem; ESTs, expressed sequence tags; GAPDH, glyceraldehyde-3-phosphate dehydrogenase; msh, muscle segment homeobox; RA, retinoic acid; TCL1, T-cell lymphoma 1.

Downloaded from http://jb.oxfordjournals.org/ at Peking University on October 2, 2012

nucleotide and/or protein databases with either BLASTN or FASTA scores higher than 200 were added to the identified ESTs (*5, 6*), and those showing BLASTX or TBLASTX scores higher than 100 were also added to the identified group (*5, 6*). The unidentified ESTs showing sequence similarities with Poisson p-values of less than 0.01 were putatively classified as the $p<0.01$ group. We used the BLASTN, BLASTX, and TBLASTX programs (*10*), and the FASTA program (version 2.0u4, February, 1996) (*11*) at the Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo. The final searches were made around August 21-31, 1996.

ii) Construction of an F9 EST database: We constructed an F9 EST database consisting of the sequences of 2132 ESTs (*5, 6*) and several other RA-inducible ESTs prepared in our lab (*7-9*) in the Genome Information Research Center, Osaka University, Osaka, and named it dbEST-F9. To construct and up-date the database, we developed and used the following programs: DBADD, to add new EST sequences to the database; DNAdb-update, to create a B-tree index file for the database; DBLST, to list the names of ESTs in the database; and DBF9EXT, to extract an entry from the database.

iii) Detection of redundant ESTs: WFASTA program: All the ESTs in the dbEST-F9 were compared with one another to detect sequence redundancy using the WFASTA program. The WFASTA program executes the FASTA program simultaneously for both plus and minus strands of a nucleotide sequence.

iv) Translation into amino acid sequences: ORFTRNS program: The nucleotide sequences of the unidentified ESTs were translated into all possible reading frames using the ORFTRNS program at the Genome Information Research Center. This program translates arbitrary numbers of ESTs, usually around 40 to 50 ESTs, with one manipulation. Since the accuracy of sequencing in our experiment was around 95% (*5*), the ORFTRNS program translates codons containing N base(s) into X, and translates all stop codons into Z.

v) Searches for similarities to protein sequences: MFASTA program: All the deduced amino acid sequences were searched for matches in the non-redundant protein sequence database (pir49+sp34), after introducing appropriate gaps, with the MFASTA program, which is a user interface for the FASTA program (version 1.7) and enables one to make searches of around 40 to 50 query sequences with one manipulation. The examined ESTs were classified as described under "*i) Database searches.*"

vi) Detection of motif patterns: DBPROSITE program: The predicted amino acid sequences were used for a motif pattern search. Motif patterns were searched for all possible amino acid sequences of the novel ESTs at the Genome Information Research Center, using the prosite database (release 13.0 November, 1995) (*12*) and the DBPROSITE program, which executes the prosite program (Hartmuth, K. and Zorn, M.D., 1992, unpublished) for multiple query sequences with one manipulation.

vii) Prediction of the coding-regions: Grail program: Each EST was examined for the presence of coding-region(s). To predict which ESTs are likely to contain a protein-coding sequence and to exhibit the proper frame and conceptual translation of the peptide sequence (*13*), we used the Coding Recognition Module (CRM) of the Gene Recognition and Analysis Internet Link (GRAIL) at the Oak Ridge National Laboratory, Tennessee, USA.

TABLE I. **Summary of database searches.**

ESTs	Numbers of ESTs (%)
Identified ESTs	216 (30)
Unidentified ESTs	500 (70)
$p<0.01$ group	307 (43)
$p>0.01$ group	109 (15)
dbEST group	84 (12)
Total ESTs examined	716 (100)

Non-redundant nucleotide and/or protein sequence databases were searched using the BLASTN, BLASTX, FASTA, and TBLASTX programs, in that order. ESTs matching known gene and/or protein sequences with one of the following scores were classified into the identified group: BLASTN scores, >200; BLASTX scores, >100; FASTA opt scores, >200; or TBLASTX scores, >100 (*5, 6*). Unidentified ESTs were further classified into the following three groups: the $p<0.01$ group, ESTs showing sequence similarities to known gene and/or protein sequences with a Poisson p-value of less than 0.01; the $p>0.01$ group, ESTs showing no significant sequence similarities to known gene and/or protein sequences; and the dbEST group, a part of the $p>0.01$ group, ESTs showing similarities to sequences of dbEST with BLASTN scores of >200.

Software tools developed specifically for this work at the Genome Information Research Center, Osaka University, upon request, are available to academic researchers.

Expression pattern analyses: Preparation of digoxigenin (DIG)-labeled cDNA probes and dot-blot analysis of RNA levels: cDNA probes of ESTs labeled with DIG were prepared using AmpliTaq DNA polymerase (Perkin Elmer), and primers Z1 (5′-GAAACAGCTATGACCATG-3′) and Z2 (5′-GTAAAACGACGGCCAGTG-3′), as described in the manuals provided with Boehringer Mannheim's DIG-labeling kits (Boehringer Mannheim, GmbH, Mannheim, Germany). The amounts and sizes of the synthesized cDNAs were determined using an anti-DIG antibody after spotting aliquots of the cDNAs and the DIG-labeled DNA standard onto a nylon membrane. The sizes of the PCR products ranged between 1 and 3 kb. About 0.2 μg aliquots of the prepared probes were used for each strip of a filter.

For dot-blot analysis, 1 μg aliquots of poly(A)$^+$RNAs were spotted onto nylon membranes (Amersham), hybridized with DIG-labeled probes, and, after binding of anti-DIG-AP (alkaline phosphatase) Fab fragments, detected with Lumi-Phos 530 (Wako). The intensities of the hybridization signals were measured with an ATTO Densitograph Lumino-CCD and a quantification program called ATTO Densitograph Ver3.01. Since the specific activity of each probe was different, the intensities of the signals were compared only within each EST, between undifferentiated F9 cells and F9 cells treated with RA for 72 h, and between undifferentiated F9 cells and each of the following five organs of adult mice; brain, liver, kidney, heart, and testes. The data were examined by means of hierarchical cluster analysis together with standardized Euclidean distances and the UPGMA method (*14*), and were used to classify ESTs.

Other procedures, such as RNA preparation, RNA analysis, construction of cDNA libraries, isolation of cDNA clones, and single pass cDNA sequencing were described elsewhere (*5, 6*). As cloning vectors, λZAPII was used for 97 ESTs, numbers 00D08 to 62H03, and λuni-ZAP for the

Downloaded from http://jb.oxfordjournals.org/ at Peking University on October 2, 2012

TABLE IIA. **Newly identified ESTs.**

EST[1]	DB:Accession #[2]	Species[3]	Putative Identification	Score	P-value	Program[4]
Cytoskeletal and Contractile Elements						
C72D10	gb:T30076	h	human EST112245 similar to actin	412	8.0e-57	Bn
C83C07	gb:HSARCP5	h	archain, involv. in cell. architectur	202	0.0002	Fa
C95C03	gpu:HSU32944_1	h	cytoplasmic dynein light chain 1	416	2.7e-52	Bx
C90E09	SW:RSP3_CHLRE	cr	RADIAL SPOKE PROTEIN 3, REGULATION	211	6.8e-61	tBx
C68H09	gb:M95787	h	SMOOTH MUSCLE PROTEIN 22 kDa	102	0.00033	tBx
Extracellular Matrix						
C68A06	gbu:D86425	h	osteonidogen, bone matrix protein	182	6.1e-24	tBx
C77F04	gb:U05823	m	pericentrin, microtubule organization	322	5.5e-52	tBx
C97A04	gp:HSTITIN2B_1	h	titin (connectin)	104	9.0e-08	Bx
Energy Metabolism						
C82E08	gbu:N67639	h	CITRATE SYNTHASE	320	5.9e-17	Bn
C66A01	gb:CGU12420	ha	mitocho. benzodiazepine receptor (MBR	243	0.00025	Fa
C74F04	gb:X16560	h	CYTOCHROME C OXIDASE, COX VIIc	182	1.1e-40	tBx
C67D04	gb:RATCYPD45	r	cytochrome P-450d gene	214	9.9e-08	Bn
CA9G09	gb:H10448	h	NADH-CYTOCHROME B5 REDUCTASE	367	3.3e-35	Bn
C46C04	PIR:S16967	bo	NADH dehydrogenase (ubiquinone) (EC 1	237	1.7e-26	tBx
C69C04	sp:PEL1_YEAST	y	PHOSPHATIDYLSERINE SYNTHASE, (EC 2.7.	149	1.7e-13	Bx
C78B06	gb:T31998	h	human EST42243, similar to LDH D	851	1.5e-65	Bn
C67C12	gbu:AA087137	m	similar, SW:UCRX_BOVIN UBIQ-CYTO C RE	151	5.4e-61	tBx
Hormone and Hormonal Regulation						
C85A11	gbu:MUSCIP21R	m	26S proteasome ATPase (CIP21)	1522	8.0e-126	Bn
C83F03	gb:MMU21103	m	mammary gland factor	212	0.00035	Fa
C84A11	gb:HUMIRS	h	RS-1, insulin receptor subunit	225	1.3e-08	Bn
C77B06	gb:HSU37146	h	silencing mediator of retinoid & thy	793	3.4e-58	Bn
Signal Transduction and Cell Regulation						
(Signal transduciton)						
C75A09	gb:MUSA4P	m	alpha 4 prot. Ig receptor-med. sig tr	411	1.5e-16	Fa
C68F01	gb:MUSA4P	m	alpha 4 prot. Ig receptor-med. sig tr	353	1.7e-72	Bn
CA6G02	gbu:HSU50078	h	guanine nucleotide exchange fact p619	114	3.1e-07	tBx
C67D05	gb:HSRHO1	h	mRNA for rho GDP-dissociation	417	6.3e-16	Fa
(Kinases and phosphatases)						
C66H10	gb:MMU35249	m	CDK-activating kinase assembly	458	1.1e-32	Bn
C93G07	PIR:JX0342	r	choline kinase (EC 2.7.1.32) R2	107	1.3e-19	tBx
CA9E10	gbu:RNMKP3	r	dual specificity phosphatase, MKP-3	912	3.3e-68	Bn
C75E02	SW:MLK1_HUMAN	h	MIXED LINEAGE KINASE 1	211	1.6e-41	Bn
C73F04	sp:YBF6_YEAST	y	51.4 KD PHOSPHATASE 2C IN SHP1	136	6.7e-16	Bx
C86G05	pir:A55346	r	phosphoprotein phosphatase (EC 3.1.	124	4.4e-09	Bx
C40D04	gb:HSPTPAA	h	phosphotyrosyl phosphatase activator	244	5.0e-07	Fa
C73H12	gb:HUMPP2A	h	protein phosphatase 2A B56-alpha	1405	3.5e-109	Bn
C85B02	gb:R74715	m	pyruvate; orthophosphate dikinase	1311	6.8e-136	Bn
C86B01	gb:HSSDS22MR	h	regulator of protein phosphatase-1	500	2.4e-90	Bn
(Cell growth and/or developmental regulation)						
CB5F04	gp:DMU31961_14	ff	bithorax complex (BX-C), complete seq	104	3.0e-16	Bx
C75D04	SW:COP9_ARATH	at	COP9 PROTEIN, SUPPRESS SEEDLING	246	4.0e-38	tBx
CB7D07	sp:EYA_DROME	ff	DEVELOPMENTAL PROTEIN EYES ABSENT	178	2.7e-21	Bx
C86A12	sp:DSXM_DROME	ff	DOUBLESEX PROTEIN, MALE-SPECIFIC	246	4.0e-38	tBx
C30B08	gbu:MMU50206	m	G-CSF induced gene	605	9.6e-27	Fa
C74D10	gb:RRU05341	r	homolog of yeast cell div. cycle prot	438	1.2e-14	Fa
C90C07	gb:MUSTGPO	m	jumonji protein.	1050	8.8e-92	Bn
C89B06	PIR:S48828	ff	lethal-3 protein, male-spec.	347	1.0e-41	tBx
C56G09	gb:MMP382G4	m	p38-2G4, varies with the cell cycle	1550	3.0e-13	Bn
CB9E05	sp:RAE1_SCHPO	y	POLYA+ RNA EXPORT PROT., REQ. CELL GR	129	1.9e-24	Bx

Downloaded from http://jb.oxfordjournals.org/ at Peking University on October 2, 2012

TABLE IIA. **Newly identified ESTs** (continued).

EST	DB:Accession #	Species	Putative Identification	Score	P-value	Program
(Oncogenes, tumor suppressors and tumor-related)						
C87B10	sp:YA65_CHICK	ch	65 KD YES-ASSOCIATED PROTEIN (YAP65	123	1.3e-09	Bx
C59G07	gb:HSHGLHOMO	h	dm giant larvae tumor suppressor homo	223	7.0e-05	Fa
C78C09	gbu:MMU42383	m	FGF inducible gene 13 (FIN13)	225	1.5e-35	tBx
CC0A03	gb:J05205	m	junD proto-oncogene	217	3.8e-09	Bn
C89C09	gb:MMU33626	m	homolog of leukemia-associated PML	279	3.9e-15	Bn
CA7H03	gb:HSTCL1	h	mRNA for T cell leukemia/lymphoma 1	119	3.6e-19	tBx
C83A02	gb:HSTCL1	h	mRNA for T cell leukemia/lymphoma 1	225	6.4e-06	Fa
C68F08	gb:HSTCL1	h	mRNA for T cell leukemia/lymphoma 1	356	2.0e-13	Fa
C73E12	gb:HSTCL1	h	mRNA for T cell leukemia/lymphoma 1	244	1.1e-10	Bn
C73F03	gb:HSTCL1	h	mRNA for T cell leukemia/lymphoma 1	253	4.9e-11	Bn
C82E01	gb:MUSNF2D	m	neurofibromatosis 2 gene	538	7.8e-72	Bn
C74C05	gbu:HSU46751	h	phosphotyrosine independent ligand p6	938	4.1e-107	Bn
CA4F02	gbu:HSU46751	h	phosphotyrosine independent ligand p6	522	1.1e-35	Bn
C67E04	gbu:HSU46752	h	phosphotyrosine independent lig. p62	978	7.4e-93	Bn
CB2G02	gbu:HSU46752	h	phosphotyrosine independent lig. p62	935	4.3e-70	Bn
C72C05	gb:H45496	h	PROTO-ONCOGENE TYR-PROTEIN KINASE FYN	484	1.8e-60	Bn
C73G03	gb:T32080	h	retinoblastoma-binding protein 1	253	3.0e-11	Bn
C04F06	SW:RSP1_MOUSE	m	RSP-1 PROTEIN, SUPPRESSING V-RAS TR	107	7.1e-06	tBx
C93D12	gb:T08049	h	similar, EGF repeat family	265	2.0e-14	Bn
C81C06	gbu:HSU25801	h	Tax1 binding protein	429	2.5e-34	Bn
CA6A11	SW:TGFB_HUMAN	h	TRANSFORMING GROWTH FACTOR BETA-1 B	384	1.1e-46	tBx
C81G04	gb:H50443	h	TRANSFORMING PROTEIN RHOC	348	4.1e-36	Bn
C83B07	gbu:HSTRKT3ON	h	TRK-T3 oncogene	423	2.0e-27	Bn
C84F01	gbu:MMU52945	m	tumor susceptibility protein TSG101	149	3.4e-12	tBx
Transcription and Translation						
(Transcription factors)						
C90A05	gb:HSU09825	h	acid finger protein (ZNF173)	226	1.7e-22	Bn
C73F11	gbu:RNU49057	r	CTD-binding SR-like protein rA9	339	1.9e-38	tBx
CA2H05	gb:HSU39360	h	DNA-binding prot. (CROC-1A), c-fos ac	283	3.1e-40	Bn
C69F10	gp:HSU39361_1	h	DNA-binding prot. (CROC-1B), c-fos ac	154	2.0e-14	Bx
C74H06	gb:HSU15306	h	DNA-binding prot. NFX1, a repressor	483	6.4e-31	Bn
C90G06	gb:HSU69127	h	FUSE binding protein 3, Transactivatio	334	9.7e-38	tBx
C82H12	gb:MUSHRS	m	Hrs, 115-kd prot. with zinc finger do	1168	6.5e-107	Bn
CB8H06	gb:MMU19106	m	imprinted zinc-finger gene (znf127)	231	6.4e-47	Bn
C28A09	gb:MMU25096	m	Kruppel-like factor LKLF	253	1.3e-32	Bn
C92D12	gbu:HSNC2ALPH	h	NC2 alpha subunit	1115	7.3e-92	Bn
C94A02	gb:SCU02598	y	nucleolar protein NOP4, pre-rRNA proc	232	0.00084	Fa
C69D11	gb:X64002_cds1	h	RAP74, initiation fact. RNA polyme. II	335	7.3e-40	tBx
C98B06	gb:RATRPIT	r	RNA polymerase II ts factor SIII p18	839	1.6e-64	Bn
C91C05	gbu:AA023950	m	similar, WP:C16C10.7, ZINC FINGER PRO	294	2.3e-34	tBx
C69G12	gb:HSU15641	h	transcription factor E2F-4	312	3.3e-18	Bn
C83B04	gbu:N43821	h	translation initiation factor eIF-4	234	2.3e-10	Bn
C73C04	gb:HSU17969	h	translation initiation factor eIF-5A	553	7.5e-40	Bn
C94E10	gbu:MMU63323	m	translation initiation factor	513	8.4e-35	Bn
CA6D06	gbu:MMU63323	m	translation initiation factor	507	8.3e-62	tBx
CB2E02	gb:HSU38864	h	zinc finger protein C2H2-150	339	1.5e-18	Bn
C67B10	gb:MDZFP30	m	zinc finger protein 30	267	3.6e-08	Fa
(Transcription and translation machinery)						
C88B04	gpu:HSU28042_1	h	DEAD box RNA helicase-like protein	120	7.9e-20	Bx
C93F04	gb:R36350	h	hetero. RN NUCL. RIBONUCLEOPROTEIN G	823	1.7e-69	Bn
C88H08	gb:R62689	h	hetero. RN NUCL. RIBONUCLEOPROTEIN G	996	1.4e-84	Bn
C83H07	pir:S57489	h	HPBRII-4, ribonucleoprotein superfami	200	4.6e-22	Bx
C89A06	gb:MUSDHM1P	m	homolog of yeast dhp1+ gene, RNA met	1267	1.1e-97	Bn
C67E02	gb:HSU28686	h	putative RNA binding protein RN	229	2.2e-08	Fa
C69D01	gb:HSU28686	h	putative RNA binding protein RN	544	1.2e-25	Fa
C69G08	gb:HUMHRH1	h	putative RNA helicase H	485	2.3e-24	Fa
C40C05	gbu:H97533	h	RE-MRNA SPLICING FAC. RNA HELICASE	863	1.3e-66	Bn

Downloaded from http://jb.oxfordjournals.org/ at Peking University on October 2, 2012

TABLE IIA. **Newly identified ESTs** (continued).

EST	DB:Accession #	Species	Putative Identification	Score	P-value	Program
CB9H02	gb:W95618	h	similar, PIR:A39564 trscrp repres pro	191	5.1e-40	tBx
C73H06	gb:N42449	h	similar, WP:R74.1, LEUCYL-TRNA SYNTHE	282	1.4e-46	tBx
C89D01	gb:T08178	h	similar to p68 RNA helicase	482	3.2e-59	Bn
C83H10	gb:W98590	m	similar to PIR:S43484 hnRNP F protein	229	4.0e-25	tBx
C74G04	gb:R75071	m	similar to splicing factor	606	7.5e-70	Bn
(Ribosomal proteins)						
C93E03	SW:RL15_MYCCA	r	50S RIBOSOMAL PROTEIN L15	298	5.9e-35	tBx
C09A01	SW:RL20_ECOLI	ec	50S RIBOSOMAL PROTEIN L20	253	1.1e-55	tBx
C99F11	SW:RL20_ECOLI	ec	50S RIBOSOMAL PROTEIN L20	121	2.3e-08	tBx
CA5E09	gb:H35836	r	Ribosomal protein L7	407	1.9e-26	Bn
CB4A02	gb:AA033298	m	similar, SW:RL2_BACST 50S RibPROT. L2	598	2.6e-76	tBx
(Heat shock proteins)						
C92C08	gb:RNHSRP	r	heat shock related protein	550	6.0e-52	Bn
C86F07	gb:HUMHSP40	h	heat shock protein 40	118	5.8e-06	tBx
Membrane-Associated						
(Receptor and membrane-associated)						
C85A03	gb:HSERGICA	h	ERGIC-53, a mem. prot. of the ER-Golg	235	0.0011	Fa
C78F03	gb:R81533	h	ERYTHROCYTE BAND 7 INTEGRAL MEMBRANE	238	1.6e-39	Bn
C74A08	gb:HSU37673	h	neuron-specific vesicle coat protein	333	8.8e-34	Bn
CA6D05	gbu:MMU41805	m	putative T1/ST2 receptor binding prot	1432	1.0e-111	Bn
C81E07	gb:MUSMARIB	m	ribophorin, RER specific membrane pro	375	2.8e-14	Fa
C82A03	PIR:S37395	r	secret. carrier membrane protein 37	103	1.6e-29	tBx
CB4B11	gb:H17043	h	similar, SP:B48580 ANTIGEN EG13	229	1.0e-23	tBx
CC0B04	gb:HSU36764	h	TGF-beta receptor interacting protein	831	1.7e-61	Bn
(Transporters)						
CB1A07	gbu:MMCMPSATR	m	CMP-sialic acid transporter	555	7.0e-71	tBx
C95H03	pir:S59302	y	Nip1 prot., required for nuclear tran	129	3.2e-13	Bx
C40B06	gbu:RNU43175	r	vacuolar ATPase subunit F	637	1.6e-69	Bn
C73B09	gbu:RNU43175	r	vacuolar ATPase subunit F	1552	1.5e-129	Bn
C30A09	gb:U19521	m	vesicle transport protein	287	2.1e-33	tBx
C60H03	emb:MM30838	m	voltage dependent anion channel	417	9.1e-43	Bn
Other-secreted Proteins						
CA7E06	gb:SMU30265	sm	cyclophylin-like protein (TNF-alpha)	248	3.4e-06	Fa
C72B02	gb:H32230	r	Interferon gamma	301	1.2e-17	Bn
Other Metabolism						
(DNA metabolism)						
C93B06	gb:U01147	h	BREAKPOINT CLUSTER REGION PROTEIN	137	1.0e-20	tBx
CA5D09	gb:U01147	h	BREAKPOINT CLUSTER REGION PROTEIN	170	1.6e-16	tBx
CA6D03	gb:HSDNALIG3	h	DNA ligase III.	377	1.0e-28	Bn
C82A08	gb:RRFE65	r	FE65 mRNA for an integrase-like prote	216	1.4e-06	Fa
C90G07	gb:MUSFEN1X	m	flap endonuclease-1 (FEN-1)	214	0.001	Fa
C85E02	gbu:HSU51166	h	G/T mismatch-spec. thymine DNA glycos	114	3.0e-14	tBx
C70A11	gbu:MMU42190	m	G/T-mismatch binding protein	816	3.2e-87	Bn
CB4F08	gbu:HSPGAMMA	h	pur alpha, single-strand-spec. bind. p	610	6.2e-43	Bn
C89A11	gb:T33139	h	rec. & DNA-damage resistance protein	541	7.8e-38	Bn
C77C09	gb:AA009145	m	similar, gb:D26090 CDC46 homolog	435	3.0e-66	tBx
(Repetitive DNA and virus-related sequences)						
C93F09	gb:HSU177E8	h	betw. markers DXS366/DXS87 on chro. X	173	2.2e-15	tBx
C90H11	gb:HSU5OF11	h	betw. markers DXS366/DXS87 on chro. X	361	1.9e-41	tBx
CA5H02	gb:SCYJL050W	y	chromo. X reading frame ORF	279	2.5e-15	Bn
C87E01	gb:HS40C2F	h	CpG DNA, forward read cpg40c2.ft1k	464	4.0e-30	Bn
C81D03	gb:HS53E4F	h	CpG DNA, clone 53e4, forward	229	0.00067	Fa
C39C06	gbu:HSU74C11	h	DNA seq., on chromo. X contains EST	276	6.2e-22	Bn

Downloaded from http://jb.oxfordjournals.org/ at Peking University on October 2, 2012

TABLE IIA. **Newly identified ESTs** (continued).

EST	DB:Accession #	Species	Putative Identification	Score	P-value	Program
CA6D10	gb:MUSENDPRO	m	endogenous provirus gag, pol	266	3.6e-08	Fa
CB8H04	gbu:W37596	h	similar, Alu repetitive element	282	5.8e-18	Bn
C78B03	gbu:W16740	h	similar, LTR6 repetitive element	345	1.2e-40	tBx
C96H02	gb:R01333	h	similar, MER22 repetitive element	445	1.8e-29	Bn
C90B02	gbu:W00604	h	similar, MER22 repetitive element	259	1.6e-29	tBx
C72D01	gb:T85206	h	similar, PTR5 repetitive element	272	1.9e-45	Bn
CA9F09	gb:DM65G7T	ff	STS determined from EMP cosmid	139	1.8e-12	tBx
C02G04	gb:DMU35403	ff	telomere 2L	112	1.9e-05	tBx
(Protease and protease inhibitors)						
C91B08	gb:H35529	r	ATP-dependent CLP protease, proteolyt	232	3.1e-05	Fa
C97A07	gb:R88213	h	cathepsin B-like protease	522	2.6e-36	Bn
C75G02	gb:HSYUBG1	h	ubiquitin (3 repeats)	207	1.7e-05	Fa
C88D03	gb:T99403	h	UBIQUITIN-CONJUGATING ENZYME E2-17 KD	268	8.1e-29	Bn
(Other metabolism)						
C85F01	gbu:HSU43573	h	alpha-N-acetylglucosaminidase (NAGLU)	659	3.0e-76	Bn
C38F11	gbu:MMU45978	m	calcium-binding prot. Cab45b, in Golg	1388	7.8e-108	Bn
CA5F01	gbu:MMU45978	m	calcium-binding prot. Cab45b, in Golg	326	6.9e-37	tBx
C88G08	gb:RATCFTRAAA	r	CFTR promoter region	274	3.8e-06	Fa
C81D06	gb:RATAKGE2	r	dihydrolipoamide succinyltransferase	200	3.6e-07	Bn
CA2F06	gb:HSU11861	h	edg-2, homo. Xenopus maternal trs G10	402	6.2e-56	Bn
C70H10	SW:SODE_ONCVO	ov	EXTRACELLULAR SUPEROXIDE DISMUTASE	229	4.4e-62	tBx
C68E12	gb:HUMFKBPA	h	FK-506 binding prot. homolog (FKBP38)	405	1.7e-53	Bn
C83G02	gb:HUMFKBPA	h	FK-506 binding prot. homolog (FKBP38)	320	2.6e-21	Bn
C88E12	gb:HSU18543	h	fragile X mental retard. protein 1	757	5.4e-101	Bn
C71A05	gb:H30434	h	GLUTATHIONE S-TRANSFERASE 2	318	4.6e-42	Bn
C57B05	PIR:S19884	h	highly charged protein	232	3.7e-17	Bn
C42E06	gb:S76337	dp	IgE-binding proteins (clone WM)	222	9.4e-11	Bn
C78G02	gb:RNMFHGENE	r	mammalian fusca gene homolog; mfh gen	611	1.0e-29	Fa
C77H04	gb:RATPPT	r	palmitoyl-protein thioesterase	377	1.9e-13	Fa
CA7D11	gb:T52623	h	processing a-glucosidase	328	3.1e-20	Bn
C83F07	SW:ACR1_YEAST	y	REGULATOR OF ACETYL-COA SYNTHETASE	121	8.1e-08	tBx
C71G09	gp:RICRAB24P_1	ri	RAB24 prot., thiol-spec. antioxidant	209	2.7e-22	Bx
C90A04	gbu:MMU47323	m	stromal cell prot. interact pre-B cel	258	3.1e-50	tBx
C93E09	gb:W10932	m	similar, WP:D2013.7 M0V-34 PROTEIN	356	6.0e-85	tBx
C66C08	gb:W33755	m	similar, SW:YSA1_YEAST YSA1 PROTEIN	297	5.5e-75	tBx
C96C05	gb:N53302	h	similar to WP:C16C10.1 CARRIER PROTEI	121	4.5e-21	tBx
CB8F07	gp:HSU21914_1	h	spinal muscular atrophy (SMA) critica	101	4.8e-05	Bx
(Not classified)						
CA0G10	gp:CELF10G7_10	ce	cosmid F10G7.1 gene product	126	5.0e-10	Bx
CA2G11	gb:T31738	ce	cosmid R08D7.3, 64.2Kda protein	182	8.7e-41	tBx
C89D05	gpu:CEZK792_1	ce	cosmid ZK792.1	141	2.6e-12	Bx
CB4F10	gp:CEZK1128_1	ce	cosmid ZK1128.1	105	6.5e-13	Bx
CA1C11	gp:CEC16C10_10	ce	C16C10.4	154	2.0e-28	Bx
CA4F05	gb:L35933	m	erythrocyte protein 4.2	210	1.0e-22	tBx
C81H04	gp:CELF57B9_8	ce	F57B9.2 gene product	104	3.7e-07	Bx
C85D12	gpu:CEF49C12_13	ce	F49C12.13	216	1.3e-24	Bx
CB5C02	SW:YHS2_YEAST	y	HYPOTHETICAL 25.7 KD PROT IN MSH1-E	411	1.6e-50	tBx
C87F01	SW:YCAT_YEAST	y	HYPOTHETICAL 27.3 KD PROT IN CAT5 5	199	1.7e-42	tBx
CA2A12	SW:YB92_YEAST	y	HYPOTHETICAL 27.6 KD PROT IN PRP5-A	119	2.1e-08	tBx
C86E01	SW:YBD6_YEAST	y	HYPOTHETICAL 29.1 KD PROT IN U.+2	160	3.8e-51	tBx
C87A03	SW:YEY6_YEAST	y	HYPOTHETICAL 38.2 KD PROT IN BEM2-S	156	1.5e-32	tBx
C68C12	SW:YB84_YEAST	y	HYPOTHETICAL 42.5 KD PROT IN PDB1-A	204	4.5e-41	tBx
C87F08	SW:YB84_YEAST	y	HYPOTHETICAL 42.5 KD PROT IN PDB1-A	471	3.7e-59	tBx
CA0D11	SW:YABA_SCHPO	y	HYPOTHETICAL 44.4 KD PROT C2G11.10C	150	1.8e-13	tBx
C73H04	pir:S54549	y	hypothetical protein YM9796.02	211	5.6e-30	Bx
C29A09	PIR:S49633	y	hypothetical protein 4 (RAD10 3' regi	400	2.8e-49	tBx
CB1G08	PIR:S52698	y	hypothetical protein YD9346.02c	281	2.7e-48	tBx

Downloaded from http://jb.oxfordjournals.org/ at Peking University on October 2, 2012

TABLE IIA. **Newly identified ESTs** (continued).

EST	DB:Accession #	Species	Putative Identification	Score	P-value	Program
CA1A10	PIR:S39974	se	hypothetical protein	371	3.1e-45	tBx
C55D07	gbu:HUMKIAA22	h	KIAA0182 protein	336	1.3e-26	Bn
CB2C09	gp:CEU19615_1	ce	LET 858	182	3.0e-18	Bx
CB7E01	gp:HUMORFKG1N_1	h	ORF	356	2.2e-44	Bx
C73C12	gp:HUMORFT_1	h	ORF	357	5.1e-43	Bx
CA0B07	gp:HUMRSC765_1	h	ORF	142	1.0e-28	Bx
CB6C11	gp:HUMRSC786_1	h	ORF, from male myeloblast KG-1 cell	118	5.7e-12	Bx
C77D10	SW:HG74_HUMAN	h	OVARIAN GRANULOSA CELL 13.0 KD PROT	190	8.0e-20	tBx
C87E04	SW:HG74_HUMAN	h	OVARIAN GRANULOSA CELL 13.0 KD PROT	224	1.2e-41	tBx
C92D11	SW:HG74_HUMAN	h	OVARIAN GRANULOSA CELL 13.0 KD PROT	113	7.2e-15	tBx
C90G09	SW:HG74_HUMAN	h	OVARIAN GRANULOSA CELL 13.0 KD PROT	178	6.7e-32	tBx
CA9G08	gbu:W09101	m	similar, WP:C09F5.2, CE01774	136	2.7e-12	tBx
CA1E09	gbu:W70754	m	similar, WP:F45E12.4, CE02740	157	3.3e-15	tBx
C73G01	gbu:W80046	m	similar, WP:ZK945.3, PUMILIO-REPEAT L	283	8.0e-50	tBx
C95A02	PIR:S42069	r	TEGT protein	172	4.2e-16	tBx
C74D06	gb:HSTEGT	h	TEGT mRNA, abundant in adult testis	213	0.0001	Fa
C92A07	PIR:S53031	y	YM8270.04 protein	334	1.8e-73	tBx

[1]In Tables II, IV, V, and VIII, the name of each EST has the prefix C. As cloning vectors, λZAPII was used for clones C00D08 to C62H03, and λuni-ZAP for clones C66A01 to CC1B12 (*5, 6*). [2]DB, databases; emb, EMBL database; gb, Genbank; gbu, Genbank-updated; gp, Genbank nucleic acid sequences translated into amino acid sequences; gpu, Genbank-updated nucleic acid sequences translated into amino acid sequences; pir or PIR, PIR protein sequence database; SW or sp, Swiss-Prot database; prf or PRF, PRF protein sequence database. When ESTs were determined to show significant similarities to uncharacterized ESTs, hypothetical proteins, and characterized gene and/or protein sequences, we cite only the database and accession number (#) related to the characterized gene or protein sequences. [3]Species: at, *Arabidopsis thaliana*; bo, bovine; ce, *Caenorhabditis elegans*; ch, chicken; cr, *Chlamydomonas reinhardtii*; dp, *Dermatophagoides pteronyssinus*; ec, *Escherichia coli*; ff, fruit fly (*Drosophila melanogaster*); h, human; ha, hamster; m, mouse; ov, *Onchocerca volvulus*; r, rat; ri, rice (*Oryza sativa*); se, *Streptococcus equisimilis*; sm, *Schistosoma mansoni*; y, yeast (fission yeast, *S. cerevisiae* or *Schizosaccharomyces pombe*). [4]Programs: Bn, BLASTN; Bx, BLASTX; tBx, TBLASTX; Fa, FASTA.

remaining 619 ESTs, numbers 66A01 to C1B12 (*5, 6*).

RESULTS AND DISCUSSION

To acquire more information and improve the assignment of the unidentified ESTs to functional categories, we tried to establish a system for characterization of the ESTs matching no known genes. For this we analyzed their sequences using several computer programs, and by examining the expression patterns of the corresponding mRNAs in F9 cells and several organs of adult mice.

Computer-Assisted Analyses—i) Database searches: We found that 716 out of the 2,132 ESTs prepared from undifferentiated F9 cells match no known gene and/or protein sequences (*6*). Since databases are up-dated on a daily basis, we repeated the database searches of all the 716 unidentified ESTs using the BLASTN, BLASTX, FASTA, and TBLASTX programs, in that order, and found that 216 of the 716 ESTs (30%) match known gene and/or protein sequences (Tables I and IIA).

Since sequence similarities identified with the BLAST and FASTA programs were considered to be statistically significant, with a p-value of less than 0.01, the sequences of the remaining 500 unidentified ESTs were further classified on the basis of the results of database searches into the following three groups: the $p<0.01$ group, matching known gene and/or protein sequences, with a p-value of <0.01, and including 307 ESTs (Tables I and IIB); the $p>0.01$ group, matching no known gene and/or protein sequences, with a p-value of <0.01, and including 109 ESTs; and the dbEST group, part of the $p>0.01$ group, matching sequences in the dbEST, and including 84 ESTs (Table I). Statistical significance does not necessarily mean functional similarity. Nevertheless, some of these similarities should lead to the characterization of some ESTs. It is obvious that the ESTs classified as the $p<0.01$ group show marginal similarities to sequences in various organisms other than those in mouse and man (Table IIB).

ii) Redundant ESTs: WFASTA program: When we examined the redundancy of the 716 unidentified ESTs in dbEST-F9 using the WFASTA program, we found that 66 corresponded to 29 groups of redundant ESTs (Table III). There are several clusters of redundant ESTs: one cluster each of 7, 5, and 3 redundant ESTs, 2 clusters of 4 redundant ESTs, and 24 clusters of 2 redundant ESTs, respectively (Table III). All these redundant ESTs showed nucleotide sequence similarities, with FASTA opt scores exceeding 200 (Table III). Since some of the ESTs come from different parts of the same mRNA, the number of redundant ESTs, estimated with the WFASTA program, is obviously the lower limit.

iii) Amino acid sequence similarities: MFASTA program: In the previous protein sequence database search, we used the BLASTX and TBLASTX programs (*6*, Table II), which do not allow the insertion of gaps during homology alignment. To detect further similarities, we translated the sequences of the 500 ESTs matching no known genes into 6 possible reading frames, using the ORFTRNS program, and searched for matches with a composite database of sequences from PIR and Swiss-Prot, using the MFASTA program after introducing appropriate gaps. The results from each frame were examined, and 27 ESTs were found

Downloaded from http://jb.oxfordjournals.org/ at Peking University on October 2, 2012

TABLE IIB. **ESTs classified into the $p < 0.01$ group.**

EST	DB:Accession #	Species	Putative Identification	Score	P-value	Program
Cytoskeletal and Contractile Elements						
C85A09	gb:HAMSHCA	ma	alpha-cardiac myosin hea	169	0.00080	Bn
C67A05	prf:1209265M	bm	chorion protein A3	48	7.0e-07	Bx
C88D10	sp:DREB_CHICK	ch	DREBRINS E1 AND E2	38	2.4e-06	Bx
C03A06	gp:YSCDHC1A_1	y	dynein	48	0.00012	Bx
C69H03	sp:GLT4_WHEAT	wht	GLUTENIN, HIGH MOLECULAR WEIGHT SUBUN	63	7.1e-07	Bx
C72C10	prf:1920252A	h	keratin 2	34	7.3e-05	Bx
C94H04	sp:MAPA_RAT	r	MICROTUBULE-ASSOCIATED PROTEIN 1A (CO	45	1.1e-06	Bx
CA0H07	pir:S40612	slm	myosin-related protein	89	3.7e-05	Bx
C94D07	prf:2121478A	r	selenoprotein W	63	8.9e-07	Bx
C93B12	pir:S20901	rab	titin	52	0.00030	Bx
C92E02	pir:A55721	m	tropoelastin	45	3.9e-05	Bx
CC1B05	sp:UTRO_HUMAN	h	UTROPHIN (DYSTROPHIN-RELATED PROTEIN)	47	0.0032	Bx
Extracellular Matrix						
C88E03	prf:2123304A	jl	chitinase	37	8.1e-05	Bx
C72D02	pir:S57243	m	collagen alpha 1(I) chain precursor	48	8.3e-05	Bx
C59H07	pir:S57243	m	collagen alpha 1(I) chain precursor	51	6.5e-09	Bx
CA4D11	pir:A20855	ch	collagen alpha 1(III) chain	37	0.0021	Bx
C71D06	gp:MMU25652_1	m	collagen type XII alpha-1 precursor	50	2.8e-05	Bx
C30A12	sp:CA1E_HUMAN	h	COLLAGEN ALPHA 1(XV) CHAIN PRECURSOR	34	0.0053	Bx
C86F06	prf:1905344A	h	collagen:SUBUNIT=alpha1:ISOTYPE=XVI	39	0.00034	Bx
C40E05	prf:1006299A	ch	collagen,cartilage specific short	44	1.8e-07	Bx
C85F06	prf:1404272A	h	collagen alpha2(IV)	48	4.6e-05	Bx
C62H03	pir:D41132	hyd	collagen-related protein 4	44	2.0e-08	Bx
C69C02	sp:CA11_CHICK	ch	PROCOLLAGEN ALPHA 1(I) CHAIN PRECURSO	64	8.2e-07	Bx
CA2A05	sp:CA13_MOUSE	m	PROCOLLAGEN ALPHA 1(III) CHAIN PRECUR	43	0.00072	Bx
C39G08	sp:CA14_MOUSE	m	PROCOLLAGEN ALPHA 1(IV) CHAIN PRECURS	46	0.0028	Bx
C77E09	gp:HSU32169_2	h	Pro-a2(XI)	44	1.8e-08	Bx
C69F11	gp:HSCOLL_1	h	Pro-alpha-2 chain	41	7.9e-07	Bx
C74E11	gp:HSCOLL_1	h	Pro-alpha-2 chain	40	2.7e-05	Bx
CA7D06	gp:DAREXTA_1	ca	extensin	31	0.0069	Bx
C57D05	sp:EXTN_TOBAC	tob	EXTENSIN PRECURSOR (CELL WALL HYDROXY	64	0.0088	Bx
C86C05	sp:EXTN_TOBAC	tob	EXTENSIN PRECURSOR (CELL WALL HYDROXY	60	0.0095	Bx
C75H03	pir:S49915	mz	extensin-like protein	61	7.3e-08	Bx
C33C11	prf:1806311A	h	heparan sulfate proteoglycan core pro	41	0.00054	Bx
C34C03	prf:1308156B	y	gene S1	59	0.0060	Bx
C42E03	gp:HSLAMAG_1	h	laminin A chain	38	5.8e-05	Bx
C73A02	prf:2203365A	m	laminin alpha5	34	0.0091	Bx
C90B04	prf:1410326A	ff	laminin B2	41	0.00025	Bx
C68A01	prf:2110381A	m	reelin	40	0.00013	Bx
C97G12	prf:2110381A	m	reelin	41	3.1e-07	Bx
CA2A06	gb:GGTENS	ch	tensin	171	0.00014	Bn
Cell/Organism Defense						
C73F07	pir:S37671	h	bat2 protein	47	0.0016	Bx
C87C11	pir:S37671	h	bat2 protein	53	4.6e-07	Bx
C74C08	pir:S37671	h	bat2 protein - or MHC BAT3	48	1.2e-07	Bx
C78D09	gb:K01144	h	HLA CLASS II ANTIGEN, GAMMA CHAIN PRE	91	0.0024	tBx
C69F02	gp:HSVHMAB61_1	h	Ig heavy chain precursor	39	0.0074	Bx
CA0A06	gp:MUSIGKAE_1	m	ig kappa unproductively rearrang	44	1.5e-05	Bx
C34C02	gp:PARSP51A4_1	pa	immobilization antigen	41	0.0040	Bx
CA5B04	gp:S79427_1	m	immunoglobulin lambda light chain var	44	0.0029	Bx
CB1G06	gp:HSU18288_1	h	MHC class II transactivator CIITA	51	0.00019	Bx
C94B07	pir:B29356	kb	MHC class III histocompatibility anti	45	1.2e-08	Bx
C86A06	pir:B35098	h	MHC class III histocompatibility anti	49	0.0034	Bx
C67C03	pir:B35098	h	MHC class III histocompatibility anti	52	0.0086	Bx
C85D02	gbu:MACMAFAAD	mf	MHC-DRB6 class II gene	86	0.00035	tBx
CA7G07	pir:A55071	m	hydrogen peroxide-inducible protein h	46	1.8e-07	Bx

Downloaded from http://jb.oxfordjournals.org/ at Peking University on October 2, 2012

TABLE IIB. **ESTs classified into the $p < 0.01$ group** (continued).

EST	DB:Accession #	Species	Putative Identification	Score	P-value	Program
C85D05	sp:PROP_MOUSE	m	PROPERDIN (FRAGMENT)	74	0.00010	Bx
Energy Metabolism						
C67F09	sp:LEU3_SOLTU	po	3-ISOPROPYLMALATE DEHYDROGENASE PRECU	41	0.0012	Bx
C88F12	gpu:PAALGYGEN_1	pa	alginate lyase	47	0.00019	Bx
C87E12	prf:2020412A	sa	carbamoyl phosphate synthase III	39	0.0041	Bx
C67E11	gp:PPCAB_1	ph	chlorophyll-a, b-binding protein	31	0.0077	Bx
C72H12	gp:MBU36763_1	mb	fatty acid synthase	41	0.00012	Bx
C87B07	gp:MBU36763_1	mb	fatty acid synthase	42	1.4e-07	Bx
C71C12	prf:1924281B	rc	hydrogenase repressor	34	0.0019	Bx
C78B11	sp:ODO1_HUMAN	h	H2-OXOGLUTARATE DEHYDROGENASE E1 COMP	43	0.00024	Bx
C89D09	pir:S57489	h	HPBRII-4 protein	36	5.7e-06	Bx
71A0C7	gp:RRFILPHR_1	r	lactase-phlorizin hydrolase precursor	36	0.0010	Bx
CA4G12	pir:S34960	co	NADH dehydrogenase (ubiquinone) (EC 1	36	5.5e-05	Bx
C94D08	gp:ARBMTFR1_3	al	ND4 gene product	41	0.0038	Bx
C67H04	gp:BPU34894_1	bp	oxygenase	33	0.0083	Bx
C05G09	sp:PPCK_DROME	ff	PHOSPHOENOLPYRUVATE CARBOXYKINASE (GT	46	0.0010	Bx
C94B06	prf:2122340A	tva	pyruvate/ferredoxin oxidoreductase	39	0.0033	Bx
Hormone and Hormonal Regulation						
C67C09	pir:S41971	fpv	3-beta-hydroxysteroid dehydrogenase/s	85	0.00013	Bx
CA7H02	gp:OCU15025_1	rab	anti-Mullerian hormone receptor precu	36	8.9e-05	Bx
CB9B09	gb:X57025_rna1	h	INSULIN-LIKE GROWTH FACT IA PRECURSOR	157	0.00095	Bn
C99A11	pir:S25113	m	insulin-like growth factor binding pr	49	0.0032	Bx
C75B04	sp:THYG_BOVIN	bo	THYROGLOBULIN PRECURSOR	39	0.00021	Bx
C89A08	gpu:XLU41839_1	xl	XL18	40	0.0025	Bx
Signal Transduction and Cell Regulation						
(Signal transduction)						
C29A03	pir:S52957	en	bimD protein	59	3.6e-05	Bx
C05G04	gpu:OSGPROTBS_1	os	g protein B subunit	54	0.0067	Bx
CB7B09	gpu:BTU35363_1	bo	latent TGF-beta binding protein-2	38	8.2e-05	Bx
C60A03	pir:A55494	h	latent transforming growth factor-bet	59	2.9e-10	Bx
C70F12	pir:A48998	m	nucleolar protein p120	39	7.3e-06	Bx
C89B07	pir:S25714	m	son-of-sevenless-2 protein	34	1.8e-06	Bx
C28C08	pir:A55117	m	tsg24 protein	45	0.0011	Bx
(Kinases and phosphatases)						
C04G11	gb:RNU36482	r	31 kDa putatative ser/thr protein kin	138	3.3e-11	Bn
C84E02	gpu:OCU26360_1	oc	AKAP78	55	1.2e-07	Bx
CB9D10	gpu:RNDENDRIN_1	r	dendrin	38	5.8e-05	Bx
C89H08	pir:A57099	h	DNA-dependent protein kinase	43	0.0045	Bx
C73B12	gpu:HSPRB4M_1	h	Po protein	54	2.8e-06	Bx
CC1A08	pir:A45617	nem	6-phosphofructokinase	39	0.0050	Bx
C62B11	pir:A55148	r	protein-tyrosine-phosphatase (EC 3.1.	37	2.9e-07	Bx
CB9H07	pir:S20898	h	titin	46	5.0e-05	Bx
C74D09	pir:S20901	rab	titin	48	3.1e-05	Bx
CB5G04	prf:2124435A	h	titin	40	0.0074	Bx
(Developmental regulation)						
C68A10	gp:GGBRMPROT_1	ch	BRM protein	41	6.7e-08	Bx
C99D08	sp:CDP_CANFA	do	CCAAT DISPLACEMENT PROTEIN (HOMEOBOX	40	6.0e-06	Bx
C73H07	prf:2001380B	h	cell proliferation-associated antigen	41	6.0e-05	Bx
C47B06	gb:HSU17105	h	cyclin F	161	0.0022	Bn
C92C12	gp:CELDEG1A_1	ce	degenerin	36	4.2e-05	Bx
C72C06	prf:2123391A	m	GT1 gene	41	2.8e-05	Bx
C92B12	pir:A39369	ff	homeotic protein BarH1	34	0.00020	Bx
C90E04	pir:A43556	m	homeotic protein Hox 1.4	46	2.2e-05	Bx
C70C09	sp:HXA5_HUMAN	h	HOMEOBOX PROTEIN HOX-A5 (HOX-1C)	39	0.0010	Bx
C00G12	gb:MMU25096	m	Kruppel-like factor LKLF	163	0.0018	Bn

Downloaded from http://jb.oxfordjournals.org/ at Peking University on October 2, 2012

TABLE IIB. **ESTs classified into the $p<0.01$ group** (continued).

EST	DB:Accession #	Species	Putative Identification	Score	P-value	Program
C89B01	prf:1908302A	m	notch-1 gene	42	2.1e-05	Bx
C85H11	gp:DVDNAPTP_1	ff	predicted trithorax protein	39	1.5e-05	Bx
C91B03	gb:T28326	h	proliferating-cell nucl. antigen P40	188	7.4e-06	Bn
C28D09	prf:2109263A	scd	prophenin 1	59	1.3e-07	Bx
C87H12	gb:DMUSVAR39	ff	Su(var)3-9 prot. regul. of homeot gen	99	0.00011	tBx
C78D05	prf:2115310B	h	tastin	51	2.1e-07	Bx
C94H11	sp:TGFB_RAT	r	TRANSFORMING GROWTH FACTOR BETA-1 MAS	49	5.3e-07	Bx
C72B10	prf:2115310A	h	trophinin	40	8.9e-05	Bx
(Oncogenes, tumor suppressors and tumor-related)						
CA2D08	prf:1304193A	h	abl gene	36	0.00043	Bx
C55H12	prf:1202241A	h	gene c-fes/fps	41	0.0035	Bx
C73F01	gb:DMMBT163	ff	MBT163 tumor-supressor	88	0.0012	tBx
C77E02	gp:HUMMTG8_1	h	MTG8 protein	36	1.1e-05	Bx
C78D10	prf:2115319A	h	mucin	55	0.00010	Bx
C88H02	gb:HUMIMUCA	h	mucin 2 (MUC2)	90	0.010	tBx
C33C02	sp:MUC2_HUMAN	h	MUCIN 2 (INTESTINAL MUCIN 2) (FRAGMEN	37	0.0028	Bx
C87F03	sp:MUC2_HUMAN	h	MUCIN 2 (INTESTINAL MUCIN 2) (FRAGMEN	41	0.00015	Bx
C66G12	sp:MUC2_HUMAN	h	MUCIN 2 (INTESTINAL MUCIN 2) (FRAGMEN	48	8.7e-07	Bx
C91A07	gb:CALMYCG	cj	Myc gene, complete cds	71	1.8e-05	tBx
C68B07	gp:HUMPKD1G08_1	h	polycystic kidney disease 1 protein	41	2.8e-07	Bx
C28G03	sp:CBL_HUMAN	h	PROTO-ONCOGENE C-CBL	40	2.7e-05	Bx
C89C11	sp:RBL2_HUMAN	h	RETINOBLASTOMA-LIKE PROTEIN 2 (130 KD	37	3.5e-05	Bx
C72B04	gbu:MMU52945	m	tumor susceptibility protein TSG101	66	8.4e-14	tBx
C88A07	pir:TVHUM2	h	transforming protein N-myc (version 2	44	2.4e-05	Bx
Transcription and Translation						
(Transcription factors)						
C73A03	gb:D13748	h	EUKARYOTIC INITIATION FACTOR 4A-I	66	1.4e-07	tBx
C93H02	pir:S44265	ch	gammaFBP-B	41	0.00055	Bx
C77G02	pir:S53611	r	MIBP1 protein	43	0.0056	Bx
C88F04	pir:S53611	r	MIBP1 protein	47	1.9e-07	Bx
C05E01	prf:1303370D	y	ORF rho11-I	37	0.0014	Bx
C96A01	sp:ICP4_HSV11	hsv	TRANS-ACTING TRANSCRIPTIONAL PROTEIN	45	4.3e-06	Bx
C88A03	sp:X_WHV59	whv	TRANS-ACTIVATING PROTEIN X	32	0.0035	Bx
C28B09	SW:GCN5_YEAST	y	TRANSCRIPTIONAL ACTIVATOR GCN5	180	7.9e-29	Bn
C74F08	gb:HEHSV1G3	hsv	transcriptional activator IE175	71	8.9e-07	tBx
C92H10	sp:TFE3_HUMAN	h	TRANSCRIPTION FACTOR E3 (FRAGMENT)	35	9.9e-05	Bx
C84G10	prf:2116442A	h	transcription factor IIIB	54	0.0071	Bx
C86D05	pir:S34416	do	transcription factor ITF-2	43	0.00047	Bx
C41D04	pir:A45690	hpv	transactivator EBNA-2	44	0.0052	Bx
C10B07	pir:S04336	m	U1 snRNP 70K protein (long form)	46	0.0013	Bx
C96C08	gp:XMU12777_1	xm	ZF1	67	5.2e-06	Bx
C34C01	gp:MUSZFP7_1	m	zinc finger protein	37	0.0056	Bx
CB2C02	pir:JC2069	h	zinc-finger protein, BR140	35	0.00037	Bx
C68B04	gp:HUMZFHP_1	h	zinc finger homeodomain protein	51	0.0019	Bx
(Transcription and translation machinery)						
C95C10	sp:RNHA_HUMAN	h	ATP-DEPENDENT RNA HELICASE A	88	5.7e-05	Bx
C02C03	sp:RPB1_DROME	ff	DNA-DIRECTED RNA POLYMERASE II LARGES	57	7.1e-07	Bx
C02B01	gb:DMHRP361	ff	hrp36.1	66	1.2e-06	tBx
CA7C10	sp:GC14_YEAST	y	GCD14 PROTEIN	67	2.2e-10	Bx
CA1A07	sp:ROU_HUMAN	h	HETEROGENOUS NUCLEAR RIBONUCLEOPROTEI	94	9.2e-10	Bx
C85C05	pir:S49782	y	MSF1 protein=Phe-tRNA synthetase	84	0.00015	Bx
CB3C12	gb:A18777	asq	SP6 RNA polymerase promoter	172	3.2e-06	Bn
(Ribosomal proteins)						
C40D01	sp:RL2_THEMA	tm	50S RIBOSOMAL PROTEIN L2	77	3.3e-09	Bx
CA5F02	SW:RL2_YERPS	yp	50S RIBOSOMAL PROTEIN L2	76	2.4e-11	tBx
CB7G11	gb:RRRPS28	r	ribosomal protein S2	169	0.0023	Fa

Downloaded from http://jb.oxfordjournals.org/ at Peking University on October 2, 2012

TABLE IIB. **ESTs classified into the $p < 0.01$ group** (continued).

EST	DB:Accession #	Species	Putative Identification	Score	P-value	Program
C03H03	pir:S49579	at	ribosomal protein L2	45	0.00012	Bx
C68B05	gb:HSITS2	h	rRNA primary transcript ITS2	86	0.0020	tBx
C82E05	gb:w65998	m	similar, WP:T04A8.11 50S RIBO PRO L16	91	4.5e-21	tBx
C01D08	gbu:PHU37526	ph	small subunit ribosomal RNA gene	172	5.0e-05	Bn
(Heat shock proteins)						
Membrane-Associated						
(Receptor and membrane-associated)						
C74A09	gp:PAR51A_1	pt	51A surface protein	36	0.0025	Bx
C87G07	sp:ADP1_MYCPN	mp	ADHESIN P1 PRECURSOR (CYTADHESIN P1)	41	5.9e-08	Bx
C86G08	pir:A55575	h	ankyrin 3, long form	60	0.0052	Bx
CB2D10	gb:HSBEIN	h	beta1 integrin	154	0.0081	Bn
C71A06	sp:TRKB_MOUSE	m	DNF/NT-3 GROWTH FACTORS RECEPTOR P	46	0.0095	Bx
CA7C02	sp:GAA1_YEAST	y	GAA1 PROTEIN	62	0.0076	Bx
C83B01	sp:NME3_MOUSE	m	GLUTAMATE (NMDA) RECEPTOR SUBUNIT EPS	39	4.1e-05	Bx
C87F11	pir:JQ1579	hbv	major surface antigen	42	4.2e-08	Bx
C87D07	pir:A37967	ch	neural cell adhesion molecule Ng-CAM	39	9.8e-06	Bx
C74H10	gp:RATNR2DA_1	r	NMDA receptor subunit NR2D	39	2.6e-07	Bx
CC0B05	pir:A40670	r	nuclear envelope protein POM 121	44	4.4e-07	Bx
CA0H09	prf:2118292A	le	receptor adenylate cyclase	42	9.4e-06	Bx
C55A08	sp:RYNR_HUMAN	h	RYANODINE RECEPTOR, SKELETAL MUSCLE	37	0.0031	Bx
C69D04	prf:2003267A	ff	ryanodine receptor/Ca release channel	42	0.00046	Bx
C41E03	pir:A54161	fr	ryanodine-binding protein alpha form	42	0.00044	Bx
C57E05	sp:ZO1_HUMAN	h	TIGHT JUNCTION PROTEIN ZO-1	36	0.00076	Bx
C88A09	gp:RNVCAM1R_1	r	Vascular cell adhesion molecule 1	73	1.1e-05	Bx
(Transporters)						
C94E04	sp:BAR3_CHITE	ct	BALBIANI RING PROTEIN 3 PRECURSOR	37	0.00023	Bx
C90E11	sp:CCB2_RABIT	rab	BRAIN CALCIUM CHANNEL BI-2 PROTEIN	51	6.2e-06	Bx
C78F10	pir:A34308	r	Ca2+-transporting ATPase (EC 3.6.1.38	36	6.6e-05	Bx
C82D04	SP:S44090	r	carnithine/acylcarnithine CARRIER PRO	95	0.00092	tBx
C10A06	sp:CIC2_RAT	r	CHLORIDE CHANNEL PROTEIN 2 (CLC-2)	37	0.0066	Bx
C88F03	sp:CICH_TORMA	tm	CHLORIDE CHANNEL PROTEIN	47	0.0071	Bx
C67A11	gbu:MMDAT1	m	dopamine transporter gene	187	7.9e-06	Bn
CB7C10	pir:A53102	ch	LDL receptor-related protein/alpha-2	39	7.9e-05	Bx
CA1C01	pir:A47437	ce	LDL-receptor-related protein	38	0.0023	Bx
CB1E08	gpu:RATCMOAT_1	r	organic anion transporter	48	0.00017	Bx
C93H12	SW:SC23_YEAST	y	PROTEIN TRANSPORT PROTEIN SEC2	191	1.2e-19	Bn
Other-secreted Proteins						
C78G10	gp:BOVAMEL_1	bo	amelogenin	36	0.00074	Bx
C78B07	prf:2110286A	ff	masquerade gene	36	0.00018	nBx
Other Metabolism						
(DNA metabolism)						
C00D08	sp:DPOL_HCMVA	cmv	DNA POLYMERASE (EC 2.7.7.7)	44	0.0068	Bx
C87H04	sp:MTDM_HUMAN	h	DNA (CYTOSINE-5)-METHYLTRANSFERASE	42	1.2e-05	Bx
C39C09	SW:NHP2_YEAST	y	HMG-LIKE NUCLEAR PROTEIN 2	95	1.2e-05	Bn
C30C01	gb:MUSHIS2AR	m	replication-dependent histone H2A.1	163	8.2e-06	Bn
CB5D05	gp:SCD9717_15	y	suppressor of MIF2 mutations	38	1.9e-06	Bx
C33A09	pir:S18106	ab	type II site-specific deoxyribonuclea	55	0.0042	Bx
(Repetitive DNA and virus-related sequences)						
C84A07	prf:0910274A	hhv	antigen, virus nuclear	43	8.4e-07	Bx
C69C03	gp:HSBBICP4A_1	bhv	BICP4	42	5.2e-05	Bx
C87C12	gb:HSA205YB1	h	DNA containing (CA) repeat	157	0.0043	Bn
C87F12	gp:HHV6AGNM_3	hhv	DR2	57	3.4e-05	Bx
C73H11	gp:HHV6AGNM_3	hhv	DR2	39	1.8e-06	Bx
C93D03	gp:HPU31778_5	hpv	E4 gene product	40	0.0021	Bx

Downloaded from http://jb.oxfordjournals.org/ at Peking University on October 2, 2012

TABLE IIB. **ESTs classified into the $p<0.01$ group** (continued).

EST	DB:Accession #	Species	Putative Identification	Score	P-value	Program
CA6G01	pir:S10206	h	early E2A DNA-binding protein	34	0.0035	Bx
C89H07	sp:E321_ADE1P	adv	EARLY E3 20.6 KD GLYCOPROTEIN	43	0.0048	Bx
CA2C03	sp:ENV_FIVPE	fiv	ENVELOPE POLYPROTEIN PRECURSOR (GP150	45	2.6e-08	Bx
C98D12	sp:ENV_MLVFP	fmv	ENV POLYPROTEIN PRECURSOR	42	0.0089	Bx
C46G08	gpu:HIV1U39254_1	hiv	envelope glycoprotein	41	2.9e-05	Bx
C42C10	gp:T1223ENVG_1	uni	env gene product	45	1.8e-05	Bx
C94E06	prf:1207326C	hiv	env gene	41	0.0025	Bx
C88C12	gp:FCGGAONC_1	fsv	feline sarcoma virus (gardner-arnstei	37	4.4e-05	Bx
C82C05	gb:HUMFLNG6PD	h	from filamin (FLN) to G6PD gene	69	0.0059	tBx
C67G03	prf:0711245A	mlv	gag/pol/env protein	41	6.7e-05	Bx
C86D12	sp:POLG_TMEVB	mev	GENOME POLYPROTEIN (COAT PROTEINS VP1	44	0.0083	Bx
C88C04	sp:POLG_DEN1S	dev	GENOME POLYPROTEIN (CONTAINS: CAPSID	40	2.4e-07	Bx
C90C09	sp:POLG_HCVJ6	hcv	GENOME POLYPROTEIN (CONTAINS: CAPSID	36	0.00062	Bx
C67E08	sp:VGLB_HSV1F	hsv	GLYCOPROTEIN B PRECURSOR	42	0.00010	Bx
CA7H07	sp:VGLH_EBV	hhv	GLYCOPROTEIN GP85 PRECURSOR	46	0.00090	Bx
C39E07	pir:S54266	chv	glycoprotein gC	37	2.6e-06	Bx
C73B03	gp:IVHAARN79_1	ifv	hemagglutinin precursor (partial)	42	0.00064	Bx
C95C02	sp:Y48K_ELV	elv	HYPOTHETICAL 48 KD PROTEIN	46	0.00043	Bx
C74B09	pir:B40505	shv	hypothetical protein	43	1.9e-05	Bx
C90C06	pir:B40505	shv	hypothetical protein	55	6.9e-08	Bx
C91A06	pir:B40505	shv	hypothetical protein	87	7.6e-11	Bx
C77H02	gpu:HHU43400_37	hhv	large tegument protein	40	0.0017	Bx
C81D04	gpu:HPVL1_1	hpv	major capsid protein	47	0.00033	Bx
C30B01	gp:OLVCG_2	ol	pol polyprotein	32	0.0026	Bx
CB4B01	sp:RPOL_EAV	eav	POL POLYPROTEIN (CONTAINS: RNA-DIRECT	43	0.00030	Bx
C29D02	prf:2015331A	hcv	polyprotein	40	0.00015	Bx
C95B02	sp:VNUA_PRVKA	prv	PROBABLE NUCLEAR ANTIGEN	55	7.6e-10	Bx
C73E08	sp:VNUA_PRVKA	prv	PROBABLE NUCLEAR ANTIGEN	49	7.5e-06	Bx
C85B04	sp:RRPO_PMV	pmv	RNA REPLICATION PROTEIN	39	0.00045	Bx
CB3D05	sp:VP2_ROTBR	bo	RNA-BINDING PROTEIN VP2 (MAJOR INTERN	77	0.00018	Bx
C77E07	gp:HS1LSSDS_1	hsv	type 1 latency-as	45	1.9e-07	Bx
C74D11	prf:1508243AP	hhv	UL36 gene	39	0.00013	Bx
C93A02	gp:BVD125IN1_1	bo	viral nonstructural protein p125	40	0.0013	Bx
(Protease and protease inhibitors)						
CB7D01	sp:IWIT_MEDSA	alf	BOWMAN-BIRK TYPE WOUND INDUCED TRYPSI	41	0.0058	Bx
C94F06	gb:HSCYSTC1	h	cystatin C exon 1	78	2.8e-08	tBx
C69D05	prf:2117219B	h	matrix metalloprotease: ISOTYPE=2	50	0.0063	Bx
CC0D07	pir:A25834	pig	plasmin (EC 3.4.21.7)	39	0.0046	Bx
CA2E12	prf:2109370A	hhv	protease	44	0.0041	Bx
C68D06	pir:JX0309	elv	proteinase inhibitor (Bowman-Birk)	40	0.0074	Bx
CA9E08	sp:UBPX_HUMAN	h	PROBABLE UBIQUITIN CARBOXYL-TERMINAL	84	0.00021	Bx
C91G08	pir:S47152	an	serine-type carboxypeptidase (EC 3.4.	42	8.9e-06	Bx
C73C09	gb:H33498	r	Ubiquitin activator	194	3.0e-08	Bn
(Other metabolism)						
CB1B12	sp:THHR_HORVU	hvu	ANTIFUNGAL PROTEIN R	36	0.00077	Bx
CC0B08	sp:CLAT_HUMAN	h	CHOLINE O-ACETYLTRANSFERASE (EC 2.3.1	43	9.3e-05	Bx
C83B11	gb:HUMGNOS48	h	endothelial nitric-oxide-synthase gen	142	7.4e-06	Bn
C94F05	pir:S55511	pv	GAM1 protein	48	1.4e-05	Bx
C74G11	pir:JC2467	ps	iron sulfür protein large chain	37	2.3e-05	Bx
C87G04	pir:A56390	sp	mannosyl-glycoprotein endo-beta-N-ace	40	2.6e-05	Bx
C83H01	sp:MT2A_RABIT	rab	METALLOTHIONEIN-IIA (MT-2A)	35	2.5e-05	Bx
C87B06	prf:1817302A	h	nitric oxide synthase	48	0.0046	Bx
C71E01	sp:OXYB_RABIT	rab	OXYSTEROL-BINDING PROTEIN	52	3.7e-07	Bx
C84E08	gb:T27705	h	phosphoglucomutase 1	153	1.9e-16	Bn
C75C12	gpu:SAPKSGENE_1	y	polyketide synthase	45	0.0052	Bx
C84G08	gpu:SPAC18G6_5	y	polyketide synthase	46	0.0028	Bx
C69E01	prf:2116303A	soc	polyketide synthase	55	1.7e-07	Bx
C74A07	gb:W71369	m	similar to WP:C08B11.7 THIOESTERASE	97	0.00098	tBx

Downloaded from http://jb.oxfordjournals.org/ at Peking University on October 2, 2012

TABLE IIB. **ESTs classified into the $p < 0.01$ group** (continued).

EST	DB:Accession #	Species	Putative Identification	Score	P-value	Program
C04E08	pir:A54964	h	spliceosome-associated protein SAP-49	87	1.6e-08	Bx
CB8F11	gp:HSSTAR8_1	h	steroidogenic acute regulatory protei	91	1.9e-05	Bx
(Not classified)						
C89A02	sp:PRP5_HUMAN	h	BASIC PROLINE-RICH PEPTIDE IB-1	52	2.5e-08	Bx
C90B07	pir:A61294	h	basic proline-rich glycoprotein	33	0.0028	Bx
C67D01	gpu:CEC01F6_2	ce	C01F6.1	53	0.00045	Bx
C93B09	gpu:CEC30F2_1	ce	C30F2.1	50	0.00032	Bx
C66D03	gp:HUMCPGL_1	h	Clone S16 gene from CpG-e	36	0.00072	Bx
C34E01	gpu:CED1046_4	ce	D1046.4	36	0.0088	Bx
C73E03	gp:CELF26A1_9	ce	F26A1.11 gene product	40	0.00072	Bx
C30B05	gpu:CEF36H1_2	ce	F36H1.5	39	0.00018	Bx
C83H09	gp:CEF37A8_3	ce	F37A8.3	39	0.0025	Bx
C33B04	gpu:CELF46C8_1	ce	F46C8.4 gene product	41	0.0084	Bx
C59A12	gp:CEF47A4_2	ce	F47A4.2	45	1.6e-05	Bx
CB5C12	gp:CELF48E8_6	ce	F48E8.2 gene product	94	2.3e-12	Bx
CA0H02	gp:HSU10991_1	h	G2	37	0.00017	Bx
C81A02	pir:S19774	tom	glycine-rich protein	45	1.6e-05	Bx
C83A12	gp:HSHCF1_1	h	host cell factor	45	2.2e-10	Bx
CA6E10	pir:S19341	y	hypothetical protein YCL014w	45	0.00031	Bx
C86F11	pir:JQ0405	ml	hypothetical 119.5K protein (uvrA reg	40	0.00072	Bx
C87C06	pir:S23689	pf	hypothetical protein 6	37	0.0010	Bx
C67A03	pir:S40713	ce	hypothetical protein	42	0.00034	Bx
C81D02	sp:YLS8_CAEEL	ce	HYPOTHETICAL 83.6 KD PROTEIN F09G8.8	36	3.6e-05	Bx
C75E11	sp:YNX3_CAEEL	ce	HYPOTHETICAL 337.6 KD PROTEIN T20G5.3	42	8.2e-07	Bx
CA4D09	sp:YQXL_BACSU	bs	HYPOTHETICAL PROTEIN IN COMG 5'REGION	36	0.0043	Bx
C73A06	sp:J1I_HCMVA	bs	HYPOTHETICAL PROTEIN HKRFX (J1I)	43	0.0036	Bx
CA0A05	gpu:CELK03A1_1	ce	K03A1.3 gene product	48	0.00095	Bx
C94D05	gpu:CEK07F5_14	ce	K07F5.14	76	2.4e-07	Bx
C94C12	gbu:HUMKG14	h	myeloblast KIAA0134 gene	70	0.00013	tBx
CA5C01	gbu:HUMKG16	h	myeloblast KIAA0136 gene	96	0.00038	tBx
C72B03	gb:HUMKG1CC	h	ORF (novel protein)	146	0.00059	Bn
CB2A03	gp:HUMKG1C_1	h	ORF	37	1.2e-06	Bx
C71A10	prf:1304244B	mc	ORF	41	0.00023	Bx
C96F10	pir:S24577	ff	ovarian protein	46	0.00063	Bx
C30A08	gb:HSB28A062	h	partial cDNA; clone 28A	172	0.00028	Bn
C89C07	pir:S24600	ff	projectin	35	0.00044	Bx
C81C12	pir:S16589	tom	proline-rich protein	41	0.00046	Bx
C91C10	gp:SC38KCXVI_11	y	putative protein	85	6.1e-08	Bx
C33E03	pir:S43565	ce	R01H10.4 protein (clone R01H10)	40	0.00023	Bx
C92D07	gpu:CELR03G5_3	ce	R03G5.3 gene product	54	2.5e-07	Bx
C95F02	gp:CELR144_7	ce	R144.2 gene product	39	0.0041	Bx
C04E01	gp:CER166_2	ce	R166.2	52	0.00042	Bx
C93A08	gb: AA003101	m	similar to WP:R13A5.12 CE01373	74	1.1e-15	tBx
C86H07	gp:HUMSPR2C_1	h	small proline-rich protein	56	0.00011	Bx
C81E08	gb:R22491	h	SP:F01F1.11 CE01224	152	3.6e-08	Bn
CA6A08	gp:CELT10E10_2	ce	T10E10.4 gene product	54	8.6e-05	Bx
CA5D11	sp:WASP_HUMAN	h	WISKOTT-ALDRICH SYNDROME PROTEIN (WAS	42	3.8e-05	Bx
C81C09	gpu:CEZK1067_5	ce	ZK1067.2	57	8.6e-05	Bx

Abbreviations: ab, *Azospirillum brasilense*; adv, human adenovirus; al, *Arbacia lixula*; alf, alfalfa; an, *Aspergillus niger*; asq, artificial sequence; bm, *Bombyx mori*; bhv, bovine herpesvirus; bp, *Bordetella pertussis*; bs, *Bacillus subtilis*; ca, carrot (*D. carota*); cj, *Callithrix jacchus*; chv, caprine herpesvirus; cmv, human cytomegalovirus; co, *Crithidia oncopelti*; ct, *Chironomus tentans* (Midge); do, dog; dev, Dengue virus; eav, Equine arteritis virus; elv, Erysimum latent virus; en, *Emericella nidulans*; eva, *Erythrina variegata*; fiv, feline immunodeficiency virus; fmv, Friend murine leukemia virus; fpv, fowlpox virus; fr, bullfrog (*Rana catesbeiana*); fsv, feline sarcoma virus; hbv, hepatitis B virus; hcv, hepatitis C virus; hhv, human herpesvirus (Epstein-Barr virus); hiv, human immunodeficiency virus; hpv, human papillomavirus; hsv, Herpes simplex virus; hvu, *Hordeum vulgare* (barley); hyd, *Hydra*; ifv, Influenza virus; jl, *Janthinobacterium lividum*; kb, kidney bean (*Phaseolus vulgaris*); le, *Leishmania*; ma, *Mesocricetus auratus*; mb, *Mycobacterium bovis*; mc, *Mycobacterium tuberculosis*; mev, Theiler's murine encephalomyelitis virus; mf, *Macaca fascicularis*; ml, *Micrococcus luteus*; mlv, Moloney murine leukemia virus; mp, *Mycoplasma pneumoniae*; mz, maize; nem, nematode (*Haemonchus contortus*); oc, *Oryctolagus cuniculus*; ol, *Ovine lentivirus*; pa, *Paramecium aurelia*; pf, *Plasmodium falc*; ph, *Polypodium hydriforme*; pmv, Papaya mosaic potexvirus; po, potato (*Solanum tuberosum*); prv, Pseudorabies virus; ps, *Pseudomonas*; pt, *Paramecium tetraure*; pv, *Plasmodium vivax*; rab, rabbit; rc, *Rhodobacter capsulatus*; sa, *Squalus acanthias* (spiny dogfish); scd, *Susscrofa domestica*; shv, suid herpesvirus; slm, slime mold (*Dictyostelium discoideum*); soc, *Sorangium cellulosum*; sp, *Streptococcus pneumoniae*; tm, *Thermotoga maritima*; tob, tobacco (*Nicotiniana tabacum*); tom, tomato; tva, *Trichomonas vaginalis*; uni, unidentified; wht, wheat; whv, Woodchuck hepatitis virus; xl, *Xenopus laevis*; xm, *Xiphophorus maculatus*; yp, *Yersinia pseudotuberculosis*. All other abbreviations were as described in the footnotes to Table IIA.

Downloaded from http://jb.oxfordjournals.org/ at Peking University on October 2, 2012

TABLE III. **Redundant ESTs prepared.**

Group[1]	Redundant EST[2] (DB: Accession #[3])
1) 29A03^^	A5B04^^
2) 40D01^^	B4A02* (gb:MUSGS00013, Ribosomal protein L2)
3) 66D03^^	71E01^^
4) 66H05^	A4D09^^
5) 67D04*	81D06*, B9B12**(gb:RNRPS14, Ribosomal protein S14)
6) 67E02*	69D01* (gb:HSU28686, RNA binding protein)
	88H08* (gb:R62689, HNRN protein)
	93F04* (gb:R36350, HNRN protein)
7) 67E04*	74C05* (gbu:HSU46751)
8) 68A10^^	90E04^^
9) 68F01*	75A09* (gb:MUSA4P, mouse alpha 4 protein)
10) 68F08*	73E12* (gb:HSTCL1), 73F03* (gb:HSTCL1)
	83A02* (gb:HSTCL1), A7H03* (gb:HSTCL1)
11) 69D04^^	78B11^^, 86A10, 94B12
12) 72B02*	86H04**(gb:H32230, interferon gamma)
13) 72C01	92H10
14) 74B09^^	86B10^
15) 74H10^^	83B10^
16) 77D10*	78D05^^, 87E04* (SW:HG74_HUMAN), 90G09* (SW:HG74_HUMAN),
	92D11* (SW:HG74_HUMAN), 92E02^^, 04A10**(gb:HUMOGC)
17) 81D02^^	82D04^^
18) 81E03	99C08
19) 82C06	94B11^
20) 83E03	84B07
21) 84G10^^	96F10^^
22) 87B08	88C05
23) 87G07^^	89A02^^
24) 88A07^^	93B12^^
25) 88D03*	B8D06**(gb:T99403, ubiquitin-conjugating enzyme)
26) 95A02*	74D06* (gb:HSTEGT, TEGT gene)
27) A4F02*	B2G02* (gbu:HSU46752)
28) A9F09*	A9F10
29) B7G11^^	B8E08**(gb: MMU11248, Ribosomal protein S28)

[1]Redundant ESTs were grouped as to sequence similarities. [2]Names of redundant ESTs with significant sequence similarities. [3]Database names and accession numbers of ESTs (see the text and Table II). The dbEST and $p<0.01$ group ESTs were denoted by ^ and ^^, respectively. ESTs denoted by ** were identified in our previous works (*5, 6*), and those denoted by * were identified in the present work. Although the groups 7 and 27 ESTs were not sharing any significant sequence similarities, these two groups are representing different parts of the same mRNA. All other abbreviations were as described in the footnotes to Tables IIA and IIB.

to match the protein database sequences with a p-value <0.01 (Table IV). Although 10 of the 27 ESTs have already been classified into the $p<0.01$ group by means of a BLASTX search (Tables IIB and IV), only 5 of the 10 putative identifications were identical between the two different searches (Table IV). By means of a MFASTA search, one EST was newly classified into the identified group, and another 16 into the $p<0.01$ group (Table IV). These results indicate that a MFASTA search is useful for detecting homologous sequences missed in a BLASTX search. By means of a MFASTA search, we found that 94B12 codes for a protein showing 30% identity in 115 amino acids, which overlaps a *Drosophila* muscle segment homeobox (msh) protein (p-value=0.002) (Table IV). Interestingly, 94B12 shared a highly homologous nucleotide sequence with 69D04, 78B11, and 86A10 (Table III).

iv) Motif pattern search: DBPROSITE program: To further classify the unidentified ESTs and to search for clues as to the functions of the coded proteins, we searched all possible open reading frames for the presence of specific motif patterns, using the DBPROSITE program. The prosite documentation file containing 889 documentation entries includes 1,167 different patterns (*12*). Although we found more than 10,000 motif patterns, more than 99% were related to post-translational modifications of proteins (data not shown). The top five motif patterns were those of the protein kinase C phosphorylation site, *N*-myristoylation site, casein kinase II phosphorylation site, *N*-glycosylation site, and cAMP- and cGMP-dependent protein kinase phosphorylation site (data not shown).

Other than the motif patterns related to the post-translational modifications of proteins, we found 10 different motif patterns (Table V). Thirty-nine of the 500 unidentified ESTs contained one or two of these motif patterns (Table V). It is apparent that one can not deduce from the motif patterns the most likely proteins for the majority of unidentified ESTs. In this motif pattern search, 95C10 was predicted to code for a protein carrying an ATP/GTP-bind-

Downloaded from http://jb.oxfordjournals.org/ at Peking University on October 2, 2012

TABLE IV. **ESTs newly identified in a MFASTA search.**

EST (C)[1]		DB|Accession #	Species[2]	Putative Identification	Score	P-value
C73H04	*	pir|S54549	y	hypothetical protein YM9796.02c	273	7.60e-19
CA0H07		pir|S40612	slm	myosin-related protein	129	3.50e-08
	Bx	pir|S40612	slm	myosin-related protein	89	3.70e-05
C95B02	*	pir|S64242	y	hypothetical protein YGL220w	143	6.90e-07
CB8F11		pir|A55455	m	steroidogenic acute regulatory	124	5.50e-06
	Bx	pir|A55455	m	steroidogenic acute regulatory	90	2.60e-05
C85C05		pir|S37758	y	MSF1 protein	172	2.50e-05
	Bx	pir:S49782	y	MSF1 protein	84	0.00015
C72H03	*	pir|A35320	rbv	nonstructural polyprotein	104	3.20e-05
C73F09(C)	*	pir|S43455	y	hypothetical protein (LEU2 3' r	109	0.00006
C94B01(C)	*	pir|S60123	y	hypothetical protein R10E11.1	100	0.0002
CA9E09(C)	*	pir|E26277	hpv	E4 protein	105	0.00046
C69H03(C)		pir|C40513	h	hypothetical protein ORF3	100	0.00049
	Bx	sp:GLT4_WHEAT	wht	GLUTENIN	63	7.1e-07
C59H07		pir|S42731	su	collagen alpha 1 chain	119	0.0005
	Bx	pir:S57243	m	collagen alpha 1(I) chain precur	51	6.50e-09
C68B04(C)		sp|P04280	h	SALIVARY PROLINE-RICH PROTEIN	110	0.0013
	Bx	gp:HUMZFHP_1	h	zinc finger homeodomain protein	51	0.0019
C54A04	*	pir|S08491	ha	hypothetical protein	113	0.0015
C66D03		sp|P54258	r	ATROPHIN-1 (DENTATORUBRAL-PALL	136	0.0018
	Bx	gp:HUMCPGL_1	h	Human (clone S16) gene from CpG-	36	0.00072
C84D06	*	pir|S09843	cmv	UL80 protein	107	0.0019
C94B12	*	pir|S55392	ff	msh protein	101	0.002
C94G10(C)	*	pir|S63389	y	hypothetical protein YNR057c	105	0.0021
C74C08(C)		pir|A31757	m	homeotic protein Hox 2.6	125	0.0022
	Bx	sp:HXB4_MOUSE	m	HOMEOBOX PROTEIN HOX-B4 (HOX-2.6	62	1.00e-06
C86B10	*	pir|C45219	r	N-methyl-D-aspartate receptor c	103	0.0028
C68C02(C)	*	sp|P41688	ca	T-CELL SURFACE GLYCOPROTEIN	107	0.0029
CB8E02(C)	*	pir|S36749	r	transcription factor HES-3	127	0.0029
C71E01		sp|P42522	slm	MYOSIN IC HEAVY CHAIN.	112	0.0046
	Bx	sp:OXYB_RABIT	rab	OXYSTEROL-BINDING PROTEIN	52	3.70e-07
C95A05		sp|P33524	bn	CRUCIFERIN BNC2 PRECURSOR (1	108	0.0067
	Bx	sp:GLT5_WHEAT	wht	GLUTENIN	58	6.00e-06
C74B10	*	sp|P49901	h	SPERM MITOCHONDRIAL CAPSULE S	109	0.007
C82E12	*	sp|P37370	y	VERPROLIN	128	0.0073
C90F01(C)	*	pir|A36325	r	epidermal growth factor recepto	107	0.0088
C94C07	*	sp|P00544	fsv	TYROSINE-PROTEIN KINASE TRANSP	106	0.0091

[1](C), amino acid sequences were predicted from the complementary strands; the Bx lines show the results of a BLASTX search, and * indicates that no homologous sequence with $p<0.01$ was detected in the search. [2]Species: bn, *Brassica napus* (rape); ca, cat; rbv, rubella virus; su, sea urchin. All other abbreviations were as described in the footnotes to Tables IIA and IIB.

ing site motif, and 55A08 for a protein carrying a growth factor and cytokine receptor family signature (Table V). These results may support the putative identification of 95C10 as ATP-dependent RNA helicase (see Table IIB, "Transcription and translation machinery"), and that of 55A08 as the ryanodine receptor (see Table IIB, "Receptor and membrane-associated").

v) Predicted reading frames: Grail program: Individual ESTs were analyzed using the GRAIL neural network (*13*) to determine the probability that each sequence encodes a protein. The predicted translations allowed further analysis of ESTs to search for new gene families, motifs and other structural features. Each of the ESTs was classified as a "coding" or "non-coding" EST, according to the probability of containing a coding sequence (Table VI). Interestingly, 44% of the 716 unidentified ESTs were predicted to contain coding regions with GRAIL, whereas 70% of the 573 identified ESTs appeared to contain coding regions (Table VI). The 716 unidentified ESTs were further characterized in this work and 109 of them were confirmed to correspond to neither known gene and/or protein database sequences nor to the dbEST sequences, *i.e.*, the $p>0.01$ group (Table I). Only 34% of the $p>0.01$ group ESTs were predicted to contain coding regions (Table VI). These results are probably due to the following two possibilities; (i) the previously identified ESTs (*5, 6*) represent more abundant transcripts and match the neural network training set used for GRAIL better than the low abundance mRNAs represented by the unidentified ESTs, and (ii) the unidentified ESTs, such as the $p>0.01$ group ones, include sequences corresponding to either the 3′-untranslated region or the transcripts of the repetitive sequences. The above results suggest that the GRAIL-predicted false negative rate for mouse ESTs is around 30%. This value is comparable to that estimated for the human ESTs, *i.e.*, 20 to 30% (*15*). About 10% of the identified ESTs isolated from the unidirectional library

Downloaded from http://jb.oxfordjournals.org/ at Peking University on October 2, 2012

TABLE V. **Prosite motif patterns detected in unidentified ESTs[1].**

EST (No.)[2]	Motif pattern	Name of motif pattern	Prosite pattern number
C33A09 (13)	R-G-D.	Cell attachment sequence	PS00016
C95C10 (8)	[AG]-x(4)-G-K-[ST]	ATP/GTP-binding site motif A (P-loop)	PS00017
C33E03 (2)	C-x-C-x(5)-G-x(2)-C	EGF-like domain cysteine pattern signature	PS00022
C68D01 (9)	L-x(6)-L-x(6)-L-x(6)-L.	Leucine zipper pattern	PS00029
C81C12 (1)	G-H-E-x(2)-G-x(5)-[GA]-x(2)-[IVAC]	Zinc-containing alcohol dehydrogenases signature	PS00059
CA9E09 (1)	[LIF]-G-x(4)-[LIVMF]-P-W.	Dihydrofolate reductase signature	PS00075
C30A08 (2)	C-{CPWHF}-{CPWR}-C-H-{CFYW}.	Cytochrome c family heme-binding site signature	PS00190
C28A03 (1)	P-[DE]-W-[FY]-[LFY](2)	Cytochrome b/b6 Qo site signature	PS00193
C29D02 (1)	[DENG]-x-[DENQGSTARK]-x(0,2)-[DENQARK]-[LIVFY]-{CP}- G-{C}-W-[FYWLRH]-x-[LIVMTA]	Lipocalin signature	PS00213
C55A08 (1)	[STGL]-x-W-[SG]-x-W-S.	Growth factor and cytokines receptors family signature 2	PS00340

[1]Pattern matches other than those related to the post-translational modification sites in the prosite database are shown. The open reading frames include those of 500 unidentified ESTs. [2]Name of the motif pattern-matched EST. When more than one EST matched the same pattern, one of the ESTs is shown as a representative. The numbers in parentheses are the numbers of ESTs having the same motif pattern.

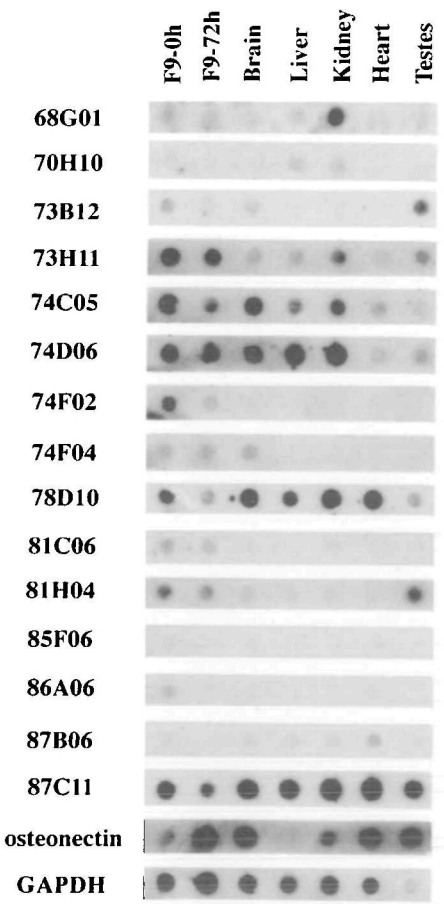

TABLE VI. **Prediction of ORFs in F9 ESTs.**

Class of ESTs	No. of ESTs examined	Predicted ORFs (%) Excellent	Good	Marginal	Total
(Previous classification)[1]					
Identified ESTs	573	311 (54)	73 (13)	22 (4)	406 (70)
Unidentified ESTs	716	184 (26)	80 (11)	53 (7)	317 (44)
(716 unidentified ESTs were further characterized in this work)[2]					
Identified ESTs	216	57 (26)	32 (15)	10 (5)	100 (46)
$p<0.01$ group	307	84 (27)	32 (10)	26 (8)	140 (46)
$p>0.01$ group	109	21 (19)	9 (8)	7 (6)	37 (34)
dbEST group	84	22 (26)	7 (8)	10 (12)	40 (48)

(GRAIL program)

[1]All of these identified ESTs were described in Ref. *5* or *6*, and all of the 716 unidentified ESTs were characterized in this work. [2]The results of characterization are summarized in Table I.

Fig. 1. **Dig dot-blot analysis of expression patterns.** Each filter strip shows seven spots corresponding to the poly(A)$^+$RNAs prepared from the following sources; undifferentiated F9 cells (F9-0h), F9 cells treated with RA for 72 h (F9-72h), and brain, liver, kidney, heart, and testes of adult mice, respectively. The signals of the spots on each strip were measured with an ATTO Densitograph Lumino-CCD and the quantification program, ATTO Densitograph Ver3.01. The names of the ESTs used as probes are given on the left of each panel. The sizes of the inserts were between 0.5 and 3.5 kb, and they were DIG-labeled by PCR, as described under "MATERIALS AND METHODS." The sizes of the PCR products ranged between 0.5 and 2 kb. The X-ray film was exposed overnight, except in the cases of the osteonectin and GAPDH cDNA probes, where the film was exposed for 5 h. Osteonectin cDNA (*6*) was used as an up-regulated gene control, and GAPDH cDNA as a housekeeping gene control.

were predicted not to contain coding frames in the expected direction but rather to contain them in the opposite orientation (data not shown). This predicted ratio is a little higher than that determined in our previous study, *i.e.*, 3 of the 170 ESTs (about 2%), cloned in the λuni-ZAP vector and classified into an identified group, carried cDNAs in the opposite orientation (*6*).

DIG Dot-Blot Analysis of the Expression Patterns of the Unidentified ESTs—The expression patterns of the unidentified ESTs should provide a starting point for elucidating

Downloaded from http://jb.oxfordjournals.org/ at Peking University on October 2, 2012

TABLE VII. **Summary: Expression patterns of ESTs.**

	Numbers of ESTs examined		
	Ided ESTs	Unided ESTs	Total
Regulation by retinoic acid in F9 cells[1)]			
Up regulation	28 (10)	51 (18)	79 (14)
No regulation	218 (81)	218 (77)	436 (79)
Down regulation	23 (9)	15 (5)	38 (7)
Total numbers of ESTs examined	269 (100)	284 (100)	553 (100)
Organ-specific expression in adult mice[2)]			
Brain	0 (0.0)	2 (0.7)	2 (0.4)
Liver	9 (3.3)	3 (1.1)	12 (2.2)
Kidney	0 (0.0)	2 (0.7)	2 (0.4)
Heart	2 (0.7)	2 (0.7)	4 (0.7)
Testes	21 (7.8)	9 (3.2)	30 (5.4)

[1)]The levels of hybridizable RNAs in undifferentiated F9 cells and F9 cells treated with retinoic acid for 72 h were compared, as described under "MATERIALS AND METHODS." The numbers indicating the levels in RA-treated F9 cells were divided by those in undifferentiated F9 cells. ESTs showing values higher than 3, between 3 and 0.3, and lower than 0.3 were classified as RA up-regulated, non-regulated, and down-regulated ESTs, respectively. The numbers in parentheses are the relative ratios (%). The levels of mRNAs were measured using the DIG system, as described under "MATERIALS AND METHODS." [2)]The levels of hybridizable RNAs in adult mouse organs were determined by densitometer reading (see "MATERIALS AND METHODS" and Fig. 1). When one of the five organs examined showed an at least 3-fold higher level of mRNA than those of four other organs, this organ was taken as showing organ-specific expression.

the functions of the coded proteins. To take advantage of F9 cells and the mouse system, we tried to classify unidentified ESTs by means of the expression patterns in F9 cells and several organs of adult mice. We also tried to screen ESTs possibly involved in the regulation of early mouse differentiation, based on this classification.

We arbitrarily picked up 393 ESTs from the 716 unidentified ESTs and 160 ESTs from the ESTs we previously identified (*6*): 109 of the 393 ESTs were identified in the present study. We estimated the levels of the mRNAs corresponding to these 553 ESTs in undifferentiated F9 cells, F9 cells treated with RA for 72 h, and in brain, liver, kidney, heart, and testes of adult mice. Several typical results of dot-blot analysis are presented in Fig. 1. The strength of each signal was quantitated with a densitometer (see "MATERIALS AND METHODS"), and the examined ESTs were classified as to the responses to RA in F9 cells into up-regulated, non-regulated, and down-regulated ESTs (Table VII). A few ESTs showed organ-specific expression patterns (Table VII).

Using these expression patterns, the ESTs were further classified by hierarchical cluster analysis (*14*). The results obtained for the RA down-regulated group are shown in Table VIII. Thirty-eight of the 553 examined ESTs belonged to the RA down-regulated group: 23 were identified ESTs and the remaining 15 unidentified ESTs (Tables VII and VIII). By means of hierarchical cluster analysis, the RA down-regulated ESTs were classified into the following five clusters (Table VIII): cluster I, showing low level expression in all five organs, and including 7 unidentified and 2 identified ESTs, such as 83A02 and 73F03 (=human T cell lymphoma 1 gene, TCL1 gene) (*16*); cluster II, expressed in heart and/or kidney, and including two unidentified and two identified ESTs, such as A6A11 (=transforming growth factor beta-1 binding protein); cluster III, expressed in liver, and including 1 unidentified and 4 identified ESTs, such as A4E06 (=ph34, down-regulated in EC cells); cluster IV, expressed in all organs examined, and including 5 identified ESTs, such as 69E11 (=cleavage stimulating factor); and cluster V, expressed in testes and several other organs, and including 4 unidentified and 10 identified ESTs, such as 96B04 (=proteasome C2 component), A5E09 (=ribosomal protein L7), A5D12 (=REX-1, zinc finger protein), 88H12 (=cripto, EGF- and heart development-related protein), and 73E10 (=DNA ligase I).

The ESTs of cluster I show high level expression in undifferentiated F9 cells (Table VIII). Their expression in RA-treated F9 cells is dramatically decreased, and is either at relatively low levels or not detectable at all in the five different organs of adult mice (Table VIII). We found, with the MFASTA program, that 94B12, one of the cluster I ESTs, may code for a protein showing weak similarity to a *Drosophila* msh protein (Table IV), and that 78B11 and 86A10, two of the cluster I ESTs, share highly homologous nucleotide sequences with that of 98B12 (Table III). Interestingly, they show almost identical expression patterns to those of 83A02 and 73F03, which are identified ESTs corresponding to the human TCL1 gene, reportedly involved in T-cell malignancy (*16*). From these results, we speculate that these ESTs in cluster I may participate in the maintenance of an undifferentiated state. The genes and functions of these ESTs in cluster I will be addressed in the following study.

ESTs have applications for identifying new genes, mapping of the genome, and identification of coding regions in genomic sequences. Clearly, the information obtained on sequence analysis increases the value of each EST and improves the assignment of genes into functional categories. In this work, we characterized 562+154 ESTs matching no known genes (*6*) by means of computer assisted programs, and demonstrated that some information possibly leading to the elucidation of their functions was obtained from at least 67% of the ESTs examined, *i.e.*, 208 were identified and 256 were classified as the $p<0.01$ group. Subsequently, we tried to characterize unidentified ESTs by examining the expression patterns of the corresponding mRNAs in F9 cells, RA-treated F9 cells, and several organs of adult mice. The strategy we used for characterizing the unidentified ESTs is useful for characterizing ESTs isolated from the mouse and not matching known gene and/or protein sequences.

We wish to thank M. Ohara for the many helpful comments on the manuscript, and M. Miyazaki, T. Sakaura, and K. Sumitomo for their expert technical assistance. We also thank the National Center for Biotechnology Information, USA, for providing access to the network BLAST server, and the Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, for that to the network BLAST and FASTA servers. We also thank the Oak Ridge National Laboratory, Tennessee, USA, for providing access to the network GRAIL server.

REFERENCES

1. Strickland, S. and Mahdavi, V. (1978) The induction of differentiation in teratocarcinoma stem cells by retinoic acid. *Cell* **15**, 393-403
2. Martin, G.R. (1980) Teratocarcinomas and mammalian embryogenesis. *Science* **209**, 768-776

Downloaded from http://jb.oxfordjournals.org/ at Peking University on October 2, 2012

TABLE VIII. **Retinoic acid down-regulated ESTs.**

EST	Putative Identification		F9	RA	Br	Lv	Kd	Hr	Ts[1]
Cluster I									
C68F08*	gb:HSTCL1	H.sapiens TCL1 gene	8527	1104	400	400	400	400	400[2]
C78B11^		unidentified, P<0.01 group	19072	2822	400	400	400	400	400
C94B12^		unidentified	10403	3093	400	400	400	400	400
C86A10		unidentified	16528	3967	400	400	400	400	400
C83A02*	gb:HSTCL1	H.sapiens TCL1 gene	6702	531	400	400	400	400	400
C81E10		unidentified	1805	400	400	400	400	400	400
C69D04^		unidentified, P<0.01 group	2873	488	400	400	400	400	400
C93B12		unidentified, P<0.01 group	10511	3077	400	400	400	400	400
C73F03*	gb:HSTCL1	H.sapiens TCL1 gene	1633	406	400	400	400	400	400
Cluster II									
C67H04		unidentified, P<0.01 group	8959	2129	400	400	3732	4380	740
CA6A11	sp:TGFB_HUMAN	TGF BETA-1 BINDING Protein	3049	400	897	400	2761	3667	1086
C90E05	sp:p19216	EF-TS	2866	400	425	400	1385	400	609
C69H03		unidentified, P<0.01 group	3017	882	626	413	958	400	1388
Cluster III									
C74G12	gb:RATRIP	Rieske iron-sulfur protein	1859	516	975	2177	808	825	963
C84D07	gb:MUSLASSB	Autoantigen La (SS-B)	6470	400	608	3960	400	400	448
CA4E06	gb:MMPH34MRA	pH 34, down-regul. in EC	8742	2651	400	595	400	400	400
CA4D09		unidentified, P<0.01 group	6484	2261	400	841	400	497	400
C71G03	gb:HUMSUIISO	Suiliso1, TI factor	11724	3878	1310	3155	743	1671	487
Cluster IV									
CA2F06	gb:H33231	similar to G10 protein	8799	2049	3765	2864	1039	914	6253
C69E11	gb:HUMCSF	Cleavage stimulation factor	8847	1125	2677	4023	1639	1165	5841
C84H09	gb:MMCCTEP	Ccte CCT	6606	1747	1499	3236	2436	1028	3728
CA2D08		unidentified, P<0.01 group	3293	1118	635	1420	1147	487	3432
C69G11	gb:MMSRP54	Docking protein, SRP54	5165	1767	2248	4316	2120	1534	3008
CA1E09	WP:F45E12.4	CE02740	3471	403	2450	3837	2234	931	4170
Cluster V									
C70A11	gbu:MMU42190	G/T-mismatch binding protei	4500	1107	400	523	400	400	785
CA5D12	gb:MUSREX1	REX-1, zinc finger protein	8028	1543	400	703	400	400	1420
C99A09	sp:RPC9_YEAST	DNA-directed RNA polymerase	3634	1083	681	1705	1082	830	4701
C94F05		unidentified, P<0.01 group	15941	4183	400	1902	476	715	3006
C96B04	gb:RATPROC2A	Proteasome C2 component	6659	2056	400	3573	777	1556	2865
CA5E09	gb:H35836	Ribosomal protein L7	9181	3166	586	3738	899	2406	4103
CA2G05		unidentified, dbEST group	5558	1735	400	1320	400	1562	5546
C93D03		unidentified, P<0.01 group	4014	470	574	470	798	1280	1534
CA6D03	gb:HSDNALIG3	DNA ligase III	15179	1105	434	400	1394	992	7293
C66C06		unidentified, dbEST group	2070	648	400	432	400	400	677
C73E10	gb:MMU04674	DNA ligase I	6393	2030	400	400	400	400	1368
C74D10	gb:RRU05341	Rat p55CDC mRNA, comple	9583	2662	689	651	416	478	5066
CA6D10	gb:MUSENDPRO	Mouse endogenous provirus	6635	2036	400	400	400	525	2564
C88H12	gb:MUSCRIPTO	Cripto	4449	400	485	400	400	400	602

[1]RNAs were prepared from the following sources: F9, uninduced F9 cells; RA, F9 cells treated with RA for 72 h; Br, adult mouse brain; Lv, liver; Kd, kidney; Hr, heart; Ts, testes. [2]The numbers indicate densitometer readings (see the text). Since the densitometer readings were proportional to the concentration of RNA between around 400 and 25,000, estimates lower than 400 are represented by 400. The ESTs denoted by * and ^ are two different groups of redundant ESTs (see Table III and the text). All other abbreviations were as described in the footnotes to Table IIA.

3. Gudas, L.J. (1991) Retinoic acid and teratocarcinoma stem cells. *Semin. Dev. Biol.* **2**, 171-179
4. Capecchi, M.R. (1989) Altering the genome by homologous recombination. *Science* **244**, 1288-1292
5. Nishiguchi, S., Joh, T., Horie, K., Zou, Z., Yasunaga, T., and Shimada, K. (1994) A survey of genes expressed in undifferentiated mouse embryonal carcinoma F9 cells: Characterization of low-abundance mRNAs. *J. Biochem.* **116**, 128-139
6. Nishiguchi, S., Sakuma, R., Nomura, M., Zou, Z., Jearanaisilavong, J., Joh, T., Yasunaga, T., and Shimada, K. (1996) A catalogue of genes in mouse embryonal carcinoma F9 cells identified with expressed sequence tags. *J. Biochem.* **119**, 749-767
7. Nomura, M., Takihara, Y., and Shimada, K. (1994) Isolation and characterization of retinoic acid-inducible cDNA clones in F9 cells: One of the early inducible clones encodes a novel protein sharing several highly homologous regions with a *Drosophila Polyhomeotic* protein. *Differentiation* **57**, 39-50

Downloaded from http://jb.oxfordjournals.org/ at Peking University on October 2, 2012

8. Nomura, M., Takihara, Y., Yasunaga, T., and Shimada, K. (1994) One of the retinoic acid-inducible cDNA clones in mouse embryonal carcinoma F9 cells encodes a novel isoenzyme of fructose 1,6-bisphosphatase. *FEBS Lett.* **348**, 201-205
9. Zou, Z., Nomura, M., Takihara, Y., Yasunaga, T., and Shimada, K. (1996) Isolation and characterization of retinoic acid-inducible cDNA clones in F9 cells: A novel cDNA family encodes cell surface proteins sharing partial homology with MHC class I molecules. *J. Biochem.* **119**, 319-328
10. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410
11. Pearson, R.P. and Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85**, 2444-2448
12. Bairoch, A. (1994) PROSITE documentation file, release 13.0 of November 1995
13. Uberbacher, E. and Mural, R. (1991) Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. USA* **88**, 11261-11265
14. Andersberg, M.R. (1973) *Cluster Analysis for Applications*, Academic Press, New York
15. Adams, M.D., Kerlavage, A.R., Fleischmann, R.D., *et al.* (1995) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377** (suppl.), 3-17
16. Virgilio, L., Narducci, M.G., Isobe, M., Billips, L.G., Cooper, M.D., Croce, C.M., and Russo, G. (1994) Identification of the TCL1 gene involved in T-cell malignancies. *Proc. Natl. Acad. Sci. USA* **91**, 12530-12534

Downloaded from http://jb.oxfordjournals.org/ at Peking University on October 2, 2012