BodyMap: A Collection of 3′ ESTs for Analysis of Human Gene Expression Information

Shoko Kawamoto,[1] Junji Yoshii,[2] Katsuya Mizuno,[2] Kouichi Ito,[1] Yasuhide Miyamoto,[1] Tadashi Ohnishi,[1] Ryo Matoba,[1] Naohiro Hori,[1] Yuhiko Matsumoto,[1] Toshiyuki Okumura,[1] Yuko Nakao,[1] Hisae Yoshii,[1] Junko Arimoto,[1] Hiroko Ohashi,[1] Hiroko Nakanishi,[1] Ikko Ohno,[1] Jun Hashimoto,[1] Kota Shimizu,[1] Kazuhisa Maeda,[1] Hiroshi Kuriyama,[1] Koji Nishida,[1] Akiyo Shimizu-Matsumoto,[1] Wakako Adachi,[1] Reiko Ito,[1] Satoshi Kawasaki,[1] K.S. Chae,[1] Katsuji Murakawa,[1] Masahiro Yokoyama,[1] Atsushi Fukushima,[1] Teruyoshi Hishiki,[1] Akihiko Nakaya,[3] Jun Sese,[3] Norikazu Monma,[3] Hitoshi Nikaido,[3] Shinichi Morishita,[3] Kenichi Matsubara,[4] and Kousaku Okubo[5]

[1]Institute for Molecular and Cellular Biology, Osaka University, Osaka 565–0871, Japan; [2]Hitachi Software Engineering Co., Ltd., Yokohama 231–0015, Japan; [3]Department of Genome Knowledge Discovery System, Institute of Medical Science, University of Tokyo, Tokyo 108–8639, Japan; [4]Internal Institute for Advanced Study, Kyoto 619–0225, Japan

BodyMap is a collection of site-directed 3′ expressed sequence tags (ESTs) (gene signatures, GSs) that contains the transcript compositions of various human tissues and was the first systematic effort to acquire gene expression data. For the construction of BodyMap, cDNA libraries were made, preserving abundance information and histologic resolutions of tissue mRNAs. By sequencing 164,000 randomly selected clones, 88,587 GSs that represent chromosomally coded transcripts have been collected from 51 human organs and tissues. They were clustered into 18,722 independent 3′ termini from transcripts, and more than 3000 of these were not found among ESTs assembled in UniGene (Build 75). Assessment of the prevalence of polyadenylation signals and comparison with GenBank cDNAs indicated that there was no significant contamination by internally primed cDNAs or genomic fragments but that there was a relatively high incidence (12%) of alternative polyadenylation sites. We evaluated the sensitivity and resolution of expression information in BodyMap by in silico Northern hybridization and selection of tissue-specific gene probes. BodyMap is a unique resource for estimation of the absolute abundance of transcripts and selection of gene probes for efficient hybridization-based gene expression profiling. [BodyMap data are available at http://bodymap.ims.u-tokyo.ac.jp.]

In the early phase of its development, the expressed sequence tag (EST) collection (Adams et al. 1993, 1995) primarily served as a catalog to be screened for clones of interest by sequence homology. In the next phase, gene coverage was pursued (Aaronson et al. 1996; Williamson 1999) by using normalized libraries and/or highly complex sources (Soares et al. 1994; Hillier et al. 1996) to use the entries as markers to create a transcript map of the human genome, after clustering redundantly accumulated ESTs into gene units (Schuler et al. 1996). As genome sequencing efforts progress, ESTs have been used for exon identification (Dunham et al. 1999; Hattori et al. 2000), and they are being mapped and organized in the framework of genome sequence at a resolution of single nucleotides. Progress in the integration of ESTs into the genomic sequence will make EST data more of an expression of gene records rather than merely a pool of nucleotide sequences. Reflecting this trend, the major EST collection projects have shifted emphasis from efficiency of identifying novel sequences to meaningful source selection, such as coverage of a majority of cancer types (Strausberg et al. 2000).

BodyMap is a collection of site-directed 3′ ESTs (gene signatures, GSs) designed as an anatomical database of human gene expression in which sequences are used as identifiers (Okubo et al. 1992). Construction of BodyMap began in 1991 (Okubo et al. 1991) and representative human tissues and organs have been incorporated. During the collection of GSs, nonstructural information about the mRNA, including transcript abundance and anatomical distribution, was pre-

[5]Corresponding author.
E-MAIL kousaku@imcb.osaka-u.ac.jp; FAX 81-6-6877-1922.
Article and publication are at www.genome.org/cgi/doi/10.1101/gr.151500.

served. The libraries were constructed from well-characterized sources by using methods that minimize the differences in cloning efficiencies among transcripts, and libraries were never amplified before sequencing (Okubo et al. 1991). Accordingly, BodyMap has characteristics distinct from those of other public EST data sets, which were generated as sequence collections at the expense of expression information (Bonaldo et al. 1996). BodyMap has been used in the isolation and characterization of tissue-specific transcripts (Nishida et al 1996; Ohno et al. 1996; Maeda et al. 1997; Shimizu-Matsumoto et al. 1997) and in disease gene identification (Irvine et al. 1997; Nishida et al. 1997). Here we describe the structure and features of 88,587 GSs from human tissues collected in BodyMap.

RESULTS

Sources and Library Construction

The numbers of informative 3′ site-directed ESTs representing chromosomally coded genes are summarized in Table 1. We refer to these 3′ ESTs, covering restricted 3′ ends in the sense direction, as gene signatures (GSs) (Okubo et al. 1992). Sources were selected to cover the most representative tissues and cell types. Emphasis was placed on pure connective tissues and epithelial cells, which are underrepresented in dbEST. In every case, tissue preparation was performed carefully, sometimes by microscopy, to minimize contamination by other cell types. For example, human epithelial cells were prepared by careful isolation of a monolayer or layers of cells free from visible contamination by connective tissues and blood cells (Ohnishi et al. 1999). As a result, for example, the sequence of the immunoglobulin λ chain transcript, which was found in 1% (11/870) of clones from colonic mucosa having a thin lining of loose connective tissue (lamina propria), was not identified in 20,440 clones from purified epithelium. Because of the elaborate manipulation steps, libraries were sometimes constructed by direct priming of less than a microgram of total RNA. Nevertheless, contamination by ribosomal RNA was very low (0.26%), probably because of the high specificity of first-strand synthesis with a low concentration vector primer.

Validation of Collected GSs

Collected GS sequences were evaluated if they represented true mRNA termini. Of 3928 independent GS sequences that matched GenBank entries, 3470 (88%) represented the most 3′ *Mbo*I fragments of the deposited cDNA sequences. The rest represented alternatively polyadenylated mRNAs or internally primed artifacts that cannot be discriminated by sequence inspection of individual cases. Thus, the presence of the poly(A) addition signals upstream of the addition site was used for the validation as mass data. Canonical signal (AATAAA) and sequences with single-base substitutions were examined 10 bp to 50 bp from the poly(A) tail in 4431 independent GS sequences (Fig. 1A).1 In 93% of GS sequences, AATAAA or a single base variant was found. The prevalence of AATAAA and single base variants was quite similar to that observed in 1123 GenBank human cDNAs with clear annotations of poly(A) sites. In the case of 3′ ESTs deposited in dbEST, the proportions of signals differed greatly between those starting with a stretch of Ts and those without them (Fig. 1A). The former has very similar signal occurrence, but in the latter the proportion of AATAAA is greatly reduced, indicating that the short region following a poly(T) stretch was trimmed before data submission as reported (Hillier et al. 1996). The 458 BodyMap GS that matched internal regions of GenBank mRNAs had similar frequencies of hexanucleotide signals, suggesting that the majority of them are also polyadenylated in vivo. Because we counted not only the well-known single-base variants, such as ATTAAA, but also all of the possible single-base substitutions, the fraction of each of these variants was compared for cDNA ends having only one candidate signal (Fig. 1B). The proportion was consistent across GenBank primate sequences, BodyMap, and untrimmed 3′ ESTs. Trimmed 3′ ESTs served as nonterminal controls. This agreement between the three sets of data suggests that some of the uncommon hexanucleotide variants, such as BATAAA and AATABA (B = T, G, C), are functional and that most of the cDNA sequences with only single-base variants in either set of data represent true 3′ termini of transcripts.

Constitution of GS Population

Without exception, the most recurrent GSs in differentiated cells or in adult tissues were from nonhousekeeping genes (Table 1). Some were unique to each tissue, and some were shared among cells of the same lineage. The fraction of the most abundant GS varied more than tenfold across tissues or cell types. There were six tissues in which more than 10% of the total ESTs were attributable to a single GS cluster (Fig. 2). They were secretory epithelia or muscular tissues. In the remaining tissues, the content varied by a small percent (mean, 2.5%; SD, 1.8%).

The characteristics of the GS population for each source group—nervous tissues, connective tissues, and epithelial tissues—are illustrated in the accumulated frequency curve (Fig. 3) in which the cumulative sums of occurrences were plotted in descending order of GS occurrence. The epithelial and connective tissues have very similar curves, whereas that of the nervous system is clearly shifted downward. The curve for neural tis-

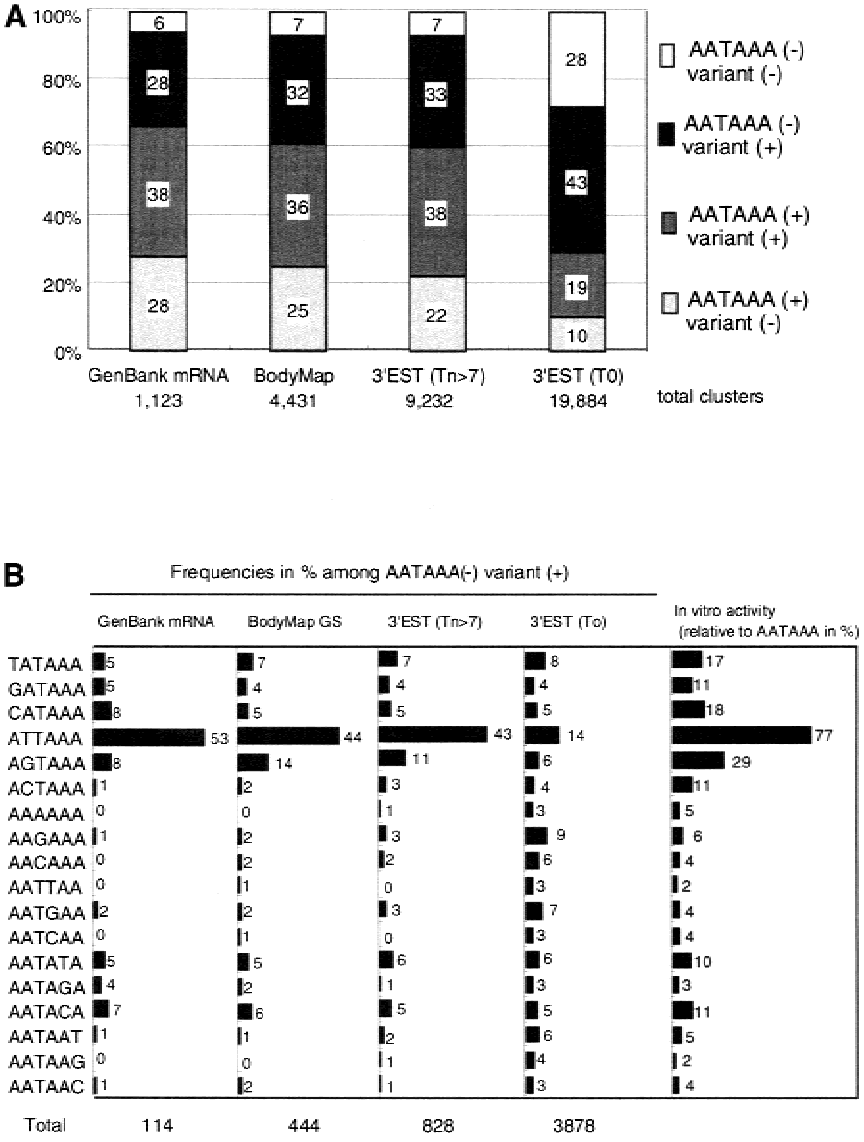

Figure 1 Distribution of AATAAA and single-base variants in 3′ ends of ESTs. (*A*) The hexanucleotide signals from the 3′ ends of four sets of cDNA sequences. The regions 10–50 bp from the polyadenylation sites were examined for the presence of AATAAA or its single-base variants. The 3′ ESTs from dbEST were divided into those starting with more than seven Ts (Tn > 7) and those without a T in the first position (T_0). (*B*) The prevalence of each single-base variant in the cDNA termini with only one variant signal and no AATAAA within the 10–50-bp region from the poly(A) tail in each of four data sets is shown. The frequencies (%) of all 18 possible variants in each data set are shown beside bars. In vitro polyadenylation activities for each variant measured in the context of the SV40 polyadenylation signal are reproduced from the literature (Wickens et al. 1984; Sheets et al. 1990)

sues did not overlap with the others at a credit level of 0.85 in the top 486 genes. As seen in Figure 2, 50% of the mass is accounted for by ~500 genes in connective and epithelial tissues but by >900 genes in nervous tissue.

Overlapping with dbEST

To further characterize BodyMap data, we compared them with dbEST entries in UniGene (Build 75) that were clustered into 72,831 physical and annotational clusters. Of 18,722 GS clusters composed of 89,831 GS tags in BodyMap, 3,382 GS clusters did not match ESTs listed.

The GS in overlapping fraction have an average redundancy of 5.6 in BodyMap, whereas it was 1.3 in the GS cluster unique to BodyMap. GS unique to BodyMap were distributed at frequencies of 1%–5% in every library (mean, 4.0%; SD, 2.2%) and had hexanucleotide signal occurrences similar to the rest (data not shown). Nervous system tissues had a slightly greater content of unique GS (mean, 5.4%; SD, 2.1%) than was found in other tissues (mean, 3.8%; SD, 1.8%). These values equaled or exceeded 10% in only two libraries: full thickness of skin (10%) and fetal neuron (11%).

In Silico RNA Experiments

The primary goal for the construction of BodyMap was to create a genes × tissues matrix of transcription level that could be used for in silico experiments such as Northern hybridization and subtraction cloning. Although the depth of the clone collection limits the sensitivity and specificity of experiments, for abundant clones these primary objectives have been achieved. The sensitivity of the present matrix was assessed by probing the data with several genes known to have moderate expression levels and known tissue specificities. As shown in Table 2, the distributions of cytoskeletal intermediate filaments and collagens suggest that this clear segregation is applicable also to anonymous genes with similar expression levels. Such pure segregation patterns are not seen in libraries constructed from complex starting materials.

Another example of an in silico experiment is selection of genes with given patterns of expression. For example, genes differentially expressed in myeloid cells, based on the criteria that frequency variation between myeloid cells and nonmyeloid cells was highly significant ($P < .005$), are shown in Table 3. By increas-

Table 1. The Most Abundant Transcripts in Human Tissues

B01.hl60	857		C08.Aortic media	1002		N01. Retina	877		X01. hepG2	740	
Ribosomal protein S8	X67247	24	Elastin	M17282	155	Opsin	K02281	14	EF-1α	X16869	17
Ribosomal protein L9	U09953	22	Osteonectin	J03040	27	Na/K ATPase β2	D87330	13	Albumin	L00133	17
Ribosomal protein L23	X53777	17	Ribosomal protein L21	X89401	16	Aldolase C	X07292	10	TPT-1	X16064	9
Ribosomal protein L7a	X52138	16	GS13325	None	16	Ribosomal protein L9	U09953	9	Ribosomal protein L31	X69181	9
B02. hl60/ DMSO	**1081**		**C09. Ventricle muscle**	**3785**		**N02. Cortex**	**2242**		**X02. Neonate liver**	**739**	
β-actin	X00351	14	Myosin heavy chain	M25139	101	Myelin basic protein	M13577	49	Albumin	L00133	227
HHCPA78 homolog	S73591	6	Ig-λ light chain	D01059	75	hng/RC3	Y15059	14	Apolipo-protein B	J02775	38
Ribosomal protein L3	M90054	5	Myoglobin	X00373	60	Apolipo-protein J	M74816	13	α2-HS-glyco-protein	M16961	21
L-plastin	L05492	5	Troponin C, skel/card	M37984	53	Aldolase C	X07292	9	Haptoglobin α 1S	X00637	16
B03. hl60/TPA	**889**		**C10. atrial muscle**	**2823**		**N03. cerebellum**	**1107**		**X03. fetal liver**	**641**	
EF-1α	X16869	26	ANF	M54951	203	GFAP	S40719	22	Albumin	L00133	109
Methionine AT-a	L43509	14	Actin, a-cardiac	J00073	88	Aldolase C	X07292	7	Haptoglobin α 1S	X00637	27
Ferritin L	M11147	14	α B-crystallin	S45630	27	Myelin basic protein	M13577	5	γ-G globin	X55656	17
TPT-1	X16064	13	Troponin T, cardiac	X74819	25	Apolipoprotein J	M74816	5	Apolipo-protein All	X04898	14
B04. granulo-cyte	**1164**		**C11. Skeletal muscle**	**4527**		**N04. Neuroblast**	**1235**		**X04. Adult liver**	**956**	
β-2-micro-globulin	M17987	25	α-actin, skeletal	M20543	301	EF-1α	X16869	19	Albumin	L00133	279
Spermidine/ spermineAT	M77693	22	Myosin heavy chain	X03741	173	H3.3 histone	M11354	17	Haptoglobin α 1S	X00637	41
HLA-Cw1	M26429	21	Myosin heavy chain	X03740	137	ribosomal protein L9	U09953	16	α-1 acid gp	M13692	20
Pre-B enhancing factor	U02020	20	Troponin C, skeletal	X07898	121	TPT-1	X16064	15	Apolipo-protein B	J02775	19
B05. CD8 T cell	**1104**		**C12. Hair follicle**	**2164**		**N05. Caudate nucl.**	**1077**		**X05. Lung**	**874**	
β-2-micro-globulin	M17987	16	Fibronectin	K00799	84	hng/RC3	Y15059	12	Pulmonary SAP	M30838	87
TPT-1	X16064	12	COL1A1	M32798	57	TALLA-1	D29808	11	Clara cells 10 kd prot.	U01101	31
EF-1a	X16869	10	EF-1α	X16869	43	Myelin basic protein	M13577	8	HLA-E heavy chain	X64881	12
Yeast rp L4 homolog	Z12962	10	Osteonectin	J03040	30	KIAA0607	AB011179	7	Fibronectin	K00799	10
B06. CD4 T cell	**1028**		**E1. Keratino-cyte**	**820**		**N06. Thalamus**	**912**		**X06. Colon mucosa**	**921**	
β-2-micro-globulin	M17987	19	Cytokeratin 14	J00124	15	Myelin basic protein	M13577	36	L-FABP	M10617	40
Ribosomal protein L11	X79234	14	Metallothio-nein	V00594	10	GFAP	S40719	8	Galectin-4	AF01483	18
TPT-1	X16064	13	Lipocortin II	D00017	9	apo J	M74816	7	CLCA1	AF03940	13
23 kD highly basic protein	X56932	12	Ribosomal protein S19	M81757	8	Sox 8	AF164104	6	Ig-λ-light chain	D01059	12

Table 1. (*Continued*)

C01. Adipose tissue	**1488**		**E02. Cornea**	**2793**		**N07. Putamen**	**871**		**X07. Small cell ca. lung**	**843**	
Gelatin BP	AB012165	19	Apolipo-protein J	M74816	73	Myelin basic protein	M13577	8	BBC1	X64707	8
Ribosomal protein S8	X67247	16	Cytokeratin 12	D78367	55	GS04506	None	8	23 kD highly basic prot.	X56932	7
apM2	D45370	15	apM2	NM006829	40	TPT-1	X16064	7	Ribosomal protein S11	X06617	7
TPT-1	X16064	14	Ferritin H	M11146	31	Na/K ATPase β2	D87330	7	Ribosomal protein L7a	X52138	6
C02. Aortic endothel*	**967**		**E03. Conjunc-tiva**	**937**		**N08. Astrocyte***	**1103**		**X08. Adeno ca. of lung**	**1183**	
Fibronectin	K00799	36	β-2-micro-globulin	M17987	23	EF-1α	X16869	26	COL3A1	X14420	20
TPT-1	X16064	14	Cytokeratin 13	X52426	23	Ribosomal protein S17	M13932	14	Thymosin β4	M17733	15
PAI-1	X13345	12	Lipocortin	X05908	8	GFAP	S40719	12	Ig-λ light chain	D01059	14
CTGF	X78947	12	EF-4All	D30655	8	Thymosin β4	M17733	10	Ig-κ light chain	M11937	13
C03. Osteo-blast*	**928**		**E04. Intest. Metaplasia**	**2192**		**N09. Schwann cell***	**975**		**X09. Squamous cell ca. lung**	**1190**	
COL1A2	J03464	25	Calcyclin	J02763	25	Ribosomal protein L10	AB007170	18	Calcyclin	J02763	23
Fibronectin	K00799	22	EF-1α	X16869	20	Ribosomal protein L9	U09953	17	Ferritin L chain	M11147	14
Osteonectin	J03040	20	Amino-peptidase N	M22324	17	Ribosomal protein S19	M81757	17	Cystatin B	L03558	11
COL3A1	X14420	18	PSCA	AF043489	17	Ribosomal protein S29	U14973	15	Cathepsin B	L16510	9
C04. Fibro-blast*	**1097**		**E05. Fundic gland**	**3304**		**N10. Fetal neuron***	**1108**		**X10. Iris**	**3314**	
Stromelysin	X05232	26	Pepsinogen	J00287	572	Ribosomal protein L37a	X66699	9	Apolipo-protein D	M16696	45
Fibronectin	K00799	22	Lysozyme	X14008	33	TPT-1	X16064	9	1-8D	X57351	43
Collagenase	X05231	15	Gastric lipase	X05997	32	Ribosomal protein L5	U14966	8	Yeast rp L41 homolog	Z12962	42
PAI-1	X13345	14	TPT-1	X16064	29	Thymosin β4	M17733	8	TPT-1	X16064	40
C05. Mesangium	**1101**		**E06. Ileum epithel**	**3675**		**N11. Corpus callosum**	**949**		**X11. Skin full thickness**	**4604**	
Fibronectin	K00799	48	β-2-micro-globulin	M17987	59	GFAP	S40719	12	Yeast rp L41 homolog	Z12962	59
Calcyclin	J02763	30	CLCA1	AF039400	51	Ribosomal protein L37a	X66699	7	GS20959	None	58
ribosomal protein S19	M81757	13	defensin 6	M98331	46	ribosomal protein S8	X67247	7	ribosomal protein S18	X69150	57

Table 1. (*Continued*)

Yeast rp S28 homolog	D14530	10	GS2706	None	34	Myelin basic protein	M13577	7	Delta-6 desaturase	AF03679	57
C06. Itoh cell	**1283**		**E07. Colon epithel**	**6451**		**N12. Substantia nigra**	**3477**		**X12. Tumor infiltrates**	**1585**	
Osteonectin	J03040	23	Galectin-4	AF014838	106	Myelin basic protein	M13577	129	β-2-micro-globulin	M17987	49
Fibronectin	K00799	18	cytokeratin 8	X12882	86	Ribosomal protein L7a	X52138	51	LD78α	D90144	31
PAI-1	X13345	16	L-FABP	M10617	78	EF-1α	X16869	48	RF-1α	X16869	25
EF-1α	X16869	16	Calcyclin	J02763	75	αB-crystallin	S45630	46	Yeast rp S28 homolog	D14530	24
C07. Bone flakes	**1042**		**E08. Pituitary**	**1015**		**N13. Fetal brain**	**3797**				
Osteonectin	J03040	25	Prolactin	M29386	181	α-Tubulin	X01703	35			
COL1A2	J03464	24	Growth hormone	M13438	20	Ribosomal protein L37a	X66699	21			
β-Globin	V00497	13	Secretogranin I	Y00064	12	EF-1α	X16869	20			
COL3A1	X14420	12	sGTP-bp	X07036	9	Stathmin	J04991	16			

The source tissue or cells and the number of total ESTs representing chromosomally encoded genes are given (shaded cells). The source groups were blood cells (B01–B06), connective and muscular tissues (C01–C12), epithelial tissues (E01–E08) and nervous tissues (N01–N13). When the source tissue was composed of multiple cell types or an uncategorizable cell type, it was categorized as complex (X01–X12). Asterisks denote primary cultured cells. The identities of the most frequently isolated tags are given along with their frequencies. (TPT-1) Translationally controlled tumor protein; (methionine AT-a) methionine adenosyltransferase-α; (spermine AT) spermine acetyltransferase; (EF-1a) elongation factor 1-α; (apM2) adipose most abundant protein-2; (PAI-1) plasminogen activator inhibitor-1; (ANF) atrial atriuretic factor; (EF-4AII) elongation factor 4AII; (CLCA1) calcium activated chloride channel 1; (L-FABP) liver fatty acid binding protein; (sGTP-BP) stimulatory GTP bonding protein; (hng/RC3) human neurogranin; (GFAP) glial fibrillary acidic protein; (TALLA-1) T-cell acute lymphoblastic leukemia associated antigen 1; (pulmonary SAP) pulmonary surfactant apoprotein.

ing the *P* value to 1%, 112 more genes were selected (data not shown).

DISCUSSION

The wide coverage of human genes in dbEST permits parallel gene expression monitoring based on prior knowledge of gene sequence (Lockhart et al. 1996; Iyer et al. 1999). However, from a practical perspective, researchers must select a set of genes suitable for target tissues to make testing efficient (Loftus et al. 1999). The well-preserved abundance information and high anatomical resolution make BodyMap a preferable source for probe selection (http://bodymap.ims.u-tokyo.ac.jp). Another unique feature of BodyMap is the absolute abundance values for transcripts for various tissues. Such information is also found in shorter tag collection, SAGE (Velculescu et al. 1995; Welle et al. 1999), and the tissue-coverage complement to each other. The abundance data covering various tissues are complementary to relative gene expression comparison (DeRisi 1996; Schena et al. 1996; Kawamoto et al. 1999) for evaluating the functions of uncharacterized genes.

Site-directed EST sequences are indispensable for identification of gene ends within genomic sequences because even the most sophisticated computer programs tend to overpredict the presence of exons (Dunham et al. 1999). The overlap of dbESTs with BodyMap indicates that there are still more transcripts to be identified in brain and other tissues. The higher complexity of transcripts in brain, as shown by the accumulated frequency curve, supports this idea. Possible overprediction of genes by using 3′ ESTs is due to cloning artifacts and alternative polyadenylation. Validation of 3′ ESTs by using hexanucleotide signals suggested that such artifacts were negligible in our data set. The 3′ ends without the AATAAA were observed at high incidences not only in BodyMap (39%), but also in human cDNAs in GenBank (37%) and qualified 3′ ESTs from dbEST (40%). In those 3′ ends, several uncommon single-base variants, such BATAAA and AATABA (B = T, G, C), plausibly responsible for poly(A) formation in these 3′ ends, were found at very similar rates. After this paper was submitted, Beaudoing et al. (2000) published similar results from an analysis of 4344 human 3′ untranslated regions (UTRs) and 3′ ESTs overlapping with them. The proportion of 3′ ends without AATAAA was 41.8% in their analysis, and uncommon single-base variants were found at significant frequencies among them. In BodyMap, upstream alternative polyadenylation was found in 12% of GenBank mRNA entries. Assuming the same incidence of downstream alternatives, our estimate of alternative polyadenylation

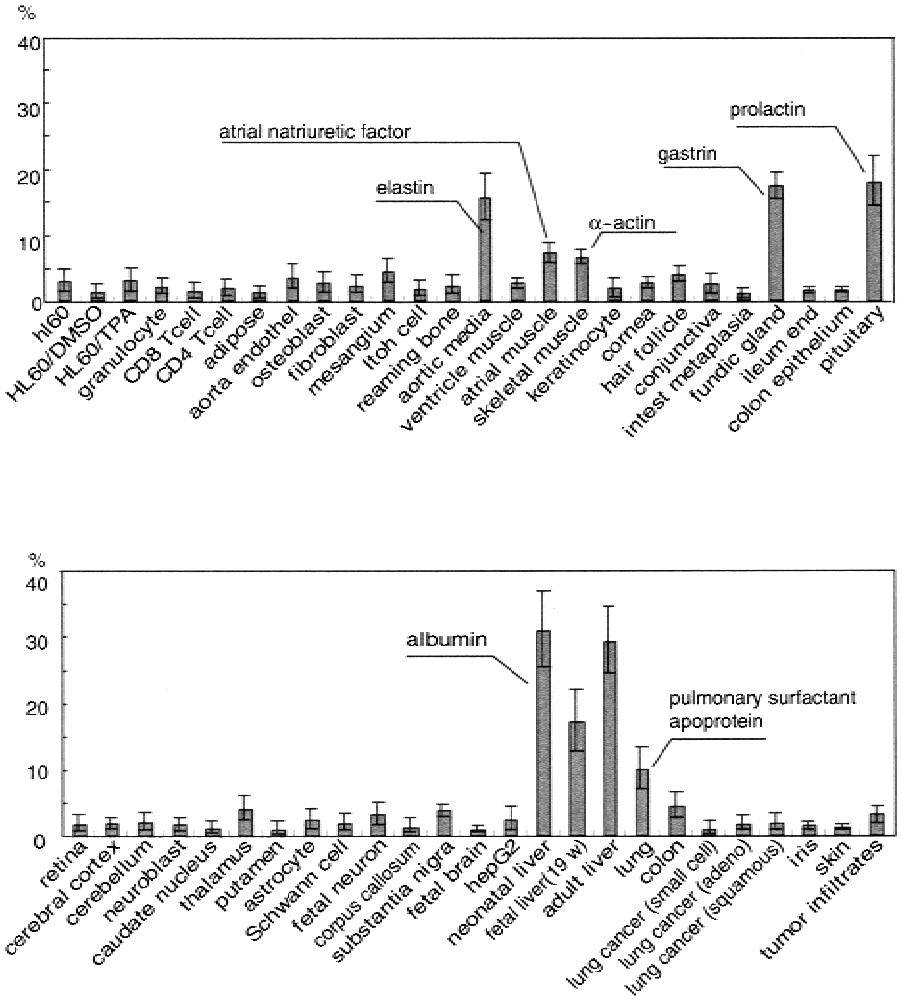

Figure 2 The relative contents of the most abundant transcripts in 51 human tissues or cell types as measured by gene signature collection. The error range indicates the *P* value of 0.1 calculated for each observed occurrence. The identities of some transcripts are given. For the identities of other transcripts, see Table 1.

is 24%, close to the reported estimates by EST clustering (16%) (Gautheret et al.) and recent 3′ UTR analysis (28.6%). Although generation of multiple 3′ ESTs from one gene may affect transcript counting by EST clustering, assigning them to genomic sequences will easily resolve this problem as long as the ESTs are not far apart.

In summary, our site-directed 3′ ESTs can serve as a resource for selection of probes for sequence-based expression profiling methods and can provide absolute levels of gene expression that are important in considering gene function. Our collection covers various rare tissues and provides information on their mRNA populations. To allow full use of BodyMap for in silico mRNA experiments, the representation frequency matrix of gene × sources and all representative sequences have been made available through our ftp site (http://bodymap.ims.u-tokyo.ac.jp/datasets/index.html).

METHODS

Library Construction

Of 51 human libraries listed (Table 1), 15 were made by direct priming of total RNA. The specimens used for RNA preparations and the methods used are described elsewhere (http://bodymap.ims.u-tokyo.ac.jp). The other libraries were made from poly(A)-selected RNA. For counting transcripts by sequencing, only the most 3′-terminal fragment left by cleaving off the bulk of the fragment with *Mbo*I from the pUC119 vector-primed cDNA was cloned, as described previously (Matsubara et al. 1993). This shortening of the inserts facilitates the unbiased representation of mRNA regardless of their original sizes at the expense of losing ~5% of gene sequences due to the absence of *Mbo*I site or its location too close to the poly(A) tail.

Data Collection and Cleansing

Starting with randomly isolated transformants, sequence templates were prepared by PCR amplification of the insert cDNA in single-stranded phage released into the culture medium. All sequences were read from the *Mbo*I site toward poly(A), which allows unambiguous identification of the original transcripts. They were referred to as GSs (Okubo et al. 1992). In half of the cases, dye primer chemistry was used, and in the remaining cases, DYEnamic ET* Terminator Cycle Sequencing Kit (Amersham Pharmacia Biotech Inc.) was used. Sequences with >5% Ns, not starting with GATC (the *Mbo*I site), or having more

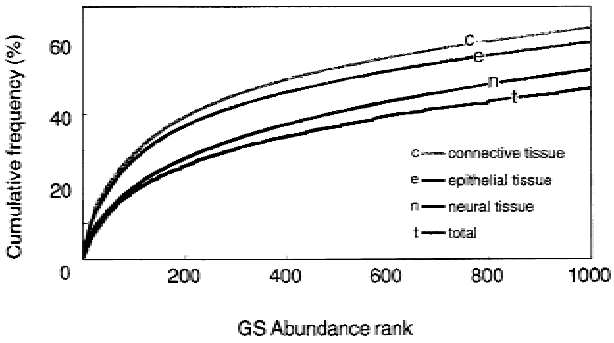

Figure 3 Cumulative frequencies of gene signature (GS) sequences. The cumulative sums calculated in descending order of GS frequencies are plotted as a percentage of total tag occurrence. Tag occurrences in each of three major tissue categories were plotted separately.

Table 2. Distribution of Transcripts for Cytoskeletal Filaments and Collagens across Libraries

		Blood						Connective/muscle												Epithelial								Neural											Cluster ID	
Gene name	GenBank	B01	B02	B03	B04	B05	B06	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12	E01	E02	E03	E04	E05	E06	E07	E08	N01	N02	N03	N04	N05	N06	N07	N08	N09	N10	N11	N12	N13	BodyMap	UniGene
Cytokeratin 4	X07695																				4																			GS08105	Hs. 3235
Cytokeratin 5	M19723																			2	7	2																		GS05829	Hs. 195850
Cytokeratin 6	L42601																			3																				GS05780	Hs. 111758
Cytokeratin 8	X12882							1				2					2				2	1	12	26	29	86														GS00223	Hs. 78271
Cytokeratin 12	D78367																				55																			GS08025	Hs. 66739
Cytokeratin 13	X52426																				2	23																		GS06283	Hs. 74070
Cytokeratin 14	J00124																			15	2																			GS05738	Hs. 117729
Cytokeratin 17	Z19574																			3																				GS05804	Hs. 2785
Cytokeratin 18	X12883																1			1			4	5	5	8														GS00243	Hs. 65114
Cytokeratin 19	Y00503																			1	4		14	13	5	25														GS05013	Hs. 182265
α-Actin, cardiac	J00073															13	88	1																						GS13600	Hs. 7768
α-Actin, skeletal	M20543															18	2	301											1	1								1		GS14552	Hs. 1288
α-Actin, vascular	X13839							3		3	2		8		5	4	5	1	23																					GS03142	Hs. 195851
β-Actin	X00351		14	1	4			2	1	2	11		2	2		1	5	3	22	4	2	2	10	7	12	17		1	6		4					3	1	3	1	8	GS00244	Hs. 180952
γ-Actin, cytoskeletal	M19283	1	5	1				4	1	3	9	9	3	5			2	1	11	2	2		4	9	3	17	3			6	1		2		4	9	8	4	2	9	GS00114	Hs. 215747
γ-Actin, enteric	D00654							1							1				6																					GS03304	Hs. 77443
Neurofilament-L	X05608																											1	6			1	5	1							GS04860	Hs. 222661
Neurofilament-M	Y00067																												3				4	1					2	1	GS05625	Hs. 71346
GFAP	S40719																												8	22		3	8			12		12	7	4	GS06290	Hs. 1447
COL1A1	M32798							1		11	4		2	9	1	1			57															1	1					GS02049	Hs. 172928
COL1A2	J03464									26	13		15	34	22	3	8		1	54																				GS02285	Hs. 179573
COL2A1	X16468													2																										GS12481	Hs. 81343
COL3A1	X14420							4	1	28	5		16	14	12	5	2		11																2					GS03074	Hs. 119571
COL4A1	M26576											1	1	1					2																				1	GS06075	Hs. 119129
COL4A2	J02760																	1																						GS16997	Hs. 75617
COL5A1	M76729									1			3					1	2																				1	GS03963	Hs. 146428
COL5A2	M11718								1		3				1				2	1																				GS03623	Hs. 82985
COL6A1	X15880							2		1									1																					GS03931	Hs. 10885
COL6A2	M34572							3		1						1	2																							GS02756	Hs. 219020
COL6A3	X52022							1		5			2		1				6																2					1	GS03837	Hs. 80988
COL9A2	M95610																																	2					1	GS09520	Hs. 37165
COL9A3	L41162														1																									GS05017	Hs. 53563
COL11A1	J04177																		1																					GS12953	Hs. 82772
COL15A1	L25286							1						1			2	1												3									GS3082	Hs. 83164
COL18A1	L22548													1		2													3							1			1	GS02482	Hs. 78409

The numbers are the observed recurrence of each GS in library grouped by tissue system. For the source materials and number of total isolates, see Table 1.

Table 3. Known Genes Selected as Uniquely Expressed in Myeloid Cells

Cluster ID	GenBank identity	#ACC	*P* value
GS08362	Granulocyte colony-stimulating factor receptor	S71484	2.56E-17
GS00697	Plasminogen activator inhibitor-2 (PAI-2)	M24657	1.13E-09
GS01724	Pleckstrin (P47)	X07743	1.13E-09
GS01024	Leukocyte adhesion protein/CD18	M15395	2.14E-08
GS01345	Bactericidal permeability increasing protein (BPI)	J04739	7.56E-06
GS08325	Phosphatidylinositol 3-kinase p110delta	U86453	7.56E-06
GS01990	Secreted protein (I-309)	M57502	7.56E-06
GS01200	ICB-1 mRNA	AF044896	1.42E-04
GS01719	Myeloid cell nuclear differentiation antigen	M81750	1.42E-04
GS01202	Neutrophil oxidase factor (NCF2)/p67-phox	U00788	1.42E-04
GS00779	Wegener's granulomatosis autoantigen proteinase 3	M97911	1.42E-04
GS01687	c-raf-1 proto-oncogene	L00212	2.68E-03
GS01000	EVI2B3P	M60830	2.68E-03
GS08337	Grancalcin (neutrophil monocyte Ca binder)	M81637	2.68E-03
GS00610	Beige protein homolog (chs)	U67615	2.68E-03
GS01164	Differentiation antigen (CD33)	M23197	2.68E-03
GS08512	Monocytic leukemia zinc finger protein	U47742	2.68E-03
GS01229	Migration inhibitory factor-related protein 8	M21005	2.68E-03
GS01963	Type II interleukin-1 receptor antagonist (IL-1ra3)	AF057168	2.68E-03

Gene signature (GS) clusters were selected by the criteria that have probability of uncontrolled expression between myeloid cells (HL60, HL60/DMSO, HL60/TPA, granulocytes) and the remaining tissues less than 0.5%. *P* values represent the probabilities of each gene with uncontrolled expression between two sets of libraries (see Methods for calculations). Along with these known genes, 28 GSs for novel genes as follows were selected: GS01371, GS01572, GS08424, GS01582, GS01553, GS01965, GS01356, GS00656, GS08595, GS01922, GS08435, GS01123, GS00963, GS01383, GS08551, GS08572, GS01352, GS01109, GS01561, GS08460, GS05157, GS00549, GS01251, GS00627, GS08477, GS08379, GS01458, GS01987. For sequences, refer to http://bodymap.ims.u-tokyo.ac.jp.

than one GATC were eliminated. We then eliminated those sequences having >90% similarity in an overlap longer than 50 bp or 70% of the sequence length with vectors and ribosomal sequences. Sequences for mitochondrial transcripts were also eliminated. When the GATC and poly(A) tail were separated by <17 bp, the sequences were eliminated from the analysis because they were not always unique enough. Lastly, sequences were compared with a library of repetitive sequences, REPBASE (Jurka 1995, ftp://ncbi.nlm.nih.gov/repository/repbase/) by using BLAST (Altschul et al. 1990), and repetitive regions were masked as previously reported (Hishiki et al. 2000). All GS sequences were submitted to the DNA DataBank of Japan (DDBJ) and made available at our web site (http://bodymap.ims.u-tokyo.ac.up).

Transcript Counting/EST Clustering

Sequences from each new library were first compared to each other with FASTA (Pearson et al. 1988). When the similarity exceeded 95% for an overlap longer than 50 bp or 70% of insert length and the overlap started at a GATC, they were considered the same tag and clustered (primary cluster). From each cluster, one representative GS was selected and compared with representative sequences from previously generated clusters. By using the same criteria, clusters of the same GS were grouped, and a new representative tag was selected from the new cluster (secondary cluster). A five-figure cluster ID, referred to as the GS number, was assigned to each independent cluster. Representative sequences for the GS clusters were compared periodically with primate sequences in GenBank (Re. 110.0) and ESTs in UniGene (Build 75, http://www.ncbi.nlm.nih.gov/UniGene/). The criteria for identity were the same as those used for clustering. The correspondence of BodyMap ID (GS) to UniGene ID (Hs) was submitted to GenBank and implemented in UniGene.

Selection of Differentially Expressed Genes

For the selection of genes preferentially expressed in a given set of tissues, for example tissues A, B, and C, libraries A–C were considered one library and the remaining 48 libraries in BodyMap another library. The probability of unregulated expression between the two hypothetical libraries was calculated for each GS by the equation reported by Audic and Claverie (Audic et al. 1997):

$$P(y|x) = \left(\frac{N2}{N1}\right)^{y} \frac{(x+y)!}{x!\,y!\left(1+\frac{N2}{N1}\right)^{(x+y+1)}}$$

Total isolation in A–C is *N1* and isolation of the relevant GS is *x*. The total isolation in the remaining libraries is *N2* and the occurrence of the relevant GS is *y*.

Analysis of Polyadenylation Signals

Among 62,710 entries of primate sequences in GenBank (Re.97), all human mRNAs with a single "poly(A)-site" listed in the features were used. From the representative sequences for all GSs, we selected those that satisfied all of the following conditions. The GS does not have matches in GenBank, is

longer than 100 bp, and ends with poly(A). The GS sequence does not contain more than 5% Ns within 100 bp of the poly(A). The GS does not contain repetitive sequences or an N in the AATAAA sequence, such as 'NATAAA'.

From dbEST (Re. 93), we selected 118,353 3′ ESTs from the Washington-U/Merck project (Hillier et al. 1996) to avoid confusion due to inconsistencies in the feature descriptions from different laboratories. EST matches to BodyMap entries and GenBank primate mRNAs were eliminated first. Those ESTs with discrepancies between clone name and definition (5′ in clone name and 3′ EST in definition), and those denoted as "possible reverse clone" were also eliminated. 3′ ESTs with a stretch of longer than seven Ts (Tn > 7) at the beginning and those starting with A, G, or C (T_0) were analyzed separately. Those ESTs starting with one to seven Ts were not used. Within each of these four categories, the 100 bases from the poly(A) site were compared with each other with BLAST N with the same criteria used for GS clustering, and the fragment containing the lowest number of Ns was selected from each cluster and used in the analysis.

ACKNOWLEDGMENTS

The authors thank Ms. Kumiko Takagi for her secretarial assistance. This work was supported in part by Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Science and Culture, and Research for the future of Japan Society for the Promotion of Science, Japan.

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

REFERENCES

Aaronson, J.S., Eckman, B., Blevins, R.A., Borkowski, J.A., Myerson, J., Imran, S., and Elliston, K.O. 1996. Toward the development of a gene index to the human genome: An assessment of the nature of high-throughput EST sequence data. *Genome Res.* **6:** 829–845.

Adams, M.D., Kerlavage, A.R., Fields, C., and Venter, J.C. 1993. 3,400 new expressed sequence tags identify diversity of transcripts in human brain. *Nat. Genet.* **4:** 256–267.

Adams, M.D., Kerlavage, A.R., Fleischmann, R.D., Fuldner, R.A., Bult, C.J., Lee, N.H., Kirkness, E.F., Weinstock, K.G., Gocayne, J.D., White, O., et al. 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377:** 3–174.

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.

Audic, S. and Claverie, J.-M. 1997. The significance of digital gene expression profiles. *Genome Res.* **7:** 986–995.

Beaudoing, E., Freier, S., Wyatt, J.R., Claverie, J-M., and Gautheret, D. 2000. Patterns of variant polyadenylation signal usage in human genes. *Genome Res.* **10:** 1001–1010.

Bonaldo, M.F., Lennon, G., and Soares, M.B. 1996. Normalization and subtraction: Two approaches to facilitate gene discovery. *Genome Res.* **6:** 791–806.

DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y., Su, Y.A., and Trent, J.M. 1996. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* **14:** 457–460.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., et al. 1999. The DNA sequence of human chromosome 22. *Nature* **402:** 489–495.

Gautheret, D., Poirot, O., Lopez, F., Audic, S., and Claverie, J.-M. 1998. Alternate polyadenylation in human mRNAs: A large-scale analysis by EST clustering. *Genome Res.* **8:** 524–530.

Hattori, M., Fujiyama, A., Taylor, T.D., Watanabe, H., Yada, T., Park, H.S., Toyoda, A., Ishii, K., Totoki, Y., Choi, D.K., et al. 2000. The DNA sequence of human chromosome 21. *Nature* **405:** 311–319.

Hillier, L.D., Lennon, G., Becker, M., Bonaldo, M.F., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, et al. 1996. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* **6:** 807–828.

Hishiki, T., Kawamoto, S., Morishita, S., and Okubo, K. 2000. BodyMap: A human and mouse gene expression database. *Nucleic Acids Res.* **28:** 136–138.

Irvine, A.D., Corden, L.D., Swensson, O., Swensson, B., Moore, J.E., Frazer, D.G., Smith, F.J., Knowlton, R.G., Christophers, E., Rochels, R., et al. 1997. Mutations in cornea-specific keratin K3 or K12 genes cause Meesmann's corneal dystrophy. *Nat. Genet.* **16:** 184–187.

Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C.F., Trent, J.M., Staudt, L.M., Hudson, J., Jr., Boguski, M.S., et al. 1999. The transcriptional program in the response of human fibroblasts to serum. *Science* **283:** 83–87.

Kawamoto, S., Ohnishi, T., Kita, H., Chisaka, O., and Okubo, K. 1999. Expression profiling by iAFLP: A PCR-based method for genome-wide gene expression profiling. *Genome Res.* **9:** 1305–1312.

Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., et al. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14:** 1675–1680.

Loftus, S.K., Chen, Y., Gooden, G., Ryan, J.F., Birznieks, G., Hilliard, M., Baxevanis, A.D., Bittner, M., Meltzer, P., Trent, J., et al. 1999. Informatic selection of a neural crest-melanocyte cDNA set for microarray analysis. *Proc. Natl. Acad. Sci.* **96:** 9277–9280.

Maeda, K., Okubo, K., Shimomura, I., Mizuno, K., Matsuzawa, Y., and Matsubara, K.1997. Analysis of an expression profile of

Matsubara, K, and Okubo, K. 1993. cDNA analyses in the human genome project. *Gene* **135:** 265–274.

Nishida, K., Adachi, W., Shimizu-Matsumoto, A., Kinoshita, S., Mizuno, K., Matsubara, K., and Okubo, K. 1996. A gene expression profile of human corneal epithelium and the isolation of human keratin 12 cDNA. *Invest. Ophthalmol. Vis. Sci.* **37:** 1800–1809.

Nishida, K., Honma, Y., Dota, A., Kawasaki, S., Adachi, W., Nakamura, T., Quantock, A. J., Hosotani, H., Yamamoto, S., Okada, M.,et al. 1997. Isolation and chromosomal localization of a cornea-specific human keratin 12 gene and detection of four mutations in Meesmann corneal epithelial dystrophy. *Am. J. Hum. Genet.* **61:** 1268–1275.

Ohnishi, T. and Okubo, K. 1999. Isolation of pure human mucosal epithelium for RNA analysis. *Biotechniques* **27:** 978– 986.

Ohno, I., Hashimoto, J., Shimizu, K., Takaoka, K., Ochi, T., Matsubara, K., and Okubo, K. 1996. A cDNA cloning of human AEBP1 from primary cultured osteoblasts and its expression in a differentiating osteoblastic cell line. *Biochem. Biophys. Res. Commun.* **228:** 411–414.

Okubo, K., Hori, N., Matoba, R., Niiyama, T., and Matsubara, K. 1991. A novel system for large-scale sequencing of cDNA by PCR amplification. *DNA Seq.* **2:** 137–144.

Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y., and Matsubara, K. 1992. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nat. Genet.* **2:** 173–179.

Pearson, W.R. and Lipman, D.J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* **85:** 2444–2448.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O., and Davis, R.W. 1996. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci.* **93:** 10614–10619.

Schuler, G.D., Boguski, M.S., Stewart, E.A., Stein, L.D., Gyapay, G., Rice, K., White, R.E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., et al. 1996. A gene map of the human genome. *Science* **274:** 540–546.

Sheets, M.D., Ogg, S.C., and Wickens, M.P. 1990. Point mutations in AAUAAA and the poly (A) addition site: Effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. *Nucleic Acids Res.* **18:** 5799–5805.

Shimizu-Matsumoto, A., Adachi, W., Mizuno, K., Inazawa, J., Nishida, K., Kinoshita, S., Matsubara, K., and Okubo, K. 1997. An expression profile of genes in human retina and isolation of a complementary DNA for a novel rod photoreceptor protein. *Invest. Ophthalmol. Vis. Sci.* **38:** 2576–2585.

Soares, M.B., Bonaldo, M.F., Jelene, P., Su, L., Lawton, L., and Efstratiadis, A. 1994. Construction and characterization of a normalized cDNA library. *Proc. Natl. Acad. Sci.* **91:** 9228–9232.

Strausberg, R.L., Buetow, K.H., Emmert-Buck, M.R., and Klausner, R.D. 2000. The cancer genome anatomy project: Building an annotated gene index. *Trends Genet.* **16:** 103–106.

Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. 1995. Serial analysis of gene expression. *Science* **270:** 484–487.

Welle, S., Bhatt, K., and Thornton, C.A. 1999. Inventory of high-abundance mRNAs in skeletal muscle of normal men. *Genome Res.* **9:** 506–513.

Wickens, M. and Stephenson, P. 1984. Role of the conserved AAUAAA sequence: Four AAUAAA point mutants prevent messenger RNA 3′ end formation. *Science* **226:** 1045–1051.

Williamson, A. R. 1999. The Merck Gene Index project. *Drug Discov. Today* **4:** 115–122.

Received June 8, 2000; accepted in revised form September 18, 2000.